OAK RIDGE NATIONAL LABORATORY

PLANT SYSTEMS BIOLOGY

Spring 2010

Plant Systems Biology Group
Oak Ridge National Laboratory

January 2010

Nancy L. Engle – B.A. Rollins College, Florida

Research interests include quantifying plant and microbe responses to perturbations using mass spectrometry-based metabolomics, structural elucidation of unknown metabolites, and natural product synthesis. Structured pedigrees and natural genotypic variation are used to identify genetic and environmental effects on metabolite production, to which advanced mass-spectometry based analysis using ion trapping and Time-of-Flight analyzers are used to aid in identification of unknowns. Results are used to better understand the metabolic effects of alterations and to develop synthesis strategies for confirmation of unknown metabolites

Lee E. Gunter – M.S. University of Georgia

Research interests include the biology and evolution of *Populus*; the discovery and characterization of genes involved in plant cell wall biosynthesis through bioinformatics, association studies and mapping; effect of phytochrome expression on crown architecture; and abiotic stress response in plants. Results are applied to design and assessment of transgenic poplar and gene expression and evolution.

Sara S. Jawdy – M.S., University of Tennessee

Research interests include transcriptomics and expression analysis of plants, both wild-type and genetically modified, that have been subjected to various experimental treatments. Various tools such as microarray analysis, 454 sequencing and quantitative PCR are used to identify individual genes and gene networks involved in controlling response(s) to treatment. The results are used to understand how both herbaceous and woody plant species respond under certain conditions. This knowledge can be used to modify plants such that they are better suited for carbon sequestration or biomass production, or to elucidate previously unknown functions of specific genes and gene networks.

Udaya C. Kalluri – Ph. D. Michigan Technological University

Research interests include molecular and phenotypic characterization and modeling of biological phenomenon across single cell to whole plant levels. Current projects involve studies of plant processes such as cell wall biosynthesis, hormone signaling, developmental biology, plant-microbe interactions under the larger research contexts of bioenergy and carbon biosequestration.

Abhijit A. Karve - Ph.D. Clemson University

Research interests include molecular characterization of plant responses to heat stress and understanding the role of plant-microbe interactions in plant adaptation to environmental stresses. The environmental stress responses are characterized by using enzyme activity assay, proteomic analysis and protoplast based transient expression assays. The role of protein interactions in plant adaptation is studied by recombinant protein expression in bacteria and protein-protein interaction assays. The results help better understand the cellular processes that are involved in determining the adaptation to environmental stresses.

Madhavi Z. Martin – Ph.D. University of California, Los Angeles

Research interests are to pursue R&D in the design, fabrication, and testing of environmental sensors specifically for biomass characterization and forensic applications. Specific examples include the evaluation of a laser-induced breakdown spectroscopy (LIBS) system capable of rapid and simultaneous elemental analysis for dendrochemistry, detection of total carbon and nitrogen in soils, including heavy metal and radionuclide detection in the environment. Results are applied to the advanced development of bioenergy crops, to understand soil carbon cycle processes, and to correlate the environmental response to fire and climate. Forensic applications have led to solving murder cases and criminal lawsuits.

Priya Ranjan – Ph.D. Michigan Technological University

Research interests include bioinformatics analysis of high throughput “omics” data, focusing on next-generation sequencing technologies, genetic association studies using single nucleotide polymorphisms (SNPs), identification of candidate genes from QTL studies related to cell wall traits. Results are applied to genetic improvement of bioenergy crops.

Timothy J. Tschaplinski – Ph.D. University of Toronto

Research interests include characterizing gene function by analysis of transgenic plants and microbes with mass spectrometry-based metabolomics, plant response to environmental stress, and linking phenotypic traits to underlying molecular mechanisms. Structured pedigrees and natural genotypic variation are used to identify genetic and environmental effects on metabolite production, drought tolerance, carbon partitioning and allocation. Results are applied for the accelerated domestication of bioenergy crops, to understand the metabolic consequences of genetic manipulation, to unravel plant-microbe signaling pathways, and to improve microbial bioprocessing to overcome the recalcitrance of lignocellulosic biomass for bioenergy production.

Gerald A. Tuskan – Ph.D. Texas A&M University

Over 18 years of experience in leading and working with the Department of Energy and the larger scientific community on the development of woody perennial bioenergy feedstocks. Currently the co-lead for the Joint Genome Institute Plant Genomic effort and the Activity Lead for the BioEnergy Science Center *Populus* Biosynthesis team. In addition, he is the co-lead PI on DOE research related to poplar genetics and genomics, and plant-microbe interactions. His research focuses on the accelerated domestication of poplar through direct genetic manipulation of targeted genes and gene families, with focus on cell wall biosynthesis. Published more that 118 peer-reviewed articles since 1990 in the areas of genetic and genomics of perennial plants; including 15 publications with nearly 550 citations exclusively related to genomics and bioenergy.

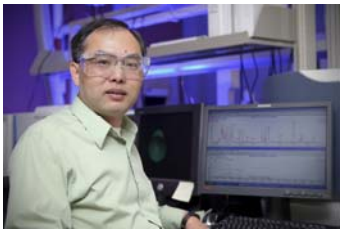
David J. Weston – Ph.D. Clemson University

The main goal of our research is to discover and validate mechanisms driving plant adaptive traits. We use a comparative systems biology approach, whereby genes, metabolites and proteins are modeled using network theory across multiple species and linked to traits of interest. Both species - conserved and - diverse aspects of the network are confirmed at molecular levels (usually reverse genetics), and scaled to higher levels of biological organization through collaborations. Current projects include linking physiological traits (especially photosynthetic properties) to the heat shock response and plant - microbe interactions.

Stan D. Wullschleger – Ph.D. University of Arkansas

Research interests include quantifying plant response to environmental stress and linking phenotypic traits to underlying molecular mechanisms. Structured pedigrees and natural genotypic variation are used to identify genetic and environmental controls on carbon partitioning and biomass distribution above- and below-ground. Results are applied to the advanced development of bioenergy crops, to understand soil carbon cycle processes, and to characterize interactions between physiology and genetics in determining plant adaptation to climate.

Xiaohan Yang – Ph.D. Cornell University



Major research interests include discovery of genes involved in plant cell wall biosynthesis, algal oil production, stress response, and intercellular signal transduction, with a focus on small protein genes. Sequencing of transcriptome and proteome is used to profile gene expression. The sequencing data are analyzed using bioinformatics to identify candidate genes, which are then characterized at biochemical, physiological, and morphological levels using molecular genetics. Radio bio-imaging technology is pursued to monitor the movement of small signaling proteins. Other interests include new genome sequencing, evolutionary dynamics, gene annotation, and phylogenetic analysis. Knowledge gained is relevant to bioethanol and biodiesel production and carbon sequestration.

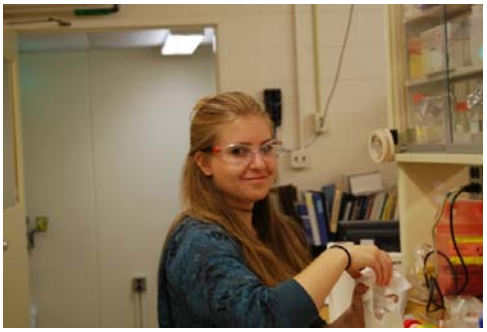
Tara Hall – Administrative Assistant

Sara Allen – Intern

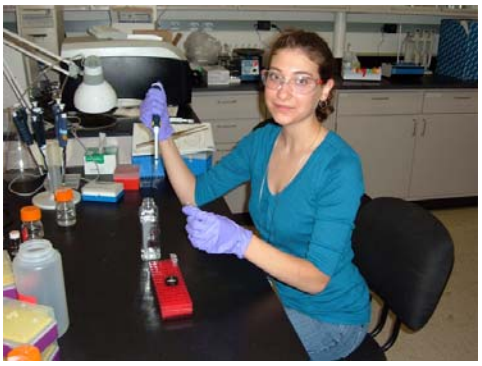
Zackary Moore – Student Intern
University of Tennessee

Maya Strahl – Student Intern

Alina Campbell – M.S. Graduate Student
University of Tennessee



Whitney McNutt – HERE Intern
University of Tennessee



Alyssa Deleon – Student Intern

Rhonda Egidy – Student Intern

Plant Systems Biology Group 2009 Citations

Chatterjee A., R. Lal, L. Wielopolski, M. Z. Martin, and M. H. Ebinger. 2009. Evaluation of different soil carbon determination methods. Plant Sciences.

Cheng, Z., and G. A. Tuskan. 2009. *Populus* community mega-genomics: coming of age. Critical Reviews in Plant Science 28: 1-3.

Islam-Faridi, M. N., C. D. Nelson, S. P. DiFazio, L. E. Gunter, and G. A. Tuskan. 2009. Cytogenetic analysis of Populus Trichocarpa-Ribosomal DNA, telomere repeat sequence, and marker-selected BACs. Cytogenetic and Genome Research 125: 74-80.

Jung, H. W., T. J. Tschaplinski, L. Wang, J. Glazebrook, J. T. Greenberg. 2009. Priming in systemic plant immunity. Science 324: 89-91.

Kalluri U. C., G. B. Hurst, P. K. Lankford, P. Ranjan, and D. A. Pelletier. 2009. Shotgun proteome profile of *Populus* developing xylem. Proteomics 9: 4871-4880.

Kang, B., L. Osburn, D. Kopsell, G. A. Tuskan, and Z. Cheng. 2009. Micropropagation of *Populus trichocarpa* 'Nisqually-1': the genotype deriving the *Populus* reference genome. Plant Cell Tissue and Organ Culture 99: 251-257.

Karve, A., and B. D. Moore. 2009. Function of *Arabidopsis* hexokinase-like1 as a negative regulator of plant growth. Journal of Experimental Botany 10: 1-13.

Leakey, A. D. B., E. A. Ainsworth, S. M. Bernard, R. J. Cody Markelz, D. R. Ort, S. A. Placella, A. Rogers, M. D. Smith, E. A. Sudderth, D. J. Weston, S. D. Wullschleger, and S. A. Yuan. 2009. Gene expression profiling: opening to the black box of plant ecosystem responses to global change. Global Change Biology 15: 1201-1213.

Ranjan, P., T. Yin, X. Zhang, U. C. Kalluri, X. Yang, S. Jawdy, and G. A. Tuskan. 2009. Bioinformatics-based identification of candidate genes from QTLs associated with cell wall traits in *Populus*. BioEnergy Research doi:10.1007/s12155-009-9060-z.

Wullschleger, S. D., D. J. Weston, and J. M. Davis. 2009. *Populus* Responses to edaphic and climatic cues: emerging evidence from systems biology research. Critical Reviews in Plant Science 28: 368-374.

Yang, S., T. J. Tschaplinski, N. L. Engle, S. L. Carroll, S. L. Martin, B. H. Davison, A. V. Palumbo, M. Rodriquez Jr., and S. D. Brown. 2009. Transcriptomic and metabolomic profiling of Zymomonas mobilis during aerobic and anaerobic fermentations. BMC Genomics.

Yang, S., I. Kataeva, S. D. Hamilton-Brehm, N. L. Engle, T. J. Tschaplinski, C. Doeppke, M. Davis, J. Westpheling, and M. W. W. Adams. 2009. Efficient degradation of Lignocellulosic plant biomass, without pretreatment by the Thermophilic Anaerobe "*Anaerocellum thermophilum*" DSM. Applied and Environmental Microbiology 75: 4762-4769.

Yang, X., S. Jawdy, T. J. Tschaplinski, and G. A. Tuskan. 2009. Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. Genomics 93: 473-480.

Yang, X., U. C. Kalluri, S. P. DiFazio, S. D. Wullschleger, T. J. Tschaplinski, Z. Cheng, and G. A. Tuskan. 2009. Poplar Genomics: State of Science. Critical Reviews in Plant Science 28: 285-308.

Yin, T. M., X. Y. Zhang, L. E. Gunter, S. X. Li, S. D. Wullschleger, M. R. Huang, and G. A. Tuskan. 2009. Miscorsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. New Phytologist 181: 498-503.

Zhao, N., J. Guan, F. Forouhar, T. J. Tschaplinski, Z. Cheng, L. Tong, and F. Chen. 2009. Two poplar methyl salicylate esterases display comparable biochemical properties but divergent expression patterns. Phytochemistry 70: 32-39.

Critical Reviews in Plant Science, 28:164–178, 2009

ISSN: 0735-2689 print / 1549-7836 online
DOI: 10.1080/07352680902776556

Taylor & Francis
Taylor & Francis Group

Evaluation of Different Soil Carbon Determination Methods

A. Chatterjee,[1] R. Lal,[1] L. Wielopolski,[2] M. Z. Martin,[3] and M. H. Ebinger[4]

[1]*Carbon Management and Sequestration Center, Ohio State University, Columbus, OH 43210*
[2]*Brookhaven National Laboratory, Environmental Sciences Department, Upton, NY 11973*
[3]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831*
[4]*Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM 87545*

Referee: Dr. John M. Kimble, USDA-NRCS (Ret.), 151 East Hill Church Road, Addison, NY 14801

Table of Contents

Determining soil carbon (C) with high precision is an essential requisite for the success of the terrestrial C sequestration program. The informed choice of management practices for different terrestrial ecosystems rests upon accurately measuring the potential for C sequestration. Numerous methods are available for assessing soil C. Chemical analysis of field-collected samples using a dry combustion method is regarded as the standard method. However, conventional sampling of soil and their subsequent chemical analysis is expensive and time consuming. Furthermore, these methods are not sufficiently sensitive to identify small changes over time in response to alterations in management practices or changes in land use. Presently, several different in situ analytic methods are being developed purportedly offering increased accuracy, precision and cost-effectiveness over traditional ex situ methods. We consider that, at this stage, a comparative discussion of different soil C determination methods will improve the understanding needed to develop a standard protocol.

Keywords soil carbon, wet oxidation, dry combustion, inelastic neutron scattering, remote sensing, laser induced breakdown spectroscopy

Address correspondence to A. Chatterjee, Carbon Management and Sequestration Center, Ohio State University, Columbus, Ohio 43210. E-mail: forestcarbon@gmail.com

I. INTRODUCTION

Rapid, accurate, and inexpensive measurement of the soil's carbon pool is essential to detect and quantify change in the ecosystem dynamics of C. A comparative assessment of present determination methods is needed urgently to identify promising techniques that reduce uncertainty in measuring the soil's C pool and flux at the farm and watershed scale. Evaluation of sustainable land-use and soil management practices to stabilize

or increase the soil's C pool also demands a sensitive analytical protocol that will pave the way to establish soil C as a tradable commodity in the global market. Accordingly, there is an undeniable need to identify method(s) to determine the rate of change in the soil's C pool over a specific period. Quantifying the site-specific ancillary benefits of soil C sequestration also necessitates establishing standard protocol for evaluating soil's organic C (SOC) pool and flux over multiple scales.

Credible estimates of soil C pool and its fluxes also are required to identify policies and site-specific management practices to increase or at least stabilize the SOC pool. Depending on such practices and land uses, the C pool can play a dominant role as either a net source or net sinks of atmospheric carbon dioxide (CO_2). To a depth of 300 cm the SOC pool comprises 2344 Pg C (1 Pg $= 10^{15}$ g), whereas the inorganic soil C (SIC) content ranges between 695–748 Pg C for the soil profile depth of 100 cm (Jobbagy and Jackson, 2000; Batjes, 1996). The former is mainly composed of (1) soluble organic compounds (sugars and proteins), (2) amorphous organic compounds (humic acid, fat, waxes, lignins and polyuronides), and, (3) organomineral complexes (Schnitzer, 1991). The latter comprises primary and secondary carbonates (Eswaran *et al.*, 1995). Most studies have focused on measuring the SOC pool and relating it to land use and soil management as this is the pool that can be easily affected by changes in agriculture practices and land use management. The SOC content can be determined directly or indirectly from the difference between the total soil C (TSC) and the SIC concentration, measured separately. The SIC can be determined quantitatively by treating a soil sample with HCl and measuring the CO_2 released from carbonates either by gas chromatography or by pressure calcimetry (Sherrod *et al.*, 2002). For soils lacking SIC, TSC value represents the SOC value. However, when the parent material is enriched in carbonaceous mineral, such as limestone and dolomite, the SIC must be measured to determine the sample's SOC.

Besides soil carbonates, soils have organic compounds such as coal and charcoal (Black C) that interfere the determination of SOC. Coal is a major concern in determining the actual potential of reclamation measures in mineland, whereas charcoal can be present in soils affected by fire. Coal-derived C can be quantitatively measured by radiocarbon (^{14}C) activity; however, this method is highly expensive and limited availability of the facilities needed for the analysis (Rumpel *et al.*, 2003). Alternatively, diffuse reflectance infrared Fourier transform (DRIFT) spectroscopy in combination with multivariate statistical analysis can be used to separate coal-derived C and this method produced a good fit with ^{14}C measurement (Rumpel *et al.*, 2001). Recently, Ussiri and Lal (2008) developed a chemi-thermal method to determine coal-derived C; this method is cost-effective and has a high recovery percentage comparable to ^{14}C measurement. There are several methods based on thermal and/or chemical oxidation to quantify charcoal C but the recovery percentages varied widely because charcoal C is a mixture of wide continuum ranging from large pieces of slightly charred biomass (1–100 μm) to submicron soot particles (30–40 nm) (Hedges *et al.*, 2000; Masiello, 2004; Hammes *et al.*, 2007).

Another problem associated with the determination of SOC content is the representation of the data. It is simple to report SOC content as mass of C per unit mass of soil (g kg^{-1}); however, for the calculation of soil C pool, concentration of C is necessarily expressed on an area basis (Mg ha^{-1}) or volume basis (Mg m^{-3}). Calculation of SOC content on an area basis requires data on soil bulk density values, depth increments for soil sampling, and rock and root fragments; and significant uncertainties are associated with the calculation of these parameters (Post *et al.*, 2001). Particularly, uncertainties associated with soil bulk density estimation arises from determining the total soil volume for a range of soils including soils with high gravel content, high organic matter content and high swell-shrink soils (Lal, 2006). Moreover, the volume of rock and coarse fragments (>2 mm) must also be estimated and subtracted from the total soil volume prior to determining the soil bulk density value. These problems have significant influence on the calculation of soil bulk density value and subsequently the SOC content on an area basis.

The common principle underlying SOC evaluation is the ex situ chemical- or high temperature-destruction of the soil organic matter (SOM) from field samples in a laboratory. However, several non-destructive, in situ methods currently being developed promise to increase the accuracy and reduce the time and cost of conventional field soil sampling and laboratory analyses. The objective of our review is to consider fully the information on ex situ and in situ methods of determining the SOC pool, and offer a critical comparative analysis of sensitivity, predictability, and time and cost-efficiency of these novel approaches.

II. EX SITU METHODS

Ex situ methods involve collecting representative soil samples and measuring the C concentration via dry or wet combustion techniques. The latter process involves the oxidation of organic matter by an acid mixture and measuring the evolved CO_2 by gravimetric, titrimetric, or manometric methods. In the 19th century, Rogers and Rogers (1848) reported that dichromate-sulfuric acid solution could oxidize organic substances. After unsuccessful attempts by Warrington and Peake (1880), and Cameron and Breazeale (1904), Ames and Gaither (1914) accomplished the higher recovery of organic substances by the dichromate-sulfuric mixture. Schollenberger (1927) introduced the titrimetric determination of unused chromic acid in the oxidation reaction with ferrous ammonium sulfate using several indicators (diphenylamine, o-phenanthroline, or N-phenylanthranillic acid (Tabatabai, 1996). Walkley and Black (1934) and Tyurin (1935) developed a complete quantification method of SOC by wet oxidation without necessitating external heating. However, Tinsley (1950) and Meibus (1960) proposed applying external heat for an extended period of time to increase the recovery of SOC.

TABLE 1
Features of ex situ soil C determination methods

Method	Principle	CO_2 determination	Advantages/Disadvantages
I. Wet combustion			
Combustion train	Sample is heated with $K_2Cr_2O_7$-H_2SO_4-H_3PO_4 mixture in a CO_2-free air stream to convert OC in CO_2.	Gravimetric/ Titirimetric	Gravimetric determination requires careful analytical techniques and titrimetric determination is less precise.
Van-Slyke-Neil apparatus	Sample is heated with $K_2Cr_2O_7$-H_2SO_4-H_3PO_4 mixture in a combustion tube attached to the apparatus to convert OC in CO_2.	Manometric	Expensive and easily damaged apparatus.
Walkley-Black	Sample is heated with $K_2Cr_2O_7$-H_2SO_4-H_3PO_4 mixture. Excess dichromate is back titrated with ferrous ammonium sulfate.	Titrimetric	Oxidation factor is needed. Variable SOC recovery. Generate hazardous byproducts such as Cr.
II. Dry combustion			
Weight-loss-on-ignition	Sample is heated to 430°C in a muffle furnace during 24 hours.	Gravimetric	Weight losses are due to moisture and volatile organic compounds. Overestimate the organic matter content.
Automated	Sample is mixed with catalysts or accelerator and heated in resistance or induction furnace in O_2 stream to convert all C in CO_2.	Thermal conductivity, gravimetric, IR absorption spectrometry	Rapid, simple, and precise but expensive. Slow release of contaminant $CO_2$from alkaline earth carbonates with resistance furnace.

Rather (1917) introduced the technique of estimating SOM from the weight loss of soils on ignition (LOI). He also suggested first destroying the hydrosilicates by treating the samples with hydrochloric and hydrofluoric acids to eliminate the loss of hydroxyl groups during heating, but invariably some SOM is prone to decompose during this treatment. Mitchell (1932) described a low temperature ignition method to remove the soil water by heating the sample at 110°C and exposing the dried soil at 350–400°C temperature for 8 hours in a furnace. Jackson (1958) recommended using an induction furnace wherein heat is generated from high-frequency electromagnetic radiation. Temperature and duration of heating have substantial effect on the loss of SOM (Schulte *et al.*, 1991). Moreover, the relation between LOI-SOC varies widely with soil depths and types (Konen *et al.*, 2002). These factors need to be considered before determining the SOC by the LOI method.

Tabatabai and Bremner (1970) introduced an automated CO_2 analyzer based on thermal conductivity measurement of the effluent gases. Current automated total C analyzers follow the principles described by Tabatabai and Bremner (1991). Currently this method is considered as the standard method to determine soil C concentration and widely accepted. In the following sections, principles, advantages, and disadvantages of different ex-situ methods are discussed and summarized in Table 1.

A. Wet Combustion

The analysis of SOC content by wet combustion long has been regarded as standard procedure since Schollenberger (1927) introduced it; most of the time it produces results in agreement with those of the dry combustion technique (Nelson and Sommers, 1996). Wet combustion involves oxidizing SOM to CO_2 with a solution containing potassium dichromate ($K_2Cr_2O_7$), sulfuric acid (H_2SO_4) and phosphoric acid (H_3PO_4), following the reaction

$$2Cr_2O_7^{-2} + 3C^0 + 16H^+ = 4Cr^{3-} + 3CO_2 + 8H_2O \qquad [1]$$

This reaction generates a temperature of 210°C and is sufficient to oxidize carbonaceous matter. The excess $Cr_2O_7^{-2}$ (not used in oxidation) is titrated with Fe $(NH_4)_2(SO_4)_2{\cdot}6H_2O$, and reduced $Cr_2O_7^{-2}$ is assumed to be equivalent to the sample's SOC content. Calculations for SOC content are based on the fact that C present in soil has an average valence of zero.

TABLE 2
Modifications in wet digestion methods for determining SOC (adapted from Nelson and Sommers, 1996)

Method	Reagent concentrations (N)		Ratio of H_2O to acid (v:v)	Digestion Conditions	C.V.%
	$K_2Cr_2O_7$	H_2SO_4			
Schollenberger (1927)	0.058	18		Tube heated by flame at 175°C for 90 s	1.4–1.9
Tyurin (1931)	0.066	9	1.00	Flask with funnel boiled at 140°C for 5 min	8.5
Walkley-Black (1934)	0.055	12	0.50	Flask with no external heat, max temp is 120°C	1.6–4.2
Tinsley (1950)	0.027	7.2	0.67	Flask with condenser refluxed for 2 h at 150°C	0.8–3.1
Mebius (1960)	0.045	10	0.42	Flask with condenser refluxed for 30 min at 159°C	1.2–1.8
Modified Mebius (1982)	0.033	10.8	0.67	Flask with condenser refluxed at 150°C for 30 min	1.0–3.6
Heans (1984)	0.055	12	0.50	Tube heated in block at 135°C for 30 min	4.1

The wet combustion method has undergone a number of modifications related to the type and concentration of the acids used and whether external heat is applied or not (Table 2). Schollenberger (1927) suggested heating the soil- H_2SO_4-$K_2Cr_2O_7$ mixture to complete SOM oxidation and thereby increase the recovery. Others soon realized that the temperature and its duration were critical and must be standardized to ensure the oxidation of a constant proportion of SOM; for that, a consistent amount of dichromate must be thermally decomposed during digestion. Tyurin (1931) incorporated a definite heating time and temperature for soil-chromic acid mixtures in a test tube. However, Walkley and Black (1934) reported satisfactory results with no heating, and suggested using a factor of 1.32 (assuming 76% recovery) to account for the incomplete digestion; however, this percent is all over depending on soil type and soil depth and mineralogy. Table 2 lists the correction factors for various soils. Meibus (1960) proposed boiling the soil-dichromate-sulfuric acid mixture for 30 min in an Erlenmeyer flask connected to a reflux condenser. Subsequently, many researchers tried to modify earlier procedures to enhance the recovery, such as proposed by Meibus (Nelson and Sommers, 1982) and Heans (1984). For dry combustion, Soon and Abboud (1991) reported that the Walkley-Black (WB), modified Tinsley, and modified Meibus methods respectively recovered 71, 95 and 98% of soil C. Thus, we conclude that external heat can improve the SOC recovery, although the WB method is far more popular than the modified methods with external heat. Assuming a recovery of 76% often leads to overestimating or underestimating SOC concentration, depending on the soil's type. The recovery percentage varies from 59% to 88%, and the corresponding correction factor from 1.69 to 1.14 (Table 3). Díaz-Zorita (1999) attributed the low recovery of SOC by the WB method in soil from a graminean pasture system to presence of a high percentage of recalcitrant SOM (e.g., phenolic and lignin compounds). The modified WB method sometimes overestimates the SOC content, while the WB method underestimates it (Brye and Slaton, 2003). Variable recovery percentage of the WB method depends on the soil's type rather than on landuses. Mikhalilova *et al.* (2003) compared four management regimes (native grassland, grazed, continuous cropping, and continuously plowed fallow) and derived a single correction factor of 1.63 independent of the management regime. De Vos *et al.* (2007) reported a strong correlation between recovery percentage (using the WB method) and the soil's textural class and pedogenetic horizons. Recovery was higher by 3 to 8% from sandy soils than from loam and silt-loam soils. Similarly, recovery from samples from eluvial horizons was significantly higher than those from A horizons, presumably due to higher SOM content in the upper than lower soil horizons.

Interferences by chloride (Cl^-), ferrous iron (Fe^{2+}), higher oxides of manganese (Mn^{3+} and Mn^{4+}) and coal particles also entail incorrect estimations of SOC content (Nelson and Sommers, 1996). Particularly, these ions participate in chromic acid-oxidation-reduction reaction, wherein Fe^{2+} and Cl^-lead to a positive error, and MnO_2 to a negative error. Large concentrations of Fe^{2+} occur in highly reduced soil and are oxidized to Fe^{3+} by $Cr_2O_7^{-2}$, giving high values for SOC content (Eq. 2). This error is more prevalent when the soil sample is not dried before analysis.

$$Cr_2O_7^{-2} + 6Fe^{2+} + 14H^+ = 2Cr^{3+} + 6Fe^{3+} + 7H_2O \quad [2]$$

In case of salt affected soils, Cl^- ion reacts with dichromate producing chromyl chloride that consumes of $Cr_2O_7^{-2}$ (Eq. 3).

$$K_2Cr_2O_7 + 4KCl + 6H_2O = CrO_2Cl_2 + 6KHSO_4 + 3H_2O \quad [3]$$

The interference of Cl^- ions can be eliminated by washing the soil with Cl^- free water, precipitating $Cl^-$by adding Ag_2SO_4 or by stoichiometric correction (Eq. 4). Heans (1984) concluded that adding Ag_2SO_4 either before or after $K_2Cr_2O_7$ failed to control Cl^- interference, and suggested separate assay and stoichiometric correction as the only permissible alternative for assessing SOC by the WB method.

$$\text{Corrected} - \text{SOC}(\%) = (\text{Estimated SOC}\%) - (\%Cl^-/12) \quad [4]$$

The higher oxides of Mn (mainly MnO_2) often produce a negative error by competing with oxidizable substances in

TABLE 3
Correction factors for soil organic C not recovered by the Walkley-Black (Walkley and Black, 1934) method

Sampling location	Recovery%	Average Correction Correction factor	Reference
Forest soils, Belgium	63	1.58	De Vos *et al.* (2007)
Calcareous soil, Italy	77	1.30	Santi *et al.* (2006)
Sierra Leone	83	1.20	Kamara *et al.* (2007)
Native Prairie, Arkansas,USA	66	1.51	Bryre and Slaton (2003)
Agricultural soil, Arkansas,USA	63	1.59	
Russian Chernozem	61	1.63	Mikhailova *et al.* (2003)
Graminean pasture, Argentina	59	1.69	Díaz-Zorita, (1999)
Mineral soils, New Zealand	80	1.25	Grewal *et al.* (1991)
Canadian Prairie	71	1.40	Soon and Abboud (1991)
Australia	88	1.14	Lowther *et al.* (1990)

soil-chromic acid mixtures (Eq. 5). This interference produces an error of small magnitude in calculations because even in highly manganiferous soils, a minute quantity of MnO_2 competes with $Cr_2O_7^{-2}$ for the oxidation of SOC (Nelson and Sommers, 1996) (Eq. 5).

$$2MnO_2 + C^0 + 8H^+ = CO_2 + Mn^{2+} + 4H_2O \qquad [5]$$

The presence of carbonized materials (e.g., charcoal, coal, coke and soot) also is responsible for poor recovery in the wet digestion process. Without applying any external heat, the percentage recovery of SOC present in carbonized materials is low; and with external heat, the recovery is variable depending on the properties of the carbonized materials (Heans, 1984; Skjemstad and Taylor, 1999; and De Vos *et al.*, 2007). Thus, Walkley (1947) reported that the WB method recovered only 2–11% of SOC present in carbonized materials. Microscopic inspection of the digested material revealed charred materials in the -remaining organic fragments. Wet digestion methodologies cannot be employed to recover carbonized materials or to separate the SOC fractions from carbonized materials because their oxidation depends on the time and temperature of heating the chromic acid mixture and the carbonized material's properties, such as bonding with organomineral complexes, and the groups present, etc. Also, there is an environmental problem associated with using and disposing of the compounds containing chromium. Many laboratories avoid the use of the chromium-based compounds.

Using colorimetric analyses rather than titrations can increase the precision of the wet combustion method (Soon and Abboud, 1991). There are two approaches to colorimetric determination: (i) determining the unreacted dichromate solution that changes color from orange to green, and (ii) measuring the absorbance of the color complex (violet) produced from the reaction of Cr^{3+} with s-diphenylcarbazide at 450 nm (Tabatabai, 1996). The Cr^{3+}ion has two broad maxima in the visible range, one near 450 nm and the other near 600 nm. The dichromate ion also has an absorption maximum near 450 nm, but not near 600 nm, and hence, it is advisable to determine the absorbance at 600 nm. Soon and Abboud (1991) measured the absorbance of clear supernatant 10-ml aliquot of soil-chromic acid mixture at 600 nm against a set of standard sucrose solutions and achieved 100% recovery by comparison with dry combustion as the reference method. Using automatic titration or digital burettes, coupled with the wet digestion process, also may improve accuracy. Nevertheless, even though the wet digestion method has limitations due to variable recovery percentage, still it is used worldwide throughout the world to measure SOC concentration because of its low cost and minimum requirements.

B. Dry Combustion

Incinerating SOM and thermal decomposing carbonate minerals generate CO_2 that is measured by (1) dry combustion followed by measuring the changes or mass loss-on-ignition (LOI), and, (2) dry oxidation of SOC, then collecting and determining the evolved CO_2 with automated instruments (Table 1). Both methods involve oxidizing the SOC at a high temperature. The LOI method entails heating the sample in a muffle furnace between 200–500°C, whereas dry oxidation via automated analyzer is accomplished between 950–1150°C.

1. Loss on Ignition (LOI)

In this method, the SOM is assessed by measuring the weight loss from a dry soil sample (oven-dried at 105°C) after high-temperature ignition of the carbonaceous compounds in a muffle furnace. Three assumptions underlie this method: (a) LOI is due only to the combustion of SOM, and, (ii) the C content of SOM is constant (Christensen and Malmros, 1982). The concentration of SOC can be computed from the LOI-SOC relationship, where SOC is determined by an autoanalyzer or by the multiplication factor of 0.58, assuming that SOM comprises 58% of the SOC. However, this conversion factor (0.58) varies with soil's type, the sampling depth, and types of organic compounds in

TABLE 4
Relationship between soil organic C (determined by wet oxidation and dry combustion) and weight loss-on-ignition (LOI)

Soil type	Temperature (°C)	Duration (h)	SOC = m*LOI + c			Reference
			m	c	(r^2)	
Forest soils, USA	300	2	0.4315	0.1603	0.69	Abella and Zimmer (2007)
Sierra Leone	375	2	−6.55	0.64	0.93	Kamara *et al.* (2007)
	550	3	0.5783	−1.2875	0.96	De Vos *et al.* (2005)
Nebraska sand hills, USA	360	2	1.414	−0.6791	0.94	Konen *et al.* (2002)
Central loess plains, USA	360	2	0.6717	−4.5359	0.94	Konen *et al.* (2002)
Southern Wisconsin & Minnesota till prairies, USA	360	2	0.5743	0.1025	0.98	Konen *et al.* (2002)
Central Iowa and Minnesota till prairies, USA	360	2	0.6824	−2.8696	0.97	Konen *et al.* (2002)
Illinois and Iowa deep loess and drift, USA	360	2	0.6094	0.1949	0.98	Konen *et al.* (2002)
Tasmanian acidic soils						
Non-basalt derived	375	17	0.726	−1.598	0.96	Wang *et al.* (1996)
Basalt derived	375	17	0.469	−0.941	0.95	Wang *et al.* (1996)
Canadian Prairie	375	16	−9.36	0.633	097	Soon and Abboud (1991)
	450	6	0.568	0	0.98	Donkin (1991)
	450	16	0.914	0	0.99	Lowther *et al.* (1990)
Various soils of U.K.	550	3	0.840	−1.68	0.98	Howard & Howard (1990)
	400	8	0.972	−0.37	0.97	Ben-Dor & Banin (1989)
	450	12	1.04	−0.03	0.92	David (1988)

the SOC. The LOI does not generally represent SOM because LOI can decompose inorganic constituents without igniting the entire SOM pool. Temperature and the duration of ignition are critical to prevent the loss of CO_2 from carbonates and the structural water from clay minerals and amorphous materials (volcanic soils), the oxidation of Fe^{2+}, and the decomposition of hydrated salts (Schulte and Hopkins, 1996 and Ben-Dor and Banin, 1989).

While some hygroscopic water is removed from the soil during ignition at 105°C, sometimes the dehydration of the sample is incomplete; thus the SOM value may be overestimated. Also, different salts present in soil release molecular water at different temperatures above 105°C. For example, gypsum ($CaSO_4 \cdot 2H_2O$) contains up to 21% water and loses 1.5 H_2O molecules at 128°C, and the remaining H_2O at 163°C. Epsom salts ($MgSO_4 \cdot 7H_2O$) loses six H_2O molecules at 150°C and the remaining one at 200°C. Four H_2O molecules are lost from $CaCl_2 \cdot 6H_2O$ at 30°C, and the remaining two molecules at 200°C (Schulte and Hopkins, 1996 and Lide, 1993). Dehydroxylation of silicates starts between 350 to 370°C, whereas Na-montmorillonite, vermiculite, gibbsite, goethite, and brucite lose crystal-lattice water between 150 to 250°C (Barshad, 1965). Schulte and Hopkins (1996) reported that gypsum is dehydrated fully at 150°C; they recommended using this temperature for soils dominated by hydrated clays. Volcanic soils have large amounts of water that can affect the results.

It is difficult to predict an optimum temperature and duration of ignition to ensure the maximum SOM recovery and avoid loss by the dehydration of clays or decomposition of other soil constituents. Abella and Zimmer (2007) reported that at 300°C, 85–89% of LOI occurred during the first 30 min and 98-99% during the first 90 min. In contrast, Ben-Dor and Banin (1989) suggested that 400°C for 8 hr is suitable, basing their suggestion on using fast and prolonged heating working with natural and synthetic soils of Israel. Donkin (1991) observed that optimum temperature for ignition is 450°C for 6 hr, and concluded that higher temperature does not provide any advantage in terms of recovery percentage. Schulte *et al.* (1991) recommended that after reaching 360°C, exposure for two hours is suitable for routine soil sample analyses.

Despite these limitations, the LOI and SOC content of soil are correlated strongly, as calculated from organic C data (Table 4). However, the slopes (m) and intercepts (c) are highly variable depending on the ignition temperature and duration, soil types, and the compounds that comprise the SOC. The data in Table 4 show that values of the slope of the regression equation <1 indicate a loss of soil constituents other than SOM during ignition, whereas those >1 represent the incomplete recovery of SOM. Konen *et al.* (2002) concluded that predictive equations developed by LOI-SOC relationship were significantly different for different Major Land Resource Areas (MLRAs) in the central United States. For all their soil samples, ignition was

accomplished at 360°C (Table 4). Different recovery of SOM during ignition was controlled by the heating time and duration and by soil texture (particularly clay %) and type. De Vos *et al.* (2005) observed that the intercept in LOI-SOC regression equation is determined significantly by the samples' clay content. Spain *et al.* (1982) reported that a 9% improvement in the predictive capability of the equation using a bivariate function of LOI and clay.

Sample size is another source of variation in LOI measurements. Schulte *et al.* (1991) reported that LOI value significantly decreased with increase in sample weight. Diffusion of oxygen within the sample inhibits oxidation in large samples, a feature that is critical for organic soils, e.g., peat and muck soil.

While the LOI is a simple, rapid, and inexpensive technique of determining SOC content, the LOI-SOC regression equation must be determined for particular soil type and depth. Inclusion of the clay percentage in the bivariate regression equation can increase the correlation between LOI and SOC content. Finally, consistency should be assured for ignition temperatures, exposure times, and the samples' size and information on these three parameters included at the time of publishing the research data (Konen *et al.*, 2002; Heiri *et al.*, 2001).

2. Automated Carbon Analyzer

Use of the automated analyzer for determining of total C has evolved to become the standard method. Following are the major steps in detecting C in an automated CN analyzer: (i) Automatic introduction of the sample into a high-temperature oxidation zone wherein soil C is converted to CO_2; (ii) carriage of CO_2 by a carrier gas (generally helium) and separated from other gasses (N_2, NO_X, H_2O vapor, SO_2) either by a gas chromatographic system, or a series of selective traps for the individual gases; and, (iii) detection of the concentration of CO_2 mainly by thermal conductivity, mass spectrometry or infrared gas analyzing methods (Smith and Tabatabai, 2004). Method of CO_2 detection varies with instruments' manufacturers and models. Table 5 is a short list of automated analyzers, the detection principles of detection, manufacturers and contact information. An automated analyzer is calibrated with glutamic acid and generally samples are replicated to ensure the quality of the run. The main advantages of automated analyzer are the following: (i) rapid and precise, (ii) no loss of soil C during combustion, (iii) potential for simultaneously measuring nitrogen and sulfur (depending on model), and, (iv) can be connected to mass spectrometer for stable isotope analysis.

Special care must be taken in homogenizing the soils and ensuring its fineness. In most cases, 100–200 mg of soil sample is used for auto analyzer analyses. Pérez *et al.* (2001) suggested that simple crushing was not sufficient to guarantee homogeneity of small soil samples, and precision generally is better for finely ground samples (<177 μm). They also concluded that 100 mg of soil sample is adequate to obtain the best results from an auto analyzer. In contrast, Jimnez and Ladha (1993) recommended soil samples of 60 mg with fineness of 150 μm; this can be achieved by roller grinding or ball milling the sample after passing it through a 2-mm sieve.

The sample size must be large enough to create detectable signals and generate representative data within the limits of its combustibility. However, sample size can not indefinitely be increased because of incomplete oxidation under a low O_2 supply and the physical limitations of the sample's container. Most automated analyzers (like Elementar Vario Macro) offer options to increase O_2 dosing and combustion time. In particular, samples of organic soil in particular must be analyzed under a sufficient O_2 supply or with a sample of lesser weight, although the latter may contribute to uncertainty in the sample's representativeness. Extremely small soil samples with low SOC content also generated very low detector signal-to-noise ratio, and hence, poor accuracy and precision (Jimnez and Ladha, 1993).

Complete combustion of the sample also depends on the temperature within the combustion furnace, generally held between 950 and 1200°C. For some models, soil samples are encapsulated in a tin foil that raises the combustion temperature to about 1800°C. Wright and Bailey (2001) compared two combustion temperature profiles, 1040 and 1300°C, concluding that 1300°C is essential for accurately measuring total soil C. They observed that under lower combustion conditions (1040°C), carbonate decomposition from samples of pure $CaCO_3$ is minimized to 5%, whereas it is maximized to 98% at higher temperatures (1300°C).

Dry combustion with auto analyzers have higher precision than wet combustion or LOI, but also costs more due to the expense of buying the analyzer (US $40,000 to over $50,000) and the associated components such as an ultra-microbalance, computer, and printer. The operating costs of auto analyzers also are slightly higher due to the required consumables and high purity gasses (He and O_2). The instrument consumes a significant amount of electricity in heating the furnace. Jimnez and Ladha (1993) estimated that the cost per sample for analyzing TSC using the Perkin-Elmer 2400 CHN analyzer ranges between $3.8 and $6.50 for running 100 samples and 10 samples, respectively, in a single operation. Analyzing few samples increases the cost of analysis because of the extended time required for their stabilization and calibration and the increase in the quantity of standard runs for each operation. Running a large batch of samples can reduce cost of analysis by economies of scale.

Comparing different ex situ methods to determine soil C reveals that high precision and low analysis cost cannot be achieved using the same method. Thus, automated dry combustion analysis provides high precision, whereas the LOI method involves low cost. The expense of assessing soil C can be lowered provided the relationship between LOI and automated dry combustion is established for a particular soil type. However, it is rare to find a strong linear relationship between the two (Abella and Zimmer, 2007). Further, wet digestion, the Walkley and Black method (REF), carries a wide variation in recovery percentages, and also does not demonstrate a strong correlation

TABLE 5
Current automated dry combustion CN analyzers, description, and operating principles

Manufacturer	Address/website	Model (s)	Operating principle/detection system
Costech Analytical Technologies	26074 Avenue Hall, Suite 14, Valencia, California-91355, USA www.costechanalytical.com	ECS 4010 CHNSO	The sample within tin capsule reacts with oxygen and combust at temperatures of 1700–1800°C. Combustion of sample generates mixture of N_2, CO_2, H_2O and SO_2. The gases are separated by gas chromatographic (GC) separation column and are detected sequentially by the TCD (thermal conductivity detector). The TCD generates a signal, which is proportional to the amount of element in the sample.
LECO Corp.	3000 Lakeview Avenue, St. Joseph, Michigan 49085–2396, USA www.leco.com	TruSpec series	Sample encapsulated in tin foil is combusted at 950°C and detection by infrared.
PerkinElmer Life and Analytical Sciences	710 Bridgeport Avenue, Shelton, Connecticut-06484–4794, USA www.perkinelmer.com	2400 Series II CHNS/O Elemental Analyzer	Based on the Pregl-Dumas method. Samples are combusted with user flexible mode and gases are separated by frontal chromatography and eluted gases are measured using TCD.
Elementar Analysensysteme GmbH	Donaustrasse 7 D-63452 Hanau, Germany www.elementar.de	vario Macro, vario Max, vario EL III	Samples are dropped into the combustion tube at user selected temperature up to 1200°C. The use of tin vessels further elevates the temperature up to 1800°C. Complete combustion is ensured with O2 jet injection. Except for N_2, other gases are retarded into specific adsorption trap. After TCD signal for N2 is received, adsorption traps are thermally desorbed and the corresponding gases detected with TCD sequentially.
Thermo Scientific (part of Thermo Fisher Scientific Corporation)	81 Wyman Street, Waltman, MA 02454, USA www.thermo.com	Flash EA 1112 NC	Detection by TCD.

with the automated dry combustion technique (De Vos *et al.* 2007) and chemical disposal is an environmental problem. In general, we conclude that automated dry combustion is the only reliable, comprehensive method to determine soil C concentration with the added benefit of also measuring N and S at the same time. With a limited budget, LOI method might be used rather than the automated technique, but the correlation factor in between them should be reported with the results.

Although soil sampling in the field and automated dry combustion is considered as the standard method, the whole process is expensive, time consuming and labor intensive. The automated analysis of prepared soil samples alone costs around $12 per sample. Moreover, without intensive soil sampling, it is hard to detect changes in soil C over large landscapes due to spatial heterogeneity (Freibauer *et al.*, 2004). All laboratory analyses use a small quantity of homogenized samples, generally between 0.1 to 1 g. These major limitations with ex situ methods instigated the development of alternative methods, particularly in situ ones, to achieve higher precision, faster analyses, and lower costs and than the present ex situ determination methods.

TABLE 6
Features of in situ soil C determination techniques

Method	Principle	Penetration in soil (cm)	Sampled volume (cm^3)	Features
Mid- and near-infrared reflectance spectroscopy (MIRS/NIRS)	NIRS (400–2500 nm) and MIR (2500–25000 nm) region utilized to quantify soil C. Based on the absorption of C-H, N-H and O-H groups found in organic constituents	0.2–1	~10	Invasive, MIR region needed KBr dilution because of strong absorptions. Strength of these absorptions may result into spectral distortions and nonlinearities.
Laser-induced breakdown spectroscopy (LIBS)	Laser is focused on sample forming microplasma that emits light characteristic of the sample elemental composition	0.1	$\sim 10^{-2}$	Able to provide data at 1 mm resolution, invasive, roots and rock fragments presence may cause C signal variability.
Inelastic neutron scattering (INS)	Based on inelastic scattering of fast, 14 MeV, neutrons from C nuclei and subsequent detection of gamma rays emitted from first C excited level	30	$\sim 10^5$	Nondestructive, multi-elemental, scanning modality, analytic response function

III. IN SITU METHODS

New in situ soil C methods promise high precision without as much sample processing time and their subsequent analysis. In situ methods mainly are based on remote sensing and spectroscopic measurements in the field (Table 6). Spectroscopic methods include infra-red reflectance near-infra-red (NIR) and mid-infra-red, laser-induced breakdown spectroscopy (LIBS) and inelastic neutron scattering (INS). Potential of these methods are being calibrated with reference to soil sampling and subsequent analysis with automated dry combustion method.

A. Infrared Reflectance Spectroscopy

Infrared reflectance spectroscopy is a rapid technique for measuring soil C based on the diffusely reflected radiation of illuminated soil (McCarty *et al.*, 2002). Within diffuse reflectance spectroscopy, both the near infrared region (NIR, 400–2500 nm), and the mid infrared (MIR, 2500–25000 nm) region have been evaluated for quantifying soil C (Morón and Cozzolino, 2002; McCarty *et al.*, 2002; Russell, 2003). NIR uses a quantitative determination of components of complex organic compounds, whereas MIR spectroscopy involves the spectral interpretation of chemical structures. McCarty *et al.* (2002) reported that organic and inorganic C pools can be measured simultaneously by spectral analysis; they observed that useful calibrations for soil C can be developed using MIR, and to a lesser extent, NIR analysis. NIR is based on the absorption of the C-H, N-H, and O-H groups found in organic compounds. These absorptions are overtones and combination bands of the much stronger absorption band seen in MIR spectra (Murray 1993; Batten, 1998; Deaville and Flinn, 2000; Reeves, 2000). Multiple regression statistics (Partial Least Square and Principal Component Analysis) relate the NIR data at selected wavelengths to reference values for calibration (Deaville and Flinn, 2000; Cozzolino and Morón, 2006). The major limitation of NIR is the continual need for calibration and quality control. Due to differences in particle size and soil mineral absorption intensities, NIR absorption by soil is not linearly related to the individual soil matrix components (Russell, 2003). The NIR has excellent performance ($R^2 = 0.961$ to 0.975) when applied to a calibration set of samples of a similar particle size distribution. However, predictability is low in samples with heterogeneous particle size and high variability in moisture content (Madari *et al.*, 2005). Veris Technologies (Salina, Kansas) developed a mobile in situ NIR device and field validation results predicted SOM with an R^2 value of 0.67 between the laboratory and NIR prediction (Christy, 2008). Accuracy of prediction will increase with the increase in area. However, NIR simultaneously measures quantitatively and qualitatively certain soil parameters (like forms of C), in addition to C content. Commercial field portable NIR instruments are commercially available and cost around $20,000 (Oceanoptics Inc., Florida), and widen the use of NIR for in situ measurement of soil C.

B. Laser-Induced Breakdown Spectroscopy (LIBS)

Laser-induced breakdown spectroscopy (LIBS) is based on atomic emission; the soil's C content is determined by

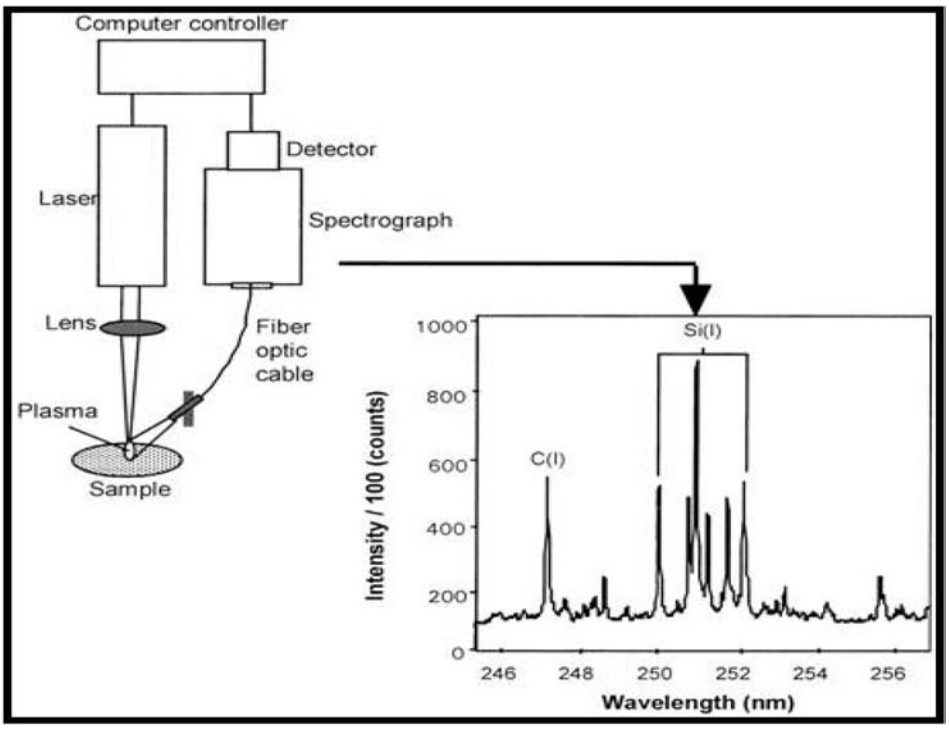


FIG. 1. Schematic presentation of the LIBS system, collection of microplasma, detection, and the spectral resolution of a sample (adapted from Cremers *et al.*, 2001).

analyzing the unique spectral signature of C (at 247.8 or 193 nm, or both). A laser beam at a specific wavelength, e.g., 1064 nm, is focused on each sample with a lens of 50 mm focal length to form microplasma that emits light that is characteristic of the sample's elemental composition (Ebinger *et al.*, 2006) (Fig. 1). The emitted light is spectrally resolved using a grated-intensified photodiode array detector. Intact soil cores or discrete, pressed samples are used for analysis; spectra are collected along a soil core or from each discrete sample. The spatial variability of C in soil profiles is accounted for by the ability to analyze and average multiple spots. Cremers *et al.* (2001) compared the data from LIBS measurements with those from dry combustion and observed a high correlation of 0.96 for soils of similar morphology. They also reported that LIBS quickly determines C (in less than a minute) with excellent instrumental detection limit of $\sim$300 mg kg^{-1} and a precision of 4–5%. The greatest advantage of LIBS is its capability for remote surface chemical analysis of samples although the utility of this feature for soil C analysis remains to be demonstrated. The rapid determination of soil C and the portability of LIBS systems afford the potential to collect and analyze thousands of measurements to characterize soil C content, its distribution and heterogeneity over a large area; nevertheless, these undoubted advantages need to be balanced against the very small volumes analyzed. In addition to soil C, LIBS measures most of the major elements (nitrogen, phosphorus and potassium) present in soil and can be widely applicable to enumerate the soil fertility status or solving soil health related problems like heavy metal contamination.

Soil properties (e.g., texture, carbonate and moisture content) influence LIBS analyses; thus, numerous calibration curves based on soil texture were required. However, this practice is unacceptable for a field deployable instrument. There is urgency in developing a "universal calibration curve," an essential tool for soil C measurements. The new approach of using multivariate analysis for quantifying soil C builds upon and extends the preliminary observations of Cremers *et al.*, 2001, and Martin *et al.*, 2002. Multivariate statistical analysis (MVA) helps control the variability in C concentrations due to the influence of the soil's matrix, which accounts for the textural dependence of calibrations. Acid washing of soil samples to remove calcium carbonate reduces the standard deviation by almost 8% after normalizing the C signal to the silicon (Si) signal. The reproducibility of LIBS analyses can be improved by (i) increasing the number of shots and averaging the spectra over more shots, (ii) applying the method of intensity ratios of C with either Si or Al and, (iii) using the MVA techniques (Martin *et al.*, 2003). Commercialization of a portable LIBS system has reduced its unit cost and might increase its employment for high-resolution soil C analyses. Cost of a LIBS unit that can detect soil C with high detection limit as well as collect other spectral features for multivariate analysis costs $100,000. Considering the number of samples that can be analyzed and labor cost associated with sample processing, LIBS will be more profitable over the time compared to automatic dry combustion technique. Future research is needed to reduce the variability in the LIBS signal caused by the presence of rock fragments, roots, and other

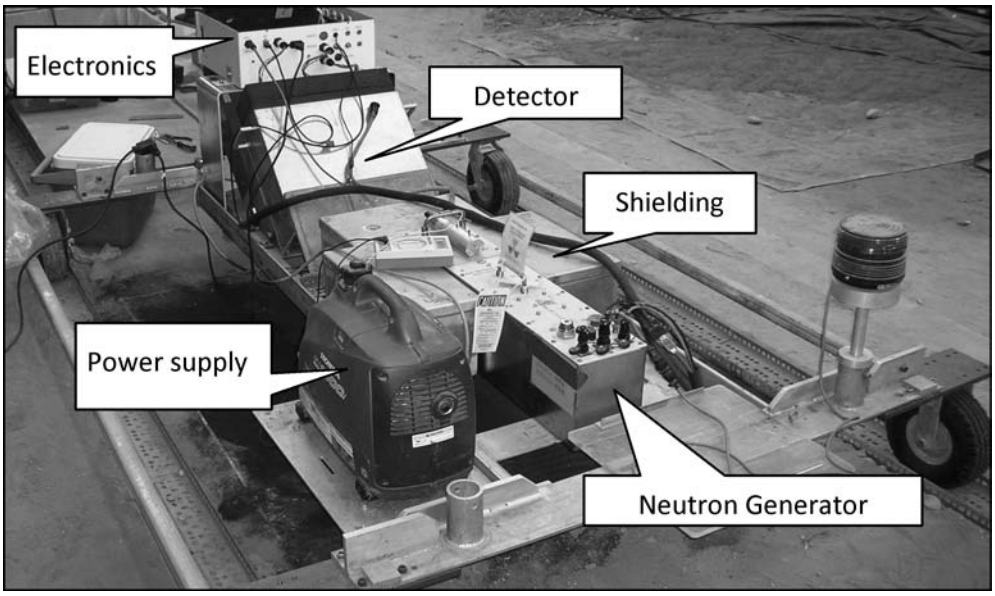


FIG. 2. Different major sections of an Inelastic Neutron Scattering (INS) alpha prototype.

materials. A collaborative soil C detection and quantification effort using LIBS was initiated between Oak Ridge National Laoratory and Los Almos National Laboratory.

C. Inelastic Neutron Scattering

The new inelastic neutron scattering (INS) system for soil C analysis is based on spectroscopy of gamma rays resulting from fast neutrons interacting with the nuclei of the elements in soil. A neutron generator, which is turned off when not in operation, generates fast neutrons that penetrate the soil and stimulate gamma rays that subsequently are detected by an array of NaI detectors. The peak areas in the measured spectra are proportional to the soil's elemental content. The key elements measured presently are C, Si, O, N, H, Al, and K. Since the INS is based on nuclear processes that are very fast, it is insensitive to the chemical configuration of the element and can be used in a scanning mode. Figure 2 shows the main components of the self-contained INS alpha prototype field unit mounted on a cart. Having placed the INS system in position, it hovers about 30 cm above the ground; data acquisition typically is set for an interval between 30 to 60 min. The INS subsequently analyzes the acquired spectra for spectral peak intensities (counts), and, using an established calibration line, reports the results instantly in units of kg C m^{-2} (Wielopolski *et al.*, 2008).

The INS system, which by and large sees a constant volume, is linear from zero up to very high levels of soil C. It has multi-elemental capabilities, for example, the H peak was calibrated against soil moisture. The following are INS's unique and invaluable characteristics: (a) Interrogation of large volumes containing over 200 kg of soil; (b) a large footprint of about 2 m^2; and, sampling the soil to a depth of about 30 cm. Since the INS response is governed by the exponential attenuation functions, Beer's law, of the neutrons penetrating into the soil and the gamma rays emanating from it, these values are not strictly defined but rather effective, or on average. Implicitly, the linear regression of the INS calibration line embeds them. Calibrating the INS system with synthetic soils in which sand was mixed with a known amount of carbon yielded an r^2 value of 0.99 (Wielopolski *et al.*, 2004). The system was also calibrated in grassland, pine forest, and hardwood forest in the Blackwood Division of the Duke Forest near Durham with an r^2 value of 0.97. The latter calibration in terns of g C cm^{-2} was accomplished using chemical analysis by dry combustion of samples taken after homogenizing an excavation pit of $40\times40\times40$ cm^3 (Wielopolski *et al.*, 2008). The INS system set up seeing a constant volume does not require the knowledge of the exact volume. To the first approximation INS is looking into a constant volume of about 0.4 m^3 in the soil in which the C signal is proportional to the number of C atoms in that volume. Small variations in the soil bulk density have negligible effect on the interrogated volume. The INS system is directly calibrated in g C cm^{-2} representing the total carbon in the column below a unit area regardless of the BD that varies with depth. Once the system is calibrated we know exactly the amount of C to be the same as determined by the conventional sampling and dry combustion analysis. Thus the C credit payment is proportional to the change.

The presence of coarse fragments reduces the amounts of soil in that volume thus reducing the C signals. When corrected for the solid fraction the data points coincided with the calibration line. INS being a nuclear method is insensitive to chemical configuration of the C. However, by measuring additional

TABLE 7
Comparison of soil C determination methods using automated dry combustion, LIBS and INS techniques

Automated dry combustion	LIBS	INS
	Sampling and processing	
Destructive soil sampling and processed to finely homogeneous powder	Destructive soil sampling using cores, no processing needed	Nondestructive
	Analysis time	
Sampling to final result needs at least week	Few minutes	One hour
	Foot print	
Core diameter (2–3.5 cm)	Laser beam diameter (200 μm)	$\sim$ 1.5 m^2
	Analysis	
In most cases, thermal conductivity of CO_2 (evolved from combustion of soil) converted to percent C using homogenized sample weight	Spectra normalized to the total detected emission and calibration by standard samples to determine total C	Spectra normalized to a monitored neutron generator output. Trapezoidal peak net areas converted instantaneously to elemental C.
	Future developments	
None	Improved sensitivity	Improved sensitivity, measuring depth profile
	Cost of unit	
$40,000–60,000	$100,000	$150,000

elements; for example calcium and magnesium, it might be possible to partition SIC and SOC. The system is not commercial at this point; however, a system cost is estimated at about $150,000 and no consumable costs are involved. The INS system is an electrical device producing radiation; as such it has to satisfy radiation regulatory requirements. However, it does not carry any radioactive sources and at the end of data acquisition is turned off. The device is well collimated and shielded without introducing any environmental hazard.

D. Remote Sensing

Since 1960, remote sensing has been explored as an alternative nondestructive method for SOC determination, at least of surface soils (Merry and Levine, 1995). Reflectance of various spectral bands was correlated with soil properties, including SOM content (Chen *et al.*, 2000; 2007). Spectral sensors were developed and examined to measure SOM (Pitts *et al.*, 1983; Griffis, 1985; Smith *et al.*, 1987; Shonk *et al.*, 1991). Sensors usually operate with wavelengths between 0.3 μm and 1 m and are divided into the following four groups: (1) visible (0.4–0.7 μm), (2) reflective infrared (0.7–3 μm), (3) thermal infrared (8–14 μm), and (4) microwave (1 mm–1m). A wavelength between 0.4 μm to 2.5 μm, is suitable for soil with $>$2% SOM content (Baumgardner *et al.*, 1970). Research shows that predictions can be made of the SOM content from light reflectance with a linear or curvilinear relationship in the visual and infrared range (Baumgardner *et al.*, 1970; Smith *et al.*, 1987; Sudduth and Hummel, 1988; Henderson *et al.*, 1992). Ben-Dor and Banin (1995) successfully correlated statistical data and Landsat TM imaging analysis with the sand, clay, and SOM content of different soils in Madison County, Alabama. Chen *et al.* (2000) proposed a process of mapping SOC with remotely sensed imagery (bare field) that includes image filtering, regression analysis, classification and reclassification. With this method, they obtained a high correlation (0.97–0.98) between predicted and measured values at field scale level (of area 115 ha) in coastal plain region of Georgia. Mapping of SOC with remote sensing has proven to be both accurate and economic; however this method requires separate sampling and mapping for each crop field. Chen *et al.* (2007) proposed to group field based on image similarity and mapping them together as one group to reduce sampling costs.

Although there is a strong relationship between remotely sensed spectral data and SOC content, prediction at different spatial scales has not been achieved. Moreover, to draw inferences of SOC content from satellite imagery on a large scale necessitates having surrogate indices such as vegetation type and species or soil moisture (Merry and Levine, 1995). Beside these shortcomings, remote sensing with its high resolution monitoring abilities is applicable for predicting SOC distribution, which is not feasible by any other means.

All methods have pros and cons and they should be matched to specific measurement needs and applications before they are selected or rejected. The choice of an instrument or measurement techniques will depend upon the researchers' need and resources, such as the project objective and funding allotted for the project.

IV. SUMMARY AND CONCLUSIONS

Considering both ex situ and in situ measurements, we suggest that three methods, automated dry combustion, LIBS and INS, have higher precision and detection limit over others. Table 7 compares analytical advantages and disadvantages of these three methods. Although LIBS and INS system have major advantages over ex situ, automated dry combustion method, but still need to consider following factors: (i) separation of SIC concentration from total soil C concentration; (ii) consideration of soil bulk density, and root and rock fragments containing C; (iii) reducing complexity associated with operation and calculation; and (iv) planning of measurement protocols for different soil types and landscape situations. In general, there are major constraints to determine SOC content on an area basis mainly due to spatial variability in SOC distribution and uncertainties associated with soil bulk density estimation. To improve the accuracy in prediction of SOC content over a landscape, it is foremost need to develop a sound soil sampling method to counteract the question of spatial variability and least cost effective and routine methods of measuring/predicting soil bulk density. In situ methods, particularly LIBS and INS have potential to solve these problems, particularly uncertainties associated with spatial variability. Until these advanced techniques are calibrated, methods for determining SOC will follow the legacy of standard field soil core sampling and automated dry combustion analysis.

ACKNOWLEDGMENTS

This papar is also referenced as Los Alamos National Laboratory unclassified report LA-UR-08-05870. Some or all LIBS applications described in this paper are protected by U.S. Patent Application held by Los Alamos National Loboratory.

REFERENCES

Abella, S. R., and Zimmer, B. W. 2007. Estimating organic carbon from loss-on-ignition in northern Arizona forest soils. *Soil Sci. Soc. Am. J.* **71**: 545–550.

Ames, W. J., and Gaither, E. W. 1914. Determination of carbon in soils and soil extracts. *J. Ind. Eng. Chem.* **6**: 561.

Barshad, I. 1965. Thermal analysis techniques for mineral identification and mineralogical composition. **In**: *Methods of Soil Analysis. Part I. Agron. Monograph 9*, pp. 699–742. Black, C. A., Ed., ASA, CSSA, and SSSA, Madison, WI.

Batjes, N. H. 1996. Total carbon and nitrogen in the soils of the world. *Eur. J. Soil. Sci.* **47** (2): 151–163.

Batten, G. D. 1998. Plant analysis using near-infrared reflectance spectroscopy: the potential and the limitations. *Aust. J. Exp. Agric.* **38**: 697–706.

Baumgardner, M. F., Kristof, S., Johannsen, C. J., and Zachary, A. 1970. Effects of organic matter on the multispectral properties of soils. *Proc. Indiana Acad. Sci.* **79**: 413–422.

Ben-Dor, E., and Banin, A. 1989. Determination of organic matter content in arid-zone soils using a simple "loss-on-ignition" method. *Commun. Soil Sci. Plant Anal.* **20**: 1675–1695.

Ben-Dor, E., and Banin, A. 1995. Near-infrared analysis as a rapid method to simultaneously evaluate several soil properties. *Soil Sci. Soc. Am. J.* **56**: 364–372.

Brye, K. R., and Slaton, N. A. 2003. Carbon and nitrogen storage in a typic Albaqualf as affected by assessment method. *Commun. Soil Sci. Plant Anal.* **34**: 1637–1655.

Cameron, F. K., and Breazeale, J. F. 1904. The organic matter in soils and subsoils. J. Am. Chem. Soc. **26**: 29–45.

Chen, F., Kissel, D. E., West, L. T., and Adkins, W. 2000. Field-scale mapping of surface soil organic carbon using remotely sensed imagery. *Soil Sci. Soc. Am. J.* **64**: 746–753.

Chen, F., Kissel, D. E., West, L. T., Adkins, W., Rickman, D., and Luvall, J. C. 2007. Mapping soil organic carbon concentration for multiple fields with image similarity analysis. *Soil Sci. Soc. Am. J.* **72**: 186–193.

Christensen, B. T., and Malmros, P. A. 1982. Loss-on-ignition and carbon content in a beech forest soil profile. *Holartic Ecol.* **5**: 376–380.

Christy, C. D. 2008. Real-time measurement of soil attributes using on-the-go near infrared reflectance spectroscopy. *Computers and Electronics in Agriculture* **61**: 10–19.

Cozzolono, D., and Morón, A. 2006. Potential of near-infrared reflectance spectroscopy and chemometrics to predict soil organic carbon fractions. *Soil Tillage Res.* **85**: 78–85.

Cremers, D. A., Ebinger, M. H., Breshears, D. D., Unkefer, P. J., Kammerdiener, S. A., Ferris, M. J., Catlett, K. M., and Brown, J. R. 2001. Measuring total soil carbon with Laser induced breakdown spectroscopy (LIBS). *J. Environ. Qual.* **30**: 2202–2206.

David, M. B. 1988. Use of loss-on-ignition to assess soil organic carbon in forest soils. *Commun. Soil Sci. Plant Anal.* **19**: 1593–1599.

De Vos, B., Letterns, S., Muys, B., and Deckers, J. A. 2007. Walkley-Black analysis of forest soil organic carbon: recovery, limitations and uncertainty. *Soil Use Manage*. **23**: 221–229.

De Vose, B., Vandecasteele, B., Deckers, J., and Muys, B. 2005. Capability of loss-on-ignition as a predictor of total organic carbon in non-calcareous forest soils. *Commun. Soil Sci. Plant Anal.* **36**: 2899–2921.

Deaville, E. R., and Flinn, P. C. 2000. Near-infrared (NIR) spectroscopy: an alternative approach for the estimation of forage quality and voluntary intake. **In:** *Forage Evaluation in Ruminant Nutrition*, pp. 301–320. Givens, D. I., Owen, E., Axford, R. F. E., and Omed, H. M., Eds., CABI Publishing, UK.

Díaz-Zorita, M. 1999. Soil organic carbon recovery by the Walkley-Black method in a typic Hapludoll. *Commun. Soil Sci. Plant Anal.* **30**: 739–745.

Donkin, M. J. 1991. Loss-on-ignition as an estimator of soil organic carbon in A-horizonforestry soils. *Commun. Soil Sci. Plant Anal.* **22**: 233–241.

Ebinger, M. H., Cremers, D. A., Meyer, C. M., and Harris, R. D. 2006. Laser-induced breakdown spectroscopy and applications for soil carbon measurement.. **In:** *Carbon Sequestration in Soils of Latin America.* pp. 407–421. Lal, R., C. C. Cerri, M. Bernoux, J. Etchveers, E. Cerri., Eds., Food Products Press, Binghamton, NY.

Eswaran, H., Van den Berg, E., Reich, P., and Kimble, J. M. 1995. Global soil carbon resources. **In:** *Soils and Global Change.* pp. 27–44. Lal, R., Kimble, J., Levine, E. and Stewart, B. A., Eds., Lewis Publishers, CRC Press, Inc., Boca Raton, FL.

Freibaur, A., Rounsevell, M. D. A., Smith, P., and Verhagen, J. 2004. Carbon sequestration in the agricultural soils of Europe. *Geoderma* **122**: 1–23.

Gehl, R. J., and Rice, C. W. 2007. Emerging technologies for in situ measurement of soil carbon. *Clim. Change* **80**: 43–54.

Grewal, K. S., Buchan, G. D., and Sherlock, R. R. 1991. A comparison of three methods of organic carbon determinations in some New Zealand soils. *J. Soil Sci.* **42**: 251–257.

Griffis, C. L. 1985. Electronic sensing of soil organic matter. *Trans. ASAE* **28**: 703–705.

Hammes, K., Hammes, K., Schmidt, M. W. I., Smernik, R. J., Currie, L. A., Ball, W. P., Nguyen, T. H., Louchouarn, P., Houel, S., Gustafsson, O., Elmquist, M., Cornelissen, G., Skjemstad, J. O., Masiello, C. A., Song, J., Peng, P., Mitra, S., Dunn, J.C., Hatcher, P. G., Hockaday, W. C., Smith, D. M., Hartkopf-Fröder, C., Böhmer, A., Lüer, B., Huebert, B. J., Amelung, W., Brodowski, S., Huang, L., Zhang, W., Gschwend, P. M., Flores-Cervantes, D. X., Largeau, C., Rouzaud, J.-N., Rumpel, C., Guggenberger, G., Kaiser, K., Rodionov, A.,

Gonzalez-Vila, F. J., Gonzalez-Perez, J. A., de la Rosa, J. M., Manning, D. A. C., Lo'pez-Cape'l, E., Ding, L. 2007. Comparison of quantification methods to measure fire-derived (black/elemental) carbon in soils and sediments using reference materials from soil, water, sediment and the atmosphere. *Global Biogeochem. Cycles* **21**: GB3016.

Heans, D. L. 1984. Determination of total organic-C in soils by an improved chromic acid digestion and spectrophotometric procedure. *Commun. Soil Sci. Plant Anal.* **15**: 1191–1213.

Hedges, J. I., et al. 2000. The molecularly-uncharacterized component of non-living organic matter in natural environments. *Org. Geochem.* **37**: 501–510.

Heiri, O., Lotter, A. F., and Lemcke, G. Loss on ignition as a method for estimating organic and carbonate content in sediments and comparability of results. *J. Paleolimnology* **25**: 101–110.

Henderson, T. L., Baumgardner, M. F., Frazeier, D. P. Stott, D. E., and Coster, D. C. 1992. High dimensional reflectance analysis of soil organic matter. *Soil Sci. Soc. Am. J.* **56**: 865–872.

Howard, P. J. A., and Howard, D. M. 1990. Use of organic carbon and loss-on-ignition to estimate soil organic matter in different soil types and horizons. *Biol. Fertil. Soils* **9**: 306–310.

Jackson, M. L. 1958. *Soil Chemical Analysis.* Prentice-Hall, Inc., Englewood Cliffs, NJ.

Jimnez, R. R., and Ladha, J. K., 1993. Automated elemental analysis: a rapid and reliable but expensive measurement of total carbon and nitrogen in plant and soil samples. *Commun. Soil Sci. Plant Anal.* **24**: 1897–1924.

Jobbagy, E. G., and Jackson, R. B. The vertical distribution of soil organic carbon and its relation to climate and vegetation. *Ecol. Appl.* **10**(2): 423–436.

Kamara, A., Rhoades, E. R., and Sawyer, P. A. 2007. Dry combustion carbon, Walkley-Black carbon, and loss on ignition for aggregate size fractions on toposequence. *Commun. Soil Sci. Plant Anal.* **38**: 2005–2012.

Konen, M. E., Jacobs, P. M., Burras, C. L., Talaga, B. J., and Mason, J. A. 2002. Equations for predicting soil organic carbon using loss-on-ignition for north central U.S. soils. *Soil Sci. Soc. Am. J.* **66**: 1878–1881.

Lal, R. 2006. Bulk density measurements for assessment of soil carbon pools. **In:** *Carbon Sequestration in Soils of Latin America*, pp. 491–516. Lal, R., Cerri, C. C., Bernoux, M., Etchveers, J., and Cerri. E., Eds., Food Products Press, Binghamton, NY.

Lide, D. R. (ed.). 1993. *Handbook of Chemistry and Physics*. CRC Press, Ann Arbor, MI.

Lowther, J. R., Smethurst, P. J., Carlyle, J. C., and Nambiar, E. K. S. 1990. Methods for determining organic carbon in podzolic sands. *Commun. Soil Sci. Plant Anal.* **21**: 457–470.

Madari, B. E., Reeves III, J. B., Coelho, M. R., Machado, P. L. O. A., and De-Polli, H. 2005. Mid- and near-infrared spectroscopic determination of carbon in diverse set of soils from the Brazilian national Soil Collection. *Spect. Lett.* **38**: 721–740.

Martin, M. Z., Wullschleger, S. D., Garten Jr., C. T., and Palumbo, A. V. 2003. Laser-induced breakdown spectroscopy for the environmental determination of total carbon and nitrogen in soils. *Applied Optics* **42**(12): 2072–2077.

Masiello, C. A. 2004. New directions in black carbon organic geochemistry. *Mar. Chem.* **92**: 201–213.

McCarty, G. W., Reeves III, J. B., Reeves, V. B., Follet, R. F., and Kimble, J. M. 2002. Mid-infrared and near-infrared diffuse reflectance spectroscopy for soil carbon measurement. *Soil Sci. Am. J.* **66**: 640–646.

Meibus, L. J. 1960. A rapid method for the determination of organic carbon in soil. *Anal. Chim. Acta.* **22**: 120–121.

Merry, C. J., and Levine, E. R. 1995. Methods to assess soil carbon using remote sensing techniquesIn: *Advances in Soil Science: Soils and Global Change*, pp. 265–274. Lal, R., Kimble, J., Levine, E., and Stewart, B. A., Eds., Lewis Publishers, CRC Press, Inc., Boca Raton, Fl.

Mikhailova, E. A., Noble, R. R. P., and Post, C. J. 2003. Comparison of soil organic carbon recovery by Walkley-Black and dry combustion methods in the Russian Chernozem. *Commun. Soil Sci. Plant Anal.* **34**: 1853–1860.

Mitchell, J. 1932. The origin, nature, and importance of soil organic constituents having base exchange properties. *J. Am. Soc. Agron.* **24**: 256–275.

Murray, I. 1993. Forage analysis by near-infrared spectroscopy. **In:** *Sward Herbage Measurement Handbook.* pp. 285–312. Davies, A., Baker, R. D. Grant, S. A., and Laidlaw, A. S. Eds., British Grassland Society, Reading, U.K.

Nelson, D. W., and Sommers, L. E. 1982. Total carbon, organic carbon, and organic matter. **In:** *Methods of Soil Analysis. Part 2. Agron. Monogr. 9.*, pp. 539–579. Page, A. L. et al., Eds., ASA and SSSA, Madison, WI.

Nelson, D. W., and Sommers, L. E. 1996. Total carbon, organic carbon, and organic matter. **In:** *Methods of Soil Analysis. Part 2. Agron. Monogr. 9.* pp. 961–1010. Sparks, D. L. et al. Eds., ASA and SSSA, Madison, WI.

Pérez, D. V., Alcantara, S., Arruda, R. J., and Menegheli, N. A. 2001. Comparing two methods for soil carbon and nitrogen determination using selected Brazilian soils. *Commun. Soil Sci. Plant Anal.* **32**: 295–309.

Pitts, M. J., Hummel, J. W., and Butler, B. J. 1983. Sensors utilizing light reflection to measure soil organic matter. Pap. 83–1011. *Am. Soc. Agric. Eng.*, St. Joseph, MI.

Rather, J. B. 1917. An accurate loss on ignition method for determination of organic matter in soils. Arkansas Agric. Exp. Stn. Bull 140.

Reeves III, J. B. 2000. Use of near-infrared reflectance spectroscopy. **In:** *Farm Animal Metabolism and Nutrition*, pp. 184–209. D'Mello, J. P. F. Ed., CABI Publishing, UK.

Rogers, R. E., and W. R. Rogers. 1848. New method of determining the carbon in native and artificial graphite, etc. *Am. J. Sci.* **2**: 352.

Rumpel, C., Balesdent, J., Grootes, P., Weber, E., and Kogel-Knabner, I. 2003. Quantification of lignite- and vegetation-derived soil carbon using C-14 activity measurements in forested chronosequence. *Geoderma* **112**: 155–166.

Rumpel, C., Janik, L. J., Skjemstad, J. O., and Kogel-Knabner, I. 2001. Quantification of carbon derived from lignite in soils using mid-infrared spectroscopy and partial least squares. *Org. Geochem.* **32**: 831–839.

Russell, C. A. 2003. Sample preparation and prediction of soil organic matter properties by near infra-red reflectance spectroscopy. *Commun. Soil Sci. Plant Anal.* **34**(11 & 12): 1557–1572.

Santi, C., Certini, G., and D'Acqui, L. P. 2006. Direct determination of organic carbon by dry combustion in soils with carbonates. *Commun. Soil Sci. Plant Anal.* **37**: 155–162.

Schnitzer, M. 1991. Soil organic matter: The next 75 years. *Soil Sci.* **151**: 41–58.

Schollenberger, C. J. 1927. A rapid approximate method for determining soil organic matter. *Soil Sci.* **24**: 65–68.

Schulte, E. E., and Hopkins, B. G., 1996. Estimation of soil organic matter by weightloss-on-ignition. **In:** *Soil Organic Matter: Analysis and Interpretation. SSSA Special Publication no. 46*, pp. 21–31. Magdoff, F. R. Ed., Soil Science Society of America, Inc. Madison, WI.

Schulte, E. E., Kauffmann, C., and Peter, J. B. 1991. The influence of sample size and heating time on soil weight loss-on-ignition. *Commun. Soil Sci. Plant Anal.* **22**: 159–168.

Sherrod, L. A., Dunn, G., Peterson, G. A., and Kolberg, R. L. 2002. Inorganic carbon analysis by modified pressure-calcimeter method. *Soil Sci. Soc. Am. J.* **66**: 299–305.

Shonk, J. L., Gaultney, L. D., Schulze, D. G., and Van Scoyoc, G. E. 1991. Spectroscopic sensing of soil organic matter content. *Trans. ASAE* **34**: 1978–1984.

Skjemstad, J. O., and Taylor, J. A. 1999. Does the Walkley-Black method determine soil charcoal? *Commun. Soil Sci. Plant Anal.* **30**: 2299–2310.

Smith, D. L., Worner, C. R., and Hummel, J. W. 1987. Soil spectral reflectance relationship to organic matter content. Pap. 87–1608. *Am. Soc. Agric. Eng.*, St. Joseph, MI.

Smith, K. A., and Tabatabai, M. A., 2004. Automated instruments for the determination of total carbon, nitrogen, sulfur, and oxygen. **In:** *Soil and Environmental Analysis: Modern Instrumental Techniques.* pp. 235–282. Smith, K. A., and Cresser, M. S., Eds., Marcel Dekker, NY.

Soon, Y. K. and Abboud, S. 1991. A comparison of some methods for soil organic carbon determination. *Commun. Soil Sci. Plant Anal.* **22**: 943–954.

Spain, A. V., Probert, M. E., Isbell, R. F., and John, R. D. 1982. Loss-on-ignition and the carbon contents of Australian soils. *Aust. J. Soil Res.* **20**: 147–152.

Sudduth, K. A., and Hummel, J. W. 1988. *Optimal signal processing for soil organic matter determination*. Pap. 88–7004. *Am. Soc. Agric. Eng.*, St. Joseph, MI.

Tabatabai, M. A. 1996. Soil organic matter testing: an overview. **In:** *Soil Organic Matter: Analysis and Interpretation. SSSA Special Publication no. 46.* pp. 1–9. Magdoff, F. R. Ed., Soil Science Society of America, Inc. Madison, WI.

Tabatabai, M. A., and Bremner, J. M. 1970. Use of the Leco automatic 70-second carbon analyzer for total carbon analysis in soils. *Soil Sci. Soc. Am. Proc.* **34**: 608–610.

Tabatabai, M. A., and Bremner, J. M. 1991. Automated instruments for determination of total carbon, nitrogen, and sulfur in soils by combustion techniques. In: *Soil Analysis*, pp. 261–286. Smith, K. A. Ed., Marcel Dekker, New York.

Tinsley, J. 1950. Determination of organic carbon in soils by dichromate mixtures. **In:** *Trans. 4th Int. Congr. Soil Sci., Vol. 1.* pp. 161–169. Hoitsemo Brothers, Gronigen, Netherlands.

Tyurin, I. V. 1931. A new modification of the volumetric method of determining soil organic matter by means of chromic acid. *Pochvovedenie*. **26**: 36–47.

Tyurin, I. V. 1935. Comparative study of the methods for the determination of organic carbon in soils and water extracts of soils. *Dokuchaive Soil Inst. Stud, Genesis Geogr. Soils.* **1935**: 139–158.

Ussiri, D. A. N., and Lal, R. 2008. Method for determining coal carbon in the reclaimed minesoils contaminated with coal. *Soil Sci. Soc. Am. J.* **72**(1): 231–237.

Walkley, A., and Black, I. A. 1934. An examination of the Degtjareff method for determininfg soil organic matter and a proposed modification of the chromic acid titration method. *Soil Sci.* **37**: 29–38.

Wang, X. J., Smethurst, P. J., and Herbert, A. M. 1996. Relationships between three measures of organic matter or carbon in soils of eucalypt plantations in Tasmania. *Aust. J. Soil Res.* **34**: 545–553.

Warrington, R., and Peake, W. A. 1880. On the determination of carbon in soils. *J. Chem. Soc.* (London) **37**: 617–625.

West, T. O., Brandt, C. C., Wilson, B. S., Hellwinckel, C. M., Tyler, D. D., Marland, G., De La Torre Ugarte, D. G., Larson, J. A., and Nelson, R. G. 2008. Estimating regional changes in soil carbon with high spatial resolution. *Soil Sci Soc. Am. J.* **72**: 285–294.

Wieloploski, L., Hendrey, G., Johnsen, K., Mitra, S., Prior, A., Rogers, H. H., and Torbert, H. A. 2008. Non-destructive system for analyzing carbon in the soil. *Soil Sci. Soc. Am. J.* **72**: 1269–1277.

Wielopolski, L., Mitra, S., Hendrey, G., Orion, I., Prior, S., Rogers, H. H., and Torbert, H. A. 2004. Non-destructive soil carbon analyzer (ND-SCA), BNL Report no. 72200–2004.

Wright, A. F., and Bailey, J. S., 2001. Organic carbon, total carbon, and total nitrogen determination in soils of variable calcium carbonate contents using a LECO CN-2000 dry combustion analyzer. *Commun. Soil Sci. Plant Anal.* **32**: 3243–3258.

Critical Reviews in Plant Science, 28:1–3, 2009

ISSN: 0735-2689 print / 1549-7836 online
DOI: 10.1080/07352680903240933

Taylor & Francis
Taylor & Francis Group

Populus Community Mega-Genomics: Coming of Age

Zong-Ming Cheng[1] and Gerald A. Tuskan[2]

[1]*Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996*
[2]*Environmental Science Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831*

X Plants, once germinated or transplanted, will live and die in the same location their entire life. And for woody perennial plants, this circumstance exists for many years, even millennia. As a result, woody perennial plants in natural ecosystems or in plantations uniquely create an environment that fosters, shapes, and preserves heterogeneous communities. That is, trees co-exist with other organisms such as microbial and fungal pathogens, endophytes, arboreal insects, and rhizospheric microorganisms. Moreover, over a life span, not only do trees have to respond to potential biotic interactions, they also have to respond to abiotic challenges such as seasonal changes from spring to summer to winter as well as interannual climatic changes. Despite such challenges, arboreal forms of plant growth thrive today and throughout evolutionary history (Groover, 2005). Still, our understanding of the mechanisms of how trees adapt to local conditions and cope with biotic and abiotic challenges remains largely obscure.

In the last ten years, *Populus* has emerged as the model tree for biological studies among all forest trees because of numerous attributes, including rapid growth, relatively short reproductive cycles, ease of vegetative propagation, general amenability to *in vitro* culture, regeneration and transformation, and extensive genetic and genomics resources such as breeding populations, genetic maps and large EST resources (Bradshaw *et al.*, 2000; Jansson and Douglas, 2007; Tuskan *et al.*, 2004; Wullschleger *et al.*, 2002). The recent release of the reference genome of *Populus trichocarpa* (Tuskan *et al.*, 2006) has provided unprecedented opportunities to advance our understanding of *Populus* biology and subsequent interactions with associated organisms. This is evidenced by a drastic increase of refereed publications related to *Populus* in the past three years, which now triples that of its close relative, *Salix* (Yang *et al.*, 2009). Since the *Populus* reference genome was released, one of the important *Populus* pathogens, *Melampsora larici-populina* (Martin *et al.*, 2004) (http://genome.jgi-psf.org/Mellp1/Mellp1.home.html), has also been sequenced. The DOE Joint Genome Institute (JGI) has also sequenced the genome of a symbiotic fungal basidiomycete, *Laccaria bicolor*, (Martin *et al.*, 2008) that is able to engage a mutualistic ecto-mycorrhizal association with the roots of *Populus* (Kohler *et al.*, 2008). Likewise, many other *Populus*-associated pathogens and endophytes have recently been sequenced (http://genome.jgi-psf.org/stema/stema.home.html). Simultaneously, advanced "omics' tools have been developed, such as transcript profiling using high-throughput short-read sequencing and proteomic and metabolomic profiling via advance GS-MS. Thus, *Populus*, as a model, provides vast opportunities for studying *Populus*/biotic, *Populus*/abiotic, and *Populus*/climatic interactions, particularly at the molecular levels.

Beyond summarizing the relevant fields of research, our main intention with this special issue is to further stimulate thinking from the perspective of *Populus* as an entire living community, a community that contains a "mega-genome," from both abiotic and biotic aspects. We certainly hope that this special issue will inspire more research in this area, not only by those who have long been involved in *Populus* genomics, but also by those who have not.

STATE OF THE SCIENCE

In this special issue, we solicited seven reviews with a theme of *Populus* community genomics. These reviews have been grouped into three areas: an update on *Populus* genomics, *Populus* interactions with the biotic organisms, and *Populus* interactions with the environment.

The first review provides an overview of *Populus* genomics. Yang *et al.* (2009) offer a comprehensive review of the state-of-the-science in *Populus* experiment-based functional genomics and computational genomics. These authors also summarize applications of various classical and "omics" technologies toward advancing *Populus* genetics and functional genomics. A major portion of the review also covers current progress in sequence-based discovery, from predicting gene function to comparative analysis of gene families to development of genomic databases and studies of the evolutionary dynamics at both the gene and genome level.

The next four reviews offer detailed progress and associated challenges in understanding *Populus* interactions with biotic organisms: insects, pathogenic organisms, endophytes, and rhizospheric microorganisms. Ralph (2009) provides an overview

Address correspondence to Corresponding authors: zcheng@utk.edu, tuskanga@ornl.gov

of emerging research strategies designed to target defense genes that directly mediate insect resistance in *Populus*, including the use of transgenesis to functionally characterize candidate defense genes, identify novel defense mechanisms through mutant population screening, and linked genetic/genomic approaches to study changes in gene expression. Duplessis *et al.* (2009) assess *Populus*-pathogen interactions at the molecular level for both the host and the pathogen. The availability of the *Populus* genome sequence has allowed comparative analyses with other sequenced plant genomes in an effort to reveal gene families that play key roles in defense response. One of the striking findings is that the NBS-LRR resistance (*R*)-gene family has expanded in *Populus* compared with other plant genomes, including *R*-gene subfamilies not previously reported in plants. This article also provides details on interactions of *Populus* with *Melampsora*, a rust disease that causes considerable damage in *Populus* plantations. Studies in this pathosystem have been advanced by sequencing of the *Melampsora larici-populina* genome, facilitating the reciprocal study of host response to pathogen attack and pathogen's reaction to the host at comparable genomic levels. Transcript profiles derived from compatible (susceptible) and incompatible (specific host-resistance) *Populus-Melampsora* interactions demonstrate that defense responses in perennial *Populus* are similar to those of annual plant species such as up-regulation of transcripts encoding pathogenesis-related proteins.

In addition to the traditional host/pest or host/pathogen interactions, this special issue also reviews the state-of-the-science related to the interactions of *Populus* and its endophytes. Endophytes are microorganisms that create complex and long-lived associations within plant tissues. Unlike many pathogens, endophytes typically have a beneficial effect on their host. Despite this benefit, the role of endophytic bacteria in growth and development of their host plants remains unresolved. Van der Lelie *et al.* (2009) review the interaction between *Populus* species and their endophytic bacteria and point to potential breakthroughs in our understanding of improvements in the productivity of *Populus*. The authors summarize direct and indirect mechanisms of improving host plant growth and development via endophytic inoculations, including applications of custom endophyte-host partnerships in *Populus* for improving productivity and establishment of *Populus* on marginal soils and phytoremediation of contaminated soils and groundwater.

An even less studied field of *Populus* community genomics is *Populus* interactions with its associated soil microbiome. Podila (2009) provides new insights gleaned from genomic-level studies involving *Populus* and its soil community, especially mutualistic and symbiotic ectomycorrhizal interactions. Special focus is given to the communication and signaling that occurs in the soil between tree roots and mycorrhizal fungi and the effect that root exudates and fungal enzymes exert on the turnover and translocation of nitrogen, mineral nutrients and soil organic matter.

The final two reviews deal with *Populus* environmental interactions with Wullschleger *et al.* (2009) reviewing light, temperature and moisture responses and Chen *et al.* (2009) emphasizing metabolic responses to biotic-abiotic factors. Although research focused on understanding the molecular processes that underpin the *Populus* growth and development has steadily increased over the last several decades, the ability to examine the basic mechanisms whereby trees respond to a changing climate and resource limitations has benefitted greatly from the recent release of the reference genome of *P. trichocarpa*. Wullschleger *et al.* (2009) summarize the literature with focuses on integrating transcriptomic, proteomic, and metabolomic analyses related to *Populus'* response to its climatic and edaphic environment. These authors highlight instances where two or more omic-scale measurements confirm and expand our inferences about mechanisms contributing to observed patterns of response. Chen *et al.* (2009) cover the genomics-level secondary metabolism in *Populus* that is involved in interactions with biotic and abiotic environments. To ensure its survival and reproduction in complex biotic and abiotic community environment, *Populus* produces a myriad of secondary metabolites as adaptation mechanisms. These authors also review how these compounds relate to certain biological/ecological processes such as defense against insects and microbial pathogens or adaptation to abiotic stresses.

CHALLENGES AND FUTURE ADVANCES

The landmark event of the release of the reference genome of *P. trichocarpa* along with the genomic sequences of some of its biotic associates has greatly elevated research activities in *Populus* and its associated communities. *Populus* biologists are now able to take a systems approach to quantifying the diversity of genes, proteins and metabolites that govern the growth and development of *Populus*. While substantial advancements have been achieved in *Populus* genome-based science, many challenges remain. To meet these challenges, a coordinated effort amongst the diverse set of research communities will be required.

The interactions between *Populus* and other biotic organisms and/or the abiotic environment are complex, involving multifaceted signal transduction networks and many genes expressed in spatial-temporal manners. As more "omics" tools become available and the costs come down, massive amount of "omics" data (e.g., transcriptomics, proteomics, metabolomics and phenomics) will be generated in an effort to deconvolute the processes and links between phenotypic responses with gene expression. Concomitantly, high-performance computational pipelines will be needed in order to handle data generation, storage and analysis. This vast amount data will need to be integrated into functional systems biology platforms which are designed to reveal the underlying mechanisms controlling *Populus'* response to biotic organisms and/or abiotic conditions. Based on the reviews in this special issue, it is anticipated that a network of interactomes will be produced in the near future.

Finally, *Populus* is considered a woody bioenergy crop and is one of the two primary bioenergy crops studied within the

three DOE bioenergy science centers (BioEnergy Science Center, http://bioenergycenter.org/). *Populus* also plays a critical role in maintaining healthy natural and managed ecosystems, contributing to carbon sequestration (Tuskan and Walsh, 2001), and providing raw materials for pulping, paper and other forest products (Tuskan, 1998). As such, there is one outcome we can surely predict, i.e., *Populus* will continue to gain importance in production forestry, environmental stewardship and renewable energy production. This popularity will demand more of systems-biology-based research, embracing both functional genomics and community mega-genomics. We anticipate that *Populus* and its associated mega-genome will continue to be a model for the study of other forest tree species as well as a new model for the study of host microbe interactions among all plant species.

ACKNOWLEDGMENTS

We would like to thank the Editors-in-Chief, Drs. Gray and Trigiano for giving us an opportunity to organize this combined special issue. We like to thank the authors of each chapter for contributing to the reviews. Special thanks goes to J. C. Tuskan for discussion on community-based genomics. Research in Cheng's lab is supported by the Department of Energy/Consortium for Plant Biotechnology Research, Inc., by the Department of Energy's BioEnergy Science Center, and by the Tennessee Agricultural Experiment Station. The research in Tuskan's lab is funded by the U.S. Department of Energy, Office of Science, Bioenergy Science Center. ORNL is managed by UT-Battelle, LLC for the U.S. Department of Energy under contract no. DE-AC05-00OR22725.

REFERENCES

Bradshaw, H. D., Ceulemans, R., Davis, J., and Stettler, R. 2000. Emerging model systems in plant biology: Poplar (Populus) as a model forest tree. *J. Plant Growth Reg.* **19**: 306–313.

Chen, F., Liu, C.-J., Tschaplinski, T. J., and Zhao, N.. 2009. Genomics of secondary metabolism in Populus: Interactions with biotic and abiotic environments. *Crit. Rev.Plant Sci.* **28**: xx–yy.

Duplessis, S., Major, I., Martin, F., and Séguin, A. 2009. Poplar and pathogen interactions: Insights from Populus genome-wide analyses of resistance and defense gene families and gene expression profiling. *Crit. Rev. Plant Sc.* **28**: xx–yy.

Groover, A. T. 2005. What genes make a tree a tree? *Trends Plant Sci.* **10**: 210–214.

Jansson, S., and Douglas, C. J. 2007. Populus: A model system for plant biology. *Ann. Rev. Plant Biol.* **58**: 435–458.

Kohler, A., Rinaldi, C., Duplessis, S., Baucher, M., Geelen, D., Duchaussoy, F., Meyers, B. C., Boerjan, W., and Martin,F. 2008. Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mo.l Bio.l* **66**: 619–636.

Martin, F., Tuskan, G. A., DiFazio, S. P., Lammers, P., Newcombe, G., and Podila, G. K. 2004. Symbiotic sequencing for the *Populus* mesocosm. *New Phytologist* **161**: 330–335.

Podila, G. K., Sreedasyam, A., and Muratet, M. A. 2009. *Populus* Rhizosphere and the Ectomycorrhizal interactome. *Crit. Re. Plant Sci.* **28**: xx–yy.

Ralph, S. G. 2009. Studying Populus defenses against insect herbivores in the post-genomic era. *Crit. Rev. Plant Sci.* **28**: xx–yy.

Tuskan, G. A. 1998. Short-rotation woody crop supply systems in the United States: What do we know and what do we need to know? *Biomass & Bioenergy* **14**: 307–315.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G. L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, Davis, R., Degroeve, J., S., Dejardin, Depamphilis, A., C., Detter, Dirks, J., B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, Jones, N., S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J. C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D. R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C. J., Uberbacher, E., Unneberg, P., *et al.* 2006. The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313**: 1596–1604.

Tuskan, G. A., DiFazio, S. P., and Teichmann, T. 2004. Poplar genomics is getting popular: The impact of the poplar genome project on tree research. *Plant Biol.* **6**: 2–4.

Tuskan, G. A., and Walsh, M. E. 2001. Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: A US case study. *Forestry Chronicle* **77**: 259–264.

van der Lelie, D., Taghavi, S., Monchy, S., Schwender, J., Miller, L., Ferrieri, R., Rogers, A., Wu, X., Zhu, W., Weyens, N., Vangronsveld, J., and Newman, L. 2009. Poplar and its bacterial endophytes: coexistence and harmony. *Crit. Rev. Plant Sci.***28**: xx–yy.

Wullschleger, S. D., Jansson, S., and Taylor, G. 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* **14**: 2651–2655.

Wullschleger, S. D., Weston, D. J., and Davis, J. M. 2009. *Populus* Responses to Edaphic and Climatic Cues: Emerging Evidence from Systems Biology Research. *Crit.l Rev. Plant Sci.* **28**: xx–yy.

Yang, X., Kalluri, U. C., DiFazio, S. P., Wullschleger, S. D., Tschaplinski, T. J., M., Cheng, Z.-M., and Tuskan, G. A. 2009. Poplar genomics: State of the science. *Crit. Rev. Plant Sci.* **28**: xx–yy.

Original Article

Cytogenetic and Genome Research

Cytogenet Genome Res 2009;125:74–80
DOI: 10.1159/000218749

Accepted after revision: November 25, 2008
by B. Friebe

Cytogenetic Analysis of *Populus trichocarpa* – Ribosomal DNA, Telomere Repeat Sequence, and Marker-selected BACs

M.N. Islam-Faridi[a] C.D. Nelson[b] S.P. DiFazio[c] L.E. Gunter[d] G.A. Tuskan[d]

[a]Southern Institute of Forest Genetics, Southern Research Station, U.S. Forest Service, Forest Tree Molecular Cytogenetics Laboratory, Department of Ecosystem Science and Management, Texas A&M University, College Station, Tex., [b]Southern Institute of Forest Genetics, Southern Research Station, U.S. Forest Service, Harrison Experimental Forest, Saucier, Miss., [c]Department of Biology, West Virginia University, Morgantown, W.Va., [d]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tenn., USA

Key Words
BAC · FISH · *Populus trichocarpa* · rDNA · Telomere

Abstract
The 18S-28S rDNA and 5S rDNA loci in *Populus trichocarpa* were localized using fluorescent in situ hybridization (FISH). Two 18S-28S rDNA sites and one 5S rDNA site were identified and located at the ends of 3 different chromosomes. FISH signals from the *Arabidopsis*-type telomere repeat sequence were observed at the distal ends of each chromosome. Six BAC clones selected from 2 linkage groups based on genome sequence assembly (LG-I and LG-VI) were localized on 2 chromosomes, as expected. BACs from LG-I hybridized to the longest chromosome in the complement. All BAC positions were found to be concordant with sequence assembly positions. BAC-FISH will be useful for delineating each of the *Populus trichocarpa* chromosomes and improving the sequence assembly of this model angiosperm tree species.

The genus *Populus* consists of many species that are ecologically and economically important across the Northern Hemisphere. Poplars (*Populus* spp.) are early successional, fast growing species that can be cultivated for high wood yields [Stettler et al., 1996]. The wood has many uses, most notably as fiber for paper products but more recently as biomass for bioenergy feedstocks [Rubin, 2008]. In addition, their fast growth rate and globally wide distribution makes poplars valuable in sequestering carbon. Therefore, wider cultivation of these species may benefit the carbon balance and assist in moderating global temperature change [Lemus and Lal, 2005]. The poplars are considered to be model forest tree species for genetics and genomics research in part because of their relatively small genome size (480 Mb/1C, 2n = 2x = 38), amenability to tissue culture and genetic transformation, relative ease of controlled-pollination and high seed set, and short generation time (<5 years). In particular this distinction has fallen on *Populus trichocarpa* (black cottonwood) as it was the 3rd plant species (after *Arabidopsis* and rice) and 1st tree species to have its genome sequenced [Tuskan et al., 2006].

Prior to the genome sequence, a number of genetic linkage (i.e., recombination-based) maps were developed

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

1424–8581/09/1251–0074$26.00/0

Accessible online at:
www.karger.com/cgr

Nurul Faridi
Forest Science Laboratory, Texas A&M University
2585-TAMU
College Station, TX, 77843-2585 (USA)
Tel. +1 979 862 3908, Fax +1 979 845 3272, E-Mail nfaridi@tamu.edu

and reported for poplars [e.g., Bradshaw et al., 1994; Grattapaglia and Sederoff, 1994; Bradshaw, 1996; Cervera et al., 1996, 2001; Yin et al., 2004]. However, little work has been completed at the cytogenetic level, especially when compared to other plant models such as *Arabidopsis* [Jackson et al., 1998; de Jong et al., 1999; Fransz et al., 2000, 2002], rice [Cheng et al., 2001, 2002; Zhao et al., 2002; Cheng et al., 2005; Tang et al., 2007], and sorghum [Islam-Faridi et al., 2002; Kim et al., 2005a, b]. Cytogenetic analysis in these plant species has been useful in aiding genome assembly by facilitating integration of linkage maps with physical maps and in determining the order of sequence scaffolds and their respective contigs [Zhao et al., 2002; Cheng et al., 2005; Kim et al., 2005a]. This has provided a more complete understanding of the structural and functional properties of these genomes.

Detailed cytogenetic analysis of poplar chromosomes should help improve the genome assembly, which currently consists of 2,447 major scaffolds >5 kb [Tuskan et al., 2006]. In this paper, we utilized fluorescent in situ hybridization (FISH) to study and localize the 18S-28S and 5S rDNA and the *Arabidopsis*-type telomere repeat sequence (ATRS) sites in *P. trichocarpa*. In addition, we used FISH to localize 6 marker-selected BAC clones that represent 2 linkage groups of *P. trichocarpa*. Our specific objectives were to 1) determine the number and location of the rDNA sites, 2) determine the distribution of the ATRS sites, and 3) study the feasibility of utilizing BAC-FISH in *Populus* for chromosome localization. The 3rd objective is important for furthering our work in establishing a cytomolecular map for each poplar chromosome and to better understand the structural details of the *Populus* genome.

Materials and Methods

Chromosome Preparation

Chromosome spreads were prepared for *P. trichocarpa* clone Nisqually-1 (383–2499), the same genotype that was used for whole genome sequencing [Tuskan et al., 2006]. Actively growing root tips about 1.5 cm long were harvested from rooted cuttings growing in potting soil in a greenhouse. Harvested root tips were immediately pre-treated with an aqueous solution of α-monobromonaphthalene (0.8%, Sigma) for 1.5 to 1.75 h at room temperature (RT) in the dark and then fixed in 4:1 ethanol:glacial acetic acid to arrest cell division at metaphase. Fixed root tips were processed enzymatically (5% cellulase Onozuka R-10, Yakult Honsha Co. Ltd., and 1.25% pectolyase Y-23, Kyowa Chemical Products Co. Ltd.) in 0.01 M citrate buffer, and the chromosome spreads were prepared as described elsewhere [Jewell and Islam-Faridi, 1994]. The chromosome spreads were checked with a phase contrast microscope (Axioskop, Carl Zeiss, Inc.), and slides containing good chromosome spreads were selected and stored at –80°C for use in FISH.

Probe DNA

The following probes were used in the current experiments: 18S-28S rDNA of maize [Zimmer et al., 1988], 5S rDNA including a spacer region of sugar beet [Schmidt et al., 1994], *Arabidopsis*-type telomere repeat sequence $(TTTAGGG)_n$ (kindly provided by Dr. T. McKnight, Department of Biology, Texas A&M University), and BAC clones from 2 *P. trichocarpa* linkage groups LG-I and LG-VI.

The BAC clones were derived from a library prepared from Nisqually-1 [Stirling et al., 2001]. Positions of the BACs were determined by alignment of end sequences to the genetic map-anchored whole genome sequence assembly. The repeat content of each BAC clone was inferred based on the frequency of constituent 16mers in the full set of 7.5 million sequence reads from the *Populus* genome sequencing project and the abundance of protein coding sequences contained in the BAC. The selected BACs were 66B19, 75P22, and 87F21 from LG-I and 78O18, 88A10, and 93N12 from LG-VI (http://www.bcgsc.ca/platform/mapping/data/?searchterm=poplar).

BAC DNA was isolated from overnight liquid cultures in selective media by alkaline lysis, digested with *Eco*RI, and followed by further purification using Plant DNeasy spin columns (QIAGEN, Valencia, CA) as described elsewhere [Childs et al., 2001]. Probe DNAs with and without whole plasmids were labeled with biotin-16-dUTP (Biotin-Nick Translation Mix, Roche Diagnostics) and/or digoxigenein-11-dUTP (dig) (Dig-Nick Translation Mix, Roche Diagnostics) by nick translation as recommended by the manufacturer. Labeled probes were dot-blotted to verify incorporation of label.

Fluorescent in situ Hybridization (FISH)

The hybridization mixture consisted of deionized formamide (50%, Fisher Chemical), dextran sulfate (10%, Fisher Chemical), 2× SSC, labeled probe DNA (30 ng/slide), carrier DNA (*E. coli* DNA, 7.5 μg/slide), and blocking DNA (poplar Nisqually-1 Cot-1 DNA, 300 to 600 ng/slide, depending on single- or dual-color FISH, respectively). The poplar Cot-1 DNA was prepared as described by Zwick et al. [1997] and was used only for FISH with BAC probes (BAC-FISH). The BAC-FISH hybridization mixture was denatured in a boiling water bath for 10 min, immediately placed on ice for 5 to 10 min, and then incubated at 37°C for 30 min to allow the Cot-1 DNA to hybridize with the repetitive DNA sequences of the BAC probes. Chromosome spreads were denatured in 70% deionized formamide at 72°C for 1.5 min in an oven on a metal block followed by dehydration through a series of ethanol (70, 85, 95, and 100%) washes at –20°C for 3 min each. The slides were dried with forced-air and on the bench at RT for 25 to 30 min prior to hybridization with probe DNA.

The hybridizations were accomplished by loading 25 μl of hybridization mixture on the chromosome spread and placing and sealing a 22 × 30 mm glass coverslip over the mixture. The coverslips were sealed with rubber cement, and the slides were placed in a humidity chamber and incubated over night at 37°C. Following hybridization, the coverslips were washed off with 2×

SSC, 37°C, using a squeeze bottle. The slides were then washed twice in 2× SSC, 30% deionized formamide, and 2× SSC for 5 min each at 40°C, followed by washing twice in 2× SSC and once in 4× SSC, 0.2% Tween 20 for 5 min each at RT. Finally, the slides were incubated in 200 μl blocking solution consisting of 5% bovine serum albumin (BSA, IgG-free, protease-free; Jackson ImmunoResearch Laboratories, Inc.) in 4× SSC, 0.2% Tween 20.

