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

G. A. Tuskan,[1,3]* S. DiFazio,[1,4]† S. Jansson,[5]† J. Bohlmann,[6]† I. Grigoriev,[9]† U. Hellsten,[9]† N. Putnam,[9]† S. Ralph,[6]† S. Rombauts,[10]† A. Salamov,[9]† J. Schein,[11]† L. Sterck,[10]† A. Aerts,[9] R. R. Bhalerao,[5] R. P. Bhalerao,[12] D. Blaudez,[13] W. Boerjan,[10] A. Brun,[13] A. Brunner,[14] V. Busov,[15] M. Campbell,[16] J. Carlson,[17] M. Chalot,[13] J. Chapman,[9] G.-L. Chen,[2] D. Cooper,[6] P. M. Coutinho,[19] J. Couturier,[13] S. Covert,[20] Q. Cronk,[7] R. Cunningham,[1] J. Davis,[22] S. Degroeve,[10] A. Déjardin,[23] C. dePamphilis,[18] J. Detter,[9] B. Dirks,[24] I. Dubchak,[9,25] S. Duplessis,[13] J. Ehlting,[7] B. Ellis,[6] K. Gendler,[26] D. Goodstein,[9] M. Gribskov,[27] J. Grimwood,[28] A. Groover,[29] L. Gunter,[1] B. Hamberger,[7] B. Heinze,[30] Y. Helariutta,[12,31,33] B. Henrissat,[19] D. Holligan,[21] R. Holt,[11] W. Huang,[9] N. Islam-Faridi,[34] S. Jones,[11] M. Jones-Rhoades,[35] R. Jorgensen,[26] C. Joshi,[15] J. Kangasjärvi,[32] J. Karlsson,[5] C. Kelleher,[6] R. Kirkpatrick,[11] M. Kirst,[22] A. Kohler,[13] U. Kalluri,[1] F. Larimer,[2] J. Leebens-Mack,[21] J.-C. Leplé,[23] P. Locascio,[2] Y. Lou,[9] S. Lucas,[9] F. Martin,[13] B. Montanini,[13] C. Napoli,[26] D. R. Nelson,[36] C. Nelson,[37] K. Nieminen,[31] O. Nilsson,[12] V. Pereda,[13] G. Peter,[22] R. Philippe,[6] G. Pilate,[23] A. Poliakov,[25] J. Razumovskaya,[2] P. Richardson,[9] C. Rinaldi,[13] K. Ritland,[8] P. Rouzé,[10] D. Ryaboy,[25] J. Schmutz,[28] J. Schrader,[38] B. Segerman,[5] H. Shin,[11] A. Siddiqui,[11] F. Sterky,[39] A. Terry,[9] C.-J. Tsai,[15] E. Uberbacher,[2] P. Unneberg,[39] J. Vahala,[32] K. Wall,[18] S. Wessler,[21] G. Yang,[21] T. Yin,[1] C. Douglas,[7]‡ M. Marra,[11]‡ G. Sandberg,[12]‡ Y. Van de Peer,[10]‡ D. Rokhsar[9,24]‡

We report the draft genome of the black cottonwood tree, *Populus trichocarpa*. Integration of shotgun sequence assembly with genetic mapping enabled chromosome-scale reconstruction of the genome. More than 45,000 putative protein-coding genes were identified. Analysis of the assembled genome revealed a whole-genome duplication event; about 8000 pairs of duplicated genes from that event survived in the *Populus* genome. A second, older duplication event is indistinguishably coincident with the divergence of the *Populus* and *Arabidopsis* lineages. Nucleotide substitution, tandem gene duplication, and gross chromosomal rearrangement appear to proceed substantially more slowly in *Populus* than in *Arabidopsis*. *Populus* has more protein-coding genes than *Arabidopsis*, ranging on average from 1.4 to 1.6 putative *Populus* homologs for each *Arabidopsis* gene. However, the relative frequency of protein domains in the two genomes is similar. Overrepresented exceptions in *Populus* include genes associated with lignocellulosic wall biosynthesis, meristem development, disease resistance, and metabolite transport.

Forests cover 30% (about 3.8 billion ha) of Earth's terrestrial surface, harbor substantial biodiversity, and provide humanity with benefits such as clean air and water, lumber, fiber, and fuels. Worldwide, one-quarter of all industrial feedstocks have their origins in forest-based resources (*1*). Large and long-lived forest trees grow in extensive wild populations across continents, and they have evolved under selective pressures unlike those of annual herbaceous plants. Their growth and development involves extensive secondary growth, coordinated signaling and distribution of water and nutrients over great distances, and strategic storage and redistribution of metabolites in concordance with interannual climatic cycles. Their need to survive and thrive in fixed locations over centuries under continually changing physical and biotic stresses also sets them apart from short-lived plants. Many of the features that distinguish trees from other organisms, especially their large sizes and long-generation times, present challenges to the study of the cellular and molecular mechanisms that underlie their unique biology. To enable and facilitate such investigations in a relatively well-studied model tree, we describe here the draft genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray), and compare it to other sequenced plant genomes.

P. trichocarpa was selected as the model forest species for genome sequencing not only because of its modest genome size but also because of its rapid growth, relative ease of experimental manipulation, and range of available genetic tools (*2*, *3*). The genus is phenotypically diverse, and interspecific hybrids facilitate the genetic mapping of economically important traits related to growth rate, stature, wood properties, and paper quality. Dozens of quantitative trait loci have already been mapped (*4*), and methods of genetic transformation have been developed (*5*). Under appropriate conditions, *Populus* can reach reproductive maturity in as few as 4 to 6 years, permitting selective breeding for large-scale sustainable plantation forestry. Finally, rapid growth of trees coupled with thermochemical or biochemical conversion of the lignocellulosic portion of the plant has the potential to provide a renewable energy resource with a concomitant reduction of greenhouse gases (*6*–*8*).

Sequencing and Assembly

A single female genotype, "Nisqually 1," was selected and used in a whole-genome shotgun

[1]Environmental Sciences Division, [2]Life Sciences Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. [3]Plant Sciences Department, University of Tennessee, TN 37996, USA. [4]Department of Biology, West Virginia University, Morgantown, WV 26506, USA. [5]Umeå Plant Science Centre, Department of Plant Physiology, Umeå University, SE-901 87, Umeå, Sweden. [6]Michael Smith Laboratories, [7]Department of Botany, [8]Department of Forest Sciences, University of British Columbia, Vancouver, BC V6T 1Z4, Canada. [9]U.S. Department of Energy, Joint Genome Institute, Walnut Creek, CA 94598, USA. [10]Department of Plant Systems Biology, Flanders Interuniversity Institute for Biotechnology (VIB), Ghent University, B-9052 Ghent, Belgium. [11]Genome Sciences Centre, 100-570 West 7th Avenue, Vancouver, BC V5Z 4S6, Canada. [12]Umeå Plant Science Centre, Department of Forest Genetics and Plant Physiology, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden. [13]Tree-Microbe Interactions Unit, Institut National de la Recherche Agronomique (INRA)–Université Henri Poincaré, INRA-Nancy, 54280 Champenoux, France. [14]Department of Forestry, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA. [15]Biotechnology Research Center, School of Forest Resources and Environmental Science, Michigan Technological University, Houghton, MI 49931, USA. [16]Department of Cell and Systems Biology, University of Toronto, 25 Willcocks Street, Toronto, Ontario, M5S 3B2 Canada. [17]School of Forest Resources and Huck Institutes of the Life Sciences, [18]Department of Biology, Institute of Molecular Evolutionary Genetics, and Huck Institutes of Life Sciences, The Pennsylvania State University, University Park, PA 16802, USA. [19]Architecture et Fonction des Macromolécules Biologiques, UMR6098, CNRS and Universities of Aix-Marseille I and II, case 932, 163 avenue de Luminy, 13288 Marseille, France. [20]Warnell School of Forest Resources, [21]Department of Plant Biology, University of Georgia, Athens, GA 30602, USA. [22]School of Forest Resources and Conservation, Genetics Institute, and Plant Molecular and Cellular Biology Program, University of Florida, Gainesville, FL 32611, USA. [23]INRA-Orléans, Unit of Forest Improvement, Genetics and Physiology, 45166 Olivet Cedex, France. [24]Center for Integrative Genomics, University of California, Berkeley, CA 94720, USA. [25]Genomics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA. [26]Department of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA. [27]Department of Biological Sciences, Purdue University, West Lafayette, IN 47907, USA. [28]The Stanford Human Genome Center and the Department of Genetics, Stanford University School of Medicine, Palo Alto, CA 94305, USA. [29]Institute of Forest Genetics, United States Department of Agriculture, Forest Service, Davis, CA 95616, USA. [30]Federal Research Centre for Forests, Hauptstrasse 7, A-1140 Vienna, Austria. [31]Plant Molecular Biology Laboratory, Institute of Biotechnology, [32]Department of Biological and Environmental Sciences, University of Helsinki, FI-00014 Helsinki, Finland. [33]Department of Biology, 200014, University of Turku, FI-20014 Turku, Finland. [34]Southern Institute of Forest Genetics, United States Department of Agriculture, Forest Service and Department of Forest Science, Texas A&M University, College Station, TX 77843, USA. [35]Whitehead Institute for Biomedical Research and Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA. [36]Department of Molecular Sciences and Center of Excellence in Genomics and Bioinformatics, University of Tennessee, Memphis, TN 38163, USA. [37]Southern Institute of Forest Genetics, United States Department of Agriculture, Forest Service, Saucier, MS 39574, USA. [38]Developmental Genetics, University of Tübingen, D-72076 Tübingen, Germany. [39]Department of Biotechnology, KTH, AlbaNova University Center, SE-106 91 Stockholm, Sweden.

