Functional Genomics of Wood Quality and Properties

Wei Tang[1]*, Xiaoyan Luo[2], Aaron Nelson[1], Hilary Collver[1], and Katherine Kinken[1]

[1] *Department of Biology, Howell Science Complex, East Carolina University, Greenville, NC 27858, USA;*
[2] *Department of Cell and Developmental Biology, University of North Carolina, Chapel Hill, NC 27599, USA.*

Genomics promises to enrich the investigations of biology and biochemistry. Current advancements in genomics have major implications for genetic improvement in animals, plants, and microorganisms, and for our understanding of cell growth, development, differentiation, and communication. Significant progress has been made in the understanding of plant genomics in recent years, and the area continues to progress rapidly. Functional genomics offers enormous potential to tree improvement and the understanding of gene expression in this area of science worldwide. In this review we focus on functional genomics of wood quality and properties in trees, mainly based on progresses made in genomics study of *Pinus* and *Populus*. The aims of this review are to summarize the current status of functional genomics including: (1) Gene discovery; (2) EST and genomic sequencing; (3) From EST to functional genomics; (4) Approaches to functional analysis; (5) Engineering lignin biosynthesis; (6) Modification of cell wall biogenesis; and (7) Molecular modelling. Functional genomics has been greatly invested worldwide and will be important in identifying candidate genes whose function is critical to all aspects of plant growth, development, differentiation, and defense. Forest biotechnology industry will significantly benefit from the advent of functional genomics of wood quality and properties.

Key words: gene discovery, genomic sequencing, lignin biosynthesis, cell wall biogenesis, expression profiling

Introduction

Genome research projects are now producing enormous quantities of sequence data (*1*, *2*). The human genome, for instance, with its sequence of about 3×10^9 bp, has been completed (3, 4). However, huge amounts of genomic data from all genomic projects are being gathered with little practical value so far. The physiological functions of genome sequences are widely unknown. Time and investment are needed before the benefits to breeding and genetic conservation can be realized (*3*). Trees represent a unique life form and have developed a perennial lifestyle that produces the majority of terrestrial biomass (*5*, *6*). There are differences between trees and annual plants in gene flow, genetic diversity, and the link among molecular genetics, physiology and yield. Trees are the majority of the forestry, and wood-processing industries depend upon them for the economies of timber, pulp, and paper (*7*–*9*). The increasing demands for forestry products is likely to require greater forest productivity and more intensive research to create novel products from wood. Forest biologists have developed strong justifications for why trees should be viewed as model systems in plant biology (*6*, *10*).

Trees are different from annual, herbaceous plants by perennial growth, large size, complex crown architecture, extensive secondary xylem, dormancy, and juvenile-mature phase changes (*11*, *12*). The genomes of trees have being sequenced in pine and poplar. After the effort to sequence the entire genome of the poplar tree, a full-scale functional genomics effort on trees will set a completely new agenda for forest research (*6*, *12*). Although efforts to identify *Populus* as a model tree began long before the time when sequencing a tree genome was a possibility, the choice of poplar was ideal in that the genome size is small (about 550 Mb). The genome size of poplar is similar to that of rice, only four times larger than that of *Arabidopsis*, yet 40 to 50 times smaller than that of pine

*** Corresponding author.**
E-mail: tangw@mail.ecu.edu

(*6*, *12*, *13*). Work on other tree species will benefit from the progress being made with poplar and with some pine species.

A challenge for forest biologists in the future is to ensure that the forest industry benefits from rapidly developing genomic and biotechnological advances (*14*, *15*). Forest biotechnology could derive enormous advantages from information generated through functional genomics approaches. As we have seen for rice and *Arabidopsis* (*16–18*), the sequencing of a tree genome promises to enrich the study of forest biology. Used in conjunction with microarrays, metabolomics, high-efficiency transformation technologies, and high-throughput phenotyping, sequence data will enable researchers to attain a truly mechanistic understanding of tree function (*5*, *6*). Information from the poplar genome will allow fundamental questions to be asked not only in tree biology, but also in forestry and the ecological sciences, because this genus is distributed widely across the northern hemisphere (*6*, *19*, *20*). The purposes of this review are to summarize the current status of genomics in forest trees, to consider potential uses of genomics in novel areas of biotechnological research, and to identify concerns likely to arise from the rapid anticipated growth of forest biotechnology industries.

Gene Discovery

Gene discovery projects can help researchers identify important genes and understand their function in a microorganism, plant, and animal species, such as how to improve productivity, resistance to disease, and environmental adaptability (*5*, *21*, *22*). The knowledge and the molecular biological techniques developed and used by gene discovery projects will benefit genomic research including developing mutant seeds, databases, and various tools that could be used to determine the function of many plant genes (*23*, *24*). With that foundation established, researchers could more efficiently pursue studies that will improve crop yields and tree production, contribute to our understanding of crop and tree genetics, and promote fundamental discoveries in plant biology (*25–27*). Two complementary approaches, expressed sequence tags (ESTs) and sequencing genomic DNA, are used in discovering novel genes. Forest genomics began when EST projects were initiated in pine and poplar, after the pioneering work of Venter and colleagues (*28*) had proven the value of EST sequencing as a cheap but efficient method of finding putative tissue-specific genes. The first publications reported about 5,000 (*14*) and 1,000 EST sequences for the poplar and pine species (*29*). Based on ESTs derived from the gene discovery projects, new bioinformatics tools and links among DNA sequences, gene expression patterns and the phenotypic consequences of mutations in specific genes will be developed by the projects (*3*, *5*).

To date, 110,622 ESTs have been sequenced from loblolly pine (*Pinus taeda* L.) and 65,981 ESTs from poplar (*Populus tremula x Populus tremuloides*) (Table 1). Furthermore, EST collections with more than 5,000 sequences have been sequenced from *Populus tremula* (31,288), *Populus balsamifera subsp. trichocarpa* (26,825), *Pinus pinaster* (15,719), *Populus tremuloides* (12,813), *Populus x canescens* (10,446), and *Populus balsamifera subsp. trichocarpa x Populus deltoides* (6,579) (Table 2). Other poplar and pine species with ESTs collection include *Populus alba x Populus glandulosa*, *Pinus radiata* (Monterey pine), *Pinus banksiana*, *Pinus patula*, and *Pinus elliottii* (Table 2). In addition to these academic efforts, impressive EST projects based on radiata pine and eucalyptus have been undertaken and reported by industrial laboratories (*5*, *6*). Using these EST resources, we may be able to elucidate the genetic basis for the great differences in wood quality observed between gymnosperms and angiosperms. Unlike earlier activities, where the objective was simply to identify the main sequences expressed in the species being considered, more recent efforts have focused on the creation and comparison of multiple cDNA libraries (*5*, *30*, *31*). These libraries were made from RNA isolated from a variety of tissues and from plants either in various developmental stages or subjected to different treatments. Nevertheless, these sequence information could provide biologists considerably more knowledge about the genetic composition of trees than we did previously (*5*, *11*).

