Single Nucleotide Polymorphisms (SNPs) Discovery and Linkage Disequilibrium (LD) in Forest Trees

Zhang De-qiang[1,2] *Zhang Zhi-yi*[1*]

[1]Key Laboratory of Genetics and Breeding in Forest Trees and Ornamental Plants, Ministry of Education, Beijing Forestry University, Beijing 100083, P. R. China

[2]Laboratory of Biotechnology, Institute of Forestry, Chinese Academy of Forestry, Beijing 100091, P. R. China

ABSTRACT With completion of the *Populus* genome sequencing project and the availability of many expressed sequence tags (ESTs) databases in forest trees, attention is now rapidly shifting towards the study of individual genetic variation in natural populations. The most abundant form of genetic variation in many eukaryotic species is represented by single nucleotide polymorphisms (SNPs), which can account for heritable inter-individual differences in complex phenotypes. Unlike humans, the linkage disequilibrium (LD) rapidly decays within candidate genes in forest trees. Thus, SNPs-based candidate gene association studies are considered to be a most effective approach to dissect the complex quantitative traits in forest trees. The present study demonstrates that LD mapping can be used to identify alleles associated with quantitative traits and suggests that this new approach could be particularly useful for performing breeding programs in forest trees. In this review, we will describe the fundamentals, patterns of SNPs distribution and frequency, summarize recent advances in SNPs discovery and LD and comment on the application of LD in the dissection of complex quantitative traits in forest tress. We also put forward the outlook for future SNPs-based association analysis of quantitative traits in forest trees.

KEY WORDS single nucleotide polymorphisms (SNPs), linkage disequilibrium (LD), quantitative traits, association studies, forest tree

[Supported by the National Natural Science Foundation of China (Grant No. 30471409) and the National "863" Project (Grant No. 2002AA241071)]

1 Introduction

As more genomes, including those of humans, *Arabidopsis*, rice and *Populus*, are completely sequenced, and more expressed sequence tags (ESTs) databases are accessed, attention is now rapidly shifting towards the study of individual genetic variation in natural populations. The most common form of genetic variation in many eukaryotic species is represented by single nucleotide polymorphisms (SNPs, fondly known as "snips"). SNPs are single base variations at a unique physical location within a locus among different individuals. SNPs have been well characterized since the beginning of DNA sequencing in the 1980s, but to discover on a large scale and genotype them rapidly in large numbers of samples was not possible until the application of faster sequencing methods and high-throughput genotyping techniques (Sanger *et al*. 1977, Botstein *et al*. 1980, Brookes 1999, Chan 2005, Hinds *et al*. 2005). It has been estimated that the human genome contains more than 10 million nucleotide positions that have common SNPs between individuals in a population (Wang *et al.* 1998, Kruglyak and Nickerson 2001). SNPs can potentially alter protein functions and mRNA structural folds if they occur in the coding region and cause a qualitative difference (Shen *et al*. 1999, Schaeffer *et al*. 2001). Alternatively, alterations in noncoding DNA sequences, including regulatory domains and intron regions, can affect the level of gene expression, RNA stability and result in quantitative variants. Both types of genetic variation may produce functional changes and affect the phenotype of an organism (Guo *et al*. 2004, Olivier 2004). SNPs are therefore rapidly usurping simple sequence repeats (SSRs or microsatellites) and other classes of DNA markers in modern genetics research, due to that they have inherent characteristics, such as an abundant, mutational stable and biallelic nature and are amenable to high-throughput genotyping.

These desirable characteristics SNPs owned suggest that they should have an important application in a diversity of fields for genetics research. Inarguably, the most promising application of SNPs will be in linkage disequilibrium (LD) mapping or association mapping based studies to identify individual genes or genomic regions for their association with complex quantitative traits. Quantitative traits, as is generally

*Author for correspondence. E-mail: zhangzy@bjfu.edu.cn, Tel: +86-10-62338502

known, are those that show a continuous variation among individuals in all organisms. Most economically important traits, such as biomass production, wood quality, biotic and abiotic stress responses are complex quantitative traits in forest trees. Many of these traits are controlled by multiple genes interacting with each other and with the environment. Each gene is assumed to have a relatively small effect, although "major genes" that control a greater portion of the variation are known to exist for some traits (Bradshaw and Stettler 1995). Genes that have a measurable effect on quantitative traits are called quantitative trait loci (QTLs). During the past two decades, QTLs mapping has been used as the key tool for identifying the genetic basis underlying these quantitative traits and contributed greatly to our understanding of complex trait architecture, for example, number, magnitude of effect, and mode of action of QTLs (Sewell and Neale 2000). However, unlike crop plant species such as rice and wheat, most trees have the characteristics of defoliation, long generation intervals and hence it is very difficult to construct near-isogenic lines large enough for high-resolution detection of QTLs. Furthermore, QTLs mapping uses only the recombinations found in the progeny of pedigrees, which are typically two to three generations. Therefore, linkage analysis can only identify the chromosomal regions closely associated with a quantitative trait and this genetic interval of 5 cM typically corresponds to 1 Mb of DNA or more in forest trees. QTLs in this situation cannot be efficiently applied to marker-assisted selection (MAS) to enhance genetic gain or to perform the positional cloning approach. However, LD approaches which emerged recently as a powerful tool to detect associations between SNPs within candidate genes and complex phenotype traits, will overcome the above impasse (Jorde 2000, Plomion *et al*. 2003).

LD mapping has, of course, already been routinely used in human genetics to test individual genes or genomic regions associated with disease phenotype, such as obesity, diabetes, hypertension, or cardiovascular disorders (Kruglyak 1999, Suha and Vijg 2005). In general, studies indicate that LD extends over large distances ranging from 5 to 500 kb in human, making genome wide LD mapping feasible, whereas there is a large variation in the extension of LD in plant species (Reich *et al*. 2001, Flint-Garcia *et al*. 2003). In selfing species such as *Arabidopsis* LD extends up to 250 kb (Nordborg *et al*. 2002) and up to 10 cM in barley (Kraakman *et al*. 2004). Conversely, in maize, an outcrossing species, LD will be decayed within one or two kb (Remington *et al*. 2001). Only in recent years, the studies of LD have been extended to forest tree species, such as some commercially important tree species like pine, poplar and eucalyptus. These studies showed that LD declines rapidly within one to several kb (Brown *et al*. 2004, Neale and Savolainen 2004, Ingvarsson 2005, Thumma *et al.* personal communication). When LD declines rapidly with distance, LD mapping is potentially very precise (Gaut and Long 2003). It suggests that in outcrossing species genome wide LD mapping may not be feasible and not necessary, but candidate genes based LD mapping could be particularly useful in breeding programs of forest trees.

In this review, we will describe the fundamentals, nature and frequency of SNPs, summarize recent advances in SNPs diversity and LD, and comment on the application of LD in the dissection of complex quantitative traits in forest trees. We also put forward the outlook for future SNPs-based association analysis of quantitative traits in forest trees.

2 Single nucleotide polymorphisms (SNPs) discovery

2.1 SNP fundamentals

The DNA molecule consists of long strands of sugar and phosphate connected by base pairs. There are four types of bases in DNA: A (adenine), T (thymine), G (guanine), and C (cytosine) and they are weakly linked in four types of pairs: A-T, G-C, T-A or C-G. Their sequence along the DNA double helix determines the structure and function of a gene. A SNP is a tiny variation in an individual's genetic code. SNPs occur when a single nucleotide, A, T, C or G in the DNA sequence in a region of the genome is altered. They are the result of mutations occurring along the branches of the genealogical tree relating the homologous copies of a particular site in the genome. A variation must occur in at least 1% of the population to be considered a SNP. Thus, SNPs also called point mutations.

