ORNL/PPA-2004/1

Oak Ridge National Laboratory

# LABORATORY DIRECTED RESEARCH AND DEVELOPMENT PROGRAM

# FY 2003 ANNUAL REPORT

**March 2004**

**Prepared by**
**OAK RIDGE NATIONAL LABORATORY**
**P.O. Box 2008**
**Oak Ridge, Tennessee 37831-6285**
**managed by**
**UT-BATTELLE LLC**
**for the**
**U.S. DEPARTMENT OF ENERGY**
**under Contract DE-AC05-00OR22725**

# Ecosystem Genomics—An Emerging Opportunity for Environmental Research

S. P. DiFazio,[1] S. Jawdy,[1] L. Gunter,[1] B. A. Wilson,[2] and A. Brunner[3]
[1]*Environmental Sciences Division*
[2]*Department of Biology, Jackson State University*
[3]*Forest Science Department, Oregon State University*

We have undertaken a research program that capitalizes on the poplar genome sequence as an entry point for studying the molecular bases of ecologically significant processes such as flowering, drought adaptation, and response to perturbations such as elevated carbon dioxide. Analytical techniques for studying molecular underpinnings of phenotypes are currently most developed for association studies involving qualitative traits such as disease incidence. Therefore, as a proof of principle, we analyzed gender in poplar because this is an easily assayed qualitative trait that is under complex genetic control.[1] Determination of gender is most likely under the control of floral meristem identity genes, which in turn are regulated by transcription factors that control the transition of a vegetative meristem to a floral meristem, and transduction of environmental signals such as day length, light quality, and cold. We identified a total of 98 polymorphisms in 4.1 kb of mostly noncoding sequence in putative regulatory regions of four floral homeotic genes. There were substantial levels of linkage disequilibrium between polymorphic sites for three of the sequenced regions, suggesting that these polymorphisms could be under selection. There were no significant associations with gender for any of the sequenced polymorphisms. This is the first study of nucleotide polymorphism and linkage disequilibrium for wild poplar trees and provides important baseline data for future association studies in this model organism.

## Introduction

Ecosystem genomics has vast potential for enhancing understanding of ecosystem functioning as genomic information, methods, and analysis tools continue to develop. However, at present these tools are in an early stage of development, and much exploratory and proof-of-principle work is needed to develop and validate approaches for extending genomic information to ecosystem scales. We have begun this research program by focusing on a single model species that plays a dominant role in many ecosystems: the poplar tree (*Populus spp.*). There are many reasons that poplar serves as an excellent bridge from current, single-species genome studies to an ecosystem scale. First, poplar shares many of the desirable characteristics of other model species: it has a small genome size (550 Mb, 4 times larger than *Arabidopsis* and comparable to rice), it is readily transformed and clonally propagated, genetic linkage maps and large pedigrees exist, controlled crosses are readily performed within and between species, and a great deal is known about poplar physiology and genetics.[2] In recognition of its utility as a model species and relevance for agency objectives, the Department of Energy has sequenced the entire poplar genome, thus placing poplar in the same class as the handful of other model eukaryotes whose genomes have been sequenced.[3]

In addition to its characteristics as a model species, poplar provides an excellent point of departure for ecosystem research. The poplar genus consists of 29 species distributed throughout the northern hemisphere across a wide range of ecological amplitude.[4] Poplars often play a keystone role in riparian ecosystems, where they are pioneers on newly formed sediments and may be the dominant tree species on the landscape (e.g., in the high desert and intermountain west).[5] Individual poplar populations harbor a tremendous amount of genetic diversity in adaptive traits due to an outcrossing breeding system and extensive potential for gene flow among populations.[6] Also, closely related, sympatric poplar species form natural hybrid zones, further increasing the range of genetic variation in wild populations and creating excellent opportunities for studying the genetic mechanisms controlling species distributions.[7] Therefore, poplar offers an immediate opportunity for tractable studies of the molecular bases of adaptation with ecosystem-scale implications.

## Association Studies for Detecting Molecular Bases of Adaptation

The study of the genetic basis of adaptation has long been of interest to evolutionary biologists and applied plant and animal breeders, and a rich array of analytical

techniques have been developed.[8,9] Early attempts at quantifying the effects of natural selection focused on spatial and temporal changes in phenotypes and, later, on changes in gene frequencies in response to environmental gradients or stimuli.[10–12] With the advent of high-density genetic maps, it became possible to identify quantitative trait loci (QTL) for traits with adaptive and/or commercial value, based on linkage with neutral molecular markers.[13,14] QTL analysis allows the identification of chromosomal segments containing one or more genes with quantifiable effects on adaptation, and a number of groups have successfully cloned genes of adaptive and commercial significance using this technique.[15]

Despite its widespread adoption and application, QTL analysis has some substantial limitations for the discovery of adaptive molecular polymorphisms. First, the size of the region identified by QTL analysis depends on the amount of linkage disequilibrium in the genome segment, which in turn is determined by the effective population size, selection, and local recombination rates.[16,17] Large amounts of linkage disequilibrium can result in identification of extensive genome segments containing hundreds of candidate genes and thousands of polymorphisms, making it nearly impossible to identify the genes and polymorphisms of adaptive significance.[15,18] Alternatively, low levels of linkage disequilibrium can make initial detection of adaptive QTL extremely difficult, requiring a very high density of mapped molecular markers.[19] Also, the range of inference of QTL analysis can be unacceptably narrow. Robust identification of QTL requires large pedigrees containing several hundred progeny, so these analyses are usually performed in one or two families only, especially in the case of forest trees for which progeny tests are large and expensive. QTL effects may not be conserved across families because of epistatic effects of the different genetic backgrounds.[8,20] Finally, adaptive polymorphisms identified in the context of controlled crosses and field trials may have little significance in wild ecosystems because of different selection pressures and genotype by environment interactions.[21,22]