When we consider the overall biodiversity represented within the EST libraries, most sequences are attributed to either model plant species (*Arabidopsis, Chlamydomonas, Physcomitrella*) or species of agricultural or agronomic interest (rice, maize, soybean ref. *5*). Represented species are restricted to just a few groups within the plant evolutionary tree (*3*, *7*). There is no evidence upon which we can consider our currently completed plant genomes or the genomes with deeply sampled EST collections (Tables 1 and 2) as being taxonomically representative beyond their

Table 1 Species Ranked by the Available Number of ESTs with More than 50,000 Sequences

Rank	Species	Number of ESTs*
1	*Homo sapiens* (human)	5,469,433
2	*Mus musculus* + *domesticus* (mouse)	4,030,839
3	*Rattus sp.* (rat)	558,402
4	*Triticum aestivum* (wheat)	549,915
5	*Ciona intestinalis*	492,511
6	*Gallus gallus* (chicken)	451,655
7	*Danio rerio* (zebrafish)	405,962
8	*Zea mays* (maize)	391,145
9	*Xenopus laevis* (African clawed frog)	357,038
10	*Hordeum vulgare* + *subsp. vulgare* (barley)	348,282
11	*Glycine max* (soybean)	344,524
12	*Bos taurus* (cattle)	331,139
13	*Silurana tropicalis*	297,086
14	*Drosophila melanogaster* (fruit fly)	267,332
15	*Oryza sativa* (rice)	266,949
16	*Saccharum officinarum*	246,301
17	*Sus scrofa* (pig)	240,001
18	*Caenorhabditis elegans* (nematode)	215,200
19	*Arabidopsis thaliana* (thale cress)	196,904
20	*Medicago truncatula* (barrel medic)	187,763
21	*Sorghum bicolor* (sorghum)	161,766
22	*Dictyostelium discoideum*	155,032
23	*Chlamydomonas reinhardtii*	154,600
24	*Lycopersicon esculentum* (tomato)	150,410
25	*Schistosoma mansoni* (blood fluke)	139,135
26	*Oncorhynchus mykiss* (rainbow trout)	137,127
27	*Vitis vinifera*	135,712
28	*Anopheles gambiae* (African malaria mosquito)	134,784
29	*Solanum tuberosum* (potato)	132,122
30	*Pinus taeda* (loblolly pine)	110,622
31	*Oryzias latipes* (Japanese medaka)	103,098
32	*Physcomitrella patens subsp. patens*	82,313
33	*Toxoplasma gondii*	72,859
34	*Lactuca sativa*	68,188
35	*Populus tremula x Populus tremuloides*	65,981
36	*Helianthus annuus*	59,841
37	*Salmo salar*	59,420
38	*Strongylocentrotus purpuratus* (purple urchin)	51,744

* A summary for all ESTs available within the NCBI dbEST database is available from http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.

Table 2 Pine and Poplar Species Ranked by the Available Number of ESTs

Species	Family	Number of ESTs*
Populus tremula	Salicaceae	31,288
Populus balsamifera subsp. trichocarpa	Salicaceae	26,825
Pinus pinaster	Pinaceae	15,719

Table 2 ***Continued***

Species	Family	Number of ESTs*
Populus tremuloides	Salicaceae	12,813
Populus x canescens	Salicaceae	10,446
Populus balsamifera subsp. trichocarpa x Populus deltoides	Salicaceae	6,579
Populus alba x Populus glandulosa	Salicaceae	519
Pinus radiata (Monterey pine)	Pinaceae	69
Pinus banksiana	Pinaceae	46
Pinus patula	Pinaceae	23
Pinus elliottii	Pinaceae	8

* A summary for all ESTs available within the NCBI dbEST database is available from http://www.ncbi.nlm.nih.gov/dbEST/dbEST_summary.html.

most immediate clades. This naturally poses limitations on the scope and types of comparative analyses that can be performed using the currently available plant EST sequences (*3*).The taxonomically rich sequence diversity already existing within and between the individual groups certainly has the potential to be used to address specific questions about the conservation of protein families between well-sampled groups (*3*). The need for a more even sampling of plant genomes has recently been discussed, and there are many genomes that could be the focus of complete genome sequencing (*32*, *33*). With the complications of complete plant genome sequencing, deep EST sampling from a broader collection of currently unsampled taxa might offer us a better glimpse of the functional and evolutionary processes that are fundamental to plant life (*3*). There are two main problems associated with EST sequences: (1) the overall representation of host genes within a library, and (2) the overall quality of any individual sequence within a collection (*3*).The uneven representation of cDNA clones within the underlying libraries, however, can be addressed. Both oligofingerprinting (*5*, *6*, *11*, *12*, *34*) and normalization/subtraction (*35*) of cDNA libraries have been used to equalize the relative occurrence of the common and rarer transcripts, and have recently accounted for a leap in the sequence diversity reflected within some cDNA libraries (*3*).

EST and Genomic Sequencing

ESTs are small DNA molecules reverse-transcribed from the cellular mRNA population (*36*, *37*). EST sequencing is the most attractive route for broad sampling of the transcriptome (*5*, *38*). ESTs provide a robust sequence resource that can be exploited for gene discovery, genome annotation and comparative genomics (*3*, *5*). Over fifteen million sequences from approximately 633 species have been deposited in the publicly available plant EST sequence databases. Many of the ESTs have been sequenced as an alternative to complete genome sequencing or as a substrate for cDNA array-based expression analyses. Among EST collections with more than 50,000 sequences from 38 species, two pine and poplar tree species are ranked 30 (*Pinus taeda* L.) and 35 (*Populus tremula x Populus tremuloides*) (Table 1). Although EST collections are certainly no substitute for a whole genome scaffold, this resource forms the core foundations for various genome-scale experiments within the genomes. The entire genome projects have been established in human, chimpanzee, mouse, rat, zebrafish, fugu, mosquito, fruitfly, *C. elegans*, and *C. briggsae*; *Arabidopsis*, wheat, rice, barley, soybean, potato, and tomato (*3*, *5*, *28*). The first entire tree genome project is the *Populus* Genome project established recently (*11*).

Although EST sequencing is a cheap and quick way to identify expressed genes, the complete genomic sequence of species will allow biologists systemically analyze the structure, function, and evolution of genetic information (*3*, *5*, *14*). The genomic sequence of a tree is necessary for several reasons. Firstly, it is highly unlikely that all the genes of any tree will be identified by EST sequencing alone (*5*, *11*). Secondly, even if there are several hundred genes unique to trees, it would be extremely useful to identify their individual contributions to the observed architectural and other differences between simple annual weeds like *Arabidopsis* and trees (*3*, *11*). Thirdly, acquisition of a full tree genome sequence would be very valuable for quantitative trait locus (QTL) analysis, marker-

assisted breeding and, importantly, the genome sequence from one tree could be used as a platform for identifying synteny among tree species, as has been done for *Arabidopsis* and *Brassica* and other species (*5*, *11*). Therefore, the news that the United States Department of Energy (DOE) has decided to sequence the genome of poplar was welcome to all tree biologists (www.ornl.gov/ipgc). The DOE's Joint Genome Institute (www.jgi.doe.gov) was expected to produce six times coverage of the entire genome sequence during 2003. However, without further advances in sequencing and bioinformatics, it seems unlikely that we will obtain genomic information from any gymnosperm in the near future. This is because gymnosperms have a massive genome with haploid DNA contents of, on average, 15,500 Mb (*39*), as compared with 125 Mb for *Arabidopsis* (*17*, *18*) and 550 Mb for poplar (*11*).

With the advent of cDNA array-based methodologies, ESTs have become a key reagent within an experiment rather than the final product (*40*). Plant genome sizes extend over at least four orders of magnitude. *Arabidopsis* and *Oryza* sativa (rice), our model plants with fully sequenced genomes, have among the smallest known genomes: 125 Mb and 430 Mb, respectively. Tomato has a genome size of 950 Mb (*41*) and maize has a genome size of 2,670 Mb. Cycad and wheat have genome sizes of 14,000 Mb and 17,000 Mb, respectively (*3*, *11*). The largest known genomes are currently those of *Fritillaria assyriaca* (125,000 Mb) and *Psilotum nudum* (250,000 Mb) (http://www.rbgkew.org.uk; ref. *3*, *42*). The expansion of genomes has mainly been the result of multiplication of retrotransposon repeat sequences. In maize, such retrotransposons have accounted for the doubling of the genome size during the past six million years (*43–45*). Retrotransposons have been shown to aggregate within the gene space and their presence has been used to explain the narrow range of GC percentages within the gene space isochors (*46*, *47*). Although the main emphasis of plant genome sequencing is currently on discovering and characterizing the range of protein-coding genes present within the genome, thousands of copies of large repeats yield no information on the proteome.

