European Journal of Human Genetics (2005) 13, 79–85

www.nature.com/ejhg

ARTICLE

Overdispersion of allele frequency differences between populations: implications for meta-analyses of genotypic disease associations

Cliona M Molony[1,2,4], Anthony P Fitzgerald[1,3,4] and Denis C Shields*[,1,2]

[1]*Department of Clinical Pharmacology, Royal College of Surgeons in Ireland, Dublin, Ireland;* [2]*Institute of Biopharmaceutical Sciences, Royal College of Surgeons in Ireland, Dublin, Ireland;* [3]*The Haughton Institute, St James's Hospital, Dublin, Ireland*

Methods correcting case–control studies of genetic polymorphisms for unmeasured genetic population substructure by modelling the variation at a number of variant loci provide no standard and easily implemented approach to meta-analysis, which is a key to understanding the effects of minor genotypic risks on complex diseases. A correction of the odds ratio estimate and its confidence interval is shown to be easy to implement using a mixed effects logistic regression. The method is shown to substantially reduce bias and to give accurate coverage even when there is substantial overdispersion of allele frequency differences between populations. Major sequence classes of single-nucleotide polymorphism (SNP) are likely to act as valid controls for each other, since CpG SNPs did not differ in the extent of population structure from other SNPs. Agreement among investigators and journals to provide these straightforward statistics in publications of polymorphism studies will enhance the ability of future investigators to perform meta-analyses of weak genetic effects across accumulated studies that allow for population structure.
European Journal of Human Genetics (2005) **13,** 79–85. doi:10.1038/sj.ejhg.5201275
Published online 6 October 2004

Keywords: association; stratification; meta-analysis; polymorphism; CpG

Introduction

Complex genetic diseases are those with multiple environmental and genetic components contributing to the cumulative risk. Case–control association studies of genetic polymorphisms in complex genetic disease are in general difficult to replicate.[1] This may in part reflect the low risks conferred by candidate genes coupled with potential publication bias of small studies. However, it is important to eliminate the possibility that some of the contribution to the heterogeneity in the findings relates to minor confounding of differences among cases and controls with genetic substructure (ie the cases are drawn from a slightly different genetic background from the controls). This can arise when disease incidence (for genetic or nongenetic reasons) varies among genetic grouping, and the noncausal association of a chromosomal region with disease is simply part of a larger trend in allele frequency differences between genetic groupings.

There are a number of methods available to correct estimates of test marker association by using information on genetic substructure obtained from the frequencies of a number of other markers. These fall into two broad categories.[2] The first are those based on a 'Genomic Control' approach[3,4] that provides an appropriate reduction of type I error by adjusting for the level of difference between affected and unaffected individuals at the other markers. The second rely on categorising individuals into genetic substrata, and estimating risk effects across these strata.[2,5–9]

*Correspondence: Dr D Shields, Department of Clinical Pharmacology, Royal College of Surgeons in Ireland, Dublin 2, Ireland. Tel: +353 1 4022381; Fax: +353 1 4022388; E-mail: dshields@rcsi.ie
[4]These authors contributed equally to this work
Received 30 January 2004; revised 29 June 2004; accepted 23 July 2004

Standard epidemiological meta-analysis of associations[10] usually relies on an estimate of association of outcome with the risk factor in each study, often expressed as an odds ratio (OR) in case–control studies and as a relative risk in prospective studies, and an estimate of the variance of that association, usually presented as a 95% confidence interval.[11] Methods correcting case–control studies of genetic polymorphisms[4–7] for unmeasured genetic population substructure by modelling the variation at a number of variant loci provide no standard and easily implemented approach to meta-analysis, which is a key to understanding the effects of minor genotypic risks on complex diseases. Often, they are confined to hypothesis testing and do not directly generate estimates of the OR and confidence interval: the Genomic Control method[3] and the χ^2-scaling method[4] simply estimate *P*-values. An alternative approach is to estimate genetic strata within the population from the genotypic data and then perform a stratified analysis.[5,6] This has the disadvantage that strata and effects are not simultaneously estimated,[7] although this bias may not be great.[2] Secondly, a mild bias in a number of studies will not be detected as a significant stratification in individual studies, but such minor biases may accumulate within a meta-analysis.

