 doi:10.1093/bfgp/elp048

Using association mapping to dissect the genetic basis of complex traits in plants

David Hall, Carolina Tegström and Pär K. Ingvarsson

Advance Access publication date 6 January 2010

Abstract

Association or linkage disequilibrium mapping has become a very popular method for dissecting the genetic basis of complex traits in plants. The benefits of association mapping, compared with traditional quantitative trait locus mapping, is, for example, a relatively detailed mapping resolution and that it is far less time consuming since no mapping populations need to be generated. The surge of interest in association mapping has been fueled by recent developments in genomics that allows for rapid identification and scoring of genetic markers which has traditionally limited mapping experiments. With the decreasing cost of genotyping future emphasis will likely focus on phenotyping, which can be both costly and time consuming but which is crucial for obtaining reliable results in association mapping studies. In addition, association mapping studies are prone to the identification of false positives, especially if the experimental design is not rigorously controlled. For example, population structure has long been known to induce many false positives and accounting for population structure has become one of the main issues when implementing association mapping in plants. Also, with increasing numbers of genetic markers used, the problem becomes separating true from false positive and this highlights the need for independent validation of identified association. With these caveats in mind, association mapping nevertheless shows great promise for helping us understand the genetic basis of complex traits of both economic and ecological importance.

Keywords: *association mapping; complex traits; genotyping; plants; population structure*

INTRODUCTION

Complex quantitative traits are usually influenced by a large number of genes as well as environmental effects. Understanding the genetic basis of complex traits have traditionally been the focus of quantitative genetics, which relies on partitioning phenotypic variation within and among individuals with known degrees of relatedness [1]. However, as the availability of useful genetic markers have increased, it has become possible to associate genome regions containing these markers to variation in complex traits. In quantitative trait locus (QTL) mapping, early generation crosses (F_1 or F_2) are used to dissect quantitative variation separating the individuals of the parental generation. QTL mapping has proven to be extremely useful in identifying many genome regions that influence complex traits in a large number of species [2–4]. However, the QTL approach suffers from a number of limitations. First, allelic variation in each cross is usually restricted because typically only two parents are used to initiate a QTL mapping populations. Second, since early generation crosses are used, the number of recombination events per chromosome is usually small, limiting the resolution of the genetic map. A typical QTL identified from a cross of

Corresponding author. Pär K. Ingvarsson, Department of Ecology and Environmental Science, Umeå Plant Science Centre, Umeå University, SE90187 Umeå, Sweden. Tel: +46-90-786-7414; Fax: +46-90-786-6705; E-mail: par.ingvarsson@emg.umu.se

David Hall is a graduate student at the Department of Ecology and Environmental Science, Umeå University a focusing on understanding the genetic basis of phenology differences in European Aspen (*Populus tremula*).

Carolina Tegström is a graduate student at the Department of Ecology and Environmental Science, Umeå University working on understanding the genetic basis of natural variation in wound-inducible herbivore defenses in European Aspen (*Populus tremula*).

Pär K. Ingvarsson is professor of Evolutionary Genetics at Department of Ecology and Environmental Science, Umeå University and a group leader at Umeå Plant Science Centre. His current research interest revolves around understanding the genetic basis of complex traits on plants on how these genes are shaped by various evolutionary processes.

consisting of a few hundred offspring can span anywhere between a few to tens of centiMorgans, which might correspond to genomic regions encompassing several megabases. Such large genome regions contain, typically, hundreds if not thousands of genes, making the process of identifying the causal gene in a QTL region, through techniques such as map-based cloning, a tedious and quite time-consuming task [5]. In addition, for many organisms the generation of mapping populations is either not possible or at least very time consuming. For instance, the long generation time of most forest trees have thus far either slowed down or completely prevented any progress in elucidating the genetic basis of complex traits using QTL mapping experiments [6].

Association or linkage disequilibrium mapping have been hailed as a more efficient way of determining genetic basis of complex traits. Association mapping relies on utilizing occurring variation in diverse germplasms and therefore does not suffer from the lack of variation that characterizes many QTL mapping populations. In addition, the naturally occurring recombination events that have occurred over evolutionary history also means that linkage blocks are substantially smaller in an association mapping population compared with a QTL mapping population and hence association mapping results in much more fine-scale mapping [7]. On the other hand, the limited extent of linkage disequilibrium suggests that a substantially greater number of genetic markers are needed to ensure adequate power to detect linkage between a marker and a causal locus [8]. Association mapping has rapidly come into focus as a very promising approach for the genetic dissection of complex traits, but it is also associated with potential problems and pitfalls. In this article, we review current aspects of using association mapping in plants, including how to initiate association mapping studies, the common methods for genotyping, phenotyping and how to ultimately analyze data to identify and verify causal association.

