PROGRESS

GENOME-WIDE ASSOCIATION STUDIES

New approaches to population stratification in genome-wide association studies

Alkes L. Price, Noah A. Zaitlen, David Reich and Nick Patterson

Abstract | Genome-wide association (GWA) studies are an effective approach for identifying genetic variants associated with disease risk. GWA studies can be confounded by population stratification — systematic ancestry differences between cases and controls — which has previously been addressed by methods that infer genetic ancestry. Those methods perform well in data sets in which population structure is the only kind of structure present but are inadequate in data sets that also contain family structure or cryptic relatedness. Here, we review recent progress on methods that correct for stratification while accounting for these additional complexities.

Genome-wide association (GWA) studies have identified hundreds of common variants associated with disease risk or related traits[1] (see the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies). These studies have overcome the dangers of population stratification, which can produce spurious associations if not properly corrected[2–3]. However, accounting for population structure is more challenging when family structure or cryptic relatedness is also present, and these limitations have motivated the development of new methods. Spurious associations have occurred primarily at markers with unusual allele frequency differences among subpopulations[2,4], so it is crucial that new methods aimed at correcting for stratification are evaluated at unusually differentiated markers.

The prevailing paradigm in recent years has been to use genomic control to measure the extent of inflation due to population stratification or other confounders, and to correct for stratification (if necessary) using methods that infer genetic ancestry, such as structured association or principal components analysis (PCA). A limitation of this strategy is that it fails to account for other types of sample structure, such as family structure or cryptic relatedness[5–6]. Modelling family structure is a necessity in studies with family-based sample ascertainment, and there is increasing evidence that cryptic relatedness may occur in a wide range of data sets (see below). Family-based association tests offer one potential solution for dealing with family structure. More recently, approaches using mixed models that incorporate the full covariance structure across individuals have been proposed.

Below, we review each of these methods, conduct simulations to evaluate their performance, discuss stratification in the specific context of low-frequency or rare variants and conclude with guidelines and recommendations.

Detecting stratification

A widely used approach to evaluate whether confounding due to population stratification exists is to compute the genomic control λ (λ_{GC}), which is defined as the median χ^2(1 degree of freedom) association statistic across SNPs divided by its theoretical median under the null distribution[7–9]. A value of $\lambda_{GC} \approx 1$ indicates no stratification, whereas $\lambda_{GC} > 1$ indicates stratification or other confounders, such as family structure or cryptic relatedness (see below), or differential bias[10]. P–P plots are a standard tool for visualization of test statistics (FIG. 1). Values of $\lambda_{GC} < 1.05$ are generally considered benign; we note that inflation in λ_{GC} is proportional to sample size.

If population stratification exists, it is important to distinguish between subpopulation differences that are due to recent genetic drift and those that arose from more ancient population divergence[11]. In the case of genetic drift, dividing association statistics by λ_{GC} will provide a sufficient correction for stratification. In the case of ancient population divergence, markers with unusual allele frequency differences that lie outside the expected distribution, which could be caused by natural selection, make stratification a much more severe problem, and dividing association statistics by λ_{GC} is likely to be inadequate. In the case of family structure or cryptic relatedness, dividing association statistics by λ_{GC} will generally produce the approximate null distribution, although a refinement to the method may be needed when there is uncertainty in the estimate of λ_{GC} (REF. 12). However, even if the appropriate null distribution is obtained, in general this approach will not maximize power to detect true associations. Other approaches to correcting for stratification, including approaches that also account for family structure and cryptic relatedness, are described below.

Inferring genetic ancestry

Structured association. Methods that explicitly infer genetic ancestry generally provide an effective correction for population stratification in data sets in which population structure is the only type of sample structure. In the structured association approach, samples are assigned to subpopulation clusters (possibly allowing fractional cluster membership) using a model-based clustering program such as STRUCTURE[13–14], and association statistics are computed by stratifying by cluster using a program such as STRAT[15]. The applicability of this approach to large genome-wide data sets has historically been limited by its high computational cost when allowing fractional cluster membership, but faster model-based approaches for inferring population structure have recently been developed (such as the ADMIXTURE software)[16]. Thus, applying structured association to both infer population structure and compute

