African Journal of Biotechnology Vol. 6 (6), pp. 650-657, 19 March 2007
Available online at http://www.academicjournals.org/AJB
ISSN 1684–5315 © 2007 Academic Journals

Review

Localizing genes using linkage disequilibrium in plants: integrating lessons from the medical genetics

Joel W. Ochieng[1*], Anne W. T. Muigai[2] and George N. Ude[3]

[1]Section of Genetics, College of Agriculture and Veterinary Sciences, University of Nairobi, P.O. Box 29053 Nairobi, 00625 Kenya.
[2]Institute for Biotechnology Research, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62000 Nairobi, 00200 Kenya.
[3]Department of Natural Sciences, Bowie State University, 14000 Jericho Park Road, Bowie, MD 20715, USA.

Accepted 9 February, 2007

Finding genes controlling quantitative traits will aid molecular breeding for crops and livestock with superior yields, growth rates, and evolutionary potential. Such genes can be located using the candidate gene approach, genome wide scans, or by within family mapping. Linkage disequilibrium (LD) or association mapping, is a candidate gene approach that relies on detecting a statistical association between the desired quantitative trait and a molecular marker allele. This approach is emerging as a leading tool for precise estimation of QTL positions, because it offers several advantages over family-based mapping: LD mapping detects associations with greater resolution, the associations detected are relevant population wide, and in plants, the use of natural populations would circumvent the need to raise large controlled crosses. However, LD approach is facing obstacles, with well over 60% of studies reporting associations in the medical genetics disapproved in subsequent tests. A large proportion of these false associations (or lack of it) result from population stratification, while the rest may be caused by other demographic and evolutionary processes that create a statistical association between a marker allele and the trait, such as bottlenecks, natural selection, hybridization and genetic drift. The problem is expected to escalate in plants, owing to the complex population structures. Regardless of the many recent methods that purport to take into account population stratification during association tests, we discuss the reasons why in plants, a priori knowledge of population structures is essential in any robust association analysis.

Key words: Association mapping, linkage disequilibrium, population structure, nonreplication, quantitative trait loci.

INTRODUCTION

Methods for identifying genes affecting traits

A quantitative trait locus (QTL) is a region in the genome that contains one or several genes affecting a quantitative trait. QTLs, like other genes, can be mapped, and the effect of an individual QTL can be estimated. The objective of genetic mapping is to identify simply inherited markers in close proximity to genetic factors affecting quantitative traits, that is, QTL. This localization relies on processes that create a statistical association between marker and QTL alleles and processes that selectively reduce that association as a function of the marker distance from the QTL. Two main family based methods of localizing genes underlying complex traits are the QTL mapping and the transmission/disequilibrium test (TDT). A third method that uses natural populations is association mapping, also referred to as linkage disequilibrium (LD) mapping. A short overview of these methods and their pit-falls precedes discussions on challenges and headways for LD mapping in plants.

Quantitative trait loci (QTL) mapping

When using crosses between inbred parents to map QTL, we create in the F1 hybrid complete association between all marker and QTL alleles that derive from the same parent. Recombination in the meioses that lead to doubled haploid, F2, or recombinant inbred lines reduces

*Corresponding author. E-mail: jochieng@uonbi.ac.ke. Tel: +61 2 6620 3961, Fax: +61 2 6622 2080.

the association between a given QTL and markers distant from it. Unfortunately, arriving at these generations of progeny requires relatively few meioses such that even markers that are far from the QTL (beyond 10 cM) remain strongly associated with it (Janoo et al., 1999; Farnir et al., 2000; Reich et al., 2001). Such long-distance associations hamper precise localization of the QTL. One approach for fine mapping is to expand the genetic map, for example through the use of advanced intercross lines, such as F6 or higher generational lines derived by continual generations of outcrossing the F2 (Darvasi and Soller, 1995). In such lines, sufficient meioses have occurred to reduce disequilibrium between moderately linked markers. The central problem with any of the family based approaches for fine mapping is the limited number of meioses that have occurred and the cost of propagating lines to allow for a sufficient number of meioses. Therefore, while QTL mapping permits the decomposition of complex traits into their Mendelian components, it does not allow the actual genes underlying trait variation to be identified since the confidence interval of a QTL is often very large.