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

†These authors contributed equally to this work as second authors.

‡These authors contributed equally to this work as senior authors.

sequence and assembly strategy (*9*). Roughly 7.6 million end-reads representing 4.2 billion high-quality (i.e., Q20 or higher) base pairs were assembled into 2447 major scaffolds containing an estimated 410 megabases (Mb) of genomic DNA (tables S1 and S2). On the basis of the depth of coverage of major scaffolds (~7.5 depth) and the total amount of nonorganellar shotgun sequence that was generated, the *Populus* genome size was estimated to be 485 ± 10 Mb (±SD), in rough agreement with previous cytogenetic estimates of about 550 Mb (*10*). The near completeness of the shotgun assembly in protein-coding regions is supported by the identification of more than 95% of known *Populus* cDNA in the assembly.

The ~75 Mb of unassembled genomic sequence is consistent with cytogenetic evidence that ~30% of the genome is heterochromatic (*9*). The amount of euchromatin contained within the *Populus* genome was estimated in parallel by subtraction on the basis of direct measurements of 4′,6′-diamidino-2-phenylindole–stained prophase and metaphase chromosomes (fig. S4). On average, 69.5 ± 0.3% of the genome consisted of euchromatin, with a significantly lower proportion of euchromatin in linkage group I (LGI) (66.4 ± 1.1%) compared with the other 18 chromosomes (69.7 ± 0.03%, $P \leq 0.05$). In contrast, *Arabidopsis* chromosomes contain roughly 93% euchromatin (*11*). The unassembled shotgun sequences were derived from variants of organellar DNA, including recent nuclear translocations; highly repetitive genomic DNA; haplotypic segments that were redundant with short subsegments of the major scaffolds (separated as a result of extensive sequence polymorphism and allelic variants); and contaminants of the template DNA, such as endophytic microbes inhabiting the leaf and root tissues used for template preparation (*12*) (fig. S1 and table S3). The end-reads corresponding to chloroplast (fig. S5) and mitochondrial genomes were assembled into circular genomes of 157 and 803 kb, respectively (*9*).

We anchored the 410 Mb of assembled scaffolds to a sequence-tagged genetic map (fig. S3). In total, 356 microsatellite markers were used to assign 155 scaffolds (335 Mb of sequence) to the 19 *P. trichocarpa* chromosome-scale linkage groups (*13*). The vast majority (91%) of the mapped microsatellite markers were colinear with the sequence assembly. At the extremes, the smallest chromosome, LGIX [79 centimorgans (cM)], is covered by two scaffolds containing 12.5 Mb of assembled sequence, whereas the largest chromosome, LGI (265 cM), contains 21 scaffolds representing 35.5 Mb (fig. S3). We also generated a physical map based on bacterial artificial chromosome (BAC) fingerprint contigs using a Nisqually-1 BAC library representing an estimated 9.5-fold genome coverage (fig. S2). Paired BAC-end sequences from most of the physical map were linked to the large-scale assembly, permitting 2460 of the physical map contigs to be positioned on the genome assembly. Combining the genetic and physical map, nearly 385 Mb of the 410 Mb of assembled sequence are placed on a linkage group.

Unlike *Arabidopsis*, where predominantly self-fertilizing ecotypes maintain low levels of allelic polymorphism, *Populus* species are predominantly dioecious, which results in obligate outcrossing. This compulsory outcrossing, along with wind pollination and wind-dispersed plumose seeds, results in high levels of gene flow and high levels of heterozygosity (that is, within individual genetic polymorphisms). Within the heterozygous Nisqually-1 genome, we identified 1,241,251 single-nucleotide polymorphisms (SNPs) or small insertion/deletion polymorphisms (indels) for an overall rate of approximately 2.6 polymorphisms per kilobase. Of these polymorphisms, the overwhelming majority (83%) occurred in noncoding portions of the genome (Table 1). Short indels and SNPs within exons resulted in frameshifts and nonsense stop codons within predicted exons, respectively, suggesting that null alleles of these genes exist in one of the haplotypes. Some of the polymorphisms may be artifacts from the assembly process, although these errors were minimized by using stringent criteria for SNP identification (*9*).

Table 1. Characterization of polymorphisms according to their positions relative to predicted coding sequences, introns, and untranslated regions (UTRs). Rate shows the percentage of potential sites of each class that were polymorphic. Most indels within exons resulted in frame shifts, but we could not quantify this due to difficulties with assembly and sequencing of regions containing indels. Nonsense mutations created stop codons within predicted exons.

Source	Number of loci	Rate (%)
Noncoding	1,027,322	0.32
INTRON	141,199	0.25
3′UTR	6,731	0.25
5′UTR	3,306	0.24
Exon	62,656	0.14
Within exons:		
Indels	2,722	0.01
Nonsense	926	0.02
Nonsynonymous	32,207	0.10

Gene Annotation

We tentatively identified a first-draft reference set of 45,555 protein-coding gene loci in the *Populus* nuclear genome (www.jgi.doe.gov/poplar) using a variety of ab initio, homology-based, and expressed sequence tag (EST)–based methods (*14*–*17*) (table S5). Similarly, 101 and 52 genes were annotated in the chloroplast and mitochondrial genomes, respectively (*9*). To aid the annotation process, 4664 full-length sequences, from full-length enriched cDNA libraries from Nisqually 1, were generated and used in training the gene-calling algorithms. Before gene prediction, repetitive sequences were characterized (fig. S15 and table S14) and masked; additional putative transposable elements were identified and subsequently removed from the reference gene set (*9*). Given the current draft nature of the genome, we expect that the gene set in *Populus* will continue to be refined.

About 89% of the predicted gene models had homology [expectation (E) value $\leq 1 \times 10^{-8}$] to the nonredundant (NR) set of proteins from the National Center for Biotechnology Information, including 60% with extensive homology that spans 75% of both model and NR protein lengths. Nearly 12% (5248) of the predicted *Populus* genes had no detectable similarity to *Arabidopsis* genes (E value $\leq 1 \times 10^{-3}$); conversely, in the more refined *Arabidopsis* set, only 9% (2321) of the predicted genes had no similarity to the *Populus* reference set. Of the 5248 *Populus* genes without *Arabidopsis* similarity, 1883 have expression evidence from the manually curated *Populus* EST data set, and of these, 274 have no hits (E value $\geq 1 \times 10^{-3}$) to the NR database (*9*). Whole-genome oligonucleotide microarray analysis provided evidence of tissue-based expression for 53% of the reference gene models (Fig. 1). In addition, a signal was detected from 20% of genes that were initially annotated and excluded from the reference set, suggesting that as many as 4000 additional genes (or gene fragments) may be present. Within the reference gene set, we identified 13,019 pairs of orthologs between

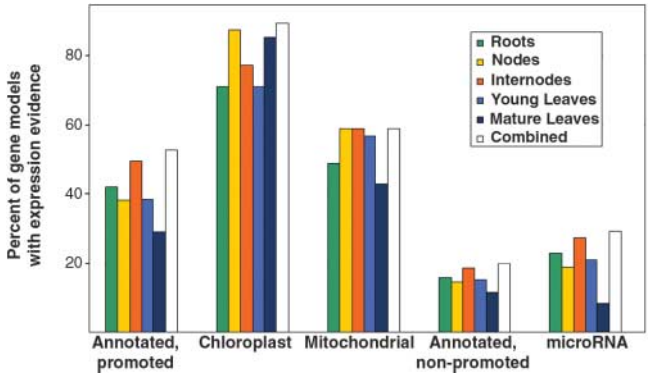

Fig. 1. Whole-genome oligonucleotide microarray expression data for all predicted gene models in *P. trichocarpa*. Values represent the proportion of genes expressed above negative controls at a 5% false discovery rate. The *x* axis represents the subsets of predicted genes that were analyzed for the annotated and promoted *P. trichocarpa* gene set (42,373 genes), chloroplast gene set (49 genes), mitochondria gene set (49 genes), annotated, nonpromoted gene set (10,875 genes), and microRNAs (48 miRNAs).

genes in *Populus* and *Arabidopsis* using the best bidirectional Basic Local Alignment Search Tool (BLAST) hits, with average mutual coverage of these alignments equal to 93%; 11,654 pairs of orthologs had greater than 90% alignment of gene lengths, whereas only 156 genes had less than 50% coverage. As of 1 June 2006, ~10% (4378) gene models have been manually validated and curated.

Genome Organization

Genome duplication in the Salicaceae. *Populus* and *Arabidopsis* lineages diverged about 100 to 120 million years ago (Ma). Analysis of the *Populus* genome provided evidence of a more recent duplication event that affected roughly 92% of the *Populus* genome. Nearly 8000 pairs of paralogous genes of similar age (excluding tandem or local duplications) were identified (Fig. 2). The relative age of the duplicate genes was estimated by the accumulated nucleotide divergence at fourfold synonymous third-codon transversion position (4DTV) values. A sharp peak in 4DTV values, corrected for multiple substitutions, representing a burst of gene duplication, is evident at 0.0916 ± 0.0004 (Fig. 3A). Comparison of 1825 *Populus* and *Salix* orthologous genes derived from *Salix* EST suggests that both genera share this whole-genome duplication event (Fig. 3B). Moreover, the parallel karyotypes and collinear genetic maps (*18*) of *Salix* and *Populus* also support the conclusion that both lineages share the same large-scale genome history.

