Heredity (2011) 106, 511–519

www.nature.com/hdy

REVIEW

Overview of techniques to account for confounding due to population stratification and cryptic relatedness in genomic data association analyses

MJ Sillanpää[1,2]
[1]Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland and [2]Department of Agricultural Sciences, University of Helsinki, Helsinki, Finland

Population-based genomic association analyses are more powerful than within-family analyses. However, population stratification (unknown or ignored origin of individuals from multiple source populations) and cryptic relatedness (unknown or ignored covariance between individuals because of their relatedness) are confounding factors in population-based genomic association analyses, which inflate the false-positive rate. As a consequence, false association signals may arise in genomic data association analyses for reasons other than true association between the tested genomic factor (marker genotype, gene or protein expression) and the study phenotype. It is therefore important to correct or account for these confounders in population-based genomic data association analyses. The common correction techniques for population stratification and cryptic relatedness problems are presented here in the phenotype–marker association analysis context, and comments on their suitability for other types of genomic association analyses (for example, phenotype–expression association) are also provided. Even though many of these techniques have originally been developed in the context of human genetics, most of them are also applicable to model organisms and breeding populations.
Heredity (2011) **106,** 511–519; doi:10.1038/hdy.2010.91; published online 14 July 2010

Keywords: association analysis; stratification; relatedness; genomic data; Bayesian statistics

Introduction

In genomic data association analysis, putative functional links can be identified by relating the measurements from single genomic data type to other data sources such as observed phenotypes, gene or protein expressions, molecular markers, functional classifications or cellular responses (for example, Risch and Merikangas, 1996; Jansen and Nap, 2001; Jansen *et al*., 2002; Bhattacharjee *et al*., 2008). Quantitative and qualitative traits are commonly studied by using the methods developed for phenotype–marker association analysis. These same methods can also be used to find regulatory pathways or patterns controlling gene expressions (eQTLs) and protein expressions (pQTLs) by treating the expression level of the gene or protein as a classical quantitative phenotype (Jansen and Nap, 2001; Bystrykh *et al*., 2005; Foss *et al*., 2007). Other possible links can be identified by using phenotype–expression and phenotype–protein association analysis as well as simultaneous association analysis of multiple data types (Hoti and Sillanpää, 2006; Bhattacharjee *et al*., 2008; Sillanpää and Noykova, 2008; Bhattacharjee and Sillanpää, 2009). Complementary evidence provided by different data sources can be joined together afterwards to study supported overlapping genomic regions (Aune *et al*., 2004; Bhattacharjee *et al*., 2008).

Genome-wide (population-based) association analysis is generally considered to be a main tool to infer causative links between genomic marker data and phenotype (Risch and Merikangas, 1996; McCarthy and Hirschhorn, 2008). This happens regardless of problems such as genetic heterogeneity (Terwilliger and Weiss, 1998; Sillanpää and Bhattacharjee, 2006), winner's curse (Lande and Thompson, 1990; Beavis, 1998; Göring *et al*., 2001; Xiao and Boehnke, 2009) and missing heritability (Maher, 2008; McCarthy and Hirschhorn, 2008; Slatkin, 2009). A particular property of marker data is the systematic spatial dependence along the chromosome. In gene- and protein-expression data, the spatial dependence of expression values at neighbouring genes (or mass-to-charge ratios) is not well established. However, it is possible that the normalization method used will induce some spatial dependence to the expression data. It is also common to assume the presence of some other form of dependence in the data, namely that the expressions of genes belonging to the same pathway are highly dependent on each other. Exception for the theme is provided by protein antibody microarrays (used in studies of cancer immune responses) in which the presence of spatial dependence is evident (Wu *et al*., 2009). It is good to keep in mind differences between data types, as they affect the applicability of the methods reviewed in this study.

Correspondence: Dr MJ Sillanpää, Department of Mathematics and Statistics, University of Helsinki, PO Box 68, Helsinki FIN-00014, Finland.
E-mail: mjs@rolf.helsinki.fi
Received 25 March 2010; revised 3 June 2010; accepted 8 June 2010; published online 14 July 2010

In a population-based phenotype–marker association study, we hope that study individuals are distantly related at the small genome regions (containing the trait loci) so that there is systematic linkage disequilibrium generated by rare occurrence of recombination events within these regions during a large number of meioses in the ancestral pedigree. At the same time, individuals in the study sample are assumed to be mutually independent (unrelated or equally related to each other). However, this assumption does not hold for individuals showing more complex relationship structures, and the potential existence of such discrepancy in the data is generally known as a cryptic relatedness problem (for example, Devlin and Roeder, 1999; Voight and Pritchard, 2005; Zhang and Deng, 2010). In cryptic relatedness, relationships between individuals in the sample are typically either completely known (for example, pedigrees or families are available) or unknown (for example, a sample of potentially related individuals). The existence of individuals originating from multiple source populations in the study sample may create another problem known as population stratification (for example, Lander and Schork, 1994; Cardon and Palmer, 2003). In addition, in this case, the population structure may be either known (for example, a sample of very distinct populations) or unknown (for example, a random sample of individuals from a single or multiple sites).

It is well known that population-based genomic data association analyses generally suffer from confounding because of population stratification (inability to divide the variance into within- and among-population components) and cryptic relatedness (inability to account for varying within-population relationships among study individuals). If not properly accounted for, spurious associations may occur in the genomic data association analyses because of these confounding factors (stratification or cryptic relatedness) rather than real association between the tested genomic factor and the trait value. Population stratification is a more widely discussed topic than cryptic relatedness. Yet, several papers on both topics can be easily found in phenotype–marker association studies (Lander and Schork, 1994; Cardon and Palmer, 2003; Yu *et al*., 2006; Kang *et al*., 2008), and also in phenotype–expression association studies (Gibson, 2003; Kraft and Horvath, 2003; Kraft *et al*., 2003; Lu *et al*., 2004) and clinical quantitative trait locus studies in which phenotype is simultaneously explained by multiple data types (Hoti and Sillanpää, 2006; Pikkuhookana and Sillanpää, 2009). It may be noted that in some cases, for example, in model organisms, population stratification and cryptic relatedness may both be needed to be corrected simultaneously (Yu *et al*., 2006; Kang *et al*., 2008; Stich *et al*., 2008).

