Modern Advances in Tree Breeding

Yousry A. El-Kassaby, Fikret Isik, and Ross W. Whetten

Abstract Traditional tree improvement programs are long-term endeavours requiring extensive resources. They require establishing mating designs, installing progeny tests on multiple sites to evaluate parents and their offspring over large geographic areas, monitoring those tests over extended periods of time, and eventual analysis of measurements to assess economic traits. Most tree breeding programs follow the classical recurrent selection scheme, resulting in the generation of multiple breeding and production populations. This process, while successful in attaining appreciable gains, remained static for a long time. The availability of plentiful, reliable, and most of all increasingly affordable genetic markers brought about drastic changes to present-day breeding methods. In this chapter, we focus on four significant genetic marker-dependent approaches with significant potential to directly or indirectly change contemporary tree breeding methods. These include pedigree reconstruction, pedigree-free models, association genetics, and genomic selection.

1 Introduction

Tree breeding programs are resource- and time-dependent endeavours. The selection and testing phases are often conducted over vast geographic areas with large trials, requiring frequent and long-time monitoring and assessment. The lowest-intensity

Y.A. El-Kassaby (✉)
Department of Forest Sciences, Faculty of Forestry,
University of British Columbia, Vancouver, BC V6T 1Z4, Canada
e-mail: y.el-kassaby@ubc.ca

F. Isik • R.W. Whetten
Department of Forestry and Environmental Resources,
North Carolina State University, Campus Box 8002, Raleigh, NC 27695, USA
e-mail: fisik@ncsu.edu; Fikret_Isik@ncsu.edu; ross_whetten@ncsu.edu

T. Fenning (ed.), *Challenges and Opportunities for the World's Forests in the 21st Century*, Forestry Sciences 81, DOI 10.1007/978-94-007-7076-8_18,

approach to tree improvement is a reciprocal transplanting-like approach known as provenance testing (Callaham 1964) for the identification of superior seed sources for reforestation. Provenance testing allowed evaluating several seed sources originating from multiple locations within the species' natural range through their field-testing over potential target planting areas. This process aided in identifying superior seed sources and their adaptability for the safe transfer of their seed to the new planting sites (Rehfeldt 1983). Provenance testing focused on acquiring precise knowledge of the seed sources and their performance over testing sites (Konig 2005). This process is a simple population improvement method, as the pedigree or genealogy of the tested material is often unknown. The main achievement of provenance testing is the delineation of areas for safe seed transfer, known as seed zones (Campbell 1986).

The first and simplest pedigree-known testing utilized wind-pollinated/open-pollinated families (also known as half-sib families because their offspring share the seed donors' genotype). Wind-pollinated testing, as a partial pedigree method, permits within and among family selection, thus it is expected to yield greater gains than provenance testing. The New Zealand radiata pine tree improvement program is the most notable program for adopting this approach (Burdon and Shelbourne 1971). The main attractive feature of this method is its simplicity and suitability for testing large number of families; however, it is often considered as a spring-board to full pedigree testing (Jayawickrama and Carson 2000). It should be stated that wind-pollinated testing is fraught with assumptions that cannot be either tested or fulfilled, and often leads to inaccuracies in estimates of individual breeding values (Namkoong 1966).

The utilization of a full pedigree (i.e., individuals with known genealogy) is the most common testing mode in tree breeding programs (White et al. 2007). The formation of a structured pedigree, created through the implementation of a mating design of controlled pollinations, provides greater control of the genealogy and the eventual accurate estimation of genetic parameters such as trait heritabilities and parent and offspring breeding values (Namkoong et al. 1988). It should be stated that the successful completion of structured pedigree is an elaborate process requiring time and substantial painstaking effort. The recurrent selection scheme is the most common breeding framework used when full pedigree is used (Allard 1960).

2 Pedigree Reconstruction

Structured pedigree designs (full- and half-sib families) constitute the backbone for most tree breeding programs, resulting in impressive gains and better management of inbreeding and genetic diversity (White et al. 2007). Lambeth et al. (2001) introduced an idea of polymix breeding and pedigree reconstruction. El-Kassaby and Lstibůrek (2009) further implemented this idea via the posterior analysis of naturally-occurred crosses among a group of parents. They coined the method "Breeding without Breeding (BwB)" and proposed the utilization of molecular markers, SSRs in this case, and pedigree reconstruction models (see Jones and

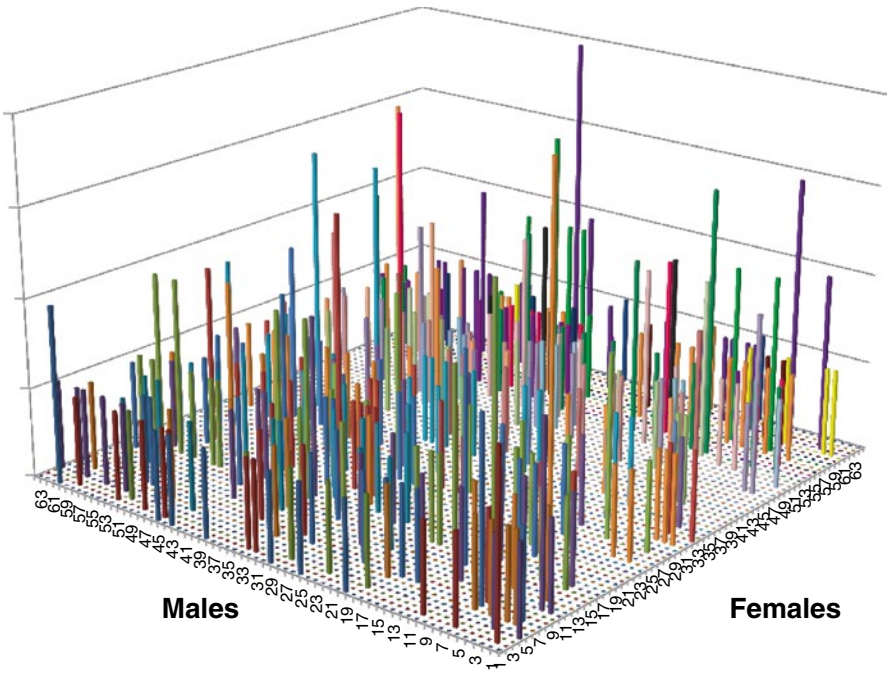

Fig. 1 Distribution of posteriorly assembled naturally-occurred crosses among 63-parent lodgepole pine seed orchard revealed by full pedigree reconstruction of bulk offspring (i.e., unknown maternal and paternal parentage) using DNA microsatellite markers (nine nuclear and six chloroplast loci) and pedigree reconstruction (El-Kassaby, unpublished)

Ardren 2003 for review) to by-pass the costly and time consuming breeding phase. The disconnected partial diallel mating scheme is often employed to create the structured pedigree for generating the offspring needed for testing (Namkoong et al. 1988). The BwB concept is illustrated using bulk seed sample from a 63-parent lodgepole pine seed orchard (El-Kassaby, unpublished), and can be compared with the disconnected partial-diallel design. With this number of parents and the implementation of a six-parent scheme, 153 full-sib families are expected to be generated (seven 6-parent and three 7-parent partial diallel units). However; when pedigree reconstruction was implemented, a total of 446 full-sib families were assembled without making any controlled crosses (Fig. 1). The resulting mating is far more efficient as many more crosses were created as compared to the classical disconnected partial diallel.