GENOTYPING

A common practice in many association genetic studies is to use unlinked and putatively neutral markers to characterize genetic variation in the accessions used in the mapping study and to account for population structure (for more on this, see below). There are many types of markers that can be used for this, including AFLP [9] and single sequence repeats (SSRs, also known as microsatellite markers). AFLP markers are easily obtained in almost any organism, even for those lacking previously existing genomics data. However, AFLP markers are almost exclusively dominant, that is the heterozygous genotype cannot be distinguished from one of the homozygous genotypes, and this introduces a number of problems when using AFLP markers for estimating, for instance, population structure [10] or for use directly in mapping [11]. SSR markers, on the other hand, are usually highly polymorphic but require a great deal of work to isolate and are rarely transferable between anything but the most closely related species [12]. The high variability of SSR markers, combined with the availability of semi-automatic detection methods, have, until recently, made them the markers of choice for use in estimating population structure or pairwise relatedness among individuals.

The development of next-generation sequencing technologies has allowed for unprecedented genotyping capabilities, even in organisms that have traditionally not been considered model organisms (see for example Varshney *et al*. [13], Simon *et al*. [14] or Nordborg and Weigel [15] for reviews of some recent applications in plants). The current next-generation sequencing technologies are capable of analyzing anywhere from hundreds of thousands to tens of millions of DNA molecules in parallel compared with hundreds at a time which is the maximum throughput of most traditional (i.e. Sanger-based) sequencing instruments. This massive increase in throughput stems from a change in methodology, from the traditional Sanger-sequencing method that produce read lengths of up to 1 kb from individual DNA clones to the current state-of-the-art sequencing technologies which produce read lengths in the range from 30 to 400 bp (although lengths are rapidly increasing) from templates consisting of beads or spots of DNA. The next step in sequencing technology will likely be the development of single molecule sequencing which will all but eliminate the need for extensive template preparation [16].

The rapid development in genomics has opened up the possibility to both identify and score a large number of genotypes in virtually any organism with relatively little effort. Next-generation sequencing technologies allows for rapid identification of a large number of genetic markers, mainly single

nucleotide polymorphisms (SNPs) [17–19]. SNP markers have both a higher genome density and a lower mutation rate than SSR markers and they are also more easily amenable to high-throughput genotyping in multiplex or microarray format [20, 21]. The mutational processes underlying SNP variation is well-understood whereas the mutational processes of other types of markers, such as SSRs and AFLPs, are poorly understood and this sometimes hampers analyses using such markers. The vast majority of SNPs are bi-allelic and the information content per marker is therefore much lower than in SSR markers. This, however, is more than compensated for by the fact that they are more widely distributed across the genome in most organisms [21]. SNP markers are therefore rapidly becoming the markers of choice for most association mapping studies in both model and non-model plant species.

One issue that has been receiving an increasing interest is how the selection of SNPs to include in an association study can potentially bias the results. Such ascertainment bias is usually attributed to the process of identifying and selecting SNPs for further use in an association study. For instance, SNP discovery panels are often small, suggesting that low-frequency mutations are more likely to go undetected. This will bias the frequency spectrum of the identified mutations compared with what would be obtained from the full sample, with relatively more SNPs occurring at intermediate frequencies [22]. The ascertainment bias introduced in the SNP selection process have important consequences for any inferences that are drawn from the data; for association mapping the most detrimental effect is an over-sampling of mutations at intermediate frequencies which results in lower levels of linkage disequilibrium (LD) than if SNPs were selected completely at random. The effect of ascertainment bias on the power of association studies is more complex, and largely depends on whether low or intermediate frequency are assumed to have a larger effect on the trait of interest [22, 23]. Another important problem to be aware of in association mapping is the genotyping error rate. While state-of-the-art SNP scoring methods are usually quite robust, the rate of genotyping errors can vary a lot between different SNPs even when scored on a single chip. This is important to remember since even low error rates (around 3% or less) can have dramatic consequences for the accuracy of estimates of LD [24] and hence also for association mapping.

CANDIDATE GENES VERSUS WHOLE GENOME SCANS

Another issue that a prospective scientist will face when embarking on an association study is whether to base the study on candidate genes or whether to apply association mapping to the whole genome of the organism of interest. The absolutely most important aspect when deciding between a candidate gene approach and a whole-genome study is the extent of LD in the organism of interest, because the extent of LD determines not only the mapping resolution that can be achieved, but also the numbers of markers that are needed for an adequate coverage of the genome in a genome-wide study [25]. When considering the extent of LD one should preferably also account for variation of recombination rates across the genome, although this may be hard to implement in organisms where regional variation in LD is poorly documented. In species where LD extends over long physical distances, relatively few markers are needed to ensure adequate genome coverage; for example the extensive LD seen in species like *Arabidopsis thaliana* or in inbred lines of barley, where LD can extend for tens or even hundreds of kilo base pairs, allow for genome-wide association mapping with a relatively low number of evenly spaced SNPs markers ([26, 27], see also Table 1). However, in many predominantly or obligately outcrossing organisms, such as maize [28] and many forest trees [6, 29], LD only extends a few hundred base pairs at the most and adequate genome-wide coverage would require several million SNPs.