If we naively calibrated the molecular clock using synonymous rates observed in the Brassicaceae (*19*) or derived from the *Arabidopsis-Oryza* divergence (*20*), we would conclude that the genome duplication in *Populus* is very recent [8 to 13 Ma, as reported by Sterk (*21*)]. Yet the fossil record shows that the *Populus* and *Salix* lineages diverged 60 to 65 Ma (*22–25*). Thus, the molecular clock in *Populus* must be ticking at only

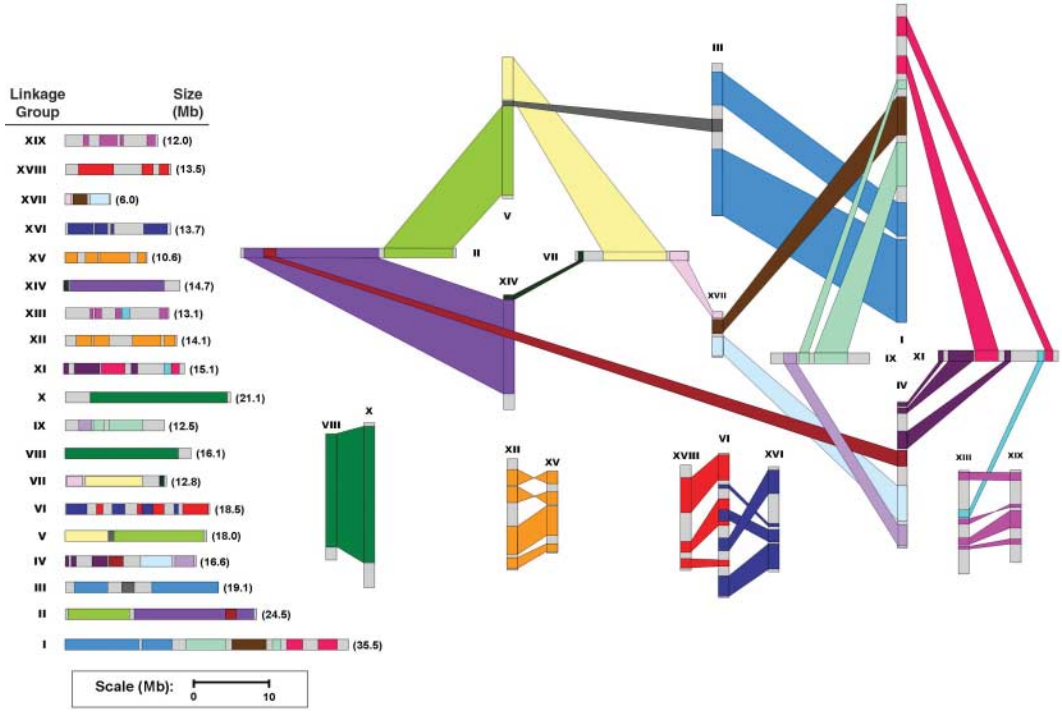

Fig. 2. Chromosome-level reorganization of the most recent genome-wide duplication event in *Populus*. Common colors refer to homologous genome blocks, presumed to have arisen from the salicoid-specific genome duplication 65 Ma, shared by two chromosomes. Chromosomes are indicated by their linkage group number (I to XIX). The diagram to the left uses the same color coding and further illustrates the chimeric nature of most linkage groups.

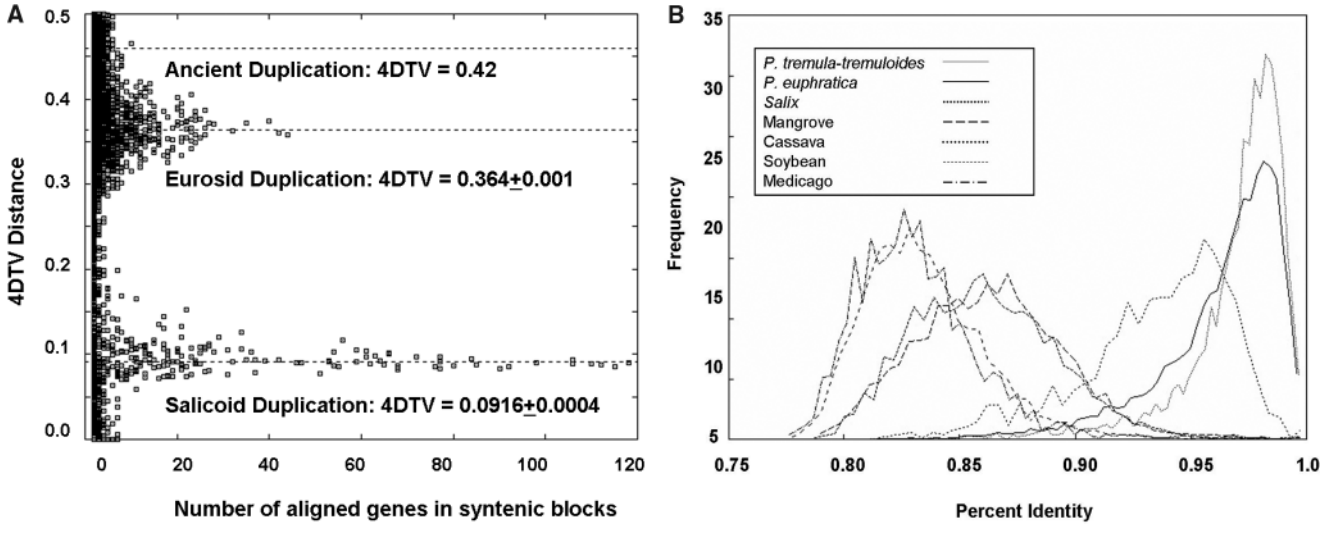

Fig. 3. (**A**) The 4DTV metrics for paralogous gene pairs in *Populus-Populus* and *Populus-Arabidopsis*. Three separate genome-wide duplications events are detectable, with the most recent event contained within the Salicaceae and the middle event apparently shared among the Eurosids. (**B**) Percent identity distributions for mutual best EST hit to *Populus trichocarpa* CDS.

one-sixth the estimated rate for *Arabidopsis* (that is, 8 to 13 Ma divided by 60 to 65 Ma). Qualitatively similar slowing of the molecular clock is found in the *Populus* chloroplast and mitochondrial genomes (*9*). Because *Populus* is a long-lived vegetatively propagated species, it has the potential to successfully contribute gametes to multiple generations. A single *Populus* genotype can persist as a clone on the landscape for millennia (*26*), and we propose that recurrent contributions of "ancient gametes" from very old individuals could account for the markedly reduced rate of sequence evolution. As a result of the slowing of the molecular clock, the *Populus* genome most likely resembles the ancestral eurosid genome.

To test whether the burst of gene creation 60 to 65 Ma was due to a single whole-genome event or to independent but near-synchronous gene duplication events, we used a variant of the algorithm of Hokamp *et al*. (*27*) to identify segments of conserved synteny within the *Populus* genome. The longest conserved syntenic block from the 4DTV ~0.09 epoch spanned 765 pairs of paralogous genes. In total, 32,577 genes were contained within syntenic blocks from the salicoid epoch; half of these genes were contained in segments longer than 142 paralogous pairs. The same algorithm, when applied to randomly shuffled genes, typically yields duplicate blocks with fewer than 8 to 9 genes, indicating that the *Populus* gene duplications occurred as a single genome-wide event. We refer to this duplication event as the "salicoid" duplication event.

Nearly every mapped segment of the *Populus* genome had a parallel "paralogous" segment elsewhere in the genome as a result of the salicoid event (Fig. 2). The pinwheel patterns can be understood as a whole-genome duplication followed by a series of reciprocal tandem terminal fusions between two separate sets of four chromosomes each—the first involving LGII, V, VII, and XIV and the second involving LGI, XI, IV, and IX. In addition, several chromosomes appear to have experienced minor reorganizational exchanges. Furthermore, LGI appears to be the result of multiple rearrangements involving three major tandem fusions. These results suggest that the progenitor of *Populus* had a base chromosome number of 10. After the whole-genome duplication event, this base chromosome number experienced a genome-wide reorganization and diploidization of the duplicated chromosomes into four pairs of complete paralogous chromosomes (LGVI, VIII, X, XII, XIII, XV, XVI, XVIII, and XIX); two sets of four chromosomes, each containing a terminal translocation (LGI, II, IV, V, VII, IX, and XI); and one chromosome containing three terminally joined chromosomes (LGIII with I or XVII with VII). The colinearity of genetic maps among multiple *Populus* species suggests that the genome reorganization occurred before the evolution of the modern taxa of *Populus*.

Genome duplication in a common ancestor of *Populus* and *Arabidopsis*. The distribution of 4DTV values for paralogous pairs of genes also shows that a large fraction of the *Populus* genome falls in a set of duplicated segments anchored by gene pairs with 4DTV at 0.364 ± 0.001, representing the residue of a more ancient, large-scale, apparently synchronous duplication event (Fig. 3A). This relatively older duplication event covers about 59% of the *Populus* genome with 16% of genes in these segments present in two copies. Because this duplication preceded and is therefore superimposed upon the salicoid event, each genomic region is potentially covered by four such segments. Similarly, the *Arabidopsis* genome experienced an older "beta" duplication that preceded the Brassicaceae-specific "alpha" event (*28*–*32*).

