Review

Population stratification and spurious allelic association

Lon R Cardon, Lyle J Palmer

Great efforts and expense have been expended in attempts to detect genetic polymorphisms contributing to susceptibility to complex human disease. Concomitantly, technology for detection and scoring of single nucleotide polymorphisms (SNPs) has undergone rapid development, extensive catalogues of SNPs across the genome have been constructed, and SNPs have been increasingly used as a means for investigation of the genetic causes of complex human diseases. For many diseases, population-based studies of unrelated individuals—in which case-control and cohort studies serve as standard designs for genetic association analysis—can be the most practical and powerful approach. However, extensive debate has arisen about optimum study design, and considerable concern has been expressed that these approaches are prone to population stratification, which can lead to biased or spurious results. Over the past decade, a great shift has been noted, away from case-control and cohort studies, towards family-based association designs. These designs have fewer problems with population stratification but have greater genotyping and sampling requirements, and data can be difficult or impossible to gather. We discuss past evidence for population stratification on genotype-phenotype association studies, review methods to detect and account for it, and present suggestions for future study design and analysis.

Prevalence of many complex human diseases such as asthma, cardiovascular disease, and diabetes has risen greatly over the past two decades in developed countries.[1,2] During the same period, the genetic causes of such diseases have been increasingly emphasised as a means to better understand their pathogenesis, with the ultimate goal of improvement of preventive strategies, diagnostic tools, and treatment.[3–8] Considerable effort is being expended in attempts to detect genetic loci contributing to complex diseases.[9] Association and linkage studies comprise the two dominant strategies: association studies aim to find disease-predisposing alleles at the population level; and linkage studies focus on familial segregation. Although both strategies have compelling strengths, association analyses are more widely done and likely to spread even further in the future, especially in the pharmacogenetics domain.[5]

Technical developments in molecular genetics facilitate these studies, as does use of gene-specific variants derived from the human genome sequencing project.[10] Furthermore, extensive catalogues of anonymous DNA sequence variants across the human genome are being compiled.[11,12] Some large-scale, population-based human samples have been, or are expected to be, gathered (eg, EPIC,[13] ISIS,[14] Million Women Study,[15] MRC/Wellcome Trust Biobank UK),[16–19] and use of DNA variants in drug development is expanding.[4] Coupling of high-throughput molecular technology, many genetic variants, and population-based samples offers unique opportunities for understanding the cause of common diseases.

Genetic variants—or polymorphisms—arise from new mutations. The simplest type of polymorphism is a single base mutation, which substitutes one nucleotide for another, referred to as a single nucleotide polymorphism (SNP). SNPs do not necessarily have any relevance to disease or outcome; they can be anonymous variants within or between genes (ie, uncharacterised with respect to protein coding or gene function), or could be functional, causal mutations. More SNPs are thought to exist in the human genome than any other type of polymorphism.[20] Nearly three million variants have been reported and are catalogued in a public database (http://www.ncbi.nlm.nih.gov/SNP/). In this review, we restrict our attention to SNPs, owing to their widespread presence and use, but the issues and principles described are general and apply to other DNA polymorphisms.

Genetic association studies aim to correlate differences in disease frequencies between groups (or in trait levels for continuously varying characters) with differences in allele frequencies at an SNP. Thus, the frequencies of the two variant forms (alleles) of an SNP are of primary interest for identification of genes affecting disease. The simplest study design for assessment of genotype-phenotype correlation is the traditional case-control approach. However, this design carries the strong assumption that any noted differences in allele frequencies actually relate to the outcome measured—ie, there are no unobserved confounding effects,[21] either directly attributable to the causal marker or through another marker that is located nearby.

Unfortunately, allele frequencies are known to vary widely within and between populations, irrespective of disease status.[22] This disparity in frequencies arises

Search strategy and selection criteria

Reference material for this review was selected on the basis of its relevance for specifically addressing the effects, outcomes, or effect of population stratification on allelic association studies. We used our own reference compilations and PubMed to identify the references cited in this work. Beyond our own material, our search terms included "population stratification", "admixture", "spurious association", "genomic control", and "pharmacogenetic association". For inclusion, recent reviews and research articles appearing in high impact journals were preferred over other sources.

Lancet 2003; **361:** 598–604

Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK (Prof L R Cardon PhD)**; Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA** (L J Palmer PhD)**; and Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH** (L J Palmer)

Correspondence to: Prof Lon Cardon
(e-mail: lon.cardon@well.ox.ac.uk)

because each population has a unique genetic and social history, and thus ancestral patterns of geographical migration, mating practices, reproductive expansions and bottlenecks, and stochastic variation all yield differences in allele frequencies between individuals,[23] yet none is necessarily associated with any particular disease. These population-frequency discrepancies are widespread throughout the genome including many genes of known medical relevance.[24,25] Consequently, the assumption of no confounding effects in genetic applications of the case-control design could be violated a priori for reasons that are at least partly outside the control of the investigator. In effect, nearly all outbred populations are confounded by genetic admixture at some level; the challenge is not merely to show that it exists, but to avoid making erroneous conclusions because of it.