Theoretically, a SNP within a locus can produce as many as four alleles, each containing one of four bases at the SNP site, e.g. A, G, C or T. There are therefore six distinct types of single base substitutions (C⇔T, G⇔A, C⇔A, G⇔T, C⇔G, G⇔C, T⇔A and A⇔T), based on double strand DNA in one bi-allelic locus. Because C⇔T (G⇔A) is an identical 'mirror image' or sequence complement of G⇔A (C⇔T), in fact only four-way replacement is valid if one considers each DNA strand to be equivalent. Nucleotide changes between purine and pyrimidine bases are called transversions, while same class changes (purine to purine

or pyrimidine to pyrimidine) are called transitions (Fig. 1). It is expected that the frequencies of these four basic SNP types should have been equal if the mutation is random in the genome. However, the occurrence probability of base substitution types is not identical in human genome, with most SNPs (about 2/3) involving the transitions (C⇔T or G⇔A) variety, while the other three transversions types occur at similar levels to each other to comprise the remainder. The real reason for this phenomena remains unclear, but one probable explanation for this bias is the highly spontaneous rate of deamination of 5-methyl cytosine (5mC) to thymidine in the CpG dinucleotide, leading to the generation of higher levels of C⇔T SNPs, seen as G⇔A SNPs on the reverse strand (Cooper and Youssoufian 1988, Cooper and Krawczak 1990).

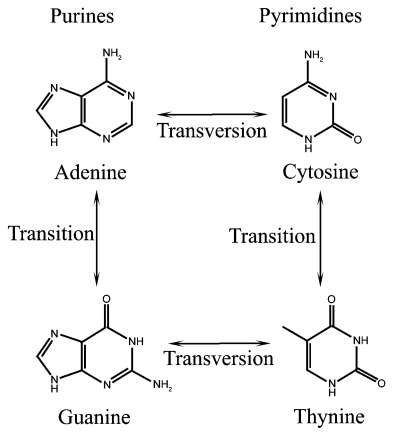

FIGURE 1 Nucleotide changes in DNA molecule

SNPs can occur in both coding (gene) and noncoding regions of the genome. In coding region, SNPs have been divided into two types, i.e. synonymous and nonsynonymous mutations. Substitutions that result in amino acid replacements are said to be nonsynonymous while substitutions that do not cause an amino acid replacement are said to be synonymous ones. The effect of SNPs on phenotype traits is variable and elusive, but is clearly dependent upon the location of SNPs in the genome. Although the majority of SNPs are likely to occur outside of actual gene encoding regions, polymorphisms located within the context of a gene need not be involved in protein encoding to result in a functional change. SNPs that occur in regions upstream of the protein-encoding gene regions might influence the binding of promoters or repressors, resulting in differential regulation of transcription (Carlson *et al*. 2005). Polymorphisms at intron/exon boundaries may affect exonic or intronic splicing enhancer or silencer positions, or especially conserved GT donor or AC acceptor positions, modifying the resulting polypeptide (Fairbrother *et al*. 2002, Maniatis and Tasic 2002). There is even demonstrated potential for phenotypic effects from non-coding or synonymous SNPs through alteration of RNA secondary structure (Shen *et al*. 1999). Similarly, untranslated distal 3′ differences may have additional effects, including interruption of poly-adenylation, which would alter the effectiveness of the template.

2.2 SNP discovery

SNP discovery is the process of finding the polymorphic sites in the genome of the species and populations of interest. Clearly, identification of large-scale SNPs are prerequisite before we begin to use them extensively as genetic tools. Currently, several approaches have been used for SNP discovery, including single strand conformation polymorphism (SSCP) analysis (Orita *et al*. 1989), heteroduplex analysis by denaturing high performance liquid chromatography (DHPLC) (Lichten and Fox 1983), direct DNA sequencing (Harding *et al*. 1997) and eSNP discovery in ESTs libraries (Garg *et al*. 1999). Below we will briefly describe and discuss the principles of each method listed in Table 1 and simply evaluate their accuracy, cost and throughput for SNP discovery in humans, plants and forest trees.

TABLE 1 Common methods for SNP discovery

Methods	Accuracy	Cost	Throughput	Reference
SSCP	Low	Low	Low	Orita *et al*. 1989
DHPLC	Medium	Low	Medium	Choy *et al*. 1999
DNA Sequencing	High	High	High	Chan 2005
eSNP	High	Low	High	Garg *et al*. 1999

2.2.1 SSCP

SSCP analysis, first described by Orita *et al*. (1989), has been used extensively in human and animal genetics to detect mutations within genes (Hayashi 1992, Barendse and Armitage 2001). For SSCP detection, the DNA fragment spanning the putative SNP is PCR amplified, denatured and run on a non-denaturing polyacrylamide gel. During the gel run, the single-stranded fragments adopt a secondary structure according to sequence. Fragments bearing SNPs are identified as a result of their aberrant migration pattern and confirmed by sequencing. Although a widely used and relatively simple technique, SSCP has a variable success rate for SNP detection, typically ranging from 70% to 95%. In addition, it is labour intensive and has relatively low throughput, although higher capacity methods are under development using capillary- rather

than gel-based detection system (Wenz *et al*. 1999).

2.2.2 DHPLC

DHPLC method, originally described by Oefner and Underhill (1995), has rapidly become popular in human SNP detection. DHPLC uses ion-pair reverse phase liquid chromatography to detect DNA heteroduplexes. Under partially denaturing conditions within a linear acetonitrile gradient, heteroduplexes denature more readily and display reduced column retention time relative to their fully complementary homoduplex counterparts, which are detected as new chromatographic peaks in less retention time. In comparison with the SSCP method, DHPLC offers the major advantage of being an automated hands-free and requiring no post-PCR sample processing. Moreover, the ion pairing agent, i.e. triethylammonium acetate, compresses the melting range of the amplicons, reducing the need for GC-clamped PCR primers. Therefore HPLC has rapidly become a popular method for heteroduplex-based SNP detection with simplicity, low cost and a high rate of detection (Choy *et al*. 1999). Analysis time is rapid, ranging from 6 to 10 min per sample for fragment sizes from 200 to 700 bp. However, with present instruments, only one sample can be analyzed at a time and parallel analysis is not yet possible. In view of this low throughput DHPLC is not very suitable for large-scale testing. It is, however, used in several clinical diagnostic laboratories with a moderate sample throughput load.

2.2.3 DNA sequencing

Direct DNA sequencing is currently the most accurate and most-used and the most direct approach for SNPs discovery with high-throughput in human, animals, plants and forest trees (Freudenberg-Hua *et al*. 2003, Schmid *et al*. 2003, Brown *et al*. 2004, Chan 2005, Kononoff *et al*. 2005). For this approach, PCR primers are designed to amplify 1 000–2 000 bp segments, based on the target gene for several unrelated individuals in a species. Once the sequencing reactions have been completed, the PCR products are sequenced directly in both directions with the Applied Biosystems 3 700 capillary system. In this way more than 1 500 DNA fragments of about 1 000 bp in 48 h can be produced with minimal human intervention. The resulting sequences are assembled and aligned and ultimately identify the true SNPs in the target gene. This method has been extensively used to exploit SNPs with candidate genes in several economically important forest trees, such as *Populus tremula*, *Eucalyptus nitens*, *Pinus taeda*, *P. radiata*, *P. pinaster*, *P. sylvestris* and *Cryptomeria japonica* (Table 2).