An alternative approach to identifying adaptive polymorphisms is to take advantage of naturally occurring variation in a species and search for statistical associations with molecular polymorphisms.[19,23] This approach has been most commonly applied to qualitative traits such as case-control studies of human disease using individual single nucleotide polymorphism (SNP) markers or population haplotypes consisting of SNP's in linkage disequilibrium.[24,25] This is an extremely active area of research, with a number of notable successes in identifying disease genes.[26] However, skeptics have questioned the value of haplotype-based association studies because of uncertainties about the level of disequilibrium and hence the number of makers required for full genome coverage,[16,27] the confounding effects of population structure, epistasis, allelic heterogeneity, and the difficulties in establishing a causative link between polymorphisms and phenotypes.[19,28,29] A solution that addresses many of these concerns is the use of candidate genes to narrow the search for adaptive polymorphisms and thus reduce the number of markers that must be screened.[15,20,23] A candidate gene approach is particularly appropriate for a species like poplar, which has an outcrossing breeding system, high levels of genetic variation within wild populations, extensive gene flow among populations, and continuously large population sizes for thousands of generations, all of which suggests that linkage disequilibrium will be quite low.[16,30] Therefore, a tremendous number of markers would likely be required for a whole-genome scan for associations with adaptive phenotypes, a prospect that would be prohibitively expensive with current technology. However, information from model organisms such as *Arabidopsis* can guide the identification of candidate genes in poplar based on sequence homology, and these can be used in hypothesis-driven searches for polymorphisms with adaptive significance.

## Pilot Study: Molecular Control of Gender Determination in Poplar

Analytical techniques are currently most developed for association studies involving qualitative traits such as disease incidence. Therefore, as a proof of principle, we analyzed gender in poplar because this is an easily assayed qualitative trait that is under complex genetic control.[1,31] Sex expression has been intensively studied in poplar because of its central importance in breeding and safety of genetically engineered plantations.[32] Also, sex expression has attracted considerable attention in model species such as *Arabidopsis*, in part because of the potential for direct improvements to food production through alteration of floral meristems, and also because of the central evolutionary importance of flowering. Determination of gender is most likely under the control of floral meristem identity genes, which in turn are regulated by transcription factors that control the transition of a vegetative meristem to a floral meristem, and transduction of environmental signals such as day length, light quality, and cold.[33] A large number of these genes have been identified in *Arabidopsis*, and a number of poplar homologs have been sequenced as well. The floral meristem identity genes are of particular interest, because they specify the formation of floral organs, and are likely to be involved in sex determination. In fact, overexpression of versions of poplar homologs of the transcription factors AGAMOUS and LEAFY resulted in production of female floral structures by a male poplar clone.[34] This suggests

that gender could be determined by a small number of sequence differences in regulatory elements and/or structural genes. This supposition is supported by results from an intensive effort aimed at finding neutral markers linked to sex determination in poplar[35] and willow (a member of the same family as poplar), which resulted in identification of only a single locus under complex genetic control.[31] By focusing on candidate genes known to affect floral morphology and gender, we should greatly enhance the prospects for elucidating the mechanism of gender formation.

## Candidate Genes for Gender Determination

The ABC model of floral organ determination is a generally accepted framework for the actions of floral homeotic genes. Briefly, A-function genes specify sepal and petal formation, B-function genes specify petals and stamens, and C-function genes specify carpels. Petals require both A and B genes, and stamens require both B and C.[36] Genes from each of these functional classes have been isolated and sequenced in poplar, and we have reason to believe that each may be involved in gender determination.

LEAFY: This gene controls transition of vegetative meristems to inflorescence meristems, so it is upstream of the "ABC" genes. As a bud is forming, a decision point determines whether it will become a branch or a flower. LEAFY expression is intimately involved in that decision, and LEAFY probably controls expression of the other floral genes involved in the ABC model. The poplar gene is called PTLF:[37] Overexpression of this gene in poplar has caused a male tree to produce hermaphroditic and female flowers.[34] There is a Sort Interspersed Nuclear Element (SINE) insertion in the promoter region of this gene, and the region around the insertion is strongly differentiated between male and female trees (A. Brunner, unpublished data).

APETALA1: This is an A-function gene that specifies petal and sepal formation. Its possible role in gender is unknown, but it regulates expression of some of the genes involved in specifying male and female structures. The poplar genes are called PTAP1-1 and PTAP1-2 (A. Brunner, unpublished). .

AGAMOUS: This is a C-function gene, and knockouts result in completely sterile flowers. The poplar versions are called PTAG1 and PTAG2.[1] Overexpression of PTAG2 has caused male poplar trees to produce female flowers.[34]

## Objectives

Objectives for this study were to explore molecular mechanisms of gender determination in poplar trees using a SNP association approach and assay allelic diversity and patterns of variation in major floral organ identity genes in poplar populations.

## Technical Approach

### *Plant Collections*

To accomplish the main objectives of this study, it was necessary to assay male and female trees that represented a cross section of the genetic variability that is present in *P. trichocarpa* populations. We therefore sampled trees across an east-west transect that represents the major axis of environmental variation in the range of *Populus trichocarpa*. We sampled trees from 7 populations, 3 of which were on the xeric east side of the Cascade mountains, and 4 from the more mesic west side of the Cascades (Table 1). Collections were made in November 2002. Many of the trees were selected because gender and microsatellite profiles had been determined for a previous study.[38] The remaining trees were haphazardly selected, taking care to avoid sampling multiple ramets from the same clone. We determined gender for trees with accessible floral buds. In total we analyzed 28 males, 26 females, and one cosexual tree that produced male, female, and hermaphroditic flowers, a fairly common occurrence in poplar.[39]

**Table 1. Locations of sample collections**
F, female; M, male; MF, cosexual.
Location is relative to the Cascade mountain range, which casts a substantial rain shadow to the east

| Site | F | M | MF | Location |
|---|---|---|---|---|
| Hood River | 0 | 3 | 0 | West |
| Marchel | 11 | 12 | 0 | West |
| River Ranch | 6 | 6 | 0 | West |
| Scappoose | 0 | 0 | 1 | West |
| John Day | 2 | 1 | 0 | East |
| Snake River | 1 | 3 | 0 | East |
| Umatilla | 6 | 3 | 0 | East |
| Total | 26 | 28 | 1 | |

### *Neutral Variation*

We assessed variation for eight microsatellite loci derived from random genomic sequencing and an enrichment[40] (Table 2). We calculated genetic distance as the sum of squared differences between loci[41] and determined relationships among genotypes using the Unweighted Pair-Group Method using Arithmetic means in the PHYLIP software package.[42] We also calculated differentiation based on provenance (east versus west) and gender using *Rst* and exact tests in the Arlequin program.[43]