Complete genome sequences have been produced for *Arabidopsis* (*17*, *18*) and rice (*16*, *22*, *48*). The complete genome scaffolds for *Zea mays, Medicago truncatula, Brassica napus* and *Populus* are either within the sequencing or preparation stage and other plant genomes will follow (*3*). ESTs really spring into the limelight when we are presented with a new complete genome sequence and wish to start annotating genes to the chromosomes. Although the underlying methods and science required for the detection and modelling of eukaryotic genes have been well described elsewhere (*49*, *50*), one universal theme is the strong value and dependence placed on ESTs, first within the identification of the gene regions for training the gene prediction algorithms and, second, within the validation and correction of genes that have been predicted using the trained gene modelling algorithms (*51*). ESTs have also demonstrated their worth in the selection of apparently unannotated proteins and putative small peptides from *Arabidopsis* (*52*, *53*). This EST and cDNA approach has also been used to annotate the UTRs of genes, to correct the boundaries of introns and exons, and to identify new introns (especially within the UTRs) and probable micro-exons. ESTs have also been used to discover non-canonical splice sites (*54*, *55*). On the basis of EST data, alternative splicing has been shown to be a rare occurrence within plants, although examples can be found (*56*). This contrasts greatly with the mammalian system, in which alternative splicing is widespread. ESTs are invaluable within genome annotation and, with the arrival of new genomes, more ESTs and full length cDNAs are sure to follow. Issues with annotation of the rice genome have interestingly been partly attributed to the lack of high quality ESTs and full length cDNAs (*54*, *55*).

From EST to Functional Genomics

ESTs represent an informative tool for gene discovery. It was reported on an extensive EST library being developed as part of the Swedish *Populus* Genome project, a joint collaboration between UPSC and the Genome Center at the Royal Institute of Technology in Stockholm (*3*, *6*). In the initial phase of this project, almost 5,700 ESTs were developed for wood-forming tissues (*14*). This resource has grown to 95,000 ESTs sequenced from 20 different cDNA libraries and from a range of tissues and developmental stages. Analyses indicate that these ESTs derive from perhaps 15,000 to 20,000 genes, a significant fraction of the 40,000 to 50,000 genes believed to be coded by the *Populus* genome (*3*, *5*, *6*). Basic Local Alignment Search Tool (BLAST) searches against sequenced ESTs are possible through the project database or data deposited in GenBank. Al-

though a functional classification of all 95,000 ESTs has not been completed, several subsets of the data have been analyzed (*3*, *6*). Table 3 shows the distribution of genes in various functional categories for young poplar leaves and leaves harvested before visible signs of senescence (*6*). According to Jansson, young leaves devote one third (36%) of their transcript pool to "energy", whereas older leaves have a high abundance of transcripts in categories such as "cell death and aging" and "protein destination", which includes functions related to proteolytic degradation (*3*, *6*). Most of the ESTs sequenced to date have been for hybrid aspen (*P. tremula x P. tremuloides*), European aspen (*P. tremula*), and *P. trichocarpa*. Functional distribution of genes according to a modified MIPS (Munich Information Center for Protein Sequences) classification scheme of 4,842 ESTs from young *Populus* leaves and 5,128 ESTs from leaves collected in autumn was listed in Table 3.

Table 3 Functional Distribution of Genes from 4,842 ESTs of Young *Populus* Leaves and 5,128 ESTs of Autumn Leaves (*6*)

Function	Young leaves	Autumn leaves
Cell rescue and defense	4%	11%
Cellular communication	3%	4%
Cellular organization	36%	8%
Energy	36%	8%
Metabolism	7%	7%
Protein destination	3%	7%
Protein synthesis	2%	4%
Transcription	2%	3%
Unclassified proteins	23%	21%
Blast<100	23%	28%
Other	2%	4%

To overcome this situation, different analysis tools have been developed in order to detect and understand the phenomena of gene regulation and physiological functions, in particular of the protein-coding genes (so-called open reading frames, ORFs; ref. *3*, *5*, *6*). Most of these tools are searching for sequence similarities comparing unknown genes with genes of known function from other organisms. This method is strictly limited to the assignment of genes with known functions (*6*, *11*). Therefore, to learn more about functionally unassigned ORFs (about 30% in the well-known microorganisms *Escherichia coli* and *Saccharomyces cerevisiae*), gene expression studies are to be combined with functional characterization assuming that under different physiological conditions individual genes may be differently expressed. Specific responses to certain stimuli, like the addition of certain natural products or the supply of certain substrates, will provide indications with respect to the functions of the induced genes (*3*, *5*, *11*). A promising approach is to analyze transcription profiles using DNA microarrays of all genes under changing conditions in connection with the available knowledge in databases. This can be described as supervised learning if knowledge is partially available and unsupervised learning if not. In the entire genome sequence of the microorganism *E. coli*, widely used in biotechnology for the production of recombinant proteins as well as in microbial research, 4,290 ORFs were identified (*3*, *11*).

When we consider individual nucleotides within an EST against their cognate genomic reference nucleotides, as many as 3% of the individual nucleotides can be incorrect (*57*), representing insertions, deletions and substitutions. The quality of individual nucleotides reflects the fidelity of the reverse transcriptase used within cDNA preparation (*58*), the fidelity of the sequencing reaction performed, and the accuracy with which the sequence has been determined from the electropherogram trace file (*59*). Full-length cDNA sequences are obtained by shotgun sequencing cDNA clones that have been selected for both 5′ and 3′ ends (*60*). Such a strategy yields many individual ESTs that can be assembled into a single contig. Bioinformatics-based sequence resources have been developed addressing the quality, redundancy and partial nature of EST sequences. Sequence resources such as the dbEST database (*40*) and the EMBL database (*61*) archive all the available ESTs and pro-

vide methods to search for individual sequences on the basis of species, clones or homology attributes. Although there are a range of methods that achieve this goal, they generally perform the same processing steps to achieve a common result. Sequences are aggressively trimmed of vectors and polylinker remnants before a fast clustering method places the ESTs into buckets of similar sequences (*62*). A final assembly step places the clustered sequences into logical contigs and singletons (*63*). The clustered sequences are typically longer than any individual EST. Cluster consensus sequences additionally merge valuable information on sequence polymorphisms that would otherwise not be observable. A collection of these sequence resources is shown in Table 4. Most of these sequence databases have added further value to the sequences by attaching additional annotation to the sequences and by providing methods to select specific sequences or groups of sequences that satisfy specific criteria (*3*). The most valuable annotations and methods are those that assign tentative function and allow retrieval and identification of sequences on the basis of tissue or challenge specificity (*3*).

Table 4 Specific Plant EST Databases with Significant Value and Large Collections of EST Sequences

Plant EST database	Websites	References
B-EST barley database	http://pgrc.ipk-gatersleben.de/est/login.php	*3*
Chlamydomonas resource centre	http://www.biology.duke.edu/chlamy_genome/	*66*
Kazusa EST database	http://www.kazusa.or.jp/en/plant/database.html	*3*
MIPS Sputniks	http://mips.gsf.de/proj/sputnik/	*3*
NCBI Unigenes	http://www.ncbi.nlm.nih.gov/UniGene/	*96*
PlantGDB	http://www.zmdb.iastate.edu/PlantGDB/	*3*
Solanaceae genomics network	http://sgn.cornell.edu/	*41*
TIGR Plant Gene Indices	http://www.tigr.org/tdb/tgi/plant.shtml	*94*
University Minnesota	http://www.ccgb.umn.edu/	*95*

ESTs as a current alternative to complete genomes could be applied as the foundation sequence of some genome-scale analyses. EST-derived cluster sequences have been widely annotated with tentative functions (*3*). Sources of functional annotation have included non-redundant protein databases (*64*), the *Arabidopsis* genome annotation (*41*), and catalogues of functionally assigned proteins (*3*, *54*). The annotations are homology based, and EST sequences or clusters inherit the annotative attributes of their matches. This approach naturally suffers from problems with the propagation of annotation errors, but manual validation of EST assignments has been shown to be consistent with such automated annotations (*65*). The surrogate annotation methods have been used to crudely dissect the overall representation and distribution of functional classes of protein both within and between genomes, and functional pie charts have become common within both genome and EST papers (*17*, *18*, *22*). With a selection of annotated proteins from a mixture of tissues from the same species, commonality can be observed among libraries. In *Chlamydomonas*, the EST resources have been used in a similar manner to select the genes that are most likely to be involved within stress responses by performing such *in silico* subtraction on genes found within abiotically challenged cells (*66*).