Furthermore, El-Kassaby et al. (2011) extended the BwB concept and increased the method's efficiency through the application of two distinct steps: (1) the use of simplified half-sib progeny testing with large sample size per parent and (2) restricting offspring sampling for DNA fingerprinting and pedigree reconstruction to a random sample of offspring from a subset of parents rather than the entire parental population. The use of half-sib families in testing is expected to simplify the

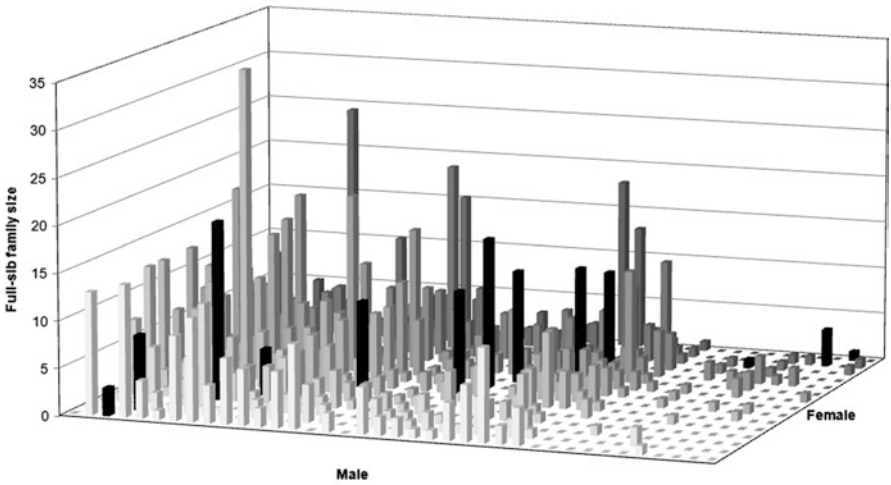

Fig. 2 Pedigree reconstruction from natural mating produced from seed collected from 15 seed donors growing in a 41-parent western larch seed orchard showing the formation of full-sib families nested within the maternal and paternal half-sib families with selfing presented as *black bars* (After El-Kassaby et al. (2011))

progeny test design as compared to multiple full-sib families. A random sample of offspring from a subset of seed parents is expected to capture most of the un-sampled parents as fathers (i.e., paternal half- and full-sib families) and therefore their breeding values can be estimated. Finally, the inclusion of all the offspring phenotypic information from both full- and half-sib families is expected to increase the estimated genetic parameters' precision; however, it should be stated that the breeding value of the half-sib individuals will be estimated with lesser precision as compared to those of full-sib families. El-Kassaby et al. (2011) empirically tested this concept and assessed offspring generated from only 15 seed-donors (i.e., half-sib families) out of a 41-parent western larch seed orchard. In this experiment, each half-sib family was represented by 400 seedlings bringing the total experiment sample size to $N \approx 6{,}000$. They randomly sampled 1,500 individuals, irrespective of their half-sib family designation, for DNA fingerprinting and pedigree reconstruction. As expected, an unbalanced mating structure was produced reflecting variation in parental reproductive output (Fig. 2).

It is interesting to note that the assembled matings produced offspring sired by all 41 parents in the orchard, indicating that the pedigree reconstruction successfully captured the un-sampled parents as pollen donors even when the offspring sampling was restricted to 15 seed-donors only. The most interesting observation from the data analyses is the congruence between height breeding values from the combined analysis (1,500 FS + 4,500 HS) and that based on the conventional full-sib families alone (1,500 individuals). This was observed for both parents and offspring (Fig. 3). The great advantage of the FS and HS combined analysis is the role

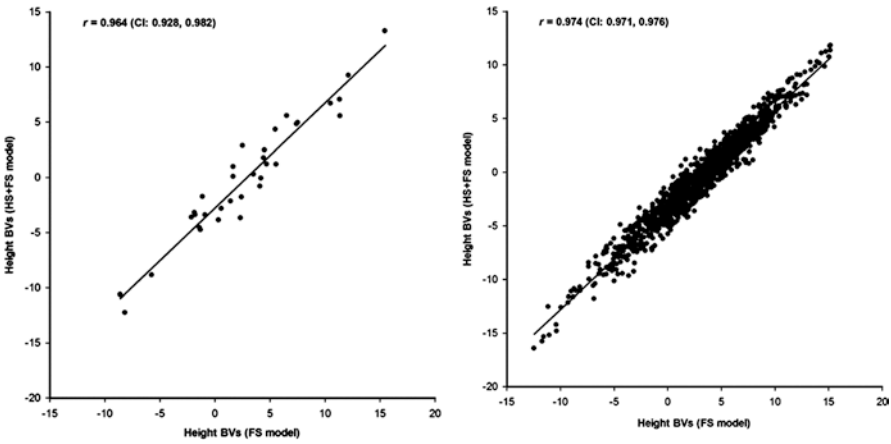

Fig. 3 Scatter plot of predicted breeding values for parents (*left*) and offspring (*right*) from the incomplete (combined HS + FS) and complete (FS) pedigree models. Pearson correlation (r) is in the left corner of each graph (After El-Kassaby et al. (2011))

played by the 1,500 FS individuals in linking the remaining 4,500 HS to the paternal and maternal parents and their half- and full-sib families (Fig. 3). Furthermore, El-Kassaby et al. (2011) demonstrated that individuals' breeding values precision did not change drastically if the random sampling of individuals for fingerprinting and pedigree reconstruction was reduced to approximately one third (i.e., less fingerprinting efforts).

Pedigree reconstruction is an effective method in situations where the posterior determination of offspring genealogy is needed or for species that do not lend themselves to controlled pollination. Using pedigree reconstruction for trees from plantation blocks that originated from seed orchards or breeding arboreta can instantaneously convert them to progeny test trials (Hansen and McKinney, 2010). While this approach requires good GIS tracking of plantations polygons over the landscape (see Ding et al. 2012), it also requires rigorous spatial analysis to account for site heterogeneity (see Cappa et al. 2011).