In the first section, we will cover the most common approaches to controlling population stratification. In the second section, we will consider approaches for cryptic relatedness. Finally, we consider the use of estimation-based variable selection and multilocus association models and their robustness to these confounding factors. In each point, the methods are presented for determining phenotype–marker association, but the suitability of each approach is also commented for other types of genomic data association analyses.

Approaches for population stratification

In principle, it is possible to minimize the risk of population stratification by carefully selecting the study material from a genetically isolated population, or by using stringent ethnic origin criteria. Otherwise, in population-based association studies, the current techniques to overcome the problem of hidden population structure (stratification) can roughly be divided into the following categories: (1) stratified analysis, (2) genomic controls, (3) structured association, (4) smoothing, (5) principal component approach, (6) matching, (7) approaches based on relationship information and, finally, (8) use of secondary samples. Even though most of these approaches have been considered only together with genetic marker data, they may be arguably applicable with small changes also for other data types (that is, use of gene and protein expressions as explanatory variables). In the following, we shortly present the underlying ideas behind these approaches.

Stratified analysis

If groups of individuals are known or have been observed before the analysis, it is possible to study within-group genomic associations, which are robust to population stratification (Clayton, 2007). Completely separate within-group (within strata) analyses will decrease the overall statistical power because of small sample sizes, but it is also possible to combine within-group information in the test statistic or joint likelihood. The family-based association test methods rely on this principle by using family as unit for a known group (Lange *et al*., 2002; Horvath *et al*., 2004).

Unfortunately, within-group membership information or families are not always easy to collect and one can try to approximate membership information (construct approximate 'families') based on some other information that is available, for example, self-identified ethnicity in human data (see Tang *et al*., 2005), known location where individuals' grandparents lived, or estimate the most likely ancestry (population assignment) or pairwise relatedness based on an independent set of molecular markers (for example, Lynch and Ritland, 1999; Pritchard *et al*., 2000a; Weir *et al*., 2006). For within-population stratified analyses, see the section 'Structured association'. It is, however, known that population-based association studies even with related individuals are statistically more powerful than family-based (within family) association studies (Teng and Risch, 1999; Havill *et al*., 2005; Aulchenko *et al*., 2007a; Hernandéz-Sánchez *et al*., 2003). Thus, other ways of correcting for population stratification than this may be more favourable.

Genomic controls

In the genomic control approach, one modifies (adjust) the threshold P-value on the basis of a neutral set of independent markers—null markers (genetic marker panel) providing information on the adequate adjustment factor (Devlin and Roeder, 1999; Banacu *et al*., 2002; Zheng *et al*., 2006). The adjustment factor (λ) describes variance inflation, that is, reduction in effective sample size ($\approx N/\lambda$), where N is the original sample size (see Hinds *et al*., 2004). The underlying assumption for this method is that all the spurious association signals are

smaller than the real signals, so that it is possible to handle the problem by adequately adjusting the threshold value. It is assumed that this external set of markers does not include any trait-associated loci. The benefit of this approach is that one does not need to make any assumptions on the number of subpopulations. Different variants of the genomic control approach have been considered and compared by Dadd *et al*. (2009). An improved version of this approach was presented by Wang (2009), in which, instead of using a single adjustment factor, one can use several of them. However, it has been argued that the genomic control approach suffers from weak statistical power when the effect of population structure is large, as is common in model organisms (Yu *et al*., 2006; Kang *et al*., 2008). In principle, by using the genomic control approach, it should be possible to construct a related genomic control adjustment factor based on the neutral gene expression data, which can then be applied for testing phenotype–expression association.

Structured association

In the structured association, unknown population membership probabilities are first estimated using population assignment methods (for example, Pritchard *et al*., 2000a; Dawson and Belkhir, 2001; Corander *et al*., 2003; Falush *et al*., 2003; Alexander *et al*., 2009). These probabilities are subsequently used in association analysis (Pritchard *et al*., 2000b; Thornsberry *et al*., 2001; Yu *et al*., 2006). Phenotype–marker association at candidate locus can be tested within subpopulations using likelihood ratio test (Pritchard *et al*., 2000b; Thornsberry *et al*., 2001). Alternatively, the association model can include the subpopulation mean terms weighted with individual membership probabilities (Yu *et al*., 2006). Versions of the method in which both of these tasks (estimation of population memberships and phenotype–marker association) are carried out simultaneously exist (for example, Satten *et al*., 2001; Ripatti *et al*., 2001; Sillanpää *et al*., 2001; Hoggart *et al*., 2003). In all of these structured association methods, population memberships are estimated on the basis of neutral set of independent markers (genetic marker panel). For exception to this, see Sillanpää and Bhattacharjee (2006). It may be noted that the structured association approach has been recently extended to genome-wide sets of marker loci (Alexander *et al*., 2009). Generally, it is easy to include the subpopulation mean terms also in other types of genomic (that is, phenotype–expression or phenotype–protein) association models or in models that consider marker and expression data jointly to explain the phenotype.