Approaches to Functional Analysis

As EST frequency gives rather rudimentary data on gene expression, to obtain a better understanding of the temporal and spatial expression patterns of different genes, information of transcript profiling is especially important for biotechnological purposes. For instance, it may be desirable to modify a metabolic pathway in a small subset of cells. In trees, a global transcript profiling was first established in poplar. The first microarray slide contained about 2,500 features and was used to investigate the molecular basis of xylem development (*30*, *31*). Recently, a new amplification technique was developed, allowing RNA to be isolated from submilligram amounts of tissues to generate probes for microarray analysis (*30*, *31*). Genetic maps of various quality have been generated for several forest tree species, using a variety of approaches. QTLs have been identified for a range of traits, such as wood density, fibre length and resis-

tance (*67*, *68*). It is apparent from work in *C. elegans*, *Arabidopsis* and yeast that development of a suitable model system provides a great tool for understanding complex biological processes. Similarly, poplar has been developed as an important model system for tree molecular genetics. The major advantages of poplar are its ease of transformation, relatively small genome size and rapid growth (*3*, *5*).

Global transcript profiling and other genomic technologies are also being used in poplar studies, and effective forward and reverse genetic screens are being designed. Collectively, these resources will make poplar not only the model for tree biology, but will also an excellent model for many unique aspects of plant biology associated with perenniality and developmental phase changes, adaptation to harsh climates, secondary growth, and secondary metabolism (*3*, *5*, *11*). While great strides have been taken towards developing poplar as a model system for tree biology, the value of *Arabidopsis* as a test bed for studying "tree" genes and their functions should not be underestimated. As more and more information is derived from exploratory experiments such as transcript profiling in poplar and other tree species, the scope for genetic analysis will be limited in trees (*5*, *69*, *70*). To date, the major focus of research in forest biotechnology has been to modify lignin and/or cellulose contents. For example, ways in which lignin levels can be modified and the consequences for wood and tree growth have been intensively studied in forest genomics (*9*, *69*, *71*).

Genomics research related to the control of cambial activity, which underlies wood production, may provide alternative approaches to enhance productivity. For example, it has been shown that overexpressing phytochrome can prevent growth cessation (*14*). Genes that play important roles in the regulation of diverse pathways were targeted, but for practical purposes it may be essential to identify genes that alter growth in a very specific manner. This is the kind of application for which genomic techniques such as transcript profiling and EST sequencing of tissue-specific libraries will be highly useful, notably for identifying candidate genes to alter cambial activity. Wood formation is a process that can be divided into a series of well-defined developmental events that are initiated in the vascular cambium (*72*). Cambial derivatives develop into xylem cells through the processes of division, expansion, secondary wall formation, lignification and, finally, programmed cell death (*5*, *13*). Therefore, wood engineering almost necessitates the use of genomics, as genomic approaches can provide information on the regulation of not just one gene or enzyme, but an entire pathway or several pathways at the same time. Recent microarray experiments by Hertzberg *et al* (*31*) demonstrated this point by providing expression profiles of over 2,300 genes across the developmental gradient during wood formation (*5*, *14*). These experiments not only clearly indicate the complexity that wood engineers will have to deal with, but also provide tempting glimpses into how regulators of specific aspects of wood development may be identified for modulation of wood properties.

Genetic analysis, mainly in *Arabidopsis*, has identified a large array of genes that appear to help determine the identity of the primordial and the SAM itself. However, owing to the limited extent of secondary growth in *Arabidopsis*, we still know very little about the genes regulating the growth of the vascular meristems (*5*, *6*, *13*). We know even less about the regulation of meristem function in trees, which show alternating cycles of growth and dormancy, as well as alternating periods of vegetative and reproductive growth. An important and apparently unique aspect of tree development is the late juvenility-to-maturity transition, which makes trees flower later than any other known plants. It is not clear whether this particular aspect of the regulation of tree flowering has a counterpart in annual plants such as *Arabidopsis* (*3*, *12*). Although genes regulating flower meristem identity appear to have conserved functions between *Arabidopsis* and trees (*73*, *74*), we still lack evidence demonstrating that any genes normally involved in regulating *Arabidopsis* flowering time also have a function in trees. It is important to characterize flower-specific genes from trees and isolate the corresponding promoters (*75–77*). One possible way of circumventing this is to test the sterility constructs in transgenic trees that have been engineered to flower early using the techniques described above. Alternatively, they could be tested in naturally early-flowering variants of otherwise late-flowering trees (*5*).

Transcript profiling can reveal patterns of gene expression during developmentally regulated events such as wood formation and leaf expansion and maturation. In trees, poplar microarrays were first used to study developmental processes involved in wood formation (*6*, *31*). In this case, microarrays helped characterize how gene expression varied throughout cell division, expansion, secondary wall formation, lig-

nification, and cell death. Preparation of a higher density *Populus* microarray based on the 95,000-EST library described above has been initiated (Table 3). As reported by Peter Nilsson from the Royal Institute of Technology in Stockholm, the Swedish *Populus* Genome project has produced a spotted EST microarray containing about 13,000 clones (*6*, *31*). The challenge, of course, will be to determine how changes in gene expression are related to altered biochemical and physiological function and, ultimately, to tree growth.

The creation of transgenic lines (Figure 1) with enhanced or reduced levels of gene expression is the most straightforward way to determine gene function. An ambitious program in which a commercial partner, SweTree Genomics, which will use RNA interference to knock out 2,000 genes involved in wood formation, and an activation tagging as an approach to the creation of dominant mutations in trees have been described (*5*, *6*). Activation tagging is a method whereby an enhancer element is inserted randomly into the genome, enhancing the expression of a nearby gene. These techniques are being used to identify genes with preferential expression in leaf vascular tissues, wood-forming tissues, roots, and adventitious root primordial (*6*, *14*). In addition to T-DNA–based methods, an alternative transposable tagging system is being explored by Sandeep Kumar and Matthias Fladung from the Institute for Forest Genetics and Forest Tree Breeding (Grosshansdorf, Germany) for use in generating mosaics of insertional and activation mutants within a single plant (*6*). It is anticipated that the number of transgenic lines produced by the methods described above will eventually exceed the capacity for their maintenance and characterization. Given that many of these techniques are just being developed, all speakers agreed that various analytical and computational challenges would be encountered in this area of investigation (*5*, *6*, *12*).

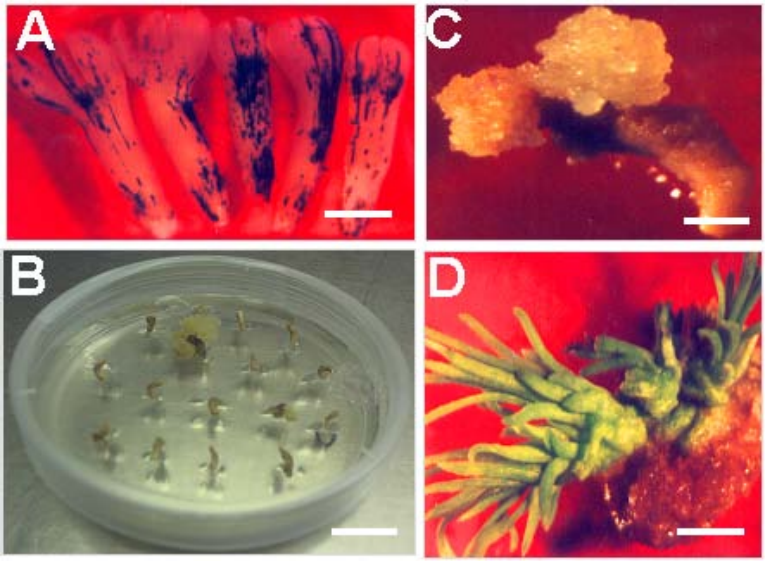

Fig. 1 Regeneration of transgenic loblolly pine from transgenic organogenic callus as a model for engineering wood quality and proterty. A. Transient *uidA* expression in transformed embryos (bar=0.4 cm); B. Establishment of kanamycin resistant calluses (bar=1.6 cm); C. Proliferation of kanamycin resistant calluses (bar=0.5 cm); D. transgenic loblolly pine shoots (bar=0.8 cm).

