Forestry Department

Food and Agriculture Organization of the United Nations

Forest Genetic Resources Working Papers

Forest Genomics for conserving adaptive genetic diversity

by

Konstantin V. Krutovskii and David B. Neale

July 2001

Forest Resources Development Service
Forest Resources Division
Forestry Department

Working Paper FGR/3 (E)
FAO, Rome (Italy)

Disclaimer

The Forest Genetic Resources Working Papers report on issues addressed in the work programme of FAO. These working papers do not reflect any official position of FAO. Please refer to the FAO website (www.fao.org/fo) for official information.

The purpose of these papers is to provide early information on on-going activities and programmes of major interest, and to stimulate discussion.

Comments and feedback are welcome.

For further information please contact:

Ms. Christel Palmberg-Lerche, Chief
Mr. Pierre Sigaud, Forestry Officer (Forest Genetic Resources)
Mr. Peter Iversen, APO (Forest Genetic Resources)
Forest Resources Development Service
Forest Resources Division
Forestry Department
FAO
Viale delle Terme di Caracalla
I-00100 Rome (Italy)
e-mail: christel.palmberg@fao.org
pierre.sigaud@fao.org
peter.iversen@fao.org

The present paper is based on a lecture by Dr. Konstantin V. Krutovskii, entitled: "Forest genomics and new molecular genetic approaches to measuring and conserving adaptive genetic diversity in forest trees", presented at a training workshop organized by the International Plant Genetic Resources Institute (IPGRI), Rome and the Austrian Federal Ministry of Agriculture and Forestry, Environment and Water Management (BMLFUW), in technical collaboration with FAO. The training workshop was held in Gmunden, Austria 30 April to 11 May 2001.

Address of the authors: USDA Forest Service, Pacific Southwest Research Station, Institute of Forest Genetics, Environmental Horticulture Department, University of California at Davis, One Shields Avenue, Davis, Ca 95616, USA

E-mail: kkrutovs@dendrome.ucdavis.edu; http://www2.psw.fs.fed.us/ifg/Staff/kostya.htm

For quotation:
FAO (2001). *Forest genomics for conserving adaptive genetic diversity*. Paper prepared by Konstantin V. Krutovskii and David B. Neale. Forest Genetic Resources Working Papers, Working Paper FGR/3 (July 2001). Forest Resources Development Service, Forest Resources Division. FAO, Rome (*unpublished*).

FOREST GENOMICS FOR CONSERVING ADAPTIVE GENETIC DIVERSITY

by

Konstantin V. Krutovskii and David B. Neale

USDA Forest Service, Pacific Southwest Research Station,
University of California at Davis, USA

1. INTRODUCTION

1.1. Why it is important to measure and save adaptive genetic diversity in forest tree populations

Genetic diversity is the basis of the ability of organisms to adapt to changes in their environment through natural selection. Populations with little genetic variation are more vulnerable to the arrival of new pests or diseases, pollution, changes in climate and habitat destruction due to human activities or other catastrophic events. The inability to adapt to changing conditions greatly increases the risk of extinction. Gene conservation management aimed to save adaptive genetic diversity should be based on the knowledge of the genetic basis of adaptation. The goal of this paper is to describe how adaptive genetic diversity can be measured using new molecular genetic approaches and achievements in forest genomics.

1.2. Traditional methods to measure adaptive genetic diversity

1.2.1. Field Experiments

Field experiments (common-garden tests) have been used traditionally to measure adaptive genetic diversity in trees. These tests continue to be used extensively in tree breeding and are very effective in identification of families and clones that are specifically adapted to particular environments or to a broad variety of environments. However, field experiments are very time consuming and relatively expensive, and more importantly, they are based solely on the phenotypes. They can estimate genetic parameters but only on measurable traits, not on individual genes. This method can neither provide information on what particular genes and how many of them are involved in adaptation nor how much of phenotypic variation can be explained by genetic variation in these genes.

1.2.2. Molecular genetic markers

Another, generally complementary, approach for estimating adaptive genetic diversity is to measure genetic variation using molecular genetic markers. However, DNA variation that resides in the non-coding genomic regions or does not lead to a change in the amino acid sequence (for example, so-called synonymous nucleotide substitutions in the second or the third positions in a codon encoding an amino acid) is unlikely to have any significant

contribution to adaptation. Many modern genetic markers belong to so-called anonymous DNA marker type such as microsatellites or simple sequence repeats (SSRs), restriction fragment length polymorphisms (RFLPs), random amplified polymorphic DNA (RAPDs), and amplified fragment length polymorphisms (AFLPs). These marker types generally measure apparently neutral DNA variation, and are very useful (with different efficiency, of course) in the analysis of phylogenetic relationships, population structure, mating system, gene flow, parental assignment, introgressive hybridization, marker-aided selection and genetic linkage. They are not useful for measuring adaptive genetic diversity.

Isozymes are another class of genetic markers widely used in forest genetics in the last several decades. Although variation revealed by these markers is caused by amino acid variation, it is unclear whether this variation is selectively neutral or has any adaptive significance. There are many studies showing great adaptive differences (in morphological or phenological characteristics) among populations of forest tree species, but no accompanying differences for the isozyme markers (see references in Boshier and Young 2000).

Markers of all kinds are used now in forest genetics — both anonymous and genic, dominant and codominant, highly and less polymorphic, expensive and inexpensive, supposedly selective and apparently neutral, abundant and less numerous. A classification of genetic markers is offered in Table 1, which takes into account their most important features. Details on the nature of these markers, their advantages and disadvantages and use in different applications are available elsewhere (see the most recent reviews by Cervera *et al.* 2000; Linhart 2000; Glaubitz and Moran 2000; Savolainen and Karhu 2000; Chapters 12-14 in Mandal and Gibson 1998).

The ideal marker for estimating adaptive variation should meet the following criteria: (1) be directly involved in the genetic control of adaptive traits; (2) have identified DNA sequence and known function; (3) be readily available for genetic analysis, and (4) have easily identifiable allelic variation. No marker fully satisfies all these criteria. However, a promising new marker, expressed sequence tag polymorphisms (ESTPs), seems to satisfy most or all of these criteria, emerged recently as a result of genomic studies.

Table 1: Comparison of commonly used genetic markers

Feature	RFLP	SSR	RAPD	AFLP	Isozymes	STS/EST
Origin	Anonymous / Genic	Anonymous	Anonymous	Anonymous	Genic	Genic
Maximum theoretical number of possible loci in analysis	Limited by the restriction site (nucleotide) polymorphism (tens of thousands)	Limited by the size of genome and number of simple repeats in a genome (tens of thousands)	Limited by the size of genome, and by nucleotide polymorphism (tens of thousands)	Limited by the restriction site (nucleotide) polymorphism (tens of thousands)	Limited by the number of enzyme genes and histochemical enzyme assays available (30-50)	Limited by the number of expressed genes (10,000-30,000)
Dominance	Codominant	Codominant	Dominant	Dominant	Codominant	Codominant
Null alleles	Rare to extremely rare	Occasional to common	Not applicable (presence/ absence type of detection)	Not applicable (presence/ absence type of detection)	Rare	Rare
Transferability	Across genera	Within genus or species	Within species	Within species	Across families and genera	Across related species
Reproducibility	High to very high	Medium to high	Low to medium	Medium to high	Very high	High
Amount of sample required per sample	2-10 mg DNA	10-20 ng DNA	2-10 ng DNA	0.2-1 µg DNA	Several mg of tissue	10-20 ng DNA
Ease of development	Difficult	Difficult	Easy	Moderate	Moderate	Moderate
Ease of assay	Difficult	Easy to moderate	Easy to moderate	Moderate to difficult	Easy to moderate	Easy to moderate
Automation / multiplexing	Difficult	Possible	Possible	Possible	Difficult	Possible
Genome and QTL mapping potential	Good	Good	Very good	Very good	Limited	Good
Comparative mapping potential	Good	Limited	Very limited	Very limited	Excellent	Good to very good
Candidate gene mapping potential	Limited	Useless	Useless	Useless	Limited	Excellent
Potential for studying adaptive genetic variation	Limited	Limited	Limited	Limited	Good	Excellent

Table 1: Comparison of commonly used genetic markers (cont.)