The hybridization sites were detected with 5 μg/ml of fluorescein (FITC)-conjugated anti-digoxigenin (Roche Diagnostics), or 0.75 μg/ml Cy3-conjugated streptavidin (Jackson ImmunoResearch Laboratories, Inc.), or both depending on the labeled probe DNA used in the hybridization mixture. All detection reactions were completed at 37°C for 25 to 30 min using the blocking solution. The slides were then washed 4 times in 4× SSC/0.2% Tween 20 for 5 min each at 37°C to remove the unbound antibodies, counterstained with 4 μg/ml 4′, 6-diamidino-2-phenylindole dihydrochloride (DAPI, Sigma) in McIlvaine buffer, pH 7.0 (9 mM citric acid, 80 mM Na_2HPO_4, 2.5 mM $MgCl_2$) for 10 to 12 min, and briefly washed in 4× SSC, 0.2% Tween 20 followed by washing in 2× SSC. The slides were dried with forced-air, and a small drop (~6 μl) of anti-fade solution (Vectashield, Vector Laboratories) was added under a glass coverslip (22 × 50 mm). Slides were incubated overnight at 4°C to stabilize the fluorochromes before observation under an epifluorescence microscope. All detection reactions and treatments were conducted in subdued light.

The FISH experiments using 18S-28S and 5S rDNA and *Arabidopsis*-type telomere repeat sequence (ATRS) probes were carried out in 3 phases. First, we used a dig-labeled 18S-28S rDNA probe and detected the hybridization sites with FITC-conjugated anti-dig to provide green signals. Second, we used biotin-labeled 18S-28S rDNA and dig-labeled 5S rDNA probes simultaneously. The 18S-28S rDNA probe was detected with Cy3-conjugated streptavidin to provide red signals, and the 5S rDNA probe was detected with FITC-conjugated anti-dig to provide green signals. Third, we used dig-labeled 18S-28S rDNA and biotin-labeled ATRS probes, respectively, with signal detection as described above.

The BAC-FISH experiments for LG-I and LG-VI were also carried out in 3 phases. First, we used 2-color FISH with 2 BACs, where BAC 1 (to simplify here we code the BACs 1 to 3) was labeled with biotin and detected with Cy3-conjugated streptavidin (red) and BAC 2 was labeled with dig and detected with FITC-conjugated anti-dig (green). Second, we used 2-color FISH with BACs 2 and 3 (labeling and detection as before for BACs 1 and 2). Third, we hybridized all 6 BACs pooled together, 3 from each LG, followed by simultaneous detection of both colors.

Digital Image Capture and Process

Digital images were recorded from an Olympus AX-70 epifluorescence microscope with suitable filter sets (Chroma Technology) using a 1.3 MP Sensys (Roper Scientific) camera and a MacProbe v4.2.3 digital image system (Applied Imaging International) and then further processed with Adobe Photoshop CS v8 (Adobe Systems).

Results

The chromosome spread technique of Jewell and Islam-Faridi [1994] with air-drying routinely provided unifocal views of the *Populus* chromosomes. These chromosome spreads were mostly free of cell walls, nuclear membranes, and cytoplasmic debris providing good accessibility for the probes and low background hybridization.

18S-28S rDNA and 5S rDNA Sites

Two major 18S-28S rDNA sites were clearly identified in *P. trichocarpa* and they are located at the ends of 2 different chromosomes (figs. 1, 2). This is in contrast to results inferred from interphase to prophase stages where numerous 18S-28S rDNA FISH signals were observed (figs. 1c–e, 2c, d). A single 5S rDNA site was found in *P. trichocarpa* and it is likewise located at the end of a 3rd chromosome (green signals, arrows, fig. 2a, b).

Arabidopsis-*type Telomere Repeat Sequence (ATRS)*

The ends of each chromosome of *P. trichocarpa* showed ATRS signals (fig. 2c), while interphase nuclei suggested numerous ATRS signals scattered throughout the chromosomal complement (fig. 2d).

LG-I and LG-VI BACs

Two-color FISH with 2 BAC clones of both LG-I and LG-VI allowed an evaluation of the relative copy number of the probes, i.e., whether either combination of the BAC clones contains unique/low copy or high copy DNA sequences. BAC 87F21 (red signal, fig. 3a) and BAC 66B19 (green signal, fig. 3a) were observed to be located on the same chromosome as expected, and it appeared to be the longest chromosome in the complement. Pairs of red and green signals were also observed in interphase nuclei (fig. 3b). BAC 87F21 was located in a low DAPI stained (euchromatic) region at the end of the long arm while BAC 66B19 was proximally located in a high DAPI stained (heterochromatic) region. All but one BAC (66B19) showed unique hybridization signals at their respective chromosomal locations. Interestingly, BAC 66B19 showed a strong hybridization signal near the putative centromere (i.e., apparent primary constriction as observed under the microscope) of the longest chromosome along with some scattered FISH signals on all of the other chromosomes, which indicates that 66B19 contains some degree of repetitive DNA. For the LG-I chromosome, BACs 87F21 and 66B19 appear to be located on the long arm, and BAC 75P22 appears to be located on the short arm

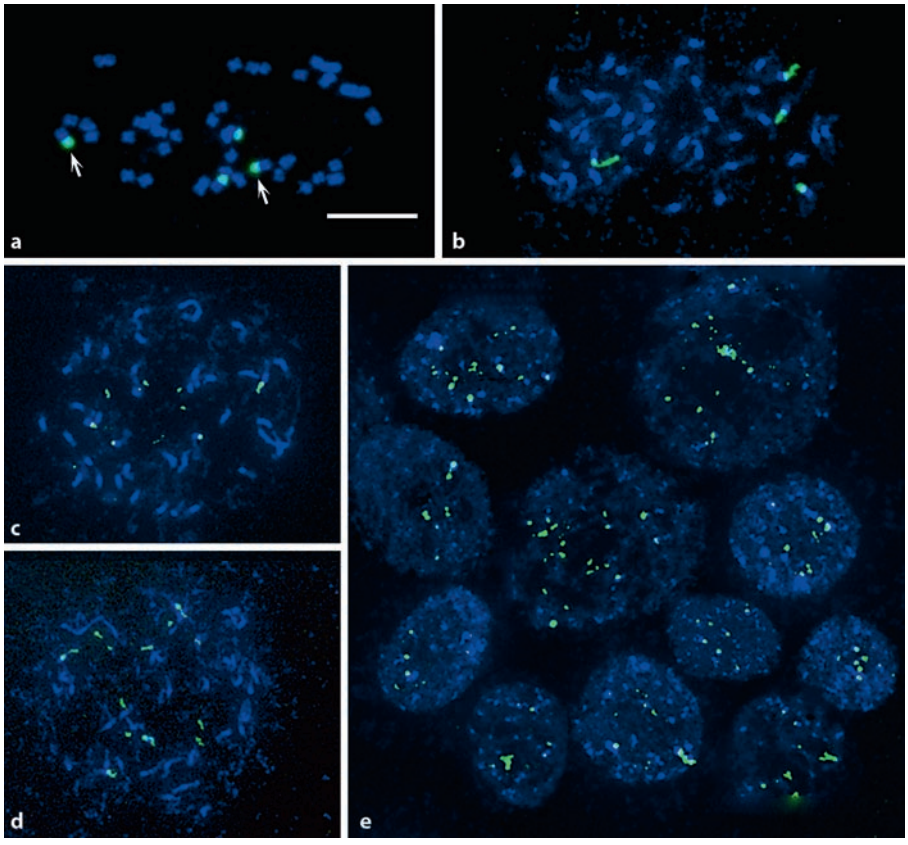


Fig. 1. FISH results using 18S-28S rDNA probe on mitotic chromosome spreads and intact nuclei in *Populus trichocarpa*. **a** Metaphase chromosome spread, arrows point to signals that are slightly stronger than the others. **b** Late prophase chromosome spread. **c** Mid-prophase chromosome spread. **d** Early prophase nuclei. **e** Interphase nuclei. Bar = 10 μm.

(fig. 3a, c, e). For the LG-VI chromosome, BACs 78O18 and 93N12 are located on one arm, and BAC 88A10 is located interstitially on the other arm (fig. 3d, f).

Discussion

Our 18S-28S rDNA FISH results support the findings of Faivre-Rampant et al. [1992] based on Southern hybridization, where they inferred the presence of two 18S-25S rDNA loci in *P. trichocarpa*. Using FISH, we observed 2 major 18S-28S rDNA sites in *P. trichocarpa*, located on 2 different chromosomes. In addition, we found one site to be slightly stronger in FISH signal intensity than the other site (arrows, fig. 1a). This difference is most likely due to variation in copy number of rDNA repeat units located at the 2 sites as suggested by Faivre-Rampant et al. [1992]. They analyzed restriction fragments of the 18S-25S rDNA loci in *P. trichocarpa* and found the 11.7-kb *Eco*RV fragment to be present in higher frequency than the 11.4-kb fragment, suggesting the presence of some minor loci. In contrast, our FISH results showed no evidence of minor 18S-28S rDNA loci in *P. trichocarpa*.

The 18S-28S rDNA loci were located at the end of each of 2 homologous pairs of chromosomes (figs. 1a, b, 2). In addition, a red Cy3 ATRS signal was observed directly distal of each of the 18S-28S rDNA signals. As expected, the telomeric end of each chromosome arm showed a pair of ATRS signals [Fuchs et al., 1995]. Furthermore, the 18S-28S rDNA loci are located in satellited regions of both chromosomes, as shown by FISH with the ATRS probe and DAPI staining (fig. 2c, d). In contrast, only one satellited chromosome was reported in *P. alba*, *P. balsamifera*, *P. deltoides*, *P. euroamericana*, and *P. nigra*, while one to three 18S-28S rDNA sites were reported for these species [Prado et al., 1996].

Numerous 18S-28S rDNA FISH signals (sometimes more than 20) were observed in interphase nuclei, and more than 4 signals were observed in early-prophase to mid-prophase. These variations of signal numbers are most likely due to the highly decondensed nature of DNA in interphase nuclei and to a lesser degree in prophase. Collecting data at these decondensed chromatin stages

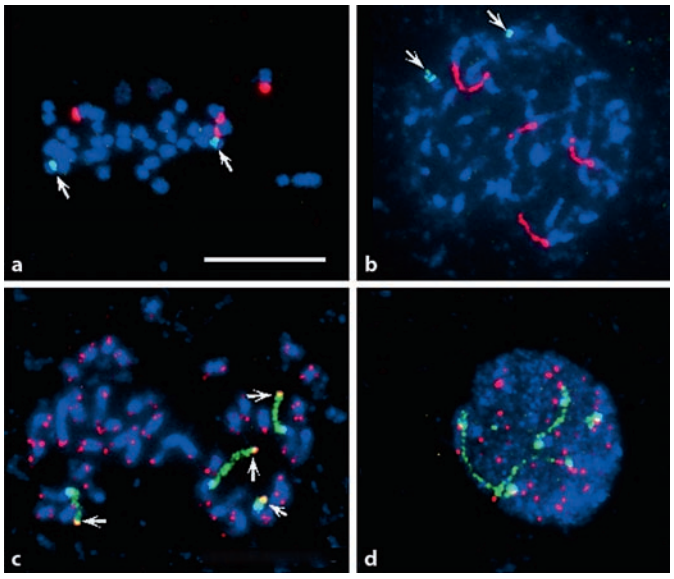


Fig. 2. FISH results using 18S-28S rDNA, 5S rDNA, and ATRS probes on somatic chromosome spreads and an interphase nucleus in *Populus trichocarpa*. **a** Metaphase chromosome spread with 18S-28S rDNA signals (red) and 5S rDNA signals (green, arrows). **b** Mid-prophase chromosome spread with 18S-28S rDNA signals (red) and 5S rDNA signals (green, arrows). **c** Prometaphase chromosome spread with 18S-28S rDNA signals (green), ATRS signals (red), and overlapping probes (arrow pointing to yellowish color). **d** Interphase nucleus with 18S-28S rDNA signals (green) and telomere repeat DNA sequence signals (red). Bar = 10 μm.

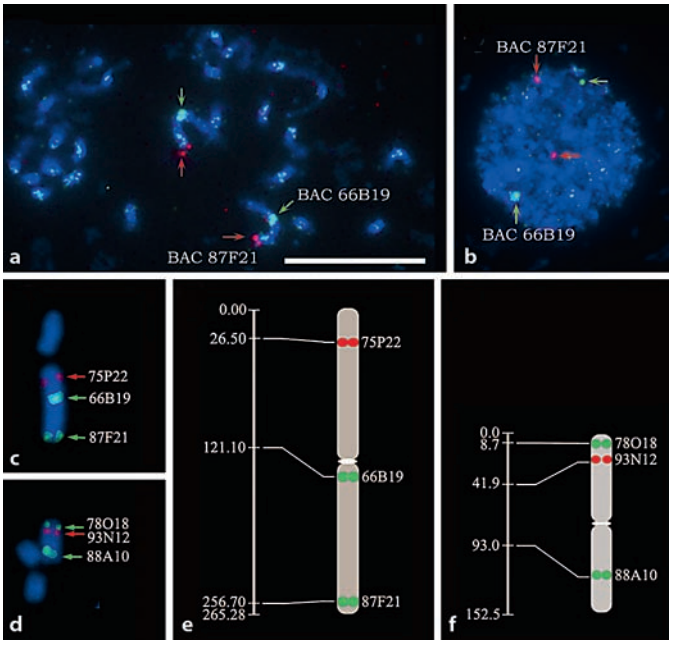


Fig. 3. Linkage group specific BAC FISH and diagrammatic representations of BAC map positions on LG-I and LG-VI in *Populus trichocarpa*. **a** LG-I BAC FISH signals (BAC 87F21 in red and BAC 66B19 in green) on chromosome complement. **b** Interphase nucleus showing red and green signals from the same BACs as in **a**. **c** LG-I BAC FISH signals (red and green) on a specific chromosome. **d** Three LG-VI BAC FISH signals (red and green) on a specific chromosome. **e** Diagrammatic representation of LG-I BAC positions on chromosomes along with linkage map. **f** Diagrammatic representation of LG-VI BAC positions along with linkage map. Primary constrictions (i.e., putative centromeres) are represented by white ellipses towards the center of the chromosomes (**e** and **f**). Bar = 10 μm.

provides an upwardly biased count of the number of tandemly repeated DNA loci such as 18S-28S rDNA. Because of this, interphase nuclei should not be used to determine the number of repetitive DNA sites such as 18S-28S rDNA. When chromatin is not fully condensed, these FISH signals, depending on the degree of condensation, can be observed like a series of dots along a filament over these loci (green signals, fig. 2d). These signals become increasingly dense when the nuclei advance from interphase to prophase (red signals, fig. 2b; green signals, fig. 2c) as the chromatin condensation process reaches a maximum level at metaphase, showing one signal (figs. 1a, 2a). Similar patterns of FISH signals also have been reported for *Brassica* [Maluszynska and Heslop-Harrison, 1993], alfalfa [Calderini et al., 1996], lentil, and peanut [Islam-Faridi, unpublished data].

Only one 5S rDNA locus was identified in *P. trichocarpa*, and it was located using two-color FISH on a chromosome different to either of the two 18S-28S rDNA loci bearing chromosomes (arrows, fig. 2a, b). In contrast, Prado et al. [1996] reported that there were two 5S rDNA loci in each of 5 diploid species of *Populus* and that they were located on 2 different chromosomes. However, they could not confirm whether the 5S rDNA and 18S-28S rDNA loci were on the same or different chromosomes. We observed that the 5S rDNA FISH signal in *P. trichocarpa* is much reduced compared to either of the 18S-28S rDNA signals, apparently because the 5S rDNA locus contains fewer repetitive units than either of the 18S-28S rDNA loci. Similar results were also reported for *P. nigra* [Ribeiro et al., 2008].

The BACs were selected from LG-I and LG-VI based on low or single copy BLAST hits against the whole-genome sequence assembly. We tested these BACs as a pilot

study to determine whether or not BAC-FISH could be useful in developing a BAC-based cytomolecular map in *Populus* as it has been for *Arabidopsis* [Fransz et al., 2000, 2002], sorghum [Islam-Fraidi et al., 2002; Kim et al., 2005a, b], and rice [Cheng et al., 2001, 2002; Zhao et al., 2002; Cheng et al., 2005; Tang et al., 2007]. All but one of the BAC clones (66B19) were observed to be located in euchromatic regions of the 2 chromosomes, and each showed almost no or very low background FISH signal. These results demonstrated that 5 of the 6 tested BAC clones (75P22, 87F21, 78O18, 93N12, and 88A10) contained low repetitive DNA compared with BAC clone 66B19. On the other hand, BAC clone 66B19 was found to be located in a heterochromatic region as determined by its co-location with a highly DAPI stained region. Since BAC clone 66B19 apparently originated from a heterochromatic region, it seems likely that it may contain some repetitive DNA, and the observed, dispersed FISH signals across most if not all chromosomes support this assertion. It is unclear why the repetitive nature of this BAC was not apparent from the frequency of 16mers in the assembled genome sequence. However, repetitive regions are notoriously difficult to assemble using a whole genome shotgun approach [Green, 2001], so some highly repetitive regions are probably not included in the assembled sequence, including the region spanned by BAC clone 66B19.

FISH with the LG-I BAC clones revealed the order of the BAC clones as telomere (of one chromosome arm) → BAC 75P22 → BAC 66B19 → BAC 87F21 → telomere of the other arm. For the BAC clones selected from LG-VI, the order observed was: telomere of one arm → BAC 78O18 → BAC 93N12 → BAC 88A10 → telomere of the other arm. The chromosomal positions of the 6 BACs clones appear to be syntenic to their inferred positions in the sequence-based assembly, but the relative positions (i.e., physical vs. genetic) vary, especially for the 3 BAC clones 66B19 (LG-I, fig. 3e), 93N12, and 88A10 (LG-VI, fig. 3f). It is not surprising that there is a discrepancy between BAC positions inferred by FISH and those inferred from the sequence assembly. The genome sequence contains many gaps, and the genetic map has relatively low resolution, being based on genotypes of only 44 progeny in many cases [Tuskan et al., 2006]. This highlights the importance of FISH for determining physical distances between markers in the *Populus* genome.

Using linkage group specific BAC clones as FISH probes, we found that the longest chromosome is associated with BAC clones from LG-I. Given these results, we propose using LG-specific BACs to identify and enumerate the poplar chromosomes. Considering this standard, we have studied the location of 3 BAC clones from each of chromosomes 1 and 6. We also found that 3 chromosomes contain rDNA loci, two 18S-28S loci and one 5S locus, and that none of these chromosomes are either chromosomes 1 or 6. Additional BAC-FISH will be needed to determine the chromosomes that contain the rDNA loci. In addition, a centromere-specific FISH probe will be needed to define the basic karyotype of *P. trichocarpa* and the other poplars. These experiments should lead to the development of a comprehensive cytomolecular map that is anchored to the genetic linkage map, the BAC-based physical map, and the whole-genome sequence. Furthermore the development and availability of a comprehensive cytomolecular map should help in closing and resolving the remaining gaps and issues in the whole-genome sequence assembly.

Acknowledgements

Funding for this research was provided by the U.S. Department of Energy, Office of Science, Basic Energy Sciences Program and by the NSF Plant Genome Program Project 0421743. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725. Additional funding was provided by Southern Research Station (U.S. Forest Service). We thank Dr. Max Cheng (University of Tennessee) for plant materials, and Dr. Thomas Byram (Texas Forest Service) and Dr. David Stelly (Texas A&M University) for greenhouse and microscopy facilities, respectively.

References

Bradshaw HD Jr: Molecular genetics of *Populus,* in Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM (eds): Biology of *Populus* and its Implications for Management and Conservation, pp 183–199 (NRC Press, Ottawa 1996).

Bradshaw HD Jr, Villar M, Watson BD, Otto KG, Stewart S, Stettler RF: Molecular genetics of growth and development in *Populus*. 3. A genetic linkage map of a hybrid poplar composed of RFLP, STS, and RAPD markers. Theor Appl Genet 89:167–178 (1994).

Calderini O, Pupilli F, Cluster PD, Mariani A, Arcioni S: Cytological studies of the nucleolus organizing regions in the *Medicago* complex: *sativa-coerulea-falcata*. Genome 39: 914–920 (1996).

Cervera MT, Gusmão J, Steenackers M, Peleman J, Storme V, et al: Identification of AFLP molecular markers for resistance against *Melampsora larici-populina* in *Populus*. Theor Appl Genet 93:733–737 (1996).

Cervera MT, Storme V, Ivens B, Gusmao J, Liu BH, et al: Dense genetic linkage maps of three *Populus* species *(Populus deltoides, P. nigra* and *P. trichocarpa)* based on AFLP and microsatellite markers. Genetics 158:787–809 (2001).

Cheng CH, Chung MC, Liu SM, Chen SK, Kao FY, et al: A fine physical map of the rice chromosome 5. Mol Gen Genomics 274:337–345 (2005).

Cheng Z, Presting GG, Buell CR, Wing RA, Jiang J: High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. Genetics 157:1749–1757 (2001).

Cheng Z, Buell C, Wing R, Jiang J: Resolution of fluorescence in situ hybridization mapping on rice mitotic prometaphase chromosomes, meiotic pachytene chromosomes and extended DNA fibers. Chromosome Res 10: 379–387 (2002).

Childs KL, Klein RR, Klein PR, Morishige DT, Mullet JE: Mapping genes on an integrated sorghum genetic and physical map using cDNA selected technology. Plant J 27:243–255 (2001).

de Jong JH, Fransz P, Zabel P: High resolution FISH in plants – techniques and applications. Trends Plant Sci 4:258–263 (1999).

Faivre-Rampant P, Jeandroz S, Lefevre F, Lemoine M, Villar M, Berville A: Ribosomal DNA studies in poplars: *Populus deltoides*, *P. nigra*, *P. trichocarpa*, *P. maximowiczii* and *P. alba*. Genome 35:733–740 (1992).

Fransz PF, Armstrong S, Hans de Jong J, Parnell LD, van Drunen C, et al: Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of heterochromatic knob and centromere region. Cell 100: 367–376 (2000).

Fransz PF, Hans de Jong J, Lysak M, Castiglione MR, Schubert I: Interphase chromosomes in *Arabidopsis* are organized as well defined chromocenters from which euchromatin loops emanate. Proc Natl Acad Sci USA 99: 14584–14589 (2002).

Fuchs J, Brandes A, Schubert I: Telomere sequence localization and karyotype evolution in higher plants. Plant Syst Evol 196:227–241 (1995).

Grattapaglia D, Sederoff RR: Genetic-linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudotestcross-mapping strategy and RAPD markers. Genetics 137: 1121–1137 (1994).

Green ED: Strategies for the systematic sequencing of complex genomes. Nat Rev Genet 2: 573–583 (2001).

Islam-Faridi MN, Childs KL, Klein PE, Hodnett G, Menz MA, et al: A molecular cytogenetic map of sorghum chromosome 1: fluorescence in situ hybridization analysis with mapped bacterial artificial chromosomes. Genetics 161:345–353 (2002).

Jackson SA, Wang ML, Goodman HM, Jiang J: Application of fiber-FISH in physical mapping of *Arabidopsis thaliana*. Genome 41: 566–572 (1998).

Jewell DC, Islam-Faridi MN: Details of a technique for somatic chromosome preparation and C-banding of maize, in Freeling M, Walbot V (eds): The Maize Handbook, pp 484–493 (Springer, New York 1994).

Kim JS, Klein PE, Klein RR, Price HJ, Mullet JE, Stelly DM: Molecular cytogenetic maps of sorghum linkage groups 2 and 8. Genetics 169:955–965 (2005a).

Kim JS, Islam-Faridi MN, Klein PE, Stelly DM, Price HJ, et al: Comprehensive molecular cytogenetics analysis of sorghum genome architecture: distribution of euchromatin, heterochromatin, genes and recombination in comparison to rice. Genetics 171:1963–1976 (2005b).

Lemus R, Lal R: Bioenergy crops and carbon sequestration. Critical Rev Plant Sci 24:1–21 (2005).

Maluszynska J, Heslop-Harrison JS: Physical mapping of rDNA loci in *Brassica* species. Genome 36:774–781 (1993).

Prado EA, Faivre-Rampant P, Schneider C, Darmency MA: Detection of variable number of ribosomal DNA loci by fluorescent in situ hybridization in *Populus* species. Genome 39:1020–1026 (1996).

Ribeiro T, Barão A, Viegas W, Morais-Cecílio L: Molecular cytogenetics of forest trees. Cytogenet Genome Res 120:220–227 (2008).

Rubin EM: Genomics of cellulosic biofuels. Nature 454:841–845 (2008).

Schmidt T, Schwarzacher T, Heslop-Harrison JS: Physical mapping of rRNA genes by fluorescent in situ hybridization and structural analysis of 5S rRNA genes and intergenic spacer sequences in sugar-beet (*Beta vulgaris* L.). Theor Appl Genet 88:629–636 (1994).

Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM: Biology of *Populus* and its implications for management and conservation. Ottawa, Ont., Canada, NRC 40337, NRC Research Press (1996).

Stirling B, Newcombe G, Vrebalov J, Bosdet I, Bradshaw HD: Suppressed recombination around the *MXC3* locus, a major gene for resistance to poplar leaf rust. Theor Appl Genet 103:1129–1137 (2001).

Tang X, Bao W, Zhang W, Cheng Z: Identification of chromosomes from multiple rice genomes using a universal molecular cytogenetic marker system. J Integrative Plant Biol 49:953–960 (2007).

Tuskan GA, Difazio S, Jansson S, Bohlmann J, Grigoriev I, et al: The genome of western black cottonwood, *Populus trichocarpa* (Torr & Gray). Science 313:1596–1604 (2006).

Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA: Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. Theor Appl Genet 109:451–463 (2004).

Zhao Q, Zhang Yu, Cheng Z, Chen M, Wang S, et al: A fine physical map of the rice chromosome 4. Genome Res 12:817–823 (2002).

Zimmer EA, Jupe ER, Walbot V: Ribosomal gene structure, variation and inheritance in maize and its ancestors. Genetics 120:1125–1136 (1988).

Zwick MS, Hanson RE, McKnight TD, Islam-Faridi MN, Stelly DM, et al: A rapid procedure for the isolation of Cot-1 DNA from plants. Genome 40:138–142 (1997).

9. J. M. Zytkow, J. Zhu, A. Hussam, in *Proceedings of the Eighth National Conference on Artificial Intelligence (AAAI-90)*, 889 (AAAI Press, Menlo Park, CA, 1990).
10. L. Hood, J. R. Heath, M. E. Phelps, B. Lin, *Science* **306**, 640 (2004).
11. L. N. Soldatova, R. D. King, *J. R. Soc. Interface* **3**, 795 (2006).
12. L. N. Soldatova, A. Clare, A. Sparkes, R. D. King, *Bioinformatics* **22**, e464 (2006).
13. R. D. King *et al.*, The Robot Scientist Project, www.aber.ac.uk/compsci/Research/bio/robotsci/ (2008).
14. J. Warringer, E. Ericson, L. Fernandez, O. Nerman, A. Blomberg, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15724 (2003).
15. K. E. Whelan, R. D. King, *BMC Bioinformatics* **9**, 97 (2008).
16. Materials and methods are available as supporting material on *Science* Online.
17. I. Horrocks, P. F. Patel-Schneider, F. van Harmelen, *Web Semantics* **1**, 7 (2003).
18. J. D. Ullman, *Principles of Database and Knowledge-Base Systems, Vol. I: Classical Database Systems* (Computer Science, New York, 1988).
19. M. J. Herrgard *et al.*, *Nat. Biotechnol.* **26**, 1155 (2008).
20. A. Turing, *Mind* **236**, 433 (1950).
21. T. M. Zabriskie, M. D. Jackson, *Nat. Prod. Rep.* **17**, 85 (2000).
22. M. Young, *J. Exp. Bot.* **24**, 1172 (1973).

23. This work was funded by grants from the Biotechnology and Biological Sciences Research Council to R.D.K., S.G.O., and M.Y.; by a SRIF 2 award to R.D.K.; by fellowships from the Royal Commission for the Great Exhibition of 1851, the Engineering and Physical Sciences Research Council, and the Royal Academy of Engineering to A.C.; and by a RC-UK Fellowship to L.N.S. We thank D. Struttman for help with yeast and M. Benway for help with Adam.

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5923/85/DC1
Materials and Methods
Figs. S1 to S3
Table S1
References

8 September 2008; accepted 10 February 2009
10.1126/science.1165620

Priming in Systemic Plant Immunity

Ho Won Jung,[1] Timothy J. Tschaplinski,[2] Lin Wang,[3]* Jane Glazebrook,[3] Jean T. Greenberg[1]†

Plants possess inducible systemic defense responses when locally infected by pathogens. Bacterial infection results in the increased accumulation of the mobile metabolite azelaic acid, a nine-carbon dicarboxylic acid, in the vascular sap of *Arabidopsis* that confers local and systemic resistance against the pathogen *Pseudomonas syringae*. Azelaic acid primes plants to accumulate salicylic acid (SA), a known defense signal, upon infection. Mutation of the *AZELAIC ACID INDUCED 1 (AZI1)* gene, which is induced by azelaic acid, results in the specific loss of systemic immunity triggered by pathogen or azelaic acid and of the priming of SA induction in plants. Furthermore, the predicted secreted protein AZI1 is also important for generating vascular sap that confers disease resistance. Thus, azelaic acid and AZI1 are components of plant systemic immunity involved in priming defenses.

Whole plant immunity, called systemic acquired resistance (SAR), often develops after localized foliar infections by diverse pathogens. In this process, leaves distal to the localized infection become primed to activate a stronger defense response upon secondary infection (*1*). Leaves infected with SAR-inducing bacteria produce vascular sap, called petiole exudate, which confers disease resistance to previously unexposed (naïve) plants (*2*, *3*). This indicates that a mobile systemic signal(s) is involved in SAR (*4*). Although the hormone jasmonic acid (JA) accumulates to a high level in petiole exudates from leaves infected with SAR-inducing bacteria, JA does not seem to be the critical signal for SAR (*5*, *6*). Instead, SAR and the production of active exudates require the DIR1 protein, a predicted secreted protein and putative signal carrier in the lipid transfer protein family, and other proteins involved in glycerolipid biosynthesis (*2*, *3*, *7*). Additionally, SAR and exudate-induced resistance appears to require the phenolic metabolite salicylic acid (SA) (*3*, *8*) and possibly methylsalicylate (MeSA) and its methyl

[1]Department of Molecular Genetics and Cell Biology, The University of Chicago, 1103 East 57th Street EBC410, Chicago, IL 60637, USA. [2]Oak Ridge National Laboratory, Environmental Sciences Division, Oak Ridge, TN 37831–6341, USA. [3]Department of Plant Biology, Microbial and Plant Genomics Institute, University of Minnesota, 1500 Gortner Avenue, St. Paul, MN 55108, USA.

*Present address: Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY 14853–1801, USA.
†To whom correspondence should be addressed. E-mail: jgreenbe@midway.uchicago.edu

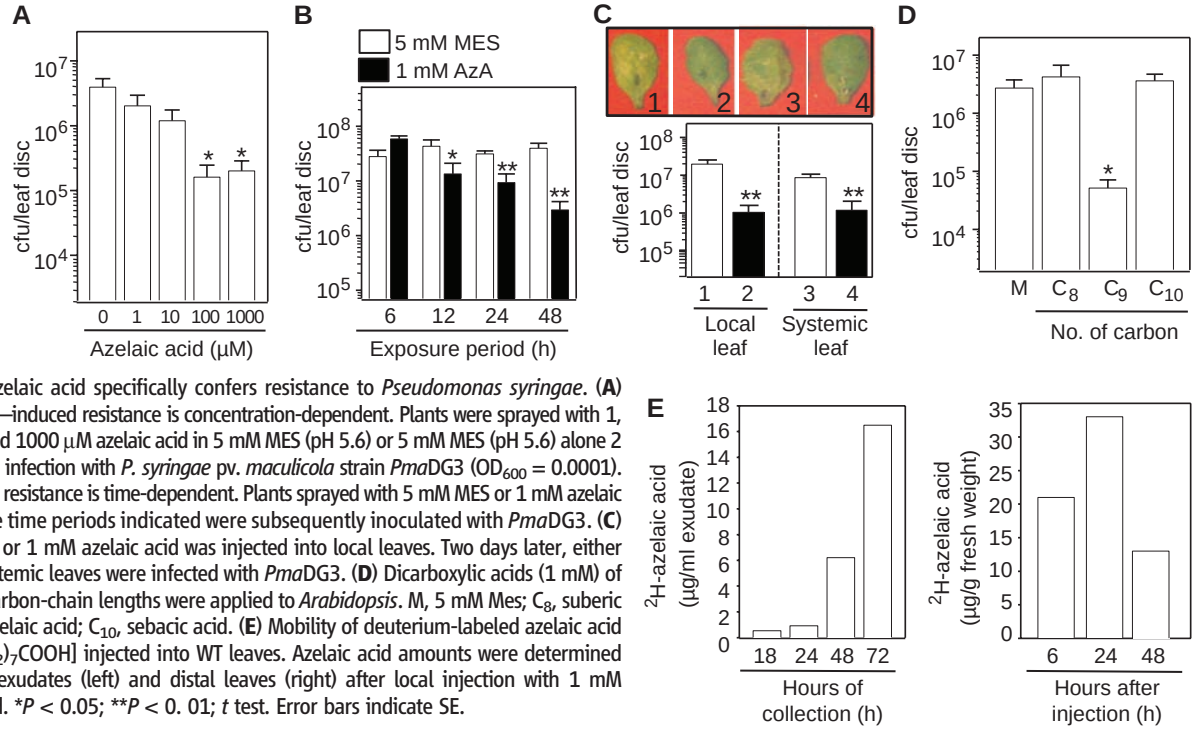


Fig. 1. Azelaic acid specifically confers resistance to *Pseudomonas syringae*. (**A**) Azelaic acid–induced resistance is concentration-dependent. Plants were sprayed with 1, 10, 100, and 1000 μM azelaic acid in 5 mM MES (pH 5.6) or 5 mM MES (pH 5.6) alone 2 days before infection with *P. syringae* pv. *maculicola* strain *Pma*DG3 (OD_{600} = 0.0001). (**B**) Induced resistance is time-dependent. Plants sprayed with 5 mM MES or 1 mM azelaic acid for the time periods indicated were subsequently inoculated with *Pma*DG3. (**C**) 5 mM MES or 1 mM azelaic acid was injected into local leaves. Two days later, either local or systemic leaves were infected with *Pma*DG3. (**D**) Dicarboxylic acids (1 mM) of different carbon-chain lengths were applied to *Arabidopsis*. M, 5 mM Mes; C_8, suberic acid; C_9, azelaic acid; C_{10}, sebacic acid. (**E**) Mobility of deuterium-labeled azelaic acid [$HOOC(CD_2)_7COOH$] injected into WT leaves. Azelaic acid amounts were determined in petiole exudates (left) and distal leaves (right) after local injection with 1 mM azelaic acid. *$P < 0.05$; **$P < 0.01$; *t* test. Error bars indicate SE.

esterase activity (which converts MeSA to SA) in distal leaves (*9*, *10*).

Effective establishment of SAR is not always correlated with elevated systemic SA accumulation, despite its necessary presence, before secondary infections (*11*). This finding implicates that there are additional signal(s) important for priming during SAR. We hypothesize that any such signal should: (i) show elevated levels in petiole exudates of tissue treated with a SAR-inducing pathogen, (ii) confer local and systemic disease resistance through a priming event, (iii) be mobile in plants, and (iv) act in a manner that depends on SA.

To identify this hypothetical signal, we obtained petiole exudates from SAR-induced plants that, unlike exudates from mock-treated plants, conferred disease resistance. We found that pathogen-induced exudates (Col-Pex), but not mock-induced exudates (Col-Mex), conferred resistance to a virulent *Pseudomonas syringae* strain (*Pma*DG3) (fig. S1). Importantly, the SAR-induced exudates did not induce disease resistance when applied to many SAR-defective mutants (fig. S2). Thus, these exudates were biologically active and induced disease resistance in a manner that requires many of the same genes important for SAR.

Because small molecules are often involved in plant signaling, we used gas chromatography coupled with mass spectrometry with electron impact ionization to scan for small molecules (70 to 550 dalton) enriched in the active (SAR-induced) versus inactive (mock-induced) exudates. In four independent experiments, the active exudates consistently had an average of 6.2-fold higher levels of the dicarboxylic acid azelaic acid ($C_9H_{16}O_4$) as compared with inactive exudates [5.1 μM (± 2.3 SEM) in mock-induced exudates versus 31.6 μM (± 10.0 SEM) in active exudates, $P = 0.042$, *t* test]. In vitro, azelaic acid showed weak antimicrobial activity against phytopathogenic bacteria but no activity against fungi (table S1).

To examine if azelaic acid induces disease resistance, we measured *Pma*DG3 growth after spraying plants with different concentrations of azelaic acid and found that more than 10 μM azelaic acid was required to confer disease resistance (Fig. 1A). Additionally, plants needed to be exposed to azelaic acid for at least 12 hours before infection, suggesting that azelaic acid does not directly inhibit *Pma*DG3 during infection (Fig. 1B). Local application of azelaic acid (1 mM) conferred both local and systemic resistance to *Pma*DG3 (Fig. 1C). The immunity-inducing properties of azelaic acid appeared to be specific, as the related C_8 and C_{10} dicarboxylic acids did not induce disease resistance (Fig. 1D). Furthermore, azelaic acid accumulated in distal systemic tissue as well as petiole exudates when it was locally applied to leaves, showing that it was mobile in plants when labeled with deuterium [$HOOC(CD_2)_7COOH$] (Fig. 1E and fig. S3).

To test which genes are necessary for azelaic acid–induced resistance, we monitored *Pma*DG3 growth in leaves of various mutants affecting SAR.

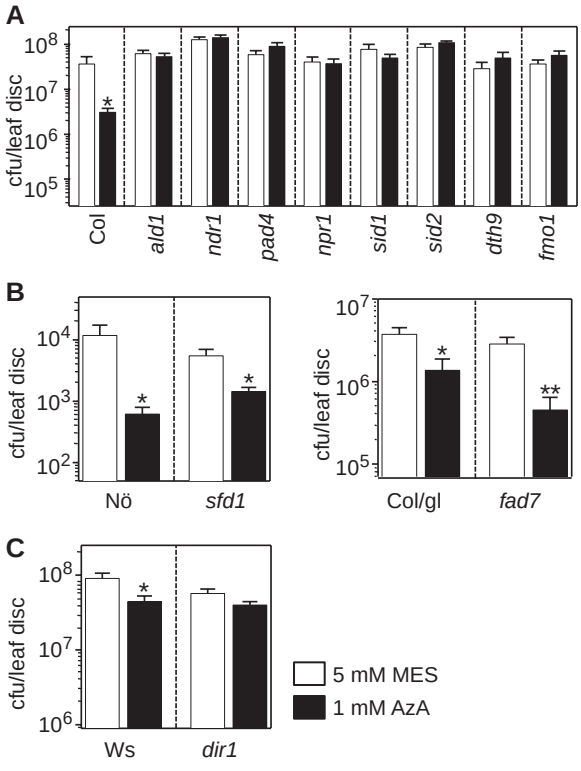


Fig. 2. Genes involved in azelaic acid–induced resistance. 1 mM azelaic acid in 5 mM MES and 5 mM MES alone were sprayed on to WT *Arabidopsis* accessions and the indicated SAR-defective SA pathway mutants (**A**) and glycerolipid mutants (**B** and **C**) 2 days before challenge inoculation with *Pma*DG3 (A and B) or *P. syringae* pv. *tomato* DC3000 (C). Bacterial growth in azelaic acid–treated plants should only be compared with those of the same genotype pre-treated with 5 mM MES, because different genotypes were grown separately. $*P < 0.05$; $**P < 0.01$; *t* test. Error bars indicate SE.

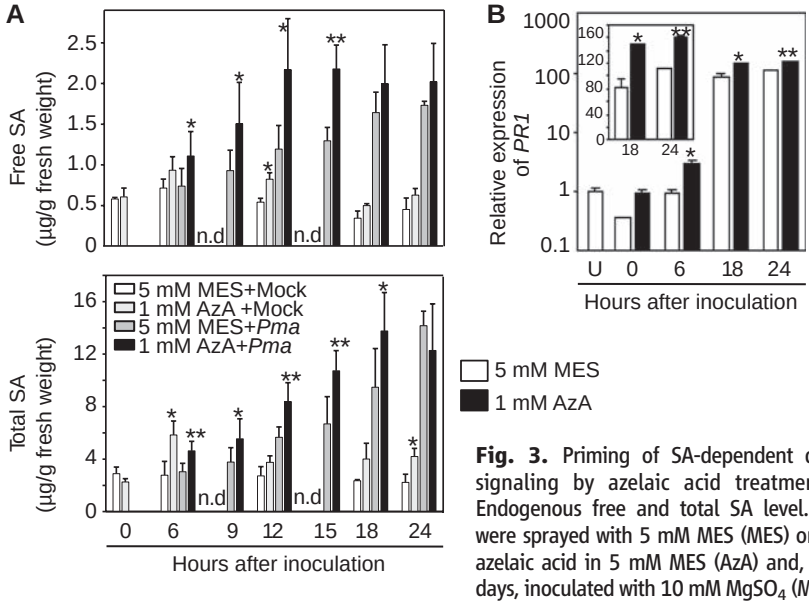


Fig. 3. Priming of SA-dependent defense signaling by azelaic acid treatment. (**A**) Endogenous free and total SA level. Plants were sprayed with 5 mM MES (MES) or 1 mM azelaic acid in 5 mM MES (AzA) and, after 2 days, inoculated with 10 mM $MgSO_4$ (Mock) or *Pma*DG3 (*Pma*). (**B**) Relative *PR1* expression in leaves treated as in (A). Expression of *PR1* is plotted on a log scale; *PR1* levels at 0 hours were not statistically different ($P = 0.065$). (Inset) Increased *PR1* levels in azelaic acid–treated plants at 18 and 24 hours after subsequent *Pma*DG3 infection. $*P < 0.05$; $**P < 0.001$; *t* test. Error bars indicate SD [number of samples analyzed per treatment (n) = 5 (A) or 3 (B)]. U, untreated; n.d., not determined.

We found that azelaic acid did not induce resistance in known SAR-defective SA pathway mutants (Fig. 2A) but did confer resistance to *sfd1* and *fad7* mutants (Fig. 2B), which lack an unidentified glycerolipid-requiring SAR signal (*3*, *7*). In contrast, the SAR-defective *dir1-1* mutant was insensitive to azelaic acid, suggesting that a DIR1-mediated signal is required for azelaic acid–induced resistance (Fig. 2C). The hormone mutants, *jar1* and *etr1*, that are not SAR-defective were responsive to azelaic acid (fig. S4).

Azelaic acid's effects may be to directly induce SA production. Although intact SA synthesis and signaling was required for azelaic acid–induced resistance, free and total SA levels were not significantly elevated up to 48 hours after treatment, compared with those of mock-treated plants (fig. S5). Alternatively, it may be that azelaic acid primes

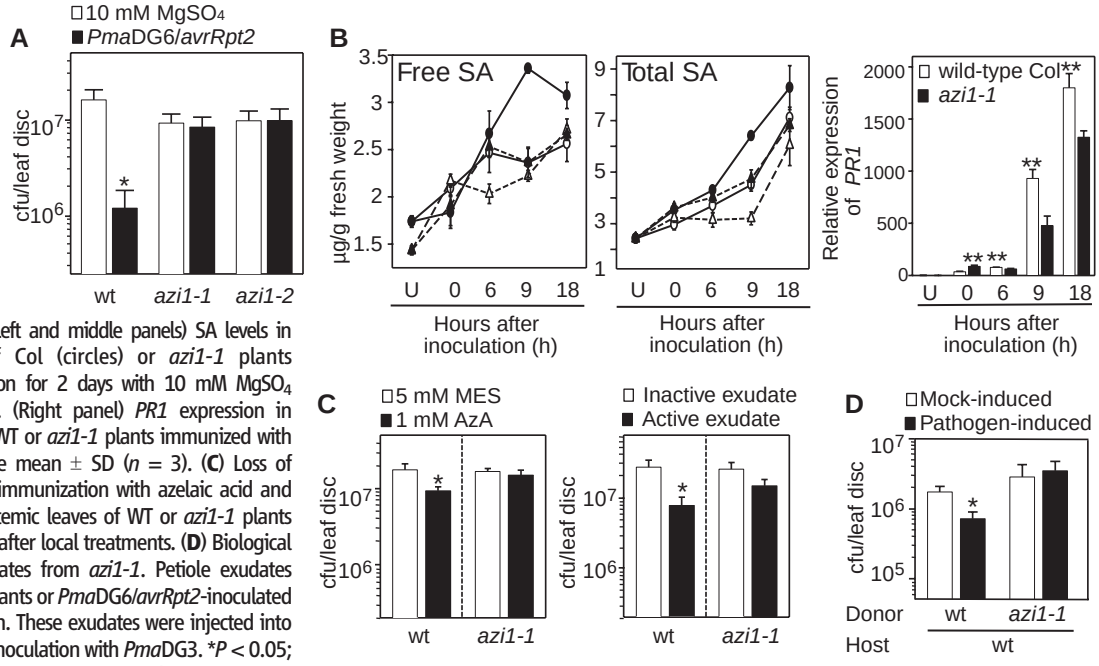


Fig. 4. Signaling defects of *azi1* mutants. (**A**) Loss of SAR as measured by *Pma*DG3 growth in immunized *azi1* plants. WT Col or *azi1* plants were immunized with 10 mM MgSO4 or *Pma*DG6/*avrRpt2* 2 days before secondary infection of distal systemic leaves with *Pma*DG3 and differences identified with a *t* test. (**B**) Reduced defense priming in distal systemic leaves of *Pma*DG6/*avrRpt2*-inoculated *azi1*. (Left and middle panels) SA levels in *Pma*DG3–infected distal leaves of Col (circles) or *azi1-1* plants (triangles) after local immunization for 2 days with 10 mM $MgSO_4$ (white) or *Pma*DG6/*avrRpt2* (black). (Right panel) *PR1* expression in *Pma*DG3–infected distal leaves of WT or *azi1-1* plants immunized with *Pma*DG6/*avrRpt2*. Data present the mean ± SD (n = 3). (**C**) Loss of systemic response of *azi1* to local immunization with azelaic acid and Col-Pex (active exudate). Distal systemic leaves of WT or *azi1-1* plants were infected with *Pma*DG3 2 days after local treatments. (**D**) Biological activity of pathogen-induced exudates from *azi1-1*. Petiole exudates were collected from mock-treated plants or *Pma*DG6/*avrRpt2*-inoculated plants for 72 hours after inoculation. These exudates were injected into WT plants 2 days before challenge inoculation with *Pma*DG3. *P < 0.05; **P < 0.001; *t* test. Error bars indicate SE. U, untreated.

plant cells to mount a faster and/or stronger defense response, including SA production, upon infection. Azelaic acid–treated plants that were subsequently infected had higher levels of both SA (Fig. 3A) and transcripts of the SA-associated signaling marker *PR1* compared with mock-treated plants (Fig. 3B). Thus, azelaic acid primes SA production upon infection upstream of both the SA-dependent SAR signaling pathway and the DIR1-dependent signal and downstream or independent of SFD1 and FAD7.

We examined possible effectors of azelaic acid with microarray analysis (*12*) but observed no statistically significant altered expression above twofold (table S2) of 464 defense-related genes. Thus, azelaic acid does not cause dramatic reprogramming of the plant transcriptome. However, *AZI1* (*AZELAIC ACID INDUCED 1*, At4g12470), which showed modest induction at 24 hours (1.8-fold, P = 0.13), was transiently and significantly induced at 3 to 6 hours by azelaic acid and active exudates (fig. S6). AZI1 encodes a predicted secreted protease inhibitor/seed storage/lipid transfer protein family protein, has no significant similarity to DIR1 (*2*), and confers disease resistance when overexpressed (*13*).

Two independent *azi1* mutants (SALK_017709 and SALK_085727, fig. S7) were found to be defective in SAR (Fig. 4A). However, *azi1* plants showed normal susceptibility to local infection with several strains of *P. syringae* (fig. S8). We observed that after local immunization with SAR-inducing bacteria, wild-type (WT) plants appeared to be primed to accumulate SA and *PR1* transcripts in distal leaves upon secondary infection (Fig. 4B). In contrast, *azi1* plants showed reduced priming during SAR (Fig. 4B), although they showed normal SA and *PR1* accumulation during local immunization (fig. S9).

SAR can be impaired because of a failure to recognize a defense/priming signal, generate the signal(s) in local infected-leaves, or translocate the signal(s) from local infected-leaves. To test if *azi1* plants fail to recognize a defense/priming signal(s), we infiltrated azelaic acid or active petiole exudate into leaves of *azi1* mutants. Two days later, we inoculated the same leaves with *Pma*DG3. *azi1* plants were resistant to subsequent local infection (fig. S10), indicating that *azi1* mutants still recognize defense/priming signal(s). To test whether AZI1 functions in long-distance signaling for systemic immunity, we examined the growth of *Pma*DG3 in systemic leaves of *azi1* plants in which azelaic acid or active petiole exudate (Col-Pex) was locally infiltrated. Azelaic acid and petioles exudates failed to induce systemic immunity in *azi1* plants, although these treatments protected WT plants against subsequent infection (Fig. 4C). Additionally, pathogen-induced exudates from *azi1* were inactive when applied to WT plants (Fig. 4D). Thus, AZI1 modulates production and/or translocation of a mobile signal(s) during SAR.

In summary, azelaic acid has the expected properties of a SAR component in that the molecule is mobile in plants, shows increased accumulation in biologically active exudates, confers pathogen resistance to local and systemic tissues, requires genes important for SA production and action to confer disease resistance, and primes SA accumulation and SA-associated gene expression. AZI1 appears to be induced by azelaic acid and regulates and/or directly translocates a SAR signal(s) from local infected tissues. The identification of previously unknown SAR components may be useful for plant protection and provides insight into how some interactions trigger systemic plant immunity.

References and Notes

1. J. A. Ryals *et al.*, *Plant Cell* **8**, 1809 (1996).
2. A. M. Maldonado, P. Doerner, R. A. Dixon, C. J. Lamb, R. K. Cameron, *Nature* **419**, 399 (2002).
3. R. Chaturvedi *et al.*, *Plant J.* **54**, 106 (2008).
4. R. A. Dean, J. Kuc, *Phytopathology* **76**, 966 (1986).
5. W. Truman, M. H. Bennett, I. Kubigsteltig, C. Turnbull, M. Grant, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1075 (2007).
6. C. M. J. Pieterse *et al.*, *Plant Cell* **10**, 1571 (1998).
7. A. Nandi, R. Welti, J. Shah, *Plant Cell* **16**, 465 (2004).
8. J. Malamy, J. P. Carr, D. F. Klessig, I. Raskin, *Science* **250**, 1002 (1990).
9. S.-W. Park, E. Kaimoyo, D. Kumar, S. Mosher, D. F. Klessig, *Science* **318**, 113 (2007).
10. A. C. Vlot *et al.*, *Plant J.* **56**, 445 (2008).
11. R. K. Cameron, N. L. Paiva, C. J. Lamb, R. A. Dixon, *Physiol. Mol. Plant Pathol.* **55**, 121 (1999).
12. M. Sato *et al.*, *Plant J.* **49**, 565 (2007).
13. C. Chassot, C. Nawrath, J.-P. Métraux, *Plant J.* **49**, 972 (2007).

14. This work was funded by NSF grants to J.T.G. (IOB-0450207) and J.G. (IOB-0419648) and by a fellowship from Korea Research Foundation to H.W.J., and T.J.T. was supported by the Office of Biological and Environmental Research of the U.S. Department of Energy. Oak Ridge National Laboratory is managed by UT-Battelle for the U.S. Department of Energy under contract DE-AC05-00OR22725. We thank J. Bergelson and L. Mets for the use of their high-performance liquid chromatographys, F. Katagiri and J. Malamy for useful discussions, R. Cameron for sharing her detailed exudate collection protocol, and N. Engle for analysis of exudates. A patent application related to this work has been filed by the University of Chicago and UT-Battelle on behalf of inventors J.T.G., H.W.J., and T.J.T. on the use of azelaic acid to confer protection in crop plants against infection.

Supporting Online Material

www.sciencemag.org/cgi/content/full/324/5923/89/DC1
SOM Text
Figs. S1 to S10
Tables S1 to S3
References

19 December 2008; accepted 9 February 2009
10.1126/science.1170025

RESEARCH ARTICLE

Shotgun proteome profile of *Populus* developing xylem

Udaya C. Kalluri[1], Gregory B. Hurst[2], Patricia K. Lankford[3], Priya Ranjan[1] and Dale A. Pelletier[3]

[1] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[2] Chemical Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA
[3] Biosciences Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA

Received: November 1, 2008
Revised: July 25, 2009
Accepted: July 27, 2009

Understanding the molecular pathways of plant cell wall biosynthesis and remodeling is central to interpreting biological mechanisms underlying plant growth and adaptation as well as leveraging that knowledge towards development of improved bioenergy feedstocks. Here, we report the application of shotgun MS/MS profiling to the proteome of *Populus* developing xylem. Nearly 6000 different proteins were identified from the xylem proteome. To identify low-abundance DNA-regulatory proteins from the developing xylem, a selective nuclear proteome profiling method was developed. Several putative transcription factors and chromatin remodeling proteins were identified using this method, such as NAC domain, CtCP-like and CHB3-SWI/SNF-related proteins. Public databases were mined to obtain information in support of subcellular localization, transcript-level expression and functional categorization of identified proteins. In addition to finding protein-level evidence of candidate cell wall biosynthesis genes from xylem (wood) tissue such as cellulose synthase, sucrose synthase and polygalacturonase, several other potentially new candidate genes in the cell wall biosynthesis pathway were discovered. Further application of such proteomics methods will aid in plant systems biology modeling efforts by enhancing the understanding not only of cell wall biosynthesis but also of other plant developmental and physiological pathways.

Keywords:
LC-MS/MS / Plant proteomics / *Populus* / Xylem

1 Introduction

The release of the *Populus trichocarpa* (Torr. & Gray) genome sequence and the development of supportive reverse genetics, population genetics and transcript profiling capabilities have facilitated gene sequence–based research in an unprecedented manner [1]. Currently, widespread efforts are being undertaken by the research community to improve woody biomass–based biofuel production. Central to the improvement of feedstock properties is a comprehensive understanding of wood development and cell wall biosynthesis properties. Wood or dead secondary xylem in tree trunks is formed as a result of division, enlargement, differentiation, maturation and programmed cell death of xylem cells that are perennially produced by the vascular cambium. Moreover, plant cell walls, which are significant to development and morphology, water and solute transport mechanisms, disease resistance and strength of a plant as a whole, are predominantly of the secondary-walled type in xylem tissue [2]. Therefore, developing xylem cells serve as useful models to investigate secondary cell wall formation. Several studies have reported single-gene to transcriptome-level changes during xylem development or progression to secondary cell wall formation [3–7]. These studies are based on EST, microarray or RT-PCR transcript-level gene expression data, which are valuable indicators but are not entirely representative of gene product activity. Protein-level measurements are significant in deciphering the post-translational abundance, regulatory status and subcellular localization of the gene product and hence in suggesting its

Correspondence: Dr. Udaya C. Kalluri, Staff Scientist, Environmental Sciences Division, Oak Ridge National Laboratory, P. O. Box 2008, Oak Ridge, TN 37831, USA
E-mail: kalluriudayc@ornl.gov
Fax: +1-865-576-9939

Abbreviations: MudPIT, multidimensional protein identification technology; **NSAF,** normalized spectral abundance factor

activity. However, protein-level information for *Populus* has thus far been primarily derived using 2-D PAGE approaches [8–11]. The ability to conduct proteome measurements in a high-throughput manner will be highly valuable in expediting research investigations as well as in making them more comprehensive.

HPLC interfaced with MS/MS offers a general and automated approach for identifying large numbers of individual protein components of the proteome. In the Multidimensional Protein Identification Technology (MudPIT) implementation of this approach [12, 13], the entire protein complement of a tissue, cell or subcellular compartment is enzymatically digested. The resulting peptides are analyzed using automated 2-D chromatographic separation (strong cation exchange followed by reverse phase) interfaced *via* electrospray with a mass spectrometer. MS/MS analysis of peptides eluting from the 2-D separation provides partial amino acid sequence information, which allows software-based identification of a peptide and further its assembly into protein identifications. The MudPIT approach has been previously applied to proteomics of rice [14] as well as to the study of ubiquitinated proteins in *Arabadopsis thaliana* [15]. To our knowledge, MudPIT has not been applied to proteomics studies of *Populus*.

Central to making sense of the large genome data sets and their predicted functions is the ability to put the proteome in a cell biology context. It is therefore highly desirable to be able to apply the shotgun profiling approaches to subcellular fractions. Here, we report the application of MudPIT to shotgun proteome profiling of *Populus*, a model bioenergy crop. Using developing xylem from *Populus* stems, we have applied subcellular fractionation techniques to obtain crude, pellet and nuclear protein fractions. LC-MS/MS analysis was followed with bioinformatics-based analysis of functional annotation, gene duplication, predicted subcellular localization and transcript-level expression support.

2 Materials and methods

2.1 Plant materials

Wild-type *Populus* plants, hybrid aspen clone 717 (*Populus tremula x alba*), were vegetatively propagated in soilless mix of perlite:peat (2:1) and grown for a minimum of 6 months under greenhouse conditions of ambient humidity and controlled light (high pressure sodium lamps, 16 h photoperiod), fertilization (weekly treatment of Jack's professional water soluble 20:10:20 fertilizer with micronutrients), temperature ($72 \pm 2^{\circ}$F) and automated drip irrigation. Green stem tops representing younger internodes were excised and removed and the woody portion of the stem was used to generate developing xylem sample. Xylem tissue was collected by peeling off the bark, scraping the juicy (developing tissue) outer layers of exposed stems and flash-freezing the material in liquid nitrogen. Such a tissue sample is expected to be composed of cell types from various phases of secondary xylem development, including expanding as well as secondary cell wall forming xylem vessels and fibers (predominant tissue component) [4, 5]. Xylem tissue pooled from several plants was used in generating two batches of samples for proteome characterization.

2.2 Protein extraction and quantification

Xylem tissue was ground under liquid nitrogen using a mortar and pestle. A 3 g sample of the ground tissue was suspended in 15 mL lysis buffer containing 125 mM Tris (pH 8.5), 10% glycerol, 50 mM DTT and 1 mM EDTA [16].

The suspension was vortexed twice for 30 s each time, then sonicated (Branson 185 sonifier, power setting of 40) on ice for three rounds of 30 s each. Large debris was removed from the highly viscous sample by centrifugation for 6 min at $1200 \times g$. The supernatant was again centrifuged for 10 min at $12\,000 \times g$, and the pellet discarded. A final centrifugation step at $100\,000 \times g$ for 1 h yielded a crude soluble protein fraction (crude/soluble fraction) and a pellet (pellet fraction). Protein determination using Lowry's method [17] indicated 135 mg total protein in the crude soluble fraction and 30 mg total protein in the pellet.

2.3 Nuclei isolation

Nuclei were extracted from xylem tissue using a CelLytic PN Isolation/Extraction Kit following the manufacturer's protocol (Sigma, St. Louis, MO, USA). Briefly, a 3.4 g sample of the ground xylem tissue was briefly suspended in 10 mL of 1 × nuclei isolation buffer containing 1 mM DTT; the sample was then gently agitated for 10 min at 4°C. The suspension was passed through filter mesh to remove debris into a 50 mL tube. The suspension was centrifuged at $1000 \times g$ in tabletop centrifuge for 10 min. The supernatant was decanted. The resulting pellet was gently resuspended in 1 mL of nuclei isolation buffer and lysed by addition of 0.3% Triton X-100. About 2 mL of cell lysate was obtained. Aliquots of 600 uL of lysed cell material was applied to a 0.8 mL cushion of 1.5 M sucrose in a 1.5 mL centrifuge tube and centrifuged for 10 min at $12\,000 \times g$. The upper green phase was aspirated leaving the nuclei pellets. The pellets were washed twice with ~0.5 mL of buffer, pelleted by centrifugation and resuspended in 150 μL of extraction buffer with 5 mM DTT and then vortexed at 4°C for 30 min. The sample was then centrifuged at $12\,000 \times g$ for 10 min yielding 150 μL of soluble extracted nuclei (nuclear fraction). Protein determination using Lowry's method [17] indicated 300 μg total protein.

2.4 Immunoblotting

Aliquots of total protein extracted from xylem cells and enriched nuclei were separated on a 4–20% Precision Protein acrylamide gel (Pierce, Rockville, IL, USA) and transferred to PVDF using the Invitrogen iBlot system (Invitrogen, Carlsbad, CA, USA). Membranes were washed overnight at 4°C in PBS containing 0.1% Tween 20 (PBS-Tween). Non-specific binding was blocked by incubation of the membrane with gentle rocking for 0.5 hour in 5% non-fat milk powder, 5% BSA in PBS-Tween (''Blotto''). The membrane was then incubated for 2 h with rocking at room temperature with rabbit anti-Histone H3 (Sigma). After extensive washing in PBS-Tween, the membrane was incubated with rocking for 1.5 h in HRP-conjugated goat anti-rabbit IgG (Bio-Rad, Hercules, CA, USA) at 1/1500 dilution in Blotto. After washing, color was developed by the addition of Immuno-Pure metal enhanced 3,3'-diaminobenzidine substrate (Pierce).

2.5 Protein digestion

The digestion protocol was adapted from similar methods previously applied in proteomics studies on a range of bacterial species [18, 19]. In preparation for MS analysis, samples were denatured with 6 M guanidine and 10 mM DTT for 1 h at 60°C. The reduced and denatured samples were diluted with 50 mM Tris-HCl, 10 mM $CaCl_2$ (pH 7.6) to bring the guanidine concentration to 1 M. Digestion was performed by adding 20 μg modified porcine trypsin (sequencing grade, Promega) to 3 mg protein (for crude and pellet fractions) or 3 μg trypsin to 300 μg protein (for nuclear fraction) at 37°C overnight, followed by a second addition of the same amount of trypsin and incubation for an additional 4 h at 37°C.

Crude and pellet samples were desalted using SepPak Plus C18 cartridges (Waters) following the manufacturer's protocol, with final elution using 100% ACN. Nuclear protein samples were similarly desalted using SepPak Lite cartridges (Waters.) A 500 μL portion of aqueous 0.1% formic acid was added to each desalted sample. ACN was removed using vacuum centrifugation (SpeedVac, Savant Instruments, Holbrook NY) to bring samples to a final volume of ~500 μL. Samples were passed through 0.45 μm Ultrafree-MC filters (Millipore, Bedford, MA, USA) to remove particulates.

2.6 LC-MS/MS

Proteins were identified from digests using MudPIT [12, 13, 20], with 2-D HPLC interfaced with MS/MS, as described previously [18, 19]. Each fraction (crude, pellet, nuclear proteins) was analyzed in duplicate.

Aliquots of 100 μL protein digest, containing ~500 μg protein from crude or pellet fractions or ~150 μg protein from nuclear protein fraction, were loaded *via* a pressure cell (New Objective) onto a ''back column'' constructed as follows. A 100 μm ID fused-silica capillary column contained a 3–4 cm length of C18 RP resin (Aqua, 5 μm particle, 200 Å pore size [Phenomenex]) upstream of a 3–4 cm length of strong cation exchange phase (SCX; Luna, 5 μm particle, 100 Å pore size [Phenomenex]). The back column was attached *via* a filter union (Upchurch) to the ''front column,'' a 100 μm ID resolving column/nanospray tip packed with C18 RP resin (Jupiter, 5 μm particles, 300 Å pore size [Phenomenex]). The assembled columns were attached to the flow from an HPLC pump (Ultimate, LCPackings/Dionex, Sunnyvale CA, USA). A total flow of 150 μL/min from the pump was split to provide a flow through the column of ~300 nL/min.

Twelve HPLC cycles were performed *per* sample. The first cycle consisted of an RP gradient from 100% solvent A (95% H_2O, 5% ACN, 0.1% formic acid) to 50% solvent B (30% H_20, 70% ACN, 0.1% formic acid) over 45 min followed by a ramp to 100% solvent B over 10 min. In cycles 2–11, 100% solvent A was applied for 5 min, followed by a 2 min salt step gradient of 400 mM ammonium acetate (10, 15, 20, 25, 30, 35, 40, 45, 50 and 60%, respectively, in solvent A), followed by 3 min of 100% solvent A, then an RP gradient (100% solvent A to 50% solvent B over 110 min). In cycle 12, 100% solvent A was applied for 5 min, followed by a 10 min salt step gradient of 400 mM ammonium acetate (100%), followed by 9 min of 100% solvent A, then an RP gradient (100% solvent A to 100% solvent B over 75 min). The back column was removed, and the front column subjected to a final wash. The equilibration step was performed by ramping from 100% solvent A to 100% solvent B over 10 min, holding 2 min at 100% solvent B, ramping to 100% solvent A over 5 min, and holding at 100% solvent A for 10 min. A single front column was used for several experiments, while a new back column was prepared for each LC-MS/MS analysis.

The LC eluent was interfaced *via* a nanospray source with the mass spectrometer (LTQ, ThermoFinnigan, San Jose CA, USA), controlled by XCalibur software. Acquisition of tandem mass spectra was triggered in a data-dependent mode provided by the XCalibur software, with collision-activated dissociation of five parent ions selected from the most intense ions in each full scan mass spectrum. Parent ions selected for MS/MS analysis more than once (*i.e.* repeat count = 1) within 1 min were placed on an exclusion list for 3 min, during which time they were not subjected to collision-activated dissociation. For both full scan mass spectra and tandem mass spectra, two microscans were averaged.

2.7 Protein identification

Peptides were identified using version 27 of the software program Sequest [21] to compare experimental tandem

mass spectra with predicted fragmentation patterns of tryptic peptides generated from the protein database for *P. trichocarpa* (version 1.1, available at http://genome.jgi-psf.org/cgi-bin/searchGM?db = Poptr1_1 on July 20, 2007, file contained 45 555 proteins in total) plus common contaminant proteins. A decoy database, containing amino acid sequence-reversed analogs of each protein, was appended to allow estimates of false discovery rates [22, 23]. Sequest searches were carried out with parent ion tolerance of 3.0 m/z units, fragment mass tolerance of 0.5 m/z units. The trypsin cleavage rule was applied, with up to four internal missed cleavage sites allowed *per* peptide. Peptide identifications from Sequest were filtered and organized into protein identifications using DTASelect version 1.9 [24]. Ambiguous peptide identifications were removed using the *–a false* option in DTASelect. Peptide identifications were retained for Sequest results of XCorr $\geq$ 1.8 ($z = 1$), XCorr $\geq$ 2.5 ($z = 2$) or XCorr $\geq$ 3.5 ($z = 3$), and DeltaCN $\geq$ 0.08. These values yielded peptide false discovery rates $<$ 0.5%. Identification of a protein required the identification of two or more peptides from that protein or the identification of a single peptide in at least two charge states [25]. False discovery rates at the protein level were $< = 1.2\%$; each known false protein identification was a 2-peptide hit, with one exception which was a 3-peptide hit. Lists of peptide sequences identified from six different runs of xylem protein extracts are available at http://compbio.ornl.gov/populus_tremula_x_alba_proteome/. Estimates of protein quantities were based on comparisons of summed spectrum counts from conserved and/or unique peptides. For a given protein, spectrum count is the number of tandem mass spectra that are assigned to peptides from that protein and provides an approximate indicator of protein abundance [26]. While quantitative estimates discussed in this paper were based on spectrum counts, we have also provided the normalized spectral abundance factor (NSAF) for every protein [27], which can be found in Supporting Information Tables 5–10 at http://compbio.ornl.gov/populus_tremula_x_alba_proteome/. NSAF value is the spectrum count for a protein divided by its length in amino acids, divided by the sum across all proteins of that same quantity. Since only unique peptides are considered, the NSAF is equal to zero for any protein for which no unique peptides were identified.

2.8 Bioinformatics analysis

The annotation information for gene models including locus ID, gene description, protein sequence, conserved domain information, gene ontology (GO annotation), EuKaryotic Orthologous Groups (KOG) information, Enzyme Commission (EC) number annotations and Kyoto Encyclopedia of Genes and Genomes (KEEG) pathway information was obtained from the Joint Genome Institute (JGI) website for *Populus* genome (http://genome.jgi-psf.org/Poptr1_1/). The sub-cellular localization of proteins was predicted using a locally downloaded version of WoLF PSORT (http://wolfpsort.org/) [28], which is an extension of the PSORT II program. WoLF PSORT predicts proteins localizing to major subcellular sites such as nuclear, cytosol, mitochondrial and extracellular regions based on sorting signals, amino acid composition and functional motifs such as DNA-binding motifs. The presence or absence of membrane spanning regions for gene models was predicted using a locally downloaded version of TMpred software (http://www.ch.embnet.org/software/TMPRED_form.html) [29]. A PERL script was written to automate the whole process and assemble the information. The presence or absence of EST support for a given gene model was inferred from NCBI poplar EST and the PopulusDB EST data sets. Nucleotide sequences of gene models were queried against these ESTs using a local BLAST database with an *e*-value cut-off of 1E-10.

The best EST category, which refers to tissue library in which ESTs corresponding to the gene model is best represented, was identified based on EST library distribution information at PopulusDB [30]. A list of highly similar paralogous gene pairs occurring within large conserved syntenic blocks that resulted from the salicoid duplication event in the *Populus* genome has been previously published [31]. This information was used to correlate the extent to which our proteome profiling method could uniquely identify or differentiate between protein products of duplicate gene pairs.

3 Results and discussion

3.1 Identification and functional classification of proteins in *Populus* xylem proteome

The complete xylem proteome data set consisted of proteins representing 5847 distinct *Populus* gene models, which were identified based on matches of tandem mass spectra to the *Populus* genome sequence-based-peptide database (Supporting Information Table 1). The proteins were obtained from developing xylem extracts of wild-type *Populus* plants and further fractionated into crude, pellet and nuclear fractions prior to LC-MS/MS analysis. While the xylem proteome set contained some proteins that were identified from more than one fraction, 1124 proteins were identified exclusively in the crude fraction, 907 were identified in only the pellet fraction and 775 were identified in only the nuclei (Fig. 1).

Sequest searches of the LC-MS/MS data identified 26 757 distinct tryptic peptide sequences. The vast majority of peptides (23 889) contained zero or one missed trypsin cleavage sites, while only 61 peptides contained four missed cleavage sites, indicating the completeness of the digestion. Most of the identified peptides (17 728 or 66%) occur in only a single locus in the annotated *Populus* genome, while 6572

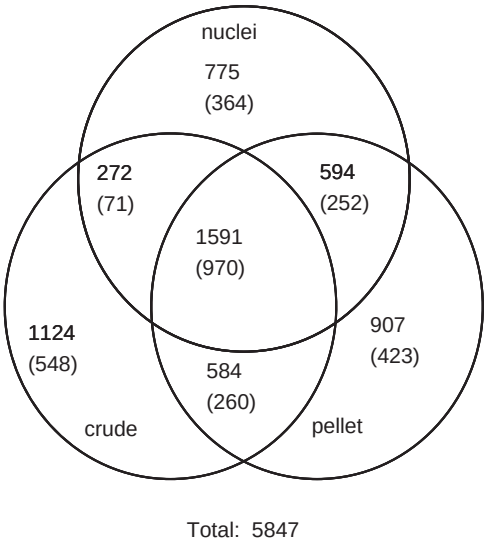


Figure 1. Summary of numbers of identified proteins in various fractions. Numbers in parentheses represent proteins detected in both LC-MS/MS experiments of relevant fractions.

(25%) occur in two loci, 1190 (4%) in three loci, 669 (3%) in four loci and the remainder (598 or 2%) in five or more loci. In total, 48% (12 748) of the peptides were identified by three or more MS/MS spectra.

Based on the annotation of the *Populus* genome sequence, these peptides were assembled into protein identifications. The nomenclature of Yang *et al.* (2004) was followed to classify identified proteins from each LC-MS/MS experiment (parsimony columns in Supporting Information Table 1) and the classification codes are summarized in (Supporting Information Table 2) [32]. *Distinct* proteins contain only peptides that are uniquely found in that locus. *Differentiable* proteins contain one or more peptides that are unique to that locus, as well as one or more peptides that also occur elsewhere in the annotated proteome. The remaining classes contain proteins that are characterized by zero identified peptides that are unique to the locus. *Indistinguishable* proteins have a set of identified peptides that is identical to the set of identified peptides for one or more other proteins in the data set. The indistinguishable proteins can be combined into groups that share sets of peptides. All identified peptides in a *subset* protein also occur in another protein, which contains additional identified peptides. Note that overlap is possible between indistinguishable and subset proteins. Finally, the identified peptides in *subsumable* proteins can be found in two or more proteins encoded by other loci in the annotated genome. Out of the 5847 total protein identifications in the present study, 4283 proteins were uniquely identified (classified as *Distinct* or *Differentiable*).

The distribution of isoelectric points predicted from the amino acid sequences of the experimentally identified proteins (Supporting Information Fig. 1, upper left) is similar to the distribution for all proteins from the predicted proteome (Supporting Information Fig. 1, lower left). The distribution of predicted molecular masses of identified proteins (Supporting Information Fig. 1, upper right) is skewed slightly to higher masses than that of all predicted proteins (Supporting Information Fig. 1, lower right), suggesting a slight bias against smaller proteins in our measurement. A similar experimental bias was noted in a large-scale proteomics study of another plant model system, *Arabidopsis* [33], which employed a significantly different workflow that combined gel electrophoretic separation, in-gel digestion and RP LC-MS/MS analysis.

The molecular clock is estimated to be ticking slowly in *Populus* relative to rice, *Arabidopsis* and human genomes owing to its perennial, long-lived and vegetatively propagating nature [31]. Due to this fact, duplicate genes arising out of the salicoid duplication event, a recent 60 mya genome-wide duplication event that has resulted in nearly two-thirds of genes in the *Populus* genome, are often very similar at the protein sequence level (~90% or higher identity). Our analysis reveals that such duplicate pairs can be differentiated at appreciable levels. For example, 47% of the proteins encoded by duplicated genes that were detected in the crude fraction were differentiable based on identification of unique representative peptides (Supporting Information Table 3). The observation that 74% of all crude fraction proteins were identified based on distinct or differentiable peptide sequences is encouraging with respect to the role proteome profiling can play in poplar biology. It is also significant to note that, among the proteins that were supported by conserved peptides (Indistinguishable peptide groups), about 64% represent proteins from duplicate gene pairs but 36% do not represent duplicates; instead, they were found to represent other closely related protein family members in the same fraction.

The xylem proteome data set could be broadly classified into 23 functional categories (Fig. 2). It was found, as expected, that proteins with housekeeping functions such as histone-fold/TFIID-TAF and histone H4 genes had very high spectrum counts (~400) and certain transcription factors such as MYB1 DNA-binding protein [eugene3.01450016], bZIP family transcription factor [eugene3.01630045], and Homeobox DNA-binding protein [fgenesh4_pg.C_LG_I002015] were found to have low (<10) spectrum counts. A genome-scale proteomics study conducted using various *Arabidopsis* plant organs also found the aforementioned distinction in the categories of proteins under- and over-represented in proteome data sets [33].

It is significant to note that BLAST search of the gene models against locally downloaded poplar ESTs revealed that our shotgun profiling effort uncovered proteins for 99 representative *Populus* gene models that previously lacked experimental validation (''EST support'' column in Supporting Information Table 1).

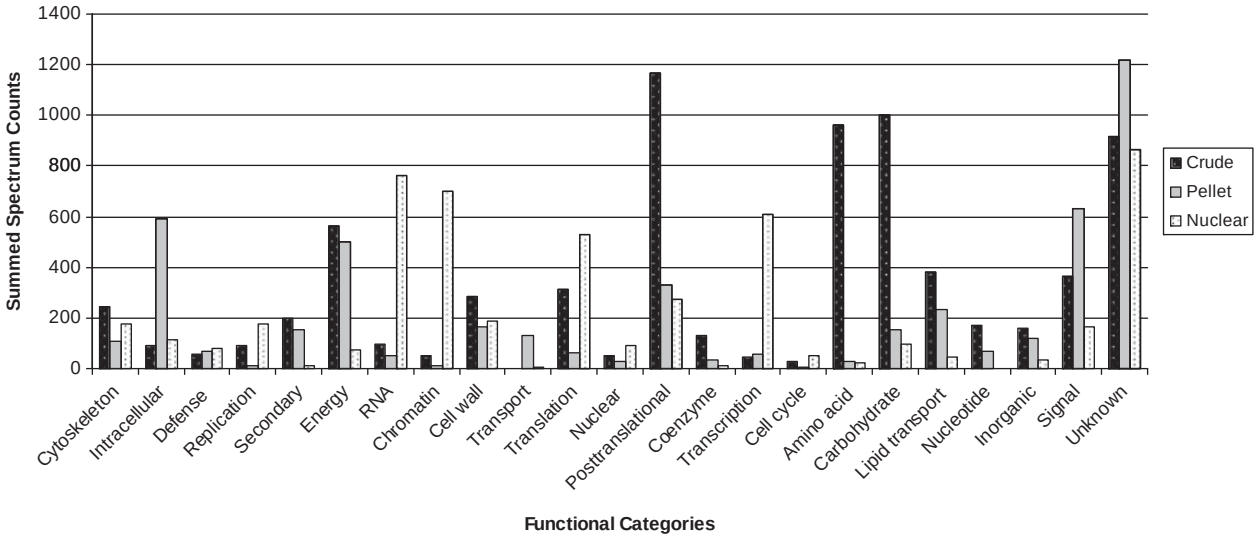


Figure 2. Functional categorization of xylem crude, pellet and nuclear proteomes. Quantitative distribution of proteins identified exclusively from either crude, pellet or nuclear xylem proteomes. For a given protein, spectrum count is the number of tandem mass spectra that are assigned to peptides from that protein, and provides an approximate indicator of protein abundance. The *x*-axis labels represent various predicted functional categories (see Section 2.8); cytoskeleton: cytoskeleton, Intracellular: intracellular trafficking, secretion, and vesicular transport, Defense: defense mechanisms, Replication: replication, recombination and repair, Secondary: secondary metabolites biosynthesis, transport and catabolism, Energy: energy production and conversion, RNA: RNA processing and modification, Chromatin: chromatin structure and dynamics, Cell wall: cell wall/membrane/envelope biogenesis, Transport: transport, Translation: translation, ribosomal structure and biogenesis, Nuclear: nuclear structure, intracellular trafficking, secretion, and vesicular transport, Posttranslational: posttranslational modification, protein turnover, chaperones, Coenzyme: coenzyme transport and metabolism, Transcription: transcription, Cell cycle: cell cycle control, cell division, chromosome partitioning, Amino acid: amino acid transport and metabolism, Carbohydrate: carbohydrate transport and metabolism, Lipid: lipid transport and metabolism, Nucleotide: nucleotide transport and metabolism, Inorganic ion: inorganic ion transport and metabolism, Unknown: unknown, Signal: signal transduction mechanisms. The *y*-axis represents the summed spectrum count of proteins in each functional category.

3.2 Functional categorization of pellet fraction proteins

The protocol applied in collecting the pellet fraction of xylem protein extracts is expected to provide enrichment of membrane proteins, but can also contain significant impurities. In this study, ~3800 proteins were experimentally identified from the pellet fraction of xylem extracts, out of which, 907 were identified exclusively from the pellet fraction. The top functional groups into which proteins found in pellet fractions are categorized are intracellular trafficking, secretion, and vesicular transport; energy production and conversion; cell wall/membrane/envelope biogenesis; post-translational modification, protein turnover, chaperones; lipid transport and metabolism; signal transduction mechanisms and unknown (Fig. 2). Of the proteins detected in our study that were predicted by TMpred to have at least three transmembrane domains (''TmPred'' column in Supporting Information Table 1), 78% were experimentally detected in the pellet fraction. Conversely, only 19% of proteins that were predicted to have at least one transmembrane domain were found exclusively from the crude fraction. Significantly higher spectrum counts were detected in the pellet fractions for the integral membrane protein,

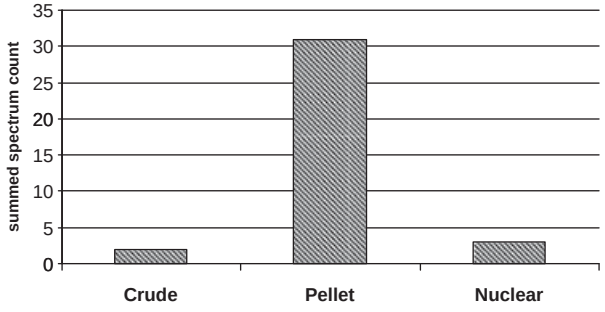


Figure 3. Quantitative estimate of cellulose synthase proteins found exclusively from crude, pellet or nuclear fractions.

cellulose synthase, as compared with crude and nuclear fractions (Fig. 3). The observed differences in spectrum count across fractions for cellulose synthase were considerably greater than that observed for trypsin, which was present in similar amounts in each LC-MS/MS analysis (6 μg for crude and pellet; 3 μg for nuclei). Spectrum counts for autolysis products of trypsin averaged 96.5 in the crude fraction, 70 in the pellet fraction and 79 in the nuclear fraction. This measurable difference in cellulose synthase is supportive of membrane enrichment in the pellet fraction.

It is known that there are several cell wall remodeling or biosynthesis genes that associate or integrate with cellular membranes. Our study shows that the proteins identified exclusively from the pellet fraction have significant over-representation in the cell wall biogenesis sub-category within the broader cell wall/membrane/envelope biogenesis functional category (Supporting Information Table 4). Such proteins include cellulose synthase (CesA) (Fig. 3), sucrose synthase (SuSy), pectinacetylesterase-pectinesterase, rhamnogalacturonatelyase, glycosyl transferases and glycosyl hydrolases (polygalacturonase) [34]. Over-representation of such proteins is further supportive of membrane enrichment in the pellet fraction.

3.3 Identification of known xylem- or cell wall development-associated proteins

Supporting Information Table 4 presents proteomics results for genes found to be enriched in xylem development or wood formation-related PopulusDB EST libraries (tension wood, shoot meristem, cambial zone, active cambium, bark, wood cell death, roots, petioles and apical shoot [30]). A sub-data set of the genes that were classified into the cell wall functional category and had strong EST support from tension wood (characterized by high cellulose content in xylem cell walls as well as higher xylem cell count) libraries contained such known cellulose biosynthesis pathway proteins as the secondary cell-wall-associated CesAs [eugene3.00040363 and gw1.XI.3218.1] and several α-TUBULIN and β-TUBULIN proteins [*e.g.* gw1.IX.2621, *Differentiable* protein] [6]. Though peptides representing several different SuSy proteins were discovered from the xylem proteome set, just two SuSy isoforms [estExt_fgenesh4_pm.C_LG_XVIII0009 and estExt_fgenesh4_pg.C_280066] were present in the sub-data set generated based on criteria of EST expression in tension wood library [6, 35–37]. Based on the high spectrum count relative to the average count of proteins in the cell wall category, both SuSy isoforms appear to be highly expressed at the protein level. Interestingly, only these two SuSy isoforms were uniquely identified (DS or DF), from the pellet fraction. It is known that these isoforms express at a higher level during enhanced cellulose biosynthesis (as occurs in xylem and under tension stress) relative to other tissue types, and it is believed that SuSy participates in primary metabolism when localized in cytosol and in secondary metabolism (cell wall synthesis), when membrane-associated [37–39]. In light of this knowledge, the identification of *Populus* sucrose synthase proteins in two replicate pellet (predominantly membrane) analyses adds credence to the theory that membrane-associated *Populus* sucrose synthase partakes in cellulose biosynthesis in xylem cells. Other previously reported wall-remodeling proteins that we identified include polygalacturonase, laccase (diphenol oxidase) and fasciclin and related adhesion glycoproteins.

3.4 Characterization of the xylem nuclear proteome

Nuclear fractionation of xylem was undertaken to identify low-abundance nuclear-localized proteins including nucleic acid binding and regulatory proteins. About 77% of WoLF PSORT-based nuclear localization predictions are correlated with experimental evidence (''wolfpsort'' column in Supporting Information Table 1). The efficiency of the nuclear enrichment method was inferred to be significantly high based on the abundance of specific nuclear marker proteins, as Fig. 4A demonstrates for histone and histone-associated proteins. Compared with crude and pellet fractions, the nuclear fraction had a significantly higher representation of proteins relating to such nuclear processes as replication, recombination and repair; nuclear structure, intracellular trafficking, secretion and vesicular transport; cell cycle control, cell division and chromosome partitioning; chromatin structure and dynamics and transcription (Fig. 5). The nuclear protein enrichment was also validated through a Western hybridization experiment using an anti-histone antibody (Fig. 4B).

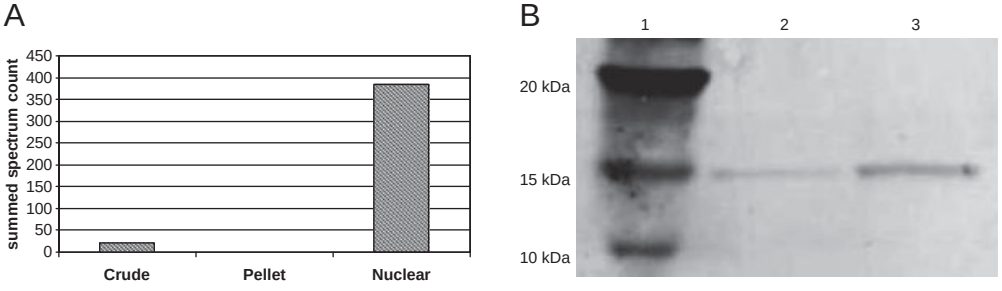


Figure 4. (A) Quantitative estimate of histone and histone-associated proteins found exclusively from crude, pellet or nuclear fractions. (B) Western blot of nuclear and crude protein fractions using anti-histone antibody. Lane 1, molecular weight standards; lane 2, 1 μg protein extracted from xylem cells; lane 3, 1 μg of protein extracted from enriched xylem nuclei. The blot was probed with anti-Histone H3 antibody.

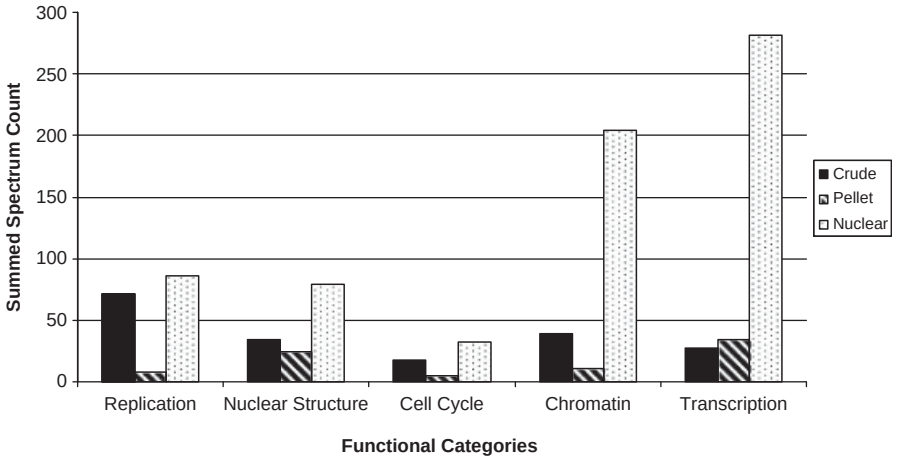


Figure 5. Distribution of exclusive crude, pellet or nuclear proteins within five functional categories that are related to nuclear processes. The data set consists of only uniquely identified proteins (DS or DF). The *x*-axis labels represent various predicted nucleus-associated functional categories. The *y*-axis represents the summed spectrum count of proteins in each functional category.

3.5 Identification of putative transcription and regulatory factors from the nuclear proteome

Transcription factors (ARF, NAC, MYB, JUMONJI, LIM, BLH3, HD-ZIP, HOMEOBOX domain proteins) and regulatory factors involved in chromatin remodeling (histone acetylates and deacetylases, SWI/SNF, SWIFT/BRCT domain proteins), RNA processing or modifying (RRM, Helicase, DEAD/DEAH, SnoRNP, PABP domain proteins) and cell cycle/replication (cyclin, retinoblastoma-associated, tudor, DNA-repair proteins) represent nearly half (44–47%) of the proteins identified exclusively from the nuclear fraction.

Interestingly, transcript-level data obtained from the *Populus* EST database (PopulusDB) suggests that many of the nuclear localized putative transcription regulators identified from xylem proteome are also preferentially represented in xylem libraries (cambium, stem, wood, tension wood, xylem and petiole libraries) (''Best EST Category'' in Supporting Information Table 1). Such candidate genes have no reported reverse or forward genetic mutant observations.

Among the proteins in the ''transcription'' category, one gene [eugene3.00140349] had a particularly high and specific EST expression support from wood-forming tissues such as cambial zone and tension wood. We found that the protein coded by this gene, a CtBP-like transcription factor, was experimentally detectable from two nuclear MS/MS analyses, as well as predicted to be nuclear by WoLF PSORT program. The closest *Arabidopsis* homolog of this gene is *ANGUSTIFOLIA* [40]. It is has been shown that the *ANGUSTIFOLIA* gene product controls the cortical tubular network and likely the expression of *MERI5* gene, a xyloglucan endotransglycosylase required in cell wall remodeling [41]. Interestingly, our xylem proteome data set also contains the closest *Arabidopsis* MERI5 homolog, gw1.XVIII.2837.1. It is reported that *Populus* xyloglucan endotransglycosylase functions in gelatinous layers of tension wood fibers even days after cell death [7]. Additionally, NAC-domain transcription factor proteins [gw1.I.5485.1 and gw1.XI.947.1] having strong homology to transcription factors known to control secondary cell wall formation in *Arabidopsis* such as SND1 [42] and NST1 [43] proteins were found exclusively in the nuclear fraction. Other interesting nuclear proteins include a Myb DNA-binding protein [gw1.IV.467.1] and a CHB3-SWI/SNF-related protein. It will be valuable to evaluate the roles of such transcription factor genes as master regulators in the *Populus* cell wall remodeling pathway.

4 Concluding remarks

The present study aimed at applying MudPIT to study *Populus* developing xylem tissue. Our results show that the technique successfully isolated and identified ~6000 proteins from xylem tissue, greatly expanding the numbers of protein identifications reported from previous *Populus* proteome studies [8–11]. Subcellular proteomics has a twofold advantage of indicating cellular contexts for functional roles as well as potentially detecting low-abundance proteins. Our attempt to enrich nuclei from xylem was successful as indicated by the presence in this fraction of a high number of detected proteins that are associated with nuclear processes. We identified in this study several candidate secondary cell wall or wood formation regulator proteins highly similar to SND1, NST1, CtCP, CHB3-SWI/SNF-related proteins. In addition, many proteins of as yet unknown function but predicted to be nuclear were experimentally found to be abundant in the nuclear proteome and also had enhanced transcript expression in wood- or xylem-related tissue contexts. Such genes are good candidates for further functional genomics investigations. Differential representation of SuSy proteins in soluble and predominantly membrane fractions was found to be consistent with the knowledge of membrane associatedness of certain SuSy isoforms.

Conserved duplicate gene pairs that originated from the salicoid duplication event in *Populus* were found to be

distinguishable at appreciable rates. Data generated by shotgun proteome profiling can also prove useful in gene annotation. Our proteome data set presented first-time-expression support for ~100 predicted *Populus* gene models, attesting to the functional validity of these gene models. Further applications of the proteomics technique will be useful in genome annotation as well as in functional classification of *Populus* genes, many of which have a representative duplicate in the genome. Moreover, the proteome data set discussed in this report, albeit from one particular plant tissue, xylem, will be useful in future efforts to predict and understand gene functions in much the same way that we have used the EST database to obtain transcript level support from wood-forming tissues.

This study presents an additional tool for systems biology investigations in *Populus*. While not explored in the current study, proteomics can also be applied to yield valuable knowledge with respect to post-translational modifications and quantitative differences in protein expression. The ability to rapidly identify and contrast whole proteomes from different cellular fractions across treatments and genetic variations along with supportive bioinformatics capabilities will be key to providing accurate systems biology models of not only cell wall biosynthesis but also other plant developmental and physiological pathways.

The authors would like to thank Manesh Shah for website assistance and Drs. Brian Davison and Stan Wullschleger for their technical reviews of this paper. The present study was enabled by research funds to U.C.K. through the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL) and the BioEnergy Science Center, a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. ORNL is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the U.S. Department of Energy.

The authors have declared no conflict of interest.

5 References

[1] Jansson, S., Douglas, C. J., *Populus:* a model system for plant biology. *Annu. Rev. Plant Biol.* 2007, *58*, 435–458.

[2] Higuchi, T., *Biochemistry and Molecular Biology of Wood*, 1st Edn., Springer-Verlag, Berlin-Heidelberg 1997.

[3] Israelsson, M., Eriksson, M. E., Hertzberg, M., Aspeborg, H. *et al*. Changes in gene expression in the wood-forming tissue of transgenic hybrid aspen with increased secondary growth. *Plant Mol. Biol.* 2003, *52*, 893–903.

[4] Karpinska, B., Karlsson, M., Srivastava, M., Stenberg, A. *et al*. MYB transcription factors are differentially expressed and regulated during secondary vascular tissue development in hybrid aspen. *Plant Mol. Biol.* 2004, *56*, 255–270.

[5] Kalluri, U. C., Joshi, C. P., Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. *Planta* 2004, *220*, 47–55.

[6] Andersson-Gunneras, S., Mellerowicz, E. J., Love, J., Segerman, B. *et al*. Biosynthesis of cellulose-enriched tension wood in *Populus:* global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J.* 2006, *45*, 144–165.

[7] Nishikubo, N., Awano, T., Banasiak, A., Bourquin, V. *et al*. Xyloglucan endo-transglycosylase (XET) functions in gelatinous layers of tension wood fibers in poplar--a glimpse into the mechanism of the balancing act of trees. *Plant Cell Physiol.* 2007, *48*, 843–855.

[8] Ferreira, S., Hjerno, K., Larsen, M., Wingsle, G. *et al*. Proteome profiling of *Populus euphratica* Oliv. upon heat stress. *Ann. Bot. (Lond.)* 2006, *98*, 361–377.

[9] Du, J., Xie, H. L., Zhang, D. Q., He, X. Q. *et al*. Regeneration of the secondary vascular system in poplar as a novel system to investigate gene expression by a proteomic approach. *Proteomics* 2006, *6*, 881–895.

[10] Plomion, C., Lalanne, C., Claverol, S., Meddour, H. *et al*. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* 2006, *6*, 6509–6527.

[11] Kieffer, P., Dommes, J., Hoffmann, L., Hausman, J. F., Renaut, J., Quantitative changes in protein expression of cadmium-exposed poplar plants. *Proteomics* 2008, *8*, 2514–2530.

[12] Link, A. J., Eng, J., Schieltz, D. M., Carmack, E. *et al*. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 1999, *17*, 676–682.

[13] Washburn, M. P., Wolters, D., Yates J. R., III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* 2001, *19*, 242–247.

[14] Koller, A., Washburn, M. P., Lange, B. M., Andon, N. L. *et al*. Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. USA* 2002, *99*, 11969–11974.

[15] Maor, R., Jones, A., Nuhse, T. S., Studholme, D. J. *et al*. Multidimensional protein identification technology (MudPIT) analysis of ubiquitinated proteins in plants. *Mol. Cell. Proteomics* 2007, *6*, 601–610.

[16] Gion, J. M., Lalanne, C., Le Provost, G., Ferry-Dumazet, H. *et al*. The proteome of maritime pine wood forming tissue. *Proteomics* 2005, *5*, 3731–3751.

[17] Lowry, O. H., Rosebrough, N. J., Farr, A. L., Randall, R. J., Protein measurement with the Folin phenol reagent. *J. Biol. Chem.* 1951, *193*, 265–275.

[18] VerBerkmoes, N. C., Shah, M. B., Lankford, P. K., Pelletier, D. A. *et al*. Determination and comparison of the baseline proteomes of the versatile microbe *Rhodopseudomonas palustris* under its major metabolic states. *J. Proteome Res.* 2006, *5*, 287–298.

[19] Mahowald, M. A., Rey, F. E., Seedorf, H., Turnbaugh, P. J. *et al*. Characterizing a model human gut microbiota

composed of members of its two dominant bacterial phyla. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 5859–5864

[20] McDonald, W. H., Ohi, R., Miyamoto, D. T., Mitchison, T. J., Yates, J. R., Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int. J. Mass Spectrom.* 2002, *219*, 245–251.

[21] Eng, J. K., Mccormack, A. L., Yates, J. R., An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[22] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[23] Moore, R. E., Young, M. K., Lee, T. D., Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* 2002, *13*, 378–386.

[24] Tabb, D. L., McDonald, W. H., Yates J. R., III, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* 2002, *1*, 21–26.

[25] Sardiu, M. E., Cai, Y., Jin, J., Swanson, S. K. *et al*. Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. *Proc. Natl. Acad. Sci. USA* 2008, *105*, 1454–1459.

[26] Liu, H., Sadygov, R. G., Yates J. R., III, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[27] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K. *et al*. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[28] Horton, P., Park, K. J., Obayashi, T., Fujita, N. *et al*. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 2007, *35*, W585–W587.

[29] Hofmann, K., Stoffel, W., Tmbase – a database of membrane spanning proteins segments. *Biol Chem. Hoppe-Seyler*, 1993, *374*, 166.

[30] Sterky, F., Bhalerao, R. R., Unneberg, P., Segerman, B. *et al*. A *Populus* EST resource for plant functional genomics. *Proc. Natl. Acad. Sci. USA* 2004, *101*:13951–13956.

[31] Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J. *et al*. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 2006, *313*, 1596–1604.

[32] Yang, X. Y., Dondeti, V., Dezube, R., Maynard, D. M. *et al*. DBParser: Web-based software for shotgun proteomic data analyses. *J. Proteome Res.* 2004, *3*, 1002–1008.

[33] Baerenfaller, K., Grossmann, J., Grobei, M. A., Hull, R. *et al*. Genome-scale proteomics reveals *Arabidopsis thaliana* gene models and proteome dynamics. *Science* 2008, *320*, 938–941.

[34] Geisler-Lee, J., Geisler, M., Coutinho, P. M., Segerman, B. *et al*. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* 2006, *140*, 946–962.

[35] Joshi, C. P., Bhandari, S., Ranjan, P., Kalluri, U. C. *et al*. Genomics of cellulose biosynthesis in poplars. *New Phytol.* 2004, *164*, 53–61.

[36] Suzuki, S., Li, L., Sun, Y. H., Chiang, V. L., The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* 2006, *142*, 1233–1245.

[37] Haigler, C. H., Ivanova-Datcheva, M., Hogan, P. S., Salnikov, V. V. *et al.* Carbon partitioning to cellulose synthesis. *Plant Mol. Biol.* 2001, *47*, 29–51.

[38] Hardin, S. C., Winter, H., Huber, S. C., Phosphorylation of the amino terminus of maize sucrose synthase in relation to membrane association and enzyme activity. *Plant Physiol.* 2004, *134*, 1427–1438.

[39] Hardin, S. C., Duncan, K. A., Huber, S. C., Determination of structural requirements and probable regulatory effectors for membrane association of maize sucrose synthase 1. *Plant Physiol.* 2006, *141*, 1106–1119.

[40] Stern, M. D., Aihara, H., Cho, K. H., Kim, G. T. *et al*. Structurally related Arabidopsis ANGUSTIFOLIA is functionally distinct from the transcriptional corepressor CtBP. *Dev. Genes Evol.* 2007, *217*, 759–769.

[41] Kim, G. T., Shoda, K., Tsuge, T., Cho, K. H. *et al.* The ANGUSTIFOLIA gene of Arabidopsis, a plant CtBP gene, regulates leaf-cell expansion, the arrangement of cortical microtubules in leaf cells and expression of a gene involved in cell-wall formation. *EMBO J.* 2002, *21*, 1267–1279.

[42] Zhong, R., Demura, T., Ye, Z. H., SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. *Plant Cell* 2006, *18*, 3158–3170.

[43] Zhong, R., Richardson, E. A., Ye, Z. H., Two NAC domain transcription factors, SND1 and NST1, function redundantly in regulation of secondary wall synthesis in fibers of Arabidopsis. *Planta* 2007, *225*, 1603–1611.

Plant Cell Tiss Organ Cult (2009) 99:251–257
DOI 10.1007/s11240-009-9596-9

ORIGINAL PAPER

Micropropagation of *Populus trichocarpa* 'Nisqually-1': the genotype deriving the *Populus* reference genome

Byung-guk Kang · Lori Osburn · Dean Kopsell · Gerald A. Tuskan · Zong-Ming Cheng

Received: 31 December 2008 / Accepted: 25 August 2009 / Published online: 10 September 2009

Abstract *Populus* serves as a model tree for biotechnology and molecular biology research due to the availability of the reference genome sequence of *Populus trichocarpa* (Torr. & Gray) genotype 'Nisqually-1'. However, 'Nisqually-1' has been shown to be very recalcitrant to micropropagation, regeneration and transformation. In this study, a highly efficient micropropagation protocol from greenhouse-grown shoot tips of 'Nisqually-1' was established. The optimal micropropagation protocol involves growing in vitro shoots in plant growth regulator-free Murashige and Skoog (MS) basal medium supplemented with 3% sucrose, 0.3% Gelrite® and 5–10 g L^{-1} of activated charcoal. Plants grown on this medium were significantly longer, and contained significantly higher concentrations of chlorophyll. This highly effective protocol provides a consistent supply of quality leaf and stem materials throughout the year for transformation experiments and other in vitro manipulations, therefore eliminating inconsistency due to seasonal and greenhouse environmental variations and the need for repetitive tissue sterilization.

Keywords Activated charcoal · Cytokinin · Gelling agent · Gelrite · Tissue culture · Poplar

B. Kang · L. Osburn · D. Kopsell · Z.-M. Cheng (✉)
Department of Plant Sciences, University of Tennessee, Knoxville, TN 37996, USA
e-mail: zcheng@utk.edu

G. A. Tuskan
Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

Abbreviations

AC	Activated charcoal
BA	6-Benzyladenine
Chl *a*	Chlorophyll *a*
MS	Murashige and Skoog
PGR	Plant growth regulator

Introduction

Populus species are widely used for wood, paper, as an energy source and for other purposes worldwide. They are now considered by the US Department of Energy to be the leading choice for dedicated woody bioenergy crops due to their fast growth, wide adaptation and ease of propagation (Tuskan 1998; Wullschleger et al. 2002). *Populus* is also regarded as the model for studying woody plant gene function because of the available reference genome of *Populus trichocarpa*, genotype 'Nisqually-1' (Tuskan et al. 2006). However, 'Nisqually-1' can be very recalcitrant to transformation with in vitro plant tissue (Ma et al. 2004), and protocols using greenhouse plants require the use of a vigorous sterilization procedure and are subject to limitation of a year-round supply of materials (Song et al. 2006). In addition, seasonal variation may lead to inconsistent plant materials which may produce variant results (Song et al. 2006).

In the process of establishing 'Nisqually-1' in vitro for regeneration and transformation research, we have encountered persistent difficulty in maintaining even basic growth of this genotype in various media used for other *Populus* species in culture (Dai et al. 2003; Ma et al. 2004). Despite its ease to root and propagate in the greenhouse and field, in vitro shoots remain green for only about 1 week and

then begin to turn chlorotic and necrotic. This limits growth and proliferation and ultimately results in plant death. Rutledge and Douglas (1988) also failed to establish in vitro cultures of *P. trichocarpa* and were unable to conduct micropropagation. Furthermore, Nadel et al. (1992) reported severe shoot tip dieback of *P. trichocarpa* and leaf yellowing in three different media, though addition of Ca-gluconic acid and 2-[*N*-morpholino] ethanesulfonic acid (MES) transiently reduced the problem. The objective of this research was to develop an effective protocol for growing 'Nisqually-1' in vitro so consistent plant material can be produced year-round for transgenic research for poplar functional genomics research, particularly for creating a mutational library and validating gene function. We tested the effects of basal medium, cytokinin concentration, gelling agent and activated charcoal (AC) on microshoot growth of 'Nisqually-1' and thus developed an optimal medium for growing and maintaining this genotype in vitro. This protocol enables us now to produce consistent and year-round materials for regeneration, transformation and other in vitro manipulation.

Materials and methods

Plant material

In vitro cultures 'Nisqually-1' were established from young shoots of greenhouse-grown plants. Vigorously growing 5-cm shoot tips were excised and then soaked sequentially in a 1% Tween-20 solution for 5 min, 70% ethanol for 1 min and in a 0.525% sodium hypochlorite (Clorox®) solution for 15 min. Explants were triple rinsed with sterile water for 5 min each time. A 2-cm shoot tip was excised from the surface sterilized shoot and placed into a 200-mL baby food jar (Sigma–Aldrich, St. Louis, MO) containing 30 mL of Murashige and Skoog (MS) (1962) basal medium supplemented with MS vitamins (PhytoTechnology Laboratories, Shawnee Mission, KS), 100 mg L^{-1} myo-inositol, 3% (w/v) sucrose, 4.4 μM N^6-benzylaminopurine (BA) and 0.8% (w/v) agar (Cat. No. BP1423, Fisher Scientific, Pittsburg, PA). The solution was adjusted to pH 5.8 prior to autoclaving at 120°C and 103.5 kPa for 20 min. This was the medium which was used to maintain in vitro aspen (*Populus* spp.) cultures in our laboratory (Dai et al. 2003). The cultures were maintained in a growth room at 25°C under a 16-h photoperiod provided by cool-white fluorescent lamps. The lamps provided a photosynthetic photon flux of 125 μmol m^{-2} s^{-1} as measured by a Licor LI-250 light meter (LI-COR Inc., Lincoln, Nebraska) held at the top of the culture vessels. All of the stock cultures in the following experiments were maintained under these conditions.

The effects of basal medium and cytokinin concentration on plant growth

To evaluate the effect of basal medium salt on plant growth, 'Nisqually-1' shoots were cultured aseptically on either MS, woody plant medium (WPM) (Lloyd and McCown 1981) or Driver and Kuniyuki walnut (DKW) (Driver and Kuniyuki 1984) medium. In each basal medium, three concentrations of BA were tested: 0.0, 2.2 and 4.4 μM (Table 1). All media were supplemented with 0.1% MS vitamins, 100 mg L^{-1} myo-inositol, 3% (w/v) sucrose and 0.8% (w/v) agar (Cat. No. BP1423, Fisher Scientific, Pittsburg, PA).

Table 1 Survival rate and performance rating of *Populus trichocarpa* 'Nisqually-1' after 4 weeks on three basal media with three concentrations of N^6-benzylaminopurine

Basal medium	BA concentration (μM)	Percentage of shoots surviving after 4 weeks	Performance rating[a,b]
MS	0	75.0	3.875 a
MS	2.2	50.0	2.750 b
MS	4.4	37.5	2.375 b
DKW	0	37.5	2.500 b
DKW	2.2	37.5	2.375 b
DKW	4.4	37.5	2.250 b
WPM	0	50.0	2.875 b
WPM	2.2	37.5	2.375 b
WPM	4.4	37.5	2.125 b

[a] Plants were rated on a 5-point scale: (5) plant survived, actively growing, no sign of senescence; (4) plant survived, limited growth, lower leaves showing senescence; (3) plant survived, no growth, showing moderate senescence; (2) plant nearly dead; and (1) plant completely dead. For each treatment, the average value was given as the performance rating

[b] The same letters in the different rows indicate that there is no significant difference ($P \leq 0.05$)

Table 2 The effect of gelling agent and N^6-benzylaminopurine concentration on the survival of *Populus trichocarpa* 'Nisqually-1' after 4 weeks

Gelling agent	BA concentration (μM)	Percentage of shoots surviving after 4 weeks	Performance rating[a,b]
Gelrite	0	100.0	4.875 a
Gelrite	2.2	62.5	3.250 bc
Gelrite	4.4	50.0	3.125 bc
Agar	0	75.0	3.875 b
Agar	2.2	50.0	2.750 c
Agar	4.4	37.5	2.375 c

[a] Plants were rated on a 5-point scale: (5) plant survived, actively growing, no sign of senescence; (4) plant survived, limited growth, lower leaves showing senescence; (3) plant survived, no growth, showing moderate senescence; (2) plant nearly dead; and (1) plant completely dead. For each treatment, the average value was given as the performance rating

[b] The same letters in the different rows indicate that there is no significant difference ($P \leq 0.05$)

The effect of gelling agent on plant growth

Once the effects of basal medium and cytokinin concentration on plant growth were evaluated, MS medium was selected as the basal medium for determining the effect of gelling agent on the growth of 'Nisqually-1' shoots. Two gelling agents were tested: 0.8% (w/v) agar (Fisher Scientific) and 0.3% (w/v) Gelrite® (PlantMedia, Dublin, OH). The culture medium contained either 0.0, 2.2 or 4.4 μM BA (Table 2). The surviving 'Nisqually-1' shoots from previous experiments were used in this experiment.