2.2.4 eSNPs discovery

ESTs are single-pass sequences generated from random sequencing of cDNA clones. Large-scale analysis of ESTs has resulted in a useful gene expression resource of both known and unknown genes. It offers a rapid and valuable first look at genes expressed in specific tissue types, under specific physiological conditions, or during specific developmental stages. At present, ESTs have been extensively used for SNP detection in a diversity of organisms and proved to be a valuable tool to discover SNPs of candidate genes in humans, animals, plants and forest trees (Picoult-Newberg *et al*. 1999, Schmid *et al*. 2003, Dantec *et al*. 2004, Zimdahl *et al*. 2004). Mining SNPs from EST sequences data generally involves the following steps: treatment of raw ESTs sequences, clustering, assembly, and detection of SNPs in aligned sequence data. Several bioinformatic tools, including PolyBayes, PolyPhred and PolyPhrap softwares, have been developed for *in silico* SNP detection. In a maritime pine, Dantec *et al*. (2004) compared automated and visual methods for SNP mining on a reference set of true SNPs. The result showed that eSNP analysis can detect 83.1% of all the SNPs and 98% of the "non-rare" SNP alleles. By using this method, a total of 1 400 SNPs have been detected from several EST databases of different tissues, resulting in a frequency of 1 SNP every 660 bp. This demonstrates the eSNP approach is an effective, cost-efficient and high-throughput way for SNPs discovery. Furthermore, SNPs frequently occur in the coding sequence and can directly influence protein structure and functions.

2.3 Patterns of SNP distribution and frequency

As a prerequisite for SNPs-based genetic analyses, it will be important to have a comprehensive understanding of the patterns of SNP distribution and frequency within and among different genes, and within and among natural populations. Many efforts have been made systematically to discover and analyze SNPs frequency within candidate genes and the dispersal pattern of SNPs across the genomic regions and have obtained striking achievements in humans, animals, plants and forest trees. The characteristics of SNPs in some species are listed in Table 2. Table 2 shows the formal characteristics of SNPs in some species with different genes. Although the methods and species samples used in each study are different, making detailed comparisons impractical, several common trends can be observed. 1) The SNPs diversity shows there is variation across genes and species, showing different selective pressures on each gene as well as different mutation and recombination rates across the genome of different species; 2) SNPs in

non-coding sequence and synonymous mutation in coding sequence are generally more common than non-synonymous mutations, reflecting mutations at positions coding amino acid identity have declined owing to greater selective pressure on them; 3) SNPs of transitional changes are more common than transversion mutation because CpG dinucleotides are known to be prone to point mutation (Cooper and Youssoufian 1988, Cooper and Krawczak 1990).

TABLE 2 Characteristics of SNPs in different species

Species	Number of genes or ESTs	SNP diversity No. (bp)	Transition/ transversion	Exon/intron	Nonsynonymous/ synonymous	Reference
Human	3 950	1/200–1/300	2.4	0.8	0.54	Salisbury *et al*. 2003
Drosophila	109	1/300	1.2	-	-	Hoskins *et al*. 2001
Plants						
Arabidopsis	10 706	1/336	-	-	0.47	Schmid *et al*. 2003
Rice	45	1/616	-	0.9	1.57	Shirasawa *et al*. 2004
Maize	21	1/27.6	1.0	-	0.68	Tenaillon *et al*. 2001
Soybean	90	1/328	0.9	0.5	3.30	Zhu *et al*. 2003
Forest trees						
P. tremula	5	1/60	-	-	0.10–0.23	Ingvarsson 2005
E. nitens	3	1/59	1.3	0.4	0.27	Author's data
P. taeda	19	1/63	-	-	0.16	Brown *et al*. 2004
P. pinaster	8	1/415	-	-	0.33	Pot *et al*. 2005
P. radiata	8	1/538	-	-	0.50	Pot *et al*. 2005
P. sylvestris	1	1/189	1.4	-	0.33	Dvornyk *et al*. 2003
P. sylvestris	2	1/187	1.3	0.2	0.67	Garcia-Gil *et al*. 2003
C. japonica	7	1/118	-	0.7	0.88	Kado *et al*. 2003

Note: "-" represents no data available.

To understand comprehensively the distribution frequency of SNPs within the complete genes, including 5′ upstream, 5′ UTR, coding region, intron and 3′ UTR, we take the human genome as an example to access it based on 3 393 genes (Fig. 2).

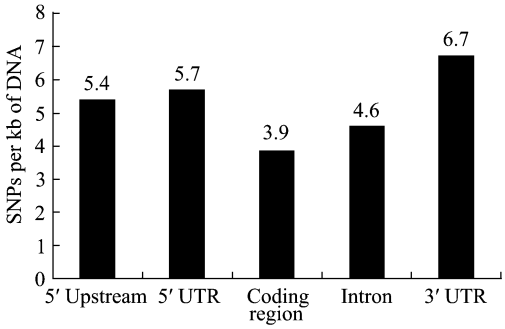

FIGURE 2 SNPs distribution per kb of functionally defined genomic region (3 393 genes, 54 829 SNPs)
Cited from Salisbury *et al*. (2003)

Fig. 2 shows the distribution of SNPs frequency (SNPs per kb of candidate gene) in different gene regions in a statistical way, in which 20.5% in 5′ upstream, 21.7% in 5′ UTR, 14.8% in exon regions, 17.5% in intron region and 25.5% in 3′ UTR. The patterns of SNPs distribution in the gene regions observed here are consistent with most previous reports using more or less gene numbers and this finding reflects a pervasive trend of SNPs distribution frequency in different gene regions in most genes (Schneider *et al*. 2003). SNPs are less abundant in coding regions than in introns and other regions, suggesting that there is a conservation of the coding region relative to the other regions under natural pressure.

The frequency and nature of SNPs in forest trees receive considerable attention in recent years. Several reports in trees have provided estimates of nucleotide diversity in these species (Table 2). They are outcrossing species with large, stable population size and long history and are expected to have high levels of SNPs diversity in natural populations. The analysis of SNPs diversity in these species has been mainly confined to single gene or candidate DNA fragments with the goal of identifying the gene structure, evolutionary relationships and SNPs-based association with quantitative traits (Dvornyk *et al*. 2002, Kado *et al*. 2003, Brown *et al*. 2004, Ingvarsson 2005, Pot *et al*. 2005, Thumma *et al*. personal communication). The nucleotide diversity, and the ratios of transition to transversion, exon to intron and non-synonymous to synonymous are different among woody species researched (Table 2). There was abundant nucleotide variation varying from 1/59 to 1/538 in these 7 tree species, but

more SNPs diversity in hardwood trees than in softwood trees can be observed (Table 2). The identical patterns of SNPs level of transition/transversion, exon/intron and non-synonymous/ synonymous in these 7 different tree species can also be made. The result shows the frequency and nature of SNPs in forest trees are consistent with those in humans and plants.

Let us take *COBL2* gene as an example to describe in detail the diversity and characteristics of SNPs in trees. To identify the SNPs of *COBL2* gene in *E. nitens*, we use the gene specific primer pairs to amplify this gene, including partial 5′ UTR and 3′ UTR, based on 18 unrelated *E. nitens* individuals selected from natural populations in Australia. A total of 17 common SNPs (frequency > 0.10) have been detected over 2 281 bp of DNA segment (Table 3). Previous literature has generally used a nomenclature for SNPs positions relative to the start codon ATG. Thus, the SNPs located in 5′ UTR have been referred to as –13, and those in 3′ UTR have been referred to as +2 183. Of the 17 common SNPs, 1 (5.9%), 6 (35.3%), 9 (52.9%) and 1 (5.9%) have been located in 5′ UTR, exon region, intron region and 3′ UTR, respectively. Of these, 10 were transitions (58.8%) and 7 transversions (41.2%). The ratio of transitions to transversions for these SNPs was about 1.4:1, which had a similar trend in human genes. In coding regions, 4 SNPs (SNP6, C⇔T; SNP11, C⇔T; SNP14, A⇔G; SNP16, T⇔G) were found to occur at the third codon position (TT**C** or TT**T** (Phe); AA**C** or AA**T** (Asn); GC**A** or GC**G** (Ala); GT**T** or GT**G** (Val)) and they cause no change in their amino acids and therefore have been referred to as synonymous mutations. In contrast, 2 SNPs (SNP2, A⇔G; SNP7, A⇔G) were found to occur at the fist and second positions (**G**TC (Val) or **A**TC (Ile); A**A**G (Lys) or A**G**G (Arg)), respectively. These two nucleotide mutations result in corresponding amino acid variation and these two SNPs have been referred to as non-synonymous mutations and may become very useful for their potential links to phenotype variation via linkage disequilibrium.