Satten *et al*[7] proposed a latent model that considers stratification and tests of gene effects simultaneously, and generates estimates of ORs and confidence intervals. This method is not currently implemented within a standard statistical package. A Markov chain Monte Carlo approach also simultaneously models structure while assessing the risk conferred by a test gene, and the authors point out that probabilities of association may be readily combined across studies in a meta-analysis, although the estimate of the extent of risk, such as the OR, is not generated by these analyses.[8] In contrast, logistic regression analyses are ideal for simple extensions considering covariates.[12] We show here that they can be easily extended to encompass analyses that allow for genetic stratification in a simple and straightforward manner, and investigate the impact of overdispersion (OD) in allele frequency differences among strata on the alternative methods.

Methods

Data were simulated in order to determine the performance of alternative methods in recovering the true OR. The simulations were performed in a manner reflecting the typical genetic analysis of common disease: considering a set of possible test markers drawn from a distribution of various allele frequencies, and a random set of unlinked markers drawn, which have a similarly varying set of allele frequencies. In order to reflect a model of the underlying process of ascertainment of test markers, markers were simulated in two strata having no effect on disease, and a second ascertainment step enriched for cases carrying the risk allele at a level appropriate for a particular OR value. We simulated two strata, A and B, both having the same underlying OR but differing in disease prevalence (0.05 and 0.2, respectively), reflecting the differences in disease incidence seen for common diseases such as hypertension in different major ethnic groups.[13] In all, 30 markers were simulated with a mean frequency difference of 10% in the two populations. For the first population, allele frequencies were drawn from normal distributions using the population-specific mean (0.55) and SD (0.03). This mean and standard deviation were the values observed in a real set of markers reported in a study[14] of 114 single-nucleotide polymorphisms (SNPs) in five genetically divergent populations. For the second population, the allele frequency was simulated as that population mean plus the mean allele frequency difference, where the mean allele frequency difference was drawn from a normal distribution using the sample mean allele frequency difference and SD. From these markers, for each simulation one was randomly chosen to be the test marker. Separate simulations were performed with OR set at 1.0, 1.1, 1.4 and 1.7. Expected frequencies for the allele of interest were calculated for a population of 500 cases and 500 controls sampled from subjects assigned carrier status based on disease risk and the allele frequency attributed a particular population. Samples were randomly drawn assuming a given probability of the marker frequency in each stratum. The remaining 29 markers were then sampled assuming an underlying OR of 1.0 in each subpopulation. For each value of test marker OR, a separate simulation was performed with a different level of OD of allele frequency differences. In simulation OD1, the allele frequency differences were identical (0.10) in all markers (ie no OD). This case is clearly not realistic and is presented for illustrative purposes only. In simulations OD2–OD5, there was an increasing variability of the underlying allele frequency differences (standard deviation of allele frequency differences of 0.01, 0.02, 0.03 and 0.04, respectively) for the two subpopulations, introducing increased OD. Each simulation was performed 1000 times. The advantage of this simulation approach is that the underlying population structure for both test and random genes is the same, while the imposed enrichment of the test allele among cases directly reflects the kind of enrichment that results from sampling of cases. Four sets of simulations were considered with different degrees of population substructure (between 7 and 15% mean differences in allele frequencies between strata).

The simulation approach we adopted here was somewhat simplistic, manually specifying mean allele frequency differences between the two populations, and then increasing the dispersion of allele frequency differences. Alternative simulations based on normal-binomial or beta-binomial distributions could provide a better fit to real

data;[15] however, these models have not been extended to incorporate a parameter representing the extent of dispersion of allele frequency differences, which we wished to investigate here.

