8 The *Populus* Genome Initiative

Stephen DiFazio
Department of Biology, West Virginia University, Morgantown, West Virginia 26508-6057 USA
Email: spdifazio@mail.wvu.edu

1 INTRODUCTION

1.1 Background on *Populus* Biology

The genus *Populus* consists of about 29 species organized into six major sections that occur primarily in the Northern Hemisphere (Eckenwalder 1996). Species from the different sections of the genus have diverse ecological characteristics. Two of the most economically important sections (Aigeiros and Tacamahaca) contain species collectively known as cottonwoods. These occur mostly in riparian zones and are characterized by primarily ruderal life history, dominating early successional stages and thriving on flood-mediated disturbance (Braatne et al. 1996; Karrenberg et al. 2002). The other major section of the genus (section Populus, also known as Leuce), contains species commonly known as aspens, which are characterized by extensive clonal growth, and which can occur in very diverse sites, from mixed upland forests to boreal regions (Barnes 1967; Peterson and Peterson 1992; Romme et al. 2001). The cottonwood sections of the genus are highly interfertile and readily hybridize, and the aspen species are also highly interfertile, but the cottonwoods are reproductively isolated from the aspens. This isolation is reflected by the strong ecological, morphological, and genetic distinctions between these groups of species (Stettler et al. 1980). All

Populus species have 38 chromosomes in their diploid genomes (Blackburn and Heslop Harrison 1924), with exceptions apparently occurring regularly within species due to meiotic nondisjunction (Einspahr et al. 1963; Bradshaw and Stettler 1993).

1.2 Selection of *Populus* for Sequencing

Populus was the first tree selected for whole-genome sequencing, and the third plant overall. At the time of its selection in 2003, *Populus* was already well-established as a model organism because of its experimental tractability, potential economic importance, and its central role in many ecosystems (Wullschleger et al. 2002). One of the primary advantages of *Populus* is the ease with which it can be vegetatively propagated using simple stem cuttings in many species without the necessity of hormonal treatments (Heilman 1999). Another key factor is that the genus consists of ecologically contrasting species that readily hybridize. This occurs naturally between sympatric *Populus* species around the world (Eckenwalder 1984; Zsuffa et al. 1996; Lexer et al. 2004), and has been a key feature of *Populus* breeding programs for decades (Stettler et al. 1996). In fact, the combination of hybridization and vegetative propagation has dominated *Populus* genetic improvement efforts thus far, and most major breeding programs have focused on making a large number of interspecific crosses and choosing clones that show a high degree of heterosis, and display characteristics that are particularly amenable to plantation culture (Stettler et al. 1996). This trait has also made *Populus* one of the premier models for genetic engineering (Taylor 2002; Bhalerao et al. 2003), because transformed lines with desirable characteristics and low somaclonal variation can be rapidly and cheaply disseminated (Strauss et al. 1995). Hybridization has also been a key feature in ecological studies, as *Populus* hybrid zones have provided a fertile area for ecological and evolutionary genetics, and have been a favorite subject in the burgeoning new field of Community Genetics (Whitham et al. 2006).

The rationale for selecting *Populus* was based in part on its importance, and in part on its tractability as a model organism for genomics research. The *Populus* whole-genome sequencing project was initially proposed by Roger Dahlman, Program Manager of the US Department of Energy's (DOE) Terrestrial Carbon Program, in January 2001. The DOE had made major infrastructure and personnel investments for the Human Genome Project, including the creation of the largest public sequencing facility in the world, the Joint Genome

Institute Production Genomics Facility (JGI) in Walnut Creek, CA. Due to continuous improvements in the efficiency of their sequencing pipeline, JGI was capable of producing in excess of 1 Gb of high-quality sequence per month at the time (Detter et al. 2002), which was orders of magnitude higher than was expected at the time of its establishment. As the DOE portion of the Human Genome Project neared completion in 2000, DOE had begun to identify organisms worthy of sequencing that would be relevant to the DOE missions of research related to energy production and its consequences. The choice of *Populus* therefore made perfect sense, because there was a long history of DOE-funded research on *Populus* as a potential bioenergy crop (Dinus et al. 2001), and a surge of interest in *Populus* as a possible solution for carbon sequestration to counter anthropogenic climate change (Tuskan and Walsh 2001). Following Dahlman's initial suggestion, Stan Wullschleger and Jerry Tuskan of DOE's Oak Ridge National Laboratory led to efforts in organizing the *Populus* genetics and genomics community, leading to the formation of the International *Populus* Genome Consortium (*http://www.ornl.gov/sci/ipgc/*). Following substantial grass-roots lobbying and the production of a series of white papers, the sequencing project was approved by the DOE's Office of Science in October 2001.

The importance of community organization and international cooperation in the establishment and success of this project cannot be overestimated. The *Populus* Genome Sequencing Project was a gargantuan effort involving 40 laboratories in eight countries, with 108 researchers directly contributing to the sequencing and subsequent analysis, with particularly notable contributions from the United States, Canada, Sweden, and Belgium. The sequencing itself took place primarily at the DOE's Joint Genome Institute (JGI) under the direction of Daniel Rokhsar, and this certainly represented the bulk of the expense of the project. However, major contributions were made by cooperating institutions around the world. Some of this work was funded by the US National Science Foundation's Plant Genome Research Program, which supported efforts to create a genome portal and to enhance *Populus* bioinformatics tools. Genome Canada supported the involvement of the University of British Columbia, which contributed full-length cDNA sequences and BAC end-sequencing, and a BAC physical map to enhance assembly. The Swedish government supported efforts by the Umea Plant Sciences Center, which spearheaded the collection and analysis of approximately 350,000 EST sequences provided by laboratories around the world. Finally, the government of Belgium and the European Union supported efforts by the University of Ghent, which

customized software for gene prediction and annotation. Countless other individual researchers around the world provided in-kind contributions to the project, in the form of physical resources like template DNA, ESTs or BAC libraries, or analytical expertise related to some aspect of the genome analysis.

1.3 Tools in Place at the Time of Selection

From a technical standpoint, many of the required features for a successful genome sequencing project were already in place for *Populus* at the time it was selected. First, enough cytogenetics and genetic mapping work had been done to establish that *Populus* had a relatively tractable genome. Genetic maps were available for multiple species (Bradshaw et al. 1994b; Cervera et al. 2001; Yin et al. 2001), establishing that segregation occurs as expected in interspecific diploid crosses (with some notable exceptions: Bradshaw and Stettler 1993, 1994; Yin et al. 2004b). Furthermore, information from genetic mapping and flow cytometry indicated that the haploid genome size of the major *Populus* species was approximately 550 Mb (Bradshaw and Stettler 1993), about four times larger than *Arabidopsis*, but a fraction of the size of most conifers (Bradshaw et al. 2000). This was one of the primary factors in the choice of *Populus* as the first sequenced tree over more commercially-important species like loblolly pine (*Pinus taeda*) and Douglas-fir (*Pseudotsuga menziesii*), which had dominated tree improvement research funding prior to the genomics era. Another major factor was the phylogenetic similarity between *Populus* and *Arabidopsis*. Prior to sequencing, approximately 150,000 ESTs were available for *Populus*, and these showed high sequence conservation with known *Arabidopsis* genes (Sterky et al. 1998, 2004b), thereby facilitating provisional annotation of predicted genes based on similarity to intensively studied model annual species. Finally, several BAC libraries were available for *Populus*, and shotgun sequencing of a handful of these BACs had revealed microsynteny between *Arabidopsis* and *Populus*, and a moderate density of repeat elements in *Populus* (Stirling et al. 2003; Lescot et al. 2004). This initial sequencing suggested that assembly of a whole-genome shotgun sequence was feasible, and that existing algorithms for gene prediction in Angiosperms would function reasonably well for *Populus* (Lescot et al. 2004).

2 THE GENOME SEQUENCING PROJECT

The results of the genome sequencing project have previously been described by Tuskan and 108 co-authors in a publication in Science in 2006 (Tuskan et al. 2006), as well as in many subsequent publications by members of the International *Populus* Genome Sequencing Consortium. Unless otherwise specified, the original citation for my descriptions of this project below is the *Science* paper.

2.1 Sequencing Strategy

At the initiation of the project, there was substantial debate within the sequencing consortium about the overall approach to the project. The primary question was whether the genome should be approached as a true whole-genome shotgun project, or whether a BAC-by-BAC approach would be more prudent. The whole-genome shotgun approach, in which the genome would be randomly sheared into small pieces, which would then be sequenced separately and subsequently assembled computationally, had previously been used successfully for a number of large whole-genome sequencing projects at JGI and elsewhere (Aparicio et al. 2002; Dehal et al. 2002; see also Chapter 5 of this volume). The alternative BAC-by-BAC approach requires a physical map of tiled BACs, which is produced with restriction fragment length polymorphism (RFLP) fingerprinting, followed by shotgun sequencing of individual BACs of known position in the physical map (Marra et al. 1997; see also Chapter 2 of this volume). A similar debate had taken place within the Human Genome Sequencing Projects, with the public sequencing consortium initially attempting a BAC-by-BAC approach (Lander et al. 2001), and a private effort led by Celera and J. Craig Venter attempting a whole-genome shotgun approach (Venter et al. 2001). The primary advantage of the whole-genome shotgun approach is speed, and when it became clear that the private sequencing effort would be completed well ahead of the public effort, the public Human Genome Sequencing project was forced to partially switch to the shotgun strategy as well (Green 2001). In the case of *Populus*, it was unclear if a whole-genome shotgun would be feasible for two reasons: (1) the repeat composition and high heterozygosity of the genome could prevent coalescence of sequence contigs into coherent scaffolds; and (2) there was some evidence of a recent genome duplication in *Populus* (Bradshaw et al. 1994a), and it was feared that this would complicate the assembly.