The effect of activated charcoal on plant growth

To evaluate the effect of AC on plant growth, 'Nisqually-1' shoots were cultured on MS basal medium without plant growth regulators (PGRs) and with 0.3% (w/v) Gelrite®. Four AC concentrations, 0, 3, 5 and 10 g L^{-1}, were tested (Fig. 1). AC was added to the culture medium after adjusting the pH to 5.8, prior to autoclaving.

Experimental design, data collection and statistical analysis

In the first two experiments (the effects of basal medium and cytokinin concentration on plant growth and the effect of gelling agent on plant growth), four 200-mL baby food jars containing one shoot per jar were used for each treatment, and both experiments were repeated three times. Plant performance was evaluated at 4 weeks for survival rate, leaf chlorosis and overall appearance according to the following numeric criteria: (5-points) plant survived, actively growing, no sign of senescence; (4-points) plant survived, limited growth, lower leaves showing senescence; (3-points) survived, but no growth, showing senescence; (2-points) nearly dead; and (1-point) completely dead. For each treatment, the average of the replicates was given as the performance rating. An additional observation was made at 6 weeks, but data are not shown because many plants died during the last 2-week period. In the experiment that tested AC on plant growth, the explants were aseptically grown in 9-cm diameter × 14-cm high (1,000 ml) culture vessels (PhytoTechnology Laboratories). Four explants were used for each treatment and the experiment was repeated three times. The length of each plant was measured every 2 weeks for 6 weeks. The experimental design was a Completely Randomized Design (CRD), and the performance rating and length data were evaluated by analysis of variance (ANOVA) using SAS version 9.1 (SAS Institute Inc, Cary, NC).

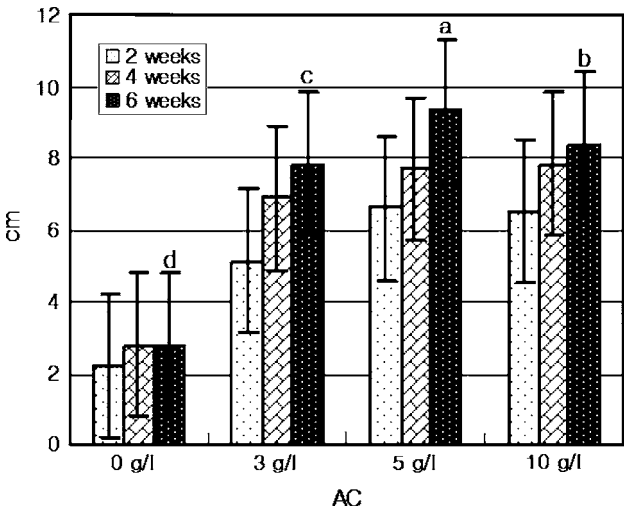


Fig. 1 Mean length of 'Nisqually-1' shoots grown on MS PGR-free Gelrite medium supplemented with either 0, 3, 5 or 10 g L^{-1} activated charcoal (AC) after 2, 4 and 6 weeks. Shoots were grown in 9 × 14 cm (1,000 ml) plastic vessels in sterile conditions at 25°C and 16 h photoperiod. Data are the means of 16 shoots. The *bar* represents the standard error of the mean. The *different letters above the bars* at 6 weeks (the final data collection date) indicate the separation of the treatment means

Chlorophyll content analysis by high-performance liquid chromatography (HPLC)

Since the plants grown in the medium supplemented with AC were significantly longer and appeared much healthier and greener in color than those grown in media without AC, leaf tissues were analyzed for chlorophyll content. Four plants from each of the four in vitro treatments were sampled. We also analyzed the chlorophyll content of plants grown on potting mix without AC in a growth chamber under fluorescent and incandescent lights. 'Nisqually-1' leaf tissues were lyophilized for no less than 48 h (Model 12 L FreeZone; LabConCo, Kansas City, MO) and stored at −80°C prior to extraction and analysis according to Kopsell et al. (2004) and analyzed according to Emenhiser et al. (1996). Briefly, a 0.1-g sample from each homogenate was re-hydrated with 0.8 mL of ultra pure H_2O and placed in a water bath set at 40°C for 20 min. After incubation, 0.8 mL of the internal standard ethyl-β-8′-apo-carotenoate (Sigma) was added to determine extraction efficiency. After sample hydration, there was an addition of 2.5 mL of tetrahydrofuran (THF) stabilized with 25 mg L^{-1} of 2,6-Di-*tert*-butyl-4-methoxyphenol (BHT). Samples were then homogenized in a Potter–Elvehjem (Kontes, Vineland, NJ) tissue grinding tube. During homogenization, the tube was immersed in ice to dissipate heat. The tube was then placed into a clinical centrifuge for 3 min at $500\times g_n$. The supernatant was decanted and the sample pellet was re-suspended in 2 mL THF and homogenized again with the same extraction technique. The procedure was repeated for a total of four extractions per each sample to obtain a colorless supernatant. The combined sample supernatants were reduced to 0.5 mL under a stream of nitrogen gas (N-EVAP 111; Organomation Inc., Berlin, MA), and brought up to a final volume of 5 mL with methanol (MeOH). A 2-mL aliquot was filtered through a 0.2-μm polytetrafluoroethylene (PTFE) filter (Model Econofilter PTFE 25/20, Agilent Technologies, Wilmington, DE.) using a 5-mL syringe (Becton, Dickinson and Company, Franklin Lakes, NJ) prior to HPLC analysis.

An Agilent 1200 series HPLC unit with a photodiode array detector (Agilent Technologies, Palo Alto, CA) was used for pigment separation. Chromatographic separations were achieved using an analytical scale (4.6 mm i.d. × 250 mm) 5 μm, 200 Å polymeric C_{30} reverse-phase column (ProntoSIL, MAC-MOD Analytical Inc., Chadds Ford, PA), which allowed for effective separation of chemically similar pigment compounds. The column was equipped with a guard cartridge (4.0 mm i.d. × 10 mm) and holder (ProntoSIL), and was maintained at 30°C using a thermostatted column compartment. All separations were achieved isocratically using a binary mobile phase of 11% methyl *tert*-butyl ethanol (MTBE), 88.9% MeOH and 0.1% triethylamine (TEA) (v/v). The flow rate was 1.0 mL min^{-1}, with a run time of 55 min, followed by a 2 min equilibration prior to the next injection. Eluted pigments and chemically similar pigment compounds from a 10 μL injection were detected at 453 (carotenoids and internal standard) and 652 [chlorophyll *a* (Chl *a*)] nm, and data were collected, recorded, and integrated using ChemStation Software (Agilent Technologies). Peak assignment for individual pigments was performed by comparing retention times and line spectra obtained from photodiode array detection using external standards (ChromaDex Inc., Irvine, CA).

Results

Basal medium and cytokinin effects

Shoots grown on MS medium without BA had higher survival rates at 4 weeks than those grown in MS medium with 2.2 or 4.4 μM BA, or in WPM or DKW medium with or without BA (Table 1). Shoots grown on MS medium without BA also appeared healthier and greener and became chlorotic and necrotic about 2 weeks later than those in other media (photographs not shown).

Gelling agent effect

It was clear that significantly more shoots grown on PGR-free medium with Gelrite as the gelling agent survived to week four and had significantly higher performance ratings than those grown in the medium solidified with agar (Table 2). The shoots remained green and produced new growth (photographs not shown). In Gelrite-containing media without BA, shoot survival rates were significantly higher than those grown on medium with either 2.2 or 4.4 μM BA, confirming that shoots grew better on PGR-free medium which was more favorable than BA-containing medium.

Activated charcoal effect

Although the shoots grown on MS basal medium solidified with Gelrite had higher survival rates, growth was still limited and not sustained. Addition of AC into Gelrite-containing medium significantly improved the growth of 'Nisqually-1'. In the medium without AC, all 'Nisqually-1' shoots survived to 4 weeks, but at 6 weeks survival was reduced to 50%. Shoots grew to an average of 2.2 and 2.8 cm at 4 and 6 weeks, respectively (Fig. 1). In contrast, all of the 'Nisqually-1' shoots grown on AC-containing media survived to week 6, and they grew significantly more than those on the AC-free medium (Fig. 1). At week 6, the 'Nisqually-1' shoots cultured on the medium with 3, 5 and

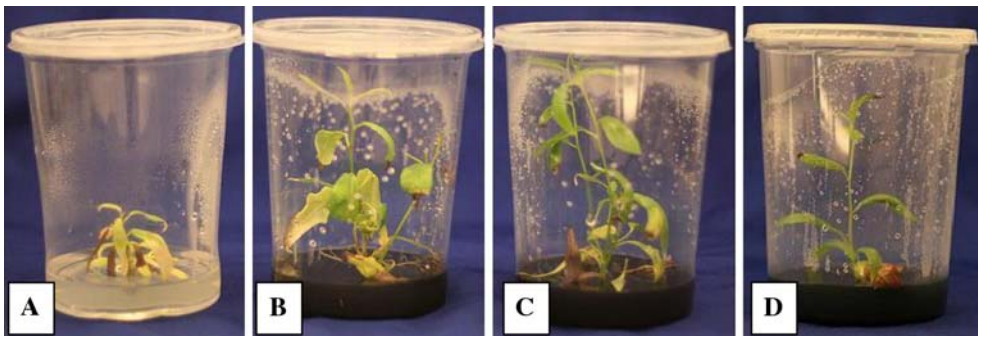


Fig. 2 'Nisqually-1' grown on MS basal medium, solidified with Gelrite and supplemented with either **a** 0, **b** 3, **c** 5 or **d** 10 g L^{-1} activated charcoal. Cultures were maintained in a growth room at 25°C and a 16-h photoperiod where fluorescent light intensity was 125 μmol $m^{-2}s^{-1}$. *Photographs* were taken at 6 weeks

10 g L^{-1} AC grew to 7.8, 9.3 and 8.4 cm, respectively (Fig. 1). All of the 'Nisqually-1' plants grown on AC-containing medium produced two to three green and healthy shoots (Fig. 2). Using this medium, we have maintained the 'Nisqually-1" shoots for more than 2 years (data not presented).

HPLC assay of chlorophyll a

'Nisqually-1' grown on basal MS medium with all AC-containing media appeared to have greener leaves (Fig. 2). This observation correlated with the chl *a* concentration of the leaves (Table 3). 'Nisqually-1' leaves from shoots grown on the medium containing 10 g L^{-1} AC had the highest concentration of chl *a* [1056.95 μg chl *a* g^{-1} dry weight (DW)], much higher than those grown on the medium containing 5 g L^{-1} AC (779.73 μg chl *a* g^{-1} DW), both being maintained in a growth room with 125 μmol m^{-2} s^{-1} fluorescent light. The control plants grown without AC in potting mix in a growth chamber had 507.68 μg chl *a* g^{-1} DW (data not presented), which was significantly less than those grown on media containing 5 or 10 g L^{-1} AC, but higher than those grown on culture media with 0 and 3 g L^{-1} AC.

Discussion

Although many *Populus* species are generally relatively easy to grow and propagate in the greenhouse as well as in tissue culture (Son et al. 2000; Dai et al. 2003), we have encountered extreme difficulty in maintaining the in vitro culture of the genotype 'Nisqually-1' of *P. trichocarpa* in a common medium that contains cytokinin and is solidified with agar (Dai et al. 2003). In previous reports with *P. trichocarpa*, extensive difficulty of in vitro propagation has also been reported (Rutledge and Douglas 1988; Nadel et al. 1992). Rutledge and Douglas (1988) failed to initiate shoots from *P. trichocarpa* meristems and were unable to perform micropropagation. Nadel et al. (1992) was able to reduce leaf yellowing and meristem dieback of *P. trichocarpa* in culture by growing on half strength medium supplemented with Ca-gluconic acid and 2-[*N*-morpholino] ethanesulfonic acid (MES), but the effect was transient only for a few subculture cycles. In this research, we have developed an effective and efficient method to propagate this important genotype of which the whole genome is sequenced. There are three key factors which contributed to the optimization of the protocol, namely, use of PGR-free MS medium, use of Gelrite as the gelling agent and addition of 5 or 10 g L^{-1} AC.

The basal medium type can affect the performance of woody plants in vitro (Mackay and Kitto 1988; Nadel et al. 1992; Cheng et al. 2000; Dai et al. 2005). Our results showed MS basal medium was more suitable than WPM and DKW for growing 'Nisqually-1'. The MS medium is known for its rich macro- and micro-elements, particularly nitrogen (Murashige and Skoog 1962). Although WPM was developed for micropropagating woody plants (Lloyd and McCown 1981) and DKW medium was developed for

Table 3 Visible absorption spectra of Chlorophyll *a* of 'Nisqually-1' by HPLC

AC treatment (g L^{-1})[a]	Sample dry wt (g)	HPLC % recovery	Recovered chlorophyll *a* (μg/g dry weight)
0	0.2812	0.91	48.91
3.0	0.0653	0.74	182.36
5.0	0.0736	0.79	779.73
10.0	0.0484	0.82	1056.95
Control (on soil, without AC)	0.0276	0.83	507.68

[a] Plants were grown on PGR-free MS basal medium, solidified with Gelrite, and supplemented with either 0, 3, 5 or 10 g L^{-1} activated charcoal (AC). Cultures were placed in a growth room where light intensity was 125 μmol m^{-2} s^{-1} provided by fluorescent light. The control plants were grown on soil without AC in a growth chamber where light intensity was 221 μmol m^{-2} s^{-1} provided by fluorescent and incandescent lighting

propagating walnut (*Juglans nigra*) (Driver and Kuniyuki 1984), both of these media were not well-suited for growing 'Nisqually-1'. Nadel (1992) also reported that MS medium was a suitable medium, although less effective than ½-strength MS.

Gelling agents can significantly affect the performance of in vitro culture (MacCrae and Van Staden 1990; Cheng and Shi 1995). The performance of 'Nisqually-1' in agar and Gelrite media were similar to that observed with Siberian elm [*Ulmus pumila* (Cheng and Shi 1995)], where shoots in agar-solidified medium deteriorated in 1 week, but fully recovered when transferred to Gelrite medium, while those in Gelrite medium deteriorated in 1 week after being transferred to agar-gelled medium. This deterioration phenomenon appears to be specific to 'Nisqually-1' because many other *Populus* species have been grown in agar-containing medium without this problem (Rutledge and Douglas 1988; Nadel et al. 1992; Son et al. 2000; Dai et al. 2003). In other species, agar medium was more favorable than Gelrite medium for shoot subculture of French tarragon [*Artemisia dracunculus* (Mackay and Kitto 1988)] and Asian white birch [*Betula platyphylla* (Cheng et al. 2000)]. Such a dramatic inhibitory effect may be attributed to the species-specific over-sensitivity of 'Nisqually-1' to microelements such as copper (Debergh 1983; Cheng and Shi 1995) in unpurified agar.

Activated charcoal clearly had a significant effect on improving growth of 'Nisqually-1'. Activated charcoal can improve development and growth of many plant species in vitro (Pan and Staden 1998), for example, establishment of lisianthus (*Eustoma grandiflorum*) protoplast culture (Kunitake et al. 1995; Teng 1997), spruce (*Picea abies*) somatic embryo development (Pullman et al. 2005), lily (*Lilium longiflorum*) bulb formation (Han et al. 2004) and embryogenesis of *Brassica oleracea* (da Silva Dias 1999). The beneficial effect of AC is thought to be attributed to its adsorption of inhibitory substances in the culture medium (Fridborg et al. 1978; Weatherhead et al. 1978, 1979). The positive effect of AC is also considered to be due to a reduction of the toxic effects of cytokinins and auxins in some cases (Weatherhead et al. 1978; Zaghmout and Torello 1988; Pan and Staden 1998), therefore, altering ratios of culture medium components and influencing plant growth in vitro (Johansson 1983; Druart and Wulf 1993). However, the enhancement of AC on 'Nisqually-1' is unlikely due to such an effect because our best performing medium lacks of exogenous hormones. One of the likely reasons for enhanced growth may be due to the higher content of chl *a*, responsible for greener leaves. Genomic analysis has revealed signaling functions among chlorophyll biosynthetic pathway intermediate compounds which regulate transcriptional production of light-harvesting chlorophyll-binding proteins such as carotene and xanthophyll carotenoids (Lohr et al. 2005). Furthermore, the biosynthesis of chlorophyll molecules is linked to the occurrence and production of light-harvesting complex polypeptides (Xu et al. 2001). Together, genomic and analytical data demonstrate the close connections between chlorophyll and carotenoid biosynthetic pathways. It is widely held that there is a positive correlation between the chlorophyll in leaves and the growth rate of plants (Gupta and Durzan 1984). Moreover, positive correlations between chlorophyll and carotenoid pigment concentrations in plants have been established (Kopsell et al. 2004). Since carotenoid pigments function in light-harvesting and photoprotection, it is also possible that elevated chlorophylls and carotenoids could impart greater fitness to the micropropagated plants. The dark environment provided by AC may also contribute to enhanced growth by promoting early root growth, thus allowing shoots to absorb nutrients early because 'Nisqually-1' produced more adventitious roots and produced roots early (data not presented).

Acknowledgments This project was supported in part by DOE-Bioenergy Center (BESC) grant, by the US Department of Energy/Oak Ridge National Laboratory (subcontract to Z.-M.C.), and by the Tennessee Agricultural Experiment Station. The BESC is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

References

Cheng Z-M, Shi N-Q (1995) Micropropagation of mature Siberian elm in two steps. Plant Cell, Tissue Organ Cult 41:197–199

Cheng Z-M, Schnurr JP, Dai WH (2000) Micropropagation by shoot tip culture and regeneration from leaf explants of *Betula platyphylla* 'Fargo'. J Environ Hortic 18:119–122

da Silva Dias JC (1999) Effect of activated charcoal on *Brassica oleracea* microspore culture embryogenesis. Euphytica 108: 65–69

Dai WH, Cheng Z-M, Sargent WA (2003) Regeneration and Agrobacterium-mediated transformation of two elite aspen hybrid clones from in vitro leaf tissues. In Vitr Cell Dev Biol Plant 39:6–11

Dai WH, Jacques V, Herman D, Cheng ZM (2005) Micropropagation of a cold hardy selection of *Cercis canadensis* L. through single-node culture. J Environ Hortic 23:54–58

Debergh PC (1983) Effects of agar brand and concentration on the tissue culture medium. Physiol Plant 59:270–276

Driver JA, Kuniyuki AH (1984) In vitro propagation of 'Paradox' walnut rootstock. HortScience 19:507–509

Druart P, Wulf O (1993) Activated charcoal catalyses sucrose hydrolysis during autoclaving. Plant Cell, Tissue Organ Cult 32:97–99

Emenhiser C, Simunovic N, Sander LC, Schwartz SJ (1996) Separation of geometric carotenoid isomers in biological extracts using a polymeric C30 column in reverse-phase liquid chromatography. J Agric Food Chem 44:3887–3893

Fridborg G, Pedersen M, Landstorm L-E, Eriksson T (1978) The effect of activated charcoal on tissue culture: absorption of metabolites inhibiting morphogenesis. Physiol Plant 43:104–106

Gupta PK, Durzan DJ (1984) Plant regeneration via somatic embryogenesis from subcultured callus of mature embryos of *Picea abies* (NORWAY SPRUCE). In Vitr Cell Dev Biol 22:685–688

Han BH, Yu HJ, Yae BW, Peak KY (2004) In vitro micropropagation of *Lilium longiflorum* 'Georgia' by shoot formation as influenced by addition of liquid medium. Sci Hortic 103:39–49

Johansson L (1983) Effects of activated charcoal in anther cultures. Physiol Plant 59:397–403

Kopsell DA, Kopsell DE, Lefsrud MG, Curran-Celentano JLD (2004) Variation in lutein, beta-carotene, and chlorophyll concentrations among *Brassica oleracea* cultigens and seasons. HortScience 39:361–364

Kunitake H, Nakashima T, Mori K, Tanaka M, Mii M (1995) Plant regeneration from mesophyll protoplasts of lisianthus (*Eustoma grandiflorum*) by adding activated charcoal into protoplast culture medium. Plant Cell, Tissue Organ Cult 43:59–65

Lloyd G, McCown B (1981) Commercially feasible micropropagation of mountain laurel, *Kalmia latifolia*, by use of shoot-tip cultures. Comb Proc Int Plant Propag Soc 30:421–426

Lohr M, Im C, Grossman AR (2005) Genome-based examination of chlorophyll and carotenoid biosynthesis in *Chlamydomonas reinhardtii*. Plant Physiol 138:490–515

Ma C, Strauss SH, Meilan R (2004) *Agrobacterium*-mediated transformation of the genome-sequenced poplar clone, 'Nisqually-1' (*Populus trichocarpa*). Plant Mol Biol Rep 22:311–312

MacCrae S, Van Staden J (1990) In vitro culture of *Eucalyptus grandid*: effect of gelling agents on micropropagation. J Plant Physiol 137:249–251

Mackay WA, Kitto SL (1988) Factors affecting in vitro shoot proliferation of French tarragon. HortScience 113:282–287

Murashige T, Skoog F (1962) A revised medium for rapid growth and bioassays with tobacco tissue cultures. Physiol Plant 15:473–497

Nadel BL, Hazen G, David R, Huttermann A, Altman A (1992) In vitro propagation of *Populus* species: response to growth regulators and media composition. Acta Hortic 314:61–68

Pan MJ, Staden JV (1998) The use of charcoal in in vitro culture—a review. Plant Growth Regul 26:155–163

Pullman GS, Gupta PK, Timmis R, Carpenter C, Kreitinger M, Welty E (2005) Improved Norway spruce somatic embryo development through the use of abscisic acid combined with activated carbon. Plant Cell Rep 24:271–279

Rutledge CB, Douglas GC (1988) Culture of meristem tips and micropropagation of 12 commercial clones of poplar in vitro. Physiol Plant 72:367–373

Son SH, Park YG, Chun YW, Hall RB (2000) Germplasm preservation of *Populus* through in vitro culture systems. In: Klopfenstein NB, Chun YW, Kim N-S, Ahuja MR (eds) Micropropagation, genetic engineering, and molecular biology of *Populus*. USDA Forest Service, pp 44–49

Song J, Lu S, Chen Z-Z, Lourenco R, Chiang VL (2006) Genetic transformation of *Populus trichocarpa* genotype Nisqually-1: a functional genomic tool for woody plants. Plant Cell Physiol 47:1582–1589

Teng WL (1997) Activated charcoal affects morphogenesis and enhances sporophyte regeneration during leaf cell suspension culture of *Platycerium bifurcatum*. Plant Cell Rep 17:77–83

Tuskan GA (1998) Short-rotation woody crop supply systems in the United States: what do we know and what do we need to know? Biomass Bioenerg 14:307–315

Tuskan GA, DiFazio SP, Hellsten U, Jansson S, Rombauts S et al (2006) The genome of western black cottonwood, *Populus trichocarpa* (Torr & Gray ex Brayshaw). Science 313:1596–1604

Weatherhead MA, Burdon J, Henshaw GG (1978) Some effects of activated charcoal as an additive to plant tissue culture media. Z Pflanzenphysiol 89:141–147

Weatherhead MA, Burdon J, Henshaw GG (1979) Effects of activated charcoal as an additive to plant tissue culture media: part 2. Z Pflanzenphysiol 94:399–405

Wullschleger SD, Jansson S, Taylor G (2002) Genomics and forest biology: *Populus* emerges as the perennial favorite. Plant Cell 14:2651–2655

Xu H, Vavilin D, Vermass W (2001) Chlorophyll *b* can serve as the major pigment in functional photosystem II complexes of cyanobacteria. Proc Natl Acad Sci USA 98:14168–14173

Zaghmout OMF, Torello WA (1988) Enhanced regeneration in long-term callus cultures of red fescue by pretreatment with activated charcoal. HortScience 23:615–616

Journal of Experimental Botany, Page 1 of 13
doi:10.1093/jxb/erp252
This paper is available online free of all access charges (see http://jxb.oxfordjournals.org/open_access.html for further details)

Journal of Experimental Botany
www.jxb.oxfordjournals.org

RESEARCH PAPER

Function of *Arabidopsis* hexokinase-like1 as a negative regulator of plant growth

Abhijit Karve* and Brandon d. Moore†

Department of Genetics and Biochemistry, Clemson University, Clemson SC 29634, USA

Received 21 January 2009; Revised 3 July 2009; Accepted 29 July 2009

Abstract

A recent analysis of the hexokinase (HXK) gene family from *Arabidopsis* revealed that three hexokinase-like (HKL) proteins lack catalytic activity, but share about 50% identity with the primary glucose (glc) sensor/transducer protein AtHXK1. Since the AtHKL1 protein is predicted to bind glc, although with a relatively decreased affinity, a reverse genetics approach was used to test whether HKL1 might have a related regulatory function in plant growth. By comparing phenotypes of an HKL1 mutant (*hkl1-1*), an HXK1 mutant (*gin2-1*), and transgenic lines that overexpress HKL1 in either wild-type or *gin2-1* genetic backgrounds, it is shown that HKL1 is a negative effector of plant growth. Interestingly, phenotypes of HKL1 overexpression lines are generally very similar to those of *gin2-1*. These are quantified, in part, as reduced seedling sensitivity to high glc concentrations and reduced seedling sensitivity to auxin-induced lateral root formation. However, commonly recognized targets of glc signalling are not apparently altered in any of the HKL1 mutant or transgenic lines. In fact, most, but not all, of the observed phenotypes associated with HKL1 overexpression occur independently of the presence of HXK1 protein. The data indicate that HKL1 mediates cross-talk between glc and other plant hormone response pathways. It is also considered Whether a possibly decreased glc binding affinity of HKL1 could possibly be a feedback mechanism to limit plant growth in the presence of excessive carbohydrate availability is further considered.

Key words: Auxin, glucose signalling, growth regulation, GUS staining, hexokinase, hexokinase-like, hypocotyl elongation, plant hormones.

Introduction

All living organisms have complex regulatory networks that enable them to sense their nutrient status and to adjust their growth and development accordingly. Glucose (glc) is an important metabolic nutrient, which also functions as a signalling molecule that regulates gene expression in a variety of organisms (Towle, 2005; Rolland *et al.*, 2006; Gancedo, 2008). In plants, glc affects the expression of more than 1000 genes involved in a diverse array of biological processes (Price *et al.*, 2004; Osuna *et al.*, 2007). Many of the glc-regulated genes are involved in phytohormone biosynthesis and response pathways which control plant growth (Gibson, 2004). Furthermore, genetic studies indicate that many mutants of plant glc signalling are alleles of genes with defined roles in ABA or ethylene biosynthesis, or their signalling networks (Leon and Sheen, 2003; Rognoni *et al.*, 2007).

Genetic and biochemical evidence indicates that two *Arabidopsis* proteins, hexokinase1 (HXK1) and the regulator of G-protein signalling1 (RGS1), have independent roles in glc sensing and phytohormone responses (Rolland *et al.*, 2006). As a glc sensor, AtHXK1 modulates plant growth at many different developmental stages (Moore *et al.*, 2003). A null mutant of AtHXK1, *gin2-1*, has reduced shoot and root growth, increased apical dominance, delayed flowering, and

* Present address: Environmental Sciences Division, Oak Ridge National Laboratory, 1 Bethel Valley Road, Oak Ridge, TN 37831, USA.
† To whom correspondence should be addressed: E-mail: moore8@clemson.edu

Abbreviations: GFP, green fluorescent protein; GUS, glucuronidase; HA, haemagglutinin; HKL, hexokinase-like; HXK, hexokinase; LD, long day; LUC, luciferase; NAA, naphthalene acetic acid; NPA, 1-naphthylphthalamic acid; PPDK, pyruvate orthophosphate dikinase; RBCS, ribulose-1,5-bisphosphate carboxylase small subunit; RT-PCR, reverse transcriptase polymerase chain reaction; SD, short day; UBQ, ubiquitin.

altered sensitivities to auxin, cytokinin, and glc (Moore *et al.*, 2003). On the other hand, AtRGS1 has been suggested to function as a glc binding protein that can attenuate cell division in primary root apical meristems through its interaction with GPA1, a heterotrimeric G-protein subunit (Chen *et al.*, 2006; Johnston *et al.*, 2007). Similar to HXK1, plant heterotrimeric G-proteins also affect a diverse array of developmental and hormone responses (Perfus-Barbeoch *et al.*, 2004). However, even though seedlings of null mutants of both AtHXK1 and AtRGS1 fail to undergo normal glc-dependent cell cycle arrest, their hypocotyl elongation responses at low light are opposite to each other (Chen *et al.*, 2003; Moore *et al.*, 2003).

Plant HXKs are encoded by a modest family of about 5–10 genes (Claeyssen and Rivoal, 2007). HXK proteins are reported to occur in the cytosol, mitochondria, plastids, nuclei, and Golgi (Balasubramanian *et al.*, 2007, and references therein). AtHXK1 is predominantly associated with the mitochondria, but also reportedly can occur in the nucleus (Cho *et al.*, 2006). There is evidence that from both locations, AtHXK1 can regulate gene and/or protein expression, but there are questions regarding both scenarios (see Balasubramanian *et al.*, 2008). In rice, OsHXK5 and OsHXK6 have been shown recently to act as glc sensors and similarly to have a predominantly mitochondrial association, but possible nuclear function (Cho *et al.*, 2009). However, in contrast to AtHXK1, both OsHXK5 and OsHXK6 do contain a predicted nuclear localization signal.

A recent analysis of the *Arabidopsis* HXK gene family revealed that three of the six members lack catalytic activity when assayed with varying concentrations of glc or fructose (Karve *et al.*, 2008). These were designated as hexokinase-like (HKL) proteins since they also are about 50% identical to AtHXK1. The basis for the lack of catalytic activity in the HKL proteins was attributed to a number of changes throughout the primary sequences and not to any specific single amino acid change. Known functional domains and key residues are reasonably well conserved in AtHKL1 (At1g50460) and AtHKL2 (At3g20040), and both proteins can probably bind glc (Karve *et al.*, 2008). However, sequence divergence in AtHKL3 (At4g37840) is so extensive that the protein might not bind either glc or ATP. Interestingly, all three *Arabidopsis* HKL proteins have a mitochondrial targeting peptide which is very similar to that of AtHXK1. Experimental evidence for their mitochondrial association has been shown by using a proteomics approach (Heazlewood *et al.*, 2004) and by examining the cellular expression of C-terminal GFP fusion proteins (Karve *et al.*, 2008).

Non-catalytic HXKs have been reported in fungi and possibly occur commonly among higher plants (A Virnig and Bd Moore, unpublished data). The fungal HKL proteins have divergent roles including one as a meiosis-specific transcription factor in *Saccharomyces cerevisiae* (Daniel, 2005) and others as regulators of a carbon starvation response in *Aspergillus nidulans* (Bernardo *et al.*, 2007). Despite the reports of the presence of HKL proteins in evolutionarily diverse species, their lack of catalytic activity has made it challenging to define their functions. In this study, a reverse genetics approach was used to determine whether AtHKL1 might have a role in plant growth, perhaps as an effector of glc signalling. Analyses of phenotypes from gain-of-function *Arabidopsis* plants and from an identified mutant line with a T-DNA insertion in HKL1, show that HKL1 is a negative regulator of plant growth and that it affects seedling growth responses to glc and auxin. However, HKL1 does not affect glc signalling, as shown in protoplast transient expression assays and by seedling candidate gene expression assays. These data indicate that AtHKL1 has an important role in plant growth and development, perhaps by mediating cross-talk between glc and hormone response pathways.

Materials and methods

Plant material and growth conditions

Seeds of *Arabidopsis thaliana* ecotype Columbia (Col-0), ecotype Landsberg *erecta* (L*er*), and a Col line with a T-DNA insertion within the *HKL1* locus (At1g50460; line WISCDSLOX383A5; hereafter designated *hkl1-1*) were obtained from the *Arabidopsis* Biological Resource Center (Ohio State University). Seeds of maize (*Zea mays* L.) were purchased (Seed Genetics, Lafayette, IN). Lines for *gin2-1*, *tir1*, and transgenic lines expressing HXK1-HA or HXK1-FLAG were as previously described (Moore *et al.*, 2003). A homozygous line containing the T-DNA insertion in the *HKL1* gene was identified by PCR genotyping using the following primers: p745 (5′-AACGTCCGCAATGTGT-TATTAAGTTG-3′) and HKL1A5RP (5′-CCGTGTT-ATCTGAGCCTTACG-3′) for the T-DNA insertion allele; and, HKL1A5LP (5′-TGCAAACAAATTTAACGGCTC-3′) and HKL1A5RP for the WT allele. The insertion position in the *hkl1-1* mutant was mapped by sequencing the PCR product obtained by the primers L1WLP (5′-TGCAAACAAATTTAACGGCTC-3′) and L1WRP (5′-CCGTGTTATCTGAGCCTTACG-3′), using *hkl1-1* genomic DNA as template.

Arabidopsis seeds were surface-sterilized and stratified for 2 d at 4 °C as in Jang *et al.* (1997). Plants grown in soil were in a growth chamber (125 μmol m^{-2} s^{-1}, 22 /20 °C day/night temperature) at either a 12 h photoperiod (normal), an 8 h photoperiod (short day, SD), or a 16 h photoperiod (long day, LD). Plants were also grown for some assays on 1× MS agar plates (modified basal medium with Gamborg vitamins; PhytoTechnology Laboratories, Shawnee Mission, KS) at pH 5.7, normally with 0.5% sucrose, and under constant light (30 μmol m^{-2} s^{-1}). For glc repression assays, seedlings were grown on 1× MS plates with a substituted carbon source as 3–7% glc or 3–7% mannitol, for 7 d under constant light. Hypocotyl elongation assays were done at reduced light and nutrients as described before (Moore *et al.*, 2003). For the assay of auxin-induced lateral root formation, *Arabidopsis* seeds were grown on 1× MS plates with 0.5% sucrose plus 5 μM 1-naphthylphthalamic acid (NPA) for 5 d and then were transferred to sucrose plates

with or without 0.1 μM naphthalene acetic acid (NAA) for 5 d (Chen *et al.*, 2003).

To perform glc signalling assays by candidate gene expression, 15–20 seedlings were grown in 125 ml flasks containing 50 ml of half-strength MS medium supplemented with 1% sucrose. Seedlings were grown on a rotary shaker at 250 rpm under constant light (70 μmol m^{-2} s^{-1}) at 22 °C for 7 d. Seedlings were then washed with sugar-free half-strength MS medium for 24 h in the dark while shaking, and subsequently transferred to the light in fresh sugar-free medium (control) or in medium plus 2% glc. Seedlings were treated under constant light with shaking for 8 h, and were then harvested by quickly blotting with filter paper before freezing in liquid N_2.

Plasmid constructs

RBCS-LUC, *PPDK-LUC*, and *UBQ10-GUS* constructs have been described previously (Schaffner and Sheen, 1991; Balasubramanian *et al.*, 2007). An available clone of *HKL1* with a double haemagglutinin (HA) tag (Karve *et al.*, 2008) was subcloned with a substituted C-terminal FLAG tag in the HBT vector (Moore et al., 2003). Each fusion gene was then transferred into the pCB302 binary vector (*bar* selection marker; Xiang *et al.*, 1999), using *Bam*HI and *Pst*I cloning sites. For cloning the HKL1 promoter, a 3098 bp fragment upstream of the start codon was PCR amplified using the following primers: L1PGUSFP (5′-CCCAAGCCTGGGCAGCGAGCTGTCAAACTGGGGA-3′) and L1PGUSRP (5′-GCTCTAGATGCCCCAAAACAGAACCAAAAAGACA-3′). The promoter was cloned into the binary vector pSMAB704 (*bar* selection marker; Igasaki *et al.*, 2002), using *Hin*dIII and *Sma*I cloning sites upstream of the β-glucuronidase (GUS) gene. The identities of all clones were verified by DNA sequencing.

Binary constructs were introduced into *Agrobacterium tumefaciens* GV3101 by electroporation. *Arabidopsis* plants of Col-0, L*er* or *gin2-1* were transformed using the floral dip method (Clough and Bent, 1998). Transformants were selected for herbicide resistance (200 μM glufosinate ammonium; Rely 200, Bayer Crop Science, Kansas City, MO). Seeds of transgenic lines segregating 3:1 for herbicide resistance in the T_2 generation were selected for isolating homozygous lines. Seeds from two or more T_3 lines homozygous for the single insert were used for experiments.

RT- PCR analysis

Total RNA was isolated from whole seedlings of different lines using the RNeasy plant kit (Qiagen, Germantown, MD). One μg of total RNA was converted to cDNA using the Protoscript II RT-PCR kit (New England BioLabs, Ipswich, MA). PCR primer sequences for *HXK1*, *HKL1*, and *UBQ5* were described previously (Karve *et al.*, 2008). The expression of a variety of candidate genes was assessed in preliminary experiments by semi-quantitative RT-PCR, based on published data from glc transcript profiling studies (Price *et al.*, 2004). Selected glc regulated genes are a subset which responded most robustly under the current treatment conditions. The PCR primer sequences for the candidate genes were generated using the AtRTPrimer public database (Han and Kim, 2006): *ASN1* (asparagine synthase1, At3g47340; 5′-TGATTCTCAGGCCAAGAGAGTTCGT-3′, 5′-CCCAACCAATGTAGAGCGAAGTGAC-3′, expected size=413 bp), *T6P* (trehalose 6-phosphate synthase8, At1g70290; 5′-AGCTCCATTGTTCAAGATCCAAGCA-3′, 5′-GCTCCCCGCGTTCTACCATTTCTC-3′, expected size=626 bp), and *GLYK* (glycerate kinase, At1g80380; 5′-TTGGTGCGAAGATCAGATTGCTTTG-3′, 5′-GGAGACAGCATCGCATTAGTTTGC-3′, expected size=544 bp). All the primers were designed to span one or more introns such that the amplicon size from cDNA would be different than that from genomic DNA. The template amounts were first titrated to balance the *UBQ5* expression in different samples (using densitometry), and corresponding template amounts were used thereafter, while varying PCR cycle numbers.

Immunoblots and gluokinase activity assays

Total soluble proteins were extracted as described by Karve *et al.* (2008). The protein concentration in the leaf extracts was measured by Coomassie Blue (Bio-Rad, Hercules, CA). Equal amounts of proteins were electrophoresed by SDS-PAGE and transferred onto Immobilon-P membrane (Millipore, Bedford, MA). The membranes were probed with monoclonal anti-HA (Roche, Indianapolis, IN) or anti-FLAG M_2 (Sigma-Aldrich, St Louis, MO) antibodies, then incubated with HRP conjugated secondary antibody, followed by chemiluminescence reagents (SuperSignal West Pico, Pierce Biotechnology, Rockford, IL) and detection by film (Blue X-ray, Phenix Research Products, Candler, NC). Glucokinase activity was measured directly from leaf extracts or from lysates of maize protoplasts transfected with the indicated plasmids (Karve *et al.*, 2008).

Protoplast transient expression assays

Leaves of greening maize seedlings or *Arabidopsis* plants (Col-0 or *hkl1-1*) were used as a source of protoplasts for protein expression and signalling assays (Jang and Sheen, 1994; Hwang and Sheen, 2001). Protoplasts were transfected (Yoo *et al.*, 2007) with promoter constructs for *RBCS-LUC* (4 μg) or *PPDK-LUC* (6 μg), and with *UBQ10-GUS* (2 μg) as an internal control (Balasubramanian *et al.*, 2007). Protoplasts were co-transfected as indicated with effectors *HXK1-HA* (6 μg) and/or *HKL1-HA* (8 μg). Transfection efficiencies were routinely >60%, as determined using WRKY-GFP (Balasubramanian *et al.*, 2007). An empty vector was included to maintain a balanced concentration of DNA during transfections. Following transfection, protoplasts were incubated in the dark for 90 min, then treated with 2 mM glc and incubated in the light for 6–8 h at 30 μmol m^{-2} s^{-1}. Protoplasts were collected by low speed centrifugation. After resuspending in lysis buffer, GUS and LUC activities were measured as described previously

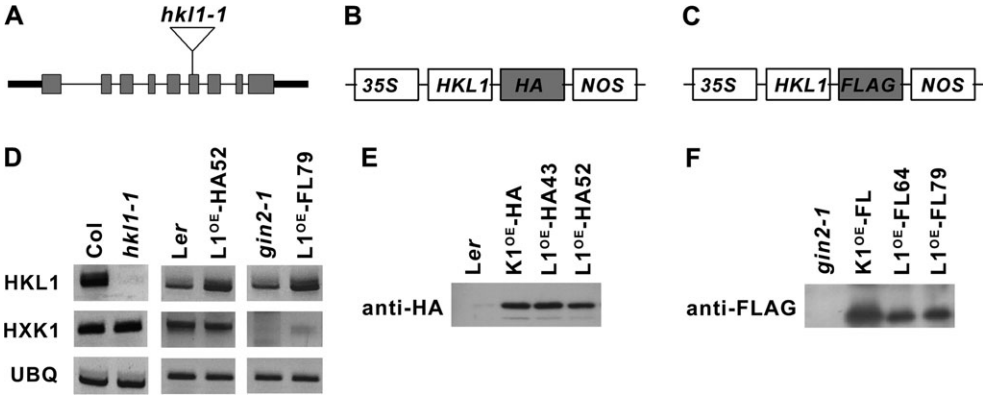


Fig. 1. Molecular characterization of *Arabidopsis* HKL1 mutant and transgenic lines. (A) Schematic diagram showing the gene structure of *HKL1* (At1g50460). Exons are indicated by grey rectangles, introns are indicated by the thinner lines. The location of the T-DNA insertion in *hkl1-1* is shown with the open triangle. (B, C) Design of plasmid constructs used to transform *Arabidopsis* lines. HKL1-HA was used to transform L*er* and HKL1-FLAG was used to transform *gin2-1*. Boxes are not drawn to scale. 35S, CaMV promoter; NOS, nopaline synthetase terminator; HA, 2 copies of the 10 amino acid haemagglutinin tag; FLAG, 1 copy of the 8 amino acid FLAG tag. (D) Transcript expression of *HKL1* and *HXK1* by semi-quantitative RT-PCR: Col and *hkl1-1*; L*er* and HKL1-HA line 52; and *gin2-1* and HKL1-FLAG line 79. At*UBQ5* mRNA was used as a control for the amount of template. PCR cycle numbers for HKL1, HXK1, and UBQ were 33, 30, and 30, respectively. $L1^{OE}$, HKL1 overexpression. (E) Immunoblot analysis using anti-HA antibody and 1 μg protein from leaf extracts of L*er*, HXK1-HA transgenic ($K1^{OE}$-HA), and two HKL1-HA lines. (F) Immunoblot analysis using anti-FLAG antibody and 1 μg protein from leaf extracts of *gin2-1*, HXK1-FLAG transgenic ($K1^{OE}$-FL), and two HKL1-FLAG lines.

(Balasubramanian *et al.*, 2007). Promoter activities are expressed as relative LUC/GUS values, normalized to control samples, which had no added glc.

Histochemical GUS staining and fluorometric GUS assays

Histochemical staining of transgenic *Arabidopsis* plants expressing the *pHKL1*-GUS fusion construct was performed as described by Crone *et al.* (2001). The plant tissue was incubated in GUS staining buffer containing 25 mg ml^{-1} of X-Glc (Gold BioTechnology, St Louis, MO) for 2–4 h and destained with 95% ethanol for 6–8 h. For measuring total extractable GUS activity, seedlings were extracted in buffer containing 50 mM NaH_2PO_4 (pH 7.0), 10 mM EDTA, 0.1% Triton X-100, 0.1% sodium lauryl sarcosine, and 10 mM β-mercaptoethanol. The enzymatic reaction was carried out in 100 μl of extraction buffer plus 1 mM 4-methyl umbelliferyl glucuronide (MUG, Sigma-Aldrich) at 37 °C for the indicated times, before stopping with 300 μl of 0.2 M Na_2CO_3. Fluorescence was measured in a 96-well microtitre plate format using a GENios spectrophotometer (Phenix Research Products) at a 360 nm excitation wavelength and a 465 nm emission wavelength. Sample GUS activities were calculated from a standard curve made using 0.1–1 μM 4-methyl-umbelliferone (Sigma-Aldrich).

In one experiment, transgenic seeds expressing *pHKL1*-GUS were grown on 1× MS plates plus 0.5% sucrose for 7 d, then transferred to liquid MS medium for 4 h with 10 μM indoleacetic acid (IAA), 1 μM abscisic acid (ABA), 50 μM 1-aminocylcopropane-1-carboxylic acid (ACC), or 10 μM zeatin (all from Sigma-Aldrich). Both treated and control seedlings were analysed for GUS staining and extractable GUS activity as described above.

Light microscopy

Light microscopy was used to view and capture images for routine seedling pictures, as well as for the GUS-stained seedlings or tissues, using a Nikon SMZ1500 stereo microscope (Nikon Instruments Inc., Melville, NY) with a MicroPublisher CCD cooled colour camera and Image Pro Plus v5.0 software (Media Cybernetics, Bethesda, MD). For measuring the hypocotyl lengths, the stereomicroscope was calibrated throughout the magnification range, using a stage micrometer.

Results

Molecular characterization of HKL1 knockout and increased expression lines

To understand the biological role of AtHKL1, a functional genomics approach was used by examining phenotypes of mutant and transgenic lines with altered HKL1 protein expression level. Seeds of a T-DNA insertion line for AtHKL1, generated by the University of Wisconsin knockout facility, were obtained through ABRC. Homozygous knockout plants with a possible single insert were identified by PCR screening. The T-DNA insertion was shown using real-time PCR and the $2^{-\Delta\Delta Ct}$ method of relative quantification (Ingham *et al.*, 2001) to be present as a single copy (see Supplementary Fig. S1 at *JXB* online), as shown by the dilution series values close to 1. The insertion site was mapped to exon VI of HKL1 (Fig. 1A). Using semi-quantitative

RT-PCR, the mutant line was found to have no detectable HKL1 transcript (Fig. 1D). This line was designated *hkl1-1*.

Transgenic *Arabidopsis* plants that constitutively express HKL1 in different genetic backgrounds were made: HKL1-HA expressed in L*er* or HKL1-FLAG expressed in *gin2-1* (Fig. 1B, C). This was done in order to distinguish possible HKL1 dependent phenotypes in relation to HXK1 expression. Three independent homozygous lines were obtained for the HKL1-HA transformants and seven lines for the HKL1-FLAG transformants. Representative transformed lines had substantially increased HKL1 transcripts, relative to each respective parental line (Fig. 1D). Notably, the HXK1 mRNA abundance was not altered in *hkl1-1* or in transformed L*er* lines which expressed HKL1-HA. The transformed lines with HKL1-FLAG did not have HXK1 transcripts, consistent with their parental background being *gin2-1*.

Western blot analysis of leaf extracts was carried out using antibodies to the introduced epitope tags (Fig. 1E, F). All of the transgenic lines expressed the corresponding tagged protein, while the parental lines did not. Positive controls included transgenic lines that expressed either HA or FLAG-tagged forms of HXK1 protein. From these assays, the two indicated lines expressing each construct were selected for further phenotypic analyses, with data presented for HKL1-HA line 52 and for HKL1-FLAG line 79.

Growth phenotypes of HKL1 knockout and increased expression lines

To test whether the HKL1 protein has a discernible function in plant growth, the different experimental lines were grown under different conditions. When grown on agar plates with 0.5% sucrose, the HKL1-HA seedlings were distinctly smaller than were the parental L*er* seedlings, as were the *gin2-1* seedlings (Fig. 2A). However, expression of HKL1 in the *gin2-1* background had no apparent affect on seedling growth. Growth of *hkl1-1* seedlings on sucrose plates resembled growth of the parental Col-0 seedlings. These results indicated that HKL1 might be a negative regulator of plant growth when overexpressed in the L*er* background.

Transgenic and mutant lines also were grown in soil under different light conditions. When grown under SD conditions, both HKL1 overexpression lines had normal growth, when compared with control plants (Fig. 2). However, the *hkl1-1* plants under SD conditions were somewhat smaller than control plants, with a rosette diameter reduced by about 20% (Fig. 2B, C). Growth of the *hkl1-1* plants under LD conditions was similar to Col-0. By contrast, growth of the overexpression lines in either L*er* or *gin2-1* backgrounds was considerably reduced under LD conditions. For example, the rosette diameter for HKL1-HA plants was 50% smaller than for L*er* plants. This resulted in mature plants of the transgenic line being about 4-fold smaller. Also, the diameter of HKL1-FLAG plants was reduced by 25% compared to *gin2-1* plants, resulting in almost a 2-fold decrease in plant size. The reduced rosette sizes were not associated with a change in leaf numbers at the time of flowering for the different transformants relative to control lines (Fig. 2D), or with a change in the time to flowering (data not shown). These observations indicate that the intrinsic developmental programme was not changed due to increased HKL1 protein expression. However, seed yield from the small plants was greatly reduced, but not seed viability. Notably then, HKL1 overexpression in the L*er* background resulted in an even smaller plant than when overexpressed in the absence of HXK1 protein in the *gin2-1* background (see Supplementary Fig. S2 at *JXB* online).

Seedling hypocotyl growth among different HKL1 expression lines

Since *Arabidopsis* hypocotyl growth is sensitive to many endogenous factors that regulate plant cell elongation (Salchert *et al.*, 1998), the hypocotyl growth of 7-d-old seedlings grown vertically on plates under constant low light conditions was measured (Fig. 3). The average hypocotyl length of HKL1-HA seedlings was about 50% less than of the parental L*er* seedlings. On the other hand, *hkl1-1* seedlings had a 40% increase in hypocotyl length relative to Col-0 seedlings. The average hypocotyl length of *gin2-1* seedlings was about 45% less than that for L*er* seedlings. However, HKL1-FLAG seedlings did not show any significant change in hypocotyl growth when compared with the parental genotype, *gin2-1*. By this assay, HKL1 again was a negative regulator of seedling growth when expressed in a WT background, which contains HXK1.

Auxin-induced lateral root formation among different Arabidopsis *lines*

The reduced hypocotyl growth of *gin2-1* seedlings was previously linked to its being relatively insensitive to auxin (Moore *et al.*, 2003). Therefore, the different transgenic and mutant lines were also tested by an auxin assay for lateral root formation. In this assay, seedling growth in the presence of the auxin transport inhibitor NPA greatly reduces the number of lateral roots (Himanen *et al.*, 2002). Lateral root formation can then be initiated after seedling transfer to plates with NAA. With these treatments, both Col-0 and L*er* seedlings showed a robust induction of lateral root formation, increasing 5-fold and 4-fold, respectively, after transfer to plates with NAA (Fig. 4). Seedlings of *hkl1-1* showed a similar increase in their number of lateral roots relative to Col-0 seedlings. However, auxin treatment induced relatively fewer lateral roots in *gin2-1*, HKL1-HA, and HKL1-FLAG seedlings, about a 2-fold increase. As a control for this assay, the same treatments of the auxin receptor mutant *tir1* (Col background) did not appreciably induce any lateral roots, with the *tir1* mutant having fewer roots even than *gin2-1* or the two HKL1 overexpression lines. These data indicate that HXK1 has a significant role in the auxin induction of lateral roots and that HKL1 blocks this induction response to a level comparable with that observed in the absence of HXK1.

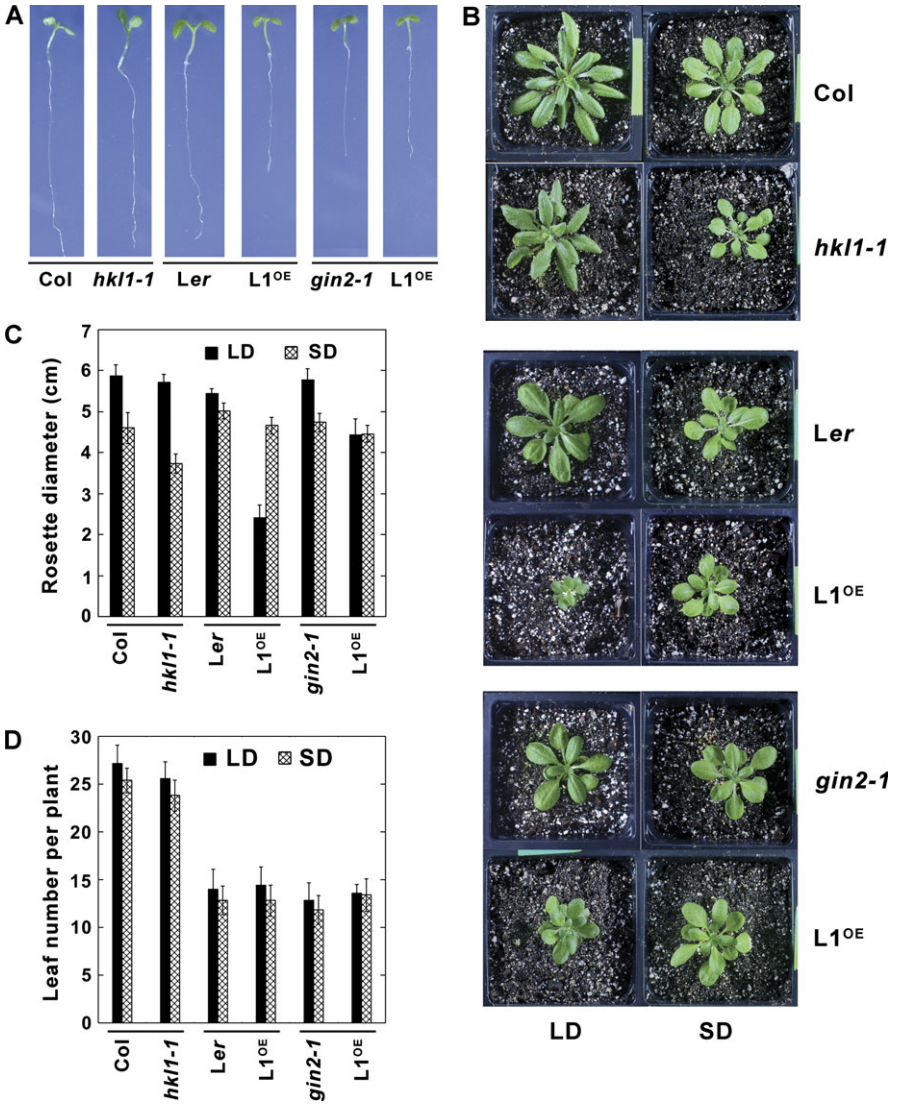


Fig. 2. Growth phenotypes of HKL1 mutant and transgenic lines. Dark bars indicate corresponding parental controls and modified lines. (A) Seven-day-old seedlings on 1× MS plates+0.5% sucrose. (B) Plants grown 30 d in a growth chamber under 8 h (short day, SD) or 16 h (long day, LD) photoperiods. (C) Average rosette diameter (cm) after 30 d ±SD, n=10. The difference in average diameters of Col and *hkl1-1* plants under SD conditions is statistically significant by a 2-tailed T test at P >0.95. (D) Average leaf number per plant at the time of bolting ±SD, n=10.

Glucokinase activities and glc signalling assays using different expression lines

The growth phenotypes of the HKL1 transgenic and mutant lines that have been described could be due to an influence of HKL1 protein on HXK1 protein catalytic activity, on HXK1 signalling functions, and/or on the function of an unknown protein. To test for the possible influence of HKL1 protein on glucokinase activity, rate measurements were carried out using leaf extracts from the different lines. There was no significant difference for enzyme activities between the transgenic lines and their respective control lines (Fig. 5A). As reported previously, HXK enzyme activity in *gin2-1* is about one-half of that in L*er* (Moore *et al.*, 2003) and HKL1-HA did not have any glc phosphorylation activity (Karve *et al.*, 2008). The possible inhibition of HXK1 by HKL1 was also tested after transiently expressing HXK1-HA and HKL1-HA in maize protoplasts. However, HKL1 protein did not affect the measured glucokinase activity (Fig. 5B).

Since HKL1 lacks glucokinase activity, but has a largely conserved glc binding domain, it is possible that, instead, the protein affects glc signalling activities. A widely used screen to identify mutants in glc signalling is based on the ability of some mutants to develop normally on otherwise inhibitory concentrations of exogenous glc (Rolland *et al.*, 2006). Therefore, seedling growth of the different lines was assessed in the presence of varying glc concentrations (Fig. 6A, B; see Supplementary Fig. S3 at *JXB* online). At relatively high glc levels, Col-0 and L*er* seedlings underwent developmental arrest, with much reduced root and shoot growth, and did not accumulate chlorophyll. The *hkl1-1* seedlings were hypersensitive to developmental arrest, showing substantial repression even on 4% glc. By contrast, the HKL1-HA

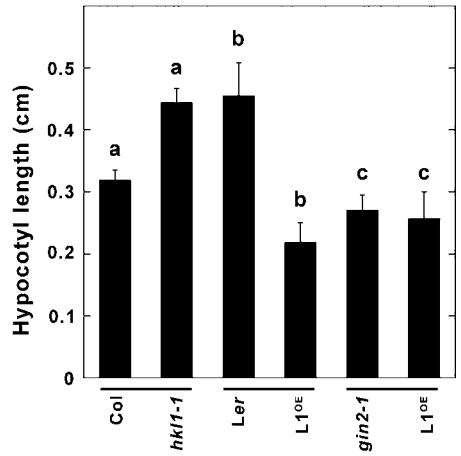


Fig. 3. Average seedling hypocotyl length of HKL1 mutant and transgenic lines. Seedlings were grown vertically for 7 d on 1/5× MS plates under constant light (15 μmol m^{-2} s^{-1}) at 22 °C. Values are means ±SD, n=15. a, b, by 2-tailed T tests, values are statistically different at P >0.95; c, values are not different at P >0.95.

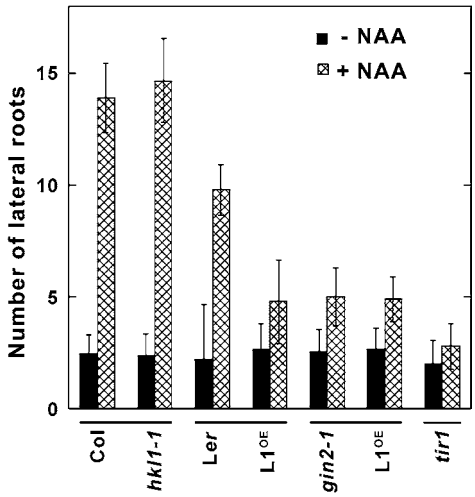


Fig. 4. Auxin-induced lateral root formation in seedlings of HKL1 mutant and transgenic lines. The number of lateral roots were counted 5 d after seedling transfer from plates with 5 μM NPA to plates with or without 0.1 μM NAA. Values are average lateral root numbers ±SD, n=10.

seedlings were glc-insensitive relative to the L*er* control line. When grown on 6% glc, >90% of the HKL1-HA seedlings have green cotyledons versus 0% of the L*er* seedlings. The responses of HKL1-FLAG seedlings were comparable with those of *gin2-1* seedlings. As an osmotic control, all lines were shown to have a similar phenotype on MS plates with 6% mannitol (Fig. 6C). Also, mannitol did not repress cotyledon greening in any of the lines (Fig. 6D). The observed glc-dependent phenotype suggested that HKL1 could be a negative regulator of glc signalling.

To test whether HKL1 might have a role in glc signalling, protoplast transient expression assays were carried out using p*RBCS*-LUC and p*PPDK*-LUC as established reporters of HXK1 signalling (Balasubramanian *et al.*, 2007). Leaf protoplasts of Col-0 and *hkl1-1* plants were used in independent assays. Relative *RBCS*-driven LUC activities expressed in protoplasts of either genotype was reduced by 25% with 2 mM glc (Fig. 7A). In both cases, co-transfection with HXK1 plus treatment with 2 mM glc reduced the reporter activity by about 55%. By contrast, transfected HKL1 did not affect the relative expressed *RBCS*-driven LUC activity with glc alone or with HXK1 plus glc, using protoplasts from either wild-type or mutant leaves. Similar results were obtained using p*PPDK*-LUC (data not shown). Notably, in all cases the expression of p*UBQ10*-GUS was not affected by co-transfection of HXK1, HKL1 and/or by addition of 2 mM glc.

To complement the transient expression assays, an alternate assay of glc signalling was carried out using seedlings grown in liquid culture and treated with or without 2% glc for 8 h. The selected *GLYK* and *T6P* genes are thought to be regulated by HXK1-dependent glc signalling, and *ASN* by a glycolysis-dependent glc signalling pathway (Price *et al.*, 2004). Supporting this interpretation, transcripts of *ASN*, *GLYK*, and *T6P* were all repressed by glc treatment of Col-0 and L*er* seedlings, while *GLYK* and *T6P* mRNA abundance were not affected by treatment of *gin2-1* seedlings (Fig. 7B). The response of these transcripts was not differentially affected in any of the tested mutant or transgenic lines, relative to the corresponding control lines. These data indicate that HKL1 probably does not affect the commonly recognized transcriptional targets of glc signalling, whether by a HXK1-dependent or a glycolysis-dependent pathway.

HKL1 promoter expression and activity assays

To improve our understanding of possible HKL1 functions, transgenic *Arabidopsis* plants were made that express an *HKL1* promoter–GUS fusion construct (p*HKL1*-GUS). At the early stages of seedling development, GUS staining was detected mainly in the root, particularly towards the root tip (Fig. 8A). With increased seedling growth, GUS staining was progressively localized to the vascular tissues of cotyledons (Fig. 8B), was relatively strong in the root and shoot meristems, but not in leaf primordia (Fig. 8C). In adult plants, GUS expression was highest in the root and leaf vascular tissue, and in the emerging lateral roots (Fig. 8D, E, F). In stem cross-sections, GUS staining was observed in phloem tissue. In flowers, GUS staining was observed in anther filaments, but not in the pistils (Fig. 8G, H). Staining was also observed broadly in developing siliques, becoming localized apparently to the funiculi of more mature seeds (Fig. 8I).

Since HKL1 overexpression reduced the sensitivity of seedlings to auxin-dependent lateral root formation (Fig. 4) and reduced the sensitivity of seedlings to glc repression of development (Fig. 6), the influence of short-term treatment of seedlings with different hormones was examined on the expression of p*HKL1*-GUS activity (Fig. 9). The effect of

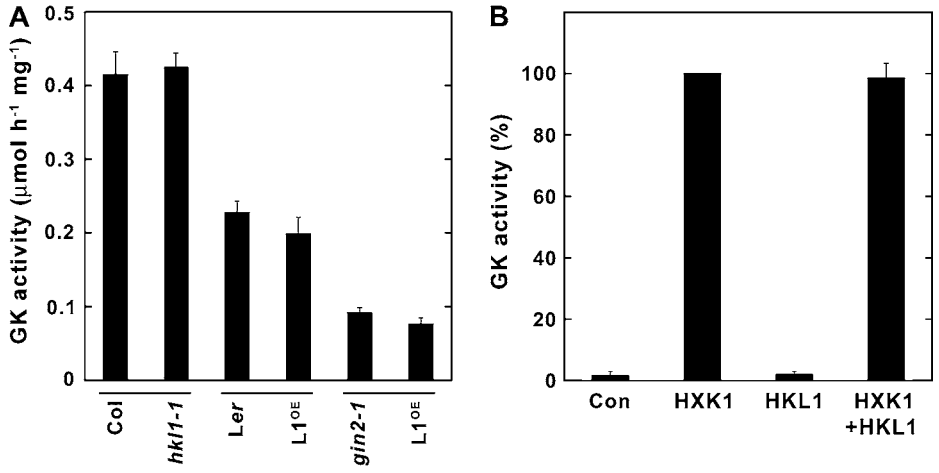


Fig. 5. Glucokinase activity of HKL1 mutant and transgenic lines. (A) Clarified leaf extracts of greenhouse-grown plants were assayed directly for enzyme activity. Values are means ±SD, n=3. (B) Maize protoplast extracts were assayed for enzyme activity after expression of plasmids with HXK1-HA and/or HKL1-HA. Protein expression was routinely monitored by labelling with [^{35}S]-methionine (data not shown; as in Karve *et al.*, 2008). Values are means ±SD, n=3, expressed relative to control protoplasts with empty vector DNA only.

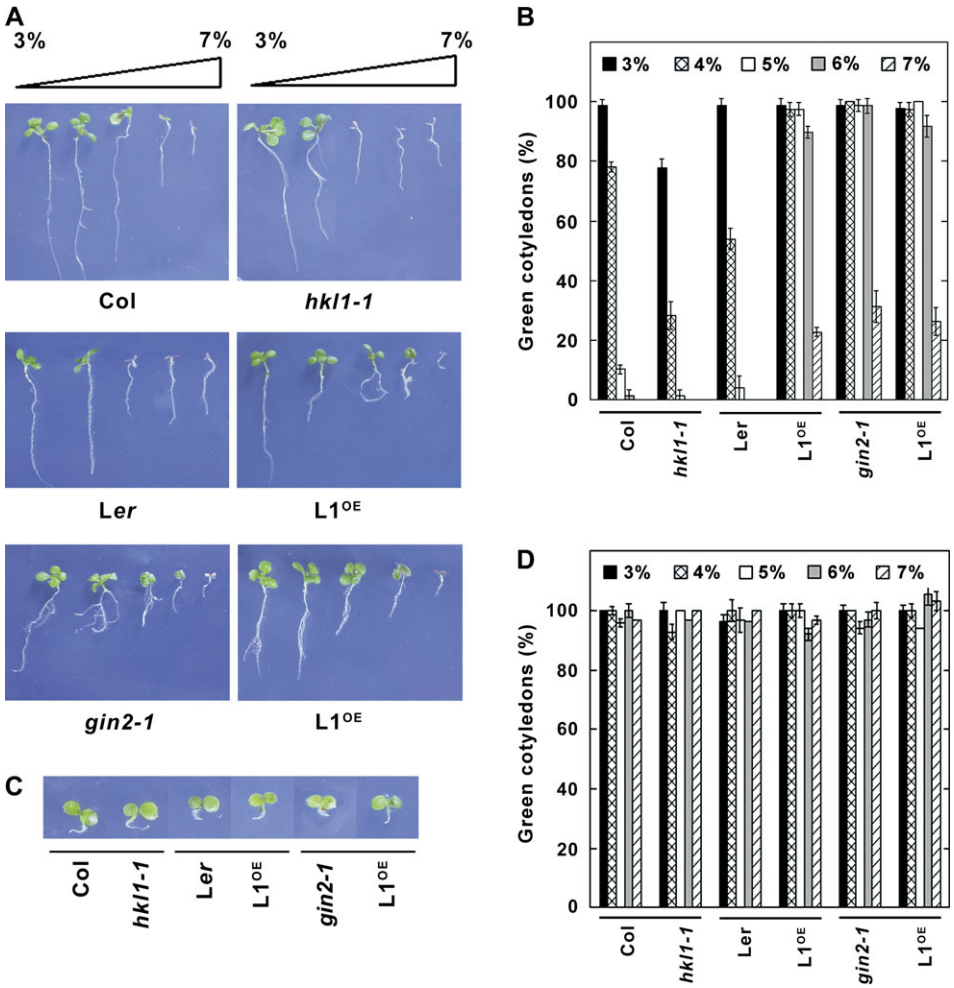


Fig. 6. Phenotypes of HKL1 mutant and transgenic lines grown on agar plates with 3–7% glc. (A) Images are representative 7-d-old seedlings. (B) Percentage of seedlings in (A) at corresponding glc concentrations which had green cotyledons. Values are expressed relative to the total number of germinated seedlings (30–40), as means ±SD, n=3. (C) Images of representative 7-d-old seedlings grown on agar plates with 6% mannitol. (D) Percentage of seedlings in (C) at corresponding mannitol concentrations which had green cotyledons. Values are expressed relative to the total number of germinated seedlings, as means ±SD, n=3.

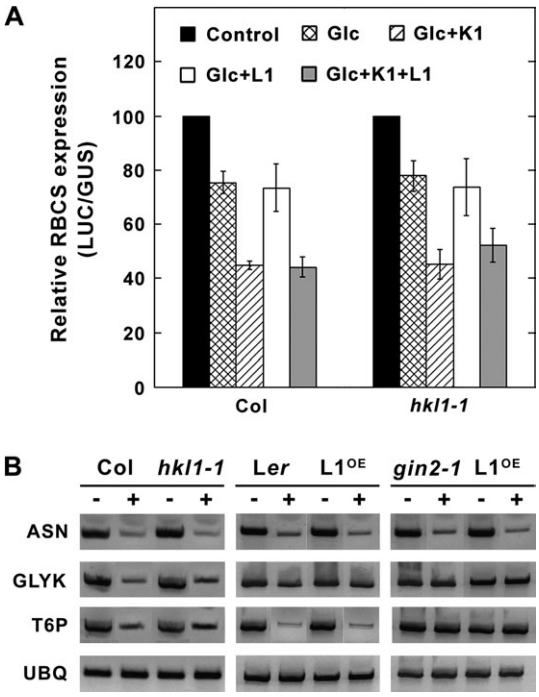


Fig. 7. Glc signalling assays. (A) Transient expression assays using leaf protoplasts from WT Col or *hkl1-1*. Protoplasts were co-transfected with p*RBCS*-LUC and an internal control, p*UBQ10*-GUS, plus or minus effectors HXK1-HA and/or HKL1-HA. Protoplast treatments include without glc or effectors (Control), with 2 mM glc (Glc), with 2 mM glc+HXK1-HA (Glc+K1), with 2 mM glc+HKL1-HA (Glc+L1), and with 2 mM glc+HXK1-HA+HKL1-HA (Glc+K1+L1). Values are means ±SD of the relative LUC units to GUS activities for replicated assays normalized to the control. GUS activity was not affected by the presence of glc or either effector. (B) Expression of glc regulated genes in HKL1 transgenic lines and mutants. Semi-quantitative RT-PCR was used to determine the transcript levels of asparagine synthase (ASN), glycerate kinase (GLYK), trehalose 6-phosphate synthase (T6P), and ubiquitin (UBQ). Seedlings grown in liquid medium were challenged without (–) or with (+) 2% glc for 8 h (see Materials and methods for further details). The number of PCR cycles was varied in each case, but for the presented data are as follows: ASN, 32 cycles, GLYK, 32 cycles, T6P, 33 cycles, UBQ, 31 cycles.

the hormone treatments was determined visually and also quantitatively after extraction and assay of GUS activity. For the latter, initial assays in the absence of added stimuli established that a 2 h reaction time with seedling extracts was within the linear range of activity (Fig. 9B). IAA treatment (or GA_3 treatment, data not shown) did not induce p*HKL1*-GUS expression or enzyme activity (Fig. 9A, C). However, ABA treatment greatly reduced seedling GUS staining and reduced the extractable GUS activity by 50%. On the other hand, zeatin or ACC treatments induced GUS expression throughout the seedling and not just in the vascular tissue. Correspondingly, the extracted GUS activities following these treatments increased up to 2-fold relative to the control treatment. The results of the GUS assays indicate that *AtHKL1* might be regulated by multiple plant hormones and, thereby, could have a regulatory role in plant growth and development.

Discussion

Non-catalytic HXKs have been identified in fungi including *S. cerevisiae* and *A. nidulans* (Daniel, 2005; Bernardo *et al.*, 2007) and also in *Arabidopsis* (Karve *et al.*, 2008). Whether non-catalytic homologues of known enzymes are commonly present in other protein families is not known. The *Arabidopsis* glutathione transferase family does include both non-catalytic as well as catalytic forms, although their relative distribution between the groups apparently has not been strictly determined (Dixon *et al.*, 2003). Recently, β-amylase4 (BAM4) of *Arabidopsis* was shown to lack apparent catalytic activity, yet somehow to facilitate starch breakdown (Fulton *et al.*, 2008). BAM4 is one of perhaps four chloroplastic isoforms within *Arabidopsis*. Also, the plant shikimate kinase gene family includes two non-catalytic homologues which have been present in all major plant lineages for over 400 million years (Fucile *et al.*, 2008). In *Arabidopsis*, these express novel functions, one of which is required for chloroplast biogenesis (Fucile *et al.*, 2008). Non-catalytic enzyme homologues might occur somewhat more often among plant gene families than what is currently appreciated, since sequence divergence levels within families are often >25%. That is, in order to transfer all four digits of an EC number at an error rate below 10%, the estimated level of sequence identity needs to be >75% (Rost *et al.*, 2003). It is suggested that when non-catalytic homologues of known enzymes do occur, they are likely to have important regulatory functions. For example, several catalytically inactive homologues of phosphoinositide 3-phosphatases have been linked to specific human diseases (Robinson and Dixon, 2006).

As one general approach to understand protein function, the tissue expression pattern and regulation of gene expression can provide an important physiological context. The AtHKL1 transcript was previously shown to be expressed in the principal plant organs (Karve *et al.* 2008). These observations have been extended in this study by demonstrating that p*HKL1*-GUS activity occurs predominantly in the vascular tissues of different sink organs such as roots, stems, and anthers (Fig. 8). In stem cross-sections, the vascular staining was associated with phloem tissue. While we are not aware of any HXK family members having been reported in surveys of the phloem proteome, nonetheless many phytohormones and a number of regulatory proteins have been detected in phloem sap (Giavalisco *et al.*, 2006). The HKL1 promoter activity was also found to be influenced by several phytohormones, including being repressed by ABA and induced by both ACC and cytokinin. Hormone induction of the HKL1 promoter occurred in both vascular and non-vascular tissues (Fig. 8). Our analysis of the HKL1 promoter sequence for known regulatory elements (Higo *et al.*, 1999; Molina and Grotewold, 2005; Obayashi *et al.*, 2007) indicates that the

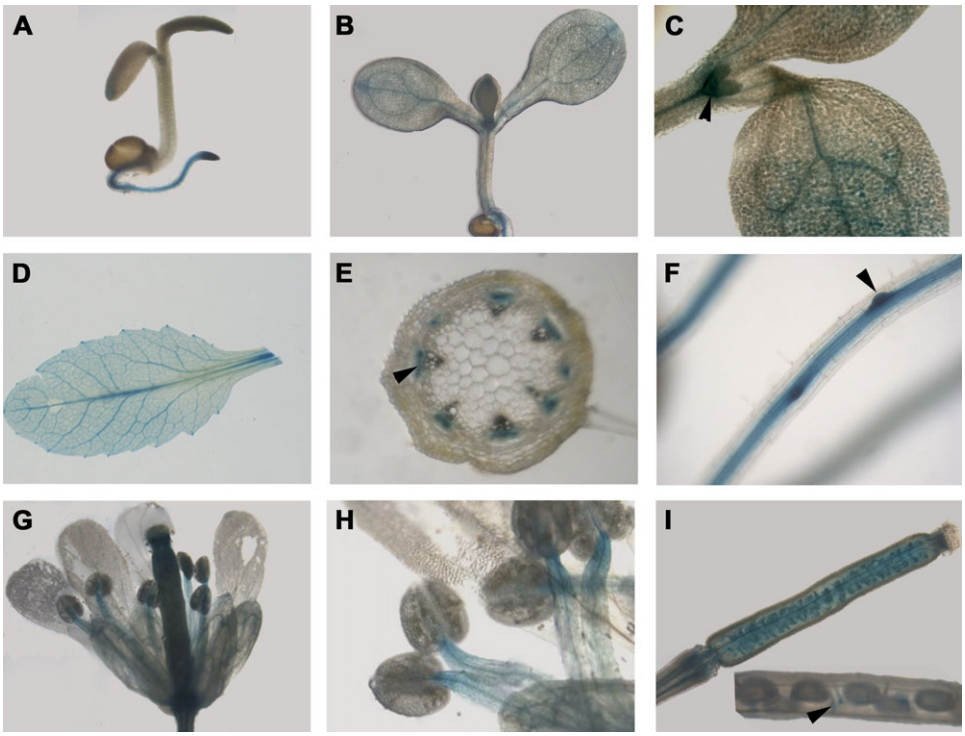


Fig. 8. Organ and tissue expression of p*HKL1*-GUS. (A) Seedlings grown for 3 d on MS plates. (B) Seedlings grown for 7 d on MS plates. (C) Shoot of a 5-d-old seedling, with the arrowhead pointing to specific stain in the meristem. (D) Leaf from a 21-d-old plant. (E) Stem cross-section, with the arrowhead pointing to staining of phloem. (F) Root of a 10-d-old seedling, with the arrowhead pointing to enhanced staining at the site of lateral root initiation. (G) Opened flower. (H) Anthers and filaments. (I) Developing silique, with insert showing a mature silique and an arrowhead pointing to the funiculus of a developing seed.

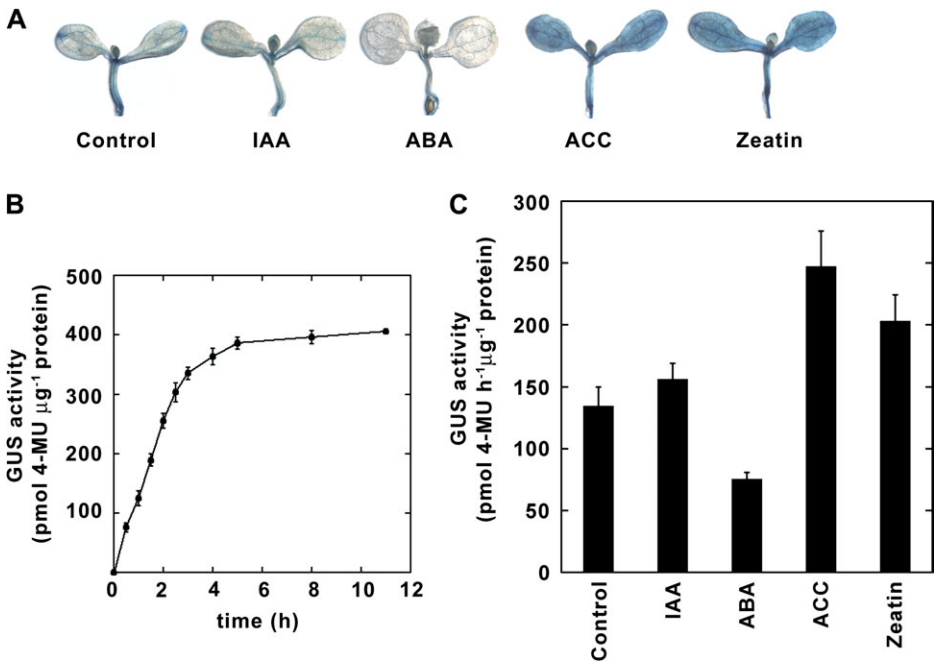


Fig. 9. Effect of different plant hormones on p*HKL1*-GUS expression. Seedlings of *pHKL1*-GUS lines were grown for 7 d on MS plates, then transferred to liquid MS medium for 4 h with different plant hormones: control (no additions), 10 μM IAA, 1 μM ABA, 50 μM ACC, and 10 μM zeatin. (A) Seedlings stained for GUS activity. (B) Reaction time-course for GUS activity assayed from control seedlings. (C) GUS activity of seedlings after a 2 h reaction. Values are means ±SD, n=3.

promoter does have motifs proposed to be regulated by several hormones (data not shown).

Phenotypes of the AtHKL1 overexpression lines provide evidence that the HKL1 protein is a negative regulator of plant growth. HKL1 overexpression in L*er* (HKL1-HA) resulted in reduced seedling growth on sucrose plates (Fig. 2), reduced hypocotyl elongation under low light conditions (Fig. 3), severely reduced rosette size under LD conditions (Fig. 2), and a decreased sensitivity to auxin-induced lateral root formation (Fig. 4). In a recent initial report, some rice HXK family members also were considered to be possible negative regulators of seedling growth (Yu and Chiang, 2008). The status of the glc binding domain in possible regulatory HXKs needs to be evaluated experimentally. It has previously been shown that AtHXK1-G173A has a 90% decrease in glc phosphorylation activity (Karve *et al.*, 2008). Since AtHKL1 has the same recognized glc binding domain as does this mutated protein, we speculate that glc binding affinity is reduced in AtHKL1, but not eliminated. Thus, for a negative regulator, decreased glc binding affinity could be a feedback mechanism to limit plant growth in the presence of excessive carbohydrate availability.

The HKL1 protein might function as a negative regulator of cell expansion, based on reduced hypocotyl growth of HKL1-HA seedlings and on increased hypocotyl growth of the *hkl1-1* seedlings (Fig. 3). Seedling hypocotyl growth by cell elongation integrates diverse signals including light, temperature, nutrients, and most plant hormones (Collett *et al.*, 2000; De Grauwe *et al.*, 2005). In *gin2-1*, reduced hypocotyl growth has been attributed to the possible insensitivity of seedlings to auxin signalling (Moore *et al.*, 2003). However, ethylene also can repress hypocotyl elongation in seedlings grown under conditions similar to those in our experiment (Smalle *et al.*, 1997). Thus, it is possible that HKL1 expression promotes ethylene sensitivity instead of attenuating auxin sensitivity. Consistent with this possibility, while lateral root formation does require auxin synthesis, transport, and/or signalling (Casimiro *et al.*, 2003), enhanced ethylene signalling has more recently been shown to repress lateral root formation by modulating auxin transport (Negi *et al.*, 2008). Thus, the observed HKL1 repression phenotype for auxin-induced root formation (Fig. 4) might instead be associated with an altered ethylene response. Further experiments are needed to clarify the mechanisms involved.

The mode of action of AtHKL1 is not known, but does merit further consideration. On the one hand, since both HXK1 and HKL1 are targeted to mitochondria (Heazlewood *et al.*, 2004; Karve *et al.*, 2008), the two proteins have the potential to interact such that HKL1 could act as a dominant negative effector. In this case, the overexpression of HKL1 in the *gin2-1* background might not result in a novel phenotype relative to its overexpression in the presence of HXK1. Assay results for hypocotyl growth (Fig. 3), for auxin induction of lateral root growth (Fig. 4), and for glc tolerance (Fig. 6) are consistent with this possibility. Furthermore, the contrasting phenotypes observed by these assays with the *hkl1-1* mutant also support this interpretation. On the other hand, the overexpression of HKL1 in WT did result in a much more diminutive plant under LD conditions than was observed in *gin2-1* (Fig. 2). This implies that HKL1 could have a more complicated mode of action by also independently affecting one or more targets possibly involved in mediating phytohormone responses.

In summary, the present results indicate that the non-catalytic AtHKL1 protein can negatively influence plant growth, possibly by somehow influencing cross-talk between glc and other plant hormone response pathways. Elucidating the functions of non-catalytic proteins will be an ongoing challenge for contemporary biologists.

Supplementary data

Supplementary data are available at *JXB* online.

Supplementary Fig. 1. The number of T-DNA insertions in *hkl1-1* as determined by real time PCR.

Supplementary Fig. S2. Growth under 16 h (long day) photoperiod conditions of transgenic *Arabidopsis* expressing HKL1 protein in different genetic backgrounds.

Supplementary Fig. S3. Growth response on agar plates with varying glc concentrations for the transgenic HKL1-Flag line 43.

Supplementary Fig. S4. Transcript abundance by semi-quantitative RT-PCR of HXK1 and HKL1 from L*er* seedlings grown on plates with 0.5% sucrose (–) or with 6% glucose (+).

Acknowledgements

We very much appreciate technical help from Ms Xiaoxia Xia. We thank Dr Zhou Li and Dr Jen Sheen for sharing their initial observations on the relative auxin insensitivity for lateral root induction in *gin2-1*. We also thank Dr J-C Jang for useful discussions on seedling glc signalling assays. This paper is technical contribution no. 5642 of the Clemson University Experiment Station. This material is based upon work supported by the CSREES/USDA, under project number SC1700190. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the view of the USDA.

References

Balasubramanian R, Karve A, Kandasamy M, Meagher RB, Moore B. 2007. A role for F-actin in hexokinase-mediated glucose signaling. *Plant Physiology* **145,** 1423–1434.

Balasubramanian R, Karve A, Moore B. 2008. Actin-based framework for cellular glucose signalling by Arabidopsis hexokinase1. *Plant Signaling and Behavior* **3,** 322–324.

Bernardo SM, Gray KA, Todd RB, Cheetham BF, Katz ME. 2007. Characterization of regulatory non-catalytic hexokinases in *Aspergillus nidulans*. *Molecular Genetics and Genomics* **277,** 519–532.

Casimiro I, Beekman T, Graham N, Bhalerao R, Zhang H, Casero P, Sandberg G, Bennett MJ. 2003. Dissecting Arabidopsis lateral root development. *Trends in Plant Science* **8,** 165–171.

Chen J-G, Willard FS, Huang J, Liang J, Chasse SA, Jones AM, Siderovski DP. 2003. A seven-transmembrane RGS protein that modulates plant cell proliferation. *Science* **301,** 1728–1731.

Chen Y, Ji F, Xie H, Liang J, Zhang J. 2006. The regulator of G-protein signalling proteins involved in sugar and abscisic acid signalling in Arabidopsis seed germination. *Plant Physiology* **140,** 302–310.

Cho JI, Ryoo N, Eom JS, *et al*. 2009. Role of the rice hexokinases *OsHXK5* and *OsHXK6* as glucose sensors. *Plant Physiology* **149,** 745–759.

Cho YH, Yoo SD, Sheen J. 2006. Regulatory functions of nuclear hexokinase1 complex in glucose signalling. *Cell* **127,** 579–589.

Claeyssen E, Rivoal J. 2007. Isozymes of plant hexokinase: occurrence, properties and functions. *Phytochemistry* **68,** 709–731.

Clough SJ, Bent AF. 1998. Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *The Plant Journal* **16,** 735–743.

Collett CE, Harberd NP, Leyser O. 2000. Hormonal interactions in the control of Arabidopsis hypocotyl elongation. *Plant Physiology* **124,** 553–562.

Crone D, Rueda J, Martin KL, Hamilton DA, Mascarenhas JP. 2001. The differential expression of a heat shock promoter in floral and reproductive tissues. *Plant, Cell and Environment* **24,** 869–874.

Daniel J. 2005. Sir-dependent downregulation of various aging processes. *Molecular Genetics and Genomics* **274,** 539–547.

De Grauwe L, Vandenbussche F, Tietz O, Palme K, Van Der Straeten D. 2005. Auxin, ethylene and brassinosteroids: tripartite control of growth in the Arabidopsis hypocotyl. *Plant and Cell Physiology* **46,** 827–836.

Dixon DP, McEwen AG, Lapthorn AJ, Edwards R. 2003. Forced evolution of a herbicide detoxifying glutathione transferase. *Journal of Biological Chemistry* **278,** 23930–23935.

Fucile G, Falconer S, Christendat D. 2008. Evolutionary diversification of plant shikimate kinase gene duplicates. *PLoS Genetics* **4,** e1000292.

Fulton DC, Stettler M, Mettler T, *et al*. 2008. β-AMYLASE4, a noncatalytic protein required for starch breakdown, acts upstream of three active β-amylases in Arabidopsis chloroplasts. *The Plant Cell* **20,** 1040–1058.

Gancedo JM. 2008. Early steps of glucose signalling in yeast. *FEMS Microbiology Review* **32,** 673–704.

Giavalisco P, Kapitza K, Kolasa A, Buhtz A, Kehr J. 2006. Towards the proteome of *Brassica napus* phloem sap. *Proteomics* **6,** 896–909.

Gibson SI. 2004. Sugar and phytohormone response pathways: navigating a signalling network. *Journal of Experimental Botany* **55,** 253–264.

Han S, Kim D. 2006. AtRTPrimer: database for Arabidopsis genome-wide homogeneous and specific RT-PCR primer-pairs. *BMC Bioinformatics* **7,** 179.

Heazlewood JL, Tonti-Filippini JS, Gout AM, Day DA, Whelan J, Millar AH. 2004. Experimental analysis of the Arabidopsis mitochondrial proteome highlights signalling and regulatory components, provides assessment of targeting prediction programs, and indicates plant-specific mitochondrial proteins. *The Plant Cell* **16,** 241–256.

Higo KY, Ugawa M, Iwamoto Korenaga T. 1999. Plant *cis*-acting regulatory DNA elements (PLACE) database. *Nucleic Acids Research* **27,** 297–300.

Himanen K, Boucheron E, Vanneste S, de Almeida Engler J, Inzé D, Beeckman T. 2002. Auxin-mediated cell cycle activation during early lateral root initiation. *The Plant Cell* **14,** 2339–2351.

Hwang I, Sheen J. 2001. Two-component circuitry in Arabidopsis cytokinin signal transduction. *Nature* **413,** 383–389.

Igasaki T, Ishida Y, Mohri T, Ichikawa H, Shnohara K. 2002. Transformation of *Populus alba* and direct selection of transformants with the herbicide bialaphos. *Bulletin of Forestry and Forest Products Research Institute* **1,** 235–240.

Ingham D, Beer S, Money S, Hansen G. 2001. Quantitative real-time PCR assay for determining copy number in transformed plants. *BioTechniques* **31,** 132–140.

Jang JC, Leon P, Zhou L, Sheen J. 1997. Hexokinase as a sugar sensor in higher plants. *The Plant Cell* **9,** 5–19.

Jang JC, Sheen J. 1994. Sugar sensing in higher plants. *The Plant Cell* **6,** 1665–1679.

Johnston CA, Taylor JP, Gao Y, Kimple AJ, Grigston JC, Chen JG, Siderovski DP, Jones AM, Willard FS. 2007. GTPase acceleration as the rate-limiting step in Arabidopsis G protein-coupled sugar signalling. *Proceedings of the National Academy of Sciences, USA* **104,** 17317–17322.

Karve A, Rauh BL, Xia X, Kandasamy M, Meagher RB, Sheen J, Moore BD. 2008. Expression and evolutionary features of the hexokinase gene family in Arabidopsis. *Planta* **228,** 411–425.

Leon P, Sheen J. 2003. Sugar and hormone connections. *Trends in Plant Science* **8,** 110–116.

Molina C, Grotewold E. 2005. Genome wide analysis of Arabidopsis core promoters. *BMC Genomics* **6,** 25.

Moore B, Zhou L, Rolland F, Hall Q, Cheng WH, Liu YX, Hwang I, Jones T, Sheen J. 2003. Role of the Arabidopsis glucose sensor HXK1 in nutrient, light, and hormonal signalling. *Science* **300,** 332–336.

Negi S, Ivanchenko MG, Muday GK. 2008. Ethylene regulates lateral root formation and auxin transport in *Arabidopsis thaliana*. *The Plant Journal* **55,** 175–187.

Obayashi T, Kinoshita K, Nakai K, Shibaoka M, Hayashi S, Saeki M, Shibata D, Saito K, Ohta H. 2007. ATTED-II: a database of co-expressed genes and *cis* elements for identifying co-regulated gene groups in *Arabidopsis*. *Nucleic Acids Research* **35,** D863–D869.

Osuna D, Usadel B, Morcuende R, *et al*. 2007. Temporal responses of transcripts, enzyme activities and metabolites after adding sucrose to carbon-deprived Arabidopsis seedlings. *The Plant Journal* **49,** 463–491.

Perfus-Barbeoch L, Jones AM, Assmann SM. 2004. Plant heterotrimeric G protein function: insights from Arabidopsis and rice mutants. *Current Opinion in Plant Biology* **7,** 719–731.

Price J, Laxmi A, St Martin SK, Jang JC. 2004. Global transcription profiling reveals multiple sugar signal transduction mechanisms in Arabidopsis. *The Plant Cell* **16,** 2128–2150.

Robinson FL, Dixon JE. 2006. Myotubularin phosphatases: policing 3-phosphoinositides. *Trends in Cell Biology* **16,** 403–412.

Rognoni S, Teng S, Arru L, Smeekens S, Perata P. 2007. Sugar effects on early seedling development in Arabidopsis. *Plant Growth Regulation* **52,** 217–228.

Rolland F, Baena-Gonzalez E, Sheen J. 2006. Sugar sensing and signalling in plants: conserved and novel mechanisms. *Annual Review of Plant Biology* **57,** 675–709.

Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. 2003. Automatic prediction of protein function. *Cellular and Molecular Life Sciences* **60,** 2637–2650.

Salchert K, Bhalerao R, Koncz-Kálmán Z, Koncz C. 1998. Control of cell elongation and stress responses by steroid hormones and carbon catabolic repression in plants. *Philosophical Transactions of the Royal Society of London, Biological Sciences* **353,** 1517–1520.

Schaffner AR, Sheen J. 1991. Maize rbcS promoter activity depends on sequence elements not found in dicot *rbcS* promoters. *The Plant Cell* **3,** 997–1012.

Smalle J, Haegman M, Kurepa J, Van Montagu M, Straeten DV. 1997. Ethylene can stimulate Arabidopsis hypocotyl elongation in the light. *Proceedings of the National Academy of Sciences, USA* **94,** 2756–2761.

Towle HC. 2005. Glucose as a regulator of eukaryotic gene transcription. *Trends in Endocrinology and Metabolism* **16,** 489–494.

Xiang C, Han P, Lutziger I, Wang K, Oliver D. 1999. A mini binary vector series for plant transformation. *Plant Molecular Biology* **40,** 711–717.

Yoo SD, Cho YH, Sheen J. 2007. Arabidopsis mesophyll protoplasts: a versatile cell system for transient gene expression analysis. *Nature Protocols* **2,** 1565–1572.

Yu SM, Chiang CM. 2008. Distinct hexokinases (HXKs) act as positive and negative regulators in sugar signalling pathways. *BMC Plant Biology* 60–61.

Global Change Biology (2009) **15**, 1201–1213, doi: 10.1111/j.1365-2486.2008.01818.x

Gene expression profiling: opening the black box of plant ecosystem responses to global change

ANDREW D. B. LEAKEY*, ELIZABETH A. AINSWORTH*†, STEPHANIE M. BERNARD‡, R. J. CODY MARKELZ*, DONALD R. ORT*†, SARAH A. PLACELLA§, ALISTAIR ROGERS¶‖, MELINDA D. SMITH**, ERIKA A. SUDDERTH††, DAVID J. WESTON‡‡, STAN D. WULLSCHLEGER‡‡ and SHENGHUA YUAN**

*Department of Plant Biology, Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801, USA, †Photosynthesis Research Unit, USDA-ARS, Urbana, IL 61801, USA, ‡Ecology Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA, §Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA 94720, USA, ¶Department of Environmental Sciences, Brookhaven National Laboratory, Upton, NY 11973, USA, ‖Department of Crop Sciences, University of Illinois at Urbana–Champaign, 1207 West Gregory Drive, Urbana, IL 61801, USA, **Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT 06520, USA, ††Department of Integrative Biology, University of California, Berkeley, CA 94720, USA, ‡‡Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6422, USA*

Abstract

The use of genomic techniques to address ecological questions is emerging as the field of genomic ecology. Experimentation under environmentally realistic conditions to investigate the molecular response of plants to meaningful changes in growth conditions and ecological interactions is the defining feature of genomic ecology. Because the impact of global change factors on plant performance are mediated by direct effects at the molecular, biochemical, and physiological scales, gene expression analysis promises important advances in understanding factors that have previously been consigned to the 'black box' of unknown mechanism. Various tools and approaches are available for assessing gene expression in model and nonmodel species as part of global change biology studies. Each approach has its own unique advantages and constraints. A first generation of genomic ecology studies in managed ecosystems and mesocosms have provided a testbed for the approach and have begun to reveal how the experimental design and data analysis of gene expression studies can be tailored for use in an ecological context.

Keywords: elevated CO_2, genomic, microarray

Received 7 July 2008 and accepted 14 September 2008

Introduction

The use of genomic techniques to address ecological questions is emerging as the important new field of genomic ecology (Jackson *et al*., 2002; Ouborg & Vriezen, 2007; Wullschleger *et al*., 2007; Roelofs *et al*., 2008; Shiu & Borevitz, 2008; Ungerer *et al*., 2008). Tools are now available to assess: (1) variation in genome sequence, (2) patterns of gene expression, and (3) gene function (Ouborg & Vriezen, 2007). The use of many of these tools, including quantitative trait loci analysis, association mapping, and genome sequencing has been reviewed previously (Lee *et al*., 2004; Straalen & Roelofs, 2006; Ouborg & Vriezen, 2007). This review focuses on how experiments investigating plant responses to elements of global change are becoming a testing ground for the use of transcript profiling, as a result of strategically targeted funding from U.S. Department of Energy's Program for Ecosystem Research (http://per.ornl.gov/PERprojects-current.html). Support for genomic ecology is timely because the new techniques available, and specifically gene expression analysis by transcript profiling, are ideal for addressing many of the major knowledge gaps in plant responses to global change. It is well recognized that our ability to predict the impact of global change on both ecosystem function and food supply is constrained by our limited

Correspondence: Andrew Leakey, tel. +1 217 244 0302, fax +1 217 244 2057, e-mail: leakey@life.uiuc.edu

understanding of plant responses to interacting elements of global change (e.g., drought × elevated CO_2), intra- and interspecific variation in response, nonlinear responses, and trophic interactions (Poorter, 1993; Wullschleger *et al.*, 2002; Fuhrer, 2003; Leakey *et al.*, 2006a; Long *et al.*, 2006; Bradley & Pregitzer, 2007; Delucia *et al.*, 2008). Because the impact of global change factors on plant performance are mediated by direct effects at the molecular, biochemical, and physiological scales, investigation of these processes promises understanding that has previously been consigned to the 'black box' of unknown mechanism (Fig. 1). This can be done in the traditional hypothesis-testing framework or in surveys designed to identify novel and unexpected aspects of response. In either case, there has been a move towards broader and more integrative thinking as transcript profiles are combined with high-throughput metabolite screening, physiological assessment, and automatic environmental data collection (Fig. 1).

Incorporation of global transcript profiling and other 'omic' approaches into ecological studies constitutes a major shift in philosophy compared with investigation of a few physiological and ecological parameters, and necessitates collaboration among scientists with diverse skill sets. An additional key feature of genomic ecology is experimentation under ecologically relevant treatments and conditions, unlike many molecular biology studies that have used shock treatments for the task of elucidating gene function. The new genomic ecology approach requires the physiologist and ecologist to learn new techniques and optimize the tools for use within the ecosystem context. This paper is the outcome of a workshop held at the University of Illinois in November 2007 to review the opportunities available for addressing important questions in global change biology using transcript profiling and associated technologies. We discuss the different approaches of studying model vs. nonmodel species, the opportunities and challenges in profiling ecologically relevant gene expression, and the value and interpretation of 'omic' data in an ecological context.

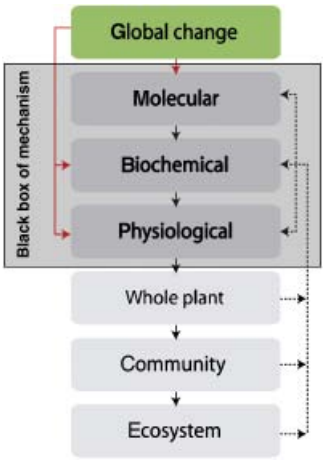


Fig. 1 Schematic describing the integration of plant, community, and ecosystem responses to an element of global change. Elements of global change directly impact molecular, biochemical, and physiological processes (red arrows), which combine to determine whole plant performance. Genotypic variation in whole plant responses drives ecological interactions that underlie community and ecosystem responses to global change. Feedbacks from larger scales of organization (dashed arrows) impact individual plant performance via effects on resource availability and disturbance that modify the direct effects on global change on plant function. Transcript profiling and high-throughput biochemical and physiological screening provide an opportunity to better understand the 'black box' of mechanisms driving plant responses to various elements of global change under field conditions.

Investigating model and nonmodel species

Busch & Lohmann (2007) classified the different methods for gene expression profiling in three categories: (1) PCR-based methods, such as quantitative real-time reverse-transcription PCR (qRT-PCR); (2) sequencing-based methods, such as cDNA-AFLP (amplified fragment length polymorphism), serial analyses of gene expression (SAGE), and massive parallel signature sequencing (MPSS); and (3) hybridization-based methods, such as microarrays. For this discussion, model species are defined as those for which a sufficiently large fraction of the genome has been sequenced to allow relatively easy transcript profiling of most or all genes by qRT-PCR or microarray analysis. Although real-time PCR can be high-throughput (Czechowski *et al.*, 2004), microarray analysis is currently the most common method of choice for transcript profiling. Microarrays are glass, plastic, or silicon chips with thousands of DNA oligonucleotides arrayed across their surface. Each oligonucleotide spot, or probe, corresponds to a specific target mRNA from a specific gene. The pool of RNA transcripts from sample tissue is extracted and labeled with a fluorescent tag before being washed over the microarray. Transcripts bind to their corresponding probes and the abundance of all transcripts is quantified by assessing the intensity of fluorescence associated with each probe. The result is information on the abundance of transcripts encoding a large fraction of the protein structures and enzymes in the sample tissue. A major assumption in interpretation of microarray data is that transcript abundance is related to protein synthesis and activity. The method does not directly assess the rate of gene expression or transcript degradation, but instead the pool size of

transcripts that is the result of the two processes. In addition, a number of posttranscriptional and posttranslation processes can disrupt the link between transcript abundance and enzyme activity. These assumptions influence the inferences that can be drawn from such datasets, but have not prevented the widespread use of this powerful technique.

To date, microarrays have been produced for at least 38 plant species (Table S1). Affymetrix is the largest commercial supplier of microarrays, and alone produces microarrays for Arabidopsis, barley, cotton, citrus species, grape, maize, *Medicago* spp., poplar, rice, soybean, sugarcane, tomato, and wheat (www.affymetrix.com). Other companies and research institutions manufacture microarrays for additional species, but these are also biased towards economically, rather than ecologically, significant species. As of February 2008, the National Center for Biological Information listed 37 land-plant species for which whole-genome sequencing was complete or in progress (http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi). Increasing numbers of ecologically and evolutionarily important species such as *Arabidopsis lyrata*, *Capsella rubella*, *Brachypodium distachyon*, *Mimulus lewisii*, and *Selaginella moellendorffii* are being sequenced. More species will rapidly become available for genomic investigation as techniques such as pyrosequencing allow smaller research groups to generate large amounts of sequence information and develop tools specifically for their own species of interest (Hudson, 2007). Alternatively, some researchers are using the technique of heterologous hybridization to profile transcripts of nonmodel species with microarrays designed for a closely related model species (e.g., Gong *et al*., 2005; Travers *et al*., 2007). These various tools and approaches for studying gene expression mean that one can choose between studying model and nonmodel species to address genomic ecology questions in global change biology; however, each approach has constraints that are important to consider.

Limitations to molecular and functional inference in model and nonmodel species

Generally, in model species that have been fully sequenced and for which microarrays have been specifically designed (e.g., Arabidopsis, *Populus* sp., and rice), data describing the abundance of nearly all transcripts can be attributed to the relevant genes with a high degree of confidence. Nonspecific binding of products from two or more genes to a single probe on the microarray, or cross-hybridization, can cause problems if genes share very high sequence similarity, but is relatively rare (Shiu & Borevitz, 2008). Even when a full genome sequence becomes available, it is not immediately possible to (1) identify all the genes capable of being expressed to produce proteins, and (2) assign all the RNA transcripts being profiled to specific genes. However, bioinformatic techniques to identify genes are becoming increasingly efficient, especially when sequences from multiple species are analyzed in parallel (Lin *et al*., 2007).

High-quality transcript profile data are also available for species for which microarrays have been produced from expressed sequence tag libraries but the full genome sequence is not available, such as maize and soybean (Wang *et al*., 2003; Vodkin *et al*., 2004). However, there is less certainty that (1) each probe sequence on the microarray is unique to a single gene, or (2) every functional gene is detected by the microarray. For example, the soybean genechip from Affymetrix probes expression of $\sim$ 38 000 unique genes, while the recent Joint Genome Institute (http://www.jgi.doe.gov/) release of the soybean genome suggests there are 58 556 loci containing protein-coding transcripts (http://www.phytozome.net/soybean). Further assembly and analysis is needed before it is known how much the disparity in these numbers is explained by partial polyploidy (Schlueter *et al*., 2007).

Using heterologous hybridization to study transcript profiles of nonmodel species causes greater uncertainty about cross-hybridization or missing genes. A preliminary analysis from hybridizing the genomic DNA of the study species to the microarray can be useful in identifying which probes have no corresponding gene and therefore can be subsequently ignored. Although this reduces the number of genes whose expression can be profiled, heterologous hybridization has been used to identify genes important to drought stress, cold stress, and heavy metal tolerance (Gong *et al*., 2005; Hammond *et al*., 2006; Sharma *et al*., 2007). In studies on a single nonmodel species, errors associated with heterologous hybridization should be common to all treatments, which limit some of the problems in interpretation. By comparison, if the transcript profiles of multiple species are assessed with a common microarray platform, then sequence divergence among species could impact the efficiency of hybridization and falsely suggest differential transcript abundance. Comparing among species the results of hybridizing genomic DNA with microarrays can help quantify the extent of this problem and again eliminate probes likely to cause problems (Shiu & Borevitz, 2008).

Functional interpretation of microarray data is dependent on correct annotation of gene function. As sequence data from plants accumulates, finding means to efficiently and effectively analyze the sequences and assign annotation remains a major challenge (Dong *et al*., 2005). Arabidopsis has been the primary subject

of studies determining gene function in plants and, therefore, more (though far from all) genes have been annotated in this species, and annotations are generally accepted with the greatest degree of confidence. Currently, ~60% of the 28 152 protein coding genes in Arabidopsis have been annotated to a Gene Ontology (GO) molecular function, with 50% annotated to a GO biological process (http://www.geneontology.org). The majority of annotations are based on a computational analysis of the gene sequence. Therefore, even in the most well studied plant species much work remains to be done to experimentally determine gene function. In other species the function of some genes may have been directly determined, but the annotation of the great majority of genes is inferred from sequence similarity to genes in Arabidopsis. An automated BLAST search (Altschul *et al*., 1997) against a protein database accomplishes this task. Top BLAST matches are typically assigned an expectation value along with a putative function and GO terms associated with similar protein sequences. The more evolutionarily distant from Arabidopsis the subject species is, the greater the likelihood the gene sequence will have diverged, which increases uncertainty in the annotation. Nonetheless, many genes are highly conserved and can be annotated with confidence in a large number of distantly related species (Frickey *et al*., 2008). The BLAST procedure has the inherent flaw of propagating annotation errors from one species to another (Gilks *et al*., 2002), but remains the most practical choice for sequence annotation. As more sequence data from various species becomes available, interspecific sequence analyses are also proving valuable for improvement of annotations, automation of annotation, and identification of novel coding regions (Windsor & Mitchell-Olds, 2006).