We fitted a mixed effects logistic analysis[16,17] in which the log of the OR varied randomly between markers with a mean and variance that we estimated from an analysis of the random markers. The bias due to hidden stratification was approximated by the estimated mean. When looking at the test marker, we used a fixed effect logistic regression but used the bias correction to adjust the estimate of the log(OR). To allow for OD among test markers, the variance of this bias-corrected estimate was increased by the estimated between-marker variation in log(OR). We calculated 95% confidence intervals for the log(OR) using the 1.96 multiplier. In our practical implementation of this, we arbitrarily assigned the allele that was more frequent among the cases as the risk-conferring allele, yielding estimates of OR equal to or exceeding 1 in all cases. The data for each of the random markers were analysed separately using an ordinary logistic regression, yielding estimates of the marker-specific log-odds β_i and its variance:

$$V_i = V\mathrm{ar}(\hat{\beta}_i)$$

Results were combined to estimate

$$\hat{\beta}_{\mathrm{random}} = \sum \hat{\beta}_i/n$$

and

$$V\mathrm{ar}(\hat{\beta}_{\mathrm{random}}) \text{ as } \sum V_i/n^2$$

A logistic regression of the test marker ignoring information from the n random markers was used to estimate the logit coefficient β_{crude} and its variance V_{crude}. The corrected logit coefficient was estimated as

$$\hat{\beta}_{\mathrm{adjust}} = \hat{\beta}_{\mathrm{crude}} - \hat{\beta}_{\mathrm{random}}$$

Ignoring $\mathrm{OD}_{\mathrm{nod}}$, the confidence interval of the adjusted OR is equivalent to that of the crude OR, inflated by the variance of the bias estimated from the random genes.

$$\mathrm{CI}_{\mathrm{nod}} = \exp\left[\hat{\beta}_{\mathrm{adjust}} \pm Z_{\alpha/2} * V_{\beta\,\mathrm{nod}}\right]$$

where

$$V_{\beta\,\mathrm{nod}} = V_{\beta\,\mathrm{crude}} + \sum V_i/n^2.$$

The confidence interval of the OR allowing for the assumption that the test marker is sampled from an OD distribution ($\mathrm{CI}_{\mathrm{od}}$) is estimated as

$$\mathrm{CI}_{\mathrm{od}} = \exp[\hat{\beta}_{\mathrm{adjust}} \pm Z_{\alpha/2} * V_{\beta\,\mathrm{adjust}}]$$

where

$$V_{\beta\,\mathrm{adjust}^{=}} = V_{\beta\,\mathrm{crude}} + \sum(\hat{\beta}_{\mathrm{random}} - \hat{\beta}_i)^2/(n-1) - \sum V_i/n + \sum V_i/n^2$$

The above models are equivalent to a mixed effects logistic regression model,[17] where gene status (fixed or random) has a fixed effect, and carrier status represents the random effect.

The Reich and Goldstein χ^2-scaling method[4] was calculated, as well as the frequentist implementation of the Genomic Control approach.[3] The two-step stratification approach of structured association method by Pritchard and Donnelly[5,6] was also attempted for a subset of simulations. Analysis used the general-purpose statistics package, STATA 8.2 (Statacorp, College Station, TX, USA; Stata Corporation, 2003). We considered the impact of the number of markers used in the correction on the inferences (15, 20 or 29 random markers). STATA code used to generate simulations and perform the analyses, along with the simulated data sets, are available on request from the authors.

Results

As expected (data not shown), the $\mathrm{OR}_{\mathrm{crude}}$ was linearly proportional to the true OR used in the simulations. When the true OR of the test marker is 1.0, there is a clear problem with the uncorrected $\mathrm{OR}_{\mathrm{crude}}$, with a rejection rate of 38% rising to a rate of 57% when there is a higher degree of simulated OD (Table 1). The five simulation conditions (OD1–OD5) represent differing degrees of OD of allele frequency differences, with OD1 representing a constant difference across all markers, and with OD3 most closely

Table 1 % Coverage of null hypothesis at $P \leq 0.05$ over 1000 simulations when the true OR is 1.0, with correction for stratification using 29 random markers

