www.sciencemag.org/cgi/content/full/313/5793/1596/DC1

Supporting Online Material for

The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray)

G. A. Tuskan,* S. DiFazio, S. Jansson, J. Bohlmann, I. Grigoriev, U. Hellsten, N. Putnam, S. Ralph, S. Rombauts, A. Salamov, J. Schein, L. Sterck, A. Aerts, R. R. Bhalerao, R. P. Bhalerao, D. Blaudez, W. Boerjan, A. Brun, A. Brunner, V. Busov, M. Campbell, J. Carlson, M. Chalot, J. Chapman, G.-L. Chen, D. Cooper, P. M. Coutinho, J. Couturier, S. Covert, Q. Cronk, R. Cunningham, J. Davis, S. Degroeve, A. Déjardin, C. dePamphilis, J. Detter, B. Dirks, I. Dubchak, S. Duplessis, J. Ehlting, B. Ellis, K. Gendler, D. Goodstein, M. Gribskov, J. Grimwood, A. Groover, L. Gunter, B. Hamberger, B. Heinze, Y. Helariutta, B. Henrissat, D. Holligan, R. Holt, W. Huang, N. Islam-Faridi, S. Jones, M. Jones-Rhoades, R. Jorgensen, C. Joshi, J. Kangasjärvi, J. Karlsson, C. Kelleher, R. Kirkpatrick, M. Kirst, A. Kohler, U. Kalluri, F. Larimer, J. Leebens-Mack, J.-C. Leplé, P. Locascio, Y. Lou, S. Lucas,F. Martin, B. Montanini, C. Napoli, D. R. Nelson, C. Nelson, K. Nieminen, O. Nilsson, V. Pereda, G. Peter, R. Philippe, G. Pilate, A. Poliakov, J. Razumovskaya, P. Richardson, C. Rinaldi, K. Ritland, P. Rouzé, D. Ryaboy, J. Schmutz, J. Schrader, B. Segerman, H. Shin, A. Siddiqui, F. Sterky, A. Terry, C. Tsai, E. Uberbacher, P. Unneberg, J. Vahala, K. Wall, S. Wessler, G. Yang, T. Yin, C. Douglas, M. Marra, G. Sandberg, Y. Van de Peer, D. Rokhsar

*To whom correspondence should be addressed. E-mail: gtk@ornl.gov

Published 15 September, *Science* **313**, 1596 (2006)
DOI: 10.1126/science.1128691

This PDF file includes:

BACKGROUND INFORMATION

Populus life history and anatomy

Populus trichocarpa, and most *Populus* species in general, in their juvenile phase of growth (0 to ca. 15 years), are characterized by an excurrent stem, lenticel covered bark possessing cortical photosynthesis and a deliquescent canopy (***1***). *Populus* wood consists of diffuse porous structure and homocellular rays, with typically subtle differentiation among annual rings. The leaves of *P. trichocarpa* are heterophyllic, alternate and typically lanceolate, with actinodromous venation and a transverse, flattened petiole. *P. trichocarpa* is dioecious, as are most members of the Salicaceae family. Male and female flowers are borne on racemose inflorescence with a reduced calyx. Pollen is wind dispersed as are the plumose seeds. A single mature female can generate over 50 million seeds per year (***1***). *Populus* are generally pioneering species, requiring open environments with exposed mineral soil for successful seed germination and seedling establishment. *Populus* species also effectively propagate themselves through soboliferous or cladoptic shoot production. *P. trichocarpa* today occurs from 62º 30' N in southern Alaska to 31º 45' N in Baja California, Mexico (***2***).

The crown taxa of the genus *Populus* -- poplars, cottonwoods and aspens -- arose during a period of global cooling in the late Miocene (5-10 million years ago (Mya)) (***3, 4***). The first definitive *Populus* fossil dates from 48 Mya (***5***); the sister genus of *Populus, Salix* (willows), shared a common ancestor approximately 65 Mya (***6, 7***). The extended relationships of *Populus* and *Salix* to other taxa have been debated and are ill-defined; consequently, these genera were classified in a digeneric family, Salicaceae. However, recent molecular phylogenetic work has revealed a close relationship between the Salicaceae and Flacourtiaceae (***8***), and under the APG classification, the family Salicaceae now consists of *Populus*, *Salix* and 53 other genera comprising some 1010 species, (http://www.mobot.org/MOBOT/research/APweb/) (***9, 10***). Within the newly circumscribed family, *Populus* and *Salix* are particularly closely related to several genera (*e.g.*, *Carrierea*, *Idesia*, *Itoa*, *Olmediella* and *Poliothyrsis*) (***8, 11***), which together form the salicoid (*i.e.*, *Salix*-related) clade (***12***) within the larger Salicaceae. All members of the Salicaceae are placed in the order Malpighiales (***13***), a diverse group that also includes cassava (*Manihot esculenta* Crantz) and mangrove (*Rhizophora* spp.). The Malpighiales belong to the eurosid I clade of eudicotyledonous angiosperms along with the Fabales (including the legumes soybean, pea and Medicago) and Rosales (including

the stone fruits and berries). In contrast, the model herbaceous plant, *Arabidopsis thaliana* (L.), a member of the Brassicaceae (broccoli, cauliflower, etc.), lies within the eurosid II clade. Comparisons between the *Populus* and Arabidopsis genomes therefore have the potential to illuminate features of their last common ancestor – the ancestral eurosid – which lived approximately 100-120 Mya (***14, 15***).

OVERVIEW OF SEQUENCING AND ASSEMBLY

Shotgun sequencing strategy and results

A whole-genome shotgun strategy (***16***) was adopted for sequencing and assembling the *Populus* genome. It was augmented by construction of a physical map based on BAC restriction fragment fingerprints, BAC-end sequencing, and extensive genetic mapping based on simple sequence repeat (SSR) length polymorphisms (***17, 18***). Since *Populus* is an obligate outcrosser, inbred strains were not available and haplotypic polymorphisms were expected. All genomic DNA was obtained from a single genotype from Washington State, designated 'Nisqually-1' (previously referred to as *P. trichocarpa* clone 383-2499) (***19***). Template DNA was initially extracted from surface disinfested leaf tissue and randomly sheared and size-fractionated to create libraries with roughly 3 kb and 8 kb inserts. Initial quality-control sequencing determined that these libraries contained a high degree of chloroplast and mitochondrial DNA contamination (see below). Therefore, a second genomic DNA preparation from the same genotype was prepared from root tips (***20***) of plants grown in tissue culture (***21***) and hydroponic culture using a sucrose gradient to separate nuclei from organelles, followed by a cesium chloride gradient centrifugation and pulsed-field gel electrophoresis (***22***). The root-derived template was used to prepare fosmid libraries, which were effectively free of organellar contamination. Prior to the initiation of the sequencing project, a BAC library (~10X genome coverage) was constructed from Nisqually-1 at Texas A&M University with partially HindIII digested genomic DNA (***23***).

BAC-end sequence (BES) reads were generated on ABI Prism 3700 DNA Analyzer. The trace data were processed by the program Phred (***24, 25***), with default parameters. Low-quality bases and vector sequence were trimmed from the reads. Trimmed reads containing ≥15 bp were retained for analysis. The BAC-end sequences are available for download at: (http://www.bcgsc.bc.ca/lab/mapping/data).

Nearly 7.5X total sequence redundancy was obtained from the 3 kb and 8 kb paired plasmid ends using standard methods, along with ca. 15X clone coverage in paired fosmid ends (ca. 36 kb average insert size) and ca. 8X clone coverage in paired BAC-ends (ca. 108 kb average insert size) (Table S1). The sequence was assembled with the JAZZ assembler. This resulted in 234 scaffolds longer than 200^{+} kb which cover 378.5 Mb of the genome; 62 scaffolds, each longer than 2.06 Mb, covering a total of 238 Mb or more than half of the whole genome. A total of 410 Mb are captured in the scaffold assembly (Table S2).

Unassembled reads

It was not possible to assemble a substantial fraction of the whole-genome shotgun reads into sequence scaffolds. To assess the nature of the sequences in this fraction of the genome, we performed wu-BLAST searches against databases containing the assembled *P. trichocarpa* mitochondrion, chloroplast, a database of repeats identified by Recon and RepeatMasker (described below), and the non-redundant nucleotide database from NCBI (Fig. S1). These results show that a substantial fraction of the unassembled DNA belongs to repetitive DNA in our database of *Populus* repeats and to possible repetitive DNA not yet characterized in *Populus* or other organisms (no hit to NCBI non-redundant database).

Although the genomic DNA template for whole-genome shotgun sequencing was prepared from surface disinfested leaves and roots, endophytic microorganisms apparently escaped removal by such approaches and their contaminating DNA contributed to the unassembled whole-genome shotgun sequences. To assess these sequences, unassembled reads and small scaffolds (<10 kb) were queried against the NCBI non-redundant database. Table S3 shows that 0.16% of the total number of end reads were from DNA from archaea, bacteria and fungi, suggesting that these organisms may in fact be *Populus* endophytes. Several taxa from Table S3 have been isolated from *Populus* in other studies (*e.g.*, *Rhizobium tropic* (***26***) and *Pseudomonas putida* (***27***)).

Completeness of the assembly of the "euchromatic" genome

Although as described in the previous section a significant fraction of shotgun reads were not assembled, the assembled regions did capture the vast majority of known genes in *Populus*. Specifically, BLAT alignment of 4,664 full-length cDNAs found

that 98.8% hit the assembly over at least 50% of their length. Similarly, 89% of the 260,809 *Populus* ESTs were mapped to the assembly.

BAC clone fingerprinting and physical map construction

BAC clones were fingerprinted with HindIII with an agarose gel-based method (***28-30***). Restriction fragment identification, fragment mobility, and size determination were performed with automated analysis software (***31***). Automated fingerprint map assembly was accomplished with FPC (***32, 33***). Additional processing of the map contigs was achieved by a combination of manual editing and automated tools. The fingerprint map is available for download in FPC format from the Genome Sciences Centre website (http://www.bcgsc.bc.ca/lab/mapping/data) and will be described in detail elsewhere (C. Kelleher *et al*., in preparation). The maps may be viewed with Internet Contig Explorer (***34***).

Comparisons of BAC-end sequence to the whole-genome shotgun assembly (JGI Poplar Genome Assembly version 1.0) were conducted with BLAST (***24, 25, 35***). Those alignments satisfying the criteria of either (i) $\geq$99% identity and E-value$\leq$1e^{-50} or (ii) $\geq$95% identity for $\geq$95% of the read length with an alignment length $\geq$50 bp were used to anchor fingerprint map contigs to the sequence assembly. A total of 94,877 alignments associated with 73,374 unique end-sequences and 42,809 unique clones passed these filters. When alignments for both end sequences of a clone were available, they were subject to additional orientation and distance filters to accept only topologically consistent paired-end alignments (Fig. S2).

To aid integration of the assembled sequence and physical map with the 19 *Populus* linkage groups (LG), a genetic map of *Populus* was constructed from 535 di-, tri- and tetranucleotide sequence-tagged markers (SSR) identified from BAC-end reads, raw shotgun data and targeted library sequences (***17***). These markers were applied to a *P*. *trichocarpa* x *P*. *deltoides* hybrid pedigree with an average recombination rate of ca. 1 centiMorgan (cM) per 200 kb.

Locations of mapped markers in the assembled genome sequence were determined, resulting in an initial assembly of 155 sequence scaffolds, representing 335 Mb mapped to the *P*. *trichocarpa* chromosomes by one or more sequence-tagged markers, with over two-thirds of the assigned scaffolds oriented by two or more markers. This formed the basis of the first publicly-released sequence assembly (available at: www.jgi.doe.gov/poplar). Subsequently, 162 scaffolds, totaling 385 Mb of genomic

sequence were assembled, taking advantage of the orientation information provided by aligning homeologous chromosome segments (see below). The order of the markers on the genetic map was consistent with the major scaffolds, corroborating the large-scale structure of the shotgun assembly (Fig. S3). Polymorphisms associated with heterozygosity within Nisqually-1 were identified by examining alignments of sequence reads produced by the JAZZ assembly program (see below).

Whole-genome DNA alignment

We used the VISTA pipeline infrastructure (***36***) for the construction of genome-wide pairwise DNA alignments between *Populus* and assemblies of *Oryza* and Arabidopsis. To align genomes we implemented new algorithms that used an efficient combination of global and local alignment methods. First, we obtained a map of large blocks of conserved synteny between the two species by applying Shuffle-LAGAN global chaining algorithm (***37***) to local alignments produced by translated BLAT (***38***). After that we used Supermap, the fully symmetric whole-genome extension to the Shuffle-LAGAN. Then, in each syntenic block we applied Shuffle-LAGAN a second time to obtain a more fine-grained map of small-scale rearrangements such as inversions. We have also extended this approach to compare duplicated segments within the genomes. 58% of the length of coding exons, 8% of UTRs and 5% of non-coding sequences of the *Populus* genome are covered by significant pair-wise alignments with Arabidosis (calculated using the techniques first applied to the human-mouse comparison (***39***)). These fractions are respectively 38%, 4% and 3% for the alignment of *Populus* with *Oryza*. The constructed genome-wide pair-wise alignments can be downloaded from: http://pipeline.lbl.gov/downloads.shtml and are accessible for browsing and various types of analysis through the VISTA browser at: http://pipeline.lbl.gov/.

FISH METHODS

Slide preparation *Populus* cuttings of Nisqually-1 were grown in a greenhouse. Healthy roots, about 1 cm long, were excised and pretreated with a saturated aqueous solution of α-bromonapthalene for 1.3 h in dark at room temperature and then fixed in 95% ethanol-glacial acetic acid (4:1 v/v). The root tips were treated enzymatically as described by Jewell and Islam-Faridi (***40***). The digested root tips were macerated and spread on a clean slide in 3:1 ethanol-glacial acetic acid with fine pointed forceps.

Probe DNA nick translation BACs for karyotyping were selected from the whole-genome shotgun assembly based on the frequency of component 16-mers in the whole-genome database. BACs were also preferentially selected based on coding sequence composition. BAC DNA was isolated by alkaline lyses, digested with HindIII, followed by further purification using Plant DNeasy spin columns (QIAGEN, Valencia, CA) using a modified protocol (***41***). BAC DNA and whole plasmids of 18S-28S and 5S rDNA were labeled with biotin-16-dUTP (Biotin-Nick Translation Mix, Roche, Germany) and/or Digoxigenin-11-dUTP (DIG-Nick Translation Mix, Roche, Germany) following instructions provided by the manufacturer. Labeled DNA was dot-blotted to verify incorporation of labeled nucleotides.

In situ **hybridization** The hybridization mixture consisted of 50% deionized formamide (Fisher molecular grade), 10% dextran sulfate, 2X SSC, labeled BAC DNA (30 ng/slide; for dual BAC-FISH, 30 ng of each BAC DNA), 30 ng 18S-28S rDNA or 5S rDNA/slide (when included in the mixture), carrier DNA (*E. coli* DNA, 5 µg/slide), and blocking *Cot-1* DNA (10-fold excess of labeled BAC DNA for single or dual BAC-FISH). The hybridization mixture was denatured in boiling water for 10 min, chilled on ice for 5-6 min and then incubated in a 37ºC water bath for 25-30 min to allow the *Cot-1* DNA to hybridize with the repetitive sequences of BAC DNA. Chromosomal DNA on slides was denatured at 72ºC in 200 µl of 70% deionized formamide/2X SSC on a hot block in an oven for 1.5 min followed by dehydration through an ethanol series (70, 85, 95 and 100%) at -20ºC for 3-4 min each. Slides were air dried for about 25 min prior to loading 25 µl of hybridization mixture, then covered with a glass cover slip and sealed with rubber cement. Following overnight incubation at 37ºC, slides were washed twice in 2X SSC for 5 min each, 30% deionized formamide (Sigma-Aldrich, St. Louis, MO) for 5 min each, 2X SSC for 5 min each at 40ºC followed by twice in 2X SSC (5 min each) and 4X SSC/0.2% Tween-20 (5 min each) at room temperature. Slides were blocked 10 min at room temperature with 5% (w/v) BSA (IgG-free, protease-free, Jackson ImmunoResearch Laboratories, West Grove, PA). The hybridization sites were detected with fluorescein isothiocyanate (FITC) conjugated anti-digoxigenin (Roche, Germany), Cy3-conjugated Streptavidin (Jackson ImmunoResearch Laboratories, USA) or both depending on labeled DNA used in the hybridization mixture. Slides were washed four times in 4X SSC/0.2% Tween-20 for 5 min each at 37ºC, then counterstained with DAPI

(2 µg/ml) in McIlvaine buffer pH 7.0 for 10 min and briefly washed in 4X SSC/0.2% Tween-20 followed by mounting with Vectashield (Vector Laboratories, Burlingame, CA, USA). Slides were stored over-night at 4ºC to stabilize the fluorochromes before viewing under epi-fluorescence microscope.