The alternative approach to take when genome-wide association mapping is precluded, is to perform a candidate gene-based association study. A candidate-gene association mapping study is more hypothesis-driven than a genome-wide study, since association mapping is restricted to relevant candidates genes thought to be involved in controlling the trait of interest [6]. The selection of candidates is not straightforward, but choices can be based on relevant information obtained from, for instance, genetic, biochemical, or physiology studies in both model and non-model plant species [6]. Candidate-gene selection is usually quite straightforward when restricted to well characterized developmental pathways, like the flowering pathways in *Arabidopsis* and other plants [30], or to traits with a well-understood biochemical basis, such as the starch-synthesis pathway in maize [31]. Candidate gene studies are less

Table 1: A sample of recent association mapping studies in both crop and wild plant species.

Plant species	Populations	Sample size	Background markers	Traits	Method	Associations found	References
Arabidopsis	Accessions	95		L + R	Genome-wide	4	Aranzana *et al*. [26]
	Accessions	96		L	Candidate gene	3	Ehrenreich *et al*. [30]
Barley	Germplasm accessions	220	EST-SSR: 25	L	Candidate gene	1	Stracke *et al*. [67]
Cotton	Germplasm accessions	335	SSR: 202	M	Genome-wide	20/trait	Abdurakhmonov *et al*. [68]
Douglas fir	Diverse families	700	SNP: 384	L	Candidate gene	30	Eckart *et al*. [30]
European aspen	Natural population	116	SNP: 42 SSR: 25	A	Candidate gene	2	Ingvarsson *et al*. [51]
Loblolly pine	Diverse clones	961	SNP: 46	A	Candidate gene	4	González-Martínez *et al*. [71]
	Lines	435	SNP: 58	M	Candidate gene	4	González-Martínez *et al*. [70]
Maize	Elite inbred lines	553	SNP: 8590	Y	Genome-wide	1	Beló *et al*. [72]
	Inbred lines	282	SSR: 47	Y	Candidate gene	4	Harjes *et al*. [73]
Pearl millet	Inbred lines/accessions	90/598	SSR: 27/25 AFLP: 306	M + L + Y	Candidate gene	3	Saïdou *et al*. [74]
Perennial ryegrass	Germplasm accessions	26	AFLP: 589	L	Genome-wide	3	Skøt *et al*. [75]
Potato	Diverse cultivars	221	AFLP: 250	M + R + Y	Candidate gene	68	D'hoop *et al*. [76]
	Diverse cultivars	123	NBS: 49	R	Candidate gene	2	Malosetti *et al*. [77]
Rice	Diverse cultivars	103	SSR: 123	M + Y	Genome-wide	25	Agrama *et al*. [78]
Soybean	Breeding lines	139/115	SSR: 84	M	Candidate gene	3	Wang *et al*. [79]
	Germplasm accessions	96	SSR: 150	Y	Genome-wide	11	Jun *et al*. [80]
Sugar beet	Inbred lines	111	SSR: 26	M	Candidate gene	4	Stich *et al*. [81]
	Elite clones	768	SSR: 49 RFLP: 9	M + Y	Genome-wide	44	Stich *et al*. [82]
Wheat	Germplasm accessions	108	SSR: 85 EST-SSR: 40	M + Y	Genome-wide	14	Yao *et al*. [83]

M, morphology; L, life history; R, resistance; Y, yield; A, adaptive.

demanding in terms of the number of markers that are required and many candidate gene association studies have successfully been completed using tens to hundreds of markers in mapping populations consisting of a few hundred individuals (Table 1). However, it is important to remember that a candidate gene approach is limited by the choice of candidate genes that are identified and hence always runs the risk of missing out on identifying causal mutations that are located in non-identified candidate genes. In addition, candidate genes are often initially discovered from loss-of-function mutations in inbred lab strains and it is not clear how well such mutations describe the variation that actually underlie quantitative trait variation in natural populations.

PHENOTYPING

As the costs of genotyping is rapidly declining, a greater fraction of the budget of any association mapping project will be spent on phenotyping. In fact, while the importance of accurate identification and scoring of genotypes have received quite a deal of attention ([24, 32, 22]; see also above), the effects of phenotyping has yet to be evaluated in any greater detail. It has been shown, however, that increasing the number of individuals phenotyped is far more efficient than increasing the number of SNPs for increasing the power in association studies [33]. Also, several new experimental designs are actively being developed that combine the best aspects of traditional QTL mapping and association mapping (e.g. nested association mapping [34]).

