Genomics is the study of the structure, location, function, regulation and interaction of large numbers of genes simultaneously. Genomics is an extension of some of the traditional sub-disciplines of genetics, such as transmission and molecular genetics, to the entire genome. Genomics is a science that is made possible and is dependent upon technologies that allow rapid analysis of hundreds or thousands of genes and often in many individuals at once, such as by automated DNA sequencing. Whereas traditional molecular genetics enabled the discovery and functional analysis of one or a small number of genes in single experiments, genomics not only seeks to understand the function of large numbers of genes in parallel, but also seeks to understand the complex interactions among genes. Genomic sciences should ultimately lead to a deeper understanding of the relationship between an organism’s total genetic composition and its complete phenotype. This will also mean a mechanistic understanding of gene action, pleiotropy, GxE interaction and epistasis.

Three major sub-areas of genomics are generally recognized: structural genomics, functional genomics and comparative genomics. Structural genomics seeks to identify all the genes in the genome and determine their locations on chromosomes. This is usually achieved by DNA sequencing of individual genes or of entire genomes and by genetic or physical mapping. Functional genomics seeks to determine the function of genes and how genes determine the phenotype. A variety of new techniques and experimental approaches have been developed to help understand how genes and genomes function. Comparative genomics seeks to understand either the structure or function of genes by making comparisons across taxa. DNA sequencing and genetic mapping techniques are most often used for comparative genomic analyses.

Genomic science technologies have largely been developed through the Human Genome Project. The pharmaceutical and biomedical industries have made significant investments in genomic science research for the discovery of new drugs and treatment of diseases. These same technologies are now being applied to agricultural plants, animals and forest trees. The large investments being made by private corporations lead to patenting of genomic information and control of intellectual property. So although genomic sciences promise to achieve a deeper understanding of genomes and the function of genes, this knowledge may not be free for all to use; the products developed through genomics may be owned by individual corporations.

In this chapter, we discuss how structural, functional and comparative genomic sciences are being applied to forest trees. In Chapters 19 and 20, we illustrate how knowledge of the structure and function of genes and genomes can be applied to develop varieties of trees for specific end-product needs. Other recent reviews of genomic sciences in forestry include Sederoff (1999), Krutovskii and Neale (2003) and Kumar and Fladung (2004).
**STRUCTURAL GENOMICS**

The genomes of forest trees, like all eukaryotes, are organized into chromosomes, each of which contains a large number of genes (Chapter 2). The total number of genes in the genome is not yet known for any tree species. Estimates from a few animal and plant species have been made as a result of complete genome sequencing (Table 18.1), and from these data it is expected that tree species have in the range of 30,000 genes. The central task of structural genomics is to build a catalog of all genes in the genome, sometimes called gene discovery, and determine their locations on chromosomes by genetic and physical mapping.

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Saccharomyces cerevisiae</em> (yeast)</td>
<td>6034</td>
</tr>
<tr>
<td><em>Caenorhabditis elegans</em> (nematode)</td>
<td>19,099</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em> (fruit fly)</td>
<td>13,061</td>
</tr>
<tr>
<td><em>Mus musculus</em> (mouse)</td>
<td>~30,000</td>
</tr>
<tr>
<td><em>Homo sapiens</em> (human)</td>
<td>~40,000</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em> (small flowering plant)</td>
<td>25,498</td>
</tr>
<tr>
<td><em>Oryza sativa</em> (rice)</td>
<td>59,855</td>
</tr>
</tbody>
</table>

**Gene Discovery**

Several different approaches are used in gene discovery. The most direct approach is to determine the DNA sequence of the entire genome, which includes DNA sequences for both gene-encoding and non-encoding portions of the genome. The individual genes can be identified from the total DNA sequence using sophisticated computational algorithms that take into account many factors related to the known structure of genes such as open reading frames, intron-exon splice sites, initiation and termination codons and much more (Chapter 2). This approach has been used for a variety of organisms, many of which have small genomes that make it experimentally and economically feasible, and it has also been used for the human genome because of its importance (IHGSC, 2001; Venter et al., 2001). The complete genome sequence of two model plant species, *Arabidopsis thaliana* (TAGI, 2000) and rice (Goff et al., 2002; Yu et al., 2002), were the first to be determined and several more are in progress. Full genome sequencing of the *Populus* genome was started in 2002 and a first draft sequence was released in Sept. 2004.

A complementary approach to complete genome sequencing, called **expressed sequence tags (ESTs)**, has been applied to scores of organisms, including several forest tree species. EST sequencing is based on identifying only the DNA sequences that code for genes that are expressed. The basic approach of EST sequencing is to construct cDNA libraries (Chapter 2) for the many genes expressed in one or more tissue-types and determine the DNA sequences of short segments of a large number of cDNA clones from the libraries (Figure 18.1). The DNA sequences are most often obtained from the 5’ end of the cDNA because there is generally greater sequence conservation at this end of genes. This raw sequence information is submitted to gene databases and compared to all other sequences in the database to identify matches (Figure 18.2). If there are matches to gene sequences whose functions have already been determined, then the likely identity of these ESTs will have been made without further investigation.
Fig. 18.1. Schematic of the important steps in the development of expressed sequence tags (ESTs). a) Tissue is harvested from the tree, such as xylem in this example; b) messenger RNA (mRNA) (see Figure 2.9) is isolated from the xylem tissue; c) mRNA is converted to complementary DNA (cDNA) (see Box 4.2); d) DNA sequences are determined from cDNA clones using automated instrumentation.

EST sequencing was first applied to identify genes expressed in the human brain (Adams et al., 1991) and has since been applied to scores of plant species. Some of the first EST sequencing projects in forest trees focused on genes expressed in wood forming tissues (Allona et al., 1998; Sterky et al., 1998; Whetten et al., 2001; Kirst et al., 2003). These projects discovered many genes involved in lignin and cellulose biosynthesis and other components of cell walls. Genes for basic metabolism in woody plant cells have also been identified (Figure 18.3). EST sequencing projects are in progress for many forest tree species and for a wide variety of tissue types (e.g. roots, leaves/needles, stems, meristems and reproductive structures) (Table 18.2). As of 2005, more than a million sequences from forest tree species were found in Genbank (http://www.ncbi.nlm.nih.gov), and this number is growing rapidly.
Figure 18.2. A report from the DNA sequence database Genbank for the cinnamyl alcohol dehydrogenase (cad) gene sequence from *Pinus taeda*. a) Sequence annotation showing accession number, gene name, species, authors and journal reference; b) Nucleotide sequence of the cad gene from *P. taeda*; c) The cad nucleotide sequence from *P. taeda* was compared to all sequences in the database and aligned with the most similar sequence found. In this case, a nucleotide sequence from *Arabidopsis thaliana* (Sbjct) was completely identical to a portion of the *Pinus taeda* sequence (Query) except for two nucleotide differences (indicated by upward and downward arrows at *Pinus taeda* positions 328 and 659, respectively).
Fig. 18.3. Functional classification of a) 4809 ESTs from *Populus* cambium and b) 833 ESTs from *Populus* xylem. Note the large proportion of ESTs that could not be assigned to a functional class. “Unknown” indicates a match to a gene of unknown function in the database and “No hits” indicates that no match was found. (from Sterky et al. 1998)

Genetic Mapping

Genetic mapping is central to genomic sciences because it provides an organizational framework to the genome. In Chapter 4, the many types of genetic markers available for use in forest trees are described. The value of genetic markers is enhanced if their location in the genome is known, i.e. their map position on a specific chromosome. Genetic markers can be positioned onto two kinds of maps: physical maps and genetic maps. Physical maps provide the exact location of genes or genetic markers on chromosomes. Techniques such as radio-labeled or fluorescent in situ hybridization are used for physical mapping (Figure 2.7). Ribosomal RNA genes and a few highly repeated DNA sequences have been physically mapped using such techniques (Brown et al., 1993; Brown and Carlson, 1997; Doudrick et al., 1995). Current technologies, however, are not sensitive enough with the large genomes of conifers to map molecular markers that are based on relatively small segments of DNA, such as cDNAs. Large fragments (~100 kb) of cloned DNA, such as bacterial artificial chromosomes (BACs) (see section on Positional Cloning of QTLs for a description of BACs), would be ideal for physical mapping in conifers.
Table 18.2. The number of ESTs found in the Genbank database for a sample of tree species.