TABLE 3 SNPs and LD in the *COBL2* gene in *E. nitens*.

Individuals	Nucleotide positions																
	–13	10	127	371	425	832	912	1 085	1 086	1 210	1 299	1 380	1 437	1 559	1 786	1 932	2 183
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	C	G	T	C	T	C	A	G	A	G	T	G	A	A	G	T	C
2	C	G	T	C	T	C	A	G	A	G	T	G	A	A	G	T	C
3	T	A	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
4	T	A	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
5	T	A	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
6	T	A	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
7	T	A	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
8	T	G	C	C	T	T	G	G	A	T	T	C	A	G	G	T	C
9	T	G	C	C	T	T	G	G	A	T	T	C	A	G	G	T	C
10	T	G	C	C	T	T	G	G	A	T	T	C	A	G	G	T	C
11	T	G	C	C	T	T	G	G	A	T	T	C	A	A	G	T	C
12	C	G	T	C	T	T	G	G	A	T	T	C	A	A	G	T	C
13	T	G	T	T	G	C	A	G	G	G	T	G	C	A	G	T	C
14	T	G	T	T	G	C	A	G	G	G	T	G	C	A	G	T	C
15	T	G	T	T	G	C	A	G	A	G	T	G	A	A	G	T	C
16	C	G	T	C	T	C	A	T	A	G	C	G	A	A	C	G	T
17	C	G	T	C	T	C	A	T	A	G	C	G	A	A	C	G	T
18	C	G	T	C	T	T	A	G	A	G	T	G	A	A	C	G	T

2.4 SNP genotyping

The association analysis or LD mapping of SNPs-based candidate genes will require to ascertain SNPs genotypes for several hundreds of individuals and numerous SNPs in each case study. The development of high-throughput technologies with accuracy, easiness and low cost has been vital to the widespread use of SNPs to dissect quantitative traits. At present, many different techniques are used for SNPs typing, mainly including hybridization, primer extension, exonuclease detection (*TaqMan*), invasive cleavage of oligonucleotide probes (invader assay), multiplex ligation-dependent probe amplification (MLPA) and so on. The articles describing the use of SNPs genotyping to perform the LD mapping in forest trees have not been reported as of now, but some forest geneticists have

cists have developed the research of SNPs genotyping in natural populations. To our knowledge, only two methods, i.e. the single-base primer extension and multiplex ligation-dependent probe amplification techniques, have been used to carry out the SNPs typing in some forest genetic labs. Thus, we will simply describe the principles, protocols, advantages and limitations of these two methods as follows.

2.4.1 Primer extension

The single-base primer extension is a method of ascertaining the precise genotypes of an interest SNP site (Nyren *et al*. 1997). It utilizes the inherent accuracy of DNA polymerase to determine the presence or absence of the specific nucleotide at the SNP site. The principle behind this method is to design a detection primer complementary to the target DNA. The 3′-terminal of detection primer ends at the base just before the target base. The detection primer hybridizes to the target sequence. DNA polymerase inserts the complementary dideoxy nucleotide terminator to the SNP site. The extended product will display different colours corresponding to each nucleotide (A, T, C, G) using a fluorescently labelled ddNTP. We will clearly know the genotypes of interest SNPs based on the colors in the monitor of sequencing machine. This method contains three major steps, i.e. template cleanup, single-base primer extension, and extended product cleanup and then runs the samples on the CEQ 8000 using a SNP-1 separation method. Now a commercial CEQ SNP-Primer Extension Kit has been available from the biotechnology company, which provides CEQ users an accurate, simple, cheap, and robust solution for SNP scoring and validation based on single-base primer extension technology. CEQ 8000 Genetic Analysis System allows the analysis of multiple SNP loci in a single capillary. Therefore this method is extensively used to ascertain the SNPs genotype in humans, plants and forest trees. For example, Thumma *et al*. have used this method to detect the genotypes of 290 individuals and successfully done the association mapping analysis with accurate SNPs data (Thumma *et al*. personal communication).

2.4.2 Multiplex ligation-dependent probe amplification

Multiplex ligation-dependent probe amplification (MLPA) is a new method to detect the copy number and single-base mutation of up to 45 nucleic acid sequences in one single reaction (Schouten *et al*. 2002). It can be applied on genomic DNA (both copy number detection and methylation quantification) as well as for mRNA profiling. With MLPA, it is possible to perform a multiplex PCR reaction in which up to 45 specific sequences are simultaneously quantified. Amplification products are separated by sequence type electrophoresis. Since only one pair of PCR primers is used, MLPA reactions result in a very reproducible gel pattern with fragments ranging from 130 to 490 bp. MLPA probes are able to discriminate between sequences that differ in only one nucleotide and only require a minimum of 20 ng of genomic DNA. Compared with other techniques, an MLPA reaction is fast and very easy to perform. The equipment required is present in most molecular biology laboratories.

MLPA probe signal is completely absent if the short probe oligonucleotide has a mismatch at the 3′ nucleotide when annealed to the target sequence. This sensitivity of the ligase for a mismatch next to the ligation site can be used to distinguish two sequences differing in only one or a few nucleotides, such as SNPs and various mutations. In order to obtain a signal from both alleles, which can be distinguished by the length of the probe amplification products, Schouten *et al*. (2002) used probes for the two alleles that have the M13-derived oligonucleotide in common. Two different short synthetic probe oligonucleotides are used that differ at both the site of the mutation/SNP and also by 3′ nucleotides in length. These two oligonucleotides will compete with each other for binding to both target sequences since a single mismatch at the end of the probe oligonucleotide will usually not be sufficient to destabilize the hybrids.

In comparison with the single-base primer extension, MLPA is a timesaving and economical, accurate, sensitive and high-throughput method for SNPs genotyping. This method is currently used to detect the copy number, deletion and duplication of exons in humans (Janssen *et al*. 2005). This method has also been successfully used for SNPs genotyping of several candidate genes in forest trees. We consider that MLPA is a most cost-effective, accurate and high-throughput and feasible method for large-scale analysis of SNPs genotype in forest trees up to date.

3 Linkage disequilibrium (LD)

3.1 The theory of linkage disequilibrium

Linkage disequilibrium (LD), simply defined, is the non-random association of alleles at different loci. For example, the SNP7, SNP10 and SNP12 are the form of LD in the region of the *COBL2* gene (Table 3). The concepts of LD can date back to the early 20th century (Jennings 1917), but the first commonly used LD measure was developed by Richard Lewontin in the 1960s (Lewontin 1964). LD is created when a new mutation occurs on a chromosome that carries a par-

ticular allele at a nearby locus, and is gradually eroded by recombination. Recurrent mutations can also lessen the association between alleles at adjacent loci. Theoretically, in a large enough, randomly mated population with loci segregating independently, but in the absence of selection, mutation, or migration, polymorphic loci will be in linkage equilibrium (Falconer and Mackay 1996). Practically, however, factors such as physical linkage, genetic drift, selection on multilocus genotypes within populations, and population admixture can also cause LD between markers and traits.