3 Pedigree-Free Models

Fundamentally, Breeding without Breeding is anchored to the utilization of pedigree reconstruction to assemble half- and full-sib families needed for conducting standard intra-class correlation analyses for estimating quantitative genetics parameters such as traits' heritabilities and parental and offspring breeding values (Falconer and Mackay 1996). In situations where pedigree reconstruction is not feasible, molecular genetic markers offer an alternative approach for estimating quantitative genetic parameters. Molecular markers can be used to estimate

"marker-based pairwise relationships" among any group of individuals irrespective of their genealogy, based on the assumption that markers identical by state are also identical by descent (Li et al.1993; Queller and Goodnight 1989; Lynch and Ritland 1999; Wang 2002). The use of "marker-based pairwise relationship" created an opportunity to studying domesticated and undomesticated species in experimental or natural setting with and without the availability of pedigree, thus permitting the estimation of genetic parameters in an unstructured population. Efficient methods have been developed for the use of high-density marker information for a group of individuals to estimate their realized relationship matrix (vanRaden 2008). This matrix is used in place of the classical pedigree-based numerator relationship matrix required in quantitative genetics analyses. This approach allows estimating quantitative genetic parameters such as narrow sense heritability and breeding values using the genomic best linear unbiased prediction method, as described in more detail below (Zapata-Valenzuela et al. 2011; El-Kassaby et al. 2012; Porth et al. 2012).

The realized relationship matrix was successfully used to estimate narrow sense heritability, breeding value and genetic and phenotypic correlations in an unstructured black cottonwood population (El-Kassaby et al. 2012; Porth et al. 2012). More interesting is the study of Klápště et al. (2013) in which a pedigree-free model was compared to a marker-based pairwise relationship model. Surprisingly, Pearson's product moment and Spearman's rank correlations between western larch offspring breeding values produced from the two approaches were highly significant, indicating that the generated DNA-based pair-wise relationship matrix is indeed a valid substitute for the classical pedigree matrix (Fig. 4). This approach was further extended to accommodate a mixture of information generated from both genetic markers and conventional pedigree by Korecký et al. (2013). This approach is unique as the combination of historical and contemporary co-ancestry generated by the genetic markers and pedigree, respectively, could not be attained by either approach individually. Thus, combining both data sets is expected to improve the accuracy of the estimated genetic parameters as the often ignored Mendelian sampling term in structured pedigree is precisely accounted for when molecular markers are used.

The availability of molecular markers is expected to effectively increase breeding efficiency. The use of densely well dispersed SNP data to estimate the realized relationship among individuals is expected to result in a greater kinship resolution and offers an opportunity improvement to classical breeding efforts.

4 Marker-Trait Association

The availability of cost-effective molecular genetic marker systems opens the door to analysis of the genetic basis of phenotypic traits measured in breeding populations. Classical quantitative genetics approaches, whether based on provenance, pedigree, or realized relationship matrices, are based on the 'infinitesimal model' proposed by

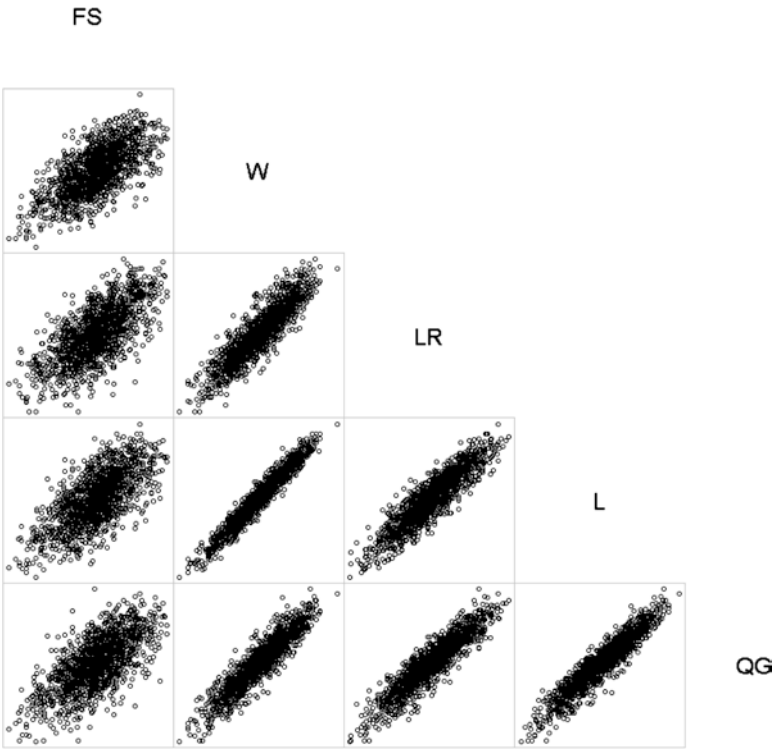

Fig. 4 Correlations of individuals' breeding values produced from pedigree-based full-sib (FS) and four molecular genetic markers-based pairwise relationship estimation methods (W: Wang (2002); LR: Lynch and Ritland (1999); L: Li et al. (1993); QG: Queller and Goodnight (1989)) (After Klápště et al. (2013))

Fisher (1918). Fisher's model reconciled the disparate views of geneticists who studied quantitative traits that show continuous variation, and geneticists who studied discrete characters controlled by single genes, by hypothesizing that continuous variation is the cumulative effect of many different genes, each with a small and approximately equal additive effect on the phenotype. This model has been extremely useful for close to a century, and recent publications have reviewed the substantial body of evidence supporting the main features of the model (Hill et al. 2008; Stranger et al. 2011). This model has important implications for efforts to understand the molecular genetic mechanisms that underlie phenotypic variation in forest tree breeding programs, and for breeders interested in accurately predicting genetic merit of individuals based on genotype information.

The analytical approach called "association genetics" was described over 15 years ago (Lander and Schork 1994; Risch and Merikangas 1996) as an alternative to family-based linkage mapping approaches to characterize the genetic basis of human disorders. Much more work has been done using association genetics in the

field of human biomedical genetics than in any other area, and much has been learned about the strengths and weaknesses of the approach (reviewed by Stranger et al. 2011; Rowe and Tenesa 2012). Neale and Savolainen (2004) reviewed key requirements for association genetics, and proposed that populations of conifers (and by extension, other wind-pollinated forest tree species) would be suitable experimental materials for association genetics. Applications of association genetics in tree breeding were described by White et al. (2007, pp. 543–547) and Wilcox et al. (2007); a brief overview will be given to set the stage for discussion of the current status.