In the end a hybrid strategy was agreed upon. The group at the

University of British Columbia and the BC Genome Sciences Center would produce a physical map by fingerprinting and end-sequencing a *Populus* BAC library, while JGI would proceed with a whole-genome shotgun for this same genotype. The end sequences were to be used to integrate the physical map with the sequence, and gaps and problematic areas of the assembled sequence were to be filled by shotgun sequencing of selected BAC clones. The whole-genome shotgun was to proceed using paired clone-end sequencing of random genomic fragments in independently-prepared libraries of three main sizes: 3 kb inserts, 8 kb inserts in standard plasmid libraries, and 40 kb inserts in phosmid libraries. The rationale for the different-sized libraries is that the bulk of the sequencing would be performed on 3 kb fragments, which would minimize cloning bias that is commonly observed in large-fragment libraries, while the larger libraries would enhance contiguity of scaffolds by bridging small repetitive regions of the genome.

2.2 Selection of Genotype for Sequencing

Another controversial decision was the selection of genotype to be sequenced. There was a strong push to select an aspen (*Populus tremula*, *P. tremuloides*, or *P. alba*), because most of the model transformation clones were derived from these species, and the majority of ESTs were also from that section of the genus. However, the cottonwoods are much more important commercially in the United States, and cottonwood hybrids were the leading candidates for high-yield plantations for bioenergy and carbon sequestration (Tuskan 1998; Perlack et al. 2005). Furthermore, most genetic maps and pedigrees were for cottonwoods (Bradshaw et al. 1994b; Cervera et al. 2001; Yin et al. 2004b, 2008), and the existing BAC libraries were also from cottonwoods (Stirling et al. 2001a; Lescot et al. 2004), so the most relevant resources for genome sequencing and assembly were already in hand. Therefore, for strategic reasons, and to accelerate the production of template for the sequencing pipeline, it was decided to sequence a black cottonwood tree (*Populus trichocarpa* Torr. & Gray). The selected genotype, clone number 383-2499, was originally collected along the Nisqually River in Washington State by Toby Bradshaw, one of the pioneers of *Populus* genomics. This clone, commonly called 'Nisqually-1', was the maternal parent for the largest pedigree produced for *Populus*, a cross with *P. deltoides* clone ILL-101 to produce a family of 2,028 F_1 progeny. The purpose of this pedigree was to isolate a major gene conferring resistance to a hybrid leaf pathogen, *Melamspora x columbiana* 3, which was segregating at an 1:1 ratio in this

pedigree (Stirling et al. 2001a). For this purpose, a 9.5× BAC library was also prepared for this pedigree by partially digesting high molecular weight genomic DNA with *Hind*III (Stirling et al. 2001a). The existence of the large pedigree and the BAC library, coupled with the availability of abundant material in clone banks at the University of Washington and elsewhere, was enough to tip the balance in favor of this genotype. There are, however, several ironies about Nisqually-1: (1) the disease resistance gene that originally piqued interest in this genotype still has not been isolated, due in part to high complexity of this genomic region, which may be linked to suppression of recombination in the large pedigree, thereby making map-based cloning nearly impossible (Stirling et al. 2001b; Yin et al. 2004a); (2) the original ortet has since been destroyed by flooding in its native habitat; and (3) Nisqually-1 has proven to be somewhat difficult to handle in tissue culture. Even though transformation protocols have been successfully developed (Ma et al. 2004; Song et al. 2006), this clone is unlikely to supplant aspen hybrids as the model of choice for functional genomics in *Populus* (Busov et al. 2005). However, there is enough sequence conservation on a genome-wide scale between the model aspens and *P. trichocarpa* that for most purposes genomic resources from one species can be used informatively for other species in the genus (Sterky et al. 2004a).

2.3 Preparation of Sequencing Template

The initial sequencing template was prepared from surface-sterilized leaves of Nisqually-1 using a CTAB-based protocol. This template was used to construct the 3 kb and 8 kb libraries that form the basis for most of the sequence data (Table 1). A second set of templates was also prepared from root tissue grown in hydroponics and tissue culture. The DNA extraction protocol involved a nucleii isolation step using a sucrose gradient followed by a cesium chloride gradient centrifugation step. This DNA was expected to be virtually free of plastid contamination, and was used to construct the fosmid libraries.

2.4 Shotgun Sequencing

A total of 7.65 million sequence reads were generated from these libraries, with 4.4 million reads coming from 3 kb libraries and 2.5 million reads from 8 kb libraries, and 650,000 reads from fosmid libraries (Table 1). In addition, 81,904 end sequences were obtained from BAC clones that averaged 100 kb in size (Kelleher et al. 2007b). This resulted in a theoretical sequence coverage of the genome of nearly 10× (i.e., an

Table 1 Description of sequencing libraries generated for the *Populus trichocarpa* genome sequencing project by JGI as of January 2004. The difference between theoretical and actual coverage of the genome is based on the cumulative length of sequence actually assembled into contigs.

Insert size (kb)	*Libraries*	*Sequences*	*Theoretical sequence coverage*	*Assembled sequence coverage*	*Clone coverage*[b]
3	4	4,427,983	5.48	3.57	13.69
8	4	2,570,799	3.18	2.14	21.20
36	3	651,211	0.81	0.62	24.17
100	1	81,904	0.11	0.10	9.49
Total	11	7,649,993	9.46	6.33	6

[a] Sequence coverage is calculated based on the total amount of sequence in each library divided by the estimated genome size of 485 Mb.
[b] Clone coverage is the total insert size of the clones (assuming the averages given in the insert size column) divided by the estimated genome size of 485 Mb.

Table 2 Kingdoms represented among unassembled sequence reads and small scaffolds from the *Populus* shotgun sequence dataset, based on WU-BLASTN searches versus the NCBI non-redundant nucleotide database.

	Unassembled reads		*Small scaffolds (<10 Kb)*	
Kingdom	*Taxa*	*Sequences*	*Taxa*	*Sequences*
Fungi	78	540	1	1
Metazoa	175	10,638	6	45
Archaea	9	54	0	0
Bacteria	291	13,656	40	231
Eukaryota	40	477	2	2
Viruses	27	407	0	0
Vector	67	1,996	5	7
Viridiplantae	723	577,511	35	2,900
Total	1,410	605,279	89	3,186

average of 10 sequences representing each nucleotide position), and an expected clone coverage of nearly 70× (i.e., the average number of clone inserts covering each position in the genome, though only the ends of the clones are represented by actual sequence). Therefore, a highly contiguous assembly was expected.

2.5 Sequence Assembly

The shotgun sequences were initially assembled based on homology and paired end read information using the JAZZ assembler (Aparicio et al. 2002). The assembly process began with identification and masking of reads derived from repetitive regions of the genome. This was accomplished by counting the number of occurrences of all 16-mer 'words' in the entire set of 7.65 million sequences, and then masking of 16-mers that occurred more than 32 times. This resulted in removal of entire reads from the assembly process and mitigated the confounding effects of repetitive DNA on shotgun assembly (Green 2001). Pairwise alignments of all sequences were then performed, and contigs were constructed by converting pairwise relationships to a graph topology and finding the most direct route through the graph. Sequence contigs were joined using similar methodology, taking advantage of linkage information contained in paired end-read information.

The initial assembly utilized 4.8 million of the sequence reads to form approximately 45,970 sequence contigs of at least 1 kb in length, resulting in approximately 427 Mb of assembled genome sequence contigs, excluding gaps. These contigs were grouped together using paired clone end information into 22,136 sequence scaffolds that covered 464 Mb of assembled sequence and 'captured' sequence gaps (the size of which was estimated based on average clone insert size). Half of this scaffold sequence was contained in 62 major scaffolds, each of which was at least 2 Mb in size. The maximum contig size was 1.7 Mb, and the maximum scaffold size was nearly 12.5 Mb.

2.6 Contamination of the Sequencing Template

Nearly 2.85 million of the original sequence reads could not be assembled into meaningful sequence contigs in the whole-genome shotgun assembly. Approximately 750,000 of these were simply low-quality or chimeric sequences that were excluded by the assembler. However, 2.1 million were high quality sequences that otherwise should have assembled. The leaf-derived sequences failed to assemble at a much higher rate (25%) than the root-derived sequences (18%), suggesting that the DNA extraction method was related to this problem. Two sets of sequences with uniform sequence depth (954× and 60× respectively) were removed and assembled into putative chloroplast and mitochondrial genomes, respectively. These accounted for approximately 300,000 of the unassembled sequences. Another 613,000 corresponded to

Populus repeat elements, as determined by high 16-mer composition and comparison to *Populus* repeat libraries using WU-BLASTN (see below). The remaining 1.1 million unassembled sequences were compared to the NCBI nonredundant nucleotide database using WU-BLASTN searches. Approximately 600,000 of these sequences showed no homology to known sequences, and are therefore of unknown origin. An additional 482,199 had significant hits to known, non-*Populus* sequences. Of these, the vast majority had hits (E <1e^{-10}) to other plants, and likely represent inexplicably unassembled portions of the *Populus* genome. However, nearly 25,000 of the remaining sequences had hits to fungi, bacteria, and viruses that were likely endophytic or pathogenic contaminants of the sequencing template, despite the fact that the leaves and roots were surface-sterilized prior to extraction. Similar trends were seen for small scaffolds from the sequencing dataset, where nearly 300 of the scaffolds <10 kb in size were apparently of microbial origin. This provides a potentially-interesting window into the invisible and largely unknown microbial associates of *Populus* (Germaine et al. 2004).