Transmission/disequilibrium test (TDT)

The problem of false associations due to population admixture in QTL mapping led to the development of The Transmission/disequilibrium test (TDT) (Spielman et al., 1993), to identify loci contributing to disease susceptibility in humans in the presence of population structure. For outbred species, the test employs family trios consisting of both parents and a progeny that is affected by disease (or, in general, that belongs to one category of a dichotomous trait). One of the parents must be heterozygous and carry one copy of the focal marker allele putatively linked to the disease susceptibility allele. In brief, the test consists of determining the frequency of transmission of the focal allele to affected progeny. A chi-square or binomial test can determine whether that frequency deviates from the expectation of 0.5. Two conditions are necessary for a significant deviation: the marker allele must be both in gametic phase disequilibrium (GPD) with and also linked to a disease susceptibility allele. In the TDT, both case and control marker alleles are in effect within the same heterozygote parent. Random Mendelian segregation therefore ensures that the distribution of the TDT statistic under the null hypothesis is unaffected by population structure or selection within the pedigree (Spielman and Ewens, 1996). Despite these efforts, the TDT is not a suitable test for population-wide association as the test may detect association that exists solely in the pedigree from which those families derive but not in the general population (Martin et al., 2000). Further, the critical interest in using association mapping is in finding tightly linked markers. A TDT based on multiple related families may detect association based on fairly distant marker-QTL pairs simply because recombination within the confines of the single pedigree evaluated will fail to reduce their association.

Allelic association and linkage

A statistical association between a neutral marker allele and the phenotype occurs when marker alleles are in gametic phase disequilibrium (GPD) with alleles at a QTL. GPD is sometimes used synonymously with Linkage disequilibrium (LD), but markers can still be in GPD without being linked. In principle, associated alleles must occur in gametes. Therefore the term GPD is deliberately used to avoid mention of linkage, a term often used loosely in medical genetics, e.g., Linkage disequilibrium (LD), the non-random association between alleles of linked markers, reflects the size of chromosomal segments remaining intact in a population (Mohlke et al., 2001). During the reproductive phase, recombination occurs between the corresponding chromosomes from the two parents in the cross. This leads to reshuffling of the genes from each parent so that the chromosomes in the offspring will consist of mixtures, with derivatives of pieces from either parent. The closer two loci are on the chromosome; the less likely it is that a crossover point will occur between them. Such loci will therefore be more likely inherited together, because they are physically linked. Because linked markers will most likely segregate together and hence show no independence between them, this non-independent segregation is termed Linkage disequilibrium. LD is therefore intended to measure the closeness between genetic markers and a QTL for a particular trait, and may be used to identify markers in close proximity to the gene(s) responsible for the trait. Although the desired cause of LD in association studies is physical linkage, several other biological and historical factors (e.g., population stratification, natural selection, bottlenecks) can cause the non-independent segregation of alleles. This is why the term GPD is used since it avoids reference to linkage. Two alleles at distinct loci are in positive GPD if they occur together more often than predicted on the basis of their individual frequencies.

Association mapping

In the simplest definition, association mapping is the utility of linkage disequilibrium, also known as gametic phase disequilibrium, in natural populations to identify markers with significant allele frequency differences between individuals with the trait of interest and a set of unrelated control individuals. A statistical association between genotypes at a marker locus and the trait of interest is considered to be evidence of close physical

linkage between the marker and the QTL controlling that trait (Pritchard et al., 2000). While classical gene mapping approaches are useful in genome-wide scan for loci controlling QTLs, association mapping is emerging as a leading tool for precise estimation of QTL positions. For example, this method has been used to identify genes for complex traits in medical genetics (Lander and Schork, 1994; Risch, 2000), and its application is gradually moving to other fields such as plant genetics. Since association mapping uses natural populations, many generations (and therefore meioses) have elapsed, thus recombination will have removed association between a QTL and any marker not tightly linked to it. Association mapping thus allows for much finer mapping than standard biparental cross approaches.