The fundamental concept in association genetics is to test for a statistical association between the allelic state at a genetic marker locus in an individual and the phenotype of that individual, for many individuals in a population. The value of such associations is that they can help to identify the molecular basis for phenotypic variation, which in turn may provide molecular markers useful for marker-assisted breeding (Neale and Savolainen 2004). The power to detect associations is a function of several parameters, including the presence of population structure (Neale and Savolainen 2004), the extent of linkage disequilibrium in the test population, the size of the test population, and the proportion of phenotypic variation accounted for by each causative genetic variant involved in the phenotype of interest. The genetic variants tested for association with phenotype may be in known genes that are believed to play a role in controlling the phenotype under study (the 'candidate gene' approach), or they may be chosen on the basis of the allele frequencies in the population and distribution in the genome (the 'genome-wide' approach). As with any statistical testing procedure, if multiple tests of the same hypothesis are conducted, false positive (Type I) errors are likely unless the significance threshold is corrected for the number of tests made. Risch and Merikangas (1996) proposed a threshold of 5×10^{-8} for genome-wide significance in an experiment testing associations of one million single-nucleotide polymorphism (SNP) loci in the human genome; more recent publications have refined this estimate slightly for different sets of human SNP loci (Li et al. 2012). Linkage disequilibrium (LD), the non-random association between allelic states at different loci, affects the independence of multiple tests, and so correction for multiple testing should take into account patterns of LD among the loci analyzed.

An early study of linkage disequilibrium in Douglas fir, based on a relatively small sample of 18 genes from 32 haploid megagametophyte samples, concluded that each gene contained 2–3 independent "haploblocks" of genetic variation, and 4–5 SNP loci per gene would be required to adequately sample the genetic variation in each gene (Krutovsky and Neale 2005). This study focused on transcribed regions, because relatively few resources were available at the time for analysis of non-transcribed regions of genomic DNA in any conifer species. The majority of SNPs identified as significantly associated with target traits in human GWA studies are in non-coding sequences (45 % in introns and 43 % in intergenic regions; Hindorff et al. 2009), suggesting that efforts to model the genetic variation underlying phenotypic variation must include analysis of non-coding genomic DNA sequences. Fortunately, reference genome sequencing projects are now underway

for loblolly pine, white spruce, and Norway spruce (searchable abstracts available on-line at https://pag.confex.com/pag/xx/webprogram/start.html), and reference genome sequences are already available for poplar (Tuskan et al. 2006) and eucalyptus (available on-line at http://phytozome.net/), so genomic sequence information will be more readily available for future efforts to model genetic variation.

Determination of the appropriate sample size and number of genetic loci to test in order to achieve a specific level of power in an association study requires evaluation of several population parameters that affect power (Ball 2005; Spencer et al. 2009). The magnitude of the genetic effect of a locus, the frequency in the population of the allele that causes an effect, and the extent of LD between the causative allele and nearby genetic markers (e.g. SNPs) are some of these parameters. Association studies in humans primarily focus on disease-related phenotypes, and the magnitude of the genetic effect is often expressed as a ratio of the likelihood of disease occurrence in a heterozygous individual to the likelihood of disease in an individual homozygous for the most common allele (genotypic risk ratio, Risch and Merikangas 1996, or relative risk per allele, Spencer et al. 2009). The structure of linkage disequilibrium in the human genome is complex enough that simulation is the most general approach to modeling the dependence of experimental power on sample size, relative risk, and allele frequency (Spencer et al. 2009). Such simulations indicate that power is lower for lower risk allele frequencies, for lower risk per allele, and for lower numbers of genetic variant loci tested; for a relative risk per allele of 1.5, an array that assays one million SNP loci provides only about 50 % power in a sample size of 5,000 when the risk allele frequency is less than 10 % (Spencer et al. 2009). A relative risk per allele of 1.5 is roughly equivalent to accounting for 5 % of phenotypic variation, although that equivalence is affected by allele frequency in the population; relatively few loci detected to date in human genome-wide association studies have effects that large (Stranger et al. 2011). This suggests that association genetics studies will not be powerful enough to detect individual genes that account for a significant proportion of phenotypic variation in complex traits in forest trees, if the infinitesimal model is accurate. Some traits of interest to tree breeding programs, such as resistance to fusiform rust disease in *Pinus taeda*, are controlled by individual genes with major effects (Wilcox et al. 1996); association genetics approaches are well-suited to analysis of such traits.

Height growth is an important phenotype in many tree breeding programs, so results of association genetics analysis of height in humans are of interest. Yang et al. (2010) reported that joint analysis of all SNPs as random effects in a mixed linear model that incorporated relationship information derived from marker genotypes explained almost half the genetic variation in height in a sample human population of less than 4,000 individuals, although all 180 loci identified by meta-analysis of association studies in a combined population of 183,727 individuals (Lango Allen et al. 2010) together explained about 14 % of the genetic variation in height. The difference between the analytical approaches taken by these two groups is that Yang et al. focused their attention on creating a predictive model, without concern for identifying specific loci, while Lango Allen et al. followed a more classical association approach using rigorous statistical methods to reduce the likelihood of

false positive results and identify loci and pathways mechanistically related to height growth. Many of the loci identified by Lango Allen et al. can be grouped into biological pathways with recognized effects on growth and development, and in many cases, multiple genetic variants were identified per gene (Lango Allen et al. 2010). This phenomenon, referred to as allelic heterogeneity, reduces power in association analyses, because the same phenotype can be due to multiple different genetic variants, even at the same functional gene. Occurrence of multiple genetic variants within genes that affect the same phenotype creates the possibility for epistatic interactions; epistatic interactions within genes or between tightly-linked genes can result in differences between the heritability estimated from closely-related individuals versus distantly-related individuals (Haig 2011; Würschum et al. 2012; Zuk et al. 2012). The approach of analyzing association genetics data by grouping variants into functional genes, organizing genes into pathways, and integrating genetic pathways with gene expression data may provide additional power for understanding phenotypic variation, if modeling approaches that can take pathway structure and gene expression patterns into account can be developed (Cookson et al. 2009; Bennett et al. 2012; Kreimer et al. 2012; O'Hagan et al. 2012). Another approach, similar to that used by Yang et al. (2010), is to incorporate all SNP loci as random effects in the association analysis; this approach has been reported to overcome disadvantages of both traditional linkage analysis and association analysis methods in livestock (Kemper et al. 2012). This type of analysis has much in common with genomic selection, discussed later in the chapter.