We are often confused by the conception between linkage and LD. Thus, it is necessary to summarize briefly the differences between them as follows. 1) Linkage focuses on a locus, but LD emphasizes an allele; 2) Linkage results from recombination events in the last 2–3 generations and LD from much earlier, ancestral recombination events; 3) Linkage measures co-segregation in a pedigree while LD measures co-segregation in a population; 4) From the point of view of a dynamic system, linkage is the "dynamic equation", LD is the "initial condition"; 5) Linkage is usually detected for markers reasonably close to the target gene (one cM or more); LD, however, is detected for markers even closer (within 0.01–0.02 cM or one gene). Therefore, unlike linkage analysis, LD analysis effectively incorporates the effects of many past generations of recombination and has often been instrumental in the final phases of gene localization (Long and Langley 1999, Jorde 2000, Nordborg 2000). These successes have fueled hopes that association studies based on LD can provide high resolution for identifying genes that may contribute to phenotypic variation.

3.2 Measuring LD

A variety of statistical approaches has been used to measure the extent of LD and these have different strengths, depending on the context (Delvin and Risch 1995, Jorde 2000). Among these, the two most common measures are the absolute value of D' and r^2. Thus, we will introduce these two methods.

The metric D is a quantitative measure of allelic association. Consider a pair of loci with alleles A and a at locus one, and B and b at locus two, with allele frequencies P_A, P_a, P_B and P_b, respectively. The resulting haplotype frequencies are P_{AB}, P_{Ab}, P_{aB} and P_{ab}. The basic component of all LD statistics is the difference between the observed and expected haplotype frequencies,

$$D_{ab} = (P_{AB} - P_A P_B).$$

The case of D'=1 is known as the complete LD. Values of D'<1 indicate that the complete ancestral LD has been disrupted. The magnitude of values of D'<1 has no clear interpretation. D incorporates information about allelic association and allele frequencies. Estimates of D' are strongly inflated in small samples. Therefore, statistically significant values of D' near 1 provide a useful indication of minimal historical recombination, but intermediate values should not be used for comparisons of the strength of LD between studies, or to measure the extent of LD.

Another measure that deserves mention is r^2, which is equal to D^2 divided by the product of the allele frequencies at the two loci. Hill and Robertson (1968) deduced that

$$E\,[r^2] = 1/(1 + 4Nc)$$

where c is the recombination rate in morgans between the two markers and N is the effective population size. This equation illustrates two important properties of LD. First, expected levels of LD are a function of recombination. The more recombination between two sites, the more they are shuffled with respect to one another, decreasing LD. Second, LD is a function of N, emphasizing that LD is a property of populations. To arrive at this equation, Hill and Roberson (1968) assumed that the population was an "ideal", large, random-mating population without natural selection and mutation.

3.3 Factors that influence LD

LD is affected by both biological and non-biological factors, such as mutation, recombination, mating system, selection, population history and population structure. Mutation provides the raw material for producing polymorphisms that will be involved in LD. Recombination is the main phenomenon that weakens intrachromosomal LD, whereas interchromosomal LD is broken down by independent assortment. Thus, mutation and recombination might have the most evident impact on LD. Recombination rates are known to vary by more than one order of magnitude across the genome. Because breakdown of LD is primarily driven by recombination, the extent of LD is expected to vary in inverse relation to the local recombination rate (Nachman 2002).

The mating system has also a profound effect on LD. Generally, LD decays more rapidly in outcrossing species compared with selfing species (Table 4). For example, the LD extended over long distances (>50 kb) in selfing species such as *Arabidopsis*, whereas it decayed rapidly in only several hundred or thousand bp in outcrossing species such as maize or forest trees (Table 3). The cause of this phenomenon is that selfing mechanism may have increased recombination

rates per meiosis (Charlesworth and Charlesworth, 1979); For example, the recombination rate per base pair is estimated to be approximately two and six times higher in selfing *Arabidopsis* than in *Drosophila* and maize, respectively (Gaut and Long 2003). However, selfing increases homozygosity, thereby limiting the number of double heterozygotes that can be shuffled by recombination. As a result, the effective rate of recombination is low in selfing species, genetic polymorphisms tend to remain correlated and LD is expected to be maintained over long physical distances.

A strong directional selection (natural or artificial) at a locus is expected to reduce diversity of alleles and increase LD in the surrounding regions. There are two primary routs by which selection can affect the extent of disequilibrium. The first is a hitchhiking effect, in which an entire haplotype that flanks a favored variant can be rapidly swept to high frequency or even fixation (Wang *et al*. 2002). Although the effect is generally milder, selection against deleterious variants can also inflate LD, as the deleterious haplotypes are swept from the population. Genetic hitchhiking is expected to affect the frequency distribution of variants at segregating sites so that derived variants will be in higher frequency than expected under a neutral equilibrium model. The second way in which selection can affect LD is through epistatic selection for combinations of alleles at two or more loci on the same chromosome (Cannon 1963). This form of selection leads to the association of particular alleles at different loci. This has provided a major motivation for historical studies of LD in *Drosophila* genetics as a means of detecting the action of natural selection.

Furthermore, the range of LD is highly dependent on the population in which it is measured. A previous study of Reich *et al*. (2001) established that LD in human is highly population-dependent. In some populations, such as Europeans, LD might extend to 60 kb, whereas in others, such as Yoruba Africans, it declines within a few kb. In this situation, the main factor determining the extent of LD appeared to be population history, particularly population bottlenecks associated with geographical expansion and population isolation.

3.4 Patterns of LD in different species

Global patterns of genomic LD are directly relevant to the design LD mapping, since fewer SNPs will be required in regions of high LD than in regions of low LD. In general, LD is expected to decrease with physical distance, since there will be more opportunity for recombination between a distant pair of SNPs than a closely-spaced pair of SNPs. Therefore, it is a precondition that we should clearly know the patterns of LD in different species before we can use it in LD mapping. Much of our understanding of how LD is shaped in natural populations initially came from research on the *Drosophila* species. In *Drosophila*, LD decays within one kb in the regions on the X chromosome owing to the impact of recent selection and less recombination in some cases (Table 4). However, in humans, studies indicated that LD extends over large distances ranging from 5 to 60 kb because of increasing recombination and increasing effective population size (Table 4). Similarly, in mammal populations, such as cattle, sheep, pigs and dogs, the LDs have been extended in the range of tens of cM, even to unlinked markers (Table 4). It has been concluded that most of the LD observed in these animal populations could be accounted for by bottlenecks caused by the globalization of semen trading.