<table>
<thead>
<tr>
<th>Species</th>
<th>ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Pinus taeda</em></td>
<td>173,680</td>
</tr>
<tr>
<td><em>Populus tremula x Populus tremuloides</em></td>
<td>65,981</td>
</tr>
<tr>
<td><em>Picea glauca</em></td>
<td>55,108</td>
</tr>
<tr>
<td><em>Populus balsamifera subsp. trichocarpa</em></td>
<td>54,660</td>
</tr>
<tr>
<td><em>Populus balsamifera subsp. trichocarpa x Populus deltoides</em></td>
<td>33,134</td>
</tr>
<tr>
<td><em>Populus tremula</em></td>
<td>31,288</td>
</tr>
<tr>
<td><em>Pinus pinaster</em></td>
<td>18,254</td>
</tr>
<tr>
<td><em>Populus deltoides</em></td>
<td>14,645</td>
</tr>
<tr>
<td><em>Populus balsamifera subsp. trichocarpa x Populus nigra</em></td>
<td>14,281</td>
</tr>
<tr>
<td><em>Populus euphratica</em></td>
<td>13,903</td>
</tr>
<tr>
<td><em>Populus tremuloides</em></td>
<td>12,813</td>
</tr>
<tr>
<td><em>Picea engelmannii x Picea sitchensis</em></td>
<td>12,127</td>
</tr>
<tr>
<td><em>Picea sitchensis</em></td>
<td>12,065</td>
</tr>
<tr>
<td><em>Populus x canescens</em></td>
<td>10,446</td>
</tr>
<tr>
<td><em>Populus euramericana</em></td>
<td>10,157</td>
</tr>
<tr>
<td><em>Pseudotsuga menziesii var. menziesii</em></td>
<td>6,721</td>
</tr>
<tr>
<td><em>Cryptomeria japonica</em></td>
<td>6,589</td>
</tr>
<tr>
<td><em>Cycas rumphii</em></td>
<td>5,952</td>
</tr>
<tr>
<td><em>Juglans regia</em></td>
<td>5,025</td>
</tr>
<tr>
<td><em>Tamarix androssowii</em></td>
<td>4,756</td>
</tr>
<tr>
<td><em>Welwitschia mirabilis</em></td>
<td>3,732</td>
</tr>
<tr>
<td><em>Betula pendula</em></td>
<td>2,545</td>
</tr>
<tr>
<td><em>Gnetum gnemon</em></td>
<td>2,128</td>
</tr>
<tr>
<td><em>Camellia sinensis</em></td>
<td>1,989</td>
</tr>
<tr>
<td><em>Ginkgo biloba</em></td>
<td>1,953</td>
</tr>
<tr>
<td><em>Pinus sylvestris</em></td>
<td>1,663</td>
</tr>
<tr>
<td><em>Eucalyptus grandis</em></td>
<td>1,574</td>
</tr>
<tr>
<td><em>Eucalyptus tereticornis</em></td>
<td>1,131</td>
</tr>
<tr>
<td><em>Populus tomentiglandulosa</em></td>
<td>1,127</td>
</tr>
</tbody>
</table>

The alternative approach is to develop **genetic maps** by segregation and linkage analysis (Chapter 3). Genetic maps identify the relative distance between two markers based on the number of recombination events between the markers. That is, physical maps show the locations of genes in base-pair distance units from each other while genetic maps show the relative locations of genes in terms of recombination units. The relationship between a physical map and a genetic map is not direct, because the amount of recombination between any two equidistant markers can vary significantly throughout the genome. Genetic linkage mapping is discussed in this chapter because this has been the most widely employed mapping technique in forest trees.

There are three major steps to constructing a genetic map: (1) Selecting an appropriate mapping population; (2) Obtaining genetic marker segregation data from the mapping population; and (3) Applying analytical methods for linkage analysis to place the segregating loci in relative positions and distances from one another. Genetic markers have already been discussed (Chapter 4); so, in the next section only mapping populations and linkage analysis are discussed. Genetic maps have been constructed for a large number of forest tree species and the types of genetic markers, mapping populations, and methods of linkage analysis all vary among mapping projects (Cervera et al., 2000; http://dendrome.ucdavis.edu/).
Genetic Mapping Populations

Genetic maps are based on segregation and linkage analysis performed on some type of segregating or mapping population. In Chapter 3, the concept of genetic linkage was illustrated using a simple example of two linked allozyme loci in *Pinus rigida* (Box 3.4). The mapping population used in that example was a sample of segregating haploid megagametophytes in open-pollinated seed from a single seed tree. The two general types of mapping populations most commonly used in forest trees are haploid conifer megagametophytes and full-sib pedigrees.

Megagametophyte mapping populations were first used with allozyme markers to estimate linkages among allozyme loci (Chapter 3, Guries et al., 1978; Rudin and Ekberg, 1978; Conkle, 1981; Adams and Joly, 1980). In the last 20 years, megagametophyte mapping populations have been used to estimate linkages among genes coding allozymes in dozens of conifer species. More recently, megagametophyte mapping populations have been employed to estimate linkages and construct genetic maps based on RAPD and AFLP markers (Chapter 4) (e.g. Tulsieram et al., 1992; Nelson et al., 1993; Binelli and Bucci, 1994; Remington et al., 1999). Clearly, the megagametophyte mapping population system will continue to be used for genetic mapping of molecular markers in conifers.

Segregating populations can also be generated by controlled crossing. These populations are essential for linkage mapping in angiosperms (which lack haploid megagametophytes) and also can serve as an alternative to megagametophyte mapping in conifers. When inbreeding depression is not a concern, such as in many crop plants that are naturally predominately self-pollinated, ideal mapping populations can be generated by first creating highly inbred lines that are homozygous at most, or all, marker loci. By crossing two different inbred lines, F1 progeny are produced that are heterozygous at many loci. Backcrossing one of the F1 individuals to one of the parental inbred lines produces a segregating population for linkage mapping. Backcrosses are ideal, because all recombination occurs in the gametes of one parent (the heterozygous F1), and the linkage phase of all doubly-heterozygous combinations of loci is known (i.e. coupling ABab or repulsion AbaB). The segregation data obtained with a backcross mapping population is similar to that derived from segregating megagametophytes, except that linkage phase is not known with megagametophyte data.