Smoothing

The idea in the smoothing approach is that the signal is smoothed along the chromosome according to an exponential decay or some other spatial function, depending on the genetic or physical map distances (Conti and Witte, 2003; Sillanpää and Bhattacharjee, 2005; Tsai *et al*., 2008). This is feasible for a tightly linked set of markers that are in considerable linkage disequilibrium with each other. In such a setting, neighbouring markers are used to strengthen the weak but real association signals and smooth the spurious association signals downwards. A similar control approach for spurious peaks has also been proposed for proteomics data sets, in which protein intensity peaks are smoothed with respect to the neighbouring locations (Du *et al*., 2006) and according to the m/z distance between the positions (Bhattacharjee *et al*., 2008). For a related smoothing approach for expression data, see Sillanpää and Noykova (2008).

Principal component approach

The use of principal component analysis to correct the stratification in structured populations has been suggested (Patterson *et al*., 2006; Price *et al*., 2006; Zhu *et al*., 2008). One proceeds by assuming that an external set of neutral molecular markers is available and any of them is not associated with the trait of interest. Principal components are first estimated from the correlation matrix of external marker genotypes of unrelated individuals (Price *et al*., 2006). Then, the first few principal components (explaining most of the underlying variation) are used as regression covariates in the association model, or are incorporated into a randomization test (see Kimmel *et al*., 2007). On minor levels of stratification, one can simply omit samples that appear as outliers in principal component approaches. In the related method of Epstein *et al*. (2007), instead of principal components, one uses components from a partial least-squares regression. Owing to their ease of implementation, these methods are very popular in human genetics, even if the correction given by them may fail, especially if the external marker set used is not large (for example, Epstein *et al*., 2007; Lee *et al*., 2008; Wang, 2009). It should be relatively easy to apply these corrections (estimated either from external marker or from expression data) to phenotype–expression and phenotype–protein–expression association studies.

Matching

During the design stage of the study, it is possible to collect pairs of individuals (full sibs or cousins; one case and one control individual) that are otherwise similar (that is, individually matched) with respect to covariates such as ethnic background, sex, age and so on (Gauderman *et al*., 1999; Zondervan *et al*., 2002). Group matching refers to a similar process in which instead of individuals, groups are matched. Even if matching could eliminate the problem of population stratification, one potential problem is over-matching (that is, reduction of statistical power by unnecessary matching on too many factors, which creates matched units that are too much alike also in their phenotypes). To consider matching in quantitative traits, phenotypically discordant sib-pair collection strategies may be used (cf. Risch and Zhang, 1995).

Special procedures for genetic ancestry matching that are applicable to existing data sets have been proposed lately on the basis of the information on non-genetic variables (Lee, 2004) and marker genotypes (Hinds *et al*., 2004; Luca *et al*., 2008; Guan *et al*., 2009). In Hinds *et al*. (2004), ancestry was estimated by population assignment methods, and in Luca *et al*. (2008) by principal components. These methods considered different strategies for genetic ancestry group matching and removing 'unmatchable outlier individuals'. Guan *et al*. (2009)

proposed the use of identity-by-state-based simple (dis)similarity measures as a tool for individual genetic ancestry matching. Luca *et al*. (2008) emphasized the use of 'control databases' in genome-wide association studies and considered a problem wherein cases are sampled in a quite different region from that of the controls.

Generalization and application of these techniques to other forms of genomic association analysis should be relatively easy.

Use of relationship information

Affected trios: The transmission and disequilibrium test (TDT) allows to control for confounding by studying association in the presence of linkage (Spielman *et al*., 1993). Originally TDT was developed as a confirmatory second test to filter real associations out of spurious signals. This original motivation is well justified because TDT has lesser power than ordinary association testing (Long and Langley, 1999). However, TDT is nowadays commonly used as a general test of association. The basic version of TDT assumes binary phenotype, biallelic loci and uses data on affected trios (unrelated cases and their parents). At a given locus, it tests whether a certain allele is transmitted from heterozygous parents to the affected offspring more often than expected under Mendelian segregation, and the observed segregation distortion is taken as evidence for the locus having something to do with the affection status. TDT has been generalized to quantitative traits (Allison, 1997; Rabinowitz, 1997; Abecasis *et al*., 2000), multiallelic markers (Sham and Curtis, 1995), marker haplotypes (Clayton, 1999) and to several data structures (Spielman and Ewens, 1998; Abecasis *et al*., 2000). As handling of nonrandom missing genotype data in parents is problematic, a robust version of TDT has also been developed (Sebastiani *et al*., 2004).

Pseudo-control data: Another relationship information-based approach to control for confounding is the so-called pseudo-control approach, in which a sample containing only affected cases (and their parents) is collected and the artificial control sample is derived indirectly on the basis of parental genotypes and haplotypes. At each locus, pseudo-control individuals have genetic material that was not transmitted from the parents to the cases (for example, Falk and Rubinstein, 1987; Terwilliger and Ott, 1992; Lander and Schork, 1994; Gauderman *et al*., 1999; Greenland, 1999). The benefit of deriving the control sample in this way is that one obtains well-matched controls and avoids spurious associations because of ethnic confounding, that is, closer kinship among the affected samples (Terwilliger and Weiss, 1998). This pseudo-control approach has been generalized also to multilocus association analysis and single-tail sampling with quantitative traits (Sillanpää and Hoti, 2007).

Pedigree data: More general approaches use pedigree data, in which linkage information and association information are combined (George *et al*., 1999; Lund *et al*., 2003; Pérez-Enciso, 2003; Meuwissen and Goddard, 2004; Meuwissen and Goddard, 2007; Gasbarra *et al*., 2009; Hernandéz-Sánchez *et al*., 2009), resulting in a strong signal at true positions. Linkage information confirms only the real associations and one obtains weaker signals at the spurious positions, which provides a way to control confounding in association studies.