The key advantages of this method over family based methods described above are: firstly, genetic associations detected are relevant for whole population as opposed to family based methods in which the association detected may be valid only for that family (Martin et al., 2000); secondly, improved precision of locating the QTL due to dissipating role of recombination, such that only tightly linked markers remain in disequilibrium, and thirdly, the low costs associated with finding natural populations as opposed to generating or sourcing pedigrees (Teng and Risch, 1999). Despite all these, association mapping is fraught with the possibility of false associations due to demographic and biological processes. These obstacles and opportunities are discussed.

Non-reproducibility of reported associations: Over the past decade, numerous research projects have reported associations between certain traits and regions of the genome (e.g., Gm3, 5, 13, 14 and Type I and Type II diabetes mellitus- Knowler et al., 1988; *Dwarf 8* polymorphism and flowering time in Maize- Thornsberry et al., 2001). Unfortunately, many of the reported associations have not been replicated in independent research. The nonreplication of these earlier findings is a concern and has caused some researchers to question the utility of association methodology in genetic studies (Gambaro et al., 2000; Holtzman, 2001; Strohman, 2002). Altsuler and others in a Nature Genetics editorial (Altshuler et al., 2000) echoed the prevailing discontent that “Genome-wide linkage scans have yielded few significant findings, and failure to reproduce published linkage results is endemic”. Some researchers have renewed efforts trying to replicate such reports (Hirschhorn et al., 2002 and references therein; Lohmueller et al., 2003). This may still be rewarding in terms of further localization of such genes in false-negative results. However, false-positive associations with candidate genes essentially become dead ends. If a high proportion of such associations turn out to be false positives, the wasted effort could be considerable.

Lohmueller et al. (2003) performed meta-analysis for several genetic association studies over a wide range of phenotypes. They reported meta-analysis for nine different markers that had been examined for association with Type II diabetes in 50 published association studies. Of the 41 studies designed to replicate the initially observed associations, only 10 produced results in agreement with the original findings. Based on their results, they concluded that out of the nine potential markers associated with Type II diabetes, only three of them have sufficient replication to support the claim of an association. Thus, a great concern has gripped the medical genetics community from the fact that few reported associations between markers and phenotypes are consistently and convincingly replicated. To lend support to this concern, Hirschhorn et al. (2002) reviewed over 600 published positive association studies and discovered that only 166 were investigated three or more times. Of the 166 studies, they concluded that only six (6) associations had been consistently replicated.

Causes for irreproducibility of reported associations: What makes so many reported associations irreproducible is a question whose answers are critical to the future of LD research. Some of the answers have been suggested in the literature, while other plausible possibilities are proposed in this review. Although the desired cause of LD in association studies is physical linkage, several biological and historical factors can cause statistical associations. The most common problems with LD in association studies are the spurious associations or false positives caused by population structure or admixture, which can lead to highly significant associations between a marker and a phenotype, even when the marker is not physically linked to any causative loci (Pritchard and Przeworski, 2001). Other factors include founder effect and population bottlenecks, in which LD in a young or recovering population has not had time to decay through recombination, self-fertilization and inbreeding, which reduce heterozygosity, hence low recombination that would dissipate LD, and natural selection acting with sufficient intensity favouring certain genotypes (Przeworski, 2002). Strong selection of a particular allele limits genetic diversity around a locus, resulting in short term increases in LD around the selected gene. It is also believed that gene-gene or gene-environment can differ between populations, for instance, if a particular genetic variant were only manifest in populations with a particular environmental background (Hirschhorn et al., 2002). Likewise, associations can be real but nonetheless not reproducible if the underlying genetic effect is weak (Hirschhorn et al., 2002). A less frequent cause can be chromosomal inversions affecting genes in or near the inversions. Apart from these acknowledged causes of false LD and hence non-replication of associations in subsequent studies, available literature to date has ignored a number of tenable possibilities. For one, we think that gene duplication (paralogs, pseudogenes) can

be responsible for some of the failures to reproduce linkage results. Small inevitable variations in PCR conditions can sometimes influence which of the paralogs are amplified. In such cases, researchers trying to replicate previous studies involving sequence polymorphisms could be amplifying alternate haplotype(s), which may show no association with the trait variant because the paralogs may differ at the associated nucleotide polymorphism. Another plausible cause is in the technicality of involving case-control choices, because pairs of cases are likely to be more closely related than are pairs of controls or case-control pairs if in fact the trait does have a common genetic basis. Hence it is clear that the possible causes of associations that cannot be replicated in subsequent and independent research are many. In this review, we focus more on population stratification, as it is currently the principal problem in LD mapping as well as the most likely nuisance in plant biotechnology.