Cost						
Feature	**RFLP**	**SSR**	**RAPD**	**AFLP**	**Isozymes**	**STS/EST**
Development	Moderate	Expensive	Inexpensive	Moderate	Inexpensive	Expensive
Assay	Moderate	Moderate	Inexpensive	Moderate to expensive	Inexpensive	Moderate
Equipment	Moderate	Moderate to expensive	Moderate	Moderate to expensive	Inexpensive	Moderate to expensive

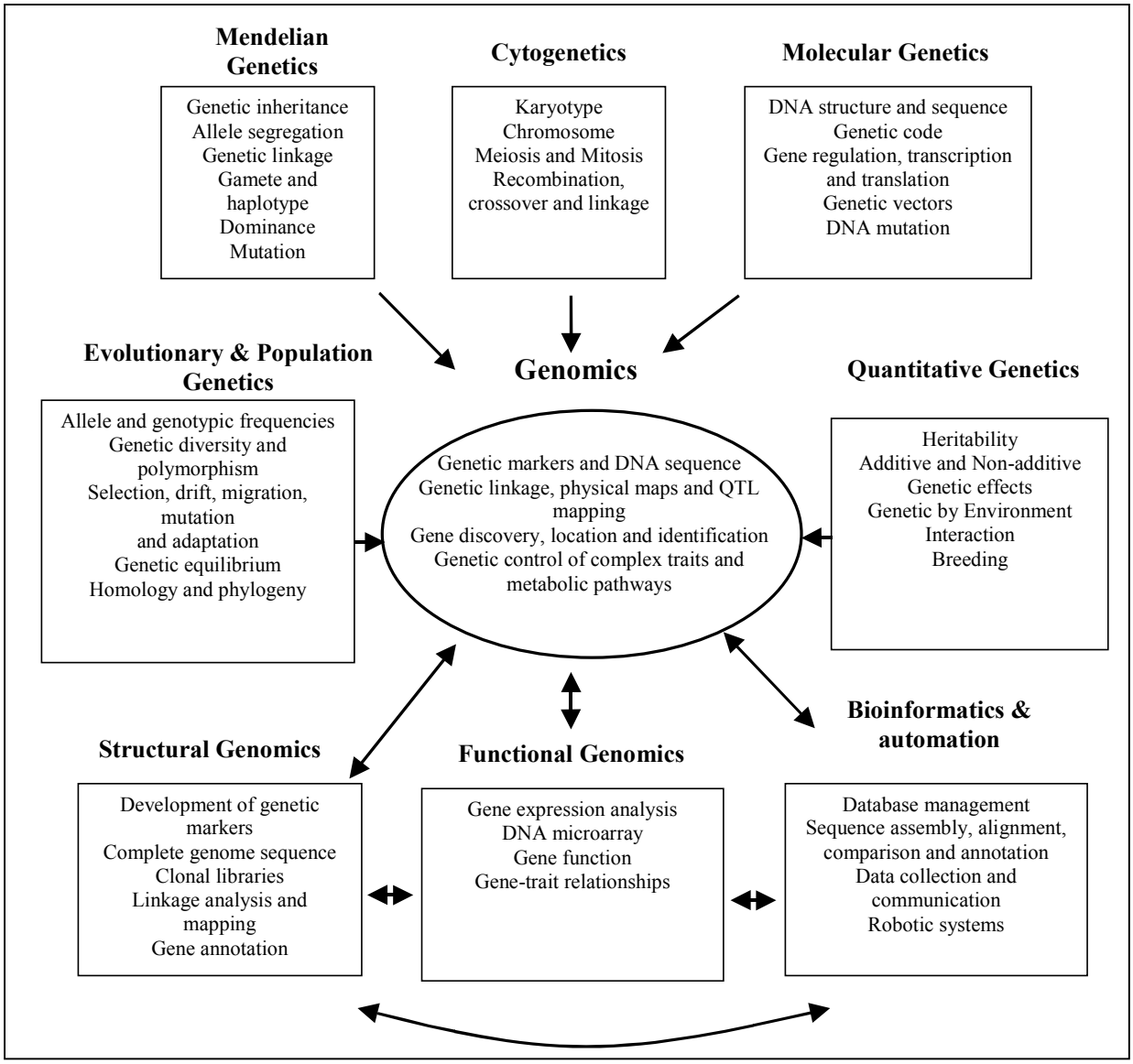
RFLP - restriction fragment length polymorphism; SSR - simple sequence repeats (microsatellites); RAPD - random amplified polymorphic DNA; AFLP - amplified fragment length polymorphism; STS - sequence tagged site; EST - expressed sequence tags.

2. HOW FOREST GENETIC CONSERVATION CAN BENEFIT FROM NEW ACHIEVEMENTS IN GENOMICS

2.1. Introduction to genomics

Genomics has arisen as a new science that studies the whole genome by integrating traditional genetic disciplines such as population, quantitative and molecular genetics with new technologies in molecular biology, DNA analysis, bioinformatics and automated robotic systems (Figure 1).

Figure 1: Genomics is a broad discipline that integrates traditional areas of genetics (adapted from Figures 1.1 and 1.2 in Liu 1998).



A number of subdisciplines of genomics can be combined to provide a powerful approach to studying adaptive genetic variation: structural, functional, comparative, statistical and associative genomics. A brief description of these subdisciplines might be useful in helping those new to the field to understand how modern genomics can affect gene conservation.

