Chapter 10

GENE-ASSISTED SELECTION: APPLICATIONS OF ASSOCIATION GENETICS FOR FOREST TREE BREEDING

Phillip L. Wilcox[1], Craig E. Echt[2], and Rowland D. Burdon[3]

SUMMARY

This chapter describes application of association genetics in forest tree species for the purposes of selection. We use the term gene-assisted selection (GAS) to denote application of marker–trait associations determined via association genetics, which we anticipate will be based on polymorphisms associated with expressed genes. The salient features of forest trees are reviewed, including existing and somewhat limited knowledge of linkage disequilibrium (LD), as well as genomic information for both conifers and hardwoods. The relatively short span of LD in largely undomesticated and outbred forest tree species offer good prospects for precisely locating quantitative trait nucleotide (QTN), but necessitates wise candidate gene selection and generation of nongenic sequences, which could be limiting, particularly for conifers. Prerequisites for successful application are discussed, and include suitable populations for detecting LD; powerful quantitative genetic and bioinformatic capabilities; large EST libraries, if not whole genomic sequences, to identify candidate genes; and other capabilities for studying functional genomics; as well as a mix of quantitative genetics, tree breeding, and molecular biology skills. Experimental designs for tree improvement applications are also described, as well as analytical methods. For existing tree improvement practice, GAS should be applicable in virtually all population strata, although careful evaluation on a case-by-case basis will be needed to determine the appropriate implementation pathway(s). Such evaluation will likely include numerical simulation. GAS also fits well with other biotechnologies used for tree improvement. A number of impediments to

[1] Cellwall Biotechnology Centre, Scion (New Zealand Forest Research Institute), Private Bag 3020, Rotorua, New Zealand

[2] USDA Forest Service, Southern Institute of Forest Genetics 23332 MS Highway 67, Saucier, MS 39574, USA

[3] Ensis Genetics, Scion (New Zealand Forest Research Institute), Private Bag 3020, Rotorua, New Zealand

application are also discussed, including institutional barriers; implementation costs; certain molecular mechanisms underpinning variation; and modes of gene action such as epistasis and genotype-environment interaction.

10.1 INTRODUCTION

Many of the generic applications of association genetics described in this book apply to forest trees as well as other plant species. However, in implementation of association genetics plantation forest tree species differ from most other plant species, because of the unique combination of physical and genetic characteristics of forest trees, as well as the state of existing genomic information. This is especially so for breeding applications, as tree breeding is often very different from breeding of other plant species, particularly annual crop plants. In this chapter we focus on association genetics specifically in the context of tree breeding applications, in part because the most frequent use of association genetics may well be in the areas of selection and breeding. We describe where association genetics can be used in existing tree breeding programs, as well as new technologies under development, and discuss experimental components necessary for demonstration of concept and, ultimately, operational implementation. We also identify some potential limitations and challenges for successful implementation of association genetics in a tree improvement context.

In this section we provide relevant background to the chapter by reviewing the salient biological features of forest tree species, as well as the current state of knowledge of genomics in forest trees, and what is known about patterns of LD in tree species. We also introduce the term "gene-assisted selection" (GAS) used to denote the application of information from association genetics in a selection context, and compare and contrast this with marker-assisted selection (MAS), which uses information from marker-trait associations in pedigreed mapping populations for within-family selection.

10.2 DISTINGUISHING FEATURES OF FOREST TREES

Key features of most forest tree species include their large size and long lifespan; predominantly outbreeding behavior; slowness to express their phenotype as well as to reach reproductive maturity; and high levels of synteny within genera, and among conifers, within orders. The size and longevity of trees has both benefits and drawbacks. In terms of the latter, size can create major complications for both conventional breeding and the application of DNA polymorphism for selection. The complications involve both delayed expression of traits, and high costs of producing and managing the genetic material. For phenotypic selection, the delayed expression of traits may preclude effective selection for a number of years. It similarly affects any cross-referencing of phenotype with either genomic markers or QTN. The size of trees, along with the lifespan, means that field-testing trees is very expensive, either for a selection population in itself or for establishing relationships between phenotypic values and DNA polymorphisms. Unless the cost is accepted, which is a problem in itself, this in turn will tend to restrict both the potential selection intensity and the quality of information available on the relationships in question. In contrast, however, a key benefit of the long lifespan is the lasting presence of genotypes across years, even decades or centuries, almost "immortalizing" populations. Such a benefit can allow for repeated measurements over time on the same populations, further leveraging genotypic data, and/or allow for repeated DNA collections and therefore continued generation of genotypic data.

Trees are predominantly outbreeding, meaning that population-wide LD between quantitative trait loci (QTL) and neutral-marker alleles will generally be lacking, unless the base population(s) is/are quite strongly structured in one or more of certain ways. Such structuring will tend to be limited in wind-pollinated species, unless breeding populations are composed of recently admixed populations from distinct progenitor provenances or species. Interspecific hybrid breeding populations represent an example of this. Within species we would expect an absence of population-wide LD between QTL- and neutral-marker alleles; therefore LD will be confined to individual families, such that detection and quantification of QTL need to be undertaken independently for each family. Given the large population sizes needed for each family, unless there are extremely large QTL effects, this creates a very powerful incentive to develop GAS, based on establishing the effects of QTN. A key point here is that for among-family selection, which is common in tree improvement, marker–trait relationships ascertained via QTL mapping may have little or no predictive value for among-family selection.

Forest tree species are also frequently slow to reproduce, resulting in breeding generations within tree improvement programs that typically exceed a decade for conifer species, or much more for some angiosperms such as certain oaks (*Quercus* spp.). This contrasts with annual crop species such as corn, where two generations per year are possible in commercial breeding programs. A compensating feature of many tree species is that once reproductive maturity is attained, the numbers of seed produced can be very large, and seed production can last for decades, albeit seasonally, therefore facilitating generation of potentially large populations for experimental purposes. Furthermore, many tree species can be clonally reproduced, allowing for more precise estimation of genotypic value as well as allowing longer-term storage of specific genotypes.

An associated feature of forest trees is the high level of genetic load with deleterious effects of inbreeding (Williams and Savolainen 1996). Related matings for the most part greatly reduce fitness and frequently lead to phenomena such as embryo lethality that result in segregation distortion (Kuang *et al.* 1999), reduced rates of growth, and abnormal phenotypes (Williams and Savolainen 1996). Such effects, combined with the slow onset of reproduction, effectively eliminate the opportunity to develop homozygous lines, therefore populations used for association genetics and QTL mapping alike are typically heterozygous, and show strong variation both phenotypically and genetically.

A further, but mitigating, characteristic of forest trees is the high level of synteny among species, and even among genera, especially in conifers. The potential advantage of this is the leveraging of sequence information across species, as well as information regarding the functional role(s) of specific genes in trait variation. Furthermore, because different species frequently produce structures that are phenotypically very similar e.g., woody tissues), opportunities are enhanced for cross-referencing genomic information among species.

10.3 STATUS OF GENOMIC INFORMATION IN TREE SPECIES

Successful application of association genetics in forest trees, like all other species, requires considerable genomic information, either in the species of interest or in some highly syntenic species. Currently, forest tree species straddle the pre- and postgenome divide, with the majority (especially conifers) in the former. Recently, the full genome sequence of a poplar (*Populus*) has been determined, a first for a forest tree species (http://www.jgi.doe.gov/poplar). A further effort is currently underway in *Eucalyptus*

(www.ieugc.up.ac.za). Extensive EST databases have been developed for a number of species within these genera, although some questions have been raised about the level of EST representation; based on gene predictions, it is estimated that as much as 75% of genes are not represented in EST databases (unpublished results cited in Plomion *et al.* 2005). In conifers, most DNA sequence information is restricted to EST databases, which exist for a number of commercially important *Pinus* and *Picea* species, in addition to Douglas-fir (*Pseudostuga menziesii*) and a number of other conifer species. Most of these resources are publicly available (e.g., http://fungen.botany.uga.edu/Projects/Pine/Pine.htm, http://dendrome.ucdavis.edu/Gen_res.htm, http://web.ahc.umn.edu/biodata/nsfpine/), although some are proprietary. EST sequences have been determined using cDNA libraries constructed from a wide range of tissues, including developing xylem and cambium, roots, floral structures, and needles/leaves. For conifers, only a limited amount of nongenic sequence has been generated, and is usually associated with genic sequences (e.g., regulatory elements). While an increased amount of gDNA sequence data is likely, the large size of conifer genomes means it is unlikely that full sequence will be available within a short timeframe. Some technologies, such as sequencing Cot-based libraries and bacterial artificial chromosomes, may facilitate generation of a limited amount of genomic sequence data.

Linkage maps have been constructed for a wide range of tree species, primarily for the purposes of QTL studies (see Sewell and Neale 2000 and references therein), based upon a wide array of commonly used marker systems, including ESTs. A number of comparative mapping studies have also been undertaken (Devey *et al.* 1999; Echt *et al.* 1999; Chagné *et al.* 2003), elucidating the synteny referred to above, particularly among conifers. Linkage maps have been constructed for the most, if not all, commercially important forest plantation species, although applications in breeding programs have not been as widespread. However, relatively few studies have been undertaken evaluating synteny of QTL across species. One such study – comparing traits of adaptive significance in *Quercus robur* and *Castanea sativa* – found conservation of QTL for timing of bud burst but not for height or carbon-isotope discrimination (Casasoli *et al.* 2006). Telfer *et al.* (2006) reported nonrandom coincidence of QTL for wood density between *Pinus radiata* and *Pinus taeda*.

Linkage maps have been used extensively for QTL detection studies, mostly in full- or half-sib families. With the notable exception of disease resistance (e.g., Kinloch *et al.* 1970; Wilcox *et al.* 1996), the vast majority of QTL for commercially relevant traits appear to be of small effect only (Wilcox *et al.* 1997; Sewell and Neale 2000; Brown *et al.* 2003; Devey *et al.* 2004), indicating a large number of genes involved in variation of a particular trait. Implications for association genetics in an applied breeding context are that large population sizes will be needed to detect such QTL in sufficient quantity. This is discussed in more detail later in this chapter.

More recently, a range of gene expression technologies have also been applied in forest tree species, in particular microarrays (Kirst *et al.* 2004; Paux *et al.* 2004), and more recently reverse transcriptase polymerase chain reaction (RT-PCR), elucidating the level of gene expression in specific tissue types. This, coupled with EST databases and a suite of bioinformatics tools available, has generated much knowledge about the relative levels of gene expression, including both temporal and spatial variation in tissue of interest for a suite of genes. Such expression studies will be useful for selecting candidate genes for association genetics studies.

As with many other plant and animal species, however, the roles of genes in trait variation are largely unknown. To date, there are no reports of QTL having been cloned from forest tree species, partly due to the large number of candidates within QTL confidence intervals, but also because of the length of time required for trait expression of transformants arising from complementation studies, as well as the largely subtle effects expected for most QTL, together requiring considerable experimental resources to confirm complementation.

10.4 LD AND NUCLEOTIDE DIVERSITY IN FOREST TREE SPECIES

LD and nucleotide diversity, insofar as the latter governs functional variation, are the two key parameters for evaluating the efficacy of association genetics. To date, there have been relatively few extensive studies of LD in forest trees (see Gupta *et al.* 2005 for a review of LD in higher plants). Studies conducted in the 1980s with relatively limited numbers of polymorphic isozyme loci indicated limited or no LD, as would be expected in outbred species with relatively large effective population sizes. Mitton *et al.* (1980) found higher-than-expected digenic LD (6 out of 30 locus pairs) in *Pinus ponderosa*. Similarly, Roberds and Brotschol (1985) found evidence for age-related differences in the incidence of LD in *Liriodendron tulipifera*. Muona and Szmidt (1985) reported no evidence of LD in either pollen or megagametophytes in *Pinus sylvestris*. A study in *Pinus contorta* by Epperson and Allard (1987) showed higher-than-expected LD, but was limited to certain locus combinations, with some closely linked loci not in LD. Geburek (1998) also reported higher-than-expected digenic LD in *Picea abies*, although most were restricted to two or less subpopulations. In most of the aforementioned isozyme-based studies, nonrandom mating and/or selection on a limited number of loci were the most frequent explanations offered for higher-than-expected observed LD.

Studies with DNA-based markers have tended to reveal similar results. Bucci and Menozzi (1995) reported no LD in a small sample of *P. abies* using RAPD markers. A later study in *P. radiata*, involving microsatellite marker loci from a range of linkage groups, also indicated very little genome-wide LD (Kumar *et al.* 2004). More recently, a number of results from DNA sequence have been reported for conifers as well as *Eucalyptus* (Thumma *et al.* 2005) and *Populus* (Yin *et al.* 2004), surveying LD patterns in relatively small regions in and around expressed genes. Results to date generally indicate very short regions of LD, particularly in conifers where r^2 values tend to decrease to zero within a few hundreds to low thousands of base pairs (Table 10.1, and associated references), although there is considerable variability even within genes. Some exceptions have been noted in the average length of LD within genera; Yin *et al.* (2004) reported significant LD in regions around the *MXC3* resistance gene in *Populus trichocarpa* in the order of 16–34 kb. These results indicate that while on average the amount of LD is confined to relatively short spans in forest tree species, variations need to be taken into account, which can only be characterized via empirical data on genes of interest.

Table 10.1. Estimates of linkage disequilibrium and nucleotide diversity in plantation forest tree species based on DNA markers and candidate genes

Genus and species	Extent of LD	Metric(s)	No. of genes	Nucleotide diversity (Synonymous or not)		References
				yes	no	
Pinus radiata	No evidence between unlinked SSR markers	r^2	N/A	N/A	N/A	Kumar *et al.* (2004)
P. radiata	Not estimated	N/A	1	0.0300	0.0043	Cato *et al.* (2006)
P. radiata	Not estimated	N/A	8	0.0008	0.00005	Pot *et al.* (2005)
P. pinaster	Not estimated	N/A	8	0.0003	0.00015	Pot *et al.* (2005)
P. sylvestris	None observed within approx. 2 kb[a]	r^2	11	0.0056	0.0022	Dvornyk *et al.* (2002)
P. taeda	2,000 bp	$r^2 \sim 0.2$	19	0.0064	0.0011	Brown *et al.* (2004b)
Pseudostuga menziesii	1,000 bp	$r^2 \sim 0.1$	18	0.0105	0.0021	Krutovsky and Neale (2005)
Picea abies	100 bp 200 bp	$r^2 \sim 0.2$	?	Not provided	Not provided	Unpublished results cited in Rafalski and Morgante (2004)
Eucalyptus nitens	"Similar results to maize and *Pinus*"	r^2	1	Not estimated	Not estimated	Thumma *et al.* (2005)
Populus trichocarpa	Up to 34 kb	Not provided	1	Not estimated	Not estimated	Yin *et al.* (2004)
Populus tremula	<500 bp	$r^2 < 0.05$	5	0.0220	0.0059	Ingvarsson (2005)

[a]Analyses based on one gene only.

Nucleotide diversity in forest tree species appears to be variable both among and within species. In most conifers, typical reported values range between ca. 10^{-2} and 10^{-4}, with some variation within species (Krutovsky and Neale 2005). Overall, forest trees appear to show more such diversity than humans, but slightly less than that observed in species such as maize (Brown *et al.* 2004b). Diversity appears to be lower in coding sequences, with nonsynonymous substitutions being less frequent than synonymous substitutions, although rarely are such differences reported as being statistically significant – for example, Brown *et al.* (2004b) found no evidence for selection in 19 genes in *P. taeda,* while Krutovsky and Neale (2005) reported evidence for selection in *P. menziesii* in three of 18 expressed genes. Cato *et al.* (2006) reported evidence for selection in a putative dehydrin gene in *P. radiata*, and found weak associations with the same gene and wood density and growth rate.