2.7 BAC Physical Map

A physical map was constructed using a 10× Nisqually-1 BAC library that had previously been constructed by the BAC Center at the Texas A&M University. The library was constructed from high molecular weight DNA extracted from leaf tissue (Stirling et al. 2001a) and partially digested with *Hind*III. Restriction enzyme fingerprinting (Marra et al. 1997) using *Hin*dIII was performed on 46,025 clones from this library with an average insert size of 100 kb, providing 9.4-fold coverage of the physical map (Kelleher et al. 2007a). This resulted in production of 3,471 contigs containing 11 BAC clones each on average. One exceptional contig consisted of over 1,200 clones, which was ultimately discovered to represent chloroplast contamination. The relative lack of contiguity in this library appears to be the result of complex haplotype structure in the *Populus* genome (Kelleher et al. 2007a). In particular, heterozygous polymorphisms at *Hind*III sites prevented haplotypes from converging in the assembly, leading to complex forking patterns in the tiling path. This is likely a problem for the sequence-based assembly as well. Another complicating factor is the apparent existence of *Hind*III 'deserts' in the *Populus* genome: large regions entirely lacking *Hind*III sites. These regions would not be represented in this BAC library, since it was constructed with a single restriction enzyme. Such complexity greatly mitigates the advantages of a BAC-by-BAC sequencing approach

in a highly heterozygous organism like *Populus*, and tips the balance strongly in favor of the more efficient whole-genome shotgun approach. Nevertheless, BAC end-sequences and the BAC physical map were extremely useful in enhancing the contiguity of the shotgun assembly, so BAC fingerprinting and mapping still plays a vital role in genome sequencing projects.

2.8 Map-Based Assembly

The large number of scaffolds in both the sequence assembly and the physical map posed substantial challenges for the analysis and application of the genome sequence. We, therefore, sought to anchor sequence contigs onto a genetic map representing the 19 *Populus* chromosomes (Yin et al. 2004b, 2008). This was accomplished by using sequence tags provided by PCR primers for 356 simple sequence repeat (SSR) markers that could be placed unambiguously in the sequence as well as on the genetic map. Location of the primer sequences in the assembled genome sequence was accomplished by performing BLASTN searches and requiring that both primers match the putative SSR locus in inverse orientation relative to each other, and at a distance that was consistent with the known size of the SSR. This resulted in linking 155 major sequence scaffolds and 335 Mb of sequence into chromosomal linkage groups (LGs) (Fig. 1). The smallest chromosome, LG_IX, was covered by two scaffolds containing 12.5 Mb of sequence. In contrast, the largest chromosome, LG_I, contained 21 scaffolds representing 35.5 Mb of sequence.

Some caveats are in order regarding the map-based assembly. First, the vast majority of the markers used for genome assembly were mapped with only 44 progeny, so the resolution of the map is quite low, and small sequence scaffolds are not always positioned or oriented accurately. This is also a problem for scaffolds that are only mapped with one marker, which is true for 75 out of 155 scaffolds. Second, the low resolution of the map can also lead to tandem assembly of scaffolds representing different haplotypes that should actually be assembled to the same position. This type of misassembly can easily be misinterpreted as large-scale tandem or segmental duplications. It is extremely difficult to distinguish between mis-assemblies of this type and true duplication events. Examples of problematic regions of the assembly are the peritelomeric (top) portions of LG_X and LG_XIX (Fig. 1). LG_XIX has been investigated in some detail, and demonstrated by intensive mapping to have very strong haplotypic divergence in this peritelomeric

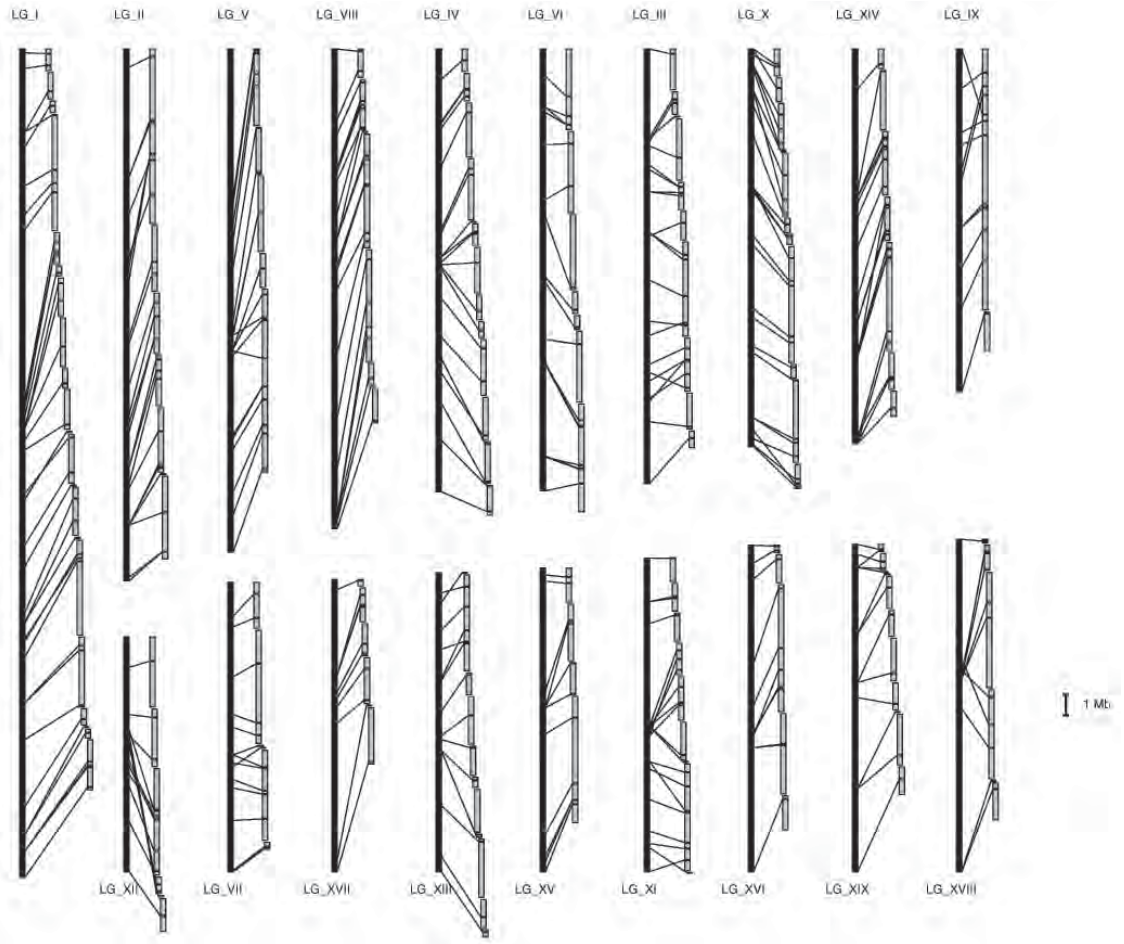

Fig. 1 Representation of a recent map-based assembly of the *Populus* genome. Black bars represent chromosomal linkage groups derived from genetic mapping, and white bars represent sequence scaffolds from the sequence-based assembly. Positions of microsatellite markers are indicated by lines connecting the linkage groups to the sequence scaffolds.

region (Yin et al. 2008). This, coupled with strong suppression of recombination and the mapping of sex determination to this region (Markussen et al. 2007; Gaudet et al. 2008), has led to the speculation that this chromosome might be in the early stages of evolving into a sex chromosome (Yin et al. 2008).

Integration of the physical and genetic maps afforded the opportunity to examine the ratio of physical to genetic distance in *Populus*. The median ratio of physical to genetic distance was 118.5 kb/centi-Morgan, based on 54 'framework' SSR markers (mapped with at least 150 progeny) located on the same sequence scaffold. There is, as expected, a substantial amount of variation in these estimates (Fig. 2), reflecting real differences in recombination frequency across the genome,

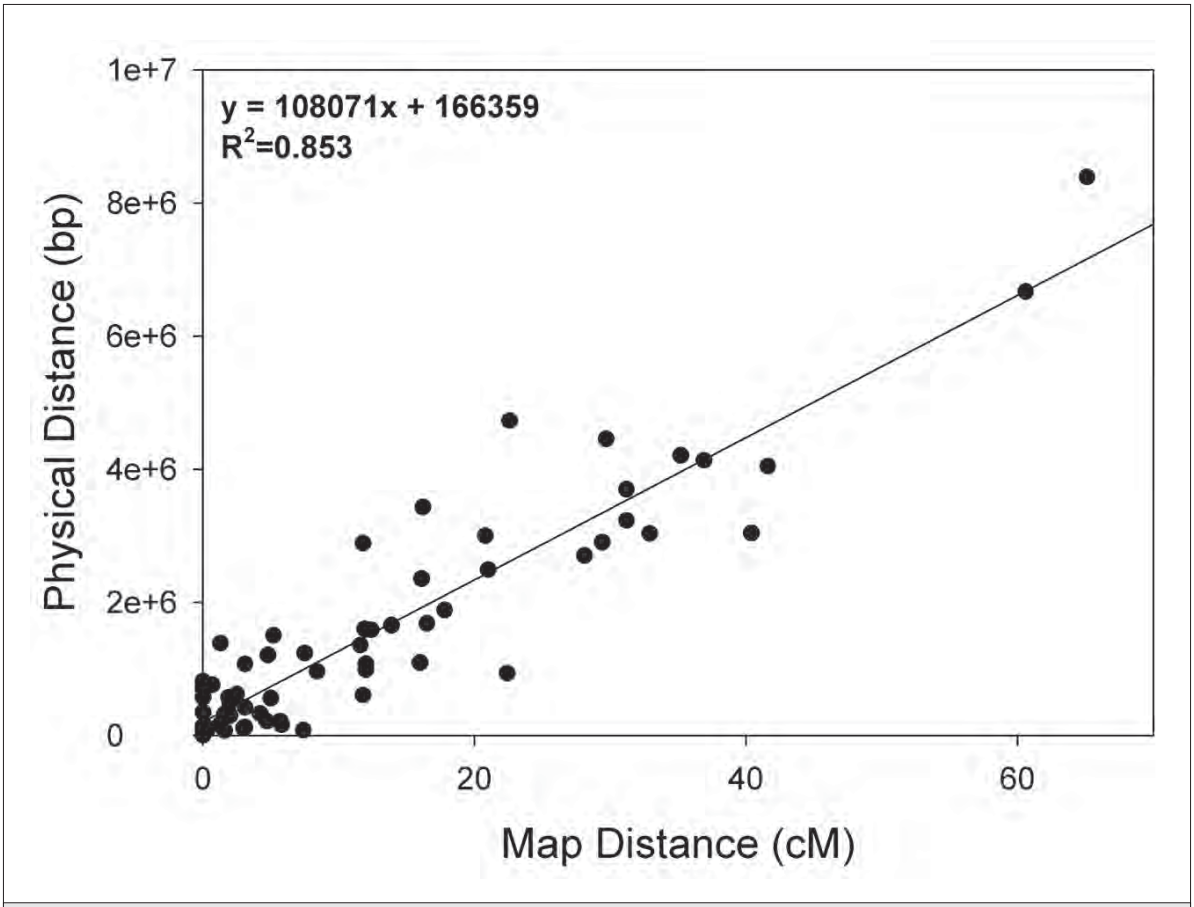

Fig. 2 Relationship between genetic distance and physical distance for *Populus* sequence scaffolds for Populus framework markers, which were mapped with over 150 progeny (Yin 2004).

as well as errors in estimation of physical genome size caused by sequence gaps, imprecision in mapping positions, and actual differences in genome composition between Nisqually-1 and the clone used in the mapping pedigree, 93-968 (Yin et al. 2004b).