TABLE 4 LD in different species

Species	Mating system	LD range	Reference
Human			
Nigerian	Outcrossing	5 kb	Reich *et al*. 2001
European	Outcrossing	60 kb	Reich *et al*. 2001
Animals			
Cattle	Outcrossing	10 cM	Farnir *et al*. 2000
Sheep	Outcrossing	30 cM	Mcrae *et al*. 2002
Pig	Outcrossing	33 cM	Nsengimana *et al*. 2004
Dog	Outcrossing	5-10 cM	Sutter *et al*. 2004
Drosophila	Outcrossing	<1 kb	Long *et al*. 1998
Plants			
Arabidopsis	Selfing	250 kb	Nordborg *et al*. 2002
Rice	Selfing	100 kb	Garris *et al*. 2003
Barley	Selfing	10 cM	Kraakman *et al*. 2004
Soybean	Selfing	50 kb	Zhu *et al*. 2003
Maize	Outcrossing	<1 kb	Remington *et al*. 2001
Forest trees			
P. tremula	Outcrossing	<500 bp	Ingvarsson 2005
E. nitens	Outcrossing	<500 bp	Author's data
P. taeda	Outcrossing	<2 kb	Brown *et al*. 2004

Among plants and forest trees, the ranges of LD extension have a sharp contrast between the selfing species and outcrossing species (Table 4). Table 4 shows that LD extends long distances up to 250 kb or 10 cM in the selfing species such as *Arabidopsis*, rice, barley and soybean because the long term selfing system results in a limited number of recombination events that have occurred over the past 200 years. Conversely, in outcrossing species, such as maize, *P. tremula*, *E. nitens* and *P. taeda*, LD extends over relatively short distances (less than 1 or 2 kb). Taking the *E. nitens* as an example, we describe the pattern of LD

within one gene in detail (Table 3). The extension of LD was rapidly decayed in less than 500 bp. Several SNPs sites have been detected in the form of LD within the *COBL2* gene (Table 3). For example, SNP4 and SNP5, SNP7, SNP10 and SNP12, SNP8 and SNP11, SNP9 and SNP13, SNP15, SNP16 and SNP17 are in the form of complete LD within 500 bp. These results imply that SNPs-based association studies are feasible within candidate genes in forest trees.

4 Applications in forest genetics

Unlike other DNA marker techniques, SNPs are highly abundant (one every hundred bases), mutationally stable, biallelic innature, easy to scorn, high-throughput genotyping, and are randomly distributed across the genome. These characteristics, along with the availability of large-scale ESTs database, make SNPs useful as genetic tools to exploit population genetics and SNPs-based association studies in forest trees.

4.1 Application of SNPs in studies of population genetics

Since the early 1990s, SSRs loci and mitochondrial DNA (mtDNA) sequences have been the tools of choice in molecular studies in population evolution in forest trees (Gillet and Scholz 1999, Rajora *et al*. 1992). Both kinds of genetic markers represent rapidly evolving DNA sequences that are informative for answering population-level questions. However, the SSRs loci typically have many alleles (5–20), null alleles and high mutation rates. Furthermore, there occurs the homoplasy in SSRs loci, that is, the occurrence of SSRs alleles of identical size but different evolutionary origins (Viard *et al*. 1998). Conversely, it is also likely that SSRs of different sizes are embedded in identical haplotypes. Homoplasy poses severe limitations on subsequent data analysis and, thus, the biological meaning and usefulness of the results. Inferences drawn from mtDNA sequences are further limited due to the fact that the mtDNA genome comprises a single maternally inherited locus. Therefore, SSRs are less suitable for genetic evolution analysis in comparison with SNPs. SNPs analyses do not need DNA separation by size, and can be automatically measured on a large scale and located in distinct genomic regions. On the other hand, SNPs are biallelic and often closely spaced and in the presence of LD. Therefore, the information provided by SNPs is most useful to define haplotypes in the region being examined completely and can accurately disclose the evolution history and structure of population in forest trees.

Analyzing the instantaneous rate of nonsynonymous (amino acid-changing) and synonymous (silent) nucleotide substitutions in protein-coding molecular sequences can give important clues to understanding how they evolved. In particular, the ratio of the non-synonymous to synonymous fixation has been used to measure the level of selective pressure on proteins. For example, the ratio of non-synonymous to synonymous diversity ranged from 0.1 to 0.5 for most genes listed in Table 2 in forest trees, indicating strong purifying selection at most codons in these genomic regions, especially in *P. tremula* and *E. nitens* (Table 2). Table 2 shows that non-synonymous substitution rates are markedly differently in different genes, suggesting that selective constraints and/or the history of adaptive evolution vary among genes listed in Table 2. The level of synonymous sites for identical gene in different species can reflect the evolution rate of nuclear genomes in the lineages. For example, the level of synonymous sites are higher in Cupressoideae than in Taxodioideae based on 10 nuclear genes, suggesting that the former species have evolved faster than the latter one (Kusumi *et al*. 2002). Synonymous mutations do not change the encoded proteins and so are often assumed to be selectively neutral. A significant excess of nonsynonymous over synonymous substitution has been used as evidence for adaptive evolution.

4.2 SNP analysis to dissect quantitative traits

The genetic dissection of quantitative traits to individual genetic components is a topic of great interest in humans, animals, plants and forest trees. Currently, SNPs-based association studies or LD mapping have successfully been used to identify the causative SNPs responsible for disease or patients' response to drugs (Nowotny *et al*. 2001, Cheng *et al*. 2005, Zollner *et al*. 2005). This approach has also been extended to model plant species *Arabidopsis* for LD mapping the flower time trait in recent years (Thornsberry *et al*. 2001, Hagenblad *et al*. 2004, Olsen *et al*. 2004). Forest geneticists have got the insights from these pioneer association studies and soon began to develop the LD mapping strategy in forest trees (Plomion *et al*. 2003). LD mapping using natural populations results in high resolution of marker-trait associations compared with multiply-generation family based QTLs analysis. Unlike humans and *Arabidopsis*, the level of LD extension is rapidly decayed within most genes in forest trees (Table 4). Therefore, candidate genes based LD mapping is feasible for most forest trees. For example, Thumma *et al*. have identified the SNPs in *Cinnamoyl CoA Reductase* (*CCR*) gene associated with variation in microfibril angle in *Eucalyptus* with LD mapping approach (Thumma *et al*., personal communication).

Their study demonstrates that candidate genes based LD mapping can be used to identify alleles associated with wood quality traits in natural tree populations. At present, we now are performing many genes association analysis with wood quality in *E*. *nitens*.

5 Future outlook for SNP discovery in forest trees

With the complete genomic sequence of *Populus trichocarp* and the availability of many ESTs database in forest trees, a large-scale SNPs discovery can be carried out in the near future. A better understanding of the LD distribution and level within candidate genes is necessary in forest trees. Concurrent with the high-throughput genotyping progress, it will become feasible to perform large-scale association studies, taking full advantage of the wealth of natural populations in the world. Population establishment and accurate phenotype measurement is required considering for realization of high resolution LD mapping in forest trees. We think that large-scale and comprehensive SNPs discovery and LD mapping will in the first place be carried out in the model tree *Populus* owing to complete genome sequence and abundant population resources available. We hope that SNPs-based association studies will provide an unprecedented opportunity to understand the regulation of genes and interaction of gene to gene and of gene to environment. It is expected that long-term sustainability of forest health, high productivity and enough ability of biotic and abiotic stress tolerance will be holding in the near future.