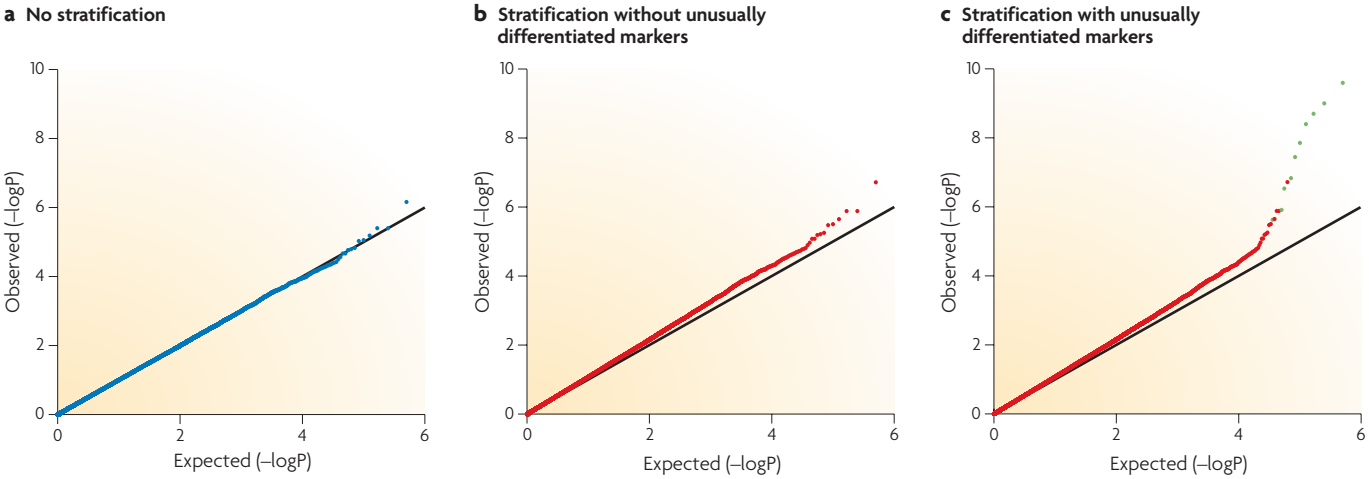

Figure 1 | **P–P plots for the visualization of stratification or other confounders.** The figure shows simulated P–P plots under three scenarios for genome-wide scans with no causal markers. **a** | No stratification: *p*-values fit the expected distribution. **b** | Stratification without unusually differentiated markers: *p*-values exhibit modest genome-wide inflation. **c** | Stratification with unusually differentiated markers: *p*-values exhibit modest genome-wide inflation and severe inflation at a small number of markers.

association statistics in genome-wide data sets is likely to become a practical approach.

Principal components analysis. PCA is a tool that has been used to infer population structure in genetic data for several decades, long before the era of GWA studies[17–20]. It should be noted that top principal components do not always reflect population structure: they may reflect family relatedness[19], long-range linkage disequilibrium (LD) (due to, for example, inversion polymorphisms[4]) or assay artefacts[10]. These effects can often be eliminated by removing related samples, regions of long-range LD or low-quality data, respectively, from the data used to compute principal components. In addition, PCA can highlight effects of differential bias that require additional quality control[21].

Using top principal components as covariates corrects for stratification in GWA studies[21,22], and this can be done using software such as EIGENSTRAT. Like structured association, PCA will appropriately apply a greater correction to markers with large differences in allele frequency across ancestral populations. Unlike initial implementations of structured association, PCA is computationally tractable in large genome-wide data sets. Related approaches, such as multidimensional scaling (MDS) and genetic matching, have also proven useful[23–25] and can be carried out using the PLINK software. When genome-wide data are not available (for example, in replication studies), structured association or PCA can infer genetic ancestry, and hence correct for stratification, using ancestry-informative markers (AIMs)[26]. A common misconception is that AIMs should be used to infer genetic ancestry even when genome-wide data are available, but in fact the best ancestry estimates are obtained using a large number of random markers.

A limitation of the above methods is that they do not model family structure or cryptic relatedness. These factors may lead to inflation in test statistics if they are not explicitly modelled because samples that are correlated are assumed to be uncorrelated. Although correcting for genetic ancestry and then dividing by the residual λ_{GC} will restore an appropriate null distribution, association statistics that explicitly account for family structure or cryptic relatedness are likely to achieve higher power owing to improved weighting of the data.