3 *POPULUS* GENOME CHARACTERISTICS

3.1 Gene Content

Gene content prediction was carried out using four different approaches, and the results were merged to provide a consensus list of gene models. The four main gene-calling algorithms were *ab initio* FgenesH, homology-based FgenesH (which uses EST evidence) (Solovyev et al. 2006), Genewise, GrailExp6, and EuGène (Foissac et al. 2008), all of which were trained with a set of over 4,664 full-length cDNA sequences (Ralph et al. 2008). These gene predictions were carried out by three

independent groups (JGI, ORNL, and the University of Ghent), and then merged by the JGI to produce consensus predictions. Most gene prediction programs provide markedly different results, and each has its own strengths and weaknesses, with major tradeoffs between specificity and sensitivity, depending on the weight given to different evidence sources (ESTs, full-length cDNAs, alignments to other genomes) in the training and analysis phases (Foissac et al. 2008). As expected, the gene finding algorithms produced quite different results for *Populus* as well (Fig. 3), and it was challenging to identify the best model at each locus, and derive a consensus set of predicted genes that are likely to be true, protein coding genes. The initial consensus set of contained 58,036 putative genes, only 25% of which were predicted by two or more algorithms. This set was quickly discovered to contain many pseudogenes and transposable elements, and was gradually reduced to the publicly-released set of 45,555. However, this set still contains approximately 370 genes with strong hits to known transposable

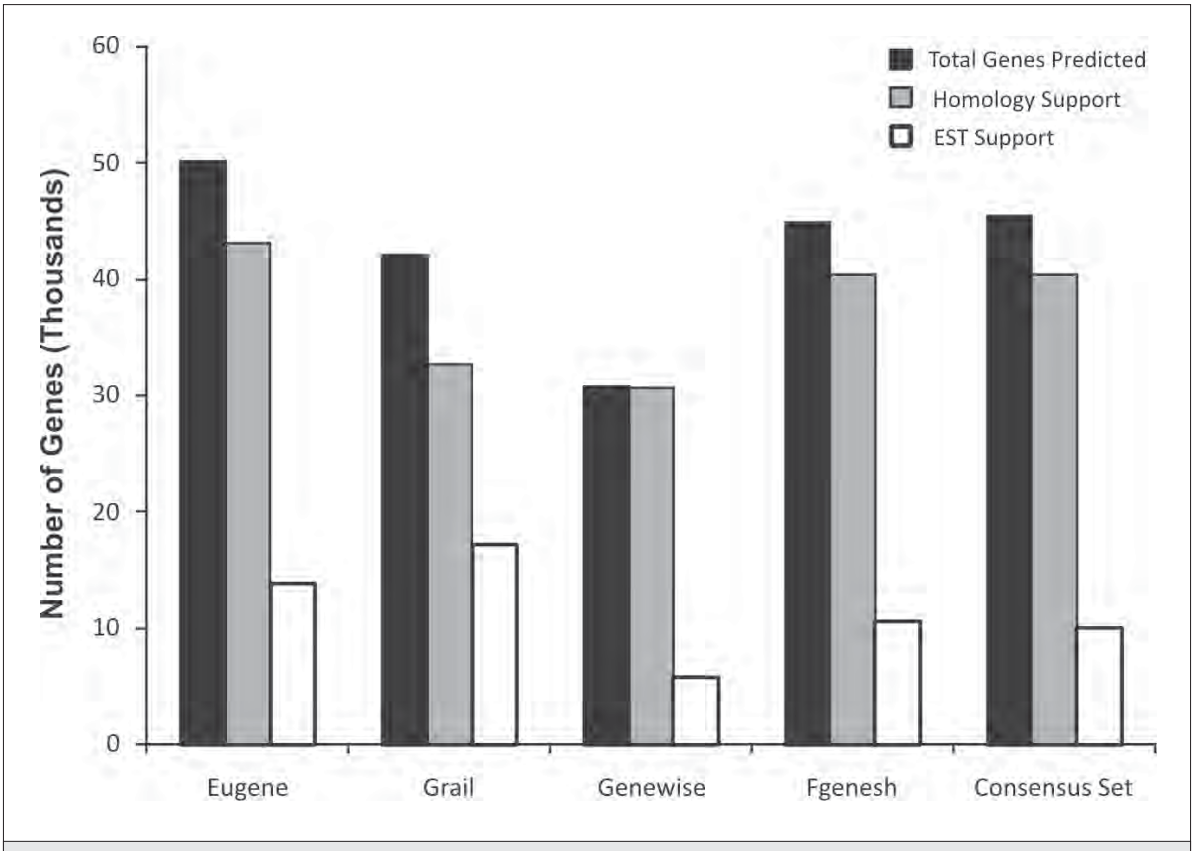

Fig. 3 Number of genes predicted by each set of gene prediction programs, as well as the consensus set of gene models released by the genome consortium in 2004. 'Homology support' indicates the number of gene models with significant BLASTN ($E < 1e^{-10}$) alignments versus know proteins in the NCBI nonredundant database, and 'EST Support' is the number of gene models with significant BLASTN alignments to one of the ~250,000 Populus ESTs that were available at the time of prediction.

elements. Furthermore, it appears that many bona fide genes from the initial set are not included in the final set of released genes. BLASTP searches revealed that 4,520 of these excluded genes have significant hits to plant proteins in the NCBI non-redundant database. Furthermore, 3,675 of these excluded genes showed some evidence of expression in whole-genome microarray experiments using diverse *Populus* tissues (Brunner et al. in prep.). Therefore, the gene content of *Populus* is still poorly determined, and efforts to improve the selection and annotation of gene models are continuing. Nevertheless, it seems clear that the final number of *Populus* genes will exceed 40,000, based on expression evidence and homology to known genes.

3.2 Comparison of Gene Content with *Arabidopsis* and Grape

In one sense, the gene content of *Populus* is surprisingly similar to that of *Arabidopsis*, despite the considerable phylogenetic distance between these taxa and their obviously contrasting biological characteristics. *Populus* is a member of the Eurosid I subclass, while *Arabidopsis* is in the Eurosid II subclass. Furthermore, *Arabidopsis* is a diminutive herbaceous annual with perfect flowers and largely selfing mating system (see also Chapter 6 of this volume), while *Populus trichocarpa* is the tallest perennial angiosperm in the northern hemisphere, with a dioecious, completely outcrossing breeding system (DeBell 1990). Nevertheless, 72% of the *Populus* genes had significant BLASTP hits to at least one *Arabidopsis* gene, with an average expectation score of 7.3×10^{-13} and an average of 59% (+/-1.6%) amino acid identity over their aligned lengths. The comparison looks even more favorable in the opposite direction, with 87.4% of *Arabidopsis* genes showing significant hits to *Populus* proteins, with an average amino acid identity of 61% (+/-1.6%) over their aligned length, with 17% of gene models having 80% or greater amino acid identity (Fig. 4). The discrepancy in the reciprocal comparisons probably reflects a higher frequency of incorrectly annotated pseudogenes in *Populus*, since the *Arabidopsis* annotation has received substantially more attention and resources than the *Populus* annotation (Swarbreck et al. 2008).

The complete genome of the wine grape (*Vitis vinifera*) was also recently published (Jaillon et al. 2007b), and this genome makes for some interesting comparisons with *Populus*. First, one might expect more similarity in coding sequences because grape is also a perennial woody plant. However, the phylogenetic position of grape relative to the rosids

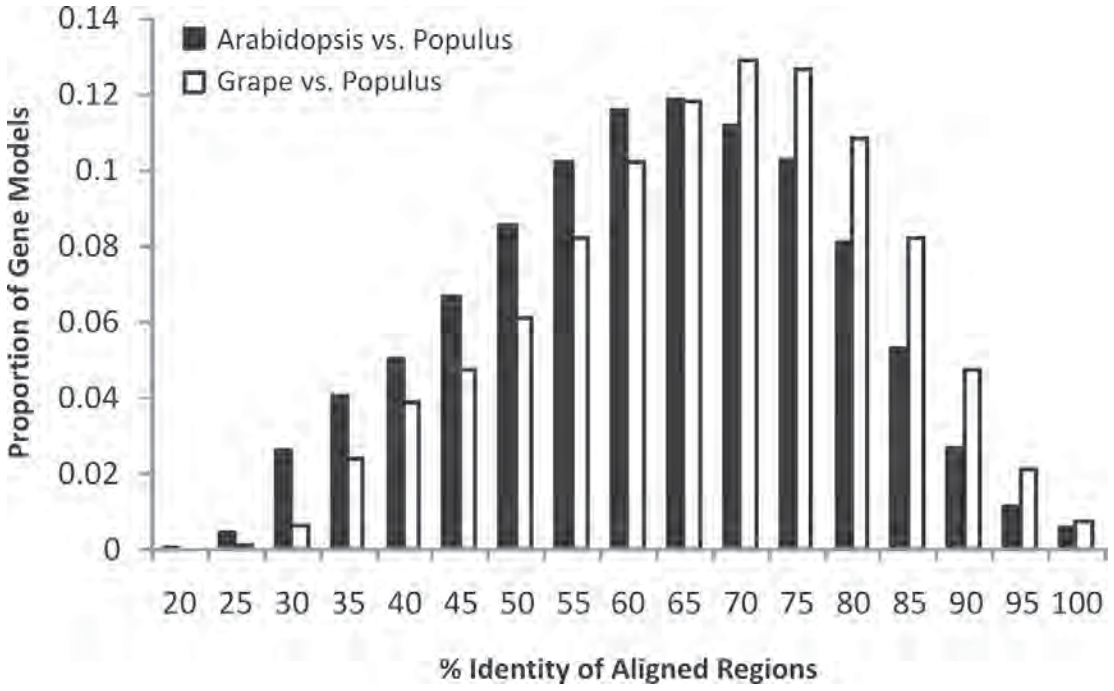

Fig. 4 Comparison of *Populus* predicted proteins to those in the *Arabidopsis* and grape genomes. Comparison only includes genes with significant BLASTP alignments (*E* score <1e-10). The % identity is the weighted average of all High Scoring Pairs from the BLASTP alignments. The *Arabidopsis* comparison is based on 26,994 significant alignments out of 30,900 proteins compared (87.4%), while the grape comparison is based on 46,288 significant alignments out of 55,990 proteins compared (82.7%).