Teams would be formed to: (1) examine genetic and genomic resources currently available to *Populus* researchers; (2) identify areas in which tools, techniques, and additional resources must be developed; and (3) assess applications and opportunities for future research associated with the completion of the *Populus* genome sequence (*11*, *13*, *14*, *24*, *25*). Applications would emphasize tree growth and development in the context of general poplar culture, basic science investigations, bio-based products and energy, carbon sequestration, and forest responses to changes in physical and chemical climates. The formation of an international consortium and the development of a science plan were endorsed strongly by symposium and workshop participants (*6*). The consortium's World Wide Web site has been established (www.ornl.gov/ipgc).

Engineering Lignin Biosynthesis

Plant metabolic engineering requires the coordinate manipulation of multiple genes. Through metabolic engineering, wood could be improved for papermaking by making lignin easier to remove from cellulose during pulping (*78*, *79*). Metabolic engineering of wood would allow more environmentally friendly processes to be used to yield a cleaner cellulose pulp and the paper produced would be less prone to yellowing as it ages in the light (*6*, *7*). Although progress has mostly been limited to modulating the expression of single genes of well-studied pathways, such as the lignin biosynthetic pathway in model plants, a recent report illustrates a new level of sophistication in metabolic engineering by overexpressing one lignin enzyme while simultaneously suppressing the expression of another lignin gene in aspen (*80*, *81*). The best improvements in lignin properties will probably be gained by manipulating genes (Table 5) that are important to lignin structure, as well as genes that control lignin content. By overexpressing one lignin biosynthetic enzyme while suppressing the expression of a second lignin biosynthetic gene in a tree species, lignin and other plant metabolic products can be engineered in one step (*7*, *82*). Although there are relatively few examples of plant metabolic engineering where multiple genes have been manipulated, this strategy provides new insight to forest biotechnology.

Table 5 Lignin Enzymes Available to Engineer Its Properties by Manipulating Genes

Enzyme (EC)	Abbreviations	Names
EC 4.3.1.5	PAL	Phenylalanine ammonia-lyase
None	TAL	Tyrosine ammonia-lyase
EC 1.14.13.11	C4H	Cinnamate 4-hydroxylase
None	C3H	Coumarate 3-hydroxylase
EC 2.1.1.68	COMT	Caffeate O-methyltransferase
EC 2.1.1.104	CCoAOMT	Caffeoyl-Coenzyme A O-methyltransferase
EC 2.1.1.104	CCOMT	Caffeoyl-CoA 3-O-methyltransferase
None	F5H	Ferulate 5-hydroxylase
EC 6.2.1.12	4CL	4-coumarate-Coenzyme A ligase
EC 1.2.1.44	CCR	Cinnamoyl-Coenzyme A reductase
EC 1.1.1.195	CAD	Cinnamyl alcohol dehydrogenase

Different transgenes can be held on distinct T-DNAs and co-transformed into plants. The outlook for more widespread adoption of the co-transformation strategy is extremely promising because there have been several elegant demonstrations of its effectiveness recently (*7*, *83*). For example, four genes including a selectable marker were introduced into rice by co-transformation and enabled the grain to accumulate β-carotene (provitamin A), and the engineered grain can alleviate vitamin A deficiency in certain regions of the world (*7*, *84*). In *Arabidopsis*, six genes including two selectable marker genes were introduced into its cells to enable the production of a biodegradable plastic co-polymer (*85*). The manipulation of multiple genes in forest tree species presents problems distinct from those encountered with most conventional crops. The long generation time of trees rules out sexual crossing as a means of combining transgenes. However, the long rotation times typical for plantation forestry mean that transgene expression needs to be stable over a full rotation (*9*, *19*, *86*). The co-transferred T-DNAs frequently integrate at the same locus (*87*). This co-insertion and linkage of independent T-DNAs is an important feature of the system (*88*). A possible disadvantage of co-transformation is that the conditions that promote co-integration might also favour the integration of high numbers of transgene copies, which typically integrate as repeat structures, potentially increasing problems with transgene silencing in subsequent generations (*89*).

Lignin is a three-dimensional polymer of phenylpropanoid alcohols (monolignols) and is always associated with cell wall cellulose and hemicellulose to provide mechanical rigidity to plant-supporting and plant-conducting tissues. There are three monolignols: p-coumaryl alcohol, coniferyl alcohol, and sinapyl alcohol (*10*, *80*). Gymnosperm lignins contain mainly coniferyl alcohol, angiosperm lignins contain both coniferyl alcohol and sinapyl alcohol, whereas

all three types of monolignols are found within lignins of grasses. The transport of monolignols to the cell wall and their lignification to lignin is poorly understood. Global demand for wood products continually increases, creating strong pressure to improve commercial forest productivity while preserving native woodlands and biodiversity (*20*, *90*, *91*). Because conventional tree breeding is such a slow process, the idea of taking elite genotypes and further enhancing them by genetic manipulation is a particularly attractive one (*78*). Manipulations of several other tree genes have proved valuable for enhancing wood for papermaking. Inhibition of cinnamyl alcohol dehydrogenase, another lignin biosynthetic gene, yielded wood that was easier to pulp (*9*). Overexpression of a pine glutamine synthetase (*92*) and constitutive suppression of 4-coumarate-CoA ligase (*69*) enhanced the wood growth. Combining these different traits by multi-gene manipulation *via* co-transformation, genetically modified for herbicide tolerance, disease resistance or sterility, is an exciting idea for the future. Functional genomics paves the way to genetic engineering in commercial forestry by significantly improving wood quality and properties (*7*, *8*).

Modification of Cell Wall Biogenesis

The plant cell wall is a highly organized composite that may contain many different polysaccharides, proteins, and aromatic substances. The importance of the plant cell wall is revealed in the shear number of genes that are likely to be involved in cell wall biogenesis, assembly, and modification. In *Arabidopsis*, over 400 proteins have been identified that reside in the wall and over 2,000 genes are likely to participate in wall biogenesis during plant development (*15*). Beyond this, some integral membrane-associated proteins, such as cellulose synthase, obviously function in cell wall biogenesis. Thus, it is likely that some 15% of the *Arabidopsis* genome is dedicated to cell wall biogenesis and modification. Of these, only small subsets have been characterized. Recently, functional genomics approaches have provided insight into the genes relevant to cell wall metabolism (*15*, *21*, *23*). Reverse genetic and molecular biological approaches, based on discovery of homologous genes from bacteria, fungal, and animal systems, have augmented the collection of recognized wall-relevant genes considerably, but the functions of many of these genes still remain elusive (*17*, *18*, *21*, *23*).