In pedigree-based linkage analysis, founder individuals are generally assumed to be unrelated. However, association information is also available in pedigrees, when founders are related. Thus, combined analysis tries to model also relationships between pedigree founders. To do so, some assumptions (for example, from effective population size) are often made from founders and/or a recent history of the population (see for example, Meuwissen and Goddard, 2001).

Correction methods using relationship information are not easy to generalize to gene- or protein-expression data because these approaches are based on the discrete nature of the marker data and the linkage concept.

Use of secondary samples

The idea behind this approach is that analysis is carried out jointly for two samples of data from the same study population, but with different study designs: one containing a population-based sample of individuals (the association signal from these data suffers from confounding) and the other sample comprising related individuals (the association signal from these data is robust to confounding; see above). As the overall signal is a synthesis of the individual signals of two data sets, it is likely to be relatively robust to confounding (for example, Epstein *et al*., 2005; Kazeem and Farrall, 2005). In addition, this 'meta-analysis' approach improves the statistical power by combining information from multiple data sets. For a review and comparison of different secondary sample approaches, see Glaser and Holmans (2009) and Infante-Rivard *et al*. (2009). Alternatively, it is possible to analyse these two samples (with association and linkage) separately and study the overlap between the results (Manenti *et al*., 2009), or carry out association testing conditionally on the linkage results (Cantor *et al*., 2005).

As the use of secondary samples to correct for population stratification relies on two separate samples, issues relating to the presence of heterogeneity cannot be fully ruled out (see Sillanpää and Auranen, 2004). In addition, as this correction method is based on the use of relationship information and marker linkage, this method is not easy to generalize to gene- or protein-expression data. A variant in which two samples are combined and population stratification is corrected for by using the principal component approach (Zhu *et al*., 2008) should also be applicable for gene- or protein-expression data.

Approaches for cryptic relatedness

It is good to keep in mind that population stratification and cryptic relatedness are two different problems and correction methods typically consider only a single problem at a time. Exceptions to this were the stratified analysis, genomic controls and the use of relationship information subsections above. Especially useful in this respect may be the approaches in which linkage information and association information are combined. Otherwise, the current techniques to overcome the problem of cryptic relatedness in population-based association analysis can be divided into the following categories: (1) infinite polygenic model, (2) regression

covariates, (3) test-statistic accounting for relatedness and (4) genomic controls. These techniques can be generalized quite easily to the other genomic data analyses. In the following, we shortly describe the ideas behind these methods.

Infinite polygenic model

The classical approach to correcting for relatedness in the sample is to include a polygenic component into the population-based genomic association analysis model (for example, George and Elston, 1987; Jannink *et al*., 2001; Lu *et al*., 2004; Yu *et al*., 2006; Bradbury *et al*., 2007; Pikkuhookana and Sillanpää, 2009). This practice is also known as the measured genotype approach. Because of the availability of large high-throughput association data sets, it is more popular to use a recent variant of this approach, called GRAMMAR (Amin *et al*., 2007; Aulchenko *et al*., 2007a, b), in which residual dependencies are first precorrected from data and repeated (phenotype–marker) association analyses are carried out for adjusted residuals using rapid methods. Even though this approach was presented for marker data, it is straightforward to use it (or measured genotype approach) in concert with other types of genomic (for example, phenotype-expression or other) association analysis. Nevertheless, the precorrection approach suffers from model misspecification. It underestimates uncertainty in polygenic effects, and may reduce the statistical power (cf. Martinez *et al*., 2005). Moreover, estimation of variance components is known to be unstable when sample sizes are small (Misztal, 1996; Burton *et al*., 1999; Pikkuhookana and Sillanpää, 2009). In any case, the GRAMMAR approach has been shown to outperform many of the competing methods introduced for pedigree-based association analysis (see Aulchenko *et al*., 2007a). For unknown relationships, a relationship matrix can be first estimated using various methods (for example, Milligan, 2003; Leutenegger *et al*., 2003; Blouin, 2003; Weir *et al*., 2006; Frentiu *et al*., 2008). Interestingly, the use of a simple identity-by-state allele-sharing matrix has recently been found to provide an efficient alternative to more sophisticated methods in correcting for cryptic relatedness-induced confounding (Zhao *et al*., 2007; Kang *et al*., 2008). See also van de Casteele *et al*. (2001) and Bink *et al*. (2008).

Regression covariate approach

An alternate approach to correcting for relatedness is to use the method of Bonney (Bonney, 1986; Thomas, 2004). This method approximates the influence of the infinite polygenic model by having phenotypes of the parents, the spouse and sibs of the participant as regression covariates in the model (Bonney, 1986). Pikkuhookana and Sillanpää (2009) compared the performance of these two approaches (infinite polygenic model and regression covariates) in Bayesian genomic association models and found the regression covariate approach to perform better for smaller genomic data sets. It may be noted that their model considered the effects of marker genotypes and gene expressions jointly in explaining the phenotype. Although this approach provides a framework for including phenotypes from ungenotyped parents into the analysis (cf. Purcell *et al*., 2005), it cannot be applied in case of the unknown relatedness, as the phenotypes of the relatives are generally not available.

Test statistic accounting for relatedness

The test statistics often used for population-based association analysis assume the independence of individuals in the sample. Test statistics for population-based phenotype–marker association testing among related individuals (correlated family data) have been introduced, for example, for (between-family) association in family-based design (Teng and Risch, 1999) or for a more general design using all the related and unrelated individuals (Slager and Schaid, 2001). A similar kind of test based on family-based (within-family) association analysis has been developed for expression data (Kraft *et al*., 2003). In case of unknown relationships, one can start by estimating the family structure (or pairwise relatedness) using an additional set of molecular markers (for example, Gasbarra *et al*., 2007; Bink *et al*., 2008). However, one should be careful here whether the test statistic is measuring population-based or within-family association. Unlike within-family association, population-based association analysis with related individuals suffers from population stratification, but has more statistical power than family-based (within family) association analysis, which was considered in the section 'Stratified analysis' (Teng and Risch, 1999; Slager and Schaid, 2001). With modifications, it is possible to derive suitable test statistics accounting for relatedness in population-based association for gene- or protein-expression data.