The moderate nucleotide diversity, coupled with the typically low LD per base pair, indicates a relatively high number of haplotypes per genic region. For example, Krutovsky and Neale (2005) found that there were approximately 2–3 haploblocks per gene, thus on average, 4–5 single nucleotide polymorphisms (SNPs) would be needed to adequately cover most single genes for association genetics applications.

What is the significance of these results for association genetics in conifers? Firstly, the observed levels of nucleotide diversity indicate there is sufficient polymorphism for association genetics studies. Secondly, the relatively small regions of LD give some cause for optimism regarding functional assignment, as the small regions of LD observed within most genes indicate the possibility of implicating genes (or even small regions within, or associated with, genes) in trait variation. The disadvantage is that relatively

detailed studies will be needed, typically assaying many polymorphisms in regions of interest, necessitating judicious targeting of candidate regions to limit the number of genes to be screened. Such detailed studies are costly and time-consuming, particularly if applied breeding is the key objective. However, short stretches of LD mean there is some potential for using association genetics to assign putative function to genes, and will be of use to those seeking to determine molecular mechanisms underpinning phenotypic variation.

To date, relatively few results have been reported from association genetics experiments, although this should change. Kumar *et al.* (2004) found only weak evidence for association between polymorphic SSR markers and a number of traits in a small female-tester mating design in *P. radiata*. Since then, Brown *et al.* (2004a) reported a putative association between an SNP within an α-tubulin, and earlywood microfibril angle, a key component influencing performance of structural-grade timber in conifers. More recently, Thumma *et al.* (2005) reported an association between polymorphism encoding a putative splice-site variant in a Cinnamoyl CoA Reductase (*CCR*) gene in *Eucalytus nitens* and microfibril angle. A mutation in the putative functional homologue of this gene in *Arabidopsis thaliana* proved to cause the *IRX4* phenotype (Jones *et al.* 2001). Cato *et al.* (2006) reported an association in *P. radiata* between polymorphisms in a putative stress-response gene, with both wood density and growth rate in large association population.

10.5 GENE-ASSISTED SELECTION VERSUS MARKER-ASSISTED SELECTION

One of the key features of outcrossing species such as forest trees is the expectation of widespread linkage equilibrium within unstructured populations, and conversely, the expectation of strong LD within specific pedigrees. The latter has been extensively utilized to date in the field of QTL mapping based on pedigreed populations (usually full-sib families), leading to the development of linkage maps for a wide range of species and demonstration of the potential for within-family MAS. This approach, however, has various limitations, including the restriction of selection to within specific families for which the marker allele–trait associations have been previously established (Strauss *et al.* 1992; Johnson *et al.* 2000; Wilcox *et al.* 2001).

From a tree breeding perspective, the key feature of association genetics is the opportunity to select both among and within families, by establishing relationships between polymorphisms and heritable trait variation outside of any family structure. However, because LD is restricted to relatively small chromosomal regions in forest tree species, we consider that the most likely polymorphisms to be associated with trait variation are those within, or associated with, expressed genes. For this reason we use the term "gene-assisted selection" to denote the application of within- and/or among-family selection based on polymorphisms shown to be associated with trait variation in unstructured populations, i.e., association genetics.

The idea of selecting genotypes based on DNA sequence variation is not new – the concept of MAS is indeed based on the same principles, i.e., selecting on the phenotypic-, and/or discrete isozyme-, and/or DNA-sequence variants that are correlated, through linkage, with phenotypic variation in commercially relevant traits. There are key differences between MAS and GAS, however (Table 10.2), from perspectives of both research and operational implementation. Here, the terms GAS and MAS are used primarily to define differences relevant to typical forest tree breeding; we refer to MAS

as a technology for within-family selection only, in contrast to GAS, where selection can in theory be applied at the family level, in addition to individual genotypes within families, without prior pedigree information. These differences are not trivial with respect to the objectives and design of the underlying experiments needed to detect and quantify marker–trait associations. For example, for MAS, marker–trait associations are generally detected using pedigreed mapping populations, thereby maximizing linkage disequilibria between neutral markers and QTL that control detectable proportions of the phenotypic variation. For GAS, researchers basically accept and work with the existing levels of (dis)equilibria, however incomplete, that prevail in populations within which there are no recognized patterns of interrelatedness. Marker systems are likely to differ also, although in limited cases there may be some overlap. For MAS, selectively neutral marker systems adequate for development of moderate-density linkage maps and high-throughput (HTP) genotyping are considered satisfactory (e.g., RAPDs, AFLPs, microsatellites). For GAS, however, we consider it is more likely that polymorphisms associated with candidate gene sequences, i.e., SNPs, and insertions/deletions (indels), would be the marker systems of choice.

Table 10.2. Comparisons of requirements for MAS based on QTL detection and GAS based on association genetics in a tree breeding context

Attribute	MAS	GAS
Detection goal	Quantitative trait *locus* – i.e., chromosomal regions within specific pedigrees within which a QTL is located	Quantitative trait *nucleotide* – i.e., maximize causative sequence(s)
Genomic resolution	Low – moderate density linkage maps only required	High disequilibria within small physical regions usually needed (<2 kb) Linkage disequilibrium experiments: unrelated
Experimental design for detection	Defined pedigrees, e.g., three and two generation pedigrees/families, half-sib families	individuals (association tests), or large numbers of small unrelated families (transmission disequilibrium tests, TDTs)
Applicable to	Within-family forwards[a] selection only, within specific families where associations detected	Plus-tree selection, among- *and* within-family forwards selection, within reference population[b]
Marker neutrality	Neutral	Non neutral
Marker specificity	Non-trait-specific	Trait-specific[c]
Marker discovery costs	Moderate	Moderate for few traits, high for many traits
Prescreening[d] for functional association required?	No	Yes[e]
Opportunity to identify co-adapted gene complexes	Moderate	Good
Number of markers required	200–300 codominant markers per genome on average	>5 prescreened markers per gene on average, likely >5 genes per trait

[a]Selection among latest generations of breeding-population off spring.
[b]As defined by populations used for detection experiments.
[c]Except when polymorphism is in disequilibria with gene(s) controlling more than one trait.
[d]Prescreening defined as the need to select candidate sequences based on some *a priori* expectation of association or causation (e.g., candidate genes).
[e]Assuming lack of genome-wide, ultra-high density marker maps.

10.6 GENERIC BENEFITS OF GAS

Stromberg *et al.* (1994) classified generic benefits relating to the use of DNA markers for selection into three areas: earlier selection; cheaper, more cost-effective selection; and increased selection intensity. In the context of GAS in a tree improvement program these also apply, but for the sake of completeness, can be expanded. The following, partly overlapping areas are where we consider most of the potential benefits will be:

(1) *Earlier selection.* Perhaps the single most important limiting factor in plantation forest tree improvement has been selection age. The vast majority of characteristics do not adequately express their genotypic value until one-quarter to one-half of rotation age, which is a key factor influencing the long generation intervals typical of most tree breeding programs. GAS, like MAS, offers the tantalizing prospect of selecting at an emergent seedling stage, rather than waiting for up to many years for adequate trait expression. Such early selection can be used as a substitute for direct phenotypic selection, or as a complement in a multistage selection procedure, or simultaneously with information on phenotype. The net effect will be to increase selection intensity (see (3)). A further benefit, particularly in the cases of plus-tree and among-family selection, is the prospect of screening individuals without need to generate and evaluate offspring, which will further reduce generation interval by directly evaluating genotype.

(2) *Cheaper, more cost-effective selection.* Knowledge of the sequence variants and their effects on phenotype offers opportunity to select based on sequence only which could reduce or perhaps ultimately eliminate need for field screening. Field testing is one of the most expensive components of tree breeding programs, and sequence-based selection is likely to be cheaper, particularly for multitrait breeding objectives where expensive-to-measure traits such as wood properties are involved. Furthermore, advances in DNA technologies offer further reductions in costs in the medium term, whereas phenotypic measurements are likely to remain relatively expensive. One factor to consider, however, is the reasonably high cost of establishing marker–trait associations, which means that a large-scale breeding operation may be needed to justify use of GAS. Nonetheless, these costs can be reduced through various means such as pooling DNA samples (e.g., Germer *et al.* 2000). Moreover, the associations are expected to hold across a number of generations, so costs can be spread accordingly provided generation intervals are short. However, sample sizes necessary for detection of marker–trait association in LD populations may require at least several thousand genotypes for small-effect QTL for even modest levels of power and ability to infer association (Ball 2005; Chapter 8).

(3) *Increased selection intensity.* This can result partly from the low cost of producing young propagules that can be screened by GAS and partly from the higher-throughput evaluation capacity. Even with current moderate- to high-throughput genotyping technologies, there is capacity to screen far more genotypes than can be field-tested, at potentially much lower cost. Thus, genetic gains are likely to increase, particularly with multitrait breeding objectives that will tend to require larger numbers of selection candidates. In fixed-resource phenotypic-screening programs the addition of another trait into a breeding objective will typically incur costs in gain for any single specific trait unless the "new" and "existing" traits are strongly

and favorably correlated. Such a cost can be reduced with increased selection intensities, but in contemporary breeding programs this usually means more phenotypic evaluations (often on progenies) and possibly introduction of new genotypes into breeding populations. GAS could be used as a surrogate selection tool in these situations, although there may be the challenge of establishing the requisite associations simultaneously in several traits.

(4) *Reduced need for phenotypic selection.* The combined result of selection that is cheaper and/or earlier and/or more intensive may mean, in theory, at least, that GAS could ultimately replace phenotypic selection. This is based on the intriguing possibility that concomitant advances in genomics, proteomics, and metabolomics could eventually lead to development of predictive models that integrate information on gene sequences with information on environmental influences to predict phenotype, thus reducing reliance on phenotypic selection, and basing genetic selection entirely upon DNA sequence. The reduced reliance on field testing has several distinct advantages, including a reduction in costs and/or a concomitant increase in effectiveness of a tree improvement program via reallocating financial resources to other components of the operational program. Field testing is one of the most costly items in a tree improvement program, not just in terms of data collection, but also trial establishment and maintenance, and to a lesser extent, analyzing data and maintaining records. While the need for various forms of field experiments will likely persist even once all genes are sufficiently well characterized with respect to effects on trait variation (e.g., genetic gain trials), significant cost reductions should become possible.

(5) *Increased flexibility for operational evaluation and selection of genotypes.* Knowledge of phenotypic value associated with specific DNA sequences that can be applied across unrelated genotypes expands the scope of potential application. GAS can be applied to plus-tree selection, as well as among- and within-family selection, in contrast to MAS, where associations between marker alleles and trait variability are family-specific, and are thus applicable to within-family selection only. Therefore, in theory a genetic value can be placed on any specific individual based on DNA sequence information, where sequence has some nonzero association with trait value. While implementation of GAS would lead to more field trialling initially because of the need to find sufficient associations between markers and traits, ultimately, GAS could reduce need for "common-garden" testing, and allows new introductions to be evaluated without progeny evaluations (as is typically practiced).

(6) *Complementary/synergistic fit with both existing and new genetics technologies to enhance genetic gains.* Because various genetic technologies are available for forest tree improvement, in addition to an array of new technologies currently being developed, there are typically alternative routes to delivery of genetic gain. GAS potentially offers an additional technological route, in that it can either complement – or possibly supplant – phenotypic selection, but in addition, fits well with newer technologies. We describe in more detail in Section 10.9 the fit with new biotechnologies.

(7) *Prediction of genotypic value and enhanced opportunities for optimizing combinations of genotypic, site, and silvicultural characteristics.* Eventually, the knowledge of DNA sequences underpinning heritable variation could be combined with knowledge of key environmental and silvicultural influences to predict phenotypic characteristics. While this is a far-reaching goal, it is a tantalizing

possibility that knowledge of the causative nucleotides in combination with the extent to which environment affects the roles in particular characteristics could be combined to design combinations of genotypes and silviculture that optimize returns to forest growers. Such a capability would be extremely beneficial for designing genotypes with particular characteristics in mind, and would also aid silviculturists in designing genotype-specific regimes to maximize value, as well as a more optimal matching of genotypes to sites.

(8) *Provision of experiments that could ultimately lead to identification of actual QTN.* Because GAS typically develops initially from correlation rather than causation, identification of causative QTN may not be necessary for selection. However, the candidate genes and experiments necessary for identifying which polymorphisms are associated with trait variation (described below) are also necessary components for identifying the actual QTN. This knowledge is a key step in elucidating molecular mechanisms underpinning quantitative variation, and this information could be used to design new strategies for creating and utilizing variation, by identifying genomic regions that, when further altered, could lead to creation of additional useful variation.

(9) *Provision of experiments to answer questions about the genetic structures of forest tree populations and provide key information that could assist in management of breeding populations.* A benefit of the experimental infrastructure established for association genetics is the opportunity to generate genome-wide information that could be used to elucidate genetic phenomena such as presence/absence of trait variation, population structure and history, and evidence of selection. The genetic architecture of trait variation can be defined as the frequencies, location, magnitude, and mode(s) of action of QTL/N effects underpinning quantitative traits. While QTL mapping has been very informative in this regard, the results are relevant only to the pedigree(s) used, rather than to whole populations. Association genetics may therefore be more relevant for understanding the genetic landscape of trait variation in forest trees. While large, essentially panmictic populations cannot be expected to have appreciable across-family linkage disequilibrium, cryptic structuring may exist which generates significant disequilibrium. For example, localized population bottlenecks, followed by coalescences, could easily cause this. Such LD could provide valuable clues to "metapopulation" history. Despite wind pollination, various factors can generate population structure in conifers (Mitton 1992). Interesting possibilities of structure exist in populations derived from recent admixture. In *P. radiata*, the exotic, domesticated "land races" still have large elements of the wild state. Interestingly, they evidently represent a genetically recent fusion of two of the native populations, Año Nuevo and Monterey (Burdon 1992; Burdon *et al.* 1998), which may provide a basis for some admixture disequilibrium. A further benefit of association genetics is that DNA sequence data derived from both genic and nongenic regions can reveal much about genetic history of those regions. Departures from Hardy–Weinberg equilibrium could reveal presence of previously undetected genetic phenomena such as presence/absence of inbreeding. Indeed, genetic variance (and gain) estimates are based on assumptions regarding relatedness of parents used in genetic tests. Such data can be used to check these assumptions and provide empirical data for more accurate estimates. Similarly, sequence data from genic regions can reveal evidence of selection (see Section 1.3 for recent examples in forest trees). Such evidence – which can be generated on a

relatively small subset of genotypes – could be an effective prescreen for genes more likely to be associated with trait variations, although some caveats apply regarding power to detect effects of selection (Wright and Gaut 2005).

10.7 PREREQUISITES FOR FEASIBILITY

10.7.1 Basic Prerequisites for Operational Implementation

Successful application of association genetics for forest tree breeding must depend on the context of a well-structured breeding program. Genetic variation for economic traits is essential, and must be proven, while important genetic correlations between different economic traits need to be at least reasonably understood. Achieving this will entail major progress towards obtaining the populations needed for detecting associations between DNA polymorphisms and phenotypes. Efficient assays, which can be used on young trees, are important for this purpose, just as they are for conventional breeding. This will generally require new measurement technology, and/or easily measured juvenile traits that are good proxies for harvest-age economic traits. For wood quality, the SilviScan instrument (e.g., Evans 1994; Evans *et al.* 1999) has been developed to measure several detailed anatomical properties, and this has been complemented by an improved understanding of how such properties affect processing- and product-performance characteristics. Resistance to certain diseases can be assayed by inoculation trials of young seedlings (e.g., Powers *et al.* 1982). Very early evaluation for growth rate, however, can be very problematic: juvenile–mature correlations can be low, physiological variables can show highly nonlinear relationships with performance, and metabolite fluxes can be far more important than metabolite concentrations.