Allele frequency of the minor allele at biallelic SNP loci has a major impact on the power of association genetics studies (Spencer et al. 2009; Stranger et al. 2011). Most SNP loci in a sample of over 3,000 SNPs assayed in over 900 loblolly pine trees had minor allele frequencies of less than 15 % (Eckert et al. 2010). Such low minor allele frequencies in samples of unrelated populations contributes to a requirement for extremely large sample sizes to achieve significance in traditional association genetics studies; only alleles with relatively large effects can be detected unless sample sizes exceed 5,000 and marker allele frequency is close to causative variant allele frequency (Ball 2005; Stranger et al. 2011). Structured populations descended from a smaller number of parents can reduce this problem by increasing the frequency of rare alleles that occur in that sample of parents. This strategy has been used to develop the maize Nested Association Mapping (NAM) population (Yu et al. 2008; McMullen et al. 2009), and methods to deal with the population structure that arises in populations produced from mating designs have also been developed (Yu et al. 2006). The combined use of the NAM population and a more typical association population of 282 inbred lines allowed identification of several SNPs that affect maize kernel composition (Cook et al. 2012). Similar strategies may become feasible in forest tree breeding programs, once reference genome sequences are available and haplotype information can be readily developed for the parents of elite breeding populations.

Understanding of molecular mechanisms underlying phenotypic variation is not the primary objective of breeding programs – instead, the objective is to create models of genetic variation in breeding populations that have predictive power to

identify individuals of high genetic merit. Studies that increase understanding molecular mechanisms can contribute to development of predictive genetic models in the long term, while studies that focus on developing models of inheritance of complex traits in breeding populations have more immediate value in the short term. Understanding molecular mechanisms can be challenging in human biomedical genetics (Peters and Musunuru 2012), and will be even more challenging for most trees of interest to breeding programs. The association genetics approach can contribute fundamental understanding of mechanisms underlying traits controlled by relatively small numbers of genes, but traits controlled by many genes of equal and small effects will be very expensive to analyze using this method.

5 Genomic Selection

5.1 Background

Many traits of interest to breeders are polygenic, being controlled by many genes each with small effect (Hill et al. 2008). These small-effect genes are crucial for the success of complex trait improvement (Crosbie et al. 2003). For many decades plant and animal breeders relied on phenotype and resemblance among relatives to capture genetic variance explained by these small effect genes. The methods used to improve complex traits were 'black box' as breeders did not know the underlying genetic architecture of complex traits, such as the number of genes controlling the trait and their location in the genome. Tree breeders have adopted these methods since 1950s. The success in improvement of tree characteristics has been relatively modest because breeding-testing-selection cycles for forest trees take many years to complete and tree breeding is logistically complex. Breeders have long looked to molecular markers to overcome challenges and improve the efficiency of selection (Neale and Savolainen 2004).

Beginning in late 1970s quantitative trait loci (QTL) mapping and later candidate gene approaches have been explored as tools to explain gene architecture of complex traits. The idea was that if alleles with large effects on the trait are traced (oligogenic model) with the markers, they could be used for selection of superior genotypes in breeding populations. This concept is called marker aided selection (MAS). However, QTL mapping and candidate gene approaches have had limited use to improve quantitative traits in most plant and animal breeding programs. Major reasons include the cost of producing large number of markers, and the observation that most quantitative traits are controlled by many QTLs, each with small effect, as predicted by the infinitesimal model. Individual QTLs often explained only a small percent (<5 %) of total variance and marker-trait associations discovered in individual families were not repeatable across the population (Goddard and Hayes 2009; Neale 2007).

QTL mapping experiments have been useful in discovering the genetic architecture of quantitative traits important in agricultural and forestry, but the focus is on

identifying genetic loci associated with phenotypes. In breeding, on the contrary, the emphasis is on predicting genetic merit of individuals or lines rather than on discovering individual genes. A good predictor of genetic merit does not have to identify the underlying genes (Goddard and Hayes 2009). What is needed is a large number of markers to populate the genome and to explore the LD between these markers and the many QTL with small effect. This approach is called genomic selection (GS) or genome-wide selection. Since the introduction of the concept by Meuwissen et al. (2001), GS has shifted the paradigm, driven by the increased efficiency in DNA sequencing technologies and computing power.

GS contrasts greatly with traditional MAS, because in GS there is no defined subset of significant markers used for selection. Instead, GS jointly analyzes all markers in a population, attempting to explain the total genetic variance with dense genome-wide marker coverage through summing marker effects to predict breeding values of individuals (Meuwissen et al. 2001). The idea is that if we populate the genome with high-density markers, we can capture the LD between markers or marker haplotypes and causal polymorphism. Such association would be consistent across different families (Meuwissen et al. 2001). With advancement in DNA sequencing technologies and efficiency in genotyping, GS has become a reality in dairy cattle breeding (Goddard and Hayes 2009). Many livestock breeding programs now routinely apply GS to market bulls (Hayes et al. 2009). Genomic selection processes start from a training population. Candidates to establish a next cycle of breeding are selected through GS. The training can be performed iteratively as new phenotype and marker data accumulate (Heffner et al. 2011).

5.2 Empirical Examples from Forest Trees

Forest tree breeding programs are still at the first stage of breeding-testing and selection cycles with little genetic difference from natural populations. If successful, the impact of genomic selection on forest tree breeding could be far greater than for other crops or animal breeding programs. A few early empirical studies on genomic selection in forest trees are encouraging. For example, in a cloned loblolly pine breeding population, accuracies of GS varied between 0.55 and 0.88, matching those achieved by conventional phenotypic selection (Resende et al. 2012). Similarly in the same species, Isik et al. (2011) reported genomic estimated breeding values with reliability as high as breeding values based on resemblance among relatives and phenotypic data. These studies estimated the individual marker effect and summed up the coefficients to estimate genomic estimated breeding values of trees.

Alternatively a smaller subset of markers can be used to estimate realized genomic relationships using frequency of alleles shared by individuals (Legarra and Misztal 2008). Then, the additive genetic relationship matrix derived from pedigree is substituted by the genomic relationship matrix to predict genomic estimated breeding values. Genomic BLUP (GBLUP) could be a powerful tool for forest tree breeding programs. Such models can capture the Mendelian segregation effect in

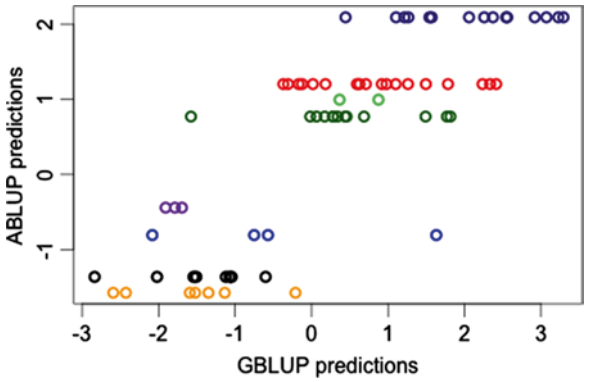

Fig. 5 Predicted breeding values of loblolly pine clones based on pedigree (*y-axis*) and genomic BLUP (*x-axis*) for eight crosses. Each cross is designated with a different color. In the absence of phenotype, the expected breeding value of sibs would be the same, which is the mid-parent value (ABLUP). However, DNA markers can capture Mendelian sampling effect within each cross as shown here, and thus, sibs can be ranked and selected without progeny testing (Zapata-Valenzuela et al. 2011)

full-sib families, which was not the case using the average additive genetic relationships. For example, Zapata-Valenzuela et al. (2011) showed that accuracies of genomic estimated breeding values using GBLUP were comparable to traditional pedigree-based BLUP methods. In the same study, breeding values of a training population were estimated using GBLUP and classical BLUP (Henderson 1984). In the absence of phenotype, sibs from a cross had the same mid-parent breeding values when classical BLUP was used (Fig. 5). However, genomic relationship matrices based on SNP markers allowed prediction of different genetic values for sibs from a single cross.