We next asked whether the *Arabidopsis* "beta" (*30*, *32*) and *Populus* 4DTV ~0.36 duplication events were (i) independent genome-wide duplications that occurred after the split from the last common eurosid ancestor (H_1) or (ii) a single shared duplication event that occurred in an ancestral lineage (i.e., before the divergence of eurosid lineages I and II) (H_2). These two hypotheses have very different implications for the interpretation of homology between *Populus* and *Arabidopsis*. Under H_1, each genomic segment in one species is homologous to four segments in the other; whereas under H_2, each segment is homologous to only two segments in the other species. These hypotheses were tested by comparing the relative distances between gene pairs sampled within and between *Populus* and *Arabidopsis*. H_2 was generally supported (*9*), but we could not reject H_1. We can only conclude that the *Populus* genome duplication occurred very close to the time of divergence of the eurosid I and II lineages (*9*), with slight support for a shared duplication. This coincident timing raises the possibility of a causal link between this duplication and rapid diversification early in eurosid (and perhaps core eudicot) history. We refer to this older *Populus*/*Arabidopsis* duplication event as the "eurosid" duplication event. We note that the salicoid duplication occurred independently of the eurosid duplication observed in the *Arabidopsis* genome.

Gene Content

Although *Populus* has substantially more protein-coding genes than *Arabidopsis*, the relative frequency of domains represented in protein databases (Prints, Prosite, Pfam, ProDom, and SMART) in the two genomes is similar (*9*). However, the most common domains occur in *Populus* compared with *Arabidopsis* in a ratio ranging from 1.4:1 to 1.8:1. Noteworthy outliers in *Populus* include genes and gene domains associated with disease and insect resistance (such as, in *Populus* versus *Arabidopsis*, respectively: leucine-rich repeats, 1271 versus 527; NB-ARC domain, 302 versus 141; and thaumatin, 55 versus 24), meristem development (such as NAC transcription factors, 157 versus 100, respectively), and metabolite and nutrient transport [such as oligopeptide transporter of the proton-dependent oligopeptide transporter (POT) and oligopeptide transporter (OPT) families, 129 versus 61, and potassium transporter, 30 versus 13, respectively].

Some domains were underrepresented in *Populus* compared with *Arabidopsis*. For example, the F-box domain was twice as prevalent in *Arabidopsis* as in *Populus* (624 versus 303, respectively). The F-box domain is involved in diverse and complex interactions involving protein degradation through the ubiquitin-26*S* proteosome pathway (*33*). Many of the ubiquitin-associated domains are underrepresented in *Populus* compared with *Arabidopsis* (for example, the *Ulp1* protease family and the C-terminal catalytic domain, 10 versus 63, respectively). Moreover, the RING-finger domains are nearly equally present in both genomes (503 versus 407, respectively), suggesting that protein degradation pathways in the two organisms are metabolically divergent.

The common eurosid gene set. The *Populus* and *Arabidopsis* gene sets were compared to infer the conserved gene complement of their common eurosid ancestor, integrating information from nucleotide divergence, synteny, and mutual best BLAST-hit analysis (*9*). The ancestral eurosid genome contained at least 11,666 protein-coding genes, along with an undetermined number that were either lost in one or both of the lineages or whose homology could not be detected. These ancestral genes were the progenitors of gene families of typically one to four descendents in each of the complete plant genomes and account for 28,257 *Populus* and 17,521 *Arabidopsis* genes. Gene family lists are accessible at www.phytozome.net. The gene predictions in these two genomes that could not be accounted for in the eurosid clusters were often fragmentary or difficult to categorize, and we could not confidently assign orthology to them. They may include previously unidentified or rapidly evolving genes in the *Populus* and/or *Arabidopsis* lineages, as well as poorly predicted genes.

Noncoding RNAs. Based on a series of publicly available RNA detection algorithms (*34*), including tRNAScan-SE, INFERNAL, and snoScan, we identified 817 putative tRNAs; 22 U1, 26 U2, 6 U4, 23 U5, and 11 U6 spliceosomal small nuclear RNAs (snRNAs); 339 putative C/D small nucleolar RNAs (snoRNAs); and 88 predicted H/ACA snoRNAs in the *Populus* assembly. All 57 possible anticodon tRNAs were found. One selenocysteine tRNA was detected and two possible suppressor tRNAs (anticodons that bind stop codons) were also identified. *Populus* has nearly 1.3 times as many tRNA genes as *Arabidopsis*. In contrast to *Arabidopsis* (fig. S7A), the copy number of tRNA in *Populus* was significantly and positively correlated with amino acid occurrence in predicted gene models (fig. S7B). The ratio of the number of snRNAs in *Populus* compared with the number in *Arabidopsis* is 1.3 to 1.0, yet U1, U2, and U5 are overrepresented in *Populus*, whereas U4 is underrepresented. Further-

more, U14 was not detected in *Arabidopsis*. The snRNAs and snoRNAs have not been experimentally verified in *Populus*.

There are 169 identified microRNA (miRNA) genes representing 21 families in *Populus* (table S7). In *Arabidopsis*, these 21 families contain 91 miRNA genes, representing a 1.9X expansion in *Populus*, primarily in miR169 and miR159/319. All 21 miRNA families have regulatory targets that appear to be conserved among *Arabidopsis* and *Populus* (table S8). Similar to the miRNA genes themselves, the number of predicted targets for these miRNA is expanded in *Populus* (147) compared with *Arabidopsis* (89). Similarly, the genes that mediate RNA interference (RNAi) are also overrepresented in *Populus* (21) compared to *Arabidopsis* (11) [e.g., AGO1 class, 7 versus 3; RNA helicase 2 versus 1; HEN, 2 versus 1; HYL1-like (double-stranded RNA binding proteins), 9 versus 5, respectively].

Tandem duplications. In *Populus* there were 1518 tandemly duplicated arrays of two or more genes based on a Smith-Waterman alignment E value $\leq 10^{-25}$ and a 100-kb window. The total number of genes in such arrays was 4839 and the total length of tandemly duplicated segments in *Populus* was 47.9 Mb, or 15.6% of the genome (fig. S8). By the same criteria, there are 1366 tandemly duplicated segments in *Arabidopsis*, covering 32.4 Mb, or 27% of the genome. By far the most common number of genes within a single array was two, with 958 such arrays in *Populus* and 805 in *Arabidopsis*. *Arabidopsis* had a larger number of arrays containing six or more genes than did *Populus*. Tandem duplications thus appear to be relatively more common in *Arabidopsis* than in *Populus*. This may in part be due to difficulties in assembling tandem repeats from a whole-genome shotgun sequencing approach, particularly when tandemly duplicated genes are highly conserved. Alternatively, the *Populus* genome may be undergoing rearrangements at a slower rate than the *Arabidopsis* genome, which is consistent with our observations of reduced chromosomal rearrangements and slower nucleotide substitution rates in *Populus*.

In some cases, genes were highly duplicated in both species, and some tandem duplications predated the *Populus*-*Arabidopsis* split (*9*). The largest number of tandem repeats in *Populus* in a single array was 24 and contained genes with high homology to S locus–specific glycoproteins. Genes of this class also occur as tandem repeats in *Arabidopsis*, with the largest segments containing 14 tandem duplicates on chromosome 1. One of the InterPro domains in this protein, IPR008271, a serine/threonine protein kinase active site, was the most frequent domain in tandemly repeated genes in both species (fig. S8). Other common domains in both species were the leucine-rich repeat (IPR007090, primarily from tandem repeats of disease resistance genes), the pentatricopeptide repeat RNA-binding proteins (IPR002885), and the uridine diphosphate (UDP)-glucuronosyl/UDP-glucosyltransferase domain (IPR002213) (table S9).

In contrast, some genes were highly expanded in tandem duplicates in one genome and not in the other (fig. S8). For example, one of the most frequent classes of tandemly duplicated genes in *Arabidopsis* was F-box genes, with a total of 342 involved in tandem duplications, the largest segment of which contained 24 F-box genes. *Populus* contains only 37 F-box genes in tandem duplications, with the largest segment containing only 3 genes.

Postduplication Gene Fate

Functional expression divergence. In *Populus*, 20 of the 66 salicoid-event duplicate gene pairs contained in 19 *Populus* EST libraries (2.3% of the total) showed differential expression (*9*) [displaying significant deviation in EST frequencies per library (Fig. 4)]. Out of 18 eurosid-event duplicate gene pairs (2.7% of the total), 11 also displayed significant deviation in EST frequencies per library. Many of the duplicate gene pairs that displayed significant overrepresentation in one or more of the 19 sampled libraries were involved in protein-protein interactions (such as annexin) or protein folding (such as cyclophilins). In the eurosid set, there was a greater divergence in the best BLAST hit among pairwise sets of genes. These results support the premise of functional expression divergence among some duplicated gene pairs in *Populus*.

To further test for variation in gene expression among duplicated genes, we examined whole-genome oligonucleotide microarray data containing the 45,555 promoted genes (*9*). There was significantly lower differential expression in the salicoid duplicated pairs of genes (mean = 5%) relative to eurosid duplications (mean = 11%), again suggesting that differential expression patterns for retained paralogous gene pairs is an ongoing process that has had more time to occur in eurosid pairs (Fig. 5). This difference could also be due to absolute expression level, which may vary systematically between the two duplication events. Moreover, differential expression was more evident in the wood-forming organs. Almost 14 and 13% (2632 pairs of genes) of eurosid duplicated genes in the nodes and internodes, respectively, displayed differential expression, compared with 8% or less in roots and young leaves (Fig. 5).