When cases and controls have different allele frequencies attributable to diversity in background population, unrelated to outcome status, a study is said to have population stratification. Two circumstances must be met for population stratification to affect genetic association studies: differences in disease prevalence must exist between cases and controls; and variations in allele frequency between groups must be present.[26] Despite the misconception that any differences in allele frequency will lead to spurious association, the presence of either of these two circumstances alone is not sufficient to cause population stratification. The same type of undetected population stratification, which causes serious concern for traditional association studies, comprises the essential information for an alternative method of gene localisation—admixture mapping.[27–31] Here, we restrict our focus to undetected population stratification as a concern for association studies, rather than other mapping tools such as admixture assessment or linkage analysis.

Population stratification is probably the most often cited reason for non-replication of genetic association results, which have unfortunately been more the rule than the exception.[32–34] Leading scientific journals have noted the importance of population stratification as a cause of non-replicated association outcomes,[35] and it is usual practice in grant applications and manuscript peer-review to demand that stratification is explicitly addressed.[36] In the past decade, the potential for this effect to yield false-positive findings has led to a great shift in association study design, away from the traditional case-control approach and towards more demanding and (genotypically) less efficient family-based designs. The primary impetus for these costly design changes was to protect against false-positive inference attributable to population stratification.

However, family-based association designs are often neither practical nor plausible,[37] either for pharmacogenetic studies (ie, development of personalised drugs), population-based cohorts, or high-efficiency studies to detect genes of modest effect. However, some recent statistical and genetic developments offer promising methods to detect stratification in population samples rather than families, needing no essential changes to study design. Furthermore, there is growing recognition that population stratification might not have been as important a problem as originally believed, and has probably been a minor or even irrelevant factor for most non-replicated association studies, albeit with notable exceptions in studies of ethnically diverse samples.[9,38] Here, we consider the effects of population stratification on association studies in terms of its potential confounding effect, the empirical evidence for its occurrence, methods for its detection and control, and its relevance to pharmacogenetic applications.

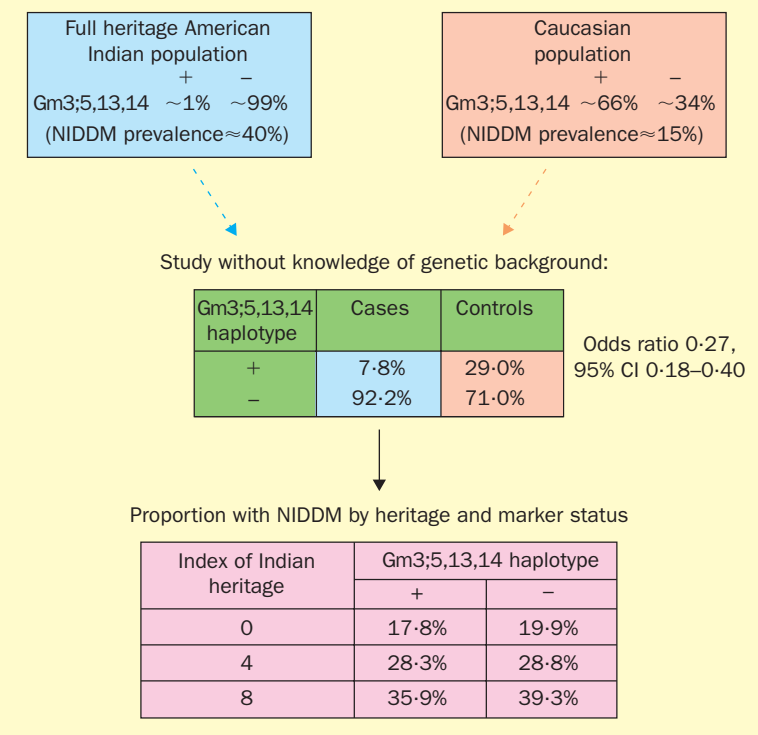
Genetic association studies

Statistical evidence for an association between an allele and a phenotype comes from one of three situations.[39] First, the allele itself might be functional and directly affect expression of the phenotype. Second, the allele might be correlated with, or be in linkage disequilibrium with, a causative allele located nearby. Third, the association could be attributable to chance or artifact—eg, confounding or selection bias.