POPULATION STRUCTURE

Population structure is the presence of mating subgroups within the population, usually resulting in differences in alleles and hence allele frequencies among these groups. Population structure is caused when there is non-random (assortative) mating in a population. In evolutionary genetics, each mating sub-group in a population is referred to as a deme. Structure can result from among others, geographic isolating factors (isolation-by-distance phenomenon), selection affecting mate choices and resource allocation. When mating is non-random, gene flow is more within each deme than it is among demes. This can make the frequency of certain alleles to be higher in one deme than it is in another. In the context of association mapping, population stratification occurs when both of the following conditions are met: First, allele frequencies under investigation must vary among subpopulations. Second, the mean trait value must vary among subpopulations. One problem with this is that it is not usually possible to tell whether mating is random or not within a population, and the empirical status may be incongruent with intuition. Furthermore, mating in some life forms such as plants can be much more complex as it often depends on secondary factors such as pollinators and in some cases under both environmental and genetic control (such as genes for flowering time).

How structure causes false associations

With the existence of cryptic demes within a supposedly random mating population, it is inevitable that some deme(s) will be overrepresented in a sample than others. When the trait frequency varies across demes, it may increase the probability that affected individuals will be sampled from particular demes. In such a case, any marker allele that will be in higher frequency in the over-represented deme will then show association with the trait (Pritchard and Rosenberg, 1999). Allele frequencies of many genes have been shown to vary substantially across populations (Perez-Lezaun et al., 1997). Moreover, the extent of variation is directly related to the genetic distance between populations, and the degree of variation is, of course, dependent on the allele being examined. What is at issue is how much variation in allele frequencies and trait levels there is between subdivisions of the major population groups. Even in the absence of confounding bias, population stratification can distort significance levels through cryptic relatedness i.e., unobserved ancestral relationships between individual cases and controls that are naively treated as independent in the standard chi-square test. In particular, pairs of cases are likely to be more closely related than are pairs of controls or case-control pairs if in fact the trait does have a common genetic basis. This will have the effect of inflating the “effective” sample size, thereby increasing the false-positive rate, even in the absence of any confounding bias. However, this effect is likely to be more important in inbred population isolates than in large outbred populations. Nevertheless, even if the magnitude of the bias attributable to either confounding or cryptic relatedness is small, the effect on significance levels is related to sample size, and hence very large case-control studies involving thousands of subjects could have considerably inflated false-positive rates. Population stratification also has the potential to confound inferences about gene-environment or gene-gene interactions, although generally to a much lesser extent (Wacholder et al., 2000).

Examples where structure has mislead LD mapping

Two genetic association studies using case-control methodology in medical genetics are commonly cited as examples of spurious findings due to population stratification. Knowler et al. (1988) reported an association between an HLA haplotype and diabetes for Pima Indians. When the analysis was repeated stratifying subjects by the amount of European ancestry, the observed association between HLA haplotype and diabetes was not observed. The other common example is Blum et al. (1990) who reported an association between alcoholism and the dopamine DRD2 allele. Gelernter et al. (1993), through a review of all published association studies of the alcoholism and the dopamine DRD2 allele, demonstrated no association between trait and allele, indicating that the original findings were likely due to population stratification because of the large ethnic variation in the prevalence of the A1 allele and alcoholism. In plants, a study of association between Dwarf8 gene and flowering time in maize (Thornsberry et al., 2001) is a practical example of how structure can cause false asso-