Microscopy Digital images were recorded from an Olympus (Center Valley, PA) AX-70 Epi-fluorescence microscope with suitable filter sets (Chroma Technology, VT), using a 1.3 MP Sensys (Roper Scientific Tucson, AZ) camera and the MacProbe v4.2.3 digital image system (Applied Imaging, Intl.). Images were processed with MacProbe v4.2.3 (UCSF Medical Center, San Fransisco, CA) and Adobe Photoshop v8 (Adobe Systems, San Jose, CA).

Euchromatin and heterochromatin measurement and analysis Seven well-spread cells each of prophase and metaphase were chosen to measure total chromosome length and heterochromatic (bright DAPI stained region) length. Euchromatic length of each chromosome was determined by subtracting the heterochromatic length from the total length. In each cell, chromosomes were numbered arbitrarily from 1 to 38. Distances from one to the other end of each chromosome and the block of heterochromatic region were measured three times. All data were collected with Optimas v6 (Pixera, Los Gatos, CA) after zooming 300% for metaphase and 200% for prophase chromosomes to minimize measurement error.

Metaphase adjusted heterochromatin and euchromatin lengths were obtained by assuming 3X higher contraction rate for euchromatin than for heterochromatin (as observed in maize) such that the overall chromatin contraction matched what was observed in *Populus* (61.5%). For these estimates the assumed contraction rates are 22.9% for heterochromatin and 68.6% for euchromatin (Fig. S4).

Telomeric repeats Putative telomeric repeats were identified from repetitive sequences found at the ends of scaffolds mapped onto chromosomes. The consensus telomeric repeat, CCCTAAA (***41***), was used as a query sequence in a BLASTN search (with dust filter disabled, E-value$\leq 1e^{-10}$) against the assembled genome, yielding hits at 20 of 38 chromosome ends (Table S4). Furthermore, 21 unassembled sequence scaffolds had BLASTN hits covering 90% or more of their lengths, plus 38 additional scaffolds had hits covering at least 30% of their length. Together these scaffolds ranged in size from 1 kb

to 62 kb (median 1.8 kb) and may correspond to portions of unassembled telomeric ends. Fluorescent *in situ* hybridization (FISH) using a telomere-specific probe indicated that all 19 chromosomes have extensive telomeric ends, with no evidence of degenerated internal telomeres (Fig. S4).

GENE CONTENT

Gene prediction and annotation

The gene prediction methods used for annotation include *ab initio* FgenesH (***42***), homology-based FgenesH^{+} (http://www.softberry.com/berry.phtml), Genewise (***43***), GrailExp6 (***44***), and EuGène (***45***). Parameters for each gene prediction method were developed independently on subsets of *Populus* and other plant genes by three separate annotation groups. A total of 4,464 full-length sequences from enriched cDNA libraries prepared from Nisqually-1 were generated and used in training the gene-calling algorithms. Repetitive elements were identified as described below and masked on the genome assembly to exclude such elements from the final model set. All predicted gene models were annotated by double-affine Smith-Waterman alignments against SwissProt, KEGG, nonredundant green plant protein database at NCBI, and known Arabidopsis proteins. Protein domains were predicted using InterProScan against various domain libraries (Prints, Prosite, Pfam, ProDom & SMART). Annotations were also assigned to Gene Ontology (***46***), eukaryotic clusters of orthologous groups (KOG (***47***)) and KEGG metabolic pathways (***48***). The composite non-redundant "reference set" of genes was promoted on the basis of I) homology to a curated set of 307,579 plant proteins, II) completeness of the model, III) homology to a manually-curated *Populus*-specific EST and IV) predicted transcript and protein size (Table S5).

Nomenclature

The 45,555 nuclear gene models that were promoted to a "Reference" set, including 4,378 models manually annotated at a community "Jamboree", are available at: www.jgi.doe.gov/poplar. Genes are provisionally designated by the reference set. This designation is a combination of the annotation gene program name and a unique alpha-numerical combination. Each gene has also a unique numeric identifier (protein_id), which defines it locus tag.

Transposable elements

Transposable element coding regions were screened out of the predicted set of gene models based on homology to known transposable elements present in the GenBank nonredundant nucleotide database. However, this set of transposable elements is incomplete and does not contain elements specific to *Populus*. Subsequent to the release of the 45,555 gene models, we identified *Populus*-specific repeats as described below. A BLASTn comparison of these newly annotated elements with the *Populus* gene set revealed that 375 of the promoted gene models had significant (E-value$\leq 1e^{-10}$) homology to putative *Populus* transposable elements and an additional 2,873 gene models had homology to unannotated *Populus* repetitive elements.

Conserved hypothetical homologs

The annotated *Populus* gene set contains conserved homologies with approximately 725 Arabidopsis gene models designated as "hypothetical." The homologous annotations in *Populus* can be used to change the Arabidopsis designation from hypothetical to unknown (*i.e.*, "conserved genes of unknown function"). Conversely, approximately 1,099 hypothetical genes in Arabidopsis were not corroborated by clear conservation in *Populus* and for now should be viewed as: 1) spurious or partial gene predictions or 2) highly divergent but still hypothetical. In addition, there were 25 hypothetical Arabidopsis genes that had a BLAST hit on almost all *Populus* chromosomes. These hits were however mostly in repetitive regions of the *Populus* genome, suggesting that these hypothetical genes may represent undiscovered transposable element families in Arabidopsis (See http://www.ornl.gov/sci/ipgc/).

WHOLE-GENOME MICROARRAY ANALYSES

The *Populus trichocarpa* oligonucleotide microarray

The *Populus* whole-genome microarray manufactured by NimbleGen (Madison, WI) contains one to three independent, non-identical, 60-mer probes per gene model. Included in the microarray are 44,133 annotated, promoted gene models – 42,373 represented by three 60-mer probes and 1,760 by one or two probes. An additional 11,661 gene models (10,875 represented by three probes), initially annotated and later discarded from the reference set due to lack of biological support (annotated, non-promoted gene models), as well as 69 chloroplast genes (49 with three probes), 58

mitochondrial genes (49 with three probes) and 48 microRNA precursors are also represented in the microarray. Approximately 1,400 annotated and promoted gene models are not specifically targeted by the array because unique probes could not be designed. A manuscript fully describing the array is in preparation (***49***).

Plant material and RNA extraction

Four clonal replicates of Nisqually-1 were planted in 8-liter pots and grown in an ebb-and-flow flood bench system under complete nutrient solution. Whole roots, stem nodes and internodes, and young and mature leaves were collected from juvenile plants, over two consecutive days (2 replicates/day) during the growth season. Tissues were immediately frozen in liquid nitrogen and RNA extraction was carried out in a GenoGrinder2000 (Spex Certiprep Inc.), followed by purification using the RNeasy Plant Mini Kit (Qiagen) (***50***). Total RNA preparations (4 biological replicates per vegetative organ) were labeled and each biological replicate was individually hybridized to the arrays by NimbleGen, using standard single-dye labeling and hybridization protocols.

Statistical analysis

Microarray data collected for each organ, in 4 biologically replicated samples, was analyzed by means of a mixed linear model, following the strategy previously outlined by Hsieh *et al*. (***51***) and Chu *et al*. (***52***) (Model II – no mismatch probe data). The data was normalized by array and $\log_2$-transformed before analysis of variance (ANOVA) was carried out in a model comprising gene (fixed) and probe (random) effects. Individual genes and negative-controls effects were estimated by least-square means, followed by pairwise comparisons of signal estimates between pairs of duplicated genes in a t-test. Genes to be compared were separated into two groups based on the age of the duplication event (2,632 pairs for eurosid and 6,968 pairs for salicoid event). A false discovery rate (***53***) of 5% was applied to define differentially expressed duplicated genes in each vegetative organ. A similar strategy was applied for identification of expressed genes. Signal intensities detected for each predicted transcriptional unit were contrasted to a set of 20 negative-control probes. Presence of expression was declared when signal intensity was significantly higher than that detected in the negative control probes (false discovery rate of 5%). These analyses were carried out for transcriptional units for which three 60-mer probes were available.

ORGANELLAR GENOMES

As described above, organellar DNA represented a substantial contaminant of the nuclear DNA preparations from leaf tissue. Putative organellar reads were identified as sequences with an unusually high depth in the initial stages of JAZZ assembly. Organellar genomes were assembled from a subset of these reads using Phrap.

The Chloroplast

The *P. trichocarpa* chloroplast genome was assembled from 139,442 sequence reads (See http://genome.ornl.gov/poplar_chloroplast/poplar_chloroplast.html). The unprecedented chloroplast assembly depth (*i.e.*, 410 high-quality (>Q40) reads per position on average) ensured a highly accurate sequence and assembly (Fig. S5). The resulting genome consists of 157,033 bp (*i.e.*, 85,129 bp in a large single-copy region, 16,600 bp in a single-copy region (which is present in both orientations) and 27,652 bp in each of two copies of an inverted repeat. The overall GC content was 36.7% (A, 31%; C, 19%; G, 18%; T, 32%). There was homology (E-value$\leq$1e^{-10}) support for 101 *ab initio* coding sequences, of which 16 contained two introns, corresponding to a gene density of 0.64 genes per kb (at a men of 1,554 bp per gene) and a coding percentage of 53.3% across the genome (coding percentage including introns: 61.0%). The resulting assembly is similar to that of other angiosperms with two exceptions -- the first in gene content (*rps16* and *rpl36* are missing in *Populus*) and the second in slight differences in the border of the inverted repeat. The hyperaccurate state of the *Populus* chloroplast genome suggests that these exceptions are not artefacts of sequencing or assembly.

The chloroplast sequence assembly also contained 150 single nucleotide polymorphisms (SNP) or indel polymorphisms, represented by at least two high-quality (>Q40) sequence reads for each polymorphism. In addition, putative nuclear-chloroplast chimeras were identified when one read from a pair aligned to the nucleus assembly, while the other aligned to the chloroplast assembly. Sequences harboring polymorphisms had a significantly higher rate of nuclear-chloroplast chimeras compared to sequences with no polymorphisms (18% vs. 6%, respectively), suggesting that these polymorphisms were primarily due to nuclear translocations of portions of the chloroplast sequence. Nuclear-chloroplast chimeras occurred across the chloroplast genome, indicating that all portions of the chloroplast have been translocated to the nucleus (Fig. S6). However, due to the difficulties of assembling repetitive regions in a shotgun

assembly, it is difficult to quantify the exact extent and size of chloroplast insertions in the nuclear genome with the current database. Evolutionary analyses based on an alignment of sixty orthologous genes from six seed plant taxa (*Populus*, Arabidopsis, spinach, rice, lotus and pine) show that the *Populus* chloroplast displays a dramatic reduction in the rate of nucleotide substitution, with a 43% reduction in rate relative to *Lotus corniculatus*.

The Mitochondrion

The *Populus* mitochondrion genome was assembled from 280,792 sequence reads, resulting in assembly of three circular molecules of 186, 280 and 336 kb. These were in turn assembled into a putative master molecule of 803 kb based on the presence of shared direct repeats (***54***). This assembly has not been subjected to experimental validation. Gene content of the mitochondrion was provisionally determined with the GrailEXP pipeline at Oak Ridge National Laboratory. In total there are 52 predicted protein coding genes.

TIMING OF WHOLE-GENOME DUPLICATIONS

The *Populus* and Arabidopsis genomes were reconstructed into segments with conserved synteny that subsequently were compared with a variant of the algorithm of Hokamp *et al*. (***55***). For *Populus* reconstruction I, a maximum of 10 non-aligning genes between aligning genes within segment pairs was allowed and at least five aligning genes per segment pair was required, there were 171 segments containing 18,308 *Populus* genes representing 28,174 of the annotated genes. These segments have 4DTV distances between 0.02 and 0.18, suggesting they are almost exclusively from the salicoid duplication. The total combined 4DTV distance is 0.0913$\pm$0.0003 or corrected for multiple substitutions (MS) 0.1008$\pm$0.0003. Reconstruction II resulted in 64 segments containing 2,914 genes responsible for 5,632 of the annotated *Populus* genes. The total combined 4DTV distance is 0.359$\pm$0.003 or corrected for MS 0.633$\pm$0.004.

Arabidopsis reconstruction I, under the same conditions described above resulted in 160 segments containing 13,118 genes responsible for 16,795 of the Arabidopsis genes. The total combined 4DTV distance is 0.2414$\pm$0.0009 or corrected for MS 0.330$\pm$0.001. Reconstruction II resulted in 28 segments containing 1,132 genes

responsible for 1,738 of present day Arabidopsis genes. The total combined 4DTV distance is 0.391±0.005 or corrected for MS 0.763±0.008.

The *Populus*-Arabidopsis reconstruction based on the two sets of twice-reconstructed segments resulted in 214 segments with 10,070 unique genes. The combined 4DTV distance is 0.377±0.001 or MS corrected 0.701±0.002. The distribution function of the 4DTV distances of the individual segments shows a single well-defined peak at 0.377 (data not shown).

Assuming that the MS corrected 4DTV distances are additive, the following relationship between the *Populus*-Arabidopsis 4DTV (PA) and the 4DTV distances between segments from the respective *Populus* and Arabidopsis duplication II (PP and AA) holds:

$$1/2\ (PP + AA) = PA + X,$$

where, $X \geq 0$ if duplication II is a single shared duplication event that occurred in an ancestral eurosid lineage (hypothesis H_1) and $X < 0$ if the two genome-wide duplications from this epoch occurred independently in each lineage (H_2). This relationship is independent of whether the transversion rate has been faster or slower in one species.

Solving for X from the above inter- and intra-genomic fourfold synonymous rates, we find X=0.012±0.013. Thus, X is indistinguishable from zero, and we cannot cleanly resolve the timing of duplication II in *Populus* and Arabidopsis. Since X is only ~1-2% of the PA divergence, our best estimate of the timing of the duplication is only a few million years before the Eurosid I (*Populus*)-Eurosid II (Arabidopsis) divergence and we cannot rule out the simultaneity of speciation and duplication.

Quartets, *i.e.,* cases where there were two surviving copies from earlier duplication in both *Populus* (eurosid) and Arabidopsis (β), were examined. From analyzing the reconstructed segments, 11.6% of the original genes survive in two copies after duplication in Arabidopsis, whereas 20.2% survives in *Populus*. Hence, if the survival rates were independent, 2.34% of the genes would be expected to be in quartets. In reality, 6.8% was observed; suggesting that the survival rates are correlated between the two species. This is possibly due to loss prior to the *Populus*-Arabidopsis split, but may also reflect a general tendency for certain types of genes to be retained or lost, which could be similar even for different lineages.

A total of 103 quartets were discovered. At most, three to five quartets come from a single reconstructed ancestral segment. If the earlier duplication happened prior to the P-A split (H_1), the quartets should show a ((P, A), (P, A)) phylogeny, whereas H_2 would support ((P, P), (A, A)). Some quartets do not contain adequate information for reliable phylogenetic determination; these genes may be too well conserved, too short or too variable. The predicted trees were recalculated to obtain a bootstrapped estimate for each tree. Of these the 60 quartets which had bootstrap support of 90% or better, 42 supported ((P, A), (P, A)), *i.e.,* H_1, whereas 18 supported H_2. However, there are two possible trees supporting PA, namely ((P1, A1), (P2, A2)) or ((P1, A2), (P2, A1)). Hence, if the earlier duplication happened at the very time of speciation, there should have been twice as many trees supporting H_1 as H_2. These results supply slightly more support for H_1, but not enough to reject H_2 if the two versions of duplication happened in both species very shortly after speciation.

GENE CONTENT COMPARISONS

The Phytozome clustering noted in the text was performed as follows: first, all-vs.-all Smith-Waterman alignments were carried out between the gene sets of *Populus*, Arabidopsis and *Oryza*, including within-genome alignments. Alignments were performed using a TimeLogic Decypher engine. To produce the *Populus*-Arabidopsis clusters, we made a “backbone” clustering of mutual-best-hits between *Populus* and Arabidopsis. Next, for genes not in these backbone clusters, we found those that were “best hits” to genes in the backbone. Such a gene was assigned to this best-hitting cluster if it met both of the following criteria: (i) the 4DTV score of the gene to its best hit is greater than 0.28 and (ii) the gene also hit one of the original backbone genes in the cluster with score greater than E-value$\leq 1e^{-10}$. This step was repeated until the clustering converged. Finally, remaining unclustered *Populus* and Arabidopsis genes were clustered separately into organism-specific gene families by single link clustering with threshold E-value$\leq 1e^{-10}$.