Genomic controls

Even though genomic control has been introduced to control for population stratification, it has also been suggested for handling cryptic relatedness (Devlin and Roeder, 1999; Banacu *et al*., 2002; Yan *et al*., 2009). In this method, variation inflation is corrected by adjusting the test statistic based on the information from unlinked null markers. This works also in case of unknown relatedness because unlike the infinite polygenic model, one does not need to estimate the pairwise relationships first. The genomic control approach has also been suggested to be useful for additional correction after the infinite polygenic model is fitted (Amin *et al*., 2007).

Use of estimation-based variable selection and multilocus models

It may sound strange that by using estimation-based variable selection and multilocus models (without a correction term), one can automatically reduce the number of false positives in genomic data association analyses. However, there is increasing evidence that Bayesian or frequentist multilocus modelling approaches are flexible enough to automatically account or self-correct for population stratification in binary traits (Setakis *et al*., 2006), ordinal and censored traits (Iwata *et al*., 2009), as well as in quantitative traits (Iwata *et al*., 2007). In case of cryptic relatedness and quantitative traits, the self-correction property of Bayesian multilocus association approach was found by Pikkuhookana and Sillanpää (2009). The robustness of the multilocus association approach to these problems results presumably

from the fact that, during the estimation (variable selection) process, other genetic components, few at a time, could capture or explain a small amount of confounding variation (see Pikkuhookana and Sillanpää, 2009). This is essentially so because in these approaches, variable selection is done simultaneously with the effect estimation (Kilpikari and Sillanpää, 2003; O'Hara and Sillanpää, 2009) and large candidate panels jointly have the potential to explain many types of variation. For a close connection between the multilocus association model and a polygenic model with realized relationship matrix, see Hayes *et al*. (2009). However, additional studies on the self-correction property of these approaches are needed before one can say anything definitive on this, for example, on how much variability in the markers is needed for self-correction approach to be effective. In two-genotype data analysis (in which there is only a single estimable effect coefficient at each locus), Iwata *et al*. (2007) found that the use of a correction term in the model still provides some additional advantages over self-correction. However, it is likely that using single-nucleotide polymorphism data and by fitting two estimable coefficients (for three genotypes) can provide more variability and again more ability for self-correction in the model. As a conclusion, I wish to emphasize that there are good reasons why future studies on genomic data association analysis should focus on or at least pay much more attention to better characterizing the benefits and pitfalls of these estimation-based multilocus approaches. It is also well known that the use of multilocus models improves statistical power and helps avoid problems due to model misspecification, such as biased position estimates, and occurrence of 'ghost QTLs' (see, for example, Sillanpää and Auranen, 2004).

Conflict of interest

The author declares no conflict of interest.

Acknowledgements

I am grateful to Crispin M Mutshinda for constructive comments on this paper. This work was supported by research grants from the Academy of Finland and University of Helsinki's research funds.

References

Abecasis GR, Cardon LR, Cookson WO (2000). A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* **66**: 279–292.

Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**: 1655–1664.

Allison DB (1997). Transmission-disequilibrium tests for quantitative traits. *Am J Hum Genet* **60**: 676–690.

Amin N, van Duijn CM, Aulchenko YS (2007). A genomic background based method for association analysis in related individuals. *PloS One* **12**: e1274.

Aulchenko YS, de Koning D-J, Haley C (2007a). Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* **177**: 577–585.

Aulchenko YS, Ripke S, Isaacs A, Van Duijn CM (2007b). GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**: 1294–1296.

Aune TM, Parker JS, Mass K, Liu Z, Olson NJ, Moore JH (2004). Co-localization of differentially expressed genes and shared susceptibility loci in human autoimmunity. *Genet Epidemiol* **27**: 162–172.

Banacu SA, Devlin B, Roeder K (2002). Association studies for quantitative traits in structured populations. *Genet Epidemiol* **22**: 78–93.

Beavis WD (1998). QTL analysis: power, precision, and accuracy. In: Paterson AH (ed.). *Molecular Dissection of Complex Traits*. CRC Press: Boca Raton, FL, pp 145–162.

Bhattacharjee M, Botting CH, Sillanpää MJ (2008). Bayesian biomarker identification based on marker-expression-proteomics data. *Genomics* **92**: 384–392.

Bhattacharjee M, Sillanpää MJ (2009). Bayesian joint disease-marker-expression analysis applied to clinical characteristics of chronic fatigue syndrome. In: McConnell P, Lim S, Cuticchia AJ (eds) *Methods of Microarray Data Analysis VI*. CreateSpace Publishing: Scotts Valley, CA, pp 15–34.

Bink MCAM, Anderson AD, van de Weg WE, Thompson EA (2008). Comparison of marker-based pairwise relatedness estimators on a pedigreed plant population. *Theor Appl Genet* **117**: 843–855.

Blouin MS (2003). DNA-based methods for pedigree reconstruction and kinship analysis in natural populations. *Trends Ecol Evol* **18**: 503–511.

Bonney GE (1986). Regressive logistic models for familial disease and other binary traits. *Bioinformatics* **42**: 611–625.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics* **23**: 2633–2635.

Burton P, Tiller K, Gurrin L, Cookson W, Musk A, Palmer LJ (1999). Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling. *Genet Epidemiol* **17**: 118–140.

Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T *et al*. (2005). Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. *Nat Genet* **37**: 225–232.

Cantor RM, Chen GK, Pajukanta P, Lange K (2005). Association testing in a linked region using large pedigrees. *Am J Hum Genet* **76**: 538–542.

Cardon LR, Palmer LJ (2003). Population stratification and spurious allelic association. *Lancet* **361**: 598–604.

Clayton DG (1999). A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* **65**: 1170–1177.

Clayton DG (2007). Population association. In: Balding DJ, Bishop M, Cannings C (eds). *Handbook of Statistical Genomics* 3rd edn, vol. 2. pp. Wiley: Chichester, UK, pp 1264–1285.

Conti DV, Witte JS (2003). Hierarchical modeling of linkage disequilibrium: genetic structure and spatial relations. *Am J Hum Genet* **72**: 351–363.

Corander J, Waldmann P, Sillanpää MJ (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**: 367–374.

Dadd T, Weale ME, Lewis CM (2009). A critical evaluation of genomic control methods for genetic association studies. *Genet Epidemiol* **33**: 290–298.

Dawson KJ, Belkhir K (2001). A Bayesian approach to the identification of panmictic populations and the assignment of individuals. *Genet Res* **78**: 59–77.

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

Du P, Kibbe WA, Lin SM (2006). Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **22**: 2059–2065.

Epstein MP, Allen AS, Satten GA (2007). A simple and improved correction for population stratification in case-control studies. *Am J Hum Genet* **80**: 921–930.

Epstein MP, Veal CD, Trembath RC, Barker JNWN, Li C, Satten GA (2005). Genetic association analysis using data from triads and unrelated subjects. *Am J Hum Genet* **76**: 592–608.
Falk CT, Rubinstein P (1987). Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Ann Hum Genet* **51**: 227–233.
Falush D, Stephens M, Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567–1587.
Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR *et al.* (2007). Genetic basis of proteome variation in yeast. *Nat Genet* **39**: 1369–1375.
Frentiu FD, Clegg SM, Chittock J, Burke T, Blows MW, Owens IPF (2008). Pedigree-free animal models: the relatedness matrix reloaded. *Proc R Soc B* **275**: 639–647.
Gasbarra D, Pirinen M, Sillanpää MJ, Arjas E (2009). Bayesian quantitative trait locus mapping based on reconstruction of recent genetic histories. *Genetics* **183**: 709–721.
Gasbarra D, Pirinen M, Sillanpää MJ, Salmela E, Arjas E (2007). Estimating geneologies from unlinked marker data: a Bayesian approach. *Theor Pop Biol* **72**: 305–322.
Gauderman WJ, Witte JS, Thomas DC (1999). Family-based association studies. *J Natl Cancer Inst Monographs* **26**: 31–37.
George V, Tiwari HK, Zhu X, Elston RC (1999). A test of transmission/disequilibrium for quantitative traits in pedigree data, by multiple regression. *Am J Hum Genet* **65**: 236–245.
George VT, Elston RC (1987). Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet Epidemiol* **4**: 193–201.
Gibson G (2003). Population genomics: celebrating individual expression. *Heredity* **90**: 1–5.
Glaser B, Holmans P (2009). Comparison of methods for combining case-control and family-based association studies. *Hum Hered* **68**: 106–116.
Greenland S (1999). A unified approach to the analysis of case-distribution (case-only) studies. *Stat Med* **18**: 1–15.
Guan W, Liang K, Boehnke M, Abecasis CR (2009). Genotype-based matching to correct for population stratification in large-scale case-control genetic association studies. *Genet Epidemiol* **33**: 508–517.
Göring HHH, Terwilliger JD, Blangero J (2001). Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* **69**: 1357–1369.
Havill LM, Dyer TD, Richardson DK, Mahaney MC, Blangero J (2005). The quantitative trait linkage disequilibrium test: a more powerful alternative to the quantitative transmission disequilibrium test for use in the absence of population stratification. *BMC Genet* **6**(Suppl 1): S91.
Hayes BJ, Visscher PM, Goddard ME (2009). Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* **91**: 47–60.
Hernandéz-Sánchez J, Grunchec J-A, Knott S (2009). A web application to perform linkage disequilibrium and linkage analyses on a computational grid. *Bioinformatics* **25**: 1377–1383.
Hernandéz-Sánchez J, Haley CS, Visscher PM (2003). Power of QTL detection using association tests with family controls. *Eur J Hum Genet* **11**: 819–827.
Hinds DA, Stokowski RP, Patil N, Konvicka K, Kershenobich D, Cox DR *et al.* (2004). Matching strategies for genetic association studies in structured populations. *Am J Hum Genet* **74**: 317–325.
Hoggart CJ, Parra EJ, Shriver MD, Bonilla C, Kittles RA, Clayton DG *et al.* (2003). Control of confounding in genetic associations in stratified populations. *Am J Hum Genet* **72**: 1492–1504.
Horvath S, Xu X, Lake SL, Silverman EK, Weiss ST, Laird NM (2004). Family-based tests for associating haplotypes with general phenotype data: application to asthma genetics. *Genet Epidemiol* **26**: 61–69.
Hoti F, Sillanpää MJ (2006). Bayesian mapping of genotype x expression interactions in quantitative and qualitative traits. *Heredity* **97**: 4–18.
Infante-Rivard C, Mirea L, Bull SB (2009). Combining case-control and case-trio data from the same population in genetic association analyses: overview of approaches and illustration with a candidate gene study. *Am J Epidemiol* **170**: 654–664.
Iwata H, Ebana K, Fukuoka S, Jannink J-L, Hayashi T (2009). Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among Oryza sativa L germplasms. *Theor Appl Genet* **118**: 865–880.
Iwata H, Uga Y, Yoshioka Y, Ebana K, Hayashi T (2007). Bayesian association mapping of multiple quantitative trait loci and its application to the analysis of genetic variation among Oryza sativa L germplasms. *Theor Appl Genet* **114**: 1437–1449.
Jannink J-L, Bink MCAM, Jansen RC (2001). Using complex plant pedigrees to map valuable genes. *Trends Plant Sci* **6**: 337–342.
Jansen R, Greenbaum D, Gerstein M (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res* **12**: 37–46.
Jansen RC, Nap J-P (2001). Genetical genomics: the added value from segregation. *Trends Genet* **17**: 388–391.
Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ *et al.* (2008). Efficient control of population structure in model organism association mapping. *Genetics* **178**: 1709–1723.
Kazeem GR, Farrall M (2005). Integrating case-control and TDT studies. *Ann Hum Genet* **69**: 329–335.
Kilpikari R, Sillanpää MJ (2003). Bayesian analysis of multilocus association in quantitative and qualitative traits. *Genet Epidemiol* **25**: 122–135.
Kimmel G, Jordan MI, Halperin E, Shamir R, Karp RM (2007). A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet* **81**: 895–905.
Kraft P, Horvath S (2003). The genetics of gene expression and gene mapping. *Trends Biotechnol* **21**: 377–378.
Kraft P, Schadt E, Aten J, Horvath S (2003). A family-based test for correlation between gene expression and trait values. *Am J Hum Genet* **72**: 1323–1330.
Lande R, Thompson R (1990). Efficiency of marker-assisted selection in the improvement of quantitative traits. *Genetics* **124**: 743–756.
Lander ES, Schork NJ (1994). Genetic dissection of complex traits. *Science* **265**: 2037–2048.
Lange C, DeMeo DL, Laird NM (2002). Power and design considerations for a general class of family-based association tests: quantitative traits. *Am J Hum Genet* **71**: 1330–1341.
Lee S, Sullivan P, Zou F, Wright F (2008). Comment on a simple and improved correction for population stratification. *Am J Hum Genet* **82**: 524–526.
Lee W-C (2004). Case-control association studies with matching and genomic controlling. *Genet Epidemiol* **27**: 1–13.
Leutenegger AL, Prum B, Génin E, Verny C, Lemainque A, Clerget-Darpoux F *et al.* (2003). Estimation of the inbreeding coefficient through use of genomic data. *Am J Hum Genet* **73**: 516–523.
Long AD, Langley CH (1999). The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* **9**: 720–731.
Lu Y, Liu P-Y, Liu Y-J, Xu F-H, Deng H-W (2004). Quantifying the relationship between gene expression and trait values in general pedigrees. *Genetics* **168**: 2395–2405.
Luca D, Ringquist S, Klei L, Lee AB, Gieger C, Wichman H-E *et al.* (2008). On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am J Hum Genet* **82**: 453–463.