Family-based association tests

Family-based studies, in which individuals are ascertained from family pedigrees, offer a unique solution to population stratification. Family-based association tests that focus on within-family information (generalizing the transmission disequilibrium test[27]) are immune to stratification, as transmitted and untransmitted alleles have the same genetic ancestry[28–30], and such tests can be performed using the FBAT and QTDT software. However, fully powered statistics for family-based studies will need to incorporate between-family information, which is still susceptible to stratification. A recent suggestion is to transform between-family information into a rank statistic before combining within-family and between-family information, guaranteeing that both sources of information are immune to stratification[31,32]. This approach performs favourably compared with previous family-based approaches[31,32], but it places an upper bound on the statistical power that can be extracted from the between-family component of the overall signal. This is because the transformed rank statistic cannot be more statistically significant than one divided by the number of samples.

Mixed models

Mixed models, which owe their roots to applications in animal breeding, can model population structure, family structure and cryptic relatedness[33,34]. The basic approach is to model phenotypes using a mixture of fixed effects and random effects. Fixed effects include the candidate SNP and optional covariates, such as gender or age, whereas random effects are based on a phenotypic covariance matrix, which is modelled as a sum of heritable and non-heritable random variation (BOX 1). Mixed models have historically been a theoretically appealing but computationally intensive approach; however, recent computational advances (such as the EMMAX and TASSEL software) have now made it possible to apply them to GWA studies[35,36]. Methods that explicitly model population structure, family structure and cryptic relatedness are expected to perform better in the presence of these complexities than methods that do not, and this has now been confirmed[35,36]. For example, in an analysis of seven Wellcome Trust Case Control Consortium phenotypes, the application of mixed models consistently

yielded values of λ_{GC} that were less than 1.01, in contrast to other approaches[34].

Population structure: a fixed or random effect? An important and unanswered question is whether population structure should be modelled as part of the set of random effects, together with family structure and cryptic relatedness, or as a separate fixed effect requiring principal component covariates and additional model parameters[35,36] (BOX 1). Inclusion in random effects is much simpler, and has been shown to provide a sufficient correction for stratification in data sets from Finland and the UK[35].

However, population structure is actually a fixed effect (that is, its effect as a function of genetic ancestry is the same for all samples), and spurious associations might result if it is modelled as a random effect based on overall covariance, particularly in the case of unusually differentiated markers. Modelling population structure as a fixed effect provides a higher level of certainty in correcting for stratification but requires running PCA (or a similar method) to infer the genetic ancestry of each sample[36]. If family structure is present, inferring genetic ancestry by PCA is a challenge because family relatedness may lead to artefactual principal components[19]. A possible solution is to compute principal components using SNP loadings inferred from a set of unrelated samples, either by using a different set of samples from those in the disease study or by using an unrelated subset of samples from the disease study[37]. This is likely to be sufficient when the set of unrelated samples used is very large relative to the magnitude of population structure effects. However, unless sample sizes are very large, principal components computed from external SNP loadings will be biased towards zero owing to statistical noise in the SNP loadings[11,38]. This motivates further work on PCA in related samples.

Modelling phenotypes as fixed. Mixed models model phenotypes using a fixed set of genotypes. However, as an alternative to mixed models, genotypes can be modelled using a fixed set of phenotypes, a theoretically appealing approach that makes fewer assumptions about phenotypic covariance structure[39,40]. Simulations in the absence of unusually differentiated markers have shown that using the genotypic covariance matrix to account for both population and family structure can effectively control spurious associations under a variety of settings[39]; this can be done using the ROADTRIPS software. However, in the case of unusually differentiated markers, normality assumptions (about genotype distributions) underlying the test statistics will be violated, and stratification may lead to confounding unless principal component covariates are used. The question of whether to model random effects only or to include principal component covariates as fixed effects is analogous to the mixed model framework. When viewing phenotypes as fixed, principal component covariates may be essential. as modelling only random effects leads to a uniform correction factor in the absence of missing data[39].