The Phytozome angiosperm clusters were produced by the following approach: first, mutual best hit backbone clustering of genes from *Populus*, *Oryza* and Arabidopsis were made. Next, genes or *Populus*-Arabidopsis clusters that were not yet assigned to angiosperm groups were allowed to join an angiosperm cluster. Both the “eurosid”

(*Populus*-Arabidopsis) and "angiosperm" (*Populus*-Arabidopsis-*Oryza*) clusters can be interactively accessed at: www.phytozome.net.

A second comparative analysis used the Inparanoid methodology (***56***) to compare the complete gene sets from *Populus* and Arabidopsis. This approach produced 9,423 groups of inparalogs representing 14,837 *Populus* genes and 12,618 Arabidopsis genes (Table S6). Gene clusters for 7,323 *Populus* genes and 5,596 Arabidopsis could not be generated because they represented unique genes. The average *Populus* to Arabidopsis ratio across all orthologous groups was 1.33. The most frequent gene-to-gene ratio was 1:1, represented by 4,607 gene pairs. The next most frequent ratio was 2:1 *Populus* to Arabidopsis. However, there were extreme ratios in both species, *e.g.*, a single gene pair with a ratio of 1:40 *Populus* to Arabidopsis for F-box domains (PF00646) vs. a second gene pair with a ratio of 20:1 *Populus* to Arabidopsis for a zinc finger (B-box type) family protein/salt tolerance-like protein (PF00643).

NON-CODING RNAS

Transfer RNAs (tRNA)

The tRNAScan-SE algorithms, as applied to the chromosome-level assembly with "relaxed" setting for tRNAscan and EufindtRNA and a cutoff of 20 bits (***57***), resulted in the identification of 817 putative tRNA in the *Populus* assembly. All 57 possible anti-codon tRNA were found. One selenocysteine tRNA was detected with these settings and two probable suppressor tRNAs (anticodon which binds stop codons) were also found. In addition, 54 tRNA pseudogenes were detected by tRNAScan-SE. As a test of the accuracy of this approach, the COVE program (http://selab.wustl.edu/cgi-bin/selab.pl?mode=software#cove) was used to scan the chloroplast without the preliminary tRNAscan and EufindtRNA analysis. Thirty tRNA were predicted in the chloroplast genome when COVE was used, while 37 tRNA had previously been manually annotated in the chloroplast. This difference, *i.e.*, seven-undetected tRNA, included tRNA that had long introns. We then performed the same analysis on the Arabidopsis genome assembly (TIGR v01212004), and found 643 tRNA, compared to 711 identified by the Arabidopsis Genome Initiative (***58***). Therefore, we are likely underestimating the number of nuclear tRNA in *Populus*. However, our estimates for

Populus and Arabidopsis can be compared directly. This comparison suggests that *Populus* has nearly 1.3 times as many tRNA as Arabidopsis (Fig. S7).

Spliceosomal RNAs (snRNA)

Using INFERNAL with the default scanning window of 200 bp, a cut off of 10 bits and models supplied by RFAM, all expected spliceosomal snRNAs were discovered. The *Populus* genome contains 22 copies of the U1, 26 copies of U2, 6 copies of U4, 23 copies of U5 and 11 copies of U6. Six copies of the pre-rRNA processing snRNAs U14 were also found. The snRNA were randomly dispersed across the genome. A similar analysis with the Arabidopsis genome revealed that *Populus* has a 1.3 to 1.0 ratio in the number of snRNA compared with Arabidopsis. Comparatively, U1, U2 and U5 are overrepresented in *Populus* while U4 is under-represented. Furthermore, U14 was not detected in Arabidopsis. The snRNA have not been experimentally verified in *Populus*.

Small Nucleolar RNAs (snoRNA)

The C/D snoRNA were predicted using snoScan with the yeast rRNA methylation sites and yeast rRNA sequences provided by the snoScan distribution (http://lowelab.ucsc.edu/snoscan/). The minimum cutoff score was based on the settings which yield a false positive rate of 25 bits. A total of 339 putative C/D snoRNAs were predicted for *Populus*. Under identical criteria, 108 Arabidopsis C/D snRNAs were predicted from the TAIR Arabidopsis database, representing a 3.1-fold expansion in *Populus*. Similarly, H/ACA snoRNAs were detected using snoGPS using the yeast score tables and target pseudouridines. Under these criteria, *Populus* contains 88 predicted H/ACA snoRNAs, compared with 38 predicted H/ACA snoRNA for Arabidopsis, representing a 2.3-fold expansion in *Populus* (http://bioinf.scri.sari.ac.uk/cgi-bin/plant_snorna/arabidopsis). The snoRNA have not been experimentally verified in *Populus*.

Ribosomal RNAs (rRNA)

The consensus size of the *Populus* 5S ribosomal RNA (rRNA) repeats was 490 bp and the consensus 18S-5.8S-26S (45S) repeat was 5,737 bp. A BLASTN analysis using these sequences as queries revealed that portions of the 5S repeat assembled to 13 chromosomes and 32 unassembled scaffolds, whereas portions of the 45S repeat assembled to 14 chromosomes and 99 unassembled scaffolds.

The number and locations of major rRNA repeats were determined by means of fluorescent *in situ* hybridization (FISH) with diagnostic BAC probes and ribosomal repeat probes (methods described above). FISH revealed one main 5S repeat cluster on LGXVII and two major 45S repeat clusters, one of which is located on LGXIV and the other of which remains undetermined (Fig. S4). These results conflict somewhat with previous studies that revealed two major 45S clusters and two 5S clusters in the closely related species *P. balsamifera* (***59***).

MicroRNA (miRNA)

Results of miRNA characterization are reported in the main text. In addition, major classes of miRNA in *Populus*, Arabidopsis and *Oryza* are indicated in Table S7. Of the 21 miRNA families conserved between Arabidopsis and *Populus*, several have been shown to be conserved outside of seed plants (***60-62***). In general it is likely that these conserved miRNA play similar roles in most plants. However, the overrepresentation of certain miRNA families and target classes suggests that some of these miRNA families may play a unique role in *Populus* development (Table S8). For example, target sites for MYB and TCP transcription factor families are underrepresented in *Populus* relative to Arabidopsis, yet the number of actual miRNAs is nearly tripled in *Populus*, suggesting either a more complicated gene regulation system in *Populus* or a simplified regulation system in Arabidopsis. Similarly, miR169, which interacts with CCAAT binding factors (HAP2-like), is overrepresented in *Populus* and has recently been shown to play a role in winter vegetative dormancy (***63***) and lateral branching (***64***), physio-morphological processes not common in Arabidopsis. Finally, the miR397 family in *Populus* is complementary to mRNAs of 26 laccase genes, whereas it has comparable complementarity to only three mRNAs in Arabidopsis. While the roles that laccases play in the biology of plants are not well understood, there is speculation that they may be involved in secondary cell wall formation (***65, 66***), a process that is likely to be more critical in a woody plant such as *Populus*. Many of the predicted miRNA have been recently verified experimentally by Lu *et al*. (***67***).

TANDEM REPEATS

Tandemly duplicated genes were identified and defined as an array of two or more promoted gene models with Smith-Waterman alignment E-value$\leq 1e^{-25}$ that were

enclosed within a 100 kb window. This analysis was performed for both *Populus* (assembled linkage groups only) and Arabidopsis to facilitate comparison of tandem duplication rates in both species. The total number of InterPro domains contained in tandemly duplicated genes was calculated for both Arabidopsis and *Populus* (Table S9; Fig. S8A, B).

FATES OF HOMEOLOGOUS GENES

Abundance of ESTs from non-normalized libraries prepared from different *Populus* species, tissue types and treatments (***68***) was compared as an indicator of differential expression patterns for duplicated genes. Raw ESTs were mapped to gene models based on best BLASTN hits.

In order to calculate the rate of false rejection of the null hypotheses for a given number of ESTs for each gene, pairs of resampled distributions were generated from the same randomly selected gene and α values were calculated for rejection of the null hypothesis that the distributions are equal:

$$\alpha = 1- \exp((-ND)^2)$$

where, N is the number of ESTs in each library, and D is the Kolmogorov-Smirnov D statistic (the maximum difference between the cumulative distributions). This test was repeated for different numbers of ESTs per library, ranging from 5 to 25, with equal numbers sampled for the pairs. Figure S9 shows the weighted mean alpha value for 1,000,000 tests for each number of sampled ESTs. A plot of the proportion of false positives versus number of ESTs gives similar results, but the curve is discontinuous due to the combinatorial nature of the test. These plots were used to determine that each library must have a minimum of 16 ESTs per pairwise comparison for a type 2 error rate of 0.05 (Fig. S9). Sixty-six duplicate pairs of genes met this criterion for the salicoid set; 18 duplicates in the eurosid set. A Kolmogorov-Smirnov goodness-of-fit test was then used to determine if the frequency of detected ESTs varied between paralogous genes. The critical D-values and associated p-values were calculated on an individual library basis and were adjusted for unequal sample size per library.

As an independent test, we used the frequency of EST observed in libraries from different tissue types and experimental treatments (***68***) to test for differential expression of duplicated pairs of genes. Comparing paralogous gene pairs resulting from either the

eurosid or salicoid duplication events in *Populus* and analyzing differences in the overall expression levels between eurosid pairs, salicoid pairs and a set of random gene pairs, significant divergence in expression levels was detected in the numbers of ESTs per library (Fig. S10A). The gene pairs in the salicoid dataset displayed a 2X difference in the number of ESTs per library in 70% of the paired comparisons (*i.e.*, 30% had more than 2X difference). Approximately 20% of the gene pairs in the eurosid dataset showed less than a 2X difference in the number of detected EST. In the random dataset, about 8% of the genes had less than 2X difference in expression. Overall, duplicated gene pairs displayed a significant decrease in shared expression patterns per tissue library over time, *i.e.*, the salicoid gene pairs had fewer paired comparison with a 2X difference or greater than did the eurosid pairs. Likewise, a correlation analysis between the expression profiles with a Pearson correlation test ($p \leq 0.01$) supports the conclusion that functional divergence is occurring in the duplicated *Populus* genome. Only 1% of the random pairs had significantly correlated expression patterns, 3% of the eurosid pairs had significantly correlated expressions and 7% of the salicoid pairs had significantly correlated expressions (Fig. S10B). Similarly, differential expression patterns were detected in the fraction of the duplicated genes that had a tissue-specific expression based on a Fisher exact test ($p \leq 0.001$). Here, approximately 4% of the genes in the random dataset appeared to have a tissue-specific expression, whereas 7% in the salicoid dataset and almost 10% in the eurosid dataset fulfilled this criterion (Fig. S10C).

SINGLE NUCLEOTIDE POLYMORPHISMS

Single nucleotide polymorphisms (SNP) were identified by examining alignments of raw sequence reads that were constructed by the JAZZ assembler. SNP were defined as loci with at least three sequence reads for each allele and only two alleles per locus. The number of identified SNP was strongly dependent on the minimum number of sequences required for each allele (Fig. S11). The number of SNP causing frameshift mutations was particularly sensitive, suggesting that many of these were artifacts of the assembly process. In contrast, SNP of other classes, including synonymous, nonsynonymous and noncoding, all responded similarly to increased stringency of coverage, suggesting assembly artifacts may not be as important for these classes of SNP. The analyses reported here disregard all frameshift mutations.

Rates of heterozygous synonymous and nonsynonymous coding sequence polymorphisms in the sequenced genotype were estimated using the yn00 program of PAML (***69***). The ratio of synonymous to nonsynonymous substitution rates ($\omega=d_N/d_S$) was calculated for all genes with at least 5 total SNP and at least one synonymous SNP.

We performed an Analysis of Variance using the GenMod procedure of SAS. The dependent variable for this analysis was ω and explanatory variables were indicator variables for retention of duplicates from the eurosid or the salicoid duplication events. Covariates included gene size, synonymous substitution rate and minimum genetic distance to the closest paralog as covariates (Table S10). To further reduce the possibility of skewing results with pseudogenes, we restricted the analysis to genes with significant BLAST hits (E-value$\leq 1e^{-10}$) to annotated plant genes and eliminated outliers with anomalously high d_N and d_S values. Genes with homeologs from the salicoid duplication have significantly lower ω, even after taking gene size (assuming pseudogenes are truncated), synonymous substation rate and minimum 4DTV distance into account (Table S11). Similar results were obtained with nonparametric Wilcoxon Rank Sum tests.

GENE FAMILY COMPARISONS

Transporter gene family

Supplemental Table S12 shows a comparative summary of the various transporter gene families found in *Populus* and Arabidopsis

Lignin biosynthetic genes

Supplemental Table S13 contains the set of 37 manually annotated *Populus* phenylpropanoid metabolite biosynthesis, including lignin, the given reference gene model names and the proposed *Populus* gene designations.

Kinases and transcription factors

Gene families coding for transcription factors and kinases generally show a high retention after a genome duplication event (***70***). This trend is found in *Populus* with three major exceptions. Unlike the other transcription factors, the MADS-box and GRAS transcription factor families appear to have lost the majority of their duplicated gene pairs (Fig. S12), with the MADS-box family in *Populus* having roughly the same number of

gene members as Arabidopsis (117 vs. 116, respectively). Also, several classes of chromatin-based transcription factors are not overrepresented in *Populus* when comapred to Arabidopsis, including histone acetyltransferases and deacetylases (28 in each species). Most non-histone chromatin proteins that bind DNA directly are also nearly equally represented in *Populus* and Arabidopsis, *e.g.*, HMG box proteins (15 vs. 14, respectively), methyl-DNA binding proteins (14 vs. 13) and DNA methyltransferases (9 vs. 11).

Disease Resistance Genes

Figure S13 presents the chromosomal localization designated by linkage groups, for disease resistance genes, genes coding for P450 enzymes, and all transcription factors.

Transporter Genes

Suppplemental Figure S14 depicts comparative numbers of transporter gene models in *Populus* and Arabidopsis.

Cytochrome P450

Cytochrome P450 (CYP) enzymes constitute a large family of proteins responsible for diverse functions (***71, 72***), including biosynthesis of signaling molecules, alkaloids, pigments, and essential oils. Comparison of *Populus* and Arabidopsis shows that all CYP families (***73, 74***) in Arabidopsis (246 full-length P450 genes and 28 pseudogenes) are present in *Populus* (363 full-length P450 genes and 203 putative pseudogenes, the largest total of any species so far) with the exception of CYP702 and CYP708, which are also missing from other plants (***75-77***). Arabidopsis is missing six CYP families present in *Populus* (CYP92, CYP727, CYP728, CYP729, CYP733 & CYP736). Perhaps, surprisingly, there was only one new P450 family found in *Populus*. This is the CYP737 family in the CYP72 clan. CYP737 is similar to the CYP734 family, which is involved in a light sensing pathway. CYP734A1 is induced by far-red light, hydroxylating castasterone and brassinolide to inactive forms, thus altering growth in a light dependent manner (***78***). Three quarters of the plant CYP families are shared between *Populus*, Arabidopsis and *Oryza*, but only three families (CYP51, CYP710 and CYP711) are shared with *Chlamydomonas*, implying that the birth of the other 60 plant

CYP families occurred after the plants colonized the land. For detailed sequence and gene nomenclature information see: http://drnelson.utmem.edu/CytochromeP450.html.

REPETITIVE ELEMENTS

The repeat composition of the *Populus* genome was estimated in three complementary ways. First, the frequency of 16-mer words was determined for all sequence reads and these word frequencies were mapped onto the assembled sequence. Approximately 41% of the assembled genome was covered by 16-mers that occurred with a frequency of 34 or greater in the raw sequence reads, which corresponded to 44% of the assembled genome. This initial identification of repeats was used to premask the assembled genome prior to gene calling, thus minimizing the number of coding regions from repetitive elements that were included in the initial set of gene models. The methods for incorporating this masking information into gene calling algorithms varied among the annotation groups.

The word-based method for identifying repeats is efficient but coarse, and many coding sequences from large gene families were included in the masked regions. We therefore also identified individual repetitive elements in the assembled genome by comparing all assembled scaffolds with each other using wu-BLASTn and processing the output using Recon (***79***). We identified 12,250 consensus sequences for repetitive elements. Only 794 of these repeat elements had homology to known repeat elements in the databases at RepBase (http://www.girinst.org/) (Table S14). We used RepeatMasker v3.1 with wu-BLAST to delineate the occurrence of these elements in the assembled genome. In total, these elements covered approximately 173 Mb, and an additional 3.3 Mb were covered by low complexity repeats, representing 42% of the assembled genome. Elements that were annotated as known TEs covered approximately 53 Mb of the assembled genome, while unidentified elements accounted for the remaining 120 Mb of repetitive sequence (Table S14).