Many study designs are available for association analyses, which can be broadly broken down into family-based designs (extended pedigrees, relative-pairs, parent-child trios, nuclear families) and non-family-based studies (case-control, cohort). Historically, case-control studies have been the workhorse of both mainstream epidemiology and genetic association studies.[40] They are recognised as being well suited for localisation of susceptibility loci,[39] and they are more powerful than family-based linkage analyses for detection of weak genetic effects.[5,41] There are several important advantages to use of a case-control design in genetic association studies: the methodology is well understood from its widespread use in epidemiology; cases and controls are convenient to enrol and offer more efficient recruitment than family-based sampling; late-onset diseases can be studied; very large samples can be gathered; disease-allele frequency, penetrance, and attributable risk can be simultaneously estimated; and unrelated controls can provide increased power over studies of genetically-related individuals (yet do not always do so).[42,43] However, although case-control genetic association studies have been widely used in attempts to identify loci that affect complex human disease, their inconsistency is a generally recognised limitation.[33,34] The absence of reproducibility is generally ascribed to inadequate statistical power, biological and phenotypic complexity, population-specific linkage disequilibrium, effect-size bias, and population stratification.[5,33,34,44,45] Undetected population stratification has caused the most concern and is an issue both for direct candidate-gene approaches and indirect association via linkage disequilibrium mapping.[46]

Because potential for spurious outcomes as a result of undetected population substructure has led to such serious concern about the case-control design, to the point of limiting its use as a standard association design, we might expect that many studies are well known to have had such substructure. In fact, few studies can be unequivocally ascribed to have yielded erroneous outcomes attributable to underlying allelic stratification.

Two examples are consistently cited as illustrative of spurious outcomes because of population stratification. The most frequently cited example comes from a study of the association between an HLA haplotype and diabetes on a Pima Indian reservation.[47] This study showed a classic case of confounding attributable to admixture of white European and Pima Indian ancestry on the association of the haplotype Gm3;5,13,14 with non-insulin-dependent diabetes mellitus (figure). The association disappeared when analysis was restricted to full-heritage Pima-Papago Indians.[47] This example, which is actually one of genetic admixture rather than of stratified subsamples of different ethnic origin, is often cited to show the perils of poor epidemiologic design because the classic conditions for confounding by ethnic origin were met—failure to control for ethnic origin introduced bias because diabetes prevalence and frequency of the haplotype of interest were both much higher in individuals of American Indian ancestry than in those of European ancestry. The second example of



Population stratification in a study of Gm3;5,13,14 haplotype in a genetically admixed sample of native Americans of the Pima and Papago tribes

Disease prevalence and allele frequencies both differ between European and native American populations; the haplotype is not found in full-blood native Americans. The noted association of the Gm3;5,13,14 haplotype with reduced risk of non-insulin-dependent diabetes mellitus (NIDDM) is attributable to ancestral population of origin rather than to linkage disequilibrium between the disease and marker loci.

population stratification includes studies of the association between alcoholism and the dopamine D2 receptor, in which both alcoholism prevalence and *DRD2* allele frequencies vary greatly by ethnic group.[48,49] In fact, there is much greater heterogeneity between these studies than between cases and controls within any one study.[48]

These two examples show that population stratification can induce important biases. However, these studies are among the only clear examples available in published work, and both are essentially examples of fundamentally flawed epidemiology rather than just poor genetic matching.[32,50] As Morton and Collins[38] have noted, if investigators adhere to basic principles of good epidemiologic design then related controls may be unnecessary.

Studies to assess population substructure with newly developed methods[51] are presently underway in many academic and industry research groups worldwide. Preliminary empirical data from large studies of different ethnic groups suggest that when the potential for sample heterogeneity is carefully addressed as part of study design, heterogeneity is usually not seen to be extensive.[52–58] These findings indicate that the extent of bias from stratification has been exaggerated.

Furthermore, even if some bias is present in a case-control study because of population stratification, the amount of bias is likely to be small apart from under extreme conditions. Wacholder and colleagues[26] have shown—empirically by cancer studies of US white populations of European origin, and theoretically with simulated data—that bias is not substantial (<1%) in case-control studies with unrelated controls unless there are major correlated differences in allele frequency and disease prevalence across ethnic groups, and the available questionnaire data on ethnic origin are not adequate to control bias. Bias will decrease as the number of ethnic strata increases; increasing the number of diverse ethnic groups and their admixture will actually decrease association bias from population stratification.[26] Further, in the presence of genuine association, the maximum potential bias will probably be associated with the effect size of the loci in question[26] and the sample size under study.[59] It has become apparent over the past decade that the common, complex human diseases of present interest—such as most cancers, asthma, and cardiovascular disease—are almost certainly under the control of many genes of modest individual effect and many non-genetic factors.[34,60]

Although population stratification seems far less of a confounding issue for population-based studies than has been proposed, for some study settings it will clearly be more troublesome: studies of groups with sparse or inaccurate knowledge of their ancestry; studies of clear recent admixture; studies of HLA alleles, whose frequencies differ strikingly by ethnic group; studies of diseases whose frequency varies strikingly by ethnic origin. Still, each of the conditions noted by Wacholder and colleagues[26] must be met for substantial bias to arise. A much bigger general problem in genetic association studies, and the probable cause of much non-replication, is the simple overinterpretation of marginal results because of absence of stringency for statistical significance.[5] Unless changes are made in accepted significance thresholds and standard practice of analysing one marker at a time, this difficulty will only worsen as the amount of available genetic marker data increases.