References

Barendse W, Armitage S M. 2001. The single strand conformational analysis of cattle and human single nucleotide polymorphisms may be biased towards specific sequence motifs that minimize local secondary structure of single strand DNA. *Anim Biotechnol*. 12: 21–8

Botstein D, White R L, Skolnick M, Davis R W. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am J Hum Genet*. 32: 314–331

Bradshaw H D, Stettler R F. 1995. Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* 139: 963–973

Brookes A J. 1999. The essence of SNPs. *Gene*. 234: 177–186

Brown G R, Gill G P, Kuntz R J, Langley C H, Neale D B. 2004. Nucleotide diversity and linkage disequilibrium in loblolly pine. *Proc Natl Acad Sci USA*. 101: 15 255–15 260

Cannon G B. 1963. The effects of natural selection on linkage disequilibrium and relative fitness in experimental population of *Drosophila melanogaster*. *Genetics*. 48: 1 201–1 216

Carlson C S, Aldred S F, Lee P K, Tracy R P, Schwartz S M, Rieder M, Liu K, Williams O D, Iribarren C, Lewis E C, Fornage M, Boerwinkle E, Gross M, Jaquish C, Nickerson D A, Myers R M, Siscovick D S, Reiner A P. 2005. Polymorphisms within the C-reactive protein (CRP) promoter region rre associated with plasma CRP levels. *Am J Hum Genet*. 77: 64–77

Chan E Y. 2005. Advances in sequencing technology. *Mutation Research*. 573: 13–40

Charlesworth B, Charlesworth D. 1979. The evolutionary genetics of sexual systems in flowering plants. *Proc R Soc Lond Ser B Biol Sci*. 205: 513–530

Cheng R, Ma J Z, Elston R C, Li M D. 2005. Fine mapping functional sites or regions from case-control data using haplotypes of multiple linked SNPs. *Annals of Human Genetics*. 69: 102–112

Choy Y S, Dabora S L, Hall F, Ramesh V, Niida Y, Franz D, Kasprzyk-Obara J, Reeve M P, Kwiatkowski D J. 1999. Superiority of denaturing high performance liquid chromatography over single-stranded conformation and conformation-sensitive gel electrophoresis for mutation detection in TSC2. *Ann Hum Genet*. 63: 383–391

Cooper D N, Youssoufian H. 1988. The CpG dinucleotide and human genetic disease. *Hum Genet*. 78: 151–155

Cooper D N, Krawczak M. 1990. The mutational spectrum of single base-pair substitutions causing human genetic disease: patterns and predictions. *Hum Genet*. 85: 55–74

Dantec L L, Chagne D, Pot D, Cantin O, Garnier-Gere P, Bedon F, Frigerio J M, Chaumeil P, Leger P, Garcia V, Laigret F, de Daruvar A, Plomion C. 2004. Automated SNP detection in expressed sequence tags: statistical considerations and application to maritime pine sequences. *Plant Molecular Biology*. 54: 461–470

Delvin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*. 29: 311–322

Dvornyk V, Sirviö A, Mikkonen M, Savolainen O. 2003. Low nudeotide diversity at the pal/locus in the widely distributed *Pinus sylvestris*. *Mol Biol Evo*. 19: 179–188

Fairbrother W G, Yeh R F, Sharp P A, Burge C B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science*. 297: 1 007–1 013

Falconer D S, Mackay T F. 1996. *Introduction of Quantitative Genetics*. Essex, UK: Longman Group Ltd. 464

Farnir F, Coppieters W, Arranz J, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M. 2000. Extensive genome-wide linkage disequilibrium in cattle. *Genetics*. 10: 220–227

Flint-Garcia S A, Thornsberry J M, Buckler IV E S. 2003.

Structure of linkage disequilibrium in plants. *Annu Rev Plant Biol*. 54: 357–374

Freudenberg-Hua Y, Freudenberg J, Kluck N, Cichon S, Propping P, Nöthen M M. 2003. Single nucleotide variation analysis in 65 candidate genes for CNS disorders in a representative sample of the European population. *Genome Research*. 13: 2 271–2 276

Garg K, Green P, Nickerson D A. 1999. Identification of candidate coding region single nucleotide polymorphisms in 165 human genes using assembled expressed sequence tags. *Genome Research*. 9: 1 087–1 092

Garris A J, McCouch S R, Kresovich S. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). *Genetics*. 165: 759–769

Gaut B S, Long A D. 2003. The lowdown on linkage disequilibrium. *The Plant Cell*. 15: 1 502–1 506

Gillet E M, Scholz F. 1999. *Which DNA Marker for Which Purpose?* European Union DGXII Biotechnology FW IV Research Programme. 1–5

Guo M, Rupe M A, Zinselneier C, Habben J, Bewen B A, Smith O S. 2004. Allelic variation of gene expression in maize hybrids. *The Plant Cell*. 16: 1 707–1 716

Hagenblad J, Tang C, Molitor J, Werner J, Zhao K, Zheng H, Marjoram P, Weigel D, Nordborg M. 2004. Haplotype structure and phenotypic associations in the chromosomal regions surrounding two *Arabidopsis thaliana* flowering time loci. *Genetics*. 168: 1 627–1 638

Harding R M, Fullerton S M, Griffiths R C, Bond J, Cox M J, Schneider J A, Moulin D S, Clegg J B. 1997. Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet*. 60: 772–789

Hayashi K. 1992. PCR-SSCP: A method for detection of mutations. *Genet Anal Tech Appl*. 3: 73–79

Hill W G, Robertson A. 1968. Linkage disequilibrium in finite populations. *Theor Appl Genet*. 38: 226–231

Hinds D A, Stuve L L, Nilsen G B, Halperin E, Eskin E, Ballinger D G, Frazer K A, Cox D R. 2005. Whole-genome patterns of common DNA variation in three human populations. *Science*. 307: 1 072–1 079

Hoskins R A, Phan A C, Naeemuddin M, Mapa F A, Ruddy D A, Ryan J J, Young L M, Wells T, Kopczynski C, Ellis M C. 2001. Single nucleotide polymorphism markers for genetic mapping in *Drosophila melanogaster*. *Genome Research*. 11: 1 100–1 113

Ingvarsson P K. 2005. Nucleotide polymorphism and linkage disequilibrium within and among natural populations of European aspen (*Populus tremula* L., Salicaceae). *Genetics*. 169: 945–953

Jennings H S. 1917. The numerical results of diverse systems of breeding, with respect to two pairs of characters, linked or independent, with special relation to the effects of linkage. *Genetics*. 2: 97–154

Janssen B, Hartmann C, Scholz V, Jauch A, Zschocke J. 2005. MLPA analysis for the detection of deletions, duplications and complex rearrangements in the dystrophin gene: potential and pitfalls. *Neurogenetics*. 6: 29–35

Jorde L B. 2000. Linkage disequilibrium and the search for complex disease genes. *Genome Research*. 10: 1 435–1 444

Kado T, Yoshimaru H, Tsumura Y, Tachida H. 2003. DNA variation in a conifer, *Cryptomeria japonica* (Cupressaceae sensu lato). *Genetics*. 164: 1 547–1 559

Kononoff P J, Deobald H M, Stewart E L, Laycock A D, Marquess L S. 2005. The effect of a leptin single nucleotide polymorphism on quality grade, yield grade, and carcass weight of beef cattle. *J Anim Sci*. 83: 927–932

Kraakman A T, Niks W R E, Van den Berg P M M M, Stam P, Van Eeuwijk F A. 2004. Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. *Genetics*. 168: 435–446

Kruglyak L. 1999. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat Genet*. 22: 139–144

Kruglyak L, Nickerson D A. 2001. Variation is the spice of life. *Nat Genet*. 27: 234–236

Kusumi J, Tsumura Y, Yoshimaru H, Tachida H. 2002. Molecular evolution of nuclear genes in Cupressacea, a group of conifer trees. *Mol Bio Evol*. 5: 736–747

Lewontin R C. 1964. The interaction of selection and linkage. I. General considerations: heterotic models. *Genetics*. 49: 49–67

Lichten M J, Fox M S. 1983. Detection of non-homology-containing heteroduplex molecules. *Nucleic Acids Res*. 11: 3 959– 3 971

Long A D, Langley C H. 1999. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Research*. 9: 720–731

Long A D, Lyman R F, Langley C H, Mackay T F. 1998. Two sites in the Delta gene region contribute to naturally occurring variation in bristle number in *Drosophila melanogaster*. *Genetics*. 149: 999–1017

Maniatis T, Tasic B. 2002. Alternative splicing pre-mRNA splicing and proteome expansion in metazoans. *Nature.* 418: 236–243

Mcrae A F, McEwan J C, Dodds K G, Wilson T, Crawford A M, Slate J. 2002. Linkage disequilibrium in domestic sheep. *Genetics*. 160: 1 113–1 122