More specific requirements for applying GAS include quantitative capabilities, both in providing appropriate material to furnish phenotypic data and in managing, analyzing, and interpreting phenotypic and genomic data; access to HTP genotyping technologies; and good marker selection. This involves selection of candidate genes that could be associated with quantitative variation, and discovery and evaluation of important polymorphisms. We discuss each of these requirements.

Operational implementation will depend not only on meeting the various technical conditions listed above, but also on meeting organizational and even institutional requirements. Between the tree breeders and the genomic scientists there need to be close communication and considerable mutual education. Allocation of resources to the various parties will be a continuing challenge. A further challenge will lie in maintaining a strategic focus, whereby GAS and other new technologies can be used to best long-term advantage.

The total scale of undertakings for successful development and application of GAS will typically require collaboration between institutions, including industry, specialist research organizations, and universities. This will need to be achieved in the face of a climate of competitive bidding for research funding and the various pressures to appropriate Intellectual Property for individual organizations' own gain.

10.7.2 Quantitative genetic skills for experimental design and analyses

Effective application of association genetics for selection applications also requires both good experimental designs and analytical skills so that sufficient numbers of QTN

can be detected and utilized. These issues are discussed in more detail elsewhere in this book (Chapters 7 and 8). We cover components relevant to application of association genetics for tree breeding.

10.7.2.1 Experimental Design

A key prerequisite for GAS is the identification of DNA polymorphisms for selection. But what kind of experiments and what analytical methods are necessary? This has been covered to some extent in Chapter 8. so here we confine our discussion to issues relevant to tree breeding.

One of the few benefits of tree breeding is that unstructured (or loosely structured) populations already exist due to the nature of breeding programs, which usually consist of breeding populations with moderate numbers of heterozygous genotypes that show considerable genetic variation, despite being subject to phenotypic selection as a prerequisite to introduction in breeding populations. Moreover, such populations have usually been extensively progeny-tested, sometimes with clonally replicated progenies, for which phenotypic records have been generated for a range of commercially important traits. In addition, most programs maintain reasonable records of the geographic locations of the original first generation selections, as well as good knowledge of the range of genetic diversity represented in the naturally occurring populations – which may or may not contribute to breeding populations.

However it is necessary to bear in mind what information is needed from association tests that could be of use to breeders. Firstly, sufficient numbers of markers associated with QTL/N are required to obtain worthwhile genetic gains, implying the necessity for moderate–high power of detection of QTL/N. In addition, the genomic location of these polymorphisms and their magnitudes of effect, as well as modes of gene action and population allele frequencies, are also key pieces of information. Furthermore, it is necessary to account for population structure, as the impact of population structure can affect both the validity of any detected associations (Pritchard and Rosenberg 1999) as well as the estimates of gene-substitution effects (Deng 2001). Methods are available to do this (Pritchard *et al.* 2000; Thornsberry *et al.* 2001; Yu *et al.* 2006), and some experimental designs can account for such admixture (Allison 1997; Wu *et al.* 2002). In the relatively few studies undertaken to date there is very little evidence of population structuring in forest trees – no evidence was found in Douglas-fir (Krutovsky and Neale 2005) or loblolly pine (Brown *et al.* 2004b) which are both wind-pollinated conifer species, nor in *E. nitens* (Thumma *et al.* 2005). However, population structure has been indicated for other species. For example, Lagercrantz and Ryman (1990) reported presence of structure among populations of *P. abies* based on both allozyme and morphological (but not genecological) variability, a result at least in part, of population disruption during the most recent glaciation.

In order to determine appropriate experimental designs, it is germane to briefly review what is known about the genetic architecture of traits of commercial value in forest tree species. Numerous studies have been conducted using QTL mapping populations usually involving full- or half-sib families for most forest tree species of commercial value. For traits such as disease and insect resistance there are well-documented examples of major genes (Devey *et al.* 1995; Wilcox *et al.* 1996) although it is unlikely that resistance to pests and pathogens is solely conferred through major genes alone. For quantitatively inherited traits, which appear to be the norm for the majority of

commercially important traits, there has been some debate regarding the true nature of the underlying variation. Early studies involving relatively small populations indicated genetic variation was dominated by a few genes of moderate effect, however, these results were difficult to repeat, even in the same families (Wilcox *et al.* 1997; Sewell and Neale 2000). Interpretations of those early studies may therefore have been erroneous in that results are also consistent with genetic architecture involving genes of small effect only, similar to that described in corn (Beavis 1994), and subsequent verification, when done, have indicated this to be the case (Wilcox *et al.* 1997; Sewell and Neale 2000; Brown *et al.* 2003). Therefore for most traits, we contend that the underlying genetic architecture is most likely to be dominated by genes of relatively small effect contributing a few percent of the variation at most (e.g., Devey *et al.* 2004). An exception may be that interspecific hybrids could involve genes of moderate–large effect (e.g., Bradshaw and Stettler 1995), although small-effect genes may also have a role. Experimental designs for association genetics will therefore need to be cognizant of these architectures, particularly genes of small effect, if selection is going to be effective.

A number of different experimental designs could be used to detect associations between QTL/N and polymorphisms, such as an unstructured population consisting of putatively unrelated (or distantly related) genotypes; or combined with information on progeny (analogous to a TDT design, except using quantitative traits); or alternatively a hybrid QTL–LD population (see Chapter 8 and references therein). Some of these approaches have been evaluated in a manner more relevant to forest trees (e.g., Wu *et al.* 2002; Ball 2005). Furthermore, some of the genetic characteristics of forest trees parallel humans (e.g., high levels of heterozygosity, adverse effects of inbreeding, longevity), for which much has been written in regard to the theory and efficacies of specific experimental designs and analytical procedures, and are therefore relevant to tree species. We review some of this literature here, and refer the reader to Chapter 8 for a more extensive review.

A number of theoretical studies have been conducted, particularly in comparing designs with and without use of information from sibs. A somewhat unclear picture has emerged to date, however, partly because of differing assumptions and input values used for simulations. Long and Langley (1999) showed that for smaller-effect QTL (~5% of phenotypic variance), unstructured or random populations were more powerful than TDT-based designs, and that power increased more when greater numbers of individuals rather than markers were used. Moreover, they concluded that unstructured populations sample sizes ≥500 individuals would suffice to detect small-effect QTL assuming a Type-1 error rate of 0.05. A further and nontrivial finding was that equally large populations would be needed to verify any detected associations.

Wu *et al.* (2002) developed theory for combined linkage- and linkage-disequilibrium mapping, based on use of genotypic information from a single parent combined with genotypic and phenotypic information from offspring, analogous to multiple half-sib families, as in often used in breeding population testing. They compared different combinations of family numbers and sizes, and compared the power to detect a segregating QTL of large effect with an unstructured population without information from progenies. In contrast to Long and Langley (1999), they found that simulation results indicated that use of information from progenies was more powerful than unstructured populations only, particularly with low disequilibrium, assuming the same number of individuals genotyped. Results also indicated that few families with many offspring per family were more powerful than many families with few offspring. A key benefit of this

approach is that the use of progenies obviates the need to independently evaluate population structure. However, because these results were based upon a single QTL with a large effect (both additive and dominance terms equal to residual error), relevance of these results may well be limited, as individual QTL effects are typically much less than residual variance. Therefore these results would need more careful evaluation using a range of QTL effects more relevant to known genetic architectures.

Most of the above studies have involved estimating power with comparison-wise Type-1 error rates in the region of 0.01–0.05. However, such values may be problematic in reality because actual results in that range of *P*-value may not be equate to strong evidence for an association. Using a Bayesian approach based on theory originally developed by Luo (1998), Ball (2005) calculated that *P*-values in the range of 0.01–0.05 actually represented weak evidence against an association for sample sizes in the 432–1,200 individuals in an unstructured population. *P*-values in the range of 10^{-4} would be more indicative of evidence for an association, assuming high prior expectation for an association (see Chapter 8). This also implies that larger sample sizes than those generally reported above would be needed.

Ball (2005) also showed that very large sample sizes are necessary for high power (0.9) of detection of QTL with small effects (explaining 1–5% of total variance) when using either candidate genes or a genome scan in an unstructured population. To obtain high power with strong posterior odds (Bayes Factor >20) with moderate disequilibium ($D' = 0.1$), sample sizes ranging from 6,800 to 40,100 would be necessary to detect QTL of 5 and 1% effect, respectively. Such sample sizes are based in part on relatively low prior odds, which may be increased through generation of additional experimental and biological information on specific genes (e.g., expression profiles, evidence of selection), therefore sample sizes could be reduced. However, even with relatively high prior odds, sample size requirements will still be relatively high. Furthermore, Ball (2005) quantified the power to detect QTL when marker and QTN frequency differed. Even with very large sample sizes (19,200 and 38,400 genotypes), there is relatively low power to detect rare QTN with intermediate marker allele frequencies, even when in almost complete disequilibria. This is an important consideration, given that long-term genetic gains are driven by low-frequency QTN, along with mutations that arise during the selection period.

What can be concluded regarding optimal experimental designs based upon the work described above, and what are the implications for tree breeding programs? Firstly, moderate- to large-effect genes are likely to be easily detected using material from existing breeding populations, as long as there are sufficient numbers (200–1,000 putatively unrelated genotypes with phenotypic records available). For smaller-effect genes, which are likely to dominate the genetic architecture of quantitative traits in particular, much larger sample sizes are likely to be needed; therefore augmentation of existing breeding populations with genotypes from natural populations may be necessary. The implication here is that such augmentation will require common-garden experimentation, which is time-consuming, and could delay or militate against use of association genetics. Furthermore, maintenance of genetic diversity of nonbreeding population genotypes is also a necessity. Optimal designs with sufficient power for detection of small-effect QTL will therefore need to be ascertained in the context of tree improvement programs, most likely necessitating numerical simulation on a case-by-case basis.

Experimental designs could nonetheless be incorporated into tree improvement programs even if additional genotypes are necessary: such populations will be useful for other purposes (such as parameter estimation for new traits), particularly if progenies are incorporated. Indeed some redesign of breeding strategies may well be necessary if genetic tests are to take effective advantage of association genetics.

10.7.2.2 Analytical Methods

A further requirement for successful implementation of GAS is the use of appropriate methods for analyzing results from association tests. Some parameters such as population structure, linkage disequilibrium, and evidence for natural selection are estimable from sequence data generated on a small subset of genotypes, which could be used as a prescreen for a larger association test. For the latter, it would be necessary to use only those polymorphisms not in LD with other polymorphisms ("haplotype-tagged" polymorphisms), which would be determined in such a prescreen.

A number of analytical approaches could be used, depending on the experimental design. For most experimental designs, population structure will need to be tested for and, if present, taken into account. After examination of evidence for population structure, a number of parameters need to be simultaneously estimated for effective application. These include gene-substitution effects, population structure, frequency of both marker and QTN, mode(s) of gene action, and genotype × environment interaction (if present). Methods for estimating such parameters are discussed more fully in Chapter 8.

Preliminary analyses for detection of marker–trait associations can be undertaken using simple regression or ANOVA-based approaches, which can be undertaken in a variety of software packages. Specific software such as PowerMarker (www.powermarker.net) and TASSEL (www.maizegenetics.net) can also undertake limited analyses. While these may be useful for indicating a potential association, more detailed analyses are required for adequate statistical inference and gain estimates. While maximum-likelihood methods have been developed to estimate key parameters (e.g., Wu and Zeng, 2001), the estimates tend to be 'prone to selection bias' if the same data are used to estimate parameters as well as detect associations (Ball 2001), and thus may be unreliable. Overestimation of some gene-substitution effects has been reported for QTL mapping (Beavis 1994; Ball 2001). Bayesian methods may be more appropriate here (see Chapter 8). Methods to reduce or eliminate selection bias have been developed for QTL mapping in pedigreed populations (e.g., Ball 2001), and extension to commonly used experimental designs for association genetics may be useful.

A further consideration for the experimental design is the actual nature of the molecular data. Data can come from haplotypes (such as directly sequencing each copy of a gene in the diploid genotypes, or genotyping haploid tissue), or directly obtaining marker genotypes at each polymorphic site without surrounding sequence information. The key difference here is that with haplotypic data, the phase relationships between polymorphic sites are known for each copy of a polymorphic region in an individual. In contrast, for marker-genotype data, phase relationships are not known. Haplotypic data are considered, by some, to be more powerful for detection of marker–trait associations, as information from multiple polymorphisms can be condensed into discrete haplotypic classes (e.g., Lynch and Walsh 1997). Long and Langley (1999) found that marker-based methods were as powerful if not more powerful in some situations, than "simple"

haplotype-based methods, and simulations suggested lower Type-I error rates. Genotypic data are sometimes cheaper to obtain, as direct sequencing is not necessary.

10.7.3 Access to Appropriate Genotyping Facilities

HTP facilities are necessary for sequencing and genotyping, for both detecting associations and operational selection. Extensive sequencing and resequencing are required, even if only a small subset of genotypes are used for initial scans of candidate gene regions. HTP genotyping is an obvious prerequisite, given the large amount of data generation necessary for adequately conducting powerful association tests. Whether or not specific breeding programs choose to develop "in-house" capacity or choose to outsource this component will be a choice made on a case-by-case basis.

10.7.4 Marker selection

Appropriate marker systems are an obvious prerequisite for detection of associations between marker and trait variation, along with HTP genotyping for selection purposes. But how many and what types of markers are needed for association genetics? Requirements for association genetics and subsequent selection applications differ substantially from those for QTL mapping (Table 10.2), primarily because disequilibrium per base pair is likely to be substantially less for apparently unstructured populations versus pedigreed QTL mapping populations. Forest trees present specific problems here. The outbreeding behavior, in particular, means that regions of LD tend to be very small, typically in the range of 0.3–2 kb (Table 10.1). In addition, gymnosperms in particular have typically large genomes (Murray 1998) adding further complications.

Several approaches could – at least in theory – be used to select polymorphisms to detect marker–trait associations. These include:

- *Use of the same markers as those developed for QTL mapping*, for example, SSR and EST markers. For most forest tree species, total number of markers used for linkage and QTL mapping is generally in the range of several hundred to low thousands, and therefore insufficient to achieve adequate resolution for association mapping given the typically small stretches of LD. Moreover, many of these loci are likely to amplify phenotypically neutral regions of the genome, or at least do not appear to be strongly correlated with trait variation even in specific pedigrees where disequilibrium is much greater, so such markers are unlikely to be adequate for association genetics. Nonetheless, these markers can be useful for revealing population structuring, which needs to be taken into account in association tests. It is also possible that a small number of loci could be in disequilibrium with QTN.
- *Whole-genome sequencing* (and resequencing), such as that undertaken in humans and a small number of important domesticated animal species. This involves complete (or near-complete) genome sequencing, followed by *in silico* polymorphism identification, after which a subset of polymorphisms are chosen for whole-genome scanning based on the patterns of observed disequilibrium. Such an approach is costly and technically challenging with existing sequencing technologies in highly repetitive and large genomes such as gymnosperms. For example, in *P. radiata*, assuming a 1C genome content of 22×10^9 bp (Murray 1998), with a 1,000 bp haplotype block size on average, we calculate that 22

million markers would be needed. To genotype a 1,000-tree population at a cost of US 5 cents per marker per genotype, would cost in excess of US $1 billion! Even for the smaller angiosperm genomes, assuming 1% of the size of the above example with a similar haploblock size, cost is still well beyond the reach of most tree breeding programs.