Simulation	*SD of allele frequency differences*	*OR*[a]*crude*	*OR*[a]*adjusted*	*Coverage of true hypothesis*				
				CI crude	*CI nods*	*CI ods*	*χ^2-scaling (ref.)*	*Genomic Control (ref.)*
OD1	0.00	1.16	1.00	62	95	93	100	94
OD2	0.01	1.17	1.00	62	90	95	99	94
OD3	0.02	1.17	1.00	59	77	94	98	94
OD4	0.03	1.19	1.01	54	63	95	97	93
OD5	0.04	1.23	1.00	43	45	93	97	94

[a]Geometric mean.

resembling the observed variation in allele frequency differences seen in Goddard *et al*.[14] Three of the methods (OR_{adjust}, χ^2-scaling and Genomic Control) that correct for population structure (using information on the random markers) give a reasonable coverage of the true hypothesis compared to the uncorrected analysis (Table 1). The χ^2-scaling method appears generally more conservative than the other two tests, which is to be expected.[4] χ^2-scaling has a slightly reduced coverage of the true hypothesis with an increase in the variance of allele frequency differences between populations. The Genomic Control method performs well regardless of the level of OD, as does the OR_{adjust} method, suggesting that both methods are reasonably robust in the presence of OD (Table 1). Taking OD3 as a realistic level of OD, OR_{adjust} has good coverage of the true hypothesis (Table 1) and minimal bias: for a true OR = 1.0, 1.1, 1.4 and 1.7, the estimated OR_{adjust} values are 1.00, 1.10, 1.40 and 1.70, whereas for the uncorrected values OR_{crude} are 1.17, 1.29, 1.63 and 1.98. For simulations at higher ORs, there is an adequate coverage of the true OR under different levels of OD, with the general trends seen in Table 1 being seen again (Table 2 and Figure 1). While the Genomic Control method may also be providing adequate coverage, this cannot be directly assessed since it does not estimate confidence intervals.

Simply correcting for effects of random markers without modelling OD (CI_{nod}) noticeably underestimated the variance in all situations except the trivial case (OD1) where allele frequency differences between populations were the same (0.10) for all markers. Thus, in simulation OD3 (Tables 1 and 2) the coverage drops from 95% (CI_{od}) to around 80% (CI_{nod}). This provides a simple illustration of the need to allow for OD in the variance in allele frequency differences when estimating the variance of the adjusted OR.

We investigated whether performance was sensitive to the number of random markers chosen. As the number of random markers falls, estimates of the between-marker variation in the log(ORs) become less precise leading to poorer coverage (Table 3). However, it is possible to improve the weakened coverage by replacing the 1.96 multiplier with the 97.5th upper percentile of a *t*-distribution with appropriate degrees of freedom (eg 2.26 for 10 random markers with nine degrees of freedom, yielding 94% coverage). For most purposes, the number of random genes is likely to be large enough so that this modification is not necessary. When less than 30 markers are used, the consequent increase in the confidence interval illustrates in part the value of choosing a large enough number of random markers. While within the situation that we simulated around 30 markers appears sufficient, it is likely that the more complex the pattern of stratification, the greater the number of markers that will be required, and recent authors have suggested at least 65 random markers.[18,19]

We also investigated whether the correction worked well when different levels of stratification were simulated. Table 4 illustrates a number of alternative scenarios that were achieved by modifying the mean allele frequency differences and disease incidence. Even when the stratification is increased, the method still appears to provide an unbiased estimate of the OR and a reasonable coverage.