In outcrossing plants, like most trees, inbred lines cannot be produced because the close inbreeding required to create these lines results in weak, infertile individuals (Chapter 5). Any two potential parent trees chosen from a natural (not inbred) population are likely to be heterozygous at many loci (Chapter 7) and if crossed, the progeny will segregate at these loci. Linkage analysis, however, is complicated because the linkage phase of pairs of markers is not known and mating types (Table 18.3) of markers can vary. Different mating types are possible because for any marker, just one or both of the parent trees can be heterozygous. This makes the estimation of linkage more complicated than in the simple inbred backcross case, because differences in mating types produce different two-locus ratios in the progeny. Still, these types of mapping populations involving crosses among heterozygous parents are widely used in forest trees out of necessity.

Linkage analysis and map construction

Once marker segregation data are obtained from a mapping population, it is possible to estimate linkages among the markers and construct a genetic map. Estimation of recombination fractions, or linkage, between a pair of markers was described in Chapter 3. If the Table 3. Informative full-sib mating types for genetic mapping.
Maternal genotype | Paternal genotype | Mating type
--- | --- | ---
A₁/A₂ | A₁/A₁ | Maternally informative
A₁/A₁ | A₁/A₂ | Paternally informative
A₁/A₂ | A₁/A₂ | Intercross
A₁/A₂ | A₁/A₃ | Fully informative
A₁/A₂ | A₃/A₄ | Fully informative

number of markers analyzed is very small (<10), as is the case with some allozyme marker data sets, it is possible to estimate all pairwise linkages and order markers manually. Some loci may appear to be closely linked (r is small), or located very far apart on the same chromosome or on different chromosomes (r approaches 0.5). However, as the number of markers increases, it becomes tedious to estimate all linkages manually and extremely difficult to manually determine marker order on linkage groups. Rather, specialized genetic mapping software is used; two commonly used programs are Mapmaker (Lander et al., 1987) and JoinMap (Stam, 1993).

All genetic mapping software follows a similar approach for constructing maps. The first step is to group all the genetic markers into linkage groups. This is done by first calculating all the pairwise linkage distances (r values) among all markers in the data set. The programs use this matrix of linkage estimates to determine which sets of markers are likely linked to one another and should be assigned to a common linkage group. Once the linkage groups are determined, markers are then ordered within each group. The statistical approach to ordering varies among programs and the extent to which the user can control the ordering analysis also varies. While in concept a linkage group is equivalent to a chromosome, the number of linkage groups often exceeds the number of chromosomes due to inadequate marker coverage or large regions with suppressed recombination (e.g. inversions) that prevent detection of linkage between regions on the same chromosome.

The Mapmaker software is most often used for megametophyte mapping population data, because this data type is very similar to the backcross and F₂ data types for which the software was designed. Genetic mapping in full-sib pedigree mapping populations in trees is significantly more complicated. The marker genotypes of the progeny result from independent segregations in both the maternal and paternal parent. Given this situation, two mapping strategies can be employed: (1) Construct individual genetic maps for each of the parents; or (2) Construct a single “sex-averaged” map using data from both parents. Groover et al. (1994) used the JoinMap program to create individual parent tree maps in Pinus taeda. Sex-averaged maps have been constructed for Pinus radiata (Devey et al., 1996), Eucalyptus nitens (Byrne et al., 1995) and Pseudotsuga menziesii (Jermstad et al., 1998) using JoinMap. Sewell et al. (1999) used Mapmaker and JoinMap to create individual parent tree maps for the four parents of two unrelated, full-sib pedigree mapping populations in Pinus taeda (Figure 18.4). JoinMap was used to merge these four maps into a consensus map.

Gene Mapping by Bulked Segregant Analysis

Once a genetic linkage map has been constructed using genetic markers, it can be used to map the position of genes of special interest. A gene controlling a qualitatively inherited trait can be mapped simply by scoring the Mendelian segregations between this gene and marker genes of known location, adding these data to the full marker data set and reconstructing the genetic map. In many cases however, the gene controlling the qualitative trait might only be segregating in a population for which a genetic map has not been constructed. In this situation, a shortcut approach to identifying markers linked just to the qualitative trait gene can be used. This method is called bulked segregant analysis (BSA).

BSA was initially developed by Michelmore et al. (1991) for mapping disease resis-
tance genes in lettuce. This method is especially amenable to dominant molecular markers such as RAPDs and AFLPs (Chapter 4), although it can also be used with codominant markers. The basic principle of BSA relies on very strong linkage (Chapter 3) between the qualitative trait gene and one or more genetic markers. To begin, two pooled DNA samples are created by combining DNA from individuals sharing each of the alternative alleles controlling the qualitative trait. For example, in a backcross for a dominant disease resistance gene, Rr x rr, two genotypes segregate in the progeny, Rr and rr. DNA from a small number (10-20) of Rr progeny is combined to form one pooled DNA sample and likewise a small number of rr progeny DNA samples are combined to form the other pooled DNA sample. In the next step, these two pooled-DNA samples are assayed for a large number of genetic markers such as RAPDs or AFLPs. The dominant markers tightly linked to the gene controlling the qualitative trait are detected by being present in one DNA pool and absent in the other DNA pool (Figure 18.5). Markers that are not tightly linked to the gene of interest are either present in both pools or absent in both DNA pools.

BSA was first used in trees by Devey et al. (1995) to map a gene for resistance to white pine blister rust (Cronartium ribicola) in Pinus lambertiana. BSA has been used to map genes for resistance to black leaf spot (Stegophora ulmea) in Ulmus parvifolis (Benet et al., 1995), to Melampsora larici-populina in Populus hybrids (Cervera et al., 1996; Villar et al., 1996), and to fusiform rust (Cronartium quercuum) in Pinus taeda (Wilcox et al., 1996). BSA has also been used to map the pendula gene controlling the narrow crown phenotype in Picea abies (Lehner et al., 1995).

**FUNCTIONAL GENOMICS**

Gene discovery and structural genomics can provide a wealth of information about the types and numbers of genes encoded in the genome, but they provide little understanding of the function of all these genes. Functional genomics seeks to understand the function of all genes in the genome using techniques that often allow study of hundreds or thousands of genes in parallel. Gene function can be assessed at the biochemical, cellular, developmental and adaptive level. Functional genomic experimental methods are developing rapidly; we discuss just a few techniques that are used in forest trees.

**Comparative Sequencing**

The simplest way to predict the biochemical function of a gene is to determine its DNA sequence and compare it to DNA sequences of genes of known function in databases. This activity is a routine component of gene discovery, as was discussed earlier. For example, if an EST from Pinus matches the DNA sequence for an alcohol dehydrogenase (ADH) gene from corn, then the predicted function of the EST is as an ADH gene. This is not absolute proof; that can only be determined through biochemical assays. The limitation of using EST database comparison to assign biochemical function to gene sequences is that only a proportion of the genes can be identified in this way. For example, only 55% of the ESTs identified in a Pinus taeda study could be assigned function (Allona et al., 1998) and only 39% in a Populus study (Sterky et al., 1998).

Once a large number of ESTs from different tissue types and/or developmental states are obtained and functions putatively assigned based on database comparison, it is possible to ask additional questions relating to gene function. For example, Whetten et al. (2001) obtained
(Figure 18.4 continued on next page)
Fig. 18.4. Pinus taeda genetic linkage map showing the positions of genetic markers on each of the 12 chromosomes of Pinus taeda. a) Linkage groups 1-4; b) Linkage groups 5-8; c) Linkage groups 9-12. Each linkage group likely corresponds to one of the 12 chromosomes in this species. (D.B. Neale, unpublished results)

22,233 ESTs from several wood-forming tissues in Pinus taeda. They found quantitative differences in the abundance of mRNAs involved in lignin and cellulose biosynthesis between compression wood samples and normal wood samples. Comparisons of gene content and gene expression across species can also be made. Kirst et al. (2003) identified gene homologs from the angiosperm Arabidopsis thaliana in a sample of 20,377 genes from P. taeda. These results suggest that genes may be highly conserved in seed plants since angiosperms and gymnosperms diverged from one another more than 300 million years ago.