has been the subject of some controversy, and recent analyses seem to place the order Vitales as a sister group to the Rosids (Jansen et al. 2006). Therefore, based on phylogeny one might expect grape proteins to be more divergent from *Populus* than *Arabidopsis* proteins. Surprisingly, the percentage of proteins showing significant hits to *Populus* proteins was somewhat higher in grape (84.9% of 30,442 proteins) compared to *Arabidopsis*, using the criteria described above. Furthermore, the amino acid identity was 65.1% (+/-1.49%), with 26.5% of models showing 80% or greater amino acid identity to *Populus* genes (Fig. 4). Therefore, grape genes have substantially higher identity to *Populus* genes than to *Arabidopsis* genes, despite the fact that *Arabidopsis* is closer evolutionarily (Velasco et al. 2007). This is probably due to the higher rate of evolution in *Arabidopsis* and other herbaceous annuals, which have many more generations per time interval than long-lived woody perennials like *Populus* and grape (Tuskan et al. 2006; Semon and Wolfe 2007).

The relative abundance of genes in functional categories provides a

more sensitive and informative measure than gross gene content differences between genomes. The Gene Ontology classification system provides a convenient means of doing this (Harris et al. 2006). Because GO classifications have not yet been performed for *Populus*, we assigned provisional GO classifications using *Arabidopsis* best hits, utilizing the simpler GO-Slim terms. There were significant differences between *Populus* and *Arabidopsis* for almost all GO-Slim categories, but only five categories were different between *Populus* and grape (Fig. 5). Interestingly, the classes that showed significant differences between grape and *Populus* were transcription factor activity and kinase activity, which were over-represented in *Populus*, and 'response to stress' and 'cell wall', which were over-represented in grape. Both of these latter classes were in turn strongly over-represented in *Populus* compared to *Arabidopsis*, so perhaps both of these classes are related to the woody, perennial habit. However, in the case of the cell wall class, a large portion of those over-represented in grape are related to flavonoid and polyphenol production, which is one of the aspects of grape chemistry that makes this species so desirable for wine production (Jaillon et al. 2007a; Velasco et al. 2007).

3.3 Genome Structure

Assembly of the genome and corresponding gene content to linkage groups made it possible to investigate the gross structure of the genome at a chromosomal scale. This revealed the striking existence of two whole genome duplication events. This was accomplished by making pairwise comparisons among all *Populus* genes using double-affine Smith-Waterman alignments. This revealed the presence of large syntenic blocks of genes on different linkage groups that had approximately concordant genetic distances (Fig. 6). Blocks of these syntenic genes were defined based on the existence of two or more genes aligning on different chromosomes, with fewer than 10 intervening, nonaligning genes. The genetic distance between these aligning genes was calculated based on the number of transversion substitutions at four-fold degenerate nucleotide sites (4DTV), which is a conservative estimate of genetic divergence that should be less susceptible to multiple substitutions than more commonly-occurring synonymous substitutions (Comeron 1995). Comparison of the size of the syntenic blocks versus the mean 4DTV value for those blocks revealed two clear groups of blocks that were of approximately uniform age (Fig. 6). The group of larger blocks centered at 4DTV = 0.068 represents the most recent whole

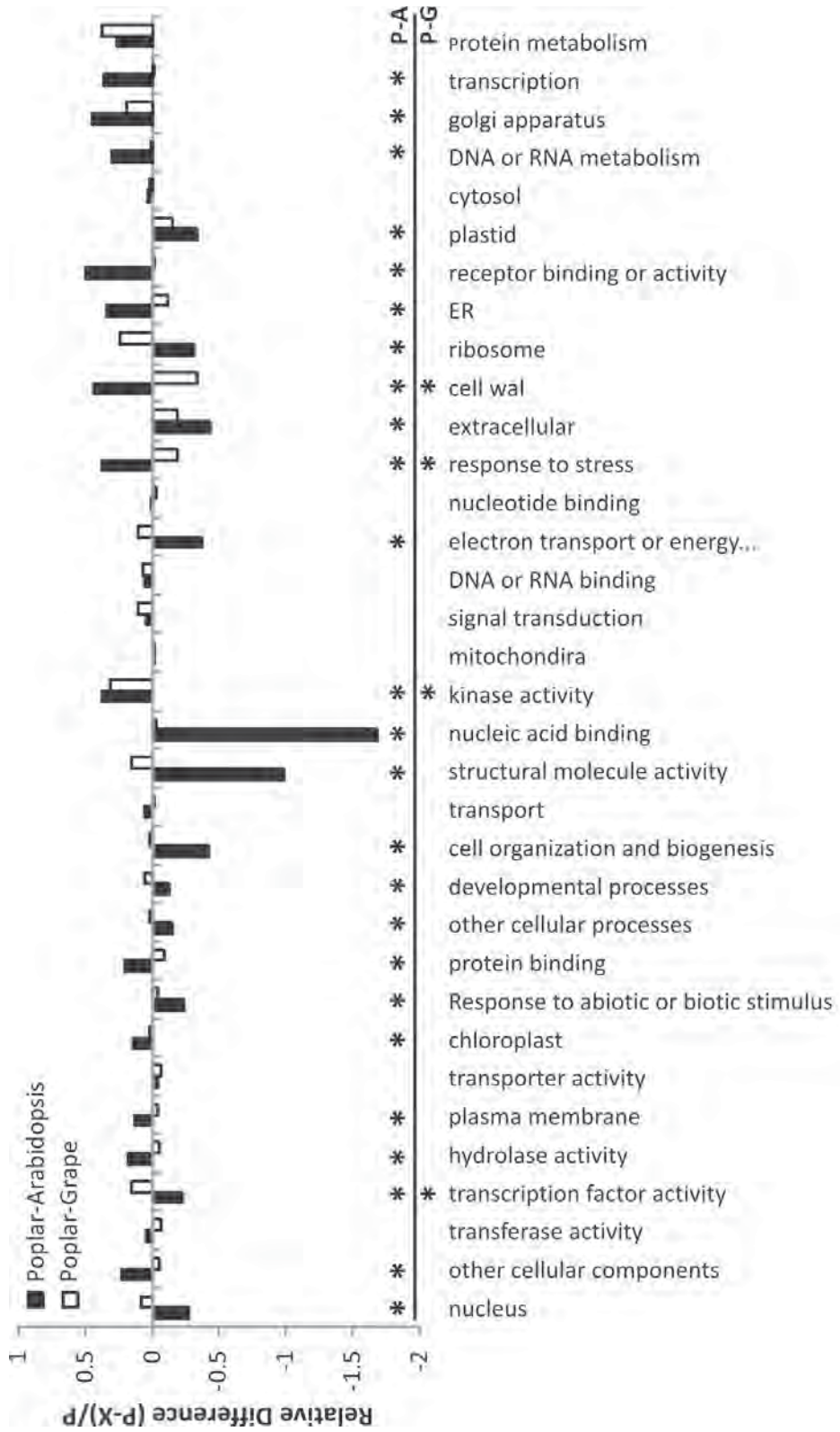

Fig. 5 Comparison of *Arabidopsis* and grape gene content to that of *Arabidopsis*, organized by GOslim categories. The relative difference in proportions of genes in each category is calculated as (P-X)/P, where P is the proportion of *Populus* genes in that category, and X is the proportion of *Arabidopsis* or grape genes in that category. Asterisks indicate significant deviations ($p < 0.001$) from expected proportions, based on a Chi-square test, with *Populus-Arabidopsis* comparisons on top, and *Arabidopsis*-grape comparisons on the bottom. GOslim groups are arranged from the most abundant (left) to the least abundant class in *Arabidopsis*.