### *Selection of Genotyping Targets*

It appears that poplars of each gender retain the ability to produce floral organs of the alternate gender, based on

**Table 2. Characteristics of microsatellite loci used for assessment of neutral variation in poplar samples**

LG, chromosomal linkage group on which marker is located;[56] N, number of trees; He, expected heterozygosity; Ho, observed heterozygosity; Fis, fixation index

| Marker | LG | N | Alleles | He | Ho | Fis |
|---|---|---|---|---|---|---|
| O15 | XVII | 33 | 25 | 0.95 | 0.94 | 0.01 |
| O349 | IV | 35 | 10 | 0.77 | 0.63 | 0.18 |
| P2515 | XIV | 36 | 15 | 0.87 | 0.78 | 0.11 |
| P2571 | X | 36 | 10 | 0.84 | 0.69 | 0.17 |
| P2585 | XV | 29 | 13 | 0.83 | 0.59 | 0.30 |
| P2610 | VIII | 33 | 10 | 0.59 | 0.45 | 0.22 |
| P2885 | XII | 30 | 11 | 0.83 | 0.67 | 0.19 |
| P649 | XIII | 32 | 37 | 0.97 | 0.75 | 0.23 |

the occurrence of cosexual genets and the conversion of gender in transgenics. We therefore hypothesized that gender was determined primarily by the patterns of expression of regulatory genes. We focused on known and suspected regulatory regions of some of the key regulatory genes involved in floral organ identity: PTAG1, PTAG2, PTLF, and PTAP1. We further refined our sequencing targets by comparing the genome sequence of the female clone, Nisqually-1, which has recently been sequenced by DOE's Joint Genome Institute, to sequences previously determined for a male clone.[1,37,44]

### *Polymorphism Detection*

We tested two main approaches for discovering polymorphisms in our candidate gene regions: cloning and sequencing of PCR fragments, and direct sequencing of PCR products. In the first approach we PCR-amplified 2-kb fragments of each gene and cloned these using the Topo-TA cloning kit (Invitrogen). We then sequenced 3 to 10 clones per individual for each fragment, using universal primers for the initial forward and reverse sequences, and then designing specific sequencing primers for internal sequence. For direct PCR sequencing we amplified 500- to 700-bp fragments from the regions of interest, purified the products with EXO-SAP kits, and then sequenced the products directly. All amplifications were accomplished with Pfu DNA polymerase (Stratagene), and all sequencing was done with Big Dye Terminator kits (ABI). All sequencing and fragment analysis was accomplished on an ABI 3700 capillary sequencer.

### *SNP Analysis*

We created alignments for each gene fragment using the Phred/Phrap/Consed suite of base calling and assembly programs.[45,46] We identified polymorphisms using Polyphred, an add-on to the above package that calls heterozygous bases for diploid sequences.[47] We processed the Polyphred output with a series of PERL scripts to identify haplotypes and format data for input into the programs for further analysis. For haploid data, derived by sequencing cloned fragments, haplotype diversity could be assessed directly using the Arlequin program, and comparison among haplotypes was accomplished using Jukes and Cantor genetic distance measure using the DNAdist program in the PHYLIP package, followed by construction of a UPGMA dendrogram. For diploid data derived by sequencing PCR products directly, haplotypes were inferred from allele frequencies using the maximum likelihood routine in the Arlequin package.[43]

We tested for linkage disequilibrium between all pairs of loci using exact tests as implemented by the Arlequin package for both haploid and diploid data. We used a Bonferroni criterion for significant linkage: the critical value for each gene fragment was calculated as 0.05/N, where N is the number of pairwise comparisons.

## Results

### *Neutral Variation*

The eight microsatellite loci were highly polymorphic in these populations, with 10 to 28 alleles observed for 37 to 44 individuals. Expected heterozygosities ranged from 0.59 to 0.97, but observed hetorozygosities were considerably lower, ranging from 0.45 to 0.94 and with FIS values between 0.09 and 0.30. These low-to-moderate FIS values may indicate that pooling of the samples has not caused a substantial Wahlund effect.[48] Furthermore, there was no evidence of differentiation between East and West trees ($R_{ST} = -0.021$, $P = 0.68 \pm 0.04$) and male and female trees ($R_{ST} = 0.083$, $P = 0.41 \pm 0.05$). Sampling was inadequate to assess pairwise population differentiation, so it is impossible to definitively assess the severity of the Wahlund effect caused by pooling these samples. However, a cluster analysis did not indicate substructure among these populations (Fig. 1), and the extremely long gene flow distances characteristic of *P. trichocarpa* have probably inhibited population differentiation for neutral loci in this species.[30,38]

## Polymorphisms and Patterns of Linkage Disequilibrium

### *Overall Patterns*

In total we produced over 2600 sequences for this project, and 832 of these were of sufficient length and quality to be included in the final alignments. We successfully surveyed 3638 bases of noncoding DNA and 498 bases of coding DNA in the vicinity of four floral homeotic genes (Table 3). We identified 98 polymorphisms, including 86 single nucleotide polymorphisms (SNPs) and 12 insertion/deletion (indel) polymorphisms. Ninety-two of the polymorphisms were