Simulations

We carried out two simulations to show the properties of the above methods in correcting for stratification at normally differentiated or unusually differentiated markers in the presence or absence of family structure. We considered a case–control study with two subpopulations, POP1 and POP2, with 300 cases and 200 controls from POP1 and 200 cases and 300 controls from POP2. We simulated 99,900 normally differentiated markers based on F_{ST}(POP1,POP2) = 0.01 (REF. 41) and 100 unusually differentiated markers based on allele frequency difference equal to 0.6 with both minor allele frequencies uniformly distributed on [0.0,0.4][21]. In simulation 1, all individuals were unrelated. In simulation 2, all individuals from POP1 were unrelated and individuals from POP2 included 80 case–case sibling pairs, 40 case–control sibling pairs and 130 control–control sibling pairs. We computed λ_{GC} for each of the following methods: an uncorrected Armitage trend test, EIGENSTRAT[21], EMMAX without principal component covariates[35], EMMAX with principal component covariates[35] and ROADTRIPS[39]. All principal component runs used only one principal component, but the additional inclusion of random principal components has little effect on results[21]. Power to detect causal variants may vary between methods, but our focus here was on correcting false-positive associations. We did not simulate the approach described in REF. 31, as this method is completely immune from stratification, ensuring a value of 1.00 in all entries of the table; this approach has appealing properties, but may have reduced power in some instances (see above). We note that the method of REF. 39 with principal component covariates incorporated is an approach of potentially high interest, but it is not currently implemented in the ROADTRIPS software.

The results of the simulations are shown in TABLE 1. EIGENSTRAT is effective in correcting for population stratification at both normally and unusually differentiated markers (simulation 1) but does not control for

Box 1 | **Mixed models**

Simple linear models

Simple linear models represent the phenotype *Y* as a function of fixed effects *X*:

$$Y = XB + \varepsilon$$

Here, *X* denotes the genotype at the candidate marker in addition to optional covariates, such as gender or age, *B* denotes coefficients of fixed effects and ε is a normally distributed noise term that accounts for unexplained variation in *Y*.

Principal components analysis (PCA) addresses the issue of population substructure by including principal component covariates in *X* to explicitly model the ancestry of each individual. If genotype is not causally related to phenotype but genotype and phenotype are both correlated to ancestry, test statistics will be inflated. Using PCA to explicitly model genetic ancestry removes this confounding effect. However, PCA only accounts for fixed effects of genetic ancestry; it does not account for relatedness between individuals, which may also cause inflation in test statistics.

Linear mixed models

Linear mixed models represent the phenotype *Y* as a function of fixed effects *X* plus random effects *u*:

$$Y = XB + u + \varepsilon$$

$$Var(u) = \sigma_g^2 K$$

Here, *u* denotes a component of the overall noise variance $u + \varepsilon$ that is distributed according to a kinship matrix *K*. Thus, *u* represents the heritable component of random variation and ε represents the non-heritable component of random variation.

The kinship matrix *K* is defined according to the pairwise genotypic similarity of individuals, so its structure is influenced by population structure, family structure and cryptic relatedness. The parameter σ_g^2 relates this structure to the phenotype *Y*. σ_g^2 captures the extent to which genetically similar individuals are phenotypically similar, thus removing confounding effects. The optimal formulation of *K*, the importance of including principal component covariates in fixed effects *X* and the effects of these choices have not yet been fully explored.

Table 1 | **Effectiveness of different approaches for correcting for stratification**

	Simulation 1, F_{ST} = 0.01	Simulation 1, Δ = 0.6	Simulation 2, F_{ST} = 0.01	Simulation 2, Δ = 0.6
Armitage trend	1.40	48.4	1.57	48.3
EIGENSTRAT	1.00	1.00	1.17	1.14
EMMAX*	1.00	2.05	1.01	1.62
EMMAX* + principal components	1.00	1.02	1.01	1.01
ROADTRIPS	1.00	48.4	1.00	48.3