According to Carpita *et al* (*15*), the major steps in wall biogenesis and modification can be divided into six specific stages: (1) the synthesis of monomer building blocks, such as nucleotide sugars and monolignols; (2) the biosynthesis of oligomers and polysaccharides at the plasma membrane and ER-Golgi apparatus; (3) the targeting and secretion of Golgi-derived materials; (4) the assembly and architectural patterning of polymers; (5) dynamic rearrangement during cell growth and differentiation; and (6) wall disassembly and catabolism of the spent polymers. To put into perspective the challenges of gene discovery and determination of function, functional genomics is predicted to make significant advances in this field. It reported a comprehensive summary of the complexities of pectin fine structure and how the use of monoclonal antibodies against pectin epitopes has revolutionized our knowledge of their cell and wall domain specificity and their dynamics during growth and development (*15*). In particular, antibodies directed against two neutral sugar side-groups, arabinans and galactans, have revealed a remarkable sub-domain distribution that will now allow more refined determinations of structural-functional and dynamic relationships of these transient components during cell growth and development (*15*, *17*, *18*, *21*).

Genomic approaches have played an important role in defining wall-relevant genes and provided a global view of gene expression related to primary and secondary cell wall synthesis. Henrissat *et al* provide a robust census of *Arabidopsis* glycosidases and glycosyltransferases derived from knowledge of the entire *Arabidopsis* genome sequence (*17*, *18*, *21*). One surprise of this census is that *Arabidopsis* encodes many more of these enzymes than does *Saccharomyces cerevisiae*, *Drosophila melanogaster* or *C. elegans*. Through expression studies, the function of some of these hydrolases involved in the turnover of storage polymers and in cell growth may be inferred (*18*, *21*). Classical means to purify and identify these enzymes relied on biochemical schemes that were difficult at best and, in many instances, impossible to accomplish. They demonstrate how bioinformatics and functional genomics can provide a powerful means to identify and evaluate candidate genes through database searches and "expression profiling" by microarray analyses. Cellulose synthase is arguably the most important enzyme involved in plant cell wall biosynthesis (*15*, *17*, *18*, *21*, *23*). Richmond and Somerville (*21*) discuss the enormity of the cel-

lulose synthase superfamily of *Arabidopsis* and how a powerful multidisciplinary approach can be used to determine gene function within this large superfamily. The genes that are at the core of cell wall biogenesis are those that encode polysaccharide synthases and glycosyl transferases (*15*). Synthases are defined as processive glycosyltransferases that iterate linkage of mono- or disaccharide units into the backbone polymer, whereas glycosyltransferases decorate the backbone with addition of specific sugars (*15*). An enormous task lies ahead to define the function of all the candidate genes that comprise this stage of wall biogenesis.

The secondary cell walls provide excellent examples of how cell wall modification confers specific properties upon a cell to allow it to fulfill specialized functions. Secondary cell walls are frequently a feature of cells that provide support for the plant body, and cells involved in the transport of water and solutes from the roots to the aerial tissues. Secondary cell walls allow these cells to resist the forces of gravity and/or the tensional forces associated with the transpirational pull on a column of water. Turner *et al* summarize how a clever mutant screen was used to define genes specifically involved in cellulose synthesis and lignification during secondary cell wall formation (*15*, *17*, *18*, *21*, *23*). As wood is essentially a collection of secondary cell walls, many cell-wall-relevant genes have also emerged from genomics research associated with wood formation. One of the few model systems to study the precise development of a single cell type *in vitro* is that of the transdifferentiation of *Zinnia* mesophyll cells into tracheary elements. The six stages of wall development might reasonably be used to classify the fundamental structural elements of the wall, but they are far from a comprehensive set of genes whose products function in the plant extracellular matrix. They undoubtedly represent the tip of the iceberg with respect to understanding how and what messages plant cells communicate (*6*, *15*, *17*, *18*, *21*, *23*).

Molecular Modelling

With the advent of genomics and biotechnology, biological researchers are investigating particular enzymes involved in cell growth, development, defense, and wall metabolism in the hope of producing crops with desired characteristics by enhancing commercially valuable traits, such as fiber production in flax, cotton, ramie and sisal, or abolishing costly ones, such as lignification in some plant tissues. At more theoretical level, potential substrate reactivities can be predicted by molecular modelling catalytic sites for individual genes and by defining the dimensions of substrate binding pockets. To date, this approach has been widely used in analyzing human P450s involved in drug metabolisms (*93–95*) and insect P450s involved in the metabolism of plant toxins (*29*). Support for individual models is derived from site-directed mutagenesis of key residues in proposed catalytic sites and analysis of alternate substrates. These same approaches are now used to define the catalytic sites of plant P450s with known functions. At present, molecular models have been constructed in the peppermint CYP71D13 protein mediating hydroxylation of limonene (*93*, *96*), the artichoke CYP73A1 (*93*, *94*) and *Arabidopsis* CYP73A5 (*93*, *94*) proteins mediating hydroxylation of *t*-cinnamic acid, the *Arabidopsis* CYP84A1 protein mediating hydroxylation of coniferaldehyde, coniferyl alcohol, and ferulate, the *Arabidopsis* CYP98A3 protein mediating hydroxylation of *p*-coumaryl shikimic and quinic acids, the *Arabidopsis* CYP75B1 protein mediating hydroxylation of narigenin and dihydrokaempferol, the licorice CYP93C2 protein catalyzing aryl migration in the formation of isoflavonoids, and the *Vicia* CYP94A2 protein mediating the hydroxylation of fatty acids (*93*). This is an excellent example of how science designed to cope with the problems associated with gene and protein function analysis is likely to be of benefit to all plant scientists.

Conclusion

EST sequencing certainly avoids the biggest problems associated with genome size and the accompanying retrotransposon repetitiveness. It provides us with potential knowledge bases to fill in knowledge gaps from the gene complement of the large plant genomes. The EST sequence resources have been proven to have a wide range of applications and novel uses in biology. However, there is no real substitute for a complete genome sequence. Only when presented with the completed chromosomes, can we dissect the gene complement and unravel the mechanistic pathways that make the plant. Until new technologies become generally available that can produce longer sequence reads more cheaply, we will be limited to incomplete solutions. The completion of the *Arabidopsis* genome

sequence culminates the first century of genetics research since the rediscovery of Mendel's experiments. We have a complete inventory of the genes sufficient to make a higher plant. The *Arabidopsis* genome has become a springboard for comparative genetics with the genomes of other plant species, including our important crop plants and trees. Although *Arabidopsis* has proven itself to be a superior model plant for genetic studies, many other species are far more suitable for cellular and biochemical studies that will unveil gene function. We estimate that about 15% of the genome is connected in some way with the biogenesis, rearrangement, and turnover of a cell wall. But only about 1,000 genes have been assigned a function by direct experimental evidence (*21*). The era of functional genomics has come to the plant biology and forest sector with several EST sequencing projects being initiated in a range of forest trees. Even more exciting is the complete sequencing of the genome of the model tree poplar. Elucidation of gene function in forest trees will be accelerated with functional genomics. Global transcript profiling is already being used in poplar. The integration of information obtained from the use of different profiling technologies will allow scientists to rapidly assess novel gene function. Functional genomics has been greatly invested (Table 6) and will be important in identifying candidate genes whose function is now being investigated and they are critical to all aspects of plant growth, development, differentiation, and defense. In the coming years, both academic research and the forest biotechnology industry are set to benefit from the advent of functional genomics of wood quality and properties.