- *Partial genome sequencing* (and resequencing) of specific (rather than entire) genomic regions. This is a more limited approach than that described above. Genomics technologies targeting gene-rich areas such as Cot-based selection methods, which target low-copy-number regions, may be an alternative to whole-genome sequencing/resequencing. Such an approach may be more financially acceptable, particularly for hardwoods which have smaller genomes than conifers. Further research is needed, however, to determine if such methods could be effective at targeting QTN, as success of this approach is predicated on whether or not the QTN are located mainly in either low-copy-number regions or regions of low methylation. Genic regions within such "short" genomic stretches would need to be identified, which could be done using gene-searching algorithms, and/or alignment with relevant EST databases. QTN discovery via this method could still be expensive, however, as high polymorphism rate and low LD per base pair mean that SNP discovery would be expensive. Moreover, gene families could further complicate this approach in the more complex, larger genomes such as in conifers.
- *Preselection of candidate genes*, followed by polymorphism discovery, within these genes as well as the surrounding regions. We consider marker selection using this approach more promising than all of the above, primarily due to cost. With this approach, nucleotide variants within the transcribed sequence and the surrounding regulatory regions would then be assayed for association with trait variation. Such candidates could be selected using various approaches, which are described in more detail in the following section.

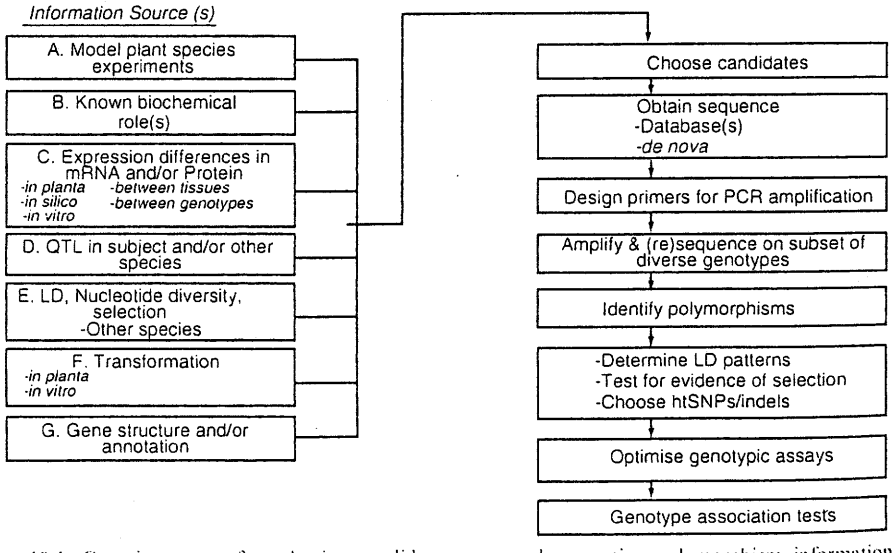


Figure 10.1. Generic process for selecting candidate genes and generating polymorphism information on association tests.

The overall process from gene selection to generation of genotypic data on association tests is described in Figure 10.1. It should be noted that this process assumes population structure has been already evaluated.

10.7.4.1 Candidate-gene selection

Generic methods for candidate gene selection are described in more detail elsewhere in this book. Here, we outline more specific approaches that could be considered, noting that except for *Populus* and *Eucalyptus*, there will be very little genome-wide data available for subject, although for most commercially important genera extensive EST sequence information is available, if not in the species of interest, then in a closely related species. Note, too, that selection of candidate genes can be based on more than a single criterion, although the relative efficacies of the various criteria are not yet known. Such criteria include:

- Choosing orthologous genes to those in model plant species that have been shown to have a role in traits of interest (Figure 10.1, Box A). For example, Thumma *et al.* (2005) found that polymorphism in an intronic region of a *CCR* gene was statistically associated with microfibril angle in *E. nitens* in a small association population. This gene was chosen because it is homologous to the *IRX4*-causing *CCR* in *A. thaliana*. However, it is not yet known to what extent and which plant model systems can predict roles of the homologous genes governing endogenous variation in forest tree species. If, in the more complex conifer genomes, there is a greater tendency for large gene families affording some degree of functional redundancy, information from short-lived angiosperms could be of limited value.
- Similar to the above, but using information on mutations and knowledge of gene sequences (and expression patterns of the sequences) from other forest tree species. For example, while an annual-plant model system could have limited applicability, a model system based on a woody perennial (e.g., *Populus*) could be more useful. In either case, the role of comparative genomics is crucial.
- Endogenous genes based on known or suspected role(s) in relevant biochemical pathways (Figure 10.1, Box B), e.g., genes involved in lignin biosynthesis as a preliminary choice to investigate natural variation in lignin chemistry. Much molecular information has been generated on this topic, and the key regulatory genes have been identified (e.g., Huntley *et al.* 2003). Such an approach has been used in mammalian systems, although with mixed success. For example, the Booroola gene in sheep (*FecB*), which causes elevated fecundity, was initially thought to be due to natural variation in FSH, a gene encoding a follicle-stimulating hormone. However, subsequent linkage analysis showed otherwise (Dodds *et al.* 1993), which was later verified by identifying the causative gene.
- Information from transcript profiling (Figure 10.1, Box C), identifying genes whose expression patterns are correlated with specific traits. A number of differential-expression technologies have been developed, including micro-arrays, cDNA–AFLP and similar approaches, and are now extensively used, although not as tools in breeding programs. Such technologies do reveal many candidates – possibly too many to be used as a screening tool alone. Moreover,

heritable variation may arise for reasons other than differential expression of allelic variants. In reviews of cloned plant QTL, only three of ten QTL whose mechanisms were determined were shown to be due to differential expression (Salvi and Tuberosa 2005). Nonetheless, combining expression-profiling technologies with QTL mapping shows considerable promise. A number of studies have shown this hybrid approach to be useful in identifying the genes potentially causing trait variation (Wayne and McIntyre 2002). For example, Kirst *et al.* (2003) reported a candidate gene underpinning a major-effect QTL in an interspecific *Eucalyptus* hybrid. Furthermore, Cato *et al.* (2006) reported a dehydrin gene associated with both wood density and growth rate in *P. radiata* that showed allelic differences in transcript abundance in different wood-forming tissues within the same genotype.

- A variant of the above, using proteomics rather than mRNA populations. The lack of complete correspondence between translation and transcription may be a useful means to eliminate those genes that are less likely to contribute to trait variation. Moreover, this approach has promise in that it may also identify gene products whose contribution to trait variation may be due to reasons other than differential expression (e.g., protein folding, etc.). Such an approach has not been extensively tried yet, at least not in forest trees.
- Expressed genes that consistently colocalize with QTL regions in multiple pedigreed QTL mapping populations, either within or across species (Figure 10.1, Box D). In practice, this could be of limited value, as confidence intervals around QTL are likely to cover much of a chromosome, particularly where sample size is limited (Dupuis and Siegmund 1999). Nonetheless, pedigreed mapping populations could be used as an additional screening step. However, caution is recommended: small–moderate size QTL mapping populations could be of limited value as they may not be sufficiently powerful to detect QTL, therefore the lack of association is not conclusive; or else the QTN may not be segregating in the particular pedigree(s) being used. If using information from another species to infer trait association in the subject species, then evidence for nonrandom colocation of QTL for traits of interest should be determined *a priori*, otherwise use of information from other species will be of little value.
- Genes that have been shown to be associated with variation in traits of interest via association genetics in other species (Figure 10.1, Box E). Caveats regarding utility of transferability of QTL across species mentioned above also apply. Nonetheless, marker–trait associations that occur in homologous sequences across species may also serve as independent validation of associations.
- Use of genetic transformation to determine potential role(s) of candidate genes (Figure 10.1, Box F). This approach involves modification of endogenous gene function in some manner, e.g., enhancer trapping, RNAi, over-expression, etc. However, for forest trees, such approaches have limited promise, particularly in species where trait expression takes years, and/or have low transformation efficiencies. Other technical problems could also be limiting, e.g., sense suppression in the case of over-expression. Regulatory issues could also impact, particularly where field trials are necessary. However, this approach may be useful in cases where *in vitro* or early-assay systems have been developed, particularly where transient expression can result in a discernable phenotype.

As we learn more about the function of specific genes alone, and in concert with other genes, other criteria are likely to be added to the above list. Moreover, as more information from each of these sources becomes available, it will be possible to evaluate the relative efficacy of each of these criteria. Suffice to say, the roles of structural and comparative genomics, proteomics, molecular biology, as well as knowledge of physiological roles of specific genes, are crucial. Very few of these skills are currently utilized by, or available within, current tree breeding programs.

Of interest too, are the identity and nature of regulatory regions associated with candidate genes (Morgante and Salamini 2003; Paran and Zamir 2003). Because trait variation could be a result of gene regulation, there is a need to ascertain – via *de novo* sequencing if necessary – regulatory sequences. This should be easily achievable for promoter sequences in close proximity to open reading frames, but may be more difficult for transacting enhancer elements, particularly if such sequences are not known *a priori*.

10.7.4.2 Polymorphism discovery and evaluation

Following selection of candidate genes, further evaluations are required (Figure 10.1). These involve resequencing on a subset of genotypes to identify specific polymorphisms, and to determine patterns of disequilibrium before choosing a subset for testing for associations with traits of interest. SNPs and indels are the most likely forms of polymorphism to be useful, although other forms (such as repeat sequences) could also be useful. Polymorphisms that need to be detected and evaluated include not only those nonsilent substitutions in coding regions, but also polymorphisms in noncoding regions such as introns, and 5′ regulatory regions, particularly if they are not in disequilibrium with polymorphisms in coding regions. Patterns of disequilibria will need to be determined on a gene-by-gene basis, unless some general patterns emerge that can be applied across all genes. The relatively short span of disequilibria observed in forest trees (Table 10.2) – at least by some statistics, such as r^2 – will necessitate extensive SNP discovery and evaluation throughout the relevant genic regions.

Detection of SNPs and evaluation of disequilibria require genomic sequence information, some of which can be obtained from EST databases, but regulatory and intronic regions will need to be sequenced from genomic DNA. This step – polymorphism detection – is likely to be very time-consuming and labor-intensive, particularly in species where little EST and/or gDNA sequence information is available, and may well limit the rate of implementation, as individual polymorphisms will need evaluation and assays for chosen SNPs will need to be optimized for large-scale genotyping. For most forest tree species, technologies are needed that expedite polymorphism detection and resolution without the need for extensive sequence information.

It may therefore be useful to implement further marker-selection criteria at this point, prior to extensive SNP optimization and/or resequencing. Possible criteria include whether the sequence data generated reveal any evidence indicating a possible role in trait variation – such as evidence of selection, which can be obtained from examining patterns of nucleotide substitution in coding and noncoding regions for example. For example, Cato *et al.* (2006) reported elevated levels of nonsynonymous substitution in a dehydrin gene in *P. radiata* that had previously been shown to colocate with wood density QTL, and was subsequently shown to be associated with both growth rate and

wood density in an association population consisting of 1,700+ genotypes. Whether or not such criteria will be broadly effective is yet to be determined, in part because some QTN may not be under natural selection, yet still of use for artificial selection.

Once polymorphisms are detected and optimized for genotyping, a subset of polymorphisms will need to be selected and screened across some form of association population for which phenotypic data are available. Numbers required per gene (and associated regulatory regions) will depend upon the number of statistically independent regions per base pair and the size of the region being evaluated. It may therefore be necessary to screen tens of polymorphisms per genic region, although Krutovsky and Neale (2005) estimated less than ten would suffice for all but large genes. Also, because the size of the populations is likely to be in the order of many hundreds to thousands (below), high-throughput SNP genotyping is likely to be necessary. A range of technologies are available for this, and technology developments in this area are ongoing. Access to such technologies is obviously required, at affordable cost.

10.8 HOW MIGHT *GAS* BE INTEGRATED INTO A TREE IMPROVEMENT PROGRAM

The generic advantages of using association genetics in tree breeding have already been stated (cf. Stromberg *et al.* 1994). For effective use there are many possibilities. Some of the issues will be common to both MAS (including marker-based and marker-assisted selection) and true GAS based on QTN, and some will be specific to one or the other. To be effective, use in tree breeding of nucleotide–trait associations derived from association genetics must be integrated with essentially the existing tree improvement practice. Such practice includes the arrangement and structuring of breeding populations, and the manner in which genetic gain is delivered into plantation forests. For the future, the practices can be modified as true GAS becomes possible.

Tree breeding differs from much traditional crop plant breeding because of various factors, including relatively little history of domestication, moderate–high levels of genetic load, and long generation intervals imposed by slowness to reach reproductive competence and/or late expression of trait values. Forest tree breeding tends therefore to take a population-based approach involving many genotypes, where populations are usually structured into a hierarchy (Burdon 1988):

- At the lowest level are unimproved *gene resources* (essentially undomesticated genotypes).
- From these, the next level, the *breeding population*, is or already has been chosen.
- From which in turn the best genotypes are chosen (usually progeny-tested) for the *production population*, from which planting stock is derived for forest plantations.

This hierarchy of populations is schematically like a pyramid with the widest genetic diversity at the base, and the narrowest genetic variation at the top level of genetic improvement. Within this scheme, there can be many variations and refinements. Movement of genetic material will tend to be very much up the hierarchy, in the nature of replenishing genetic diversity in the upper levels.

At the start of a breeding program, before any progeny testing, the production population and the breeding population are often one and the same. Thereafter, the

breeding population becomes the "engine room" for cumulative genetic advance, building up frequencies of favorable alleles through successive cycles of mating, genetic recombination, and selection. For clonal forestry, clonal selection will typically be done within crosses between top-ranked parents which may be common to both the breeding population and existing seed orchards.

To complicate matters, tree breeding typically involves multitrait breeding objectives, and some programs also develop specific breeds that focus on improving differing sets of traits (Jayawickrama and Carson 2000). Application of GAS in tree improvement programs needs to fit into this general framework in a cost-effective manner. We will now consider potential applications of GAS in the context of such population hierarchies.

10.8.1 Plus-Tree Selection Applications

In programs where new plus-tree selections are required, GAS may be useful as a prescreening tool either to increase selection intensity, or to cull candidates down to those of sufficient promise to warrant costs of testing, and of forwards selection among offspring. Here, GAS has, in theory, the advantage of favoring selection well before full phenotypic expression, therefore increasing the available number of selection candidates. However, this may be constrained by the cost of phenotyping relative to genotyping, plus the desideratum of ascertaining marker–trait associations for the multiple traits that comprise a breeding goal. Nonetheless, marker–trait associations could be accumulated over time from association tests, and utilized as they become available, thereby increasing scope for adding new material into breeding populations. Similarly, genotypes could be identified for immediate deployment, in addition to incorporating them into breeding populations – assuming propagation systems exist to cost-effectively multiply selected genotypes without detrimental effects of maturation. For instance, in response to a biotic crisis (e.g., outbreak of a new disease or pest) GAS could be directly applied to identify genotypes more likely to be resistant to the pathogen or pest, rather than undertake laborious phenotypic screening. Specific genes could then be integrated more quickly into the relevant populations. Prospects for widespread application of GAS for plus-tree selection may be limited in practice; however, as population sizes for detecting associations would most likely exceed those required for breeding population advancement. Moreover, knowledge of nucleotide–trait associations may come to hand too late for fresh plus-tree selection, especially with traits of late expression.