Concerns associated with the potential for confounding by ethnic origin have had a great effect on genetic association study design, and as a result, on funding for genetics research and published work. As discussed below, these concerns have led to development and widespread use of family-based controls as replacements for case-control studies. In view of the limited empirical support for undetected population stratification as a major cause of false positive reports, and development of these new methods of detection, it is not clear that the design changes were warranted; it is even less clear that they remain worthwhile now.

Treatment of population stratification in association studies

The problem of population stratification can be viewed essentially as one of sample matching. In general, for any well-designed epidemiological case-control study, the source population from which controls are sampled should be that from which cases are also sampled.[21] Population stratification can arise when the genetic background of the source populations differs between cases and controls.

One obvious solution to the difficulty of stratification is to carefully match cases and controls on the basis of genetic background (and study-specific environmental factors), and thereby keep irrelevant allelic differences in groups to a minimum. Inappropriate matching is a frequently cited criticism of genetic epidemiology, and

careful attention to standard epidemiological matching could help to avoid the problem entirely.[50] However, to match for all genetic differences in a population is impossible, and even concerted attempts to match on surrogate indices such as geographic proximity, physical characteristics, or self-reported family ancestry are not certain to control for unmeasured, unknown population ancestry differences. Epidemiological matching is necessary, but not always sufficient, to control for population stratification,[61] so further matching is needed. Virtually all present methods for dealing with stratification attempt to genetically match cases and controls or, in the absence of matching, at least estimate the genetic differences so that they can be addressed statistically.

Controlling for stratification with families

The most widespread study design for genetic matching includes use of relatives as controls. There are many family-based matching designs and corresponding statistical methods for discrete and continuous traits.[46,62,63] The most popular method, and that from which most others are derived, is the transmission-disequilibrium test (TDT).[64,65] The TDT design requires an affected individual and his or her parents, and uses the mendelian principle that for any polymorphic marker, each parent contributes one allele to an offspring. TDT simply involves establishment of a pseudo case-control study, in which cases are the parental alleles transmitted to the affected proband, and controls are those that were not transmitted. Protection from stratification comes from matching of each case-control pair within a family, so that any population-level allele frequency differences become irrelevant.

Concern for false positives attributable to population stratification has led to remarkably broad usage and endorsement of the TDT design in human genetics research over the past decade,[35–37] as well as considerable effort in methodological development. Theoretically and empirically, TDT controls well for population stratification; however, at least three properties of this approach are worthy of emphasis. First, for each case-control pair, DNA samples and genotypes are needed from three people (two parents and one proband). The design thus has two-thirds the genotyping efficiency of one that uses all outcome and genetic information from each participant.[66,67] Second, to provide a test of marker-outcome association, information is only conveyed when parents are heterozygous at the marker—ie, they have non-matching alleles. For an SNP, this occurrence happens at most 50% of the time, and thus, at least half the data for each parental sample are disregarded in every study. Third, by design, the TDT has a burden of family ascertainment, which is very difficult or impossible with late-onset disorders and with some psychiatric conditions; it is also impractical and potentially ethically sensitive in pharmacogenetic or clinical trials. Furthermore, statistically conditioning on irrelevant data (in this case, conditioning on parental genotypes in the absence of important stratification) will always lead to reduced power. Thus, if stratification is not of sufficient importance to affect a particular study, the study loses power. It is also unclear whether the drawback of non-replication of genetic associations is reduced for family-based designs compared with population-based tests.[44] While family-based designs offer several key strengths for genetic studies, including the potential to conduct linkage analysis, assessment of parent-of-origin differences, genotype-phase inference for haplotyping, and assessment of genotyping errors (although the TDT design is suboptimal in this respect),[68,69] their use as a means to protect against stratification in association studies comes at a high cost of genotyping and sampling efficiency. This cost will be compounded as studies begin to make use of the large and ever-increasing collections of identified SNPs.

Although there are definite situations in which familial controls will be useful or even essential, such as studies of subgroups defined by a rare allele or gene-environment interaction involving a rare allele,[42] or situations of subtle background genetic differences that arise in specific chromosomal regions rather than across the genome at large, it is generally the case that use of such controls will introduce great logistical and financial complications, potentially introduce biases attributable to overmatching, and act to reduce power to detect genetic associations.

Controlling for stratification with anonymous genetic markers

There are several methods that protect against population stratification-related drawbacks but do not need family samples. Pritchard and Rosenberg[51] popularised the notion of using anonymous genetic markers scattered throughout the genome as indicators of the amount of background diversity in cases and controls. They reasoned that as long as the markers were independent of those affecting the disease of interest, and largely did not correlate with each other, they should reflect baseline genetic differences between cases and controls. In this way, the background level of population differences can be formally quantified and tested, and in the absence of any differences, the case-control study then proceeds. As few as 30 SNPs were proposed as sufficient to detect population substructure. In view of present molecular technology and decreasing assay costs, this method is cost effective and feasible for all but the very largest population collections.