Limitations to ecological inference in model and nonmodel species

The vast majority of species for which substantial sequence information and transcript profiling tools are available have been selected because of their economic importance (Table S1). This has created enormous potential for investigating the mechanisms underlying the impacts of global change on crop yield and agroecosystem function. Transcript profiling can reveal changes in gene expression that drive physiological and ecological responses, and in doing so improve understanding of mechanism at all scales (Fig. 1). Managed ecosystems and mesocosms incorporating model species are an excellent test bed for genomic ecology because their low genetic and environmental heterogeneity increases the statistical power of field experiments and facilitates detection of subtle treatment differences (Ainsworth *et al*., 2006; Casteel *et al*., 2008; Leakey *et al*., 2009; Zavala *et al*., 2008). In addition, the current group of model species incorporates considerable diversity including angiosperms and gymnosperms, herbaceous plants and trees, C_3 and C_4 species, legumes and nonlegumes, and tropical and temperate species. This allows further fundamental biological questions to be asked regarding variation in response to global change of major functional and phylogenetic groups. However, these species are not always ideal subjects for addressing a number of important ecological and evolutionary questions in global change biology. The majority are crops bred for rapid growth and reproductive output on annual growth cycles. This means that the mechanisms underlying their responses to resource availability, disturbance, and competition may differ from those of other species adapted to diverse habitats in natural communities. Custom-made transcript profiling tools are not currently available for multiple plant species from even one natural community. This limits characterization of species-specific gene expression patterns and its contribution to driving the species interactions that control community and ecosystem responses (Fig. 1). One solution would be to accept the limitations and assumptions of heterologous hybridization in order to assess diversity of gene expression responses across a larger number of species (Travers *et al*., 2007). Alternatively, custom genomic tools could be developed for the species comprising a 'model' ecosystem or species possessing ecological traits of particular interest. Such an approach is becoming increasingly feasible with continued advances in the development of high-throughput sequencing technologies (Hudson, 2007).

Expectations, design, and analysis of ecologically relevant transcript profiling experiments

Expectations of gene expression responses to global change scenarios

Experimentation under environmentally realistic conditions to investigate the molecular response of plants to meaningful changes in growth conditions and ecological interactions is the defining feature of the genomic ecology approach. A typical laboratory-based microarray study aiming to elucidate the functions of genes will subject plants to an acute treatment that precipitates many-fold changes in the transcript abundance of thousands of genes. In contrast, results from a typical genomic ecology experiment will reveal markedly smaller magnitude changes in the abundance of transcripts from a smaller number of genes. This probably has two main causes: (1) the imposed treatments are less severe, and (2) the focus is often on plants that have

acclimated to the treatments, in many cases spending their entire lifecycle exposed to the given treatment. In field studies there are the additional distinguishing factors of greater noise in gene expression resulting from the variable growth conditions and the greater resilience of field grown plants than laboratory grown plants to perturbation.

Treatments in global change biology experiments are typically mild (e.g., a 40% difference in [CO_2]), because they aim to test the impact of changes between average field conditions today and those expected for later this century. By comparison, many molecular studies aiming to identify stress responsive genes have ensured significant treatment effects would be observed by imposing extreme conditions, such as supply of strong (200–500 mM) salt solutions (Bohnert *et al*., 2001), exposure to high (300 ppb) ozone concentrations (Tosti *et al*., 2006), and withholding water from plants in small pots of rapidly drying growth media (Talame *et al*., 2007). Important data have been generated from such experiments, but the results may not always inform us about the mechanisms controlling plant performance in the field. For example, a cell-death response leading to lesions on leaves has been identified as an important component of response in plants exposed to >300 ppb [O_3] (500% above background), but growth at <100 ppb [O_3] (60% above background) impairs productivity without causing visible damage to the plant (reviewed by Long & Naidu, 2002).

When plants experience a change in growing conditions (e.g., transfer from moderate to high temperature), they display a progression of responses. First, the altered condition is sensed, activating a signal transduction pathway, which typically drives metabolic adjustments and concludes with adoption of a new acclimated state. Well-studied examples are the time courses of cellular response to ozone exposure and attack by pathogens (Lamb & Dixon, 1997; Kangasjarvi *et al*., 2005). The changes in gene expression immediately and shortly after the change in condition are substantial in number and magnitude. Most studies aiming to understand the molecular basis of plant responses to abiotic and biotic stimuli have focused on characterizing the responses to short-term changes in conditions. This is very important for understanding the sensing and signaling processes that control the response. Also, in combination with strong treatments, the brief shock generates an easy-to-detect response. However, these short-term changes in gene expression do not reveal all the important controls of plant performance upon acclimation to the growth conditions. For example, when assessed using a high-density maize oligonucleotide array, far fewer ($<2\%$ vs. 27%) genes showed differential expression in maize ear tissue under a gradually developing stress than under a sudden stress (Campos *et al*., 2004).

Many genomic ecology studies are building on information from experiments employing acute treatments to determine gene function by characterizing the more subtle changes in gene expression that differentiate fully acclimated plant performance in different experimental treatments. This forces microarray studies to be designed and analyzed differently. For example, it is very logical to focus primarily on changes in transcript abundance of >1.5-fold if the objective is to identify components of a signal transduction pathway a specific number of hours following a stimulus (e.g., Tosti *et al*., 2006). Equally, fivefold changes in transcript abundance for metabolic genes are unlikely to be observed in plants that are fully acclimated to growth in two mildly different treatments. For example, in Free-Air Concentration Enrichment (FACE) experiments where, in many cases plants have been grown for their entire life cycles at current and elevated [CO_2], the largest fold changes in transcript abundance due to the CO_2 treatment are typically ca. twofold (Gupta *et al*., 2005; Taylor *et al*., 2005; Ainsworth *et al*., 2006; Leakey *et al*., 2009). Identifying these moderate changes can give considerable insight into alterations in metabolic pathways and allocation to biosynthetic pathways that occur over time in response to elements of global change. But the genomic ecologist is faced with the problem of balancing the cost of transcript profiling with the need for adequate replication to gain sufficient statistical power to detect small fold changes in transcript abundance.

By comparison with controlled environment facilities, field conditions can provide growing conditions for plants that are simultaneously more variable, more resource rich, and more stressful. For example, many habitats provide high light and unlimited rooting volume but also periods of water deficit and disease. This appears to reduce the sensitivity with which gene expression responds to stress treatments. For example, application of benzo(1,2,3)-thiadiazole-7-carbothioic acid *S*-methylester (BTH) to induce systemic resistance against pathogens in wheat caused substantial upregulation of defense-related genes in a greenhouse trial. However, when the experiment was repeated under field conditions, defense-related gene expression was constitutively high and did not increase further with the BTH treatment (Pasquer *et al*., 2005).

Experiments that investigate the response of plants to treatments simulating global change over long time periods are informative because they can generate understanding of (1) impacts over the entire life histories of the subject species, (2) slow ecological responses such as competition and succession, and (3) complex feedbacks

from ecological and ecosystem scale to whole plant performance (Fig. 1). Fewer space restrictions allow long-term experiments to be done in the field more successfully than under controlled environment conditions. However, plants in the field, and especially those in long-term studies, experience variable growth conditions on scales from minutes, hours, and days to months and seasons. Many of the parameters of ecological interest, for example, biomass, yield, and fecundity integrate these growth conditions over long periods of time. In contrast, transcript profiles in plants are known to respond rapidly and extensively to temperature (Seki *et al*., 2002) and light (Bertrand *et al*., 2005), show circadian rhythms (Michael & McClung, 2003; Blasing *et al*., 2005) and vary with development (Taylor *et al*., 2005; Ainsworth *et al*., 2006). Because a single sampling point only represents a snapshot view, it is important to distinguish responses of the transcriptome that are due to the experimental manipulation vs. time or weather-dependent changes (Miyazaki *et al*., 2004). This discrimination can be achieved to a significant extent by sampling at the same time each day, sampling on multiple occasions over the duration of an experiment and interpreting treatment effects on gene expression in the context of environmental data. In addition, efforts to sample homogenous tissue that is at the same developmental stage and growing under the same environmental conditions minimize unwanted variability that could prevent detection of treatment effects. In some cases the impact of natural variation in growth conditions on gene expression can provide novel understanding of the mechanisms underlying plant–environment interactions. For example, transcript profiling of pine trees grown in multiple field sites in Europe indicated that cold tolerance develops in response to combined photoperiodic and temperature cues (Joosen *et al*., 2006).

Design of experiments assessing gene expression responses to global change scenarios

Nettleton (2006) reviewed how the basic principles of experimental design apply to transcript profiling experiments, with emphasis on random assignment of experimental units to treatments, use of the maximum affordable replication and applying blocking. These issues are familiar to ecologists and physiologists, and have been extensively reviewed (e.g., Scheiner & Gurevitch, 2001). The more specific importance of understanding the distinction between, and value of, technical and biological replication in transcript profiling experiments has been highlighted by Allison *et al*. (2006) and Nettleton (2006). Technical replication provides multiple measures of a single sample from a single experimental unit. Biological replication involves measurements of multiple experimental units each of which is independently exposed to control or treatment conditions. Without biological replication it is not possible to statistically attribute observed changes in transcript abundance to the effects of a treatment. Most experiments are limited by the funds available for transcript profiling. The power to detect treatment effects will be maximized if the transcripts from each experimental unit at a given time are profiled with only one microarray (Nettleton, 2006). However, if the number of biological replicates is limited (e.g., at a Free-Air CO_2 Enrichment experiment) and there is significant measurement error, averaging across technical replicates can reduce variability and provide some gain in statistical power (Nettleton, 2006).

In ecological experiments and especially those in the field, variation in gene expression responses to experimental treatments over time are of great interest with respect to circadian/diel rhythms, interactions with climate, acclimation, and development. With a limited supply of microarrays, this creates both challenges and opportunities. If the primary aim of the experiment is to characterize treatment effects on gene expression at a single time point (e.g., a single development event such as flowering), then adding biological replicates will provide the most statistical power. If the primary aim of the experiment is to characterize the average treatment effects on gene expression (e.g., over a growing season), then it may be desirable to compromise technical or biological replication in order to allow additional sampling points over time. Of course, such trade-offs need to be determined on a case-by-case basis. Even for studies on the same species at a single field site, some experiments may necessitate technical replication (e.g., Ainsworth *et al*., 2006), while others benefit most from multiple measurements in time (e.g., Casteel *et al*., 2008).

Subsampling is often used to overcome the variation among individuals within a replicate plot in field experiments. For instance, averaging the rates of photosynthesis of four different sun leaves within individual plots of maize exposed to either ambient or elevated [CO_2] reduced variation among replicate plots and ensured there was sufficient statistical power to characterize a subtle, episodic treatment effect (Leakey *et al*., 2004, 2006a). In genomic ecology studies, one solution to the need for sampling variation within replicate plots without depleting microarray resources needed for sampling multiple biological replicates or time points is to pool mRNA from multiple samples collected within a single plot (Allison *et al*., 2006). Hybridizing this mixed mRNA sample to a single microarray will reduce between plot variance when biological variability is

high relative to measurement error (Kendziorski *et al*., 2005). This approach has been used successfully in transcript profiling studies of poplar and soybean responses to elevated [CO_2] in the field (e.g., Gupta *et al*., 2005; Taylor *et al*., 2005; Ainsworth *et al*., 2006).

Analysis of gene expression responses to global change treatments

One of the greatest challenges of transcript profiling is the data analysis. This is partly due to the large size of the datasets compared with most physiological or ecological experiments. Selecting from the large number of rapidly developing analysis tools and techniques that are available is also challenging. Although it is impossible to comprehensively discuss the advantages and disadvantages of all the available options here, it is worth briefly reviewing the major steps in the analysis process and highlighting a number of specialist reviews on the subject (e.g., Allison *et al*., 2006; Nettleton, 2006).

The first analysis step involves processing the images of the fluorescent spots on each microarray. Many approaches have been developed, and the service facilities that perform the hybridization and scanning of microarrays for most investigators can assist in making the appropriate choices. Before proceeding with data analysis, it is important to perform quality control steps and remove or replace data from defective slides or images. One simple method for eliminating poor quality data is to discount data from microarrays that do not meet threshold values of the Pearson correlation coefficient (e.g., 0.9) or kappa statistic (e.g., 0.75) when pairwise comparisons are made between microarrays from a given treatment and time point (Fleiss, 1981; McIntyre *et al*., 2006).

A number of microarrays include probes for more than one species. For example, the Affymetrix soybean genechip includes probes for genes from soybean, a nematode species and the phytopthera pathogen. If only transcripts from soybean are to be profiled, the data for probes specific to the other species should be disregarded. The Affymetrix genechip platform provides a statistic estimating whether each individual transcript is considered to have been present or absent from the sample (Affymetrix, 2002). This allows the investigator to discount data from probes for which a transcript was not considered present in a sufficient number of samples for meaningful replication to be achieved. This reduces the number of tests to be performed and prevents misinterpretation of results from probes for which there is not sufficient statistical power for meaningful testing.

Some methods can combine image analysis algorithms with the next analysis step, which is normalization. Normalization is the process that makes adjustments to minimize the influence of technical variability across different microarrays and experiments. The simplest approach involves normalizing the fluorescence intensity for individual genes by the median fluorescence intensity on an individual microarray basis (e.g., Ainsworth *et al*., 2006; Fung *et al*., 2008). This approach has the philosophical advantage of maintaining the independence of data from individual replicates, and the practical advantage of requiring a single normalization to be performed on a given microarray, even if the data are to be analyzed as part of more than one experiment. Alternatively, more complex procedures have also been developed, some of which incorporate information from all the chips in an experiment as part of the normalization process. Normalization is an area of ongoing research in which there is an unresolved debate about which method performs the best, and even how good performance should be defined (Irizarry *et al*., 2003; Bolstad *et al*., 2003; Choe *et al*., 2005; Allison *et al*., 2006). After normalization, log transformation of the data is performed in nearly all cases to ensure that the data are normally distributed.

The majority of published microarray studies use mixed-effects linear models to identify treatment effects on transcript abundance, with an independent analysis being performed for each probe in the dataset (Nettleton, 2006). This has the advantage of allowing the physiologist or ecologist to use familiar statistical tests and software packages. An additional reason for the approach is that different genes display different levels of variation in expression, creating heterogeneity that a single 'global' model has difficulty representing. However, it has been suggested that analyzing each gene independently is inefficient (Allison *et al*., 2006). Simulation studies have indicated that an intermediate approach, called variance shrinkage, which combines data from specific genes and all genes may perform better than gene-by-gene testing (Cui *et al*., 2005), although optimization of the technique is still required (Allison *et al*., 2006).

Because the analysis of most microarray experiments necessitates tens of thousands of statistical tests on individual probes, there is a greater likelihood of making type I errors (falsely identifying the abundance of transcripts as responsive to the treatment when in fact they are not) than in most physiological or ecological experiments. Consequently, techniques have been developed that quantify the false discovery rate (FDR) and allow it to be controlled (Benjamini & Hochberg, 1995; Storey & Tibshirani, 2003). Most commonly, this is done by adjusting the probability threshold at which treatment effects on transcript abundance are considered to be statistically significant, taking into account the

number of tests performed and the initial *P*-value returned for each transcript by the mixed-effects linear model. Importantly, while applying increasingly strict FDR corrections reduces the number of transcripts falsely identified to respond significantly to the treatment, it also increases the number of transcripts falsely identified *not* to respond significantly to the treatment (type II errors; Nettleton, 2006). In other words, there is a trade-off between identifying fewer genes than actually responded to the treatment, but with a high degree of confidence (strict FDR) vs. more genes that actually responded to the treatment, plus some that did not (relaxed FDR). During the experimental design and analysis processes, each researcher must select the FDR correction level that allows the most meaningful interpretation of the data.

In many global change biology experiments, where treatment effects can be small, applying strict FDR can result in few transcripts being identified as responding to the treatment. If more relaxed FDR are applied, other techniques are necessary to increase the confidence with which 'responsive' transcripts are identified. For instance, visualization of transcript data in the context of known metabolic pathways and signal transduction cascades can indicate when many transcripts associated with a common function or response display consistent responses to an experimental treatment (e.g. Leakey *et al*., 2009). If transcripts are identified as a result of random variation and not a true treatment effect, then positive and negative responses should be equal in number. However, for example, if the abundance of $\geq 50\%$ of all transcripts-encoding enzymes involved in the synthesis of flavonoids are greater when soybean grows at elevated [O_3], and no transcripts show the opposite result (Casteel *et al*., 2008), there is a good probability that the result is real rather than the result of random chance.

Difficulties associated with performing many tests can also be dealt with by putting transcripts into functional groups and performing a Fisher's exact test or chi-square test on each group. These two tests allow identification of groups within which a greater fraction of transcripts respond significantly to the treatment than on average across all transcripts, that is, functional groups of transcripts which disproportionately contribute to the overall transcriptional response. For example, the transcriptional response of soybean to growth at elevated [CO_2] was assessed by assigning each of the profiled transcripts into one of 32 functional groups (Leakey *et al*., 2009). A Fisher's exact test determined that the fraction of CO_2-responsive transcripts in functional groups related to respiration was significantly greater than the fraction of CO_2-responsive transcripts across all other functional groups.

The standard procedure of repeating an experiment can also be used to increase confidence in identification of 'responsive' transcripts. Transcripts whose abundance changes as a result of real treatment effects are more likely to display consistent changes in abundance of a similar magnitude and in the same direction. In contrast, false positives that have low *P*-values from the initial analysis of variance (ANOVA) as a result of random variation are equally likely to respond positively or negatively to the treatment in any given experiment. Varying the FDR threshold has a substantial impact on identification of transcripts which respond consistently in soybean grown at ambient and elevated [CO_2] over two consecutive growing seasons (Fig. 2). At an FDR of 0.2, 76 transcripts responded consistently in the 2 years, and no transcript displayed opposite responses in the 2 years. By contrast, applying an FDR of 0.5 to the same data identified 615 transcripts that responded consistently and 12 transcripts displaying opposite responses in the 2 years. The researcher has to choose between identifying treatment effects on 76 transcripts with a higher degree of confidence from a more conservative FDR correction or 615 transcripts from a less conservative FDR correction, plus the knowledge the transcript responded to the treatment in the same direction, and to a similar magnitude, in 2 consecutive years. Given the

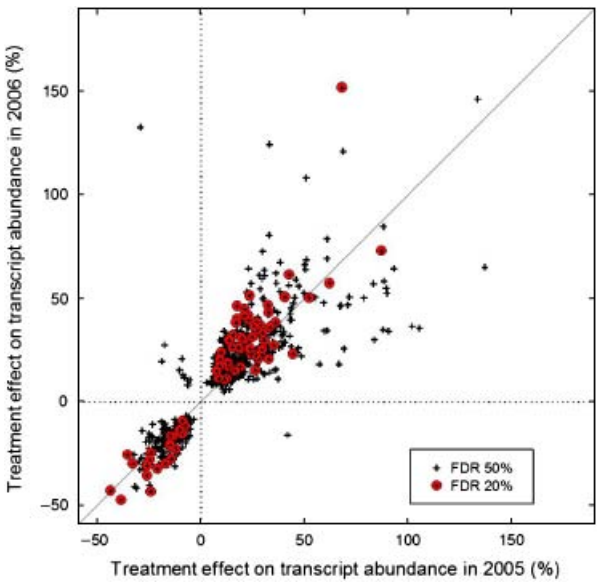


Fig. 2 Comparison of changes in transcript abundance in soybean leaves as a result of growth at ambient [CO_2] vs. elevated [CO_2] at SoyFACE during the 2005 and 2006 growing seasons. At a false discovery rate (FDR) of 0.2, 76 transcripts responded consistently in the 2 years and no transcript displayed opposite responses in the 2 years. By contrast, applying an FDR of 0.5 to the same data identified 615 transcripts that responded consistently and 12 transcripts displaying opposite responses in the 2 years. Data adapted from Leakey *et al*. (2009).

need to demonstrate that changes in transcript abundance have an impact on biochemical or physiological processes, there will be subsequent opportunity to eliminate false positives that have passed this initial analysis.

It is a common practice to validate the quantification of transcript abundance by microarrays using qRT-PCR on a subset of genes from the original experimental samples (Rajeevan *et al*., 2001). However, this practice has recently been suggested to be of little benefit because, while qRT-PCR probably provides a more accurate measure of transcript abundance, there is no reason to expect the new data will eliminate the types of errors that cause identification of false positives (Allison *et al*., 2006). This may not yet be a consensus view, but seems to be consistent with most physiological and ecological practices. For example, measurements of stomatal conductance using IRGA-based gas exchange systems are not typically validated with measurements using a porometer (e.g., Jones, 1999; Leakey *et al*., 2006b). On the other hand, qRT-PCR is more sensitive to changes in transcript abundance than microarrays and it is incredibly valuable and cost effective if transcripts identified in a microarray study are profiled by quantitative-PCR in samples from additional biological replicates, other tissues, or other time points. Such follow-up studies are vital to extend investigation from broad profiling analyses to detailed understanding of specific gene responses.

Linkages from gene expression to physiology and ecology

The ability to measure gene-specific and genome-wide patterns of transcript abundance provides a new opportunity to improve our understanding of how organisms and ecosystems respond to environmental change. Different elements of global change can elicit distinct changes in gene expression (e.g., drought vs. heat; Roelofs *et al*., 2008); therefore, the contribution of two simultaneous treatments in impacting physiological performance could start to be dissected by the molecular phenotypes. Because transcript profiling with microarrays potentially provides information on a large proportion of metabolic and signaling components, it is an ideal technique to broadly survey intra- and interspecific variation in response to a given treatment (e.g., Gong *et al*., 2005). Identifying different response pathways or magnitudes of response within a pathway at the molecular level identifies a smaller group of candidate mechanisms that can then be more easily examined at the physiological and ecological scale. For example, Leakey *et al*. (2009) used microarrays to characterize a transcriptionally driven acclimation of soybean to growth at elevated [CO_2], which led to stimulated foliar respiration. The transcript profiling also allowed a survey of biosynthetic metabolism to identify pathways that were transcriptionally upregulated coincident with the enhanced supply of energy and carbon skeletons from respiration. The ecological significance of these changes can now be evaluated in more detailed analyses. By comparison, previous methods would probably have involved laborious and less systematic investigation of individual biosynthetic pathways in different species by different research groups.

Interpretation of transcript abundance depends on assumptions about the relationship between the levels of transcripts and the functional activity of the proteins they encode. This is difficult to predict because post-transcriptional and posttranslational regulation can significantly alter the response predicted from transcript data alone (Scheible *et al*., 1997; Kaiser & Huber, 2001; Hendriks *et al*., 2003). In addition, the impact of changes in transcript abundance on a biological response depends on the turnover rate of the encoded proteins, their contribution to the control of metabolic pathways and the levels of metabolites associated with those pathways, which in turn can regulate the expression of the given gene. Genome-wide transcript profiling and analyses of enzyme activities have shown that transcript levels undergo marked and rapid changes during the diurnal cycles whereas changes in enzyme activities are often smaller and delayed, and appear to integrate changes in transcript levels over several diurnal cycles (Gibon *et al*., 2006; Morcuende *et al*., 2007; Stitt *et al*., 2007). Because transcripts, enzymes, and metabolites integrate information over different time scales, measuring their response provides a wider physiological snapshot than transcript abundance alone. Fortunately, unlike transcriptomics (and proteomics) which relies to a great extent on genomic information, metabolomics is widely applicable with only minimal time required to reoptimize protocols for a new species (Schauer & Fernie, 2006). High-throughput analysis of activity from >20 enzymes is now a reality (Gibon *et al*., 2004) and early indications suggest that these methods can also be transferred relatively easily among species (Rogers & Gibon, 2009). Although still a nascent field of investigation, techniques to model metabolic networks (Sweetlove & Fernie, 2005) and a diversity of bioinformatics tools are becoming available to aid in identifying genes that underlie important biological functions.

For transcriptomic and metabolomic data, visualization of the results in a biologically meaningful way is another challenge to functional interpretation. Thimm *et al*. (2004) introduced MAPMAN, a user-driven visualization tool for displaying transcript, metabolite, and enzyme activity datasets on plant-specific biological pathways. MAPMAN

is a flexible program that classifies genes into specific functional bins (e.g., photosynthesis, glycolysis, secondary metabolism), originally developed for Arabidopsis. It has since been extended to Solanaceous species (Urbanczyk-Wochniak *et al*., 2006) and legumes (Goffard & Weiller, 2006; Leakey *et al*., 2009) based on BLAST hits to the Arabidopsis proteome and the nonredundant protein database at NCBI. MAPMAN is but one example of a biologically relevant visualization tool. Such resources to interpret gene expression results are becoming ever more sophisticated and available for an increasing number of species. Many of the genes involved in photosynthesis, respiration, and nutrient acquisition can be identified using such software and results subsequently related to the response of plants to altered environmental conditions. Microarrays and bioinformatics, therefore, make a compelling combination to characterize mechanisms responsible for how plants respond to experimental manipulations of temperature, water, ozone, and CO_2 concentration (Watkinson *et al*., 2003; Ainsworth *et al*., 2006; Li *et al*., 2006; Weston *et al*., 2008).

Modeling gene expression data in the context of existing biochemical frameworks is useful, but requires that we understand *a priori* relationships between variables used to connect genes to physiology and beyond to ecosystem-scale processes. One challenge with this modeling approach is in reducing the dimensionality of the gene expression data before linking with the rest of the model. There are a number of approaches to accomplish this, such as the use of gene function ontologies to define a subset of genes whose expression values would subsequently be included in the model. Gene ontologies are insightful, but the functions of many genes are still not fully understood and extrapolation of model gene function to nonmodel genes is potentially problematic. Therefore, unsupervised approaches for delineating gene expression into functional clusters are promising. Weighted gene coexpression network analysis is encouraging in this regard because it is an unsupervised approach for clustering genes that share highly correlated expression patterns across treatment (Zhang & Horvath, 2005). Furthermore, the input data for this network approach are from normalized raw intensity values and thereby avoid multiple testing errors commonly associated with most expression array analytical techniques. Using this technique, Weston *et al*. (2008) were able to cluster Arabidopsis genes into functionally relevant stress responsive clusters (modules) that were then correlated to phenotypic characteristics. A similar statistical approach could be used to investigate module gene correlations with metabolites and enzyme activities of interest to strengthen understanding of the metabolic pathways governing phenotype.

Linking the large-scale datasets of genomic ecology to other predictors of plant responses to global change, including soil properties, biotic interactions, and climate conditions, presents several challenges. First, the spatial and temporal resolution of data collected across different levels of biological organization (i.e., molecular, organismal, community, and ecosystem) can vary significantly. In addition, accounting for the hierarchical structure of data can improve predictive accuracy when using multiple variables to explain observed plant responses. Statistical methods based on *probabilistic graphical models* provide a natural framework for modeling responses to environmental treatments. In this approach, the probabilistic relationships defining a complex system are specified via a sequence of nodes that represent random variables, and edges that encode direct physical or statistical dependencies (Jordan, 2004). The ability of graphical models to include *latent* or *hidden* variables to explicitly model unobserved relationships is particularly useful in biological research. A variety of computational methods developed by the statistics and machine-learning communities have been used to effectively analyze biological data with complex spatial and temporal structure. Directed graphical models, or Bayesian networks, are commonly used in systems biology to learn the structure of complex genetic networks (Blanchard, 2004; Friedman, 2004). Related multivariate modeling approaches such as structural equation modeling (SEM) have been used to identify the environmental and biotic predictors that influence plant response to various global change factors (Grace, 2006; Clark *et al*., 2007). Bayesian networks and SEM are only two examples of tools being used to analyze complex ecological responses to global change. These approaches provide powerful statistical tools that can be used to model plant responses to global change across levels of biological organization.

Conclusion

In summary, the technology to assess gene expression through transcript profiling is now available for model and nonmodel species. Managed ecosystems and mesocosms are proving to be good testbeds for the genomic ecology approach. Major advances in understanding natural communities are also promised by the increasing number of species for which transcript profiling tools are available and the accelerating advances in sequencing technology. This represents a significant new opportunity to assess the mechanisms underlying the responses of plants to elements of global change. The studies that have been performed to date have revealed some important distinctions between transcript profiling in ecological studies vs. molecular stu-

dies of gene function. This experience has allowed us to identify the strengths and weaknesses of various experimental design and analysis options available to the genomic ecologist. Possibly the biggest change resulting from the use of genomic tools is a new, integrative approach to investigating the abiotic and biotic interactions of plants. Using genomic ecology to understand the mechanisms currently consigned to the 'black box' of plant function will significantly advance analysis of future global change, its impacts on ecosystems and how we should respond to it.

Acknowledgements

We acknowledge support from the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research program as part of its Program for Ecosystem Research. A. D. B. L., E. A. A., and D. R. O. were supported by grant no. DE-FG02-04ER63849. S. A. P., S. M. B., and E. A. S. were supported by contract no. DE-AC03-76SF00098 to Lawrence Berkeley National Laboratory. A. R. was supported by contract no. DE-AC02-98CH10886 to Brookhaven National Laboratory. D. J. W. and S. D. W. were supported by contract DE-AC05-00OR22725 to UT-Battelle, LLC, which manages Oak Ridge National Laboratory for the DOE.

References

Affymetrix (2002) *Statistical algorithms description document*. Technical Report. http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf.

Ainsworth EA, Rogers A, Vodkin LO, Walter A, Schurr U (2006) The effects of elevated CO_2 concentration on soybean gene expression. An analysis of growing and mature leaves. *Plant Physiology*, **142**, 135–147.

Allison DB, Cui XQ, Page GP, Sabripour M (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nature Reviews Genetics*, **7**, 55–65.

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B – Methodological*, **57**, 289–300.

Bertrand C, Benhamed M, Li YF *et al*. (2005) ArabidopsisHAF2 gene encoding TATA-binding protein (TBP)-associated factor TAF1 is required to integrate light signals to regulate gene expression and growth. *Journal of Biological Chemistry*, **280**, 1465–1473.

Blanchard JL (2004) Bioinformatics and systems biology, rapidly evolving tools for interpreting plant response to global change. *Field Crops Research*, **90**, 117–131.

Blasing OE, Gibon Y, Gunther M *et al*. (2005) Sugars and circadian regulation make major contributions to the global regulation of diurnal gene expression in Arabidopsis. *Plant Cell*, **17**, 3257–3281.

Bohnert HJ, Ayoubi P, Borchert C *et al*. (2001) A genomics approach towards salt stress tolerance. *Plant Physiology and Biochemistry*, **39**, 295–311.

Bolstad BM, Irizarry RA, Astrand M, Speed TP (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, **19**, 185–193.

Bradley KL, Pregitzer KS (2007) Ecosystem assembly and terrestrial carbon balance under elevated CO_2. *Trends in Ecology and Evolution*, **22**, 538–547.

Busch W, Lohmann JU (2007) Profiling a plant: expression analysis in Arabidopsis. *Current Opinion in Plant Biology*, **10**, 136–141.

Campos H, Cooper A, Habben JE, Edmeades GO, Schussler JR (2004) Improving drought tolerance in maize: a view from industry. *Field Crops Research*, **90**, 19–34.

Casteel CL, O'Neill BF, Zavala JA, Bilgin DD, Berenbaum MR, DeLucia EH (2008) Transcriptional profiling reveals elevated CO_2 and elevated O_3 alter resistance of soybean (*Glycine max*) to Japanese beetles (*Popillia japonica*). *Plant, Cell and Environment*, **31**, 419–434.

Choe SE, Boutros M, Michelson AM, Church GM, Halfon MS (2005) Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, **6**, R16.

Clark CM, Cleland EE, Collins SL *et al*. (2007) Environmental and plant community determinants of species loss following nitrogen enrichment. *Ecology Letters*, **10**, 596–607.

Cui XG, Hwang JTG, Qiu J, Blades NJ, Churchill GA (2005) Improved statistical tests for differential gene expression by shrinking variance components estimates. *Biostatistics*, **6**, 59–75.

Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK (2004) Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant Journal*, **38**, 366–379.

DeLucia EH, Casteel CL, Nabity PD, O'Neill BF (2008) Insects take a bigger bite out of plants in a warmer, higher carbon dioxide world. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 1781–1782.

Dong QF, Lawrence CJ, Schlueter SD, Wilkerson MD, Kurtz S, Lushbough C, Brendel V (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiology*, **139**, 610–618.

Fleiss JL (1981) *Statistical Methods for Rates and Proportions*. John Wiley and Sons, New York.

Frickey T, Benedito VA, Udvardi M, Weiller G (2008) AffyTrees: facilitating comparative analysis of affymetrix plant microarray chips. *Plant Physiology*, **146**, 377–386.

Friedman N (2004) Inferring cellular networks using probabilistic graphical models. *Science*, **303**, 799–805.

Fuhrer J (2003) Agroecosystem responses to combinations of elevated CO_2, ozone, and global climate change. *Agriculture Ecosystems and Environment*, **97**, 1–20.

Fung RWM, Gonzalo M, Fekete C *et al*. (2008) Powdery mildew induces defense-oriented reprogramming of the transcriptome in a susceptible but not in a resistant grapevine. *Plant Physiology*, **146**, 236–249.

Gibon Y, Blaesing OE, Hannemann J *et al*. (2004) A robot-based platform to measure multiple enzyme activities in Arabidopsis

using a set of cycling assays: comparison of changes of enzyme activities and transcript levels during diurnal cycles and in prolonged darkness. *Plant Cell*, **16**, 3304–3325.

Gibon Y, Usadel B, Blaesing OE, Kamlage B, Hoehne M, Trethewey R, Stitt M (2006) Integration of metabolite with transcript and enzyme activity profiling during diurnal cycles in Arabidopsis rosettes. *Genome Biology*, **7**, 76.

Gilks WR, Audit B, De Angelis D, Tsoka S, Ouzounis CA (2002) Modeling the percolation of annotation errors in a database of protein sequences. *Bioinformatics*, **18**, 1641–1649.

Goffard N, Weiller G (2006) Extending MapMan: application to legume genome arrays. *Bioinformatics*, **22**, 2958–2959.

Gong QQ, Li PH, Ma SS, Rupassara SI, Bohnert HJ (2005) Salinity stress adaptation competence in the extremophile *Thellungiella halophila* in comparison with its relative Arabidopsis. *Plant Journal*, **44**, 826–839.

Grace JB (2006) *Structural Equation Modeling and Natural Systems*. Cambridge University Press, Cambridge.

Gupta P, Duplessis S, White H, Karnosky DF, Martin F, Podila GK (2005) Gene expression patterns of trembling aspen trees following long-term exposure to interacting elevated CO_2 and tropospheric O_3. *New Phytologist*, **167**, 129–142.

Hammond JP, Bowen HC, White PJ *et al*. (2006) A comparison of the *Thlaspi caerulescens* and *Thlaspi arvense* shoot transcriptomes. *New Phytologist*, **170**, 239–260.

Hendriks JHM, Kolbe A, Gibon Y, Stitt M, Geigenberger P (2003) ADP-glucose pyrophosphorylase is activated by posttranslational redox-modification in response to light and to sugars in leaves of Arabidopsis and other plant species. *Plant Physiology*, **133**, 838–849.

Hudson ME (2007) Sequencing breakthroughs for genomic ecology and evolutionary biology. *Molecular Ecology Resources*, **8**, 3–17.

Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of affymetrix GeneChip probe level data. *Nucleic Acids Research*, **31**, e15.

Jackson RB, Linder CR, Lynch M, Purugganan M, Somerville S, Thayer SS (2002) Linking molecular insight and ecological research. *Trends in Ecology and Evolution*, **17**, 409–414.

Jones HG (1999) Use of thermography for quantitative studies of spatial and temporal variation of stomatal conductance over leaf surfaces. *Plant, Cell and Environment*, **22**, 1043–1055.

Joosen RVL, Lammers M, Balk PA *et al*. (2006) Correlating gene expression to physiological parameters and environmental conditions during cold acclimation of *Pinus sylvestris*, identification of molecular markers using cDNA microarrays. *Tree Physiology*, **26**, 1297–1313.

Jordan MI (2004) Graphical Models. *Statistical Science*, **19**, 140–155.

Kaiser WM, Huber SC (2001) Post-translational regulation of nitrate reductase: mechanism, physiological relevance and environmental triggers. *Journal of Experimental Botany*, **52**, 1981–1989.

Kangasjarvi J, Jaspers P, Kollist H (2005) Signalling and cell death in ozone-exposed plants. *Plant, Cell and Environment*, **28**, 1021–1036.

Kendziorski C, Irizarry RA, Chen KS, Haag JD, Gould MN (2005) On the utility of pooling biological samples in microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 4252–4257.

Lamb C, Dixon RA (1997) The oxidative burst in plant disease resistance. *Annual Review of Plant Physiology and Plant Molecular Biology*, **48**, 251–275.

Leakey ADB, Bernacchi CJ, Dohleman FG, Ort DR, Long SP (2004) Will photosynthesis of maize (*Zea mays*) in the US Corn Belt increase in future [CO_2] rich atmospheres? An analysis of diurnal courses of CO_2 uptake under Free-Air Concentration Enrichment (FACE). *Global Change Biology*, **10**, 951–962.

Leakey ADB, Bernacchi CJ, Ort DR, Long SP (2006b) Long-term growth of soybean at elevated [CO_2] does not cause acclimation of stomatal conductance under fully open-air conditions. *Plant, Cell and Environment*, **29**, 1794–1800.

Leakey ADB, Uribelarrea M, Ainsworth EA, Naidu SL, Rogers A, Ort DR, Long SP (2006a) Photosynthesis, productivity, and yield of maize are not affected by open-air elevation of CO_2 concentration in the absence of drought. *Plant Physiology*, **140**, 779–790.

Leakey ADB, Xu F, Gillespie KM, McGrath JM, Ainsworth EA, Ort DR (2009) The genomic basis for stimulated respiratory carbon loss to the atmosphere by plants growing under elevated [CO_2]. *Proceedings of the National Academy of Sciences*, in press.

Li PH, Mane SP, Sioson AA, Robinet CV, Heath LS, Bohnert HJ, Grene R (2006) Effects of chronic ozone exposure on gene expression in Arabidopsis ecotypes and in *Thellungielia halophila*. *Plant, Cell and Environment*, **29**, 854–868.

Lin MF, Carlson JW, Crosby MA *et al*. (2007) Revisiting the protein-coding gene catalog of *Drosophila melanogaster* using 12 fly genomes. *Genome Research*, **17**, 1823–1836.

Long SP, Ainsworth EA, Leakey ADB, Nosberger J, Ort DR (2006) Food for thought: lower-than-expected crop yield stimulation with rising CO_2 concentrations. *Science*, **312**, 1918–1921.

Long SP, Naidu SL (2002) Effects of oxidants at the biochemical, cell and physiological levels, with particular reference to ozone. In: *Air Pollution and Plant Life* (eds Bell JNB, Treshow M), pp. 69–88. John Wiley and Sons, Chichester.

McIntyre LM, Bono LM, Genissel A *et al*. (2006) Sex-specific expression of alternative transcripts in *Drosophila*. *Genome Biology*, **7**.

Michael TP, McClung CR (2003) Enhancer trapping reveals widespread circadian clock transcriptional control in Arabidopsis. *Plant Physiology*, **132**, 629–639.

Miyazaki S, Fredricksen M, Hollis KC *et al*. (2004) Transcript expression profiles of Arabidopsis grown under controlled conditions and open-air elevated concentrations of CO_2 and of O_3. *Field Crops Research*, **90**, 47–59.

Morcuende R, Bari R, Gibon Y *et al*. (2007) Genome-wide reprogramming of metabolism and regulatory networks of Arabidopsis in response to phosphorus. *Plant, Cell and Environment*, **30**, 85–112.

Nettleton D (2006) A discussion of statistical methods for design and analysis of microarray experiments for plant scientists. *Plant Cell*, **18**, 2112–2121.

Ouborg NJ, Vriezen WH (2007) An ecologist's guide to ecogenomics. *Journal of Ecology*, **95**, 8–16.

Pasquer F, Isidore E, Zarn J, Keller B (2005) Specific patterns of changes in wheat gene expression after treatment with three antifungal compounds. *Plant Molecular Biology*, **57**, 693–707.

Poorter H (1993) Interspecific variation in the growth-response of plants to an elevated ambient CO_2 concentration. *Vegetatio*, **104**, 77–97.

Rajeevan MS, Ranamukhaarachchi DG, Vernon SD, Unger ER (2001) Use of real-time quantitative PCR to validate the results of cDNA array and differential display PCR technologies. *Methods*, **25**, 443–451.

Roelofs D, Aarts MGM, Schat H, van Straalen NM (2008) Functional ecological genomics to demonstrate general and specific responses to abiotic stress. *Functional Ecology*, **22**, 8–18.

Rogers A, Gibon Y (2009) Enzyme kinetics: theory and practice. In: *Plant Metabolic Networks* (ed. Schwender J) Springer, Berlin (in press)

Schauer N, Fernie AR (2006) Plant metabolomics: towards biological function and mechanism. *Trends in Plant Science*, **11**, 508–516.

Scheible WR, GonzalezFontes A, Morcuende R *et al*. (1997) Tobacco mutants with a decreased number of functional nia genes compensate by modifying the diurnal regulation of transcription, post-translational modification and turnover of nitrate reductase. *Planta*, **203**, 304–319.

Scheiner SM, Gurevitch J (2001) *Design and Analysis of Ecological Experiments*, 2nd edn. Oxford University Press, Oxford.

Schlueter JA, Lin JY, Schlueter SD *et al*. (2007) Gene duplication and paleopolyploidy in soybean and the implications for whole genome sequencing. *BMC Genomics*, **8**.

Seki M, Narusaka M, Ishida J *et al*. (2002) Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant Journal*, **31**, 279–292.

Sharma N, Cram D, Huebert T, Zhou N, Parkin IAP (2007) Exploiting the wild crucifer *Thlaspi arvense* to identify conserved and novel genes expressed during a plant's response to cold stress. *Plant Molecular Biology*, **63**, 171–184.

Shiu SH, Borevitz JO (2008) The next generation of microarray research: applications in evolutionary and ecological genomics. *Heredity*, **100**, 141–149.

Stitt M, Gibon Y, Lunn JE, Piques M (2007) Multilevel genomics analysis of carbon signalling during low carbon availability: coordinating the supply and utilisation of carbon in a fluctuating environment. *Functional Plant Biology*, **34**, 526–549.

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 9440–9445.

Straalen NM van, Roelofs D (2006) *An Introduction to Ecological Genomics*. Oxford University Press, pp. 299.

Sweetlove LJ, Fernie AR (2005) Regulation of metabolic networks: understanding metabolic complexity in the systems biology era. *New Phytologist*, **168**, 9–23.

Talame V, Ozturk NZ, Bohnert HJ, Tuberosa R (2007) Barley transcript profiles under dehydration shock and drought stress treatments: a comparative analysis. *Journal of Experimental Botany*, **58**, 229–240.

Taylor G, Street NR, Tricker PJ *et al*. (2005) The transcriptome of *Populus* in elevated CO_2. *New Phytologist*, **167**, 143–154.

Thimm O, Blasing O, Gibon Y *et al*. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant Journal*, **37**, 914–939.

Tosti N, Pasqualini S, Borgogni A, Ederli L, Falistocco E, Crispi S, Paolocci F (2006) Gene expression profiles of O_3 treated Arabidopsis plants. *Plant, Cell and Environment*, **29**, 1686–1702.

Travers SE, Smith MD, Bai JF *et al*. (2007) Ecological genomics: making the leap from model systems in the lab to native populations in the field. *Frontiers in Ecology and the Environment*, **5**, 19–24.

Ungerer MC, Johnson LC, Herman MA (2008) Ecological genomics: understanding gene and genome function in the natural environment. *Heredity*, **100**, 178–183.

Urbanczyk-Wochniak E, Usadel B, Thimm O *et al*. (2006) Conversion of MapMan to allow the analysis of transcript data from Solanaceous species: effects of genetic and environmental alterations in energy metabolism in the leaf. *Plant Molecular Biology*, **60**, 773–792.

Vodkin LO, Khanna A, Shealy R *et al*. (2004) Microarrays for global expression constructed with a low redundancy set of 27,500 sequenced cDNAs representing an array of developmental stages and physiological conditions of the soybean plant. *BMC Genomics*, **5**.

Wang H, Miyazaki S, Kawai K, Deyholos M, Galbraith DW, Bohnert HJ (2003) Temporal progression of gene expression responses to salt shock in maize roots. *Plant Molecular Biology*, **52**, 873–891.

Watkinson JI, Sioson AA, Vasquez-Robinet C *et al*. (2003) Photosynthetic acclimation is reflected in specific patterns of gene expression in drought-stressed loblolly pine. *Plant Physiology*, **133**, 1702–1716.

Weston DJ, Gunter LE, Rogers A, Wullschleger SD (2008) Connecting genes, coexpression modules, and molecular signatures to environmental stress phenotypes in plants. *BMC Systems Biology*, **2**, 16.

Windsor AJ, Mitchell-Olds T (2006) Comparative genomics as a tool for gene discovery. *Current Opinion in Biotechnology*, **17**, 161–167.

Wullschleger SD, Leakey ADB, St Clair SB (2007) Functional genomics and ecology – a tale of two scales. *New Phytologist*, **176**, 735–739.

Wullschleger SD, Tschaplinski TJ, Norby RJ (2002) Plant water relations at elevated CO_2 – implications for water-limited environments. *Plant, Cell and Environment*, **25**, 319–331.

Zavala JA, Casteel CL, DeLucia EH, Berenbaum MR (2008) Anthropogenic increase in carbon dioxide compromises plant defense against invasive insects. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 5129–5133.

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, 17.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Currently available resources for microarray analysis of plant gene expression.

Please note: Wiley–Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Bioenerg. Res.
DOI 10.1007/s12155-009-9060-z

Bioinformatics-Based Identification of Candidate Genes from QTLs Associated with Cell Wall Traits in *Populus*

Priya Ranjan · Tongming Yin · Xinye Zhang · Udaya C. Kalluri · Xiaohan Yang · Sara Jawdy · Gerald A. Tuskan

Abstract Quantitative trait locus (QTL) studies are an integral part of plant research and are used to characterize the genetic basis of phenotypic variation observed in structured populations and inform marker-assisted breeding efforts. These QTL intervals can span large physical regions on a chromosome comprising hundreds of genes, thereby hampering candidate gene identification. Genome history, evolution, and expression evidence can be used to narrow the genes in the interval to a smaller list that is manageable for detailed downstream functional genomics characterization. Our primary motivation for the present study was to address the need for a research methodology that identifies candidate genes within a broad QTL interval. Here we present a bioinformatics-based approach for subdividing candidate genes within QTL intervals into alternate groups of high probability candidates. Application of this approach in the context of studying cell wall traits, specifically lignin content and S/G ratios of stem and root in *Populus* plants, resulted in manageable sets of genes of both known and putative cell wall biosynthetic function. These results provide a roadmap for future experimental work leading to identification of new genes controlling cell wall recalcitrance and, ultimately, in the utility of plant biomass as an energy feedstock.

Electronic supplementary material The online version of this article (doi:10.1007/s12155-009-9060-z) contains supplementary material, which is available to authorized users.

P. Ranjan (✉) · T. Yin · X. Zhang · U. C. Kalluri · X. Yang · S. Jawdy · G. A. Tuskan
Environmental Sciences Division,
Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA
e-mail: ranjanp@ornl.gov

P. Ranjan · T. Yin · X. Zhang · U. C. Kalluri · X. Yang · S. Jawdy · G. A. Tuskan
The Bioenergy Science Center, Oak Ridge National Laboratory,
Oak Ridge, TN 37831, USA

Keywords *Populus* · Whole-genome duplication · Quantitative trait loci · Wood chemistry · Syringyl lignin · Guaiacyl lignin · Biofuels

Abbreviations

QTL	Quantitative trait loci
S/G	Syringyl and guaiacyl ratio
RL	Root lignin content
RSG	Root S/G ratio
SL	Stem lignin content
SSG	Stem S/G ratio
SSR	Simple sequence repeats

Introduction

Plant biomass has recently been promoted as a source of renewable feedstock for the conversion to liquid transportation fuels [13, 15, 22, 29]. Plant cell walls can be biochemically or thermochemically deconstructed into the primary subcomponents (cellulose, hemicellulose, and lignin) necessary for this conversion [14, 19, 20, 26]. The carbohydrate fractions are used as feedstocks for sugar and ultimately ethanol production, and lignins are typically separated and used in combustion processes to fuel the reactions. The resistance of lignin, an amorphic polymer, to separate from the carbohydrate fractions during the deconstruction phase has made lignin a target for overcoming recalcitrance [12].

Lignin, a complex polyphenolic polymer, is one of the most abundant polymers on earth. Lignin content of the cell wall influences the cell rigidity, drought tolerance, and insect and disease resistance [7]. The biochemical pathway for lignin biosynthesis is fairly well characterized and involves approximately 12–15 enzyme-regulated steps controlling the conversion of single aldehyde to syringyl

Published online: 06 November 2009

and guaiacyl precursors [5, 37]. Lignin content varies across the tissue types and organs of a plant with developmental age and environmental interactions [32]. These responses are genetically controlled and heritability for lignin is moderately high [20]. The rate limiting/critical steps in lignin formation are not yet fully determined though several studies have used reverse genetic approaches and expression analysis to modify and/or characterize lignin composition in transgenic plant materials [9, 10, 23, 24].

Lignin and other cell wall traits display a pattern of continuous phenotypic distribution rather than discrete, Mendelian distribution. Such traits are typically polygenic in nature and are influenced by the environment in which they occur. Genetic mapping can be used to compare the inheritance pattern of a trait and establish the chromosomal regions associated with such phenotypes. These chromosomal intervals may encompass one or more genes responsible for the trait and are known as quantitative trait loci (QTLs).

After the identification of QTL intervals, filtering the list of genes down to a subset of likely candidates is a difficult task. The length of the QTL intervals may be in mega base pairs (Mbp) and include hundreds of genes. One approach to reducing the number of candidate genes is to conduct further experiments using larger numbers of segregating progeny to reduce the QTL interval. Then, classical methods such as positional cloning [25, 27] and insertional mutagenesis [3, 30] can be used to identify influential genes. A complementary approach would be to use the bioinformatics tools and genome information to assign genes in the QTL interval to bins of higher probability than other candidates. This gives a smaller number of candidate genes that can be verified using transgenesis.

The recent availability of several draft and fully sequenced plant genomes have shed light on the evolutionary history of genome structure, and the role whole-genome duplication events have played in determining genome structure and gene family evolution. It is becoming apparent that nearly all plant genomes have experienced at least one whole-genome duplication event [18, 39]. These events have influenced gene family evolution and created opportunities for paralogous genes to experience neo-functionalization and/or sub-functionalization within all gene families [8, 28, 36].

The *Populus* genome contains three whole-genome duplication events [35]. The most recent, the Salicoid duplication, is found only in members of the Salicoid family and is present in approximately 8,000 paralogous gene pairs. The second duplication event, shared by *Populus* and *Arabidopsis*, is found in 3,500 paralogous gene pairs in *Populus*. In addition, the molecular clock in *Populus* is ticking a rate that is six times slower than in *Arabidopsis*, creating a duplicated molecular preservation of the ancestral genome within the extant *Populus* genome [35]. Together these genomic features can complicate genome assembly, annotation as well as map-based cloning of individual gene(s) responsible for specific phenotypes.

We use a combination of traditional QTL mapping, comparative intragenomic analysis, estimates of gene divergence, and differential expression evidence to identify regions of the *Populus* genome that contain genes controlling lignin content in shoots and roots and demonstrate that this combinational approach can be used to filter a candidate gene list to a substantially smaller subset of genes within a fixed confidence interval.

Materials and Methods

Description of QTLs

An F2 inbred interspecific hybrid poplar family was used to create a comprehensive genetic map containing 848 markers based on 293 segregating progeny as described by Yin et al. [40]. The overall observed genetic length was 1,927.6 cM. Phenotypic data was collected for all progeny using pyMBMS to obtain estimates of root lignin, root S/G ratio, stem lignin content, and stem S/G ratio [11, 32, 33]. MapQTL 5.0 was used to detect the underlying QTLs [38]. The establishment of genetic map, phenotying of lignin content, and S/G ratio of mapping individuals have been described by Yin et al. (in review).

Assigning Physical Position to the SSR Markers

Populus genome sequence, gene models, and functional categories for genes were downloaded from the JGI Populus Genome portal (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). The SSR primer resource, available at the International Poplar Genome Consortium website (http://www.ornl.gov/sci/ipgc/ssr_resource.htm; [34, 41]), was used to obtain the sequence information for the SSR primers and predicted SSR length. The physical position of each interval in the *Populus* genome was assigned based on BLAST results of the SSR primer nucleotide sequence against the genomic sequence. Additionally, the number of base pairs between the start of the left primer and end of right primer (according to BLAST result) had to be equivalent to the predicted length of the SSR marker. Perl script was used to automate this process. In total, 210 markers were successfully assigned physical position in the genome.

Assigning Physical Position to the QTLs

Assigning QTLs to physical positions in the genome was challenging as linear relationship between the physical and

genetic maps vary by position within the genome due to non-homogeneous distribution of chiasma across the genome. Thus, SSR markers flanking the QTLs were initially identified based on genetic positions. The relationship between genetic and physical distance for each QTL was then obtained by the ratio of physical distance between the markers and the genetic distance between them. This relationship was used to obtain the physical coordinates of the QTL in the genome by subtracting the difference between the ends of the QTL and the flanking marker.

Identification of Duplicated Genes Corresponding to the QTL Interval

Around 8,000 pairs of paralogous genes of similar age (excluding tandem duplications) were identified in the *Populus* genome. All genes in each QTL interval were identified based on the position of genes on each chromosome/linkage group (LG). The duplicated interval and corresponding duplicated gene information were then identified. Next, percent identity between a gene and its paralog was calculated using BLAST to align each pair. Finally, the best match *Arabidopsis* genes were identified by reciprocal blasting BLAST of the *Arabidopsis* gene set (TAIR Version 9) and *Populus* gene set to identify the top pair in each case.

Data Mining of Microarray Expression Profiles of Genes and Duplicated Genes

Populus balsamifera Affymetrix microarray datasets containing developmental tissue series (GSE13990 series) in GEO database at NCBI were used to examine the transcriptome level attributes of roots and differentiating xylem. We used this dataset to identify differences in gene expression between root and stem. The 50,848 probe sets with genome match correspond to 40,236 unique JGI *Populus trichocarpa* gene models. Cross-hybridizing and redundant probes for gene models as well as probes for alternatively spliced version of genes were eliminated in the analysis.

Identification of Differentially Expressed Genes

We used the RankProd package [16] to analyze the expression array data to identify differentially expressed genes. RankProd utilizes a rank product non-parametric method [6] to identify up- or down-regulated genes under differential conditions, e.g., two treatments, two tissue types, etc. The false discovery rate (FDR) value obtained was based on 10,000 random permutations [16]. The genes that had FDR values less than or equal to 0.10 were considered as differentially expressed.

Results

QTL Intervals

The QTL intervals for lignin and S/G ratio are located on seven linkage groups in the *Populus* genome (Fig. 1). The genetic position of each QTL interval is shown in Table 1. The QTL intervals for root lignin content were observed on LG II, LG VI, LG X, and LG XIV; for stem lignin content on LG II, LG VI, and LG X; for root S/G ratio on LG III, LG VI, LG X, and LG XIII; and for stem S/G ratio on LG II, LG III, LG VI, LG XIV, and LG VIII. These QTL intervals generally do not overlap, except on LG VI where QTL intervals for root and stem lignin content and root and stem S/G ratio co-localize and on LG X where QTL intervals for root and stem lignin content co-localize (Fig.1). The length of the QTL intervals ranged from 0.4 to 11 Mbp (Table 2); the majority of the QTL intervals were less than 2 Mbp in length. Correspondingly, the number of genes in the QTL intervals ranged from 44 to 1,501. The total number of genes in all the intervals was 4,530 (Supplementary Table 1). As there were two regions of overlapping QTLs, some genes were common to those QTLs, and the number of unique genes from all the QTL intervals was 3,788.

Duplication in *Populus* Genome and Duplicated Regions in QTL Interval

The *Populus* genome has undergone a recent genome-wide duplication event that has resulted in a conserved linear order of most of the genes within the duplicated chromosomal segments. QTLs on LG II have duplicated intervals on LG V; QTLs on LG III have duplicated intervals on LG V and scaffold_29; QTLs on LG VI have duplicated intervals on LG XVI and LG XVIII; QTLs on LG X have duplicated intervals on LG VIII (Fig. 2). Some intervals had higher numbers of genes conserved in the duplicated region as compared to the others. Across all intervals, on average, more than 53% of genes had retained a paralog in the duplicated interval and ranged from 25% to 80% (Table 3).

Comparison of Expression of Genes that Lie in the QTL Interval and Their Paralogs

Based on microarray evidence, 13 out of 19 QTL intervals were tissue specific, i.e., the QTL intervals corresponding to lignin content or S/G ratio were unique for either root or stem. Four of these QTL intervals, root lignin content (RL-1), stem lignin content (SL-3), root S/G ratio (RSG-1), and stem S/G ratio (SSG-5), were selected for a detailed analysis

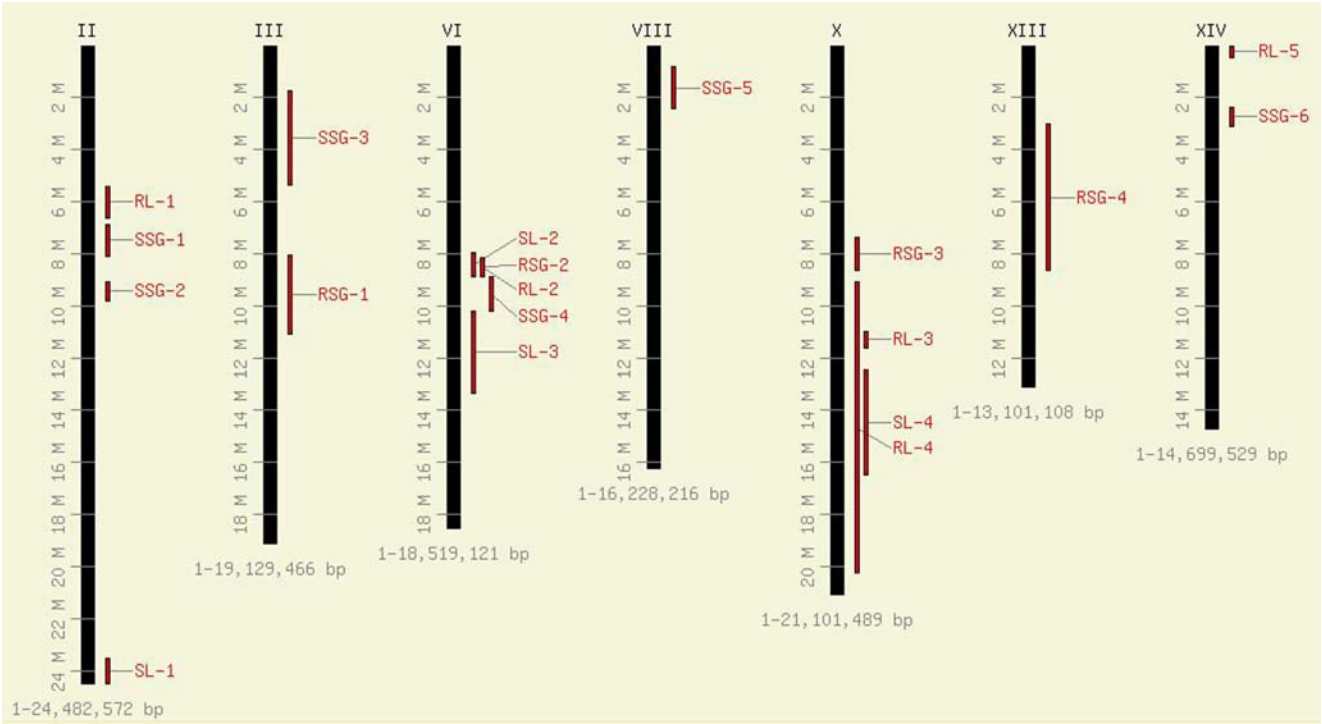


Fig. 1 Nineteen QTL interval intervals distributed on seven linkage groups (LG II, LG III, LG VI, LG VIII, LG X, LG XIII, and LG XIV) in *Populus*. *RL* root lignin QTL intervals, *RSG* root S/G ratio, *SL* stem lignin content, *SSG* stem S/G ratio

Table 1 Location of QTL intervals based on genetic map

QTL short name	QTL full name	LOD	% explanation	Linkage group	Peak position (cM)	1 LOD confidence (left, cM)	1 LOD confidence (right, cM)
RL-1	Root lignin content-1	2.77	18.2	II	62.13	55.13	67.785
RL-2	Root lignin content-2	5.06	8.9	VI	66.68	61.067	71.158
RL-3	Root lignin content-3	2.91	5.2	X	67.72	63.867	70.043
RL-4	Root lignin content-4	3.81	11.1	X	130.17	90.634	130.174
RL-5	Root lignin content-5	2.71	8.4	XIV	5.66	0	9.353
RSG-1	Root S/G ratio-1	5.19	14.9	III	57.82	48.823	65.823
RSG-2	Root S/G ratio-2	2.82	5.4	VI	71.16	62.624	71.158
RSG-3	Root S/G ratio-3	3.19	6	X	48.51	47.244	52.508
RSG-4	Root S/G ratio-4	3.21	6.6	XIII	50.39	39.936	60.298
SL-1	Stem lignin content-1	2.85	6.9	II	175.44	167.589	175.44
SL-2	Stem lignin content-2	4.29	7.6	VI	71.16	60.067	71.108
SL-3	Stem lignin content-3	4.21	7.3	VI	82.47	78.8	105.578
SL-4	Stem lignin content-4	2.71	4.6	X	80.15	75.202	105.634
SSG-1	Stem S/G ratio-1	5.02	14.7	II	76.8	71.759	82.132
SSG-2	Stem S/G ratio-2	5.51	12.5	II	90.76	88.759	95.09
SSG-3	Stem S/G ratio-3	3.22	6.4	III	23.62	10.724	32.197
SSG-4	Stem S/G ratio-4	2.58	5.2	VI	80.74	71.158	85.787
SSG-5	Stem S/G ratio-5	2.69	5.2	VIII	25.74	10.097	27.222
SSG-6	Stem S/G ratio-6	3.31	6.7	XIV	50.37	40.764	57.363

Table 2 Details of QTL in terms of physical distances

QTL short name	QTL full name	Linkage group	Final left flank physical position (bp)	Final right flank physical position (bp)	Length of interval (Mbp)
RL-1	Root lignin content-1	II	5,493,283	6,716,085	1.22
RL-2	Root lignin content-2	VI	8,063,866	8,911,035	0.85
RL-3	Root lignin content-3	X	10,873,565	11,681,744	0.81
RL-4	Root lignin content-4	X	9,141,424	20,293,167	11.15
RL-5	Root lignin content-5	XIV	1	485,878	0.49
RSG-1	Root S/G ratio-1	III	8,099,415	11,113,515	3.01
RSG-2	Root S/G ratio-2	VI	8,194,580	8,911,035	0.72
RSG-3	Root S/G ratio-3	X	7,291,775	8,702,806	1.41
RSG-4	Root S/G ratio-4	XIII	3,005,790	8,702,467	5.7
SL-1	Stem lignin content-1	II	23,535,453	24,482,567	0.95
SL-2	Stem lignin content-2	VI	7,979,913	8,906,838	0.93
SL-3	Stem lignin content-3	VI	10,229,422	13,356,289	3.13
SL-4	Stem lignin content-4	X	12,428,930	16,367,409	3.94
SSG-1	Stem S/G ratio-1	II	6,929,755	7,993,901	1.06
SSG-2	Stem S/G ratio-2	II	9,110,752	9,878,601	0.77
SSG-3	Stem S/G ratio-3	III	1,810,361	5,435,304	3.62
SSG-4	Stem S/G ratio-4	VI	8,911,035	10,139,184	1.23
SSG-5	Stem S/G ratio-5	VIII	875,268	2,359,766	1.48
SSG-6	Stem S/G ratio-6	XIV	2,282,351	3,052,827	0.77

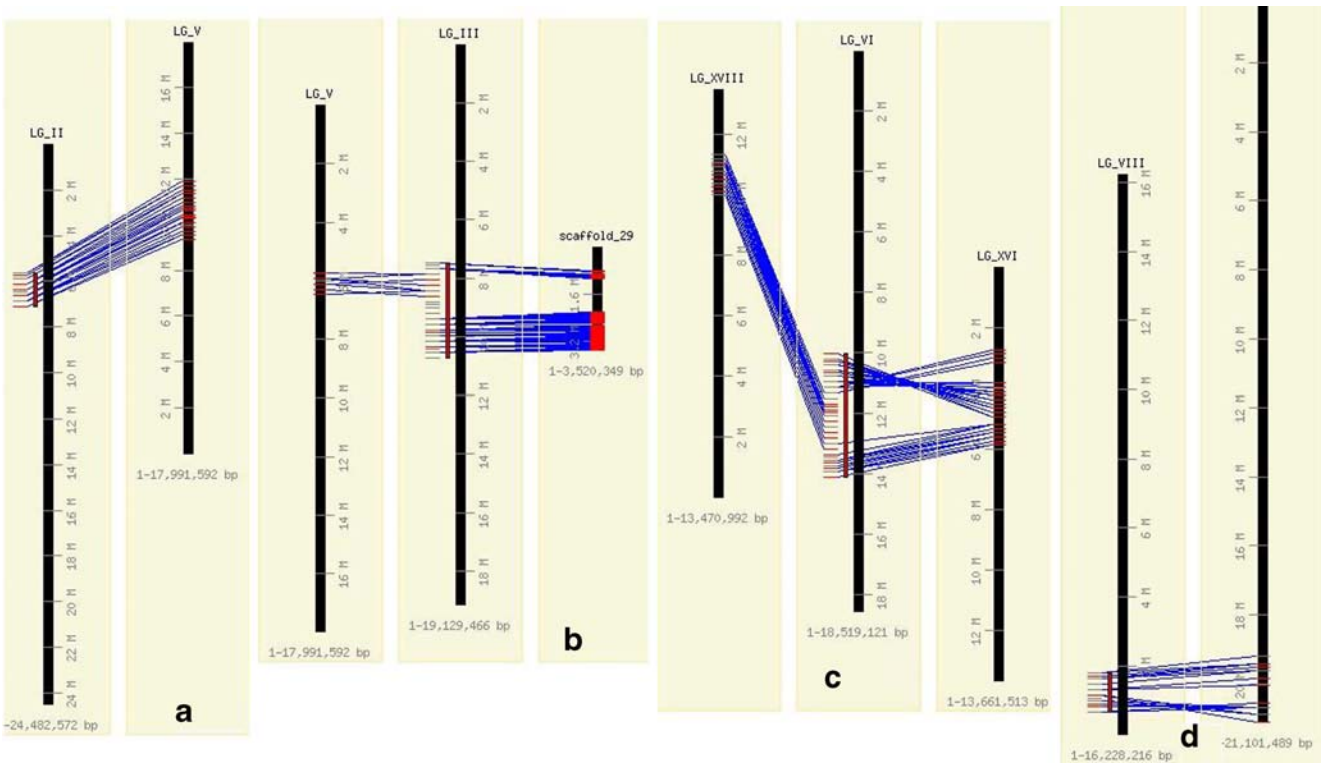


Fig. 2 QTL interval and display of duplication in genes in the interval for **a** root lignin content-1, **b** root S/G ratio-1, **c** stem lignin content-3, and **d** stem S/G ratio. Each *blue line* represents a gene and its paralog in the duplicated region

Table 3 Details of duplication and differences in expression

QTL	Total number of genes	Has paralogs	% id>90	% id>90 and expression in tissue of QTL	% id<90	% id<90 and expression in tissue of QTL	Missing paralogs and expression in tissue of QTL
Number of genes in each category							
RSG-3	118	51	16	2	35	5	13
SL-3	247	171	65	3	106	6	8
SL-2	61	14	7		7		3
SSG-3	222	136	50	3	86	3	6
RSG-4	454	209	80	6	129	14	33
SL-1	88	25	7		18		7
SSG-6	115	69	26	3	43	4	2
RSG-1	278	170	63	1	107	4	8
RL-3	98	71	30	2	41	5	5
SSG-5	226	155	67	7	88	10	6
SSG-4	66	29	10		19		3
SL-4	548	332	119	11	213	14	16
RL-4	1,501	981	391	33	590	62	74
RL-2	52	14	7		7		12
RL-5	54	27	10	2	17	3	4
RL-1	138	94	39	4	55	5	2
SSG-1	123	51	20	3	31		8
SSG-2	97	77	24	2	53	2	
RSG-2	44	14	7		7		11

because they occurred in paralogous regions and contained differential tissue data from microarray experiments.

In order to use the above data to filter the candidate gene list within each of the selected QTL intervals, three alternative approaches were used to integrate duplication information and differential expression of paralogs (Fig. 3). First, filtering was based on non-duplicated genes within a QTL and those which have higher expression in tissue related to the QTL. Second, differential microarray results were used to identify genes within the interval with expression evidence in the identified tissue whose paralogous genes did not display expression in the corresponding tissue. For example, we identified genes, present in QTL interval for stem lignin content, that show higher expression in xylem relative to root as compared to gene expression of paralogs. In addition, the genes within the QTL interval that have a predicted role in cell wall biosynthesis (e.g., PAL, 4CL, etc.) were promoted to the candidate gene list.

Genes in the QTL Interval

Root Lignin Content The total number of genes in the RL-1 interval on LG II was 138. Out of these, 94 had paralogs and 44 genes did not retain paralogs in the duplicated interval (Table 3). Two of these non-duplicated genes with higher expression in root were calmodulin (eugene3.00020820) and a signal transduction response regulator gene (gw1.II.42.1; Table 4). Five duplicated genes with <90% similarity and higher expression in root were replication factor (estExt_fgenesh4_pg.C_LG_II0728), an auxin response factor (fgenesh4_pg.C_LG_II000830), a glycosyl hydrolase hydrolyzing *o*-glycosyl compound (fgenesh4_pg.C_LG_II000867), a zinc finger transcription factor (grail3.0003072701), and an exoribonuclease (gw1.II.1849.1; Table 5). Four duplicated genes with >90% similarity and higher expression in root were sulfate transporter (eugene3.00020855), alcohol dehydrogenase (fgenesh4_pg.C_LG_II000742), a NAC domain protein (grail3.0003068301), and a nodulin-like protein (gw1.II.1386.1; Table 6). In total, the number of genes within the interval was filtered down from 138 equally likely candidates to 11 with supportive duplication and/or expression evidence.

Stem Lignin Content The total number of genes in the SL-3 interval on LG VI was 247. Out of these, 171 had paralogs and 76 genes did not retain paralogs in the duplicated interval (Table 3). Eight of these non-duplicated genes with higher expression in xylem were hypothetical protein (estExt_fgenesh4_pm.C_LG_VI0468), proteins with no known function and unique to Populus (eugene3.00061181, fgenesh4_pg.C_LG_VI001243), a nucleic acid binding

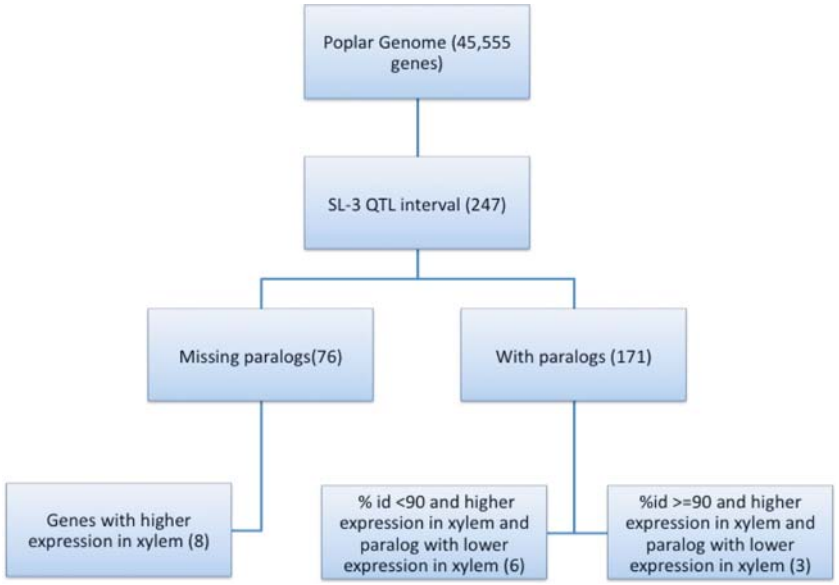


Fig. 3 Process of filtering genes. *SL-3* stem lignin QTL interval

(eugene3.00061373), protein associated with CCR4 transcription complex (grail3.0030015901), a plastocyanin-like domain-containing protein (gw1.VI.2580.1), a ribosomal protein (gw1.VI.2649.1), and a peptidyl-prolyl *cis–trans* isomerase, cyclophilin type protein (gw1.VI.847.1; Table 4). Six duplicated genes with <90% similarity and higher expression in xylem were kinesin protein involved in microtubule-based movement (estExt_fgenesh4_pm.

Table 4 List of genes that did not have a paralogous gene model in duplicated region and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	eugene3.00020820	NA	Calmodulin	≤0.001
RL-1	gw1.II.42.1	NA	Two-component response regulator	≤0.1
RSG-1	estExt_fgenesh4_pg.C_LG_III0677	NA	Oligopeptide transporter	≤0.01
RSG-1	estExt_Genewise1_v1.C_LG_III1770	NA	RNA helicase	≤0.1
RSG-1	eugene3.00030584	NA	Peroxidase	≤0.1
RSG-1	eugene3.00030600	NA	Formin-like protein	≤0.1
RSG-1	fgenesh4_pg.C_LG_III000886	NA	Expressed protein	≤0.1
RSG-1	grail3.0018015801	NA	ATP-dependent RNA helicase	≤0.01
RSG-1	gw1.III.1044.1	NA	DNA J protein	≤0.1
RSG-1	gw1.III.2608.1	NA	Transporter-like protein	≤0.001
SL-3	estExt_fgenesh4_pm.C_LG_VI0468	NA	Hypothetical protein	≤0.01
SL-3	eugene3.00061181	NA	No hits	≤0.1
SL-3	eugene3.00061373	NA	Nucleic acid-binding	≤0.1
SL-3	fgenesh4_pg.C_LG_VI001243	NA	No hits	≤0.1
SL-3	grail3.0030015901	NA	Associated with CCR4 transcription complex	≤0.01
SL-3	gw1.VI.2580.1	NA	Plastocyanin-like domain-containing protein	≤0.01
SL-3	gw1.VI.2649.1	NA	Ribosomal protein	≤0.1
SL-3	gw1.VI.847.1	NA	Peptidyl-prolyl *cis–trans* isomerase	≤0.01
SSG-5	eugene3.00080178	NA	Unknown protein	≤0.01
SSG-5	eugene3.00080195	NA	No hits	≤0.1
SSG-5	eugene3.00080203	NA	Ankyrin repeat family, transmembrane transport	≤0.001
SSG-5	eugene3.00080273	NA	No hits	≤0.001
SSG-5	fgenesh4_pg.C_LG_VIII000212	NA	Cytochrome P450	≤0.001
SSG-5	gw1.VIII.950.1	NA	Vacuolar protein	≤0.1

Table 5 List of genes that have a paralogous gene model; and the % identity is less than 90%, and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	estExt_fgenesh4_pg.C_LG_II0728	59	Replication factor	≤0.1
RL-1	fgenesh4_pg.C_LG_II000830	84	Auxin response factor	≤0.1
RL-1	fgenesh4_pg.C_LG_II000867	75	Glycosyl hydrolase (xylosidase)	≤0.1
RL-1	grail3.0003072701	69	Zinc finger transcription factor	≤0.01
RL-1	gw1.II.1849.1	89	Exoribonuclease	≤0.1
RSG-1	fgenesh4_pg.C_LG_III000669	87	calcium transporting ATPase	≤0.01
RSG-1	grail3.0018007101	67	Glucosyl transferase	≤0.001
RSG-1	grail3.0018018701	80	Tetratricopeptide-containing protein	≤0.01
RSG-1	gw1.III.1613.1	78	Proline-rich protein	≤0.01
SL-3	estExt_fgenesh4_pm.C_LG_VI0481	88	Kinesin (microtubule-based movement)	≤0.01
SL-3	estExt_fgenesh4_pm.C_LG_VI0500	51	Hypothetical protein	≤0.001
SL-3	estExt_Genewise1_v1.C_LG_VI2154	44	Senescence-associated protein	≤0.1
SL-3	grail3.0030003201	73	Unknown protein	≤0.01
SL-3	grail3.0030006902	81	Unknown protein	≤0.001
SL-3	gw1.VI.781.1	85	Nodulin-like protein	≤0.001
SSG-5	eugene3.00080177	75	Germin-like protein	≤0.1
SSG-5	eugene3.00080251	68	Acyl-CoA-binding family protein	≤0.1
SSG-5	eugene3.00080330	81	GATA zinc fringe protein	≤0.001
SSG-5	fgenesh4_pg.C_LG_VIII000250	67	Unknown protein	≤0.01
SSG-5	fgenesh4_pg.C_LG_VIII000264	77	Unknown protein	≤0.1
SSG-5	fgenesh4_pm.C_LG_VIII000069	81	Unknown protein	≤0.001
SSG-5	fgenesh4_pm.C_LG_VIII000111	81	Pumilio-family RNA-binding protein	≤0.01
SSG-5	grail3.0049006403	88	Chorismate synthase	≤0.1
SSG-5	gw1.VIII.1321.1	83	Pectate lyase-like protein	≤0.001
SSG-5	gw1.VIII.1497.1	85	Acyl-CoA-binding family protein	≤0.1

Table 6 List of genes that have a paralogous gene model; and the % identity is greater than 90%, and that showed higher expression in the tissue related to the QTL

QTL	Gene	% id with paralog	Function	FDR
RL-1	eugene3.00020855	90	Sulfate transporter	≤0.01
RL-1	fgenesh4_pg.C_LG_II000742	94	Alcohol dehydrogenase	≤0.001
RL-1	grail3.0003068301	91	NAC domain protein	≤0.001
RL-1	gw1.II.1386.1	93	Nodulin-like protein	≤0.001
RSG-1	fgenesh4_pg.C_LG_III000900	93	WRKY family transcription factor	≤0.001
SL-3	estExt_fgenesh4_pg.C_LG_VI1102	90	Endonuclease/exonuclease hydrolase activity	≤0.01
SL-3	eugene3.00061209	96	Expressed protein	≤0.01
SL-3	eugene3.00061339	93	UDP-D-Glucuronate 4-epimerase, nucleotide sugar interconversion pathway	≤0.1
SSG-5	estExt_fgenesh4_pg.C_LG_VIII0179	90	expressed protein	≤0.01
SSG-5	estExt_fgenesh4_pm.C_LG_VIII0087	91	Glucosyl transferase, cellulose synthase-like	≤0.001
SSG-5	eugene3.00080299	92	Vacuolar protein, vacuolar biogenesis	≤0.1
SSG-5	eugene3.00080329	90	Pleckstrin homology (PH) domain-containing protein	≤0.01
SSG-5	grail3.0049010802	97	CCAAT-box-binding transcription factor	≤0.1
SSG-5	gw1.VIII.1083.1	91	Photoreceptor-interacting protein	≤0.001
SSG-5	gw1.VIII.2327.1	93	Exostosin family, GT 47	≤0.01

C_LG_VI0481), a hypothetical protein (estExt_fgenesh4_pm.C_LG_VI0500), a senescence associated protein (estExt_Genewise1_v1.C_LG_VI2154) and unknown proteins (grail3.0030003201, grail3.0030006902), and a nodulin-like protein (gw1.VI.781.1; Table 5). Three duplicated genes with >90% similarity and higher expression in xylem were protein with hydrolase activity (estExt_fgenesh4_pg.C_LG_VI1102), an expressed protein with no known function (eugene3.00061209), and a UDP-D-glucuronate 4-epimerase involved in nucleotide sugar interconversion pathway (eugene3.00061339; Table 6). In total, the number of genes within the interval was filtered down from 247 equally likely candidates to 17 with supportive duplication and/or expression evidence.

Root S/G ratio The total number of genes in the RSG-1 interval on LG III was 278. Out of these, 170 had paralogs and 108 genes did not retain paralogs in the duplicated interval (Table 3). Eight of these non-duplicated genes with higher expression in root were oligopeptide transporter (estExt_fgenesh4_pg.C_LG_III0677), a RNA helicase protein (estExt_Genewise1_v1.C_LG_III1770, grail3.0018015801), a peroxidase (eugene3.00030584), a formin-like protein (eugene3.00030600), an expressed protein (fgenesh4_pg.C_LG_III000886), a DNA J protein (gw1.III.1044.1), and a transporter-like protein (gw1.III.2608.1; Table 4). Four duplicated genes with <90% similarity and higher expression in root were calcium transporting ATPase (fgenesh4_pg.C_LG_III000669), a glucosyl transferase (grail3.0018007101), a tetratricopeptide-containing protein (grail3.0018018701), and a proline-rich protein (gw1.III.1613.1; Table 5). One duplicated gene with >90% similarity and higher expression in root was WRKY family transcription factor (fgenesh4_pg.C_LG_III000900; Table 6). In total, the number of genes within the interval was filtered down from 278 equally likely candidates to 13 with supportive duplication and/or expression evidence.

Stem S/G ratio The total number of genes in the SSG-5 interval on LG VIII was 226. Out of these, 155 had paralogs and 71 genes did not retain paralogs in the duplicated interval (Table 3). 2Six of these non-duplicated genes with higher expression in xylem were unknown protein (eugene3.00080178), a protein unique to *Populus* (eugene3.00080195, eugene3.00080273), an ankyrin repeat family involved in transmembrane transport (eugene3.00080203), a cytochrome P450 protein (fgenesh4_pg.C_LG_VIII000212), and a vacuolar protein (gw1.VIII.950.1; Table 4). Ten duplicated genes with <90% similarity and higher expression in xylem were germin-like protein (eugene3.00080177), an acyl-CoA-binding family protein (eugene3.00080251), a GATA zinc finger protein (eugene3.00080330), unknown proteins (fgenesh4_pg.C_LG_VIII000250, fgenesh4_pg.C_LG_VIII000264, fgenesh4_pm.C_LG_VIII000069), a pumilio-family RNA-binding protein (fgenesh4_pm.C_LG_VIII000111), a chorismate synthase (grail3.0049006403), a pectate lyase-like protein (gw1.VIII.1321.1), and an acyl-CoA-binding family protein (gw1.VIII.1497.1; Table 5). Seven duplicated gene with >90% similarity and higher expression in xylem were expressed protein (estExt_fgenesh4_pg.C_LG_VIII0179), a glucosyl transferase also annotated as cellulose synthase-like (estExt_fgenesh4_pm.C_LG_VIII0087), a vacuolar protein (eugene3.00080299), pleckstrin homology domain-containing protein (eugene3.00080329), a CCAAT-box-binding transcription factor (grail3.0049010802), a photoreceptor-interacting protein (gw1.VIII.1083.1), and an exostosin family protein also annotated as glucoside transferase 47 (gw1.VIII.2327.1; Table 6). In total, the number of genes within the interval was filtered down from 226 equally likely candidates to 23 with supportive duplication and/or expression evidence.

Discussion

Whole-genome duplication events, followed by extensive genome reorganization, chromosomal rearrangements, and gene loss, have been widespread during the evolution of plants [31]. As a consequence of duplication, paralogs created in the genome may have one of several possible fates, including non-functionalization, neo-functionalization, and sub-functionalization [18]. The most acknowledged mode is non-functionalization where one of the copies loses function or is silenced, resulting in a pseudogene [1]. The process of neo-functionalization, where one ancestral copy retains its function and the other is free to accumulate mutations, results in acquisition of novel function. During the process of sub-functionalization the sister copies, i.e. paralogs, show different but overlapping functions [18]. Here, some duplicated genes show differential expression among organs within a single plant. In recent allopolyploidization in cotton some genes were silenced in one organ with respect to another. Similar outcomes were detected in artificial allopolyploidization [2]. In *Arabidopsis* there is evidence of sub-functionalization where clusters of duplicated genes show evidence of concerted divergence in their expression in an organ-specific expression [4].

The *Populus* genome has undergone multiple genome-wide duplications [35]. The Salicoid duplication currently contains around 8,000 pairs of genes that are syntenous across mega-base regions of the genome. As a result, almost every segment in the *Populus* genome has a parallel paralogous interval somewhere else in the genome. Yet,

QTL intervals for many stem and root lignin and S/G ratio phenotypes are present in only one position (Fig. 1). This suggests that different sets of genes to root and stem QTLs providing an opportunity to leverage the segmental duplication information. Along with gene expression of paralogous gene intervals, higher likelihood values can be assigned to genes or gene sets that are functionally related to the measured phenotype.

This expansive gene declaration results in each QTL interval having hundreds to thousands of genes. Our filtering approach led to a reduced set of genes, most not previously reported to play a direct role in monolignol biosynthesis. These filtered genes included regulatory proteins that may have roles in cell wall formation, vascular transport, and unknown function. For example, in the root lignin interval, a NAC domain transcription factor (grail3.0003068301) is present, and NAC domain transcription factors have been implicated as key regulator of secondary cell wall synthesis in *Arabidopsis* [43]. A signal transduction response regulator (gw1.II.42.1) was also identified in this interval and is an ideal candidate for further transgenic work as is kinesin (estExt_fgenesh4_pm.C_LG_VI0481), which is involved in the oriented deposition of cellulose microfibrils in *Arabidopsis* [42]. In the root S/G ratio interval two proteins seem very promising. One is glycosyl hydrolase (fgenesh4_pg.C_LG_II000867) and the other is UDP-D-glucuronate 4-epimerase (eugene3.00061339) involved in nucleotide sugar interconversion pathways. In the stem S/G content interval a cytochrome P450 (fgenesh4_pg.C_LG_ VIII000212) is a good candidate for further experimental work. Exostosin gene (gw1.VIII.2327.1) has also been shown to be more highly co-regulated with cellulose synthase genes in *Arabidopsis* [21].

The filtered gene set provides a feasible opportunity to determine gene function via functional genetics work. That is, based on the computational approach described above, a set of 15–20 candidate genes can be used in RNAi knockdown experiments, mutant complementation experiments, and in association genetics studies correlating single nucleotide polymorphisms frequency and measured phenotypes. An integrated approach that combines QTL mapping with fine-scale mapping using association mapping would require investigating SNPs associated with trait using SNP arrays [17]. Multiple SNPs need to be assayed per gene as linkage disequilibrium decays rapidly in *Populus*. The filtration strategy discussed in this paper can be used to select candidate genes to assay SNPs and uncover the underlying DNA polymorphism associated with lignin content and lignin S/G ratio.

The unique approach of filtering for genes based on duplication evidence and expression data of paralogs has its limitations. The assembly of the *Populus* genome is still in a draft state and has numerous captured gaps where the length of the missing fragment is known and non-captured gaps where the length of the missing fragment is not known. Moreover, the lack of microarray datasets for *Populus* compared to *Arabidopsis* is also a limiting factor. As more microarray datasets become available, more robust statistical analyses will be feasible. The design of Affymetrix microarray adds to the challenge. Due to the overlapping nature of the Affymetrix probe sets it is frequently difficult to distinguish paralogs. Due to the lack of the microarray datasets, we based our analysis on microarray datasets from *P. balsamifera* on Affymetrix chips, whereas the QTL intervals were obtained from *P. trichocarpa*. Future studies of this nature should use the expression data from individuals with extreme phenotypes in the population used to detect QTL.

Conclusions

This paper provides a computational approach for integrating QTLs with expression data and *Populus* genome duplication information to assign higher likelihood values to candidate genes with greater precision than other. The analytical approach was successful in identifying both genes of suspected cell wall biosynthetic function as well as genes of putative cell wall biosynthetic function. Genes of unknown or putative functions would most likely not have been examined without such an approach. In total, the list of genes in QTL intervals was reduced from hundreds or thousands of genes to 15–20 genes. These results provide a roadmap for future experimental work attempting to discover cell wall recalcitrance genes and the ultimate utility of plant biomass as an energy feedstock.

Acknowledgments The authors would like to thank Dr. Stan Wullschleger, David Weston, Lee Gunter, and Manojit Basu for their technical reviews of this paper. The present study was enabled by research funds through the BioEnergy Science Center, which is a US Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC, under contract DE-AC05-00OR22725 for the US Department of Energy

References

1. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141
2. Adams KL, Percifield R, Wendel JF (2004) Organ-specific silencing of duplicated genes in a newly synthesized cotton allotetraploid. Genetics 168:2217–2226

3. Bechtold N, Ellis J, Pelletier G (1993) In planta Agrobacterium-mediated gene transfer by infiltration of adult Arabidopsis thaliana plants. C R Acad Sci Ser III Sci Vie 316:1194–1199
4. Blanc G, Wolfe KH (2004) Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell 16:1679–1691
5. Boerjan W, Ralph J, Baucher M (2003) Lignin biosynthesis. Annu Rev Plant Biol 54:519–546
6. Breitling R, Armengaud P, Amtmann A, Herzyk P (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. FEBS Lett 573:83–92
7. Brodeur-Campbell SE, Vucetich JA, Richter DL, Waite TA, Rosemier JN, Tsai CJ (2006) Insect herbivory on low-lignin transgenic aspen. Environ Entomol 35:1696–1701
8. Byrne KP, Wolfe KH (2007) Consistent patterns of rate asymmetry and gene loss indicate widespread neofunctionalization of yeast genes after whole-genome duplication. Genetics 175:1341–1350
9. Chen F, Dixon RA (2007) Lignin modification improves fermentable sugar yields for biofuel production. Nat Biotechnol 25:759–761
10. Chiang VL (2006) Monolignol biosynthesis and genetic engineering of lignin in trees, a review. Environmental Chemistry Letters 4 (3):143–146
11. Davis MF, Tuskan GA, Payne P, Tschaplinski TJ, Meilan R (2006) Assessment of *Populus* wood chemistry following the introduction of a Bt toxin gene. Tree Physiology 26:557–564
12. Davison BH, Drescher SR, Tuskan GA, Davis MF, Nghiem NP (2006) Variation of S/G ratio and lignin content in a *Populus* family influences the release of xylose by dilute acid hydrolysis. Appl Biochem Biotechnol 130(1–3):427–435
13. Dinus RJ, Payne P, Sewell NM, Chiang VL, Tuskan GA (2001) Genetic modification of short rotation poplar wood: properties for ethanol fuel and fiber productions. Crit Rev Plant Sci 20:51–69
14. Farrokhi N, Burton RA, Brownfield L, Hrmova M, Wilson SM, Bacic A, Fincher GB (2006) Plant cell wall biosynthesis: genetic, biochemical and functional genomics approaches to the identification of key genes. Plant Biotechnol J 4:145–167
15. Himmel ME, Ding SY, Johnson DK, Adney WS, Nimlos MR, Brady JW et al (2007) Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science 315(5813):804–807
16. Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, Chory J (2006) RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics 22:2825–2827
17. Ingvarsson PK, Garcia V, Luquez V, Hall D, Jansson S (2008) Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 locus in European aspen (*Populus tremula*, Salicaceae). Genetics 178:2217–2226
18. Jaillon O, Aury JM, Wincker P (2009) "Changing by doubling", the impact of whole genome duplications in the evolution of eukaryotes. Comptes Rendus Biologies 332:241–253
19. Kalluri UC, Joshi CP (2004) Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. Planta 220:47–55
20. Leple JC, Dauwe R, Morreel K, Storme V, Lapierre C, Pollet B et al (2007) Downregulation of cinnamoyl-coenzyme A reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. Plant Cell 19:3669–3691
21. Persson S, Wei H, Milne J, Page GP, Somerville CR (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. Proc Natl Acad Sci USA 102:8633–8638
22. Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA et al (2006) The path forward for biofuels and biomaterials. Science 311(5760):484–489
23. Ralph J, Akiyama T, Kim H, Lu FC, Schatz PF, Marita JM et al (2006) Effects of coumarate 3-hydroxylase down-regulation on lignin structure. J Biol Chem 281:8843–8853
24. Ranjan P, Kao YY, Jiang H, Joshi CP, Harding SA, Tsai CJ (2004) Suppression subtractive hybridization-mediated transcriptome analysis from multiple tissues of aspen (*Populus tremuloides*) altered in phenylpropanoid metabolism. Planta 219:694–704
25. Rikke BA, Johnson TE (1998) Towards the cloning of genes underlying murine QTLs. Mamm Genome 9:963–968
26. Roberts AW, Bushoven JT (2007) The cellulose synthase (CESA) gene superfamily of the moss *Physcomitrella patens*. Plant Mol Biol 63:207–219
27. Ron M, Weller JI (2007) From QTL to QTN identification in livestock—winning by points rather than knock-out: a review. Anim Genet 38:429–439
28. Roth C, Rastogi S, Arvestad L, Dittmar K, Light S, Ekman D et al (2007) Evolution after gene duplication: models, mechanisms, sequences, systems, and organisms. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution 308B:58–73
29. Rubin EM (2008) Genomics of cellulosic biofuels. Nature 454 (7206):841–845
30. Salvi S, Tuberosa R (2005) To clone or not to clone plant QTLs: present and future challenges. Trends Plant Sci 10:297–304
31. Semon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev 17:505–512
32. Sykes R, Kodrzycki B, Tuskan G, Foutz K, Davis M (2008) Within tree variability of lignin composition in *Populus*. Wood Sci Technol 42:649–661
33. Tuskan GA, West D, Bradshaw HD, Neale D, Sewell M, Wheeler N et al (1999) Two high-throughput techniques for determining wood properties as part of a molecular genetics analysis of loblolly pine and hybrid poplar. Appl Biochem Biotech 77–79:1–11
34. Tuskan GA, Gunter LE, Yang ZMK, Yin TM, Sewell MM, DiFazio SP (2004) Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*. Can J For Res 34:85–93
35. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313:1596–1604
36. Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. Nat Rev, Genet 5:752–763
37. Vanholme R, Morreel K, Ralph J, Boerjan W (2008) Lignin engineering. Curr Opin Plant Biol 11:278–285
38. Van Ooijen JW (2004) MapQTL 5, software for the mapping of quantitative trait loci in experimental populations. Kyazma B.V, Wageningen
39. Yang XH, Jawdy S, Tschaplinski TJ, Tuskan GA (2009) Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. Genomics 93:473–480
40. Yin TM, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA (2004) Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. Theor Appl Genet 109:451–463
41. Yin TM, Zhang XY, Gunter LE, Li SX, Wullschleger SD, Huang MR et al (2009) Microsatellite primer resource for Populus developed from the mapped sequence scaffolds of the Nisqually-1 genome. New Phytol 181(2):498–503
42. Zhong R, Burk DH, Morrison WH, Ye ZH (2002) A kinesin-like protein is essential for oriented deposition of cellulose microfibrils and cell wall strength. Plant Cell 14:3101–3117
43. Zhong R, Demura T, Ye ZH (2006) SND1, a NAC domain transcription factor, is a key regulator of secondary wall synthesis in fibers of Arabidopsis. Plant Cell 18:3158–3170

Critical Reviews in Plant Science, 28:368–374, 2009

ISSN: 0735-2689 print / 1549-7836 online
DOI: 10.1080/07352680903241246

Taylor & Francis
Taylor & Francis Group

Populus Responses to Edaphic and Climatic Cues: Emerging Evidence from Systems Biology Research

Stan D. Wullschleger,[1] David J. Weston,[1] and John M. Davis[2]
[1]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 6442, USA*
[2]*School of Forest Resources and Conservation, University of Florida, Gainesville, FL 32611, USA*

Table of Contents

The emergence of *Populus* as a model system for tree biology continues to be driven by a community of scientists dedicated to developing the resources needed to undertake genetic and functional genomic studies in this genus. As a result, understanding the molecular processes that underpin the growth and development of cottonwood, aspen, and hybrid poplar has steadily increased over the last several decades. Recently, our ability to examine the basic mechanisms whereby trees respond to a changing climate and resource limitations has benefited greatly from the sequencing of the *P. trichocarpa* genome. This landmark event has laid a solid foundation upon which biologists can now quantify, in breathtaking and unprecedented detail, the diversity of genes, proteins, and metabolites that govern the growth and development of some of the longest living and tallest growing organisms on Earth. Although the challenges likely to be encountered by scientists who work with trees are many, recent literature provides a few examples where a systems approach, one that focuses on integrating transcriptomic, proteomic, and metabolomic analyses, is beginning to provide insights into the molecular-scale response of poplars to their climatic and edaphic environment. In this review, our objectives are to look at evidence from studies that examine the molecular response of poplar to edaphic and climatic cues and highlight instances where two or more omic-scale measurements confirm and hopefully expand our inferences about mechanisms contributing to observed patterns of response. Based on conclusions drawn from these studies, we propose that three requirements will be essential as systems biology in poplar moves to reveal unique insights. These include use of genetically-defined individuals (e.g., pedigrees or transgenics) in studies; incorporation of modeling as a complement to transcriptomic, proteomic and metabolomic data; and inclusion of whole-tree and stand-level phenotypes to place molecular-scale insights into a real-world context.

Address correspondence to Stan D. Wullschleger, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 6442, USA. E-mail: wullschlegsd@ornl.gov

Keywords Environmental stress, forestry, genomics, molecular biology, nutrients, trees

I. SETTING THE STAGE FOR SYSTEMS BIOLOGY IN *POPULUS*

Trees, like other multicellular plants, carry out the processes of growth, development, and reproduction in a constantly changing environment. They must possess and marshal a suite of physiological capabilities to cope with harsh climatic conditions and limited availability of nutrient resources, while at the same time ensuring their survival from one season to the next. The coordination of such events must be accomplished by orchestrating a series of complex molecular events in organisms that are distinguished from annuals by their perennial growth, complex crown architecture, dormancy, and juvenile-mature phase changes (Bradshaw *et al.*, 2000; Li *et al.*, 2006; Taylor, 2002; Wu *et al.*, 2000; Wullschleger *et al.*, 2002b). Many of these characteristics arise as a result of, or are facilitated by, long-distance signaling and distribution of water and nutrients and the storage and redistribution of resources, as modified by diurnal, seasonal, and intra-annual variation in climate (Lough and Lucas, 2006).

There have been many attempts to understand the molecular, physiological, and morphological processes by which trees

tackle these challenges, yet, to date, answers are few. A number of molecular studies on stress-responsive gene expression, protein, or metabolite profiling in trees, including *Populus*, have been undertaken. Much of this activity focused on single genes, proteins, or metabolite classes in the years pre-dating the new genetic and genomic resources for trees. While the utility of poplar as an experimental organism was widely recognized (Bradshaw *et al.*, 2000), the capacity to measure molecular-scale processes was a bottleneck. This has changed with the emergence of new technologies made possible by the draft sequencing of the *P. trichocarpa* genome (Tuskan *et al.*, 2006). In theory this should enable scientists to undertake a more comprehensive documentation of genes, proteins, and metabolites in poplar that are responsive to water and nutrient stress, variation in daily and seasonal cues, and encourage better interpretation and integration of such information.

In the wake of this landmark activity it is appropriate to examine how tree biologists have taken advantage of the opportunities created by sequencing the first tree genome. Jansson and Douglas (2007) have already shown that the number of citations for *Populus* has increased nearly 10-fold since the first EST data set was published for poplar (Sterky *et al.*, 1998). While this trend continues, it is additionally appropriate to ask what insights have been gained from our initial investments in sequencing the poplar genome and the increased availability of genomic and molecular tools for this model woody organism (Wullschleger *et al.*, 2002a; Tuskan *et al.*, 2004). Not surprisingly, the community has quickly embraced the opportunities at hand and research is currently being conducted to document the response of genes, proteins, and metabolites to several environmental and edaphic stresses (Azri *et al.*, 2009; Ferreira *et al.*, 2006; He *et al.*, 2008; Kieffer *et al.*, 2009; Lu *et al.*, 2008; Nanjo *et al.*, 2004; Xiao *et al.*, 2009). It is most encouraging that although the field is still in its infancy we are also beginning to see evidence of multiple omic-scale measurements in poplar, including the co-analysis of genes and metabolites, transcript and protein profiling, and metabolite and gene expression dynamics. It is particularly powerful to couple broad surveys of molecules with parallel observations of transgenic lines known to be perturbed in key aspects of the plant response, or with genome sequence-enabled dense marker coverage for QTL analysis. These studies mark the emergence of systems biology in poplar and are the focus of this article.

One often-noted promise of systems biology is that it will enhance our understanding of individual and collective plant functions and thereby provide a more integrated view of plant physiological responses to stress (Guy *et al.*, 2008; Hammer *et al.*, 2004). Some important clues regarding the way poplar copes with climatic stress and resource limitations are beginning to emerge and these are discussed here in relation to water and nutrient availability, and seasonal cues. Our focus is on the non-targeted analysis of genes, proteins, and metabolites, and on insights and understanding derived from global analysis of omic-scale responses in combination with other ecophysiological or genetic observations. Specifically, our objectives are to look at evidence from studies that examine the molecular response of poplar to edaphic and climatic cues and highlight instances where two or more omic-scale measurements either confirm or hopefully expand our inferences about mechanisms contributing to the observed patterns of response. We conclude with an assessment of where systems biology currently stands for poplar and make several recommendations on the future needs of this field.

II. RESPONSES TO WATER AND NUTRIENT AVAILABILITY

A lack of soil water constitutes a frequent limitation to plant growth and productivity, especially in poplar. Fast-growing trees in this genus produce maximum biomass when grown under irrigation (Coyle *et al.*, 2006; Samuelson *et al.*, 2007) or when they otherwise occupy riparian habitats in natural populations (Brunner *et al.*, 2004; DiFazio, 2005). As a result, productivity in poplar relies heavily on having readily available source of soil water (Dickmann *et al.*, 1992; Tschaplinski *et al.*, 1998). Previous research has shown that tolerance to water deficits varies widely among poplar genotypes, both inter- and intraspecifically, suggesting that the genus provides a good model in which to investigate the molecular and genetic basis of traits associated with drought tolerance (Monclus *et al.*, 2005; Tschaplinski *et al.*, 2006; Street *et al.*, 2006).

Although the biochemical, physiological and morphological behavior of important tree species, including poplar, to drought is well documented, only a handful of studies have characterized treatment or genotypic differences using a systems biology approach. Street *et al.* (2006) were among the first to combine classic ecophysiological measurements of plant water relations with quantitative trait loci (QTL) analysis and microarray experiments to examine adaptive strategies to drought in poplar. These authors used a full-sib F_2 mapping population that was previously developed from a cross between *P. trichocarpa* and *P. deltoides* (Bradshaw and Stettler, 1993) and subjected the grandparents and 167 of the F_2 progeny to a 14-d period of soil drying. Two microarray experiments using POP1 cDNA arrays (Andersson *et al.*, 2004) characterized the transcriptome in response to drought. One experiment focused on the two grandparents, while the second focused on a subset of extreme genotypes either sensitive or insensitive to drought on the basis of leaf abscission. In both of the experiments, drought stress resulted in a profound remodeling of the transcriptome. In particular, Street *et al.* (2006) showed that the divergent drought response of two poplar species exhibited segregation within an F_2 population, and that this results in the emergence of highly contrasting adaptive drought responses. Furthermore, comparing the transcriptional response of a set of high- and low-abscission genotypes revealed a striking and surprising degree of separation, suggesting that an expanded understanding of the molecular processes that contribute to leaf abscission could lead to improved biomass

productivity through conventional plant breeding or advanced genetic approaches targeting increased drought tolerance and/or length of growing season in water-limited environments.

While transcript profiling is a powerful and useful indicator of plant response to drought, it provides only one layer of information. Indeed, transcript profiling does not provide information related to protein turnover, sub-cellular localization of proteins or the complex interactions between proteins (Plomion *et al.*, 2006). To understand the major mechanisms that *P. trichocarpa* x *P. deltoides* cv. Beaupré evokes to tolerate drought stress, Plomion *et al.* (2006) investigated stress-induced gene regulation at transcript and protein levels. Both transcriptional and protein expression profiles revealed a general stress response that was consistent with the physiological data that was simultaneously collected. About 1300 and 1600 proteins (i.e., spots) from roots and leaves, respectively were resolved on Coomassie-stained 2D gels. However, only a handful of drought-induced proteins in the leaves and roots showed an increased level of their transcripts. This limited overlap between drought-regulated proteins and drought-induced transcripts likely reflects the different physical and chemical properties of the proteins investigated and the somewhat restricted set of genes (i.e., 2500) represented on the cDNA arrays used in this study. To their credit, what Plomion *et al.* (2006) did show was that a change in protein levels can occur with little or no detectable change in transcript abundance and vice versa. This demonstrates nicely the complementary nature of the transcriptomic and proteomic approaches, and the necessity to combine the two methods to reach full insights into the molecular plasticity response to drought or any other environmental cues.

Interesting and relevant research has sought to address the drought response of *P. euphratica*, a species that, unlike other members of the genus, grows in semiarid regions and is known to tolerate soils with high salinity (Chen *et al.*, 2003). In a series of studies that were conducted in natural (Brosché *et al.*, 2005) and controlled (Bogeat-Triboulot *et al.*, 2007) environments, the expression profiles of ca. 6,340 genes and of proteins and metabolites were recorded for roots and leaves. In the case of the controlled studies, in which young plants were subjected to increasing water deficits for 4 weeks, less than 1.5% of the genes on the arrays displayed significant changes in transcript levels; 70 genes in leaves and 40 genes in the roots. Moreover, the expression profile in roots was very different from that of leaves. Changes in the roots occurred earlier, at lower stress intensity, and predominately consisted of decreased, not increased, transcript abundances. In leaves and roots, most genes displaying altered expression during water deficit returned to control levels within a few days after the plants were re-watered and allowed to recover. Surprisingly, in contrast to the transcriptional response in leaves, the number of proteins whose abundance was modified by water deficit showed no correlation with water stress intensity. Bogeat-Triboulot *et al.* (2007) conclude that molecular response to water deficit in *P. euphratica* involves the regulation of different gene networks in roots and shoots.

Responses to carbon and nitrogen resource availability occur at multiple scales within trees and ecosystems (Millard *et al.*, 2007; Cooke and Weih, 2005). Novaes *et al.* (2009) used an integrated approach to test the hypothesis that responses to increased nitrogen resource availability are under genetic control in poplar. Novaes *et al.* (2009) grew clonal propagules from the pseudo-backcross family 52-124, generated from a [*P. trichocarpa* x *P. deltoides*] x *P. deltoides* cross, under conditions of low and high nitrogen availability. Cell wall metabolites reflecting abundance of syringyl and guiacyl lignin, C5 sugars and cellulose were quantified, as well as biomass traits. Strikingly, 45 of the 51 QTL identified in this study were specific to one condition of nitrogen resource availability. In particular, all of the genes regulating wood chemistry traits appeared to be highly responsive to nutrient availability, since no QTL were co-located under both nitrogen levels. The phenotypic plasticity of poplars in response to increased nitrogen resource availability had been described in a single clone (Cooke *et al.*, 2003) and a feed-forward conceptual model put forward that integrates nitrogen and carbon resource availability (Cooke *et al.*, 2005), and now the analysis of Novaes *et al.* (2009) has revealed the genomic regions governing these plastic responses to nitrogen availability on cell wall metabolites as well as growth traits.

III. RESPONSES TO SEASONAL CUES

Integrated, genome sequence-enabled approaches to dissect poplar storage and redistribution of resources induced by environmental cues are beginning to emerge. The seasonal recurrent transition to and from dormancy is a distinct feature of perennial plants and poplar has long been used to investigate such processes. Previous studies have shown that photoperiod and light quality (Howe *et al.*, 1996) as well as temperature, both heat (Wisniewski *et al.*, 1997) and cold (Rohde and Bhalerao, 2007), are critical for driving dormancy. Adversely cold temperatures ($<20^{\circ}$C) constrain the full genetic potential of plants by inhibiting metabolic reactions directly and indirectly through cold-induced drought resulting in reduced water uptake and cellular dehydration (Chinnusamy *et al.*, 2007). Although much work has been conducted on cold stress signaling and regulation in herbaceous species such as *Arabidopsis* (Zhu *et al.*, 2007), relatively few studies have investigated such processes in poplar. Benedict *et al.* (2006) report one of the few examples of such an approach and investigate the role of C-repeat binding factor (CBF) on acquiring freezing tolerance. Previous research in herbaceous annuals shows that CBF plays an important role in binding to the cis-elements of cold responsive gene promoters and orchestrating transcriptional cascades leading to increased freezing tolerance (Jaglo *et al.*, 2001; Lee *et al.*, 2005). Benedict *et al.* (2006) used ectopic expression of *Arabidopsis At*CBF1 and the POP1 cDNA microarray platform (Andersson *et al.*, 2004) to investigate CBF-mediated low temperature signaling. The authors found that ecotopic expression of *At*CBF1 increases freezing tolerance of nonacclimated

Populus and that comparative transcript profiles between *Populus* and *Arabidopsis* showed strong conservation in CBF regulation. However, there are some distinct differences. In contrast to herbaceous plants, for example, there was differential expression of CBF paralogs between perennial stem tissue and ephemeral leaf tissue. Functional analysis on differential genes between stem and leaf tissues were also evident leading the authors to suggest that perennial driven evolution may have led to specific roles for annual and perennial tissues.