10.8.2 Breeding Population Applications

Breeding populations in forest trees tend to comprise many genotypes, sometimes exceeding 1,000 parents, most of which are putatively unrelated plus trees and/or their offspring. Co-ancestry is usually minimized, to avoid deleterious and sometimes unpredictable effects of inbreeding, often via use of sublines (Burdon and Namkoong 1983). Substructuring of breeding populations is often undertaken, utilizing "main" and "elite" populations, generally with more intensive data gathering and selection in the smaller elite populations, to secure genetic gains sooner than in the main populations. Phenotypic evaluation in breeding populations is usually done on offspring that are planted in common-garden genetic tests, which allow breeding population advancement

by forwards selection for the multitrait criteria. Backwards selection, from progeny-test results, is also used to rank parents, particularly for production populations.

For breeding population advancement, the same marker–trait associations as might be used for plus-tree selection described above could be used for selecting among and within families, to increase selection intensity, as an early selection tool, and/or to reduce costs. However, even within breeding populations, specific applications will be context-dependent. For example, in main populations, which are generally less intensively managed than elite populations, GAS could be used as a surrogate for more expensive-to-measure traits. Here, phenotypic data could be generated on cheap-to-assay traits (e.g., growth rate) and GAS used for more expensive or later-expressing traits (e.g., certain wood properties). However, for the time being, DNA polymorphisms are likely to characterize less additive genetic variation than phenotypic records, resulting in potentially less gain for traits selected just on marker information. Such a reduction could be offset by increasing selection intensity among, and particularly, within families. Trade-offs will need to be carefully evaluated, initially at least via simulation.

For any breeding, an ideal is saving rare or low-frequency QTN that have current or contingently favorable additive effects. Such alleles can be the key to longer-term genetic gain and/or coping with a biotic crisis. For detecting, preserving and increasing the frequencies of these QTN, instead of losing them to genetic drift, GAS may be crucial. However, such a pursuit may well be deemed too expensive for breeding programs that are dominated by shorter-term financial imperatives.

In elite populations, with the fewer families for intensive measurement and selection, opportunities may exist for more intensive selection and faster turnover of generations. For combined among- and within-family selection, there is more scope to increase selection intensity within families. Because association tests identify markers in strong disequilibria with QTN, it may be relatively easy to detect pedigrees within which the predominant linkage phase is reversed. Undetected reverse-phase linkages are likely to be serious within small elite populations, or any other small breeding groups within the breeding population; simulation would again be helpful in quantifying potential reductions in gain.

Reducing generation intervals through use of GAS would depend on the trees becoming reproductively competent before trait expression. However, if markers or actual QTN were used as a surrogate for trait expression, genotypes could be screened as soon as sufficient tissue can be spared for DNA assays, even in germinating seedlings. Some conifers, in particular, are typically reproductively competent before selection age for at least some commercially important traits, creating a real potential for use of GAS to shorten generation interval. However, this would require marker–trait associations that explain substantial additive genetic variance for at least some important breeding goal traits. While this could one day be achieved, it is currently more likely to have associations that explain only a proportion of additive variance for just subset of traits. Thus, trade-offs between expected gain per generation and rate of generation turnover will need to be carefully evaluated.

It is more likely that, in the shorter term at least, selection in elite breeding populations would be implemented in a multistage approach, using marker information as an early screening tool, followed by phenotypic records. Such an approach could either increase selection intensity (by screening more genotypes), or reduce costs of phenotypic evaluation by short-listing genotypes for field testing, to achieve the same gain. Alternatively, using GAS to select for later-onset traits – if the nucleotide–trait

associations are established – could reduce generation interval, by concomitantly using phenotypic records on the earlier-expressed traits. A simple example could be in breeding objectives that incorporate both growth rate (if it is only expressed well at an advanced age) and, say, resistance to a disease for which empirical phenotypic screening is possible in very young seedlings.

There are other generic breeding population applications for GAS, which apply alike to both main and elite populations. These include more powerful selection via correlation breakers,[4] reselection, and as a surrogate for later-onset and/or expensive-to-measure traits. Such applications, while generic in nature, seem appropriate for where the need is greatest – more likely in elite populations.

Selection for recombinants of known QTN that break adverse genetic correlations between breeding goal traits is especially attractive. Detection of such recombinants would not require field testing, and can involve many more genotypes than could be field-tested, thus raising the probability of encountering the desired correlation breakers. Such genotypes would then need field testing, as confirmation, which would be done anyway in breeding population advancement.

A challenge will exist in applying GAS to new breeding goal traits in breeding populations. Tree breeding not only usually involves multitrait breeding goals, but also new traits are sometimes added to breeding goals in response to changes in market perceptions and values. Information for establishing the requisite associations for using GAS may be already available, even if the trait was not originally part of the breeding goal; otherwise, the major effort of fresh association tests may be needed. Alternatively, existing association tests may be screened for those new traits, and any subsequent association used for backwards and/or forwards selection, rather than extensively screening multiple progeny tests over successive generations for the same traits. For selecting a new trait, the greater selection intensity allowed by forwards selection would be very attractive, but at the risk of a new generation's decay of LD. As usual, correct choices of candidate genes will be key to making this approach cost-effective, especially finding the polymorphisms in strong LD with the QTN if not the actual QTN.

Related to this, is the potential to use GAS as a surrogate for phenotypic evaluations that are either expensive or involve destructive sampling. While establishing associations between markers and traits would of course require expensive phenotypic evaluations as part of the operational development; it may well be cheaper to use this route than to continue "trawling" numerous genetic tests over several cycles of breeding. Where assessment is necessarily destructive, there may be limited opportunity to measure progeny tests because of their inherent value for assaying other traits; therefore, GAS could be used as a surrogate for destructively sampled traits – if the requisite associations have already been established. Clonal replication of individual offspring, however, would effectively avert loss of material to destructive sampling.

QTN conferring resistance or tolerance to specific pests or pathogens may be particularly amenable to GAS. Pathotype-specific resistance genes of large effect are known in forest tree pathosystems (Kinloch *et al.* 1970; Wilcox *et al.* 1996), and in some cases are of great commercial potential despite their specificity. Identifying the QTN underpinning such pathotype-specific resistance, or finding polymorphisms in strong

[4] The type of correlations that can in principle be attacked effectively in this way would be correlations resulting from important chromosomal linkages that are persisting following fusion of differentiated ancestral populations, rather than correlations stemming from pleiotropic effects

disequilibrium with these QTN, has the benefit of obviating the need for screening families with specific pathotypes, to determine which families carry which resistance genes. Combining or "pyramiding" different resistance genes, preferably within the same individuals, can promise resistance that is durable against mutations and genetic shifts in the pathogen (Burdon 2001). Thus, phenotyping costs can be much reduced, as well as time required for manipulation of frequencies. This may be a great advantage in the event of a biotic crisis where low-frequency resistance is required to quickly combat a new disease or pest. The advantage would be increased by the desirability of pyramiding different resistance factors. Genotypes carrying such QTN can be identified in the breeding population (including directly estimating QTN frequencies), enabling among- and within-family selection to be carried out over a large proportion of the breeding population. In such circumstances, it is likely that at least some of the resistant genotypes will be suboptimal for other traits, so GAS might be used to select for other properties to reduce the loss in genetic gain.

Despite the prevalence of inbreeding depression in forest trees, use of inbreeding as a breeding tool has attractions because it can theoretically amplify the expression of additive gene effects (e.g., Burdon and Russell 1999; Russell *et al.* 2003). In most species, however, the challenge will be to "purge" highly deleterious recessive alleles ("hard" genetic load) that threaten viability and/or often mask the expression of favorable additive gene effects in inbred lines (e.g., Williams and Savolainen 1996). MAS has promise for such purging, because QTL effects of hard load should be relatively easy to detect in individual pedigrees in order to purge such alleles even in the heterozygous state (cf. Kuang *et al.* 1999). Use of GAS in this way, however, may not really work, because such genetic load almost certainly represents alleles that are individually rare but occur at very many loci and are therefore very unlikely to be involved in any general LD.

10.8.3 Production and Deployment Populations

Production populations comprise the genotypes that either provide seed for deployment into plantation forests, or are used for large-scale vegetative propagation for clonal forestry. These populations usually have a few tens of genotypes at any one time, and actually represent subsets of the breeding populations and are subject to most of the same considerations as the breeding populations for the applications of GAS. As subsets, they represent a relatively narrow genetic base compared to the breeding- and gene-resource populations. Related matings are avoided as far as possible, to avert inbreeding depression. Various systems are used to deliver commercial planting stock. Some programs use open-pollinated seed orchards, to produce seedlings. Other programs use control-pollination technologies, where top genotypes are pollinated with pollens from either single or multiple parents. Seed from these either provides seedlings for planting stock, or is vegetatively multiplied as nursery cuttings or as plantlets raised from *in vitro* culture, but, despite the average level of genetic improvement, this still produces uncharacterized segregating offspring genotypes. For clonal forestry, genotypes produced by intercrossing top parents are subject to a further round of testing and selection, before identifying and mass-propagating top clones for deployment.

Production populations are of key importance, as it is these populations from which seed and plant producers obtain most of their revenues, thus additional costs associated with this form of selection can be offset in a shorter time period than the breeding

population applications, as few if any products are delivered to forest growers directly from breeding populations. Furthermore, there is continual pressure on breeding programs to deliver gains to commercial plantations faster and/or at greater rates. Production populations are therefore more likely to be target populations for applying GAS, at least in the shorter term.

GAS, along with its variants, has obvious possibilities for selecting individual offspring for clonal forestry and/or subsequent vegetative amplification of a narrow range of genotypes – such as in situations where "family forestry" is combined with vegetative amplification. The parents – while they may already have been selected with the aid of GAS – will almost certainly still be highly heterozygous, so the expected genetic variation within any sort of family will be considerable for most quantitatively inherited traits. Where GAS is based on markers in LD with the QTN rather than on the QTN itself, response to selection of a limited number of clones in a limited number of families could be very vulnerable to reversals of the prevailing linkage phase, especially as this material will represent one more generation for decay of LD to occur in. On the other hand, the small number of families should make it relatively easy to verify linkage phases in individual pedigrees. The results of Wilcox *et al.* (2001) indicate that this scenario could be cost-effective in the context of within-family selection (MAS) based on neutral DNA markers.

In selecting clones for clonal forestry the potential of GAS for selecting rare recombinants, especially involving QTN, looks particularly attractive, because such recombinants could not be produced reliably through sexual reproduction within any reasonable timeframe.

Where new traits must be addressed in the breeding goal, the emphasis in selection for production populations is likely to shift in favor of forwards selection over backwards selection, which is likely to favor use of GAS if the appropriate associations can be established.

For disease resistance (and possibly some cases of insect-pest resistance), the potential of GAS for advantageous pyramiding of resistance factors looks especially valuable. This could be all the more important where durability of resistance may depend on certain individual resistance alleles remaining at minority frequencies, in pyramiding at the level of the population rather than the individual genotype.

10.8.4 Summary: Selection Application in Forest Tree Species

This section has outlined generic applications of marker–trait associations obtained from association genetics for tree breeding programs. Overall, GAS can be applied at the various strata and substrata in the genetic hierarchy of a classical tree breeding program. Within each of these strata there are opportunities to increase genetic gains by increasing selection intensity, more accurate selection, reduced costs of field testing and phenotypic evaluation, and possibly to speed up responses to changes in breeding objectives. Specific applications would, however, need to be carefully and quantitatively evaluated on a case-by-case basis, particularly in light of the fact that results from association tests will most likely come from a limited range of traits where only a proportion of the extant variation is accounted for by assayable polymorphisms, at least in the short term. Furthermore, because of the additional costs of this form of selection compared to phenotypic selection alone, it is likely that the initial application will be in the production populations, where

investment in nearer-to-market selection applications are likely to have more immediate pay-back.

10.9 FIT WITH OTHER BIOTECHNOLOGIES USED IN TREE IMPROVEMENT

As already stated, a key feature of GAS is the complementary fit with other genetic technologies, including those currently under development. For these new technologies to be applicable, they need to be more cost-effective at delivering genetic gains than conventional technologies. A number of new technologies are under development, and are at various stages of readiness for implementation in tree breeding programs. Here, we consider examples of new technologies that can be used to complement GAS and greatly enhance its effectiveness.

10.9.1 Within-Family Selection Based on DNA Marker–QTL Associations (MAS)

Scope exists for integrating GAS strategy with that of MAS. Because most commercially important tree breeding programs are now well into advanced-generation selection, there is significant emphasis on within-family selection in order to maintain the breadth of genetic base and avoid undue build-up of co-ancestry. MAS could be used for within-family selection although some limitations have been noted (Strauss *et al.* 1992; Kerr and Goddard 1997; Johnson *et al.* 2000), including the need for large individual family sizes necessary for achieving genetic gain for most quantitatively inherited characteristics (Wilcox *et al.* 2001). Given the high cost of detection of marker–trait associations for MAS on a family-by-family basis, it is likely that in breeding programs using MAS, detection of marker–trait associations will have been undertaken in only a subset of families in their respective breeding programs. Here, GAS could be used both as an aid to among-family selection and to augment MAS for within-family selection where family-specific marker–trait associations for MAS are not available. There are two potential benefits in doing this: firstly, increased genetic gains for reasons outlined above, and secondly, alleviation of the accelerated build-up of co-ancestry that could occur with the operational dependence on MAS. With MAS, accelerated co-ancestry could arise through MAS being available only for a small proportion of pedigrees which could therefore contribute disproportionate numbers of selections. More broadly applicable marker–trait associations (i.e., GAS), by facilitating selection from all pedigrees, would not be conducive to the same build-up of co-ancestry. Given the large sample sizes per family that are needed to detect QTL so as to achieve moderate genetic gains from MAS (Wilcox *et al.* 2001), practicing MAS across large numbers of essentially unrelated families becomes prohibitive. In comparison, GAS requires much lower sample sizes when averaged across the number of parents in breeding populations (discussed below). However, this advantage could be offset to some extent by the need to identify and assay many more polymorphisms per candidate gene, although there is potential to reduce sampling costs due to techniques such as pooling DNA samples from phenotypic extremes (Michelmore *et al.* 1991). Moreover, in specific cases such as dominant major genes for disease and insect resistance (cf. Bus *et al.* 2000), which do not require large sample sizes for detection, MAS is likely to be an effective means of obtaining gain; when thus detected, such genes may then be amenable to use of GAS, with the help

of comparative genomics based on DNA sequences in other plants. Similarly, family-specific effects associated with inbreeding (e.g., lethals, loci contributing to reduced vigor and general fitness status) may be better dealt with via MAS on a pedigree-specific basis as co-ancestry builds up, with GAS used to select for other characteristics. This could be especially important for purging highly deleterious alleles if aggressive inbreeding were to be adopted as a breeding tool.

Combined use of experimental infrastructure for both MAS and GAS has potential benefits also. Pedigreed QTL detection populations (as would be used for MAS) with association genetics population (as used for GAS) have been evaluated as a means of fine-mapping QTL (see Chapter 8 and references therein). Such an approach could be used to reduce confidence intervals around QTL location, thereby narrowing the range of potential candidates and effectively increasing the probability of choosing the appropriate genes.

10.9.2 Genetic Engineering

For operational use of genetic engineering, it is always important to do transformations on carefully chosen recipient genotypes. This is partly because inherently poor recipients will remain poor even after transformation, and partly because transformation costs are still high because of both the inherent costs of the protocols and the low success rate resulting from the inexact nature of contemporary transformation technologies. Selection of recipient genotypes, however, may be constrained by the fact that transformation may need to be done on embryogenic material. This creates a special attraction for the sort of very early selection that GAS can afford, by using DNA data (along with prior family information) to identify top candidates for transformation.