Although the Pritchard and Rosenberg approach offers promise for detection of underlying population differences, it does not offer a means to proceed if such differences are indeed detected. At least two solutions to this difficulty have been proposed. One, termed genomic control,[70-74] suggests that in the presence of population substructure, the standard χ^2 statistic used in case-control studies is inflated by a multiplicative factor, which is proportional to the degree of stratification. This multiplicative factor can be estimated with the set of unlinked genetic markers scattered around the genome. It can then be incorporated into the disease-marker association tests (by rescaling the χ^2 test statistic) to correct for background population differences.[70,75] Because power to detect stratification rises with sample size, modest population differences are more pronounced in large than in small samples.[59]

Another approach, termed structure assessment,[59,76-78] also uses unlinked genetic markers to detect stratification, but instead of estimating a scaling factor, it attempts to define underlying subgroups in stratified samples on the basis of the set of genomic markers. Subdivision into homogeneous population groups means that subgroups can be matched appropriately and the disease-marker association test done in each matched subgroup. The total test for disease association is then a statistical combination of results from each component subgroup. A strength of this particular approach is that it makes use of existing markers that are known to differ in frequency between ethnically diverse samples and thereby keeps genotyping efficiency to a maximum.

The theoretical basis for these methods is becoming increasingly well-developed,[59,72] and each of them has its respective merits—eg, genomic control is easy to implement and flexible in accommodating different phenotypes and DNA samples, but can be conservative and thus lose power in some applications. Structure assessment has a strong population genetics basis and allows for the effects of natural selection, but depends critically on clear detection of substructure for accurate false positive rates. The specific attributes of these methods are likely to change as they are further refined, though they already seem adequate to detect large frequency differences.[61,79] Although it remains unclear how many markers or samples are needed to detect subtle population differences,[26] or indeed, whether or when subtle differences are important factors in association studies,[80] availability of the methods coincides very well with the widescale availability of genetic markers and the means to assay them rapidly and efficiently. These methods for stratification detection offer promising and practical alternatives to the use of family-based controls for allelic association assessment.

All extant stratification tests and designs, family-based and genome-based, have been developed with the express aim of keeping false positive results from population stratification to a minimum. However, population stratification is rarely noted not only to yield false evidence in favour of association but also to mask real effects.[81,82]

Implications for pharmacogenetic studies

An expanding area of interest in application of SNPs to investigations of disease pathophysiology is stratification of populations by their genetically determined response to therapeutic drugs (pharmacogenetics).[83] Ideally, one would like to be able to stratify a population needing treatment into those likely, or unlikely, to respond to treatment and those likely, or unlikely, to have adverse side-effects. One of the primary goals of pharmacogenetics is to understand the role that sequence variation in individuals and populations has in variability of responses to drugs, with the aims of improvement of the efficacy of drug-based interventions and expediting of targeted drug discovery and development. Pharmacogenetic initiatives are presently an area of very active research in complex human diseases.[84–89] Confirmation of initial findings could mark the beginning of clinical use of genotyping at an individual level as an adjunct to pharmacotherapy for many diseases.

The issue of population stratification is of particular relevance to pharmacogenetic studies, because these are, almost without exception, case-control studies. For many diseases, the response to pharmacotherapy varies with age, and the number and type of drugs is changing rapidly. There are also important ethical and legal issues associated with collection of family data in a clinical trial setting. These considerations effectively preclude availability of family-based treatment data in the foreseeable future for most complex human diseases.[86] In the absence of these data, case-control or cohort association studies are the approaches of choice. Furthermore, most clinical trials are presently undertaken in highly admixed and heterogeneous populations from Europe or North America, increasing the chance that population substructure might be present. Assessments of selected pharmacogenetic loci have already indicated significant population variability in allele frequencies.[24,25,61]

As discussed above, several methods have been developed to assess population stratification and, if necessary, to correct an association test for the presence of such stratification in population-based samples. However, neither systematic testing for population stratification nor application of these new statistical methods has yet been incorporated into published pharmacogenetic studies. Empirical assessment of any potential biases attributable to substructure will be an important consideration for future pharmacogenetic studies in admixed populations.[90]

An important issue that must be considered in assessment of possible substructure is the size of the detectable effect. Although extensive stratification leading to large biases in a pharmacogenetic association analysis are likely to be detectable with as few as 30–100 random SNPs typed across the genome,[51] it is not clear what level of detectable stratification will be important with respect to bias. Knowing how much population substructure is too much will partly depend on the disease or drug target. From a clinical viewpoint, one might reasonably expect that there will be more flexibility in diseases such as colorectal cancer, for which even marginal increases in personalised medicine could have a large effect.[91] For chronic disorders with great adverse response profiles to pharmacotherapies, however, a more conservative approach might be needed, and a lower level of stratification defined as acceptable.