Single-nucleotide polymorphisms. *Populus* is a highly polymorphic taxon and substantial numbers of SNPs are present even within a single individual (Table 1). The ratio of nonsynonymous to synonymous substitution rate ($\omega = dN/dS$) was calculated as an index of selective constraints for alleles of individual genes (*9*). The overall average dN across all genes was 0.0014, whereas the dS value was 0.0035, for a total ω of 0.40, suggesting that the majority of coding regions in the *Populus* genome are subject to purifying selection. There was a significant, negative correlation between ω and the 4DTV distance to the most closely related paralog ($r = -0.034$, $P = 0.028$), which is consistent with the expectation of higher levels of nonsynonymous polymorphism in recently duplicated genes as a result of functional redundancy (*20*, *35*). Similarly, genes with recent tandem duplicates (4DTV ≤ 0.2) had significantly higher ω than did genes with no recent tandem duplicates (Wilcoxon rank sum $Z = 8.65$, $P \leq 0.0001$) (table S10).

The results for tandemly duplicated genes were consistent with expectations for accelerated evolution of duplicated genes (*20*). However, this expectation was not upheld for paralogous pairs of genes from the whole-genome duplication events. Relative rates of nonsynonymous substitution were actually lower for genes with paralogs from the salicoid and eurosid whole-genome duplication events than for genes with no paralogs (table S11). One possible explanation for this

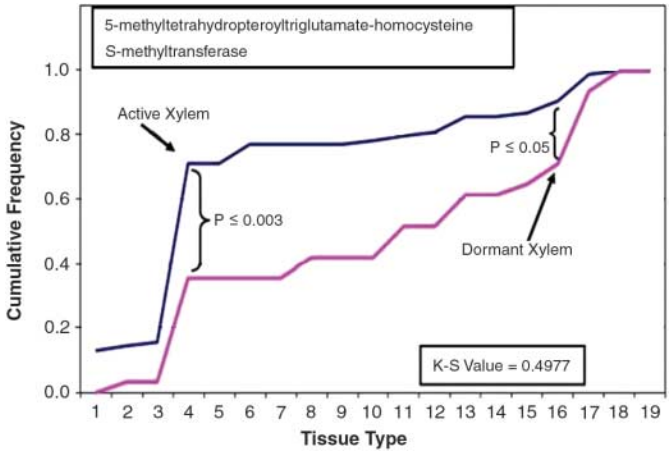

Fig. 4. Kolmogorov-Smirnov (K-S) test for differential expression for 5-methyltetrahydropteroyltriglutamate-homocysteine *S*-methyltransferase genes [for descriptions of the EST data set, see Sterky *et al.* (*79*)]. Results suggest that the duplicated genes in *Populus* are differentially expressed in alternate tissues. Tissue types include: cambial zone (1), young leaves (2), flower buds (3), tension wood (4), senescing leaves (5), apical shoot (6), dormant cambium (7), active cambium (8), cold stressed leaves (9), roots (10), bark (11), shoot meristem (12), male catkins (13), dormant buds (14), female catkins (15), petioles (16), wood cell death (17), imbibed seeds (18) and infected leaves (19).

discrepancy is that the apparent single-copy genes have a corresponding overrepresentation of rapidly evolving pseudogenes. However, this does not appear to be the case, as demonstrated by an analysis of gene size, synonymous substitution rate, and minimum genetic distance to the closest paralog as covariates in an analysis of variance with ω as the response variable (table S11). Therefore, genes with no paralogs from the salicoid and eurosid duplication events seem to be under lower selective constraints, and purifying selection is apparently stronger for genes with paralogs retained from the whole-genome duplications. Chapman *et al*. (*36*) have recently proposed the concept of functional buffering to account for similar reduction in detected mutations in paralogs from whole-genome duplications in *Arabidopsis* and *Oryza*. The vegetative propagation habit of *Populus* may also favor the conservation of nucleotide sequences among duplicated genes, in that complementation among duplicate pairs of genes would minimize loss of gene function associated with the accumulation of deleterious somatic mutations.

Gene family evolution. The expansion of several gene families has contributed to the evolution of *Populus* biology.

Lignocellulosic wall formation. Among the processes unique to tree biology, one of the most obvious is the yearly development of secondary xylem from the vascular cambium. We identified *Populus* orthologs of the approximately 20 *Arabidopsis* genes and gene families involved in or associated with cellulose biosynthesis. The *Populus* genome has 93 cellulose synthesis–related genes compared with 78 in *Arabidopsis*. The *Arabidopsis* genome encodes 10 *CesA* genes belonging to six classes known to participate in cellulose microfibril biosynthesis (*37*). *Populus* has 18 *CesA* genes (*38*), including duplicate copies of *CesA7* and *CesA8* homologs. *Populus* homologs of *Arabidopsis CesA4*, *CesA7*, and *CesA8* are coexpressed during xylem development and tension wood formation (*39*). Furthermore, one pair of *CesA* genes appears unique to *Populus*, with no homologs found in *Arabidopsis* (*40*). Many other types of genes associated with cellulose biosynthesis, such as *KOR*, *SuSY*, *COBRA*, and *FRA2*, occur in duplicate pairs in *Populus* relative to single-copy *Arabidopsis* genes (*39*). For example, *COBRA*, a regulator of cellulose biogenesis (*41*), is a single-copy gene in *Arabidopsis*, but in *Populus* there are four copies.

The repertoire of acknowledged hemicellulose biosynthetic genes in *Populus* is generally similar to that in *Arabidopsis*. However, *Populus* has more genes encoding α-L-fucosidases and fewer genes encoding α-L-fucosyltransferases than does *Arabidopsis*, which is consistent with the lower xyloglucan fucose content (*42*) in *Populus* relative to *Arabidopsis*.

Lignin, the second most abundant cell wall polymer after cellulose, is a complex polymer of monolignols (hydroxycinnamyl alcohols) that encrusts and interacts with the cellulose/hemicellulose matrix of the secondary cell wall (*43*). The full set of 34 *Populus* phenylpropanoid and lignin biosynthetic genes (table S13) was identified by sequence alignment to the known *Arabidopsis* phenylpropanoid and lignin genes (*44*, *45*). The size of the *Populus* gene families that encode these enzymes is generally larger than in *Arabidopsis* (34 versus 18, respectively). The only exception is cinnamyl alcohol dehydrogenase (CAD), which is encoded by a single gene in *Populus* and two genes in *Arabidopsis* (Fig. 6C); CAD is also encoded by only a single gene in *Pinus taeda* (*46*, *47*). Two lignin-related *Populus C4H* genes are strongly coexpressed in tissues related to wood formation, whereas the three *Populus C3H* genes show reciprocally exclusive expression patterns (*48*) (Fig. 6, A and B).

Secondary metabolism. Populus trees produce a broad array of nonstructural, carbon-rich secondary metabolites that exhibit wide variation in abundance, stress inducibility, and effects on tree growth and host-pest interactions (*49*–*53*). Shikimate-phenylpropanoid–derived phenolic esters, phenolic glycosides, and condensed tannins and their flavonoid precursors comprise the largest classes of these metabolites. Phenolic glycosides and condensed tannins alone can constitute up to 35% leaf dry weight and are abundant in buds, bark, and roots of *Populus* (*50*, *54*, *55*).

The flavonoid biosynthetic genes are well annotated in *Arabidopsis* (*56*) and almost all (with the exception of flavonol synthase) are encoded by single-copy genes. In contrast, all but three such enzymes (chalcone isomerase, flavonoid 3′-hydroxylase, and flavanone 3-hydroxylase) are encoded by multiple genes in *Populus* (*53*). For example, the chalcone synthase, controlling the committed step to flavonoid biosynthesis, has expanded to at least six genes in *Populus*. In addition, *Populus* contains two genes each for flavone synthase II (cytochrome accession number CYP98B) and flavonoid 3′,5′-hydroxylase (CYP75A12 and CYP75A13), both of which are absent in *Arabidopsis*. Furthermore, three *Populus* genes encode leucoanthocyanidin reductase, required for the synthesis of condensed tannin precursor 2,3-*trans*-flavan-3-ols, a stereochemical configuration also lacking in *Arabidopsis* (*57*). In contrast to the 32 terpenoid synthase (*TPS*) genes of secondary metabolism identified in the *Arabidopsis* genome (*58*), the *Populus* genome contains at least 47 *TPS* genes, suggesting a wide-ranging capacity for the formation of terpenoid secondary metabolites.

A number of phenylpropanoid-like enzymes have been annotated in the *Arabidopsis* genome (*44*, *45*, *59*–*61*). One example is the family encoding *CAD*. In addition to the single *Populus CAD* gene involved in lignin biosynthesis, several other clades of *CAD-like* (*CADL*) genes are present, most of which fall within larger subfamilies containing enzymes related to multifunctional alcohol dehydrogenases (Fig. 6). This comparative analysis makes it clear that there has been selective expansion and retention of *Populus CADL* gene families. For example, *Populus* contains seven *CADL* genes (*PoptrCADL1* to *PoptrCADL7*; Fig. 6C) encoding enzymes related to the *Arabidopsis* BAD1 and BAD2 enzymes with apparent benzyl alcohol dehydrogenase activities (*62*). BAD1 and BAD2 are known to be pathogen inducible, suggesting that this group of *Populus* genes, including the *Populus SAD* gene, previously characterized as encoding a sinapaldehyde-specific CAD enzyme (*63*), may be involved in chemical defense.

Disease resistance. The likelihood that a perennial plant will encounter a pathogen or herbivore before reproduction is near unity. The long-generation intervals for trees make it difficult for such plants to match the evolutionary rates of a microbial or insect pest. Aside from the formation of thickened cell walls and the synthesis of secondary metabolites that constitute a first line of defense against microbial and insect pests, plants use a variety of disease-resistance (*R*) genes.