Nachman M W. 2002. Variation in recombination rate across the genome: evidence and implications. *Curr Opin Genet Dev*. 12: 657–663

Neale D B, Savolainen O. 2004. Association genetics of complex traits in conifers. *Trends in Plant Science*. 9: 325–330

Nordborg M. 2000. Linkage disequilibrium, gene trees and selfing: an ancestral recombination graph with partial

self-fertilization. *Genetics*. 154: 923–929

Nordborg M, Borevitz J O, Bergelson J, Berry C C, Chory J, Hagenblad J, Kreitman M, Maloof J N, Noyes T, Oefner P J, Stahl E A, Weigel D. 2002. The extent of linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet*. 30: 190–193

Nowotny P, Kwon J M, Goate A M. 2001. SNP analysis to dissect human traits. *Current Opinion in Neurobiology*. 11: 637–641

Nsengimana J, Baret P, Haley C S, Visscher P M. 2004. Linkage disequilibrium in the domesticated pig. *Genetics*. 166: 1 395–1 404

Nyren P, Karamohamed S, Ronaghi M. 1997. Detection of single-base changes using a bioluminometric primer extension assay. *Analytical Biochemistry*. 244: 367–373

Oefner P, Underhill P A. 1995. Comparative DNA sequencing by denaturing high-performance liquid chromatography. *Am J Hum Genet*. 57S: 755–761

Olsen K M, Halldorsdottir S S, Stinchcombe J R, Weinig C, Schmitt J, Purugganan M D. 2004. Linkage disequilibrium mapping of *Arabidopsis CRY2* flowering time alleles. *Genetics*. 167: 1 361–1 369

Olivier M. 2004. From SNPs to function: the effect of sequence variation on gene expression. *Physiol Genomics*. 16: 182–183

Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. *Proc Natl Acad Sci USA*. 86: 2 766–2 770

Picoult-Newberg L, Ideker T E, Pohl M G, Taylor S L, Donaldson M A, Nickerson D A, Boyce-Jacino M. 1999. Mining SNPs from EST databases. *Genome Res*. 9: 167–174

Plomion C, Cooke J, Richardson T, Mackay J, Tuskan G. 2003. Report on the forest trees workshop at the plant and animal genome conference. *Comp Funct Genom*. 4: 229–238

Pot D, McMillan L, Echt C, Procost G L, Garnier-Gere P, Cato S, Plomion C. 2005. Nucleotide variation in genes involved in wood formation in two pine species. *New Phytologist*. 167: 101–112

Rajora O P, Barrett J W, Dancik B P, Strobeck C. 1992. Maternal transmission of mitochondrial DNA in interspecific hybrids of *Populus*. *Curr Genet*. 22: 141–145

Reich D E, Cargill M, Bolk S, Ireland J, Sabeti P C, Richter D J, Lavery T, Kouyoumjian R, Farhadian S F, Ward R, Lander E S. 2001. Linkage disequilibrium in the human genome. *Nature*. 411: 199–204

Remington D L, Thornsberry J M, MatsuokaY, Wilson L M, Whitt S R, Doebley J, Kresovich S, Goodman M M, Buckler E S. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA*. 98: 11 479–11 484

Sanger F, Coulson S, Coulson A R. 1977. DNA sequencing with chain-termination inhibitors. *Proc Natl Acad Sci USA*. 74: 5 463–5 467

Schaeffer S W, Walthour C S, Toleno D M, Olek A T, Miller E L. 2001. Protein variation in ADH and ADH-RELATED in *Drosophila pseudoobscura*: linkage disequilibrium between single nucleotide polymorphisms and protein alleles. *Genetics*. 159: 673–687

Schmid K J, Sorensen T R, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, Weisshaar B. 2003. Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for mapping in *Arabidopsis thaliana*. *Genome Res*. 13: 1 250–1 257

Schneider J A, Pungliya M S, Choi J Y, Jiang R, Sun X J, Salisbury B A, Stephens J C. 2003. DNA variability of human genes. *Mechanism of Ageing and Development*. 124: 17–25

Schouten J P, McElgunn C J, Waaijer R, Zwijnenburg D, Diepvens F, Pals G. 2002. Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Research*. 30: e57

Sewell M M, Neale D B. 2000. Mapping quantitative traits in forest trees. In: *Molecular Biology of Woody Plants* (Jain S M, Minocha S C eds). New York: Kluwer Academic Publishers. 407–423,

Shen L X, Basilion J P, Stanton Jr. V P. 1999. Single-nucleotide polymorphisms can cause different structural folds of mRNA. *Proc Natl Acad Sci USA*. 96: 7 871–7 876

Shirasawa K, Monna L, Kishitani S, Nishio T. 2004. Single nucleotide polymorphisms in randomly selected genes among japonica rice (*Oryza sativa* L.) varieties identified by PCR-RF-SSCP. *DNA Research*. 117: 275–283

Suha Y, Vijg J. 2005. SNP discovery in associating genetic variation with human disease phenotypes. *Mutation Research*. 573: 41–53

Sutter N B, Eberle M A, Parker H G, Pullar B J, Kirkness E F, Kruglyak L, Ostrander E A. 2004. Extensive and breed-specific linkage disequilibrium in *Canis familiaris*. *Genome Research*. 14: 2 388–2 396

Tenaillon M, Sawkins M C, Long A D, Gaut R L, Doebley J F, Gaut B S. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc Natl Acad Sci USA*. 98: 9 161–9 166

Thornsberry J M, Goodman M M, Doebley J, Kresovich S, Nielsen D, Buckler E S. 2001. *Dwarf8* polymorphisms associate with variation in flowering time. *Nat Genet*. 28: 286–289

Viard F, Franck P, Dubois M P, Estoup A, Jarne P. 1998. Variation of microsatellite size homoplasy across electromorphs, loci, and populations in three invertebrate species. *J Mol Evol*. 47: 42–51

Wang D G, Fan J B, Siao C J, Berno A, Young P, Sapolsky R, Ghandour G, Perkins N, Winchester E, Spencer J,

Kruglyak L, Stein L, Hsie L, Topaloglou T, Hubbell E, Robinson E, Mittmann M, Morris M S, Shen N, Kilburn D, Rioux J, Nusbaum C, Rozen S, Hudson T J, Lipshutz R, Chee M, Lander E S. 1998. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science*. 280: 1 077–1 082

Wang W, Thornton K, Berry A, Long M. 2002. Nucleotide variation along the *Drosophila melanogaster* fourth chromosome. *Science*. 295: 134–137

Wenz H M, Baumhueter S, Ramachandra S, Worwood M. 1999. A rapid automated SSCP multiplex capillary electrophoresis protocol that detects the two common mutations implicated in hereditary hemochromatosis (HH). *Hum Genet*. 104: 29–35

Zhu Y L, Song Q J, Hyten D L, Van Tassell C P, Matukumalli L K, Grimm D R, Hyatt S M, Fickus E W, Young N D, Cregan P B. 2003. Single-nucleotide polymorphisms in soybean. *Genetics*. 163: 1 123–1 134

Zimdahl H, Nyakatura G, Brandt P, Schulz H, Hummel O, Fartmann B, Brett D, Droege M, Monti J, Lee Y A, Sun Y, Zhao S, Winter E E, Ponting C P, Chen Y, Kasprzyk A, Birney E, Ganten D, Hubner N. 2004. A SNP map of the rat genome generated from cDNA sequences. *Science*. 6: 807

Zollner S, Wen X, Pritchard J K. 2005. Association mapping and fine mapping with Tree LD. *Bioinformatics*. 21: 3 168–3 170