The third method for characterizing repeat composition of the genome was to use conserved portions of known transposable element (TE) coding regions as query sequences in TBLASTN searches of the assembled genome, followed by examination of flanking sequence for characteristic signatures of each element. As the largest fraction of most eukaryotic genomes, TEs are the most abundant component of genomic sequencing projects (***80***). Not surprisingly, this is also true of the *Populus* genomic

sequence (Fig. S15). All previously identified major TE types are present in *Populus*. The most abundant TE are Class 1 elements (Copia-like, Gypsy-like and LINE) which are collectively represented by over ~5000 copies. Class 2 elements of all superfamilies (PIF, Pong, CACTA, MULE and hAT) account for ~1000 copies each. Comparison of the major TE types between *Populus* and Arabidopsis revealed that all TE are more abundant in *Populus* except for MULE which are roughly three times more abundant in Arabidopsis. Because the Helitrons have distinct structural features that are not readily detected by computer-assisted approach this group of elements was not examined in this study.

Supporting References and Notes

1. D. I. Dickmann, in *Poplar culture in North America*, D. I. I. J. G. E. J. E. R. J. Dickmann, C. C. National Research, Poplar Council of Canada., Poplar Council of the United States., Eds. (NRC Research Press, Ottawa, 2001), pp. 1-42.

2. D. S. DeBell, in *Silvics of North America, Volume 2, Hardwoods*, R. M. Burns, B. H. Honkala, United States., Forest Service., Eds. (U.S. Dept. of Agriculture, Forest Service, Washington, D.C., 1990), pp. 570-576.

3. M. Schoell, S. Schouten, J. S. S. Damste, J. W. Deleeuw, R. E. Summons, *Science* **263**, 1122 (1994).

4. A. E. Shevenell, J. P. Kennett, D. W. Lea, *Science* **305**, 1766 (2004).

5. S. R. Manchester, D. L. Dilcher, W. D. Tidwell, *American Journal of Botany* **73**, 156 (1986).

6. M. E. Collinson, *Proceedings of the Royal Society of Edinburgh Section B-Biological Sciences* **98**, 155 (1992).

7. J. E. Eckenwalder, in *Biology of Populus and its implications for management and conservation*, R. F. Stettler, Jr. H. D. 1. Bradshaw, P. E. Heilman, T. M. Hinckley, Eds. (NRC Research Press, Ottawa, 1996), Chap. 1, pp. 7-32.

8. M. W. Chase *et al.*, *Kew Bulletin* **57**, 141 (2002).

9. B. Bremer *et al.*, *Botanical Journal of the Linnean Society* **141**, 399 (2003).

10. P. F. Stevens. Angiosperm Phylogeny Website, Version 6. 2005. Ref Type: Internet Communication

11. M. H. Alford, Cornell University (2005).

12. Q. C. B. Cronk, *New Phytologist* **166**, 39 (2005).

13. C. C. Davis, C. O. Webb, K. J. Wurdack, C. A. Jaramillo, M. J. Donoghue, *American Naturalist* **165**, E36 (2005).

14. N. Wikstrom, V. Savolainen, M. W. Chase, *Proceedings of the Royal Society of London Series B-Biological Sciences* **268**, 2211 (2001).

15. M. J. Sanderson, J. L. Thorne, N. Wikstrom, K. Bremer, *American Journal of Botany* **91**, 1656 (2004).

16. G. Myers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 202 (1999).

17. G. A. Tuskan *et al.*, *Canadian Journal of Forest Research* **34**, 85 (2004).

18. T. M. Yin, S. P. DiFazio, L. E. Gunter, D. Riemenschneider, G. A. Tuskan, *Theoretical and Applied Genetics* **109**, 451 (2004).

19. B. Stirling, G. Newcombe, J. Vrebalov, I. Bosdet, H. D. Bradshaw, *Theoretical and Applied Genetics* **103**, 1129 (2001).

20. H. B. Zhang, X. P. Zhao, X. L. Ding, A. H. Paterson, R. A. Wing, *Plant Journal* **7**, 175 (1995).

21. C. Ma, S. H. Strauss, R. Meilan, *Plant Molecular Biology Reporter* **22**, 311a (2004).

22. K. Weising, *DNA Fingerprinting in Plants and Fungi* (CRC Press, Boca Raton, 1995).

23. B. Stirling, Z. K. Yang, L. E. Gunter, G. A. Tuskan, H. D. Bradshaw, *Canadian Journal of Forest Research* **33**, 2245 (2003).

24. B. Ewing, P. Green, *Genome Research* **8**, 186 (1998).

25. B. Ewing, L. Hillier, M. C. Wendl, P. Green, *Genome Research* **8**, 175 (1998).

26. S. L. Doty *et al.*, *Symbiosis* **39**, 27 (2005).

27. K. Germaine *et al.*, *Fems Microbiology Ecology* **48**, 109 (2004).

28. M. A. Marra *et al.*, *Genome Research* **7**, 1072 (1997).

29. J. D. McPherson *et al.*, *Nature* **409**, 934 (2001).

30. J. Schein *et al.*, *Methods Mol. Biol.* **255**, 143 (2004).

31. D. R. Fuhrmann *et al.*, *Genome Research* **13**, 940 (2003).

32. C. Soderlund, I. Longden, R. Mott, *Computer Applications in the Biosciences* **13**, 523 (1997).

33. C. Soderlund, S. Humphray, A. Dunham, L. French, *Genome Research* **10**, 1772 (2000).

34. C. D. Fjell, I. Bosdet, J. E. Schein, S. J. M. Jones, M. A. Marra, *Genome Research* **13**, 1244 (2003).

35. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, *Journal of Molecular Biology* **215**, 403 (1990).

36. K. A. Frazer *et al.*, *Nucleic Acids Research* **32**, W273 (2004).

37. M. Brudno *et al.*, *Genome Research* **13**, 721 (2003).

38. W. J. Kent, *Genome Research* **12**, 656 (2002).

39. S. Schwartz *et al.*, *Genome Research* **13**, 103 (2003).

40. D. C. Jewell, M. N. Islam-Faridi, in *The Maize Handbook*, M. Freeling, V. Walbot, Eds. (Springer-Verlag, New York, 1994), pp. 484-493.

41. K. L. Childs *et al.*, *Plant J.* **27**, 243 (2001).

42. A. A. Salamov, V. V. Solovyev, *Genome Research* **10**, 516 (2000).

43. E. Birney, R. Durbin, *Genome Research* **10**, 547 (2000).

44. Y. Xu, E. C. Uberbacher, *Journal of Computational Biology* **4**, 325 (1997).

45. T. Schiex, A. Moisan, P. Rouze, in *Computational Biology: selected papers from JOBIM'2000 number 2066 in LNCS*, Springer-Verlag, Ed. 2001), pp. 118-133.

46. The Gene Ontology Consortium, *Nature Genetics* **25**, 25 (2000).

47. E. V. Koonin *et al.*, *Genome Biology* **5**, (2004).

48. M. Kanehisa *et al.*, *Nucleic Acids Research* **32**, D277 (2004).

49. A. M. Brunner, S. P. DiFazio, Dharmawardhana, P., et al., unpublished data.

50. S. Chang, J. Puryear, J. Cairney, *Plant Molecular Biology Reporter* **11**, 113 (1993).

51. W. P. Hsieh, T. M. Chu, R. D. Wolfinger, G. Gibson, *Genetics* **165**, 747 (2003).

52. T. J. Chu, C. Glymour, R. Scheines, P. Spirtes, *Bioinformatics* **19**, 1147 (2003).

53. J. D. Storey, R. Tibshirani, *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440 (2003).

54. S. W. Clifton *et al.*, *Plant Physiology* **136**, 3486 (2004).

55. K. Hokamp, A. McLysaght, K. H. Wolfe, *J. Struct. Funct. Genomics* **3**, 95 (2003).

56. K. P. O'Brien *et al.*, *Nucleic Acids Research* **33**, D476 (2005).

57. S. Griffiths-Jones *et al.*, *Nucleic Acids Research* **33**, D121 (2005).

58. Arabidopsis Genome Initiative, *Nature* **408**, 796 (2000).

59. E. A. Prado *et al.*, *Genome* **39**, 1020 (1996).

60. M. W. Jones-Rhoades, D. P. Bartel, *Molecular Cell* **14**, 787 (2004).

61. M. J. Axtell, D. P. Bartel, *Plant Cell* **17**, 1658 (2005).

62. S. K. Floyd, J. L. Bowman, *Nature* **428**, 485 (2004).

63. J. Schrader *et al.*, *Plant Cell* **16**, 2278 (2004).

64. D. V. Dugas, B. Bartel, *Curr. Opin. Plant Biol.* **7**, 512 (2004).

65. W. Bao, D. M. O'Malley, R. Whetten, R. R. Sederoff, *Science* **260**, 672 (1993).

66. P. Ranocha *et al.*, *Plant Physiology* **129**, 145 (2002).

67. S. F. Lu *et al.*, *Plant Cell* **17**, 2186 (2005).

68. F. Sterky *et al.*, *Proceedings of the National Academy of Sciences of the United States of America* **101**, 13951 (2004).

69. Z. Yang, R. Nielsen, *Mol. Biol. Evol.* **17**, 32 (2000).

70. S. De Bodt, S. Maere, Y. Van de Peer, *Trends in Ecology & Evolution* **20**, 591 (2005).

71. D. R. Nelson, *Archives of Biochemistry and Biophysics* **369**, 1 (1999).

72. D. R. Nelson, M. A. Schuler, S. M. Paquette, D. Werck-Reichhart, S. Bak, *Plant Physiology* **135**, 756 (2004).

73. D. R. Nelson *et al.*, *Dna and Cell Biology* **12**, 1 (1993).

74. D. R. Nelson *et al.*, *Pharmacogenetics* **6**, 1 (1996).

75. S. M. Paquette, S. Bak, R. Feyereisen, *DNA and Cell Biology* **19**, 307 (2000).

76. D. B. S. P. S. M. Werck-Reichhart, in *The Arabidopsis Book*, C. R. Somerville, E. M. Meyerowitz, Eds. (American Society of Plant Biologists, Rockville, MD, 2002), pp. 1-28.

77. M. A. Schuler, D. Werck-Reichhart, *Annual Review of Plant Biology* **54**, 629 (2003).

78. E. M. Turk *et al.*, *Plant Journal* **42**, 23 (2005).

79. Z. R. Bao, S. R. Eddy, *Genome Research* **12**, 1269 (2002).

80. X. Zhang, S. R. Wessler, *Proc. Natl. Acad. Sci. U. S A* **101**, 5589 (2004).

81. Acknowledgments – The authors wish to thank U.S. Department of Energy, Office of Science for supporting the sequencing and assembly portion of this study, Genome Canada and the Province of British Columbia for providing support for the BAC end, BAC genotyping, and full-length cDNA portions of this study, the Swedish Agricultural University for supporting the EST assembly and annotation portion of this study, the membership of the International *Populus* Genome Consortium for supplying genetic and genomics resources used in the assembly and annotation of the genome, the National Science Foundation, Plant Genome Program for supporting the development of web-based tools, H.D. Bradshaw and R. Stettler for input and reviews on draft copies of the manuscript, J.M. Tuskan for guidance and input during the analysis and writing of the manuscript, and to the anonymous reviewers who provided critical input and recommendations on the manuscript. GenBank Accession Number: AARH00000000.

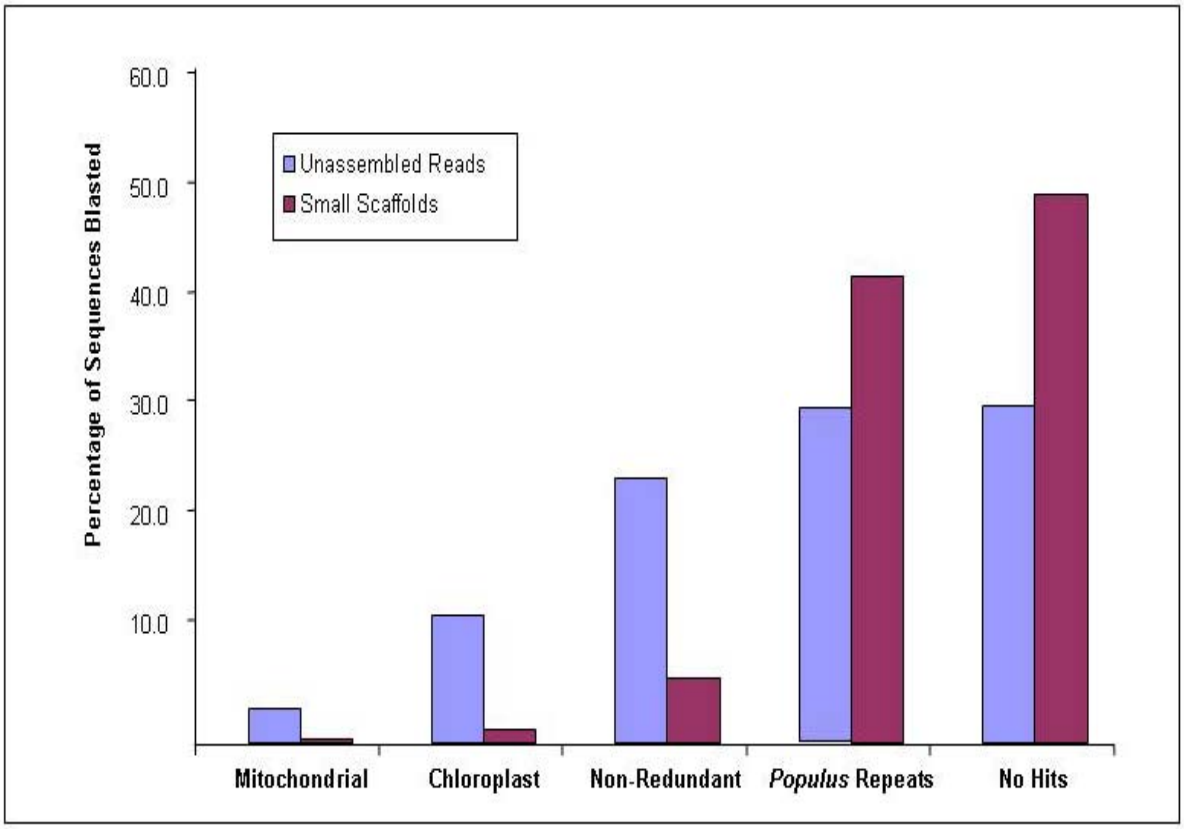


Figure S1. Putative origins of unassembled sequence reads and small scaffolds (<10 kb). Sequences were assigned to different categories based on wu-BLAST hits (E-value≤$1e^{-10}$) to databases containing the assembled *Populus trichocarpa* mitochondrion, chloroplast, a database of repeats identified by Recon and RepeatMasker or the non-redundant nucleotide database from NCBI.