ciations. This flagship study involved phenotyping maize lines and sequencing of the candidate gene in 92 inbred lines of maize representing both tropical and North American lines, and genotyped well over 100 SSR loci from across the maize genome. An excessive proportion of SSR alleles were found to be associated with the two phenotypes, due to population structure. When they accounted for structure, the association reduced to a nominal level. Rapid decay of linkage disequilibrium at the gene locus enhanced the resolution power of association tests, allowing resolution within a few thousand base pairs when they used regression analysis that included population structure. However, they were unable to associate the individual polymorphisms with the phenotype because the polymorphisms were in linkage disequilibrium with each other and they had only used maize inbred lines. Perhaps the pitfall for this experiment was the lack of natural populations. The study by Thornsberry et al. (2001) made three important contributions: firstly, it demonstrated the utility of association mapping in plants; secondly, it demonstrated the practicability of the new methods that account for structure, and thirdly it showed the importance of using natural populations in association mapping. These are only but a few examples that highlight the magnitude of the problem of confounding population structures in scientific efforts aimed at localizing genes underlying complex traits through association mapping.

Overcoming the problem of population structure

Confounding effects of population stratification can be minimized or altogether eliminated by appropriate experimental designs. We suggest a few of such designs:

Better measures of populations: Within the context of population stratification, more detailed information on individual demes than broad conventional categories such as population or species is needed. There are two dimensions to this challenge:

a. Individuals must be allocated to the finest population of origin or deme categories that can reliably be determined.
b. Individuals from mixed-ancestry families (hybrids or admixed) must be treated appropriately (Hartl, 2000).

If cases and controls are individually matched, as is generally desirable to allow for various confounding variables in addition to multiple origins, it may not always be possible to obtain an exact match on population or demes, but at least an approximate match should be attempted, with further adjustments made in the analysis. Admixed families pose greater challenges (Spielman et al., 1993). It can be very difficult, if not impossible, to find a matching control for an individual with mixed origins, and it forces one to rely on multivariate models for adjustment. Rather than allocate the entire individual to a single stratum in the analysis as conventionally done, one can construct a covariate for each stratum, giving the proportion of ancestry derived from each group and include these covariates as adjustment variables in a multiple logistic regression.

Use of Family-member controls: It has been proposed that the confounding effect of population stratification and associated difficulty of finding homogenous population controls can be completely overcome by use of family member controls (Teng and Risch, 1999). The most commonly used familial case-control designs involve the use of siblings or parents as controls (Witte et al., 1999). Sibling controls are derived from exactly the same gene pool as the cases and thus represent exactly matched controls, but they pose other practical and statistical difficulties. The major practical difficulty is that not every case will have an available sibling. Also, if sibship size or other determinants of availability are associated with genotype, selection bias will result, which could go in either direction, depending upon the direction of the association, and could increase the risk of spurious associations with candidate genes that are associated with such selective factors. A second difficulty is that controls should generally be selected from siblings who have already survived to the age that can be diagnostically concluded to be case-free for the trait. In practice, this will generally tend to limit control eligibility to older siblings, which can lead to confounding by factors related to, for example, age. Siblings are also more likely to have the same genotype as the case than are unrelated controls, thereby leading to some loss of statistical efficiency (i.e., larger sample sizes required to attain the same statistical precision). Particularly, this method would present more difficulty in medical genetics if a late onset disease were to be studied because, first parents must be recruited and genotyped, whence they qualify for further analysis only if they are heterozygous at marker alleles. We are not stating here that TDT is altogether a very bad method, only that using TDT, the gains do not outweigh the investment and that the practicality of its design can be difficult, if not impossible, for certain tests such as late trait onset like dementia in humans or wood hardness in trees.

ASSOCIATION MAPPING IN PLANT BIOTECHNOLOGY

An assessment between genes controlling various quantitative traits such as yield and growth rates, and characters in plant species is a breeding approach that will ensure early maturity and quality in much younger individuals to cut on costs and time, resulting in a more economical land use. However, leading experts agree that marker assisted selection (MAS) are of little relevance in QTLs detection in non-hybrid populations (Strauss et al.,

1992; Plomion et al., 2003). Likewise, family based methods for dissecting these QTLs usually detects associations that are family specific (Plomion et al., 2003). LD mapping approach is specifically suited for localizing genes responsible for crop yield and growth rates in natural populations, as it has many advantages over family based methods as discussed in previous sections. Hence LD mapping holds promise in crop improvement programmes. However, the complexity of mating systems in plants poses a serious obstacle to this method.