Lund MS, Sorensen P, Guldbrandtsen B, Sorensen DA (2003). Multitrait fine mapping of quantitative trait loci using combined linkage disequilibria and linkage analysis. *Genetics* **163**: 405–410.

Lynch M, Ritland K (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152**: 1753–1766.

Maher BS (2008). The case of the missing heritability. *Nature* **456**: 18–21.

Manenti G, Galvan A, Pettinicchio A, Trincucci G, Spada E, Zolin A *et al.* (2009). Mouse genome-wide association mapping needs linkage analysis to avoid false-positive loci. *PloS Genet* **5**: e1000331.

Martinez V, Thorgaard G, Robison B, Sillanpää MJ (2005). An application of Bayesian QTL mapping to early development in double haploid lines of rainbow trout including environmental effects. *Genet Res* **85**: 209–221.

McCarthy MI, Hirschhorn JN (2008). Genome-wide association studies: potential next steps on a genetic journey. *Hum Mol Genet* **17**: R156–R165.

Meuwissen THE, Goddard ME (2001). Prediction of identity by descent probabilities from marker-haplotypes. *Genet Sel Evol* **33**: 605–634.

Meuwissen THE, Goddard ME (2004). Mapping multiple QTL using linkage disequilibrium and linkage analysis information and multitrait data. *Genet Sel Evol* **36**: 261–279.

Meuwissen THE, Goddard ME (2007). Multipoint identity-by-descent prediction using dense markers to map quantitative trait loci and estimate effective population size. *Genetics* **176**: 2551–2560.

Milligan BG (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163**: 1153–1167.

Misztal I (1996). Estimation of variance components with large-scale dominance models. *J Dairy Sci* **80**: 965–974.

O'Hara RB, Sillanpää MJ (2009). A review of Bayesian variable selection methods: what, how and which. *Bayesian Anal* **4**: 85–118.

Patterson N, Price AL, Reich D (2006). Population structure and eigen analysis. *PloS Genet* **2**: e190.

Pérez-Enciso M (2003). Fine mapping of complex trait genes combining pedigree and linkage disequilibrium information: a Bayesian unified framework. *Genetics* **163**: 1497–1510.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006). Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**: 904–909.

Pritchard JK, Stephens M, Donnelly P (2000b). Inference of population structure using multilocus genotype data. *Genetics* **155**: 845–959.

Pritchard JK, Stephens M, Rosenberg NA, Donnelly P (2000a). Association mapping in structured populations. *Am J Hum Genet* **67**: 170–181.

Pikkuhookana P, Sillanpää MJ (2009). Correcting for relatedness in Bayesian models for genomic data association analysis. *Heredity* **103**: 223–237.

Purcell S, Sham P, Daly MJ (2005). Parental phenotypes in family-based association analysis. *Am J Hum Genet* **76**: 249–259.