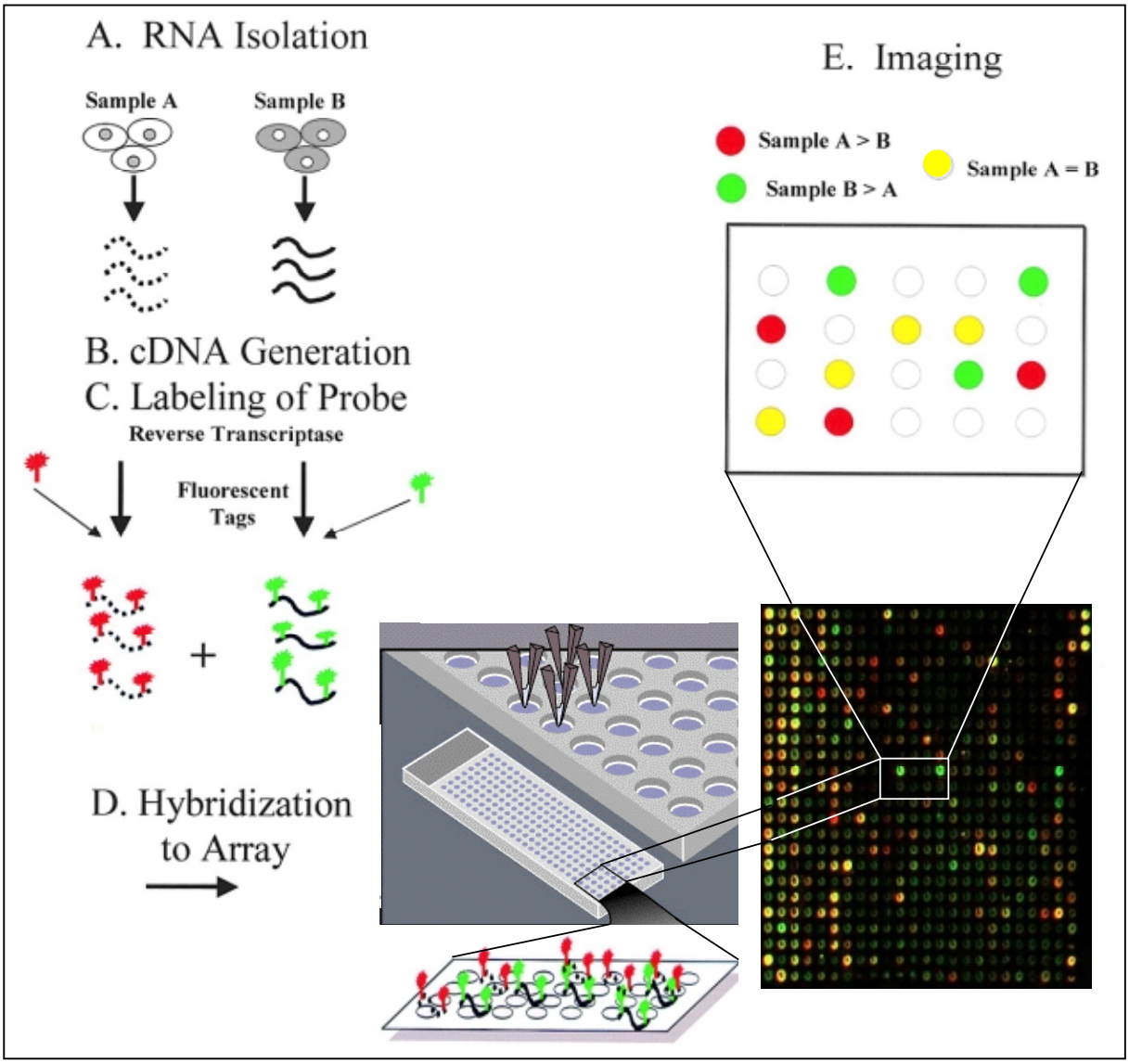
2.1.1. Structural genomics

Structural genomics attempts to identify all the genes in the genome, sometimes called gene discovery, and to determine their locations on the chromosomes. This goal is achieved by sequencing individual genes, gene segments or entire genomes. The individual genes are identified from the DNA sequence using sophisticated computer algorithms. The biochemical function of a gene is deduced via comparison of the DNA sequence with the sequences of genes of known function in the databases. When complete sequence of an entire genome is not available the location of genes can be determined either by direct physical mapping or by genetic mapping of the entire genome using numerous genetic markers. One of the most prominent applications of structural genomics for the study of adaptive genetic variation is quantitative trait loci (QTL) analysis via genome mapping. However, this approach aims to explain genomic structure and gene interaction at the genomic rather than functional level, unlike functional genomics.

2.1.2. Functional genomics

Functional genomics seeks to understand the function of genes and how they determine phenotypes. One of the major advances in functional genomics is using DNA microarrays (also known as “DNA chips”) to measure the specific expression of thousands of genes simultaneously. DNA microarray contains thousands of DNA samples or oligonucleotide sequences printed or synthesized onto nylon membrane filter or microscope glass slide in a precise and known pattern and representing thousands of genes in the genome respectively. Each DNA spot represents a unique gene that is used to quantitatively measure mRNA (messenger RNA) expression by hybridizing to fluorescent labelled mRNA (Figure 2). Adaptive response to different environmental stresses and treatments can be studied for many genes simultaneously or in parallel by analysis of differential responses of thousands of genes using DNA microarrays.

Figure 2: The use of DNA microarrays in differential gene expression analysis (adapted from Albelda and Sheppard 2000). Comparative hybridization experiment involves isolation of messenger RNA (mRNA) from two separate samples (A). The mRNA from each sample is treated with revers transcriptase (B) and labelled with a distinct fluorescent tag (C). The two pools of labelled RNA are mixed, hybridised to the DNA microarray containing a full set of thousands or tens of thousands of DNA sequences based on genome or complimentary DNA (cDNA) sequences, and washed (D). The microarray array is scanned using a specialised fluorimager, and the colour of each spot is determined (E). In this example, genes expressed only in Sample A would be red in colour, genes expressed only in Sample B would be green and those genes expressed equally in both samples would be yellow. This allows researchers to determine genes that are specifically expressed in response to the specific treatment or disease, or tissue-specific genes that are expressed in one tissue, but not in other.



Comparative genomics uses information from different species and assists in understanding gene organization and expression and evolutionary differences. It takes

advantage of the high level of gene conservatism[1] in structure and function (i.e., little variation across diverse taxa) and applies this principle in an interspecific manner in the search for functional genes and their genomic organization. Comparative genomic studies are also enhanced by examining a diversity of model organisms in which physiological, developmental or biochemical traits are readily studied.

DNA sequence comparison and comparative genetic mapping are the most often used methods in comparative genomics. It is likely that the study of adaptation in forest tree species will greatly benefit from comparative genomic studies using different models as well as other well-studied species. In particular, genomic studies in a small flowering Brassica plant, *Arabidopsis thaliana,* widely used as a model species, have already yielded complete genome sequence data. A complete genome sequence of poplar also should soon be available for comparative genomic analysis.

2.1.4. Associative genomics

Associative genomics searches for mutations in populations via linkage disequilibrium analysis and via direct assessment of association between alleles and phenotypes. This approach can be effectively used in the search for adaptive mutations such as disease resistance, drought tolerance, cold hardiness, etc. DNA variants or mutations (inherited differences in DNA sequence) can either directly contribute to phenotypic variation, influencing an individual's phenotypic characteristics (e.g., risk of disease and response to the environment), or can be tightly linked to the genes causing this variation. In the latter case, the alleles serve as markers of the selective genes and can be in linkage disequilibrium with alleles of this gene due to the limited population size, recent origin, low recombination rate and/or strong selection acting on alleles of the linked selective gene. Once candidate alleles responsible for adaptive traits are detected via QTL, candidate and comparative mapping, it will be possible to perform association studies to estimate effects of alleles or haplotypes[2] on phenotypes. It should be practical to define common haplotypes using a dense set of polymorphic markers, and to evaluate each haplotype for association with disease or any particular adaptive trait. Single nucleotide polymorphisms (SNPs) are the most appropriate markers to characterise haplotypes

[1] Unchanged gene location in chromosomes among closely related species

[2] A particular combination of alleles or sequence variations that are closely linked — that is, are likely to be inherited together — on the same chromosome.

and to achieve the required density of markers. Most sequence variation is attributable to SNPs, with the rest attributable to insertions or deletions of one or more bases, repeat length polymorphisms and rearrangements. SNPs occur (on average) every 1,000–2,000 bases when two individual sequences are compared, and are thus present at sufficient density for comprehensive haplotype analysis. SNPs are binary, and thus well suited to automated, efficient and fast genotyping. It is likely that soon a SNP map with sufficient density will be created for forest tree species, and will be used in the associative genomic study of adaptive variation. Such studies should help to find haplotypes and genetic variants that are either directly involved in the genetic control of adaptive traits or have non-random association with these traits due to a tight linkage and linkage disequilibrium.

2.1.5. Statistical genomics

Statistical genomics is an integrative sub-discipline and serves all other areas of genomics. It provides statistical tools for genome and QTL mapping in structural genomics, bioinformatics tools for gene search, comparison and annotation in functional genomics, and statistical population genetic methods in associative genomics. Statistical genomics is also very important in developing computerised comprehensive interactive biological databases. New computer tools are required to compose genetic data at all levels of biological organization—from gene to population, species and ecosystems—for multiple purposes, including gene conservation.

Certainly, the division of genomics into these subdisciplines is rather arbitrary. Often the distinctions are vague or overlapping, but may be useful in helping those new to the field to understand modern genomics. In fact, genomics is a synthetic discipline that combines many methods and approaches of molecular biology, population and evolutionary genetics and bioinformatics (Figure 1). The purpose of genomics is to study the structure, function and evolution of genome as a whole via complete genome sequencing, creating functional genetic maps for entire genomes and simultaneous analysis of patterns of differential expression of all or thousands of genes in the genome representing different cells and tissues and / or different treatments and conditions. It facilitates understanding genomes at both a molecular and a phenotypic level. It is likely that soon we will have a catalogue of all or most of genes expressed in plant and animal genomes and those that play essential roles in species- and population-level adaptation. Identifying and understanding the function of these genes, we

can associate genetic variation with phenotypes and study adaptive genetic variation in different populations.