Some recent studies have integrated transcriptome and metabolome profiling to generate new insights into shifts in resource allocation associated with seasonal cues. Such insights were not discernible prior to the availability of the poplar genome sequence. Most notably are the works of Ruttink *et al.* (2007) and Druart *et al.* (2007) and the multi-omics approach that led them to propose multistep-models for dormancy induction in poplar based on the systems integration of transcriptomic and metabolomic data sets.

Ruttink *et al.* (2007) examined poplar trees grown under contrasting long-day and short-day (SD) photoperiods, an inducing treatment known from previous studies to promote bud formation (Goffinet and Larson, 1981) and storage compound accumulation (Nelson and Dickson, 1981; Dickson and Nelson, 1982; Coleman *et al.*, 1992). Ruttink *et al.* (2007) evaluated apical shoot developmental stages using electron microscopy, which identified the cell types involved and the relative timing of storage reserve deposition. They monitored over one-third of the transcriptome at weekly intervals using the POP2 cDNA array (www.populus.db.umu.se), which contains 24,735 probes for 16,494 genes, or 34.6% of the 45,550 genes predicted in the poplar genome. In their transcriptome monitoring experiments, Ruttink *et al.* (2007) included transgenic lines known from previous studies to be perturbed in bud formation (Rohde *et al.* 2002). Specifically a line overexpressing and a line underexpressing *ABSCISIC ACID-INSENSITIVE3* (a transcription factor whose Arabidopsis and maize homologs also play dormancy-associated roles) were included to define mechanisms of *ABI3* action. The experimental design was an interconnected, balanced loop design of the three lines (two transgenic and one nontransgenic) and seven time points, which generated a great deal of statistical power to detect significantly regulated transcripts and metabolites. They evaluated transcripts and metabolites that varied at least four-fold with FDR$<$0.0001 (1091 genes) and FDR$<$0.01 (162 compounds) and found progressive shifts in anatomy, transcriptome and metabolome. Analysis of sequential time points suggested two major coordinated phases of shifts in transcript and metabolite abundance, one within the first week of SD treatments and another after 3 weeks of SD when budset normally occurs. The authors took advantage of additional insights gained previously from studies of poplar lines expressing a mutant version of *ETHYLENE TRIPLE RESPONSE1* (in which bud formation is disrupted; Ruonala *et al.*, 2006), which they correlated with shifts in ethylene-associated transcript abundance. The ethylene-associated effects were detected as a third stage, between the two major shifts in metabolite abundance. Thus the authors proposed three partially overlapping stages in bud development, specifically autumn bud formation, acquisition of desiccation and cold tolerance, and dormancy development. Dormancy development is distinct from bud formation since all of the transgenic lines were perturbed in bud formation but were dormant after six weeks of SD. Perhaps most importantly for future research, genes and metabolites were identified that are markers for discrete stages of this complex developmental process. It is now feasible to identify naturally occurring alleles in poplar populations that are genetically differentiated with respect to dormancy induction, and evaluate the robustness of the multistage model in the context of these naturally-occurring alleles in field environments.

Druart *et al.* (2007) also proposed a multistep model for dormancy induction in the vascular cambium based on time course experiments in which transcript abundance shifts and metabolite shifts were correlated in field-grown trees. The specific targeting of cambial cells is a particular strength of the approach, which employed thin sectioning to recover specific cell types prior to transcript (using POP1 cDNA arrays) and metabolite profiling (using GC-MS after extraction and derivatization). A role for ABA was hypothesized for coordination of late-stage acquisition of cold hardiness based on its peak abundance after the induction of transcripts associated with early-stage cold hardiness. Furthermore, gibberellin (GA) biosynthesis and action were proposed to coordinate transitions out of, and into cambial dormancy, respectively. This was based primarily on the observed accumulation of *REPRESSOR OF GA1-3* transcripts during growth cessation, and the transient induction of *GA-20 OXIDASE* during cambial reactivation. An interesting potential role for chromatin remodeling in dormancy transitions was inferred based on observed expression patterns of poplar homologs of *FERTILISATION INDEPENDENT ENDOSPERM* and *ENHANCER OF ZESTE*, known components of transcriptional repression complexes in *Drosophila*.

IV. CONCLUSIONS AND RECOMMENDATIONS

The *Populus* genome sequence is a landmark in the establishment of a genomics toolkit for forest trees and, as we have hopefully shown in this review, a gateway to systems biology research opportunities. The genome sequence has been leveraged in multiple ways to generate the tools required for integrated, systems-level research. For example, in a short period of time, the assembly of the genome sequence obtained from a single reference genotype created the template required for identifying allelic variants in other genotypes, enabling the construction and use of high-density maps for QTL mapping purposes (Woolbright *et al.*, 2008; Wullschleger *et al.*, 2005; Yin *et al.*, 2009). Furthermore, the identification of gene models obtained by annotation of the reference sequence enabled substantial portions of the transcriptome to be assayed in poplar trees. In this review we describe what may be the genesis of

systems biology research in poplar, manifest by the integration of information from two or more categories of traits (transcriptomic, proteomic, metabolomic) most of which was collected within a genetic framework (a genetically defined population, or transgenic contrasts).

It will be exciting in the next several years to witness the explosion of research and the new syntheses that are likely to emerge from integrative approaches. As mentioned above, the sequencing of the poplar genome has enabled high-density QTLs, which greatly reduces the chromosomal regions explaining the percentage of trait variation. The challenge now is how to explain the underlying processes by which allelic polymorphisms affect such QTLs (Benfey and Mitchell-Olds, 2008), and how best to scale those mechanisms through biological and ecological complexity. Network approaches developed in the biomedical arena have proven to be powerful ways to identify relationships among pathway and process components that are not obvious when transcripts or metabolites are analyzed separately. This systems-oriented approach groups individual genes into functionally relevant modules, which reduces the dimensionality of the omics-data from tens of thousands of genes to a few modules in a biologically meaningful way (Zhang and Horvath, 2005). Modules are then associated with QTL regions and markers, such as single-nucleotide polymorphisms (SNPs), or directly to traits of interest (Ghazalpour *et al.*, 2006; Weston *et al.*, 2008). Such an approach allows the identification of the chromosomal regions or markers influencing module expression and possible mechanisms associated with QTL regions.

Finally, having now reviewed studies in which integrated approaches were taken to dissect poplar responses to the environment, we envision that three features will be essential in development of new insights: 1) explicit inclusion of genetically defined individuals in the experimental framework. The inclusion of genotype information effectively "anchors" the phenotypes collected to allelic or gene expression variation. The ability to clonally propagate poplars allows these genotypes to be immortalized, shared and re-measured in the future; 2) explicit modeling of genetic polymorphisms (allelic variation) as a variable along with transcriptomic, proteomic and metabolomic data. Network approaches developed in the biomedical arena have proven to be powerful ways to identify relationships among pathway and process components that are not obvious when transcripts or metabolites are analyzed separately; and 3) explicit inclusion of whole-tree and stand-level phenotypes to place the molecular-level information in a real-world context. Systems biology can be an organizational framework for understanding tree processes at multiple temporal and spatial scales, but this will require that higher-level growth and yield data are meaningfully integrated into experimental analyses.

ACKNOWLEDGEMENTS

Support provided by the U.S. Department of Energy (DOE), Office of Science, Biological and Environmental Research (BER). Oak Ridge National Laboratory is managed by UT-Battelle, LLC, for the DOE under contract DE-AC05-00OR22725.

REFERENCES

Andersson, A., Keskitalo, J., Sjodin, A., Bhalerao, R., Sterky, F., Wissel, K., Tandre, K., Aspeborg, H., Moyle, R., Ohmiya, Y., Bhalerao, R., Brunner, A., Gustafsson, P., Karlsson, J., Lundeberg, J., Nilsson, O., Sandberg, G., Strauss, S., Sundberg, B., Uhlen, M., Jansson, S., and Nilsson, P. 2004. A transcriptional timetable of autumn senescence. *Genome Biol*. **5**: R24.

Azri, W., Chambon, C., Herbette, S., Brunel, N., Coutand, C., Leple, J.-C., Rejeb, I. B., Ammar, S., Julien, J.-L., and Roeckel-Drevet, P. 2009. Proteome analysis of apical and basal regions of poplar stems under gravitrophic stimulation. *Physiol. Plant.* **136**:193–208.

Benedict, C., Skinner, J. S., Meng, R., Chang, Y., Bhalerao, R., Huner, N.P.A., Finn, C. E., Chen, T.H.H., and Hurry, V. 2006. The CBF1-dependent low temperature signaling pathway, regulon and increase in freeze tolerance are conserved in *Populus* spp. *Plant, Cell Environ*. **29**: 1259–1272.

Benfey, P. N., and Mitchell-Olds, T. 2008. From genotype to phenotype: Systems biology meets natural variation. *Science* **320**: 495–497.

Bogeat-Triboulot, M. B., Brosché, M., Renaut, J., Jouve, L., Le Thiec, D., Fayyaz, P., Vinocur, B., Witters, E., Laukens, K., Teichmann, T., Altman, A., Hausman, J. F., Polle, A., Kangasjarvi, J., and Dreyer, E. 2007. Gradual soil water depletion results in reversible change of gene expression, protein profiles, ecophysiology, and growth performance in *Populus euphratica*, a poplar growing in arid regions. *Plant Physiol*. **143**: 876–892.

Bradshaw, H. D., Ceulemans, R., Davis, J., and Stettler, R. 2000. Emerging model systems in plant biology: Poplar (*Populus*) as a model forest tree. *J. Plant. Growth Regul.* **19**: 306–313.

Bradshaw, H. D., and Stettler, R. F. 1993. Molecular-genetics of growth and development in *Populus*. 1. Triploidy in hybrid poplars. *Theor. Appl. Genet*. **86**: 301–307.

Brosché, M., Vinocur, B., Alatalo, E. R., Lamminmaki, A., Teichmann, T., Ottow, E. A., Djilianov, D., Afif, D., Bogeat-Triboulot, M. B., Altman, A., Polle, A., Dreyer, E., Rudd, S., Lars, P., Auvinen, P., and Kangasjarvi, J. 2005. Gene expression and metabolite profiling of *Populus euphratica* growing in the Negev desert. *Genome Biol*. **6**: R101.

Brunner, A. M., Busov, V. B., and Strauss, S. H. 2004. Poplar genome sequence: functional genomics in an ecologically dominant plant species. *Trends Plant Sci*. **9**: 49–56.

Chen, S., Li, J., Wang, S., Fritz, E., Huttermann, A., and Altman, A. 2003. Effects of NaCl on shoot growth, transpiration, ion compartmentalization, and transport in regenerated plants of *Populus euphratica* and *Populus tomentosa*. *Can. J. For. Res*. **33**: 967–975.

Chinnusamy, V., Zhu, J., and Zhu J. K. 2007. Cold stress regulation of gene expression in plants. *Trends in Plant Sci.* **12**: 444–451.

Coleman, G. D., Chen, T.H.H., and Fuchigami, L. H. 1992. Complementary DNA cloning of poplar bark storage protein and control of its expression by photoperiod. *Plant Physiol*. **98**: 687–693.

Cooke, J.E.K., Brown, K. A., Wu, R., and Davis, J. M. 2003. Gene expression associated with N-induced shifts in resource allocation in poplar. *Plant, Cell Environ*. **26**: 757–770.

Cooke, J.E.K., Martin, T.A., and Davis, J.M. 2005. Short-term physiological and developmental responses to nitrogen availability in hybrid poplar. *New Phytol*. **167**: 41–52.

Cooke, J.E.K., and Weih, M. 2005. Nitrogen storage and seasonal nitrogen cycling in Populus: bridging molecular physiology and ecophysiology. *New Phytol*. **167**: 19–30.

Coyle, D. R., Coleman, M. D., Durant, J. A., and Newman, L. A. 2006. Survival and growth of 31 *Populus* clones in South Carolina. *Biomass Bioenergy* **30**: 750–758.

Dickmann, D. I., Liu, Z. J., and Nguyen, P. V. 1992. Photosynthesis, water relations, and growth of 2 hybrid *Populus* genotypes during a severe drought. *Can. J. For. Res*. **22**: 1094–1106.

Dickson, R. E., and Nelson, E. A. 1982. Fixation and distribution of ^{14}C in *Populus deltoides* during dormancy induction. *Physiol. Plant.* **54**: 393–401.

DiFazio, S. P. 2005. A pioneer perspective on adaptation. *New Phytol*. **165**: 661–664.

Druart, N., Johansson, A., Baba, K., Schrader, J., Sjodin, A., Bhalerao, R. R., Resman, L., Trygg, J., Moritz, T., and Bhalerao, R. P. 2007. Environmental and hormonal regulation of the activity-dormancy cycle in the cambial meristem involves stagespecific modulation of transcriptional and metabolic networks. *Plant J*. **50**: 557–573.

Ferreira, S., HjernØ, K., Larsen, M., Wingsle, G., Larsen, P., Fey, S., Roepstorff, P., and Salomé Pais, M. 2006. Proteome profiling of *Populus euphratica* Oliv. upon heat stress. *Ann. Bot*. **98**: 361–377.

Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, A. B., Schadt, E. E., Drake, T. A., Lusis, A., and Horvath, S. 2006. Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS* **2**:e130.

Goffinet, M. C. and Larson, P. R. 1981. Structural changes in *Populus deltoides* terminal buds and in the vascular transition zone of the stems during dormancy induction. *Amer. J. Bot.* **68**: 118–129.

Guy, C., Kaplan, F., Kopka, J., Selbig, J., and Hincha, D. K. 2008. Metabolomics of temperature stress. *Physiol. Plant*. **132**: 220–235.

Kieffer, P., Planchon, S., Oufir, M., Ziebel, J., Dommes, J., Hoffmann, L., Hausman, J.-F., and Renaut, J. 2009. Combining proteomics and metabolite analysis to unravel cadmium stress-response in poplar leaves. *J. Proteome Res*. **8**: 400–417.

Hammer, G. L., Sinclair, T. R., Chapman, S. C., and van Oosterom, E. 2004. On systems thinking, systems biology, and the *in silico* plant. *Plant Physiol*. **134**: 909–911.

He, C., Zhang, J., Duan, A., Zheng, S., Sun, H., and Fu, L. 2008. Proteins responding to drought and high-temperature stress in *Populus* x *euramericana* cv. '74/76". *Trees* **22**: 803–813.

Howe, G. T., Gardner, G., Hackett, W. P., and Furnier, G. R. 1996. Phytochrome control of short-day-induced bud set in black cottonwood. *Physiol. Plant.* **97**: 95–103.

Jaglo, K. R., Kleff, S., Amundsen, K. L., Zhang, X., Haake, V., Zhang, J. Z., Deits, T., and Thomashow, M. F. 2001. Components of the *Arabidopsis* C-repeat/dehydration-responsive element binding factor cold-responsive pathway are conserved in *Brassica napus* and other plant species. *Plant Physiol.* **127**: 910–917.

Jansson, S., and Douglas, C. J. 2007. *Populus*: A model system for plant biology. *Annu. Rev. Plant Biol*. **58**: 435–458.

Lee, B. H., Henderson, D. A., and Zhu, J. K. 2005. The Arabidopsis cold-responsive transcriptome and its regulation by ICE1. *Plant Cell* **17**: 3155–3175.

Li, L. G., Lu, S. F., and Chiang, V. 2006. A genomic and molecular view of wood formation. *Crit. Rev. Plant Sci.* **25**: 215–233.

Lough, T. J., and Lucas, W. J. 2006. Integrative plant biology: Role of phloem long-distance macromolecular trafficking. *Annu. Rev. Plant Biol*. **57**: 203–232.

Lu, S., Sun, Y.-H., and Chang, V. L. 2008. Stress-responsive microRNAs in *Populus*. *Plant J.* **55**: 131–151.

Millard, P., Sommerkorn, M., and Grelet, G.-A. 2007. Environmental change and carbon limitation in trees: a biochemical, ecophysiological and ecosystem appraisal. *New Phytol*. **175**: 11–28.

Monclus, R., Dreyer, E., Villar, M., Delmotte, F. M., Delay, D., Petit, J.-M., Barbaroux, C., Le Thiec, D. Bréchet, C., and Brignolas, F. 2005. Impact of drought on productivity and water use efficiency in 29 genotypes of *Populus deltoides* x *Populus nigra*. *New Phytol*. **169**: 765–777.

Nanjo, T., Futamura, N., Nishiguchi, M., Igasaki, T., Shinozaki, K., and Shinohara, K. 2004. Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. *Plant Cell Physiol*. **45**: 1738–1748.

Nelson E. A. and Dickson, R. D. 1981 Accumulation of food reserves in cottonwood stems during dormancy induction. *Can. J. For. Res.* **11**: 145–154.

Novaes, E., Osorio, L., Drost, D. R., Miles, B. L., Novaes, C.R.D.B., Benedict, C., Dervinis, C., Yu, Q., Sykes, R., Davis, M., Martin, T. A., Peter, G. F., and Kirst, M. 2009. Genetic analysis of nitrogen effect on biomass and wood chemistry of Populus identifies a major pleiotropic locus that links tree growth and wood properties. *New Phytol*. (in press).

Plomion, C., Lalanne, C., Claverol, S., Meddour, H., Kohler, A., Bogeat-Triboulot, M. B., Barre, A., Le Provost, G., Dumazet, H., Jacob, D., Bastien, C., Dreyer, E., de Daruvar, A., Guehl, J.M., Schmitter, J. M., Martin, F., and Bonneu, M. 2006. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* **6**: 6509–6527.

Rohde, A. and Bhalerao, R. P. 2007. Plant dormancy in the perennial context. *Trends Plant Sci.* **12**: 217–223.

Rohde, A., Prinsen, E., De Rycke, R., Engler, G., Van Montagu, M., and Boerjan, W. 2002. PtABI3 impinges on the growth and differentiation of embryonic leaves during bud set in poplar. *Plant Cell* **14**: 1885–1901.

Ruonala, R., Rinne, P.L.H., Baghour, M., Moritz, T., Tuominen, H., and Kangajarvi, J. 2006. Transitions in the functioning of the shoot apical meristem in birch (*Betula pendula*) involve ethylene. *Plant J*. **46**: 628–640.

Ruttink, T., Arend, M., Morreel, K., Storme, V., Rombauts, S., Fromm, J., Bhalerao, R. P., Boerjan, W., and Rohde, A. 2007. A molecular timetable for apical bud formation and dormancy induction in poplar. *Plant Cell* **19**: 2370–2390.

Samuelson, L. J., Stokes, T. A., and Coleman, M. D. 2007. Influence of irrigation and fertilization on transpiration and hydraulic properties of *Populus deltoides*. *Tree Physiol*. **27**: 765–774.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., van Montagu, M., Sandberg, G., Olsson, O., Teeri, T.T., Boerjan, W., Gustafsson, P., Uhlen, M., Sundberg, B., and Lundeberg, J. 1998. Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. *Proc. Nat. Acad. Sci. USA* **22**: 13330–13335.

Street, N. R., Skogström, O., Sjödin, A., Tucker, J., Rodríguez-Acosta, M., Nilsson, P., Jansson, S., and Taylor, G. 2006. The genetics and genomics of the drought response in *Populus*. *Plant J.* **48**: 321–341.

Taylor, G. 2002. *Populus*: Arabidopsis for forestry: Do we need a model tree? *Ann. Bot*. **90**: 681–689.

Tschaplinski, T. J., Tuskan, G. A., Gebre, G. M., and Todd, D. E. 1998. Drought resistance of two hybrid *Populus* clones grown in a large-scale plantation. *Tree Physiol*. **18**: 653–658.

Tschaplinski, T. J., Tuskan, G. A., Sewell, M. M., Gebre, G. M., Todd, D. E., and Pendley, C. D. 2006. Phenotypic variation and quantitative trait loci identification for osmotic potential in an interspecific hybrid inbred F_2 poplar pedigree grown in contrasting environments. *Tree Phyiol*. **26**: 595–604.

Tuskan, G. A., DiFazio, S. P., and Teichmann, T. 2004. Poplar genomics is getting poplar: The impact of the poplar genome project on tree research. *Plant Biol*. **6**: 2–4.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P. M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., Depamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C.,

Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., Van de Peer, Y., and Rokhsar, D. 2006. The genome of black cottonwood *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Weston, D. J., Gunter, L. E., Rogers, A., and Wullschleger, S. D. 2008. Connecting genes, coexpression modular signatures to environmental stress phenotypes in plants. *BMC Sys. Biol*. **2**: 16.

Wisniewski, M. Suater, J. Fuchigami, L., and Stepien, V. 1997. Effects of near-lethal heat stress on bud break, heat-shock proteins and ubiquitin in dormant poplar (*Populus nigra Charkowiensis x P. nigra incrassata*). *Tree Physiol.* **17**: 453–460.

Woolbright, S. A., DiFazio, S. P., Yin, T., Martinsen, G. D., Zhang, X., Allan, G.J ., Whitham, T. G., and Keim, P. 2008. A dense linkage map of hybrid cottonwood (*Populus fremontii x P. angustifolia*) contributes to long-term ecological research and comparison mapping in a model forest tree. *Heredity* **100**: 59–70.

Wu, R. L., Hu, X. S., and Han, Y. F. 2000. Molecular genetics and developmental physiology: Implications for designing better forest crops. *Crit. Rev. Plant Sci*. **5**: 377–393.

Wullschleger, S. D., Jansson, S., and Taylor, G. 2002a. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* **14**: 2651–2655.

Wullschleger, S. D., Tuskan, G. A., and DiFazio, S. P. 2002b. Genomics and the tree physiologist. *Tree Physiol*. **22**: 1273–1276.

Wullschleger, S. D., Yin, T. M., DiFazio, S. P., Tschaplinski, T. J., Gunter, L. E., Davis, M. F., and Tuskan, G. A. 2005. Phenotypic variation in growth and biomass distribution for two advanced-generation pedigrees of hybrid poplar. *Can. J. For. Res*. **35**: 1779–1789.

Xiao, X., Yang, F., Zhang, S., Korpelainen, H., and Li, C. 2009. Physiological and proteomic responses of two contrasting *Populus cathayana* populations to drought stress. *Physiol. Plant*. **136**: 150–168.

Yin, T. M., Zhang, X. Y., Gunter, L. E., Li, S. X., Wullschleger, S. D., Huang, M. R., and Tuskan, G. A. 2009. Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome. *New Phytol*. **181**: 498–503.

Zhang, B. and Horvath, S. 2005. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol*. **4**:Article17.

Zhu, J., Dong, C. H., and Zhu, J. K. 2007. Interplay between cold-responsive gene regulation, metabolism and RNA processing during plant cold acclimation. C*ur. Opin. Plant Biol.* **10**: 290–295.

BMC Genomics

BioMed Central

Research article

Transcriptomic and metabolomic profiling of *Zymomonas mobilis* during aerobic and anaerobic fermentations

Shihui Yang[1], Timothy J Tschaplinski[1], Nancy L Engle[1], Sue L Carroll[1], Stanton L Martin[2], Brian H Davison[1], Anthony V Palumbo[1], Miguel Rodriguez Jr[1] and Steven D Brown*[1]

Address: [1]Biosciences Division and BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA and [2]North Carolina State University, 840 Main Campus Drive, Raleigh, NC 27606, USA

Email: Shihui Yang - yangs1@ornl.gov; Timothy J Tschaplinski - tschaplinstj@ornl.gov; Nancy L Engle - englenl@ornl.gov; Sue L Carroll - carrollsl@ornl.gov; Stanton L Martin - martins@ornl.gov; Brian H Davison - davisonbh@ornl.gov; Anthony V Palumbo - palumboav@ornl.gov; Miguel Rodriguez - rodriguezmjr@ornl.gov; Steven D Brown* - brownsd@ornl.gov

* Corresponding author

Published: 20 January 2009

BMC Genomics 2009, **10**:34 doi:10.1186/1471-2164-10-34

Received: 3 July 2008
Accepted: 20 January 2009

This article is available from: http://www.biomedcentral.com/1471-2164/10/34

Abstract

Background: *Zymomonas mobilis* ZM4 (ZM4) produces near theoretical yields of ethanol with high specific productivity and recombinant strains are able to ferment both C-5 and C-6 sugars. *Z. mobilis* performs best under anaerobic conditions, but is an aerotolerant organism. However, the genetic and physiological basis of ZM4's response to various stresses is understood poorly.

Results: In this study, transcriptomic and metabolomic profiles for ZM4 aerobic and anaerobic fermentations were elucidated by microarray analysis and by high-performance liquid chromatography (HPLC), gas chromatography (GC) and gas chromatography-mass spectrometry (GC-MS) analyses. In the absence of oxygen, ZM4 consumed glucose more rapidly, had a higher growth rate, and ethanol was the major end-product. Greater amounts of other end-products such as acetate, lactate, and acetoin were detected under aerobic conditions and at 26 h there was only 1.7% of the amount of ethanol present aerobically as there was anaerobically. In the early exponential growth phase, significant differences in gene expression were not observed between aerobic and anaerobic conditions via microarray analysis. HPLC and GC analyses revealed minor differences in extracellular metabolite profiles at the corresponding early exponential phase time point.

Differences in extracellular metabolite profiles between conditions became greater as the fermentations progressed. GC-MS analysis of stationary phase intracellular metabolites indicated that ZM4 contained lower levels of amino acids such as alanine, valine and lysine, and other metabolites like lactate, ribitol, and 4-hydroxybutanoate under anaerobic conditions relative to aerobic conditions. Stationary phase microarray analysis revealed that 166 genes were significantly differentially expressed by more than two-fold. Transcripts for Entner-Doudoroff (ED) pathway genes (*glk, zwf, pgl, pgk, and eno*) and gene *pdc*, encoding a key enzyme leading to ethanol production, were at least 30-fold more abundant under anaerobic conditions in the stationary phase based on quantitative-PCR results. We also identified differentially expressed ZM4 genes predicted by The Institute for Genomic Research (TIGR) that were not predicted in the primary annotation.

Conclusion: High oxygen concentrations present during *Z. mobilis* fermentations negatively influence fermentation performance. The maximum specific growth rates were not dramatically different between aerobic and anaerobic conditions, yet oxygen did affect the physiology of the cells leading to the buildup of metabolic byproducts that ultimately led to greater differences in transcriptomic profiles in stationary phase.

Background

Recent high oil prices, concerns over energy security, and environmental goals have reawakened interest in producing alternative fuels via large-scale industrial fermentations. The potential and challenges involved in supplanting a substantial amount of petroleum derived transportation fuels with fuels derived from renewable resources such as ethanol from lignocellulosic materials has been the focus of several studies and reviews [1-4]. The development and deployment of ethanologenic microorganisms will be one critical component in the successful production of fuel ethanol in industrial-scale quantities.

Essential traits for an industrial microorganism include high ethanol yield, tolerance, and productivity (> 90% of theoretical, > 40 g L^{-1}, > 1 g L^{-1} h^{-1}, respectively); robust growth with simple, inexpensive growth requirements in conditions that retard contaminants (eg higher temperatures); and inhibitor tolerance, as reviewed previously [5]. Higher tolerance, productivity values and other positive industrial attributes have been reported for *Z. mobilis*, as reviewed previously [6]. Ethanol tolerance comparable up to 85 g L^{-1}(11% v/v) have been reported for *Z. mobilis* continuous culture and up to 127 g L^{-1} (16% v/v) in batch culture and productivities of 120–200 gL^{-1} h^{-1} in continuous processes with cell recycle [6]. *Saccharomyces* yeasts have been the preferred industrial biocatalyst for fuel ethanol production, although genetically engineered bacterial species such as Gram-negative bacteria *Escherichia coli*, *Zymomonas mobilis*, and *Klebsiella oxytoca* as well as Gram-positive bacteria *Bacillus subtilis* and *Corynebacterium glutamicum* are in development to address commercially important inoculum requirements [5,7,8]. Indeed, a newly formed partnership between the DuPont and Broin companies will utilize recombinant strains of *Z. mobilis* for bio-ethanol fermentation from the lignocellulosic residues such as corn stover [9].

Z. mobilis ferments glucose, fructose, and sucrose producing ethanol and carbon dioxide via the Entner-Doudoroff (ED) pathway, utilizing pyruvate decarboxylase and alcohol dehydrogenase enzymes (see [6,10-12] for reviews). *Z. mobilis* is not a classic facultative organism, rather it is aerotolerant, negating oxygen requirements in fermentations and the need for expensive oxygen transfer. The unusual physiology of *Z. mobilis* generates only one mole of ATP per mole of glucose, which results in low biomass production and greater carbon being available for fermentation products under anaerobic conditions. Its desirable ethanologenic attributes also include: high sugar uptake rates, near theoretical ethanol yields, high ethanol tolerance and generally regarded as safe (GRAS) status. Wild-type *Z. mobilis* can only utilize a limited range of substrates; however, it is amenable to genetic manipulation and recombinant strains have been developed to ferment pentose sugars such as xylose and arabinose [13-15]. Seo et al (2005) reported the first genome sequence for *Z. mobilis* ZM4 [16] and the U. S. Department of Energy's Joint Genome Institute (JGI) http://www.jgi.doe.gov/ has announced plans to sequence the genomes of additional *Z. mobilis* strains in the near future. The ZM4 genome sequence provides new opportunities for fundamental insight into the physiology and gene function and regulation of this unique microorganism and likely improvements in strain development [17]. The presence of oxygen during *Z. mobilis* fermentations has been observed to negatively affect cell and ethanol yields with acetic acid and acetaldehyde accumulating in the medium [12,18-20]. However, there are also reports that respiration inhibition by cyanide stimulates *Z. mobilis* growth aerobically [21]. Many aspects of end-product inhibition and basic aspects of *Z. mobilis* physiology remain unexplored [11].

In this study, we combined transcriptomic profiling with metabolomic measurements of aerobic and anaerobic *Z. mobilis* fermentations to elucidate the molecular mechanisms of ZM4s growth with and without oxygen. Metabolomics is the study of the low molecular weight metabolites present in and around a biological organism at a given time [22,23]. To date, there are few reports of bacterial mRNA expression-profiling using whole genome microarray analysis conducted in conjunction with metabolomics [24]. We confirmed maintenance of anaerobiosis was important for maximizing ethanol yields since acetaldehyde, acetate, lactate and acetoin accumulated and ethanol was present in lower concentrations under controlled aerobic conditions. Our data revealed alterations in mRNA expression profiles, differences in intra- and extracellular metabolites, as well as identification of differential expression for coding sequences predicted by TIGR http://cmr.jcvi.org/tigr-scripts/CMR/GenomePage.cgi?org=ntzm01 and not predicted in the primary annotation. These data provide global insight into potential molecular mechanisms of *Z. mobilis* aerobic and end-product stress responses.

Results

ZM4 grows and consumes glucose faster under anaerobic conditions

The presence of oxygen negatively affected glucose consumption and growth in *Z. mobilis* ZM4 fermentations (Fig. 1). Anaerobic fermentation led to a maximal culture density of 7.0 OD_{600} units approximately 9 h post-inoculation, while *Z. mobilis* did not reach its highest culture density of 6.5 OD_{600} units until 13 h post-inoculation under aerobic conditions despite initial inocula concentrations being slightly greater than the former condition. *Z. mobilis* also consumed glucose more slowly under aerobic conditions, with more than half of the initial aerobic

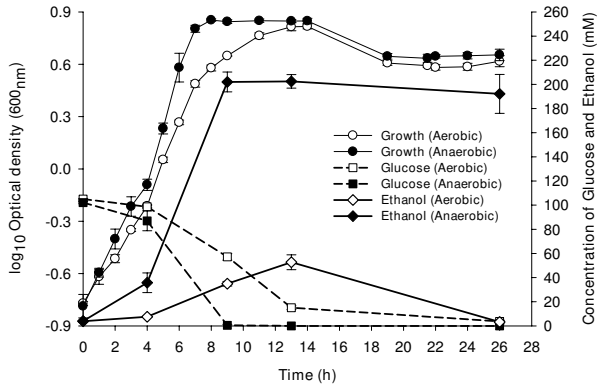


Figure 1
***Z. mobilis* fermentations under anaerobic and aerobic conditions**. Mean values for triplicate fermentors are shown for each condition ± standard deviation (bars).

glucose concentration (105 mM) remaining 9 h post-inoculation. Under anaerobic conditions 99.5% of the glucose had been utilized at this time point. When *Z. mobilis* growth reached its peak after 13 h under aerobic conditions 14% of the glucose remained with the remainder consumed without cell growth (Fig. 1). Despite these differences and differences in extracellular metabolite production (below), the maximum specific growth rates were not dramatically different, which were estimated to be 0.45 h^{-1} and 0.55 h^{-1} between aerobic and anaerobic conditions, respectively. Fermentor pH, dissolved O_2 tension (DOT), and agitation speed were well-controlled, which was indicated by the mean fermentor pH, DOT, and agitation values for each condition (see Additional file 1), and DOT trend data for each condition (see Additional file 2).

Z. mobilis *extracellular metabolite production*

Previous reports have indicated that ethanol production was decreased and other end-products such as acetaldehyde, acetate and acetoin were increased under aerobic conditions [12,18-20,25]. GC and HPLC were used to quantify and compare the kinetics of ethanol, acetate, acetaldehyde, lactate and acetoin production during aerobic and anaerobic fermentation processes and extracellular metabolites were often measured by more than one approach, which confirmed the observed trends. The more rapid production of ethanol under anaerobic conditions also corresponded with increased glucose uptake and growth under these conditions (Fig. 1). The ethanol concentration remained relatively stable post-peak production in anaerobic fermentations. In contrast, at 13 h the ethanol concentration dropped sharply from 52.7 mM to 3.2 mM at the end of the 26 h fermentation during aerobic fermentation. The decrease in ethanol concentration during this time was matched in nearly stoichiometric increases of acetate production, which went from 8.4 mM at 13 h to 72.1 mM at 26 h (Fig. 2). GC data also showed several other minor unidentified metabolites were being produced during aerobic fermentations relative to anaerobic conditions (data not shown). The concentration of acetaldehyde increased during the exponential growth and dropped appreciably during stationary phase, while acetoin was detected during stationary phase (Fig. 2). Lactate trended similarly to ethanol and acetoin profiles for respective conditions. The anaerobic ethanol yield at 13 h was 0.497 g/g of glucose or 97% theoretical and less than 1% of the yield went to solvent products other than ethanol or CO_2 (Fig. 1). In contrast, under aerobic conditions at 13 h 0.14 g/g of glucose or 27% theoretical yield was obtained for ethanol and by 26 h around 3.2 mM ethanol remained (Fig. 1). The total measured solvents produced aerobically was 101 g/L, which was approximately 50% of theoretical total solvent yield at 13 h and at 26 h these figures had dropped to 76 g/L or approximately 37% of theoretical total solvent yield. As cell biomass (as measured by optical density) was approximately equivalent under the two conditions more carbon went to maintenance energy under aerobic conditions.

Z. mobilis *intracellular metabolomic profiles*

The physiological status of ZM4 was investigated further by GC-MS analysis of stationary phase intracellular metabolomic profiles for each respective condition. GC-MS identified a large number of metabolites, however, we present those that showed differences in relative abundance and showed some statistical significance. We observed differences in relative abundance profiles for metabolites related to central carbon metabolism, although only several were considered significantly different during the stationary phase time point we examined. We present a relative comparison of the metabolite difference from two conditions in keeping with current practices, while recognizing accessing metabolites like ATP will require modified procedures [22]. Metabolite identification and analysis gave 20 metabolites that were different between anaerobic and aerobic cultures, with five out of the 20 less abundant in anaerobic *Z. mobilis*. Twelve metabolites were considered significantly different with a *p*-value less than 0.1 (Table 1). The inclusion of a greater number of replicates would have increased analysis power and possibly increased the number of metabolites considered significantly different between anaerobic cells and those exposed to oxygen.

Glucose-6-phosphate and 2-phosphoglycerate were 2.7 and 3.8 fold more abundant under anaerobic conditions ($p = 0.089$ and 0.053, respectively), while others in the ED pathway may have been at greater levels under aerobic conditions. The ability to detect labile phosphorylated

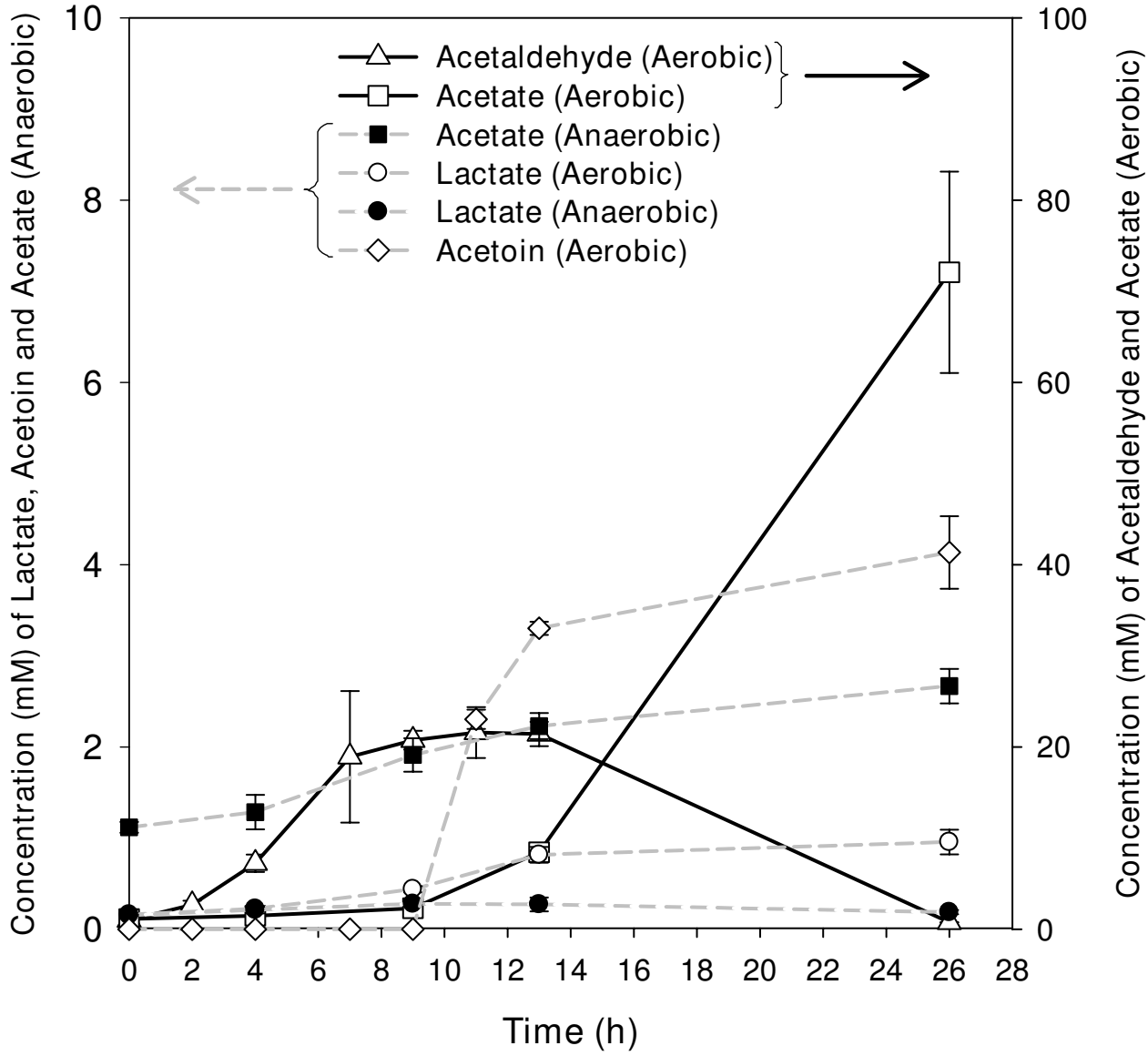


Figure 2
***Z. mobilis* extracellular fermentation product analysis**. The mean value for each metabolite identified by GC analysis from three independent fermentors for each condition is presented ± standard deviation (bars).

intermediates such as these, in the absence of accumulation phosphate was indicative of adequate sample preparation for GC-MS analysis of the metabolites analyzed in the present study. Aerobic *Z. mobilis* contained more of the glucose substrate, as well as glycerate, and gluconate as compared to the anaerobic cells, while other intermediate sugar metabolites that included glucose-6-phosphate, 2-phosphoglycerate and mannose-6-phosphate were more abundant within the anaerobic fermenting *Z. mobilis* (Table 1). The transcriptomic profiles of Entner-Dondoroff and pyruvate metabolic pathways showed higher expression values for these genes under anaerobic conditions even though extracellular glucose had been consumed by 26 h (Fig. 1, see Additional file 3). All of the differentially detected amino acids were at lower concentrations within anaerobic fermenting *Z. mobilis* (Table 1). The other metabolites showing differences included lactate, 4-hydroxybutanoate, ribitol, trehalose and unknown metabolites. Trehalose was 2.9-fold more abundant within aerobic *Z. mobilis* compared to anaerobic cultures, and 4-hydroxybutanoate was the most abundant metabo-

lite detected within anaerobic condition at 7.6-fold higher levels compared to aerobic conditions (Table 1).

Transcriptome comparison of Z. mobilis *aerobic and anaerobic fermentations*

The global transcriptional response of *Z. mobilis* ZM4 to aerobic stress was examined in a time series DNA microarray experiment for early exponential phase (3 h post-inoculation) and stationary phase (26 h post-inoculation) under aerobic and anaerobic conditions with whole-genome microarrays. In this study, we have presented experimental data of 166 differentially expressed genes at statistically significant values ($-\log_{10}(p) = 5.43$) with a number of 11 putative protein-coding sequences in strain ZM4 (ATCC31821) that were not originally described in the primary genome annotation (see Additional file 4, 5). We have deposited the entire microarray dataset including exponential and stationary phase at Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) database with the accession number of GSE10302 so interested parties can conduct their analyses.

Gene expression during the early exponential phase was surprisingly similar between anaerobic and aerobic conditions used in the present study (Fig. 3A). Eight genes were discovered that were considered differentially expressed at a significant level utilizing the ANOVA model using the False Discovery Rate (data not shown). Of these, only one gene (ZMO1752, encoding a hypothetical protein) showed a greater than a two-fold difference in relative expression levels. To confirm the microarray results, seventeen genes involving in ED and pyruvate pathways and different cellular functions were chosen for the qPCR analysis (see Additional file 6, 7). The data showed that qPCR was more sensitive and showed greater fold change differences compared to the microarray analysis, which was in keeping with previous reports [26]. We describe here only a rigorous analysis of stationary phase genes in this study and allow interested parties to conduct their own analyses on the entire dataset, which is available publicly through the GEO database. In the stationary phase 166 genes were significantly differentially expressed between anaerobic and aerobic conditions (Fig. 3B, 4, see Additional file 4, 5). This time point also showed the largest differences in extracellular metabolite profiles (Fig. 2). Fifty-five genes were up-regulated at 26 h post-inoculation under aerobic conditions and 111 genes were down-regulated (see Additional file 4, 5). Approximately two thirds of the genes down-regulated in the presence of oxygen for this time point were related to metabolism (see Additional file 4). In the presence of oxygen, genes related to regulation, cell processes, transport, and unknown function showed greater expression as compared to anaerobic conditions. Nearly half of the genes showing greater expression aerobically remain uncharacterized (see Additional file 5).

Table 1: Intracellular metabolomic profile of ZM4 during anaerobic and aerobic fermentation at 26 h.

Metabolite	Ratio (Anaerobic/Aerobic)	*p*-value
Glucose 6-P	2.7	0.089
Mannose 6-P	3.2	0.009
Glycerate	0.65	0.10
Glucose	0.16	0.14
Gluconate	0.17	0.33
2-Phosphoglycerate	3.8	0.053
Alanine	0.33	0.015
Valine	0.32	0.031
Lysine	0.35	0.097
Isoleucine	0.32	0.17
Leucine	0.29	0.13
Phenylalanine	0.31	0.23
Serine	0.17	0.19
Threonine	0.27	0.21
lactate	0.50	0.01
4-hydroxybutanoate	7.60	0.029
Ribitol	0.06	0.034
Trehalose	0.34	0.13
Unknown 285 18.19	3.0	0.074
Unknown 348 11.22 AA	0.72	0.024

In the stationary phase, ED pathway mRNAs such as glk, zwf, pgl, pgk, and eno as well as ethanol fermentation gene transcripts like pdc and adhB were shown to be more abundant (at least two-fold) under anaerobic conditions by microarray analysis (see Additional file 3, 4, 7). In this study we observed that one arginyl-tRNA synthetase involved in ribosome-mediated polypeptide synthesis, eight genes related to ribosomal protein synthesis and amino acid and co-factor biosynthetic genes such as leuC, trpB, argC, ilvI, ilvC, thrC, thiC, and ribC showed greater expression under anaerobic conditions (see Additional file 4). Metabolomics data showed that a number of detected amino acids were generally less abundant in anaerobic fermenting cells compared to aerobic cells (Table 1).

Our microarray data identified a cysteine desulfurase (ZMO1022), thiosulfate sulfurtransferase (ZMO1460) and a [2Fe-2S] binding domain family protein (NT01ZM1467) as being up-regulated under aerobic conditions (see Additional file 5) and suggested the metabolism of sulfur compounds was impacted by aerobic conditions. However, we did not observe any trends related to sulfur-containing metabolites in our metabolomics dataset. Genes related to sensing and responding to environmental signals including three chemotaxis

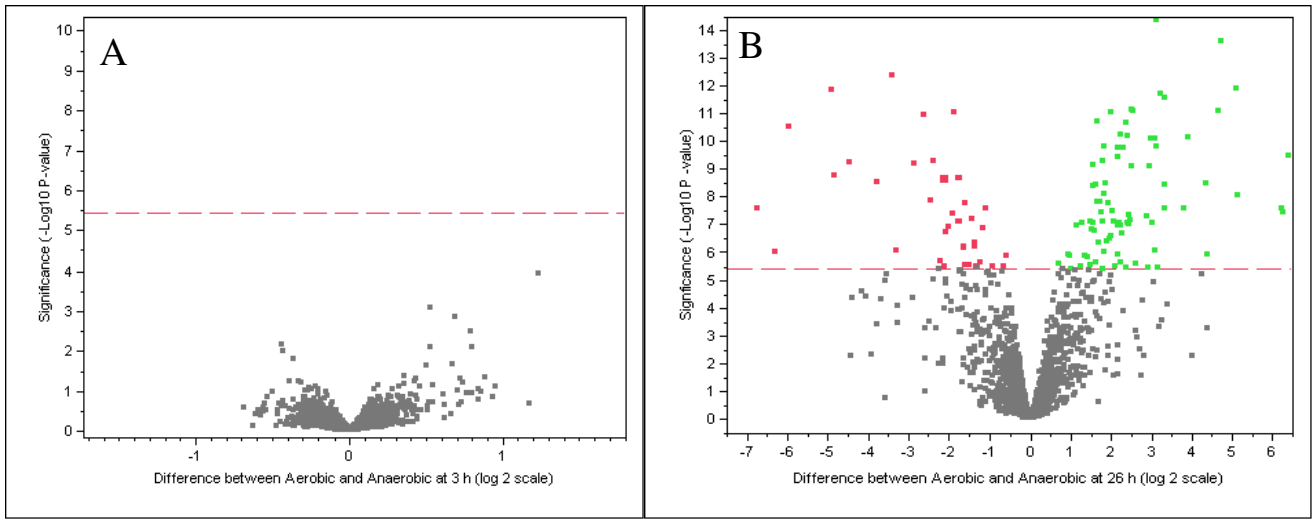


Figure 3
Volcano plot result from JMP Genomics analysis showing significantly differentially expressed genes at 3 h (A) and 26 h (B) post-inoculation at 30°C. Green dots indicate oxygen up-regulated genes and red dots indicate the oxygen down-regulated genes. Grey colored dots were not considered significantly differentially expressed. The X axis shows the difference values between aerobic and anaerobic fermentations based on a $\log_2$ scale. The Y axis shows statistical significance values for expression values, based on a $-\log_{10} p$-value. The red dashed line shows the statistical significance cut-off used in this study.

genes *cheX* (ZMO0084), *motD* (ZMO0641) and *fliD* (ZMO0651), six transcriptional regulators including a two-component signal transduction (TCSTS) histidine kinase (ZMO1216) and response regulator (ZMO1387) were up-regulated during aerobic fermentation (see Additional file 5). We also observed that several ATP synthase subunit genes (alpha and beta) were expressed less under aerobic conditions. The expression of NAD synthetase gene (*nadE*), involved in nicotinamide adenine dinucleotide *de novo* biosynthesis and salvage pathways was approximately 9-fold greater in the presence of oxygen (see Additional file 5). Several flavoprotein transcripts, nitroreductase (*tdsD*) and flavodoxin (*nifF*), were approximately 9-fold more abundant under aerobic conditions.

Expression of a number of stress response genes was found to be greater in the presence of oxygen or by-products in the stationary phase. Under aerobic conditions ZMO1097 encoding a thioredoxin was induced approximately four-fold, ZMO1830 (*fdxB*) encoding a ferredoxin showed six-fold induction, ZMO1732 (*ahpC*) encoding an alkyl hydroperoxide reductase showed 18-fold greater expression, ZMO0279 encoding a cold-shock protein was induced two-fold, and a glutathione S-transferase family protein encoded by ZMO1118 was induced eight-fold. The *E. coli* alternative sigma factor *rpoH* plays an important role in overcoming oxidative stress responses [27] and *rpoH* (ZMO0749) was induced approximately 32-fold under aerobic conditions via q-PCR in our study (see Additional file 7). Other regulators with greater expression levels under aerobic condition include ZMO1121 encoding a MerR family regulator, ZMO1216 encoding a two-component signal transduction histidine kinase, ZMO1387 encoding a two-component response regulator, and ZMO1063 (*pspA*) encoding a sigma 54-dependent transcription suppressor (see Additional file 5). A number of these microarray expression values were confirmed by qPCR (see Additional file 5, 7).

The microarray data also indicated that a number of CDS predicted by TIGR but not present in the primary annotation [16] were differentially expressed between the two conditions. Transcripts for NT01ZM1467 encoding a Fe-S binding domain family protein and nine other ORFs with unknown functions were more abundant under aerobic conditions representing approximately 9% of all the genes up-regulated under these conditions (see Additional file 5). The only example of an up-regulated gene (~six-fold) without primary locus identification during anaerobic fermentation was NT01ZM0869, which encodes a putative arginyl-tRNA synthetase (see Additional file 4).

Discussion

Z. mobilis has a number of positive attributes as an ethanogen and is a leading current generation candidate microorganism for use in commercial scale fermentations to produce fuel ethanol [6,10]. In the present study, we confirmed oxygen levels were important in *Z. mobilis*

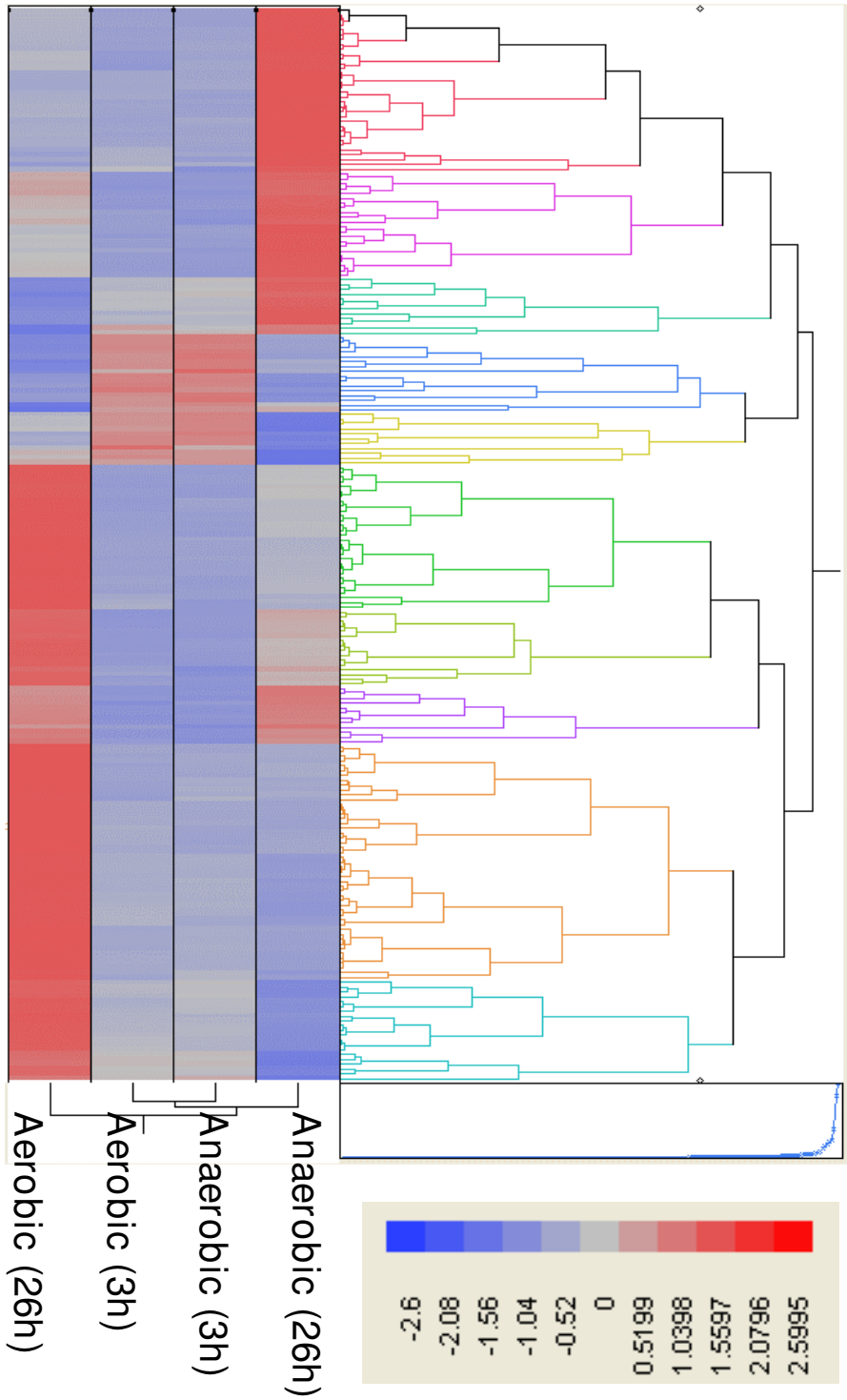


Figure 4
Hierarchical cluster analysis of significantly differentially expressed ZM4 genes for aerobic and anaerobic fermentations at 3 and 26 h. Gene expression values were clustered based on their $\log_2$ based expression values using JMP Genomics 3.0. Negative numbers (colored blue) indicates less relative gene expression aerobically, and positive numbers (colored red) indicate greater relative gene expression anaerobically.

growth rates, ethanol production and altered metabolite pools (Fig. 1, 2, Table 1). Anaerobic *Z. mobilis* fermentations utilized glucose more rapidly and grew more quickly with concomitant increases in ethanol productivity and yield as compared to aerobic *Z. mobilis* cultures (Fig. 1). Both intra- and extracellular lactate was identified as being more abundant in aerobic stationary phase *Z. mobilis* (Table 1, Fig. 2). Our data were also consistent with previous reports [28,29] showing metabolites such as lactate, acetate, and acetoin were more abundant during aerobic fermentation (Table 1, Fig. 2). Ishikawa et al. [19] proposed a model to explain the effect of oxygen on the NADH reducing power pool for the conversion of glucose into ethanol as the major end-product. They suggest that under aerobic conditions NADH is limiting due to it being oxidized by the NADH oxidase and therefore unavailable for reduction of acetaldehyde to ethanol. These data combined with lower intracellular glucose and ED pathway intermediates (Table 1) would agree with such a model and previous reports [30,31]. While we did not obtain the data for NAD^+, NADH, ATP metabolites, altered redox balance with concomitant increases in other metabolites produced such as acetate, lactate, acetoin, and acetaldehyde in the presence of oxygen may account for lower ethanol production (Table 1, Fig. 2). We were able to observe differences in gene expression between exponential phase conditions by microarray, but not at levels considered significant using the Bonferroni multiple testing method ($p < 0.05$). Quantitative PCR assays showed greater sensitivity than array analyses but the majority of genes showed little difference in expression levels between aerobic and anaerobic conditions in early exponential phase. Each culture had more than doubled by the early exponential growth phase sampling point and this time point should have mitigated any potential inocula effects. The observation that there was an absence of a lag phase indirectly supports the idea that there was neither a large perturbation nor large regulatory response by the microbe upon inoculation into the fermentors (Fig. 1). The large differences in exponential and stationary phase transcriptomic profiles imply ZM4 is aerotolerant and oxygen affected the physiology of the cells leading to the buildup of metabolic byproducts, which ultimately led to greater expression differences in stationary phase.

Seo et al (2005) describe the *Z. mobilis* ZM4 (ATCC31821) genome as consisting of a single chromosome [16], which we utilized for probe design and microarray fabrication. The same *Z. mobilis* ZM4 (ATCC31821) strain utilized in this study contained plasmid DNA, which was in keeping with an earlier report describing the nomenclature and derivation of *Z. mobilis* strains [32]. Therefore, the array data in the present study may not fully represent the differences between aerobic and anaerobic conditions since these plasmid DNA sequences were unavailable. We used a multiplex array format with an average probe length of 36 nucleotides and were able to detect significantly differentially expressed genes. However, the use of shorter probes likely affected the sensitivity of the microarray analyses [33]. The proper identification of putative coding sequences and the use of systems biology tools will be important in identifying genome features that can be altered for improved ethanol production. The JGI's plans to sequence the genomes for additional *Zymomonas mobilis* strains, including strain ZM4 will likely improve gene prediction models and elucidate potential coding sequences present on plasmid DNA.

The effects of oxygen on *Z. mobilis* fermentation performance parameters and physiology have been examined in a number of previous studies [18-21,25,31,34-37]. However, this is the first report we are aware of that examines its effects on whole genome transcriptomic and metabolomic profiles of the bacterium under rigorously controlled conditions. The *Z. mobilis* strain CP4 *glf, zwf, edd* and *glk* genes for glucose uptake and utilization form an operon and coordinate gene expression appears to result from a complex pattern of mRNA degradation [38-40]. The ZM4 *glf, zwf, edd* and *glk* genes are similarly organized and all showed greater expression in stationary phase under anaerobic conditions, which was about 20, 9, 2, and 10-fold respectively more than that of aerobic condition, respectively. However *edd* differential expression did not meet our stringent significance criterion with a $-\log_{10}(p) = 5.2$. The *glf* gene encodes a glucose-facilitated diffusion transporter protein, which showed greater expression (approximately 20-fold) under anaerobic conditions even in the absence of detectable amounts of glucose in the medium at 26 h (Fig. 1). ZMO1859 showed 5-fold greater expression in the absence of oxygen. ZMO1859 is a putative OprB family carbohydrate-selective porin containing a pfam04966 (OprB family) conserved domain with an E-value score of $5e^{-64}$ and 27% identity to *Pseudomonas aeruginosa* outer membrane protein OprB [41]. *P. aeruginosa* OprB transports glucose, mannitol, glycerol, and fructose. The *E. coli* expression of pathways for hexose uptake and metabolism has been shown to be up-regulated in the response to oxygen deprivation [30]. A number of ED and pyruvate biosynthetic pathway genes (*gntK, edd, eda, gap*) were showed higher expression levels under anaerobic conditions at 26 h when compared to the equivalent aerobic time point, but at levels not considered significant (see Additional file 3). Previous physiological studies have indicated that enzymatic activities for glucokinase, glucose-6-phosphate dehydrogenase, glyceraldehyde-3-phosphate dehydrogenase, ATPase, pyruvate kinase and alcohol dehydrogenase are decreased when oxygen partial pressure increases [31]. The intracellular metabolomic data suggests glucose present in anaerobic cells rapidly becomes phosphor-

ylated, providing substrate for further glucose utilization (see Additional file 3). Expression of the *pdc* and *adhB* genes involved in *Z. mobilis* ethanol production was increased at least two-fold during anaerobic fermentation and was consistent with fermentation product and growth data (Fig. 1, 2, see Additional file 4).

These data suggest oxygen affects how carbon is utilized and partitioned in *Z. mobilis* under highly aerobic fermentations leading to the buildup of inhibitors such as acetaldehyde and acetate and to the accumulation of non-reducing sugars like trehalose that contribute to overall lower ethanol yields. The accumulation of acetate at the end of the aerobic fermentation was much more pronounced than seen in a number of previous studies, although the highly aerated fermentors with DOT measurement used in this study likely meant the culture conditions were quite different from previous batch fermentations without DOT measurement. In this study, we used an agitation rate of 700 rpm in conjunction with sparging air at 2.5 L/min to maintain fully aerobic conditions as previous experiments had demonstrated a lower agitation rate and this airflow rate was insufficient to maintain fully aerobic conditions as cell density increased (data not shown). Putative aldehyde dehydrogenase genes such as *ssdA* (ZMO1754) did not show differential expression in our study, but may potentially play a role in conversion of acetaldehyde to acetate in conjunction with possible reverse alcohol dehydrogenase reactions. However, much more detailed studies are required to understand the large accumulation of acetate under highly aerobic conditions.

Kalnenieks [11] reviewed the physiology of *Z. mobilis*, identified a number of unanswered questions about its energy metabolism, and determined that the structure and physiological role of the *Z. mobilis* respiratory chain remains to be fully elucidated. Kalnenieks et al [37] concluded that *Z. mobilis* contains a single NAD(P)H dehydrogenase encoded by *ndh* gene, and that its inhibition results in decreased respiration and improved aerobic growth. While we were not able to measure NAD, we did observe that *ndh* (ZMO1113) encoding NADH dehydrogenase was down-regulated in aerobic stationary phase cultures (see Additional file 4). A number of other respiratory genes were also down-regulated under these conditions. These included ZMO0022, ZMO1571, ZMO1572 and ZMO1844 encoding a putative Fe-S oxidoreductase, NADH dehydrogenase, cytochrome bd-type quinol oxidase subunits 1 and 2, and oxidoreductase genes, respectively (see Additional file 4). The stimulation of aerobic growth is proposed to be due to redirection of the NADH flux from respiration to ethanol synthesis resulting in less accumulation of toxic intermediates [11]. Several models have been proposed to account for the distribution of reducing equivalents between putative respiratory gene products and alcohol dehydrogenase reactions [11]. In one model both alcohol dehydrogenase isozymes catalyze ethanol synthesis and oxidation back to acetaldehyde and in another, the two isozymes operate in opposite directions. Under aerobic conditions, ethanol yield peaked around 13 h followed by net ethanol and acetaldehyde oxidation with concomitant increases in acetate and acetoin concentrations while approximately 0.66 mM of glucose remained in the medium at 26 h (Fig. 1, 2). We observed that expression of *adhB* (alcohol dehydrogenase II) in anaerobically cultured cells was approximately three-fold greater than when compared to aerobic cells at 26 h (see Additional file 4). *adhA* (alcohol dehydrogenase I) showed approximately 1.5-fold higher expression aerobically when compared to anaerobic cells, but did not meet our stringent criterion of a $-\log_{10}$(p-value) of 5.4. *Z. mobilis* contains two alcohol dehydrogenases, AdhB, an iron-containing alcohol dehydrogenase inactivated by oxygen and AdhA, a zinc-containing alcohol dehydrogenase that is resistant to oxidation [36]. Our *adhB* expression data were consistent with potential iron limitation and/or differential inactivation of AdhB under aerobic conditions.

We identified, for the first time to our knowledge, transcripts for the putative respiratory gene ZMO1814 (*rnfA*), encoding a putative NADH:ubiquinone oxidoreductase subunit that were expressed more greatly (3.3-fold) under aerobic conditions. The *Rhodobacter capsulatus rnf* gene products form a membrane complex involved in electron transfer to nitrogenase, however no physiological information is available for the *rnf* gene products of *Z. mobilis*, which does not fix nitrogen [42]. Although Sootsuwan et al. [42] demonstrate *Z. mobilis* has a cyanide-resistant terminal oxidase and could detect little difference between aerobic and anaerobic exponential phase ubiquinol oxidase activities, they documented significantly more ubiquinol oxidase activity in stationary phase cell membrane prepared anaerobically. We observed that expression of the *Z. mobilis* cytochrome bd-type quinol oxidase subunits genes (*cydA/B*) was slightly greater in stationary phase under anaerobic conditions. Trace amounts of oxygen, below detection limits, might have been present in anaerobic fermentors during stationary phase (see Additional file 1, 2). However, *E. coli cydAB* genes show greater expression under anaerobic conditions due to the change in negative supercoiling status and it has been suggested this may provide a mechanism for increasing cytochrome *bd* levels in response to environmental stress [43]. Sootsuwan et al. [42] proposed a number of possible physiological roles for a *Z. mobilis* respiratory chain including: controlling NADH levels to control ethanol production, reducing intracellular oxygen concentration and maintenance of NADH levels inside cells. The present study iden-

tified a number of putative genes involved in respiration as being expressed differentially depending upon conditions. However, it is clear much remains to be done to fully elucidate their function in redox balance and overall regulation. The contribution of individual genes in electron transfer and metabolic processes need to be followed up with targeted deletion mutant studies. The most surprising finding was the minimal impact of oxygen on the transcriptomic profiles in early exponential. Mutant studies would also be useful in assessing the physiology of early exponential growth phase aerobic *Z. mobilis* cells.

Ribitol was the most abundant differentially detected intracellular metabolite by GC-MS, at approximately 17-fold greater levels within aerobic cells compared to anaerobic cells (Table 1). The addition of various straight-chain alditols including ribitol has been reported to lead to the formation of novel *E. coli* phospholipids [44]. However, we did not conduct detailed phospholipid fatty acid (PLFA) analysis, nor target intact membranes. Ribitol and adenosine (latter was not considered significantly differentially expressed) are also both flavin adenine dinucleotide (FAD) components, which may reflect differences in redox potential and electron transfer between the two different conditions. Further studies are required to elucidate the role of ribitol and other metabolites like 4-hydroxybutanoate in *Z. mobilis* physiology and identify the unknown metabolites we detected. GC-MS based metabolomics provides a useful platform to examine a broad range of metabolites such as sugars, sugar alcohols, sugar acids, organic acids, amino acids, fatty acids, sterols, secondary metabolites, phenolic glycosides, alkaloids, purines, pyrimidines, and nucleosides [22,23,45,46]. However, a number of classes of metabolites cannot be detected by GC-MS and ribitol may not be the largest metabolite difference between these states. The differences of intracellular lactate indicated this metabolite was more abundant under aerobic conditions (Table 1), which was consistent with levels detected in the medium supernatant (Fig. 2).

Several ZM4 genes related to oxidative stress responses were up-regulated under aerobic, stationary phase conditions, including *ahpC* encoding alkyl hydroperoxide reductase, ZMO1097 encoding thioredoxin, and ZMO1118 encoding a glutathione S-transferase family protein (see Additional file 4). However, many systems involved in oxygen detoxification and redox homeostasis were not differentially expressed. There were no significant gene expression differences for classical oxidative stress response genes such as peroxidase (ZMO1573), catalase (ZMO0928), and iron-dependent superoxide dismutase (ZMO1060) when aerobic cultures were compared to anaerobic cultures. The glutathione system related genes such as *gor* (ZMO1211), *gshB* (ZMO1913) and ZMO1556 were also not affected significantly (data not shown). *Pyrococcus furiosus* exposed to gamma irradiation, which generates hydroxyl radicals and oxidative stress showed many systems involved in oxygen detoxification and redox homeostasis appeared to be constitutively expressed [47]. Furthermore, genes encoding superoxide dismutase, catalase, nonspecific peroxidases or thioredoxin reductase were not significantly expressed in response to air exposure in the obligate anaerobe *Methanosarcina barkeri* [48]. The *Z. mobilis* transcriptome analyses also suggested a relationship between aerobic conditions (or metabolic end-products) and cellular iron requirement (or limitation) as indicated by higher levels of expression of genes involved in iron sequestration and uptake. Transcripts for *Z. mobilis* iron-uptake and scavenging genes including *pbuA* (ZMO0188), *feoB* (ZMO1541), ZMO1463, and ZMO1847 (see Additional file 5) were all more abundant under aerobic conditions in stationary phase. A number of other studies such as those described by Brown et al [26] and Williams et al [47] have shown linkages between oxidative stress and expression of genes related to iron-uptake and storage.

The *Z. mobilis* ZMO1107 and ZMO0347 genes were up-regulated under anaerobic conditions (see Additional file 4) and are annotated as transcriptional regulator and hypothetical proteins, respectively [16]. BLASTP analyses indicated that ZMO1107 is similar to the *E. coli* global regulator leucine-responsive regulatory protein (Lrp) and ZMO0347 is similar to the *E. coli* global regulator Hfq, sharing about 40% and 60% identity to the *E. coli* Lrp and Hfq proteins, respectively. *E. coli* Lrp regulates transcription of many Lrp regulon genes with leucine as a co-regulator and also acts as a determinant of chromosome structure to coordinate cellular metabolism with the environmental nutritional state [49,50]. We observed different levels of intracellular leucine in the present study, but not at highly significant levels (Table 1). The best-studied Lrp family regulator is the *E. coli* Lrp, which affects the expression of at least 10% of all *E. coli* genes. Microarray analysis of gene expression profiles for *E. coli* Lrp^+ and Lrp^- strains indicates more than 400 genes are significantly Lrp-responsive with expression for 147 genes lower in Lrp^+ compared to Lrp^- cells [50,51]. In addition, Lrp has been suggested to play an important role in *E. coli* stationary phase regulation. Out of the 200 genes induced upon entrance into stationary phase, Lrp affects nearly three-quarters of them in *E. coli* [50]. It is well known that *E. coli* RpoS alternative sigma factor controls stationary phase gene expression [52,53]. The genome sequence shows that *Z. mobilis* has *rpoH* (ZMO0749), *rpoE* (ZMO1404), *rpoN* (ZMO0274), and *fliA* (ZMO0626, sigma F) genes but no identifiable *rpoS* gene [16]. ZMO1107 may play a role in stationary phase and stress regulation in *Z. mobilis*, but further studies are required to determine what func-

tional overlap, if any, this protein has with RpoS-like functions in this bacterium.

A putative *Z. mobilis* Hfq global regulator was induced during anaerobic fermentation (Table 2). Hfq is a bacterial member of the Sm family of RNA-binding proteins, which acts by base-pairing with target mRNAs and functions as a chaperone for non-coding small RNA (sRNA) in *E. coli* [54-56]. Hfq is involved in regulating various processes and deletion of *hfq* has pleiotropic phenotypes, including slow growth, osmosensitivity, increased oxidation of carbon sources, and altered patterns of protein synthesis in *E. coli* [54,57]. *E. coli* Hfq has also been reported to affect genes involved in amino acid biosynthesis, sugar uptake, metabolism and energetics [58]. The expression of thirteen ribosomal genes was down-regulated in *hfq* mutant background in *E. coli* [58]. Hfq also up-regulated sugar uptake transporters and enzymes involved in glycolysis and fermentation such as *pgk* and *pykA*, and *adhE* [58]. Hfq is also involved in regulation of general stress responses that are mediated by alternative sigma factors such as RpoS, RpoE and RpoH. An *hfq* null mutant of *E. coli* exhibits strongly reduced RpoS levels by impairing *rpoS* translation [59]. Cells lacking Hfq induce the RpoE-mediated envelope stress response and *rpoH* is also induced in cells lacking Hfq in *E. coli* [58]. In our study, Hfq was less abundant in aerobic fermentation condition in ZM4 at 26 h post-inoculation (see Additional file 4) with eight ribosomal protein genes including *rplF*, *rplY*, and *rpmE* down-regulated (see Additional file 4). Meanwhile, glucose uptake genes such as *glf* and *oprB*, ED pathway genes such as *glk*, *zwf*, *pgl*, *eno*, and ethanol fermentation genes *pdc* and *adhB* were less abundant under aerobic condition in *Z. mobilis* (see Additional file 4). In addition, our microarray and qPCR data indicate that *rpoH* is also induced in aerobic condition (see Additional file 6) with low abundance of *hfq* transcripts (see Additional file 4). The association of Hfq with the expression of alternative sigma factor RpoH, ribosomal proteins and glycolytic enzyme in *Z. mobilis* is similar to that of *E. coli* as discussed above, which may indicate a similar conserved role of Hfq in both *E. coli* and *Z. mobilis*. However, further studies such as mutagenesis are required to elucidate the *Z. mobilis* Hfq regulon.

Conclusion

Our study has provided insights into transcriptomic and metabolic profiles of the model ethanogenic bacterium *Z. mobilis* during aerobic and anaerobic fermentation under controlled fermentation conditions for the first time. It is unlikely the oxygen levels we used in the present study would be as high in an industrial setting, even in a major perturbation to the fermentation. However, our study showed oxygen affects maintenance energy, how carbon is utilized and partitioned in *Z. mobilis* fermentations and can lead to the buildup of inhibitors such as acetaldehyde and acetate under highly aerobic conditions, which contributed to lower ethanol yields. The most surprising finding was that while high oxygen concentrations resulted in slower growth, there seemed to be minimal impact on early exponential growth phase transcriptomic profiles. However, differences in energy-spilling pathways due to oxygen and uncoupled growth later led to large differences in ethanol yield and transcriptomic profiles between anaerobic and aerobic stationary phase cultures. Our study also identified a range of genes such as *cydA/B*, *rpoH* or ZMO0347 that could be targeted for deletion to better understand *Z. mobilis* physiology and coordinate regulation. Future studies examining the transcriptomic profiles for aerobic, microoxic and anaerobic steady states, the dynamics of transition between steady states and the use of mutant strains are warranted to provide greater insights into *Z. mobilis* physiology and gene regulation. Finally, our microarray based on the TIGR annotation has identified a number of genes that were not originally annotated as coding sequences in 2005 when the first *Z. mobilis* genome sequence was published [16], which included about 9% of the aerobic up-regulated genes (see Additional file 4, 5). The rapid accumulation of genome sequence information and sequencing technology advances make improving gene coding models and annotations even more important.

Methods

Bacterial strains and fermentation conditions

Z. mobilis ZM4 was obtained from the American Type Culture Collection (ATCC31821) and cultured in RM medium [60] at 30°C. For the inoculum preparation a single colony of ZM4 was added to a test tube containing 5 mL RM broth and cultured aerobically at 30°C until it reached late exponential or early stationary phase. A 1/100 dilution was added into the pre-warmed RM broth (10 mL culture into 1000 mL RM), which was then cultured aerobically at 30°C with shaking at 150 rpm for approximately 12 h. The optical density was measured with a spectrophotometer at 600_{nm} and the inoculum was added to each fermentor so that the initial $OD600_{nm}$ was approximately 0.17 in each fermentor. Batch fermentations were conducted in approximately 2.5 L of RM medium in 7.5-L BioFlo110 bioreactors (New Brunswick Scientific, Edison, NJ) fitted with agitation, pH, temperature and DOT probes and controls. Culture pH was monitored using a pH electrode (Mettler-Toledo, Columbus, OH) and the pH control set point was maintained at 6.0 by automatic titration with 3 N KOH. Temperature was maintained automatically at 30°C and the vented gases exiting fermentors were passed through condenser units, chilled by a NESLAB Merlin M-150 refrigerated recirculator (Thermo Fisher Scientific, Newington, NH) to a vented hood via a water trap. DOT was monitored by using InPro 6800 series polarographic O_2 sensors (Mettler-Toledo). Three anaerobic fermentors were sparged overnight with

Table 2: Primer pairs used for q-PCR analysis with target gene information.

Primers[1]	Sequences (5' to 3')	PCR product size (bp)
ZMO0084_F	TGCAAGCATTGCCTACAAAG	100
ZMO0084_R	CCATTGAGGTGAACCCATCT	
ZMO0178_F	TGATCATCGGTGGTGGTATG	120
ZMO0178_R	TTTCAGCAGCAGCGAAAATA	
ZMO0367_F	ACAGCCTGATGAAACCATCC	111
ZMO0367_R	AACACATCCGTGAGGGAAAG	
ZMO0369_F	GCGTTTCTCTATTGCGGAAG	110
ZMO0369_R	CGAAACGTTCCCAAGCTAAC	
ZMO0749_F	TATCCTGCGGTCTTGGAGTC	108
ZMO0749_R	CCGTCTTCAAAGGCATTCAT	
ZMO0817_F	GCACCGAAATCAGCAAAAAT	107
ZMO0817_R	GTGTTGGGACTGGGTTCATC	
ZMO0976_F	CAGCAGATGGACGAGTTCAA	113
ZMO0976_R	TCGTTTTTCTTGGCATAGGG	
ZMO1062_F	AGGCCAATGACGGTTTACAG	115
ZMO1062_R	GCTTCCTGATCCAAAAGCTG	
ZMO1118_F	CGCCTGTTATTCTGGTGGAT	109
ZMO1118_R	CGCCTTATTTCAGCCCTATG	
ZMO1129_F	AGTGAAACCGACTGGCCTAA	104
ZMO1129_R	ATGGTTTCAATCGCAGCTCT	
ZMO1360_F	TGGTCTCAAGCATCACTTCG	114
ZMO1360_R	CCGCAGTTCAGTTCGTTACA	
ZMO1478_F	AGTAGCGGTGCTGAACTGGT	109
ZMO1478_R	CGGCAGGCCATTTATAAGAA	
ZMO1608_F	ACATGGCTAACGACGCTTCT	111
ZMO1608_R	TCTTGATCTGACCGCAGTTG	
ZMO1660_F	ATGAGCGTCCAGCAATTCTT	102
ZMO1660_R	TTCCCCGACATGACTCAGTA	
ZMO1863_F	GGAACTCTGGCAGAAACAGC	111
ZMO1863_R	GGATAACCCAAGACGAGCAA	
ZMO1876_F	ATCAGGATTTGACGCTGGAA	100
ZMO1876_R	ACCATCGCTTCGACAATAGC	
ZMO1959_F	ACAAGGCTGCCGATTTATTG	104
ZMO1959_R	TTGGCTCAGCAGATGTTGTC	

[1]PCR primer pair nomenclature. F: forward primer, R: reverse primer for each respective gene.

filter-sterilized N_2 gas and for approximately one hour post-inoculation and the three aerobic fermentors were continually sparged with filter-sterilized air at 2.5 L/min to maintain fully aerobic conditions. The agitation rate was 700 rpm in each vessel.

Growth, glucose and fermentation product analysis

Growth was monitored turbidometrically by measuring optical density at 600_{nm} with a model 8453 spectrophotometer (Hewlett-Packard, Palo Alto, CA.). Fermentation media and fermentation products from filter-sterilized

cell-free spent media were compositionally analyzed by either both the Amplex red glucose assay kit (Invitrogen, Carlsbad, CA) and high-performance liquid chromatography (HPLC) for glucose or by gas chromatography and/or HPLC for ethanol, acetate, lactate, acetaldehyde, and acetoin determinations. Glucose assays were incubated at room temperature for 30 minutes and performed according the manufacturer's instructions. Absorbance was detected using a microplate reader (model HTS 7000, PerkinElmer, Waltham, MA) at 590_{nm}. Background absorbance was corrected by subtracting the value derived from the no-glucose control.

Gas chromatography (GC)

Ethanol, acetate, acetaldehyde and acetoin concentrations in the medium supernatant were determined by flame ionization gas chromatography. Culture samples (1 mL) and standards were prepared by filtration and acidification with hydrochloric acid. The samples and standards were quantified by injecting 1 μL of each into a model 6890 Agilent Technologies equipped with a DB-FFAP 30 m × 0.53 mm × 1.5 μm film thickness capillary column (Agilent, Santa Clara, CA). The column operated with an initial temperature of 60°C and ramping 10°C to a final temperature of 180°C, while detector was at 250°C and injector temperature was 130°C with a post-injection dwell time of one minute. The carrier gas was helium at a constant flow rate of 5 mL/min.

High-performance liquid chromatography (HPLC)

HPLC analysis was used for the measurements of the concentration of glucose, acetate, ethanol, and lactate in the 0.2 μm-filtered samples taken at different time points during fermentation. The diluted fermentation samples (1:1 with 8.98 mM sulfuric acid) were separated and quantified by HPLC using a LaChrom Elite System (Hitachi High Technologies America, Inc., San Jose, CA). Analysis was performed with an oven (Model L-2350) set at 60°C, and a pump (Model L-2130) set with a flow rate of 0.5 mL/min in 5 mM H_2SO_4. The run time for each sample was set for 35 minutes (Injector Model L-2200). Eluted compounds were registered and quantified by a refractive index detector (Model L-2490) equipped with a computer-powered integrator. Soluble fermentation products were identified by comparison with retention times and peak areas of corresponding standards. Metabolites were separated on an Aminex HPX-87H, 300 × 7.8 mm column (Bio-Rad, Hercules, CA).

RNA isolation and preparation of fluorescein-labeled cDNA

RNA was isolated essentially described previously [26]. Briefly, samples from aerobic and anaerobic fermentors were harvested by centrifugation and the TRIzol reagent (Invitrogen, Carlsbad, CA) was used to extract total cellular RNA. Each total RNA preparation was treated with RNase-free DNase I (Ambion, Austin, TX) to digest residual chromosomal DNA and subsequently purified with the Qiagen RNeasy Mini kit in accordance with the instructions from the manufacturer. Total cellular RNA was quantified at OD_{260} and OD_{280} with a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE). The purified RNA from each sample was used as the template to generate cDNA copies labeled with either Cy3-dUTP or Cy5-dUTP (GE Healthcare Bio-Sciences Corp, Piscataway, NJ). In a duplicate set of cDNA synthesis reactions the fluorescent dyes were reversed for each sample so that the effects of a specific dye were minimized. Labeling reaction components and incubation conditions as well as cDNA probe purification and concentration determination have been described previously [26].

Microarray hybridization, scanning, image quantification, and data analyses

Z. mobilis microarrays were constructed by CombiMatrix Corporation (Mukilteo, WA) using coding sequences predicted by The Institute for Genomic Research (TIGR, http://www.tigr.org/) as 101 more genes were predicted in this later genome annotation. Both the primary and TIGR annotations are presented in the present study. Microarray hybridizations were performed according to the manufacturer's instructions. Briefly, gene expression analysis was performed using six independent microarray experiments (two dye reversal reactions × three biological replicates) with each microarray containing one to two probes per predicted coding sequence each. The two separately labeled cDNA pools (i.e., the aerobic and the corresponding anaerobic time point) to be compared were mixed together in a hybridization solution containing 50% (v/v) formamide, applied to microarrays and incubated overnight (16 h) at 50°C. Microarrays were washed using buffers of increasing stringency according to the manufacturer's instructions, scanned with a ScanArray 5000 Microarray Analysis System (PerkinElmer Life Sciences Inc, Boston, MA) and the images were quantified using Microarray Imager software (CombiMatrix). Raw data was $\log_2$ transformed and imported into the statistical analysis software JMP Genomics 3.0 (SAS Institute, Cary, NC). A distribution analysis and data correlation analysis were done as a quality control step. The overlayed kernel density estimates derived from the distribution analysis allowed the visualization of sources of variation based on strain, as well as variation attributed to technical factors such as array and dye. The data were subsequently normalized using the standard normalization algorithm within JMP Genomics. An analysis of variance (ANOVA) was performed to determine differential expression levels

between conditions and time points using the Bonferroni multiple testing method ($p < 0.05$).

Quantitative-PCR (qPCR) analysis

Microarray data were validated using real-time qPCR as described previously [26]. Seventeen genes representing different functional categories and the range gene expression values, based on microarray hybridizations were analyzed using qPCR from cDNA derived from stationary phase samples. Primer pairs were designed as described previously, and the oligonucleotide sequences of the seventeen genes selected for qPCR analysis are listed in Table 2.

Gas chromatography-mass spectrometry (GC-MS) metabolite analysis

Culture samples were rapidly pelleted by centrifugation, supernatants removed, cell pellets snap-frozen in liquid nitrogen and then stored at -80°C until analysis. Analyses were performed on microbial pellets collected in the stationary phase. A 5 mL aliquot of 80% ethanol (aqueous) was added to each pellet and cells disrupted using a sonicator 3000 (Misonix, Inc., Farmingdale, NY) operated at power level 5 for 6 times with a 1 min processing time and 1 min interval among each processing. An internal standard of 200 μL of sorbitol (1 mg/mL aqueous solution) was then added to each tube and 2 mL aliquots then dried in a helium stream. The internal standard was added to correct for differences in derivatization efficiency and changes in sample volume during heating. Dried exudates were dissolved in 500 μL of silylation-grade acetonitrile followed by the addition of 500 μL N-methyl-N-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (Pierce Chemical Co., Rockford, IL), and samples then heated for 1 h at 70°C to generate trimethylsilyl (TMS) derivatives. After 1 day, 1-μL aliquots were injected into a ThermoFisher DSQII GC-MS, fitted with an Rtx-5MS (crosslinked 5% PH ME Siloxane) 30 m × 0.25 mm × 0.25 μm film thickness capillary column (Restek, Bellefonte, PA). The standard quadrupole GC-MS was operated in electron impact (70 eV) ionization mode, with 6 full-spectrum (70–650 Da) scans per second. Gas (helium) flow was set at 1.1 mL per minute with in injection port configured in the splitless mode. The injection port and detector temperatures were set to 220°C and 300°C, respectively. The initial oven temperature was held at 50°C for 2 min and was programmed to increase at 20°C per min to 325°C and held for another 11.25 min, before cycling back to the initial conditions. Quantified metabolites of interest were extracted using a key selected m/z that was characteristic for each metabolite, rather than the total ion chromatogram, to minimize integration of co-eluting metabolites. Peaks were quantified by area integration and the concentrations were normalized to the quantity of the internal standard (sorbitol) recovered, amount of sample extracted, derivitized, and injected. Two technical replicates were analyzed for two biological samples from each condition. Metabolite data of ZM4 under aerobic and anaerobic condition were averaged and presented as relative responses between ZM4 under aerobic fermentation versus ZM4 under anaerobic fermentation.

Authors' contributions

SDB designed the microarray and conceived the experiment. SY and SDB performed the fermentation and sample collection. SY carried out the RNA extraction, microarray, qPCR, sample preparation for HPLC and GC-MS. MRJ performed HPLC analysis. TJT and NLE performed the GC-MS analysis. SLC performed GC-MS analysis. SLM assisted with microarray data analysis. SY and SDB analyzed the data and wrote the manuscript. BHD and AVP provided input on *Z. mobilis* metabolism and manuscript revision.

Additional material

Additional file 1

Summary of fermentor parameters during 26 h fermentations. *Mean (± S.D.) agitation (rpm), temperature (°C), pH and dissolved oxygen tension values for three aerobic fermentors and three anaerobic fermentors averaged over the entire experiment.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S1.doc]

Additional file 2

Dissolved oxygen tension during* Z. mobilis *fermentations. *Mean dissolved oxygen tension data for three aerobic fermentors and three anaerobic fermentors over 26 h. The bars represent the standard error of the mean data for each condition.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S2.ppt]

Additional file 3

Entner-Dondoroff and pyruvate metabolic pathways showing metabolomic and transcriptomic data at 26 h. *Summary of transcriptomic and metabolomic profiling data between aerobic and anaerobic conditions at 26 h.*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S3.ppt]

Additional file 4

Aerobic down-regulated genes 26 h post inoculation. *Expression profiles for significantly differentially expressed genes that were down-regulated under aerobic conditions at 26 h as detected by microarrays and real-time qPCR. The modified gene function categories are based on MultiFun categories [Serres MH and Riley M:* Microb Comp Genomics *2000, 5(4):18].*

Click here for file

[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S4.doc]

Additional file 5

Aerobic up-regulated genes 26 h post inoculation. *Expression profiles for significantly differentially expressed genes that were up-regulated under aerobic conditions at 26 h as detected by microarrays and real-time qPCR. The modified gene function categories are based on MultiFun categories [Serres MH and Riley M:* Microb Comp Genomics *2000, 5(4):18].*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S5.doc]

Additional file 6

Comparison of exponential growth phase gene expression measurements by microarray and qPCR. *The gene expression ratios for wild-type Z.* mobilis *ZM4 under aerobic and anaerobic conditions after 3 h fermentation were log transformed in base 2 (log_2<aerobic/anaerobic>). The microarray log_2 ratio values (log_2<aerobic/anaerobic>) were plotted against the qPCR log_2 values. Comparison of the two methods indicated a level of concordance of R = 0.62.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S6.ppt]

Additional file 7

Comparison of stationary growth phase gene expression measurements by microarray and qPCR. *The gene expression ratios for wild-type Z.* mobilis *ZM4 under aerobic and anaerobic conditions after 26 h fermentation were log transformed in base 2 (log_2<aerobic/anaerobic>). The microarray log_2 ratio values (log_2<aerobic/anaerobic>) were plotted against the qPCR log_2 values. Comparison of the two methods indicated a high level of concordance (R = 0.92).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2164-10-34-S7.ppt]

Acknowledgements

This work is sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL), managed by UT-Battelle, LLC for the U. S. Department of Energy under Contract No. DE-AC05-00OR22725. The BioEnergy Science Center is a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

References

1. Farrell AE, Plevin RJ, Turner BT, Jones AD, O'Hare M, Kammen DM: **Ethanol can contribute to energy and environmental goals.** *Science* 2006, **311(5760):**506-508.
2. Hahn-Hagerdal B, Galbe M, Gorwa-Grauslund MF, Liden G, Zacchi G: **Bio-ethanol – the fuel of tomorrow from the residues of today.** *Trends Biotechnol* 2006, **24:**549-556.
3. Koonin SE: **Getting serious about biofuels.** *Science* 2006, **311:**435.
4. Ragauskas AJ, Williams CK, Davison BH, Britovsek G, Cairney J, Eckert CA, Frederick WJ, Hallett JP, Leak DJ, Liotta CL, Mielenz JR, Murphy R, Templer R, Tschaplinski T: **The path forward for biofuels and biomaterials.** *Science* 2006, **311:**484-489.
5. Dien BS, Cotta MA, Jeffries TW: **Bacteria engineered for fuel ethanol production: current status.** *Appl Microbiol Biotechnol* 2003, **63:**258-266.
6. Rogers PL, Jeon YJ, Lee KJ, Lawford HG: ***Zymomonas mobilis* for fuel ethanol and higher value products.** *Adv Biochem Eng Biotechnol* 2007, **108:**263-288.
7. Romero S, Merino E, Bolívar F, Gosset G, Martinez A: **Metabolic engineering of *Bacillus subtilis* for ethanol production: lactate dehydrogenase plays a key role in fermentative metabolism.** *Appl Environ Microbiol* 2007, **73:**5190-5198.
8. Inui M, Kawaguchi H, Murakami S, Vertès AA, Yukawa H: **Metabolic engineering of *Corynebacterium glutamicum* for fuel ethanol production under oxygen-deprivation conditions.** *J Mol Microbiol Biotechnol* 2004, **8:**243-254.
9. Reisch MS: **Fuels of the future chemistry and agriculture join to make a new generation of renewable fuels.** *Chem Eng News* 2006:3.
10. Doelle HW, Kirk L, Crittenden R, Toh H, Doelle MB: ***Zymomonas mobilis* – science and industrial application.** *Crit Rev Biotechnol* 1993, **13:**57-98.
11. Kalnenieks U: **Physiology of *Zymomonas mobilis*: some unanswered questions.** *Adv Microb Physiol* 2006, **51:**73-117.
12. Swings J, De Ley J: **The biology of *Zymomonas*.** *Bacteriol Rev* 1977, **41:**1-46.
13. Deanda K, Zhang M, Eddy C, Picataggio S: **Development of an arabinose-fermenting *Zymomonas mobilis* strain by metabolic pathway engineering.** *Appl Environ Microbiol* 1996, **62(12):**4465-4470.
14. Mohagheghi A, Evans K, Chou YC, Zhang M: **Cofermentation of glucose, xylose, and arabinose by genomic DNA-integrated xylose/arabinose fermenting strain of *Zymomonas mobilis* AX101.** *Appl Biochem Biotech* 2002, **98–100:**885-898.
15. Zhang M, Eddy C, Deanda K, Finkelstein M, Picataggio S: **Metabolic engineering of a pentose metabolism pathway in ethanologenic *Zymomonas mobilis*.** *Science* 1995, **267:**240-243.
16. Seo JS, Chong HY, Park HS, Yoon KO, Jung C, Kim JJ, Hong JH, Kim H, Kim JH, Kil JI, Park CJ, Oh HM, Lee JS, Jin SJ, Um HW, Lee HJ, Oh SJ, Kim JY, Kang HL, Lee SY, Lee KJ, Kang HS: **The genome sequence of the ethanologenic bacterium *Zymomonas mobilis* ZM4.** *Nature Biotechnol* 2005, **23:**63-68.
17. Jeffries TW: **Ethanol fermentation on the move.** *Nature Biotechnol* 2005, **23:**40-41.
18. Bringer SFR, Sahm H: **Effect of oxygen on the metabolism of *Zymomonas mobilis*.** *Arch Microbiol* 1984, **139:**376-381.
19. Ishikawa H, Nobayashi H, Tanaka H: **Mechanism of fermentation performance of *Zymomonas mobilis* under oxygen supply in batch culture: fermentation performance of *Zymomonas mobilis* against oxygen supply (II).** *J Ferment Bioeng* 1990, **70:**34-40.
20. Pankova LM, Shvinka JE, Beker MJ: **Regulation of intracellular H+ balance in *Zymomonas mobilis* 113 during the shift from anaerobic to aerobic conditions.** *Appl Microbiol Biotech* 1988, **28:**583-588.
21. Kalnenieks U, Galinina N, Toma MM, Poole RK: **Cyanide inhibits respiration yet stimulates aerobic growth of *Zymomonas mobilis*.** *Microbiology* 2000, **146:**1259-1266.
22. Mashego MR, Rumbold K, De Mey M, Vandamme E, Soetaert W, Heijnen JJ: **Microbial metabolomics: past, present and future methodologies.** *Biotech Lett* 2007, **29:**1-16.
23. Oldiges M, Lutz S, Pflug S, Schroer K, Stein N, Wiendahl C: **Metabolomics: current state and evolving methodologies and tools.** *Appl Microbiol Biotech* 2007, **76:**495-511.
24. Trauger SA, Kalisak E, Kalisiak J, Morita H, Weinberg MV, Menon AL, Poole FL 2nd, Adams MW, Siuzdak G: **Correlating the transcriptome, proteome, and metabolome in the environmental adaptation of a hyperthermophile.** *J proteome Res* 2008, **7:**1027-1035.
25. Belauich JP, Senez JC: **Influence of aeration and of pantothenate on growth yields of *Zymomonas mobilis*.** *J Bacteriol* 1965, **89:**1195-1200.
26. Brown SD, Thompson MR, Verberkmoes NC, Chourey K, Shah M, Zhou J, Hettich RL, Thompson DK: **Molecular dynamics of the *Shewanella oneidensis* response to chromate stress.** *Mol Cell Proteomics* 2006, **5:**1054-1071.
27. Kogoma T, Yura T: **Sensitization of *Escherichia coli* cells to oxidative stress by deletion of the *rpoH* gene, which encodes the heat shock sigma factor.** *J Bacteriol* 1992, **174:**630-632.
28. Lawford HG, Rousseau JD: **Steady-state measurements of lactic acid production in a wild-type and a putative D-lactic acid dehydrogenase-negative mutant of *Zymomonas mobilis*: influence of glycolytic flux.** *Appl Biochem Biotech* 2002, **98–100:**215-228.

29. Tsantili IC, Karim MN, Klapa MI: **Quantifying the metabolic capabilities of engineered *Zymomonas mobilis* using linear programming analysis.** *Microb Cell Fact* 2007, **6:**8.
30. Schramm G, Zapatka M, Eils R, Konig R: **Using gene expression data and network topology to detect substantial pathways, clusters and switches during oxygen deprivation of *Escherichia coli*.** *BMC Bioinformatics* 2007, **8:**149.
31. Toh H, Doelle H: **Changes in the growth and enzyme level of *Zymomonas mobilis* under oxygen-limited conditions at low glucose concentration.** *Arch Microbiol* 1997, **168:**46-52.
32. Yablonsky MD, Goodman AE, Stevnsborg N, Delima OG, Demorais JOF, Lawford HG, Rogers PL, Eveleigh DE: ***Zymomonas mobilis* CP4 – a clarification of strains via plasmid profiles.** *J Biotechnol* 1988, **9:**71-79.
33. He ZL, Wu LY, Fields MW, Zhou JZ: **Use of microarrays with different probe sizes for monitoring gene expression.** *Appl Environ Microbiol* 2005, **71:**5154-5162.
34. Tanaka H, Ishikawa H, Osuga K, Takagi Y: **Fermentation ability of *Zymomonas mobilis* under various oxygen-supply conditions in batch culture.** *J Ferment Bioeng* 1990, **69:**234-239.
35. Moreau RA, Powell MJ, Fett WF, Whitaker BD: **The effect of ethanol and oxygen on the growth of *Zymomonas mobilis* and the levels of hopanoids and other membrane lipids.** *Curr Microbiol* 1997, **35:**124-128.
36. Tamarit J, Cabiscol E, Aguilar J, Ros J: **Differential inactivation of alcohol dehydrogenase isoenzymes in *Zymomonas mobilis* by oxygen.** *J Bacteriol* 1997, **179:**1102-1104.
37. Kalnenieks U, Galinina N, Strazdina I, Kravale Z, Pickford JL, Rutkis R, Poole RK: **NADH dehydrogenase deficiency results in low respiration rate and improved aerobic growth of *Zymomonas mobilis*.** *Microbiology* 2008, **154:**989-994.
38. Barnell WO, Liu J, Hesman TL, O'Neill MC, Conway T: **The *Zymomonas mobilis glf, zwf, edd*, and *glk* genes form an operon: localization of the promoter and identification of a conserved sequence in the regulatory region.** *J Bacteriol* 1992, **174:**2816-2823.
39. Burnett ME, Liu J, Conway T: **Molecular characterization of the *Zymomonas mobilis* enolase (*eno*) gene.** *J Bacteriol* 1992, **174(20):**6548-6553.
40. Liu J, Barnell WO, Conway T: **The polycistronic mRNA of the *Zymomonas mobilis glf-zwf-edd-glk* operon is subject to complex transcript processing.** *J Bacteriol* 1992, **174:**2824-2833.
41. Wylie JL, Worobec EA: **The OprB porin plays a central role in carbohydrate uptake in *Pseudomonas aeruginosa*.** *J Bacteriol* 1995, **177:**3021-3026.
42. Sootsuwan K, Lertwattanasakul N, Thanonkeo P, Matsushita K, Yamada M: **Analysis of the respiratory chain in ethanologenic *Zymomonas mobilis* with a cyanide-resistant bd-type ubiquinol oxidase as the only terminal oxidase and its possible physiological roles.** *J Mol Microbiol Biotechnol* 2008, **14:**163-175.
43. Bebbington KJ, Williams HD: **A role for DNA supercoiling in the regulation of the cytochrome bd oxidase of *Escherichia coli*.** *Microbiology* 2001, **147:**591-598.
44. Shibuya I, Yamagoe S, Miyazaki C, Matsuzaki H, Ohta A: **Biosynthesis of novel acidic phospholipid analogs in *Escherichia coli*.** *J Bacteriol* 1985, **161:**473-477.
45. Kopka J: **Current challenges and developments in GC-MS based metabolite profiling technology.** *J Biotechnol* 2006, **124:**312-322.
46. Schauer N, Steinhauser D, Strelkov S, Schomburg D, Allison G, Moritz T, Lundgren K, Roessner-Tunali U, Forbes MG, Willmitzer L, Fernie AR, Kopka J: **GC-MS libraries for the rapid identification of metabolites in complex biological samples.** *FEBS Lett* 2005, **579:**1332-1337.
47. Williams E, Lowe TM, Savas J, DiRuggiero J: **Microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus* exposed to gamma irradiation.** *Extremophiles* 2007, **11:**19-29.
48. Zhang W, Culley DE, Nie L, Brockman FJ: **DNA microarray analysis of anaerobic *Methanosarcina barkeri* reveals responses to heat shock and air exposure.** *J Ind Microbiol Biotechnol* 2006, **33:**784-790.
49. Newman EB, Lin R: **Leucine-responsive regulatory protein: a global regulator of gene expression in *E. coli*.** *Annu Rev Microbiol* 1995, **49:**747-775.
50. Tani TH, Khodursky A, Blumenthal RM, Brown PO, Matthews RG: **Adaptation to famine: a family of stationary-phase genes revealed by microarray analysis.** *Proc Natl Acad Sci USA* 2002, **99:**13471-13476.
51. Hung SP, Baldi P, Hatfield GW: **Global gene expression profiling in *Escherichia coli* K12. The effects of leucine-responsive regulatory protein.** *J Biol Chem* 2002, **277:**40309-40323.
52. Loewen PC, Hu B, Strutinsky J, Sparling R: **Regulation in the *rpoS* regulon of *Escherichia coli*.** *Can J Microbiol* 1998, **44:**707-717.
53. Patten CL, Kirchhof MG, Schertzberg MR, Morton RA, Schellhorn HE: **Microarray analysis of RpoS-mediated gene expression in *Escherichia coli* K-12.** *Mol Genet Genomics* 2004, **272:**580-591.
54. Valentin-Hansen P, Eriksen M, Udesen C: **The bacterial Sm-like protein Hfq: a key player in RNA transactions.** *Mol Microbiol* 2004, **51:**1525-1533.
55. Zhang A, Wassarman KM, Ortega J, Steven AC, Storz G: **The Sm-like Hfq protein increases OxyS RNA interaction with target mRNAs.** *Mol cell* 2002, **9:**11-22.
56. Zhang A, Wassarman KM, Rosenow C, Tjaden BC, Storz G, Gottesman S: **Global analysis of small RNA and mRNA targets of Hfq.** *Mol Microbiol* 2003, **50:**1111-1124.
57. Tsui HC, Leung HC, Winkler ME: **Characterization of broadly pleiotropic phenotypes caused by an *hfq* insertion mutation in *Escherichia coli* K-12.** *Mol Microbiol* 1994, **13(1):**35-49.
58. Guisbert E, Rhodius VA, Ahuja N, Witkin E, Gross CA: **Hfq modulates the sigmaE-mediated envelope stress response and the sigma32-mediated cytoplasmic stress response in *Escherichia coli*.** *J Bacteriol* 2007, **189:**1963-1973.
59. Muffler A, Fischer D, Hengge-Aronis R: **The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for *rpoS* translation in *Escherichia coli*.** *Genes Dev* 1996, **10:**1143-1151.
60. Goodman AD, Rogers PL, Skotnicki ML: **Minimal medium for isolation of auxotrophic *Zymomonas* mutants.** *Appl Environ Microbiol* 1982, **44:**496-498.

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 2009, p. 4762–4769 Vol. 75, No. 14
0099-2240/09/$08.00+0 doi:10.1128/AEM.00236-09

Efficient Degradation of Lignocellulosic Plant Biomass, without Pretreatment, by the Thermophilic Anaerobe "*Anaerocellum thermophilum*" DSM 6725▽

Sung-Jae Yang,[1,2]# Irina Kataeva,[1,2]# Scott D. Hamilton-Brehm,[2] Nancy L. Engle,[2] Timothy J. Tschaplinski,[2] Crissa Doeppke,[2,3] Mark Davis,[2,3] Janet Westpheling,[2,4] and Michael W. W. Adams[1,2]*

Departments of Biochemistry & Molecular Biology[1] and Genetics,[4] University of Georgia, Athens, Georgia 30602; BioEnergy Science Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831[2]; and National Renewable Energy Laboratory, Golden, Colorado 80401[3]

Received 31 January 2009/Accepted 15 May 2009

Very few cultivated microorganisms can degrade lignocellulosic biomass without chemical pretreatment. We show here that "*Anaerocellum thermophilum*" DSM 6725, an anaerobic bacterium that grows optimally at 75°C, efficiently utilizes various types of untreated plant biomass, as well as crystalline cellulose and xylan. These include hardwoods such as poplar, low-lignin grasses such as napier and Bermuda grasses, and high-lignin grasses such as switchgrass. The organism did not utilize only the soluble fraction of the untreated biomass, since insoluble plant biomass (as well as cellulose and xylan) obtained after washing at 75°C for 18 h also served as a growth substrate. The predominant end products from all growth substrates were hydrogen, acetate, and lactate. Glucose and cellobiose (on crystalline cellulose) and xylose and xylobiose (on xylan) also accumulated in the growth media during growth on the defined substrates but not during growth on the plant biomass. *A. thermophilum* DSM 6725 grew well on first- and second-spent biomass derived from poplar and switchgrass, where spent biomass is defined as the insoluble growth substrate recovered after the organism has reached late stationary phase. No evidence was found for the direct attachment of *A. thermophilum* DSM 6725 to the plant biomass. This organism differs from the closely related strain *A. thermophilum* Z-1320 in its ability to grow on xylose and pectin. *Caldicellulosiruptor saccharolyticus* DSM 8903 (optimum growth temperature, 70°C), a close relative of *A. thermophilum* DSM 6725, grew well on switchgrass but not on poplar, indicating a significant difference in the biomass-degrading abilities of these two otherwise very similar organisms.

Utilization of lignocellulosic biomass derived from renewable plant material to produce ethanol and other fuels is viewed as a major alternative to petroleum-based energy sources (19). The efficient conversion of plant biomass to fermentable sugars remains a formidable challenge, however, due to the recalcitrance of the insoluble starting materials (13, 21, 36). Thermal and chemical pretreatments must be used to solubilize and release the sugars, but such processes are costly and not very efficient (17, 28). Most pretreatments utilize acids, alkali, or organic solvents (39). Moreover, the plant feedstocks vary considerably in their compositions. The main components of plant biomass and the sources of the fermentable sugars, cellulose and hemicellulose, are combined with lignin, which can occupy 20% (wt/wt) or more of the plant cell wall. The development of technologies to efficiently degrade plant biomass therefore faces considerable obstacles. The discovery or engineering of new microorganisms with the ability to convert the components of lignocellulosic biomass into sugars is therefore of high priority.

Not many microorganisms are able to degrade pure crystalline cellulose, and the cellulose in plant biomass has a high order of crystallinity and is even less accessible to microbial or enzymatic attack (1, 12–14). Aerobic cellulolytic microorganisms usually secrete (hemi)cellulolytic enzymes containing carbohydrate-binding modules that serve to bind the catalytic domains to insoluble substrates. On the other hand, some anaerobic bacteria and fungi produce a large extracellular multienzyme complex called the cellulosome. This binds to and efficiently degrades cellulose and other polysaccharides, although it has a limited distribution in nature (3, 7). The rate at which microorganisms degrade cellulose increases dramatically with temperature (20), but the most thermophilic cellulosome-producing bacterium that has been characterized, *Clostridium thermocellum*, grows optimally near only 60°C (3, 9). A few anaerobic thermophiles are known that are able to grow on crystalline cellulose even though they lack cellulosomes, and in those cases the highest optimum growth temperature is 75°C (4, 32). Biomass conversion by thermophilic anaerobic microorganisms has many potential advantages over fermentation at lower temperatures. In particular, the organisms tend to have high rates of growth and metabolism, and the processes are less prone to contamination (30).

The gram-positive bacterium "*Anaerocellum thermophilum*" strain Z-1320 is among the most thermophilic of the cellulolytic anaerobes (32). It grows optimally at 75°C at neutral pH and utilizes both simple and complex polysaccharides, although it does not grow on xylose or pectin (32). The end

* Corresponding author. Mailing address: Departments of Biochemistry and Molecular Biology, Life Sciences Building, University of Georgia, Athens, GA 30602. Phone: (706) 542-2060. Fax: (706) 542-0229. E-mail: adams@bmb.uga.edu.
# These authors contributed equally to this work.
▽ Published ahead of print on 22 May 2009.

products of fermentation are lactate, ethanol, acetate, CO_2, and hydrogen. Although *A. thermophilum* Z-1320 grows very rapidly on crystalline cellulose (4), surprisingly, it has been studied very little since its discovery (32). We report here on the physiology of a very closely related strain, *A. thermophilum* DSM 6725, the genome of which was recently sequenced (16). The ability of *A. thermophilum* DSM 6725 to grow on different types of defined and complex substrates was investigated with a focus on switchgrass and poplar. These high-lignin plants have been selected as models for biomass-to-biofuel conversion by the BioEnergy Science Center (funded by the U.S. Department of Energy; http://bioenergycenter.org/). We show that *A. thermophilum* DSM 6725 is able to grow efficiently on both types of plant substrate without a chemical pretreatment step.