Rabinowitz D (1997). A transmission disequilibrium test for quantitative trait loci. *Hum Hered* **47**: 342–350.

Ripatti S, Pitkäniemi J, Sillanpää MJ (2001). Joint modeling of genetic association and population stratification using latent class models. *Genet Epidemiol* **21**(Suppl 1): S401–S414.

Risch N, Merikangas K (1996). The future of genetic studies of complex human diseases. *Science* **273**: 1616–1617.

Risch N, Zhang H (1995). Extreme discordant sib pairs for mapping quantitative trait loci in humans. *Science* **268**: 1584–1589.

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.

Sebastiani P, Abad MM, Alpargu G, Ramoni MF (2004). Robust transmission/disequilibrium test for incomplete family genotypes. *Genetics* **168**: 2329–2337.

Setakis E, Stirnadel H, Balding DJ (2006). Logistic regression protects against population structure in genetic association studies. *Genome Res* **16**: 290–296.

Sham PC, Curtis D (1995). An extended transmission/disequilibrium test (TDT) for multi-allele marker loci. *Ann Hum Genet* **59**: 323–336.

Sillanpää MJ, Auranen K (2004). Replication in genetic studies of complex traits. *Ann Hum Genet* **68**: 646–657.

Sillanpää MJ, Bhattacharjee M (2005). Bayesian association-based fine mapping in small chromosomal segments. *Genetics* **169**: 427–439.

Sillanpää MJ, Bhattacharjee M (2006). Association mapping of complex trait loci with context-dependent effects and unknown context variable. *Genetics* **174**: 1597–1611.

Sillanpää MJ, Hoti (2007). Mapping quantitative trait loci from a single-tail sample of the phenotype distribution including survival data. *Genetics* **177**: 2361–2377.

Sillanpää MJ, Kilpikari R, Ripatti S, Onkamo P, Uimari P (2001). Bayesian association mapping for quantitative traits in a mixture of two populations. *Genet Epidemiol* **21** (Suppl 1): S692–S699.

Sillanpää MJ, Noykova N (2008). Hierarchical modelling of clinical and expression quantitative trait loci. *Heredity* **101**: 271–284.

Slager SL, Schaid DJ (2001). Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am J Hum Genet* **68**: 1457–1462.

Slatkin M (2009). Epigenetic inheritance and the missing heritability problem. *Genetics* **182**: 845–850.

Spielman SR, Ewens WJ (1998). A sibship test for linkage in presence of association: the sib transmission/disequilibrium test. *Am J Hum Genet* **62**: 450–458.

Spielman RS, McGinnis RE, Ewens WJ (1993). Transmission test for linkage disequilibrium: the insulin gene region ans insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52**: 506–516.

Stich B, Möhring J, Piepho H-P, Heckenberger M, Buckler ES, Melchinger AE (2008). Comparison of mixed-model approaches for association mapping. *Genetics* **178**: 1745–1754.

Tang H, Quertermous T, Rodriguez B, Kardia SLR, Zhu X, Brown A *et al.* (2005). Genetic structure, self-identified race/ethnicity, and confounding in case-control association studies. *Am J Hum Genet* **76**: 268–275.

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 Res* **9**: 234–241.

Terwilliger JD, Ott J (1992). A haplotype-based 'haplotype relative risk' approach to detecting allelic associations. *Hum Hered* **42**: 337–346.

Terwilliger JD, Weiss KM (1998). Linkage disequilibrium mapping of complex trait: fantasy or reality? *Curr Opin Biotechnol* **9**: 578–594.

Thomas DC (2004). *Statistical Methods in Genetic Epidemiology.* Oxford University Press: New York.

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.

Tsai MY, Hsiao CK, Wen SH (2008). A Bayesian spatial multimarker genetic random-effect model for fine-scale mapping. *Ann Hum Genet* **72**: 658–669.

van de Casteele T, Galbusera P, Matthysen E (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Mol Ecol* **10**: 1539–1549.

Voight BF, Pritchard JK (2005). Confounding from cryptic relatedness in case-control association studies. *PloS Genet* **1**: e32.

Wang K (2009). Testing for genetic association in the presence of population stratification in genome-wide association studies. *Genet Epidemiol* **33**: 637–645.

Weir BS, Anderson AD, Hepler AB (2006). Genetic relatedness analysis: modern data and new challenges. *Nat Rev Genet* **7**: 771–780.

Wu J, Patwa TH, Lubman DM, Ghosh D (2009). Identification of differentially expressed spatial clusters using humoral response microarray data. *Comput Stat Data Anal* **53**: 3094–3102.

Xiao R, Boehnke M (2009). Quantifying and correcting for the winner's curse in genetic association studies. *Genet Epidemiol* **33**: 453–462.

Yan T, Hou B, Yang Y (2009). Correcting for cryptic relatedness by a regression-based genomic control method. *BMC Genet* **10**: 78.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF *et al.* (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* **38**: 203–208.

Zhang F, Deng H-W (2010). Correcting for cryptic relatedness in population-based association studies of continuous traits. *Hum Hered* **69**: 28–33.

Zhao K, Aranzana MJ, Kim S, Lister C, Shindo C, Tang C *et al.* (2007). An *Arabidopsis* example of association mapping in structured samples. *PloS Genet* **3**: e4.

Zheng G, Freidlin B, Gastwirth JL (2006). Robust genomic control for association studies. *Am J Hum Genet* **78**: 350–356.

Zhu X, Li S, Cooper RS, Elston RC (2008). A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* **82**: 352–365.

Zondervan KT, Cardon LR, Kennedy SH (2002). What makes a good case-control study. *Hum Reprod* **17**: 1415–1423.