A typical association mapping study usually involves a diverse set of accessions (see Table 1 for some recent examples), and phenotypic scoring with adequate accuracy can be both costly and time-consuming. Replication of individual accessions within a site is usually needed to increase precision in phenotypic measurements, by eliminating environmentally induced noise and measurement errors. Data on replicates of each accession can then be combined to produce an estimate of the 'mean' phenotype of the accession which is less influenced by environment or measurement errors. One example of such an approach is the estimation of breeding values which is common practice in quantitative genetics and breeding [1]. These breeding values are used as dependent traits in an association analysis in an attempt to dissect the genetic basis of the trait in question [35].

An additional benefit of replication can be achieved if the entire association mapping collection is replicated across multiple environments. Such a design can provide important information on the robustness of positive associations across environments and on the importance of genotype by environment interactions in shaping allelic contributions to the trait of interest [1].

CONTROLLING FOR POPULATION STRUCTURE

One of the main hurdles for using association mapping to dissect the genetic architecture of complex traits in plants is the risk of incurring false positives due to population structure [36, 37]. The problem of population structure arises because any phenotypic trait that is also correlated with the underlying population structure at neutral loci will show an inflated number of positive associations. The problem of population structure is well known and many methods have, not surprisingly, been developed to deal with this problem. Several of these methods are also implemented in software packages that are freely available (Table 2).

One of the first methods proposed was the method of 'genomic control' (GC) developed by Devlin and Roeder [38]. The rationale for GC is to estimate association using a large number of putative neutral markers or markers not thought to be involved in controlling the trait of interest. The distribution of the test statistic of interest is then calculated from these associations and a critical value corresponding to the desired Type I error rate is chosen from this distribution. While GC is straightforward to perform computationally, it requires a large number of control loci to accurately capture the extent of variation in population structure across the genome of an organism. Furthermore, in some situations it is possible for GC to 'overcorrect' for population structure effects resulting in a loss of power to detect true associations [39, 40].

Another method that is commonly used to control for population structure is structured associations (SA) [41]. The idea of SA builds on the method of Pritchard *et al*. [36] who infer details of population structure and the ancestry of sampled individuals using a set of unlinked genetic markers. This information is then used to identify populations within which mating is random. Markers are then tested for associations within these sub-populations identified by the genetic markers [41].

The most recent, and most promising approach, for correcting the spurious effects of population structure is the mixed-model approach outlined by Yu *et al*. [42; see also 43, 44]. Mixed-model methods use information on both population structure and more cryptic relatedness among members of an association study to correct for the spurious effects of populations structure and relatedness. These two types of population structure are incorporated into a matrix of population effects (Q) and a matrix describing the relative kinship of individuals in a sample (K) and a model is then fitted using the mixed-model framework developed in, for instance, animal breeding [45]. The Q matrix consists of one or more vectors describing the underlying

Table 2: Non-commercial computer packages for performing population structure or kinship estimation and for performing association mapping

Software package	Website	Citation
General purpose		
R	http://www.r-project.org/	R Development Core Team [66]
Population structure and relatedness		
STRUCTURE	http://pritch.bsd.uchicago.edu/structure.html	Pritchard *et al*. [36]
EIGENSOFT	http://genepath.med.harvard.edu/~reich/Software.htm	Price *et al*. [47]
BAPS	http://web.abo.fi/fak/mnf//mate/jc/software/baps.html	Corander *et al* [62]
ADMIXTURE	http://www.genetics.ucla.edu/software/admixture/	Alexander *et al*. [63]
SPAGeDi	http://www.ulb.ac.be/sciences/ecoevol/spagedi.html	Hardy and Vekemans [64]
InStruct	http://cbsuapps.tc.cornell.edu/InStruct.aspx	Gao *et al*. [59]
Association analysis		
EMMA	http://mouse.cs.ucla.edu/emma/	Kang *et al*. [43]
STRAT	http://pritch.bsd.uchicago.edu/software/STRAT.html	Pritchard *et al* [41]
TASSEL	http://www.maizegenetics.net	Bradbury *et al*. [65]

population structure and this matrix can be estimated in several ways. For example, one common approach is to use the method implemented in STRUCTURE [36] or by using principle component analysis (PCA) of the complete genotype data [46]. Using PCA to estimate population structure is especially appealing since it is far less computationally demanding than analyses based on STRUCTURE [46, 47]. A similar approach to PCA is to use non-metric multidimensional scaling (nMDS, Zhu and Yu [61]) which have been shown to reduce the false positive rate compared to other methods in structured populations.

The kinship matrix (K), on the other hand, can be estimated from pedigree data or, for non-model species where pedigree information is usually lacking, using relative kinship coefficients estimated using genetic marker data (e.g. [35, 48, 49]). The strength of the mixed-model approach is that it handles and performs well under many types of population structure [42, 43]. For instance, in a genome-wide association study in *Arabidopsis thaliana*, the mixed-model provided the most accurate control of the false-positive rate among the methods tested, despite a very complex sub-structuring of the association population [37].