Mating systems

Utility of Association mapping methods in plants have been limited by the fear of spurious associations that may result from population structure (Pritchard, 2001). Stratification (demes) within plant populations is more definite than just likely, and mating in plants is quite different from animal systems. Hence their genetic structures cannot be predicted as it varies among species and even among populations within a species. Pollination here depends on agents, such as insects, birds, water and wind, so mating is determined by a combination of the plants themselves and their agents, and we cannot observe these events most of the time. The other complexity is that because plants are immotile, distant gene dispersal is largely uniparental (through pollen).

Perhaps forest trees are a best demonstration of mating system in plants. The frequency of hybridization in forest tree genera is amongst the highest. For instance, 123 hybrid combinations have been recorded for the 23 British willow (Salix) species (Potts et al., 2003). The authors noted that there is usually low outcrossing in Eucalypt species, possibly due to differing flowering regimes or their protandrous nature. Mating in Eucalypts is rarely panmictic and studies consistently report positive values for Wright's fixation index (F) in open-pollinated populations of seed or germinant, indicating a deficit of heterozygotes at early stage of the life cycle and marked deviation from random mating (Potts and Wiltshire, 1997). In the USSR, where 125 plant families are represented, 25 are interspecific hybrids and introgressants of mostly trees. This deficit could result from inbreeding or population sub-structuring into gene pools differing in allele frequencies (the Wahlund effect), and from temporal and spatial variation in the allele frequencies in the population, like variation in fecundity and flowering phenology of differing genotypes, where the population may be divided into groups of trees with similar peak flowering times.

Genetic diversity in plants

Widespread plant species generally have high level of genetic diversity. There emerges a general relationship between genetic architecture and distribution patterns. Grouped by geographical range, widespread species have greater overall total heterozygosity (HT) and within population (HS) genetic diversity, and a lower proportion of localised alleles than regional and localised species. Differentiation is greater among populations of regional species, where the species with disjunct distributions have much higher FST values than those with continuous distributions (Moran, 1992). Within populations, non-random association of alleles from different loci (disequilibrium) may arise through inbreeding or asexual reproduction, linkage, genetic drift, mutation, migration, hybridization, epistatic selection, and other factors. Since plant species are generally immotile, they are predisposed to undergo adaptive divergence than their animal counterparts. Evolutionary history of most plants predisposes them to factors that could cause adaptive divergence, drift and bottlenecks. Most of the phenotypic variations in a common environment generally occurs within populations, but may be due to a combination of genetic and environmental effects. Specific sampling across steep environmental gradients regularly reveals marked genetic differentiation over short distances, even within continuous stands (Potts and Wiltshire, 1997).

WHY A PRIORI KNOWLEDGE OF STRUCTURE IS NEEDED IN PLANTS

Statistical approaches that use independent loci to control for the effects of structure and admixture by detecting and correcting for them have recently been developed (Devlin and Roeder, 1999; Pritchard and Rosenberg, 1999; Pritchard et al., 2000; Satten et al., 2001; Reich and Goldstein, 2001). A simulation comparing these methods to a "better" method by Chen et al. (2003) found that the earlier methods could not deal with complex structures. Recently, a group from diverse institutions in the USA published a mixed-model approach that can handle more complex situations than all existing models in terms of correcting for structure (Yu et al., 2005). The authors agreed that the methods described above do not account for complex families, pedigrees, founding effects and structures. Despite these splendid efforts, all these approaches will still have limitations, as acknowledged in a review by Flint and Mott (2001). Better (presumed or real) methods will emerge year after year. However, the basic issue is that all these "better" methods utilize a sampling strategy that is not based on sound knowledge of population structure. Pritchard and Rosenberg (1999) suggest that inference follow a two-step process; first by using a panel of markers to test for stratification followed by evaluating the candidate gene association only if homogeneity is not rejected. By simulation, they showed that the method performs well using a panel of couple dozen markers. However, it is not clear what would be inferred should the hypothesis of homogeneity be rejected in the first stage.

In Pritchard-Rosenberg method, most geneticists would be devastated to learn that a study they have laboriously conducted should simply be discarded because of the existence of population stratification at a panel of markers in which they possibly had no interest. We emphasize that a sound mapping strategy should always be preceded by an analysis of population structure, so that sampling for the mapping population can be based on a population of known gene flow pattern. This makes it essential to have prior information on the population assignment in the greatest detail that is practically possible.