2.2. DNA sequencing of entire genomes

Complete sequencing of genomes of several important and model species is a significant achievement of genomics, which provides the basis for comparative and functional analysis. Answers to questions such as (1) the number, location and distribution of genes in genome; (2) gene organization and their function; (3) what genes are the same or highly conserved across different species; and (4) what genes are responsible for species adaptation and evolution can now be obtained. Complete genome sequences have been determined for the yeast *Saccharomyces cerevisiae* (May 1997), the nematode *Caenorhabditis elegans* (December 1998), the fruit fly *Drosophila melanogaster* (March 2000), the annual plant arabidopsis (December 2000), the human (February 2001), and will also become available soon for mice, mouse, rice and maize. It now seems experimentally possible to determine the complete sequence of a forest tree genome such as Populus (500 million bp[3]) or *Eucalyptus* (340-580 million bp).

The number of genes in a genome is limited and turns out to be not as high as expected earlier (for instance, only ~26,000 in plants and animals vs. ~6,000 in baker's or budding yeast, Table 2). Moreover, many genes are common across different species and are practically unchanged from the distant evolutionary past. For instance, only 94 of 1278 protein families in the human genome appear to be specific to vertebrates. The most elementary of cellular functions — basic metabolism, transcription of DNA into RNA, translation of RNA into protein, DNA replication and the like — evolved just once and have remained almost unchanged since the evolution of single-celled yeast and bacteria.

[3] Nucleotide base pairs.

Table 2: Genome size of several species for comparison

Taxonomic rank	Latin name	Common name	Haploid chromosome	Nucleotide base pairs (x 10^6)	Genes (x 10^3)
		Prokaryotae			
Archae	12 [1]	archael microorganisms	-	1.6-3.0	1.5-2.7
Bacteria	40 [1]	bacterial microorganisms	-	0.6-7.0	0.5-6.6
Bacteria	*Escherichia coli* [2]	no common name	-	4.6	4.3
		Eukaryotae			
Yeast	*Saccharomyces cerevisiae* [2]	baker's or budding yeast	16	12	6
Worm	*Caenorhabditis elegans* [2]	nematode	5/6	97	19
Insect	*Drosophila melanogaster* [2]	fruit fly	4	180	13.6
Annual plant / Angiosperm	*Arabidopsis thaliana* [2]	arabidopsis	5	125	25.5
Annual plant/ Angiosperm	*Oryza sativa* [2]	rice	12	400	?
Annual plant/ Angiosperm	*Zea mays* [2]	maize	10	2,400-3,200	?
Perennial plant / Angiosperm	*Lycopersicon esculentum*	tomato	12	900	?
Forest tree / Angiosperm	*Eucalyptus* [3]	eucalypts	11	340-580	?
Forest tree / Angiosperm	*Populus* [3]	poplars	19	500	?
Forest tree / Gymnosperm	*Pinus* [3]	pines	12	20,000-30,000	?
Mammals / Rodent	*Mus musculus* [2]	mouse	20	3,500	21-30
Mammals / Primate	*Homo sapiens* [2]	human	23	3,400	26-31

[1] Number of species with completely sequenced genomes.

[2] Species with completely or almost completely sequenced genome.

[3] Data are based on several species.

Comparative genomics explores such gene conservatism, which helps to understand and infer the function of a particular gene from the data obtained for similar homologous genes studied in other organisms. Much about forest tree gene functions can be learned from the data obtained in other organisms, such as *Arabidopsis*. Complete genome sequences are not yet available for any forest tree species, although advances in sequencing technology should make it possible in the near future. Among various challenges are the complexity and large size of tree genomes (Table 2). The size of the pine genome (20,000-30,000 million bp), for example, is 6 to 8 times larger than the human genome (3,400 million bp), and 150 to 200 times larger than the genome of *Arabidopsis* (125 million bp). Even the relatively small physical size of the *Populus* genome (500 million bp), which is 40 times smaller than the best-studied conifer, *Pinus taeda*, and, therefore, will be likely the first forest tree genome to be entirely sequenced, is still about 4 times as large as *Arabidopsis* (although similar to rice and 6 times smaller than maize, both of which are almost completely sequenced).

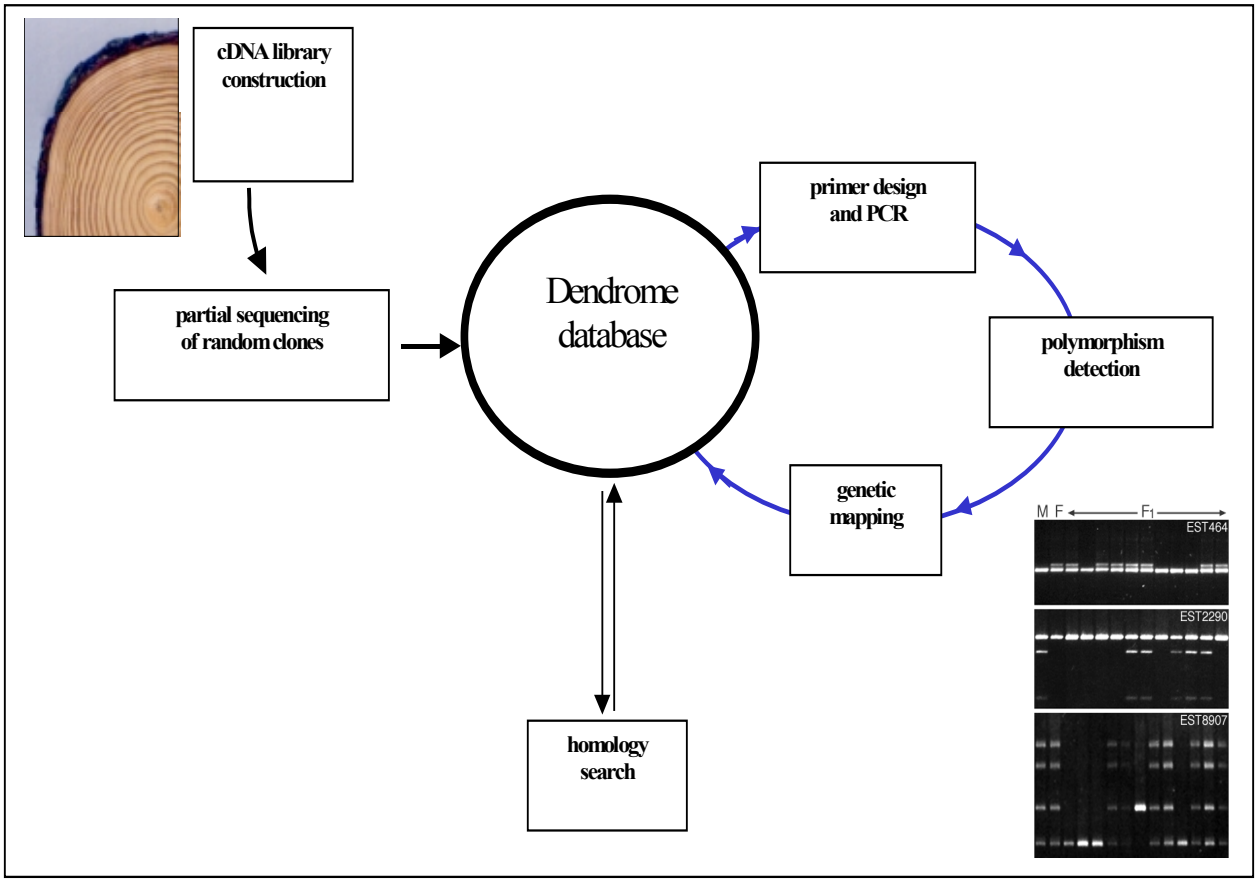
2.3. Gene discovery and expressed sequence tag polymorphisms (ESTPs)

An alternative to complete genome sequencing for discovering genes is being used in trees and many other organisms, which is based on identifying only the DNA that code for genes that are expressed in the genome. These sequences are called expressed sequence tags (ESTs). They are partial or complete sequences of complementary DNA (cDNA) obtained from mRNA isolated from different tissues and therefore represent genes expressed in these tissues with often known or suggested function (Figure 3). EST sequences are compared to all other sequences in gene databases to identify matches likely representing highly homologous genes. If there is a high similarity (homology) to some other gene sequence whose identity has been determined, then the identity of the EST can be immediately inferred. ESTs can be used as a source for identifying candidate genes for QTLs involved in genetic control of adaptive traits. Large libraries of partial or complete sequences of thousands of expressed genes have already been obtained, and large databases of EST sequences are available for many animal and plant species, including several forest tree species, such as Monterey or radiata pine (*Pinus radiata*), loblolly pine[4] (*P. taeda*), Norway spruce (*Picea abies*), *Eucalyptus* and *Populus*.