The original intent of the Q and K matrices is to capture different types of population structure [42] and several studies have found that including either the Q or the K matrix alone is not sufficient to control for all aspects of the underlying population structure of the data. However, the relative utility of the two matrices depends on the actual pattern of the underlying population structure. Both STRUCTURE [36] and the PCA-based analyses [46] have problems identifying low levels of population structure when a low to moderate number of markers are used. Patterson *et al*. [46] even defined a minimum study design that is needed to effectively evaluate population structure and showed that for a given design there exists a minimum level of population structure that can be detected. For example, a STRUCTURE-based analysis failed to identify any obvious signs of population structure in European aspen (*Populus tremula*), despite evidence for significant population structure and isolation-by-distance based on population groupings chosen *a priori* [50]. However, the same set of markers, when used to estimate the K for the sampled trees, provided a reasonable control of the underlying weak, but nevertheless significant, isolation by distance [51].

REPLICATION AND VALIDATION

Association mapping techniques are increasingly being used to dissect quantitative trait variation in both economically and ecologically important traits (for a collection of recent studies, see Table 1). However, as the number of studies documenting alleles showing significant associations with quantitative trait variation, there is an increasing need to replicate findings and to validate estimates of allelic effects. These issues are being highlighted in the human genetics community, where guidelines for conducting both initial and replication studies are being devised [52]. Replication of genotype-phenotype association are crucial for separating true from false positives and to provide less biased estimates of allelic effect sizes. However, failure to replicate a previously documented association can occur because of a large number of issues, both in the initial and the replication study, including factors like difficulties in replicating the environment, small sample size, poor study design or lack of rigorous phenotype scoring [23]. The literature on association mapping in plants does, however, include a few cases where associations have been replicated in independent mapping experiments. For example, Thornsberry *et al*. [53] found that mutations in the gene *Dwarf 8* affect the quantitative variation of flowering time and plant height in maize (*Zea mays*). This association has subsequently been verified in a larger maize association mapping population containing a different set of maize inbred lines [54]. Finally, it is worth pointing out that verification of genotype–phenotype associations does not necessarily have to come from replicate association studies, but can include validation of biological function through transgenic experiments and other molecular biology techniques [55].

Another concern is that allelic effects of previously documented associations usually decline in replication studies. This phenomenon is known as the 'Beavis effect' [56] in the QTL mapping literature and occurs because significant associations are reported only when test statistics exceed a predetermined critical threshold. The estimated effects of detected associations are therefore sampled from a truncated distribution, and the weaker the initial effect the more serious this overestimation is [57]. The Beavis effect has also been shown to occur in association mapping studies. For instance, Ingvarsson *et al*. [51] showed that naïve estimates of the effects of mutations in the photoreceptor gene *PHYB2* were

overestimated by ~2- to 3-fold. The Beavis effect is known to be weaker when the mapping population used in the experiment is larger [56], hence careful consideration of the power of the prospective association study should be taken early on in the experiment, so that things like the Beavis effect can be minimized or eliminated.

CONCLUSIONS

Earlier, the largest hurdle to clear in the search for the molecular basis of complex phenotypes has been the generation of genetic markers and the scoring of genotypes. However, with the rapidly dropping costs of modern sequencing and genotyping technologies generation of genotype data is no longer the limiting factor for most studies. This has resulted in a need for new refined statistical methods for association analysis that cover entire genomes and the greatest costs are utilized towards rigorous phenotyping instead of generation of genotypic data. A large fraction of the cost is associated with the establishment of association mapping collections, housed in, for instance establishing common gardens to minimize environmental influence and possible epigenetic effects. The current status of association mapping in plants largely draws from two fields, and those have proven to be valuable for finding associations between molecular markers and traits [36]. First, the human genetics research community is actively developing sophisticated statistical methods for handling genome-wide association studies with massive amounts of data. Second, animal breeding methods are being developed that partition phenotypic variation into genetic variance components using detailed information on relatedness between individuals. These tools together, combined into robust mixed model approaches [36, 43, 44], which account for different levels of relatedness and population stratification decrease the number of false positives which would otherwise be a problem with the rapid increase in the number of associations tested per study. However, individual alleles or QTLs identified in association studies usually explain only a few percent of the variation in traits studied and even when many loci associated to a trait are taken into account the proportion of variation explained is usually far below than prediction-based heritabilities of the traits, a phenomenon highlighted in the human-genetics community as the 'missing heritability problem' [23]. The problem of 'missing heritability' has several likely causes that are poorly accounted for in current association mapping studies, such as low-frequency alleles of large effect, allelic interactions (i.e. epistasis), copy number variation and possible epigenetic effects [23]. As more and more putative causative alleles are identified, it becomes increasingly necessary for methods that can deal with associations across gene networks of interacting genes and across developmental pathways (i.e. epistasis; [57, 58]). An additional question that should be addressed is about how the effects of individual alleles vary across different environments. Given the ubiquity of genotype environment for many traits in plants [1], to what degree QTL effects vary across environments has important implications, e.g. the utility of QTLs in breeding applications.