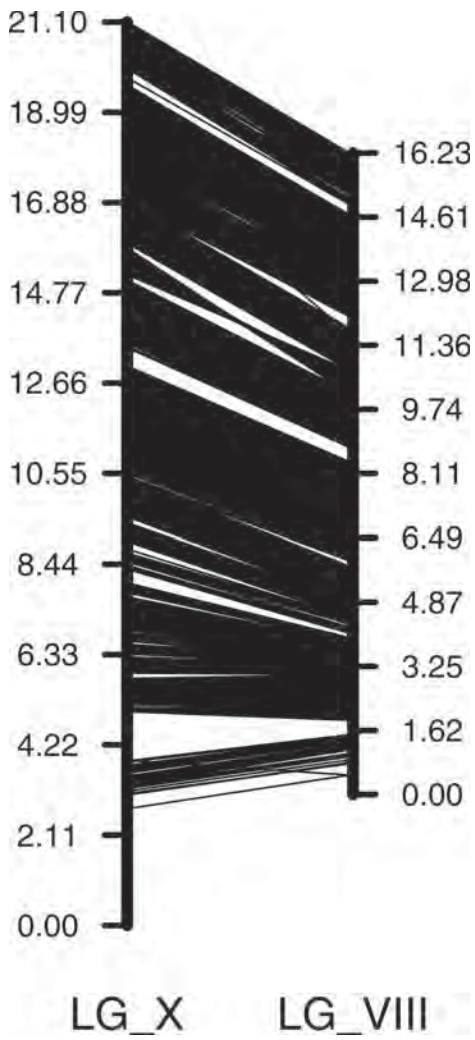

Fig. 6 Comparison of genes with significant alignments between two linkage groups, chromosome X (left) and Chromosome VIII (right). Genes are represented by black lines connecting the heavy lines representing the chromosomes. Position on the linkage group is given in megabases to the side of each group. Genetic distances between genes in these large syntenic blocks are highly concordant, indicating that these syntenic chromosome blocks originated from whole genome duplication events.

genome duplication in *Populus*, while the group of smaller blocks at 4DTV = 0.31 represents a more ancient duplication event (Sterck et al. 2005). Dating of these events is difficult, because the *Populus* genome is evolving considerably slower than genomes that have previously been used to calibrate the angiosperm molecular clock. Using the molecular clock as calibrated by fossil records for the Brassicaceae, for example, the most recent duplication dates to 8 million years ago (Sterck et al. 2005). However, the *Populus* genus has been in existence for at least 50 million years (Eckenwalder 1996), and the genome duplication is shared by

many species in the genus (Sterck et al. 2005), so the *Populus* genome is clearly evolving at a much slower rate than herbaceous angiosperms, which is to be expected based on generation time (Bell et al. 2005).

The *Arabidopsis* genome also shows evidence of at least two whole-genome duplication events (Blanc et al. 2003; see also Chapter 6 of this volume), but following these events the genome has become substantially rearranged, making it difficult to reconstruct the older events (Blanc et al. 2003). A similar rearrangement has occurred in *Populus*, but much less severe (see figure in Tuskan et al. 2006). Extensive rearrangements following genome doubling is a common component of the diploidization process (Adams and Wendel 2005; Semon and Wolfe 2007). The structural complexity of these two genomes, coupled with the high rates of gene evolution in *Arabidopsis*, make it particularly difficult to establish orthology and determine whether the ancient duplication event in *Populus* is shared with *Arabidopsis*. The timing of the event is similar to the timing of the split of the *Arabidopsis* and *Populus* lineages, as determined by pairwise comparisons of genetic distances between *Populus* duplicated genes, *Arabidopsis* duplicated genes, and between putative *Arabidopsis* and *Populus* orthologs. Given the close timing of these events, it is tempting to speculate that the genome duplication was a primary driver of the diversification of the rosids (Lynch and Conery 2003).

In contrast to *Populus* and *Arabidopsis*, the grape genome has been comparatively quiescent, with minimal rearrangements, and equivocal evidence of a single whole-genome duplication that could be shared with *Populus* and *Arabidopsis* (Jaillon et al. 2007b; Velasco et al. 2007). This structural simplicity has allowed reconstruction of the truly ancient whole genome duplication event that is shared by all angiosperms. It appears that this event resulted in hexaploidy in the ancient angiosperm progenitor, as suggested by the presence of three syntenic blocks in rice, *Populus*, and *Arabidopsis* for every one block in grape (Jaillon et al. 2007a). However, evidence for this event is still weak, because the genetic distance is too great to allow relative dating with nucleotide substitution rates, and it is possible to confound two different duplication events that occurred at very different times, followed by massive rearrangement and gene loss. This same analysis (Jaillon et al. 2007a) suggested that only one duplication event had occurred in the *Populus* genome, despite the existence of virtually unequivocal evidence for two events when relative levels of divergence are taken into account (Fig. 7). Part of the problem is the confusion about the proper

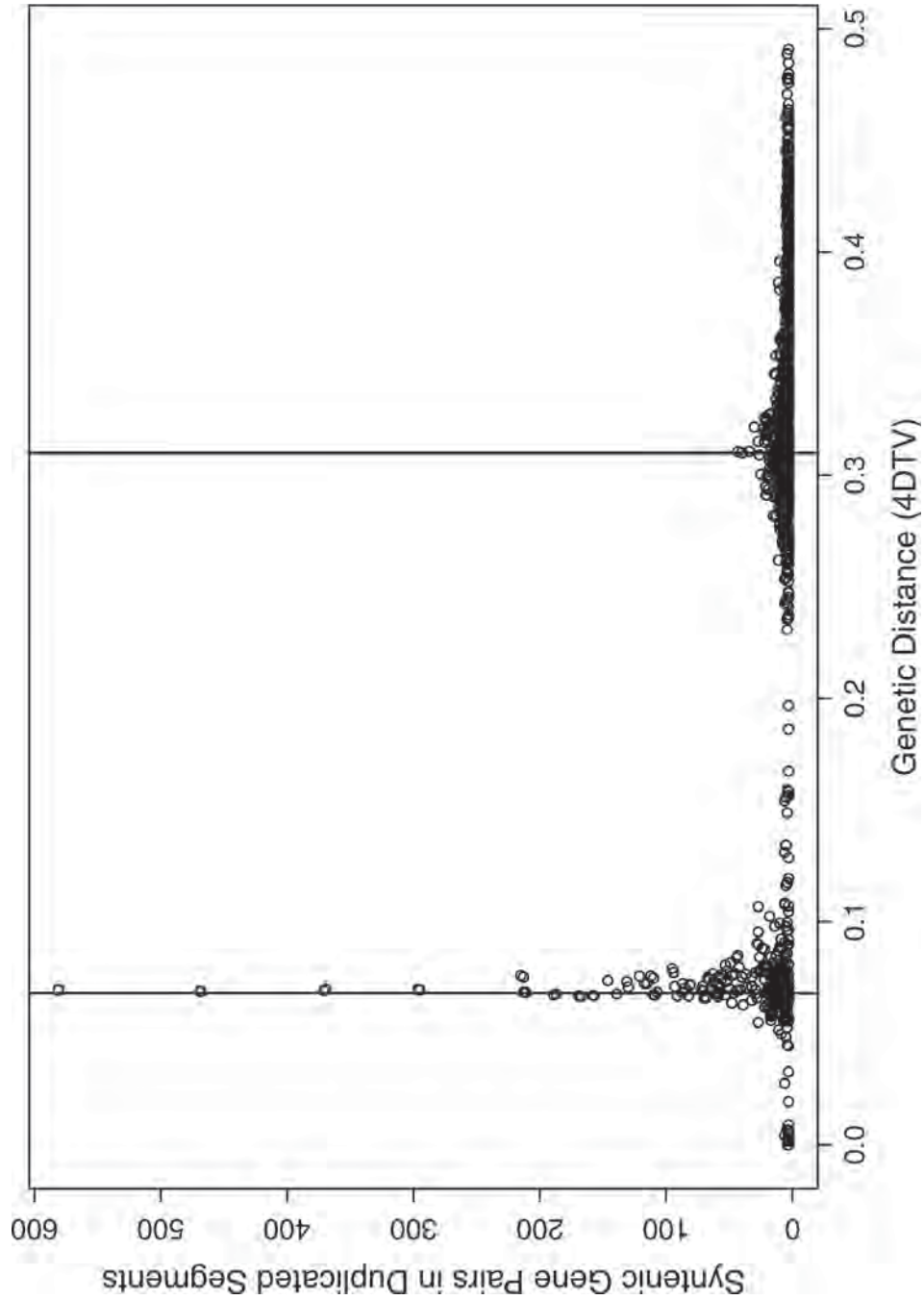

Fig. 7 Observed number of genes aligned in approximately syntenic blocks between chromosomes versus the genetic distance between genes, as measured by the rate of transversions at four-fold degenerate nucleotide sites (4DTV). Peaks corresponding to the putative Salicoid and Eurosid whole-genome duplication events are indicated as peaks at the centers of clusters of similar 4DTV values, indicating the existence of large numbers of duplicated blocks of approximately the same age.

relationship of grape to the Rosids. If grape is taken as on outgroup, then a clear model that incorporates both *Populus* duplications can be accommodated (Velasco et al. 2007). However, if grape is assumed to be closer to *Populus* than *Arabidopsis*, as sequence similarity suggests, then grape would have to share the duplication event that occurred near the time of the split with *Populus*. However, this event was not detected in grape or *Populus*, based on an analysis using reciprocal BLASTP hits, without regard to degree of divergence of putative orthologs (Jaillon et al. 2007b). Fortunately, many more plant genomes are currently in sequencing pipelines, so the duplication history of the angiosperms will soon become much clearer.

4 IMPLICATIONS AND APPLICATIONS OF THE GENOME SEQUENCE

The genome sequence was truly a watershed event for the tree genetics community, and the impacts have reverberated throughout forest science and even into other parts of plant science. The genome sequence has provided a nearly complete catalog of all genes and regulatory elements in this model tree, thus opening up a whole realm of research that was not possible before the sequencing project. One index of the impact of the sequence is the number of citations of journal articles related to *Populus* has more than doubled since 2004, the year that the sequence was first publicly released, and the number of *Populus* publications has nearly doubled since 2000. The main article describing the genome sequence (Tuskan et al. 2006) has been cited over 250 times since it was published in September, 2006.

The genome sequence has already had extensive applications in applied science. For example, *Populus* is currently the focus of three major bioenergy projects, two in the US and one in Canada, with a total committed funding of more than $20 million over the next few years. One of these projects, the DOE Bioenergy Science Center headed by Oak Ridge National Lab, is resequencing 18 *Populus* genotypes using next-generation sequencing technology. The project will characterize single nucleotide polymorphisms (SNPs) across the genome for the purpose of genetic association studies to identify genes underlying cell wall biosynthesis, with the ultimate goal of reducing the recalcitrance of lignocellulosic feedstocks to cellulose extraction (Rubin 2008). The project will then use Illumina Bead Arrays to assay 20,000 SNPs for over 1,000 trees collected across the range of *Populus trichocarpa*. These trees will be established in three different common gardens encompassing

most of the range of the species (California, Oregon, and British Columbia), and phenotyping will be performed for a large number of traits. This project, and others like it, will therefore propel *Populus* from the realm of comparative genomics, and almost complete reliance on gene homology to herbaceous models for functional annotation, to direct functional characterization of a large fraction of the genes in the genome. *Populus* will thus be solidly established as a premier model organism for functional genomics.