The largest class of characterized *R* genes encodes intracellular proteins that contain a

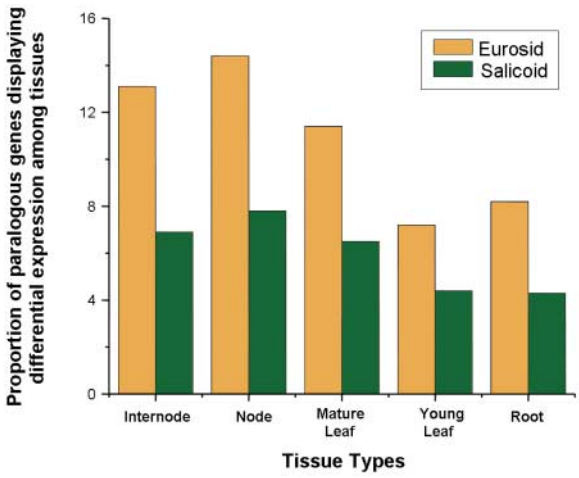

Fig. 5. Proportion of eurosid and salicoid duplicated gene sets differentially expressed in stems (nodes and internode), leaves (young and mature), and whole roots. Samples from four biological replicates collected from the reference genotype Nisqually 1 were individually hybridized to whole-genome oligonucleotide microarrays containing three 60-oligomer oligonucleotide probes for each gene. Differential expression between duplicated genes was evaluated in *t* tests and declared significant at a 5% false discovery rate (*9*).

nucleotide-binding site (NBS) and carboxy-terminal leucine-rich-repeats (LRR) (*64*). The NBS-coding *R* gene family is one of the largest in *Populus*, with 399 members, approximately twice as high as in *Arabidopsis*. The NBS family can be divided into multiple subfamilies with distinct domain organizations, including 64 TIR-NBS-LRR genes, 10 truncated TIR-NBS that lack an LRR, 233 non–TIR-NBS-LRR genes, and 17 unusual TIR-NBS–containing genes that have not been identified previously in *Arabidopsis* (TNLT, TNLN, or TCNL domains) (Table 2). Five gene models coding for TNL proteins contained a predicted N-terminal nuclear localization signal (*65*). The number of non–TIR-NBS-LRR genes in *Populus* is also much higher than that in *Arabidopsis* (209 versus 57, respectively). Notably, 40 non–TIR-NBS genes, not found in *Arabidopsis*, carry an N-terminal BED DNA-binding zinc finger domain that was also found in the *Oryza Xa1* gene. These findings suggest that domain cooption occurred in *Populus*. Most NBS-LRR (about 65%) in *Populus* occur as singletons or in tandem duplications, and the distribution of pairwise genetic distances among these genes suggests a recent expansion of this family. That is, only 10% of the NBS-LRR genes are associated with the eurosid and salicoid duplication events, compared with 55% of the extracellular LRR receptor-like kinase genes (for example, fig. S10).

Several conserved signaling components such as RAR1, EDS1, PAD4, and NPR1, known to be recruited by *R* genes, also contain multiple homologs in *Populus*. For example, two copies of the *PAD4* gene, which functions upstream of salicylic acid accumulation, and five copies of the *NPR1* gene, an important regulator of responses downstream of salicylic acid, are found in *Populus*. Nearly all genes known to control disease resistance signaling in *Arabidopsis* have putative orthologs in *Populus*. *Populus* has a larger number of β-1,3-glucanase and chitinase genes than does *Arabidopsis* (131 versus 73, respectively). In summary, the structural and genetic diversity that exists among *R* genes and their signaling components in *Populus* is remarkable and suggests that unlike the rest of the genome, contemporary diversifying selection has played an important role in the evolution of disease resistance genes in *Populus*. Such diversification suggests that enhanced ability to detect and respond to biotic challenges through *R* gene–mediated signaling may be critical over a decades-long life span of this genus.

Membrane transporters. Attributes of *Populus* biology such as massive interannual, seasonal, and diurnal metabolic shifts and redeployment of carbon and nitrogen may require an elaborate array of transporters. Investigation of gene families coding for transporter proteins (http://plantst.genomics.purdue.edu/) in the *Populus* genome revealed a general expansion relative to *Arabidopsis* (1722 versus 959, in *Populus* versus *Arabidopsis*, respectively) (table S12). Five gene families, coding for adenosine 5′-triphosphate–binding cassette proteins (ABC transporters, 226 gene models), major facilitator superfamily proteins (187 genes), drug/metabolite transporters (108 genes), amino acid/auxin permeases (95 genes), and POT transporters (90 genes), accounted for more than 40% of the total number of transporter gene models (fig. S14). Some large families such as those encoding POT (4.3X relative to *Arabidopsis*), glutamate-gated ion channels (3.7X), potassium uptake permeases (2.3X), and ABC transporters (1.9X) are expanded in *Populus*. We identified a subfamily of five putative aquaporins, lacking in the *Arabidopsis*. *Populus* also harbors seven transmembrane receptor genes that have previously only been found in fungi, and two genes, identified as mycorrhizal-specific phosphate transporters, that confirm that the mycorrhizal symbiosis may have an impact on the mineral nutrition of this long-lived species. This expanded inventory of transporters could conceivably play a role in adaptation to nutrient-limited forest soils, long-distance transport and storage of water and metabolites, secretion and movement of secondary metabolites, and/or mediation of resistance to pathogen-produced secondary metabolites or other toxic compounds.

Phytohormones. Both physiological and molecular studies have indicated the importance of

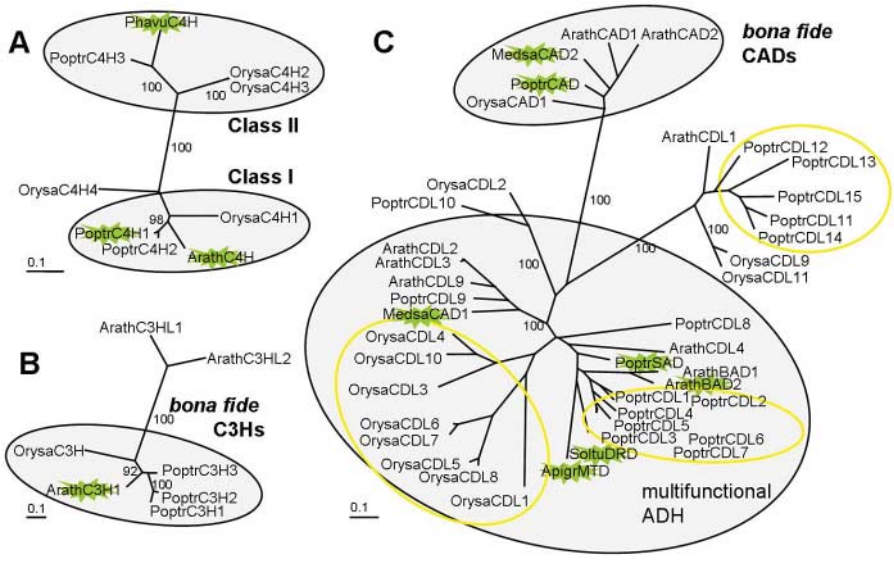

Fig. 6. Phylogenetic analysis of gene families in *Populus*, *Arabidopsis*, and *Oryza* encoding selected lignin biosynthetic and related enzymes. (**A**) Cinnamate-4-hydroxylase (*C4H*) gene family. (**B**) 4-coumaroyl-shikimate/quinate-3-hydroxlase (*C3H*) gene family. (**C**) Cinnamyl alcohol dehydrogenase (*CAD*) and related multifunctional alcohol dehydrogenase gene family. *Arabidopsis* gene names are the same as those in Ehlting *et al*. (*80*). *Populus* and *Oryza* gene names were arbitrarily assigned; corresponding gene models are listed in table S13. Genes encoding enzymes for which biochemical data are available are highlighted with a green flash. Yellow circles indicate monospecific clusters of gene family members.

Table 2. Numbers of genes that encode domains similar to plant *R* proteins in *Populus*, *Arabidopsis* (*81*), and *Oryza* (*82*). *, BED finger and/or DUF1544 domain; CC, coiled coil; –, not detected.

Predicted protein domains	Letter code	*Populus*	*Arabidopsis*	*Oryza*
TIR-NBS	TN	10	21	–
TIR-NBS-LRR	TNL	64	83	–
TIR-NBS-LRR-TIR	TNLT	13	–	–
TIR-NBS-LRR-NBS	TNLN	1	–	–
NBS-LRR-TIR	NLT	1	–	–
TIR-CC-NBS-LRR	TCNL	2	–	–
CC-NBS	CN	19	4	7
CC-NBS-LRR	CNL	119	51	159
BED/DUF1544*-NBS	BN	5	–	–
NBS-BED/DUF1544*	NB	1	–	–
BED/DUF1544*-NBS-LRR	BNL	24	–	–
NBS-LRR	NL	90	6	40
NBS	N	49	1	45
Others	–	–	41	284
Total NBS genes		398	207	535

hormonal regulation underlying plant development. Auxin, gibberellin, cytokinin, and ethylene responses are of particular interest in tree biology.

Many auxin responses (*66*–*71*) are controlled by auxin response factor (*ARF*) transcription factors, which work together with cognate AUX/IAA repressor proteins to regulate auxin-responsive target genes (*72*, *73*). A phylogenetic analysis using the known and predicted *ARF* protein sequences showed that *Populus* and *Arabidopsis ARF* gene families have expanded independently since they diverged from their common ancestor. Six duplicate *ARF* genes in *Populus* encode paralogs of *ARF* genes that are single-copy *Arabidopsis* genes, including *ARF5* (*MONOPTEROS*), an important gene required for auxin-mediated signal transduction and xylem development. Furthermore, five *Arabidopsis ARF* genes have four or more predicted *Populus ARF* gene paralogs. In contrast to *ARF* genes, *Populus* does not contain a notably expanded repertoire of *AUX/IAA* genes relative to *Arabidopsis* (35 versus 29, respectively) (*74*). Interestingly, there is a group of four *Arabidopsis AUX/IAA* genes with no apparent *Populus* orthologs, suggesting *Arabidopsis*-specific functions.