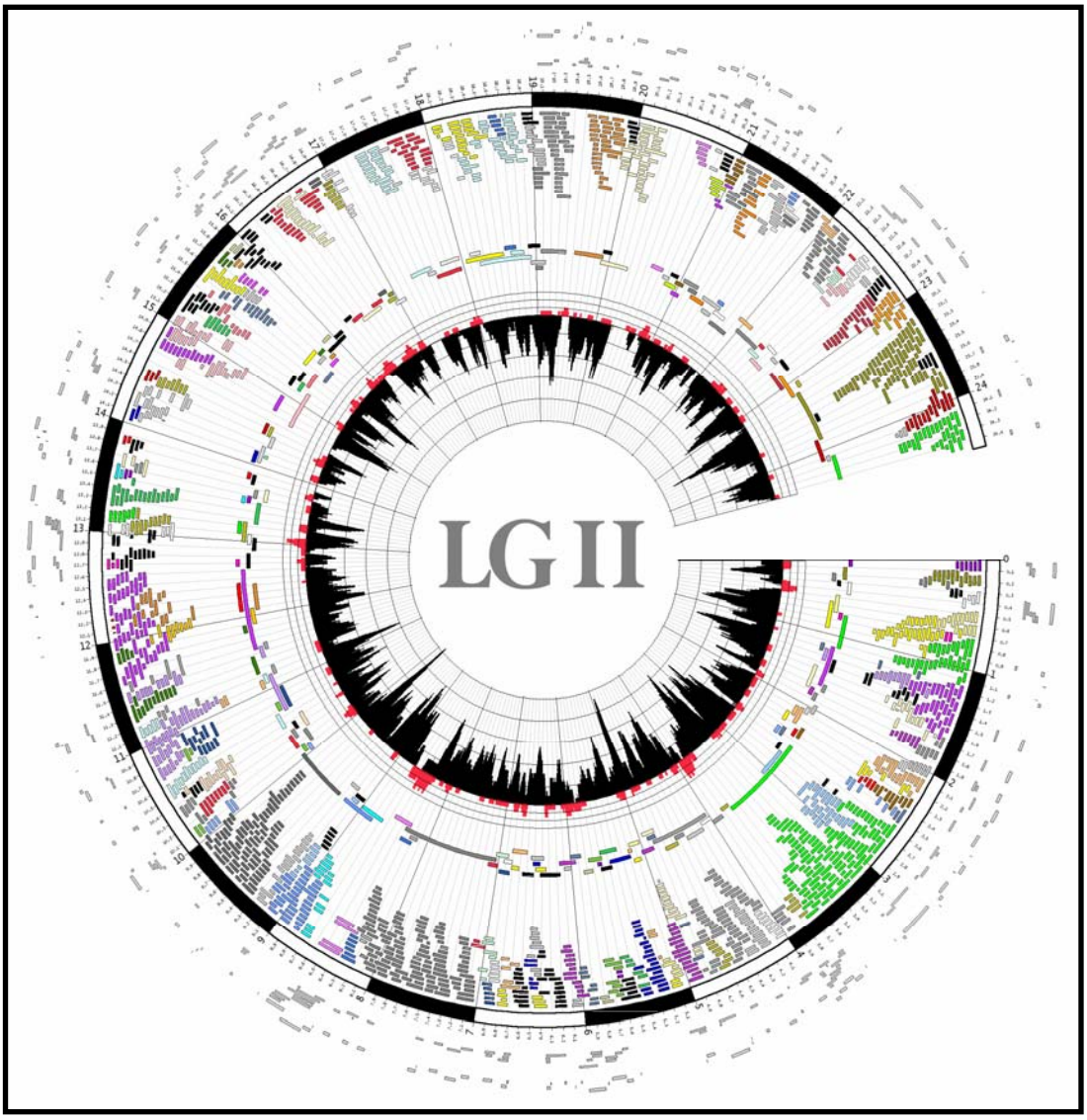


Figure S2. Fingerprint clone and contig layout on assembly of LGII. The ideogram of LGII is composed circularly, with 1 Mb spans colored in alternating black and white strips. The innermost histogram track (**black**) shows the fingerprint map clone coverage, with each concentric circle representing a 5X clone depth. The next outer histogram track (**red**) shows the coverage provided by fingerprint map clones not assigned to contigs (singletons). The next track shows the extent of anchored contigs, coded with an alternating color scheme. The final track inside the ideogram circle shows the sequence position of individual anchored clones in each contig, colored by map contig assignment. The first outer track shows the sequence position of clones that lack map contig assignment. The second outer track shows the coverage provided by the singletons.

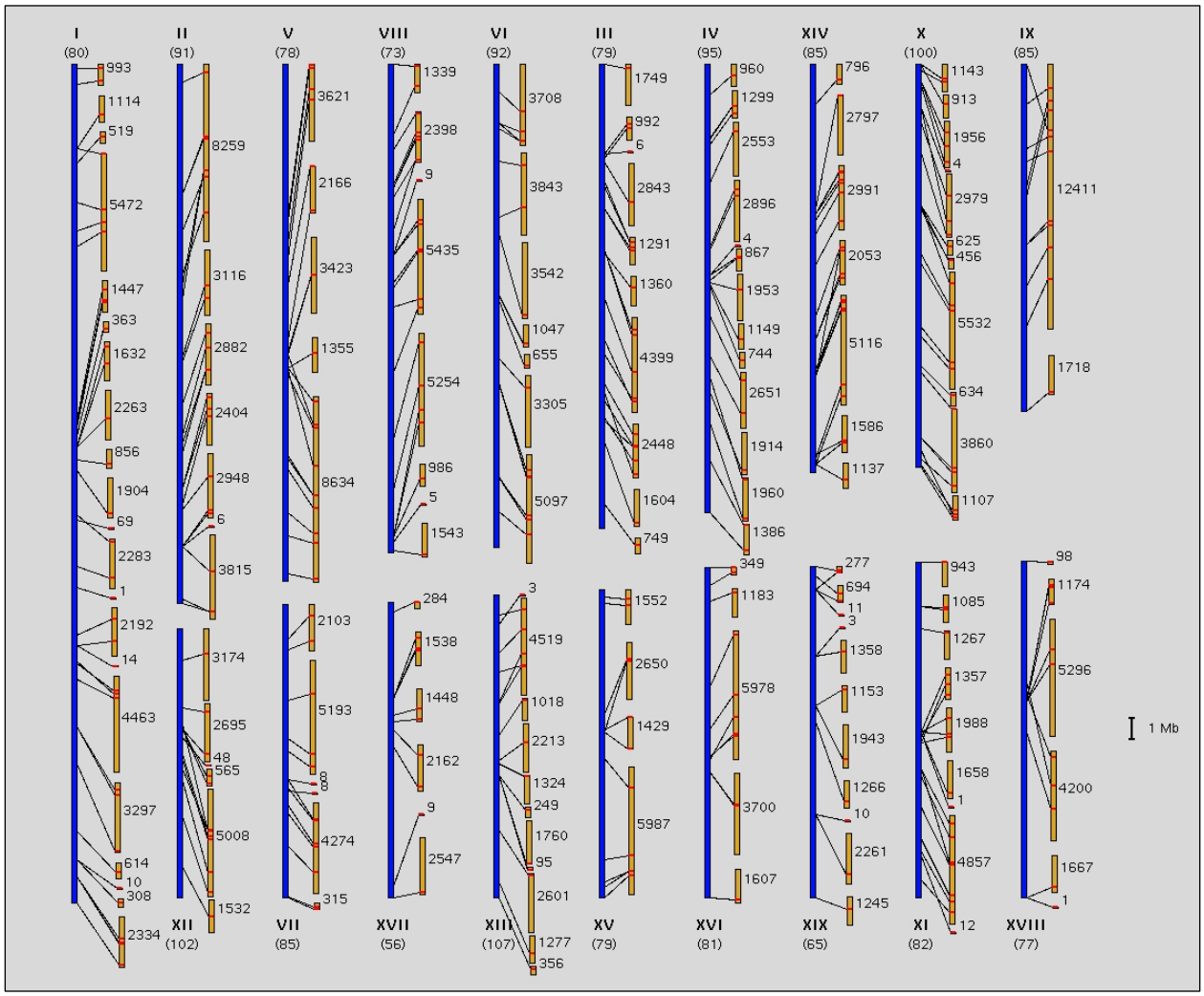


Figure S3. Representation of the 335 Mb of *Populus* genomic sequence contained in 155 scaffolds aligned and oriented to a genetic map of the 19 *Populus* linkage groups (indicated by Roman numerals **I**-**XIX**). Each scaffold (**yellow** bars) was mapped to a chromosome (**blue** bars) using microsatellite markers with unique sequence locations (**red** lines). Numbers in parentheses are estimates of the percent of the linkage group covered by assembled sequence (assuming uniform physical: genetic distance across the genome). Approximate size (in kb) is indicated to the right of each scaffold. Gaps between scaffolds are of unknown size. This assembly includes improvements since the version that was publicly released and used for most genome analyses.

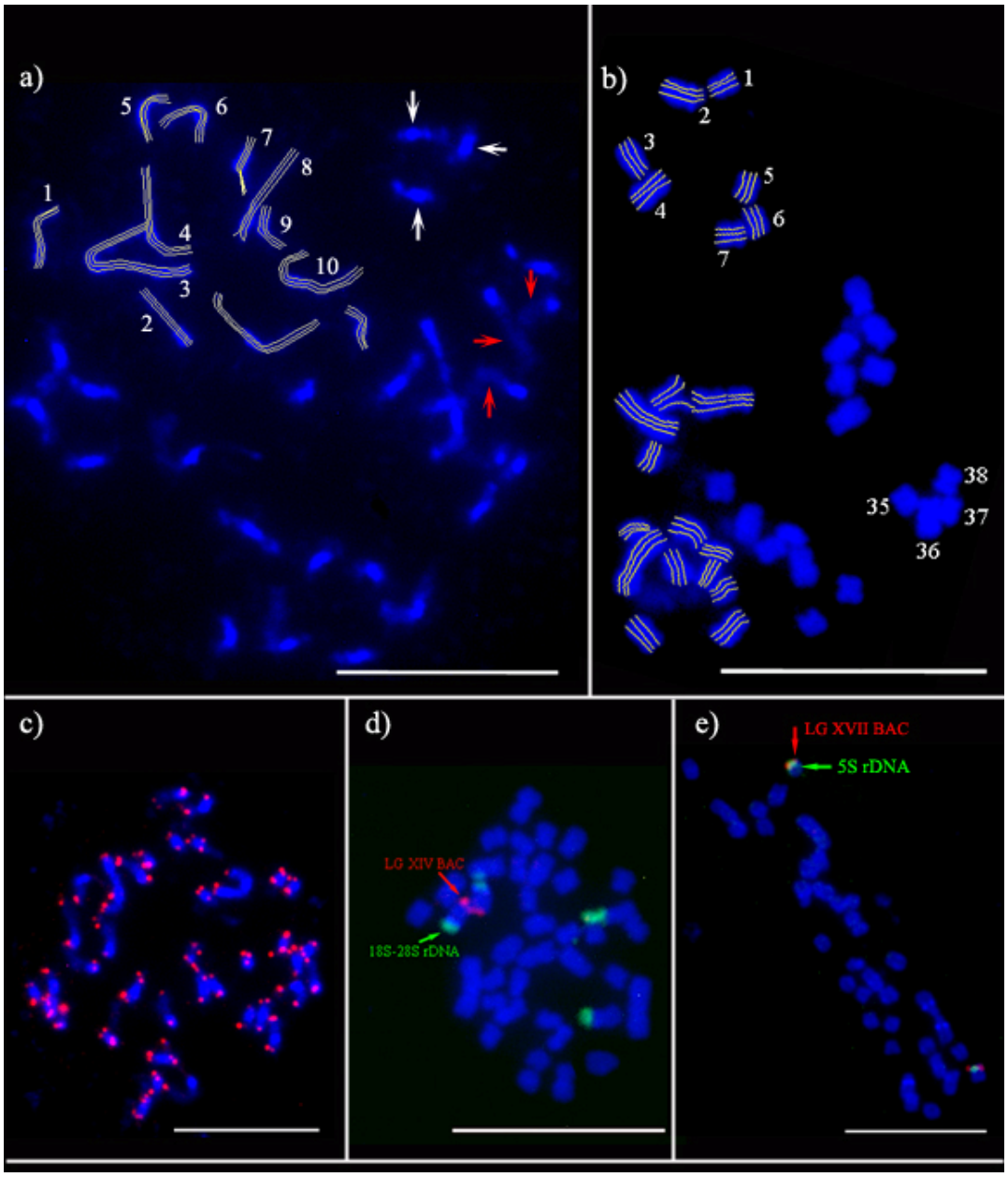


Figure S4. DAPI stained **a)** prophase and **b)** metaphase *Populus* somatic chromosomes and FISH using Arabidopsis-type telomere repeat sequence (A-type TRS), 18S-28S rDNA, 5S rDNA, and linkage group (LG) specific *Populus* BAC clones as probes. White and **red** arrows **a)** show heterochromatic (A-T rich, brightly stained) and euchromatic regions, respectively. For data collection, chromosomes were numbered arbitrarily from 1 to 38 in each cell and chromosome length was measured three times per chromosome (shown by white trace lines) using Optimas v6. **c)** A-type TRS FISH signals are observed at the end all chromosome arms, **d)** BACs from LGXIV are found to be co-localized with an 18S-28S rDNA site and **e)** BACs from LGXVII are found to be co-localized with the 5S rDNA site. Bar is 10 µm.

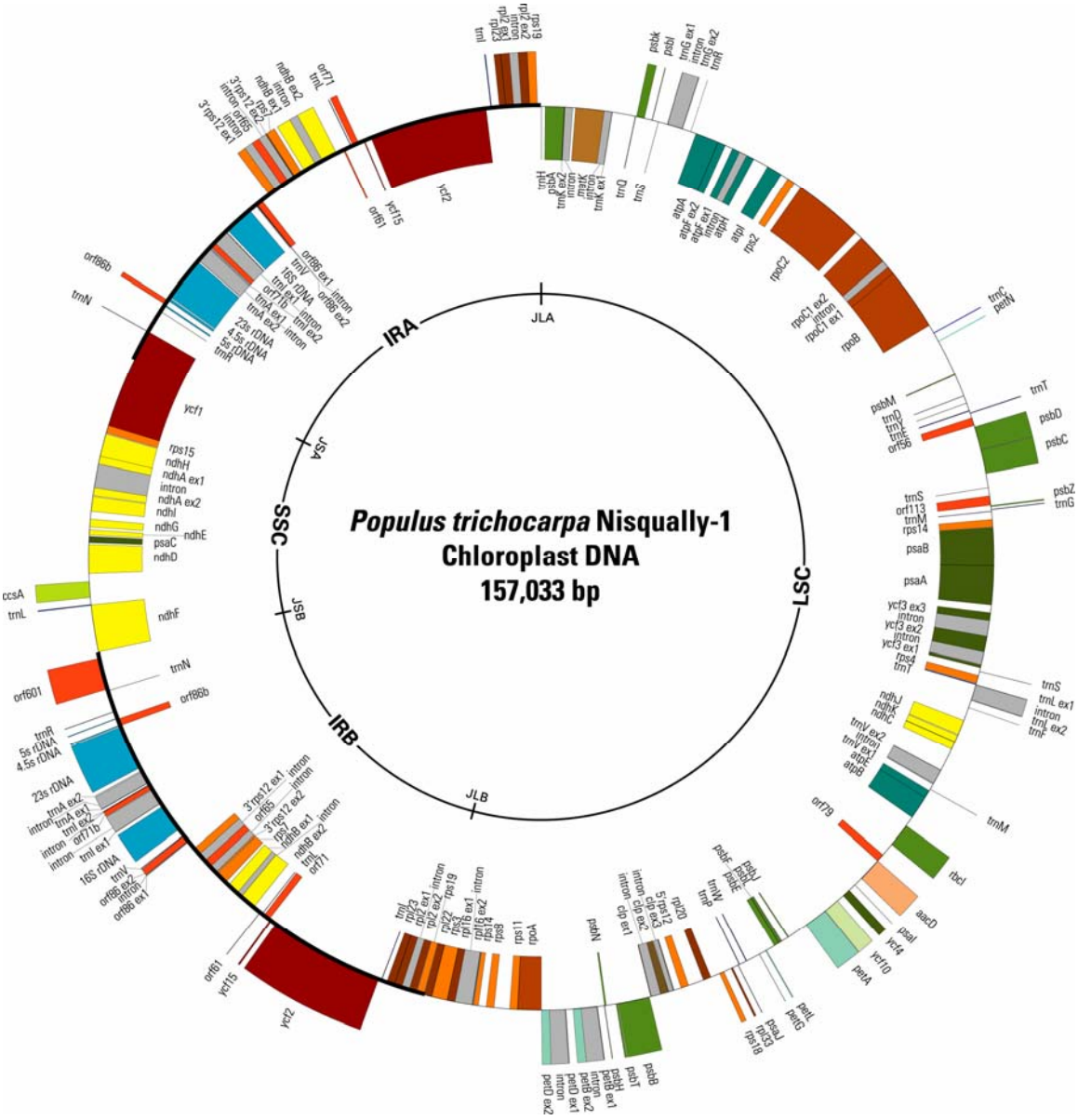


Figure S5. Graphic representation of the *de novo* whole-genome shotgun sequence assembly and annotation for the *Populus trichocarpa* chloroplast. Each nucleotide is represented by an average of 410 sequence reads at a quality score of 40 or higher. Gene models were predicted based on the Glimmer program at Oak Ridge Ntiaonal Laboratory.