We list the genomic control λ (λ_{GC}) of each method for normally differentiated markers (F_{ST} = 0.01) and unusually differentiated markers (Δ = 0.6) in simulation 1 and simulation 2. In each case, λ_{GC} was computed as the median χ^2 (1 degree of freedom) statistic (restricting to the subclass of markers tested) divided by 0.455. EIGENSTRAT corrects for population structure (simulation 1), EMMAX and ROADTRIPS correct for family structure and for population structure at normally differentiated markers (F_{ST} = 0.01), and EMMAX + principal components corrects for family structure and for population structure at normally or highly differentiated markers (F_{ST} = 0.01 or Δ = 0.6). We note that the approach of REF. 31 is immune to all of these confounders, implying a value of λ_{GC} = 1.00 for each column of the table. *EMMAX can use either the identity by descent (IBS) or Balding–Nichols estimate of the kinship matrix[35]. Results for IBS are shown in the table, and results for Balding–Nichols are (from left to right) 1.00, 1.91, 1.00 and 1.28 for EMMAX and 1.00, 1.03, 1.00 and 0.99 for EMMAX + principal components.

family structure (simulation 2). EMMAX corrects for both stratification and population structure except for a modest residual inflation at unusually differentiated markers, which is completely removed by EMMAX with principal component covariates; if the number of unusually differentiated markers is small, modest inflation at such markers may not be a major concern. ROADTRIPS corrects for family structure but not for population stratification at unusually differentiated markers, although the incorporation of principal component covariates could potentially address this limitation. We note that for each method, dividing association statistics by residual λ_{GC} is guaranteed to produce statistics with λ_{GC} = 1, but this approach may be inadequate for spurious associations at unusually differentiated markers and/or may not maximize power if family structure (or cryptic relatedness) is not fully modelled.

Low-frequency and rare variants

GWA studies have largely focused on common variants, but because most genetic heritability[34] remains unexplained, future work will increasingly focus on variants of low minor-allele frequency (0.5% < MAF < 5%) or rare variants (MAF < 0.5%)[42]. First, new low-frequency variants will be identified by the 1000 Genomes Project and included in next-generation genotyping arrays. Here, the issues are generally similar to those involving common variants, except that deviation from model specification is more likely — for example, if normality assumptions are violated or the genotypic variance of a SNP varies across subpopulations[43]. Second, exome resequencing projects will aim to identify genes in which individuals with extreme phenotypes have an aggregate excess or deficiency of rare non-synonymous variants[44]. Differences in the range of allele frequencies across ancestral populations make stratification a potential concern, but genetic ancestry

Glossary

Ancestry-informative markers
Genetic markers ascertained for large differences in allele frequency between subpopulations that are genotyped to infer genetic ancestry in new samples.

Armitage trend test
A standard χ^2(1 degree of freedom) association test computed as the number of samples times the squared correlation between genotype and phenotype.

Cryptic relatedness
Sample structure due to distant relatedness among samples with no known family relationships.

Differential bias
Spurious differences in allele frequencies between cases and controls due to differences in sample collection, sample preparation and/or genotyping assay procedures.

Exome resequencing
A study design in which exon capture technologies are used to obtain resequencing data covering all exonic regions for each individual in the study.

Family-based association tests
A class of association tests that uses families with one or more affected children as the subjects rather than unrelated cases or controls. The analysis treats the allele that is transmitted to (one or more) affected children from each parent as a 'case' and the untransmitted alleles as 'controls' to avoid the effects of population structure.

Family structure
Sample structure due to familial relatedness among samples.

F_{ST}
A measure of the genetic distance between two populations that describes the proportion of overall genetic variation that is due to differences between populations.

Genetic drift
Random fluctuations in allele frequencies over time due to sampling effects, particularly in small populations.

Genetic heritability
The proportion of the total phenotypic variation in a given characteristic that can be attributed to additive genetic effects. In the broad sense, heritability involves all additive and non-additive genetic variance, whereas in the narrow sense, it involves only additive genetic variance.

Genetic matching
A method of association testing in which cases and controls are matched for genetic ancestry, as inferred by principal components analysis or other methods.

Genomic control
A method for detecting (or detecting and correcting for) stratification based on the genome-wide inflation of association statistics.

Mixed models
A class of models in which phenotypes are modelled using both fixed effects (candidate SNPs and fixed covariates) and random effects (the phenotypic covariance matrix).

Multidimensional scaling
A dimensionality reduction technique, similar to principal components analysis, in which points in a high-dimensional space are projected into a lower-dimensional space while approximately preserving the distance between points.