Gene Expression Analysis

Traditionally, levels of transcription are determined from Northern blot analysis (Chapter 2) where separate assays are performed to assess mRNA abundance for each gene. PCR-based techniques have also been devised to measure mRNA abundance, but like Northern blot analysis, only a small number of genes can be analyzed at once. A new technique, called DNA microarray analysis, makes it possible to study differential gene expression for thousands of genes at once (Schena et al., 1995; Schenk et al., 2000).

Small amounts of DNA from 1000 or more ESTs are first arrayed onto glass slides (Figure 18.6). RNA is then isolated from two samples for which differences in gene expression will be measured. For example, one might wish to know which genes are expressed in xylem tissue compared to phloem tissue or in trees resistant versus susceptible to a particular disease. The RNA isolated from each of these samples is labelled with a different fluorescent dye (e.g. one with red, the other with green) and the two samples are hybridized to the same glass slide containing the ESTs spotted in a grid pattern. Specialized equipment is then used to measure the amount of fluorescence at each position on the grid. The software then combines results from both fluorescent scans into a single display. EST spots showing the color of one of the sample types (red vs. green) reveals a gene more abundantly expressed in one tissue versus the other tissue. A spot showing intermediate color (yellow) represents a gene whose level of expression is roughly equivalent in both samples.
Fig. 18.5. Bulked segregant analysis to identify markers linked to a dominant resistant gene to white pine blister rust (*Cronartium ribicola*) in *Pinus lambertiana*. a) Pooled DNA samples from resistant (R) and susceptible (S) seedlings were assayed with several RAPD markers (E14, E15, etc). RAPD marker E16 clearly shows the presence of a band in the R pool that is absent in the S pool. This indicates that the E16 marker may be linked to the resistance gene. b) Assay of the DNA samples from individual genotypes comprising the R and S pools confirms that the E16 RAPD marker is linked to the resistance gene. (from Devey *et al.*, 1995)

By quantifying differential gene expression for large numbers of genes in parallel, it becomes possible to identify coordinated patterns of gene expression and regulatory networks. Microarray analysis is a very powerful tool for understanding patterns of differential gene expression between tissues, during development, in response to stresses, and among genotypes. The first application of DNA microarray analysis in a forest tree was by Hertzberg *et al.* (2001) who measured differences in the expression of 2995 genes among different developmental stages of xylem formation in a *Populus* hybrid. DNA microarray analysis has also been used to study patterns of gene expression in xylem tissues over a seasonal cycle (Egertsdotter *et al.*, 2004). In addition, patterns of gene expression during embryogenesis (van Zyl *et al.*, 2002; Stasolla *et al.*, 2003) and during drought stress (Heath *et al.*, 2002) have been investigated. Clearly, the application of this technology has just begun and our understanding of the coordinated patterns of gene expression in trees will develop rapidly in the future.

**Forward and Reverse Genetic Approaches**

There are two general approaches to understanding the function of a gene: forward genetics and reverse genetics (Figure 18.7). Forward genetic approaches begin with a well-characterized phenotype, such as a tree resistant to a disease, and then work toward identify-
Fig. 18.6. Gene expression analysis using DNA microarrays: a) mRNA samples A and B are prepared from individuals subjected to different treatments, of different genotype, from different developmental stages, different tissue types, etc. b) cDNA is synthesized from mRNA. c) cDNA is labeled with fluorescent dyes, red (dark) in sample A and green (light) in sample B. d) labeled cDNAs are hybridized to arrays. e) the amount of fluorescence at each position in the array is evaluated to determine gene expression differences between the samples. Open circles represent positions where no hybridization occurred.

Reverse genetic approaches begin with a gene, for example a protein kinase, and work toward determining which phenotype(s) it determines. Forward genetics approaches such as T-DNA tagging, transposon tagging and gene or enhancer traps require inserting foreign DNA into the host tree genome (i.e. gene transformation, see Chapter 20). These methods alter the expression of the target gene in some manner such that it then reveals a relationship with a specific phenotype that has been altered by the insertion of the foreign gene. Genetic mapping approaches such as quantitative trait locus (QTL) mapping and association mapping, discussed later in this chapter, are also forward genetic approaches and are often used because gene transfer is not required. Reverse genetic approaches such as gene silencing by RNA interference (RNAi)
or anti-sense RNA (see Chapter 20) are methods whereby foreign DNA of some kind is introduced into the host genome and the expression of individual genes is in some way disrupted. In some cases, the disruption of a gene can cause a visible mutation in the plant. Some of these reverse genetic approaches can be applied on a large scale thus enabling the assignment of many genes to specific functions. These approaches, however, require the ability to genetically transform the host plant (Chapter 20), so they are likely to be used in just a few tree species, such as in \textit{Populus}, that can easily be transformed. Reverse or forward genetic approaches requiring transformation are less likely to be used in conifers in the near future because of the difficulties with transforming conifers.

\section*{Quantitative Trait Locus (QTL) Mapping}

One of the most common applications of genetic maps in forestry is for QTL mapping. As the name implies, the goal of QTL mapping is to identify the locations in the genome of the individual loci controlling a quantitative trait. Sax (1923) developed the theoretical basis for QTL mapping and the method was first empirically demonstrated by mapping bristle number genes in \textit{Drosophila} (Thoday, 1961). It was not until the advent of molecular genetic maps in the late 1980s that QTL mapping was applied to a variety of plant and animal species.

There are four basic components to all QTL mapping experiments: (1) Segregating or mapping population(s); (2) Phenotypic measurements of the quantitative trait(s) for all members of the mapping population; (3) Genetic marker data; and (4) Statistical analysis for mapping QTLs and estimating the magnitude of their influence on phenotype (Box 18.1). Quantitative trait phenotypes (\textit{e.g.} tree size) and marker genotypes (\textit{e.g.} BB, Bb and bb) are scored on all members of a mapping population and then a statistical analysis is performed to associate phenotype with marker genotype which provides evidence for the existence of a QTL. For the example shown in Box 18.1, the B locus appears to be associated with tree size because BB homozygotes are large trees, Bb heterozygotes are immediate size trees and bb homozygotes are small trees. It is easy to see how the effects of alternate alleles at the B locus on tree size phenotype can be estimated.

Each of the four components of QTL mapping is discussed below in the context of
QTL mapping in forest trees. The fundamental difference between QTL mapping in agronomic crops, such as corn, soybean and tomato, versus forest trees, is that inbred lines generally do not exist for trees. This difference affects the types of mapping populations, choice of genetic marker systems and statistical methods that can be used for detecting QTLs.

**Mapping populations for QTL estimation**

A variety of mapping population types are used in QTL mapping of forest trees. The most important consideration in choosing a mapping population is to maximize segregation of the QTLs. Due to the long generation times in many forest tree species, mapping populations are often selected from among existing populations, such as those used in breeding programs. Mapping populations can be derived from: (1) Inter- or intra-specific crosses; (2) Outbred or inbred pedigrees; (3) Single- or multi-generation pedigrees; (4) Full-sib, half-sib or open-pollinated families; and (5) Combinations of all these family structures.