Expressed sequence tag polymorphisms (ESTPs) are derived from ESTs. Using EST sequences polymerase chain reaction (PCR) primers are designed to amplify ESTs from

individual genomic DNA (Harry *et al*. 1998). Allelic polymorphism in the amplification product (ESTPs) can be revealed using different modern methods for detection and visualisation of DNA alterations (Kristensen *et al.* 2001).

Figure 3: The development of expressed sequence tag (EST) markers in forest trees that can be used in comparative and candidate gene mapping. EST markers are derived from partial or complete sequences of complimentary DNA (cDNA) libraries that obtained from messenger RNA (mRNA) isolated from different tissues (for instance, xylem). EST sequences are submitted to gene databases and compared to all other sequences in the databases to identify matches likely representing highly homologous genes. Polymerase chain reaction (PCR) primers based on the EST sequences are designed to amplify these genes. If these genes are polymorphic and segregate in the experimental population or progeny, they can be used in the genome and quantitative trait loci (QTL) mapping. They are good candidate genes for QTLs.



ESTPs mostly reveal genetic variation within genes, although variation can be found in both coding and non-coding regions of genes. Thus, ESTPs are the most informative markers in terms of gene function among the most recently developed one and are the first genetic markers that offer real potential for detecting adaptive genetic diversity broadly.

2.4. Physical and genetic mapping of the whole genome using numerous genetic markers

Genetic linkage mapping is central to genomics. It allows the positioning of genes and genetic markers on a specific chromosome. There are two kinds of maps: physical and genetic. Physical maps provide the exact location of genes or genetic markers on chromosomes. These maps are either assembled from the complete genome sequences, BAC[5] contigs[6], or based on *in situ* hybridization or other methods. However, as long as the complete genome sequences of forest tree species are not available the alternative approach is to develop genetic linkage maps by segregation and linkage analysis. Genetic maps identify the distance and order between markers based on the number of recombination events between them. Genetic maps have been already constructed for many different forest tree species using a variety of genetic marker types (see Table 1; Neale and Sederoff 1996; Krutovskii *et al*. 1998 and Cervera *et al*. 2000 for review). A complete sequence alone is not sufficient to understand the genetic control of adaptive traits. These traits are usually very complex, have quantitative inheritance and are controlled by many genes each with relatively small effects, which are called quantitative trait loci (QTL). Genetic maps can be used to study the number, location and distribution of QTLs in a genome via their genetic linkage mapping with molecular markers. Following this approach, a new genomic technique called QTL mapping has been relatively recently developed.

[5] Bacterial artificial chromosome (BAC): A chromosome-like structure, constructed by genetic engineering that carries large segments of DNA—100000 to 200000 bases—from another species cloned into bacteria. Once the foreign DNA has been cloned into the host bacteria, many copies of it can be made.

[6] A group of clones representing overlapping regions of a genome.

2.5. Analysis of genetic control of complex adaptive traits via quantitative trait loci (QTL) mapping

The method for finding and locating QTLs is called QTL mapping. The conceptual basis of this method is comparatively simple but it requires relatively dense genetic maps with evenly distributed markers covering the entire genome, appropriate statistical tools, and an experimental population of sufficient size segregating for both genetic markers and phenotypic traits (e.g., Paterson, 1998). First, multi-locus genotypes (molecular markers) and phenotypes (quantitative traits) are measured on all individuals of segregating population. Then, phenotypic values are statistically associated with genotypic values, usually using multiple regression or maximum likelihood methods to identify markers that have a strong association (joint segregation) with the quantitative trait. Such association can be the result of either tight linkage of the genetic marker and QTL (i.e., because they reside in the same region of the chromosome) or direct involvement of this particular marker(s) in genetic control of the trait. QTLs have been already detected and mapped in forest trees for such adaptive traits as growth rhythm, phenology, form, wood quality, disease resistance, cold hardiness, drought tolerance, and others (see Neale 1998, Sewell and Neale 2000 and Neale *et al*. 2002 for review). Once a QTL controlling an adaptive trait has been precisely mapped, it then may become possible to clone the gene underlying the QTL based solely on the knowledge of its genetic map position and without knowing its function or DNA sequence. This is known as positional or map-based cloning.

Numerous recently developed PCR-based markers (e.g., RAPD, AFLP, SSR, STS, etc.) are used in QTL mapping (e.g., Sewell and Neale 2000 and Neale *et al*. 2002 for review). However, many of these markers are either dominant or anonymous, and their functions are unknown. There are three important aspects to consider when choosing a genetic marker system for QTL mapping: the outbred nature of forest tree pedigrees (1), the potential for comparative (2) and candidate gene (3) mapping. First, each parent of an outbreed pedigree is typically a different, highly heterozygous individual, where the transmission of up to four different alleles must be followed from the parents to progeny. Therefore, multiallelic codominant markers are best suited to detect the maximum number of polymorphisms found in the heterozygous parents. Second, comparative mapping, both within and among species, is an important tool for relating results from different mapping experiments. Therefore a subset

of the markers used in a mapping experiment should be orthologous[7] across pedigrees and species. Third, to identify actual genes controlling a quantitative trait, genes with known or suggested function should be used in QTL mapping. Complete or partial cDNA sequences (ESTs) allow now researchers to design ESTP markers that take into account all these aspects and can be used for genetic mapping of the entire genome and for measuring adaptive genetic diversity via QTL mapping analysis (e.g., Harry *et al.* 1998; Temesgen *et al*. 2001; Neale *et al*. 2002). These are the most informative markers for adaptive trait candidate gene mapping that is now used in animal and plant species, mostly agriculture stocks and crop species, to identify genes for different yield and quality traits including also adaptive traits such as biomass, growth rate, fecundity and other reproductive traits, disease resistance, etc.

2.6. Candidate gene mapping of adaptive genes

Candidate gene mapping is based on the assumption that a gene with known or assumed function that may affect genetic control of a trait can be considered a 'candidate gene' for this trait (e.g., Gion *et al*. 2000; Neale *et al*. 2002). Furthermore, it is assumed that if this gene is also mapped to the same genomic region as a QTL for this trait, then this gene is very likely to be this QTL that directly controls the trait, although the likelihood depends on marker density, precision of QTL map and genome size.

Large forest tree EST projects will identify and provide DNA sequences that give researchers enough material to develop genetic markers for an unlimited number of genes that can be used as a source of possible candidate genes to target particular adaptive traits (Temesgen *et al.* 2001; Neale *et al*. 2002). Different subsets of specific EST markers can be used in mapping adaptive gene. For instance, EST markers derived from genes that are supposedly related to the cell defence mechanism can be used to map QTLs controlling disease resistance; EST markers derived from genes that are involved in the wood formation can be efficiently used in QTL mapping of wood related traits, etc. If function of genes used to derive ESTs is unknown but they represent cDNA isolated from a specific tissue or obtained from the cells that undergone a specific treatment, they still can be used as candidate genes in QTL mapping. For instance, heat shock genes expressed during experimental heat stress can be used to map genes related to drought resistance. The use of such meaningful

[7] Loci in two species that have arisen from the same locus of their common ancestor.