Conclusions

Failure to replicate genetic association studies is a genuine concern,[9,34,44] yet more often it involves poor study design and execution—in particular an absence of appreciation for the sample sizes needed to detect modest genetic effects and overinterpretation of marginal results—than undetected population stratification. For most complex human diseases, the reality of multiple disease-predisposing genes of modest individual effect, gene-gene interactions, gene-environment interactions, interpopulation heterogeneity of genetic and environmental determinants of disease, and the concomitant low statistical power mean that both initial detection and replication will probably be very difficult.[9,34,37] Add to these concerns the issues of multiple testing, laboratory and other measurement error, and positive publication and investigator-reporting biases, and it becomes apparent that population stratification is one of many possible reasons for non-replication of association results. We must re-emphasise that, when substantial population substructure does exist, promising methods for detection and correction for it are now available.[59,72–78,92] Growing empirical and theoretical evidence suggests that well-designed, well-conducted, and appropriately analysed and interpreted population-based studies with unrelated controls are largely robust to bias from population stratification. Available data indicate that the most practical and efficient way to acquire large enough sample sizes to map the genetic determinants of many complex human diseases by allelic association is by collection of very large case-control or cohort samples.

We anticipate that use of genomic controls in genetic association studies will become widespread, and could pave the way for genome-wide association studies,[5] which has positive and negative aspects. Adoption of genomic controls is especially important at this time, since we stand at the threshold of availability of several very large cohort opportunities in North America and Europe. For example the Biobank UK project, the joint MRC and Wellcome Trust initiative, will contain phenotypic information from 500 000 representatives.[16–19] One use of this resource could involve genotyping the entire cohort with a small sample of stratification-designed markers for later subselection of genetically matched controls against

each new case sample. In this way, genotyping the cohort once offers reusable matching for future studies without additional cohort genotyping. Another potential use of genomic controls is in the area of retrospective clarification—genomic control could allow follow-up of past findings that were promising but not replicated by other groups to see if population substructure was responsible for the initial evidence for association, or indeed, if it was accountable for masking the effect in the replication sample.

In hindsight, the fear of population stratification has probably been exaggerated. The pervasive reliance and even insistence on family-based association studies to protect against stratification has limited the breadth and depth of appropriate study samples, and has reduced power to detect true associations. A great deal of research effort seems to have been compromised to protect against a confounding factor that never realised its potential to bias allelic association in complex traits. Although many hurdles to detection of genes affecting common diseases remain, important genes could have been missed because of excessive fear of population stratification.

Conflict of interest statement
None declared.

Acknowledgments
This work was supported in part by a Wellcome Trust Principal Research Fellowship and via U01 HL65899 from the National Heart, Lung, and Blood Institute of the NIH. The sponsors had no role in study design, data collection, data analysis, data interpretation, or writing of the report.