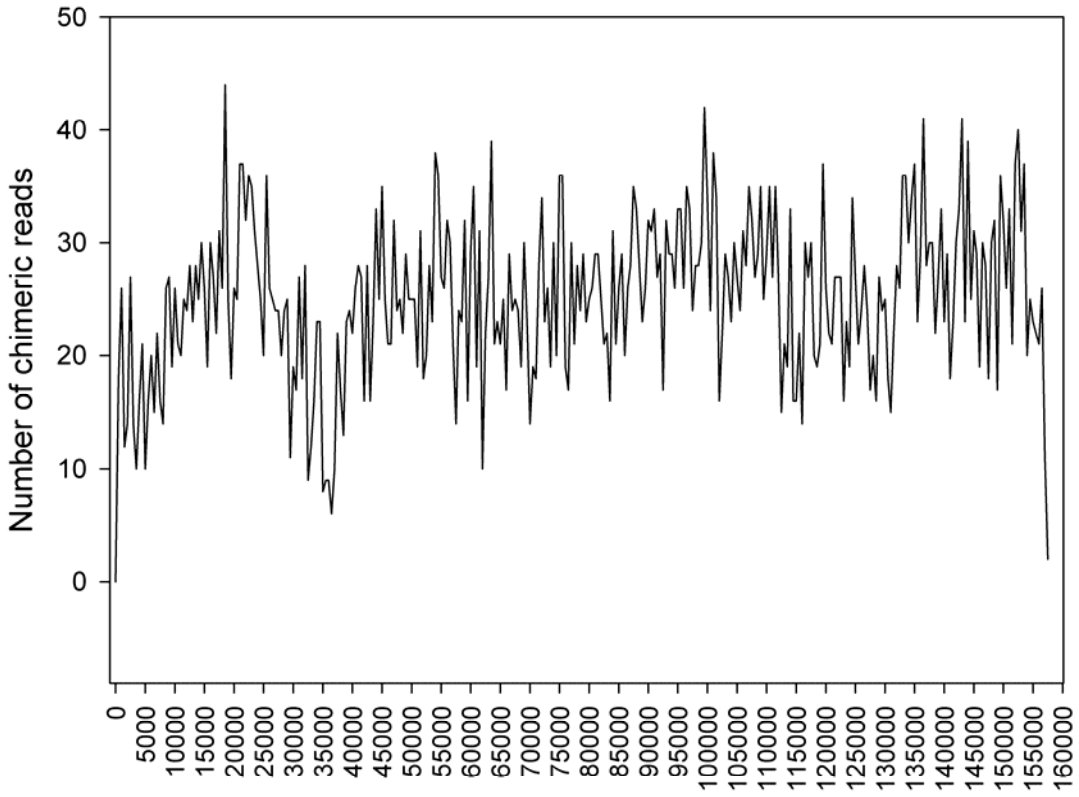


Figure S6. Frequency of chimeric reads across the *Populus* chloroplast genome. Chimeras were identified from as clones for which one end read matched the nucleus and the other matched the chloroplast.

34

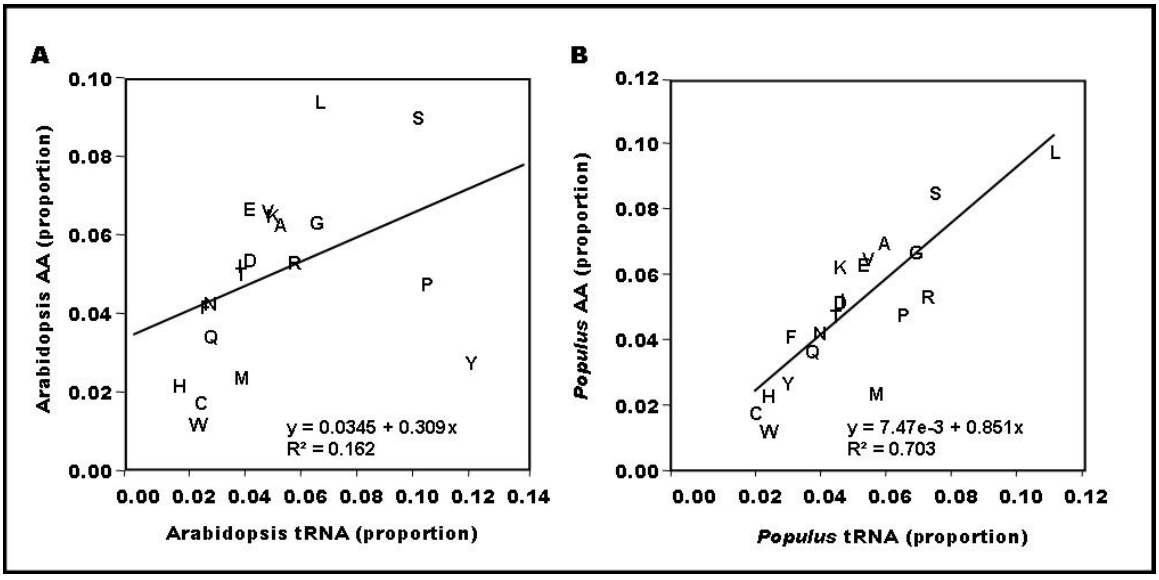


Figure S7. Relationship between amino acid abundance in the full set of predicted proteins and tRNA abundance in the **A**) Arabidopsis and **B**) *Populus* genomes. Single letters represent standard codes for amino acids. Equation and R-square values are from a simple, least square regression analysis.

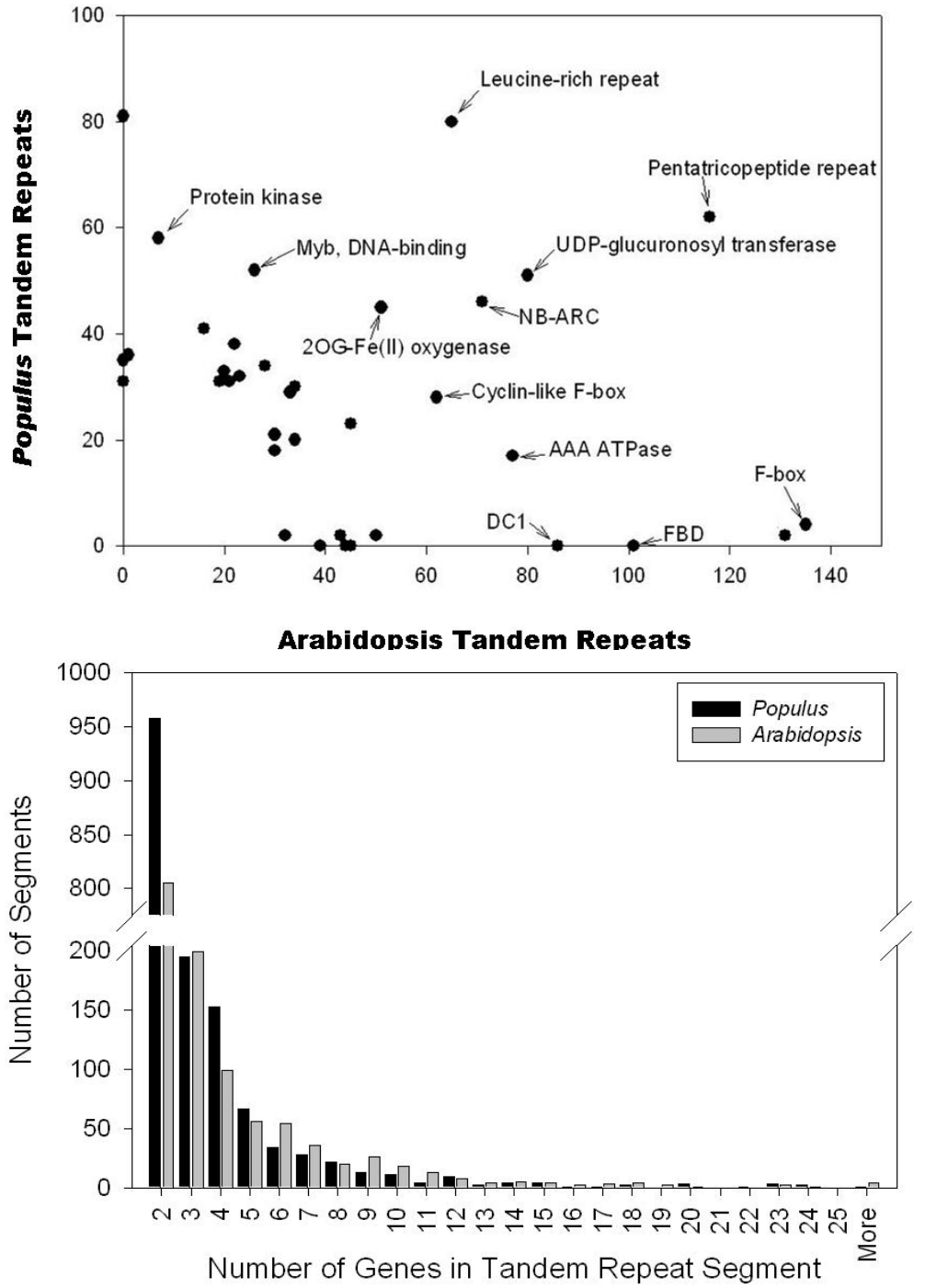
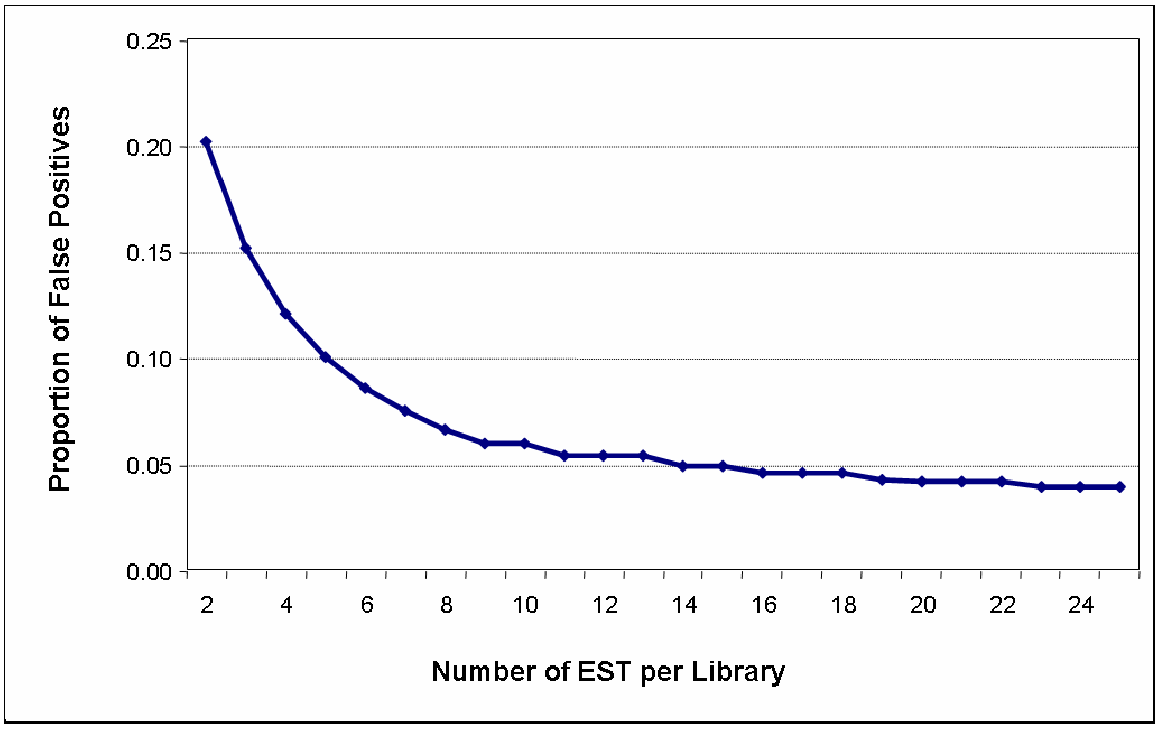


Figure S8. A) Frequency of InterPro domains for tandemly repeated genes in *Populus* and Arabidopsis. The serine/threonine protein kinase active site was highly abundant in both species (312 for *Populus* and 287 for Arabidopsis, primarily in S-locus genes) and are not shown on this figure. **B)** Distributions of total number of genes per 100 kb repeat segment in *Populus* and Arabidopsis. Genes were counted if they aligned with at least one other gene in the segment with a Smith-Waterman expectation score of E-value $\leq 1e^{-25}$ or less. Note the break in the axis.



24

Figure S9. Results of Monte-Carlo Simulation based on resampling different numbers of ESTs from the same gene, and comparing the resulting distributions using a Kolmogorov-Smirnov goodness of fit test.

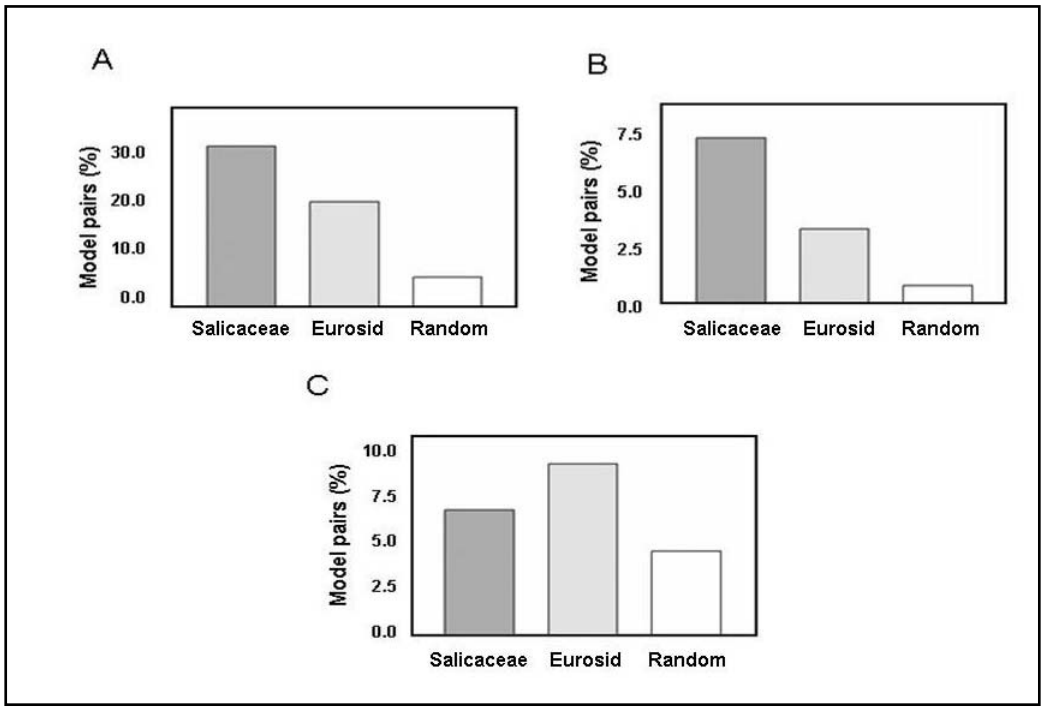


Figure S10. **A**) Models with less than a 2-fold difference in sequenced EST from various tissue-derived EST libraries. **B**) Models showing significant expression pattern correlations among tissue types/libraries. **C**) Models displaying significant over representation in one or more tissue libraries. (See (**68**) for descriptions of each tissue type/library).

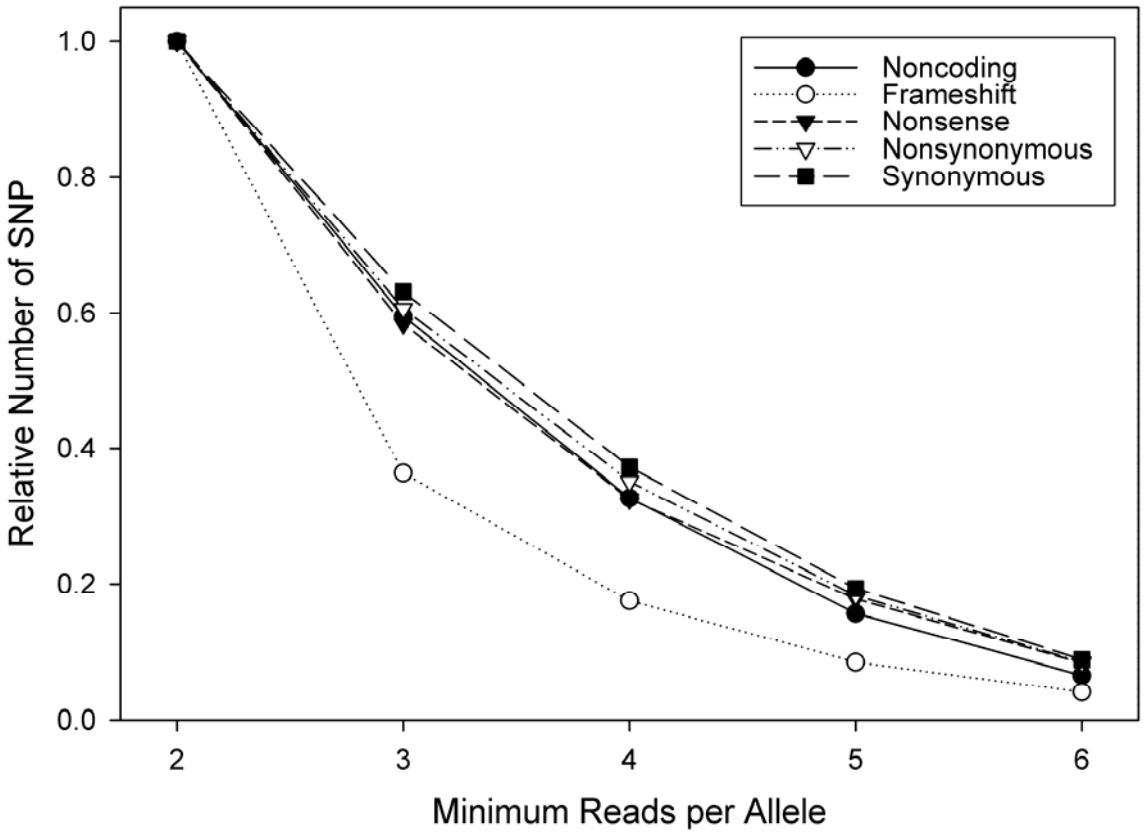


Figure S11. Minimum number of sequence reads per allele versus relative number of SNP detected in the *Populus* genome. Numbers are normalized by the maximum value observed (set to 1 for each category of SNP). Categories of SNP were determined based on positions relative to coding sequences. Raw values for 3 reads per allele are provided in the main manuscript.