In addition to the operational use of genetic engineering there is the role of genetic engineering to establish the roles of candidate genes, which may serve to inform conventional breeding via indicating which genes are likely to result in phenotypic effects. In practice, this could be limited because of the time and expense of genetic modification, although some genes could be identified in this manner (see Section 3.4.1).

10.9.3 *In Vitro* Propagation Technologies

With the various technologies for *in vitro* propagation (e.g., organogenesis and somatic embryogenesis), the opportunity for early identification of top genotypes has benefits when both amplifying limited quantities of top genetic material, as well as for development of material for clonal testing and deployment. This form of early selection not only increases selection intensity, but also could be used to increase the efficiency of tissue culture by identifying genotypes more likely to propagate well – although having to select for propagation behavior is liable to be at the expense of potential genetic gain in other directions. This also applies to *in vivo* vegetative propagation. However, with a number of propagation technologies in various commercially important forest tree species, further development of propagation technologies may be required to fully utilize the potential from GAS.

10.9.4 Accelerated Flowering and Rejuvenation Technologies

Accelerated flowering technologies may be crucial to realizing at least some of the benefits of GAS and MAS. Such technologies can make it possible to capitalize on the early selection afforded by GAS to dramatically reduce length of breeding cycles and the lead time for deployment of genetic gains, thereby achieving more effective utilization of endogenous variability. For example, breeding cycles in contemporary commercially important conifer species are still 14–20 years in length, with selection requiring 4–10 years, and flower induction and seed production requiring a further 5–8 years. Flowering-on-command, coupled with selection based on DNA sequence information, could reduce the time for identification of top genotypes dramatically – in theory to much less than a year. Indeed, reducing the time required for floral induction, fertilization, and seed production could increase rates of gain by as much as three times, depending upon the reduction of generation interval.

Rejuvenation technologies achieve the opposite to accelerated flowering in operational breeding. The prospective benefits of rejuvenation for realizing genetic gain are great (Burdon 1982; Bonga and von Aderkas 1993), but they generally interplay less specifically with GAS than do the benefits of accelerated flowering.

10.9.5 Technologies to Study Pathways of Gene Action

GAS experiments (LD populations) are also useful as screening populations for identifying potential causative QTN, allowing integration of molecular and selection technologies by sharing common experimental platforms. The potential offered by association genetics experiments to identify candidates offers molecular biologists the opportunity to use genetics to inform roles and functions, thereby elucidating the particular roles of specific genes and the manners in which they might interact at a whole-organism level, either informing or complementing *in vitro* or model plant studies. Benefits arising from identification of causal mechanisms and pathways, apart from improved understanding of the molecular basis for heritable variation, include identifying genes (and methods) to create and exploit variation based on understanding the causal mechanisms (including potential pleiotropic effects). In the shorter term, a further benefit includes the identification of which and what type of genes could be targeted to create new "mutations" (via transformation) of potentially larger effect (Section 10.9.2 and above).

10.10 LIMITATIONS AND CHALLENGES

While the potential for GAS in tree breeding looks positive, implementation in commercial breeding programs faces a number of key obstacles. These include the high cost of implementation, institutional barriers, and technical impediments due to certain molecular mechanisms underpinning trait variation. We briefly discuss each of these below.

A key impediment to uptake is the high up-front cost of implementation, which is particularly important given that most commercial breeding programs need to bear most or all of the entire costs, whereas the benefits of genetic gain tend to accrue further down

the forestry value chain, which can take decades to materialize. Reasons for high implementation costs include:

- *High cost of establishing marker–QTN associations.* In order to achieve adequate experimental power, large experiments are likely to be needed (above). Furthermore, such experiments are likely to be costly to measure, particularly as most breeding objectives involve multiple traits, and typically include expensive-to-assess wood-property traits.
- *Costs associated with polymorphism discovery and genotyping.* Polymorphism discovery consists of extensive amounts of resequencing, followed by elucidation of disequilibrium patterns after which subsets of SNPs are chosen for genotyping in association tests (Figure 10.1). Because a number of polymorphisms per gene will be needed as well as several genes per QTL interval (unless prior information indicates a clear choice), there is substantially more evaluation and genotyping required per QTN than compared with MAS using pedigreed QTL mapping populations, although with the latter marker–trait associations need to be ascertained on a pedigree-by-pedigree basis. Such costs are not trivial, and may only be offset by investment from public funding agencies or by collaborations with organizations undertaking association genetics studies for purposes other than selection. Associated with sequencing and genotyping costs is the necessity to access facilities to undertake such work, although access to technologies could be attained through service providers and existing laboratories.
- *Additional skills needed for operational implementation in breeding programs.* These include competency in marker technologies, genomics and cellular biochemistry (primarily for candidate gene selection), and quantitative genetics methods relevant to detecting and estimating linkage disequilibria. Such skills usually require teams rather than single individuals, which therefore requires additional investment to establish and maintain an infrastructure associated with such teams, unless such skills can be acquired via collaboration.
- *Occurrence of genotype × environment interaction.* This will increase the number of experimental populations that will need to be deployed, although deploying cloned experimental populations could minimize additional genotyping. Even if selection for specific environments is not needed, good coverage in terms of test environments may still be needed (cf. Johnson and Burdon 1990).
- *Intergenerational changes in relationship between QTN and phenotype.* These could arise for example with disease/pest resistance genes, where shifts in the pathogen/pest population could change predictive value of QTN(s). Similarly, changes over time in environments, or even silvicultural practices, could likewise change the nature and/or extent of the causative associations in a manner that may not be easy to predict. Such changes would be likely to make certain costs recurring.

High costs mean GAS is unlikely to be an attractive option for species and/or breeding objectives with low commercial value. Even for species with greater commercial value, the additional investment may not be considered affordable, particularly for existing operational programs that lack additional financial resources with

which to develop and implement the operational infrastructure necessary for GAS. Therefore careful evaluation of specific implementation strategies and including costs and benefits are most likely to be necessary.

Certain mechanisms underpinning trait variation could also prevent effective development of GAS. An example particularly relevant to species with limited commercial value and/or relatively limited availability of nongenic DNA sequence (particularly those with large genomes) is where causative QTN occur many kilobases distal to expressed genes. Such is the case for the *Vgt1* locus in corn, which has been shown via association genetics to map to a 2 kb region that is 70 kb away from the nearest open reading frame (Salvi *et al.* 2006). If such distal transacting regulatory factors dominate trait variability, then extensive amounts of gDNA resequencing will be required. This would significantly add to costs, as well as reduce efficacy, particularly for large-genome species, effectively precluding application in gymnosperms, as well as a number of hardwood species. Another example is where trait architecture is predominantly composed of clusters of small-effect QTN per QTL. Such architecture is theoretically possible, and further experimentation will reveal whether or not this is the case. Experiments of sufficient power will be necessary, increasing cost and time required to detect QTN. Furthermore, genotyping costs per unit of gain will be greater, potentially offsetting expected benefits.

Another technical limitation is the predictive value of associations in the light of potential modes of gene action, particularly epistasis. Nucleotide substitution effects would usually be estimated by averaging over allelic combinations sampled in association tests. However, the selected variants may not be well represented in association tests, so the predictive value of multilocus QTN could be limited in the presence of epistasis. Evidence from genetic tests in conifers indicates that large-effect epistasis is unlikely to be prevalent, but does not rule out smaller epistatic effects. Such interactions are plausible, given the nature of interdependent biosynthetic pathways that give rise to phenotype, but may not be observed (or even important) in large outbred deployment populations that are typically derived from open- and control-pollinated seed orchards. Conversely, for clonal forestry, where GAS could potentially be used to identify candidates for further testing, such interactions could be important, particularly if candidates available to be screened are unlikely to include optimal multitrait genotypes because of biological limitations on the numbers of seed that could be produced for screening.

A specific, potentially important class of epistasis, is co-adapted gene complexes. This phenomenon is possible in forest trees, although some surprising cases have been observed of essentially independent inheritance of traits that would seem to have common adaptive significance (Howe *et al.* 2003). If, however, such complexes do exist, they must be considered when generating and selecting new variants, necessitating the detection and if necessary, management of, haplotypic complexes. Fortunately, further experimentation to detect such complexes may be unnecessary, as existing technologies combined with association test populations may well be adequate. We envision that such research will be undertaken over the next few years. If present, means of managing co-adapted complexes in tree breeding programs will need to be implemented; although this may not be difficult in theory, it may present major logistical challenges.

GAS may have little or no utility for backwards selection and reselection within existing breeding and production populations, particularly where progeny tests are already established and measured for other traits. Such instances may not be rare, as

breeding objectives and strategies are frequently being revised, and new traits are often introduced into breeding programs in response to factors such as new biological pressures and/or market signals. In these cases, it may be more cost-effective to screen extant families for new properties. In breeding programs with limited resources, the short-term cost-effectiveness of such approaches may restrict or prevent investment in technologies such as GAS which are longer-term in delivery of improved germplasm, unless marker–trait relationships can be easily undertaken in association tests that result in a significant proportion of trait variation being explained by markers.

Institutional barriers to implementation also exist. In the case of breeding cooperatives and companies whose programs are based on phenotypic selection, barriers can exist to understanding the nature and complexities of molecular genetics applications as most programs have tended not to use such tools routinely, and when done, usually in some conceptually easy application such as verification of parentage or clonal identity. Convincing such organizations, which tend to be conservative, to implement this technology, may be difficult particularly in light of the few results to date that clearly demonstrate ease of detecting associations let alone actual genetic gains. Furthermore, fluctuations in the relative economics of plantation forestry and frequent ownership changes can prevent adequate investment from nongovernment sources to appropriately develop and implement the technology. This may be particularly important where plantation ownership is dominated by investors with short-term financial goals, therefore unwilling to participate in more longer-term activities such as association genetics.

For reasons described above, we foresee that GAS is most likely to be implemented in breeding programs where there are good operational links between molecular geneticists and tree breeders (as well as others), either moderate to high product values or sufficient scale to allow costs to be widely spread, and sufficient investment over the requisite period of time to enable discovery of suitable numbers of marker–QTN relationships.

10.11 CONCLUSIONS

Application of association genetics in plantation forest tree species has the potential to increase genetic gains from among- and/or within-family selection via a number of routes such as increased selection intensities and/or earlier selection. Such selection can be applied to virtually all strata of hierarchically structured populations used in tree improvement, although it is likely that the most immediate applications will be in populations used to provide seed for commercial plantations, owing to the relatively shorter timeframe to recover additional costs associated with detecting marker–trait associations. Other potential benefits include cheaper selection, reduced need for phenotypic selection, and complementary fit with other biotechnologies used either commercially or in research, as well as use of the same experimental infrastructure for purposes other than selection.

The few studies to date of LD in forest trees indicate relatively short spans of LD, implying that finding disequilibria between causative QTN will need to be undertaken via judiciously chosen candidate genes (hence use of the term "gene-assisted selection"), particularly in conifers where large genomes effectively preclude cost-effective whole genome resequencing.

There are a number of important prerequisites for GAS to be successful. These include effective integration of existing tree breeding skills with molecular genetics,

genomics, and bioinformatics, as well as relevant statistical skills. In addition, access to adequate populations with which to detect sufficient numbers of small-effect QTN are a key requirement. Access to genomics and genotyping facilities are also critical, as are accessed to technologies that will improve the ability to choose appropriate candidate genes.

There are, however, some potential impediments to implementation of association genetics in tree breeding. These include the high costs of detecting marker–trait associations relative to product value and long rotation lengths of forest trees; certain modes of gene action which may preclude effective detection of associations, particularly in conifers; and institutional barriers associated with understanding and investing in new technologies.

10.11 REFERENCES

Allison, D.B., 1997, Transmission-disequilibrium tests for quantitative traits. Genetics 60:676–690.

Ball, R.D., 2001, Bayesian methods for quantitative trait loci mapping based on model selection: approximate analysis using the Bayesian Information Criterion. Genetics 159:1351–1364.

Ball, R.D., 2005, Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873.

Beavis, W.D., 1994, The power and deceit of QTL experiments: lessons from comparative QTL studies. pp. 250–266. In: Proceedings of the 49th Annual Corn and Sorghum Industry Research Conference. American Seed Trade Association, Washington, DC.

Bonga, J.M., von Aderkas, P., 1993, Rejuvenation of tissues from mature conifersand its implications for propagation *in vitro*. In: Clonal Forestry (Eds. M.R. Ahuja, W.J. Libby) pp. 182–199. Springer-Verlag, Berlin Heidelberg.

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.

Brown, G.R., Bassoni, D.L., Gill, G.P., Fontana, J.R., Wheeler, N.C., Megraw, R.A., Davis, M.F., Sewell, M.M., Tuskan, G.A., Neale, D.B., 2003, Identification of quantitative trait loci influencing wood property traits in loblolly pine (*Pinus taeda* L.) III. QTL Verification and candidate gene mapping. Genetics 164:1537–1546.

Brown, G.R., Gill, G.P., Kuntz, R.J., Beal, J.A., Nelson, C.D., Wheeler, N.C., Penttila, B., Roers, J., Neale, D.B., 2004a, Associations of candidate gene single nucleotide polymorphism with wood property phenotypes in loblolly pine (Abstr.). Plant and Animal Genome XII, 10–14 January 2006, San Diego, CA.

Brown, G.R., Gill, G.P., Kuntz, R.J., Langley, C.H., Neale, D.B., 2004b, Nucleotide diversity and linkage disequilibrium in loblolly pine. Proceedings of the National Academy of Sciences of the United States of America 101:15255–15260.

Bucci, G., Menozzi, P., 1995, Genetic variation of RAPD markers in a *Picea abies* Karst. population. Heredity 75:188–197.

Burdon, R.D., 1982, The Roles and Optimal Place of Vegetative Propagation in Tree Breeding Strategies. In: Proceedings of IUFRO Meeting on Genetics and Breeding Strategies pp. 66–83. Sensenstein, Germany.

Burdon, R.D., 1988, Recruitment for breeding populations: objectives, genetics, and implementation. In: Proceedings of Second International Conference on Quantitative Genetics (Eds. B.S. Weir, E.J. Eisen, M.M. Goodman, G. Namkoong) pp. 555–572. Sinauer, Sunderland, MA.

Burdon, R.D., 1992, Genetic survey of *Pinus radiata*. 9: general discussion and implications for genetic management. New Zealand Journal of Forest Science 22:174–198.

Burdon, R.D., 2001, Genetic diversity and disease resistance: some considerations for research, breeding and deployment. Canadian Journal of Forest Research 32:596–606.

Burdon, R.D., Namkoong, G., 1983, Multiple populations and sublines. Silvae Genetica 32:221–222.

Burdon, R.D., Russell, J.H., 1999, Inbreeding depression in selfing experiments: statistical issues. Forest Genetics 5:179–189.

Burdon, R.D., Firth, A., Low, C.B., Miller, M.A., 1998, Multi-site provenance trials of *Pinus radiata* in New Zealand. Forest Genetic Resources No 26. pp. 3–8. FAO, Rome.

Bus, V.G., Gardiner, S.E., Bassett, H.C.M., Ranarunga, C., Rikkerink, E.H.A., 2000, Marker assisted selection for pest and disease resistance in the New Zealand apple breeding programme. Acta Horticulture 538:541–547.