References

1 Wilson PW. Diabetes mellitus and coronary heart disease. *Endocrinol Metab Clin North Am* 2001; **30:** 857–81.
2 Weiss KB, Sullivan SD. The health economics of asthma and rhinitis: I—assessing the economic impact. *J Allergy Clin Immunol* 2001; **107:** 3–8.
3 Drews J, Ryser S. The role of innovation in drug development. *Nat Biotechnol* 1997; **15:** 1318–19.
4 Roses AD. Pharmacogenetics and the practice of medicine. *Nature* 2000; **405:** 857–65.
5 Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; **405:** 847–56.
6 Palmer LJ, Cookson WOCM. Atopy and asthma. In: Sham PC, Bishop T, eds. Genetic analysis of multifactorial diseases. London: BIOS Scientific Publishers, 2000: 215–37.
7 Palmer LJ, Cookson WOCM. Genomic approaches to understanding asthma. *Genome Res* 2000; **10:** 1280–87.
8 McCarthy MI. Susceptibility gene discovery for common metabolic and endocrine traits. *J Mol Endocrinol* 2002; **28:** 1–17.
9 Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001; **2:** 91–99.
10 Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. *Nature* 2001; **409:** 860–921.
11 Collins FS, Guyer MS, Charkravarti A. Variations on a theme: cataloging human DNA sequence variation. *Science* 1997; **278:** 1580–81.
12 Sachidanandam R, Weissman D, Schmidt SC, et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001; **409:** 928–33.
13 Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol* 1992; **3:** 783–91.
14 ISIS-1 (First International Study of Infarct Survival Collaborative Group). Randomised trial of intravenous atenolol among 16 027 cases of suspected acute myocardial infarction. *Lancet* 1986; **2:** 57–66.
15 The Million Women Study. Design and characteristics of the study population. *Breast Cancer Res* 1999; **1:** 73–80.
16 Clayton D, McKeigue PM. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* 2001; **358:** 1356–60.
17 Radda G, Dexter TM, Meade T. The need for independent scientific peer review of Biobank UK. *Lancet* 2002; **359:** 2282.
18 Wallace H. The need for independent scientific peer review of Biobank UK. *Lancet* 2002; **359:** 2282.
19 Wright AF, Carothers AD, Campbell H. Gene-environment interactions: the BioBank UK study. *Pharm J* 2002; **2:** 75–82.
20 Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 1998; **280:** 1077–82.
21 Schlesselman JJ. Case-control studies: design, conduct, analysis. Oxford: Oxford University Press, 1982.
22 Cavalli-Sforza LL, Menozzi P, Piazza A. History and geography of human genes. Princeton: Princeton University Press, 1994.
23 Slatkin M. Inbreeding coefficients and coalescence times. *Genet Res* 1991; **58:** 167–75.
24 Goddard KA, Hopkins PJ, Hall JM, Witte JS. Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000; **66:** 216–34.
25 Stephens JC, Schneider JA, Tanguay DA, et al. Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293:** 489–93.
26 Wacholder S, Rothman N, Caporaso N. Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. *J Natl Cancer Inst* 2000; **92:** 1151–58.
27 Chakraborty R, Weiss KM. Admixture as a tool for finding linked genes and detecting that difference from allelic association between loci. *Proc Natl Acad Sci USA* 1988; **85:** 9119–23.
28 Stephens JC, Briscoe D, O'Brien SJ. Mapping by admixture linkage disequilibrium in human populations: limits and guidelines. *Am J Hum Genet* 1994; **55:** 809–24.
29 McKeigue PM. Mapping genes underlying ethnic differences in disease risk by linkage disequilibrium in recently admixed populations. *Am J Hum Genet* 1997; **60:** 188–96.
30 McKeigue PM. Mapping genes that underlie ethnic differences in disease risk: methods for detecting linkage in admixed populations, by conditioning on parental admixture. *Am J Hum Genet* 1998; **63:** 241–51.
31 McKeigue PM, Carpenter JR, Parra EJ, Shriver MD. Estimation of admixture and detection of linkage in admixed populations by a Bayesian approach: application to African-American populations. *Ann Hum Genet* 2000; **64:** 171–86.
32 Tabor HK, Risch NJ. Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; **3:** 391–97.
33 Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000; **26:** 151–57.
34 Terwilliger JD, Goring HH. Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000; **72:** 63–132.
35 Anon. Freely associating. *Nat Genet* 1999; **22:** 1–2.
36 Gauderman WJ, Witte JS, Thomas DC. Family-based association studies. *J Natl Cancer Inst Monogr* 1999; **26:** 31–37.
37 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273:** 1516–17.
38 Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci USA* 1998; **95:** 11389–93.
39 Silverman EK, Palmer LJ. Case-control association studies for the genetics of complex respiratory diseases. *Am J Respir Cell Mol Biol* 2000; **22:** 645–48.
40 Wacholder S, Hartge P, Palmer LJ. Case-control study. In: Elston RC, Olson JM, Palmer LJ, eds. Biostatistical genetics and genetic epidemiology. London. Chichester: Wiley, 2002.
41 Risch N, Teng J. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases, I: DNA pooling. *Genome Res* 1998; **8:** 1273–88.
42 Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999; **149:** 693–705.
43 Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases, II: individual genotyping. *Genome Res* 1999; **9:** 234–41.
44 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:** 306–09.
45 Goring HH, Terwilliger JD, Blangero J. Large upward bias in estimation of locus-specific effects from genomewide scans. *Am J Hum Genet* 2001; **69:** 1357–69.
46 Jorde LB. Linkage disequilibrium and the search for complex disease genes. *Genome Res* 2000; **10:** 1435–44.
47 Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet* 1988; **43:** 520–26.

48 Gelernter J, Goldman D, Risch N. The A1 allele at the D2 dopamine receptor gene and alcoholism: a reappraisal. *JAMA* 1993; **269:** 1673–77.

49 Pato CN, Macciardi F, Pato MT, Verga M, Kennedy JL. Review of the putative association of dopamine D2 receptor and alcoholism: a meta-analysis. *Am J Med Genet* 1993; **48:** 78–82.

50 Zondervan KT, Cardon LR, Kennedy SH. What makes a good case-control study? Design issues for complex traits such as endometriosis. *Hum Reprod* 2002; **17:** 1415–23.

51 Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65:** 220–28.

52 Ardlie KG, Lunetta KL, Seielstad M. Testing for population subdivision and association in four case-control studies. *Am J Hum Genet* 2002; **71:** 304–11.

53 Martin NC, Broxholme SJ, Cardon LR. Assessment of stratification in linkage and association samples. *Am J Hum Genet* 2001; **69** (suppl)**:** 419.

54 Pankow JS, Province MA, Hunt SC, Arnett DK. Little evidence for bias due to population stratification in a population-based study of non-Hispanic Blacks and Whites. *Am J Hum Genet* 2001; **69** (suppl)**:** 412.