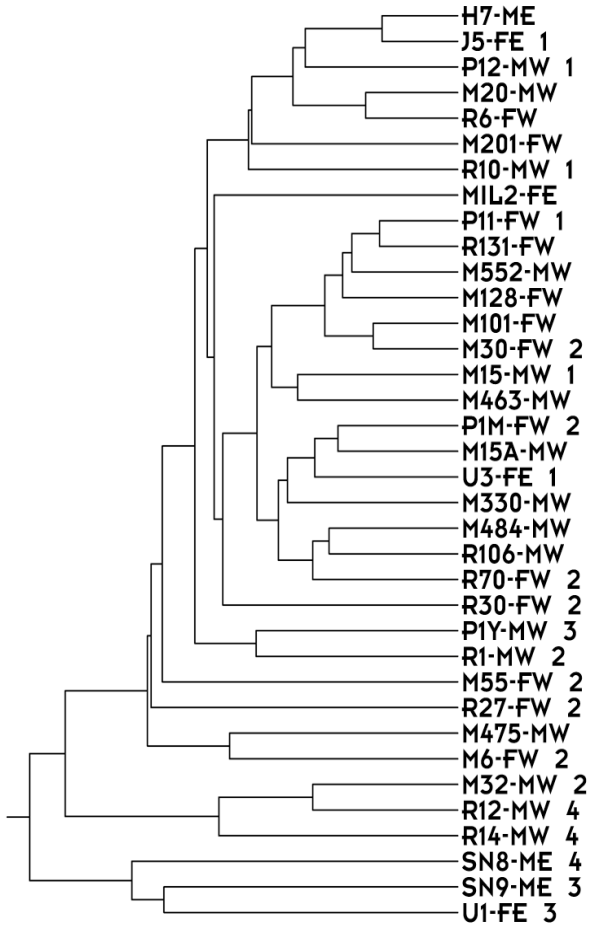


*Fig. 1. UPGMA dendrogram depicting relationships among multilocus microsatellite genotypes. Distance measurement was the sum of squared pairwise differences in allele sizes (Slatkin 1995).*

located in noncoding DNA, with 79 polymorphisms in upstream regions and 13 in introns. The only polymorphisms identified in coding regions were in the 5' untranslated region (UTR) of the PTAG1 gene. No polymorphisms were observed in 229 bp of exon sequence. In total, 169 sites showed significant LD, the vast majority of which were in the 5' region of PTAG1 (Table 4). There was no apparent relationship between LD and distance between sites (Fig. 2).

### *PTAG1*

We sequenced portions of a cloned 2-kb fragment that began 1750 bp of upstream of the transcription start site, and extended 269 bases into the 5' UTR region. The first 1800 bp of sequence corresponds to a previously undescribed Ty1-copia class retrotransposon,[49] designated PtOP1. We observed a total of 51 SNPs in this region and no indels (Fig. 3). There were 36 different haplotypes observed among the 62 sequenced chromatids, and these ranged in frequency from 1.6% to 20%. There was

**Table 3. Frequency of polymorphisms identified in coding and noncoding regions**
P, number of polymorphisms observed

| | Length | P | Rate (snp/kb) |
|---|---|---|---|
| | Noncoding | | |
| Upstream | 2605 | 79 | 30.3 |
| Intron | 1033 | 13 | 12.6 |
| Total | 3638 | 92 | 25.3 |
| | Coding | | |
| 5' UTR | 269 | 6 | 22.3 |
| Exon | 229 | 0 | 0.00 |
| Total | 498 | 6 | 12.1 |
| Total | 4136 | 98 | 23.7 |

significant linkage disequilibrium between 166 pairs of loci in this region (Fig. 4; Table 4). There were no significant differences in patterns of polymorphisms between males and females in this region, and patterns of linkage disequilibrium were similar between the two genders (Fig. 4). There were also no apparent patterns of relationships among haplotypes to suggest an association between these polymorphisms and gender (Fig. 5).

We also sequenced the 292 bases immediately upstream of the transcriptional start site, and 269 bases of the 5' untranslated region (UTR) (Table 4). In total we observed 13 polymorphisms, including two indels, and 22 haplotypes among the 56 assayed chromatids. Six of the SNPs were in the 5' UTR. No linkage disequilibrium was detected between the 156 pairs of polymorphic loci in the 561 bp region flanking the start site (Table 4).

Finally, we sequenced portions of intron 5 and exon 5 of the PTAG 1 gene, directly from a PCR product spanning positions 7060 to 7419. We observed 23 polymorphisms in this region and one indel, with a total of 23 haplotypes among the 25 PCR products examined. All of the polymorphisms were in the intron, and there were no fixed differences between males and females. We observed no significant LD between any of these polymorphisms.

### *PTAG2*

For the PTAG2 gene we focused on a region 5' of the transcriptional start site that contained a large (426 bp) insertion/deletion polymorphism that differed between the initially sequenced male and female clones. This indel was particularly intriguing due to the presence of a 30 bp tandem repeat that could be indicative of a protein binding site, suggesting a possible regulatory role. A PCR assay using primers flanking the site of the insertion revealed that 22 out of 24 trees possessed at least one copy of the insertion, and four trees were heterozygous for the insertion polymorphism. The polymorphism was approximately equally divided between males and females, with one homozygote of each gender, and 2 male and one female

**Table 4. Polymorphisms, linkage disequilibrium, and differentiation between males and females for sequenced fragments**
T, type of fragment: H, haploid, D, diploid; N, number of trees sequenced; F, females; M, Males; A, All trees, B, polymorphism shared by both; M-F Dist, genetic distance between male and female clones; P Exact, value from Exact test

| Gene | T | N | F | M | Start | End | Haplotypes | | | Polymorphisms | | | Indels | | | LD | | | | M-F Dist | P Exact | Location |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | A | F | M | A | F | M | A | F | M | A | F | M | B | | | |
| PTAG1 | H | 31 | 15 | 14 | –1745 | –611 | 36 | 18 | 25 | 51 | 36 | 27 | 0 | 0 | 0 | 166 | 83 | 64 | 35 | –0.061 | 0.12 | 5' |
| PTAG1 | H | 28 | 13 | 14 | –292 | 269 | 22 | 8 | 15 | 13 | 6 | 8 | 2 | 2 | 2 | 0 | 0 | 0 | 0 | –0.978 | 0.06 | 5'/UTR |
| PTAG1 | D | 25 | 14 | 11 | 7060 | 7419 | 23 | 14 | 12 | 8 | 8 | 8 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | –0.006 | 0.79 | Intron/Exon5 |
| PTAG2 | D | 17 | 11 | 6 | –908 | –686 | 10 | 9 | 3 | 12 | 8 | 0 | 5 | 5 | 1 | 0 | 0 | 0 | 0 | 0.139 | 1.00 | 5' |
| PTAP1-1 | D | 29 | 15 | 14 | 657 | 1560 | 16 | 14 | 11 | 5 | 5 | 5 | 0 | 0 | 0 | 3 | 4 | 3 | 3 | –0.048 | 0.07 | Intron/Exon6 |
| PTLF | D | 21 | 12 | 9 | –1630 | –1119 | 36 | 19 | 16 | 9 | 8 | 7 | 4 | 4 | 3 | 0 | 4 | 3 | 3 | –0.112 | 1.00 | 5' |

heterozygotes. In addition, the cosexual tree was heterozygous for the polymorphism.