Population structure
Sample structure due to differences in genetic ancestry among samples.

Principal components analysis
A dimensionality reduction technique used to infer continuous axes of variation in genetic data, often representing genetic ancestry.

Rank statistic
A statistic describing the rank, across markers, of association of each marker. Rank statistics can be transformed into quantiles of a standard normal distribution that can be combined with other statistics.

SNP loadings
The correlations of each SNP to a given principal component in principal components analysis. The principal component coordinates of each sample are proportional to the sum of normalized genotypes weighted by SNP loadings.

Structured association
A method for correcting for stratification in which samples are assigned to subpopulation clusters and evidence of association is stratified by cluster.

Transmission disequilibrium test
A family-based association test involving case–parent trios in which alleles transmitted from parents to children are compared with untransmitted alleles.

can be inferred from genotyping array data from the same samples, if available, and included as a covariate. Finally, the advent of whole-exome or whole-genome resequencing raises the question of whether rare variants can be used to infer genetic ancestry with greater precision, perhaps using different methods from those currently applied to common variants.

Conclusion

Many different methods of correcting for stratification have been developed, and all of these methods have important advantages. Although mixed models are relatively new and untested, they seem to offer a practical and comprehensive approach for simultaneously addressing confounding due to population stratification, family structure and cryptic relatedness.

In studies in which stratification is not a serious concern, an appealing and simple approach is to use mixed models without including principal component covariates. This approach could be applied in studies of populations of homogeneous ancestry, studies of structured populations in which structure is due to very recent genetic drift and studies of any population in which PCA or related methods, applied to either the entire sample or a subset of unrelated samples, indicate that there is no substantial stratification (that is, phenotypes are not highly correlated with any of the top principal components).

For studies that do not meet any of the above criteria, an alternative approach is to use mixed models with principal component covariates. In family-based studies in which the within-family component contributes much of the overall statistical power, the approach of REF. 31 may also prove useful. In data sets that do not contain family structure or cryptic relatedness, simpler association tests (with or without principal component correction, based on above criteria) will probably be sufficient[21,23].

Alkes L. Price, Noah A. Zaitlen, David Reich and Nick Patterson are at the Program in Medical and Population Genetics, Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA.

Alkes L. Price and Noah A. Zaitlen are also at the Department of Epidemiology and Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, USA.