55 Cutler DJ, Zwick ME, Yohn CT, et al. Human population substructure and its influence on association studies. *Am J Hum Genet* 2001; **69** (suppl)**:** 181.

56 Carvajal-Carmona LG, Ophoff R, Service S, et al. Genetic diversity and population structure of two historically related Latin-American populations. *Am J Hum Genet* 2001; **69** (suppl)**:** 418.

57 Yamada R, Kawakami H, Yamaguchi M, et al. Analysis of population structure of Japanese population using genotype data of genome wide hundreds of SNPs. *Am J Hum Genet* 2001; **69** (suppl)**:** 423.

58 Hill L, Jawaid A, Sham P, Plomin R. DNA pooling, genome control and association studies. *Am J Hum Genet* 2001; **69** (suppl)**:** 423.

59 Pritchard JK, Donnelly P. Case-control studies of association in structured or admixed populations. *Theor Popul Biol* 2001; **60:** 227–37.

60 Weeks D, Lathrop G. Polygenic disease: methods for mapping complex disease traits. *Trends Genet* 1995; **11:** 513–19.

61 Wilson JF, Weale ME, Smith AC, et al. Population genetic structure of variable drug response. *Nat Genet* 2001; **29:** 265–69.

62 Schulze TG, McMahon FJ. Genetic association mapping at the crossroads: which test and why? Overview and practical guidelines. *Am J Med Genet* 2002; **114:** 1–11.

63 Abecasis GR, Cookson WO, Cardon LR. Pedigree tests of transmission disequilibrium. *Eur J Hum Genet* 2000; **8:** 545–51.

64 Spielman R, McGinnis R, Ewens W. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993; **52:** 506–16.

65 Ewens WJ, Spielman RS. The transmission/disequilibrium test: history, subdivision, and admixture. *Am J Hum Genet* 1995; **57:** 455–64.

66 McGinnis R, Shifman S, Darvasi A. Power and efficiency of the TDT and case-control design for association scans. *Behav Genet* 2002; **32:** 135–44.

67 Nielsen DM, Weir BS. Association studies under general disease models. *Theor Popul Biol* 2001; **60:** 253–63.

68 Gordon D, Heath SC, Liu X, Ott J. A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am J Hum Genet* 2001; **69:** 371–80.

69 Kirk KM, Cardon LR. The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet* 2002; **10:** 616–22.

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

71 Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; **66:** 1933–44.

72 Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol* 2001; **60:** 155–66.

73 Devlin B, Roeder K, Bacanu SA. Unbiased methods for population-based association studies. *Genet Epidemiol* 2001; **21:** 273–84.

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

75 Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001; **20:** 4–16.

76 Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. *Genetics* 2000; **155:** 945–59.

77 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000; **67:** 170–81.

78 Satten GA, Flanders WD, Yang Q. Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am J Hum Genet* 2001; **68:** 466–77.

79 Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES IV. Dwarf8 polymorphisms associate with variation in flowering time. *Nat Genet* 2001; **28:** 286–89.

80 Jiang R, Dong J. When is the bias caused by population stratification negligible? *Am J Hum Genet* 2001; **69:** 409.

81 Deng HW. Population admixture may appear to mask, change or reverse genetic effects of genes underlying complex traits. *Genetics* 2001; **159:** 1319–23.

82 Cardon LR. Practical barriers to mapping complex trait loci. In: Plomin R, DeFries JC, Craig I, Crabbe J, eds. The future of behavioural genetics in the postgenomics era. New York: APA Books, 2002.

83 Roses AD. Pharmacogenetics and future drug development and delivery. *Lancet* 2000; **355:** 1358–61.

84 Ball S, Borman N. Pharmacogenetics and drug metabolism. *Nat Biotechnol* 1997; **15:** 925–26.

85 Poolsup N, Li Wan Po A, Knight TL. Pharmacogenetics and psychopharmacotherapy. *J Clin Pharm Ther* 2000; **25:** 197–220.

86 McCarthy JJ, Hilfiker R. The use of single-nucleotide polymorphism maps in pharmacogenomics. *Nat Biotechnol* 2000; **18:** 505–08.

87 March R. Pharmacogenomics: the genomics of drug response. *Yeast* 2000; **17:** 16–21.

88 Palmer LJ, Silverman ES, Weiss ST, Drazen JM. Pharmacogenetics of asthma. *Am J Respir Crit Care Med* 2002; **165:** 861–66.

89 Pettipher R, Cardon LR. The application of genetics to the discovery of better medicines. *Pharmacogenomics* 2002; **3:** 257–63.

90 McLeod HL. Pharmacogenetics: more than skin deep. *Nat Genet* 2001; **29:** 247–48.

91 Walker J, Quirke P. Biology and genetics of colorectal cancer. *Eur J Cancer* 2001; **37** (suppl 7)**:** S163–73.

92 Overall AD, Nichols RA. A method for distinguishing consanguinity and population substructure using multilocus genotype data. *Mol Biol Evol* 2001; **18:** 2048–56.