The impact in areas outside of genomics has been equally profound. The fields of community genetics and ecological genomics are flourishing, with *Populus* as one of the primary model organisms, driven by the availability of the genome sequence and the central importance of *Populus* in many ecosystems (Whitham et al. 2006, 2008). Multiple large-scale ecological genomics studies have been funded in *Populus* since the publication of the genome sequence, including an NSF-FIBR project (Whitham et al.), two Plant Genome Research Program projects (Buerkle et al. and Olson et al.), and several large projects in Europe (Taylor et al.; Lexer et al.). The genome sequence is allowing exploration of diverse questions, such as exploration of the genetic architecture of species barriers, based on patterns of introgression across hybrid zones (Lexer et al. 2007), or the genetic basis of sexual selection (Yin et al. 2008). Furthermore, additional species associated with *Populus* have also been sequenced (Martin et al. 2004, 2008), and many more are in progress. We are truly on the threshold of a brave new era in which genome sequencing of entire communities will become entirely plausible, potentially allowing elucidation of fundamental truths about the mechanisms of the assemblage and persistence of ecological communities (Whitham et al. 2008). This is likely to fundamentally change the way we approach ecological and evolutionary research.

Acknowledgements

The *Populus* Genome consortium, led by Jerry Tuskan, Dan Rokhsar, Carl Douglas, Stefan Jansson, Goran Sandberg, and Yves Van de Peer, made all of the work reported in this chapter possible. In particular, I have directly co-opted analyses performed by Uffe Hellsten, Nik Putnam, and Igor Grigoriev. This work was supported by the US Department of Energy, Office of Science, Biological and Environmental Research, NSF-FIBR, and the NSF Plant Genome Research Program.

References

Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8: 135-141

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

Barnes BV (1969) Indications of possible mid-Cenozoic hybridization in aspens of the Columbia Plateau. Rhodora 69: 70-81

Bell CD, Soltis DE, Soltis PS (2005) The age of the angiosperms: A molecular timescale without a clock. Evolution 59: 1245-1258

Bhalerao R, Nilsson O, Sandberg G (2003) Out of the woods: forest biotechnology enters the genomic era. Curr Opin Biotechnol 14: 206-213

Blackburn KB, Heslop Harrison JW (1924) A preliminary account of the chromosomes and chromosome behaviour in the Salicaceae. Ann Bot 38: 361-378

Blanc G, Hokamp K, Wolfe KH (2003) A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. Genom Res 13: 137-144

Braatne JH, Rood SB, Heilman PE (1996) Life history, ecology, and reproduction of riparian cottonwoods in North America. In: Stettler RF, Bradshaw HD Jr., Heilman PE, Hinckley TM (eds) Biology of Populus and its Implications for Management and Conservation. NRC Res Press, Ottawa, Canada, pp. 57-85

Bradshaw HD, Jr, Stettler RF (1993) Molecular genetics of growth and development in Populus. I. Triploidy in hybrid poplars. Theor Appl Genet 86: 301-307

Bradshaw HD, Stettler RF (1994) Molecular genetics of growth and development in Populus. II. Segregation distortion due to genetic load. Theor Appl Genet 89: 551-558

Bradshaw HD, Villar M, Watson BD, Otto KG, Stewart S, Stettler RF (1994a) 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

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

Bradshaw HD, Ceulemans RE, Davis J, Stettler RF (2000) Emerging model

systems in plant biology: poplar (*Populus*) as a model forest tree. J Plant Growth Reg 19: 306-313

Busov VB, Brunner AM, Meilan R, Filichkin S, Ganio L, Gandhi S, Strauss SH (2005) Genetic transformation: a powerful tool for dissection of adaptive traits in trees. New Phytol 167: 9-18

Cervera MT, Storme V, Ivens B, Gusmao J, Liu BH, Hostyn V, Slycken JV, Montagu MV, Boerjan W (2001) 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

Comeron JM (1995) A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. J Mol Evol 41: 1152-1159

DeBell DS (1990) *Populus trichocarpa* Torr. & Gray, Black Cottonwood. In: Burns RM, Honkala BH (eds) Silvics of North America. Vol 2. Hardwoods. USDA For Serv Agri Handbook USDA For Serv, Washington DC, USA, pp. 570-576

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

Detter JC, Jett JM, Lucas SM, Dalin E, Arellano AR, Wang M, Nelson JR, Chapman J, Lou YI, Rokhsar D, Hawkins TL, Richardson PM (2002) Isothermal strand-displacement amplification applications for high-throughput genomics. Genomics 80: 691-698

Dinus RJ, Payne P, Sewell MM, 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

Eckenwalder JE (1984) Natural intersectional hybridization between North American species of *Populus* (*Salicaceae*) in sections *Aigeiros* and *Tacamahaca*. II. Taxonomy. Can J Bot 62: 336-342

Eckenwalder JE (1996) Systematics and evolution of *Populus*. In: Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM (eds) Biology of Populus and its Implications for Management and Conservation. NRC Res Press, Ottawa, Canada, pp. 7-32

Einspahr D, Benson MK, Peckham JR (1963) Natural variation and heritability in triploid aspen. Silvae Genet 12: 51-58

Foissac S, Gouzy J, Rombauts S, Mathe C, Amselem J, Sterck L, Van de Peer Y, Rouze P, Schiex T (2008) Genome annotation in plants and fungi: EuGene as a model platform. Curr Bioinform 3: 87-97

Gaudet M, Jorge V, Paolucci I, Beritognolo I, Mugnozza GS, Sabatti M (2008) Genetic linkage maps of *Populus nigra* L. including AFLPs, SSRs, SNPs, and sex trait. Tree Genet Genom 4: 25-36

Germaine K, Keogh E, Garcia-Cabellos G, Borremans B, van der Lelie D, Barac T, Oeyen L, Vangronsveld J, Moore FP, Moore ERB, Campbell CD, Ryan D, Dowling DN (2004) Colonisation of poplar trees by gfp expressing bacterial endophytes. Fems Microbiol Ecology 48: 109-118

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

Harris MA, Clark JI, Ireland A, Lomax J, Ashburner M, Collins R, Eilbeck K, Lewis S, Mungall C, Richter J, Rubin GM, Shu SQ, Blake JA, Bult CJ, Diehl AD, Dolan ME, Drabkin HJ, Eppig JT, Hill DP, Ni L, Ringwald M, Balakrishnan R, Binkley G, Cherry JM, Christie KR, Costanzo MC, Dong Q, Engel SR, Fisk DG, Hirschman JE, Hitz BC, Hong EL, Lane C, Miyasato S, Nash R, Sethuraman A, Skrzypek M, Theesfeld CL, Weng SA, Botstein D, Dolinski K, Oughtred R, Berardini T, Mundodi S, Rhee SY, Apweiler R, Barrell D, Camon E, Dimmer E, Mulder N, Chisholm R, Fey P, Gaudet P, Kibbe W, Pilcher K, Bastiani CA, Kishore R, Schwarz EM, Sternberg P, Van Auken K, Gwinn M, Hannick L, Wortman J, Aslett M, Berriman M, Wood V, Bromberg S, Foote C, Jacob H, Pasko D, Petri V, Reilly D, Seiler K, Shimoyama M, Smith J, Twigger S, Jaiswal P, Seigfried T, Collmer C, Howe D, Westerfield M (2006) The Gene Ontology (GO) project in 2006. Nucl Acids Res 34: D322-D326

Heilman PE (1999) Planted forests: poplars. New for 17/18: 89-93

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, Vezzi A, Legeai F, Hugueney P, Dasilva C, Horner D, Mica E, Jublot D, Poulain J, Bruyere C, Billault A, Segurens B, Gouyvenoux M, Ugarte E, Cattonaro F, Anthouard V, Vico V, Del Fabbro C, Alaux M, Di Gaspero G, Dumas V, Felice N, Paillard S, Juman I, Moroldo M, Scalabrin S, Canaguier A, Le Clainche I, Malacrida G, Durand E, Pesole G, Laucou V, Chatelet P, Merdinoglu D, Delledonne M, Pezzotti M, Lecharny A, Scarpelli C, Artiguenave F, Pe ME, Valle G, Morgante M, Caboche M, Adam-Blondon AF, Weissenbach J, Quetier F, Wincker P (2007a) The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature 449: 463-467.