We also sequenced this PCR product for 17 trees. There were 12 polymorphisms in all, but nine of these occurred in just one female individual, and 5 were indels. The sequence of the large indel and flanking region was invariant for the 6 male trees we sequenced. No significant linkage disequilibrium was detected among the three polymorphisms with sufficiently high variation to allow testing.

### *PTAP1-1*

There has been some uncertainty about the number of AP1 homologs in poplar due to the detection of multiple cDNAs corresponding to PTAP1-1, during initial cloning (A. Brunner, unpublished). One of the cDNA's had the entire 3' portion of the gene deleted following exon 5. It was unclear if this was due to alternate splicing or presence of a third gene. Furthermore, Southern blots suggested that there might be two copies of PTAP1-1 in females but not in males (A. Brunner, unpublished data). We therefore amplified the region between exons 5 and 6 to determine if a length polymorphism existed that differentiated male and female trees. Amplification of 30 trees, 15 males, 14 females, and one cosexual, failed to reveal any such polymorphisms, suggesting that alternative splicing accounts for the heterogeneous cDNAs rather than

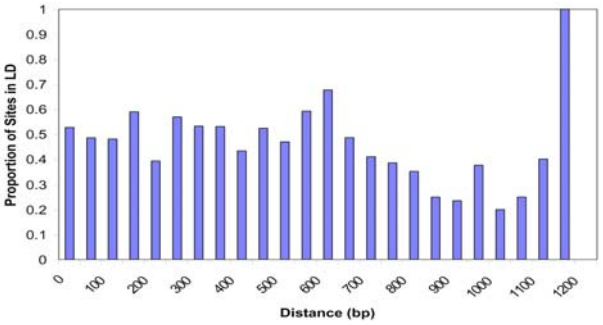


*Fig. 2. Relationship between pairwise distance between polymorphisms and the occurrence of linkage disequilibrium for all sequenced fragments.*

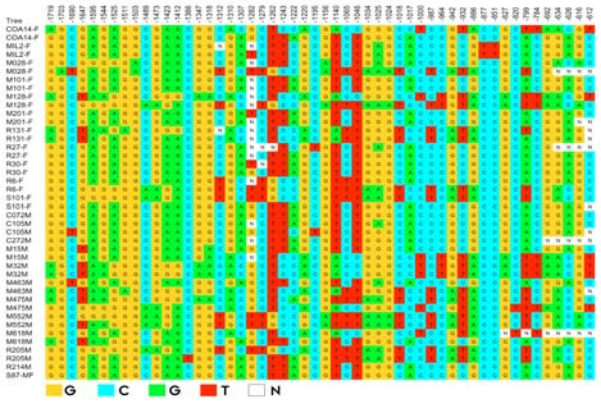


*Fig. 3. Polymorphisms in a portion of a cloned fragment of the upstream region of the PTAG1 gene, which corresponds to a portion of the retrotransposon PtOP1. Individual trees are in rows, with a trailing F or M indicating female or male, respectively, and MF indicating a cosexual tree. Position relative to the start site is indicated in the first row.*

duplicated loci. We also sequenced these PCR products and discovered five polymorphisms that were present in both male and female trees. All polymorphisms were located in introns. There was significant LD between 3 pairs of these polymorphic sites (Table 4).

### *PTLF*

We focused on the 5' region of the PTLF gene that contained a SINE (short interspersed nuclear element) which also showed substantial similarity to a SINE in the 5' region of the PTAP1-1 gene. We identified nine polymorphisms in this region, four of which were indels. Three of these polymorphisms showed significant LD in males and four in female trees. There was no evidence of differentiation between males and females, however, and three of four pairs showed LD in both genders.

## Discussion

Rates of polymorphism were moderately high in the regions we surveyed, more than an order of magnitude greater than polymorphisms observed in Arabidopsis (1 per kb,[50] about five times greater than Fugu (4 per kb[51]),

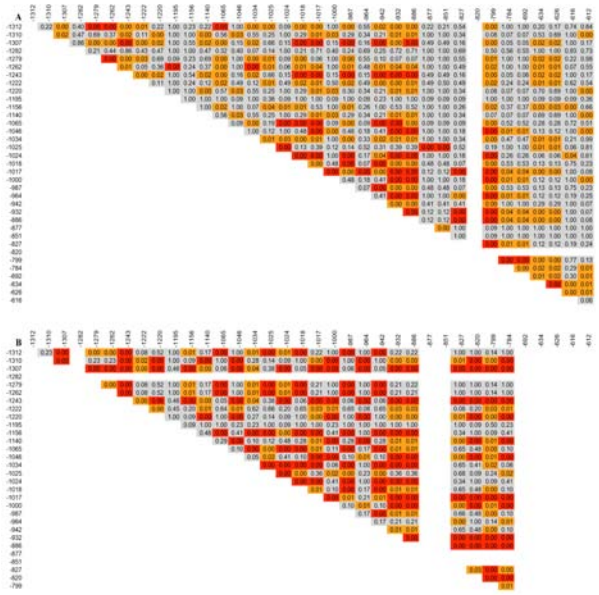

*Fig. 4. Linkage disequilibrium between pairs of polymorphic sites in the retrotransposon 5' to the PTAG1 transcriptional start site. A. Female trees, upstream fragment. B. Male trees, upstream fragment. Blank rows and columns are loci that are fixed in male or female trees. Orange indicates significance at the 0.05 level. Red indicates significance with a Bonferroni correction (P < 0.0001 in this case).*

comparable to sea squirt (12 per kb in a single individual[52]), and maize (32 per kb[53]). The patterns in this limited sample matched expectations: noncoding regions were greater than introns which in turn were greater than 5' UTR's. No polymorphisms were detected in exons, but the sample was extremely small. Much more robust estimates will soon be available for the individual clone that is being sequenced by the U.S. DOE.