MATERIALS AND METHODS

Microorganisms. *Anaerocellum thermophilum* strain DSM 6725 was obtained from the DSMZ (http://www.dsmz.de/index.htm). *Caldicellulosiruptor saccharolyticus* DSM 8903 was a gift from Robert Kelly of North Carolina State University.

Growth substrates. The following growth substrates with the indicated sources were used: D-(+)-cellobiose (catalog no. C7252), D-(+)-xylose (X1500), oat spelt xylan (X0627), and pectin (P9135; all from Sigma, St. Louis, MO); Avicel PH-101 (catalog no. 11365; Fluka, Switzerland); poplar and switchgrass (sieved −20/+80-mesh fraction; Brian Davison, Oak Ridge National Laboratory, Oak Ridge, TN); and Tifton 85 Bermuda grass and napier grass (sieved, −20/+80-mesh fraction; Joy Peterson, Department of Microbiology, University of Georgia). Samples of plant biomass were used as received without chemical or physical treatments and are referred to as untreated biomass (or without pretreatment).

Growth medium. *A. thermophilum* DSM 6725 and *C. saccharolyticus* DSM 8903 were grown in 516 medium (32) except that vitamin and trace mineral solutions were modified as follows. The mineral solution contained the following (per liter): NH_4Cl, 0.33 g; KH_2PO_4, 0.33 g; KCl, 0.33 g; $MgCl_2 \cdot 6H_2O$, 0.33 g; $CaCl_2 \cdot 2H_2O$, 0.33 g; yeast extract, 0.5 g; resazurin, 0.5 mg; vitamin solution, 5 ml; trace minerals solution, 1 ml. The vitamin solution contained the following (in mg/liter): biotin, 4; folic acid, 4; pyridoxine-HCl, 20; thiamine-HCl, 10; riboflavin, 10; nicotinic acid, 10; calcium pantothenate, 10; vitamin B_{12}, 0.2; *p*-aminobenzoic acid, 10; lipoic acid, 10. The trace mineral solution contained the following (in g/liter): $FeCl_3$, 2; $ZnCl_2$, 0.05; $MnCl_2 \cdot 4H_2O$, 0.05; H_3BO_3, 0.05; $CoCl_2 \cdot 6H_2O$, 0.05; $CuCl_2 \cdot 2H_2O$, 0.03; $NiCl_2 \cdot 6H_2O$, 0.05; Na_4EDTA (tetrasodium salt), 0.5; $(NH_4)_2MoO_4$, 0.05; $AlK(SO_4)_2 \cdot 12H_2O$, 0.05. The medium was prepared anaerobically under a N_2–CO_2 (80:20) atmosphere, $NaHCO_3$ (1 g/liter) was added, and the mixture was reduced using (per liter) 0.5 g cysteine and 0.5 g N_2S. The final pH was 7.2. All soluble and insoluble biomass and defined substrates were used at a final concentration of 0.5% (wt/vol). Growth was at 75°C (*A. thermophilum*) or at 70°C (*C. saccharolyticus*) as static cultures in 100-ml serum bottles or with shaking (150 rpm) in 0.5- or 1.0-liter flasks. All media were filter sterilized using a 0.22-μm-pore-size sterile filter (Millipore Filter Corp., Bedford, MA). Insoluble substrates were added directly to sterilized culture bottles, followed by the addition of the filter-sterilized medium. The culture media containing the insoluble substrates without inoculation were used as negative controls.

Growth on spent substrate. The residual substrate was collected in late stationary phase. The residual substrate was separated from the cells by filtering it through glass filters (pore size, 40 to 60 μm), washed with distilled water to remove cells and media, and vacuum dried at 23°C for 18 h. This spent substrate was then used to grow new cell cultures. Unspent substrate was the unused and unwashed biomass (from the package) that was used in the first culture; first-spent substrate was that which remained after the first culture growth; and second-spent substrate was that which remained after growth of a culture on first-spent substrate.

Conversion of insoluble substrate. Conversion of insoluble substrate was calculated based on the amount of substrate remaining after cell growth had reached stationary phase (residual substrate). The residual substrate was determined by weight. So-called insoluble substrates derived from switchgrass, poplar, Avicel, and xylan were prepared by washing with water at 75°C (the growth temperature of *A. thermophilum* DSM 6725). Each substrate was suspended in distilled water (1 g/50 ml), stirred overnight at 75°C, and then washed twice with an equal volume of water at 75°C using a coarse glass filter (pore size, 40 to 60 μm). The substrate that remained on the filter was dried overnight at 50°C and was used for the growth experiments as insoluble substrate. The residual substrate was washed and dried similarly. The amount of insoluble substrate was measured after drying at 105°C overnight to a constant weight.

Cell growth. Cell density was monitored by cell count using a phase-contrast microscope with 40× magnification and expressed as cells per ml. To determine the extent to which *A. thermophilum* DSM 6725 adhered to insoluble substrate, the culture was shaken (150 rpm) for various time periods at 75°C in a closed 100-ml serum bottle with 50 ml of the mineral medium (pH 7.3) containing 0.5% (wt/vol) washed switchgrass in which the gas phase was replaced with N_2–CO_2 (80/20, vol/vol). The cultures were allowed to settle at room temperature for 15 min; then, 2 ml of the supernatant was withdrawn for the planktonic (free-floating) cell suspension. Both planktonic and substrate-bound cells were harvested by centrifuging the entire culture at 10,000 × *g* for 30 min. The centrifuged pellets were suspended in 50 ml of 50 mM Tris-HCl (pH 8.0), and the suspensions were sonicated on ice (six times for 30 s each time with 30-s intervals at 30 W). Cultures incubated under the same conditions without inoculation were used as the control for measuring cell protein concentration.

Determination of structural carbohydrates and acid-soluble lignin. Structural carbohydrates and acid-soluble lignin were determined using standard procedures at the National Renewable Resources Laboratory (http://www.nrel.gov/biomass/analytical_procedures.html). After the removal of water and ethanol extractives from switchgrass and poplar, the amount of lignin was estimated based on the absorption at 197 nm of the hydrolysate. Structural carbohydrates were determined by high-performance liquid chromatography (31).

Product analyses. Acetate and lactate were measured using a high-performance liquid chromatography apparatus (model 2690 separations module; Waters) equipped with an Aminex HPX-87H column (300 mm by 7.8 mm; Bio-Rad, Hercules, CA) at 40°C with 5 mM H_2SO_4 as the mobile phase at a flow rate of 0.6 ml min^{-1} with a refractive index detector (model 2410; Waters, Milford, MA). Ethanol was measured enzymatically using an ethanol kit (Megazyme, Wicklow, Ireland). Hydrogen was determined by a gas chromatograph (model GC-8A; Shimadzu, Kyoto, Japan) equipped with a thermal conductivity detector and a molecular sieve column (model 5A 80/100; Alltech, Deerfield, IL) with argon as the carrier gas. Reducing sugars were determined as described previously (22). Gas chromatography-mass spectrometry (GC-MS) was used to quantify the relative concentrations of targeted metabolites in the culture supernatants (μg sorbitol equivalents/ml). Sample preparation GC-MS operating conditions were as described elsewhere (37). Briefly, 100 μl of sorbitol (1 mg/ml aqueous solution) was added to each 2-ml sample as an internal standard, and samples were then dried in a helium stream. The internal standard was added to correct for differences in derivatization efficiency and changes in sample volume during heating. Dried exudates were dissolved in 500 μl of silylation-grade acetonitrile, followed by the addition of 500 μl *N*-methyl-*N*-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (Pierce Chemical Co., Rockford, IL), and samples were then heated for 1 h at 70°C to generate trimethylsilyl derivatives. After 5 days, 2-μl aliquots were injected into a DSQII (Thermo Fisher Scientific, Waltham, MA) GC-MS, fitted with an Rtx-5MS (Crossbond 5% diphenyl–95% dimethyl polysiloxane) capillary column (film thickness, 30 m by 0.25 mm by 0.25 μm; Restek, Bellefonte, PA). The standard quadrupole GC-MS was operated in electron impact (70 eV) ionization mode, with six full-spectrum (70- to 650-Da) scans per second. Gas (helium) flow was set at 1.1 ml per minute with an injection port configured in the splitless mode. The injection port and detector temperatures were set to 220°C and 300°C, respectively. The initial oven temperature was held at 50°C for 2 min and was programmed to increase at 20°C per min to 325°C and held for another 11.25 min before cycling back to the initial conditions. The target metabolites were integrated using a key selected ion (and confirmed by three additional characteristic *m/z* fragments), rather than the total ion current, to minimize the quantification of interfering metabolites. Extracted peaks were quantified by area integration and the areas scaled to the total ion current using correction factors for each metabolite. The concentrations were normalized to the quantity of the internal standard (sorbitol) recovered.

RESULTS

Growth on defined carbohydrates. *A. thermophilum* DSM 6725 grew well at 75°C with pectin, xylose, xylan, cellobiose, or crystalline cellulose as the carbon and energy sources. No significant growth of the organisms was observed in the standard medium in the absence of an added carbon source. As shown

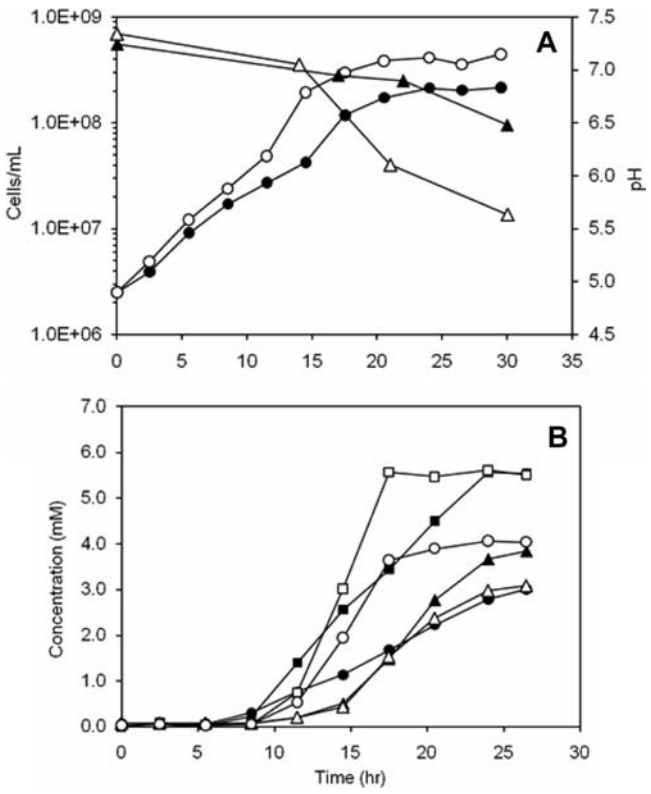


FIG. 1. Growth of *A. thermophilum* DSM 6725 on crystalline cellulose and xylan. Cell growth on unprocessed crystalline cellulose (solid symbols) and xylan (open symbols) was monitored by measuring cell density (circles) and pH (triangles) (A) and hydrogen (squares), lactate (triangles), and acetate (circles) (B).

in Fig. 1A, cells reached stationary phase on cellulose at 75°C within approximately 20 h with a cell density of 1.5×10^8 cells/ml and a decrease in pH from 7.1 to 6.4. The predominant product after this time was hydrogen, with smaller amounts of lactate. Ethanol was not detected (<100 μM). The ratio of H_2 to acetate produced after 21 h of growth on cellulose was 2.0, indicating that acetate was produced entirely as an end product of fermentation. In contrast, this ratio dropped to 1.4 after a similar growth period on xylan. This shift to more acetate formation is presumably because xylan contains acetyl substituents which are removed as part of the xylan degradation process. *A. thermophilum* DSM 6725 exhibited similar growth kinetics when cellobiose was used as the carbon source, although the end products differed, with more lactate produced than hydrogen (data not shown). Contrary to what was previously reported for *A. thermophilum* strain Z-1230 (32), strain DSM 6725 also grew well on xylose and pectin. As shown in Fig. 1A, growth on xylan was slightly better than on crystalline cellulose, with a higher cell density and a greater decrease in pH (to 5.5). The predominant end product was also hydrogen rather than lactate (Fig. 1B), and no ethanol was detected. Growth on xylose and pectin was similar to that observed on xylan (data not shown). When *A. thermophilum* DSM 6725 was grown on crystalline cellulose in the presence of acetate (50 mM) or lactate (50 mM) or under hydrogen (1 atm), there was little effect on the growth kinetics of the organism, although slightly lower cell densities were obtained (data not shown).

Growth on untreated plant biomass. Three grasses and one hardwood were selected as plant biomass substrates for growth. Tifton Bermuda grass and napier grass have relatively low lignin contents (3 to 4%), with cellulose and hemicellulose constituting 20 to 28% and 29 to 42% (wt/wt), respectively (15, 25). The high-lignin plants were switchgrass (acid-soluble lignin was measured at 17.8%, wt/wt) and the hardwood poplar (21.8% acid-soluble lignin). Chemical analyses of the biomass also indicated that, compared to poplar, switchgrass contains more xylose (19.2 versus 14.8%), arabinose (3.3 versus 0.4%), and galactose (1.8 versus 1.0%) but less glucose (31.0 versus 46.2%) and no mannose (which is found in poplar at 2.8%). Poplar and switchgrass also differed in relative amounts of water extractives (2.2 and 14.5%, respectively) and ethanol extractives (3.7 and 1.4%, respectively).

A. thermophilum DSM 6725 was able to grow on all four types of plant material when each was added to the standard growth medium without any pretreatment (the plant substrates were used as received and were added to filter-sterilized growth media). In closed static cultures (50 ml), growth on all plant materials was similar to that seen with the defined substrates, with cell densities reaching approximately 1.8×10^8 cells/ml within 20 h (data not shown). In closed stirred cultures (500 ml), *A. thermophilum* DSM 6725 grew on switchgrass and poplar, with cell densities after 21 h of 1.3×10^8 and 1.1×10^8 cells/ml, respectively (Fig. 2A). In all cases, stationary phase was reached after approximately 10 h of growth and the growth was accompanied by slight acidification of the media. As shown in Fig. 2B, hydrogen was the predominant end product. The ratios of hydrogen to acetate produced during growth on switchgrass and poplar after 21 h of growth were also less than that (2.0) found using cellulose. The values were 0.97 and 1.3, respectively, indicating that about half of the acetate that is produced originates from these highly acetylated plant materials. Chemical analyses of the residual switchgrass and poplar at periodic times throughout the growth phase up to 15 h revealed that the proportions of the constituents of the two types of biomass did not change significantly from those described above (data not shown).

Growth on insoluble plant biomass. Both switchgrass and poplar contain significant amounts of water-soluble material (14.5 and 2.2%, wt/wt, respectively), in addition to potential growth substrates such as protein. To determine if the ability of *A. thermophilum* DSM 6725 to grow on the untreated plant biomass, as well as on crystalline cellulose and xylan, was due to soluble rather than to insoluble substrates, a simple washing procedure was utilized. Both switchgrass and poplar, as well as crystalline cellulose and xylan, were incubated with unbuffered water at 75°C for 18 h to remove soluble material. The amounts of insoluble material remaining after this treatment were 72, 93, 84, and 34% (by weight) of the starting material, respectively. Thus, xylan had the majority of hot-water-extractable components (66%) and poplar the least (7%). These remaining materials were designated insoluble substrates. For the plant materials, chemical analyses revealed that the proportions of the constituents of the biomass did not change significantly after they had been incubated for 15 h at 75°C (in the absence of the organism; data not shown).

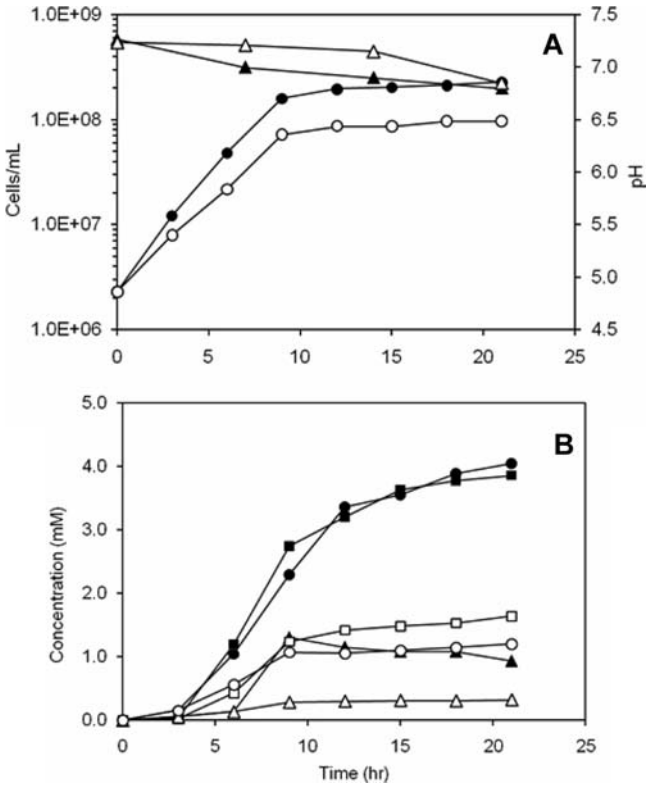


FIG. 2. Growth of *A. thermophilum* DSM 6725 on unprocessed switchgrass and poplar. Cell growth on unprocessed switchgrass (solid symbols) and poplar (open symbols) was monitored by measuring cell density (circles) and pH (triangles) (A) and hydrogen (squares), lactate (triangles), and acetate (circles) (B).

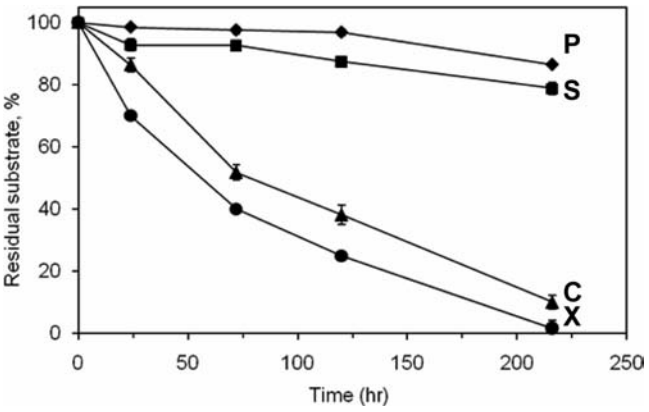


FIG. 3. Utilization of the insoluble forms of poplar, switchgrass, xylan, and crystalline cellulose by *A. thermophilum* DSM 6725. The amounts of substrate remaining after cell growth on the insoluble forms of poplar (P; diamonds), switchgrass (S; squares), xylan (X; circles), and crystalline cellulose (C; triangles) were determined by dry weight.

A. *thermophilum* DSM 6725 was able to utilize the insoluble material derived from poplar and switchgrass, as well as from crystalline cellulose and xylan, as sources of carbon and energy. The growth kinetics on each of the substrates were very similar to those observed on the unwashed (untreated) substrates, with cell densities of $\sim 2 \times 10^8$ cells/ml after 20 h or so (data not shown). To investigate the mechanism by which *A. thermophilum* DSM 6725 degraded the insoluble plant biomass, hot-water-washed insoluble switchgrass was used as the carbon and energy source, and the total amount of protein that was generated during the growth phase was determined for the planktonic cells and for all sedimented material, which included both planktonic cells and those adhered to the plant biomass. There were no significant differences between the two sets of measurements, indicating that a significant fraction of the cells is not complexed with the undegraded biomass (data not shown). Consequently, the cell densities determined in the experiments reported herein are an accurate estimate of cell growth on the insoluble plant biomass. Moreover, *A. thermophilum* DSM 6725 is predominantly in the planktonic state when it degrades biomass, and direct and permanent attachment to the insoluble substrate is apparently not necessary.

Production of reducing sugars from insoluble substrates. While the growth kinetics of *A. thermophilum* DSM 6725 on all four insoluble substrates were similar, there was an important difference in the responses to the insoluble plant material and the insoluble defined substrates upon prolonged incubation. As shown in Fig. 3, after almost 10 days, most of the xylan (98%) and cellulose (90%) had been solubilized by the organism, but the conversions of switchgrass and poplar were less extensive, with 26% and 15%, respectively, being utilized. Accordingly, chemical analysis of the growth media for reducing sugars revealed that insoluble crystalline cellulose was continuously degraded throughout the 10-day period, as shown by the continuous production of reducing sugars that approached 20 mM in concentration (after 10 days). In contrast, only low concentrations (<1 mM) of reducing sugars were produced from insoluble switchgrass and poplar, even after 10 days (data not shown). As shown in Table 1, metabolomic analyses revealed that after a 90-h incubation with *A. thermophilum* DSM 6725, high concentrations of glucose and cellobiose and, to a lesser extent, cellotriose were generated from crystalline cellulose, with comparable amounts of xylose and xylobiose and, to a lesser extent, xylotriose released from xylan. In contrast, only trace amounts of glucose were produced from poplar (Table 1). Trace amounts of cellobiose, galactose, xylose, and xylobiose were released from switchgrass, and in this case the amount of glucose produced was significant, reaching about 27% of that released on crystalline cellulose. Clearly, there is a difference in the mechanisms by which the organism metabolizes the two insoluble plant materials, and, more importantly, these also differ from those that are used to degrade the defined polysaccharide substrates.

The difference between how *A. thermophilum* DSM 6725 degrades defined polysaccharides and plant biomass was also evident from a kinetic analysis of end products. As shown in Fig. 4, the production of hydrogen and lactate closely followed cell growth, reaching a maximum in less than 20 h and showing only a slow increase even over 20 days. Note that both phenomena are seemingly independent of glucose production, which continues for about 10 days (Table 1), at which time only 10% of the cellulose remains (Fig. 3). In contrast, even though the growth kinetics on poplar and switchgrass are similar to those observed on cellulose, with a cell density of $>10^8$ cells/ml

TABLE 1. Production of simple sugars by *A. thermophilum* DSM 6725

Growth substrate	Concn (mM)[a]						
	Glucose	Cellobiose	Cellotriose	Galactose	Xylose	Xylobiose	Xylotriose
Poplar	0.06	ND	ND	ND	ND	ND	ND
Switchgrass	2.62	0.22	ND	1.05	0.40	0.31	ND
Cellulose	9.69	4.89	0.19	1.48	0.09	0.05	ND
Xylan	0.04	ND	ND	ND	9.26	4.00	0.09

[a] After 90 h of growth on insoluble forms of poplar, switchgrass, crystalline cellulose, and xylan, the concentrations were determined by GC-MS as described in Materials and Methods. ND, not detected.

reached within 20 h, hydrogen and lactate are produced continuously over the 20-day period. Moreover, there is a dramatic difference in the ratio of reduced products. Cellulose degradation results in the formation predominantly of lactate (the hydrogen/lactate ratio is 0.55 after 10 days), while hydrogen is the main product during growth on both switchgrass and poplar (the hydrogen/lactate ratio is 9.0 after 10 days).

Growth on spent insoluble substrates. The results shown in Fig. 3 prompted the following questions. Why does *A. thermophilum* DSM 6725 cease to grow significantly after 20 h or so on the insoluble plant substrates? Is it the recalcitrance of the material that remains? To address this issue, the insoluble material that remained after *A. thermophilum* DSM 6725 had reached stationary-phase growth on insoluble switchgrass and poplar was recovered and washed, and this so-called spent insoluble biomass was used as a carbon and energy source for a new *A. thermophilum* DSM 6725 culture. As shown in Fig. 5, the organism grew as well on spent switchgrass as, and even better on spent poplar than, it did on the unspent insoluble materials, with similar growth rates and cell densities. In addition, the amounts of the predominantly reduced products (hydrogen and lactate) were also virtually identical (data not shown). Moreover, the insoluble material that was left after the growth of the second culture was recovered and washed, and this second-spent insoluble biomass was used for a third fresh culture of *A. thermophilum* DSM 6725 (in all cases the initial concentration of the growth substrate was 5.0 g/liter). As shown in Fig. 5, the third culture supported similarly rapid growth and a cell density that approached the results obtained with the first and second cultures. The end products were once more comparable to those measured with the other cultures. Apparently, the mechanisms by which *A. thermophilum* DSM 6725 degrades and utilizes unspent, first-spent, and second-spent insoluble biomass from poplar and switchgrass are virtually identical. Compared to the original unspent insoluble biomass used in the first culture, the amounts of switchgrass and poplar that were solubilized after the third culture were 65.2% and 36.6%, respectively.

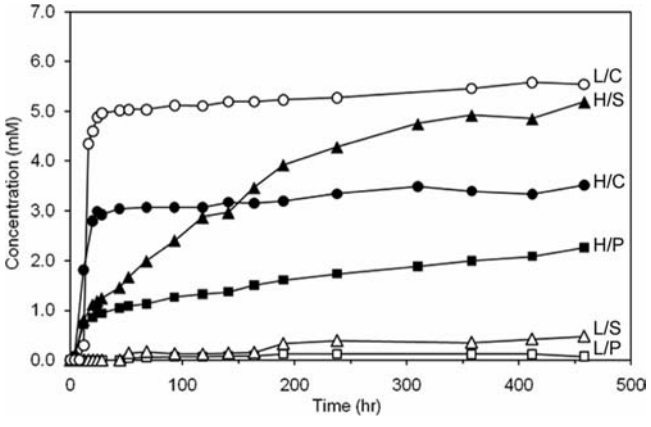


FIG. 4. End product analyses after prolonged growth of *A. thermophilum* DSM 6725 on the insoluble forms of poplar, switchgrass, and crystalline cellulose. Hydrogen (solid symbols) and lactate (open symbols) in cultures grown on the insoluble forms of poplar (squares), switchgrass (triangles), and crystalline cellulose (circles) were measured. H, hydrogen; L, lactate; P, poplar; S, switchgrass; C, crystalline cellulose.

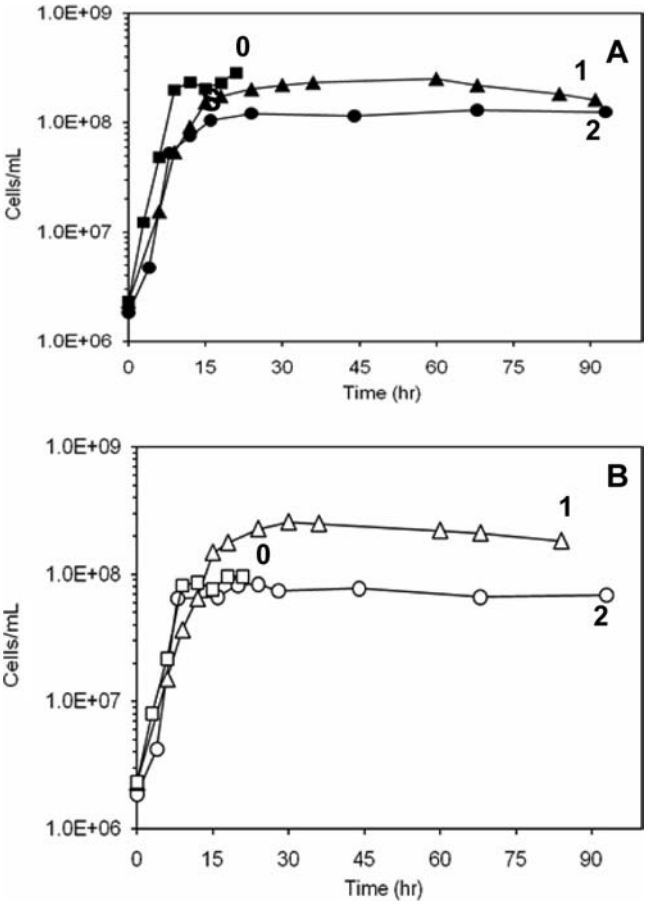


FIG. 5. Growth of *A. thermophilum* DSM 6725 on unspent, first-spent, and second-spent insoluble switchgrass (A) and insoluble poplar (B). Cells were grown on unspent (0; squares), first-spent (1; triangles), and second-spent (2; circles) switchgrass or poplar.

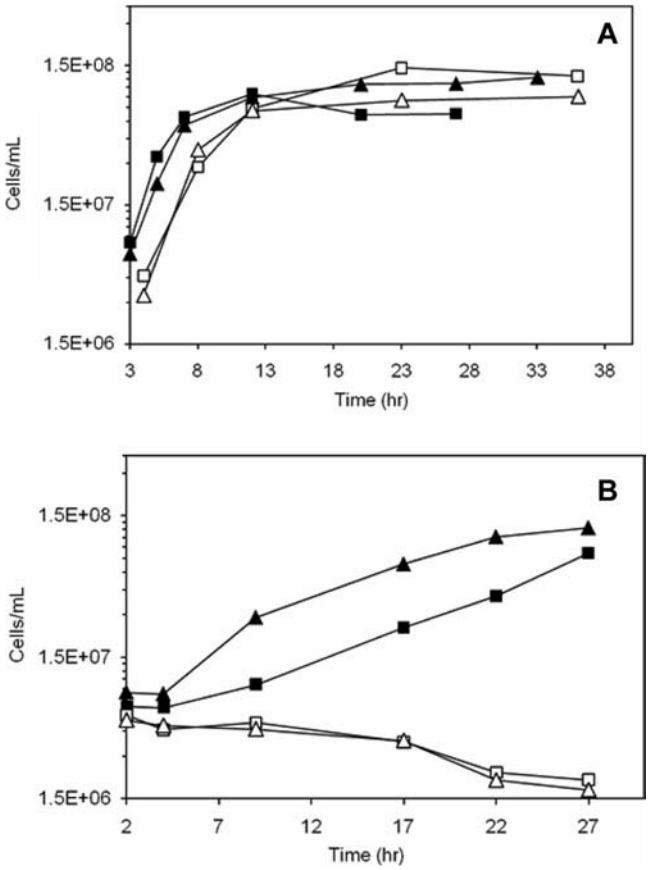


FIG. 6. Growth of *A. thermophilum* DSM 6725 and *C. saccharolyticus* DSM 8903 on insoluble and soluble fractions of switchgrass (A) and poplar (B). *A. thermophilum* (solid symbols) and *C. saccharolyticus* (open symbols) were grown on the insoluble (triangles) and soluble (squares) fractions of switchgrass or poplar.

Growth of *Caldicellulosiruptor saccharolyticus* DSM 8903. *C. saccharolyticus* is a close relative of *A. thermophilum* DSM 6725 and is known to degrade crystalline cellulose (26). As shown in Fig. 6, it was able to grow at 70°C on insoluble switchgrass in a manner similar to that of *A. thermophilum* strain 6725. The two organisms also exhibited comparable growth kinetics on so-called soluble switchgrass, which is the material released after the hot water wash of the plant material. Interestingly, however, while *A. thermophilum* DSM 6725 also grew on soluble poplar, as well as on the insoluble poplar material, *C. saccharolyticus* DSM 8903 did not. No significant growth of *C. saccharolyticus* DSM 8903 was detected on either the soluble or insoluble poplar, suggesting a significant difference in the biomass-degrading abilities of the two organisms.

DISCUSSION

A. thermophilum DSM 6725 grew very well on crystalline cellulose and xylan, the two main components of plant biomass. The ability of *A. thermophilum* strain DSM 6725 to grow on xylose and pectin, and also the lack of detectable ethanol as a fermentation product, is in contrast to what was previously reported with *A. thermophilum* strain Z-1320 (32), showing that the two strains are not identical. This conclusion was also reached from a comparison of the limited gene sequence data available for *A. thermophilum* strain Z-1320 with the relevant genes in the complete genome sequence of *A. thermophilum* strain DSM 6725 (16). Specifically, their 16S rRNA sequences have 14 mismatches, insertions, or deletions and the nucleotide sequences of their CelA genes have 23 mismatches (16, 27). In our hands, *A. thermophilum* DSM 6725 behaved as a stable, pure strain, and in all cultures examined by phase-contrast and scanning electron microscopy, we observed only one type of rod-shaped cell (data not shown). Similarly, all experiments with strain DSM 6725 were reproducible, and two cultures obtained from the DSMZ culture collection more than 2 years apart showed identical properties, including the ability to grow on xylan, xylose, and pectin. In addition, the genome sequence of *A. thermophilum* DSM 6725 was readily assembled (16), consistent with its having a pure culture as the DNA source. *A. thermophilum* DSM 6725 (16) and *A. thermophilum* Z-1320 (26, 32, 41) are therefore closely related but are not identical strains.

We also show here that *A. thermophilum* DSM 6725 is able to efficiently utilize untreated forms of both low-lignin (napier and Bermuda) and high-lignin (switchgrass) grasses and a hardwood (poplar) as carbon and energy sources, with cell densities of $>10^8$ cells/ml obtained in 20 h. Significant growth of an anaerobic thermophile such as *A. thermophilum* DSM 6725 on untreated poplar was unexpected, given that this hardwood contains a large amount of lignin and highly crystalline cellulose and it would be expected to be even more recalcitrant to microbial conversion than switchgrass. For example, softwood species contain cellulose of 52 to 62% crystallinity (1, 24) and the value for switchgrass is 55% (12), which compares with a value of 65% for poplar (38). This higher value is close to the range (66 to 75%) for the form of cellulose (Avicel) (18) used as a model substrate in the growth studies reported here. *A. thermophilum* DSM 6725 degraded more than 90% of this crystalline cellulose over a 10-day period (Fig. 3). The organism is comparable to the well-studied *Clostridium thermocellum* in its cellulose-degrading ability but has the advantage of a higher optimum growth temperature (75°C rather than 60°C) and the ability to hydrolyze xylan and consume xylose, an end product of xylan hydrolysis, which *C. thermocellum* lacks (11, 34). Like *C. thermocellum* (34, 40), *A. thermophilum* DSM 6725 generated high concentrations of glucose and cellobiose from cellulose, and similarly, xylan was converted mainly to xylobiose and xylose. These products are typical for cellulose and xylan hydrolysis by many other microorganisms, although the ratios may differ (2, 6, 40).

The concern that the ability of *A. thermophilum* DSM 6725 to grow on untreated or unprocessed plant biomass was due at least in part to its utilization of the more readily accessible, water-extractable components was found to be unwarranted by the demonstration that the organism grows just as well on what we term insoluble biomass, which is that remaining after an 18-h wash with water at 75°C (Fig. 4). Similarly, the recalcitrance of the biomass remaining at the end of the growth phase is not the reason why the organism ceases to grow, as the so-called first-spent and second-spent biomass substrates were as efficiently utilized as the unspent material (Fig. 5). The overall conversion of switchgrass (65%) and poplar (36%) after the third culture is an excellent starting point for cell

immobilization studies or the use of recycled bioreactors that might ultimately lead to almost complete solubilization of the plant material (10, 23, 35). What is not clear, however, is the fate of lignin, the other major component of plant cell walls. Lignin constitutes approximately 20% of both switchgrass and poplar biomass, and at present no anaerobic organism is known that can degrade lignin. Presumably, in the case of switchgrass, the 35% of the initial biomass that remains after the third culture is enriched with lignin and contains more recalcitrant cellulose and other components embedded into a lignin network than does the unspent switchgrass, although further analyses will be required to substantiate this.

Analysis of the end products formed upon growth on different substrates showed that on crystalline cellulose, xylan, switchgrass, and poplar, hydrogen was the predominant product, compared to lactate over the first 20 h or so. However, as shown in Fig. 4, continued incubation led to more lactate than hydrogen from cellulose, but then little of either product was produced after 30 h, even though accumulation of glucose continued (Table 1). In contrast, upon prolonged incubation on poplar and switchgrass, hydrogen remained the predominant product and production continued for up to 20 days. Changes in the ratio of hydrogen to lactate during the later stages of growth can originate from inhibition of hydrogenase by H_2 or by regulation of other enzymes involved in pyruvate conversion to lactate (8, 33). Hydrogen is clearly the predominant product when *A. thermophilum* DSM 6725 grows on plant biomass. Thus, for practical applications, the bacterium has the potential to be a hydrogen rather than ethanol producer.

C. saccharolyticus DSM 8903 could also grow on switchgrass (both soluble and insoluble fractions) but differed from *A. thermophilum* DSM 6725 in its response to poplar. *A. thermophilum* DSM 6725 grew on this substrate as well as on its water-extractive (soluble poplar) and extractive-free (insoluble poplar) fractions, while *C. saccharolyticus* DSM 8903 did not grow on either of these fractions. This may be because the insoluble fraction of poplar is too recalcitrant for this bacterium; because poplar has a higher lignin content, a higher relative amount of cellulose, and a higher crystallinity than switchgrass; and/or because mannan is present in poplar but not in switchgrass. A comparison of the genome sequences of the two organisms might indicate genes unique to *A. thermophilum* DSM 6725 that allow this bacterium to grow on untreated hardwood. Alternatively, it is known that the water-extractable part of hardwoods such as poplar, which contains alkaloids, tannins, sesquiterpenes, and lignans, can be toxic to microorganisms (5, 17, 28, 29). Chemical pretreatment of biomass, which is considered at present to be a necessary step in any applied biomass-to-biofuel conversion process, can lead to the release of additional potential inhibitors, such as furfural, metal ions, and various lignin degradation products (17, 28). The design of less severe pretreatment steps, or even avoidance of the pretreatment step altogether, is therefore of great importance. Presumably, a microorganism such as *A. thermophilum* DSM 6725 that utilizes untreated plant biomass has a great advantage.

In summary, *A. thermophilum* DSM 6725 has the ability to grow on plant biomass with a high lignin content and high crystallinity of cellulose; it is insensitive to inhibitors present in poplar biomass; its cells remain vital and produce hydrogen, which is an alternative biofuel to ethanol, for prolonged periods (20 days); it is able to hydrolyze highly crystalline cellulose almost completely with glucose and cellobiose as major products; and it grows on spent biomass efficiently. These unique properties might be of utility in any applied biomass conversion process.

ACKNOWLEDGMENTS

We thank Brian Davison for many helpful discussions.

This work was supported by grant DE-PS02-06ER64304 from the BioEnergy Science Center, Oak Ridge National Laboratory, a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science.

REFERENCES

1. **Andersson, S., H. Wikberg, E. Pesonen, S. L. Maunu, and R. Serimaa.** 2004. Studies of crystallinity of Scots pine and Norway spruce cellulose. Trees **18:**346–353.
2. **Bastawde, K. B.** 1992. Xylan structure, microbial xylanases, and their mode of action. World J. Microbiol. Biotechnol. **8:**353–368.
3. **Bayer, E. A., L. J. Shimon, Y. Shoham, and R. Lamed.** 1998. Cellulosomes—structure and ultrastructure. J. Struct. Biol. **124:**221–234.
4. **Blumer-Schuette, S. E., I. Kataeva, J. Westpheling, M. W. Adams, and R. M. Kelly.** 2008. Extremely thermophilic microorganisms for biomass conversion: status and prospects. Curr. Opin. Biotechnol. **19:**210–217.
5. **Chen, S. F., R. A. Mowery, C. J. Scarlata, and C. K. Chambliss.** 2007. Compositional analysis of water-soluble materials in corn stover. J. Agric. Food Chem. **55:**5912–5918.
6. **Cotta, M. A., and R. L. Zeltwanger.** 1995. Degradation and utilization of xylan by the ruminal bacteria *Butyrivibrio fibrisolvens* and *Selenomonas ruminantium*. Appl. Environ. Microbiol. **61:**4396–4402.
7. **Demain, A. L., M. Newcomb, and J. H. Wu.** 2005. Cellulase, clostridia, and ethanol. Microbiol. Mol. Biol. Rev. **69:**124–154.
8. **de Vrije, T., A. E. Mars, M. A. Budde, M. H. Lai, C. Dijkema, P. de Waard, and P. A. Claassen.** 2007. Glycolytic pathway and hydrogen yield studies of the extreme thermophile *Caldicellulosiruptor saccharolyticus*. Appl. Microbiol. Biotechnol. **74:**1358–1367.
9. **Freier, D., C. P. Mothershed, and J. Wiegel.** 1988. Characterization of *Clostridium thermocellum* JW20. Appl. Environ. Microbiol. **54:**204–211.
10. **Fukuda, H., S. Hama, S. Tamalampudi, and H. Noda.** 2008. Whole-cell biocatalysts for biodiesel fuel production. Trends Biotechnol. **26:**668–673.
11. **Garcia-Martinez, D. V., A. Shinmyo, A. Madia, and A. L. Demain.** 1980. Studies on cellulase production by *Clostridium thermocellum*. Eur. J. Appl. Microbiol. Biotechnol. **9:**189–197.
12. **Harris, D., and S. DeBolt.** 2008. Relative crystallinity of plant biomass: studies on assembly, adaptation and acclimation. PLoS ONE **3:**e2897.
13. **Himmel, M. E., S. Y. Ding, D. K. Johnson, W. S. Adney, M. R. Nimlos, J. W. Brady, and T. D. Foust.** 2007. Biomass recalcitrance: engineering plants and enzymes for biofuels production. Science **315:**804–807.
14. **Jeoh, T., C. I. Ishizawa, M. F. Davis, M. E. Himmel, W. S. Adney, and D. K. Johnson.** 2007. Cellulase digestibility of pretreated biomass is limited by cellulose accessibility. Biotechnol. Bioeng. **98:**112–122.
15. **Johnson, W. L., J. Guerrero, and D. Pezo.** 1973. Cell-wall constituents and *in vitro* digestability of Napier grass (*Pennisetum purpureum*). J. Anim. Sci. **37:**1255–1261.
16. **Kataeva, I. A., S.-J. Yang, P. Dam, F. L. Poole II, Y. Yin, F. Zhou, W.-C. Chou, Y. Xu, L. Goodwin, D. R. Sims, J. C. Detter, L. J. Hauser, J. Westpheling, and M. W. W. Adams.** 2009. Genome sequence of the anaerobic, thermophilic and cellulolytic bacterium "*Anaerocellum thermophilum*" DSM 6725. J. Bacteriol. **191:**3760–3761.
17. **Kostamo, A., B. Holmbom, and J. V. Kukkonen.** 2004. Fate of wood extractives in wastewater treatment plants at kraft pulp mills and mechanical pulp mills. Water Res. **38:**972–982.
18. **Kumar, V., S. H. Kothari, and G. S. Banker.** 2001. Compression, compaction, and disintegration properties of low crystallinity celluloses produced using different agitation rates during their regeneration from phosphoric acid solutions. AAPS PharmSciTech. **2:**E7.
19. **Lynd, L. R., M. S. Laser, D. Brandsby, B. E. Dale, B. Davison, R. Hamilton, M. Himmel, M. Keller, J. D. McMillan, J. Sheehan, and C. E. Wyman.** 2008. How biotech can transform biofuels. Nat. Biotechnol. **26:**169–172.
20. **Lynd, L. R., P. J. Weimer, W. H. van Zyl, and I. S. Pretorius.** 2002. Microbial cellulose utilization: fundamentals and biotechnology. Microbiol. Mol. Biol. Rev. **66:**506–577.
21. **McCann, M. C., and N. C. Carpita.** 2008. Designing the deconstruction of plant cell walls. Curr. Opin. Plant Biol. **11:**314–320.
22. **Miller, G. L.** 1959. Use of dinitrosalicylic acid reagent for determination of reducing sugar. Anal. Chem. **31:**426–428.

23. **Nedovic, V., and R. Willaert.** 2005. Applications of cell immobilization biotechnology. Springer-Verlag, New York, NY.
24. **Newman, R. H., and J. A. Hemmingson.** 1990. Determination of the degree of cellulose crystallinity in wood by carbon-13 nuclear magnetic resonance spectroscopy. Holzforschung **44:**351–355.
25. **Premazzi, L. M., F. A. Monteiro, and J. E. Corrente.** 2004. Tillering of Tifton 85 bermudagrass in response to nitrogen rates and time of application after cutting. Sci. Agric. **60:**565–571.
26. **Rainey, F. A., A. M. Donnison, P. H. Janssen, D. Saul, A. Rodrigo, P. L. Bergquist, R. M. Daniel, E. Stackebrandt, and H. W. Morgan.** 1994. Description of *Caldicellulosiruptor saccharolyticus* gen. nov., sp. nov: an obligately anaerobic, extremely thermophilic, cellulolytic bacterium. FEMS Microbiol. Lett. **120:**263–266.
27. **Rainey, F. A., N. L. Ward, H. W. Morgan, R. Toalster, and E. Stackebrandt.** 1993. Phylogenetic analysis of anaerobic thermophilic bacteria: aid for their reclassification. J. Bacteriol. **175:**4772–4779.
28. **Ranatunga, T. D., J. Jervis, R. F. Helm, J. D. McMillan, and C. Hatzis.** 1997. Identification of inhibitory components toxic toward *Zymomonas mobilis* CP4(pZB5) xylose fermentation. Appl. Biochem. Biotechnol. **67:**185–198.
29. **Rowe, J. W., and A. H. Conner.** 1979. Extractives in eastern hardwoods—a review. Gen. Tech. Rep. FPL 18. USDA Forest Products Laboratory, Madison, WI.
30. **Skinner, K. A., and T. D. Leathers.** 2004. Bacterial contaminants of fuel ethanol production. J. Ind. Microbiol. Biotechnol. **31:**401–408.
31. **Sluiter, A., B. Hames, R. Ruiz, C. Scarlata, J. Sluiter, D. Templeton, and D. Crocker.** 2008. Determination of structural carbohydrates and lignin in biomass. Technical report NREL/TP-510-42618. National Renewable Energy Laboratory, Golden, CO. http://www.nrel.gov/biomass/pdfs/42618.pdf.
32. **Svetlichnyi, V. A., T. P. Svetlichnaya, N. A. Chernykh, and G. A. Zavarzin.** 1990. *Anaerocellum thermophilum* gen. nov., sp. nov., an extremely thermophilic cellulolytic eubacterium isolated from hot-springs in the valley of Geysers. Microbiology **59:**598–604.
33. **van Niel, E. W., P. A. Claassen, and A. J. Stams.** 2003. Substrate and product inhibition of hydrogen production by the extreme thermophile, *Caldicellulosiruptor saccharolyticus*. Biotechnol. Bioeng. **81:**255–262.
34. **Wiegel, J., C. P. Mothershed, and J. Puls.** 1985. Differences in xylan degradation by various noncellulolytic thermophilic anaerobes and *Clostridium thermocellum*. Appl. Environ. Microbiol. **49:**656–659.
35. **Willaert, R.** 2007. Cell immobilization and its applications in biotechnology: current trends and future prospects, p. 287–332. *In* E. M. T. El-Mansi, C. F. A. Bryce, A. L. Demain, and A. R. Allman (ed.), Fermentation microbiology and biotechnology, 2nd ed. CRC Press, Boca Raton, FL.
36. **Wilson, D. B.** 2008. Three microbial strategies for plant cell wall degradation. Ann. N. Y. Acad. Sci. **1125:**289–297.
37. **Yang, S., T. J. Tschaplinski, N. L. Engle, S. L. Carroll, S. L. Martin, B. H. Davison, A. V. Palumbo, and S. D. Brown.** 2009. Transcriptomic and metabolomic profiling of *Zymomonas mobilis* during aerobic and anaerobic fermentations. BMC Genomics **10:**34.
38. **Zhang, W., M. Liang, and C. Lu.** 2007. Morphological and structural development of hardwood cellulose during mechanochemical pretreatment in solid state through pan-milling. Cellulose **14:**447–456.
39. **Zhang, Y. H. P., S. Y. Ding, J. R. Mielenz, J. B. Cui, R. T. Elander, M. Laser, M. E. Himmel, J. R. McMillan, and L. R. Lynd.** 2007. Fractionating recalcitrant lignocellulose at modest reaction conditions. Biotechnol. Bioeng. **97:**214–223.
40. **Zhang, Y. H. P., and L. R. Lynd.** 2005. Cellulose utilization by *Clostridium thermocellum*: bioenergetics and hydrolysis product assimilation. Proc. Natl. Acad. Sci. USA **102:**7321–7325.
41. **Zverlov, V., S. Mahr, K. Riedel, and K. Bronnenmeier.** 1998. Properties and gene structure of a bifunctional cellulolytic enzyme (CelA) from the extreme thermophile "*Anaerocellum thermophilum*" with separate glycosyl hydrolase family 9 and 48 catalytic domains. Microbiology **144:**457–465.

Genomics 93 (2009) 473–480

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*

Xiaohan Yang, Sara Jawdy, Timothy J. Tschaplinski, Gerald A. Tuskan *

Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6422, USA

ARTICLE INFO

Article history:
Received 8 October 2008
Accepted 15 January 2009
Available online 29 January 2009

Keywords:
Lineage-specific genes
Gene expression
Arabidopsis
Rice
Poplar
Perennial plants
Woody plants

ABSTRACT

Protein sequences were compared among *Arabidopsis*, *Oryza* and *Populus* to identify differential gene (**DG**) sets that are in one but not the other two genomes. The **DG** sets were screened against a plant transcript database, the NR protein database and six newly-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*) to identify a set of species-specific genes (**SS**). Gene expression, protein motif and intron number were examined. 165, 638 and 109 **SS** genes were identified in *Arabidopsis*, *Oryza* and *Populus*, respectively. Some **SS** genes were preferentially expressed in flowers, roots, xylem and cambium or up-regulated by stress. Six conserved motifs in *Arabidopsis* and *Oryza* **SS** proteins were found in other distant lineages. The **SS** gene sets were enriched with intronless genes.

The results reflect functional and/or anatomical differences between monocots and eudicots or between herbaceous and woody plants. The *Populus*-specific genes are candidates for carbon sequestration and biofuel research.

The identification of taxon specific genes has both scientific and practical values [1]. Recently, computational approaches were used to identify genes unique to bacteria [1], virus [2] and Poaceae [3]. Three of the fully-sequenced plant species, *Arabidopsis* [4], *Oryza* [5–7] and *Populus* [8], represent three major types of higher plants: annual eudicots, annual monocots and perennial eudicots, respectively. Identification of species-specific genes in these taxa may provide insights into the molecular features distinguishing monocot from eudicot or herbaceous from woody plants. Additionally, transcript assemblies have been coalesced from expressed sequences collected from the NCBI GenBank Nucleotide database for more than 250 plant species representing a wide range of the evolutionary lineages [9]. Finally, genome sequences have been published for *Carica* [10] and *Vitis* [11] and draft/partial genome sequences are available in the public domain for *Glycine* (http://www.phytozome.net/soybean), *Medicago* (http://www.medicago.org/), *Sorghum* (http://genome.jgi-psf.org/) and *Zea* (http://www.maizesequence.org). These genomic data provide a broad and robust comparative resource for identification of species-specific genes in plants.

In this study, we initially identified three differential gene (**DG**) sets in the context of *Arabidopsis*, *Oryza* and *Populus*, i.e., 1) *Arabidopsis* genes without homologs in *Oryza* or *Populus*, 2) *Oryza* genes without homologs in *Arabidopsis* or *Populus*, and 3) *Populus* genes without homologs in *Arabidopsis* or *Oryza*. Then we used these three **DG** sets to query a customized database containing more than 250 plant transcript assemblies [9] followed by a query against the NR protein database and a query against a customized database containing annotated protein sequences from six recently-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*). The **DG** genes that have no homologs in the other species revealed three sets of species-specific (**SS**) genes in *Arabidopsis*, *Oryza* and *Populus*. To gain insights into the functions of the **SS** genes, we compared their expression pattern using microarray/digital northern data and identified conserved protein motifs that were over-represented in each taxon. The exon–intron structures were also examined to aid in the understanding of differential gene evolution.

Results

Differential genes in Arabidopsis, Oryza and Populus

Using a BLASTp search with an *e*-value cutoff of 0.1, we identified three differential gene (**DG**) sets that contained expression evidence from EST or full-length cDNA (FL-cDNA) in the context of *Arabidopsis*, *Oryza* and *Populus* (Fig. 1). The *Arabidopsis* **DG** set contained 917 genes without homologs in *Oryza* or *Populus*; the *Oryza* **DG** set contained 2781 genes without homologs in *Arabidopsis* or *Populus*; the *Populus* **DG** set contained 594 *Populus* genes without homologs in *Arabidopsis* or *Oryza* (Fig. 1).

To investigate the relationship between the three **DG** sets and genes in other plant species, we used a tBLASTn search (with an *e*-value cutoff of 0.1) to query a customized database containing the transcript assemblies [9]. The 250 plant species represented in the database were manually divided into 11 sub-datasets: algae, moss, fern, herbaceous gymnosperms (gymnosperm_herb), woody

* Corresponding author. Fax: +1 865 576 9939.
E-mail address: tuskanga@ornl.gov (G.A. Tuskan).

0888-7543/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2009.01.002

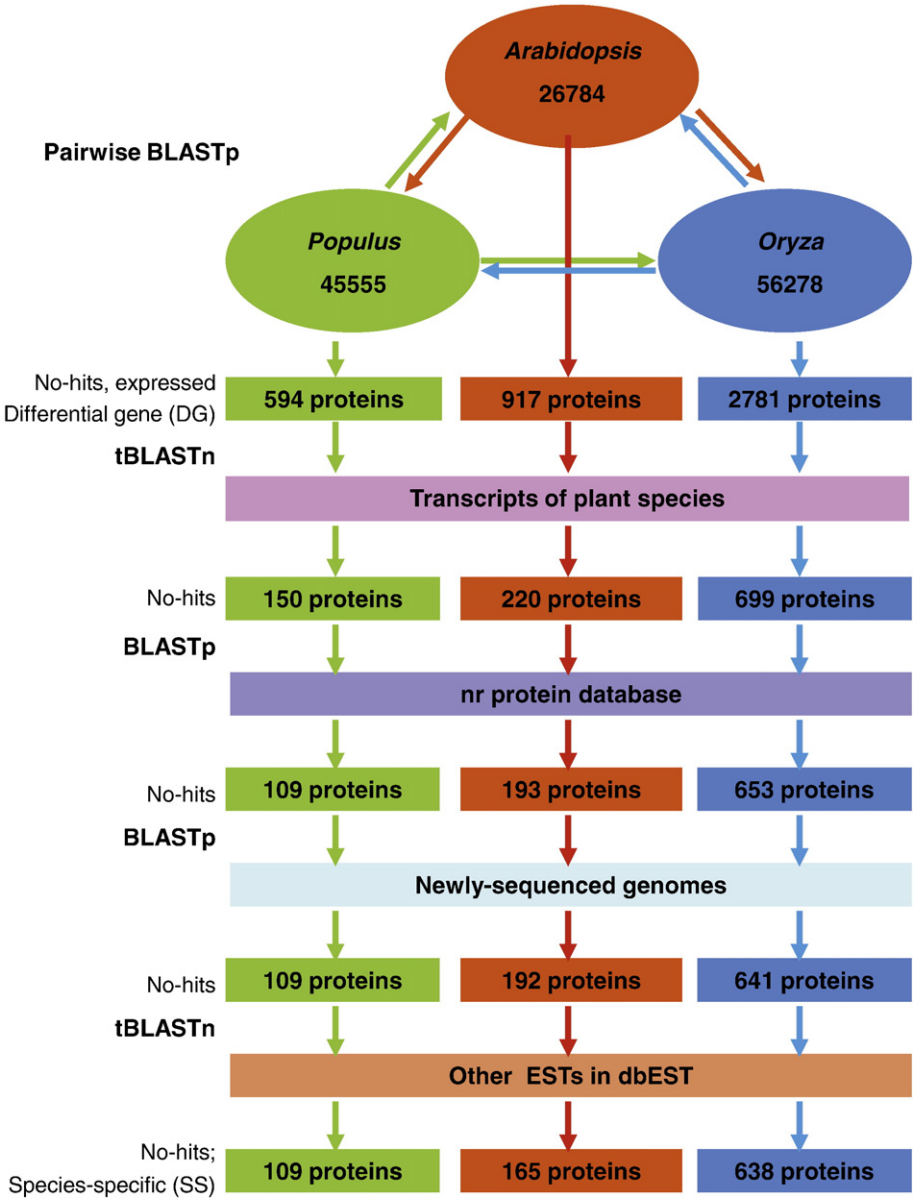


Fig. 1. Procedure for identifying species-specific genes in *Arabidopsis*, *Oryza* and *Populus*. The scoring matrix Blossum62, 80, Pam70, 30 were used for all the blast searches. The newly-sequenced genomes include *Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*.

gymnosperms (gymnosperm_woody), herbaceous basal angiosperms (angiosperm_basal_herb), woody basal angiosperms (angiosperm_basal_woody), herbaceous monocots (angiosperm_monocot_herb), woody monocots (angiosperm_monocot_woody), herbaceous eudicots (angiosperm_eudicot_herb) and woody eudicots (angiosperm_eudicot_woody). Results of the tBLASTn search revealed that the percentage of the *Arabidopsis* **DG** set with homology to herbaceous eudicots is higher than that of the *Oryza* **DG** set (90% vs. 26%, respectively; Fig. 2), whereas the percentage of the *Arabidopsis* **DG** set with homology to herbaceous monocots is lower than that of the *Oryza* **DG** set (21% vs. 92%, respectively; Fig. 2). These differences may be related to genes and processes that distinguish herbaceous monocots from herbaceous eudicots. The percentage of the *Populus* **DG** set with homology to woody eudicots is higher than that of the *Arabidopsis* **DG** set (77% vs. 26%, respectively; Fig. 2), whereas the percentage of the *Populus* **DG** set with homology to herbaceous eudicots is lower than that of the *Arabidopsis* **DG** set (71% vs. 90%, respectively; Fig. 2). These differences may be related to genes and processes that distinguish woody and herbaceous properties in eudicot plants.

Expression of genes in the differential gene sets

Gene expression in the *Arabidopsis* **DG** set that is associated with developmental and environmental responses was examined using a *K*-means clustering analysis of several *Arabidopsis* microarray datasets. For the developmental data set, using the whole seedling as the baseline reference, one cluster of 39 genes in the *Arabidopsis* **DG** set showed up-regulation in stamen (~8-fold) and pollen (~128-fold) (Supplementary Fig. S1 (Cluster 01)), one cluster of 27 genes showed up-regulated expression in root (~180-fold) (Supplementary Fig. S1 (Cluster 08)), one cluster of 49 genes showed up-regulated expression in flower tissues (~4000-fold) (Supplementary Fig. S1 (Cluster 10)) and one cluster of 88 genes showed up-regulated expression in mature pollen (~12-fold) (Supplementary Fig. S1 (Cluster 22)). Also, several clusters of genes exhibited down-regulated expression when

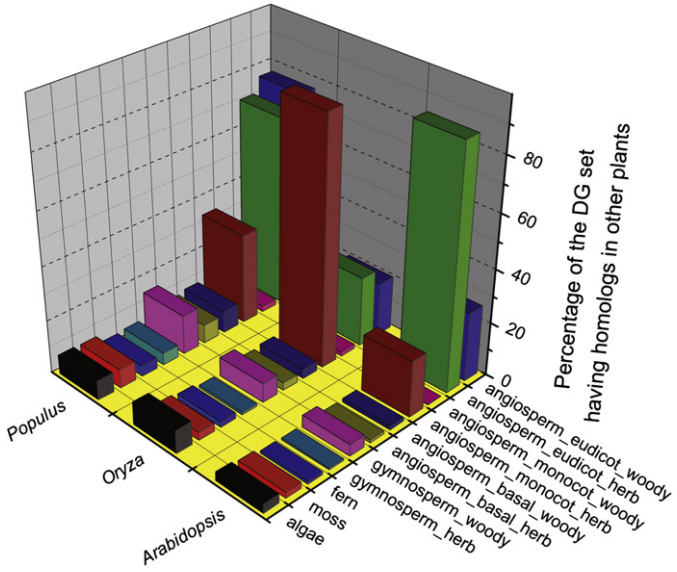


Fig. 2. The percentage of the differential gene sets in *Arabidopsis*, *Oryza* and *Populus* showing homology to other plant species as revealed by tBLASTn search against the transcript assemblies (TA) of more than 250 plant species (Childs et al., 2007), which were manually divided into 11 sub-datasets: algae, moss, fern, herbaceous gymnosperm (gymnosperm_herb), woody gymnosperm (gymnosperm_woody), herbaceous basal angiosperm (angiosperm_basal_herb), woody basal angiosperm (angiosperm_basal_woody), herbaceous monocot (angiosperm_monocot_herb), woody monocot (angiosperm_monocot_woody), herbaceous eudicot (angiosperm_eudicot_herb), and woody eudicot (angiosperm_eudicot_woody).

compared with the whole seedling, e.g. one cluster of 25 genes down-regulated in majority of the tissue types sampled (Supplementary Fig. S1 (Cluster 02)), one cluster of 6 genes down-regulated in shoot apex (~10-fold), carpel and pollen (~30-fold) (Supplementary Fig. S1 (Cluster 05)), and one cluster of 8 genes down-regulated in root (~70-fold), seed (~70-fold), stamen and pollen (~50-fold) (Supplementary Fig. S1 (Cluster 21)).

For the stress dataset, using untreated plants as the baseline reference, one cluster of 39 genes showed up-regulation under UV-B and biotic stress (~32-fold) (Supplementary Fig. S2 (Cluster 08)) and one cluster of 21 genes showed up-regulated expression under light treatments (~8-fold) (Supplementary Fig. S2 (Cluster 15)).

One cluster of 212 genes in the *Oryza* **DG** set showed relatively high levels of expression in the leaf tissue (Supplementary Fig. S3 (Cluster 20)) and one cluster of 239 genes showed relatively high levels of expression in panicle tissue (Supplementary Fig. S3 (Cluster 21)). Likewise, one cluster of 9 genes in the *Populus* **DG** set showed relatively high levels of expression in the male flowers (Supplementary Figs. S4 (Cluster 07) and S5A), one cluster of 20 genes showed relatively high levels of expression in the female flowers (Supplementary Figs. S4 (Cluster 10) and S5B), one cluster of 31 genes showed relatively high levels of expression in the xylem tissue (Supplementary Figs. S4 (Cluster 18) and S6) and one cluster of 65 genes showed relatively high levels of expression in the flower buds (Supplementary Fig. S4 (Cluster 21)). There is an overlap between the *Populus* differential genes up-regulated in woody tissues and the *Populus* differential genes having homology to other woody plants. Specifically, 23 of the 31 genes that were preferentially expressed in xylem have blast hits in other woody plants (Supplementary Table S4).

Species-specific genes in Arabidopsis, Oryza and Populus

Using the three **DG** sets to query the transcript assemblies [9] by tBLASTn (with an *e*-value cutoff of 0.1) followed by querying the NR protein database and the six newly-sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*) using BLASTp (with an *e*-value cutoff of 0.1), we identified 192 *Arabidopsis*-, 651 *Oryza*- and 145 *Populus*-specific genes that had no homologs in other species. Of the 145 *Populus*-specific genes, 36 were incorrectly annotated, i.e., truncated (without start codon or stop codon) or interrupted by internal stop codon and consequently they were excluded from the final *Populus*-specific gene list. Of the 651 *Oryza* **SS** genes, 10 were transposable elements and they were excluded from the final *Oryza*-specific gene list. Further search (tblastn) against the ESTs from other species in the dbEST database revealed that 27 *Arabidopsis*-specific proteins had hits in other lineages (mostly in *Brassica*), and 3 *Oryza*-specific proteins had hits in other lineages in Poacea. Therefore, the final species-specific gene list includes 165 *Arabidopsis*-, 638 *Oryza*- and 109 *Populus*-specific genes (Supplementary Tables S1–S3).

Expression of the species-specific genes

Expression of the *Arabidopsis*-specific genes associated with developmental and environmental responses was investigated using *K*-means clustering analysis of *Arabidopsis* microarray data.

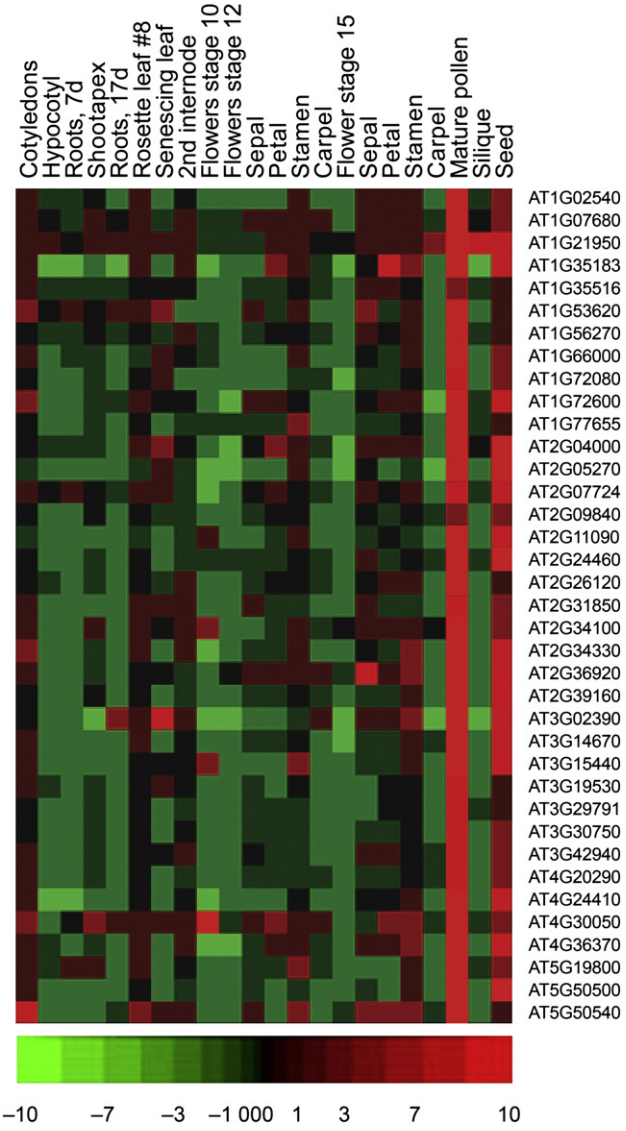


Fig. 3. Expression pattern of Cluster 01 (see Supplementary Fig. S7) of the *Arabidopsis* species-specific set as revealed by clustering analysis of the developmental microarray data. The whole seedling was used as a reference for comparison. The color scheme represents log ratio, with red indicating up-regulated and green down-regulated expression.

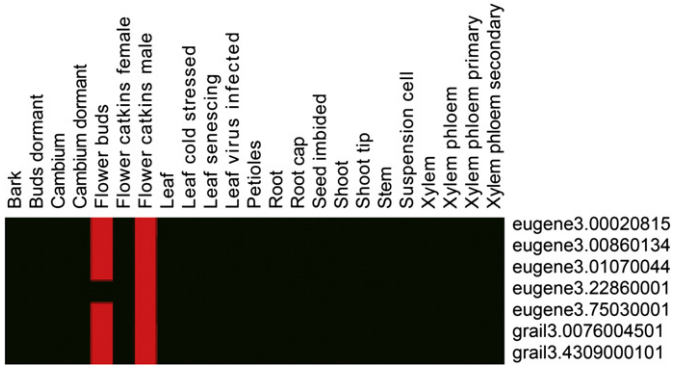


Fig. 4. Expression pattern of Cluster 06 (see Supplementary Fig. S10) of the *Populus* species-specific genes as revealed by clustering analysis of the digital northern data. Black represents no EST counts. Red indicates that EST counts are ≥1, with the brighter color the more EST counts.

For the developmental data set using the whole seedling as the baseline reference, one cluster of 37 *Arabidopsis*-specific genes showed up-regulated expression in mature pollen (~10-fold) (Fig. 3; Supplementary Fig. S7 (Cluster 01)), one cluster of 7 genes showed up-regulated expression in root (~500-fold) (Supplementary Fig. S7 (Cluster 05)) and one cluster of 5 genes showed up-regulated expression in stamen (~16-fold) and pollen (~256-fold) (Supplementary Fig. S7 (Cluster 07)). For the stress dataset using untreated plants as references, only a single cluster of 12 genes showed up-regulated expression under heat stress (~64-fold) (Supplementary Fig. S8 (Cluster 19)).

Expression of the *Oryza*-specific genes was evaluated using *K*-means clustering analysis of the *Oryza* digital northern data. One cluster of 39 *Oryza*-specific genes showed relatively high levels of expression in the flower tissue (Supplementary Fig. S9 (Cluster 14)), one cluster of 47 genes showed relatively high levels of expression in the pistil tissue (Supplementary Fig. S9 (Cluster 16)) and one cluster of 39 genes showed relatively high levels of expression in the root tissue (Supplementary Fig. S9 (Cluster 17)). Similarly, one cluster of 7 *Populus*-specific genes showed relatively high levels of expression in the female flowers (Fig. 4, Supplementary Fig. S10 (Cluster 06)), one cluster of 12 genes showed relatively high levels of expression in the xylem tissue (Supplementary Fig. S10 (Cluster 14)), one cluster of 15 genes showed relatively high levels of expression in the cambium tissue (Supplementary Fig. S10 (Cluster 15)) and one cluster of 5 genes showed relatively high levels of expression in the leaf tissue (Supplementary Fig. S10 (Cluster 18)).

Table 1
Conserved motifs identified by MEME using amino acid sequences of specific genes from *Arabidopsis* (Supplementary Table S1) and *Oryza* (Supplementary Table S2)

Motif ID	Motif consensus	Genes
Motif 1	D[E]-E-E-E-E-E-E[D]-E[R]-D[E]-D [E]-D[ELR]-E[R]	6 genes from *Arabidopsis* specific group 1
Motif 2	E-K[R]-x-E[R]-E[R]-G[R]-E-E[K]-E-K [E]-E-E-E-x-D[E]	22 genes from *Oryza* specific group 1
Motif 3	L[V]-R-F-N-D-E[K]-P[L]-R-G-D[NL]-L [PQS]-L[S]-L[W]-S[K]-P[QS]-V[G]-M [ET]-A[FHLP]-T[HP]-P[QT]-K[IN]	4 genes from *Oryza* specific group 1
Motif 4	H[D]-H-H-H-R[CG]-H-H-H	8 genes from *Oryza* specific group 1
Motif 5	K[ERS]-G[KN]-E[GIKL]-D[NRI]-S [KNQ]-D[KMT]-D[KE]-D[LS]-D[KLY]-D [EGQ]-S[TV]-D[FVY]-G[HSTW]-S[DK]-S [DNW]-N[GTV]-D[EKNV]-D[NG]-L [NSY]-D[EFG]-S[GT]-D[Q]-D[HY]-D[FL]-D [EKMS]-D[T]-A[DKSV]-M[DNS]-K[HLM]-D [CHK]-K[DH]-I[LD]-S[FGT]-D[ES]-L[SQ]-F [DHY]-K[HI]-D[EHLY]-K[CGS]-D[I]	4 genes from *Oryza* specific group 1
Motif 6	E[K]-E[R]-R[EG]-E[GL]-E[DR]-D[E]-G[S]-R [EK]-K-K[E]-K[L]-E[DK]-E[R]-E-K-E[EKRT]-K	4 genes from *Oryza* specific gene expression Cluster 18

Table 2
Other species containing motifs listed in Table 1, as revealed by MAST search of the NR protein database

Motif	Organism		Protein	
	Taxonomy	Count	Name	Count
Motif 1				
	Eukaryota; Metazoa	6	Hypothetical protein	5
	Eukaryota; Alveolata	3	Unknown protein	1
	Eukaryota; Fungi	1	Hmgb1-prov protein	1
			Casein kinase II beta subunit CKB1	1
			alpha 2B adrenergic receptor	1
			MGC85274 protein	1
Motif 2				
	Eukaryota; Metazoa	9	Hypothetical protein	9
	Archaea; Euryarchaeota	1	Similar to melanoma antigen family A	1
Motif 3				
	Bacteria; Proteobacteria	1	Hypothetical protein	1
	Viruses; dsDNA viruses	9	lef-8	9
Motif 4				
	Bacteria; Firmicutes	1	Hypothetical protein	5
	Eukaryota; Alveolata	2	Histidine-rich protein	2
	Eukaryota; Metazoa	4	Unknown protein	1
	Bacteria; Proteobacteria	2	SJCHGC08151 protein	1
	Eukaryota; Mycetozoa	1	Proline-rich region	1
Motif 5				
	Eukaryota; Alveolata	5	Hypothetical protein	8
	Eukaryota; Mycetozoa	4	MIF4G domain protein	1
	Eukaryota; Fungi	1	Adenylate and guanylate cyclase catalytic Domain containing protein	1
Motif 6				
	Eukaryota; Metazoa	6	Hypothetical protein	6
	Bacteria; Firmicutes	1	LOC566027 protein	1
	Archaea; Crenarchaeota	1	LRRGT00096	1
	Eukaryota; Viridiplantae;	2	Unknown protein	1
			lti45 (low-temperature-induced protein)	1
			LTI45	

Only the top ten hits are listed.

Over-represented protein motifs in the species-specific genes

To identify the over-represented protein motifs, the **SS** gene set in each species were divided into groups based on protein sequences using the CLUSS program [12]. The **SS** genes in *Arabidopsis*, *Oryza* and *Populus* were divided into 6 (Supplementary Table S1), 5 (Supplementary Table S2) and 4 groups (Supplementary Table S3), respectively. Five protein motifs (Motifs 1–5) were identified by the MEME program from the group 1 sequences in *Arabidopsis* and *Oryza* (Table 1). In addition, one motif was identified by MEME from the *Oryza* **SS** gene expression Cluster 18 (Table 1; Supplementary Fig. S9). To investigate the evolutionary history of the protein motifs, the motif scoring matrix were used to query the NR protein database using the MAST program. The over-represented motifs identified from the **SS** sets in *Arabidopsis* and *Oryza* were also found in unknown or hypothetical proteins of other distantly-related organisms, most of which are in the early branches of the three of life, such as Archaea, Bacteria, Fungi and Metazoan (Table 2).

Real-time PCR confirmation of Populus gene expression

Because the annotation of the newly-sequenced *Populus* genome is in draft form and may contain annotation errors, we performed real-time RT-PCR analysis of 15 **SS** genes and 38 **DG** genes in five-tissues (i.e., bark, leaf, root, shoot tip and stem) from *Populus*. Fifty of the 53 *Populus* genes (94.3%) showed detectable expression in the tissues

Table 3
Multiple-tissue real-time RT-PCR analysis of expression of *Populus* specific (**SS**) and differential (**DG**) genes

Category	Gene name	Bark	SEM	Leaf	SEM	Root	SEM	Shoot	SEM	Stem	SEM
SS	estExt_fgenesh4_kg.C_LG_I0055[a]	130.6	39.9	14.2	3.8	140.0	38.6	1.0	0.6	91.8	29.8
SS	estExt_fgenesh4_pg.C_2630012	1.1	0.0	2.4	0.4	1.5	0.2	1.1	0.1	1.0	0.2
SS	eugene3.00020815	1.0	0.3	12.2	2.7	1.5	0.2	1.6	0.4	7.6	1.6
SS	eugene3.00021257	1.0	0.1	4.3	0.4	1.6	0.2	1.2	0.1	2.5	0.3
SS	eugene3.00091626	1.4	0.2	3.5	0.3	1.2	0.1	2.5	0.4	1.0	0.1
SS	eugene3.00102359	1.7	0.4	78.7	15.4	6.8	0.6	1.0	0.2	31.8	5.7
SS	eugene3.00120934	1.0	0.4	27.5	4.5	11.7	1.8	7.6	0.0	10.8	1.9
SS	eugene3.00280166	2.2	0.1	9.1	1.2	9.4	0.9	1.0	0.1	14.6	2.1
SS	eugene3.00400046[a]	7.4	0.9	2.8	0.4	19.8	4.9	12.5	1.8	1.0	0.2
SS	eugene3.01070044	1.6	0.4	11.2	2.7	1.0	0.2	5.1	1.1	0.0	0.0
SS	eugene3.02230021	18.2	3.3	4.8	0.3	2.1	0.5	1.0	0.2	0.0	0.0
SS	eugene3.04600002	1.2	0.1	5.3	0.5	1.0	0.1	2.4	0.2	1.8	0.2
SS	eugene3.14950001	1.4	0.1	5.5	0.5	1.9	0.2	1.0	0.1	2.0	0.2
SS	grail3.0043007301	3.1	0.8	11.2	1.7	1.0	0.2	2.0	0.4	2.8	0.6
SS	grail3.0169000301	1.4	0.2	2.8	0.2	1.6	0.2	1.3	0.1	1.0	0.1
DG	estExt_fgenesh4_kg.C_280013	6.0	1.1	95.7	31.9	1.0	0.3	6.2	0.7	16.8	2.6
DG	estExt_fgenesh4_pg.C_14770002	1.0	0.2	18.3	2.3	11.3	1.3	6.6	0.0	10.0	1.5
DG	estExt_fgenesh4_pg.C_290045	16.8	2.2	228.0	63.2	1.0	0.4	47.2	7.3	13.1	2.1
DG	estExt_fgenesh4_pg.C_LG_II1111[a]	1.0	0.3	292.4	66.6	29.6	3.8	2.7	0.5	103.8	27.4
DG	estExt_fgenesh4_pg.C_LG_VIII0722[a]	1.0	0.1	18.3	2.4	22.7	3.3	6.1	1.0	7.8	1.2
DG	estExt_fgenesh4_pg.C_LG_X0606	16.5	2.1	11.8	1.9	1.0	0.2	1.4	0.4	16.5	3.2
DG	estExt_fgenesh4_pg.C_LG_XVIII0908	40.1	6.5	29.8	5.6	4.1	0.5	1.0	0.3	5.3	0.8
DG	eugene3.00011774[a]	1.5	0.2	34.0	4.5	14.9	1.6	1.0	0.1	41.0	7.5
DG	eugene3.00051028[a]	1.0	0.2	22.0	2.5	11.5	1.4	15.2	0.9	6.0	0.9
DG	eugene3.00051604	0.0	0.0	1.5	0.4	1.1	0.2	1.0	0.0	1.1	0.1
DG	eugene3.00060029	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.1
DG	eugene3.00060132	1.0	0.3	21.0	2.6	2.3	0.4	7.1	8.2	6.7	1.0
DG	eugene3.00060513	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00061307	1.3	0.1	1.0	0.2	1.9	0.2	1.1	0.3	1.0	0.1
DG	eugene3.00070428	7.7	0.7	36.1	7.1	1.0	0.2	1.3	0.3	19.6	3.7
DG	eugene3.00080951	31.8	4.0	1.0	0.2	7.2	0.6	19.9	1.5	1.8	0.3
DG	eugene3.00081749[a]	47.1	4.4	92.0	12.9	160.6	31.0	1.0	0.3	194.0	35.0
DG	eugene3.00090211	11.9	1.6	14.6	1.2	1.0	0.1	2.8	0.3	7.8	1.0
DG	eugene3.00101292	1.9	0.2	3.7	0.3	1.0	0.1	2.1	0.1	1.7	0.2
DG	eugene3.00120867	6.0	0.6	7.5	0.7	1.0	0.1	2.3	0.3	5.3	0.6
DG	eugene3.00121045	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00180855	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
DG	eugene3.00880016	2.2	0.2	4.5	0.5	1.2	0.1	2.4	0.1	1.0	0.1
DG	eugene3.02650003	11.7	1.8	3.9	0.5	1.9	0.3	1.7	0.0	1.0	0.2
DG	eugene3.08380002	1.6	0.2	4.4	0.6	4.1	0.4	1.0	0.0	1.8	0.3

For each gene, the gene expression level in the lowest expressed tissue was set to 1.0 as a reference for comparison and the gene expression levels in the other tissues were ratios relative to the reference.

[a] These genes showing relatively high expression in the root were placed into our *Populus* transformation pipeline for constitutive over-expression and RNAi knockdown.

sampled, indicating that the *Populus* **SS** and **DG** gene sets are functional (Table 3).

Exon–intron structure

To study gene structure, we examined the intron composition of the **SS** genes by dividing gene structures into 4 types: intronless, 1 intron, 2 introns and 3-or-more introns per gene. The **SS** sets in all three studied species (*Arabidopsis*, *Oryza* and *Populus*) contain more intronless genes than expected by chance alone when compared all other genes in each genomes ($P<1\times10^{-5}$) (Fig. 5). We performed GC content analysis for lineage-specific genes and found that the GC content in the coding region of the intronless genes (57%) is significantly higher than that of intron-containing genes (53%) ($P<1\times10^{-6}$). There are intronless genes that are also found in the protein motif groups. Specifically, 44% (=23/48) of the genes in the conserved protein motif groups lack introns. In particular, 59% (=13/22) of the genes in the conserved protein motif #2 (Table 1) group in *Oryza* lack introns.

Discussion

While flowering plants share many common aspects in growth and development, they have distinguishing features at the taxonomic level. Phylogenetic analyses based on both morphological and molecular data have obviously separated *Oryza* (monocot) from the eudicot species *Arabidopsis* and *Populus* [13]. Similarly, our results demonstrate that the percentage of the *Arabidopsis* or *Populus* **DG** showing homology to eudicots is higher than that of the *Oryza* **DG** set, whereas the percentage of the *Arabidopsis* or *Populus* **DG** showing homology to monocots is lower than that of the *Oryza* **DG** set. Although both *Arabidopsis* and *Populus* are eudicots, they also have distinct growth and developmental habits, an herbaceous annual vs. a woody perennial, respectively. Some of the molecular elements responsible for the differences between *Arabidopsis* and *Populus* are revealed by our results where the percentage of the *Populus* **DG** set with homology to woody eudicots is higher than that of the *Arabidopsis* **DG** set and the percentage of the *Populus* **DG** set with homology to herbaceous eudicots is lower than that of the *Arabidopsis* **DG** set. In combination these results demonstrated that some genes in either **DG** set are preferentially expressed in root, flower or xylem tissue. The differences in gene expression may be reflective of the differences in reproductive or stem anatomy characteristics in *Arabidopsis* and *Populus*.

Although the species-specific genes in *Arabidopsis* and *Oryza* have no homologs in other species, they do share some motifs with other organisms, mostly in the early branches of the tree-of-life. It is interesting that the proteins in other organisms sharing these motifs are largely functionally-unknown or hypothetical. Future molecular and biochemical experiments will be needed to investigate the functions of the unknown protein motifs. It is possible that the

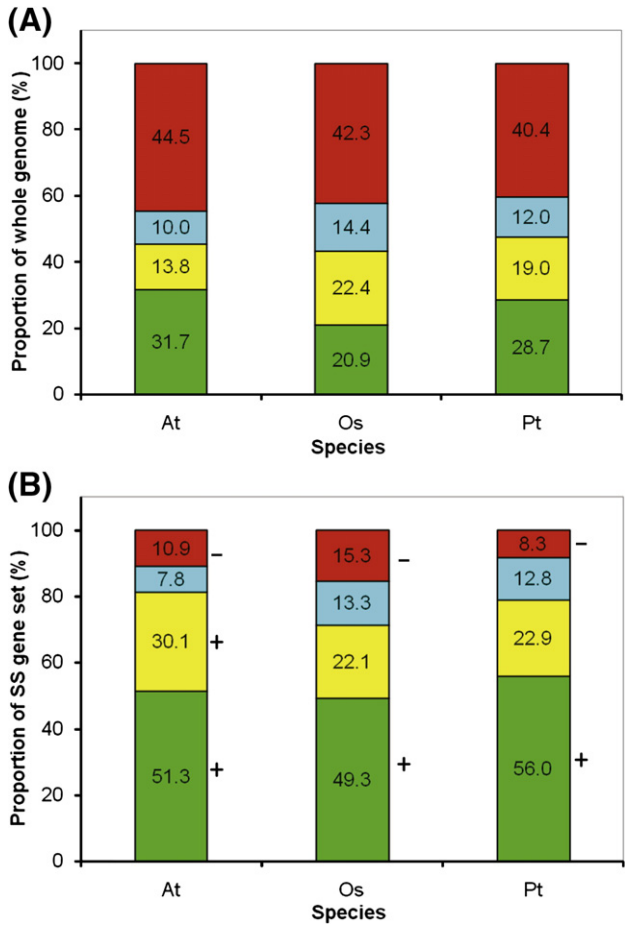


Fig. 5. Number of introns per gene in *Arabidopsis* (At), *Oryza* (Os) and *Populus* (Pt). (A) All genes in the whole genome; (B) Species-specific (**SS**) genes. "+" indicates that species-specific genes are over-represented and "−" under-represented at $P<1\times10^{-5}$, as compared with all genes in the whole genome. Gene structures were divided into 4 types: 0 (green), 1 (yellow), 2 (blue), and 3-or-more introns (red) per gene.

conserved motifs in the species-specific proteins of *Arabidopsis* and *Oryza* may have resulted from domain co-option from other organisms. Intraspecific protein domain co-option has been reported in *Populus* gene evolution [8,14]. It is also possible that the shared motifs are contained in ancestral genes that occurred early in gene evolution and have been subsequently lost in most higher plants.

Carmel et al. [15] have inferred that high intron density was reached in the early evolutionary history of plants and the last common ancestor of multicellular life forms harbored approximately 3.4 introns per kb, a greater intron density than in most of the extant fungi and in some animals. A recent report also implies that rates of intron creation were higher during earlier periods of plant evolution [16]. We recently reported that lineage-specific F-box gene is over-represented by intronless gene structure [17]. In this study, we also found that intronless genes were enriched in the species-specific gene sets as compared with the whole-genome annotation gene set. It is tempting to hypothesize that these species-specific sets resulted from recent lineage-specific expansion. Further studies are needed to test this hypothesis. We found that the GC content in the coding region of the intronless lineage-specific genes is significantly higher than that of intron-containing genes. This is consistent with previous reports revealing that high GC content class was enriched with intronless genes in plants [18,19].

The expression pattern of the species-specific genes indicates that some of these genes are associated with flower, root, leaf, or xylem development as well as stress response. We suggest that these tissue-specific or stress-responsive species-specific genes are involved in the molecular processes underlying the taxa phenotypic features in these plant species. As such, the species-specific genes will be valuable source of information for understanding the molecular mechanism underlying the distinguishing features in growth and development in the three model plant species. Future experiments involving over-expression and/or RNAi knockdown are needed to decipher the functions of these specific genes. Genes preferentially expressed in the *Populus* root tissue are potential candidate genes for carbon sequestration research. Eight *Populus* **DG**/**SS** genes showing relatively high expression in the root tissue have been placed in our *Populus* transformation pipeline for functional studies using the over-expression and RNAi knockdown strategy.

The number of **SS** genes in *Oryza* (638) is much higher than that in *Arabidopsis* (165) or *Populus* (109), even though the number of genes in the *Oryza* genome (42,653) is equivalent to that in the *Populus* genome (45,555) and 1.5 times that in the *Arabidopsis* genome (27,000) [8,20,21]. Even though the **SS** gene set in *Oryza* was identified by query against 1) the transcript assemblies from algae, moss, fern, gymnosperm to angiosperm including both monocot and eudicot, 2) the NR protein database which contains proteins identified in many diverse organisms, and 3) the *Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea* genomes, we cannot conclusively determine that all of these genes arose from lineage-specific expansion after speciation. Annotation and assembly errors may account for some of the predicted gene models in the examined species. These issues will be resolved with additional genomic sequence, publicly available expression data sets, and functional characterization of the **SS** gene set.

Materials and methods

Genome sequences

The *Arabidopsis* protein sequences were downloaded from TAIR release 7 (http://www.arabidopsis.org/). The *Oryza* protein sequences were downloaded from TIGR release 5 (http://rice.plantbiology.msu.edu/). The *Populus* protein sequences were downloaded from JGI release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). The TIGR Plant Transcript Assemblies were downloaded from: http://plantta.tigr.org/index.shtml, which were built from the expressed transcripts in more than 250 plant species collected from dbEST (ESTs) and the NCBI GenBank nucleotide database (full-length and partial cDNAs) [9]. *Carica* protein sequences were downloaded from ftp://asgpb.mhpcc.hawaii.edu/papaya/annotation.genbank_submission/. *Glycine* protein sequences were downloaded from ftp://ftp.jgi-psf.org/pub/JGI_data/Glycine_max/. *Medicago* protein sequences were downloaded from http://www.medicago.org/genome/. *Sorghum* protein sequences were downloaded from http://genome.jgi-psf.org/Sorbi1/Sorbi1.home.html. *Vitis* protein sequences were downloaded from http://www.genoscope.cns.fr/spip/Vitis-vinifera-whole-genome.html. *Zea* protein sequences were downloaded from http://www.maizesequence.org/index.html.

Homolog search

The species-specific genes were identified in a pipeline (Fig. 1) based on a homolog search using BLASTp or BLASTn [22] with an *e*-value cutoff of 0.1 and scoring matrix of Blossum62, Blossum80, Pam70 and Pam30. Because the *e*-value of BLAST search is influenced by the database size, we use the standardized NCBI NR protein database size for all the BLAST searches in this study.

Expression evidence

The expression evidence from EST or full-length cDNA (FL-cDNA) for *Arabidopsis* genes were obtained from TAIR release 7 (http://www.arabidopsis.org/). The expression evidence from EST or FL-cDNA

for *Oryza* genes were obtained from TIGR release 5 (http://rice.plantbiology.msu.edu/). The expression evidence from EST or FL-cDNA for *Populus* genes was determined by minimal 97% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences [17].

Analysis of gene expression

Two *Arabidopsis* microarray datasets were compiled from AtGenExpress [23,24]. The developmental data set contains cotyledons, hypocotyl, roots (7 or 17 d), shoot apex (vegetative), rosette leaf (# 8), senescing leaves, 2nd internode, flowers (stage 10/11, 12 or 15), sepals (flowers stage 12 or 15), petals (flowers stage 12 or 15), stamens (flowers stage 12 or 15), carpels (flowers stage 12 or 15), mature pollen), siliques (with seeds stage 5; late heart to mid torpedo embryos) and seeds (stage 10, without siliques; green cotyledons embryos). The gene expression levels are expressed as $LOG_2(x/y)$, where x is the detection signal from the above tissue types and y is the detection signal from seedling (green parts).

The environmental data set contains cold (4 °C; 1 or 3 h), salt (150 mM NaCl; 1 or 3 h), drought treatments (1 or 3 h after 15 min dry air stream leading to 10% loss of fresh weight), oxidative treatments (10 μM methyl viologen; 1 h or 3 h), UV-B (1 or 3 h after 15 min exposure to ultraviolet-B light, 1.18 W/m^2 Philips TL40W/12), heat (38 °C; 1 or 3 h), pathogen (*Phytophthora infestans*; 6, 12 or 24 h; in control treatments, H_2O was applied to leaves), blue light treatment (4 h), far-red light treatment (4 h), red light treatment (4 h) and white light treatment (4 h). Dark treatment (4 h) was used as a control for the light experiments. *K*-means clustering of the *Arabidopsis* microarray data was performed using EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) with correlation distance (uncentered).

For *Oryza* and *Populus* EST analysis, the coding sequences were used to search the dbEST using BLASTn. Expression evidence from EST sequences was determined by minimal 97% identity over an alignment of at least 100 bp and at least 80% length of the shorter sequences [17]. *K*-means clustering of the square-root transformed EST data (EST counts per tissue) was performed using EPCLUST (http://ep.ebi.ac.uk/EP/EPCLUST/) with Euclidean distance.

Protein classification

The specific protein sequences were clustered into groups using the CLUSS program [12].

Motif identification

Protein motifs of specific genes were identified statistically using MEME [25] with motif length set as 6 to 100, maximum motif number < 100, and *e*-value < 0.005. The MAST program [26] was used to search protein motifs.

Real-time PCR

P. trichocarpa 'Nisqually-1' stem and leaf tissues were taken from plants grown *in vitro* on media containing Murashige and Skoog salts [27], 3% sucrose and 0.25% Gelrite (PhytoTechnology Laboratories) at 23 °C ± 1 °C under cool-white fluorescent light (approximately 125 mmol m^{-2} s^{-1}, 16-h photoperiod). Root, shoot tip, petiole and bark tissues were taken from plants grown in a greenhouse under natural lighting and temperatures ranging from 25 °C to 35 °C. Total RNA was extracted from root, stem, shoot tip, petiole, leaf and bark using the Spectrum Plant Total RNA kit (Sigma-Aldrich) and then treated with AMPD1 DNase I (Sigma-Aldrich) to eliminate DNA, according to the manufacturer's instructions. RNA purity was determined spectrophotometrically and quality was determined by examining rRNA bands on agarose gels. cDNA was synthesized from 2 μg of RNA using the PowerScript PrePrimed Single Shots with random hexamers as primer (CLONTECH Laboratories) in a 20 μl reaction.

For real-time RT-PCR analysis using gene-specific primers the cDNA was diluted 50-fold. Amplification reactions (25.0 μl) were carried out using iQ™ SYBR® Green Supermix according to the instructions provided by Bio-Rad Laboratories. Each reaction contained a cDNA template (1.0 μl), SYBR® Green supermix (12.5 μl), sterile water (8.5 μl) and the appropriate forward and reverse 5 μM primer pair (1.5 μl each). The gene used as a control to normalize the data for differences in input RNA and efficiency of reverse transcription between the samples was an actin gene expressed at a constant rate across tissue types. PCR amplification reactions were performed in triplicate. The thermal cycling conditions took place on an iCycler Real Time PCR detection system (Bio-Rad Laboratories 2005) and included 3 min at 95 °C, 40 cycles of 95 °C for 15 s, 55 °C for 20 s and 72 °C for 20 s, 1 min at 95 °C, 80 cycles at 55 °C for 10 s with the temperature increasing by 0.5 °C after each cycle and then held at 4 °C until plates were removed from the machine. Data analysis was carried out using DART-PCR version 1.0 [28] and qBASE [29]. DART-PCR version 1.0 was used to calculate primer efficiency. This information was then used in qBASE, along with cycle threshold values, to calculate fold change in expression of each gene as compared to its expression in the tissue where transcript levels were the lowest.

Intron analysis

Information about the number of introns per gene was obtained from *Arabidopsis* genome annotation release 7 (http://www.arabidopsis.org/), *Oryza* genome annotation release 5 (http://rice.plantbiology.msu.edu/) and *Populus* genome annotation release 1.1 (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html).

GC content analysis

The GC content of coding sequences were performed using EMBOSS [30]. A two-tailed *T*-test was used to compare the mean GC content of coding sequence between the intronless and intron-containing genes.

Acknowledgments

We thank Stan Wullschleger and Udaya Kalluri for reviewing the manuscript and valuable comments, and J.C. Tuskan for input on the design of the study. Funding for this research was provided by the U.S. Department of Energy, Office of Science, Biological and Environmental Research Carbon Sequestration Program. ORNL is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2009.01.002.

References

[1] R. Mazumder, D.A. Natale, S. Murthy, R. Thiagarajan, C.H. Wu, Computational identification of strain-, species- and genus-specific proteins, BMC Bioinformatics 6 (2005) 279.
[2] H. Ogata, J.M. Claverie, Unique genes in giant viruses: regular substitution pattern and anomalously short size, Genome Res. 17 (2007) 1353–1361.
[3] M.A. Campbell, W. Zhu, N. Jiang, H. Lin, S. Ouyang, K.L. Childs, B.J. Haas, J.P. Hamilton, C.R. Buell, Identification and characterization of lineage-specific genes within the Poaceae, Plant Physiol. 145 (2007) 1311–1322.
[4] Arabidopsis Genome Initiative, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, Nature 408 (2000) 796–815.
[5] S.A. Goff, D. Ricke, T.H. Lan, G. Presting, R. Wang, M. Dunn, J. Glazebrook, A. Sessions, P. Oeller, H. Varma, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. japonica), Science 296 (2002) 92–100.

[6] International Rice Genome Sequencing Project, The map-based sequence of the rice genome, Nature 436 (2005) 793–800.
[7] J. Yu, S. Hu, J. Wang, G.K. Wong, S. Li, B. Liu, Y. Deng, L. Dai, Y. Zhou, X. Zhang, et al., A draft sequence of the rice genome (*Oryza sativa* L. ssp. indica), Science 296 (2002) 79–92.
[8] G.A. Tuskan, S. Difazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, et al., The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), Science 313 (2006) 1596–1604.
[9] K.L. Childs, J.P. Hamilton, W. Zhu, E. Ly, F. Cheung, H. Wu, P.D. Rabinowicz, C.D. Town, C.R. Buell, A.P. Chan, The TIGR plant transcript assemblies database, Nucleic Acids Res. 35 (2007) D846–D851.
[10] R. Ming, S. Hou, Y. Feng, Q. Yu, A. Dionne-Laporte, J.H. Saw, P. Senin, W. Wang, B.V. Ly, K.L. Lewis, et al., The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya* Linnaeus), Nature 452 (2008) 991–996.
[11] O. Jaillon, J.M. Aury, B. Noel, A. Policriti, C. Clepet, A. Casagrande, N. Choisne, S. Aubourg, N. Vitulo, C. Jubin, et al., The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla, Nature 449 (2007) 463–467.
[12] A. Kelil, S. Wang, R. Brzezinski, A. Fleury, CLUSS: clustering of protein sequences based on a new similarity measure, BMC. Bioinformatics. 8 (2007) 286.
[13] APG II, An update of the Angiosperm Phylogeny Group classification for the orders and families of flowering plants: APG II, Bot. J. Linn. Soc. 141 (2003) 399–436.
[14] X. Yang, G.A. Tuskan, M.Z. Cheng, Divergence of the Dof gene families in poplar, *Arabidopsis*, and rice suggests multiple modes of gene evolution after duplication, Plant Physiol. 142 (2006) 820–830.
[15] L. Carmel, Y.I. Wolf, I.B. Rogozin, E.V. Koonin, Three distinct modes of intron dynamics in the evolution of eukaryotes, Genome Res. 17 (2007) 1034–1044.
[16] S.W. Roy, D. Penny, Patterns of intron loss and gain in plants: intron loss-dominated evolution and genome-wide comparison of *O. sativa* and *A. thaliana*, Mol. Biol. Evol. 24 (2007) 171–181.
[17] X. Yang, U.C. Kalluri, S. Jawdy, L.E. Gunter, T. Yin, T.J. Tschaplinski, D.J. Weston, P. Ranjan, G.A. Tuskan, F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants, Plant Physiol. 148 (2008) 1189–1200.
[18] N. Carels, G. Bernardi, Two classes of genes in plants, Genetics 154 (2000) 1819–1825.
[19] N.N. Alexandrov, V.V. Brover, S. Freidin, M.E. Troukhan, T.V. Tatarinova, H. Zhang, T.J. Swaller, Y.P. Lu, J. Bouck, R.B. Flavell, et al., Insights into corn genes derived from large-scale cDNA sequencing, Plant Mol. Biol. 69 (2009) 179–194.
[20] S. Ouyang, W. Zhu, J. Hamilton, H. Lin, M. Campbell, K. Childs, F. Thibaud-Nissen, R.L. Malek, Y. Lee, L. Zheng, et al., The TIGR rice genome annotation resource: improvements and new features, Nucleic Acids Res. 35 (2007) D883–D887.
[21] B.J. Haas, A.L. Delcher, S.M. Mount, J.R. Wortman, R.K. Smith Jr., L.I. Hannick, R. Maiti, C.M. Ronning, D.B. Rusch, C.D. Town, et al., Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies, Nucleic Acids Res. 31 (2003) 5654–5666.
[22] S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410.
[23] J. Kilian, D. Whitehead, J. Horak, D. Wanke, S. Weinl, O. Batistic, C. D'Angelo, E. Bornberg-Bauer, J. Kudla, K. Harter, The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses, Plant J. 50 (2007) 347–363.
[24] M. Schmid, T.S. Davison, S.R. Henz, U.J. Pape, M. Demar, M. Vingron, B. Scholkopf, D. Weigel, J.U. Lohmann, A gene expression map of *Arabidopsis thaliana* development, Nat. Genet. 37 (2005) 501–506.
[25] T.L. Bailey, C. Elkan, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, Proc. Int. Conf. Intell. Syst. Mol. Biol. 2 (1994) 28–36.
[26] T.L. Bailey, M. Gribskov, Combining evidence using p-values: application to sequence homology searches, Bioinformatics 14 (1998) 48–54.
[27] T. Murashige, F. Skoog, A revised medium for rapid growth and bioassay with tobacco tissue cultures, Physiol. Plant. 15 (1962) 473–497.
[28] S.N. Peirson, J.N. Butler, R.G. Foster, Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis, Nucleic Acids Res. 31 (2003) e73.
[29] J. Hellemans, G. Mortier, A. De Paepe, F. Speleman, J. Vandesompele, qBase relative quantification framework and software for management and automated analysis of real-time quantitative PCR data, Genome Biol. 8 (2007) R19.
[30] P. Rice, I. Longden, A. Bleasby, EMBOSS: the European Molecular Biology Open Software Suite, Trends Genet. 16 (2000) 276–277.

Critical Reviews in Plant Science, 28:285–308, 2009

ISSN: 0735-2689 print / 1549-7836 online
DOI: 10.1080/07352680903241014

Poplar Genomics: State of the Science

Xiaohan Yang,[1] Udaya C. Kalluri,[1] Stephen P. DiFazio,[2] Stan D. Wullschleger,[1] Timothy J. Tschaplinski,[1] (Max) Zong-Ming Cheng,[3] and Gerald A. Tuskan[1]

[1]*Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*
[2]*Department of Biology, West Virginia University, Morgantown, West Virginia 26506, USA*
[3]*Department of Plant Sciences, University of Tennessee, Knoxville, Tennessee 37996, USA*

Table of Contents

Address correspondence to Xiaohan Yang, Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA. E-mail: yangx@ornl.gov

Recent advances in *Populus* genomics have greatly expanded its popularity as a reference for fundamental as well as applied research in woody plants. In this review, we provide an overview of the state-of-the-science in *Populus* genomics research, including experimental and computational genomics. We have surveyed and summarized the following: 1) pioneering as well as more recent reports of genetics- and genomics-based investigations in *Populus*, 2) the positive impact of technological improvements, 3) findings from phylogenetic analyses of gene families, and 4) genomic databases. In the area of *Populus* experimental genomics, genetic approaches have been advanced to the genome scale with resolution to the gene and/or single nucleotide level. On the other hand, the modern "omics" approaches have been successfully applied to analysis of gene function, such as transcriptome profiling using microarrays as well as the next-generation DNA sequencing technology, and characterization of proteome and metabolome using modern instruments. In the area of *Populus* computational genomics, significant progresses have been made in sequence-based discovery of predicted gene function, comparative analysis of gene families, development of genomic databases, and studies of the evolutionary dynamics at both the gene and genome level. While significant advancements have been achieved in *Populus* genome-based science, several challenges need to be addressed, such as 1) better annotation of the *Populus* genome, 2) robust technology for large-scale molecular profiling, 3) efficient system for genome-wide mutagenesis, and 4) high-performance computational pipelines to keep up with the pace of the rapid accumulation of data and to integrate "omics" data into functional systems biology platforms.

Keywords *Populus*, genomics, bioinformatics, gene expression, protein, metabolomics, evolution, microRNA, microarray, gene family, systems biology

I. INTRODUCTION

Populus plants are fast-growing angiosperm trees. Owing to the perennial growth habit and wide-ranging habitat, *Populus* serves as a model for ecological genomics as well as a reference for fundamental scientific investigations of wood formation, secondary cell wall development, and morpho-physiological changes associated with seasonal variations. *Populus* has been a subject of intensive research for its end use in timber and paper-pulp industries, and recently it also garnered worldwide recognition as an important model bioenergy crop. Since the U.S. Department of Energy (DOE) announced plans to sequence the *Populus* genome in early 2002 (Wullschleger *et al.*, 2002), genomics research in *Populus* has been greatly accelerated, as reflected by the number of genomics-related papers published on *Populus* (Fig. 1). The most important hallmark of *Populus* genomics research was the publication of the *Populus* genome sequence (Tuskan *et al.*, 2006), which has already been cited more than 380 times. The availability of the *Populus* genome sequence enabled the application of high-throughput genomics technology and facilitated comparative and evolutionary genomics studies, solidifying the role of *Populus* as a reference organism for tree biology. The present review provides an overview of the state-of-the-science in *Populus* genomics research, including both experimental and functional genomics as well as the newly emerged field of computational genomics.

II. EXPERIMENTALLY-BASED FUNCTIONAL GENOMICS

A. Genetic Approaches

A primary goal of functional genomics is to identify the molecular and genetic bases of phenotypes. In *Populus* functional genomics research, genetic approaches such as random mutagenesis, identification of quantitative trait loci (QTL), characterization of nucleotide polymorphisms, and creation of transgenic lines with up- or down-regulated gene expression have helped link genetic loci and/or genes to phenotypes.

1. Mutagenesis

Insertional mutagenesis is complicated for organisms with long generation times like trees because of the difficulty in creating homozygous lines. Therefore, efforts have focused on

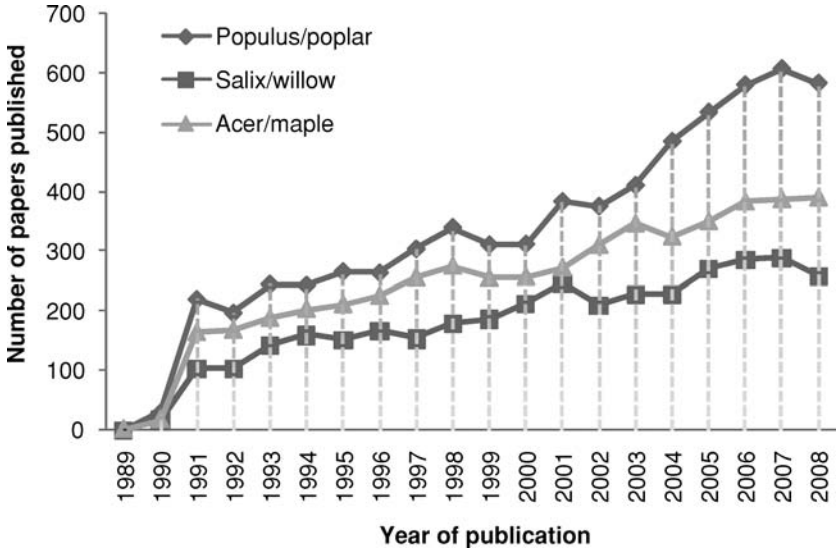


FIG. 1. Number of scientific papers related to genomics in *Populus*, *Salix*, and *Acer* published during a 20-year period from 1989 to 2008.

producing dominant mutations that can be observed in the first generation. One such approach is activation tagging, which was first developed in *Nicotiana tabacum* by Hayashi *et al.* (1992) and successfully applied to *Arabidopsis* by Weigel and colleagues (Weigel *et al.*, 2000). The general procedure of activation tagging includes: 1) introducing an enhancer cassette (typically four copies of the cauliflower mosaic virus 35S enhancer) into a plant genome by *Agrobacterium*-mediated transformation; 2) screening transgenic plants for mutants that exhibit interesting phenotypes; 3) identifying candidate gene(s) in the vicinity of the insertion site using either tail-PCR or plasmid rescue; and 4) functional characterization of the candidate gene(s) by knockdown and/or overexpression. Busov *et al.* (2003) generated 627 independent activation-tagged *Populus* lines, of which nine exhibited an obvious morphological phenotype. Recently, Harrison *et al.* (2007) produced approximately 1,800 independent activation-tagged *Populus* lines. Of the first 1,000 lines screened for developmental abnormalities, 2.4% exhibit alterations in leaf and stem structure as well as overall stature, mostly representing new phenotypes that have not previously been identified in *Populus* and/or other plants. While the activation-tagged lines provide a solid genetic basis for elucidating gene function, they are far less than required to characterize the majority of *Populus* genes, as the genome contains over 45,000 genes (Tuskan *et al.*, 2006). Therefore, a more efficient system needs to be developed to achieve saturation mutagenesis of the *Populus* genome, or substantially more resources need to be invested in activation tagging.