Gibberellins (GAs) are thought to regulate multiple processes during wood and root development, including xylem fiber length (*75*). Among all gibberellin biosynthesis and signaling genes, the *Populus* GA20-oxidase gene family is the only family with approximately two times the number of genes relative to *Arabidopsis*, indicating that most of the duplicated genes that arose from the salicoid duplication event have been lost. GA20-oxidase appears to control flux in the biosynthetic pathway leading to the bioactive gibberellins GA_1 and GA_4. The higher complement of GA20-oxidase genes may have biological importance in *Populus* with respect to secondary xylem and fiber cell development.

Cytokinins are thought to control the identity and proliferation of cell types relevant for wood formation as well as general cell division (*67*). The total number of members in gene families encoding cytokinin homeostasis related isopentenyl transferases (IPT) and cytokinin oxidases is roughly similar between *Populus* and *Arabidopsis*, although there appears to be lineage-specific expansion of IPT subfamilies. The cytokinin signal transduction pathway represents a two-component phosphorelay system, in which a two-component hybrid receptor initiates a phosphotransfer by means of histidine-containing phosphotransmitters (HPt) to phospho-accepting response regulators (RR). One family of genes, encoding the two-component receptors (such as CKI1), is notably expanded in *Populus* (four versus one in *Populus* and *Arabidopsis*, respectively) (*76*). Gene families coding for recently identified pseudo-*HPt* and atypical *RR* are overrepresented in *Populus* relative to *Arabidopsis* (2.5- and 4.0-fold increase in *Populus*, respectively). Both of these gene families have been implicated in the negative regulation of cytokinin signaling (*67*, *77*), which is consistent with the idea of increased complexity in regulation of cytokinin signal transduction in *Populus*.

Populus and *Arabidopsis* genomes contain almost identical numbers of genes for the three enzymes of ethylene biosynthesis, whereas the number of genes for proteins involved in ethylene perception and signaling is higher in *Populus*. For example, *Populus* has seven predicted genes for ethylene receptor proteins and *Arabidopsis* has five; the constitutive triple response kinase that acts just downstream of the receptor is encoded by four genes in *Populus* and only one in *Arabidopsis* (*78*). The number of ethylene-responsive element binding factor (ERF) proteins (a subfamily of the AP2/ERF family) is higher in *Populus* than in *Arabidopsis* (172 versus 122, respectively). The increased variation in the number of ERF transcription factors may be involved in the ethylene-dependent processes specific to trees, such as tension wood formation (*68*) and the establishment of dormancy (*71*).

Conclusions

Our initial analyses provide a flavor of the opportunities for comparative plant genomics made possible by the generation of the *Populus* genome sequence. A complex history of whole-genome duplications, chromosomal rearrangements and tandem duplications has shaped the genome that we observe today. The differences in gene content between *Populus* and *Arabidopsis* have provided some tantalizing insights into the possible molecular bases of their strongly contrasting life histories, although factors unrelated to gene content (such as regulatory elements, miRNAs, posttranslational modification, or epigenetic modifications) may ultimately be of equal or greater importance. With the sequence of *Populus*, researchers can now go beyond what could be learned from *Arabidopsis* alone and explore hypotheses to linking genome sequence features to wood development, nutrient and water movement, crown development, and disease resistance in perennial plants. The availability of the *Populus* genome sequence will enable continuing comparative genomics studies among species that will shed new light on genome reorganization and gene family evolution. Furthermore, the genetics and population biology of *Populus* make it an immense source of allelic variation. Because *Populus* is an obligate outcrossing species, recessive alleles tend to be maintained in a heterozygous state. Informatics tools enabled by the sequence, assembly, and annotation of the *Populus* genome will facilitate the characterization of allelic variation in wild *Populus* populations adapted to a wide range of environmental conditions and gradients over large portions of the Northern Hemisphere. Such variants represent a rich reservoir of molecular resources useful in biotechnological applications, development of alternative energy sources, and mitigation of anthropogenic environmental problems. Finally, the keystone role of *Populus* in many ecosystems provides the first opportunity for the application of genomics approaches to questions with ecosystem-scale implications.

References and Notes

1. Food and Agricultural Organization of the United Nations, *State of the World's Forests 2003* (FAO, Rome, 2003).
2. R. F. Stettler, H. D. Bradshaw Jr., in *Biology of* Populus *and Its Implications for Management and Conservation*, R. F. Stettler, H. D. Bradshaw Jr., P. E. Heilman, T. M. Hinckley, Eds. (NRC Research Press, Ottawa, 1996), pp. 1–7.
3. G. A. Tuskan, S. P. DiFazio, T. Teichmann, *Plant Biol.* **6**, 2 (2004).
4. T. M. Yin, S. P. DiFazio, L. E. Gunter, D. Riemenschneider, G. A. Tuskan, *Theor. Appl. Genet.* **109**, 451 (2004).
5. R. Meilan, C. Ma, in *Agrobacterium Protocols*, vol. 344 of *Methods in Molecular Biology*, K. Wang, Ed. (Humana Press, Totowa, NJ, 2006), pp. 143–151.
6. G. A. Tuskan, *Biomass Bioenerg.* **14**, 307 (1998).
7. G. A. Tuskan, M. Walsh, *For. Chron.* **77**, 259 (2001).
8. S. Wullschleger *et al.*, *Can. J. For. Res.* **35**, 1779 (2005).
9. Materials and methods are available as supporting material on *Science* Online.
10. H. D. Bradshaw, R. F. Stettler, *Theor. Appl. Genet.* **86**, 301 (1993).
11. M. Koornneef, P. Fransz, H. de Jong, *Chromosome Res.* **11**, 183 (2003).
12. O. Santamaria, J. J. Diez, *For. Pathol.* **35**, 95 (2005).
13. G. A. Tuskan *et al.*, *Can. J. For. Res.* **34**, 85 (2004).
14. A. A. Salamov, V. V. Solovyev, *Genome Res.* **10**, 516 (2000).
15. E. Birney, R. Durbin, *Genome Res.* **10**, 547 (2000).
16. T. Schiex, A. Moisan, P. Rouzé, in *Computational Biology: Selected Papers from JOBIM'2000, number 2066 in LNCS* (Springer-Verlag, Heidelberg, Germany, 2001), pp. 118–133.
17. Y. Xu, E. C. Uberbacher, *J. Comput. Biol.* **4**, 325 (1997).
18. S. J. Hanley, M. D. Mallott, A. Karp, *Tree Genet. Genomes*, in press.
19. M. A. Koch, B. Haubold, T. Mitchell-Olds, *Mol. Biol. Evol.* **17**, 1483 (2000).
20. M. Lynch, J. S. Conery, *Science* **290**, 1151 (2000).
21. L. Sterck *et al.*, *New Phytol.* **167**, 165 (2005).
22. L. A. Dode, *Bull. Soc. Hist. Nat. Autun* **18**, 161 (1905).
23. R. Regnier, in *Revue des Sociétés Savantes de Normandie* (Rouen, France, 1956), vol. 1, pp. 1–36.
24. M. E. Collinson, *Proc. R. Soc. Edinburgh B Bio. Sci.* **98**, 155 (1992).
25. J. E. Eckenwalder, in *Biology of Populus and Its Implications for Management and Conservation*, R. F. Stettler, H. D. Bradshaw Jr., P. E. Heilman, T. M. Hinckley, Eds. (NRC Research Press, Ottawa, 1996), chap. 1.
26. J. B. Mitton, M. C. Grant, *Bioscience* **46**, 25 (1996).
27. K. Hokamp, A. McLysaght, K. H. Wolfe, *J. Struct. Funct. Genomics* **3**, 95 (2003).
28. J. E. Bowers, B. A. Chapman, J. K. Rong, A. H. Paterson, *Nature* **422**, 433 (2003).
29. L. M. Zahn, J. Leebens-Mack, C. W. dePamphilis, H. Ma, G. Theissen, *J. Hered.* **96**, 225 (2005).
30. S. De Bodt, S. Maere, Y. Van de Peer, *Trends Ecol. Evol.* **20**, 591 (2005).
31. K. L. Adams, J. F. Wendel, *Trends Genet.* **21**, 539 (2005).
32. G. Blanc, K. Hokamp, K. H. Wolfe, *Genome Res.* **13**, 137 (2003).
33. B. A. Schulman *et al.*, *Nature* **408**, 381 (2000).
34. S. Griffiths-Jones *et al.*, *Nucleic Acids Res.* **33**, D121 (2005).
35. S. Lockton, B. S. Gaut, *Trends Genet.* **21**, 60 (2005).
36. B. A. Chapman, J. E. Bowers, F. A. Feltus, A. H. Paterson, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 2730 (2006).
37. T. A. Richmond, C. R. Somerville, *Plant Physiol.* **124**, 495 (2000).
38. S. Djerbi, M. Lindskog, L. Arvestad, F. Sterky, T. T. Teeri, *Planta* **221**, 739 (2005).