These relatively high rates of polymorphism will provide abundant variation that can be exploited for association studies such as was attempted for the present study. However, LD was relatively low for most of the surveyed regions. If this pattern holds true on a genomic scale, this will make genome-wide association studies virtually impossible because an unrealistically large number of markers will be required to survey the genome.[16,27] At the same time, low LD will enhance the confidence of associations that are discovered between phenotypes and candidate gene polymorphisms, because low LD decreases the possibility of spurious associations.[19,28,29] Low LD also reduces the practical utility of QTL detected in experimental crosses, because marker-trait associations will be unpredictably disrupted outside of the mapped families.[54] Therefore, the present study provisionally supports the use of candidate gene association studies as a means of functional characterization in poplar.

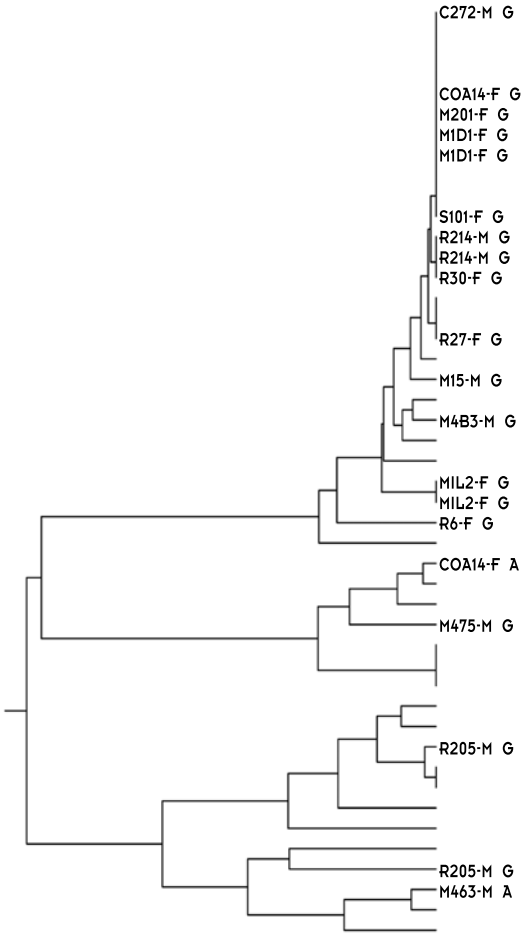


*Fig. 5. UPGMA dendrogram for haplotypes for the same portions of the PTAG1 gene that are pictured in Figs. 3 and 4. All haplotypes of sequenced individuals are represented, and those connected by vertical lines are identical.*

There are two important caveats about the LD estimates reported in this study. First, power was relatively low for most of the fragments due to the diploid nature of the data and relatively small number of individuals. Most LD was detected for the PTAG1 5' retrotransposon region, and this was also the data set for which the most power existed for detecting LD because linkage phase was known unequivocally, and the largest number of sequences were from this fragment. It is possible that LD would have been detected in several of the other fragments if more individuals had been sampled and if linkage phase had been determined. The other caveat is that LD was only tested over relatively short stretches, and there was no indication that rates of significant LD were declining with distance between tested markers, and LD was scattered throughout the fragments in which it occurred. This suggests that the LD is not due to repressed recombination

in these regions, because this would result in contiguous blocks of linkage.[55] Rather, this pattern may be indicative of coordinate selection on the paired polymorphisms. Such LD could occur over very large distances and illustrates that generalizations about LD will be difficult in the absence of direct tests of much larger numbers of loci.

The "proof-of-principle" portion of this study failed to uncover the hypothesized gender-associated polymorphisms. The observed polymorphisms occurred in both males and females, and LD generally occurred between the same pairs of loci in males and females (except in cases where spurious associations occurred due to splitting the data set and thereby reducing the frequency of individual polymorphisms). It should not be surprising that no associations were uncovered with this limited survey, since we only examined a small fraction of the possible regions of differentiation between the genders, and previous studies had indicated that gender determination was likely complex in the Salicaceae.[31,35] Furthermore, it is quite possible that epigenetic differences could account for much of the differential gene regulation that results in production of alternate sex organs in poplar, and these would not be detected with simple sequence surveys. Nevertheless, this approach is a valid one and warrants further effort, particularly as the finished poplar genome sequence emerges, and this can be readily compared on a large scale to ESTs and genomic sequence from male clones.

Finally, this study provided some insight into methodological issues surrounding SNP genotyping. The relatively high rates of polymorphism and low LD argue against SNP genotyping techniques that target individual polymorphisms. In fact, sequencing is currently one of the most cost-effective ways to survey polymorphisms in poplar because of the high density of polymorphisms. Furthermore, direct sequencing of PCR products is a viable strategy for identifying many poplar SNPs, but this technique suffers from significant shortcomings: (1) linkage phase cannot be determined, so power for detecting LD is greatly reduced; (2) the occurrence of indels renders diploid sequences uninterpretable; (3) identification of heterozygotes is tenuous except for the highest quality sequence regions; and (4) it is nearly impossible to sequence duplicated genome regions this way because mixtures of PCR products provide very poor sequence. Therefore, if SNP discovery and assessment of LD are the goals, cloning of mid-sized PCR fragments (2–4 Kb) is a viable but expensive strategy. Ultimately, shotgun sequencing of tiled BACs from multiple individuals may be the best way to determine SNP diversity and LD over large regions.

## Summary and Conclusions

This proof-of-principle study has provided valuable initial data on SNP variation in poplar, a species with tremendous economic and ecological importance. The knowledge and techniques that we have gained from this work will be directly applied in future studies of adaptation of plant populations, and will therefore add to our understanding of ecosystem composition and aid in predicting ecosystem responses to climatic change. This will have direct benefits for multiple U.S. government agencies that are concerned with the effects of climatic change, including DOE, NASA, EPA, and DOD.

This work has already led to collaborations that will enhance the competitiveness of ORNL in plant genetic research. We have been involved in two proposals that build directly off of this work. One was for the NSF Frontiers in Integrative Biological Research. This multimillion dollar project is a collaboration between Northern Arizona University, ORNL, and the University of Wisconsin. The preproposal recently passed competitive review, and we have been invited to submit a full proposal this year. In addition, we have been invited to participate in another proposal for the NSF Systematic and Population Biology Program, which would bring several hundred thousand dollars to the lab if funded. Finally, this work has positioned us to pursue future opportunities with DOE programs such as the Program for Ecosystem Research and the Basic Energy Sciences program.