David Reich is also at the Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

Correspondence to A.L.P.
e-mail: aprice@hsph.harvard.edu

doi:10.1038/nrg2813
Published online 15 June 2010

1. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
2. Campbell, C. D. *et al.* Demonstrating stratification in a European American population. *Nature Genet.* **37**, 868–872 (2005).
3. Tian, C., Gregersen, P. K. & Seldin, M. F. Accounting for ancestry: population substructure and genome-wide association studies. *Hum. Mol. Genet.* **17**, R143–R150 (2008).
4. Tian, C. *et al.* Analysis and application of European genetic substructure using 300 K SNP information. *PLoS Genet.* **4**, e4 (2008).
5. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case–control association studies. *PLoS Genet.* **1**, e32 (2005).
6. Weir, B. S., Anderson, A. D. & Hepler, A. B. Genetic relatedness analysis: modern data and new challenges. *Nature Rev. Genet.* **7**, 771–780 (2006).
7. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
8. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
9. Reich, D. E. & Goldstein, D. B. Detecting association in a case–control study while correcting for population stratification. *Genet. Epidemiol.* **20**, 4–16 (2001).
10. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case–control association study. *Nature Genet.* **37**, 1243–1246 (2005).
11. Price, A. L. *et al.* The impact of divergence time on the nature of population structure: an example from Iceland. *PLoS Genet.* **5**, e1000505 (2009).
12. Devlin, B., Bacanu, S. A. & Roeder, K. Genomic control to the extreme. *Nature Genet.* **36**, 1129–1130 (2004); author reply in **36**, 1131 (2004).
13. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
14. Rosenberg, N. A. *et al.* Genetic structure of human populations. *Science* **298**, 2381–2385 (2002).
15. Pritchard, J. K., Stephens, M., Rosenberg, N. A. & Donnelly, P. Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
16. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).
17. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
18. Cavalli-Sforza L. L., Menozzi P. & Piazza A. *The History and Geography of Human Genes* (Princeton Univ. Press,1994).
19. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
20. Novembre, J. & Stephens, M. Interpreting principal component analyses of spatial population genetic variation. *Nature Genet.* **40**, 646–649 (2008).
21. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
22. Zhu, X., Zhang, S., Zhao, H. & Cooper, R. S. Association mapping, using a mixture model for complex traits. *Genet. Epidemiol.* **23**, 181–196 (2002).
23. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
24. Luca, D. *et al.* On the use of general control samples for genome-wide association studies: genetic matching highlights causal variants. *Am. J. Hum. Genet.* **82**, 453–463 (2008).
25. Lee, A. B., Luca, D., Klei, L., Devlin, B. & Roeder. K. Discovering genetic ancestry using spectral graph theory. *Genet. Epidemiol.* **34**, 51–59 (2010).
26. Seldin, M. F. & Price, A. L. Application of ancestry informative markers to association studies in European Americans. *PLoS Genet.* **4**, e5 (2008).
27. Spielman, R. S., McGinnis, R. E. & Ewens, W. J. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* **52**, 506–516 (1993).
28. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Rev. Genet.* **7**, 385–394 (2006).
29. Abecasis, G. R., Cardon, L. R. & Cookson, W. O. A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* **66**, 279–292 (2000).
30. Lange, C., DeMeo, D. L. & Laird, N. M. Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.* **71**, 1330–1341 (2002).
31. Won, S. *et al.* On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet.* **5**, e1000741 (2009).
32. Lasky-Su, J. *et al.* On genome-wide association studies for family-based designs: an integrative analysis approach combining ascertained family samples with unselected controls. *Am. J. Hum. Genet.* **86**, 573–580 (2010).
33. Yu, J. *et al.* A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genet.* **38**, 203–208 (2006).
34. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
35. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genet.* **42**, 348–354 (2010).
36. Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nature Genet.* **42**, 355–360 (2010).
37. Zhu, X., Li, S., Cooper, R. S. & Elston, R. C. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am. J. Hum. Genet.* **82**, 352–365 (2008).
38. Lee S., Zou F. & Wright F. A. Convergence and prediction of principal component scores in high-dimensional settings. *Ann. Stat*. (in the press).
39. Thornton, T. & McPeek, M. S. ROADTRIPS: case–control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* **86**, 172–184 (2010).
40. Rakovski, C. S. & Stram, D. O. A kinship-based modification of the Armitage trend test to address hidden population structure and small differential genotyping errors. *PLoS ONE* **4**, e5825 (2009).
41. Holsinger, K. E. & Weir, B. S. Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Rev. Genet.* **10**, 639–650 (2009).
42. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
43. Abney, M. & McPeek, M. S. Association testing with principal-components-based correction for population stratification [abstract number 58]. *Proc. of the 58th Annual Meeting of The American Soc. of Human Genetics* [online], http://www.ashg.org/2008meeting/abstracts/fulltext/ (2008).
44. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 2869–2872 (2004).

Competing interests statement

The authors declare no competing financial interests.

FURTHER INFORMATION

1000 Genomes Project: http://www.1000genomes.org
ADMIXTURE software: http://www.genetics.ucla.edu/software/admixture
EIGENSTRAT, implemented in the EIGENSOFT software: http://www.hsph.harvard.edu/faculty/alkes-price/software
EMMAX software: http://genetics.cs.ucla.edu/emmax
FBAT software: http://biosun1.harvard.edu/~fbat/fbat.htm
***Nature Reviews Genetics* article series on Genome-wide association studies:** http://www.nature.com/nrg/series/gwas/index.html
NHGRI Catalog of Published Genome-Wide Association Studies: http://genome.gov/gwastudies
PLINK software: http://pngu.mgh.harvard.edu/~purcell/plink
QTDT software: http://www.sph.umich.edu/csg/abecasis/QTDT
ROADTRIPS software: http://www.stat.uchicago.edu/~mcpeek/software/index.html
STRUCTURE and STRAT software: http://pritch.bsd.uchicago.edu/software.html
TASSEL software: http://www.maizegenetics.net

ALL LINKS ARE ACTIVE IN THE ONLINE PDF