5.3 Statistical Machinery

Classical linear mixed models are not efficient to handle large number of markers as predictors because the number of predictors (p) is larger than the number of data points (n) to explain variance in the phenotype. Such large p and small n effect causes lack of degrees of freedom. Statistical analysis of large number of markers has been a very active area of research in recent years, and many statistical methods have been proposed in the literature (Gianola et al. 2009). The effect of markers or haplotypes can be estimated by simultaneously including all markers in a model, but the challenge is to estimate the variances of marker effects. The best linear unbiased prediction (BLUP) method and ridge regression approaches have been proposed to estimate individual marker effects (Meuwissen et al. 2001; Whittaker et al. 2000). These methods make the assumption that markers are sampled from a population with expectation $N \sim \left(0, \sigma_g^2\right)$ and each marker explain the same (σ_g^2 / n)

amount of genetic variance. Rather than categorizing markers as either significant or as having no effect, ridge regression and BLUP shrink all marker effects toward zero (Meuwissen et al. 2001). This is not a realistic assumption because regardless of association of markers with the trait loci, all the markers are shrunk towards the mean at the same level. Bayesian methods have a natural way of taking into account uncertainty about all unknowns in a model (e.g., Gianola et al. 2009) and, when coupled with the power and flexibility of Markov Chain Monte Carlo, Bayesian methods can be applied to almost any parametric statistical model. Meuwissen et al. (2001) introduced BayesA and BayesB and compared them with BLUP method in their original paper on GS. In BayesA, all the markers explain a fraction of genetic variance and the variance explained by each marker can vary based on the scaled inverted chi-square distribution as prior. Method BayesB corrects the shortcoming of BayesA by shrinking a high proportion (π) of markers to zero. Bayes C, Cπ, and D and Dπ were introduced to address the undesirable effect of priors on estimations observed for BayesA and BayesB. Habier et al. (2011) concluded that accuracies of the alternative Bayesian methods were similar and none of them outperformed all others across all traits and training data sizes. The choice of statistical methods for GS is sometimes is a matter of practicality, time and ease of application. Examples on empirical and simulated data suggest that Bayesian approaches are efficient to increase accuracy of predictions but the increase is usually minimal unless a large fraction of genetic variance in the trait in question is controlled by a few loci.

5.4 Challenges of GS in Forest Tree Breeding

Despite advances in the efficiency of genotyping technologies, genotyping is still costly for forest trees. For example, the Illumina SNP genotyping platform costs about $150 per sample for loblolly pine as of 2012, though the cost is decreasing. Several labs in the USA and other countries are working on alternative genotyping technologies, such as genotyping by sequencing (Baird et al. 2008; Elshire et al. 2011; Peterson et al. 2012; Poland et al. 2012; Truong et al. 2012), and we expect that the cost of genotyping could be less than $50 as of 2013.

GS has been successful in cattle breeding because the number of founders in these populations is relatively small (<30) and the LD between markers and trait loci are large, thanks to deep pedigree in the populations and small effective population size. Tree breeding populations still are at their infancy. The pedigree structures are still shallow with very low linkage disequilibrium (Neale and Savolainen 2004). Marker-trait phase detected in one generation may not hold in a subsequent generation because of meiotic recombination. For GS to be successful, well-structured populations (small effective population size, multiple generations) are needed.

Conifers are major targets of breeding programs in the northern hemisphere, and they have large and complex genomes. GS require dense coverage of whole genome to trace many QTLs associated with phenotype. Many more markers

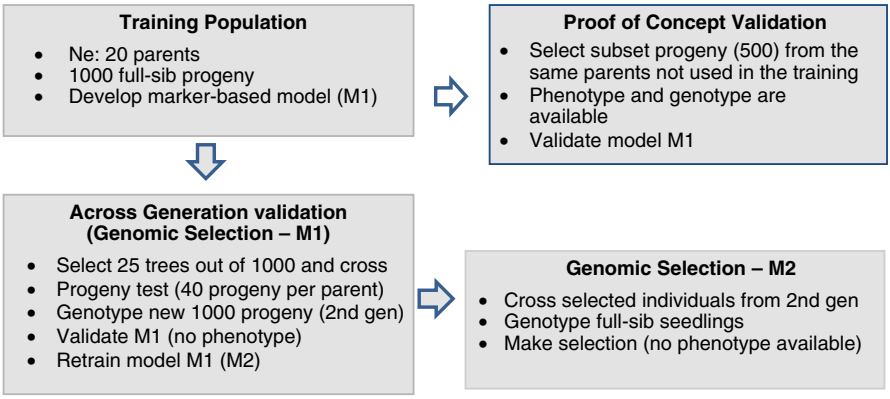

Fig. 6 Genomic selection process for an elite breeding population. A marker-based prediction model is retrained across multiple generations. Such process would make the model more powerful for genomic estimated breeding values to trace LD of markers and QTLs

might be needed to populate genome of conifers. Grattapaglia and Resende (2011) suggested that 20 markers/cM are needed for an effective population size of greater than 30.

Forest trees have some advantages in implementation of GS. A large population can be put together easily. Each family can be represented by large number of progeny (several hundreds) with little investment and time. Phenotyping can be quite accurate thanks to efficient experimental designs and cloning of individuals.

An example GS plan has been proposed for a loblolly pine breeding population within the North Carolina State University Tree Improvement Program in the USA (Fig. 6). In the diagram given in Fig. 6, the process starts with creating a training population with an effective population size (Ne) smaller than 50 parents. In this example, 20 parents are used. Relatedness among the 20 founders is desirable, because that will make the marker-based model more powerful to predict GEBV by tracing historical LD in the population. From full-sib crosses of 20 parents, about 1,000 individuals can be genotyped. This progeny population is field-tested and breeding values are obtained Deregressed breeding values of 1,000 individuals or phenotypic values adjusted for fixed effects can be obtained to use as new 'phenotype' for development of a marker-based model (M1).