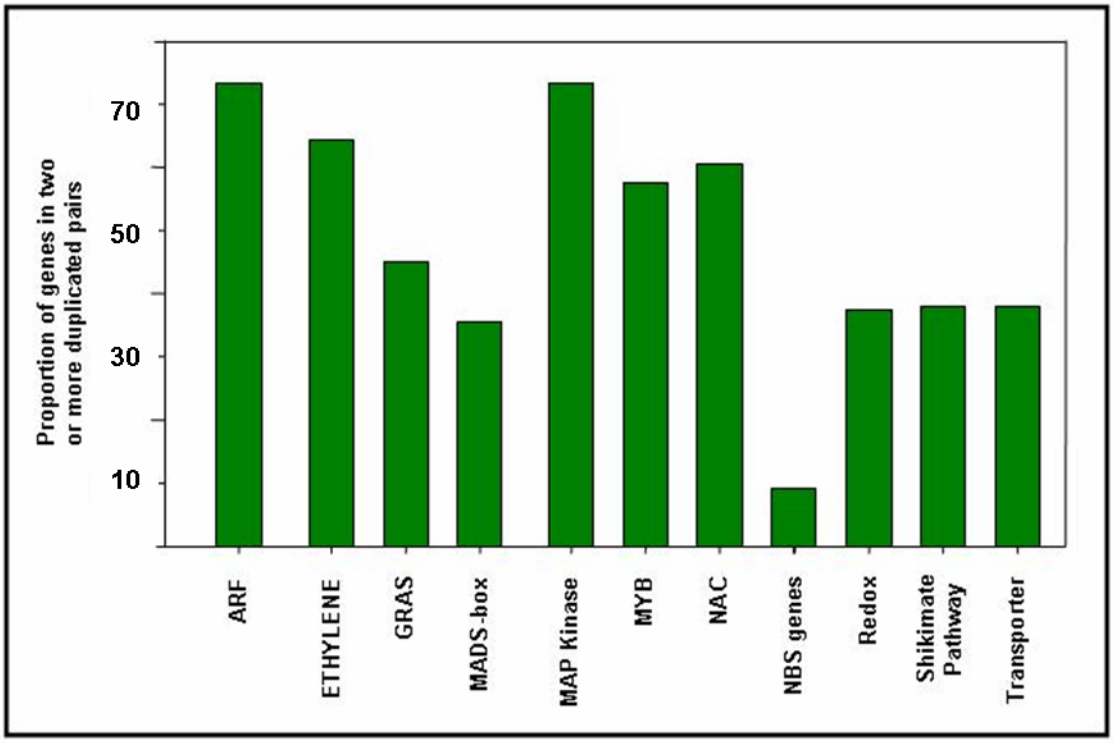


Figure S12. Comparative retention rates among various gene families following a recent genome wide duplication event. These rates reflect the comparatively high maintenance of duplicated genes in transcription factors relative to kinases, disease resistance, and general metabolism genes.

2

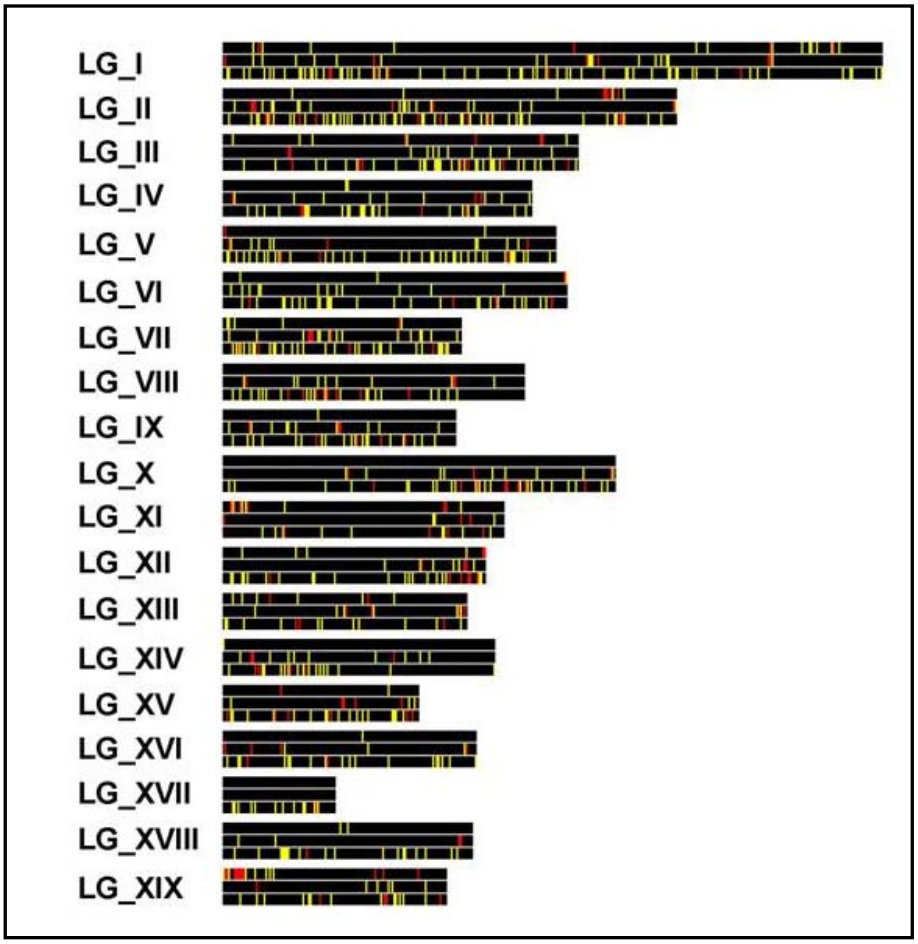


Figure S13. Chromosomal localization designated by linkage groups (LG), for disease resistance genes (top), genes coding for P450 enzymes (middle) and transcription factors (bottom). **Yellow** denotes a single gene in a 100 kb window, **red** 2 or more genes in a 100 bp window.

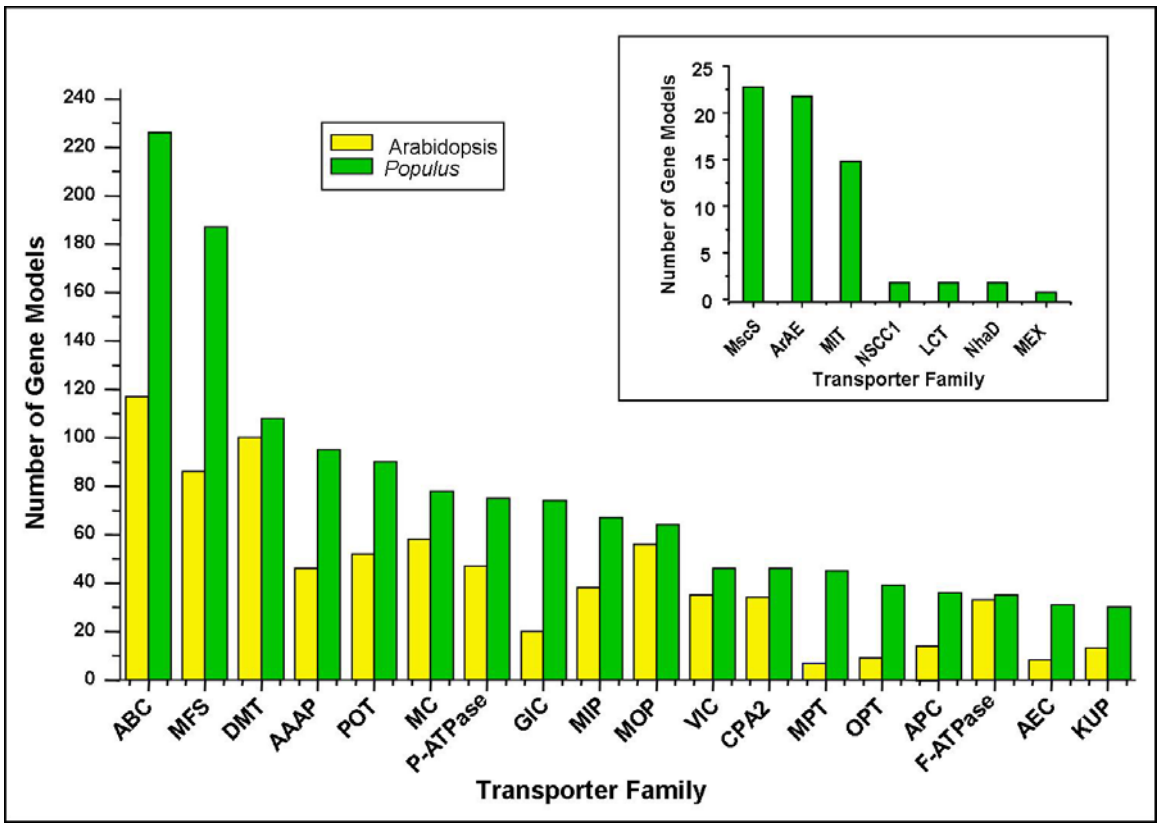


Figure S14. Comparative number of transporter gene models in *Populus* and Arabidopsis. Family name abbreviations are found in the "Supplemental Material" section and are based on the Transport Classification Database: http://www.tcdb.org/). Data from Arabidopsis can be found at the PlantsT site: (http://plantst.genomics.purdue.edu/). Data in the insert represents the number of gene models in *Populus* that do not have an Arabidopsis homolog.

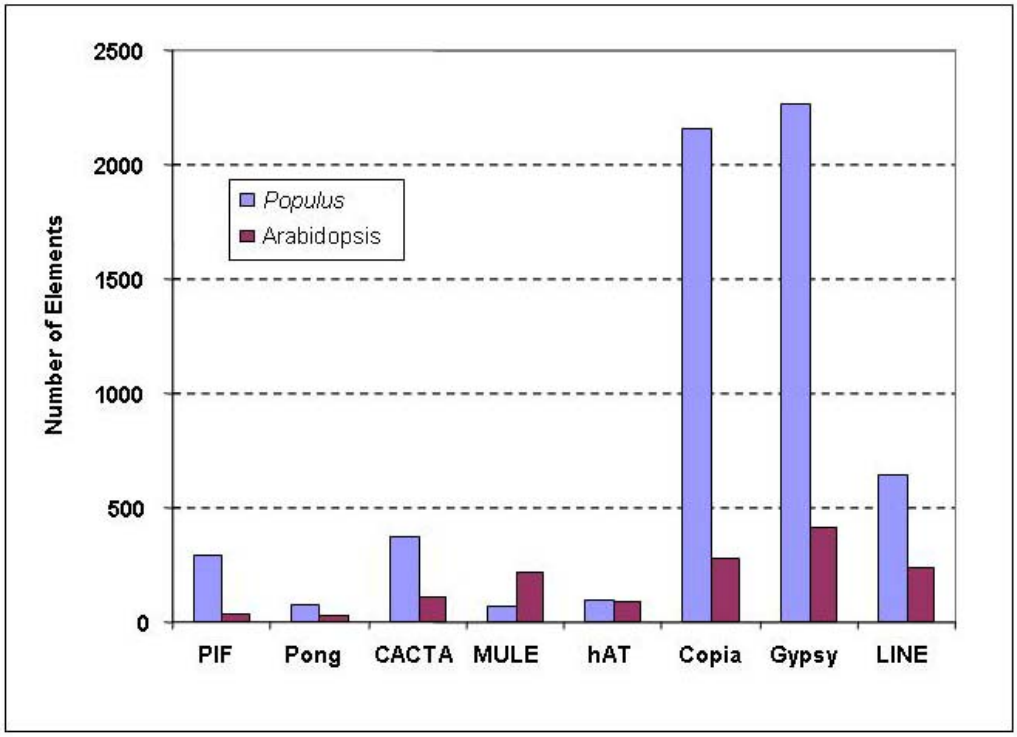


Figure S15. Comparative depiction of transposable elements found in *Populus* and Arabidopsis. Consensus amino acid sequences generated from previously identified coding regions of Arabidopsis and *Lotus japonicus* were used as queries in local TBLASTN searches against the available *Populus* database to identify all *Populus* homologs.

Table S1. Clones and end-read statistics for all sequenced insertion libraries used in the whole-genome shotgun *Populus trichocarpa* draft assembly.

			Reads (millions)		Bases(billions)	
Insert size (kb)	**Vector**	**Number of Libraries**	**All**	**Used[a]**	**Number ≥ Q20**	**Trimmed**
2.0-4.0	plasmid	4	4.45	2.75	2.76	1.73
4.5-7.5	plasmid	4	2.58	1.62	1.78	1.04
38-41	fosmid	3	0.65	0.43	0.41	0.30
Total		11	7.69	4.80	4.95	3.07

[a] Number of sequences included in *Populus* genome assembly version 1.0. This includes sequences included in scaffolds ≤ 1 kb that are not included in the statistics in Table S2.

Table S2. Sequencing and assembly statistics of the *Populus* draft sequence assembly.

Scaffolds	Reads (1000s)	Bases (Mb)	Scaffolds Number	Scaffolds N50 (kb)	Scaffolds Gaps (Mb)	Contigs Number	Contigs N50 (kb)	Contigs Gaps (Mb)
Anchored[a]	3,696	321	155	3,100	12	11,362	126	793
Unanchored >20 kb	690	93	682	389	25	8,245	29	127
Unanchored 1-20 kb	251	50	21,299	4	9	26,363	2	21

[a] Scaffolds anchored to the genetic map using sequence-tagged markers.

Table S3. Most frequently observed hits to the NR database among unassembled reads and small scaffolds (<10 kb), organized by Kingdom or Superkingdom (where Kingdom is undefined). Origin: A, possible plant associate; C, likely contaminant, U, unknown.

Kingdom	Genus	Species	Sequences	Origin
Archaea	Thermoplasma	2	13	U
Archaea	Picrophilus	1	11	U
Archaea	Haloarcula	1	1	U
Archaea	Methanosaeta	1	1	U
Archaea	Methanothermobacter	1	1	U
Bacteria	Ralstonia	7	2938	A
Bacteria	Escherichia	1	2031	C
Bacteria	Bordetella	3	1192	C
Bacteria	Cupriavidus	3	975	A
Bacteria	Xanthomonas	6	822	A
Bacteria	Pseudomonas	15	673	A
Bacteria	Rhodobacter	2	594	A
Bacteria	Burkholderia	10	558	A
Bacteria	Chromobacterium	1	359	A
Bacteria	Mesorhizobium	4	312	A
Bacteria	Acinetobacter	6	252	A
Bacteria	Acetobacter	1	212	A
Bacteria	Azoarcus	3	179	A
Bacteria	Bradyrhizobium	2	177	A
Bacteria	Sinorhizobium	1	163	A
Bacteria	Haemophilus	2	149	C
Bacteria	Pasteurella	1	135	C
Bacteria	Brucella	1	125	C
Bacteria	Caulobacter	1	115	A
Bacteria	Leptospira	1	22	U
Bacteria	Treponema	1	3	U
Bacteria	Borrelia	1	2	U
Fungi	Ustilago	1	91	A
Fungi	Gibberella	2	57	A
Fungi	Neurospora	1	40	U
Fungi	Crinipellis	1	35	U
Fungi	Suillus	4	29	A
Fungi	Schizophyllum	1	24	U
Fungi	Magnaporthe	1	20	A
Fungi	Armillaria	7	15	A
Fungi	Phanerochaete	1	15	U
Fungi	Agrocybe	1	14	U

Table S4. Putative location of telomeric repeats within the chromosome-scale assembly of the *Populus* genome. Positions were determined by BLASTN comparisons of consensus telomeric repeats with the assembled genome sequence. Because repeats were sometimes dispersed, all hits were totaled within 50 kb windows to enhance detection of putative telomere traces.