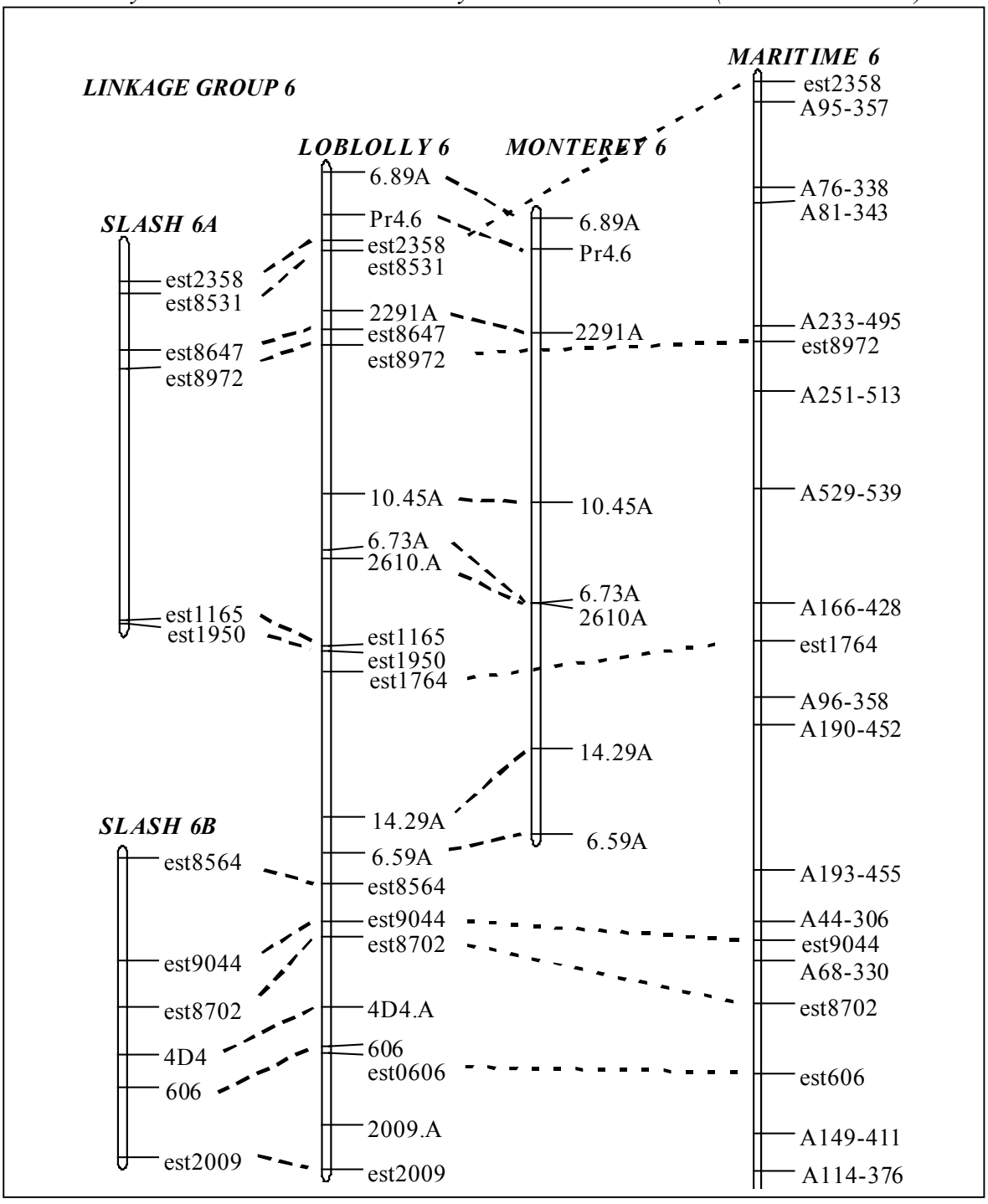
markers as ESTs directly in genetic mapping makes analysis of adaptive variation more efficient and focused. In addition, highly efficient and sensitive methods are now being developed to detect allelic differences between these genes that can be used for mapping (e.g., SNP detection).

Identifying candidate genes for QTLs controlling adaptive traits in trees would ultimately provide the diagnostic tools to screen large amounts of wild germplasm for individuals carrying alleles worthy of conserving. The challenge is to identify DNA polymorphisms within candidate genes that will distinguish alleles and then associate alleles with differences among phenotypes. This can be accomplished through SNP discovery and association studies. The approach is to identify SNPs within regions of candidate genes involved in the control of a trait, to genotype a large number of individuals from the natural population at these SNPs, and to test for associations between SNPs and phenotypes. This approach will soon be available for application in forest tree conservation programs because of the intensity and progress of research and development activities.

2.7. Comparative mapping of adaptive genes

Comparative mapping is one aspect of comparative genomics and another very promising genomic approach for discovering adaptive genes. It takes advantage of high similarity in gene location in chromosomes of closely related species and applies it across different species to search for functional genes and their genomic organization. Comparative mapping in various species has shown that gene content and gene order are conserved over long chromosomal regions among related species of animals or angiosperm plants. These results strongly suggest that similar studies can be effectively done in the forest trees. The genetic maps of closely related species can be directly compared due to synteny (i.e., co-occurrence of two or more genes on the same chromosome) among the genomes of these species. Indeed, the high levels of co-linearity among, for instance, pine species (e.g., Brown *et al.* 2001) means that genetic information from one species can be applied to others (Figure 4).

Figure 4: Comparative genetic linkage maps of linkage group 6 of loblolly, slash, Monterey and maritime pines aligned using common expressed sequence tag (EST) markers and illustrating the potential utility of loblolly pine ESTs as anchored reference loci. Loci connected by a dotted line were detected by the same EST marker (Brown et al. 2001).



The most valuable alleles of adaptive genes can be identified from the pool of all species and possibly incorporated into breeding and conservation strategies. Furthermore, the controls and interactions affecting adaptive trait expression can be studied. Further studies

should show whether comparative mapping between distantly related forest trees, for example between *Populus* or *Eucalyptus* and *Arabidopsis*, is also possible.

The development of genetic resources for comparative genomic analysis in forest trees would have significant impacts in many areas of forest gene conservation research. Comparative mapping would facilitate: (1) verification of QTLs controlling adaptive traits, (2) identification of candidate genes and (3) the understanding of evolutionary relationships. The emphasis in forest gene conservation is not on a single species, but on many, each with its own regional economic and ecological distinctions. Comparative genetic mapping in pines and other conifers follows this paradigm, focusing not only on the creation of individual species maps but also on the consensus maps to identify the genomic locations of genes affecting quantitatively inherited adaptive phenotypes, resistance to pathogens, and other biological and physiological characteristics.

Comparative mapping is possible if orthologous[8] genetic markers have been mapped to each of the species maps to be 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.

Most of the anonymous markers (e.g., RAPD, AFLP, and SSR) cannot be used for comparative mapping because they are not orthologus among species. Genetic markers that are based on genic DNA sequences, such as RFLPs and ESTPs, are more suited for comparative mapping. For example, RFLP loci from both *Pinus taeda* and *P. radiata* have been used to construct comparative maps between these species (Devey *et al.* 1999). However, because RFLP markers do not easily distinguish between orthologs and paralogs and because they are difficult to apply, they are unlikely to be used widely for comparative mapping. ESTPs are the most useful markers for comparative mapping and have been already used in genetic mapping in conifers (Tsumura *et al*. 1997; Perry and Bousquet 1998; Cato *et al.* 2001; Temesgen *et al.* 2001). ESTPs reveal orthologs among species and only occasionally paralogs. ESTP markers from *Pinus taeda* have been used to construct comparative maps for this species and slash pine, *Pinus elliottii* (Brown *et al.* 2001).