Key Points

- Association mapping has recently become the main method for dissecting the genetic basis of complex traits in plants.
- The interest in applying association mapping has been fueled by rapidly declining genotyping costs that allow for genotyping with a dense cover across the genome, facilitating mapping with high resolution.
- Association mapping also is relatively fast, compared with traditional QTL mapping, since mapping relies on a diverse set of accessions and no segregating mapping-populations have to be created.
- A great deal of effort is currently being devoted to solving problems associated with separating false from true positive associations, including methods for controlling the underlying effects of population structure and for replicating and validating associations.

FUNDING

Swedish Research Council and the Research School in Forest Genetics and Breeding (to P.K.I).

References

1. Lynch M, Walsh B. *Genetics and Analysis of Quantitative Traits*. Sunderland, MA: Sinauer Associates, 1998.
2. Mauricio R. Mapping quantitative trait loci in plants: uses and caveats for evolutionary biology. *Nat Rev Genet* 2001;**2**: 370–81.
3. Doerge RW. Mapping and analysis of quantitative trait loci in experimental populations. *Nat Rev Genet* 2002;**3**:43–52.
4. Holland JB. Genetic architecture of complex traits in plants. *Curr Opin Plant Biol* 2007;**10**:156–61.
5. Price AH. Believe it or not, QTLs are accurate!. *Trends Plant Sci* 2006;**11**:213–6.
6. Neale DB, Savolainen O. Association genetics of complex traits in conifers. *Trends Plant Sci* 2004;**9**:325–30.

7. Nordborg M, Tavaré S. Linkage disequilibrium: what history has to tell us. *Trends Genet* 2002;**18**:83–90.
8. Yu J, Buckler ES. Genetic association mapping and genome organization of maize. *Curr Opin Biotechnol* 2006; **17**:155–60.
9. Vos P, Hogers R, Bleeker M, *et al*. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res* 1995;**23**:4407–14.
10. Ritland K. Multilocus estimation of pairwise relatedness with dominant markers. *Mol Ecol* 2005;**14**:3157–65.
11. Liu BH. *Statistical Genomics: Linkage, Mapping and QTL Analysis*. Boca Raton, FL: CRC Press, 1998.
12. Goldstein DB, Schlötterer C. *Microsatellites: Evolution and Applications*. Oxford, UK: Oxford University Press, 1999.
13. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. *Trends Biotechnol* 2009;**27**: 522–30.
14. Simon SA, Zhai J, Sekhar Nanderty R, *et al*. Short-read sequencing technologies for transcriptional analyses. *Annu Rev Plant Biol* 2009;**60**:305–33.
15. Nordborg M, Wigel D. Next-generation genetics in plants. *Nature* 2009;**456**:720–3.
16. Pushkarev D, Neff NF, Quake SR. Single-molecule sequencing of an individual human genome. *Nat Biotechnol* 2009;**27**:847–50.
17. Ossowski S, Schneeberger K, Clark RM, *et al*. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. *Genome Res* 2008;**18**:2024–33.
18. Ganal MW, Altmann T, Röder MS. SNP identification in crop plants. *Curr Opin Plant Biol* 2009;**12**:211–17.
19. Imelfort M, Duran C, Batley J, Edwards D. Discovering genetic polymorphisms in next-generation sequencing data. *Plant Biotechnol J* 2009;**7**:312–17.
20. Syvänen AC. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat Rev Genet* 2001;**2**:930–42.
21. Syvänen AC. Toward genome-wide SNP genotyping. *Nat Genet* 2005;**37**:S5–S10.
22. Clark AG, Hubisz MJ, Bustamante CD, *et al*. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005;**15**:1496–502.