Casasoli, M., Derory, J., Morera-Dutrey, C., Brendel, O., Porth, I., Guehl, J.M., Villani, F., Kremer, A., 2006, Comparison of quantitative trait loci for adaptive traits between oak and chestnut based on an expressed sequence tag consensus map. Genetics 172:533–546.

Cato, S.A., Pot, D., Kumar, S., Douglas, J., Gardner, R.C., Wilcox, P.L., 2006, Balancing selection in a dehydrin gene associated with increased wood density and decreased radial growth in *Pinus radiata* (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

Chagné, D., Brown, G., Lalanne, C., Madur, D., Pot, D., Neale, D., Plomion, C., 2003, Comparative genome and QTL mapping between maritime and loblolly pines. Molecular Breeding 12:185–195.

Deng, H.-W., 2001, Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. Genetics 159:1319–1323.

Devey, M.E., Delfino-Mix, A., Kinloch, B.B., Neale, D.B., 1995, Random amplified polymorphic DNA markers tightly linked to a gene for resistance to white pine blister rust in sugar pine. Proceedings of the National Academy of Sciences of the United States of America 92:2066–2070.

Devey, M.E., Sewell, M.M., Uren, T.L., Neale, D.B., 1999, Comparative mapping in loblolly and radiata pine using RFLP and microsatellite markers. Theoretical and Applied Genetics 99:656–662.

Devey, M.E., Groom, K.A., Nolan, M.F., Bell, J.C., Dudzinski, M.J., Old, K.M., Matheson, A.C., Moran, G.F., 2004, Detection and verification of quantitative trait loci for resistance to Dothistroma needle blight in *Pinus radiata*. Theoretical and Applied Genetics 108:1056–1063.

Dodds, K.G., Montgomery, G.W., Tate, M.L., 1993, Testing for linkage between a marker locus and a major gene locus in half-sib families. Journal of Heredity 84:43–48.

Dupuis, J., Siegmund, D., 1999, Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373–386.

Dvornyk, V., Sirviö, A., Mikkonen, M., Savolainen, O., 2002, Low nucleotide diversity at the *pal1* locus in the widely distributed *Pinus sylvestris*. Molecular Biology and Evolution 19:179–188.

Echt, C.S., Vendramin, C.D., Nelson, C.D., Marquardt, P., 1999, Microsatellite DNA as shared genetic markers among conifer species. Canadian Journal of Forest Research 29:365–371.

Epperson, B.K., Allard, R.W., 1987, Linkage disequilibrium between allozymes in natural populations of lodgepole pine. Genetics 115:341–352.

Evans, R., 1994, Rapid measurement of transverse measurements of tracheids in radial wood specimens of *Pinus radiata*. Holzforschung 48:168–172.

Evans, R., Kibblewhite, R.P., Stringer, S., 1999, Variation of microfibril angle, density and fibre orientation in twenty-nine *Eucalyptus nitens* trees. Appita Journal 50:487–494.

Geburek, T., 1998, Genetic variation of Norway spruce (*Picea abies* [L.] Karst.) populations in Austria 1. Digenic disequilibrium and microspatial patterns derived from allozymes. Forest Genetics 5:221–230.

Germer, S., Holland, M.J., Higuchi, R., 2000, High-throughput SNP allele-frequency determination in pooled DNA samples by kinetic PCR. Genome Research 10:258–266.

Gupta, P.K., Rustgi, S., Kulwal, P.L., 2005, Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Molecular Biology 57:461–485.

Howe, G.T., Aitken, S.N., Neale, D.B., Jermstad, K.D., Wheeler, N.C., Chen, T.H.H., 2003, From genotype to phenotype: unravelling the complexities of cold adaptation in forest trees. Canadian Journal of Forest Research 33:1247–1266.

Huntley, S.K., Ellis, D., Gilbert, M., Chapple, C., Mansfield, S.D., 2003, Significant increases in pulping efficiency in C4H–F5H transformed poplars: improved chemical savings and reduced environmental toxins. Journal of Agricultural Food Chemicals 51:6178–6183.

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.

Jayawickrama, K.J.S., Carson, M.J., 2000, A breeding strategy for the New Zealand Radiata Pine Breeding Co-operative. Silvae Genetica 49:82–90.

Johnson, G.R., Burdon, R.D., 1990, Family-site interaction in *Pinus radiata*: implications for progeny testing strategy and regionalised breeding in New Zealand. Silvae Genetica 39:55–62.

Johnson, G.R., Wheeler, N.C., Strauss, S.H., 2000, Financial feasibility of marker-aided selection in Douglas-fir. Canadian Journal of Forest Research 30:1942–1952.

Jones, L., Ennos, A.R., Turner, S.R., 2001, Cloning and characterization of irregular xylem4 (irx4): a severely lignin-deficient mutant of *Arabidopsis*. The Plant Journal 26:205–216.

Kerr, R.J., Goddard, M.E., 1997, Comparison between the use of MAS and clonal tests in tree breeding programmes. In: IUFRO '97 Genetics of Radiata Pine (Eds. R.D. Burdon, J.M. Moore) pp. 297–303. Proceedings of NZFRI/IUFRO Conference 1–4 December and Workshop 5 December, Rotorua, New Zealand FRI Bulletin No. 203.

Kinloch, B.B., Parks, G.K., Flower, C.W., 1970, White pine blister rust: simply inherited resistance in sugar pine. Science 167:193–195.

Kirst, M.E., Myburg, A.A., Sederoff, R.R., 2003. Genetical genomics of *Eucalyptus*: combining expression profiling and genetic segregation analysis (Abstr.). Plant and Animal Genome XI, 11–15 January 2003, San Diego, CA.

Kirst, M., Myers, R.M., De León, J.P.G., Kirst, M.E., Scott, J., Sederoff, R., 2004. Coordinated genetic regulation of growth and lignin revealed by quantitative trait locus analysis of cDNA microarray data in an interspecific backcross of eucalyptus. Plant Physiology 135:2368–2378.

Krutovsky, K.V., Neale, D.B., 2005. Nucleotide diversity and linkage disequilibrium in cold hardiness and wood quality related candidate genes in Douglas-fir. Genetics 171:2029–2041.

Kuang, H., Richardson, T.E., Carson, S.D., Bongarten, B., 1999. Genetic analysis of inbreeding depression in plus tree 850.55 of *Pinus radiata* D. Don. II. Genetics of viability genes. Theoretical and Applied Genetics 99:140–146.

Kumar, S., Echt, C.S., Wilcox, P.L., Richardson, T.E., 2004. Testing for linkage disequilibrium in the New Zealand radiata pine breeding population. Theoretical and Applied Genetics 108:292–298.

Lagercrantz, U., Ryman, N., 1990. Genetic structure of Norway spruce (*Picea abies*): concordance of morphological and allozymic variation. Evolution 44:38–53.

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.

Luo, Z.W., 1998. Detecting linkage disequilibrium between a polymorphic marker locus and a trait locus in natural populations. Heredity 80:198–208.

Lynch, M., Walsh, B., 1997. Genetics and analysis of quantitative traits. Sinauer Associates, Sunderland, MA.

Michelmore, R.W., Paran, I., Kesseli, R.V., 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proceedings of the National Academy of Sciences of the United States of America 88:9828–9832.

Mitton, J.B., 1992. The dynamic mating system of conifers. New Forests 6:187–216.

Mitton, J.B., Sturgeon, K.B., Davis, M.L., 1980. Genetic differentiation in ponderosa pine along a steep elevational transect. Silvae Genetica 29:100–103.

Morgante, M., Salamini, F., 2003. From plant genomics to breeding practice. Current Opinion in Biotechnology 14:214–219.

Muona, O., Szmidt, A.E., 1985. A multilocus study of natural populations of *Pinus sylvestris*. In: Lecture notes in Bioinformatics. (Ed H.-R. Gregorious) pp. 226–240. Springer Verlag, Berlin.

Murray, B.G., 1998. Nuclear DNA amounts in gymnosperms. Annals of Botany 82(Supplement A):3–15.

Paran, I., Zamir, D., 2003. Quantitative traits in plants: beyond the QTL. Trends in Genetics 19:303–306.

Paux, E., Tamasloukht, M.B., Ladouce, N., Sivadon, P., Grima-Pettenati, J., 2004. Identification of genes preferentially expressed during wood formation in *Eucalyptus*. Plant Molecular Biology 55:263–280.

Plomion, C., Richardson, T.E., MacKay, J., 2005. Advances in forest tree genomics: forest trees workshop, plant and animal genome XIII conference, San Diego, CA, January 2005. New Phytologist 166:713–717.

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

Powers, H.R., Hubbard, S.D., Anderson, R.L., 1982. Resistance to diseases and pests in forest trees. In: Proceedings of Third International Workshop on Genetics of Host-Parasite Interactions in Forestry (Eds. H.M. Heybroek, B.R. Stephan, K. von Weissenberg). pp. 427–434. Pudoc, Wageningen, The Netherlands.

Pritchard, J.K., Rosenberg, N.A., 1999. Use of unlinked genetic markers to detect population stratification in association studies. American Journal of Human Genetics 65:220–228.

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P., 2000. Association mapping in structured populations. Genetics 67:170–181.

Rafalski, J.A., Morgante, M., 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends in Genetics 20:103–111.

Roberds, J.H., Brotschol, J.V., 1985. Linkage disequilibrium among allozyme loci in natural populations of *Liriodendron tulipifera* L. Silvae Genetica 34:137–141.

Russell, J.H., Burdon, R.D., Yanchuk, A.D., 2003. Inbreeding depression and variance structures for height and adaptation in self- and outcross *Thuja plicata* families in varying environments. Forest Genetics 10:171–184.

Salvi, S., Tuberosa, R., 2005. To clone or not to clone plant QTLs: present and future challenges. Trends in Plant Sciences 10:1360–1385.

Salvi, S., Sponza, G., Morgante, M., Tomes, D., Tuberosa, R., 2006. Confirmation of the maize flowering time QTL *Vgt1* by association mapping (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

Sewell, M.M., Neale, D.B., 2000. Mapping quantitative traits in forest trees. In: Molecular biology of woody plants, forestry Sciences (Eds. S.M. Jain, S.C. Minocha) pp. 407–433. Kluwer Academic Publishers, The Netherlands.

Strauss, S.H., Lande, R., Namkoong, G., 1992. Limitations of molecular marker-aided selection in forest tree breeding. Canadian Journal of Forest Research 22:1050–1061.

Stromberg, L.D., Dudley, J.D., Rufener, G.K., 1994, Comparing conventional early generation selection with molecular marker assisted selection in maize. Crop Science 34:1221–1225.

Telfer, E.J., Echt, C.S., Nelson, C.D., Wilcox, P.L., 2006, Comparative mapping in *Pinus radiata* and *P. taeda* reveals co-location of wood density-related QTL (Abstr.). Plant and Animal Genome XIV, 14–18 January 2006, San Diego, CA.

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

Thumma, B.R., Nolan, M.F., Evans, R., Moran, G.F., 2005, Polymorphisms in *Cinnamoyl CoA Reductase (CCR)* are associated with variation in microfibril angle in *Eucalyptus* spp. Genetics 171:1257–1265.

Wayne, M.L., McIntyre, L.M., 2002, Combining mapping and arraying: an approach to candidate gene identification. Proceedings of the National Academy of Sciences of the United States of America 99:14903–14906.

Wilcox, P.L., Amerson, H.V., Kuhlman, E.G., Liu, B.-H., O'Malley, D.M., Sederoff, R.R., 1996. Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proceedings of the National Academy of Sciences of the United States of America 93:3859–3864.

Wilcox, P.L., Richardson, T.E., Carson, S.D., 1997. Nature of quantitative trait variation in *Pinus radiata*: insights from QTL detection experiments. In: IUFRO '97 Genetics of Radiata Pine (Eds. R.D. Burdon, J.M. Moore) pp. 304–312. Proceedings of NZFRI/IUFRO Conference 1–4 December and Workshop 5 December, Rotorua, New Zealand FRI Bulletin No. 203.

Wilcox, P.L., Carson, S.D., Richardson, T.E., Ball, R.D., Horgan, G.P., Carter, P., 2001, Benefit-cost analysis of DNA marker-based selection in progenies of *Pinus radiata* seed orchard parents. Canadian Journal of Forest Research 31:2213–2224.

Williams, C.G., Savolainen, O., 1996, Inbreeding depression in conifers: implications for breeding strategy. Forest Science 42:102–117.

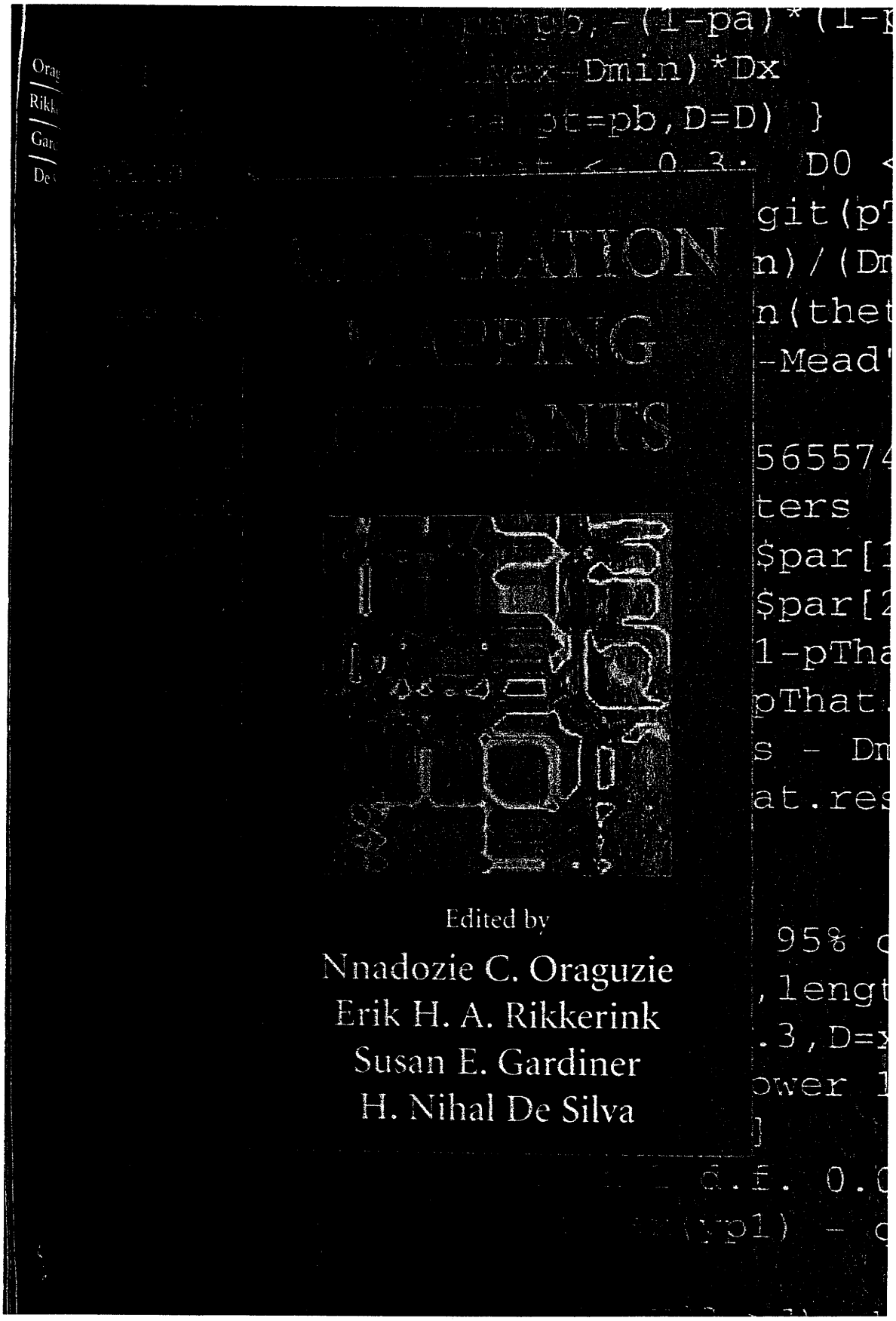
Wright, S.I., Gaut, B.S., 2005, Molecular population genetics and the search for adaptive evolution in plants. Molecular Biology and Evolution 22:506–519.