It is possible that certain classes of SNP may display greater allele frequency differences between populations,

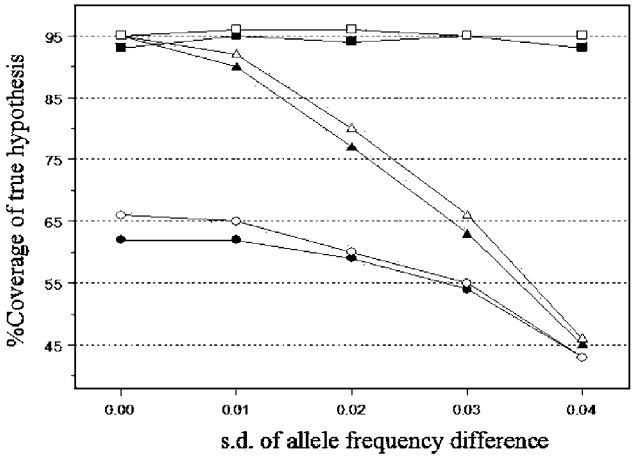

Figure 1 % Coverage of the true hypothesis by the 95% confidence intervals for estimates of the OR. Circles: crude OR; triangles: adjusted OR, without modelling OD in estimates of confidence intervals; squares: adjusted OR, allowing for OD. Results for simulations of true OR = 1.0 (solid symbols) and of true OR = 1.7 (white symbols).

Table 2 % Coverage of true hypothesis (that OR = 1.4) at $P \leq 0.05$ over 1000 simulations when the true OR is 1.4, with correction for stratification using 29 random markers

Simulation	*SD of allele frequency differences*	*OR*[a] *crude*	*OR*[a] *adjusted*	*Coverage of true hypothesis*		
				CI crude	*CI nod*	*CI od*
OD1	0.00	1.62	1.40	64	95	94
OD2	0.01	1.63	1.40	63	91	95
OD3	0.02	1.63	1.40	60	80	95
OD4	0.03	1.65	1.40	55	65	95
OD5	0.04	1.69	1.38	42	46	94

[a]Geometric mean.

Table 3 Impact of reducing the number of random markers on the estimation and coverage of OR (simulated OR = 1.0)

Simulation	*Number of random markers considered*	*OR^a_{crude}*	*$OR^a_{adjusted}$*	*% Coverage of true hypothesis CI_{od}*
OD3	29	1.17	1.00	94
OD3	20	1.19	1.00	93
OD3	15	1.20	1.02	91
OD3	10	1.18	1.00	91

[a]Geometric mean.

Table 4 Impact of different levels of stratification on the estimation and coverage of adjusted OR (simulated OR = 1.0).

Simulation level	*Allele frequency difference*		*F_{st}*	*OR^a_{crude}*	*$OR^a_{adjusted}$*	*% Coverage of true hypothesis CI_{od}*
	Mean	*SD*				
OD3	0.151	0.02	0.040	1.290	0.999	94.1
OD3	0.110	0.02	0.022	1.180	0.998	94.2
OD3	0.074	0.02	0.011	1.020	0.998	93.4
OD3	0.071	0.02	0.009	1.057	0.998	93.4

[a]Geometric mean.

reflecting differences in the dynamics of their mutation and selection constraints over the evolutionary history of populations. Allowing for any such identified heterogeneity could permit matching of an SNP to control genes with similar population structure dynamics. We looked at one class of SNPs: those involving a CpG dinucleotide polymorphism. CpG dinucleotides are hot-spots for mutational change,[20] and therefore it is possible that C/T variant SNPs followed by a G may show differing levels of stratification, given the likely differences in their population history dynamics. The inbreeding coefficient F_{st}[21] provides an index of the extent to which a variant is specific to a population. We compared African-Americans and Caucasians for the frequency distribution of F_{st} values for each of 13 802 clearly biallelic SNPs type in Caucasian, Asian and African-American populations from the June 2002 release of the Allele Frequency Project of the SNP consortium.[22] Of these SNPs, 4329 were associated with the loss or gain of a CpG dinucleotide. We could not detect any significant difference in F_{st} values for Caucasian and African-Americans between CpG and non-CpG SNPs, as determined by Wilcoxon's rank-sum test ($P = 0.85$). This finding suggests that population structure differences are dominated by drift, rather than by mutation. Given the large sample size, it is likely that different sequence classes of SNPs show similar levels of population substructure, and it is acceptable to use CpG SNPs to control for the population structure of non-CpG SNPs, and *vice versa*.