The most commonly used QTL mapping population types are: (1) The three-generation outbred pedigree; (2) The open-pollinated family; and (3) The two-generation "pseudotestcross". Groover et al. (1994) first used the three-generation outbred pedigree to map QTLs for wood specific gravity in *Pinus taeda*. This mapping population type most resembles the F$_2$ used in inbred crops (Figure 18.8). Two pairs of grandparents (first generation) are chosen: one grandparent of each pair being from the opposite end of the phenotypic distribution for the quantitative trait within the population from which they were selected. This helps to ensure that the two grandparents within a pair will differ genetically at the QTL locus. A single F$_1$ individual of intermediate phenotypic value is chosen from each of these matings (second generation) and mated; their offspring (third generation) form the segregating mapping population.

Several variations on the three-generation outbred type have been used, such as the three-generation interspecific *Populus* hybrid mapping populations (Bradshaw and Stettler, 1995; Frewen et al., 2000; Howe et al., 2000). In general, a three-generation pedigree, combined with highly informative genetic markers, can track up to four different QTL alleles potentially segregating at any one QTL; however, three-generation pedigrees with large families do not exist for most forest tree species and can take many years to develop.

Alternatives to multi-generation pedigrees are open-pollinated or half-sib families and two-generation full-sib families. In conifers, a population of seed from an individual tree can be used for QTL mapping, where the megagametophytes are employed for genetic marker analysis of segregation in gametes from the maternal parent. Most often, dominant markers (e.g. RAPDs and AFLPs) have been used with this approach. Phenotypes are measured on diploid offspring of open-pollinated or half-sib progeny. Because genetic marker data are obtained from megagametophytes, which represents only meioses in the maternal parent, there is no genetic marker information from the paternal parent of the progeny, QTLs contributed from the pollen parents cannot be estimated. This approach has less power to estimate QTLs, but can be easily applied to any tree where megagametophytes can be assayed in offspring, and does not require crossing.

In angiosperm forest trees such as *Populus* and *Eucalyptus*, the haploid megagametophyte system is not possible; however, a similar approach, called the pseudotestcross, has been used (Grattapaglia et al., 1995). Phenotyping and marker genotyping are performed on the progeny of a full-sib cross. However, dominant markers in the backcross configuration only (e.g., Aa x aa or aa x Aa) are used to track QTL segregation in the parents. Therefore, QTLs segregating in each of the parents are estimated independently and it is not possible to simultaneously estimate QTLs that are segregating in both parents, which in
crosses made between forest trees is often the case. This approach can be applied to any full-sib cross, but again has low power.

**Phenotypic measurements on quantitative traits**

The success of a QTL mapping experiment is in part dependent upon the genetic control of the quantitative trait. The first consideration often made is to estimate the number of genes controlling a trait. Theoretical methods have been developed for estimating the number of genes controlling a complex trait, but in practice these have not been used prior to QTL mapping experiments in forest trees. The question is often whether a trait is purely polygenic (controlled by many genes of small effect) or is oligogenic (controlled by a smaller number of genes of larger effect). Intuitively, QTLs for a trait under oligogenic control should be easier to detect than for a trait under polygenic control. In practice, QTL mapping experiments must be performed without this knowledge and draw inference of genetic control following the outcome of the experiment. As will be described later, QTL mapping experiments have contributed significantly to the understanding of the genetic control of complex traits in forest trees.

Factors that affect power and precision of QTL estimation are much the same as those that affect quantitative genetic parameter estimation (Chapter 15). Minimizing environmental variation and therefore increasing heritability, increases the power to detect QTLs. Clonal testing of the progeny in the mapping population is the most efficient way to minimize the effect of environmental variation and obtain better estimates of the phenotypic value because data on multiple randomized copies (ramets) of a given genotype are averaged together (Bradshaw and Foster, 1992).

The quantitative traits chosen for QTL mapping in forest trees are much the same as those involved in tree management because of the interest in marker-aided breeding (Chapter 19). Stem growth, wood quality, hardness, disease resistance, and reproductive traits have been included in QTL mapping experiments (Sewell and Neale, 1999). QTL mapping might also help to understand the genetic architecture of non-commercial traits and provide insights into genome evolution and speciation.

**Box 18.1. Quantitative trait locus (QTL) mapping in forest trees.**

a) Mate two parent trees at the extremes of the phenotypic distribution of the trait (e.g. a large tree and a small tree) that are also different for a large number of genetic markers (e.g. AA versus aa). Mate an F₁ progeny that is intermediate for the phenotype and heterozygous at marker loci with a similar F₁ (F₁'). This can be either a full-sib mating or more often a cross between two unrelated F₁’s. The progeny resulting from the F₁ x F₁' mating segregate for both the phenotype and the genetic markers. b) Each of the progeny is genotyped at all marker loci. Many different marker types can be used (Chapter 4). In the example the B locus marker is codominant so both homozygotes (BB and bb) and the heterozygote (Bb) can be scored. c) A statistical test is then performed to test for differences in mean phenotypic values among genotypic classes. In this example, BB genotypes are associated with large trees, Bb heterozygotes with intermediate size trees and bb homozygotes with small trees. The inference that can be drawn is that a QTL for tree size resides on the chromosome somewhere near the B genetic marker. This analysis is performed for all markers on all chromosomes to discovery QTLs controlling the phenotype. (Box 18.1 continued on next page)
Genetic markers and maps

The types of genetic markers available for use in forest trees were described in detail in Chapter 4. Allozymes, RFLPs, RAPDs, AFLPs, SSRs, and ESTPs are all useful for QTL mapping in forest trees.

**Figure 1.** A schematic description of the basic approach to quantitative trait loci (QTL) mapping in an outbreeding forest tree.
mapping, although some are better suited than others for application with different mapping population types. This is due to the extent of **marker informativeness** of different marker types (Table 18.3). Informativeness is a general term, which describes how completely a genetic marker “marks” the segregating variation in the region of the genome where it is located. For example, codominant markers (e.g. RFLPs, SSRs and ESTPs) are more informative than dominant markers (e.g. RAPDs and AFLPs) because the heterozygote (Aa) can be distinguished from the dominant homozygote (AA). However, there is no loss of information using dominant markers in conifer haploid megagametophyte mapping populations, because genotyping is performed on gametes (A or a) and not on zygotes (AA, Aa or aa).

Genetic marker data need not be organized into the form of a genetic map to perform QTL detection (see single-factor approach below). However, the goal of most QTL experiments is to both detect the presence of QTLs and determine their location in the genome. Thus, genetic maps are usually constructed from marker data prior to QTL mapping. One type of QTL detection approach, called interval mapping (see discussion of interval mapping below) requires that markers be mapped.

**Statistical methods for detecting QTLs**

In theory, the basic statistical approach for detecting QTLs is simple: all methods test for a relationship between the quantitative trait values and the marker genotypic classes in the mapping population. In practice, statistical methods and computational procedures can be quite complex, with two general approaches distinguished: (1) The single-factor approach;
and (2) The multi-factor or interval approach.

The single-factor approach uses analysis of variance or regression analysis to test for differences in the quantitative trait means among marker genotypic classes (Edwards et al., 1987). For example, does the height value mean of individuals in the AA genotype class differ from that of individuals in the aa genotypic class. Analyses are performed one marker at a time. A statistically significant association (i.e. significant F-value) is considered evidence for the presence of a QTL mapping somewhere near the genetic marker. The single-factor approach requires only basic statistics and can be performed using standard statistical analysis software packages. The major limitation of this approach is that the exact position of the QTL, relative to the genetic marker, cannot be easily determined. This limits the power of detection and precision in the estimation of the magnitude of effects of the QTL on the phenotype (Lander and Botstein, 1989).