There are different methods to validate the predictive ability of markers. An additional 500 progeny from the same crosses (with known phenotype and genotypes) can be used as a validation population. Alternatively, random sampling of a small subset of progeny or selection of subset of progeny within each full-sib family can be used to validate model M1. This step is a proof of concept to show that the model has predictive power, and is not necessarily an application of GS. In order to utilize the benefit of GS approaches, we need to breed the selected individuals from the training population, obtain seeds, and use M1 to make selection decisions. This can be called 'across generation' GS application. The M1 model can be retrained

when more genotypic and phenotypic data become available as breeding progresses (M2). GS training models would have more reliability as new data are included and can be used for multiple generations.

6 Conclusions

The availability of cost-effective genetic markers in forest tree species is expanding rapidly due to advances in DNA sequencing technology and investment in determining the reference genome sequences for several commercially-important species of forest trees. These resources are likely to fundamentally change the way tree breeding programs characterize genetic variation in their breeding populations, and several research groups are actively working to develop methods for applications of these tools in practical breeding programs. Molecular markers are already useful tools for population management applications such as validation of crosses, pedigree reconstruction, and unambiguous identification of clones. Association genetics results have already been reported for several traits in various species of forest trees, and application of these results in practical breeding programs may follow soon. Development of more sophisticated analytical methods capable of integrating the analysis of genetic variation detected by SNP assays with variation in gene expression patterns, metabolite levels, and phenotypic measurements may provide new tools capable of more accurate prediction of genetic value based on molecular assays. Predictive modeling of genetic value is the central objective of genomic selection methods, which have shown considerable promise in livestock and crop species that have appropriate patterns of LD in breeding populations. Forest tree breeding populations are likely to have very different patterns of LD than livestock or crop species, and new approaches to genomic selection may be required in order for this method to reach its full potential in applied tree breeding programs.

References

Allard RW (1960) Principles of plant breeding. Wiley, New Year

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA, Selker EU, Cresko WA, Johnson EA (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376

Ball RD (2005) Experimental designs for reliable detection of linkage disequilibrium in unstructured random population association studies. Genetics 170:859–873

Bennett BD, Xiong Q, Mukherjee S, Furey TS (2012) A predictive framework for integrating disparate genomic data types using sample-specific gene set enrichment analysis and multitask learning. PLoS One 7(9):e44635

Burdon RD, Shelbourne CJA (1971) Breeding populations for recurrent selection: conflicts and possible solutions. N Z J For Sci 1:174–193

Callaham RZ (1964) Provenance research: investigation of genetic diversity associated with geography. Unasylva 18:40–50

Campbell RK (1986) Mapped genetic variation of Douglas-fir to guide seed transfer in southwest Oregon. Silvae Genet 35:85–95
Cappa EP, Lstiburek M, Yanchuk AD, El-Kassaby YA (2011) Two-dimensional penalized splines via Gibbs sampling to account for spatial variability in forest genetic trials with small amount of information available. Silvae Genet 60:25–35
Cook JP, McMullen MD, Holland JB, Tian F, Bradbury P, Ross-Ibarra J, Buckler ES, Flint-Garcia SA (2012) Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. Plant Physiol 158:824–834
Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M (2009) Mapping complex disease traits with global gene expression. Nat Rev Genet 10(3):184–194
Crosbie, TM, Eathington, SR, Johnson, GR, Edwards, M, Reiter, R, Stark, S et al (2003) Plant breeding: past, present, and future. In: Lamkey, KR, Lee, M (eds) Plant breeding: the Arnel R. Hallauer International Symposium, Mexico City. 17–23 Aug 2003. Blackwell, Oxford, UK, pp 1–50
Ding C, McAuley L, Meitner MJ, El-Kassaby YA (2012) Evaluating interior spruce seed deployment with GIS-based modeling using British Columbia's Prince George seed planning zone as a model. Silvae Genet 61:271–279
Eckert AJ, van Heerwaarden J, Wegrzyn JL, Nelson CD, Ross-Ibarra J, González-Martínez SC, Neale DB (2010) Patterns of population structure and environmental associations to aridity across the range of loblolly pine (*Pinus taeda*L., Pinaceae). Genetics 185:969–982
El-Kassaby YA, Lstibůrek M (2009) Breeding without breeding. Genet Res 91:111–120
El-Kassaby YA, Cappa EP, Liewlaksaneeyanawin C, Klápšte J, Lstiburek M (2011) Breeding without breeding: is a complete pedigree necessary for efficient breeding? PLoS One 6:e25737
El-Kassaby YA, Klápšte J, Guy RD (2012) Breeding without Breeding: selection using the genomic best linear unbiased predictor method (GBLUP). New For 43:631–637. doi:10.1007/s11056-012-9338-4
Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379
Falconer DS, Mackay TFC (1996) Introduction to quantitative genetics. Longman, New York
Fisher RA (1918) The correlation between relatives on the supposition of Mendelian inheritance. Trans Roy Soc Edinb 52:399–433
Gianola D, de los Campos G, Hill WG, Manfredi E, Fernando R (2009) Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363
Goddard ME, Hayes BJ (2009) Mapping genes for complex traits in domestic animals and their use inbreeding programmes. Nat Rev Genet 10:381–391
Grattapaglia D, Resende MDV (2011) Genomic selection in forest tree breeding. Tree Genet Genomes 7:241–255
Habier D, Fernando RL, Kizilkaya K, Garrick DJ (2011) Extension of the Bayesian alphabet for genomic selection. BMC Bioinforma 12:186
Haig D (2011) Does heritability hide in epistasis between linked SNPs? Eur J Hum Genet 19:123
Hansen OK, McKinney LV (2010) Establishment of a quasi-field trial in *Abies nordmanniana* – test of a new approach to forest tree breeding. Tree Genet Genomes 6:345–355
Hayes BJ, Bowman PJ, Chamberlain AJ, Goddard ME (2009) Genomic selection in dairy cattle: progress and challenges. J Dairy Sci 92:433–443
Heffner EL, Jannink J-L, Sorrells ME (2011) Genomic selection accuracy using multifamily prediction models in a wheat breeding program. The Plant Genome 4:65–75
Henderson CR (1984) Applications of linear models in animal breeding. University of Guelph, Ontario
Hill WG, Goddard ME, Visscher PM (2008) Data and theory point to mainly additive genetic variance for complex traits. PLoS Genet 4:e1000008
Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci USA 106:9362–9367
Isik F, Whetten R, Zapata-Valenzuela J, Ogut F, McKeand S (2011) Genomic selection in loblolly pine – from lab to field. From IUFRO tree biotechnology conference 2011: from genomes to integration and delivery. BMC Proc 5(Suppl 7):I8