2. *Marker-Based Approaches*

QTLs are genomic regions associated with quantitative traits, such as yield, wood quality and resistance to abiotic and biotic stresses. QTL analyses have been invaluable in forest trees because of the insights they provide about the genetic architecture of quantitative traits (Wu and Lin, 2006). In theory, mapping of QTLs also provides the foundation for marker-assisted selection and gene discovery via positional cloning. However, several factors in forest trees work against the realization of the promise of QTLs. First, due to the aforementioned long generation times that preclude advanced generation breeding, the essential tools for positional cloning such as recombinant inbred lines and near isogenic lines are not practical in forest trees. Second, the low levels of linkage disequilibrium typical of most forest tree populations erode marker-trait associations in unstructured populations, and therefore inhibit applications of marker-assisted selection (Strauss *et al.*, 1992; Neale and Savolainen, 2004). Nevertheless, substantial insights have been gained by identifying genomic regions associated with adaptive traits, and availability of the map-anchored *Populus* genome sequence opens the possibility of linking QTLs to candidate genes, thereby linking genetic and genomic approaches to elucidating gene functions (Yin *et al.*, 2008).

Early attempts at QTL mapping in *Populus* focused on morphological and productivity-related traits in an inbred F_2 mapping population (Family 331) derived from a cross between *P. trichocarpa* 93–968 and *P. deltoides* ILL-129 (Bradshaw *et al.*, 1994), the two taxa from which most commercially-important hybrid clones have been derived. These attempts provided the first views of genetic control of growth and development, revealing a relatively small number of loci with major effects controlling growth and development (Bradshaw and Stettler, 1995; Wu, 1998). This outcome can be attributed in part to a bias against detecting loci with small effects (Beavis, 1998). QTL mapping also helped deconvolute the genetic bases of hybrid vigor in *Populus*, revealing, for example, that superior height growth was primarily conferred by the *P. trichocarpa* parent while enhanced diameter growth was provided by alleles derived from the *P. deltoids* parent (Bradshaw and Stettler,

1995). Similar observations extended to leaf morphology (Wu *et al.*, 1997) and stomatal density (Ferris *et al.*, 2002). More recent studies have largely upheld these initial findings and extended them to multiple families and environments. For example, Rae *et al.* (2008) recently performed a QTL analysis to develop new high-yield genotypes for wide-scale planting and identified 82 QTLs for eight stem and biomass traits. Directly dealing with one of the major factors limiting the deployment of QTLs in breeding, they assessed the stability of these QTLs across three contrasting environments, and found a number of "stability QTLs" that appeared at all sites (Rae *et al.*, 2008).

Important insights have also been gained from the relative positions of QTLs in *Populus*. For example, it is commonly observed that multiple related traits map to the same location in the genome, suggesting that some loci have pleiotropic effects on *Populus* development and productivity (Bradshaw and Stettler, 1995; Wullschleger *et al.*, 2005; Rae *et al.*, 2008). Co-location of QTLs for biomass production and for production of sylleptic branches (those originating from meristems produced within the same growing season) provided some evidence for the importance of sylleptic branching in heterosis of hybrid clones (Bradshaw and Stettler, 1995), a hypothesis that has since been upheld by physiological studies (Scarascia-Mugnozza *et al.*, 1999). Conversely, lack of co-location has also provided important insights. For example, Wullschleger *et al.* (2005) found separate QTLs for production of above- and below-ground biomass, suggesting separate genetic control for these components of productivity.

One of the more striking cases of map-based inferences about phenotypic traits is the discovery of a possible incipient sex chromosome in *Populus*. Although *Populus* is typically dioecious, with separate male and female sexes, morphologically distinct sex chromosomes are not apparent, leading to speculation that sex is determined by multiple autosomal loci (Alstrom-Rapaport *et al.*, 1998). However, multiple studies mapped sex determination to the same chromosomal region recently. Gaudet *et al.* (2008) applied the pseudo-test-cross strategy for linkage analysis of a F_1 pedigree obtained by crossing two genotypes of *P. nigra* from contrasting natural Italian populations, and mapped sex determination to linkage group XIX of the male parent map. Markussen *et al.* (2007) independently mapped sex determination to a similar position in a hybrid F_1 aspen (*P. tremula* x *P. tremuloides*) pedigree, demonstrating the generality of the genetic basis of this trait in the *Populus* genus. Yin *et al.* (2008) mapped sex determination to the same region in a *P. trichocarpa* x *P. deltoides* pedigree, and combined genetic, genomic, and mapping information to demonstrate that this region of linkage group XIX displays multiple characteristics of a sex chromosome, including haplotype divergence, and suppressed recombination. They suggested that *Populus* is in the process of evolving a ZW system of sex determination, with the females containing heteromorphic sex chromosomes (Yin *et al.*, 2008).

QTL studies in *Populus* have also been extended to understand plant adaptation and to pinpoint likely evolutionary responses to future climatic change. A series of experiments have examined the above- and below-ground responses of Family 331 to ambient CO_2 and elevated CO_2. Ferris *et al.* (2002) identified a total of 18 QTLs related to leaf traits, including leaf size and stomatal density. Rae *et al.* (2007) identified three QTLs on linkage groups I, IX and XII for aboveground growth and another three QTLs on linkage groups IV, XVI and XIX for root growth. To find QTLs for drought stress responses in the *Populus* genome, Street *et al.* (2006) mapped 25 QTLs in family 331 under controlled conditions, including 44 in drought. Tschaplinski *et al.* (2006) performed a large-scale drought treatment on family 331 and identified seven QTLs for osmotic potential, and two of these QTLs were later confirmed in approximately the same genomic positions by Street *et al.* (2006), who mapped 25 QTLs for physiological data in family 331 under control conditions and 44 QTLs under drought.

Progress has also been made in mapping QTLs for biotic interactions in *Populus*. In order to map the QTLs controlling resistance to leaf rust, Jorge *et al.* (2005) evaluated *P. deltoides* x *P. trichocarpa* F_1 progeny for quantitative resistance to *Melampsora larici-populina*, and detected nine QTLs, of which two had large, broad-spectrum effects. Interestingly, one of these QTLs is co-located with a QTL for formation of an ectomycorrhizal symbiosis with the fungus *Laccaria bicolor* (Tagu *et al.*, 2005). The availability of genome sequences for both of these fungi promises exciting developments in the characterization of these biotic interactions in the coming years (Martin *et al.*, 2004; Martin *et al.*, 2008; Whitham *et al.*, 2008).

One of the challenges of the 'omics era is the integration of different types of genome-scale data sets. One promising approach is the integration of metabolite profiles with QTL analysis to reveal loci that control complex metabolic pathway of closely related compounds. Tschaplinski *et al.* (2005) proposed combining metabolite profiling with QTL analysis as a novel approach to identify metabolite (m)QTL (i.e., loci that control metabolite abundance). In a preliminary assessment, the metabolite concentrations of fine roots of progeny of an interspecific backcross between *P. trichocarpa* × *deltoides* '52–225' × *P. deltoides* 'D124' were subjected to QTL analysis. mQTLs were identified for a number of secondary metabolites, including a large-effect mQTL that explained >10% of the phenotypic variation in the concentration of trichocarpinene, a secondary metabolite, and its glucoside, trichocarpin. The approach was further validated by Morreel *et al.* (2006), who identified mQTLs that control flavonoid biosynthesis in two F_1 families, *P. deltoides* cv. S9–2 x *P. nigra* cv. Ghoy and *P. deltoides* cv. S9-2 x *P. trichocarpa* cv. V24. Based on multi-trait and single-trait analyses, they identified one mQTL that explained 24% of the variation in the concentration of pinobanksin 3-acetate, one mQTL that explained 19% and 14% of the variation in the concentrations of quercetin and quercetin 3-methyl ether, respectively, and one mQTL that controlled the level of an unknown flavanone.

A similar approach relies on integration of whole-genome microarray analysis with QTL mapping. Kirst *et al.* (2005)

developed an eQTL approach to map cis- or transregulatory elements in *Eucalyptus*. Recently, Kirst and colleagues performed eQTL analysis in a *P. trichocarpa* x *P. deltoides* pseudo-backcross pedigree to identify transcriptional networks and their regulators across multiple tissues (Benedict *et al.*, 2009). While the eQTL approach looks very promising, it has not yet been widely used in *Populus*, due to a lack of cost-effective, high-performance technology for transcriptome profiling.

Due to limited recombination events in the genome, traditional QTL analysis cannot achieve a resolution at the gene and/or single nucleotide level. A promising alternative to traditional QTL analysis is association mapping based on single nucleotide polymorphisms (SNP), which takes advantage of the low linkage disequilibrium typically observed in unstructured *Populus* populations (Neale and Savolainen, 2004). In one of the first published assays of nucleotide variation in natural *Populus* populations, Gilchrist *et al.* (2006) analyzed SNPs in *Populus* using an ecotilling technique, which detects natural DNA polymorphisms using a mismatch-specific endonuclease. Specifically, they examined DNA variation in nine different genes among individuals from 41 different populations of *P. trichocarpa*, and showed that genes examined varied considerably in their level of variation, from one SNP to more than 23 SNPs per 1000-bp region. Ingvarsson (2008) demonstrated the feasibility of association studies in *P. tremula* by analyzing nucleotide polymorphism and linkage disequilibrium using multi-loci data from 77 sequence fragments, 550 bp on average. An approximate Bayesian computation was used to evaluate a number of different demographic scenarios and to estimate parameters for the best-fitting model. His analysis showed that *P. tremula* harbors substantial nucleotide polymorphism across loci and recombination rates are likely to be two to ten times higher than the mutation rate (Ingvarsson, 2008). In the first demonstration of a phenotypic association with candidate gene polymorphisms, Ingvarsson *et al.* (2008) surveyed SNPs at the phytochrome B_2 locus and identified two nonsynonymous SNPs that were independently associated with variation in the timing of bud set, explaining 1.5–5.0% of the observed phenotypic variation. These studies indicate that association studies could be a very powerful genetic approach to delineate molecular mechanisms at the single nucleotide level. An ideal resource for association studies would be a collection of genome sequencing data from each individual plant in a genetic population, which is an intimidating task due to the high cost of current genome sequencing technology. However, the future development of low-cost high-throughput $1,000 genome sequence technology (Service, 2006) would unleash the potential of association studies in *Populus* for genome-scale analysis with resolution at the single nucleotide level.

3. Transgenic Manipulation of Candidate Genes

In general, verification of gene function involves transformation to over-express and/or knockdown candidate genes in transgenic plants. One of the features that sets *Populus* apart from other tree species is the ease with which it can be manipulated in tissue culture (Taylor, 2002). Consequently, *Populus* was one of the first forest trees to be routinely transformed using *Agrobacterium*, and protocols have been developed for transformation of multiple species (Han *et al.*, 1996). Recently, Ma *et al.* (2004) and Song *et al.* (2006) independently established *Agrobacterium*-mediated transformation systems for the sequenced *P. trichocarpa* genotype Nisqually-1, in which *A. tumefaciens* C58 produced transgenic calli with regeneration efficiency of up to 13% five months after co-cultivation. More recently, Cseke *et al.* (2007) developed a transformation system for *P. tremuloides* that is suitable for high-throughput transformations using *A. tumefaciens*. This system uses *Agrobacterium*-inoculated aspen seedling hypocotyls followed by direct thidiazuron-mediated shoot regeneration on selective media, allowing fully formed transgenic trees to be generated in only three to four months.

Genetic transformation has been a primary tool for determination of gene function and genetic improvement in *Populus*, and many studies have been published since the technique was first used in the mid 1980s. One of the primary targets has been elucidation of the mechanisms of cell wall formation and lignin biosynthesis, due to the commercial importance of these traits. In the late 1990s, transgenic *P. tremuloides* plants were produced to down-regulate a caffeic acid O-methyltransferase (CAOMT) gene by homologous sense suppression (Tsai *et al.*, 1998) and a 4-coumarate:coenzyme A ligase gene by antisense inhibition (Hu *et al.*, 1999). Downregulation of cinnamoyl-CoA reductase (CCR) was achieved by Leple *et al.* (2007) in transgenic *P. tremula* x *P. alba* hybrid using antisense and sense construct via an *A. tumefaciens* procedure, with the levels of target transcript reduced down to 3 to 4% of wild-type levels. They showed that the downregulation of CCR was associated with up to 50% reduced lignin content, an orange-brown, often patchy, coloration of the outer xylem, reduced biosynthesis, and increased breakdown or remodeling of non-cellulosic cell wall polymers. Up to now, genetic engineering in *Populus* has been largely restricted to manipulation of one gene at a time. However, biological pathways generally involve multiple genes. Thus, to study gene functions at the pathway level in *Populus*, efficient and stable genetic engineering systems for manipulating multiple genes at a time need to be established in the future.

B. "Omics" Approaches

Compared to genetic approaches, "omics" approaches are relatively new in *Populus* functional genomics research. With the emergence of new life science technologies, high-throughput approaches have been utilized in *Populus* functional genomics studies, such as gene expression analysis using large-scale sequencing of expressed sequence tags (EST), microarray expression studies, protein and metabolite analysis using state-of-the-art instruments, and genome-wide identification of microRNAs using modern sequencing techniques.

1. Transcriptome Sequencing

EST profiling plays important roles in functional genomics efforts such as gene discovery, genome annotation, cDNA microarray design, and *in silico* transcript profiling. The pioneering effort in *Populus* EST sequencing was spearheaded by the Umea Plant Sciences Center in Sweden, with a primary motivation of gaining an understanding of the molecular mechanisms of cell wall formation (Sterky *et al.*, 1998). This led to the establishment of the first comprehensive public *Populus* EST source, consisting of 102,019 ESTs clustered into 11,885 clusters and 12,759 singletons, generated from 19 cDNA libraries each originating from different tissues (Sterky *et al.*, 2004). In another study aimed at understanding wood formation, Andersson-Gunneras *et al.* (2006) sequenced 5,723 ESTs from cellulose-enriched tension wood forming tissues in a *P. tremula* x *P. tremuloides* hybrid. To understand the molecular bases of the enhanced growth caused by downregulation of an enzyme in the lignin biosynthetic pathway, 4-coumarate:coenzyme A ligase in transgenic *P. tremuloides*, Ranjan *et al.* (2004) sequenced 11,308 ESTs from shoot apex, young leaf, young stem and root tip libraries, enriched by a PCR-based suppression subtractive hybridization between control and transgenic plants. The ESTs were clustered and assembled into 5,028 nonredundant transcripts, with a large number of ESTs (16%) associated with signal transduction in transgenic leaves, as well as some homologs of transposable elements upregulated in transgenic tissues. This effort was extended with 5,410 additional ESTs from two hybrids of additional *Populus* species, *P. angustifolia* and *P. fremontii*, which differ markedly in their phenylpropanoid profiles (Harding *et al.*, 2005).

EST sequencing has also been used to gain insights into responses to abiotic and biotic stresses in *Populus*. Nanjo *et al.* (2004) sequenced over 30,000 ESTs from *P. nigra* leaves treated with dehydration, chilling, high salinity, heat, ABA or H_2O_2, and discovered over 4,500 nonredundant full-length cDNAs that were responsive to these stresses. To facilitate gene discovery related to water and nutrient uptake and assimilation in roots, Kohler *et al.* (2003) sequenced 7,013 ESTs representing 4,874 unique transcripts (1,347 clusters and 3,527 singletons) in the roots of *P. trichocarpa* x *P. deltoides*, with 6% of the ESTs assumed to be associated with root functions. Brosché *et al.* (2005) sequenced 13,838 ESTs from 17 libraries derived from *P. euphratica* trees exposed to a variety of stresses. This study was particularly interesting because *P. euphratica* is a unique species in the genus, occurring in desert environments where it is exposed to drought and salinity stress, in stark contrast to the mesic environments where most *Populus* species occur. This study resulted in the identification of 7,841 unigene clusters, 26% of which were novel *Populus* transcripts (Brosche *et al.*, 2005).

To characterize inducible defenses against insect herbivory in *Populus*, Ralph *et al.* (2006) developed an EST resource as a complement of the existing *Populus* genome sequence and *Populus* ESTs by focusing on herbivore- and elicitor-treated tissues and incorporating normalization methods to capture rare transcripts. They generated 139,007 3′- or 5′-end sequenced ESTs from 15 cDNA libraries and assembled the 107,519 3'-end ESTs into 14,451 contigs and 20,560 singletons, representing 35,011 putative unique transcripts. This resource was recently expanded using full-length cDNAs derived from insect-attacked leaves of the *P. trichocarpa* x *P. deltoides* hybrid (Ralph *et al.*, 2008).

Other *Populus* EST resources have been created primarily to enhance gene prediction and annotation. Unneberg *et al.* (2005) created nine cDNA libraries from *P. tremula* and *P. trichocarpa*, and sequenced 70,747 ESTs that were clustered into 14,213 putative genes. Nanjo *et al.* (2007) sequenced female *P. nigra* var. *italica* full-length cDNA libraries, and generated about 116,000 5′- end or 3′-end ESTs that were assembled into 19,841 nonredundant full-length clones. Another 4,664 full-length cDNAs were generated recently by Ralph *et al.* (2008) using the biotinylated CAP trapper method from xylem, phloem and cambium, green shoot tips, and leaves from the *P. trichocarpa* genotype Nisqually-1.

Using public *Populus* EST resources, Moreau *et al.* (2005) performed *in silico* transcript profiling in the woody tissues in relation to programmed death of xylem fibers and identified a large number of previously uncharacterized transcripts possibly related to the death of xylem fibers. To identify expression of genes encoding carbohydrate-active enzymes in the *P. trichocarpa*, Geisler-Lee *et al.* (2006) compared EST frequencies in a collection of 100,000 ESTs from 17 different tissues, and showed that genes involved in pectin and hemicellulose metabolism were expressed in all tissues, indicating that these genes are essential for the development of cell wall matrix. The EST data also indicated that sucrose synthase genes were highly expressed in wood-forming tissues along with cellulose synthase and homologs of KORRIGAN and ELP1 whereas the expression levels of genes related to starch metabolism were low during wood formation, indicating the preferential flux of carbon to cell wall biosynthesis.

The aforementioned traditional EST sequencing approach is based on vector-cloning, which generally misses a significant portion (∼40%) of transcripts even by sequencing millions of cDNA clones from multiple tissues (Sun *et al.*, 2004; Gowda *et al.*, 2006). It is also expensive due to the use of the Sanger sequencing method. Fortunately, next-generation DNA sequencing technology such as 454 sequencing makes it feasible to directly sequence plant transcriptomes in a cost-effective and efficient manner (Emrich *et al.*, 2007). In a collaboration between Oak Ridge National Laboratory and the DOE Joint Genome Institute, transcriptome sequencing using 454 DNA sequencing technology was performed to profile gene expression associated with dehydration response in *Populus deltoides*. About 2.6 million sequencing reads were obtained from six leaf cDNA libraries. These 454 sequencing reads were mapped to the JGI *Populus* genome browser (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.info.html), providing experimental

support for approximately 50% of the *ab initio* gene models in *Populus*. A set of genes which were upregulated by dehydration were identified by bioinformatic analysis of the 454 sequencing reads using a computational pipeline developed at Oak Ridge National Laboratory. The gene ontology analysis revealed that the genes relevant to drought response were enriched in this gene set (Yang *et al.*, unpublished). The success of this 454 sequencing project demonstrated that sequencing-based profiling is an excellent approach for the quantitative analysis of *Populus* gene expression.

2. Hybridization-Based Transcript Profiling

Microarrays provide high-throughput transcript profiling based on hybridization of labeled transcripts to glass slides. The two major classes of arrays are cDNA arrays, based on spotting portions of cDNA clones directly onto slides, and oligonucleotide arrays, which have shorter probes that may be synthesized in situ or spotted. The first generation of *Populus* arrays was a cDNA microarray, the 13K POP1 array, developed by Andersson *et al.* (2004) based on 13,490 unigenes assembled from 36,354 ESTs. The POP1 array was used in transcript profiling across the wood-forming meristem to identify potential regulators of cambial stem cell identity (Schrader *et al.*, 2004), during tension wood formation to identify the genes responsible for the change in carbon flow into various cell wall components (Andersson-Gunneras *et al.*, 2006), to characterize drought responses (Street *et al.*, 2006), in isolated cambial meristem cells during dormancy to better understand the environmental and hormonal regulation of this process (Druart *et al.*, 2007), and in studying the auxin-regulated wood formation process (Nilsson *et al.*, 2008). The second generation *Populus* cDNA microarray, the 25K POP2 array, was developed by Moreau *et al.* (2005) based on 24,735 different cDNA fragments. It was used in transcript profiling in leaf tissue of *P. deltoides* to monitor nocturnal changes in gene expression during leaf growth (Matsubara *et al.*, 2006), in a subset of extreme genotypes exhibiting extreme sensitivity and insensitivity to drought in the mapping population Family 331 (Street *et al.*, 2006), in apical bud formation and dormancy induction (Ruttink *et al.*, 2007), in a transgenic *P. tremula* x *P. alba* hybrid with downregulated expression of cinnamoylcoenzyme A reductase to investigate how CCR downregulation impacted metabolism and the biosynthesis of other cell wall polymers (Leple *et al.*, 2007), and in leaves of free-growing *P. tremula* throughout multiple growing seasons (Sjodin *et al.*, 2008).

Brosché *et al.* (2005) constructed a 6K *P. euphratica* cDNA microarray representing 6,340 unigenes assembled from the EST resource described above. This array was used to assess gene expression in adult *P. euphratica* trees growing in the desert (Brosche *et al.*, 2005; Bogeat-Triboulot *et al.*, 2007) and young plants submitted to a gradually increasing water deficit for four weeks in a greenhouse (Bogeat-Triboulot *et al.*, 2007).

Ralph *et al.* (2006) developed a 15.5K *Populus* cDNA microarray containing 15,496 unigenes assembled from 107,519 3'-end ESTs obtained from 15 cDNA libraries, and utilized it to monitor gene expression in *Populus* leaves in response to herbivory by forest tent caterpillars (*Malacosoma disstria*). The 15.5K *Populus* microarray was also used by Miranda *et al.* (2007) to study the transcriptional response of *P. trichocarpa* x *P. deltoides* hybrid to infection by leaf rust (*Melampsora medusa*).

Multiple whole-genome oligonucleotide microarrays have been developed for *Populus* (reviewed by Tsai *et al.*, 2009). The first whole genome oligonucleotide microarray was designed by Oak Ridge National Laboratory in collaboration with NimbleGen (Madison, WI, USA). This array contained three different 60-mer probes for every predicted gene model in the *P. trichocarpa* genome, as well as nearly 10,000 divergent transcripts from the *P. tremula*, *P. tremuloides*, and *P. alba* EST resources described above (Groover *et al.*, 2006). This array has been used in transcript profiling in *P. trichocarpa* to study expression of invertase genes (Bocock *et al.*, 2008), in 14 different tissues of *P. trichocarpa* genotype Nisqually-1 to study expression of auxin response regulators genes (Kalluri *et al.*, 2007), in vegetative organs of the *P. trichocarpa* genotype Nisqually-1 for comparative analysis of the transcriptomes of *P. trichocarpa* and *Arabidopsis thaliana* (Quesada *et al.*, 2008), in five different tissues (young leaves, mature leaves, nodes, internodes and roots) of *P. trichocarpa* Nisqually-1 to study the expression pattern of the cytokinin response regulator gene family (Ramirez-Carvajal *et al.*, 2008), and in *Populus* leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina* (Rinaldi *et al.*, 2007).

Affymetrix has also produced a *Populus* whole-genome array based on their photolithographic fabrication technique. This array targets 61,251 predicted genes, including 47,835 from the *P. trichocarpa* genome sequence, and the remainder representing divergent unigenes predicted from an assembly of over 260,000 ESTs from 13 *Populus* species. The array has eleven 25-mer probes targeting each predicted gene, preferentially selected from the 3′ end of each predicted transcript. This array has been used to characterize responses to nitrogen stress (Qin *et al.*, 2008), and to explore patterns of expression of R2R3 Myb transcription factors in *Populus* (Wilkins *et al.*, 2009).

A third *Populus* whole-genome array has recently been produced by Agilent in a four-plex format targeting 43,803 genes with a single 60-mer probe per gene. This subset of the *Populus* gene models was selected by excluding gene models that showed high homology to transposable elements or to bacterial, fungal, or mammalian sequences that were likely contaminants of the original sequencing template (Tuskan *et al.*, 2006). As of this writing, there are not yet any published studies using this array, but several are in progress (C.J. Tsai, personal communication).

The diversity of *Populus* array designs presents challenges for the integration of data across platforms. A web-based tool, PopArray (http://aspendb.uga.edu) has recently been created to facilitate cross-referencing of probes across platforms (Tsai *et al.*, 2009). However, each of these microarray platforms/systems still suffers from several limitations. First is

a lack of full-genome coverage because the probes were designed according to either limited EST information or the first version of the *Populus* genome annotation, which contains many incomplete gene models. Second, some of the probes are not gene-specific due to high sequence identity of recently duplicated genes in the genome. Finally, splice variants are not well represented. This latter problem is compensated somewhat by a new version of the *Populus* NimbleGen array that has seven independent 60-mer probes per gene target (http://www.nimblegen.com/products/exp/custom.html). Future array designs will benefit from deep transcriptome sequencing data and the release of the next *Populus* genome annotation in the near future.

3. PCR-Based Transcript Profiling

Traditional techniques like differential display still have a role to play in the genomics era (Liang, 2002), even in a genus like *Populus* in which a species has been fully sequenced, and large numbers of ESTs are available. These enrichment techniques allow efficient use of sequencing resources, allowing researchers to focus on transcripts that are up- or down-regulated in response to specific treatments. For example, Caruso *et al.* (2008) searched for genes differentially expressed in response to drought in young rooted cuttings under PEG 6000 treatment using the differential display technique and identified 36 differentially expressed leaf cDNAs between stressed and control conditions. Another useful technique is cDNA-amplified fragment length polymorphism (AFLP) transcript profiling. It has been used to map differential gene expression during dormancy induction, dormancy, dormancy release by chilling, and subsequent bud break in apical buds of a *P. tremula* × *P. alba* hybrid, revealing novel genes linked to a crucial transitory step in dormancy induction, and to dormancy release through chilling (Rohde *et al.*, 2007). Finally, Zhuang and Adams (2007) used a clever application of the SNaPshot SNP genotyping assay (Applied Biosystems) to reveal allelic variation in gene expression in a *P. trichocarpa* × *P. deltoides* F_1 hybrid. Using this single-base primer extension assay with extension primers designed to anneal to the amplified DNA adjacent to the SNP site, they identified cis-regulation for six genes, trans-regulation for one gene, and combined cis- and trans-regulation for nine genes, demonstrating that species-specific alleles can have variable expression patterns depending on the genetic background (Zhuang and Adams, 2007).

4. Protein Profiling

Protein profiling offers multiple advantages over transcript profiling for characterizing the functional state of an organism at a given point in time. First, a given transcript can give rise to different proteins due to post-translational modifications, which will only be apparent with a direct observation of the protein. Furthermore, mRNA levels may be a poor indicator of protein levels due to differential rates of turnover of mRNA compared to proteins (e.g., a long-lived protein may be in relatively high abundance, even though the mRNA might be in low abundance due to degradation of the transcript) (Pandey and Mann, 2000). Therefore, extensive efforts have been undertaken to characterize protein responses at the whole genome scale in model organisms, including *Populus*.

Two studies have examined the proteomic responses of *P. euphratica* to experimentally-induced heat and drought stress. Ferreira *et al.* (2006) studied protein accumulation profiles of leaves from young *P. euphratica* plants submitted to 42/37°C for three days using two-dimensional electrophoresis. They detected up- or down-regulation of 45% of the 1,355 spots assayed by 2-dimensional gel electrophoresis (2-DE), and identified 51 out of 62 selected spots using matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) analysis. They showed that short-term up-regulated proteins were related to membrane destabilization and cytoskeleton restructuring, sulfur assimilation, thiamine and hydrophobic amino acid biosynthesis, and protein stability; long-term upregulated proteins were involved in redox homeostasis and photosynthesis. Late downregulated proteins were involved mainly in carbon metabolism, indicating that moderate heat response involves proteins related to lipid biogenesis, cytoskeleton structure, sulfate assimilation, thiamine and hydrophobic amino acid biosynthesis, and nuclear transport. Bogeat-Triboulot *et al.* (2007) determined the expression profiles of proteins in mature leaves at four stress levels and after recovery in young, vegetatively propagated *P. euphratica* plants that were submitted to a gradually increasing water deficit for four weeks in a greenhouse and were allowed to recover for ten days after full re-irrigation. Using 2-DE, they identified 375 spots that were responsive to the treatments, and characterized 100 of these using MALDI-TOF-MS and MALDI-TOF/TOF-MS analysis, resulting in identification of 39 drought-responsive proteins, including proteins related to energy and C metabolism, and proteins involved in glycolysis. Surprisingly, protein levels were not correlated with stress levels, in contrast to transcript abundance as determined by microarrays, which were directly correlated with the level of water deficit (Bogeat-Triboulot *et al.*, 2007).

Plomion *et al.* (2006) performed a more comprehensive analysis of the *Populus* proteome. They first created highly reproducible and well-resolved 2-DE maps of proteins for eight tissues/organs of adult *P. trichocarpa* (both male and female) plants and 2-month-old rooted cuttings of a *P. trichocarpa* x *P. deltoides* hybrid. They excised 398 spots from the 2-DE gels and identified 363 proteins (∼91.2%) by nanospray LC-MS/MS, based on comparison with 260,000 *Populus* ESTs. In order to assess the resolution of protein detection based on peptide mass fingerprinting (PMF) and compare the identification rate to that obtained by LC-MS/MS, they generated MALDI-TOF-MS profiles for 320 spots and obtained reliable PMFs for only 163 spots (51%), from which about half (83 spots) positively matched gene models from the *P. trichocarpa* genome sequence.

Du *et al.* (2006) analyzed proteins expressed in different wood regeneration stages in a system that can mimic the

initiation and differentiation of cambium cells for *P. tomentosa*. They obtained PMFs for 244 differentially-expressed proteins and assigned putative functions for 199 of these proteins. They showed that regulatory genes for cell cycle progression, differentiation and cell fate were expressed during formation of cambial tissue, while genes involved in secondary wall formation were predominantly found in the xylem developmental stage, indicating that changes in patterns of gene expression correspond to developmental stages of the secondary vascular system.

Sasaki *et al.* (2007) surveyed the localization of anionic peroxidase isoenzymes, which are important for lignification, in various organs of *P. alba* using 2-DE followed by PMF analysis. They showed that the expression profile of each isoenzyme was quite different, suggesting that individual anionic isoenzymes are differently regulated at transcription, translation, or posttranslational level.

Availability of a high-throughput, broad-spectrum proteome profiling method is highly desirable to improve turnaround time and comprehensiveness of a proteomics approach. A shotgun proteome profiling method has recently been reported for *Populus* (Kalluri *et al.*, 2009, in review). In this study, MudPIT technique was applied to profiling subcellular protein fractions of developing xylem. Identification of nearly 6,000 proteins using this method significantly increases the number of proteins validated from *Populus* beyond what was previously reported from 2-DE-based approaches.

5. *MicroRNA Profiling*

MicroRNAs (miRNA) are small RNAs approximately 21 nucleotides in length that negatively control gene expression by cleaving or inhibiting the translation of target gene transcripts (Barakat *et al.*, 2007). To test whether miRNAs play roles in the regulation of wood development in tree species, Lu *et al.* (2005) isolated small RNAs from the developing xylem of *P. trichocarpa* stems and cloned 22 miRNAs, which are the founding members of 21 miRNA gene families for 48 miRNA sequences. Their computational prediction revealed that a majority of these miRNAs potentially target developmental- and stress/defense-related genes. Ko *et al.* (2006) cloned the microRNA 166 (Pta-miR166) families from *P. tremula* × *P. alba* hybrid using a combination of *in silico* and PCR-based methods. They showed the expression of class III HD-Zip transcription factor (PtaHB1) was inversely correlated with the level of Pta-miR166, which directed the cleavage of PtaHB1 *in vivo*, as confirmed using modified 5′-rapid amplification of cDNA ends. They also found that the expression of Pta-miR166 was much higher in the winter than during the growing season, suggesting seasonal and developmental regulation of microRNA. Recently, a high throughput pyrosequencing technique was used by Barakat *et al.* (2007) to profile small RNAs from leaves and vegetative buds of *Populus*. After analysis of 80K small RNA reads, they identified 123 new sequences belonging to previously identified miRNA families as well as 48 new miRNA families that could be *Populus*-specific. They also identified putative targets of nonconserved miRNA including both previously identified targets as well as several new putative target genes involved in development, resistance to stress, and other cellular processes. They showed that almost half of the genes predicted to be targeted by nonconserved miRNAs appear to be *Populus*-specific.

6. *Metabolite Profiling*

The metabolome is the quantitative complement of all of the low-molecular weight molecules present in a cell in a given physiological state, representing the products of cellular biochemical processes. As such, analysis of the metabolome potentially provides an even finer snapshot of an organism's physiological state than proteomic analysis (Fiehn, 2002). Advances in metabolite profiling have proceeded rapidly in the past decade, raising the possibility of simultaneously assaying the levels of thousands of compounds in a high-throughput manner (Weckwerth, 2003). This great promise is beginning to be realized in model organisms like *Populus*, as tools and metabolite databases rapidly accumulate, and metabolomic analyses are being used to dissect phenotypes with unprecedented precision.

A major use of metabolomics is the characterization of biochemical changes that occur following experimental down- or up-regulation of candidate gene expression. For example, Busov *et al.* (2006) performed metabolic profiling to gain insight into the biochemical changes associated with the dramatic architectural changes of the dwarfed transgenic plants overexpressing *Arabidopsis gai* and *rgl1* using gas chromatography–mass spectrometry (GC-MS). They showed that transgenic plants had increased concentrations of citric acid and several amino acids, including asparagine and arginine, and two unidentified glucosides, but reduced concentrations of monosaccharides in roots. The combined responses were indicative of increased respiratory consumption of monosaccharides to generate Krebs cycle organic acids that are required for amino acid synthesis and root growth. In leaves, the concurrent decline in glutamine and other N-containing metabolites, including phenylalanine, was consistent with increased N allocation to roots via perturbations in the secondary carbon pathways. The transgenic dwarf plants displayed increased concentrations of various products and intermediates of the phenylpropanoid biosynthetic pathway, such as the accumulation of syringin (sinapyl alcohol glucoside), likely reflected the reduced shoot growth of the transgenics, leading to a buildup of a storage form of the monolignol precursor. Phenolic glucosides that are associated with defense, including salicin and tremulacin, were similarly found at much higher levels in both *gai* and *rgl1* expressing plants. The accumulation of 3-O-caffeoylquinic acid that results from the conjugation of a key phenolic acid precursor of monolignol biosynthesis with an upstream organic acid intermediate of the shikimic acid pathway, coupled with declines of the monomers including quinic acid and other phenolic acid conjugates, may have been indicative of the reduced carbon flux through the lignin biosynthetic pathway. A similar shift in carbon partitioning away from the

upstream lignin precursor conjugates was revealed by metabolite profiling of poplar transgenic plants expressing a bacterial nahG gene encoding salicylate hydroxylase that converts salicylic acid to catechol (Morse *et al.*, 2007). Expression of nahG decreased quinic acid-phenolic acid conjugates and phenolic acid-glucosides, including salicylate glucoside, but increased catechol glucoside, while exerting little effect on levels of salicylic acid and catechol, the substrate and product, respectively, of the nahG enzyme.

To investigate the effects of downregulating cinnamoyl-CoA reductase (CCR) on metabolism in transgenic *P. tremula* x *P. alba*, Leple *et al.* (2007) analyzed metabolome of young developing xylem of wild-type and CCR-downregulated lines using GC-MS followed by principal component analysis. They identified 20 known metabolites that were accumulated differentially in the CCR-downregulated lines compared with the wild type, with the largest fraction of differential metabolites being carbohydrates such as glucose, mannose, galactose, myo-inositol, raffinose, and melezitose, reflecting changes in central carbohydrate metabolism. Most notably, CCR-downregulated mutants had large accumulations in 4-O-β-D-glucopyranosyl sinapic acid and 4-O-β-D-glucopyranosyl vanillic acid. These were the same phenolic acid glucosides that accumulated in caffeoyl-CoA O-methyltransferase (CCoAOMT) mutants (Meyermans *et al.*, 2000). To identify the metabolites that control bud development in *P. tremula* x *P. alba*, Ruttink *et al.* (2007) analyzed developing buds of wild-type and transgenic plants that upregulate or downregulate the ABI3 transcription factor at weekly intervals during 6 weeks of short-day treatment with GC-MS, and quantified 8,852 m/z peaks, of which 1,702 m/z peaks have significant changes either between any two genotypes at a given time, or between any two sampling points for a given genotype. The 1,702 m/z peaks corresponded to 176 metabolites, of which 162 had more than a fourfold differential accumulation, including 110 unidentified compounds (67.9%), 13 organic acids (8.0%), 16 amino acids (9.9%), and 14 sugars or sugar alcohols (8.6%).

Metabolic profiling has also been a valuable tool for characterizing developmental processes in *Populus*. For example, Andersson-Gunneras *et al.* (2006) performed metabolite analysis in developing *P. tremula* tension wood using gas chromatography/time-of-flight mass spectrometry. They detected more than 350 peaks and revealed that 26 metabolites were significantly changed, among which sucrose, arabinose, inositol, shikimate, monolignols and gamma-butyric acid were decreased in tension wood, while xylose, xylitol and two fatty acid metabolites were more abundant in tension wood. Gou *et al.* (2008) analyzed the cell-wall acylesters of *P. trichocarpa* with liquid chromatography-mass spectrometry, Fourier transform-infrared microspectroscopy, and synchrotron infrared imaging facility. They showed that the cell wall of *Populus* contained a considerable amount of acylesters, primarily acetyl and p-hydroxycinnamoyl molecules and the "wall-bound" acetate and phenolics displayed a distinct tissue specific-, bending stress responsible- and developmental-accumulation pattern, indicating that different "wall-bound" acylesters play distinct roles in *Populus* cell wall structural construction and/or metabolism of cell wall matrix components. To characterize the environmental and hormonal regulation of dormancy in perennial plants, Druart *et al.* (2007) performed metabolite profiling of isolated cambial meristem cells during the course of their activity-dormancy cycle using GC-MS analysis. They detected more than 1,000 peaks, of which 227 changed significantly in one or more pairwise sample class comparisons. The responsive metabolites were classified into four groups: carbohydrates, organic acids, amines (amines and amino acids) and sterols.

7. "Omics" Application in Wood Formation and Secondary Cell Wall Formation

The occurrence of extensive secondary xylem development (wood formation) where xylem developmental phase transitions can be distinguished spatially across various radial growth layers makes/renders *Populus* as an excellent model to study molecular dynamics that underlie the processes of secondary cell wall formation and wood development. Several studies based on targeted expression profiling techniques of RT-PCR and in-situ hybridization (Kalluri and Joshi, 2004) as well as broad spectrum expression profiling based on ESTs (Sterky *et al.*, 2004) and microarray technologies (Nilsson *et al.*, 2008) have provided insights into genes that are important to secondary xylem development. Some of the relevant gene families have also been reported in greater detail elsewhere in this review, including those involved in carbohydrate biosynthesis process (Geisler-Lee *et al.*, 2006), phenyl propanoid pathway (Tsai *et al.*, 2006), transcription regulation (Arnaud *et al.*, 2007), cell cycle (Espinosa-Ruiz *et al.*, 2004), and structural proteins (Oakley *et al.*, 2007).

III. COMPUTATIONAL GENOMICS

A. Sequence-based Discovery

The availability of transcriptome and genome sequencing data opened the door for new discoveries using computational approaches, such as analysis of differential gene expression using *in silico* transcript profiling, large-scale prediction of cis-elements in genomic regions upstream of the transcription start sites, identification of lineage-specific motifs and genes by comparing gene sequences among different species.

1. Identification of Putative Cis-regulatory Sequences

To identify putative cis-regulatory sequences in the cellulose synthase (CesA) gene family which encodes the catalytic subunits of a large protein complex responsible for the deposition of cellulose into plant cell walls, Creux *et al.* (2008) carried out a comparative sequence analysis of orthologous CesA promoters from *Arabidopsis*, *Populus* and *Eucalyptus*, and identified 71 conserved sequence motifs, of which 66 were

significantly over-represented in either primary or secondary wall-associated promoters. Krom and Ramakrishna (2008) studied the co-expression and interspecies conservation of divergent and convergent gene pairs in the *Oryza*, *Arabidopsis* and *Populus* genomes. They showed that strongly correlated expression levels between divergent and convergent genes were quite common in all three species, suggesting that shared as well as unique mechanisms operate in shaping the organization and function of divergent and convergent gene pairs in different plant species. They also identified 56 known regulatory elements overrepresented in the intergenic regions of divergent genes (separated by 1 kb or less) with correlated expression in *Arabidopsis* (39 elements), *Oryza* (16 elements), and *Populus* (1 element) (Krom and Ramakrishna, 2008).

2. Analysis of Alternative Splicing in Transcripts

Baek *et al.* (2008) analyzed the structure of 5′-splice junctions in *Medicago truncatula*, *P. trichocarpa*, *A. thaliana*, and *O. sativa* and observed commonalities between the species. They found that for *M. truncatula*, *P. trichocarpa* and *A. thaliana*, but not in *O. sativa*, alternative splicing was most prevalent for introns with decreased UA content.

3. Identification of Lineage-specific Motifs and Genes

Although proteins lacking currently defined motifs or domains (POFs) appear similar to proteins with experimentally defined domains or motifs (PDFs) in their relative contribution to biological functions, the POFs have more predicted disordered structure than the PDFs, implying that they may exhibit preferential involvement in species-specific regulatory and signaling networks (Gollery *et al.*, 2006). Gollery *et al.* (2007) performed a comparative analysis of POFs and PDFs in the predicted proteomes derived from *Arabidopsis*, *Oryza* and *Populus*, finding that $>$26% of the *Populus* proteome was comprised of POFs, compared with 19% and 33% of the *Arabidopsis* and *Oryza* proteomes, respectively. In a comparison among the three plant proteomes, $\sim$75% of the unique proteins of *Populus* were POFs. Due to higher proportion of disordered structures in POFs, the shorter length of POFs compared with PDFs, and the low level of sequence similarity between POFs from different organisms, most POFs exist in plants as singletons. It was therefore hypothesized that POFs could represent newly evolving genes or genes that are evolving much faster than the genome average, suggesting that these unique *Populus* proteins could be lynchpins of the evolutionary process in this genus (Gollery *et al.*, 2006; Gollery *et al.*, 2007).

Recently, Yang *et al.* (2009) conducted a genome-wide analysis of lineage specific genes in *Arabidopsis*, *Oryza* and *Populus* and identified three differential gene (DG) sets: i) 917 *Arabidopsis* genes without homologues in *Oryza* or *Populus*, ii) 2,781 *Oryza* genes without homologues in *Arabidopsis* or *Populus*, and iii) 594 *Populus* genes without homologues in *Arabidopsis* or *Oryza*. Furthermore, they used the DG sets to search against a plant transcript database, NR protein database, NCBI dbEST and six newly sequenced genomes (*Carica*, *Glycine*, *Medicago*, *Sorghum*, *Vitis* and *Zea*) and identified 165, 638 and 109 species-specific genes (SS) genes in *Arabidopsis*, *Oryza* and *Populus*, respectively. They showed that some SS genes were preferentially expressed in flowers, roots, xylem and cambium or upregulated by stress, reflecting functional and/or anatomical differences between monocots and eudicots or between herbaceous and woody plants.

B. Phylogenetic Analysis of Gene Families

Phylogenetic analyses of individual gene families have been playing a very important role in *Populus* genomics since the draft genome sequence became available. Thus far analyses of about 50 gene families containing more than 2,400 genes in *Populus* have been published (Table 1). These gene families are involved in various biological processes, as discussed in the following sections.

1. Transcription Regulation

MADS-box transcription factor genes control diverse developmental processes in flowering plants ranging from root to flower and fruit development in plants (Becker and Theissen, 2003). Leseberg *et al.* (2006) identified 105 putative functional MADS-box genes and 12 pseudogenes in the *Populus* genome, comparable to those in *Arabidopsis* (107 genes) (Parenicova *et al.*, 2003). They constructed a phylogenetic tree using all MADS-box genes from *Arabidopsis* and *Populus* and classified the *Populus* MADS-box genes using the *Arabidopsis* MADS-box gene dataset. Their phylogenetic analysis revealed that *Populus* has 64 type II MADS-box genes, implying a higher birth rate when compared with *Arabidopsis* (64 vs. 47). In contrast, there were 41 putative functional type I genes and 9 type I pseudogenes in *Populus*, suggesting that the *Populus* type I MADS-box genes have experienced a high death rate, but a relatively lower birth rate, leading to a smaller number of type I genes in *Populus* than in *Arabidopsis* (41 vs. 60). They also found that the *Populus* MADS-box gene family has become expanded through tandem gene duplication and segmental duplication events. Recently, Diaz-Riquelme *et al.* (2009) performed a phylogenetic analysis of full-length *Vitis*, *Arabidopsis* and *Populus* MIKC^C-type MADS box protein sequences, dividing the MIKC gene family into 13 subfamilies. All *Vitis* MIKC genes were grouped with their *Arabidopsis* and *Populus* counterparts, and in most cases, two *Populus* genes were found for every homolog in *Vitis* or *Arabidopsis*, consistent with the recent discovery that *Populus* experienced a recent genome-wide duplication event (Tuskan *et al.*, 2006; Tang *et al.*, 2008a; 2008b).

The two related transcription factor gene families, Auxin/Indole-3-Acetic Acid (Aux/IAA) and Auxin Response Factor (ARF), regulate auxin-induced gene expression, each with unique localized functions as well as overlapping redundant functions (Liscum and Reed, 2002). Kalluri *et al.* (2007) identified 35 Aux/IAA and 39 ARF genes in the *Populus* genome.

TABLE 1
Populus gene familes studied by phylogenetics analysis

Category	Gene family	References
Biosynthesis of structural components	alpha- and beta-tubulin	Oakley *et al.* (2007)
Biosynthesis of carbohydrates	Starch branching enzymes	Han *et al.* (2007)
Biosynthesis of carbohydrates	Invertase	Bocock *et al.* (2008)
Biosynthesis of carbohydrates	endo-beta-mannanase	Yuan *et al.* (2007)
Biosynthesis of carbohydrates	Cellulose synthase	Suzuki *et al.* (2006)
Biosynthesis of carbohydrates	Xyloglucan endo-transglycosylases	Baumann *et al.* (2007)
Biosynthesis of carbohydrates	COBRA	Ye *et al.* (2009a)
Biosynthesis of hormone	YUCCA	Ye *et al.* (2009b)
Cell cycle	NIMA-related kinase	Vigneault *et al.* (2007)
Cell cycle	D-type cyclins	Menges *et al.* (2007)
Disease resistance	NBS resistance	Kohler *et al.* (2008)
Fatty acid metabolism	acyl:coenzyme A synthetase	Souza Cde *et al.* (2008)
Flower development	FT/TFL1	Igasaki *et al.* (2008)
Photorespiration	Glycine decarboxylase complex	Rajinikanth *et al.* (2007)
Protein degradation	F-box	Yang *et al.* (2008)
Protein degradation	Kunitz trypsin inhibitor	Major and Constabel (2008)
Protein degradation	Protease	Garcia-Lorenzo *et al.* (2006)
Regulating transcription	AP2/ERF	Zhuang *et al.* (2008)
Regulating transcription	MADS-box	Leseberg *et al.* (2006); Diaz-Riquelme *et al.* (2009)
Regulating transcription	Aux/IAA	Kalluri *et al.* (2007)
Regulating transcription	ARF	Kalluri *et al.* (2007)
Regulating transcription	LIM	Arnaud *et al.* (2007)
Regulating transcription	DOF	Yang *et al.* (2006)
Regulating transcription	Cytokinin response regulator	Ramirez-Carvajal *et al.* (2008)
Regulating transcription	R2R3-MYB	Wilkins *et al.* (2008)
Signal transduction	LysM kinase	Zhang *et al.* (2007)
Signal transduction	Calcineurin B-Like	Zhang *et al.* (2008)
Phenylpropanoid metabolism	Arogenate dehydratase (ADT)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	beta-Alanine N-methyltransferase (NMT)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Flavonoid 3′,5′-hydroxylase (F3′5′H)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Flavone synthase II (FNSII)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Flavanone 3-hydroxylase (F3H)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Flavonol synthase (FLS)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Dihydroflavonol 4-reductase (DFR)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Anthocyanidin synthase (ANS)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Anthocyanidin reductase (ANR/BAN)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Leucoanthocyanidin reductase (LAR)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	Flavonoid O-methyltransferase (FOMT)	Tsai *et al.* (2006)
Phenylpropanoid metabolism	trans-Cinnamate 4-hydroxylase (C4H)	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Coumarate 3-hydroxylase (C3H)	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Caffeic acid O-methyltransferase (COMT)	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Ferulate 5-hydroxylase (F5H)	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Caffeoyl-CoA O-methyltransferase	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	hydroxycinnamoyltransferase (HCT)	Tsai *et al.* (2006); Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Aldehyde dehydrogenase (ALDH)	Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Cinnamoyl CoA reductase (CCR)	Hamberger *et al.* (2007)
Phenylpropanoid metabolism	NADPH-cytochrome P450 oxydoreductase	Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Cinnamyl alcohol dehydrogenase-related	Hamberger *et al.* (2007)
Phenylpropanoid metabolism	Phenylalanine amonnia lyase (PAL)	Hamberger *et al.* (2007)
Phenylpropanoid metabolism	4-Coumarate:CoA Ligase (4CL)	Hamberger *et al.* (2007)

From the phylogenetic tree reconstructed from *Populus*, *Arabidopsis* and *Oryza* Aux/IAA amino acid sequences, they found that four groups of *Populus* Aux/IAAs (PoptrIAA3, 16, 27 and 29) expanded to contain three or more members each. Similarly, the phylogeny of *Populus*, *Arabidopsis* and *Oryza* ARF protein sequences revealed differential expansion or contraction between *Arabidopsis* and *Populus*. For example, the number of activator ARFs (defined by the Q-rich middle region) in *Populus* is 2.6 times that in *Arabidopsis* whereas the ratio of repressor and other ARFs between *Arabidopsis* and *Populus* is 1:1.4. Kalluri *et al.* (2007) also showed that the differential expansion or contraction in Aux/IAA and ARF between *Arabidopsis* and *Populus* was caused by high segmental and low tandem duplication events in *Populus*. Furthermore, their expression studies showed that genes in the expanded PoptrIAA3 subgroup display differential expression.

Genes in the AP2/Ethylene Response Factor (ERF) family encode transcriptional regulators with a variety of functions involved in developmental and physiological processes in plants (Okamuro *et al.*, 1997; Nakano *et al.*, 2006). Zhuang *et al.* (2008) identified 200 AP2/ERF genes in the *Populus* genome. According to a phylogeny created from the AP2/ERF domains of 200 AP2/ERF proteins in *P. trichocarpa* and 145 AP2/ERF proteins in *A. thaliana*, they divided the AP2/ERF family into four subfamilies (AP2, DREB, ERF, and RAV), and revealed that two subfamilies (DREB and ERF) were expanded in *P. trichocarpa* in comparison with *A. thaliana*.

The LIM domain is an evolutionarily-conserved double-zinc finger motif found in a variety of proteins exhibiting diverse biological roles that act as modular protein-binding interfaces mediating protein-protein interactions in the cytoplasm and the nucleus (Khurana *et al.*, 2002). In tobacco, one LIM protein (Ntlim1) is a transcription factor affecting gene expression of lignin biosynthesis (Kawaoka and Ebinuma, 2001). Arnaud *et al.* (2007) identified 12 LIM gene models in the *P. trichocarpa* genome and their phylogenetic analysis of 24 LIM domain proteins (12 *P. trichocarpa* + six *A. thaliana* + six *O. sativa* proteins) revealed that plant LIM proteins have undergone one or several duplication events during their evolution. They classified the plant LIM proteins into four groups: alphaLIM1, betaLIM1, gammaLIM2 and deltaLIM2.

Cytokinin is one key hormone regulating many aspects of plant growth and development. Response Regulators (RRs) are transcription factors that function in the final step of the two-component cytokinin signaling system, which involves histidine kinase receptors that perceive cytokinin and transmit the signal via a multistep phosphorelay, with Type-A and Type-B RRs as negative and positive regulators of cytokinin, respectively (Mason *et al.*, 2005; To *et al.*, 2007). Ramirez-Carvajal *et al.* (2008) identified 33 cytokinin RR genes in *Populus*: 11 type As, 11 type Bs, and 11 pseudo-RRs. Their phylogenetic analysis using the conserved receiver domains of RR gene family members in *Populus*, *Arabidopsis* and *Oryza* revealed that a significant number of the RRs (26 in *Populus*, 20 in *Arabidopsis* and 14 in *Oryza*) grouped in species-specific pairs. Their expression analysis showed that *Populus* RR type As and type Bs appear to be preferentially expressed in nodes, while the pseudo-RRs are preferentially expressed in mature leaves.

Plant-specific R2R3-type MYB transcription factors control plant secondary metabolism as well as the identity and fate of plant cells (Stracke *et al.*, 2001). Wilkins *et al.* (2009) performed phylogenetic analysis of the predicted R2R3-MYB protein sequences in *Populus* (192 sequences), *Vitis* (119 sequences), *Arabidopsis* (126 sequences) and other plant species (54 sequences) and divided the R2R3-MYB family into 49 clades (C1 – C49). The phylogeny revealed unequal representation of *Populus*, *Vitis* and *Arabidopsis* R2R3-MYB proteins within individual clades. For example, clades C3 and C25 contained more *Populus* genes, whereas clades C13 and C37 did not include any *Populus* genes. This differential expansion/contraction indicates lineage-specific gene duplication and gene loss. Furthermore, they showed that the expanded *Populus* R2R3-MYB members were associated with wood formation and reproductive development.

2. *Flowering*

The FLOWERING LOCUS T (FT) and TERMINAL FLOWER1 (TFL1) genes function, respectively, as a promoter and a repressor of the floral transition in *Arabidopsis* (Danilevskaya *et al.*, 2008). There is increasing evidence that the FT protein is a major component of flower signal transduction which causes changes in the gene expression that reprograms the shoot apical meristem (SAM) to form flowers instead of leaves (Turck *et al.*, 2008). Igasaki *et al.* (2008) performed phylogenetic analysis of 32 FT/TFL1 proteins in *P. nigra* var. *italica* (nine sequences), *A. thaliana* (seven sequences), tomato (six sequences), *Vitis* (five sequences), apple (three sequences) and citrus (two sequences). Their phylogenetic tree revealed that the genes fall into four different clades: the TFL1 clade, the FT clade, the MOTHER OF FT AND TFL1 clade, and the BROTHER OF FT AND TFL1 clade. Their gene expression and transgenic studies suggest that one *Populus* gene in the TFL1 clade, PnTFL1, represses flowering and two *Populus* genes in the FT clade, PnFT1 and PnFT2, promote flowering.

3. *Signal Transduction*

Calcineurin B-like (CBL) proteins have been implicated as important sensors in signaling of calcium which plays a crucial role as a second messenger in mediating various defense responses under environmental stresses (Gu *et al.*, 2008). Zhang *et al.* (2008) identified 10 CBL candidate genes (PtCBLs) in the *P. trichocarpa* genome. Their phylogenetic analysis of *Arabidopsis*, *Oryza* and *Populus* CBL genes divided the CBL gene family into four groups. Their comparative analyses indicate that the duplication events in *Populus* might have contributed to the expansion of the CBL family. Furthermore, they cloned nine CBL genes (PeCBLs) from *P. euphratica*, a mostly salt- and drought-tolerant *Populus* species, and performed gene

specific RT-PCR analysis which suggests that seven CBL gene members may play an important role in responding to specific external stimuli.

The combination of the lysin motif (LysM) and receptor kinase domains is present exclusively in plants (Zhang *et al.*, 2007). LysM domain-containing receptor-like kinases (LYK) family members are critical for both nod factor and chitin signaling (Wan *et al.*, 2008). Zhang *et al.* (2007) identified a total of 48 LYK genes in *Arabidopsis* (five genes), *Oryza* (six genes), *M. truncatula* (eight genes), *L. japonicus* (six genes), *P. trichocarpa* (eleven genes) and *Glycine max* (twelve genes), and created two congruent plant LYK phylogenies using LysM domain sequences (all LysM motifs sequences + spacer sequences) and the full-protein sequences (LysM + kinase domain), respectively. The phylogenies revealed that the plant LYK proteins fall into three major clades and two minor clades. They identified six distinct types of LysM motifs in plant LYK proteins and five additional types of LysM motifs in non-kinase plant LysM proteins. Their genomic analysis revealed that the plant LYK gene family has evolved through local and segmental duplications. Their expression data showed that most plant LysM kinase genes were expressed predominantly in the root (Zhang *et al.*, 2007).

4. Protein Degradation

Proteases play key roles in the regulation of biological processes in plants, maintaining strict protein quality control and degrading specific sets of proteins in response to diverse environmental (i.e., defense responses to pathogens and pests) and developmental stimuli (van der Hoorn and Jones, 2004; Garcia-Lorenzo *et al.*, 2006). Garcia-Lorenzo *et al.* (2006) performed a comparative analysis of protease genes in *A. thaliana* and *P. trichocarpa*. They showed that most protease families were larger in *Populus* than in *Arabidopsis*, reflecting the recent genome duplication. They also showed that different *Populus* tissues expressed unique suites of protease genes and that the mRNA levels of different classes of proteases changed along a developmental gradient.

The F-box gene family is involved in posttranslational regulation of gene expression by selective degradation of proteins. In plants, F-box genes influence a variety of biological processes such as long-distance signaling, floral development, shoot branching, leaf senescence, root proliferation, cell cycle, and responses to biotic and abiotic stresses (Gagne *et al.*, 2002; Jain *et al.*, 2007; Yang *et al.*, 2008). Yang *et al.* (2008) recently performed a comparative genomics analysis of F-box gene sequences from *Arabidopsis*, *Oryza*, *Populus*, *Vitis* and papaya. The data revealed that the F-box gene family is expanded in herbaceous annuals (*Arabidopsis* and *Oryza*) relative to woody perennials (*Populus, Vitis* and papaya), supporting the hypothesis that compared to long-lived plants like trees, short-lived herbaceous annuals require a more diverse set of protein degradation mechanisms to successfully complete development over a short life cycle.

Programmed cell death, a central regulatory process in both plant development and in plant responses to pathogens, involves various caspase-like cysteine proteases as well as serine proteases such as Kunitz trypsin inhibitors (KTI) that have specific inhibitory activity solely against trypsin proteases (Li *et al.*, 2008). Major and Constabel (2008) performed phylogenetic analysis of 22 Kunitz trypsin inhibitor (KTI) protein sequences of *P. trichocarpa* x *P. deltoides*, and divided them into three clades A (eight sequences), B (four sequences), and C (eight sequences), with the clade A divided further into three subclades A1, A2 and A3. They selected five wound- and herbivore-induced genes representing the phylogenetic clades of the KTI gene family for functional analysis and cloned them into *Escherichia coli* to produce active KTI proteins. The recombinant KTI proteins were all biochemically distinct and showed clear differences in efficacy against trypsin-, chymotrypsin- and elastase-type proteases, suggesting that functional specialization of different members of this gene family is consistent with phylogenetic diversity (Major and Constabel, 2008).

5. Cell Cycle

In both animals and plants, it appears that D-type cyclins (CYCD) play an important role in cell cycle responses to external signals, by forming the regulatory subunit of cyclin-dependent kinase complexes (Meijer and Murray, 2000). Menges *et al.* (2007) constructed a phylogeny using 46 CYCD protein sequences from *Arabidopsis* (10 sequences), *Oryza* (14 sequences), and *Populus* (22 sequences), along with six CYCD sequences from moss (*P. patens*) and algae (*C. reinhardtii* and *O. tauri*) as outgroups. They divided the CYCDs into six clades according to the phylogenetic tree and identified remarkable conservation in intron/exon boundaries and in the location of potential cyclin-dependent kinase phosphorylation sites within CYCD proteins. A promoter sequence analysis and global expression correlation analysis revealed that the phylogenetic clade structure was supported by conserved regulatory elements and by the distinct expression patterns.

The NIMA-related family of serine/threonine kinases (Neks), defined by similarity in their N-terminal catalytic domains to the founding member, Never In Mitosis A from the fungus *Aspergillus*, play essential roles in cell cycle regulation and/or localize to centrosomes in fungi and mammals (Parker *et al.*, 2007). Vigneault *et al.* (2007) identified 22 Neks in *Arabidopsis* (seven sequences), *Populus* (nine sequences) and *Oryza* (six sequences). Their phylogenetic analysis showed that plant Neks were closely related to each other and contained paralogous genes. They examined the chromosomal distribution and exon-intron structure of these genes, concluding that the plant Nek family was derived from a single representative followed by large segmental duplication events.

6. Biosynthesis of Structural Components

Tubulins, as the major structural component of microtubules, consist of alpha/beta heterodimers (Jost *et al.*, 2004). Oakley

et al. (2007) identified eight alpha-TUBULIN (TUA) and 20 beta-TUBULIN (TUB) genes in the *Populus* genome. Their phylogenetic analysis of representative algal and plant full-length TUA proteins defined two distinct classes, with the eight *Populus*, six *Arabidopsis* and four *Oryza* isoforms evenly distributed between the two classes. According to a phylogenetic tree created from TUB proteins, Oakley *et al.* (2007) defined at least four distinct classes of plant TUBs: a Class I and Class I-like group containing half of the *Populus* (10 out of 20) and known maize (4 out of eight8) TUB families, along with a single *Arabidopsis* member (ArathTUB6); Class II containing primarily dicot TUBs, including six *Populus* and four *Arabidopsis* isoforms and one each from *Oryza*, *Zea* and green foxtail; Class III and Class IV each containing two *Populus* and two *Arabidopsis* isoforms arising from genome-wide duplications. They also reported that a number of features, including gene number, alpha:beta gene representation, amino acid changes at the C terminus, and transcript abundance in wood-forming tissue, distinguish the *Populus* tubulin suite from that of *Arabidopsis*.

7. Biosynthesis of Carbohydrates

The cellulose synthase gene superfamily of *Arabidopsis* and other seed plants is comprised of the cellulose synthase (CesA) family, which encodes the catalytic subunits of cellulose synthase, and eight families of CesA-like (Csl) genes, which have been proposed to encode processive beta-glycosyl transferases that synthesize noncellulosic cell wall polysaccharides (Roberts and Bushoven, 2007). Suzuki *et al.* (2006) identified 48 members of the cellulose synthase superfamily, which includes CesA and Csl genes in the *Populus* genome. Based on phylogenetic analysis, they divided the 48 *Populus* CesA/Csl protein sequences into nine groups, of which eight groups contained a pair of CesA genes with a nearly identical sequence and the remaining group contained 30 Csl gene family members that were further classified into PtCslA, B, C, D, E, and G subfamilies according to their protein sequence homology with the 29 known AtCsl members that define these subfamilies. They also examined the absolute transcript copy numbers of cellulose synthase superfamily genes in *Populus* and showed that 37 genes were expressed in various tissues, with seven CesA and four Csl genes being xylem specific. Geisler-Lee *et al.* (2006) have carried out a more comprehensive analysis of carbohydrate-active enzyme gene families. Their study based on glycosyltransferases, glycoside hydrolases, carbohydrate esterases, polysaccharide lyases, and expansins suggests that differential mechanisms of carbon flux regulation exist between a tree and an herb.

Endo-ß-mannanase is a hemicellulase that cleaves the ß-1,4 links between the mannose residues in the backbone of mannans, the widespread hemicellulosic polysaccharides in *Plant Cell* walls (Mo and Bewley, 2003; Yuan *et al.*, 2007). Yuan *et al.* (2007) identified 28 endo-beta-mannanase genes in the genomes of *Arabidopsis* (eight genes), *Oryza* (nine genes), and *Populus* (eleven genes). Their phylogenetic analysis of the endo-beta-mannanases from the three plant species implies that the existence of endo-β-mannanases predates the divergence of monocots and dicots, and in each orthologous group, the *Arabidopsis* gene(s) is/are more related to the *Populus* gene(s) than to the *Oryza* gene(s), consistent with the evolutionary relationships between the three plant species.

The polysaccharide xyloglucan (XG) plays an important structural role in the primary cell wall of dicotyledons. The xyloglucan endotransglucosylase/hydrolase (XTH) family proteins are known to have xyloglucan endotransglucosylase (XET) activity and xyloglucan endohydrolase activity (Rose *et al.*, 2002). XET enzymes play a key role in plant morphogenesis: non-hydrolytic cleavage and re-ligation of XG in the cell wall allowing transient, turgor-driven expansion and plant endoxyloglucanases are associated with the hydrolytic mobilization of seed storage XG during germination (Gilbert *et al.*, 2008). Baumann *et al.* (2007) performed phylogenetic analysis on ~130 full-length protein sequences that reflected the diversity of XETs and xyloglucanases in the archetypal glycoside hydrolase family 16 (GH16), including all genome-derived sequences from *A. thaliana* (33 sequences), *O. sativa* (29 sequences) and *P. trichocarpa* (36 sequences), full-length ESTs from *Solanum lycopersicum* (16 sequences) and *P. tremula* × *P. tremuloides* (13 sequences), as well as individual sequences from papaya, *Tropaeolum majus*, *Litchi chinensis* and *Vitis labrusca* × *V. vinifera*. Their phylogenetic analysis, together with kinetic data, suggest that xyloglucanase activity has evolved as a gain-of-function in an ancestral GH16 XET to meet specific biological requirements during seed germination, fruit ripening and rapid wall expansion.

The starch granule is composed of two structurally distinct homopolymers: amylase, which is essentially linear, and amylopectin, which is a moderately branched macromolecule (usually 6% of α-1,6 bonds within the polymer). Starch branching enzymes (SBEs) catalyze the formation of the α-1,6 linkages within amylopectin (Dumez *et al.*, 2006). Han *et al.* (2007) performed phylogenetic analysis of 47 SBE amino acid sequences of various plant species, and identified three orthologs encoding SBEs in *A. thaliana* (AtSBEIII), *O. sativa* (OsSBEIII) and *P. trichocarpa* (PtSBEIII), which represent a new SBE family (SBEIII) with structural features quite different from those of genes of both SBEI and SBEII families in plants.

Bocock *et al.* (2008) identified 24 invertase genes in the *Populus* genome including eight acid invertase genes and 16 neutral/alkaline invertase genes. The phylogenetic tree constructed using protein sequences of *Populus* and *Arabidopsis* acid invertases classified the eight acid invertase genes in *Populus* into two clades: cell wall invertases and vacuolar invertases. The phylogenetic tree constructed using protein sequences of *Populus* and *Arabidopsis* neutral/alkaline invertases classified the neutral/alkaline invertase genes in *Populus* into two clades.

8. Disease Resistance

The majority of disease resistance genes in plants encode nucleotide-binding site leucine-rich repeat (NBS-LRR)

proteins. This large family is encoded by hundreds of diverse genes per genome and can be divided into the functionally distinct TIR-domain-containing and CC-domain-containing subfamilies (McHale *et al.*, 2006). Kohler *et al.* (2008) identified 402 *Populus* NBS genes in *Populus*, which were distributed over 228 loci, with 170 sequences located on linkage groups and 232 genes on as yet unmapped scaffolds. According to the phylogenetic tree constructed using the NBS portion of the predicted protein sequences of 117 selected *Populus* R proteins, 11 *Arabidopsis* sequences and five *Oryza* sequences, they divided the NBS family into multiple subfamilies with distinct domain organizations, including Coiled-Coil-NBS-LRR genes, TIR-NBS-LRR genes and BED-finger-NBS-LRR, as well as truncated and unusual NBS- and NBS-LRR-containing genes. Surprisingly, the expansion of R genes in *Populus* relative to Arabidopsis cannot be accounted for by the large-scale duplication of the *Populus* genome. Instead, it appears that R genes have been expanding due to small-scale tandem duplications, and this rapidly-evolving gene family appears to be under strong diversifying selection, perhaps driven by the complex biotic interactions that develop over the long lifespan of a tree (Kohler *et al.*, 2008).

9. Fatty Acid Metabolism

Acyl-coenzyme A synthetases (ACS) catalyze the fundamental, initial reaction in fatty acid metabolism (Watkins *et al.*, 2007). 4-Coumarate:CoA ligase (4CL) is a branch point enzyme of plant phenylpropanoid metabolism, catalyzing the activation of 4-coumaric acid and various other hydroxylated and methoxylated cinnamic acid derivatives to the corresponding CoA esters in a two-step reaction (Pietrowska-Borek *et al.*, 2003). Souza Cde *et al.* (2008) performed phylogenetic analysis of 104 ACS related gene sequences from various organisms, including *bona fide* 4CL sequences from *Arabidopsis*, *Populus* and *Oryza*. Their phylogenetic tree revealed two general groups of adenylate-forming proteins: one large group containing representatives from all organisms analyzed, including bacteria, fungi, *Chlamydomonas*, *Physcomitrella* and angiosperm plants and a second group containing land plant-specific ACS proteins including both *bona fide* 4CL proteins and *Arabidopsis* 4CL-like ACS proteins. Further, they performed a separate phylogenetic analysis of the plant-specific ACSs protein sequences from *Arabidopsis* (nine sequences), *Populus* (13 sequences) and *Oryza* (12 sequences) and were able to divide the plant-specific ACSs into five clades. The phylogenetic tree revealed that the number of plant-specific ACS genes in each clade varied between species while the total number of plant-specific ACS genes within each genome was similar. They hypothesized that ACS genes have undergone differential expansion in each angiosperm lineage, perhaps reflecting differences in life histories that placed varying selective pressures on the elaboration of biochemical pathways requiring ACS activity (Souza Cde *et al.*, 2008).

10. Phenylpropanoid Pathway

Populus produces a rich array of natural products such as the secondary metabolites derived from phenylalanine via phenol and phenylpropanoid metabolism (Hamberger *et al.*, 2007). More than 100 *Populus* phenylpropanoid pathway genes were phylogenetically compared with homologs in *Arabidopsis* and *Oryza* by Tsai *et al.* (2006) and Hamberger *et al.* (2007). Gene families included in their analyses were: arogenate dehydratase, beta-alanine N-methyltransferase, flavonoid 3′,5′-hydroxylase, flavone synthase II, flavanone 3-hydroxylase, flavonol synthase, dihydroflavonol 4-reductase, anthocyanidin synthase, anthocyanidin reductase, leucoanthocyanidin reductase, flavonoid O-methyltransferase, trans-cinnamate 4-hydroxylase, coumarate 3-hydroxylase, caffeic acid O-methyltransferase, ferulate 5-hydroxylase, caffeoyl-CoA O-methyltransferase, hydroxycinnamoyltransferase, aldehyde dehydrogenase, cinnamoyl CoA reductase, NADPH-cytochrome P450 oxydoreductase, cinnamyl alcohol dehydrogenase-related, phenylalanine ammonia lyase, and 4-coumarate:CoA ligase.

11. Photorespiration

The glycine decarboxylase multi-enzyme complex (GDC) catalyzes in a multi-step reaction the rapid 'cracking' of glycine molecules flooding out of the peroxisomes during the course of photorespiration (Douce *et al.*, 2001). Rajinikanth *et al.* (2007) identified ten GDC proteins in *Populus*, of which eight are localized in mitochondria and two in plastids. Their phylogenetic analysis of GDC protein sequences from representative dicot, monocot and gymnosperm species clearly revealed two distinct GDC classes, with class I corresponding to photorespiratory isoforms and class II associated with one-carbon metabolism, indicating that the functional uniqueness is consistent with phylogenetic classification.

C. Databases

Just like other plant genomics models such as *Arabidopsis* and *Oryza*, *Populus* has become a new theme of genomics databases. To date, more than ten public genomics databases have been established for *Populus* or related to *Populus* (Table 2).

1. Populus-Specific Databases

The official release of the *Populus* genome information is provided by the DOE's Joint Genome Institute (JGI) *Populus* genome browser (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). In addition, Sterky *et al.* (2004) established a public web-based EST database (POPULUSDB) which provides digital expression profiles for 18 tissues that comprise the majority of differentiated organs (http://www.populus.db.umu.se/). The RIKEN *Populus* database (http://rpop.psc.riken.jp) contains information covering 10 *Populus* species (*P. deltoides*, *P. euphratica*, *P. tremula*, *P. tremula* × *P. alba*, *P. tremula* × *P. tremuloides*, *P.*

TABLE 2
Populus genomics databases

Category	Database	URL	Reference
Populus-specific	PopulusDB	http://www.populus.db.umu.se/	Sterky *et al.* (2004)
Populus-specific	RPOPDB	http://rpop.psc.riken.jp/	
Populus-specific	RepPop	http://csbl.bmb.uga.edu/∼ffzhou/RepPop/	Zhou and Xu (2009)
Populus-specific	Transcription Factors	http://dptf.cbi.pku.edu.cn/	Zhu *et al.* (2007)
Populus-specific	POPARRAY	http://popgenome.ag.utk.edu/mdb/index.php	
Populus-specific	PopGenIE	http://www.popgenie.db.umu.se/popgenie/	Sjödin *et al.* (2008)
Populus-specific	JGI *Populus* genome	http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html	Tuskan *et al.* (2006)
Comprehensive	ChromDB	http://www.chromdb.org/org_specific.html?o = POPTR	Gendler *et al.* (2008)
Comprehensive	PlantGDB	http://www.plantgdb.org/	Duvick *et al.* (2008)
Comprehensive	Gramene	http://www.gramene.org/	Liang *et al.* (2008)
Comprehensive	PLEXdb	http://www.plexdb.org/index.php	Wise *et al.* (2007)
Comprehensive	Phytozome	http://www.phytozome.net/index.php	

tremuloides, *P. trichocarpa*, *P. trichocarpa* × *P. deltoides*, *P. trichocarpa* × *P. nigra*, and *P.* × *canadensis*). To facilitate functional studies on the regulation of gene expression in *Populus*, Zhu *et al.* (2007) created a *Populus* transcription factor (TF) database (DPTF; http://dptf.cbi.pku.edu.cn/) containing 2576 putative *Populus* TFs, distributed in 64 families. It provides comprehensive information for the *Populus* TFs such as sequence features, functional domains, GO assignment, expression evidence, phylogenetic tree of each family, and homologs in *Arabidopsis* and *Oryza*. Basing on the genome-wide analysis of 9,623 repetitive elements in the *P. trichocarpa* genome, Zhou and Xu (2009) created a web-browsable database, RepPop (http://csbl.bmb.uga.edu/∼ffzhou/RepPop/), which offers resources on DNA transposons, RNA retrotransposons, Miniature Inverted-repeat Transposable Elements (MITE), Simple Sequence Repeats (SSR), segmental duplications, etc. This database also provides various search capabilities and a Wiki system to facilitate functional annotation and curation of the repetitive elements. To facilitate the exploration of genes and gene function in *Populus*, Sjödin *et al.* (2008) developed an integrative *Populus* functional genomics database, PopGenIE (www.popgenie.org), which includes several browser tools for viewing genome sequence, QTL and synteny; an expression tool for displaying multiple-tissue expression patterns based on microarray experiments, Digital Northern analysis of EST data, and co-expressed genes; a category tool for information about gene families, pathways, GO annotations; and a sequence tool for similarity searches and data downloads.

2. *Comprehensive Databases Involving Populus*

To facilitate comparative genomic studies amongst green plants, JGI and the Center for Integrative Genomics developed Phytozome (http://www.phytozome.net/index.php), which provides access to eleven sequenced and annotated higher plant genomes, eight of which have been clustered into gene families at six evolutionarily significant nodes, along with PFAM, KOG, KEGG and PANTHER assignments, offering resources for clusters of orthologous and paralogous genes that represent the modern descendents of ancestral gene sets, as well as clade specific genes and gene expansions. Duvick *et al.* (2008) developed a comprehensive plant genomics database, PlantGDB (http://www.plantgdb.org/) which provides annotated transcript assemblies for >100 plant species including *Populus*, with transcripts mapped to their cognate genomic context integrated with a variety of sequence analysis tools and web services. PlantGDB also hosts a plant genomics research outreach portal that facilitates access to a large number of training resources. Another plant comparative genomics database is Gramene (Liang *et al.*, 2008), which contains genomic information for *Oryza sativa* var. *indica* and *japonica*, *O. glaberrima*, *O. rufipogon*, *Zea mays*, *Sorghum bicolor*, *Arabidopsis thaliana*, *Vitis vinifera* and *P. trichocarpa*, including genome assembly and annotations, cDNA/mRNA sequences, genetic and physical maps/markers, genes, quantitative trait loci, proteins, and comparative ontologies. To display sets of curated plant genes predicted to encode proteins associated with chromatin remodeling, Gendler *et al.* (2008) established the ChromDB database (http://www.chromdb.org) which displays chromatin-associated proteins, including RNAi-associated proteins, for a broad range of organisms such as *A. thaliana*, *O. sativa* ssp. *japonica*, *P. trichocarpa*, *Zea mays*, and model animal and fungal species. ChromDB contains three types of sequences: genomic-based (predominantly plant sequences); transcript-based (EST contigs or cDNAs for plants lacking a sequenced genome) and NCBI RefSeq sequences for a variety of model animal organisms. To facilitate large-scale gene expression analysis for plants, Wise *et al.* (2007) created PLEXdb (http://plexdb.org/), which offers a unified web interface to support the functional interpretation of highly parallel microarray experiments integrated with traditional structural genomics and phenotypic data. PLEXdb contains information for 13 plant species, such as *Arabidopsis*, barley, *Citrus*, cotton, *Vitis*, *Zea*,

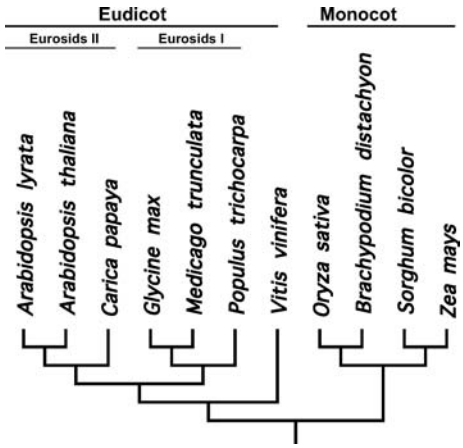


FIG. 2. Higher plant species whose genomes have been sequenced.

Medicago, *Populus*, *Oryza*, *Glycine*, sugarcane, tomato, and wheat.

D. Evolutionary Genomics

1. Genome Evolution

Knowledge of plant evolutionary history at the genome level forms an excellent context for understanding genomics at subgenome levels. Genomes of nine higher plant species have been sequenced (Fig. 2). Genome-wide sequence resources have facilitated large-scale computational analysis of genome evolution. Tuskan *et al.*, 2006 reported that *Populus* experienced two rounds of genome duplication, with the most recent event ('salicoid' duplication) contained within the Salicaceae and a second whole-genome event ('eurosid' duplication) apparently shared among the Eurosids. Using a robust computational framework that combines information from multiple orthologous and duplicated regions to construct local syntenic networks, Tang *et al.* (2008b) showed that a shared ancient hexaploidy event (γ triplication) can be inferred based on the genome sequences of *Arabidopsis*, *Carica*, *Populus*, *Vitis* and *Oryza*. They conclude that "paleo-hexaploidy" clearly preceded the rosid-asterid split, but it remains equivocal whether it also affected monocots. In short, three known genome duplication events have been inferred in *Populus*: salicoid, eurosid, and γ triplication (Fig. 3).

2. Evolutionary Dynamics of Duplicated Genes

Gene duplication and diversification contribute to the novelty of molecular functions. To study the evolutionary dynamics of gene families in plants, Yang *et al.* (2006) identified 27 pairs of paralogous Dof (DNA binding with one finger) genes in a phylogenetic tree constructed from protein sequences in *P. trichocarpa*, *A. thaliana* and *O. sativa*. The comparison of protein motif structure of the Dof paralogs and their ancestors revealed six different gene fates after gene duplication as well as epigenetic modification via protein methylation. Multiple modes of evolutionary dynamics including neo/nonfunctionalization and subfunctionalization were also reported at the promoter level by De Bodt *et al.* (2006), who investigated the evolution of MADS-box genes in *Arabidopsis* and *Populus* using phylogenetic footprinting. This study showed that many genes have diverged in their regulatory sequences after duplication and/or speciation. Based on microarray and genome sequence information, Kalluri *et al.* (2007) showed that segmental duplicate gene pairs in the *Populus* ARF

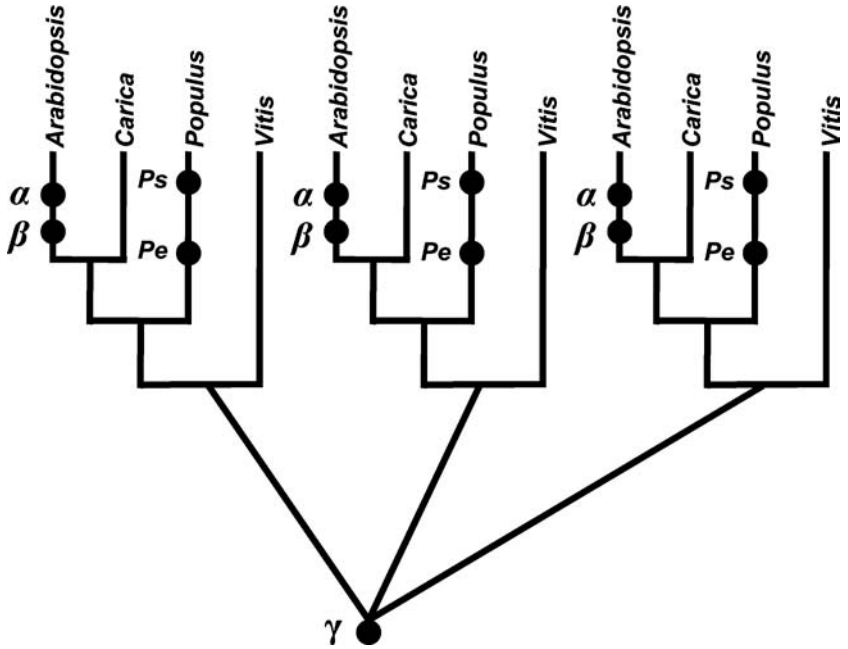


FIG. 3. The duplication events in the plant genomes. Adapted from Tang *et al.* (2008a, 2008b) and Tuskan *et al.* (2006).

gene family, *PoptrARF6.1–6.2* and *PoptrARF6.4–6.5*, have differential expression profiles. For example, *PoptrARF6.1* was expressed preferentially in mature leaves and phloem-cortex samples whereas *PoptrARF6.2* was expressed strongly in xylem, phloem, and vegetative and reproductive meristems, suggesting subfunctionalization after gene duplication. Subfunctionalization in expression of the duplicated genes has also been revealed in the some gene families involved in phenylpropanoid metabolism, such as phenylalanine ammonia lyase, coumaroyl 3-hydroxylase, cinnamyl alcohol dehydrogenase-related and caffeoyl-CoA 3-O-methyltransferase (Hamberger *et al.*, 2007), in the *Populus* MONOPTEROS gene family in which two duplicated MONOPTEROS genes, PoptrMP1 and PoptrMP2, have overlapping but distinct expression patterns (Johnson and Douglas, 2007), and in *Populus* LHY/CCA1 gene family in which two LHYs produced by the Salicoid polyploidy event showed asymmetric expressions (Takata *et al.*, 2009).

IV. CONCLUSION AND FUTURE PERSPECTIVES

A. Conclusion

Our primary objective in writing this review was to highlight how the recent sequencing of the *Populus* genome (Tuskan *et al.*, 2006) has spurred applications of state-of-the-art technologies in *Populus* research, including both experiment-based functional genomics and newly developed computational genomics. Technologies are in place for the high-throughput analysis of gene expression, coupled with protein and metabolite profiling and verification of gene function via plant transformation and incorporation of candidate genes into transgenic *Populus* plants. These capabilities, mostly developed in the last five years, provide much-needed insights into the growth, morphology and reproductive strategies for some of the largest and longest-living organisms on Earth.

Equally impressive have been the major inroads made in developing *in silico* resources for use in comparative genomics. Central data repositories and on-line bioinformatics tools now make possible identification of lineage-specific motifs and genes and improved understanding of gene family evolution and regulation. Studies based largely on *in silico* analyses are already demonstrating how gene duplication and diversification can contribute to the novelty of molecular functions, and how gene families develop and expand along unique trajectories in different phylogenetic lineages. This information highlights how plant evolutionary history at the genome level forms a context for understanding genome organization, evolution and function at the sub-genome level. Insights derived from functional and computational genomics will continue to yield improved understanding at the organismal, population, community and ecosystem scales. It will be in this arena that molecular biologists, physiologists and ecologists will derive maximum mutual benefits from investing in the development of genetic and genomics resources for *Populus* (DiFazio, 2005; Whitham *et al.*, 2008).

Finally, several articles published while the *Populus* genome was being sequenced (Taylor, 2002; Wullschleger *et al.*, 2002; Wullschleger *et al.*, 2002; Tuskan *et al.*, 2004) highlighted the notion that embracing new technology and new research paradigms will not be easy. To our surprise, and as evidenced in this review, the community has quickly risen to the challenge provided by the development and application of technologies in *Populus* genomics research. We see no reason why this enthusiasm should wane. Only time will tell, however, how our investments in functional and computational genomics will aid in identifying the suite of genes and gene families that underlie plant growth and development. Evidence exists that insights derived over a relatively short time are already being applied to enhance production of short-rotation bioenergy crops as a renewable source of biomass for transportation fuels. This advancement will potentially lessen our dependence on foreign oil and mitigating rising CO_2 concentrations in the atmosphere (Tuskan and Walsh, 2001).

B. Future Perspectives

The trend in biological investigations is shifting from reductionism to evolutionary system-inclusive biology. This trend is driven in part by the capabilities that genomic science is now providing for powerful cross-species comparative studies. Comparative genomics will be an essential component of investigating genetic features and molecular processes underlying core conserved, as well as unique divergent, plant properties. Comparative genomics studies of *Populus*, *Arabidopsis*, *Oryza*, papaya, *Vitis*, *Sorghum*, *Glycine*, *Brachypodium* and others have already begun shedding light on the evolutionary events that have resulted in divergent and conserved patterns at the genetic, molecular and organism levels. Moreover, comparative genomics research is expected to move beyond the narrow scope of studying isolated plants to the broader context of a community (i.e., the "metagenome") (Tringe and Rubin, 2005; Whitham *et al.*, 2008). The identity as well as roles of *Populus* endophytic and rhizosphere microbiota will also be important co-considerations in sustainable development and the establishment of plantations of suitably improved plants (Martin *et al.*, 2004).

It is not hard to foresee that the availability of genome sequences from dozens of plant species will propel science towards predictive outcomes facilitated by the 1) creation of open-source community resource and data repositories, 2) sharing of resources and discoveries, 3) emergence of new technological innovations, and 4) synergy in large interdisciplinary collaborative efforts. The new era of plant research will witness innovative hypotheses, better experimental design, and more sophisticated toolsets.

It is clear that plant biotechnology will play an increasingly large role in meeting soaring demands for food, fiber and energy. Therefore, it is important that the research community raise its awareness regarding future regulatory and consumer acceptance

issues. In order to realize the potential of the long-range agro- and forest biotechnology goals, molecular geneticists will need to understand the economic, environmental and social contexts, as well as biological implications, of their research (Strauss, 2003).

The attractiveness of *Populus* as a model woody crop has been acknowledged, and efforts are underway worldwide to elucidate genetic features that are key to generating suitably tailored plants. The 'book of life' awaits future examination and interpretations directed towards basic and applied research related to the development of tree-based strategies for fiber production, carbon sequestration, phytoremediation, and accelerated and sustainable domestication.

ACKNOWLEDGMENTS

We thank David J. Weston for the constructive comments on the manuscript and Tara A. Hall for assisting with proofreading. The writing of this review was supported by the BioEnergy Science Center (BESC), a U.S. Department of Energy Bioenergy Research Center supported by the Office of Biological and Environmental Research in the DOE Office of Science. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract Number DE–AC05–00OR22725.

REFERENCES

Alstrom-Rapaport, C. L. M., Wang, Y. C., Roberts, G., and Tuskan, G. A. 1998. Identification of a RAPD marker linked to sex determination in the basket willow (*Salix viminalis* L.). *J. Heredity* **89**: 44–49.

Andersson, A., Keskitalo, J., Sjodin, A., Bhalerao, R., Sterky, F., Wissel, K., Tandre, K., Aspeborg, H., Moyle, R., Ohmiya, Y., *et al*. 2004. A transcriptional timetable of autumn senescence. *Genome Biol.* **5**: R24.

Andersson-Gunneras, S., Mellerowicz, E. J., Love, J., Segerman, B., Ohmiya, Y., Coutinho, P. M., Nilsson, P., Henrissat, B., Moritz, T., and Sundberg, B. 2006. Biosynthesis of cellulose-enriched tension wood in *Populus*: global analysis of transcripts and metabolites identifies biochemical and developmental regulators in secondary wall biosynthesis. *Plant J.* **45**: 144–165.

Arnaud, D., Dejardin, A., Leple, J. C., Lesage-Descauses, M. C., and Pilate, G. 2007. Genome-wide analysis of LIM gene family in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*. *DNA Res.* **14**: 103–116.

Baek, J. M., Han, P., Iandolino, A., and Cook, D. R. 2008. Characterization and comparison of intron structure and alternative splicing between *Medicago truncatula*, *Populus trichocarpa*, *Arabidopsis* and rice. *Plant Mol. Biol.* **67**: 499–510.

Barakat, A., Wall, P. K., Diloreto, S., dePamphilis, C. W., and Carlson, J. E. 2007. Conservation and divergence of microRNAs in *Populus*. *BMC Genomics* **8**: 481.

Baumann, M. J., Eklof, J. M., Michel, G., Kallas, A. M., Teeri, T. T., Czjzek, M., and Brumer, H., 3rd. 2007. Structural evidence for the evolution of xyloglucanase activity from xyloglucan endo-transglycosylases: biological implications for cell wall metabolism. *Plant Cell* **19**: 1947–1963.

Beavis, W. D. 1998. QTL analyses: Power, precision, and accuracy. **In:** AH Paterson, ed, *Molecular Dissection of Complex Traits*. CRC Press, Boca Raton, FL, pp. 145–162.

Becker, A., and Theissen, G. 2003. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Mol. Phylo. Evol.* **29**: 464–489.

Benedict, C., Drost, D., Novaes, E., Boaventura-Novaes, C. D., Yu, Q., Dervinis, C., Maia, J. M., Yap, J., Miles, B., Peter, G. F., *et al.* 2009. Genome-wide eQTL mapping in xylem, leaf and root identifies tissue-specific hotspots in a *Populus trichocarpa* x *deltoides* pedigree. Plant & Animal Genomes XVII Conference W181.

Bocock, P. N., Morse, A. M., Dervinis, C., and Davis, J. M. 2008. Evolution and diversity of invertase genes in *Populus trichocarpa*. *Planta* **227**: 565–576.

Bogeat-Triboulot, M. B., Brosche, M., Renaut, J., Jouve, L., Le Thiec, D., Fayyaz, P., Vinocur, B., Witters, E., Laukens, K., Teichmann, T., *et al.* 2007. Gradual soil water depletion results in reversible changes of gene expression, protein profiles, ecophysiology, and growth performance in *Populus euphratica*, a poplar growing in arid regions. *Plant Physiol.* **143**: 876–892.

Bradshaw, H. D., and Stettler, R. F. 1995. Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* **139**: 963–973.

Bradshaw, H. D., Villar, M., Watson, B. D., Otto, K. G., Stewart, S., and Stettler, R. F. 1994. Molecular-genetics of growth and development in *Populus*. III. A genetic-linkage map of a hybrid poplar composed of RFLP, STS, and RAPD markers. *Theoreti. Applied Genetics* **89**: 167–178.

Brosche, M., Vinocur, B., Alatalo, E. R., Lamminmaki, A., Teichmann, T., Ottow, E. A., Djilianov, D., Afif, D., Bogeat-Triboulot, M. B., Altman, A., *et al*. 2005. Gene expression and metabolite profiling of *Populus euphratica* growing in the Negev desert. *Genome Biol.* 6: –

Busov, V., Meilan, R., Pearce, D. W., Rood, S. B., Ma, C., Tschaplinski, T. J., and Strauss, S. H. 2006. Transgenic modification of gai or rgl1 causes dwarfing and alters gibberellins, root growth, and metabolite profiles in *Populus*. *Planta* **224**: 288–299.

Busov, V. B., Meilan, R., Pearce, D. W., Ma, C., Rood, S. B., and Strauss, S. H. 2003. Activation tagging of a dominant gibberellin catabolism gene (GA 2-oxidase) from poplar that regulates tree stature. *Plant Physiol.* **132**: 1283–1291.

Caruso, A., Chefdor, F., Carpin, S., Depierreux, C., Delmotte, F. M., Kahlem, G., and Morabito, D. 2008. Physiological characterization and identification of genes differentially expressed in response to drought induced by PEG 6000 in *Populus canadensis* leaves. *J. Plant Physiol.* **165**: 932–941.

Creux, N. M., Ranik, M., Berger, D. K., and Myburg, A. A. 2008. Comparative analysis of orthologous cellulose synthase promoters from *Arabidopsis*, *Populus* and *Eucalyptus*: evidence of conserved regulatory elements in angiosperms. *New Phytol.* **179**: 722–737.

Cseke, L. J., Cseke, S. B., and Podila, G. K. 2007. High efficiency poplar transformation. *Plant Cell Rep.* **26**: 1529–1538.

Danilevskaya, O. N., Meng, X., Hou, Z. L., Ananiev, E. V., and Simmons, C. R. 2008. A genomic and expression compendium of the expanded PEBP gene family from maize. *Plant Physiol.* **146**: 250–264.

De Bodt, S., Theissen, G., and Van de Peer, Y. 2006. Promoter analysis of MADS-box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* **23**: 1293–1303.

Diaz-Riquelme, J., Lijavetzky, D., Martinez-Zapater, J. M., and Carmona, M. J. 2009. Genome-wide analysis of MIKCC-type MADS box genes in grapevine. *Plant Physiol.* **149**: 354–369.

DiFazio, S. P. 2005. A pioneer perspective on adaptation. *New Phytol.* 165**:** 661–664.

Douce, R., Bourguignon, J., Neuburger, M., and Rebeille, F. 2001. The glycine decarboxylase system: a fascinating complex. *Trends Plant Sci.* **6**: 167–176.

Druart, N., Johansson, A., Baba, K., Schrader, J., Sjodin, A., Bhalerao, R. R., Resman, L., Trygg, J., Moritz, T., and Bhalerao, R. P. 2007. Environmental and hormonal regulation of the activity-dormancy cycle in the cambial meristem involves stage-specific modulation of transcriptional and metabolic networks. *Plant J.* **50**: 557–573.

Du, J., Xie, H. L., Zhang, D. Q., He, X. Q., Wang, M. J., Li, Y. Z., Cui, K. M., and Lu, M. Z. 2006. Regeneration of the secondary vascular system in poplar as a novel system to investigate gene expression by a proteomic approach. *Proteomics* **6**: 881–895.

Dumez, S., Wattebled, F., Dauvillee, D., Delvalle, D., Planchot, V., Ball, S. G., and D'Hulst, C. 2006. Mutants of *Arabidopsis* lacking starch branching enzyme II substitute plastidial starch synthesis by cytoplasmic maltose accumulation. *Plant Cell* **18**: 2694–2709.

Duvick, J., Fu, A., Muppirala, U., Sabharwal, M., Wilkerson, M. D., Lawrence, C. J., Lushbough, C., and Brendel, V. 2008. PlantGDB: a resource for comparative plant genomics. *Nucleic Acids Res.* **36**: D959–965.

Emrich, S. J., Barbazuk, W. B., Li, L., and Schnable, P. S. 2007. Gene discovery and annotation using LCM-454 transcriptome sequencing. *Genome Res.* **17**: 69–73.

Espinosa-Ruiz, A., Saxena, S., Schmidt, J., Mellerowicz, E., Miskolczi, P., Bako, L., and Bhalerao, R. P. 2004. Differential stage-specific regulation of cyclin-dependent kinases during cambial dormancy in hybrid aspen. *Plant J.* **38**: 603–615.

Ferreira, S., Hjerno, K., Larsen, M., Wingsle, G., Larsen, P., Fey, S., Roepstorff, P., and Salome Pais, M. 2006. Proteome profiling of *Populus euphratica* Oliv. upon heat stress. *Ann. Bot.* (Lond) **98**: 361–377.

Ferris, R., Long, L., Bunn, S. M., Robinson, K. M., Bradshaw, H. D., Rae, A. M., and Taylor, G. 2002. Leaf stomatal and epidermal cell development: identification of putative quantitative trait loci in relation to elevated carbon dioxide concentration in poplar. *Tree Physiol.* **22**: 633–640.

Fiehn, O. 2002. Metabolomics – the link between genotypes and phenotypes. *Plant Mol. Biol.* 48**:** 155–171.

Gagne, J. M., Downes, B. P., Shiu, S. H., Durski, A. M., and Vierstra, R. D. 2002. The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **99**: 11519–11524.

Garcia-Lorenzo, M., Sjodin, A., Jansson, S., and Funk, C. 2006. Protease gene families in *Populus* and *Arabidopsis*. *BMC Plant Biol.* 6**:** 30.

Gaudet, M., Jorge, V., Paolucci, I., Beritognolo, I., Mugnozza, G. S., and Sabatti, M. 2008. Genetic linkage maps of *Populus nigra* L. including AFLPs, SSRs, SNPs, and sex trait. *Tree Genetics & Genomes* **4**: 25–36.

Geisler-Lee, J., Geisler, M., Coutinho, P. M., Segerman, B., Nishikubo, N., Takahashi, J., Aspeborg, H., Djerbi, S., Master, E., Andersson-Gunneras, S., *et al.* 2006. Poplar carbohydrate-active enzymes. Gene identification and expression analyses. *Plant Physiol.* **140**: 946–962.

Gendler, K., Paulsen, T., and Napoli, C. 2008. ChromDB: the chromatin database. *Nucleic Acids Res.* **36**: D298–302.

Gilbert, H. J., Stalbrand, H., and Brurner, H. 2008. How the walls come crumbling down: recent structural biochemistry of plant polysaccharide degradation. *Curr. Opin. Plant Biol.* **11**: 338–348.

Gilchrist, E. J., Haughn, G. W., Ying, C. C., Otto, S. P., Zhuang, J., Cheung, D., Hamberger, B., Aboutorabi, F., Kalynyak, T., Johnson, L., *et al.* 2006. Use of Ecotilling as an efficient SNP discovery tool to survey genetic variation in wild populations of *Populus trichocarpa*. *Mol. Ecol.* **15**: 1367–1378.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J. K., Bailey-Serres, J., and Mittler, R. 2006. What makes species unique? The contribution of proteins with obscure features. *Genome Biol.* 7: R57.

Gollery, M., Harper, J., Cushman, J., Mittler, T., and Mittler, R. 2007. POFs: what we don't know can hurt us. *Trends Plant Sci.* **12**: 492–496.

Gou, J. Y., Park, S., Yu, X. H., Miller, L. M., and Liu, C. J. 2008. Compositional characterization and imaging of "wall-bound" acylesters of *Populus trichocarpa* reveal differential accumulation of acyl molecules in normal and reactive woods. *Planta* **229**: 15–24.

Gowda, M., Li, H., Alessi, J., Chen, F., Pratt, R., and Wang, G. L. 2006. Robust analysis of 5′-transcript ends (5′-RATE): a novel technique for transcriptome analysis and genome annotation. *Nucleic Acids Res.* **34**: e126.

Groover, A. T., Mansfield, S. D., DiFazio, S. P., Dupper, G., Fontana, J. R., Millar, R., and Wang, Y. 2006. The *Populus* homeobox gene ARBORKNOX1 reveals overlapping mechanisms regulating the shoot apical meristem and the vascular cambium. *Plant Mol. Biol.* **61**: 917–932.

Gu, Z. M., Ma, B. J., Jiang, Y., Chen, Z. W., Su, X., and Zhang, H. S. 2008. Expression analysis of the calcineurin B-like gene family in rice (*Oryza sativa* L.) under environmental stresses. *Gene* **415**: 1–12.

Hamberger, B., Ellis, M., Friedmann, M., Souza, C. D. A., Barbazuk, B., and Douglas, C. J. 2007. Genome-wide analyses of phenylpropanoid-related genes in *Populus trichocarpa*, *Arabidopsis thaliana*, and *Oryza sativa*: the *Populus* lignin toolbox and conservation and diversification of angiosperm gene families. *Can. J. Bot.* **85**: 1182–1201.

Han, K.-H., Gordon, M. P., and Strauss, S. H. 1996. Cellular and molecular biology of *Agrobacterium*-mediated transformation of plants and its application to genetic transformation of *Populus*. **In**: Stettler, R. F., Bradshaw, H. D., Heilman, P. E. Hinckley, and T. M., Eds., *Biology of* Populus *and Its Implications for Management and Conservation.* National Research Council of Canada, Ottawa, Ontario, Canada, pp. 201–222.

Han, Y., Sun, F. J., Rosales-Mendoza, S., and Korban, S. S. 2007. Three orthologs in rice, *Arabidopsis*, and *Populus* encoding starch branching enzymes (SBEs) are different from other SBE gene families in plants. *Gene* **401**: 123–130.

Harding, S. A., Jiang, H. Y., Jeong, M. L., Casado, F. L., Lin, H. W., and Tsai, C. J. 2005. Functional genomics analysis of foliar condensed tannin and phenolic glycoside regulation in natural cottonwood hybrids. *Tree Physiol.* **25**: 1475–1486.

Harrison, E. J., Bush, M., Plett, J. M., McPhee, D. P., Vitez, R., O'Malley, B., Sharma, V., Bosnich, W., Seguin, A., and MacKay, J., *et al.* 2007. Diverse developmental mutants revealed in an activation-tagged population of poplar. *Can. J. Bot.* **85**: 1071–1081.

Hayashi, H., Czaja, I., Lubenow, H., Schell, J., and Walden, R. 1992. Activation of a plant gene by T-DNA tagging: auxin-independent growth in vitro. *Science* **258**: 1350–1353.

Hu, W. J., Harding, S. A., Lung, J., Popko, J. L., Ralph, J., Stokke, D. D., Tsai, C. J., and Chiang, V. L. 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nat. Biotech.* **17**: 808–812.

Igasaki, T., Watanabe, Y., Nishiguchi, M., and Kotoda, N. 2008. The FLOWERING LOCUS T/TERMINAL FLOWER 1 family in Lombardy poplar. *Plant Cell Physiol.*, **49**: 291–300.

Ingvarsson, P. K. 2008. Multilocus patterns of nucleotide polymorphism and the demographic history of *Populus tremula*. *Genetics* **180**: 329–340.

Ingvarsson, P. K., Garcia, M. V., Luquez, V., Hall, D., and Jansson, S. 2008. Nucleotide polymorphism and phenotypic associations within and around the phytochrome B2 Locus in European aspen (*Populus tremula*, Salicaceae). *Genetics* **178**: 2217–2226.

Jain, M., Nijhawan, A., Arora, R., Agarwal, P., Ray, S., Sharma, P., Kapoor, S., Tyagi, A. K., and Khurana, J. P. 2007. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. *Plant Physiol.* **143**: 1467–1483.

Johnson, L. A., and Douglas, C. J. 2007. *Populus trichocarpa* MONOPTEROS/AUXIN RESPONSE FACTOR5 (ARF5) genes: comparative structure, subfunctionalization, and *Populus Arabidopsis* microsynteny. *Can. J. Bot.* **85**: 1058–1070.

Jorge, V., Dowkiw, A., Faivre-Rampant, P., and Bastien, C. 2005. Genetic architecture of qualitative and quantitative *Melampsora larici-populina* leaf rust resistance in hybrid poplar: genetic mapping and QTL detection. *New Phytol.* **167**: 113–127.

Jost, W., Baur, A., Nick, P., Reski, R., and Gorr, G. 2004. A large plant beta-tubulin family with minimal C-terminal variation but differences in expression. *Gene* **340**: 151–160.

Kalluri, U. C., DiFazio, S. P., Brunner, A. M., and Tuskan, G. A. 2007. Genome-wide analysis of Aux/IAA and ARF gene families in *Populus trichocarpa*. *BMC Plant Biol.* **7**: 59.

Kalluri, U. C., and Joshi, C. P. 2004. Differential expression patterns of two cellulose synthase genes are associated with primary and secondary cell wall development in aspen trees. *Planta* **220**: 47–55.

Kawaoka, A., and Ebinuma, H. 2001. Transcriptional control of lignin biosynthesis by tobacco LIM protein. *Phytochemistry* **57**: 1149–1157.

Khurana, T., Khurana, B., and Noegel, A. A. 2002. LIM proteins: association with the actin cytoskeleton. *Protoplasma* **219**: 1–12.

Kirst, M., Basten, C. J., Myburg, A. A., Zeng, Z. B., and Sederoff, R. R. 2005. Genetic architecture of transcript-level variation in differentiating xylem of a *Eucalyptus* hybrid. *Genetics* **169**: 2295–2303.

Ko, J. H., Prassinos, C., and Han, K. H. 2006. Developmental and seasonal expression of PtaHB1, a *Populus* gene encoding a class III HD-Zip protein, is closely associated with secondary growth and inversely correlated with the level of microRNA (miR166). *New Phytol.* **169**: 469–478.

Kohler, A., Delaruelle, C., Martin, D., Encelot, N., and Martin, F. 2003. The poplar root transcriptome: analysis of 7000 expressed sequence tags. *Febs Lett.* **542**: 37–41.

Kohler, A., Rinaldi, C., Duplessis, S., Baucher, M., Geelen, D., Duchaussoy, F., Meyers, B. C., Boerjan, W., and Martin, F. 2008. Genome-wide identification of NBS resistance genes in *Populus trichocarpa*. *Plant Mol. Biol.* **66**: 619–636.

Krom, N., and Ramakrishna, W. 2008. Comparative analysis of divergent and convergent gene pairs and their expression patterns in rice, *Arabidopsis*, and *Populus*. *Plant Physiol.* **147**: 1763–1773.

Leple, J. C., Dauwe, R., Morreel, K., Storme, V., Lapierre, C., Pollet, B., Naumann, A., Kang, K. Y., Kim, H., Ruel, K., *et al.* 2007. Downregulation of cinnamoyl-coenzyme A reductase in poplar: multiple-level phenotyping reveals effects on cell wall polymer metabolism and structure. *Plant Cell* **19**: 3669–3691.

Leseberg, C. H., Li, A., Kang, H., Duvall, M., and Mao, L. 2006. Genome-wide analysis of the MADS-box gene family in *Populus trichocarpa*. *Gene* 378: 84–94.

Li, J., Brader, G., and Palva, E. T. 2008. Kunitz trypsin inhibitor: An antagonist of cell death triggered by phytopathogens and fumonisin B1 in *Arabidopsis*. *Mol. Plant* **1**: 482–495.

Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E. S., Casstevens, T., Hurwitz, B., McCouch, S., Ni, J., Pujar, A., *et al.* 2008. Gramene: a growing plant comparative genomics resource. *Nucleic Acids Res.* **36**: D947–953.

Liang, P. 2002. A decade of differential display. *Biotechniques* **33**: 338–344, 346.

Liscum, E., and Reed, J. W. 2002. Genetics of Aux/IAA and ARF action in plant growth and development. *Plant Mol. Biol.* **49**: 387–400.

Lu, S., Sun, Y. H., Shi, R., Clark, C., Li, L., and Chiang, V. L. 2005. Novel and mechanical stress-responsive MicroRNAs in *Populus trichocarpa* that are absent from *Arabidopsis*. *Plant Cell* **17**: 2186–2203.

Ma, C., Strauss, S. H., and Meilan, R. 2004. *Agrobacterium*-mediated transformation of the genome-sequenced poplar clone, Nisqually-1 (*Populus trichocarpa*). *Plant Mol. Biol. Reporter* **22**: 311–312.