23. Manolio TA, Collins FS, Cox NJ, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
24. Akey JM, Zhang K, Xiong M, *et al*. The effect that genotyping errors have on the robustness of common linkage-disequilibrium measures. *Am J Human Genet* 2001;**68**: 1447–56.
25. Whitt SR, Buckler EB. Using natural allelic diversity to evaluate gene function. *Methods Mol Biol* 2003;**236**:123–40.
26. Aranzana MJ, Kim S, Zhao K, *et al*. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 2005;**1**:e60.
27. Rostoks N, Ramsay L, Mackenzie L, *et al*. Recent history of artificial outcrossing facilitates whole-genome association mapping in elite inbred crop varieties. *Proc Natl Acad Sci USA* 2006;**103**:18656–61.
28. Remington DL, Thornsberry JM, Matsuoka , *et al*. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 2001;**98**: 11479–84.
29. Neale DB, Ingvarsson PK. Population, quantitative and comparative genomics of adaptation in forest trees. *Curr Opin Plant Biol* 2008;**11**:149–55.
30. Ehrenreich IM, Hanzawa Y, Chou L, *et al*. Candidate gene association mapping of *Arabidopsis* flowering time. *Genetics* 2009;**183**:325–35.
31. Wilson LM, Whitt SR, Rocheford TR, *et al*. Dissection of maize kernel composition and starch production by candidate gene association. *Plant Cell* 2004;**16**:2719–33.
32. Sobel E, Papp JC, Lange K. Detection and integration of genotyping errors in statistical genetics. *Am J Human Genet* 2002;**70**:496–508.
33. Long AD, Langley CH. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999;**9**:720–31.
34. Yu J, Holland JB, McMullen MD, Buckler ES. Genetic design and statistical power of nested association mapping in maize. *Genetics* 2008;**178**:539–51.
35. Stich B, Möhring J, Piepho HP, *et al*. Comparison of mixed-model approaches for association mapping. *Genetics* 2008;**178**:1745–54.
36. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000;**155**:945–59.
37. Zhao K, Aranzana MJ, Kim S, *et al*. An *Arabidopsis* example of association mapping in structured samples. *PLoS Genetics* 2007;**3**:e4.
38. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;**55**:997–1004.
39. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genetic Epidemiol* 2001;**20**:4–16.
40. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004;**36**:512–17.
41. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Human Genet* 2000;**67**:170–81.
42. Yu J, Pressoir G, Briggs WH, *et al*. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 2006;**38**:203–8.
43. Kang HM, Zaitlen NA, Wade CM, *et al*. Efficient control of population structure in model organism association mapping. *Genetics* 2008;**178**:1709–23.
44. Stich B, Melchinger AE. Comparison of mixed-model approaches for association mapping in rapeseed, potato, sugar beet, maize and Arabidopsis. *BMC Genomics* 2009;**10**:94.
45. Mrode RA. *Linear Models for the Prediction of Animal Breeding Values*. Oxfordshire, UK: CABI Publishing, 2005.
46. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.
47. Price AL, Patterson NJ, Plenge RM, *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9.
48. Loiselle BA, Sork VL, Nason J, Graham C. Spatial genetic structure of a tropical understory shrub, *Psychotria officinalis* (Rubiaceae). *Am J Botany* 1995;**82**:1420–5.
49. Ritland K. Estimators for pairwise relatedness and individual inbreeding coefficients. *Genet Res* 1996;**67**:175–85.