39. C. P. Joshi *et al.*, *New Phytol.* **164**, 53 (2004).
40. A. Samuga, C. P. Joshi, *Gene* **334**, 73 (2004).
41. F. Roudier *et al.*, *Plant Cell* **17**, 1749 (2005).
42. R. M. Perrin *et al.*, *Science* **284**, 1976 (1999).
43. R. W. Whetten, J. J. Mackay, R. R. Sederoff, *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **49**, 585 (1998).
44. J. Ehlting *et al.*, *Plant J.* **42**, 618 (2005).
45. J. Raes, A. Rohde, J. H. Christensen, Y. Van de Peer, W. Boerjan, *Plant Physiol.* **133**, 1051 (2003).
46. D. M. O'Malley, S. Porter, R. R. Sederoff, *Plant Physiol.* **98**, 1364 (1992).
47. J. J. Mackay, W. W. Liu, R. Whetten, R. R. Sederoff, D. M. Omalley, *Mol. Gen. Genet.* **247**, 537 (1995).
48. J. Schrader *et al.*, *Plant Cell* **16**, 2278 (2004).
49. S. Whitham, S. McCormick, B. Baker, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 8776 (1996).
50. G. M. Gebre, T. J. Tschaplinski, G. A. Tuskan, D. E. Todd, *Tree Physiol.* **18**, 645 (1998).
51. G. Arimura, D. P. W. Huber, J. Bohlmann, *Plant J.* **37**, 603 (2004).
52. D. J. Peters, C. P. Constabel, *Plant J.* **32**, 701 (2002).
53. C.-J. Tsai, S. A. Harding, T. J. Tschaplinski, R. L. Lindroth, Y. Yuan, *New Phytol.* **172**, 47 (2006).
54. M. M. De Sá, R. Subramaniam, F. E. Williams, C. J. Douglas, *Plant Physiol.* **98**, 728 (1992).
55. R. L. Lindroth, S. Y. Hwang, *Biochem. Syst. Ecol.* **24**, 357 (1996).
56. B. Winkel-Shirley, *Curr. Opin. Plant Biol.* **5**, 218 (2002).
57. G. J. Tanner *et al.*, *J. Biol. Chem.* **278**, 31647 (2003).
58. S. Aubourg, A. Lecharny, J. Bohlmann, *Mol. Genet. Genomics* **267**, 730 (2002).
59. M. A. Costa *et al.*, *Phytochemistry* **64**, 1097 (2003).
60. D. Cukovic, J. Ehlting, J. A. VanZiffle, C. J. Douglas, *Biol. Chem.* **382**, 645 (2001).
61. J. M. Shockey, M. S. Fulda, J. Browse, *Plant Physiol.* **132**, 1065 (2003).
62. I. E. Somssich, P. Wernert, S. Kiedrowski, K. Hahlbrock, *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14199 (1996).
63. L. Li *et al.*, *Plant Cell* **13**, 1567 (2001).
64. B. C. Meyers, S. Kaushik, R. S. Nandety, *Curr. Opin. Plant Biol.* **8**, 129 (2005).
65. L. Deslandes *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 2404 (2002).
66. E. J. Mellerowicz, M. Baucher, B. Sundberg, W. Boerjan, *Plant Mol. Biol.* **47**, 239 (2001).
67. A. P. Mähönen *et al.*, *Science* **311**, 94 (2006).
68. S. Andersson-Gunneras *et al.*, *Plant J.* **34**, 339 (2003).
69. J. M. Hellgren, K. Olofsson, B. Sundberg, *Plant Physiol.* **135**, 212 (2004).
70. M. G. Cline, K. Dong-Il, *Ann. Bot. (London)* **90**, 417 (2002).
71. R. Ruonala, P. Rinne, M. Baghour, H. Tuominen, J. Kangasjärvi, *Plant J.*, in press.
72. R. Moyle *et al.*, *Plant J.* **31**, 675 (2002).
73. D. Weijers *et al.*, *EMBO J.* **24**, 1874 (2005).
74. G. Hagen, T. Guilfoyle, *Plant Mol. Biol.* **49**, 373 (2002).
75. M. E. Eriksson, M. Israelsson, O. Olsson, T. Moritz, *Nat. Biotechnol.* **18**, 784 (2000).
76. T. Kakimoto, *Science* **274**, 982 (1996).
77. T. Kiba, K. Aoki, H. Sakakibara, T. Mizuno, *Plant Cell Physiol.* **45**, 1063 (2004).
78. T. Nakano, K. Suzuki, T. Fujimura, H. Shinshi, *Plant Physiol.* **140**, 411 (2006).
79. F. Sterky *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 13951 (2004).
80. J. Ehlting *et al.*, *Plant J.* **42**, 618 (2005).
81. B. C. Meyers, S. Kaushik, R. S. Nandety, *Curr. Opin. Plant Biol.* **8**, 129 (2005).
82. M. W. Jones-Rhoades, D. P. Bartel, *Mol. Cell* **14**, 787 (2004).

83. We thank the 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 Umeå University and the Royal Technological Institute (KTH) in Stockholm 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 NSF Plant Genome Program for supporting the development of Web-based tools; T. 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 the anonymous reviewers who provided critical input and recommendations on the manuscript. GenBank Accession Number: AARH00000000.

Supporting Online Material

www.sciencemag.org/cgi/content/full/313/5793/1596/DC1
Materials and Methods
Figs. S1 to S15
Tables S1 to S14
References

13 April 2006; accepted 9 August 2006
10.1126/science.1128691

Opposing Activities Protect Against Age-Onset Proteotoxicity

Ehud Cohen,[1]* Jan Bieschke,[2]* Rhonda M. Perciavalle,[1] Jeffery W. Kelly,[2] Andrew Dillin[1]†

Aberrant protein aggregation is a common feature of late-onset neurodegenerative diseases, including Alzheimer's disease, which is associated with the misassembly of the $A\beta_{1-42}$ peptide. Aggregation-mediated $A\beta_{1-42}$ toxicity was reduced in *Caenorhabiditis elegans* when aging was slowed by decreased insulin/insulin growth factor–1–like signaling (IIS). The downstream transcription factors, heat shock factor 1, and DAF-16 regulate opposing disaggregation and aggregation activities to promote cellular survival in response to constitutive toxic protein aggregation. Because the IIS pathway is central to the regulation of longevity and youthfulness in worms, flies, and mammals, these results suggest a mechanistic link between the aging process and aggregation-mediated proteotoxicity.

Late-onset human neurodegenerative diseases including Alzheimer's (AD), Huntington's, and Parkinson's diseases are genetically and pathologically linked to aberrant protein aggregation (*1*, *2*). In AD, formation of aggregation-prone peptides, particularly $A\beta_{1-42}$, by endoproteolysis of the amyloid precursor protein (APP) is associated with the disease through an unknown mechanism (*3*, *4*). Whether intracellular accumulation or extracellular deposition of $A\beta_{1-42}$ initiates the pathological process is a key unanswered question (*5*). Typically, individuals who carry AD-linked mutations present with clinical symptoms during their fifth or sixth decade, whereas sporadic cases appear after the seventh decade. Why aggregation-mediated toxicity emerges late in life and whether it is mechanistically linked to the aging process remain unclear.

Perhaps the most prominent pathway that regulates life span and youthfulness in worms, flies, and mammals is the insulin/insulin growth factor (IGF)–1–like signaling (IIS) pathway (*6*). In the nematode *Caenorhabditis elegans*, the sole insulin/IGF-1 receptor, DAF-2 (*7*), initiates the transduction of a signal that causes the phosphorylation of the FOXO transcription factor, DAF-16 (*8*, *9*), preventing its translocation to the nucleus (*10*). This negative regulation of DAF-16 compromises expression of its target genes, decreases stress resistance, and shortens the worm's life span. Thus, inhibition of *daf-2* expression creates long-lived, youthful, stress-resistant worms (*11*). Similarly, suppression of the mouse DAF-2 ortholog, IGF1-R, creates long-lived mice (*12*). Recent studies indicate that, in worms, life-span extension due to reduced *daf-2* activity is also dependent upon heat shock factor 1 (HSF-1). Moreover, increased expression of *hsf-1* extends worm life span in a *daf-16*–dependent manner (*13*). That the DAF-16 and HSF-1 transcriptomes result in the expression of numerous chaperones (*13*, *14*) suggests that the integrity of protein folding could play a key role in life-span determination and the amelioration of aggregation-associated proteotoxicity. Indeed, amelioration of Huntington-associated proteotoxicity by slowing the aging process in worms has been reported (*13*, *15*, *16*).

Reduced IIS activity lowers $A\beta_{1-42}$ toxicity. One hypothesis to explain late-onset aggregation-associated toxicity posits that the deposition of toxic aggregates is a stochastic process, governed by a nucleated polymerization and requiring many years to initiate disease. Alternatively, aging could enable constitutive aggregation to become toxic as a result of declining detoxification activities. To distinguish between these two possibilities, we asked what role the aging process plays in $A\beta_{1-42}$ aggregation-mediated toxicity in a *C. elegans* model featuring intracellular $A\beta_{1-42}$ expression (*17*). If $A\beta_{1-42}$ toxicity results from a non-age-related nucleated polymerization, animals that express $A\beta_{1-42}$ and whose life span has been extended would be expected to succumb to $A\beta_{1-42}$ toxicity at the same rate as those with a natural life span. However, if the aging process plays a role in detoxifying an ongoing protein aggregation process, alteration of the aging program

[1]Molecular and Cell Biology Laboratory, Salk Institute for Biological Studies, 10010 North Torrey Pines Road, La Jolla, CA 92037, USA. [2]Department of Chemistry and Skaggs Institute of Chemical Biology, Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed E-mail: dillin@salk.edu