Table 6 Functional Genomics Projects Awarded by NSF in USA in 1999 and 2002 (http://www.nsf.gov)

Year	Title	Total Award (USD$)
1999-2002	Tools for Potato Structural and Functional Genomics	$5,300,000
1999-2004	Functional Genomics of Maize Centromeres	$2,510,000
1999-2002	Functional and Comparative Genomics of Disease Resistance Gene Homologs	$2,530,000
1999-2002	Functional Genomics of Hemicellulose Biosynthesis	$2,250,000
1999-2002	Genomics of Wood Formation in Loblolly Pine	$4,450,000
1999-2002	Functional Genomics of Plant Phosphorylation	$2,983,734
2002-2007	Potato Functional Genomics: Application to Analysis of Growth, Development, Metabolism and Responses to Biotic and Abiotic Stress	$7,618,912
2002-2005	Genomics of Loblolly Pine Embryogenesis	$1,380,910
2002-2006	Functional Genomics of Host-Virus Interactions	$3,363,177
2002-2006	Functional Genomics of Phytophthora-Plant Interactions	$1,891,617
2002-2007	Functional Genomics of Hemicellulose Biosynthesis	$4,945,077
2002-2005	Transcriptome Responses to Environmental Conditions in Loblolly Pine Roots	$1,651,752
2002-2005	Functional genomic analysis of fruit flavor and nutrition pathways	$1,159,280
2002-2006	Functional Analyses of Plant Gamete Gene Expression	$1,135,486
2002-2007	Functional genomics of root growth and root signaling under drought	$4,549,050

References

1. Adams, M.D., *et al.* 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252: 1651-1656.
2. Mayer, K. and Mewes, H.W. 2002. How can we deliver the large plant genomes? Strategies and perspectives. *Curr. Opin. Plant Biol.* 5: 173-177.
3. Rudd, S. 2003. Expressed sequence tags: alternative or complement to whole genome sequences? *Trends Plant Sci.* 8: 321-329.
4. Ewing, R.M., *et al.* 1999. Large-scale statistical analyses of rice ESTs reveal correlated patterns of gene expression. *Genome Res.* 9: 950-959.
5. Bhalerao, R., *et al.* 2003. Out of the woods: forest biotechnology enters the genomic era. *Curr. Opin. Biotech.* 14: 206-213.
6. Wullschleger, S.D., *et al.* 2002. Genomics and forest biology: *Populus* emerges as the perennial favorite. *Plant Cell* 14: 2651-1655.
7. Halpin, C. and Boerjan, W. 2003. Stacking transgenes in forest trees. *Trends Plant Sci.* 8: 363-365.
8. Tang, W. and Newton, R.J. 2003. Genetic transfor-

mation of conifers and its application in forest biotechnology. *Plant Cell Rep.* 22: 1-15.
9. Pilate, G., *et al.* 2002. Field and pulping performances of transgenic trees with altered lignification. *Nat. Biotechnol.* 20: 607-612.
10. Pena, L. and Seguin, A. 2001. Recent advances in the genetic transformation of trees. *Trends Biotechnol.* 19: 500-506.
11. Bradshaw, H.D., *et al.* 2000. Emerging model systems in plant biology: Poplar (*Populus*) as a model forest tree. *J. Plant Growth Regul.* 19: 306-313.
12. Taylor, G. 2002. *Populus*: *Arabidopsis* for forestry. Do we need a model tree? *Ann. Bot.* 90: 677-687.
13. Hertzberg, M., *et al.* 2001. A transcriptional roadmap to wood formation. *Proc. Natl. Acad. Sci. USA.* 98: 14732-14737.
14. Sterky, F., *et al.* 1998. Gene discovery in the wood forming tissues of poplar: analysis of 5,692 expressed sequence tags. *Proc. Natl. Acad. Sci. USA.* 95: 13330-13335.
15. Carpita, N., *et al.* 2001. Molecular biology of the plant cell wall: searching for the genes that define structure, architecture and dynamics. *Plant Mol. Biol.* 47: 1-5.
16. Xue, Y., *et al.* 2003. Recent highlights of the China Rice Functional Genomics Program. *Trends Genet.* 19: 390-394.
17. *Arabidopsis* Genome Project. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 823-826.
18. *Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408: 796-815.
19. Strauss, S.H. 2003. Genomics, genetic engineering, and domestication of crops. *Science* 300: 61-62.
20. Fenning, T.M. and Gershenzon, J. 2002. Where will the wood come from? Plantation forests and the role of biotechnology. *Trends Biotechnol.* 20: 291-296.
21. Somerville, C. and Dangl, J. 2000. Plant biology in 2010. *Science* 290: 2077-2078.
22. Goff, S.A., *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296: 92-100.
23. Chory, J., *et al.* 2000. Functional genomics and the virtual plant. A blueprint for understanding how plants are built and how to improve them. *Plant Physiol.* 123: 423-425.
24. Fiehn, O. 2002. Metabolomics: the link between genotypes and phenotypes. *Plant Mol. Biol.* 48: 155-171.
25. Sheppard, L.A., *et al.* 2000. A DEFICIENS homolog from the dioecious tree black cottonwood is expressed in both female and male floral meristems of the two-whorled, unisexual flowers. *Plant Physiol.* 124: 627-639.
26. Raizada, M.N., *et al.* 2001. Somatic and germinal mobility of the RescueMu transposon in transgenic maize. *Plant Cell* 13: 1587-1608.
27. Rabinowicz, P.D., *et al.* 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* 23: 305-308.
28. Adams, M.D., *et al.* 1993. 3,400 new expressed sequence tags identify diversity of transcripts in human brain. *Nat. Genet.* 3: 256-257.
29. Allona, I., *et al.* 1998. Analysis of xylemformation in pine by cDNA sequencing. *Proc. Natl. Acad. Sci. USA.* 95: 9693-9698.
30. Israelsson, M., *et al.* 2003. Changes in gene expression in the wood-forming tissue of transgenic hybrid aspen with increased secondary growth. *Plant Mol. Biol.* 52: 893-903.
31. Hertzberg, M., *et al.* 2001. cDNA microarray analysis of small plant tissue samples using a cDNA tag target amplification protocol. *Plant J.* 25: 585-591.
32. Daly, D.C., *et al.* 2001. Plant systematics in the age of genomics. *Plant Physiol.* 127: 1328-1333.
33. Pryer, K.M., *et al.* 2002. Deciding among green plants for whole genome studies. *Trends Plant Sci.* 7: 550-554.
34. Herwig, R., *et al.* 2002. Construction of a "unigene" cDNA clone set by oligonucleotide fingerprinting allows access to 25,000 potential sugar beet genes. *Plant J.* 32: 845-857.
35. Bonaldo, M.F., *et al.* 1996. Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res.* 6: 791-806.
36. MacIntosh, G.C., *et al.* 2001. Identification and analysis of *Arabidopsis* expressed sequence tags characteristic of non-coding RNAs. *Plant Physiol.* 127: 765-776.
37. Chandler, V.L. and Brendel, V. 2002. The maize genome sequencing project. *Plant Physiol.* 130: 1594-1597.
38. Haas, B.J., *et al.* 2002. Full-length messenger RNA sequences greatly improve genome annotation. *Genome Biol.* 3: Research 0029.1-0029.12.
39. Leitch, I.J. 2001. Nuclear DNA C-values complete familial representation in Gymnosperms. *Ann. Botany* 88: 843-849.
40. Gress, T.M., *et al.* 1992. Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mamm. Genome* 3: 609-619.
41. van der Hoeven, R., *et al.* 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. *Plant Cell* 14: 1441-1456.
42. Obermayer, R., *et al.* 2002. Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. *Ann. Bot.* 90: 209-217.