## References


[1]A. M. Brunner, W. H. Rottmann, L. A. Sheppard, K. Krutovskii, S. P. DiFazio, S. Leonardi, and S. H. Strauss, "Structure and expression of duplicate AGAMOUS orthologues in poplar," *Plant Mol Biol.* **44**, 634 (2000).

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

[3]C. C. Mann, and M. L. Plummer, "Biotechnology: Forest Biotech Edges Out of the Lab," *Science* **295**, 1626–1629 (2002).

[4]J. E. Eckenwalder, "Systematics and evolution of *Populus*." In *Biology of Populus and Its Implications for Management and Conservation*, pp. 7–32. Editors: R. F. Stettler, H. D. Bradshaw, Jr., P. E. Heilman, and T. M. Hinckley. NRC Research Press, Ottawa, Canada, 1996.

[5]J. H. Braatne, S. B. Rood, and P.E. Heilman, "Life history, ecology, and reproduction of riparian cottonwoods in North America." In *Biology of Populus and Its Implications for Management and Conservation*, pp. 57–85. Editors: R. F. Stettler, H. D. Bradshaw, Jr., P. E. Heilman, and T. M. Hinckley. NRC Research Press, Ottawa, Canada, 1996.

[6]R. E. Farmer, Jr. (1996) The genecology of *Populus*. In: Biology

of Populus and its implications for management and conservation, pp. 33–55. Editors: R. F. Stettler, H. D. Bradshaw, Jr., P. E. Heilman, and T. M. Hinckley. NRC Research Press, Ottawa, Canada.

[7]L. H. Rieseberg and S. E. Carney, "Tansley Review no. 102 plant hybridization," *New Phytol.* **140**, 599–624 (1998).

[8]N. H. Barton and P. D. Keightley, "Understanding quantitative genetic variation," *Nature Reviews Genetics* **3**, 11–21 (2002).

[9]P. W. Hedrick, M. E. Ginevan, and E. P. Ewing, "Genetic polymorphism in heterogeneous environments," *Annual Review of Ecology and Systematics* **7**, 1–32 (1976).

[10]R. Lande and S. J. Arnold, "The measurement of selection on correlated characters," *Evolution* **36**, 1210–1226 (1983).

[11]J. G. Kingsolver, H. E. Hoekstra, J. M. Hoekstra, D. Berrigan, S. N. Vignieri, C. E. Hill, and A. Hoang, "The strength of phenotypic selection in natural populations," *American Naturalist* **157**, 245–261 (2001).

[12]K. Ritland, "Marker-inferred relatedness as a tool for detecting heritability in nature," *Molecular Ecology* **9**, 1195–1204 (2000).

[13]R. Doerge, "Mapping and analysis of quantitative trait loci in experimental populations," *Nature Reviews Genetics* **3**, 43–52 (2002).

[14]M. Lynch and B. Walsh, Genetics and Analysis of Quantitative Traits. Sinauer Associates, Sunderland, MA, 1998.

[15]D. L. Remington, M. C. Ungerer, and M. Purugganan, "Map-based cloning of quantitative trait loci: progress and prospects," *Genet.Res.* **78**, 213–218 (2001).

[16]J. D. Terwilliger and K. M. Weiss, "Linkage disequilibrium mapping of complex disease: fantasy or reality?," *Curr. Opin. Biotechnol.* **9**, 578–594 (1998).

[17]J. K. Pritchard and M. Przeworski, "Linkage disequilibrium in humans: Models and data," *American Journal of Human Genetics* **69**, 1–14 (2001).

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

[19]L. R. Cardon and J. I. Bell, "Association study designs for complex diseases," *Nat. Rev. Genet.* **2**, 91–99 (2001).

[20]B. Walsh, "Quantitative genetics in the age of genomics," *Theoretical Population Biology* **59**, 175–184 (2001).

[21]S. H. Strauss, R. Lande, and G. Namkoong, "Limitations of Molecular-Marker-Aided Selection in Forest Tree Breeding," *Canadian Journal of Forest Research* **22**(7), 1050–1061, 1050–1061 (1992).

[22]M. Pigliucci, "Ecological and evolutionary genetics of *Arabidopsis*," *Trends in Plant Science* **3**, 485–489 (1998).

[23]N. J. Risch, "Searching for genetic determinants in the new millennium," *Nature* **405**, 847–856 (2000).

[24]D. E. Reich and E. S. Lander, "On the allelic spectrum of human disease," *Trends in Genetics* **17**, 502–510 (2001).

[25]G. C. Johnson, L. Esposito, B. J. Barratt, A. N. Smith, J. Heward, G. Di Genova, H. Ueda, H. J. Cordell, I. A. Eaves, F. Dudbridge, R. C. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S. C. Gough, D. G. Clayton, and J. A. Todd, "Haplotype tagging for the identification of common disease genes," *Nat. Genet.* **29**, 233–237 (2001).

[26]E. R. Martin, E. H. Lai, J. R. Gilbert, A. R. Rogala, A. J. Afshari, J. Riley, K. L. Finch, J. F. Stevens, K. J. Livak, B. D. Slotterbeck, S. H. Slifer, L. L. Warren, P. M. Conneally, D. E. Schmechel, I. Purvis, M. A. Pericak-Vance, A. D. Roses, and J. M. Vance, "SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around APOE in Alzheimer disease," *Am. J. Hum. Genet.* **67**, 383–394 (2000).

[27]L. Kruglyak, "Prospects for whole-genome linkage disequilibrium mapping of common disease genes," *Nat. Genet.* **22**, 139–144 (1999).

[28]K. M. Weiss and J. D. Terwilliger, "How many diseases does it take to map a gene with SNPs?," *Nat. Genet.* **26**, 151–157 (2000).

[29]D. Altshuler, M. Daly, and L. Kruglyak, "Guilt by association," *Nature Genetics* **26**, 135–137 (2000).

[30]J. C. Weber and R. F. Stettler, "Isoenzyme variation among ten [riparian] populations of Populus trichocarpa Torr. et Gray in the Pacific Northwest," *Silvae Genet.* **30**, 82–87 (1981).

[31]C. L. M. Alstrom-Rapaport, Y. C. Wang, G. Roberts, and G. A. Tuskan, "Identification of a RAPD marker linked to sex determination in the basket willow (*Salix viminalis* L.)," *Journal of Heredity* **89**, 44–49 (1998).