3. BIOINFORMATICS AND GENOMIC DATABASES

The highly efficient, fast and productive technologies of genomic studies enable the collection of overwhelming amounts of data. The primary genomic data types are DNA and protein sequences, genetic mapping data and data resulting from functional analysis. Much of the data are freely available to the public via the Internet and World-Wide-Web (WWW). Everybody benefits from public access to the genomic databases, but especially researchers with a small research budget who can still do efficient computer data analysis and gene discovery. The National Centre for Biotechnology Information (NCBI) in the USA and European Molecular Biology Laboratory (EMBL) in Europe are the primary sites for DNA sequence databases and DNA sequence analysis tools. The primary databases are called GenBank and EMBL. They also provide on-line access to the BLAST (Basic Local Alignment Search Tool) programs, which are the primary tools used to search the databases and identify matches among sequences. 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 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 adaptive phenotype.

4. CONCLUSIONS

The study of adaptation is fundamental to forestry and forest genetic conservation. Forest geneticists have long used common-garden experiments and, to a lesser extent, molecular markers to study patterns of adaptation in forest trees. Phenotypic assessments are time consuming and expensive, and provide no information about variation in the genes controlling adaptive variations. There are numerous molecular marker technologies available, but most measure either neutral or highly conservative genetic variation of limited adaptive value. There is a need for developing rapid and informative diagnostic techniques for evaluating large numbers of adaptive genes and prospective trees for *in situ* conservation. Genomics provides new tools to study adaptation in trees. Forest geneticists can use

[8] Similarity in DNA or protein sequences between different species due to common ancestry. Describes the evolutionary origin of a locus. Loci in two species are said to be orthologous when they have arisen from the

automated, highly efficient, fast and productive technologies to determine DNA sequences and to genotype large numbers of individuals. They can ultimately identify genes responsible for forest tree adaptation via EST sequencing, QTL and candidate gene mapping. Then, using modern genotyping technologies and association studies they can determine allelic diversity for these candidate genes in forest tree populations and directly measure adaptive allelic diversity for thousands of genes simultaneously.

Despite remarkable progress much work remains to be done to understand the nature of genetic variation that underlies adaptive forest tree phenotypes. Comprehensive understanding will first require discovering, annotating and cataloging all genes in the forest tree genome. One approach towards achieving this goal is to determine the DNA sequence of the entire genome and infer the genes from the DNA sequence. This approach is currently not feasible in all forest trees because of their large genome size, but *Populus*—with a relatively small genome size—can serve as a model species. An alternative (or parallel) approach is to determine the DNA sequences for the gene-coding regions only. This can be accomplished by isolating mRNA, preparing cDNA libraries from this mRNA and sequencing cDNA. These EST sequences are submitted to databases and compared to all other sequences in the databases to see if they match to genes whose function has been determined. EST databases of tens of thousands of ESTs have been already produced and are publicly available for *Pinus*, *Picea, Populus*, and *Eucalyptus*.

The second step towards understanding adaptation involves construction of genome, QTL, comparative and consensus linkage maps for most forest tree species (e.g., Sewell *et al.* 1999). Genetic maps show the position of genes and are valuable for understanding genome organization and evolution. Maps are extremely useful tools for identifying genes controlling interesting phenotypes. Loci controlling quantitatively inherited traits, so-called QTLs, have been already identified in many forest trees for a variety of growth, wood quality, and other economic and adaptive traits. These data are immediately useful for tree improvement and gene conservation.

Next, DNA microarray analysis can be used to study the expression patterns of genes, and to understand the function of all genes and their interactions. The relationship between

---

same locus of their common ancestor.

the vast amount of allelic diversity in genes and the array of different phenotypes found in forest tree populations can be studied. A catalogue of common coding–sequence variants in forest tree genes can be created and tested for association with a phenotype. Genome-wide high-resolution maps of known polymorphisms can be used to scan the genome for marker-adaptive trait associations.

The analysis need not be limited to coding sequences. It may be that the majority of relevant mutations reside in regulatory regions. Thus, it is important to identify variants in at least the proximal and distal regulatory sequences as our poor understanding of 'regulatory' elements dictates the need for a more global approach. An approach in which marker-trait associations are sought will require the construction of a high-resolution map of genetic variants. SNPs are the natural candidates for this map because they are abundant, have a smaller mutation rate than microsatellites and can be genotyped en masse using microarray technology.

A map-based association search for multiple adaptive loci, each contributing to the total phenotype in a small yet measurable way, is feasible via haplotype analysis. The alleles of these loci can be indirectly recognized by their historical associations with other genetic variants (e.g., SNPs) in their neighborhood. The non-random association of variants with one another (linkage disequilibrium) is a well-known feature of the plant and animal genomes. DNA microarrays will have a major role in genotyping thousands of genes simultaneously, in the creation of fine maps and in mapping the components of complex adaptive phenotypes. Forest genomics has a bright future and awaits exiting applications in forest tree management and gene conservation.

ACKNOWLEDGEMENTS

We would like to thank Garth Brown (Institute of Forest Genetics) and Deborah Rogers (University of California at Davis) for providing useful comments on the manuscript, and Christel Palmberg-Lerche (Forest Resources Development Service, FAO) for encouraging the writing of this review.

REFERENCES

Albelda, S.M. and D. Sheppard (2000) Functional genomics and expression profiling: be there or be square. American Journal of Respiratory Cell and Molecular Biology 23:265-269.

Boshier, D.H and A.G. Young (2000) Forest conservation genetics: limitations and future directions. In: Forest conservation genetics: Principles and practice (A. Young, D. Boshier and T. Boyle, eds), pp. 289-297. CABI Publishing, United Kingdom.

Brown, G.R., Kadel, E.E., Bassoni, D.A., Temesgen, B., van Buijtenen, J.P., Marshall, K.A., and D.B. Neale (2001) Anchored reference loci in loblolly pine (*Pinus taeda* L.) for integrating pine genomics. Genetics (in press)

Cato, S.A., R.C. Gardner, J. Kent, and T.E. Richardson (2001) A rapid PCR-based method for genetically mapping ESTs. Theoretical and Applied Genetics 102:296-306.

Cervera, M.T., Plomion, C. and C. Malpica (2000) Molecular markers and genome mapping in woody plants. In: Molecular biology of woody plants. Forestry Sciences, Volume 64 (S. M. Jain and S. C. Minocha, eds), pp. 375-394. Kluwer Academic Publishers, The Netherlands.

Gion, J.-M., Rech, Ph., Grima-Pettenati, J., Verhaegen, D., and C. Plomion. (2000) Mapping candidate genes in *Eucalyptus* with emphasis on lignification genes. Molecular Breeding 6: 441–449.

Glaubitz, J. and G.F. Moran (2000) Genetic tools: the use of biochemical and molecular markers. In: Forest conservation genetics: Principles and practice (Young, A., D. Boshier and T. Boyle, eds), pp. 39-59. CABI Publishing, United Kingdom.

Harry, D.E., B. Temesgen and D.B. Neale (1998) Codominant PCR-based markers for *Pinus taeda* developed from mapped cDNA clones. Theoretical and Applied Genetics 97:327-336.

Human genome project: Nature 409 (6822), February 15, 2001 and Science 291 (5507), February 16, 2001.

Jain, S.M. and S.C. Minocha (eds) (2000) Molecular Biology of Woody Plants. Forestry Sciences, Volume 64. Kluwer Academic Publishers, The Netherlands.

Kristensen, V.N., D. Kelefiotis, T. Kristensen and A.-L. Børresen-Dale (2001) High-throughput methods for detection of genetic variation. BioTechniques 30:318-332.

Krutovskii, K.V., S.S. Vollmer, F.C. Sorensen, W.T. Adams, S.J. Knapp, and S.H. Strauss (1998) RAPD genome maps of Douglas-fir. J. Heredity 89:197-205.