The multi-factor or interval approach was developed to overcome the limitations of the single-factor approach (Lander and Botstein, 1989). QTLs are detected using the information in pairs of markers that flank a segment of a chromosome. The statistical estimators are complex, often using maximum likelihood procedures and computationally demanding solutions. The interval mapping software Mapmaker/QTL has been widely used in forest trees, even though it was written for use with inbred species (Lincoln et al., 1993). The interval mapping method of Knott et al. (1997) was designed specifically for use in outbred forest trees.

QTL discovery in forest trees

QTL mapping research in forest trees has provided new insight into the genetic control of complex traits by: (1) Estimating the numbers and magnitude of effects of QTLs; (2) Revealing the genomic locations of QTLs; and (3) Describing environmental and developmental patterns of QTL expression. Knowledge of such factors can, in some cases, be used directly in tree improvement through marker-aided breeding (Chapter 19) or indirectly through influencing the choice of breeding designs.

QTLs have been identified for a variety of traits and in a variety of species and were reviewed in a paper by Sewell and Neale (1999). Recent examples of QTL studies in some representative forest tree genera include; *Pinus* (Hurme et al., 2000; Sewell et al., 2002; Brown et al., 2003; Devey et al., 2004), *Pseudotsuga* (Jermstad et al., 2003), *Populus* (Freden et al., 2000), *Eucalyptus* (Thamarus et al., 2004; Kirst et al., 2005), *Quercus* (Saintagne et al., 2004; Scotti-Saintagne et al., 2004) and *Fagus* (Scalfi et al., 2004). In general the number of QTLs detected and the relative magnitude of their effects vary little among traits (Sewell and Neale, 1999). With a few exceptions, individual QTLs for growth and development, wood quality, adaptability and reproduction each account for no more than 5-10% of the phenotypic variance in their respective traits and the number of QTLs detected per trait is generally less than ten. These results suggest that the polygenic mode of inheritance is most likely for quantitative traits of commercial interest in forest trees. QTLs fitting the oligogenic model might eventually be detected in specific crosses involving trees carrying major genes, but it seems safe to conclude that few major QTLs will be detected such that a single QTL, alone, determines a large proportion of the variation in a quantitative trait.

Determination of the chromosomal location of QTLs in genomes will be important to enable use of transgenic technology in tree improvement (Chapter 20). Eventually it may become possible to genetically engineer the promoter regions of genes coding for QTLs or even replace endogenous genes with engineered genes (Chapter 20). These types of genetic manipulations, however, require precise knowledge of the location of genes to be modified. The molecular basis of pleiotropy may also become apparent as QTLs for many
different traits are localized on genetic maps.

QTLs, just as the quantitative traits themselves, can vary in their expression developmentally and when trees are grown in different environments. However, QTLs that have stable expression across time and space will be most valuable for marker-aided breeding applications. Developmental stability of QTL expression has been shown for growth traits in *Pinus pinaster* (Plomion and Durel, 1996) and wood quality traits in *Eucalyptus* (Verhaegen et al., 1997) and *Pinus taeda* (Sewell et al., 2000; Brown et al., 2003), but not for growth traits in *Pinus taeda* (Kaya et al., 1999) or *Populus* (Bradshaw and Stettler, 1995). Likewise, QTL by environment interactions have been shown for wood density in *Pinus taeda* (Groover et al., 1994) and for bud flush in *Pseudotsuga menziesii* (Jermstad et al., 2001, 2003). Clearly, patterns of QTL expression are complex and experiments need to be repeated across environments and years to fully understand how specific QTLs determine quantitative trait phenotypes.

### Positional Cloning of QTLs

Once a QTL or simple Mendelian trait has been precisely mapped, it then may be possible to clone the underlying gene based solely on the knowledge of its genetic map position. Once cloned, its DNA sequence and ultimately its biochemical function can be determined. This is known as **positional or map-based cloning**. The first step is to identify a large piece of cloned genomic DNA that contains two mapped genetic markers which flank, on opposite sides, the QTL or gene coding a simple Mendelian trait. Large-insert DNA libraries from the host genome are constructed in yeast artificial chromosomes (YACs) or in bacterial artificial chromosomes (BACs) (Figure 18.9).

In the past, genetic maps were not sufficiently dense with genetic markers such that it could be expected that the genetic markers flanking the qualitative trait gene would reside on the same YAC or BAC clone. A technique called “chromosome walking” was employed to identify the target gene. A physical map was constructed in the region including the qualitative trait gene by ordering the overlapping YAC or BAC clones based on genetic markers or DNA sequences contained within the clones. Eventually a single YAC or BAC clone could be identified that contained the target gene and its DNA sequence could be determined. Several genes may be encoded on the YAC or BAC clone, so it is yet another step to identify the true target gene. With the development of very dense genetic maps the “walking” step can now often be eliminated, if it is likely that the flanking markers reside on the same YAC or BAC. This approach is called “chromosome landing” (Tanksley et al., 1995) and has been used to clone a number of disease resistance genes in crops. A final step that is required to confirm positional cloning of the gene is to insert the cloned gene into a host genome that does not have the phenotype, for example, inserting a cloned disease resistance gene into a susceptible host. Strong constitutive or inducible promoters are usually attached to the target gene to ensure that it is expressed once integrated into the new host genome. This confirmation test is called a complementation test.

The feasibility of positional cloning of genes is highly dependent on the size of the genome. Many genes have been cloned in this way from species with small genomes such as *Arabidopsis* and rice, but the task is much more difficult and expensive for species with large genomes. No genes have yet been positionally cloned from forest trees. However, several efforts in *Populus* are underway (Stirling et al., 2001; Zhang et al., 2001). Positional cloning of genes from conifers is a daunting task and success using this approach may have to await the development of new technologies.
Fig. 18.9. Bacterial artificial chromosome cloning: a) High molecular DNA is isolated. b) DNA is cut into large fragments using a restriction enzyme. c) The bacterial chromosome vector is prepared to accept the large insert DNA. d) The DNA fragments are inserted into the bacterial chromosome. e) The bacterial colonies containing the cloned DNA are plated out onto agar plates. f) DNA is isolated from the bacterial chromosomes and put into microtiter plates.

**Association Genetics**

QTL mapping studies are very useful for estimating the number, size of effects and approximate location in the genome of genes controlling complex traits. As we discussed in the last section, however, it is a very difficult task to identify the exact gene underlying the QTL. Another approach, called association mapping, can be used to more precisely determine the exact genes controlling complex traits, and therefore ultimately identify mutations responsible for phenotypic differences among individuals for a quantitative trait. Each functional variant (i.e. mutation) is referred to as a quantitative trait nucleotide (QTN).

The fundamental difference between QTL mapping and association genetics is that the former relies on genetic linkage (Chapter 3) following a generation or two of crossing,
whereas the later takes advantage of population-level linkage disequilibrium (LD) (Chapter 5) between genetic markers and QTNs following many generations in a large intermating population (Figure 18.10). The association mapping approach has been developed for identifying genes controlling complex traits in humans (Risch, 2000; Cardon and Bell, 2001; Weiss and Clark, 2002) and has only recently been applied to plants (Remington et al., 2001; Rafalski, 2002; Neale and Savolainen, 2004).