50. Hall D, Luquez V, St. Onge KR, *et al*. Adaptive population differentiation in bud phenology across a latitudinal gradient in European aspen (*Populus tremula*, L., Salicaeae): a comparison of neutral markers, candidate genes and quantitative traits. *Evolution* 2007;**61**:2849–60.
51. Ingvarsson PK, Garcia MV, Luquez V, *et al*. Nucleotide polymorphism and phenotypic associations within and around the *phytochromeB2* locus in European aspen (*Populus tremula*, Salicaceae). *Genetics* 2008;**178**:2217–26.
52. Chanock SJ, Maniolo T, Boehnke M, *et al*. Replicating genotype-phenotype associations. *Nature* 2007;**447**:655–60.
53. Thornsberry JM, Goodman MM, Doebley J, *et al*. *Dwarf8* polymorphisms associate with variation in flowering time. *Nat Genet* 2001;**28**:286–9.
54. Camus-Kulandaivelu L, Veyrieras JB, Madur D, *et al*. Maize adaptation to temperate climate: relationship between population structure and polymorphism in the *Dwarf8* gene. *Genetics* 2006;**172**:2449–63.
55. Koornneef M, Alonso-Blanco C, Vreugdenhil D. Naturally occurring genetic variation in *Arabidopsis thaliana*. *Annu Rev Plant Biol* 2004;**55**:141–72.
56. Beavis WD. QTL analyses: power, precision, and accuracy. In: Paterson AH (ed). *Molecular Dissection of Complex Traits*. New York: CRC Press, 1998:145–62.
57. Rockman MV. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* 2008;**456**:738–44.
58. Xu S. Theoretical basis of the Beavis effect. *Genetics* 2003; **165**:2259–68.
59. Gao H, Williamson S, Bustamante CD. A Markov chain Monte Carlo approach for joint inference of population structure and inbreeding rates from multilocus genotype data. *Genetics*\ 2007;**176**:1635–51.
60. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Human Genet* 1999;**65**:220–8.
61. Zhu C, Yu J. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. *Genetics* 2009;**182**:875–88.
62. Corander J, Marttinen P, Sirén J, Tang J. Enhanced bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 2008;**9**:539–53.
63. Alexander DH, Noembrve J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* 2009;**19**:1655–64.
64. Hardy OJ, Vekemans X. SPAGeDi: a versatile computer program to analyse spatialgenetic structure at the individual or population levels. *Mol Ecol Notes* 2002;**2**:618–20.
65. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* 2007;**23**:2633–5.
66. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing 2009 Development Vienna Austria. ISBN 3-900051-07-0, URL http://www.R-project.org
67. Stracke S, Haseneyer G, Geiger J-B, Veyrieras HH, Sauer S, Graner A, H-P Piepho. Association mapping reveals gene action and interactions in the determination of flowering time in barley. *Theor Appl Genet* 2009;**118**: 259–27.
68. Abdurakhmonov IY, Saha S, Jenkins JN, Buriev ZT, Shermatov SE, Scheffler BE, Pepper AE, Yu JZ, Kohel RJ, Abdukarimov A. Linkage disequilibrium based association mapping of fiber quality traits in G. hirsutum L. variety germplasm. *Genetica* 2009;**136**:40–417.
69. Eckert AJ, Bower AD, Wegrzyn JL, Pande B, Jermstad KD, Krutovsky KV, St Clair JB, Neale DB. Association genetics of coastal douglas fir (Pseudotsuga menziesii var. menziesii, Pinaceae). I. Cold-hardiness related traits. *Genetics* 2009;**182**: 1289–302.
70. González-Martínez SC, Wheeler NC, Ersoz E, Nelson CD, Neale DB. Association genetics in Pinus taeda L. I. Wood property traits. *Genetics* 2007;**175**:399–409.
71. González-Martínez SC, Huber D, Ersoz E, Davis JM, Neale DB. Association genetics in Pinus taeda L. II. Carbon isotope discrimination. *Heredity* 2008;**101**:19–26.
72. Beló A, Zheng P, Luck S, Shen B, Meyer DJ, Li B, Tingey S, Rafalski A. Whole genome scan detects an allelic variant of fad2 associated with increased oleic acid levels in maize. *Molecular Genetics and Genomics* 2008;**279**:1–10.
73. Harjes CE, Rocheford TR, Bai L, Brutnell TP, Kandianis CB, Sowinski SG, Stapleton AE, Vallabhaneni R, Williams M, Wurtzel ET, Yan JB, Buckler ES. Natural genetic variation in lycopene epsilon cyclase tapped for maize biofortification. *Science* 2008;**319**:330–3.
74. Saïdou AA, Mariac C, Luong V, Pham JL, Bezançon G, Vigouroux Y. Association studies identify natural variation at PHYC linked to flowering time and morphological variation in pearl millet. *Genetics* 2009;**182**:899–910.
75. Skøt L, Humphreys MO, Armstead I, Heywood S, Skøt KP, Sanderson R, Thomas ID, Chorlton KH, Sackville Hamilton NR. An association mapping approach to identify flowering time genes in natural populations of *Lolium perenne* (L). *Mol Breeding* 2005;**15**:233–45.
76. D'hoop BB, Paulo MJ, Mank RA, van Eck HJ, van Eeuwijk FA. Association mapping of quality traits in potato (*Solarium tuberosum* L). *Euphytica* 2008;**161**:47–60.
77. Malosetti M, van der Linden CG, Vosman B, van Eeuwijk FA. A mixed-model approach to association mapping using pedigree information with an illustration of resistance to *Phytophthora infestans* in potato. *Genetics* 2007;**175**:879–89.
78. Agrama HA, Eizenga GC, Yan W. Association mapping of yield and its components in rice cultivars. *Mol Breeding* 2007; **19**:341–56.
79. Wang J, McClean PE, Lee R, Goos J, Helms T. Association mapping of iron deficiency chlorosis loci in soybean (Glycine max L. Merr.) advanced breeding lines. *Theor Appl Genet* 2008;**116**:777–87.
80. Jun T-H, Van K, Kim MY, Lee HS, Walker DR. Association analysis using SSR markers to find QTL for seed protein content in soybean. *Euphytica* 2008;**162**: 179–91.
81. Stich B, Piepho H-P, Schulz B, Melchinger AE. Multi-trait association mapping in sugar beet (Beta vulgaris L). *Theor Appl Genet* 2008;**117**:947–54.
82. Stich B, Melchinger AE, Mbhring MJ, Schechert A, H-P Piepho. Association mapping in multiple segregating populations of sugar beet (Beta vulgaris L.). *Theor Appl Genet* 2008;**117**:1167–79.
83. Yao J, Wang L, Liu L, Zhao C, Zheng Y. Association mapping of agronomic traits on chromosome 2A of wheat. *Genetica* 2009;**137**:67–75.