Wu, R., Zeng, Z.-B., 2001, Joint linkage and linkage disequilibrium mapping in natural populations. Genetics 157:899–909.

Wu, R., Ma, C.-X., Casella, G., 2002, Joint linkage and linkage disequilibrium mapping of quantitative trait loci in natural populations. Genetics 160:779–792.

Yin, T.M., DiFrazio, S.P., Gunter, L.E., Jawdy, S.S., Boerjan, W., Tuskan, G.A., 2004. Genetic and physical mapping of *Melampsora* rust resistance genes in *Populus* and characterization of linkage disequilibrium and flanking genomic sequence. New Phytologist 164:95–105.

Yu, J., Pressoir, G., Briggs, W.H., Bi, I.V., Yamasaki, M., Doebley, J.F., McMullen, M.D., Gaut, B.S., Nielsen, D.M., Kresovich, S., Buckler, E.S., 2006, A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nature Genetics 38:203–208.


Edited by
Nnadozie C. Oraguzie
Erik H. A. Rikkerink
Susan E. Gardiner
H. Nihal De Silva

Association Mapping in Plants

Association Mapping in Plants

Edited by

Nnadozie C. Oraguzie
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Havelock North, New Zealand

Erik H.A. Rikkerink
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Auckland, New Zealand

Susan E. Gardiner
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Palmerston North, New Zealand

H. Nihal De Silva
The Horticulture and Food Research Institute of New Zealand Ltd (HortResearch)
Auckland, New Zealand

Springer

Dr Nnadozie C. Oraguzie
HortResearch
Hawkes Bay Research Centre
Private Bag 1401
Havelock North
New Zealand

Dr Erik H.A. Rikkerink
HortResearch
Mt. Albert Research Centre
Private Bag 92169
Auckland
New Zealand

Dr Susan E. Gardiner
HortResearch
Palmerston North Research Centre
Private Bag 11030
Palmerston North
New Zealand

Dr H. Nihal De Silva
HortResearch
Mt. Albert Research Centre
Private Bag 92169
Auckland
New Zealand

Library of Congress Control Number: 2006928327

ISBN-10: 0-387-35844-7 e-ISBN-10: 0-387-36011-5
ISBN-13: 978-0387-35844-4 e-ISBN-13: 978-0-387-36011-9

Printed on acid-free paper.

© 2007 Springer Science+Business Media, LLC.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC., 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

9 8 7 6 5 4 3 2 1

springer.com

Preface

The approach taken for locating the genes that underlie human diseases has shifted from pedigree-based linkage studies to population-based association studies. In both cases the proximity of a genetic marker to a susceptibility locus is inferred from statistical measures that reflect the number of recombination events between them: in a disease pedigree there are no more than a few hundred opportunities for recombination so that recombination rates less than about one percent cannot be estimated and genes can be located only coarsely on a genetic map with that approach. The linkage disequilibrium detected in an association study, however, reflects the actions of many thousands of recombination events since the initial disease mutation and the expectation is that susceptibility genes can then be mapped more accurately.

The editors of this volume have recognized the need for parallel activity in plant species. For the past 20 years, the genes that affect plant economic traits have usually been mapped with data collected from "pedigrees" of populations formed by crossing inbred lines. These Quantitative Trait Loci have been mapped on a coarse scale, and a QTL is likely to refer to several genes in a region. The move to population-based association studies was therefore as necessary in plants as it was in humans, and readers will find this book to be a useful review of the marker technology, statistical methodology, and progress to date. Although one of the authors fears that "plant genetics can be considered as less advanced than human genetics" the chapters suggest that if that is the case it will not be so for long.

The recent increased activity in association mapping in humans has rested on the development of efficient and affordable methods for discovering and employing Single Nucleotide Polymorphism markers. Plant geneticists cannot command the resources available to their human geneticist colleagues, but they can anticipate benefiting from the success of the International HapMap Project. The improvement in marker technology from such large projects will inevitably be imported to plant studies. The editors have provided helpful guides to the use of SNPs in association studies.

Along with the substantial increase in the volume of data when large numbers of individuals are typed at millions of SNPs there are substantial challenges in the statistical interpretation of the data. This book contains a valuable account of the issues of multiple testing and an accessible account of False Discovery Rates. The more basic concepts of linkage disequilibrium and case-control versus family-based association tests are also discussed. It is often the case that geneticists do not receive extensive statistical training and the coverage of the theory of estimation and testing is therefore welcome. Readers will notice a greater use of Bayesian methods than is usually found in statistical genetics books. Such methods are appearing more frequently in scientific papers.

I congratulate the editors and all the authors on this timely and comprehensive treatment of association mapping in plants. The importance of food and fiber for human welfare cannot be overstated, and progress in plant improvement will rest in no small part on the work described in these pages. On a personal level, I am delighted by the leadership shown by my fellow antipodeans.

B.S. Weir
Professor and Chair
Department of Biostatistics
University of Washington

Acknowledgements

The motivation to write this book came from numerous correspondences and finally a meeting with Keri Witman, former senior editor in plant sciences at Springer Science+Business Media, New York, USA (who continually stressed the need for a book in this field of research), in Tulln Austria in 2004 at an EUCARPIA Quantitative genetics conference. The recommendation and suggestions by anonymous referees consulted by Springer to review the original proposal kept up the initial enthusiasm. This book would not have been possible without HortResearch's support. In particular, we would like to thank the following people; Drs Vincent Bus (Science Leader, Pipfruit and Summerfruit breeding), Andrew Granger (Future Fruit Group Leader) and Bruce Campbell (General manager science operations) for their encouragement and moral support, Stuart Ritchie for assistance with contract negotiation, Sharlene Cookson for assistance with book cover design, Dr David Chagné for his immense contribution right from the proposal stage of the book till it went into press, and the SPU team especially, Dr Anne Gunson and Christine Lamont, for editorial and technical assistance. We are grateful to the following for permission to reproduce copyright material: Trustees of the Royal Botanic Gardens, Kew; Elsevier (Figure 1 in Trends in Genetics 11: 83-90(2002) and Tables 1 & 2 in Trends in Genetics 20(2): 105(2004)); Swedish National Biobanking program, Wallenberg Consortium North (http://www.meb.ki.se/genestat.htm/-pairwise D` for 45 SNPs within a linked region); The University of Chicago press (Table 3, American Journal of Human Genetics 60(3): 676-690, 1997); Annual Reviews (Table 1 in Annual Review of Plant Biology 54, 2003); Nature Publishing group (Table 3 in Nature Genetics 28:286-289); and MacMillan publishing Inc. (Tables 2, 3, 4, 5 & 6 in Genetics 170:859-873, 2005). Finally we thank the following colleagues who reviewed the manuscripts and made suggestions that significantly improved the quality of the chapters: Drs David B. Neale (Professor of Forest Genetics, Dept of Plant Sciences, UCLA), Mark E. Sorrells (Professor, Dept of Plant Breeding and Genetics, Cornell University, Ithaca, NY), Trudy F.C. Mackay (WNR Distinguished Professor of Genetics, North Carolina State

University), Christophe Plomion (Molecular Geneticist, INRA, Cedex, France), David Pot (Molecular Geneticist, CIRAD, Montpellier, Cedex, France), Pauline Garnier-Géré (Forest Genetics Group, INRA, Cedex, France), Fred van Eeuwijk (Associate Professor in Statistics, Laboratory of Plant Breeding, Wageningen University and Research center, Wageningen, The Netherlands), Rasmus Neilsen (Ole RØmer Fellow/Professor in Statistical Genomics, University of Copenhagen, Denmark), Diane Mather (Professorial Research Fellow, School of Agriculture and Wine Science, University of Adelaide/Program Leader-New Molecular technologies, Molecular plant breeding cooperative Research Center, Plant Genomics Center, Glen Osmond SA, Australia), Martin Lascoux (Professor-Program in Evolutionary Functional Genomics, Evolutionary Biology Center, Uppsala University, Sweden), Hans van Buijtenen (Professor emeritus-Genetics, Dept of Forest Science, Texas A & M University), and Slade Lee (Associate Professor/Deputy Director, Center for Conservation Genetics/Sub-Program Leader (Phenomics), Grain Foods CRC Limited, Southern Cross University, Lismore NSW, Australia.

Contents

Contributors xi

Introduction xiii

1. **An Overview of Association Mapping** 1
   Nnadozie C. Oraguzie and Phillip L. Wilcox

2. **Linkage Disequilibrium** 11
   Nnadozie C. Oraguzie, Phillip L. Wilcox, Erik H.A. Rikkerink, and H. Nihal de Silva

3. **What Are SNPs?** 41
   David Edwards, John W. Forster, David Chagné, and Jacqueline Batley

4. **Single Nucleotide Polymorphism Discovery** 53
   David Edwards, John W. Forster, Noel O.I. Cogan, Jacqueline Batley, and David Chagné

5. **Single Nucleotide Polymorphism Genotyping in Plants** 77
   David Chagné, Jacqueline Batley, David Edwards, and John W. Forster

6. **SNP Applications in Plants** 95
   Jacqueline Batley and David Edwards

7. **Linkage Disequilibrium Mapping Concepts** 103
   H. Nihal de Silva and Roderick D. Ball

8. Statistical Analysis and Experimental Design 133
Roderick D. Ball

9. Linkage Disequilibrium-Based Association Mapping in Forage Species 197
Mark P. Dobrowolski and John W. Forster

10. Gene-Assisted Selection
Applications of Association Genetics for Forest Tree Breeding 211
Phillip L. Wilcox, Craig E. Echt, and Rowland D. Burdon

11. Prospects of Association Mapping in Perennial Horticultural Crops 249
Erik H.A. Rikkerink, Nnadozie C. Oraguzie, and Susan E. Gardiner

Index 271

Contributors

Roderick D. Ball
Ensis Wood Quality, Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, Private Bag 3020, Rotorua, New Zealand

Jacqueline Batley
Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Rowland D. Burdon
Ensis Genetics, Scion (New Zealand Forest Research Institute Limited), 49 Sala Street, Private Bag 3020, Rotorua, New Zealand

David Chagné
HortResearch, Plant Gene Mapping group, Private Bag 11030, Palmerston North, New Zealand

Noel O.I. Cogan
Primary Industries Research Victoria, Victorian AgriBiosciences Centre, La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Craig E. Echt
USDA Forest Service, Southern Institute of Forest Genetics 23332 MS Highway 67, Saucier, MS 39574, USA

David Edwards
Primary Industries Research Victoria, Victorian AgriBiosciences Centre,
La Trobe R&D Park, Bundoora, Victoria 3083, Australia

John W. Forster
Primary Industries Research Victoria, Victorian AgriBiosciences Centre,
La Trobe R&D Park, Bundoora, Victoria 3083, Australia

Mark P. Dobrowolski
Primary Industries Research Victoria, Plant Genetics and Genomics Platform, Hamilton
Centre, Mt Napier Road Hamilton, Victoria 3300, Australia

Susan E. Gardiner
HortResearch, Palmerston North Research Centre, Private Bag 11030, Palmerston North,
New Zealand

Nnadozie C. Oraguzie
HortResearch, Hawkes Bay Research Centre, Cnr. Crosses and St. George's Roads,
Private Bag 1401, Havelock North, New Zealand

Erik H.A. Rikkerink
HortResearch, Mt Albert Research Centre, 120 Mt Albert Road, Private Bag 92169,
Auckland, New Zealand

H. Nihal De Silva
HortResearch, Mt Albert Research Centre, 120 Mt Albert Road, Private Bag 92169,
Auckland, New Zealand

Phillip L. Wilcox
Cell wall Biotechnology Centre, Scion (New Zealand Forest Research Institute Limited),
49 Sala Street, Private Bag 3020, Rotorua, New Zealand

Introduction

Most traits we deal with on a daily basis have complex inheritance patterns that complicate the ability of existing mapping technologies to detect the underlying genetic factors. In the last decade or so, we have seen the successful use of conventional map-based strategies in identification and cloning of quantitative trait loci (QTLs) in model plant species including tomato and Arabidopsis. However, efficient gene discovery with this method will probably continue to be largely limited to those loci that have large effects on quantitative trait variation. Techniques are also needed to more rapidly identify genes that play a modest role in regulating quantitative trait variation. Association mapping via linkage disequilibrium or LD (non-random association of alleles at different loci) offers promise in this area. The traditional approach of linkage/QTL mapping reliant on developing large mapping populations continues to suffer from lack of mapping resolution inherent in samples with limited meiotic cross-over events. These problems are exacerbated in tree crops, where very large populations are impractical from a plant management point of view. In association mapping, there may not be any need to make crosses initially to generate segregating populations. The natural variation that exists in the available germplasm can be utilized for mapping straightaway.

Association genetics via LD mapping is an emerging field of genetic mapping that has the potential for resolution to the level of individual genes (alleles) underlying quantitative traits. LD mapping is a technology that can take full advantage of the phenomenal leaps and bounds in technology development in the area of molecular biology and marry it with our increasing understanding of the molecular basis of inheritance and molecular tools recently developed in terms of molecular markers and genetic maps in a way that could have a significant practical impact on breeding. The convergence of improved statistical methods, availability of growing plant genomics databases and improvements in the affordability and potential scale of sequencing and

ASSOCIATION MAPPING IN PLANTS

Edited by
Nnadozie C. Oraguzie, Erik H. A. Rikkerink,
Susan E. Gardiner, and H. Nihal De Silva

For the past decade, there has been success in using conventional map-based strategies in identification and cloning of quantitative trait loci (QTL) in model plant species including tomato and Arabidopsis. These quantitative traits are generally the products of many loci with varying degrees of effect upon the observed phenotypes. Recently, a new approach to genetic mapping has emerged called association mapping. This new technique takes into account the thousands of genes to evaluate for QTL effect and is a more efficient approach that does not require generation of segregating populations/large numbers of progeny. As it can utilize all of the historic recombination events in a diverse population of individuals, it can generate higher resolution genetic maps, and is needed to complement current map based cloning methods.

Association Mapping in Plants provides both basic and advanced understanding of association mapping and an awareness of population genomics tools to facilitate mapping and identification of the underlying causes of quantitative trait variation in plants. It acts as a useful review of the marker technology, the statistical methodology, and the progress to date. It also offers guides to the use of single nucleotide polymorphisms (SNPs) in association studies.

This book will appeal to all those with an interest in plant genetics, plant breeding, and plant genomics.

About the Editors:

Dr. Nnadozie C. Oraguzie *is a Senior Scientist in Genetics at the Horticulture and Food Research Institute of New Zealand Ltd (HortResearch).*

Dr. Erik H. A. Rikkerink *is a Science Leader at HortResearch, New Zealand.*

Dr. Susan E. Gardiner *is a Principal Scientist and leader of the Gene Mapping research team at HortResearch, New Zealand.*

Dr. H. Nihal De Silva *is a Senior Scientist of Biometrics at HortResearch, New Zealand.*

LIFE SCIENCES

ISBN 978-0-387-35844-4

EAN 9 780387 358444 >

Springer

> springer.com