nature genetics

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Assessing the impact of population stratification on genetic association studies

Matthew L Freedman[1–3,15], David Reich[3,4,15], Kathryn L Penney[1–3], Gavin J McDonald[3,4], Andre A Mignault[4], Nick Patterson[3], Stacey B Gabriel[3], Eric J Topol[5], Jordan W Smoller[6,7], Carlos N Pato[8,9], Michele T Pato[8,9], Tracey L Petryshen[3], Laurence N Kolonel[10], Eric S Lander[3,11], Pamela Sklar[3,6,7], Brian Henderson[12], Joel N Hirschhorn[3,4,13] & David Altshuler[1,3,4,14]

Population stratification refers to differences in allele frequencies between cases and controls due to systematic differences in ancestry rather than association of genes with disease. It has been proposed that false positive associations due to stratification can be controlled by genotyping a few dozen unlinked genetic markers. To assess stratification empirically, we analyzed data from 11 case-control and case-cohort association studies. We did not detect statistically significant evidence for stratification but did observe that assessments based on a few dozen markers lack power to rule out moderate levels of stratification that could cause false positive associations in studies designed to detect modest genetic risk factors. After increasing the number of markers and samples in a case-cohort study (the design most immune to stratification), we found that stratification was in fact present. Our results suggest that modest amounts of stratification can exist even in well designed studies.

There has been much debate[1–4] but limited data[5–7] about the impact of population stratification on case-control association studies. Systematic differences in the ancestry of cases and controls are one source of false positive associations[8,9], but the fraction of published associations that is attributable to stratification is unknown[10]. It has been argued that the effects of stratification can be eliminated simply by carefully matching cases and controls according to self-reported ancestry and geographical origin[2]. Recently, empirical methods to detect stratification based on genotypes at unlinked markers have been described[11]. The largest application of such methods involved genotyping 44 unlinked markers in four case-control studies[5]. Stratification was detected in one study, although the signal was no longer apparent after more stringent matching of cases and controls based on the birthplaces of the individuals' grandparents. This has been interpreted as evidence that stratification may be less of a concern than originally anticipated.

We assessed stratification empirically by analyzing data from 24–48 unlinked single-nucleotide polymorphisms (SNPs) in 11 association studies spanning a range of disease states and self-reported ancestries and three different epidemiological designs. These studies included seven ongoing studies in our laboratory and reanalysis of data from the four studies previously reported[5]. We assessed stratification first by testing for statistically significant evidence of differentiation between cases and controls using the method of Pritchard and Rosenberg[11] and second by estimating the magnitude of stratification consistent with the data using Genomic Control[12,13].

None of the 11 studies showed significant evidence of stratification after correcting for multiple hypothesis testing, consistent with previous studies[5,6] (**Table 1**). Comparing cases and controls from different studies with the same self-reported ancestry (European American), we found no significant evidence of stratification in nine pairwise comparisons using 33–43 SNPs (**Supplementary Table 1** online).

We next applied the method of Genomic Control[12,13] to estimate quantitatively the amount of stratification consistent with the data for each of the 11 studies. Genomic Control is conceptually simple: the method examines the distribution of association statistics (χ^2) between unlinked genetic variants typed in cases and controls. The statistic at a candidate allele being tested for association can then be compared with the genome-wide distribution of statistics for markers

Departments of [1]Medicine and Molecular Biology and [2]Hematology-Oncology, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA. [3]Program in Medical and Population Genetics, Broad Institute, One Kendall Square, Building 300, Cambridge, Massachusetts 02139, USA. [4]Department of Genetics, Harvard Medical School, New Research Building, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [5]Department of Cardiovascular Medicine, Cleveland Clinic Foundation, Cleveland, Ohio, USA. [6]Department of Psychiatry, Harvard Medical School, Boston, Massachusetts, USA. [7]Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, 149 13th Street, Charlestown, Massachusetts, USA. [8]Veterans Administration, Syracuse, New York, USA. [9]Center for Psychiatric and Molecular Genetics, SUNY/Upstate Medical University, Syracuse, New York, USA. [10]Cancer Etiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii, USA. [11]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [12]Department of Preventative Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California, USA. [13]Divisions of Genetics and Endocrinology, Children's Hospital and Department of Pediatrics, Harvard Medical School, Boston, Massachusetts, USA. [14]Diabetes Unit, Massachusetts General Hospital, 55 Fruit Street, Boston, Massachusetts 02114, USA. [15]These authors contributed equally to this work. Correspondence should be addressed to D.R. (reich@receptor.med.harvard.edu).

Published online 28 March 2004; doi:10.1038/ng1333

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Table 1 Assessment of population stratification in 11 epidemiological studies

Study	Disease	Source of controls	Self-declared ancestry	Cases	Controls	SNPs used to assess stratification	Significance of stratification based on ref. 11	Estimate of λ projected to 100 cases per 100 controls (upper bound on λ)	Estimate of λ projected to 1,000 cases per 1,000 controls (upper bound on λ)
1	Hypertension	Matched controls	African American	236	236	24	$P < 0.19$	1.11 (<1.81)	2.1 (<9.1)
2		Matched controls	European American	500	500	32	$P < 0.16$	1.05 (<1.31)	1.5 (<4.1)
3	Type II diabetes	Matched controls	European American	500	355	32	$P < 0.66$	1 (<1.20)	1 (<3.0)
4		Matched controls	Polish	500	500	32	$P < 0.72$	1 (<1.16)	1 (<2.6)
5	Prostate cancer	Cohort	African American	90	69	48	$P < 0.75$	1 (<1.73)	1 (<8.3)
6		Cohort	European American	110	97	42	$P < 0.08$	1.3 (<2.4)	4.0 (<15.0)
7		Cohort	Hispanic American	142	124	40	$P < 0.84$	1 (<1.42)	1 (<5.2)
8		Cohort	Japanese American	121	106	33	$P < 0.97$	1 (<1.43)	1 (<5.3)
9	Coronary artery	GeneQuest[a]	European American	83	80	37	$P < 0.67$	1 (<1.9)	1 (<10.0)
10	Bipolar disorder	GeneQuest[a]	European American	93	80	34	$P < 0.23$	1.18 (<2.54)	2.8 (<16.4))
11	Schizophrenia	Matched controls	Portuguese	149	152	46	$P < 0.21$	1.1 (<1.72)	2 (<8.2)