Discussion

This study supports the views of previous commentators[23–25] that population substructure need not be a major problem in genetic case–control studies. OD of allele frequencies can be adequately modelled by the Genomic Control method[3] if only significance testing is required, while calculation of OR_{adjust} and CI_{od} are appropriate to provide estimators of adjusted risk and confidence interval, which are required for meta-analyses. Major sequence classes of SNPs appear to have similar population structure, and can therefore be corrected for in a similar manner . Thus, OD of allele frequency differences can be adequately accounted for using standard statistical approaches. The one context in which careful modelling is most important is in the analysis of very modest genetic risks. In this situation, conclusions can usually only be drawn after meta-analyses of many studies.[10] Therefore, any statistics reporting adjusted risks corrected for population structure should be in the form of ORs and confidence intervals, such as the OR_{adjust} and CI_{od} values proposed here, which may then form the basis for future meta-analyses.[10]

While allele frequency differences among subpopulations are likely to be dominated by the effects of genetic drift, selection processes may distinguish some subsets of SNPs. Thus, certain groups of candidate genes may show more marked frequency differences between populations. Genes involved in pathogen responses (such as HLA variants[26]) show marked allele frequency differences among populations, and where a candidate has been drawn from such a group of genes, a random set of control genes may be less appropriate than a parallel group of genes displaying similar levels of genetic or geographic stratification. While this is ideal, it may not be easy to define, and the first-order correction with random genes is probably sufficient to reduce the genetic confounding of association studies by population substructure to an outside possibility. Rare polymorphism frequencies may be

more subject to drift than common polymorphisms, and it will be of interest to determine to what extent, if any, rare polymorphisms exhibit greater population stratification and whether this will provide a more critical situation for confounding of genetic association by population structure.

We have demonstrated that OR_{adjust} with CI_{od} provides efficient estimates of significance and has good coverage. Since this can be readily calculated by statisticians without extensive training in specialised genetics software, we propose that calculation of adjusted ORs using the methods outlined here or through careful robust modelling of population structure[7] be adopted as a standard. Presentation of such statistics provides sufficient information for future meta-analyses[10] of test marker main effects on disease, without the need to reanalyse the random markers within the meta-analysis. We considered a straightforward balanced design with equal numbers of observations per marker in which carrier status was the only risk factor. However, mixed effects regression can be used for complex analysis with unbalanced designs, multiple fixed and random effects and data missing at random.[27] Several software packages implement mixed logistic regression including SAS Proc NLMixed, MIXNO and MlwiN.[17,28–30]

More recently, it has been indicated that sophisticated Markov chain Monte Carlo modelling of population structure effects provides an opportunity to permit meta-analyses,[8] in principle, with consideration of covariates. The main advantage of the simpler approach suggested here is that the model fitting is clearer to mainstream statisticians, allowing the usual modelling of covariates, and may be more readily interpreted by nonstatistical geneticists.

Future meta-analysis of genetic association studies is best served by relatively straightforward statistical estimates with a clear basis. Thus, the methods proposed here can serve to facilitate the maximum value of meta-analyses of multiple diverse data sets. We propose that authors of publications and journals should favour the routine reporting of OR_{adjust} and $CI_{od,}$ in order to facilitate future literature-based meta-analyses[10] across various studies. Such meta-analysis is likely to be reasonably robust in the face of divergent population structures among studies, as well as different choices of random or control markers. It will be of interest in future evaluations of such methods to compare the performance of OR_{adjust} and structure-based methods (eg Hoggart *et al*[8]) in meta-analysis, ideally of a number of large real data sets as these become available.

Acknowledgements

This work was supported by the Health Research Board, Ireland, and by the Programme for Research in Third Level Institutions, administered by the Higher Education Authority, Ireland.