There are two basic approaches to association genetics; (1) genome scan and (2) candidate gene. In the genome scan approach, genetic markers are populated throughout the genome so that the entire genome can be searched for QTNs. In the candidate gene approach, genetic markers are used only within individual genes that are thought to be involved in determining the phenotype. The genome scan approach is thus more exhaustive in its search but is also more expensive to conduct.

Figure 18.10. Association genetics of complex traits in forest trees. Association genetics is similar to QTL mapping (Box 18.1) except that chromosomes being mapped are more highly recombined than those of QTL mapping populations due to the accumulation of recombination events over evolutionary time. This enables greater resolution of genes controlling complex traits. In this example, a T/C single nucleotide polymorphism (SNP) is associated with tree size; therefore, this region of the chromosome is presumed to include a gene that in part determines tree size.
Association genetics requires four components similar to those necessary for QTL mapping: (1) A mapping population; (2) Phenotypic measurements of the quantitative trait(s) for all members of the mapping population; (3) Genetic marker data for all members of the mapping population; and (4) A statistical method for associating genotype with phenotype. Issues related to phenotyping are the same as with QTL mapping and need no further discussion; however, there are many differences for the other three components.

Association mapping populations

An association population is generally constructed by sampling individuals from a large random mating population. However, some association statistical tests such as the transmission disequilibrium test (TDT) require family structure (Lynch and Walsh, 1998). The informational content of an association genetics population depends on the amount of population-level LD between genetic markers and QTNs. The amount of LD in turn depends on population history, such as the occurrence and degree of past bottlenecks (Chapter 5) and subsequent recombination. For example, a population that has undergone a recent bottleneck and/or has a very low rate of recombination would have high LD relative to a large random mating population with high recombination. Most, but not all, natural forest tree populations are expected to fall into the latter category. However, artificially constructed populations, such as breeding populations, might be expected to have much higher LD. The greater the LD, the more likely that associations between genetic markers and QTNs will be detected. This is advantageous for breeding applications (Chapter 19); however, if the goal is to exactly pinpoint the QTN, somewhat less LD would be required. This is because there could be one or more genetic markers in complete LD with the QTN and thus impossible to ascertain which polymorphism is truly the QTN.

Simulation studies show that association mapping populations should include at least 500 individuals (Long and Langley, 1999) and that measures such as clonal replication of each sampled individual should be taken to increase precision of evaluating phenotypes.

Single nucleotide polymorphisms

Association genetics is based on LD over short physical distances along chromosomes and therefore requires detection of polymorphisms over these short distances. Most of the genetic marker-types described in Chapter 4 are not sensitive enough for this purpose. Therefore, markers derived from single nucleotide polymorphisms (SNPs) are primarily utilized in association genetics. Methods based on DNA sequencing are used to discover SNPs. In Pinus taeda, it is estimated that a SNP occurs once every 60 base-pairs (bp) (Brown et al., 2004). In contrast, SNPs are found approximately 1/1000 bp in humans; therefore SNPs are clearly abundant in some forest trees and their discovery is straightforward. SNPs are usually bi-allelic, although very rarely (<1%) tri- or tetra-allelic SNPs are found. This is because the likelihood of a second mutation occurring at the same nucleotide position in a population is exceedingly small.

SNPs can be discovered either by multiple alignment and comparison of ESTs in electronic databases (electronic SNPs) or by de novo sequencing of a sample of individuals taken from a population. Haplotype-based association tests sometimes have more power to detect associations with phenotypes; therefore, it is desirable to either infer or determine directly the haplotypes of alleles in the population. A haplotype is defined as the distinct combination of SNPs on a single chromosome. In conifers, this can be easily accomplished by using haploid megagametophyte tissue. For example, there were approximately nine different haplotypes among a sample of 31 loblolly pine megagametophyte DNA samples
for the AGP6 gene (Figure 18.11). These SNPs and haplotypes were determined by DNA sequencing.

Once SNPs have been discovered in a small sample, SNP genotypes of all members of the association population must be determined. In forest trees where nucleotide diversity is generally quite high, it is necessary to prioritize SNPs for genotyping in large association populations because it would be cost-prohibitive to type all discovered SNPs. Priorities can be assigned based on the haplotype structure, functional potential of the SNP (e.g. synonymous or nonsynonymous) or tests of selective neutrality e.g. whether a SNP is under selection or selectively neutral. There are many different ways to genotype specific SNPs, each method with its own specialized chemistries and instrumentation (see review by Syvanen, 2001) but all require high throughput technologies to enable genotyping hundreds or thousands of trees at hundreds or thousands of SNPs. SNP genotype methods are evolving rapidly and there is intense competition in the life sciences industry to lower assay costs and increase throughput.

Fig. 18.11. An example of direct determination of haplotypes in conifers using In seed megagametophytes. a) Seed megagametophytes are excised from the seed coat and embryo and DNA is isolated from the megagametophyte tissue; b) Gene structure of the AGP6 gene from Pinus taeda: positions < -41 make up to the flanking region, -41 to 622 is the 1st exon, 622 to 725 is an intron, 725 to 813 is the 2nd exon and >1079 is 3’ flanking region. Two fragments (amplicons) were selected for DNA sequencing –12 to 548 and 575 to 990; c) Results of sequencing 31 megagametophyte DNA samples for both amplicons. Only nucleotide positions that were polymorphic among megagametophytes are shown. These are called single nucleotide polymorphisms (SNPs). In total, 16 haplotypes are revealed.

Association tests
A statistical test for an association between a genetic marker (genotype) and a quantitative
trait (phenotype) has similar properties to QTL detection. Members of the association
mapping population are classified based on their genotype and standard statistical analyses
(ANOVA or regression) can be performed to test for differences in the phenotypic means
among genotypic classes. SNP genotypic classes can be assigned based on individual hap-
loid or diploid SNPs or based on a multi-SNP haplotypes.

**Association mapping in forest trees**

Association mapping experiments in forest trees have only recently been implemented.
Neale and Brown (2005) reported preliminary results from a candidate gene approach to
search for associations between genes involved in lignin and cellulose synthesis and wood
property traits in *Pinus taeda*. SNP discovery was performed in a sample of 32 megagame-
tophytes (Brown et al., 2004). Linkage disequilibrium between SNPs within genes was
detected but the rate of decay of LD is fairly rapid (~2000 base pairs) (Figure 18.12). An
association population of 425 *Pinus taeda* clones was evaluated for several wood property
traits including wood specific gravity, microfibril angle and percent lignin and cellulose.
Associations between SNP genotypes in a few of the lignin biosynthetic pathway candidate
genes and wood property phenotypes were found. This preliminary study demonstrates that
it is possible to use an association mapping approach to identify individual genes control-
ling complex traits in forest trees. This approach enables application of marker-assisted
breeding for both within-family selection and between-family selection (*Chapter 19*).

![Fig. 18.12. Plot of the measure of linkage disequilibrium, r², versus distance in base pairs among polymorphic SNPs in 19 *Pinus taeda* genes. The fitted curve shows that linkage disequilibrium decays rather rapidly in *Pinus taeda*. (from Brown et al., 2004)](image_url)
Comparative genomics has developed into an experimental science of comparing genomes among species. Comparisons are most often made at the level of DNA sequences and genetic maps. There are several plant and animal species which are rich in gene discovery and genetic map information, such as human, mouse, Arabidopsis, rice and poplar. These model organisms have: (1) Very dense genetic maps; (2) Thousands of ESTs sequenced and mapped; and (3) Genomes that are completely sequenced. For a number of plant model species, every gene and its genetic map position will soon be known. The function of all genes in model species is also well on the way to being understood.