Jansen RK, Kaittanis C, Saski C, Lee SB, Tomkins J, Alverson AJ, Daniell H (2006) Phylogenetic analyses of Vitis (Vitaceae) based on complete chloroplast genome sequences: effects of taxon sampling and phylogenetic methods on resolving relationships among rosids. BMC Evol Biol 6:

Karrenberg S, Edwards PJ, Kollmann J (2002) The life history of Salicaceae living in the active zone of floodplains. Freshwater Biol 47: 733-748

Kelleher CT, Chiu R, Shin H, Bosdet IE, Krzywinski MI, Fjell CD, Wilkin J, Yin TM, DiFazio SP, Ali J, Asano JK, Chan S, Cloutier A, Girn N, Leach S, Lee D, Mathewson CA, Olson T, O'Connor K, Prabhu AL, Smailus DE, Stott JM, Tsai M, Wye NH, Yang GS, Zhuang J, Holt RA, Putnam NH, Vrebalov J, Giovannoni JJ, Grimwood J, Schmutz J, Rokhsar D, Jones SJM, Marra MA, Tuskan GA, Bohlmann J, Ellis BE, Ritland K, Douglas CJ, Schein JE (2007a) A physical map of the highly heterozygous Populus genome: integration with the genome sequence and genetic map and analysis of haplotype variation. Plant J 50: 1063-1078

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang HM, Yu J, Wang J, Huang GY, Gu J, Hood L, Rowen L, Madan A, Qin SZ, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Raymond C, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan HQ, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JGR, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang WH, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz JR, Slater

G, Smit AFA, Stupka E, Szustakowki J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ (2001) Initial sequencing and analysis of the human genome. Nature 409: 860-921

Lescot M, Rombauts S, Zhang J, Aubourg S, Mathe C, Jansson S, Rouze P, Boerjan W (2004) Annotation of a 95-kb Populus deltoides genomic sequence reveals a disease resistance gene cluster and novel class I and class II transposable elements. Theor Appl Genet 109: 10-22

Lexer C, Heinze B, Alia R, Rieseberg LH (2004) Hybrid zones as a tool for identifying adaptive genetic variation in outbreeding forest trees: lessons from wild annual sunflowers (Helianthus spp.). For Ecol Manag 197: 49-64

Lexer C, Buerkle CA, Joseph JA, Heinze B, Fay MF (2007) Admixture in European Populus hybrid zones makes feasible the mapping of loci that contribute to reproductive isolation and trait differences. Heredity 98: 74-84

Lynch M, Conery JS (2003) The evolutionary demography of duplicate genes. J Struct Funct Genom 3: 35-44

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

Markussen T, Pakull B, 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 Genet 56: 180-184

Marra MA, Kucaba TA, Dietrich NL, Green ED, Brownstein B, Wilson RK, McDonald KM, Hillier LW, McPherson JD, Waterston RH (1997) High throughput fingerprint analysis of large-insert clones. Genom Res 7: 1072-1084

Martin F, Tuskan GA, DiFazio SP, Lammers P, Newcombe G, Podila GK (2004) Symbiotic sequencing for the *Populus mesocosm*. New Phytol 161: 330-335i

Martin F, Aerts A, Ahren D, Brun A, Danchin EGJ, Duchaussoy F, Gibon J, Kohler A, Lindquist E, Pereda V, Salamov A, Shapiro HJ, Wuyts J, Blaudez D, Buee M, Brokstein P, Canback B, Cohen D, Courty PE, Coutinho PM, Delaruelle C, Detter JC, Deveau A, DiFazio S, Duplessis S, Fraissinet-Tachet L, Lucic E, Frey-Klett P, Fourrey C, Feussner I, Gay G, Grimwood J, Hoegger PJ, Jain P, Kilaru S, Labbe J, Lin YC, Legue V, Le Tacon F, Marmeisse R, Melayah D, Montanini B, Muratet M, Nehls U, Niculita-Hirzel H, Oudot-Le Secq MP, Peter M, Quesneville H, Rajashekar B, Reich M, Rouhier N, Schmutz J, Yin T, Chalot M, Henrissat B, Kues U, Lucas S, de Peer YV, Podila GK, Polle A, Pukkila PJ, Richardson PM, Rouze P, Sanders IR, Stajich JE, Tunlid A, Tuskan G, Grigoriev IV (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452: 88-92

Perlack RD, Wright LL, Turhollow AF, Graham RL, Stokes BJ, Erbach DC (2005)

Biomass as Feedstock for a Bioenergy and Bioproducts Industry: The Technical Feasibility of a Billion-Ton Annual Supply. Oak Ridge Natl Lab, Oak Ridge, TN, USA

Peterson EB, Peterson NM (1992) Ecology, Management, and Use of Aspen and Balsam Poplar in the Prairie Provinces, Canada. Forestry Canada, Victoria, British Columbia, Canada.

Ralph SG, Chun HJE, Cooper D, Kirkpatrick R, Kolosova N, Gunter L, Tuskan GA, Douglas CJ, Holt RA, Jones SJM, Marra MA, Bohlmann J (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 Genom 9:

Romme WH, Floyd-Hanna L, Hanna DD, Bartlett E (2001) Aspen's ecological role in the West. In: Shepperd WD, Binkley D, Bartos DL, Stohlgren TJ, Eskew LG (eds) Sustaining Aspen in Western Landscapes: Symposium Proceedings; 13-15 June 2000. RMRS-P-18., US Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, USA, pp. 243-259.

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

Semon M, Wolfe KH (2007) Consequences of genome duplication. Curr Opin Genet Dev 17: 505-512

Solovyev V, Kosarev P, Seledsov I, Vorobyev D (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genom Biol 7:

Song JY, Lu SF, Chen ZZ, 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

Sterck L, Rombauts S, Jansson S, Sterky F, Rouze P, Van de Peer Y (2005) EST data suggest that poplar is an ancient polyploid. New Phytol 167: 165-170

Sterky F, Regan S, Karlsson J, Hertzberg M, Rohde A, Holmberg A, Amini B, Bhalerao R, Larsson M, 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

Sterky F, Bhalerao RR, Unneberg P, Segerman B, Nilsson P, Brunner AM, Charbonnel-Campaa L, Lindvall JJ, Tandre K, Strauss SH, Sundberg B, Gustafsson P, Uhlen M, Bhalerao RP, Nilsson O, Sandberg G, Karlsson J, Lundeberg J, Jansson S (2004b) A Populus EST resource for plant functional genomics. Proc Natl Acad Sci USA 101: 13951-13956

Stettler RF, Koster R, Steenackers V (1980) Interspecific crossability studies in poplars. Theor Appl Genet 58: 273-282

Stettler RF, Zsuffa L, Wu R (1996) The role of hybridization in the genetic manipulation of Populus. In: Stettler RF, Bradshaw HD, Heilman PE, Hinckley TM (eds) Biology of Populus and its Implications for Management and Conservation. NRC Research Press, Ottowa, Canada, pp. 87-112

Stirling B, Newcombe G, Vrebalov J, Bosdet I, Bradshaw HD (2001a) Suppressed

recombination around the MXC3 locus, a major gene for resistance to poplar leaf rust. Theor Appl Genet 103: 1129-1137

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

Stirling B, Yang ZK, Gunter LE, Tuskan GA, Bradshaw HD (2003) Comparative sequence analysis between orthologous regions of the Arabidopsis and Populus genomes reveals substantial synteny and microcollinearity. Can J For Res 33: 2245-2251

Strauss SH, Rottmann WH, Brunner AM, Sheppard LA (1995) Genetic engineering of reproductive sterility in forest trees. Mol Breed 1: 5-26

Swarbreck D, Wilks C, Lamesch P, Berardini TZ, Garcia-Hernandez M, Foerster H, Li D, Meyer T, Muller R, Ploetz L, Radenbaugh A, Singh S, Swing V, Tissier C, Zhang P, Huala E (2008) The Arabidopsis Information Resource (TAIR): gene structure and function annotation. Nucl Acids Res 36: D1009-D1014

Taylor G (2002) Populus: Arabidopsis for forestry. Do we need a model tree? Ann Bot 90: 681-689

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, Walsh ME (2001) Short-rotation woody crop systems, atmospheric carbon dioxide and carbon management: A US case study. For Chron 77: 259-264

Tuskan GA, 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 RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen GL, Cooper D, Coutinho PM, 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 JC, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, 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 CJ, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313: 1596-1604

Velasco R, Zharkikh A, Troggio M, Cartwright DA, Cestaro A, Pruss D, Pindo

M, FitzGerald LM, Vezzulli S, Reid J, Malacarne G, Iliev D, Coppola G, Wardell B, Micheletti D, Macalma T, Facci M, Mitchell JT, Perazzolli M, Eldredge G, Gatto P, Oyzerski R, Moretto M, Gutin N, Stefanini M, Chen Y, Segala C, Davenport C, Dematt+¿ L, Mraz A, Battilana J, Stormo K, Costa F, Tao Q, Si-Ammour A, Harkins T, Lackey A, Perbost C, Taillon B, Stella A, Solovyev V, Fawcett JA, Sterck L, Vandepoele K, Grando SM, Toppo S, Moser C, Lanchbury J, Bogden R, Skolnick M, Sgaramella V, Bhatnagar SK, Fontana P, Gutin A, Van de Peer Y, Salamini F, Viola R (2007) A high quality draft consensus sequence of the genome of a heterozygous grapevine variety. PLoS ONE 2: e1326

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di F, V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigo R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D,

Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M (2001) The sequence of the human genome. Science 291: 1304-1351

Whitham TG, Bailey JK, Schweitzer JA, Shuster SM, Bangert RK, LeRoy CJ, Lonsdorf EV, Allan GJ, DiFazio SP, Potts BM, Fischer DG, Gehring CA, Lindroth RL, Marks JC, Hart SC, Wimp GM, Wooley SC (2006) A framework for community and ecosystem genetics: from genes to ecosystems. Nat Rev Genet 7: 510-523

Whitham TG, DiFazio SP, Schweitzer JA, Shuster SM, Allan GJ, Bailey JK, Woolbright SA (2008) Perspective—Extending genomics to natural communities and ecosystems. Science 320: 492-495

Wullschleger SD, Janssson S, Taylor G (2002) Genomics and forest biology: *Populus* Emerges as the Perennial Favortie. Plant Cell 14: 2651-2655

Yin TM, Huang MR, Wang MX, Zhu LH, Zeng ZB, Wu RL (2001) Preliminary interspecific genetic maps of the Populus genome constructed from RAPD markers. Genome 44: 602-609

Yin TM, DiFazio SP, Gunter LE, Jawdy SS, Boerjan W, Tuskan GA (2004a) Genetic and physical mapping of Melampsora rust resistance genes in Populus and characterization of linkage disequilibrium and flanking genomic sequence. New Phytol 164: 95-105

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

Yin TM, DiFazio SP, Gunter LE, Zhang X, Sewell MM, Woolbright SA, Allan GJ, Kelleher CT, Douglas CJ, Wang M, Tuskan GA (2008) Genome structure and emerging evidence of an incipient sex chromosome in Populus. Genom Res 18: 422-430

Zsuffa L, Giordano E, Pryor LD, Stettler RF (1996) Trends in poplar culture: som global and regional perspectives. In: Stettler RF, Bradshaw HD Jr, Heilman PE, Hinckley TM (eds) Biology of Populus and its implications for management and conservation. NRC Res Press, Ottawa, Canada, pp. 515-539