References

1 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K: A comprehensive review of genetic association studies. *Genet Med* 2002; **4**: 45–61.
2 Pritchard JK, Donnelly P: Case–control studies of association in structured or admixed populations. *Theor Popul Biol* 2001; **60**: 227–237.
3 Devlin B, Roeder K: Genomic control for association studies. *Biometrics* 1999; **55**: 997–1004.
4 Reich DE, Goldstein DB: Detecting association in a case–control study while correcting for population stratification. *Genet Epidemiol* 2001; **20**: 4–16.
5 Pritchard JK, Rosenberg NA: Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999; **65**: 220–228.
6 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–181.
7 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**: 466–477.
8 Hoggart CJ, Parra EJ, Shriver MD *et al*: Control of confounding of genetic associations in stratified populations. *Am J Hum Genet* 2003; **72**: 1492–1504.
9 Ripatti S, Pitkaniemi J, Sillanpaa MJ: Joint modeling of genetic association and population stratification using latent class models. *Genet Epidemiol* 2001; **21** (Suppl 1): S409–S414.
10 Ioannidis JP, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG: Genetic associations in large *versus* small studies: an empirical assessment. *Lancet* 2003; **361**: 567–571.
11 Clayton D, Hills M: *Statistical Models in Epidemiology*. Oxford: Oxford University Press, 1993.
12 Peter Armitage P, Berry G, Matthews J: *Statistical Methods in Medical Research*. Oxford: Blackwell Science, 2001.
13 Ferdinand KC, Bakris GL, Douglas JG, Sowers JR: Hypertension-related disease in African-Americans. *Postgraduate Med* 2002; **112**: 24–48.
14 Goddard KA, Hall PJ, Hopkins JM, Witte JS: Linkage disequilibrium and allele-frequency distributions for 114 single-nucleotide polymorphisms in five populations. *Am J Hum Genet* 2000; **66**: 216–234.
15 Nicholson G, Smith AV, Jónsson F, Gústafsson O, Stefánsson K, Donnelly P: Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J R Stat Soc Ser B* 2002; **64**: 695–715.
16 Snijders T, Bosker R: *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. Sage: Thousand Oakes, CA, 1999.
17 Hedeker D: *MIXNO A Computer Program for Mixed-Effects Nominal Logistic Regression. Manual*. Chicago: Divison of Epidemiology and Biostatistics and Health Policy and Research Centers, School of Public Health, University of Illinois at Chicago, 1999.
18 Turakulov R, Easteal S: Number of SNPS loci needed to detect population structure. *Hum Hered* 2003; **55**: 37–45.
19 Freedman ML, Reich D, Penney KL *et al*: Assessing the impact of population stratification on genetic association studies. *Nat Genet* 2004; **36**: 388–393.
20 Sved J, Bird A: The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proc Natl Acad Sci USA* 1990; **87**: 4692–4696.
21 Wright S: Evolution in Mendelian populations. *Genetics* 1931; **16**: 97–159.
22 Holden AL: The SNP consortium: summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques* 2002; (Suppl 1): 22–24, 26.
23 Morton NE, Collins A: Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci USA* 1998; **95**: 11389–11393.
24 Cardon LR, Bell JI: Association study designs for complex diseases. *Nat Rev Genet* 2001; **2**: 91–99.

25 Colhoun HM, McKeigue PM, Davey Smith G: Problems of reporting genetic associations with complex outcomes. *Lancet* 2003; **361**: 865–872.
26 Cavalli-Sforza L, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton, NJ: Princeton University Press, 1994.
27 Pinheiro JC, Bates DM: Approximations to the log-likelihood function in the non-linear mixed effects model. *J Comput Graph Stat* 1995; **4**: 12–35.
28 Goldstein H, Browne W, Rasbash J: Multilevel modelling of medical data. *Stat Med* 2002; **21**: 3291–3315.
29 Hosmer DW, Lemeshow S: *Applied Logistic Regression, Textbook and Solutions Manual* 2001, 2nd edn. New York: Wiley-Interscience.
30 Sullivan LM, Dukes KA, Losina E: Tutorial in biostatistics. An introduction to hierarchical linear modelling. *Stat Med* 1999; **18**: 855–888.