In this review, we have shown that plant species are expected to have a population structure that cannot be predicted without empirical evaluation. Many crops and natural plant species have muddled demographic histories characterised by range expansions-/contractions, disjunct distributions leading to bottle-necks, founder events and inbreeding. They are also characterised by biological processes such as adaptive and human selection, mixed mating systems (Leimu, 2004), widespread hybridization and reticulation, and shifts in pollinators (Schmidt-Adam et al., 2000), leading to complex population structures. We emphasize that for a robust LD mapping for genes of economic importance in plants, and to avoid the ever-growing list of false associations as demonstrated in other fields, population structure and assignment tests should be conducted before commencing work on association analysis. This can be implemented using neutral molecular markers such as micro satellites (Krutovsky et al., 2006; Ochieng et al., 2006; Steane et al., 2006). In this way judicious sampling strategy and design based on a sound knowledge can ensure that an association detected between a marker allele and a trait is valid, if structure alone were the problem.

ACKNOWLEDGEMENT

This work was done while the first author was on research leave under a Scholarship from the Commonwealth of Australia.

REFERENCES

Altshuler D, Daly M, Kruglyak L (2000). Guilt by association. Nat. Genet. 26: 135-137.

Blum K, Noble EP, Sheridan PJ, Mongtomery A, Ritchie T, Jagadeeswarns P, Nogami H, Briggs AH, Cohn JB (1990). Allelic association of human dopamine D(2) receptor gene in alcoholism. JAMA 263: 2055–2060.

Chen HS, Zhu X, Zhao H, Zhang S (2003). Qualitative Semi-Parametric Test for Genetic Associations in Case-Control Designs Under Structured Populations. Ann. Hum. Genet. 67: 250-253.

Darvasi A, Soller M (1995). Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141: 1199-1207.

Devlin B, Roeder K (1999). Genomic control for association studies. Biometrics 55: 997-1004.

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M (2000). Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 10: 220–27

Flint J, Mott R (2001). Finding the molecular basis of quantitative traits: successes and pitfalls. Nat. Rev. 2: 437-445.

Gambaro G, Anglani F, D'Angelo A (2000). Association study designs of complex diseases. Lancet 355: 308–311.

Gelernter J, Goldman D, Risch N (1993). The A1 allele at the D2 dopamine receptor gene and alcoholism. A reappraisal. JAMA 269: 1673-1677.

Hartl DL (2000). A primer of population genetics, third edn. Sinauer Associates Inc., Sunderland.

Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K (2002). A comprehensive review of genetic association studies. Genet. Med. 4: 45-61.

Holland JB, Kresovich S, and Buckler E (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203 - 208.

Holtzman NA (2001). Putting the search for genes in perspective. Int. J. Health Service 31: 445–461.

Jannoo N, Grivet L, Dookun A, D'Hont A, Glaszmann JC (1999). Linkage disequilibrium among modern sugarcane cultivars. Theor. Appl. Genet. 99:1053–60.

Knowler WC, Williams RC, Pettit DJ, Steinberg AG (1988). Gm3,5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. Am. J. Hum. Med. 43: 520-526.

Krutovsky KV, St. Clair JB, Saich R, Hipkins DV, Neale DB (2006). Estimation of Population Structure In The Douglas-Fir Association Mapping Study. Plant and Animal Genomes Conference XIV PAG, San Diego, USA: W119.

Lander ES, Schork NJ (1994). Genetic Dissection of Complex Traits. Science 265: 2037-2048.

Leimu R (2004). Variation in the Mating System of Vincetoxicum hirundinaria (Asclepiadaceae) in Peripherial Island Populations. Ann. Bot. 93: 107-113.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN (2003). Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. Nat. Genet. 33: 177–182.

Martin ER, Monks SA, Warren LL, Kaplan NL (2000). A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am. J. Hum. Genet. 67: 146-154.