Jayawickrama KJS, Carson MJ (2000) A breeding strategy for the New Zealand radiata pine breeding cooperative. Silvae Genet 49:82–90
Jones AG, Ardren WR (2003) Methods of parentage analysis in natural populations. Mol Ecol 12:2511–2523
Kemper KE, Daetwyler HD, Visscher PM, Goddard ME (2012) Comparing linkage and association analyses in sheep points to a better way of doing GWAS. Genet Res (Camb) 94:191–203
Klápště J, Lstibůrek M and El-Kassaby YA (2013) Estimates of genetic parameters and breeding values from western larch open-pollinated families using marker-based relationship. Tree Genet Genome (in press)
Konig AO (2005) Provenance research: evaluating the spatial pattern of genetic variation. In: Geburek TH, Turok J (eds) Conservation and management of forest genetic resources in Europe. Arbora Publishers, Zvolen, pp 275–333
Korecký J, Klápště J, Lstibůrek M, Kobliha J, Nelson CD El-Kassaby YA (2013) Comparison of genetic parameters from marker-based relationship, sibship, and combined models in Scots pine multi-site open-pollinated tests. Tree Genet Genomes. doi:10.1007/s11295-013-0630-z
Kreimer A, Litvin O, Hao K, Molony C, Pe'er D, Pe'er I (2012) Inference of modules associated to eQTLs. Nucleic Acids Res 40:e98
Krutovsky KV, Neale DB (2005) Nucleotide diversity and linkage disequilibrium in cold-hardiness and wood quality-traits candidate genes in Douglas-fir. Genetics 171:2029–2041
Lambeth C, Lee B-C, O'Malley D, Wheeler N (2001) Polymix breeding with parental analysis of progeny: an alternative to full-sib breeding and testing. Theor Appl Genet 103:930–943
Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265(5181): 2037–2048
Lango Allen H et al (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832–838, 292 additional co-authors
Legarra A, Misztal I (2008) Technical note: computing strategies in genome-wide selection. J Dairy Sci 91:360–366
Li CC, Weeks DE, Chakravarti A (1993) Similarity of DNA fingerprints due to chance and relatedness. Hum Hered 43:45–52
Li MX, Yeung JM, Cherny SS, Sham PC (2012) Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet 131(5):747–756
Lynch M, Ritland K (1999) Estimation of pairwise relatedness with molecular markers. Genetics 152:1753–1766
McMullen MD et al (2009) Genetic properties of the maize nested association mapping population. Science 325(5941):737–740, 31 additional co-authors
Meuwissen THE, Goddard ME, Hayes BJ (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
Namkoong G (1966) Inbreeding effects on estimation of genetic additive variance. For Sci 12:8–13
Namkoong G, Kang HC, Brouard JS (1988) Tree breeding: principles and strategies. Springer, New York, Monograph, Theor Appl Genet 11
Neale D (2007) Genomics to tree breeding and forest health. Curr Opin Genet Dev 17:539–544
Neale D, Savolainen O (2004) Association genetics of complex traits in conifers. Trends Plant Sci 9(7):325–330, ISSN 1360–1385, 07/2004
O'Hagan S, Knowles J, Kell DB (2012) Exploiting genomic knowledge in optimising molecular breeding programmes: algorithms from evolutionary computing. PLoS One 7:e48862
Peters DT, Musunuru K (2012) Functional evaluation of genetic variation in complex human traits. Hum Mol Genet. doi:10.1093/hmg/dds363
Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS One 7:e37135
Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253

Porth I, Klápště J, Skyba O, Lai BSK, Geraldes A, Muchero W, Tuskan GA, Douglas CJ, El-Kassaby YA, Mansfield SD (2012) Populus trichocarpa cell wall chemistry and ultrastructure trait variation, genetic control, and genetic correlations. New Phytol 197:777–790

Queller DC, Goodnight KF (1989) Estimating relatedness using genetic markers. Evolution 43:258–275

Rehfeldt GE (1983) Seed transfer guidelines for Douglas-fir in Central Idaho. U. S. For Serv Res Note INT-337

Resende MFR, Muñoz P, Acosta JJ, Peter GF, Davis JM, Grattapaglia D, Resende MDV, Kirst M (2012) Accelerating the domestication of trees using genomic selection: accuracy of prediction models across ages and environments. New Phytol 193:617–624

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281):1516–1517

Rowe SJ, Tenesa A (2012) Human complex trait genetics: lifting the lid of the genomics toolbox – from pathways to prediction. Curr Genom 13:213–224

Spencer CC, Su Z, Donnelly P, Marchini J (2009) Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. PLoS Genet 5:e1000477

Stranger BE, Stahl EA, Raj T (2011) Progress and promise of genome-wide association studies for human complex trait genetics. Genetics 187:367–383

Truong HT, Ramos AM, Yalcin F, de Ruiter M, van der Poel HJA, Huvenaars KHJ, Hogers RCJ, van Enckevort LJG, Janssen A, van Orsouw NJ, van Eijk MJT (2012) Sequence-based genotyping for marker discovery and co-dominant scoring in germplasm and populations. PLoS One 7:e37565

Tuskan GA et al (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793):1596–1604, 109 additional co-authors

vanRaden PM (2008) Efficient methods to compute genomic predictions. J Dairy Sci 91:4414–4423

Wang JL (2002) An estimator for pairwise relatedness using molecular markers. Genetics 160:1203–1215

White TL, Adams WT, Neale DB (2007) Forest genetics. CABI Publishing, Cambridge, MA

Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252

Wilcox PL, Amerson HV, Kuhlman EG, Liu B-H, O'Malley DM, Sederoff RR (1996) Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc Natl Acad Sci USA 93:3859–3864

Wilcox PL, Echt CE, Burdon RD (2007) Gene-assisted selection: applications of association genetics for forest tree breeding (Ch 10). In: Oraguzie NC, Rikkerink EHA, Gardiner SE, de Silva HN (eds) Association mapping in plants. Springer, New York, p 278

Würschum T, Maurer HP, Dreyer F, Reif JC (2012) Effect of inter- and intragenic epistasis on the heritability of oil content in rapeseed (*Brassica napus* L.). Theor Appl Genet 126:435–441. doi:10.1007/s00122-012-1991-7

Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565–569

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208

Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551

Zapata-Valenzuela J, Isik F, Maltecca C, Wegryzn J, Neale D, McKeand S, Whetten R (2011) BMC Proc 5(Suppl 7):P60

Zuk O, Hechter E, Sunyaev SR, Lander ES (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci USA 109:1193–1198