Linkage Group	LG Size (bp)	Position[1] (bp)	Hit Length[2]	Location
LGI	35,571,569	0	1019	End
		35,550,000	316	End
LGII	24,482,572	0	455	End
		24,450,000	792	End
LGIII	19,129,466	18,300,000	1180	End[3]
LGV	17,991,592	17,950,000	512	End
LGVI	18,519,121	0	1057	End
LGVII	12,805,987	0	437	End
		12,750,000	1099	End
LGVIII	16,228,216	0	1011	End
LGIX	12,525,049	0	245	End
LGX	21,101,489	0	1037	End
		19,900,000	429	End
LGXII	14,142,880	1,900,000	609	Interior
		3,300,000	908	Interior
		12,550,000	1525	End[3]
LGXIII	13,101,108	100,000	1115	End[4]
		8,550,000	1384	Interior
		10,700,000	9606	Interior
		13,050,000	209	End
LGXV	10,599,685	0	774	End
		10,550,000	1635	End
LGXVI	13,661,513	13,650,000	627	End
LGXVIII	13,470,992	0	732	End
LGXIX	12,003,701	10,650,000	1357	Interior
		11,600,000	387	Interior

[1] The beginning of a 50 kb window containing one or more hits.

[2] Total length of hits (bp) within 50 kb window (E-value $\leq 1e^{-10}$).

[3] End of linkage group contains a scaffold in apparently inverted orientation, because insufficient mapping information was available to determine true orientation.

[4] A small scaffold was erroneously mapped to the beginning of this linkage group in the initial assembly.

Table S5. Characteristics of gene models predicted using four different independently trained gene calling algorithms. The "Reference" set combines representative models (one per each locus) selected from all these predictions.

	EuGène	GrailEXP6	FgenesH(+)	Genewise	Reference Set
Total number of models	50,221	42,171	44,946	30,812	45,555
Gene length (bp)	2149	2212	2613	2687	2300
Exons per gene	4.0	3.5	4.8	5.0	4.3
Exon length (bp)	279	226	257	239	254
Intron length (bp)	351	569	363	380	379
Transcript length (bp)	1106	790	1232	1183	1079
CDS length (bp)	969	546	1161	1113	987
UTR length (bp)	137	244	71	70	92
Protein length (AA)	323	182	387	371	329
Models with NR hits	43,114 (86%)	32,718 (78%)	39,002 (87%)	30,758 (10%)	40,448 (89%)
Alignment coverage in model (%)	85	89	85	97	91
Alignment coverage in NR hit (%)	65	50	72	81	72
Models with EST support	13,926 (29%)	17,235 (41%)	10,646 (24%)	5,843 (19%)	10,092 (22%)

Table S6. Gene counts based on tribe analysis of *Populus* and Arabidopsis genes. Titles of columns and rows indicate the sizes of tribes in Arabidopsis and *Populus*, respectively.

		Arabidopsis Gene Count																
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**16**	**24**	**37**	**60**	**Total**
***Populus* Gene Count**	**1**	4607	844	99	25	7	3	1	2	2	3	1	1	1				5596
	2	2309	761	54	16	7	2	1	1	1	1	1	1	1	1			3157
	3	257	90	21	8	2	3	2	1	1	1	2	1	1	1	1	1	393
	4	80	33	8	6	1	1											129
	5	27	16	2	2	1	1	1	1									51
	6	14	3	1	3	2	2	1	1									27
	7	6	3	1	1	1												12
	8	2	4	1	2	1	1	1	1									13
	9	2	1	2														5
	10	3	2	1	1	1												8
	11	4	2															6
	12	1	1	1	1	1	1											6
	13	1	2	1	1													5
	14	1	1															2
	15	1																1
	16	1																1
	17	1																1
	18	1																1
	20	1																1
	21	1	1	1														3
	24	1	1	1														3
	27	1																1
	30	1																1
Total		7323	1765	194	66	24	14	7	7	4	5	4	3	3	2	1	1	9423

Table S7. Predicted regulatory targets of conserved miRNAs.

miRNA Family	**Target Family**	**Arabidopsis**	***Oryza***	***Populus***
miR156	*SBP-like transcription factors*	11	9	16
miR159/319	*MYB transcription factors*	8	6	5
miR159/319	*TCP transcription factors*	5	4	7
miR160	*Auxin Response Factors*	3	5	9
miR164	*NAC domain transcription factors*	6	6	6
miR166	*HD-Zip transcription factors*	5	4	9
miR167	*Auxin Response Factors*	2	4	7
miR169	*CCAAT binding factors (HAP2-like)*	8	7	9
miR171	*SCARECROW-like transcription factors*	3	5	9
miR172	*APETELA2-like transcription factors*	6	5	6
miR393	*bZIP transcription factors**	1	1	1
miR396	*Growth Regulating Factor*	7	9	9
	Total, transcription factors	65	65	93
miR162	*DICER-LIKE1*	1	1	1
miR168	*ARGONAUTE*	1	6	2
miR393	*F-box proteins*	4	2	5
miR394	*F-box proteins*	1	1	2
miR395	*ATP sulfurylases*	3	1	2
miR395	Sulfate transporters	1	2	3
miR396	Rhodenase-like proteins	1	1	1
miR397	*Laccases*	3	15	26
miR398	*Copper superoxide dismutases**	2	2	2
miR398	*Cytochrome C oxidases**	1	1	0
miR399	Phosphate transporters	1	4	4
miR399	E2 Ubiquitin conjugating enzymes	1	1	2
miR408	Laccases	3	2	3
	Total, non-transcription factors	23	39	53

The number of genes predicted to be targets of each miRNA family in three plant species is listed. Target families listed in *italics* have been confirmed experimentally in Arabidopsis. To be counted, a potential target must contain a complementary site to at least one member of the indicated miRNA family with a score of three or less, with the exception of the target families marked with *****, for which some targets with more relaxed complementarity were included.

Table S8. Number of conserved plant miRNA families in three plant species. The number of identified genes in each family of miRNAs is indicated, only miRNA families for which members could be identified in the *Populus* genome are listed.

miRNA family	**Arabidopsis**	***Oryza***	***Populus***
miR156	12	12	11
miR159/319	6	8	15
miR160	3	6	8
miR162	2	2	3
miR164	3	5	6
miR166	9	12	17
miR167	4	9	8
miR168	2	2	2
miR169	14	17	32
miR171	4	7	10
miR172	5	3	9
miR390	2	1	4
miR393	2	2	4
miR394	2	1	2
miR395	6	19	10
miR396	2	3	7
miR397	2	2	3
miR398	3	2	3
miR399	6	11	12
miR403	1	0	2
miR408	1	1	1
Total	**91**	**125**	**169**

Table S9. Frequency of InterPro domains in tandemly repeated genes in the *Populus* and Arabidopsis genomes. (Excel File)

Table S10. Comparison of Least-Squared Means from an ANOVA of genes from different duplication epochs.

Duplication Epoch					
Tandem	**Salicoid**	**Eurosid**	**N**	**Mean**	**Groups[1]**
1	1	0	302	0.6569	A
1	0	0	291	0.6532	AB
1	1	1	21	0.5931	ABC
0	0	1	93	0.5534	ABCD
0	0	0	1428	0.5214	ABCD
1	0	1	14	0.4579	BCD
0	1	0	1584	0.4425	CD
0	1	1	226	0.3934	D

[1] Uniform groups determined by Duncan's Multiple Range Test of Least Square Means in SAS.

Table S11. ANOVA of ω versus an indicator variable for the presence of a homeolog from the salicoid duplication (D), with synonymous substitution rate (d_S), and size of the predicted coding sequence (S) as covariates. Distance from the closest paralog (Min4DTV). All interactions were also tested, but only significant effects are presented.

Source	d.f.	Sum of Squares	Mean Square	F-value	Pr > F
Model	6	164.21	27.36	109.43	<.0001
Error	3950	987.93	0.250		
Corrected Total	3956	1152.14			
	R^2	Coeff. Var.	Root MSE	Mean	
	0.14	99.36	0.5001	0.5032	
Source	d.f.	Type III SS	MS	F-value	Pr > F
Duplication (D)	1	12.46	12.46	49.85	<.0001
Size (S)	1	13.07	13.07	52.29	<.0001
S*D	1	5.251	5.251	20.99	<.0001
d_S	1	25.58	25.58	102.31	<.0001
D*d_S	1	7.759	7.759	31.02	<.0001
S*d_S	1	6.361	6.361	25.44	<.0001

Table S12. Number of genes in membrane transporter families in *Populus* and Arabidopsis.

TC # - Transporter Families	Arabidopsis	*Populus*
1. Ion Channels (%)	13	16
Voltage-gated Ion Channel (VIC)	35	46
Major Intrinsic Protein (MIP)	38	67
Glutamate-gated Ion Channel (GIC)	20	74
Chloride Channel (ClC)	7	12
Non-selective Cation Channel-1 (NSCC1)	--	2
Chloroplast Envelope Anion Channel-forming (Tic110)	1	2
gp91phox Phagocyte NADPH Oxidase-associated Cytochrome b558 H^+-channel	18	19
Small Conductance Mechanosensitive Ion Channel (MscS)	--	23
CorA Metal Ion Transporter (MIT)	--	15
Mitochondrial and Plastid Porin (MPP)	6	14
Total	**125**	**274**
2. Secondary Transporter (%)	65	61
Major Facilitator (MFS)	86	187
Glycoside-Pentoside-Hexuronide (GPH):Cation Symporter	9	6
Amino Acid-Polyamine-Organocation (APC)	14	36
Cation Diffusion Facilitator (CDF)	12	18
Zinc (Zn^{2+})-Iron (Fe^{2+}) Permease (ZIP)	14	22
Resistance-Nodulation-Cell Division (RND)	2	3
Drug/Metabolite Transporter (DMT)	100	108
Cytochrome Oxidase Biogenesis (Oxa1)	5	11
ATP:ADP Antiporter (AAA)	2	2
Telurite-resistance/Dicarboxylate Transporter (TDT)	4	8
Proton-dependent Oligopeptide Transporter (POT)	52	90
Amino Acid/Auxin Permease (AAAP)	46	95
Ca^{2+}:Cation Antiporter (CaCA)	12	20
Inorganic Phosphate Transporter (PiT)	1	2
Solute:Sodium Symporter (SSS)	1	1
Bile Acid:Na^+ Symporter (BASS)	5	8
Mitochondrial Carrier (MC)	58	78
Cation-Chloride Cotransporter (CCC)	1	2
Anion Exchanger (AE)	7	10
Monovalent Cation:Proton Antiporter-1 (CPA1)	8	9
Monovalent Cation:Proton Antiporter-2 (CPA2)	34	46
K^+ Transporter (Trk)	1	2
Nucleobase:Cation Symporter-2 (NCS2)	11	17
Lysosomal Cystine Transporter (LCT)	--	2
Divalent Anion:Na^+ Symporter (DASS)	6	5
Ammonium Transporter (Amt)	6	14
Glycerol Uptake (GUP)	1	1
Sulfate Permease (SulP)	12	22
Metal Ion (Mn^{2+}-iron) Transporter (Nramp)	7	8
Equilibrative Nucleoside Transporter (ENT)	7	11
Organo Anion Transporter (OAT)	2	2
NhaD Na^+:H^+ Antiporter (NhaD)	--	2
Multidrug/Oligosaccharidyl-lipid/Polysaccharide (MOP) Flippase (inc. MATE)	56	64
Oligopeptide Transporter (OPT)	9	39
Auxin Efflux Carrier (AEC)	8	31
Folate-Biopterin Transporter (FBT)	9	14

K^+ Uptake Permease (KUP)	13	30
Chloroplast Maltose Exporter (MEX)	--	1
Aromatic Acid Exporter (ArAE)	--	22
Total	**621**	**1049**
3. ATP-Dependent (%)	22	23
ATP-binding Cassette (ABC)	117	226
H^+- or Na^+-translocating F-type V-type and A-type ATPase (F-ATPase)	33	35
P-type ATPase (P-ATPase)	47	75
General Secretory Pathway (Sec)	6	8
Mitochondrial Protein Translocase (MPT)	7	45
H^+-translocating Pyrophosphatase (H^+-PPase)	3	10
Total	**213**	**399**
Total Transporter Proteins	**959**	**1722**

Table S13. Annotated *Populus* phenylpropanoid metabolite biosynthesis genes, including lignin.

Gene Name[1]	Reference gene model[2]	Protein name
PAL1	estExt_Genewise1_v1.C_280658	Phenylalanine amonnia lyase
PAL2	estExt_fgenesh4_pg.C_LGVIII0293	Phenylalanine amonnia lyase
PAL3	grail3.0004045401	Phenylalanine amonnia lyase
PAL4	estExt_fgenesh4_pg.C_LGX2023	Phenylalanine amonnia lyase
PAL5	gw1.X.2713.1	Phenylalanine amonnia lyase
C4H1	estExt_fgenesh4_pg.C_LGXIII0519	Trans-cinnamate 4-monooxygenase
C4H2	grail3.0094002901	Trans-cinnamate 4-monooxygenase
C4H3	eugene3.01640067	Trans-cinnamate 4-monooxygenase
4CL1	estExt_fgenesh4_pg.C_1210004	4-Coumarate:CoA Ligase
4CL2	gw1.XVIII.2818.1	4-Coumarate:CoA Ligase
4CL3	grail3.0100002702	4-Coumarate:CoA Ligase
4CL4	grail3.0099003002	4-Coumarate:CoA Ligase
4CL5	fgenesh4_pg.C_LGIII001773	4-Coumarate:CoA Ligase
C3H1	eugene3.36160002	p-Coumaroyl shikimate 3'-hydroxylase/Coumaroyl 3-hydroxylase
C3H2	eugene3.00160247	p-Coumaroyl shikimate 3'-hydroxylase/Coumaroyl 3-hydroxylase
C3H3	estExt_fgenesh4_pm.C_LGVI0096	p-Coumaroyl shikimate 3'-hydroxylase/Coumaroyl 3-hydroxylase
F5H1	estExt_fgenesh4_pm.C_570058	Coniferylaldehyde 5-hydroxylase/Ferulate 5-hydroxylase
F5H2	eugene3.00071182	Coniferylaldehyde 5-hydroxylase/Ferulate 5-hydroxylase
CCR1	estExt_fgenesh4_pg.C_2080034	Cinnamoyl CoA reductase
CCR2	estExt_fgenesh4_kg.C_LGIII0056	Cinnamoyl CoA reductase
CCR3	fgenesh4_pg.C_scaffold_208000040	Cinnamoyl CoA reductase
CCR4	eugene3.02080031	Cinnamoyl CoA reductase
CCR5	estExt_fgenesh4_pg.C_2080041	Cinnamoyl CoA reductase
CCR6	estExt_fgenesh4_pg.C_LGI0389	Cinnamoyl CoA reductase
CCR7	gw1.I.7401.1	Cinnamoyl CoA reductase
CAD	estExt_Genewise1_v1.C_LGIX2359	Cinnamyl alcohol dehydrogenase
CCOMT1	grail3.0001059501	Trans-caffeoyl-CoA 3-O-methyltransferase
CCOMT2	estExt_fgenesh4_pm.C_LGI1023	Trans-caffeoyl-CoA 3-O-methyltransferase
COMT1	estExt_fgenesh4_pg.C_LGXV0035	Caffeic acid 3-O-methyltransferase
COMT2	estExt_fgenesh4_pm.C_LGXII0129	Caffeic acid 3-O-methyltransferase

HCT1	eugene3.00031532	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT2	estExt_fgenesh4_pm.C_LGXVIII0344	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT3	estExt_fgenesh4_pg.C_LGXVIII0910	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT4	eugene3.00180947	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT5	eugene3.18780002	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT6	eugene3.02080010	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase
HCT7	fgenesh4_pg.C_scaffold_133000007	Hydroxycinnamoyl CoA shikimate/quinate hydroxycinnamoyltransferase

[1] Proposed gene designation, to be preceded by "Poptr", *e.g.,* PoptrPAL1

[2] Gene model name in the *Populus trichocarpa* Genome Browser v. 1.1 (http://genome.jgi-psf.org/Poptr1/Poptr1.home.html)

Table S14. Characterization of repeat elements in the assembled *Populus* genome. Elements were first identified by Recon analysis of all-versus-all BLAST comparisons of the assembled genome, followed by BLAST comparisons to known repetitive elements in RepBase. Elements were then delineated in the assembled genome using RepeatMasker with wu-BLAST.

Type	Sum of Length	Count of Class
DNA/Cacta	5,566,968	70
DNA/Centromeric	256,360	4
DNA/Ds	94,695	22
DNA/Helitron	311,669	2
DNA/JT	994,325	12
DNA/MuLE	2,249,544	4
DNA/PIF	850,943	9
DNA/Pong	4,414	1
LINE	2,338,298	54
Low_complexity	8,010,929	2
LTR/Copia	7,750,064	224
LTR/Gypsy	23,884,303	435
Retroelement	100,969	9
rRNA	107,742	4
Simple_repeat	3,283,925	153
Unknown	125,419,489	11,456
Total	181,224,637	12,461