43. Gaut, B.S., *et al.* 2000. Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci. USA.* 97: 7008-7015.
44. Heslop-Harrison, J.S. 2000. Comparative genome organization in plants: from sequence and markers to chromatin and chromosomes. *Plant Cell* 12: 617-636.
45. Bennetzen, J.L. 2002. Mechanisms and rates of genome expansion and contraction in flowering plants. *Genetica* 115: 29-36.
46. Carels, N., *et al.* 1995. The gene distribution of the maize genome. *Proc. Natl. Acad. Sci. USA.* 92: 11057-11060.
47. Barakat, A., *et al.* 1997. The distribution of genes in the genomes of Gramineae. *Proc. Natl. Acad. Sci. USA.* 94: 6857-6861.
48. Yu, J., *et al.* 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296: 79-92.
49. Stormo, G.D., 2000. Gene-finding approaches for eukaryotes. *Genome Res.* 10: 394-397.
50. Mathe, C., *et al.* 2002. Current methods of gene prediction, their strengths and weaknesses. *Nucleic Acids Res.* 30: 4103-4117.
51. Reese, M.G., *et al.* 2000. Genome annotation assessment in *Drosophila melanogaster*. Genome Res. 10: 483-501.
52. Brendel, V. and Zhu, W. 2002. Computational modeling of gene structure in *Arabidopsis thaliana*. *Plant Mol. Biol.* 48: 49-58.
53. Zhu, W., *et al.* 2003. Refined annotation of the *Arabidopsis thaliana* genome by complete EST mapping. *Plant Physiol.* 132: 469-484.
54. Rudd, S., *et al.* 2003. Sputnik: a database platform for comparative plant genomics. *Nucleic Acids Res.* 31: 128-132.
55. Schoof, H. and Karlowski, W. 2003. Comparison of rice and *Arabidopsis* annotation. *Curr. Opin. Plant Biol.* 6: 1-7.
56. Burke, J., *et al.* 1998. Alternative gene form discovery and candidate gene selection from gene indexing projects. *Genome Res.* 8: 276-290.
57. Hillier, L.D., *et al.* 1996. Generation and analysis of 280,000 human expressed sequence tags. *Genome Res.* 6: 807-828.
58. Curry, J. and Glickman, B.W. 1997. Moloney murine leukemia reverse transcriptase suspect in the production of multiple misincorporations during hprt cDNA synthesis. *Mutat. Res.* 374: 145-148.
59. Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* 8: 186-194.
60. Seki, M., *et al.* 1998. High-efficiency cloning of *Arabidopsis* full-length cDNA by biotinylated CAP trapper. *Plant J.* 15: 707-720.
61. Stoesser, G., *et al.* 2003. The EMBL nucleotide sequence database: major new developments. *Nucleic Acids Res.* 31: 17-22.
62. Heumann, K. and Mewes, H.W. 1996. The hashed position tree (HPT): a suffix tree variant for large data sets stored on slow mass storage devices. In *Proceedings of the Third South American Workshop on String Processing* (eds. Ziviani, N., *et al.*), pp.101-115, Carlton University Press, Ottawa, Canada.
63. Gordon, D., *et al.* 1998. Consed: a graphical tool for sequence finishing. *Genome Res.* 8: 195-202.
64. Ronning, C.M., *et al.* 2003. Comparative analyses of potato expressed sequence tag libraries. *Plant Physiol.* 131: 419-429.
65. Fulton, T.M., *et al.* 2002. Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. *Plant Cell* 14: 1457-1467.
66. Shrager, J., *et al.* 2003. *Chlamydomonas reinhardtii* genome project. A guide to the generation and use of the cDNA information. *Plant Physiol.* 131: 401-408.
67. Cervera, M.T., *et al.* 2001. Dense genetic linkage maps of three *Populus* species (*Populus deltoides*, *P. nigra* and *P. trichocarpa*) based on AFLP and microsatellite markers. *Genetics* 158: 787-809.
68. Wu, R., *et al.* 2001. Mapping epigenetic quantitative trait loci (QTL) altering a developmental trajectory. *Genome* 1: 28-33.
69. Hu, W.J., *et al.* 1999. Repression of lignin biosynthesis promotes cellulose accumulation and growth in transgenic trees. *Nat. Biotechnol.* 17: 808-812.
70. Li, L., *et al.* 2001. The last step of syringyl monolignol biosynthesis in angiosperms is regulated by a novel gene encoding sinapyl alcohol dehydrogenase. *Plant Cell* 7: 1567-1586.
71. Moyle, R., *et al.* 2002. Environmental and auxin regulation of wood formation involves members of the Aux/IAA gene family in hybrid aspen. *Plant J.* 6: 675-685.
72. Mellerowicz, E.J., *et al.* 2001. Unravelling cell wall formation in the woody dicot stem. *Plant Mol. Biol.* 47: 239-274.
73. Weigel, D. and Nilsson, O. 1995. A developmental switch sufficient for flower initiation in diverse plants. *Nature* 377: 495-500.
74. Pena, L., *et al.* 2001. Constitutive expression of *Arabidopsis* LEAFY or APETALA1 genes in citrus reduces their generation time. *Nat. Biotechnol.* 19: 263-267.
75. Kyozuka, J., *et al.* 1997. Eucalyptus has functional equivalents of the *Arabidopsis* AP1 gene. *Plant Mol. Biol.* 35: 573-584.
76. Sung, S.K., *et al.* 1999. Characterization of MdMADS2, a member of the SQUAMOSA subfamily of genes, in apple. *Plant Physiol.* 120: 969-978.
77. Elo,A., *et al.* 2001. Three MADS-box genes similar to APETALA1 and FRUITFULL from silver birch (*Be-*

tula pendula). *Physiol. Plant* 112: 95-103.
78. Campbell, M.M., *et al.* 2003. Forestry's fertile crescent: the application of biotechnology to forest trees. *Plant Biotechnol.* 1: 141-154.
79. Eriksson, M.E. 2000. Increased gibberellin biosynthesis in transgenic trees promotes growth, biomass production and xylem fiber length. *Nat. Biotechnol.* 18: 784-788.
80. Franke, R., *et al.* 2000. Modified lignin in tobacco and poplar plants over-expressing the *Arabidopsis* gene encoding ferulate 5-hydroxylase. *Plant J.* 22: 223-234.
81. Boerjan, W., *et al.* 2003. Lignin biosynthesis. *Annu. Rev. Plant Biol.* 54: 519-546.
82. Li, L., *et al.* 2003. Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *Proc. Natl. Acad. Sci. USA*. 100: 4939-4944.
83. McCormac, A.C., *et al.* 2001. Efficient cotransformation of *Nicotiana tabacum* by two independent T-DNAs, the effect of T-DNA size and implications for genetic separation. *Transgenic Res.* 10: 143-155.
84. Ye, X., *et al.* 2000. Engineering the provitamin A (β-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. *Science* 287: 303-305.
85. Slater, S., *et al.* 1999. Metabolic engineering of *Arabidopsis* and *Brassica* for poly (3-hydroxybutyrate-co-3-hydroxyvalerate) copolymer production. Nat. Biotechnol. 17: 1011-1016.
86. Halpin, C., *et al.* 2001. Enabling technologies for manipulating multiple genes on complex pathways. *Plant Mol. Biol.* 47: 295-310.
87. de Neve, M., *et al.* 1997. T-DNA integration patterns in cotransformed plant cells suggest that T-DNA repeats originate from co-integration of separate T-DNAs. *Plant J.* 11: 15-29.
88. Komari, T., *et al.* 1996. Vectors carrying two separate T-DNAs for cotransformation of higher plants mediated by *Agrobacterium tumefaciens* and segregation of transformants free from selection markers. *Plant J.* 10: 165-174.
89. Muskens, M.W.M., *et al.* 2000. Role of inverted DNA repeats in transcriptional and post-transcriptional gene silencing. *Plant Mol. Biol.* 43: 243-260.
90. Meilan, R., *et al.* 2002. The CP4 transgene provides high levels of tolerance to Roundupw herbicide in field-grown hybrid poplars. *Can. J. For. Res.* 32: 967-976.
91. Rugh, C.L., *et al.* 1998. Development of transgenic yellow poplar for mercury phytoremediation. *Nat. Biotechnol.* 16: 925-928.
92. Gallardo, F., *et al.* 1999. Expression of a conifer glutamine synthetase gene in transgenic poplar. *Planta* 210: 19-26.
93. Schuler, M.A. and Werck-Reichhart, D. 2003. Functional genomics of P450s. *Annu. Rev. Plant Biol.* 54: 629-667.
94. Quackenbush, J., *et al.* 2001. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nucleic Acids Res.* 29: 159-164.
95. Lamblin, A.F., *et al.* 2003. MtDB: a database for personalized data mining of the model legume *Medicago truncatula* transcriptome. *Nucleic Acids Res.* 31: 196-201.
96. Wheeler, D.L., *et al.* 2003. Database resources of the National Center for Biotechnology. *Nucleic Acids Res.* 31: 28-33.