For the vast majority of species, including tree species, much less genomic information will exist in the near future. However, the information from model species can be directly “accessed” if genetic maps can be compared. This situation is already true in several major species groups, most notably in the grasses (Bennetzen and Freeling, 1993). The genetic maps of rice, corn, sorghum, barley, wheat, rye, millet, and sugarcane can be directly compared due to the discovery of synteny among the genomes of these related species. Synteny means the conservation of gene order on linkage groups or chromosomes following speciation and evolution. Although chromosome number and ploidy levels can vary among related species, large chromosomal segments with genes in the same order can be identified. Candidates for qualitative trait genes known only by phenotype or genes underlying QTLs in non-model species can be identified by comparing to map positions in model species. This approach of identifying genes from related species based on their similar location in the genome can facilitate the “chromosome landing” method described in a previous section (see Positional Cloning of QTLs).

Comparative mapping between distantly related plant groups should be possible in the future, for example between Populus or Eucalyptus and Arabidopsis, but for the moment comparative mapping in trees is limited to comparisons between species within a genus or family. Comparative maps have been constructed in the Pinaceae and Fagaceae families and for the genus Eucalyptus.

Comparative mapping requires orthologous genetic markers to be mapped in each of the species being compared. Orthologs are genes that have descended from a common ancestral locus, whereas paralogs are loci that have originated by gene duplications within an individual species (Figure 9.7). Most of the anonymous marker types (RAPD, AFLP, SSR) cannot be used for comparative mapping because loci are not orthologous across species. However, SSR markers have been used for comparative mapping in Eucalyptus (Marques et al., 2002) and Fagaceae (Barreneche et al., 2003).

Genetic markers based on genic DNA sequences, such as RFLPs and ESTPs, are most useful for comparative mapping. Because RFLPs are assayed by Southern hybridization, both orthologs and paralogs are revealed, thus RFLPs can be used for comparative mapping as long as orthologs are identified. Devey et al. (1999) used RFLP loci from both Pinus taeda and Pinus radiata to construct comparative maps between these species. However, because orthologs and paralogs are not easily distinguished in RFLP markers and because they are difficult to apply, RFLPs are unlikely to be used widely for comparative mapping. ESTP markers have many positive attributes needed for comparative mapping (Temegsen et al., 2001); ESTPs among species are usually orthologous and only occasionally paralogous. ESTP markers from Pinus taeda have been used to construct comparative maps between Pinus taeda and three other important species in the genus: Pinus elliottii (Brown et al., 2001), Pinus pinaster (Chagne et al., 2003) and Pinus sylvestris (Komulainen et al., 2003). It is now possible to treat Pinus as a single genetic system and perform comparative genomic analyses among species within the genus (Figure 18.13). A comparative map between Pinus taeda and Pseudotsuga menziesii has also been constructed (Krutovsky et al., 2004), thus extending comparative mapping between genera within the Pinaceae.
**Bioinformatics and Databases**

The high-throughput technologies of genomic sciences enable the collection of overwhelming amounts of data. **Bioinformatics** combines aspects of biotechnology, statistics, information science, and computational biology to devise new ways to analyze and extract knowledge from the vast amount of genomic data. The primary types of genomic data are; (1) DNA sequences, (2) Genetic mapping data and (3) Data resulting from functional analyses, e.g. from DNA microarray experiments. Aside from bioinformatic methods to analyze genomic data, databases must be developed, curated and made accessible to other researchers. Some of the more commonly used bioinformatic tools and database structures used in forest tree genomics are briefly described in this section.

The National Center for Biotechnology Information (NCBI) is the primary site in the United States for DNA sequence databases and DNA sequence analysis tools. The primary database is called GenBank. NCBI also provides on-line access to the Basic Local Alignment Search Tool (BLAST) programs, which are the primary tools used to search the databases and identify matches among sequences. All of these resources are free and publicly available through the World-Wide-Web.

The basic principles of linkage analysis and genetic mapping were described in Chapter 4 and earlier in this chapter. Descriptions of linkage and QTL mapping methods can be found in Liu (1998). Rockefeller University hosts a web site of genetic analysis software, including linkage and QTL mapping programs (http://linkage.rockefeller.edu/soft/list.html). A number of software packages have been developed which provide programs for both linkage map-
ping and QTL analysis. Two such suites of programs are the Mapmaker/EXP (Lander et al., 1987) and Mapmaker/QTL programs (Lincoln et al., 1993) and the JoinMap (Stam and van Ooijen, 1995) and MapQTL (van Ooijen, 2004) programs. These and other genetic mapping programs have been used extensively with forest trees, although none are designed to optimally analyze forest tree data.

The primary repository of forest tree genomic data is the TreeGenes Database that is maintained by the Dendrome Project at the Institute of Forest Genetics, Davis, California (http://dendrome.ucdavis.edu). TreeGenes contains a variety of data-types (Figure 18.14) and is an object-oriented database that allows complex queries and searches. Through the use of databases and bioinformatic tools it is possible to perform experiments in silico and begin to understand all the complex relationships among genes and how they work together to determine the phenotype.

Figure 18.14. Homepage for the TreeGenes Database. TreeGenes is the repository for all data resulting from genome research in forest trees.

**SUMMARY AND CONCLUSIONS**

**Genomics** is the study of the structure, location, function, regulation and interaction of large numbers of genes simultaneously. Genomics is an extension of some of the traditional sub-disciplines of genetics, such as transmission and molecular genetics, to the entire genome. Genomics is a science that is made possible and is dependent upon technologies that allow rapid analysis of hundreds or thousands of genes and often in many individuals at once, such as by automated DNA sequencing.

The discovery and cataloging of all genes in the genome is an integral component of genomics. One approach is to determine the DNA sequence of the entire genome and infer genes from the DNA sequence. This approach has been applied to Populus, but is currently not feasible in conifers because of their large genome sizes. An alternative approach is to determine the DNA sequences for just the gene-coding regions. This can be accomplished by sequencing cDNA derived from mRNA of genes being expressed at the time of the experiment. These sequences are called expressed sequence tags (ESTs). ESTs are submitted to databases and compared to all other sequences in the database to see if they match genes whose function has already been determined. EST databases of tens of thousands of ESTs have been developed for Pinus, Populus, and Eucalyptus.

The construction of genetic linkage maps is another integral component of genomics. Genetic maps show the position of genes relative to one another on chromosomes and are valuable for understanding genome organization and evolution. Maps are extremely useful tools for identifying genes controlling interesting phenotypes. Qualitatively inherited traits, such as disease resistance genes, can be located on maps and then cloned based on knowledge of their map position. The map positions of the individual genes controlling quantitatively inherited traits, called QTLs, can also be determined from analyses using genetic maps. QTLs for a variety of growth, wood quality, and other economic traits have been identified. Knowledge of the number and size of effects of QTLs controlling a quantitative trait can assist tree breeders.

Cataloging and mapping all the genes is, however, only an initial step in genomics. The ultimate goal is to understand the function of all genes and their interactions. Tech-
niques such as microarray analysis are used to study the expression patterns of genes. Eventually all genes in all biochemical pathways will be known, as well as how these genes and gene products interact. Functional genomic studies also seek to determine the relationship between the vast amount of allelic diversity in genes and the array of different phenotypes found in populations.

Comparative genomic analysis is possible in a few forest tree genera (*Pinus* and *Eucalyptus*) following the development of comparative genetic maps based on orthologous genetic markers. Complete genome sequencing in *Populus* will greatly enhance comparative genomic analysis in this taxa.