[32]S. H. Strauss, W. H. Rottmann, A. M. Brunner, and L. A. Sheppard, "Genetic engineering of reproductive sterility in forest trees," *Molecular Breeding* **1**, 5–26 (1995).

[33]G. G. Simpson, A. R. Gendall, and C. Dean, "When to switch to flowering," *Annu. Rev. Cell Dev. Biol.* **15**, 519–550 (1999).

[34]S. H. Strauss, R. Meilan, S. P. DiFazio, A. M. Brunner, and J. Carson, Tree Genetic Engineering Research Cooperative (TGERC) Annual Report:2000–2001. Forest Research Laboratory, Oregon State University, Corvallis, Oregon, 2001.

[35]D. N. McLetchie and G. A. Tuskan, "Gender determination im Populus," *Norwegian Journal of Agricultural Sciences Supplement* **18**, 57–66 (1994).

[36]A. L. Lawton-Rauh, E. S. Buckler, and M. D. Purugganan, "Patterns of molecular evolution among paralogous floral homeotic genes 8," *Mol. Biol. Evol.* **16**, 1037–1045 (1999).

[37]W. H. Rottmann, R. Meilan, L. A. Sheppard, A. M. Brunner, J. S. Skinner, C. Ma, S. Cheng, L. Jouanin, G. Pilate, S. H. Strauss, C. P. Ma, and S. P. Cheng, "Diverse effects of overexpression of LEAFY and PTLF, a poplar (Populus) homolog of LEAFY/FLORICAULA, in transgenic poplar and Arabidopsis," *Plant J.* **22**, 235–245 (2000).

[38]S. P. DiFazio, Measuring and Modeling Gene Flow from Hybrid Poplar Plantations: Implications for Transgenic Risk Assessment. PhD Thesis, Oregon State University, Corvallis, OR, 2002. http://www.fsl.orst.edu/tgerc/dif_thesis/difaz_thesis.pdf.

[39]R. F. Stettler, "Variation in sex expression of Black Cottonwood

and related hybrids," *Silvae Genet.* 42–46 (1971).

[40]G. A. Tuskan, L. E. Gunter, Z. Yang, T. M. Yin, M. Sewell, and S. P. DiFazio, "Characterization of Microsatellites Revealed by Genomic Sequencing of *Populus trichocarpa*," *Can. J. Forest Res.* in press (2004).

[41]M. Slatkin, "A Measure of Population Subdivision Based on Microsatellite Allele Frequencies," *Genetics* **139**, 457–462 (1995).

[42]J. Felsenstein (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. Produced and Distributed by Author.

[43]M. Slatkin and L. Excoffier, "Testing for linkage disequilibrium in genotypic data using the expectation-maximization algorithm," *Heredity* **76**, 377–383 (1996).

[44]L. A. Sheppard, A. M. Brunner, K. V. Krutovskii, W. H. Rottmann, J. S. Skinner, S. S. Vollmer, and S. H. Strauss, "A DEFICIENS homolog from the dioecious tree black cottonwood is expressed in female and male floral meristems of the two-whorled, unisexual flowers," *Plant Physiol.* **124**, 627–639 (2000).

[45]D. Gordon, C. Abajian, and P. Green, "Consed: a graphical tool for sequence finishing," *Genome Res.* **8**, 195–202 (1998).

[46]B. Ewing and P. Green, "Base-calling of automated sequencer traces using phred. II. Error probabilities," *Genome Research* **8**, 186–194 (1998).

[47]D. A. Nickerson, V. O. Tobe, and S. L. Taylor, "PolyPhred: Automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing," *Nucleic Acids Res.* **25**, 2745–2751 (1997).

[48]D. Hartl and A. G. Clark, Principles of Population Genetics. Sinauer Associates, Inc., Sunderland, MA, 1997.

[49]A. J. Flavell, S. R. Pearce, J. S. P. Heslop-Harrison, and A. Kumar, "The evolution of Ty1-copia group retrotransposons in eukaryote genomes," *Genetica* **100**, 185–195 (1997).

[50]F. J. Cho, M. Mindrinos, D. R. Richards, R. J. Sapolsky, M. Anderson, E. Drenkard, L. Dewdney, T. L. Reuber, M. Stammers, N. Federspiel, A. Theologis, W. H. Yang, E. Hubbell, M. Au, E. Y. Chung, D. Lashkari, B. Lemieux, C. Dean, R. J. Lipshutz, F. M. Ausubel, R. W. Davis, and P. J. Oefner, "Genome-wide mapping with biallelic markers in Arabidopsis thaliana," *Nature Genetics* **23**, 203–207 (1999).

[51]S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. S. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. K. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. Hawkins, B. Venkatesh, D. Rokhsar, and S. Brenner, "Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes," *Science* **297**, 1301–1310 (2002).

[52]P. Dehal, Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. De Tomaso, B. Davidson, A. Di Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. M. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Necula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H. G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. S. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B. I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh, and D. S. Rokhsar, "The draft genome of Ciona intestinalis: Insights into chordate and vertebrate origins," *Science* **298**, 2157–2167 (2002).

[53]A. Ching, K. S. Caldwell, M. Jung, M. Dolan, O. S. Smith, S. Tingey, M. Morgante, and A. J. Rafalski, "SNP frequency, haplotype structure and linkage disequilibrium in elite maize inbred lines," *Bmc Genetics* **3** (2002).

[54]S. Strauss, R. Lande, and G. Namkoong, "Limitations of molecular-marker-aided selection in forest tree breeding," *Canadian Journal of Forest Research* **22**, 1050–1061 (1992).

[55]D. E. Reich, M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti, D. J. Richter, T. Lavery, R. Kouyoumjian, S. F. Farhadian, R. Ward, and E. S. Lander, "Linkage disequilibrium in the human genome," *Nature* **411**, 199–204 (2001).

[56]T. M. Yin, S. P. DiFazio, L. E. Gunter, D. E. Riemenschneider, and G. A. Tuskan, "Large-scale Heterospecific Segregation Distortion in Populus Revealed by a Dense Genetic Map," *Theoretical & Applied Genetics* (accepted) (2004).