Major, I. T., and Constabel, C. P. 2008. Functional analysis of the Kunitz trypsin inhibitor family in poplar reveals biochemical diversity and multiplicity in defense against herbivores. *Plant Physiol.* **146**: 888–903.

Markussen, T., Pakull, B., and Fladung, M. 2007. Positioning of sex-correlated markers for *Populus* in a AFLP- and SSR-Marker based genetic map of *Populus tremula* x *tremuloides*. *Silvae Genetica* **56**: 180–184.

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E. G. J., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., *et al.* 2008. The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature*. **452**: 88–92.

Martin, F., Tuskan, G. A., DiFazio, S. P., Lammers, P., Newcombe, G., and Podila, G. K. 2004. Symbiotic sequencing for the *Populus mesocosm*. *New Phytol.* **161**: 330–335.

Mason, M. G., Mathews, D. E., Argyros, D. A., Maxwell, B. B., Kieber, J. J., Alonso, J. M., Ecker, J. R., and Schaller, G. E. 2005. Multiple type-B response regulators mediate cytokinin signal transduction in *Arabidopsis*. *Plant Cell* **17**: 3007–3018.

Matsubara, S., Hurry, V., Druart, N., Benedict, C., Janzik, I., Chavarria-Krauser, A., Walter, A., and Schurr, U. 2006. Nocturnal changes in leaf growth of *Populus deltoides* are controlled by cytoplasmic growth. *Planta* **223**: 1315–1328.

McHale, L., Tan, X. P., Koehl, P., and Michelmore, R. W. 2006. Plant NBS-LRR proteins: adaptable guards. *Genome Biol.* **7**: 212.

Meijer, M., and Murray, J. A. H. 2000. The role and regulation of D-type cyclins in the *Plant Cell* cycle. *Plant Mol. Biol.* **43**: 621–633.

Menges, M., Pavesi, G., Morandini, P., Bogre, L., and Murray, J. A. 2007. Genomic organization and evolutionary conservation of plant D-type cyclins. *Plant Physiol.* **145**: 1558–1576.

Meyermans, H., Morreel, K., Lapierre, C., Pollet, B., De Bruyn, A., Busson, R., Herdewijn, P., Devreese, B., Van Beeumen, J., Marita, J. M., *et al.* 2000. Modifications in lignin and accumulation of phenolic glucosides in poplar xylem upon down-regulation of caffeoyl-coenzyme A O-methyltransferase, an enzyme involved in lignin biosynthesis. *J. Biol. Chem.* **275**: 36899–36909.

Miranda, M., Ralph, S. G., Mellway, R., White, R., Heath, M. C., Bohlmann, J., and Constabel, C. P. 2007. The transcriptional response of hybrid poplar (*Populus trichocarpa* × *P. deltoides*) to infection by *Melampsora medusae* leaf rust involves induction of flavonoid pathway genes leading to the accumulation of proanthocyanidins. *Mol. Plant-Microbe Interact.* **20**: 816–831.

Mo, B., and Bewley, J. D. 2003. The relationship between beta-mannosidase and endo-beta-mannanase activities in tomato seeds during and following germination: a comparison of seed populations and individual seeds. *J. Exp. Bot.* **54**: 2503–2510.

Moreau, C., Aksenov, N., Lorenzo, M. G., Segerman, B., Funk, C., Nilsson, P., Jansson, S., and Tuominen, H. 2005. A genomic approach to investigate developmental cell death in woody tissues of *Populus* trees. *Genome Biol.* **6**: R34.

Morreel, K., Goeminne, G., Storme, V., Sterck, L., Ralph, J., Coppieters, W., Breyne, P., Steenackers, M., Georges, M., Messens, E., *et al.* 2006. Genetical metabolomics of flavonoid biosynthesis in *Populus*: a case study. *Plant J.* **47**: 224–237.

Morse, A. M., Tschaplinski, T. J., Dervinis, C., Pijut, P. M., Schmelz, E. A., Day, W., and Davis, J. M. 2007. A salicylate hydroxylase transgene in poplar induces compensatory mechanisms in the shikimate and phenylpropanoid pathways. *Phytochemistry* **68**: 2043–2052.

Nakano, T., Suzuki, K., Fujimura, T., and Shinshi, H. 2006. Genome-wide analysis of the ERF gene family in *Arabidopsis* and rice. *Plant Physiol.* **140**: 411–432.

Nanjo, T., Futamura, N., Nishiguchi, M., Igasaki, T., Shinozaki, K., and Shinohara, K. 2004. Characterization of full-length enriched expressed sequence tags of stress-treated poplar leaves. *Plant Cell Physiol.* **45**: 1738–1748.

Nanjo, T., Sakurai, T., Totoki, Y., Toyoda, A., Nishiguchi, M., Kado, T., Igasaki, T., Futamura, N., Seki, M., Sakaki, Y., *et al.* 2007. Functional annotation of 19,841 *Populus nigra* full-length enriched cDNA clones. *BMC Genomics* 8:448.

Neale, D. B., and Savolainen, O. 2004. Association genetics of complex traits in conifers. *Trends Plant Sci.* **9**: 325–330.

Nilsson, J., Karlberg, A., Antti, H., Lopez-Vernaza, M., Mellerowicz, E., Perrot-Rechenmann, C., Sandberg, G., and Bhalerao, R. P. 2008. Dissecting the molecular basis of the regulation of wood formation by auxin in hybrid aspen. *Plant Cell* **20**: 843–855.

Oakley, R. V., Wang, Y. S., Ramakrishna, W., Harding, S. A., and Tsai, C. J. 2007. Differential expansion and expression of alpha- and beta-tubulin gene families in *Populus*. *Plant Physiol.* **145**: 961–973.

Okamuro, J. K., Caster, B., Villarroel, R., Van Montagu, M., and Jofuku, K. D. 1997. The AP2 domain of APETALA2 defines a large new family of DNA binding proteins in *Arabidopsis*. *Proc. Natl. Acad. Sci. USA* **94**: 7076–7081.

Pandey, A., and Mann, M. 2000. Proteomics to study genes and genomes. *Nature*. **405**: 837–846.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D. S., Favalli, C., Busscher, J., Cook, H. E., Ingram, R. M., Kater, M. M., Davies, B., *et al.* 2003. Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in *Arabidopsis*: New openings to the MADS world. *Plant Cell* **15**: 1538–1551.

Parker, J. D., Bradley, B. A., Mooers, A. O., and Quarmby, L. M. 2007. Phylogenetic analysis of the Neks reveals early diversification of ciliary-cell cycle kinases. PLoS ONE **2**: e1076.

Pietrowska-Borek, M., Stuible, H. P., Kombrink, E., and Guranowski, A. 2003. 4-Coumarate: coenzyme A ligase has the catalytic capacity to

synthesize and reuse various (di)adenosine polyphosphates. *Plant Physiol* **131**: 1401–1410.

Plomion, C., Lalanne, C., Claverol, S., Meddour, H., Kohler, A., Bogeat-Triboulot, M. B., Barre, A., Le Provost, G., Dumazet, H., Jacob, D., *et al.* 2006. Mapping the proteome of poplar and application to the discovery of drought-stress responsive proteins. *Proteomics* 6**:** 6509–6527.

Qin, H., Feng, T., Harding, S. A., Tsai, C. J., and Zhang, S. 2008. An efficient method to identify differentially expressed genes in microarray experiments. *Bioinformatics* **24**: 1583–1589.

Quesada, T., Li, Z., Dervinis, C., Li, Y., Bocock, P. N., Tuskan, G. A., Casella, G., Davis, J. M., and Kirst, M. 2008. Comparative analysis of the transcriptomes of *Populus trichocarpa* and *Arabidopsis thaliana* suggests extensive evolution of gene expression regulation in angiosperms. *New Phytol.* **180**: 408–420.

Rae, A. M., Pinel, M. P. C., Bastien, C., Sabatti, M., Street, N. R., Tucker, J., Dixon, C., Marron, N., Dillen, S. Y., and Taylor, G. 2008. QTL for yield in bioenergy *Populus*: identifying G×E interactions from growth at three contrasting sites. *Tree Genetics & Genomes* **4**: 97–112.

Rae, A. M., Tricker, P. J., Bunn, S. M., and Taylor, G. 2007. Adaptation of tree growth to elevated CO2: quantitative trait loci for biomass in *Populus*. *New Phytol.* **175**: 59–69.

Rajinikanth, M., Harding, S. A., and Tsai, C. J. 2007. The glycine decarboxylase complex multienzyme family in *Populus*. *J. Exp. Bot.* **58**: 1761–1770.

Ralph, S., Oddy, C., Cooper, D., Yueh, H., Jancsik, S., Kolosova, N., Philippe, R. N., Aeschliman, D., White, R., Huber, D., *et al.* 2006. Genomics of hybrid poplar (*Populus trichocarpa* × *deltoides*) interacting with forest tent caterpillars (*Malacosoma disstria*): normalized and full-length cDNA libraries, expressed sequence tags, and a cDNA microarray for the study of insect-induced defences in poplar. *Mol. Ecol.* **15**: 1275–1297.

Ralph, S. G., Chun, H. J. E., Cooper, D., Kirkpatrick, R., Kolosova, N., Gunter, L., Tuskan, G. A., Douglas, C. J., Holt, R. A., Jones, S. J. M., *et al.* 2008. Analysis of 4,664 high-quality sequence-finished poplar full-length cDNA clones and their utility for the discovery of genes responding to insect feeding. *BMC Genomics* **9**: 57.

Ramirez-Carvajal, G. A., Morse, A. M., and Davis, J. M. 2008. Transcript profiles of the cytokinin response regulator gene family in *Populus* imply diverse roles in plant development. *New Phytol.* **177**: 77–89.

Ranjan, P., Kao, Y. Y., Jiang, H., Joshi, C. P., Harding, S. A., and Tsai, C. J. 2004. Suppression subtractive hybridization-mediated transcriptome analysis from multiple tissues of aspen (*Populus tremuloides*) altered in phenylpropanoid metabolism. *Planta* **219**: 694–704.

Rinaldi, C., Kohler, A., Frey, P., Duchaussoy, F., Ningre, N., Couloux, A., Wincker, P., Le Thiec, D., Fluch, S., Martin, F., *et al.* 2007. Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina*. *Plant Physiol.* **144**: 347–366.

Roberts, A. W., and Bushoven, J. T. 2007. The cellulose synthase (CESA) gene superfamily of the moss *Physcomitrella patens*. *Plant Mol. Biol.* **63**: 207–219.

Rohde, A., Ruttink, T., Hostyn, V., Sterck, L., Van Driessche, K., and Boerjan, W. 2007. Gene expression during the induction, maintenance, and release of dormancy in apical buds of poplar. *J. Exp. Bot.* **58**: 4047–4060.

Rose, J. K. C., Braam, J., Fry, S. C., and Nishitani, K. 2002. The XTH family of enzymes involved in xyloglucan endotransglucosylation and endohydrolysis: Current perspectives and a new unifying nomenclature. *Plant Cell Physiol.* **43**: 1421–1435.

Ruttink, T., Arend, M., Morreel, K., Storme, V., Rombauts, S., Fromm, J., Bhalerao, R. P., Boerjan, W., and Rohde, A. 2007. A molecular timetable for apical bud formation and dormancy induction in poplar. *Plant Cell* **19**: 2370–2390.

Sasaki, S., Shimizu, M., Wariishi, H., Tsutsumi, Y., and Kondo, R. 2007. Transcriptional and translational analyses of poplar anionic peroxidase isoenzymes. *J. Wood Sci*. **53**: 427–435.

Scarascia-Mugnozza, G. E., Hinckley, T. M., Stettler, R. F., Heilman, P. E., and Isebrands, J. G. 1999. Production physiology and morphology of *Populus* species and their hybrids grown under short rotation. III. Seasonal carbon allocation patterns from branches. *Can. J. Forest Res*. **29**: 1419–1432.

Schrader, J., Nilsson, J., Mellerowicz, E., Berglund, A., Nilsson, P., Hertzberg, M., and Sandberg, G. 2004. A high-resolution transcript profile across the wood-forming meristem of poplar identifies potential regulators of cambial stem cell identity. *Plant Cell* **16**: 2278–2292.

Service, R. F. 2006. Gene sequencing. The race for the $1000 genome. *Science* **311**: 1544–1546.

Sjödin, A., Street, N., Sandberg, G., Gustafsson, P., and Jansson, S. 2008. PopGenIE: The *Populus* Genome Integrative Explorer. A new tool for exploring the *Populus* genome. www.popgenie.org.

Sjodin, A., Wissel, K., Bylesjo, M., Trygg, J., and Jansson, S. 2008. Global expression profiling in leaves of free-growing aspen. *BMC Plant Biol.* **8**: 61.

Song, J., Lu, S., Chen, Z. Z., Lourenco, R., and Chiang, V. L. 2006. Genetic transformation of *Populus trichocarpa* genotype Nisqually-1: a functional genomic tool for woody plants. *Plant Cell Physiol*. **47**: 1582–1589.

Souza Cde, A., Barbazuk, B., Ralph, S. G., Bohlmann, J., Hamberger, B., and Douglas, C. J. 2008. Genome-wide analysis of a land plant-specific acyl:coenzyme A synthetase (ACS) gene family in *Arabidopsis*, poplar, rice and *Physcomitrella*. *New Phytol*. **179**: 987–1003.

Sterky, F., Bhalerao, R. R., Unneberg, P., Segerman, B., Nilsson, P., Brunner, A. M., Charbonnel-Campaa, L., Lindvall, J. J., Tandre, K., Strauss, S. H., *et al.* 2004. A *Populus* EST resource for plant functional genomics. *Proc. Nat. Acad. Sciences USA* **101**: 13951–13956.

Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., and Villarroel, R. 1998. Gene discovery in the wood-forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proc. Natl. Acad. Sci. USA* **95**: 13330–13335.

Stracke, R., Werber, M., and Weisshaar, B. 2001. The R2R3-MYB gene family in *Arabidopsis thaliana*. *Curr. Opin. Plant Biol.* **4**: 447–456.

Strauss, S. H. 2003. Genetic technologies – Genomics, genetic engineering, and domestication of crops. *Science* **300**: 61–62.

Strauss, S. H., Lande, R., and Namkoong, G. 1992. Limitations of molecular-marker-aided selection in forest tree breeding. *Can. J. Forest Res.* **22**: 1050–1061.

Street, N. R., Skogstrom, O., Sjodin, A., Tucker, J., Rodriguez-Acosta, M., Nilsson, P., Jansson, S., and Taylor, G. 2006. The genetics and genomics of the drought response in *Populus*. *Plant J.* **48**: 321–341.

Sun, M., Zhou, G., Lee, S., Chen, J., Shi, R. Z., and Wang, S. M. 2004. SAGE is far more sensitive than EST for detecting low-abundance transcripts. *BMC Genomics* **5**: 1.

Suzuki, S., Li, L., Sun, Y. H., and Chiang, V. L. 2006. The cellulose synthase gene superfamily and biochemical functions of xylem-specific cellulose synthase-like genes in *Populus trichocarpa*. *Plant Physiol.* **142**: 1233–1245.

Tagu, D., Bastien, C., Faivre-Rampant, P., Garbaye, J., Vion, P., Villar, M., and Martin, F. 2005. Genetic analysis of phenotypic variation for ectomycorrhiza formation in an interspecific F1 poplar full-sib family. *Mycorrhiza* **15**: 87–91.

Takata, N., Saito, S., Tanaka Saito, C., Nanjo, T., Shinohara, K., and Uemura, M. 2009. Molecular phylogeny and expression of poplar circadian clock genes, LHY1 and LHY2. *New Phytol.* **181**: 808–819.

Tang, H., Bowers, J. E., Wang, X., Ming, R., Alam, M., and Paterson, A. H. 2008a. Synteny and collinearity in plant genomes. *Science* **320**: 486–488.

Tang, H., Wang, X., Bowers, J. E., Ming, R., Alam, M., and Paterson, A. H. 2008b. Unraveling ancient hexaploidy through multiply-aligned angiosperm gene maps. *Genome Res*. **18**: 1944–1954.

Taylor, G. 2002. *Populus*: Arabidopsis for forestry. Do we need a model tree? *Ann. Bot.* **90**: 681–689.

To, J. P., Deruere, J., Maxwell, B. B., Morris, V. F., Hutchison, C. E., Ferreira, F. J., Schaller, G. E., and Kieber, J. J. 2007. Cytokinin regulates type-A *Arabidopsis* Response Regulator activity and protein stability via two-component phosphorelay. *Plant Cell* **19**: 3901–3914.

Tringe, S. G., and Rubin, E. M. 2005. Metagenomics: DNA sequencing of environmental samples. *Nature Reviews Genetics* **6**: 805–814.

Tsai, C. J., Harding, S. A., Tschaplinski, T. J., Lindroth, R. L., and Yuan, Y. N. 2006. Genome-wide analysis of the structural genes regulating defense phenylpropanoid metabolism in *Populus*. *New Phytol.* **172**: 47–62.

Tsai, C. J., Popko, J. L., Mielke, M. R., Hu, W. J., Podila, G. K., and Chiang, V. L. 1998. Suppression of O-methyltransferase gene by homologous sense transgene in quaking aspen causes red-brown wood phenotypes. *Plant Physiol.* **117**: 101–112.

Tsai, C. J., Ranjan, P., DiFazio, S. P., Tuskan, G. A., Johnson, V., and Joshi, C. P. 2009. Poplar genome microarrays. **In**: Joshi, C. P., and DiFazio, S. P., Eds., *Genetics, Genomics and Breeding of Crop Plants: Poplar.* Science Publishers, Enfield, New Hampshire.

Tschaplinski, T. J., Tuskan, G. A., Sewell, M. M., Gebre, G. M., Donald, E. T. I., and Pendleyi, C. 2006. Phenotypic variation and quantitative trait locus identification for osmotic potential in an interspecific hybrid inbred F-2 poplar pedigree grown in contrasting environments. *Tree Physiol.* **26**: 595–604.

Tschaplinski, T. J., Yin, T.-M., and Engle, N. 2005. Combining metabolomics and QTL analysis for identifying mQTL and gene discovery in poplar. Breakthrough Technologies Forum of IUFRO Tree Biotechnology Meeting, Pretoria, South Africa, November 6–11, 2005. Abstract S3.5.

Turck, F., Fornara, F., and Coupland, G. 2008. Regulation and identity of florigen: FLOWERING LOCUS T moves center stage. *Ann. Rev. Plant Biol*. **59**: 573–594.

Tuskan, G., and Walsh, M. 2001. Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: a US case study. *Forestry Chronicle* **77**: 259–264.

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* **313**: 1596–1604.

Tuskan, G. A., DiFazio, S. P., and Teichmann, T. 2004. Poplar genomics is getting popular: The impact of the poplar genome project on tree research. *Plant Biol.* **6**: 2–4.

Unneberg, P., Stromberg, M., Lundeberg, J., Jansson, S., and Sterky, F. 2005. Analysis of 70,000 EST sequences to study divergence between two closely related *Populus* species. *Tree Genetics & Genomes* **1**: 109–115.

van der Hoorn, R. A. L., and Jones, J. D. 2004. The plant proteolytic machinery and its role in defence. *Curr. Opin. Plant Biol.* **7**: 400–407.

Vigneault, F., Lachance, D., Cloutier, M., Pelletier, G., Levasseur, C., and Seguin, A. 2007. Members of the plant NIMA-related kinases are involved in organ development and vascularization in poplar, Arabidopsis and rice. *Plant J.* **51**: 575–588.

Wan, J., Zhang, X. C., Neece, D., Ramonell, K. M., Clough, S., Kim, S. Y., Stacey, M. G., and Stacey, G. 2008. A LysM receptor-like kinase plays a critical role in chitin signaling and fungal resistance in *Arabidopsis*. *Plant Cell* **20**: 471–481.

Watkins, P. A., Maiguel, D., Jia, Z., and Pevsner, J. 2007. Evidence for 26 distinct acyl-coenzyme A synthetase genes in the human genome. *J. Lipid Res.* **48**: 2736–2750.

Weckwerth, W. 2003. Metabolomics in systems biology. *Ann. Rev. Plant Biol.* **54**: 669–689.

Weigel, D., Ahn, J. H., Blazquez, M. A., Borevitz, J. O., Christensen, S. K., Fankhauser, C., Ferrandiz, C., Kardailsky, I., Malancharuvil, E. J., Neff, M. M., *et al.* 2000. Activation tagging in *Arabidopsis*. *Plant Physiol.* **122**: 1003–1013.

Whitham, T. G., DiFazio, S. P., Schweitzer, J. A., Shuster, S. M., Allan, G. J., Bailey, J. K., and Woolbright, S. A. 2008. Perspective – Extending genomics to natural communities and ecosystems. *Science* **320**: 492–495.

Wilkins, O., Nahal, H., Foong, J., Provart, N. J., and Campbell, M. M. 2009. Expansion and diversification of the *Populus* R2R3-MYB family of transcription factors. *Plant Physiol.* **149**: 981–993.

Wise, R. P., Caldo, R. A., Hong, L., Shen, L., Cannon, E., and Dickerson, J. A. 2007. BarleyBase/PLEXdb. *Methods Mol. Biol.* **406**: 347–363.

Wu, R., Bradshaw, H. D. J., and Stettler, R. F. 1997. Molecular genetics of growth and development in *Populus* (Salicaceae). V. Mapping quantitative trait loci affecting leaf variation. *American Journal of Beauty*. **84**: 143–153.

Wu, R. L. 1998. Genetic mapping of QTLs affecting tree growth and architecture in *Populus*: implication for ideotype breeding 3079. *Theoret. Applied Genet.* **96**: 447–457.

Wu, R. L., and Lin, M. 2006. Opinion – Functional mapping – how to map and study the genetic architecture of dynamic complex traits. *Nature Rev. Genetics* **7**: 229–237.

Wullschleger, S., Yin, T. M., DiFazio, S. P., Tschaplinski, T. J., Gunter, L. E., Davis, M. F., and Tuskan, G. A. 2005. Phenotypic variation in growth and biomass distribution for two advanced-generation pedigrees of hybrid poplar. *Can. J. Forest Res.* **35**: 1779–1789.

Wullschleger, S. D., Jansson, S., and Taylor, G. 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* **14**: 2651–2655.

Wullschleger, S. D., Tuskan, G. A., and DiFazio, S. P. 2002. Genomics and the tree physiologist. *Tree Physiol.* **22**: 1273–1276.

Yang, X. H., Jawdy, S., Tschaplinski, T. J., and Tuskan, G. A. 2009. Genome-wide identification of lineage-specific genes in *Arabidopsis*, *Oryza* and *Populus*. *Genomics* **93**: 473–480.

Yang, X. H., Kalluri, U. C., Jawdy, S., Gunter, L. E., Yin, T. M., Tschaplinski, T. J., Weston, D. J., Ranjan, P., and Tuskan, G. A. 2008. The F-box gene family is expanded in herbaceous annual plants relative to woody perennial plants. *Plant Physiol.* **148**: 1189–1200.

Yang, X. H., Tuskan, G. A., and Cheng, Z. M. 2006. Divergence of the Dof gene families in poplar, Arabidopsis, and rice suggests multiple modes of gene evolution after duplication. *Plant Physiol.* **142**: 820–830.

Ye, X., Kang, B. G., Osburn, L. D., and Cheng, Z. M. 2009a. The COBRA gene family in *Populus* and gene expression in vegetative organs and in response to hormones and environmental stresses. *Plant Growth Regulation* **58**: 211–223.

Ye, X., Kang, B. G., Osburn, L. D., Li, Y., and Cheng, Z. M. 2009b. Identification of the flavindependent monooxygenase-encoding YUCCA gene family in *Populus trichocarpa* and their expression in vegetative tissues and in response to hormone and environmental stresses. *Plant Cell Tissue and Organ Culture* **97**: 271–283.

Yin, T., DiFazio, S. P., Gunter, L. E., Zhang, X., Sewell, M. M., Woolbright, S. A., Allan, G. J., Kelleher, C. T., Douglas, C. J., Wang, M., *et al.* 2008. Genome structure and emerging evidence of an incipient sex chromosome in *Populus*. *Genome Res.* **18**: 422–430.

Yuan, J. S., Yang, X. H., Lai, J. R., Lin, H., Cheng, Z. M., Nonogaki, H., and Chen, F. 2007. The endo-beta-mannanase gene families in *Arabidopsis*, rice, and poplar. *Funct. Integ. Genomics* **7**: 1–16.

Zhang, H. C., Yin, W. L., and Xia, X. L. 2008. Calcineurin B-Like family in *Populus*: comparative genome analysis and expression pattern under cold, drought and salt stress treatment. *Plant Growth Reg.* **56**: 129–140.

Zhang, X. C., Wu, X., Findley, S., Wan, J., Libault, M., Nguyen, H. T., Cannon, S. B., and Stacey, G. 2007. Molecular evolution of lysin motif-type receptor-like kinases in plants. *Plant Physiol.* **144**: 623–636.

Zhou, F., and Xu, Y. 2009. RepPop: a database for repetitive elements in *Populus trichocarpa*. *BMC Genomics* **10**: 14.

Zhu, Q. H., Guo, A. Y., Gao, G., Zhong, Y. F., Xu, M., Huang, M., and Luo, J. 2007. DPTF: a database of poplar transcription factors. *Bioinformatics* **23**: 1307–1308.

Zhuang, J., Cai, B., Peng, R. H., Zhu, B., Jin, X. F., Xue, Y., Gao, F., Fu, X. Y., Tian, Y. S., Zhao, W., *et al.* 2008. Genome-wide analysis of the AP2/ERF gene family in *Populus trichocarpa*. *Biochem. Biophys. Res. Commun.* **371**: 468–474.

Zhuang, Y., and Adams, K. L. 2007. Extensive allelic variation in gene expression in *Populus* F_1 hybrids. *Genetics* **177**: 1987–1996.

New Phytologist

Methods

Microsatellite primer resource for *Populus* developed from the mapped sequence scaffolds of the Nisqually-1 genome

T. M. Yin[1], X. Y. Zhang[1], L. E. Gunter[1], S. X. Li[2], S. D. Wullschleger[1], M. R. Huang[2] and G. A. Tuskan[1]

[1]Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6422, USA; [2]The Key Lab of Forest Genetics and Gene Engineering, Nanjing Forestry University, Nanjing, China

Author for correspondence:
Gerald A. Tuskan
Tel: +1 (865) 576-8141
Fax: +1 (865) 576-9939
Email: gtk@ornl.gov

Received: *22 July 2008*
Accepted: *9 September 2008*

New Phytologist (2009) **181**: 498–503
doi: 10.1111/j.1469-8137.2008.02663.x

Key words: allelic variability, amplification rate, microsatellites, *Populus*, simple sequence repeat (SSR) primers.

Summary

- In this study, 148 428 simple sequence repeat (SSR) primer pairs were designed from the unambiguously mapped sequence scaffolds of the Nisqually-1 genome. The physical position of the priming sites were identified along each of the 19 *Populus* chromosomes, and it was specified whether the priming sequences belong to intronic, intergenic, exonic or UTR regions.
- A subset of 150 SSR loci were amplified and a high amplification success rate (72%) was obtained in *P. tremuloides,* which belongs to a divergent subgenus of *Populus* relative to Nisqually-1. PCR reactions showed that the amplification success rate of exonic primer pairs was much higher than that of the intronic/intergenic primer pairs.
- Applying ANOVA and regression analyses to the flanking sequences of microsatellites, the repeat lengths, the GC contents of the repeats, the repeat motif numbers, the repeat motif length and the base composition of the repeat motif, it was determined that only the base composition of the repeat motif and the repeat motif length significantly affect the microsatellite variability in *P. tremuloides* samples.
- The SSR primer resource developed in this study provides a database for selecting highly transferable SSR markers with known physical position in the *Populus* genome and provides a comprehensive genetic tool to extend the genome sequence of Nisqually-1 to genetic studies in different *Populus* species.

Introduction

The genus *Populus* possesses many characteristics that are conducive to functional genomic studies and as such it has been widely accepted as a model system in tree genomic research (Wullschleger *et al.*, 2002). Under the efforts of numerous scientists worldwide, the genome of a black cottonwood (*Populus trichocarpa* Torr. & Gray ex Brayshaw), clone 383–2499, 'Nisqually-1', has been sequenced and publicly released (Tuskan *et al.*, 2006). It is the first sequence of a woody perennial plant. However, the applicability of the Nisqually-1 genome sequence to studies of alternate *Populus* genotypes and species remains undetermined.

Microsatellites or simple sequence repeats (SSRs) have been shown to be among the most powerful genetic markers for aligning the genome of different species (Yin *et al.*, 2004), genetic fingerprinting (Schlotterer, 2001), linkage analysis (Dib *et al.*, 1996), population genetics (Wyman *et al.*, 2003) and clonal fidelity (Rajora & Rahman, 2004). Earlier studies suggest that SSRs are potentially transferable across genera of *Salicaceae* (Tuskan *et al.*, 2004; Hanley *et al.*, 2006). Moreover, the *Populus* genome project revealed that the chromosomal structure in modern *Populus* arose from an ancient whole-genome duplication event known as 'salicoid' duplication (Tuskan *et al.*, 2006) and our recent comparative mapping study demonstrated that genomes of alternate *Populus* species maintained the basic genome structure after salicoid duplication (Yin *et al.*, 2008). Therefore, it may be feasible to use SSRs to build a platform to study all *Populus* taxa as a macrogenetic system and to validate genetic findings across different *Populus* species.

To date, SSR primers for *Populus* have been designed from sequences that were randomly selected based on either library enrichment or shotgun sequencing strategies from various *Populus* species (Tuskan *et al.*, 2004). Many of these primers' sequences show low to no homology to the genome sequence of Nisqually-1 and thus no reliable physical position can be deduced for these loci, impairing their utility and application. As a resource for the international *Populus* community, we developed primers that amplified microsatellites consisting of repetitive motifs of 2–5 bp from the unambiguously mapped sequence scaffolds of the Nisqually-1 genome. Our primary objective was to create a publicly available comprehensive genetic resource for *Populus*; our secondary objective was to test the utility and allelic variability of these SSR loci in *P. tremuloides* (Michx.), a member of a divergent subgenus within *Populus*.

Materials and Methods

Sputnik program coding with C language (C. Abajian, University of Washington, USA) was used to search DNA sequence files in Fasta format for microsatellite repeats. SSR primers were subsequently designed by Primer 3 (http://frodo.wi.mit.edu/cgi-bin/primer3/primer3_www.cgi). SSR repeat and primer distributions, which are classified into intronic, intergenic, exonic, or UTR regions, were derived from a Fortran coding program created by the authors based on the comparison of the physical locations of SSRs and genes annotated at the US Department of Energy's Joint Genome Institute homepage (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html). SSR density was calculated as the number of SSR loci within a moving 2 Mb window divided by the effective A, T, G, C reads within the same region.

To validate SSR interspecific utility, 100 SSR primer pairs in intronic/intergenic genomic space and 50 in exonic genomic space were selected from across the Nisqually-1 genome to amplify *P. tremuloides* template DNA in order to evaluate amplification success rate and allelic variability. Initially, a single *P. tremuloides* genotype was used to measure amplification success rate, and then 10 alternate *P. tremuloides* genotypes were used to test allelic variability. PCR reactions were performed as described by Tuskan *et al.* (2004) and Yin *et al.* (2004). The electrophoresis conditions were controlled by the default module for microsatellite genotyping on the ABI 3730xl (Applied Biosystems, Foster City, CA, USA). Regression and ANOVA analyses were performed to test the influence of different factors on the allelic variability. An Anderson-Darling test was used to test for normality of SSR distribution at significance levels $P \leq 0.05$ and $P \leq 0.01$.

In order to compare the amplification success rates across members of different subgenus in *Salicaceae*, we recorded PCR amplification success from 100 randomly selected SSR primer pairs for two *P. trichocarpa*, *P. deltoides* and *P. fremontii* genotypes and a single genotype for all other species within the genus.

Results

A total of 148 428 SSR primer pairs were designed from the unambiguously mapped sequence scaffolds of the Nisqually-1 genome. The complete information for all SSR primers is listed in the Supporting Information, Table S1. The principal metrics in this table include a description of the SSR sequence, the physical position of SSR, the melting temperature (T_m) of each primer, the GC content of the primer sequences, and the expected PCR product sizes in Nisqually-1. The visual representation of each SSR primer position per chromosome is shown in Fig. S1.

An Anderson-Darling test for normality across the whole genome shows that the SSR numbers (mean = 524, SD = 55, adjusted $A^2 = 0.445$) in each 2 Mb window were normally distributed at $\alpha = 0.05$ level (critical $A^2 = 0.752$). There were, however, four windows that had SSRs that exceeded the expected number and three windows that had fewer than expected numbers of SSRs (Table S2). Overall these results indicate that there are no large physical gaps among the SSR priming sites within the genome. Thus this primer resource will facilitate the generation of evenly distributed SSR markers across the *Populus* genome.

On average, SSRs occurred approximately every 2.5 kb within the *Populus* genome. At the subchromosomal level, SSR location varied by genic region, with 85.4% found in intergenic regions, 10.7% in introns, 2.7% in exons and 1.2% in UTR. Interestingly, the frequency of SSRs within exons varied by chromosome, from zero SSRs in exons on chromosome V to 765 on chromosome I, averaging 316 SSRs within exons per chromosome across the genome (Fig. 1). Based on the even distribution assumption, chromosome V would be expected to contain 242 microsatellite repeats in exonic regions. Furthermore, chromosome V shares large duplicated segments with chromosomes II, III and VII (Tuskan *et al.*, 2006) and exons in paralogous genes found on these homologous chromosomes do contain SSRs. Thus, for undetermined reasons, it appears that the loss of exonic SSRs is unique to chromosome V and occurred after the salicoid duplication event.

Single sequence repeat amplification success rate across different *Populus* and *Salix* species showed that SSR primer amplification rates were higher among taxa closely related to *P. trichocarpa*, including species of *Leucoides*, *Aigeiros* and *Tacamahaca* subgenera, and lower in *P. tremuloides* (a member of *Leuce*), and lowest among members of *Turanga* section (Fig. 2).

Among the 150 SSR primer pairs tested in a single *P. tremuloides* genotype, 103 produced amplified product (Table S3). The amplification test confirmed that 63% of SSRs with priming sites in intronic/intergenic regions were successfully amplified; by contrast, amplification success rate was 80% for primers designed from exonic sequences (Table S3). We then tested 57 primer pairs, including 25 with priming sites in exonic or UTR sequences and 32 with priming sites in

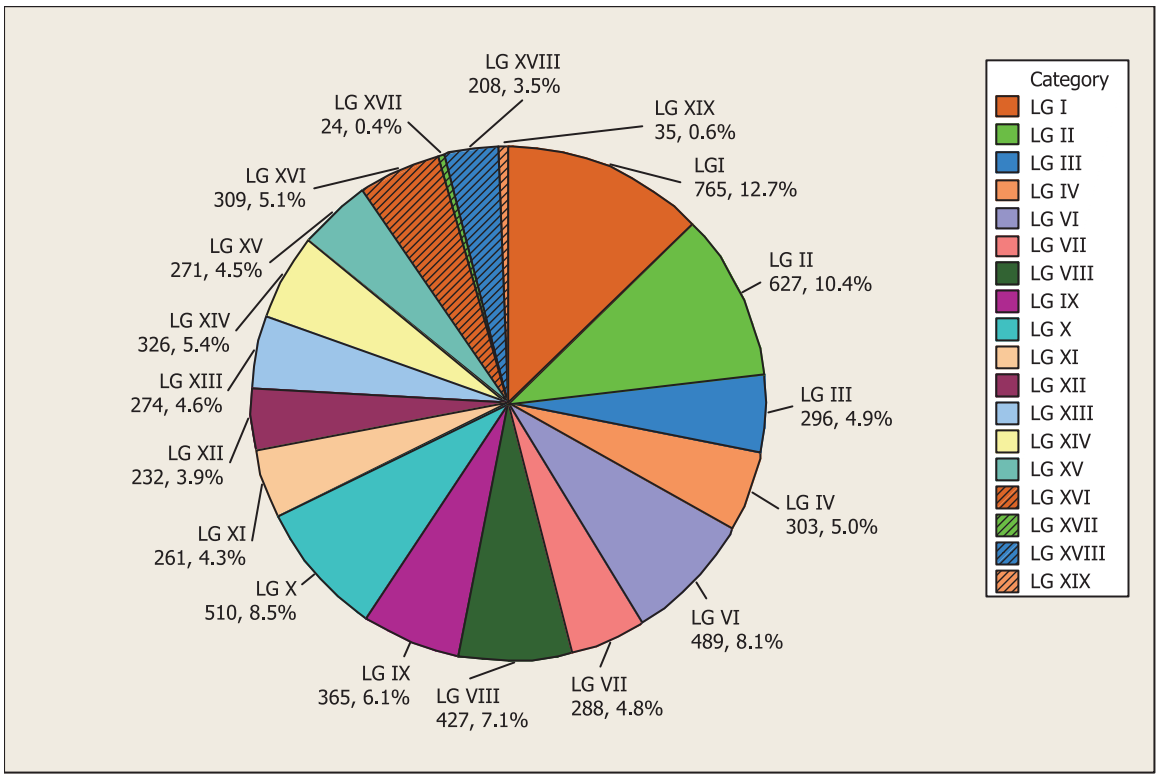


Fig. 1 The distribution of simple sequence repeats (SSRs) in transcribed regions per chromosome (labeled as linkage groups, LG) in the *Populus* genome. Note: no SSRs were found in transcribed regions on chromosome V. Based on the even distribution assumption, chromosome V would be expected to contain 242 microsatellite repeats. The first number underneath the linkage group designation is the number of SSRs located in transcribed regions on each chromosome followed by the percentage of all SSRs located on each chromosome.

intronic/intergenic sequences, among 10 randomly selected *P. tremuloides* genotypes to determine allelic variability among primer pairs. The average allele number revealed by the exonic SSR loci was higher than that obtained per intronic/intergenic SSR loci (4.25 vs 3.25 alleles); however, ANOVA analysis indicated that this difference was not significant ($P > 0.05$). Therefore, it appears that the location of the priming sites significantly influences amplification rate but not allelic variability.

Based on 100 selected primer pairs, allelic variability across members of the genus *Populus* did not significantly vary with SSR repeat length, GC content of the repeats or repeat motif numbers ($F = 0.23$, $F = 0.94$ and $F = 0.502$, respectively, critical $F = 3.84$ at $P \leq 0.05$). However, when we analyzed the allelic variability by repeat motif length, a significant negative correlation was detected such that the average allele number decreased from 4.29, 2.91 and 2.00 in di-, tri- and tetranucleotide repeats, respectively. Therefore, of the tested parameters, only repeat motif length significantly affected the SSR allelic variability among members of the genus.

We also compared the variability of SSR with repeat motifs of [AAT]/[TTA], [AC]/[TG], [AT]/[TA] and [AG]/[TC]. Among these repeat motifs, the [AG]/[TC] motif results in the highest polymorphism rate; [AAT]/[TTA] yields the lowest. Significant differences in allelic variability were detected among $[AAT]_n/[TTA]_n$ vs $[AG]_n/[TC]_n$ ($\alpha \leq 0.01$) and $[AC]_n/[TG]_n$ vs $[AG]_n/[TC]_n$ ($\alpha \leq 0.05$). Thus, the base composition of the SSR repeat motifs significantly affected allelic variability among the SSR primer pairs tested in this study.

Discussion

Our study demonstrates that the microsatellite markers derived from a single clone of *Populus trichocarpa*, Nisqually-1, have relatively high amplification rate in *P. tremuloides*, a member of the *Leuce* subgenera. The genus *Populus* contains six subgenera, including *Abaso*, *Leuce*, *Leucoides*, *Aigeiros*, *Turanga* and *Tacamahaca* (Eckenwalder, 1996; Shi *et al.*, 2001), of which the subgenera *Leuce* is more dissimilar to Nisqually-1 (a member of *Tacamahaca*), than are members of any other subgenus. The amplification success rates were higher in members of all other subgenera, expect for *P. euphratica*, which is the sole representative of the *Turanga* subgenera (Fig. 2). However, the phylogenetic position of *Turanga* is controversial. According to Eckenwalder's consensus cladogram (1996), *Turanga* is the most distant from *Tacamahaca* among all the

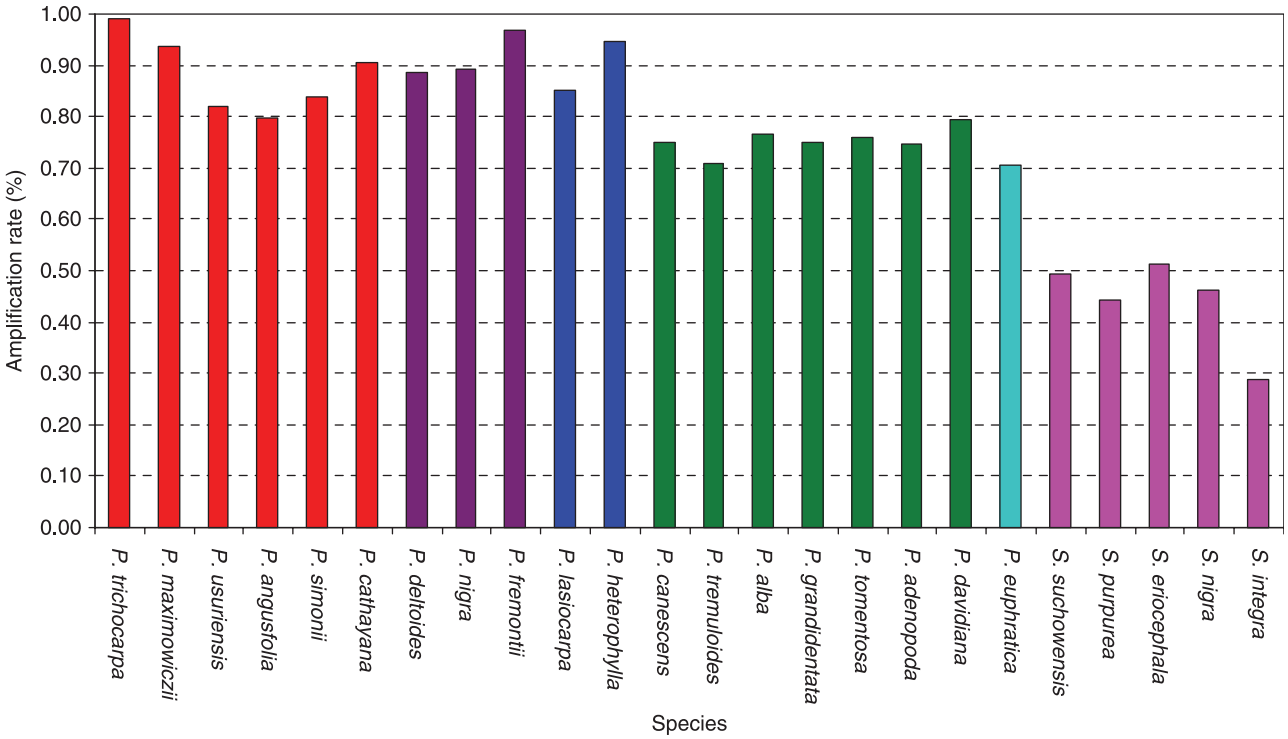


Fig. 2 Amplification rates of simple sequence repeat (SSR) primers across different *Populus* and *Salix* species. These estimates are based on 100 randomly chosen microsatellite primers. The *x*-axis indicates the species and the *y*-axis indicates the amplification rate. Red bars, species that belong to section *Tacamahaca*; purple bars, species that belong to section *Aigeiros*; blue bars, species that belong to section *Leucoides*; green bars, species that belong to section *Leuce*; light blue bar, species that belong to section *Turanga*; and pink bars, *Salix* species. The appearance of subgenus from left to right is ordered according to the increment of their phylogenetic divergence from *Tacamahaca*. For primers and origins of these samples, refer to Tuskan *et al*. (2004) and Yin *et al*. (2008).

subgenera of *Populus*. Nonetheless, we achieved moderate amplification success rate in *P. euphratica* in the *Turanga* subgenus and in members of *Salix*. In general, the amplification success rates in different species are positively correlated with their phylogenetic divergence from Nisqually-1.

It should be noted that the tested primers were randomly selected from an overall list of potential primer sequences in the Nisqually-1 genome. Data in Fig. 2 indicate that the amplification success rate was less than 100% for *P. trichocarpa*. These results suggest that the 72% transferability estimated for *P. tremuloides* and all other tested species is probably an underestimation. Our results indicate that the amplification rate in *P. trichocarpa* is high (*c*. 99%), and therefore the underestimation should be minor.

It is universally recognized that coding sequences are better conserved than noncoding sequences. In this study, we verified that amplification success rates of SSR primers were dramatically influenced by their priming site position (i.e. exon vs intron). Our experimental data demonstrated that primers located in exonic regions of Nisqually-1 have significantly higher amplification rates in a genetically divergent *Populus* species than primers designed from Nisqually-1 noncoding space. Therefore, the exonic primers would be especially useful in supplying a common language for the *Populus* community to communicate and validate findings among different *Populus* studies. In contrast to the amplification rate, the subchromosomal locations of the priming sites did not significantly affect the allelic variability of microsatellites. Evidence of the influence of flanking sequences on microsatellite variability from any organism is limited and inconclusive. Glenn *et al.* (1996) detected significant influence of the flanking sequence in alligator; however, no such effects were detected in studies by Balloux *et al.* (1998) on shrews or by Bachtrog *et al.* (2000) on *Drosophila*. In the only comparative SSR study in plants, Gao & Xu (2008) found that mutation rates of microsatellites did not significantly differ among motifs of di-, tri- and tetranucleotide repeats in four subspecies of cultivated rice *O. sativa* and its three relatives, *O. rufipogon*, *O. glaberrima* and *O. officinalis*.

Despite the wide occurrence of microsatellites, the basis of their variability is still not well understood. In this study, we found that the SSR variability in *P. tremuloides* was significantly influenced by the repeat motif length and the base compositions of the repeat motif, but not by the number of repeat units. Consistent with our findings, Bachtrog *et al.* (2000) reported that the base composition of repeat motif significantly influenced microsatellite mutation rates. However, whereas the repeat number has been observed to be positively

correlated with microsatellite variability in a variety of organisms (Jin *et al*., 1996; Wierdl *et al*., 1997; Schlotterer *et al*., 1998; Schug *et al*., 1998), this trend was not apparent in our study, and it is reasonable to speculate that the repeat number of SSRs is not conserved among highly diverged species. Although we might expect SSRs with longer repeat lengths to reveal higher polymorphism among genotypes, based on our study, this expectation is not warranted. In support of this, the findings among different studies for the influence of the repeat motif length on microsatellite variability are not consistent. There is a trend for a higher mutation rate for SSRs with dinucleotide repeat motifs than SSRs with longer repeat motifs (Chakraborty *et al*., 1997; Kruglyak *et al*., 1998; Schug *et al*., 1998), which is consistent with our results. By contrast, Weber & Wong (1993) observed higher mutation rates for tetranucleotide repeats than for dinucleotide repeats in humans.

The SSR markers developed in this study provide a comprehensive genetic resource that can be used to link findings from other *Populus* studies to the sequenced genome. These primers also represent a valuable resource for the selection of genetic markers for studying population structure, genetic vs geographic variation, and sequence-dependent evolution in *Populus*. The prospective markers can also be used to explore the distribution of recombination hotspots in the *Populus* genome and promote sharing of associated data, such as QTL position validation across unrelated pedigrees. The reported SSRs are especially useful for generating lists of candidate genes occurring in QTL intervals. Moreover, the SSR primers developed in this study can be used to selectively target regions of the whole genome for efficiently closing gaps in the genetic map.

Although the parameters we supplied cannot measure SSR polymorphism *per se*, our survey of the influence of different factors on microsatellite variability provides a reference for selecting SSR loci that potentially yield greater allelic variability. Our study confirmed that SSRs with priming sites in exons had greater utility across species than those with priming sites in introns and intergenic regions. Without such a resource, researchers would have to randomly test several hundred primer pairs to obtain a marker in a target region. This study provides practical information for selecting SSR primers and can thus reduce the time and cost associated with the development of highly transferable, highly polymorphic markers for *Populus* that can be applied to genetic mapping efforts, allelic variation discovery studies, and molecular breeding efforts related to fiber and energy production, carbon sequestration and bioremediation.

Acknowledgements

We thank M. Schuster at the University of Tennessee for establishing the web resources, Dr J. Armento in Oak Ridge for his comments and editing for this manuscript, and R. M. Tuskan for perspectives on objectives and procedures. Special thanks go to the editor and anonymous reviewers for their help in formulating the final revision. Funding for this research was provided by the Educational Department of China (NCET-04-0516), the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory (ORNL) and the US Department of Energy, Office of Science, Biological and Environmental Research Carbon Sequestration Program and Bioenergy Science Center. ORNL is managed by UT-Battelle, LLC for the US Department of Energy under contract no. DE-AC05-00OR22725.

References

Bachtrog D, Agis M, Imhof M, Schlotterer C. 2000. Microsatellite variability differs between dinucleotide repeat motifs: evidence from *Drosophila melanogaster*. *Molecular Biology and Evolution* **17**: 1277–1285.

Balloux F, Ecoffey E, Fumagalli L, Goudet J, Wyttenbach A, Hausser J. 1998. Microsatellite conservation, polymorphism, and GC content in shrews of the genus *Sorex* (*Insectivora, Mammalia*). *Molecular Biology and Evolution* **15**: 473–475.

Chakraborty R, Kimmel M, Stivers D, Davison L, Deka R. 1997. Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proceedings of the National Academy of Sciences, USA* **94**: 1041–1046.

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Kazan J, Seboun E *et al.* 1996. A comprehensive genetic map of the human genome based on 5264 microsatellites. *Nature* **380**: 152–154.

Eckenwalder JE. 1996. Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM, eds. *Biology of* Populus *and its implications for management and conservation*. Ottawa, Canada: NRC Research Press, 7–32.

Gao LZ, Xu HY. 2008. Comparisons of mutation rate variation at genome-wide microsatellites: evolutionary insights from two cultivated rice and their wild relatives. *BMC Evolutionary Biology* **8**: 11. doi:10.1186/1471-2148-8-11.

Glenn TC, Stephan W, Dessauer HC, Braun MJ. 1996. Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability. *Molecular Biology and Evolution* **13**: 1151–1154.

Hanley SJ, Mallott MD, Karp A. 2006. Alignment of a *Salix* linkage map to the *Populus* genomic sequence reveals macrosynteny between willow and *Populus* genomes. *Tree Genetics and Genomes* **3**: 35–48.

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E. 1996. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. *Proceedings of the National Academy of Sciences, USA* **93**: 15285–15288.

Kruglyak S, Durrett RT, Schug MD, Aquadro CF. 1998. Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proceedings of the National Academy of Sciences, USA* **95**: 10774–10778.

Rajora O, Rahman M. 2004. Microsatellite DNA and RAPD fingerprinting, identification and genetic relationships of hybrid poplar (*Populus* × *canadensis*) cultivars. *Theoretical and Applied Genetics* **106**: 470–477.

Schlotterer C. 2001. Genealogical inference of closely related species based on microsatellites. *Genetic Research* **78**: 209–212.

Schlotterer C, Ritter R, Harr B, Brem G. 1998. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. *Molecular Biology and Evolution* **15**: 1269–1274.

Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TF, Aquadro CF. 1998. The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*. *Molecular Biology and Evolution* **15**: 1751–1760.

Shi QL, Zhuge Q, Huang MR, Wang MX. 2001. Phylogenetic relationship of *Populus* sections by ITS sequence analysis. *Acta Botanica Sinica* **43**: 323–325.

Tuskan GA, DiFazio SP, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A *et al.* 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray ex Brayshaw). *Science* **313**: 1596–1604.

Tuskan GA, Gunter LE, Yang ZM, Yin TM, Sewell MM, DiFazio SP. 2004. Characterization of microsatellites revealed by genomic sequencing of *Populus trichocarpa*. *Canadian Journal of Forest Research* **34**: 5–93.

Weber JL, Wong C. 1993. Mutation of human short tandem repeats. *Human Molecular Genetics* **2**: 1123–1128.

Wierdl M, Dominska M, Petes TD. 1997. Microsatellite instability in yeast: dependence on the length of the microsatellite. *Genetics* **146**: 769–779.

Wullschleger SD, Jansson S, Taylor G. 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* **14**: 2651–2655.

Wyman J, Bruneau A, Tremblay MF. 2003. Microsatellite analysis of genetic diversity in four populations of *Populus tremuloides* in Quebec. *Canadian Journal of Botany* **81**: 367.

Yin T, DiFazio SP, Gunter LE, Riemenschneider D, Tuskan GA. 2004. Large-scale heterospecific segregation distortion in *Populus* revealed by a dense genetic map. *Theoretical and Applied Genetics* **109**: 451–463.

Yin TM, DiFazio SP, Gunter LE, Zhang XY, Swell MM, Woolbright S, Allan GJ, Kelleher CT, Douglas CJ, Tuskan GA. 2008. Genome structure and primitive sex chromosome revealed in *Populus*. *Genome Research* **18**: 422–430.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Distribution of microsatellites, including the SSR primers along each *Populus* chromosome.

Table S1 Database of SSR primer sequences and parameters developed from the mapped sequence scaffolds on each *Populus* chromosome

Table S2 The distribution test of SSR densities in a 2 Mb sliding window along each *Populus* chromosome

Table S3 PCR amplification results and primer pair information used in testing SSR amplification success rate and SSR allelic variability in *P. tremuloides*

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the *New Phytologist* Central Office.

Phytochemistry 70 (2009) 32–39

Contents lists available at ScienceDirect

Phytochemistry

journal homepage: www.elsevier.com/locate/phytochem

Two poplar methyl salicylate esterases display comparable biochemical properties but divergent expression patterns

Nan Zhao [a], Ju Guan [a], Farhad Forouhar [b], Timothy J. Tschaplinski [c], Zong-Ming Cheng [a], Liang Tong [b], Feng Chen [a,*]

[a] Department of Plant Sciences, University of Tennessee, 252 Ellington Plant Science Bldg., 2431 Joe Johnson Drive, Knoxville, TN 37996, USA
[b] Department of Biological Sciences, Northeast Structural Genomics Consortium, Columbia University, New York, NY 10027, USA
[c] Environmental Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA

ARTICLE INFO

Article history:
Received 3 June 2008
Received in revised form 27 October 2008
Available online 10 January 2009

Keywords:
Black cottonwood
Populus trichocarpa
Methyl esterase
SABP2
Methyl salicylate
Salicylic acid
Gene family
Molecular modeling

ABSTRACT

Two genes encoding proteins of 98% sequence identity that are highly homologous to tobacco methyl salicylate (MeSA) esterase (SABP2) were identified and cloned from poplar. Proteins encoded by these two genes displayed specific esterase activities towards MeSA to produce salicylic acid, and are named PtSABP2-1 and PtSABP2-2, respectively. Recombinant PtSABP2-1 and PtSABP2-2 exhibited apparent Km values of 68.2 ± 3.8 μM and 24.6 ± 1 μM with MeSA, respectively. Structural modeling using the three-dimensional structure of tobacco SABP2 as a template indicated that the active sites of PtSABP2-1 and PtSABP2-2 were highly similar to that of tobacco SABP2. Under normal growing conditions, *PtSABP2-1* showed the highest level of expression in leaves and *PtSABP2-2* was most highly expressed in roots. In leaf tissues of poplar plants under stress conditions, the expression of *PtSABP2-1* was significantly down-regulated by two stress factors, whereas the expression of *PtSABP2-2* was significantly up-regulated by four stress factors. The plausible mechanisms leading to these two highly homologous MeSA esterase genes involved in divergent biological processes in poplar are discussed.

1. Introduction

Plants respond to pathogen infection by activating a variety of defense mechanisms (Dangl and Jones, 2001). At the infection site and surrounding area, cells usually undergo programmed cell death, which leads to the formation of necrotic lesions (Mittler et al., 1996). In addition to local responses, plants frequently develop resistance in other parts of the plant. This resistance, called systemic acquired resistance (SAR), is broad-spectrum and long-lasting (Ryals et al., 1996). Salicylic acid (SA) (**1**) (Fig. 1) is a critical signal for activating both local and systemic defenses (Sticher et al., 1997). After pathogen invasion, the endogenous levels of SA (**1**) in the infection site and systemic tissues increase, which induces the expression of defense-associated genes including those encoding pathogenesis-related (PR) proteins (Lamb and Dixon, 1997). In recent years, molecular genetic and biochemical studies led to identification of a number of key components of the SA (**1**) signaling pathway (Durrant and Dong, 2004). The SA-binding protein 2 (SABP2) is one of them.

SABP2 was originally identified from tobacco using a biochemical approach. It is present in low abundance in tobacco cells and can specifically bind SA (**1**) with high affinity (Kumar and Klessig, 2003). The gene encoding SABP2 was cloned and its deduced protein sequence showed high similarity to α/β fold hydrolases (Kumar and Klessig, 2003). A genetic study demonstrated that *SABP2* is a vital component of the SA (**1**) signaling pathway. When *SABP2* expression was silenced in tobacco, several SA-dependent responses, including development of local resistance to tobacco mosaic virus (TMV) and induction of PR gene expression, were impeded (Kumar and Klessig, 2003; Kumar et al., 2006). The three-dimensional structure of SABP2 was recently determined, which showed typical structural features of an α/β fold hydrolase superfamily (Forouhar et al., 2005). Further biochemical study demonstrated that SABP2 has strong esterase activity using methyl salicylate (MeSA) (**2**) as substrate (Forouhar et al., 2005). MeSA (**2**) is an ubiquitous compound in plants. It is produced from SA (**1**) by the action of SA methyltransferase (SAMT) (Ross et al., 1999). The hydrophobic nature of MeSA (**2**) makes itself a diffusible intercellular signal transducer. A recent study demonstrated that MeSA (**2**) is a mobile signal for SAR in tobacco (Park et al., 2007). Produced from SA (**1**) at the local infection site by the action of SAMT, MeSA (**2**) is readily translocated through the phloem. In systemic tissues, MeSA (**2**) is perceived by SABP2 and converted back to SA (**1**), which is the active signal (Park et al., 2007).

The role of the SA-dependent signaling pathway in plant defenses has been relatively well studied in a variety of annual species, such as tobacco (Malamy et al., 1990), Arabidopsis (Cao et al.,

* Corresponding author. Tel.: +1 865 974 8521; fax: +1 865 974 1947.
E-mail address: fengc@utk.edu (F. Chen).

0031-9422/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.phytochem.2008.11.014

1994; Clarke et al., 2000) and rice (Chern et al., 2005; Iwai et al., 2007). In contrast, little is yet known about the role of the SA (**1**) signaling pathway in plant defenses in perennial woody species. Such species are long lived and therefore, more prone to attacks by pathogens during their life cycle (Rinaldi et al., 2007). SA (**1**) treatment can induce expression of defense genes in trees such as pine (Davis et al., 2002), suggesting that the SA-dependent signaling pathway is also important for defense responses in perennial woody species.

We have undertaken a project to study plant defense mechanisms against biotic stresses in perennial woody species using *Populus* as a model. There are about 40 species within the genus *Populus*. Members of the *Populus* are among the fast-growing tree species in the world. That is one important reason why *Populus* has been considered a candidate as bioenergy crop. Because of its small genome, relative ease of genetic manipulation, and rapid growth, *Populus* has also been established as a model for forest tree genetics and genomics (Jansson and Douglas, 2007). Black cottonwood (*Populus trichocarpa*, Torr. & Gray) is the first tree species whose genome has been fully sequenced (Tuskan et al., 2006), which provides unprecedented opportunities for the study of tree biology and physiology. Despite the many advantages of using poplar as a tree model as well as a bioenergy crop, poplar trees are susceptible to many pathogens (Ostry and McNabb, 1985). Elucidating the natural defense mechanisms of poplar is therefore important for both basic and applied sciences. The current study concerns the SA (**1**) signaling pathway in poplar.

Previous studies showed that some common characteristics of the SA (**1**) signaling pathway are shared by poplar and some herbaceous plants. For example, SA (**1**) treatment can induce expression of defense genes in both poplar (Koch et al., 2000) and tobacco (Uknes et al., 1993). However, poplar also has distinct features. For example, the basal levels of SA (**1**) in poplar are much higher than those in tobacco and Arabidopsis (Koch et al., 2000), posing a question on the uniqueness of the molecular mechanism of the SA (**1**) signaling pathway in poplar. In this paper, we were particularly interested in determining whether a key component of the SA (**1**) signaling pathway, SABP2, is conserved in poplar. Gene cloning, biochemical function characterization and expression analysis of two SABP2 genes from black cottonwood are reported.

2. Results

2.1. Identification and sequence analysis of putative poplar SABP2 genes

When the protein sequence of tobacco SABP2 (Kumar and Klessig, 2003) was used to blast search the sequenced poplar genome (Tuskan et al., 2006) using e-10 as a cut-off E-value, 30 proteins significantly homologous to SABP2 were identified. Two proteins, estExt_fgenesh4_pm.C_LG_VII0354 and eugene3.00070971, are most similar to tobacco SABP2 with an E-value lower than e-100. The two proteins displayed specific MeSA esterase activity (see Section 2). Thus, estExt_fgenesh4_pm.C_LG_VII0354 and eugene3.00070971 were named PtSABP2-1 and PtSABP2-2, respectively.

Both PtSABP2-1 and PtSABP2-2 are localized on chromosome VII. They are about 34 kbs from each other. Both genes contain two introns and three exons. Sequence comparison showed that the first introns of the two genes are 100% identical and the second introns 97% identical (data not shown). The promoter regions of the two genes were also compared. Significant sequence similarity was detected in the regions from start codon to about 300 bps upstream of the start codon. In contrast, the promoter regions between 300 and 800 bps upstream of the start codon of the two genes share no significant sequence similarity (see Supplementary data).

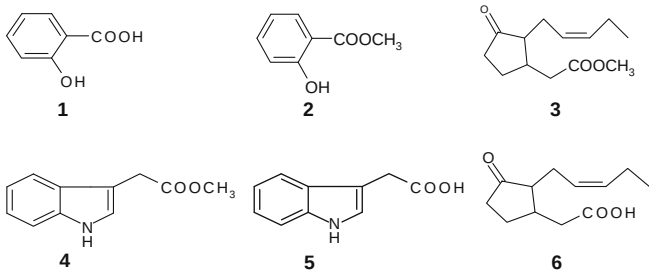


Fig. 1. Structures of compounds **1**–**6**.

Full-length cDNAs of PtSABP2-1 and PtSABP2-2 were cloned via RT–PCR from poplar leaf and root cDNAs, respectively. Both PtSABP2-1 and PtSABP2-2 encode a protein of 263 amino acids. Whereas PtSABP2-1 and PtSABP2-2 are 98% identical to each other, their sequences are 77% identical to that of tobacco SABP2 (Fig. 2). The putative pIs of PtSABP2-1, PtSABP2-2 and NtSABP2 are 5.01, 4.78 and 5.27, respectively.

2.2. PtSABP2-1 and PtSABP2-2 display MeSA esterase activity

Recombinant proteins of PtSABP2-1 and PtSABP2-2 were expressed in *Escherichia coli* from their corresponding full-length cDNAs. In the vector pET100/D-TOPO, the full-length cDNAs of PtSABP2-1 and PtSABP2-2 were fused to an N-terminal sequence containing codons for six histidine residues. His-tagged PtSABP2-1 and PtSABP2-2 were expressed in *E. coli* then purified using Ni-NTA agarose to near electrophoretic homogeneity (see Supplementary data).

Purified PtSABP2-1 and PtSABP2-2 were tested for esterase activity using MeSA as substrate. While the amounts of MeSA (**2**) decreased, the amounts of SA (**1**) increased in the same reaction, indicating that both PtSABP2-1 and PtSABP2-2 (not shown) catalyzed the hydrolysis of the methyl group of MeSA (**2**) to form SA (**1**). Both of PtSABP2-1 and PtSABP2-2 had the optimum pH at 7.0 (see Supplementary data). Under steady-state conditions, PtSABP2-1 exhibited an apparent Km value 68.2 ± 3.8 μM with MeSA and a V_{max} value of 4.2 ± 0.04 pmol S^{-1} (Fig. 3A). PtSABP2-2 had a Km value 24.6 ± 1 μM with MeSA (**2**) and a V_{max} value of 0.35 ± 0.03 pmol S^{-1} (Fig. 3B).

2.3. Substrate specificity analysis of PtSABP2-1 and PtSABP2-2

In *in vitro* enzyme assays, tobacco SABP2 displayed esterase activity with multiple substrates, including MeSA (**2**), MeJA (**3**) and MeIAA (**4**), with different specific activities (Forouhar et al., 2005). For comparison, PtSABP2-1 and PtSABP2-2 were also analyzed for esterase activity with MeJA (**3**) and MeIAA (**4**). At 10 μM and 100 μM concentrations of substrates, PtSABP2-1 and PtSABP2-2 showed no or very low activity with MeJA (**3**) and MeIAA (**4**) (Fig. 4). At 1 mM concentration of substrates, PtSABP2-1 and PtSABP2-2 showed significant activity with MeJA (**3**). However, the esterase activity of PtSABP2-1 and PtSABP2-2 with MeJA (**3**) is only 9% and 14% of their corresponding activities with MeSA (**2**), respectively (Fig. 4). These results indicate that the two poplar SABP2 proteins are highly specific for MeSA (**2**) among the three substrates tested at both physiologically relevant (10 μM) and irrelevant (1 mM) concentrations.

2.4. Structural features of PtSABP2-1 and PtSABP2-2

The crystal structure of tobacco SABP2 has been experimentally determined (Forouhar et al., 2005). Since tobacco SABP2 and the

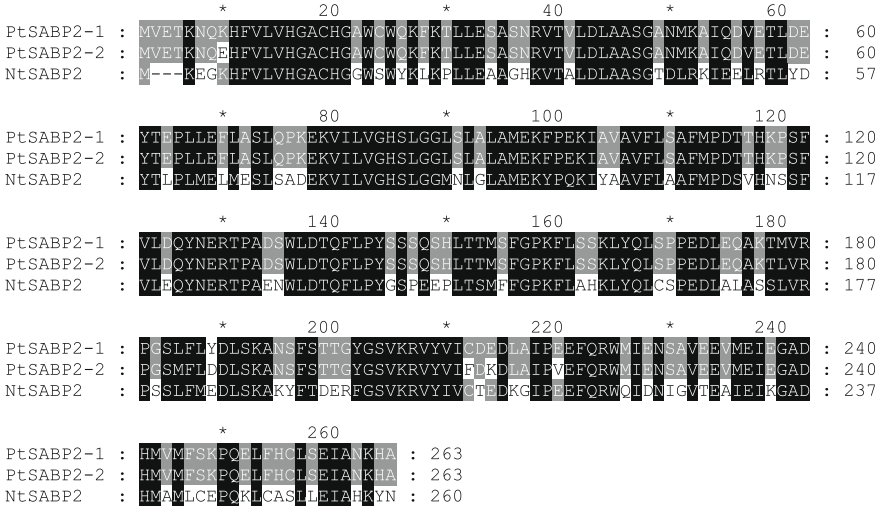


Fig. 2. Multiple sequence alignment of deduced amino acid sequences of PtSABP2-1 and PtSABP2-2 and the protein sequence of tobacco SABP2 (NtSABP2).

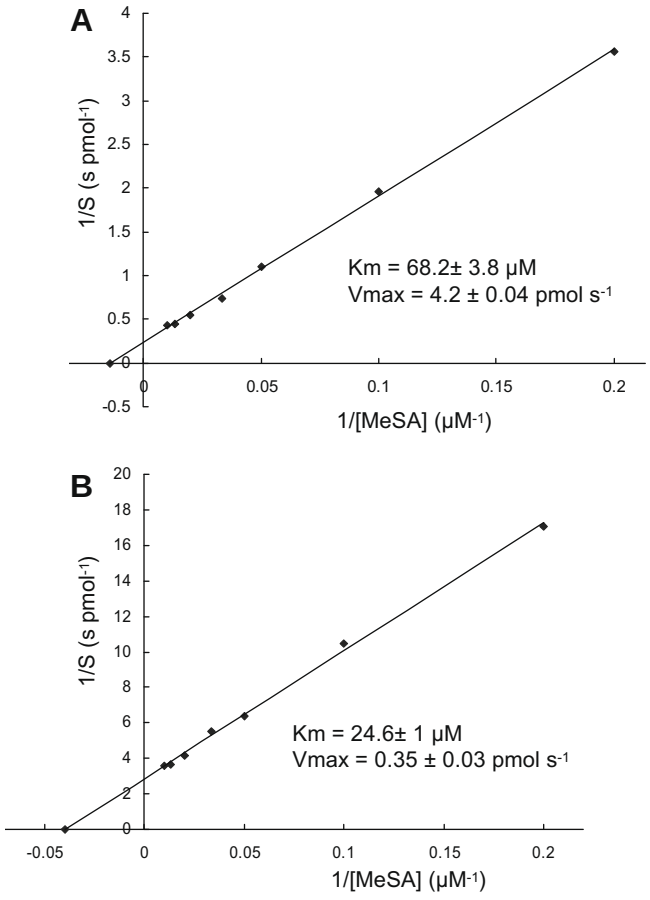


Fig. 3. Biochemical properties of PtSABP2-1 and PtSABP2-2. Steady-state kinetics of PtSABP2-1 and PtSABP2-2 was measured using MeSA (**2**) as substrate. Examples of the Lineweaver–Burk plot for PtSABP2-1 (A) and PtSABP2-2 (B) were shown.

two poplar SABP2s possess 77% sequence identity, reliable three-dimensional models of PtSABP2-1 and PtSABP2-2, including their active sites, were constructed using tobacco SABP2 structure as a template (Fig. 5). The overall structures of PtSABP2-1 and PtSABP2-2 are very similar to that of tobacco SABP2, consistent with their high sequence conservation. In the active site region, PtSABP2-1 has only one substitution with respect to tobacco SABP2: G212 of tobacco SABP2 is replaced by A215 in PtSABP2-1 (Fig. 5A). The extra methyl group in PtSABP2-1 resulted from the G to A change is not expected to affect the binding of SA, in agreement with the kinetic studies. PtSABP2-2 has two substitutions: G212 and L181 of tobacco SABP2 are respectively replaced by A215 and M184 in PtSABP2-2 (Fig. 5B). The introduction of a bulkier Met side chain in PtSABP2-2 probably has an impact on the observed lower Km value of this enzyme (Fig. 3).

2.5. Expression analysis of PtSABP2-1 and PtSABP2-2

RT–PCR analysis was performed to determine the expression patterns of *PtSABP2-1* and *PtSABP2-2*. Gene specific primers were designed to discriminate the transcripts of the two genes (Fig. 6A). Total RNA was isolated from young leaves, old leaves, stems, and roots of one year-old poplar trees and used for semi-quantitative RT–PCR analysis. Gene expression analysis (Fig. 6B) showed that *PtSABP2-1* had the highest level of expression in leaves and a low level of expression in stems and roots. In contrast, *PtSABP2-2* had the highest level of expression in roots, a moderate level of expression in stems and a low level of expression in leaves.

To understand the potential roles of *PtSABP2-1* and *PtSABP2-2* in plant defense, expression of the two genes under stress conditions was also analyzed. Poplar plants were either physically injured or treated with SA (**1**), MeJA (**3**), or a fungal elicitor alamethicin. Leaves of the treated poplar plants and appropriate control plants were collected for extraction of RNAs, which were used RT-PCR analysis. The expression of *PtSABP2-1* (**1**) was not significantly affected by the stress factors tested, except for 2 h wounding and SA (**1**) treatment for 24 h, which significantly suppressed the expression of *PtSABP2-1* (Fig. 6C). In contrast, the expression of *PtSABP2-2* was significantly up-regulated by a number of treatments, including 24 h wounding, 2 h SA (**1**) treatment, 4 h MeJA (**3**) treatment and 2 h alamethicin treatment. Treatments by wounding, SA (1), and MeJA (**3**) at other time points did not significantly affect the expression of *PtSABP2-2* (Fig. 6C).

3. Discussion

SAR has been shown to be activated by a vascular-mobile signal that moves throughout the plant from the infected leaves (Durrant and Dong, 2004). The nature of the mobile signal for SAR, however,

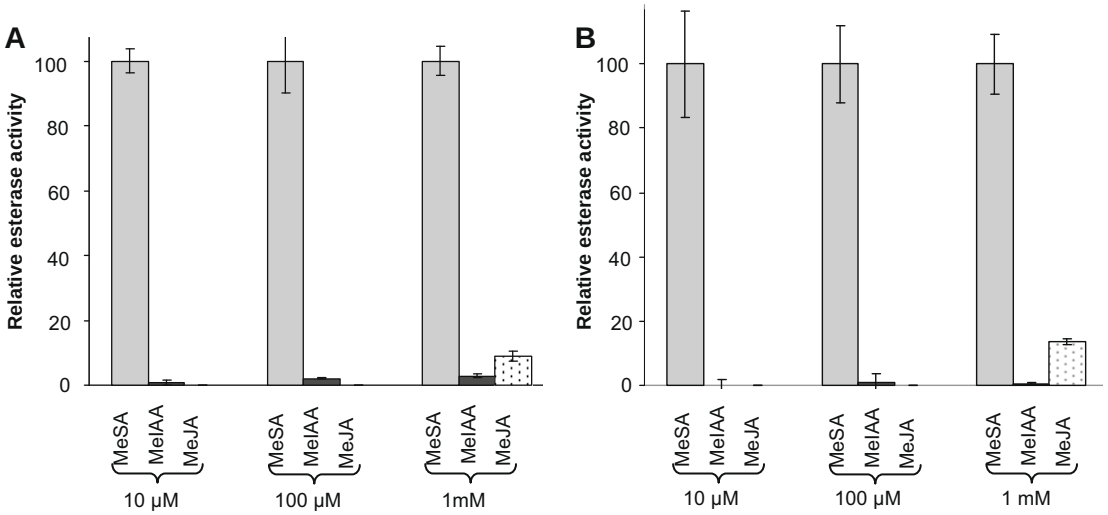


Fig. 4. Substrate specificity of PtSABP2-1 and PtSABP2-2. Relative methyl esterase activity of PtSABP2-1 (A) and PtSABP2-2 (B) was measured with MeSA (**2**), MeIAA (**4**), and MeJA (**3**) at three different concentrations (10 μM, 100 μM, and 1 mM). The activity with MeSA (**2**) at each of the substrate concentrations was set at 100%. The activity ratios between MeSA (**2**) and the other substrates are indicated.

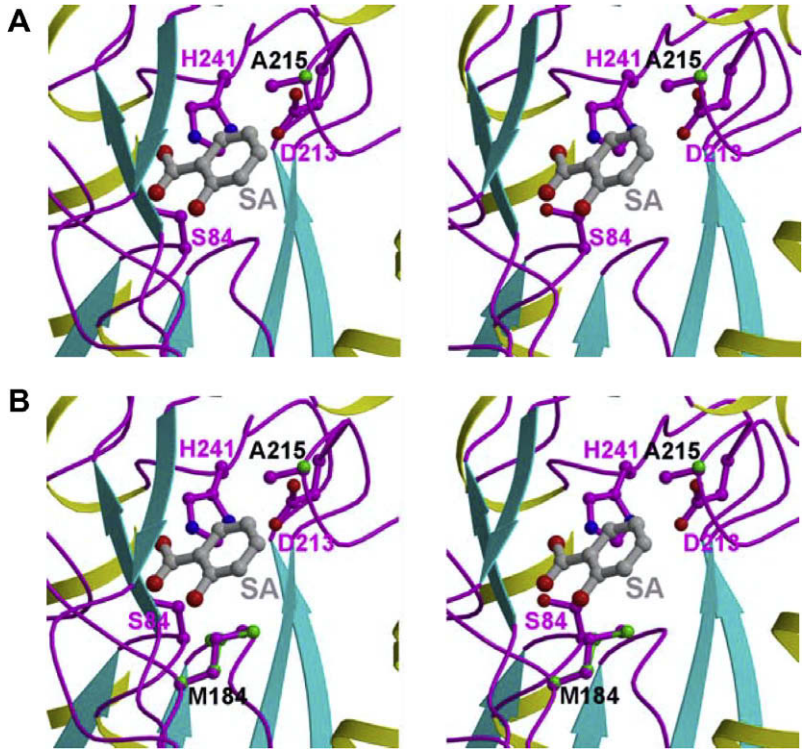


Fig. 5. Active sites of PtSABP2-1 and PtSABP2-2. (A) Stereo-view representation of the active site of PtSABP2-1; the SA (light grey) and the side chains of the catalytic triads, S84, D213, and H241 (all in magenta) are shown as ball-and-stick models. A215 (magenta) is labeled in black so as to distinguish it from the corresponding residue G212 of tobacco SABP2, which is depicted in green ball; and (B) stereo-view representation of the active site of PtSABP2-2. The two substitutions, A215 and M184, in the active site of PtSABP2-2, corresponding to G213 and L181 (both shown in green) of tobacco SABP2, are shown in ball-and-stick models and labeled in black.

has been a subject of controversy. SA (**1**) has been demonstrated to be essential for the activation of SAR (Gaffney et al., 1993). SA (**1**) itself, however, has been disassociated as the mobile signal (Vernooij et al., 1994). MeSA, the methyl ester of SA (**1**), was recently demonstrated to be a mobile signal for SAR in tobacco (Park et al., 2007). With MeSA (**2**) being the inactive mobile signal, it is critical that MeSA (**2**) is converted back to SA (**1**) at the site of action. This reaction is catalyzed by MeSA esterase. Prior to this study, the only MeSA esterase that has been functionally characterized is SABP2 from tobacco (Forouhar et al., 2005). Here, we showed that poplar, a woody perennial species, contains two functional *SABP2* genes. PtSABP2-1 and PtSABP2-2 were identified from the poplar genome based on their high sequence similarity to tobacco SABP2 (Fig. 2). In *in vitro* biochemical studies, like tobacco SABP2, both PtSABP2-1 and PtSABP2-2 displayed highly specific esterase activity towards MeSA (**2**) (Figs. 3 and 4), supporting that MeSA (**2**) is the in vivo substrate. Whether PtSABP2-1 and PtSABP2-2 accept substrates other than MeSA (**2**) in planta, however, remains to be determined. The Km value of tobacco SABP2 was reported to be 8.6 μM (Forouhar et al., 2005). The Km values

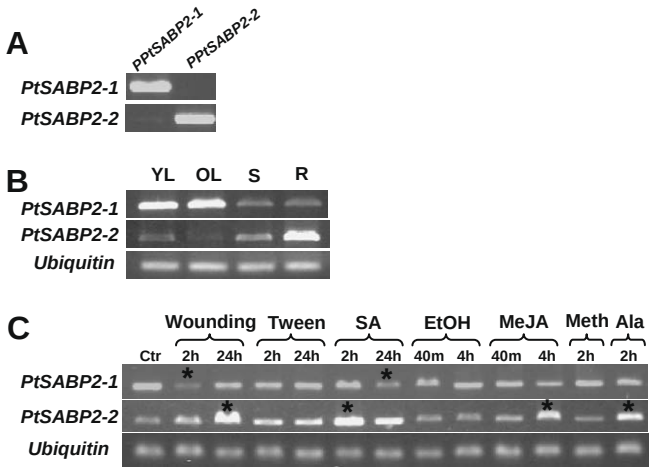


Fig. 6. Expression analysis of *PtSABP2-1* and *PtSABP2-2*. (A) Plasmids *PPtSABP2-1* and *PPtSABP2-2* containing full-length cDNAs of *PtSABP2-1* and PtSABP2-2, respectively, were used as templates in PCRs with gene-specific primers for PtSABP2-1 and PtSABP2-2; (B) tissue-specific expression of PtSABP2-1 and PtSABP2-2. Young leaves (YL), old leaves (OL), stems (S), and roots (R) were colleted from one year-old poplar trees grown in the greenhouse; and (C) expression of PtSABP2-1 and PtSABP2-2 in leaves of poplar plants that were either physically injured (wounding, 2 h and 24 h), or treated with one of the following chemicals: salicylic acid (SA) (**1**), methyl jasmonate (MeJA) (**3**) and alamethicin (Ala). Treatments with 0.1% tween (Tween), ethanol (EtOH) and methanol (Meth) were used as controls for SA (**1**), MeJA (**3**) and Ala treatments, respectively. Bands marked with a asterisk (*) indicate that the gene showed significant difference in expression levels in treated vs. control plants ($P < 0.05$). In both (B) and (C), PCR with primers for *Ubiquitin* was used to judge equality of the concentration of cDNA templates in different samples.

of PtSABP2-1 and PtSABP2-2 are about 10 times and three times higher than that of tobacco SABP2. It will be interesting to determine whether the endogenous levels of MeSA (**2**) in poplar tissues under stresses are higher than that in tobacco tissues under stresses. The overall three-dimensional structures and active sites of the three proteins are highly conserved (Fig. 5). Despite their similar biochemical properties, the two poplar *SABP2* genes display divergent expression patterns. They showed different tissue-specificity under normal growing conditions and different responses to stress factors (Fig. 6), suggesting that these two genes have divergent biological roles.

Studying the potential involvement of SABP2s in poplar-pathogen interactions will be important for elucidation of the biological roles of poplar SABP2s, especially in the context of a SA metabolism that is somehow different from that in tobacco. Whereas the SA (**1**) levels in healthy leaves of tobacco are low, in healthy tissues of poplar they are high (Koch et al., 2000; Morse et al., 2007). In poplar, overexpressing a bacterial SA hydroxylase gene *nahG* in poplar did not cause a significant reduction of the SA (**1**) levels (Morse et al., 2007), which is different from the study with tobacco (Delaney et al., 1994). How PtSABP2-1 and PtSABP2-2 function in poplar with respect to SA metabolism as compared with that in tobacco remains to be determined. Poplar mosaic virus (PopMV) is the best characterized tree virus (Cooper, 1993). The infection dynamics of PopMV has been investigated using different genotypes of *Populus* (Smith and Campbell, 2004). The poplar-PopMV system, therefore, presents a good model for determining whether PtSABP2s function in SAR in this perennial woody species. Such study will provide further evidence on whether SABP2-dependent SAR mediated by SA (**1**) is a conserved defense pathway.

In addition to its involvement in SAR, MeSA (**2**), the substrate of SABP2, has been implicated in a number of other biological/ecological processes. MeSA (**2**) is often emitted as a volatile compound from the plants that are being challenged by stress factors, including insect feeding (Chen et al., 2003a), virus infection (Shulaev et al., 1997), and elicitor treatment (Chen et al., 2003a). MeSA (**2**) produced under such stress conditions has been suggested to have a number of biological roles. For example, insect-induced volatile MeSA (**2**) has been suggested to be involved in attracting natural enemies of the feeding insects (Dicke et al., 1990). Virus-induced air-borne MeSA (**2**) has been suggested to act as a signal to activate defense responses in nearby healthy plants (Shulaev et al., 1997). In all these cases, if SA (**1**) is the active signal, then inactive MeSA (**2**) would need to be converted back to SA (**1**). If true, it would suggest that SABP2 has multiple biological functions in addition to its role in SAR, which is supported by induction of expression of poplar *SABP2* genes by multiple stress factors (Fig. 6).

The presence of two MeSA esterase genes with a same biochemical function but seemingly divergent biological roles is intriguing. It is tempting to speculate that this reflects the specific nature of natural defenses of perennial woody species. Due to their long generation time, perennial woody species can not match the evolutionary rates of microbial pathogens that go through several generations every year (Rinaldi et al., 2007). Therefore, positive adaptation of defense mechanism may be necessary for perennial woody species to survive. Overrepresentation of defense-related genes, such as NBS-LRR in the first sequenced tree genome, supports this notion (Tuskan et al., 2006). The presence of two *SABP2* genes with divergent expression patterns may provide poplar trees an advantage in defense against pathogens.

4. Concluding remarks

The mechanism that leads to functional divergence of PtSABP2-1 and PtSABP2-2 is also intriguing. The two SABP2s are highly homologous to each other, suggesting they are the consequence of gene duplication. Gene duplication may result from whole genome duplication, segmental duplication of chromosomes, or local duplication caused by unequal crossover (Zhang, 2003; Yang et al., 2006). *PtSABP2-1* and *PtSABP2-2* are localized on chromosome VII and are 30 kbs from each other. They are therefore likely the consequence of local gene duplication. The high sequence similarity between the two genes, even in introns, implies that the duplication event occurred relatively recently. The promoter regions of the two genes, however, show relatively high degree of divergence, suggesting that promoter divergence is probably a major determinant of functional divergence of *PtSABP2-1* and *PtSABP2-2*.

5. Experimental

5.1. Plant material and chemicals

The female black cottonwood clone 'Nisqually-1', which was used for whole genome sequencing (Tuskan et al., 2006), was used for gene cloning and expression analysis in this study. The tissues used for gene expression analysis at normal growing conditions, including young leaves, old leaves, stems, and roots, were collected from one year-old poplar trees grown in the greenhouse. Leaf tissues used for gene expression analysis under stress conditions were collected from poplar plants at the eight-leaf stage that were vegetatively propagated grown on MS medium. Poplar plants were either physically injured or treated with several chemicals, including SA (**1**), methyl jasmonate (MeJA) (**3**) or a fungal elicitor alamethicin. For physical injury, leaves were cut with a sterile razor blade to produce three lateral incisions on each side of the midvein and the wounded leaves were collected at 2 h and 24 h after wounding. For SA (**1**) treatment, poplar plants were sprayed with either 5 mM SA (**1**) solution in 0.1% Tween (pH 7.0) or 0.1% Tween only as control. Leaves were collected at 2 h and 24 h after initiation of the

treatment. For MeJA (**3**) treatment, poplar plants were placed in a 1 L glass jar containing a cotton tip applied with 1 μl MeJA (**3**) dissolved in EtOH (200 μl) ethanol. Poplar plants placed in another 1 L glass jar containing only EtOH (200 μl) were used as control. The glass jar was quickly sealed. Leaves were collected after 40 min and 4 h after initiation of the treatment. For alamethicin treatment, leaves were cut-off from the base of the petiole and the detached leaves then placed apright in a small glass beaker containing 10 ml of 5 μg/ml alamethicin (dissolved 1000-fold in water from a 5 mg/ml stock solution in 100% MeOH). As a control, detached leaves were submerged in 0.1% MeOH in H_2O. Only the petiole of each leaf was submerged in the solution. The glass beaker was then sealed with Saran wrap and placed in a growth chamber. Leaves were collected 2 h after the treatment. Two replicates were performed for each treatment. All chemicals were purchased from Sigma–Aldrich (St. Louis, MO, USA).

5.2. Database search and sequence analysis

To identify putative poplar *SABP2*-like esterase genes, the protein sequence of tobacco *SABP2* (accession: AY485932) was used as a query to search the genome sequence database of poplar (http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.home.html) using the BlastP algorithm (Altschul et al., 1990). Two poplar genes encoding proteins with the highest level of sequence similarities (77%) to tobacco SABP2 were chosen for further analysis.

Nucleic acid and protein sequence alignments were made using the ClustalX program (Thompson et al., 1997), and displayed using GeneDoc (http://www.psc.edu/biomed/genedoc/).

5.3. Cloning full-length cDNA of PtSABP2-1 and PtSABP2-2

Total RNA was extracted from the leaf and root tissues using the RNeasy Plant Mini Kit (Qiagen, Valencia, CA, USA) and DNA contamination was removed with an on-column DNase (Qiagen, Valencia, CA, USA) treatment. Total RNA (1.5 μg) was reverse-transcribed into first-strand cDNA in a 15 μL reaction volume using the first-strand cDNA synthesis Kit (Amersham Biosciences, Piscataway, NJ, USA) as previously described (Chen et al., 2003b). *PtSABP2-1* full-length cDNA was amplified with the leaf cDNA using the forward primer 5′-CACCATGGTAGAGACCAAGAATCAGA-3′ and the reverse primer 5′-CAAACCACTATAAGTAAGGGTGCTGA-3′. *PtSABP2-2* full-length cDNA was amplified with the root cDNA using the forward primer 5′-CACCATGGTAGAGACCAAGAATCAGG-3′ and the reverse primer 5′-TATAACTAAGGGTGCTGATAAGGACA-3′. PCR was set as follows: 94 °C for 2 min followed by 30 cycles at 94 °C for 30 s, 60 °C for 45 s and 72 °C for 1 min 30 s, and a final extension at 72 °C for 10 min. The PCR product was separated on 1.0% agrose gel. The target band was sliced from the gel and purified using QIAquick Gel Extraction Kit (Qiagen, Valencia, CA, USA). The PCR product was cloned into pET100/D-TOPO vector using the protocol recommended by the vendor (Invitrogen, Carlsband, CA, USA). The cloned cDNAs in pET100/D-TOPO vector were fully sequenced.

5.4. Purification from E. coli-expressed recombinant proteins

To express PtSABP2-1 and PtSABP2-2, the corresponding protein expression constructs were transformed into *E. coli* strain BL21 (DE3) CodonPlus (Stratagene, La Jolla, CA, USA). Protein expression was induced by IPTG for 18 h at 22 °C and cells were lysed by sonication. His-tagged PtSABP2-1 and PtSABP2-2 proteins were purified from *E. coli* cell lysate using Ni-NTA agarose following the manufacturer's instruction (Invitrogen, Carlsband, CA, USA). The purity of the protein was verified by SDS–PAGE and the concentration of the protein was determined by the Bradford assay (Bradford, 1976).

5.5. MeSA esterase activity assay

In a single assay, the reaction was performed in 90 μl volume containing 50 mM Tris–HCl, pH 7.5, 4 μg purified enzyme, and 600 μM MeSA (**2**) at 25 °C. The individual assays were terminated by addition of 10 μl 2 N HCl at one of the following time points: 0 h, 1 h, 2 h, and 4 h after initiation of the reaction. Then individual assays were divided into two halves. The first half (50 μl) was extracted with EtOH (100 μl) and the organic phase analyzed for the amount of MeSA (**2**) using GC–MS as previously described (Zhao et al., 2008). The second half was used for identification and quantification of SA (**1**). Aliquots (50 μl) of crude assays were quantitatively transferred to scintillation vials and dried down in a He stream. The dried samples were dissolved in 500 μL of silylation–grade CH_3CN followed by addition of 500 μL *N*-methyl-*N*-trimethylsilyltrifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) (Pierce Chemical Co., Rockford, IL, USA), and then heated for 1 h at 70 °C to generate trimethylsilyl (TMS) derivatives. After 2 days, 1 μL aliquots were injected into a ThermoFisher DSQII GC–MS, fitted with an Rtx-5MS (crosslinked 5% PH ME Siloxane) 30 m × 0.25 mm × 0.25 μm film thickness capillary column (Restek, Bellefonte, PA, USA). The standard quadrupole GC–MS was operated in electron impact (70 eV) ionization mode, with six full-spectrum (70–650 Da) scans per second. Gas (helium) flow was set at 1.1 mL/min with the injection port configured in the splitless mode. The injection port and detector temperatures were set to 220 °C and 300 °C, respectively. The initial oven temperature was held at 50 °C for 2 min and was programmed to increase at 20 °C/min to 325 °C and held for another 11.25 min, before cycling back to the initial conditions. The SA (**1**) peak was quantified by extracting 267 m/z to minimize integration of co-eluting metabolites. Peaks were quantified by area integration and the concentrations were derived from an external calibration curve of amount of SA (**1**) injected versus peak area integration of the extracted m/z. Final values are an average of three independent measurements.

5.6. Determination of relative specific activity of PtSABP2-1 and PtSABP2-2 with three substrates

A two-step radiochemical esterase assay was performed to determine the substrate specificity of PtSABP2-1 and PtSABP2-2 following a protocol previously reported (Forouhar et al., 2005). The reaction of the first step was performed with a 40 μl volume containing 50 mM Tris–HCl, pH 7.5, one of the three substrates (MeSA, MeJA and methyl indole-3-acetate (MeIAA)) at 10 μM, 100 μM or 1 mM concentrations and 0.5 μg purified PtSABP2-1 or PtSABP2-2 for 30 min at 25 °C. After that, the samples were boiled for 5 min to stop the reaction and denature the enzyme. The second step reaction started with the addition of 3 μM ^{14}C-adenosyl-L-methionine (SAM) with a specific activity of 51.4 mCi/mmol (Perkin Elmer, Boston, MA, USA), 120 μM SAM and purified methyltransferase (SAMT from Arabidopsis (Chen et al., 2003a), indole-3-acetic acid (**5**) methyltransferase from poplar (Zhao et al., 2007), or jasmonic acid (**6**) methyltransferase from Arabidopsis (Seo et al., 2001)). The reaction was allowed to proceed for 30 min at 25 °C. Radiolabeled products were extracted and radioactivity counted using a liquid scintillation counter (Beckman Coulter, Fullerton, CA, USA) as previously described (D'Auria et al., 2002). Three independent assays were performed for each substrate.

5.7. pH optimum for PtSABP2-1 and PtSABP2-2 activities

Both PtSABP2-1 and PtSABP2-2 activities were determined in 50 mM Bis–Tris propane buffer for the pH range across 6.0–10.0 using the radiochemical assay described above. Data presented are the average of three independent assays.

5.8. Determination of kinetic parameters of PtSABP2-1 and PtSABP2-2

The increase in reaction rate with increasing concentrations of MeSA (**2**) was evaluated with the radiochemical assay described above and was found to obey Michaelis–Mention kinetics. Appropriate enzyme concentrations and incubation times were determined in time-course assays so that the reaction velocity was linear during the assay period. To determine the Km for MeSA (**2**), the concentrations of MeSA (**2**) were independently varied in the range from 5 μM to 100 μM. Lineweaver–Burk plots were made to obtain apparent Km values and maximum velocity values, as previously described (Chen et al., 2003a). Final values are an average of three independent measurements.

5.9. Molecular modeling

The tobacco SABP2 and the two poplar MeSA esterases, PtSABP2-1 and PtSABP2-2, share approximately 77% identity with no gap in their sequence alignment (Fig. 2). This allowed us to use the crystal structure of tobacco SABP2 in a complex with SA (**1**) at 2.1 Å resolution (PDB accession code 1Y7I, Forouhar et al., 2005) as a reliable template to generate structural models for both PtSABP2-1 and PtSABP2-2. Using the XtalView program (McRee, 1999), the models were built manually and were subject to two cycles of refinement by CNS, primarily for energy minimization of both the protein and the bound SA (**1**) ligand (Brünger et al., 1998).

5.10. Determination of gene expression using semi-quantitative RT–PCR

Total RNA extraction from young leaves, old leaves, stems, and roots of one year-old poplar trees and leaves of poplar plants treated with various stress factors and subsequent first-strand cDNA synthesis were performed essentially the same as described for full-length cDNA cloning. Forward primer 5′-ACAATGGTAGAGACCAAGAATCAGA-3′ and reverse primer 5′-TGGTATCGCTAAATCTTCATCGC-3′ were used as gene-specific primers for PtSABP2-1. Forward primer 5′-ACAATGGTAGAGACCAAGAATCAGG-3′ and reverse primer 5′-ACTGGTATCGCTAAATCTTTGTCAA-3′ were used as gene-specific primers for PtSABP2-2. The two primers used for the PCR amplification of *Ubiquitin* were designed as previously described (Kohler et al., 2004): forward primer 5′-CAGGGAAACAGTGAGGAAGG-3′ and reverse primer 5′-TGGACTCACGAGGACAG-3′. Initially, PCR was performed with PtSABP2-specific primers using 0.1 μL, 0.2 μL, 0.5 μL and 1.0 μL cDNA from each sample as template. The PCR program was set as follows: 94 °C for 2 min followed by 30 cycles (PtSABP2-1) or 32 cycles (PtSABP2-2) at 94 °C for 30 s, 60 °C for 30 s and 72 °C for 1 min 30 s, and a final extension at 72 °C for 10 min. Amplified products were separated on a 1.0% agarose gel and stained with ethidium bromide. Gels were visualized under UV-light and quantified using the Bio-Rad Quantity One software (Bio-Rad, Hercules, CA, USA). Analysis showed that the amounts of amplified products with gene-specific primers increased linearly with increasing amounts of template cDNA. Therefore, 0.2 μL cDNA was chosen as the optimal amount of template for final data collection. For the *Ubiquitin* gene, PCR was performed in separate tubes under conditions similar to those for PtSABP2 genes except that the reactions were performed for 25 cycles. The levels of the *Ubiquitin* gene were used for normalization. All PCRs were replicated three times using the first-strand cDNA made from two independent RNA preparations. The *t*-test using the SPSS software was performed to identify significant differences in expression levels in control and treated samples.

Acknowledgements

We are grateful to Byung-Guk Kang for providing poplar plants for gene expression analysis. This research was supported in part by the DOE Office Biological and Environmental Research – Genome to Life Program through the BioEnergy Science Center (BESC), and by the Tennessee Agricultural Experiment Station.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.phytochem.2008.11.014.

References

Altschul, S.F., Stephen, F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.
Bradford, M.M., 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein–dye binding. Anal. Biochem. 72, 248–254.
Brünger, A.T., Adams, P.D., Clore, G.M., Delano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Warren, G.L., 1998. Crystallography and NMR system: a software suite for macromolecular structure determination. Acta Crystallogr. D54, 905–921.
Cao, H., Bowling, S.A., Gordon, A.S., Dong, X., 1994. Characterization of an Arabidopsis mutant that is nonresponsive to inducers of systemic acquired resistance. Plant Cell 6, 1583–1592.
Chen, F., D'Auria, J.C., Tholl, D., Ross, J.R., Gershenzon, J., Noel, J.P., Pichersky, E., 2003a. An Arabidopsis gene for methylsalicylate biosynthesis, identified by a biochemical genomics approach, has a role in defense. Plant J. 36, 577–588.
Chen, F., Tholl, D., D'Auria, J.C., Farooq, A., Pichersky, E., Gershenzon, J., 2003b. Biosynthesis and emission of terpenoid volatiles from Arabidopsis flowers. Plant Cell 15, 481–494.
Chern, M.S., Fitzgerald, H.A., Canlas, P., Ronald, P., 2005. Over-expression of a rice NPR1 homologue leads to disease resistance, activation of defense gene expression, and a lesion mimic phenotype. Mol. Plant Microbe Interact. 18, 511–520.
Clarke, J.D., Volko, S.M., Ledford, H., Ausubel, F.M., Dong, X., 2000. Roles of salicylic acid, jasmonic acid, and ethylene in cpr-induced resistance in Arabidopsis. Plant Cell 12, 2175–2190.
Cooper, J.I., 1993. Virus Diseases of Trees and Shrubs. Chapman and Hall, London.
Dangl, J.L., Jones, J.D., 2001. Plant pathogens and integrated defence responses to infection. Nature 411, 826–833.
Davis, J.M., Wu, H.G., Cooke, J.E.K., Reed, J.M., Luce, K.S., Michler, C.H., 2002. Pathogen challenge, salicylic acid, and jasmonic acid regulate expression of chitinase gene homologs in pine. Mol. Plant Microbe Interact. 15, 380–387.
Delaney, T.P., Uknes, S., Bernooij, B., Friedrich, L., Weymann, K., Negrotto, D., Gaffney, T., Gut-Rella, M., Kessmann, H., Ward, E., 1994. A central role of salicylic acid in plant disease resistance. Science 266, 1247–1250.
Dicke, M., Vanbeek, T.A., Posthumus, M.A., Bendom, N., Vanbokhoven, H., Degroot, A.E., 1990. Isolation and identification of volatile kairomone that affects acarine predator-prey interactions-involvement of host plant in its production. J. Chem. Ecol. 16, 381–396.
Durrant, W.E., Dong, X., 2004. Systemic acquired resistance. Annu. Rev. Phytopathol. 42, 185–209.
D'Auria, J.C., Chen, F., Pichersky, E., 2002. Characterization of an acyltransferase capable of synthesizing benzylbenzoate and other volatile esters in flowers and damaged leaves of *Clarkia breweri*. Plant Physiol. 130, 466–476.
Forouhar, F., Yang, Y., Kumar, D., Chen, Y., Fridman, E., Park, S.W., Chiang, Y., Acton, T.B., Montelione, G.T., Pichersky, E., Klessig, D.F., Tong, L., 2005. Structural and biochemical studies identify tobacco SABP2 as a MeSA esterase and implicate it in plant innate immunity. Proc. Natl. Acad. Sci. USA 102, 1773–1778.
Gaffney, T., Friedrich, L., Vernooij, B., Negmtto, D., Nye, G., Uknes, S., Ward, E., Kessmann, H., Ryals, J., 1993. Requirement of salicylic acid for the induction of systemic acquired resistance. Science 261, 754–756.
Iwai, T., Seo, S., Mitsuhara, I., Ohashi, Y., 2007. Probenazole-induced accumulation of salicylic acid confers resistance to *Magnaporthe grisea* in adult rice plants. Plant Cell Physiol. 48, 915–924.
Jansson, S., Douglas, C.J., 2007. *Populus*: a model system for plant biology. Annu. Rev. Plant Biol. 58, 435–458.
Koch, J.R., Creelman, R.A., Eshita, S.M., Seskar, M., Mullet, J.E., Davis, K.R., 2000. Ozone sensitivity in hybrid poplar correlates with insensitivity to both salicylic acid and jasmonic acid. The role of programmed cell death in lesion formation. Plant Physiol. 123, 487–496.
Kohler, A., Blaudez, D., Chalot, M., Martin, F., 2004. Cloning and expression of multiple metallothioneins from hybrid poplar. New Phytol. 164, 83–93.
Kumar, D., Klessig, D.F., 2003. High-affinity salicylic acid-binding protein 2 is required for plant innate immunity and has salicylic acid-stimulated lipase activity. Proc. Natl. Acad. Sci. USA 100, 16101–16106.

Kumar, D., Gustafsson, C., Klessig, D.F., 2006. Validation of RNAi silencing specificity using synthetic genes: salicylic acid-binding protein 2 is required for innate immunity in plants. Plant J. 45, 863–868.

Lamb, C., Dixon, R.A., 1997. The oxidative burst in plant disease resistance. Annu. Rev. Plant Physiol. Plant Mol. Biol. 48, 251–275.

Malamy, J., Carr, J.P., Klessig, D.F., Raskin, I., 1990. Salicylic acid: a likely endogenous signal in the resistance response of tobacco to viral infection. Science 250, 1002–1004.

McRee, D.E., 1999. XtalView/Xfit – a versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. 125, 156–165.

Mittler, R., Shulaev, V., Seskar, M., Lam, E., 1996. Inhibition of programmed cell death in tobacco plants during a pathogen-induced hypersensitive response at low oxygen pressure. Plant Cell 8, 1991–2001.

Morse, A.M., Tschaplinski, T.J., Dervinis, C., Pijut, P.M., Schmelz, E.A., Day, W., Davis, J.M., 2007. Salicylate and catechol levels are maintained in nahG transgenic poplar. Phytochemistry 68, 2043–2052.

Ostry, M.E., McNabb, H.S., 1985. Susceptibility of *Populus* species and hybrids to disease in the north-central United States. Plant Dis. 69, 755–777.

Park, S.W., Kaimoyo, E., Kumar, D., Mosher, S., Klessig, D.F., 2007. Methyl salicylate is a critical mobile signal for plant systemic acquired resistance. Science 318, 113–116.

Rinaldi, C., Kohler, A., Frey, P., Duchaussoy, F., Ningre, N., Couloux, A., Wincker, P., Thiec, D.L., Fluch, S., Martin, F., Duplessis, S., 2007. Transcript profiling of poplar leaves upon infection with compatible and incompatible strains of the foliar rust *Melampsora larici-populina*. Plant Physiol. 144, 347–366.

Ross, J.R., Nam, K.H., D'Auria, J.C., Pichersky, E., 1999. *S*-adenosyl-L-methionine: salicylic acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant defense, represents a class of plant methyltransferases. Arch. Biochem. Biophys. 367, 9–16.

Ryals, J.A., Neuenschwander, U.H., Willits, M.G., Molina, A., Steiner, H.Y., Hunt, M.D., 1996. Systemic acquired resistance. Plant Cell 8, 1809–1819.

Seo, H.S., Song, J.T., Cheong, J.J., Lee, Y.H., Lee, Y.W., Hwang, I., Lee, J.S., Choi, Y.D., 2001. Jasmonic acid carboxyl methyltransferase: a key enzyme for jasmonate-regulated plant responses. Proc. Natl. Acad. Sci. USA 98, 4788–4793.

Shulaev, V., Silverman, P., Raskin, I., 1997. Airborne signalling by MeSA in plant pathogen resistance. Nature 385, 718–721.

Smith, C.M., Campbell, M.M., 2004. *Populus* genotypes differ in infection by, and systemic spread of, poplar mosaic virus. Plant Pathol. 53, 780–787.

Sticher, L., Mauch-Mani, B., Métraux, J.P., 1997. Systemic acquired resistance. Annu. Rev. Phytopathol. 35, 235–270.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The clustalx windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24, 4876–4882.

Tuskan, G.A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R.R., Bhalerao Rprao, R.P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., Busov, V., Campbell, M., Carlson, J., Chalot, M., Chapman, J., Chen, G.-L., Cooper, D., Coutinho, P.M., Couturier, J., Covert, S., Cronk, Q., Cunningham, R., Davis, J., Degroeve, S., Dejardin, A., DePamphilis, C., Detter, J., Dirks, B., Dubchak, I., Duplessis, S., Ehlting, J., Ellis, B., Gendler, K., Goodstein, D., Gribskov, M., Grimwood, J., Groover, A., Gunter, L., Hamberger, B., Heinze, B., Helariutta, Y., Henrissat, B., Holligan, D., Holt, R., Huang, W., Islam-Faridi, N., Jones, S., Jones-Rhoades, M., Jorgensen, R., Joshi, C., Kangasjarvi, J., Karlsson, J., Kelleher, C., Kirkpatrick, R., Kirst, M., Kohler, A., Kalluri, U., Larimer, F., Leebens-Mack, J., Leple, J.-C., Locascio, P., Lou, Y., Lucas, S., Martin, F., Montanini, B., Napoli, C., Nelson, D.R., Nelson, C., Nieminen, K., Nilsson, O., Pereda, V., Peter, G., Philippe, R., Pilate, G., Poliakov, A., Razumovskaya, J., Richardson, P., Rinaldi, C., Ritland, K., Rouze, P., Ryaboy, D., Schmutz, J., Schrader, J., Segerman, B., Shin, H., Siddiqui, A., Sterky, F., Terry, A., Tsai, C.-J., Uberbacher, E., Unneberg, P., Vahala, J., Wall, K., Wessler, S., Yang, G., Yin, T., Douglas, C., Marra, M., Sandberg, G., de Peer, Y., Van, Rokhsar, D., 2006. The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313, 1596–1604.

Uknes, S., Dincher, S., Friedrich, L., Negrotto, D., Williams, S., Thompson-Taylor, H., Potter, S., Ward, E., Ryals, J., 1993. Regulation of pathogenesis-related protein-1a gene expression in tobacco. Plant Cell 5, 159–169.

Vernooij, B., Friedrich, L., Morse, A., Reist, R., Kolditz-Jawhar, R., Ward, E., Uknes, S., Kessmann, H., Ryals, J., 1994. Salicylic acid is not the translocated signal responsible for inducing systemic acquired resistance but is required in signal transduction. Plant Cell 6, 959–965.

Yang, X., Tuskan, G., Cheng, Z.M., 2006. Divergence of the dof gene families in poplar, Arabidopsis, and rice suggests multiple modes of gene evolution after duplication. Plant Physiol. 142, 820–830.

Zhang, J., 2003. Evolution by gene duplication: an update. Trends Ecol. Evol. 18, 292–298.

Zhao, N., Ferrer, J.L., Ross, J., Guan, J., Yang, Y., Pichersky, E., Noel, J.P., Chen, F., 2008. Structural, biochemical and phylogenetic analyses suggest that indole-3-acetic acid methyltransferase is an evolutionarily ancient member of the SABATH family. Plant Physiol. 146, 455–467.

Zhao, N., Guan, J., Lin, H., Chen, F., 2007. Molecular cloning and biochemical characterization of indole-3-acetic acid methyltransferase from poplar. Phytochemistry 68, 1537–1544.