Linhart, Y. B. (2000) Variation in woody plants: molecular markers, evolutionary processes and conservation biology. In: Molecular biology of woody plants. Forestry Sciences, Volume 64 (S. M. Jain and S. C. Minocha, eds), pp. 341-374. Kluwer Academic Publishers, The Netherlands.

Liu, B.-H. (1998) Statistical Genomics: linkage, mapping and QTL analysis. CRC Press, New York.

Mandal, A.K. and G.L. Gibson (eds.) (1998) Forest genetics and tree breeding. CBS Publishers, New Delhi, India.

Neale, D.B. (1998) Molecular genetic approaches to measuring and conserving adaptive genetic diversity. In: The proceedings of international symposium on *In Situ* Conservation of Plant Genetic Diversity, pp. 385-390. November 4-8, 1996, Antalya, Turkey, CRIFC, Turkey.

Neale, D.B. and R.R. Sederoff (1996) Genome mapping in gymnosperms: A case study in loblolly pine (*Pinus taeda* L.). In: Genome mapping in plants (A. H. Paterson, ed.), pp. 407-423. Academic Press, New York.

Neale, D.B., M.M. Sewell, and G.R. Brown (2002) Molecular dissection of the quantitative inheritance of wood property traits in loblolly pine. Annals of Forest Science (in press)

Paterson, A. H. (ed.) (1998) Molecular dissection of complex traits. CRC Press, New York.

Perry, D.J. and J. Bousquet. (1998) Sequence-tagged-site (STS) markers of arbitrary genes: development, characterization and analysis of linkage in black spruce. Genetics 149: 1089-1098

Savolainen, O. and A. Karhu (2000) Assessment of biodiversity with molecular tools in forest trees. In: Molecular biology of woody plants. Forestry Sciences, Volume 64 (S. M. Jain and S. C. Minocha, eds), pp. 395-406. Kluwer Academic Publishers, The Netherlands.

Sewell, M.M. and D.B. Neale (2000) Mapping quantitative traits in forest trees. In: Molecular biology of woody plants. Forestry Sciences, Volume 64 (S. M. Jain and S. C. Minocha, eds), pp. 407-423. Kluwer Academic Publishers, The Netherlands.

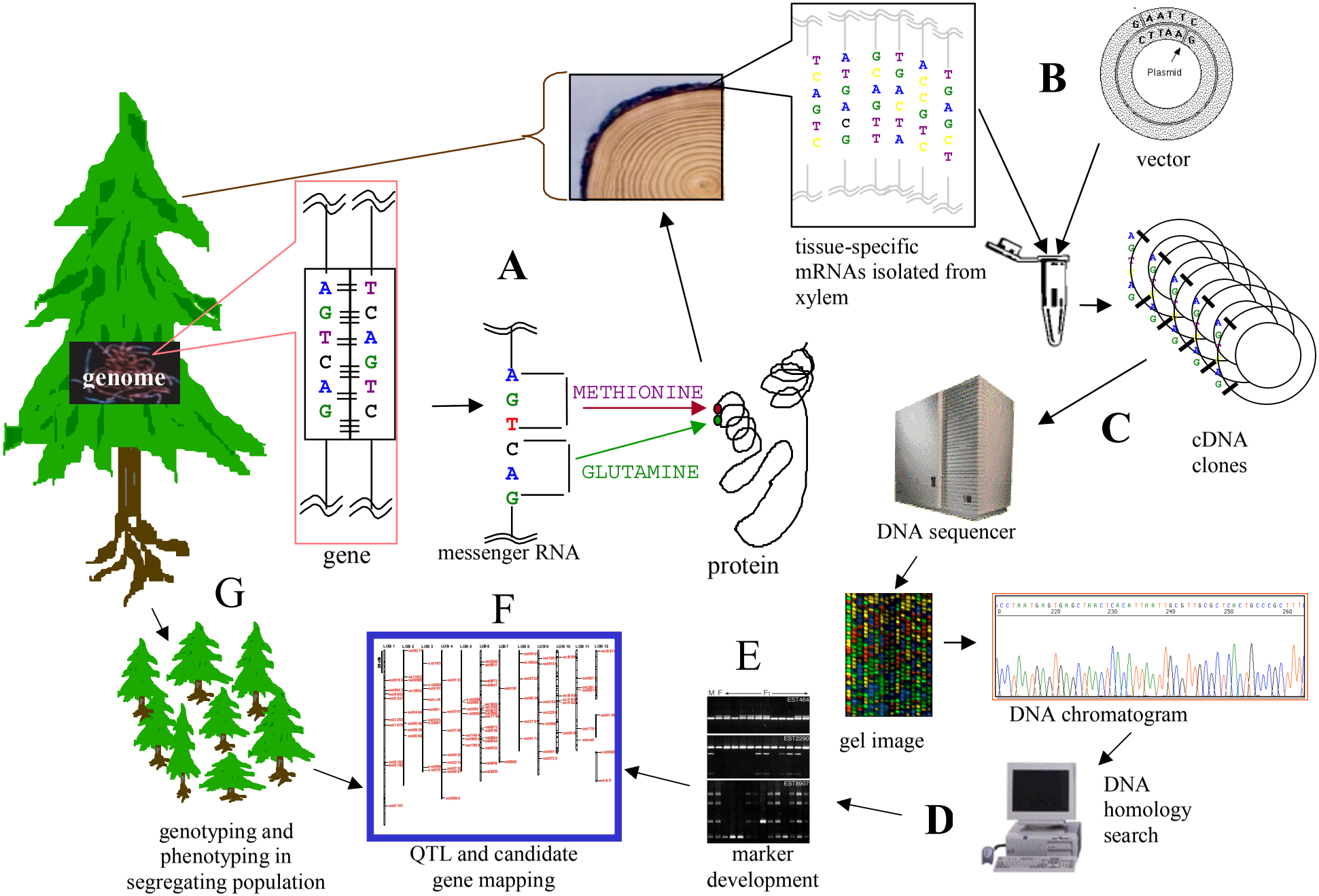
Temesgen, B., G.R. Brown, D.E. Harry, C.S. Kinlaw and D.B. Neale (2001) Genetic mapping of expressed sequence tag polymorphism (ESTP) markers in *Pinus taeda* L.).

Tsumura, Y., Y. Suyama, K. Yoshimura, N. Shirato and Y. Mukai (1997) Sequence-tagged-sites (STSs) of cDNA clones in *Cryptomeria japonica* and their evaluation as molecular markers in conifers. Theoretical and Applied Genetics 94:764-772.

Young, A., D. Boshier and T. Boyle (eds) (2000) Forest Conservation Genetics: Principles and Practice. CABI Publishing, United Kingdom.

Annex 1: Discovery of adaptive genes. ***A****: DNA in genes encode different proteins, for example, those with specific cellular functions related to the growth, wood quality and other adaptive traits;* ***B****: Sets of these genes are expressed according to unique patterns in time and location via messenger RNA (mRNA) that can be isolated, for instance, from xylem tissue and captured in vitro as complimentary DNA (cDNA);* ***C****: cDNA are sequenced, and* ***D****: cDNA are compared to database sequences to suggest gene functions. Large-scale partial cDNA sequencing can identify many genes within a genome. Specific sets of genes expressed to produce specific structures (e.g., wood) or specific physiological responses (e.g., disease resistance) can be identified;* ***E****: expressed sequence tag (EST) genetic markers can be developed from these gene sequences;* ***F****: ESTs can be genotyped and mapped in the experimental population or progeny segregating for both genetic marker and adaptive trait (****G****).*

ANNEX 1


genome
gene
A
messenger RNA
METHIONINE
GLUTAMINE
protein
tissue-specific
mRNAs isolated from
xylem
B
Plasmid
vector
C
cDNA
clones
DNA sequencer
gel image
DNA chromatogram
D
DNA
homology
search
E
marker
development
F
QTL and candidate
gene mapping
G
genotyping and
phenotyping in
segregating population