Mohlke KL, Lange EM, Valle TT, Ghosh S, Magnuson VL, Silander K, Watanabe RM, Chines PS, Bergman RN, Tuomilehto J, Collins FS, Boehnke M (2001). Linkage Disequilibrium Between Microsatellite Markers Extends Beyond 1 cM on Chromosome 20 in Finns. Genome Research 11: 1221-1226.

Moran GF (1992). Patterns of genetic diversity in Australian tree species. New Forests 6: 49-66.

Ochieng JW, Baverstock PR, Henry RJ, Shepherd M (2006). Population panmixia in a spotted gum species complex (Mytaceae): Implications for association mapping. Plant and Animal Genomes Conference XIV PAG, San Diego, USA: p.657.

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Bosch E, Bertranpetit J (1997). Allele frequencies for 20 microsatellites in a worldwide population survey. Hum. Hered. 47: 189-196.

Plomion C, Cooke J, Richardson T, MacKay J, Tuskan G (2003). Report on the Forest Trees Workshop at the Plant and Animal Genome Conference. Comp. Funct. Genomics 4: 229-238.

Potts BM, Wiltshire JE (1997). Eucalypt genetics and genecology. Cambridge University Press, Cambridge. pp.???

Potts BM, Barbour RC, Hingston AB, Vaillancourt RE (2003). Genetic Pollution of native eucalypt gene pools- identifying the risks. Aust. J. Bot. 51: 1-25.

Pritchard JK (2001). Deconstructing maize population structure. Nat. Genet. 28: 203-204.

Przeworski M (2002).The signature of positive selection at randomly chosen loci. Genetics 160: 1179-1189.

Pritchard JK, Rosenberg NA (1999). Use of Unlinked Genetic Markers to Detect Population Stratification in Association Studies. Am. J. Hum. Genet. 65: 220–228.

Pritchard JK, Przeworski M (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. 69: 1-14.
Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000). Association mapping in structured populations. Am. J. Hum. Genet. 67: 170-181.
Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, Lander ES (2001). Linkage disequilibrium in the human genome. Nature 411: 199–204.
Reich DE, Goldstein DB (2001). Detecting association in a case-control study while correcting for population stratification. Genet. Epidemiol. 20: 4-16.
Risch NJ (2000). Searching for genetic determinants in the new millennium. Nature 405: 847-856.
Satten GA, Flanders WD, Yang Q (2001). Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. Am. J. Hum. Genet. 68: 466-477.
Gabriele SA, Young AG, Murray BG (2000). Low Outcrossing Rates and Shift in Pollinators in New Zealand Pohutukawa (Metrosideros excelsa; Myrtaceae) Am. J. Bot. 87(9): 1265-1271
Spielman RS, Ewens WJ (1996). The TDT and other family-based tests for linkage disequilibrium and association. Am. J. Hum. Genet. 59: 983-989.
Spielman RS, Mcginnis RE, Ewens WJ (1993). Transmission Test for Linkage Disequilibrium - the Insulin Gene Region and Insulin-Dependent Diabetes-Mellitus (Iddm). Am. J. Hum. Genet. 52: 506-516.
Steane DA, Conod N, Jones RC, Vaillancourt RE, Potts BM (2006). A comparative analysis of population structure of a forest tree, *Eucalyptus globulus* (Myrtaceae), using microsatellite markers and quantitative traits. Tree Genet. Genomes 2: 30-38
Strauss SH, Lande R, Namkoong G (1992). Limitation of molecular marker-aided selection in forest tree breeding. Can. J. For. Res. 22: 1050-1061.
Strohman R (2002). Maneuvering in the complex path from genotype to phenotype. Science 296: 701-703.
Teng J, Risch N (1999). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. Genome Research 9: 234-241.
Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES (2001). Dwarf8 polymorphisms associate with variation in flowering time. Nat. Genet. 28: 286-289.
Yu JM, Pressoir G, Bi I, Yamasaki M, Doebley JF, McMullen M, Gaut BS, Nielsen D, Holland JB, Kresovich S, Buckler E (2005). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203 - 208.
Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N (2000). Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. J. Natl. Canc. Inst. 96: 434 – 42.
Witte JS, Gauderman WJ, Thomas DC (1999). Asymptotic bia s and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. Am. J. Epidemiol. 148: 693-705.