Follow-up for study 5 (prostate cancer in African Americans) with more SNPs and samples

Cases and controls[b]	All samples analyzed	Excluding samples who claim some non-African ancestry
90 cases, 69 controls (no AIMs included)	$P < 0.75$ (48 SNPs)	
469 cases, 268 controls (no AIMs included)	$P < 0.04$ (114 SNPs)	$P < 0.03$ (114 SNPs)
469 cases, 268 controls (AIMs included)	$P < 0.01$ (210 SNPs)	$P < 0.005$ (210 SNPs)
474 cases, 476 controls (AIMs included)[c]	$P < 0.0001$ (211 SNPs)	$P < 0.000001$ (211 SNPs)

Follow-up for study 6 (prostate cancer in European Americans) with more SNPs and samples

Cases and controls	All samples analyzed
110 cases, 97 controls	$P < 0.08$ (42 SNPs)
391 cases, 456 controls	$P < 0.10$ (79 SNPs)

[a]No controls were collected along with cases. To assess the effect of matching cases to controls based only on self-reported ancestry, we used a set of controls obtained by the GeneQuest coronary artery disease study by random-digit-dialing in Atlanta, Georgia, USA. [b]AIMs refers to markers chosen to have very different frequencies between Africans and Europeans[7,15]. [c]One more SNP was analyzed for the larger sample because it met the criteria for inclusion in the study.

that are probably unrelated to disease to assess whether the candidate allele stands out. In the absence of stratification, association between unlinked genetic variants and disease should follow a χ^2 distribution with 1 degree of freedom[12,13]. In the presence of stratification, the distribution of association statistics should be inflated by a value termed λ, which becomes larger with increasing of sample size (**Fig. 1**).

We estimated stratification for each of the 11 data sets and report the inflation of association statistics that would be expected in a study of 1,000 cases and 1,000 controls, called λ_{1000}. (It is simple to extrapolate from λ_{1000} to the inflation factor due to stratification for any sample size[12].) Consistent with the fact that the 11 data sets showed no significant evidence for stratification, the confidence intervals for

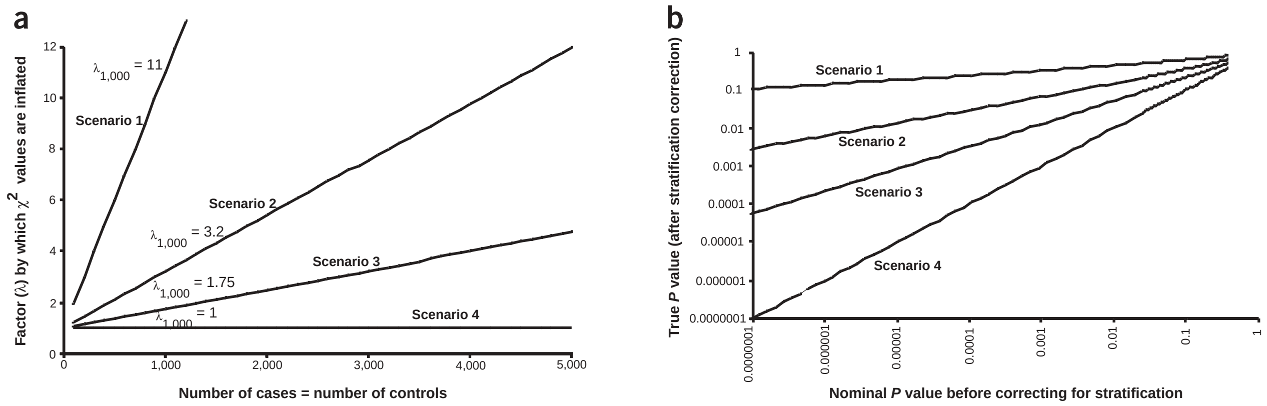

Figure 1 The effect of stratification on association studies. (**a**) Stratification inflates χ^2 association statistics by a factor λ, which changes depending on the sample size. Scenario 1 corresponds to gross stratification; scenarios 2 and 3 correspond to the range of stratification estimated in the African American prostate cancer study; and scenario 4 corresponds to no stratification. (**b**) Comparison of the nominal *P* values with those corrected for stratification shows that stratification that is difficult to detect in a study of hundreds of cases and controls can cause many false positive signals in a study of thousands of samples.

λ_{1000} overlapped 1 in every study. Nevertheless, we found that the confidence intervals were sufficiently broad that substantial levels of stratification could not be excluded. For example, the 95th percentile upper bound on λ_{1000} in the studies averaged 7.9 (**Table 1** and **Fig. 2**).

We increased power to detect stratification by increasing the number of SNPs and samples examined in one of the 11 studies that initially showed no significance evidence for stratification, the African American prostate cancer study ($P < 0.75$). For the follow-up, we approximately quadrupled the number of markers and increased the sample size by a factor of 5–6 (474 prostate cancer cases and 476 cohort controls). The new markers consisted of a collection of missense SNPs, which we treated as being in the same class as the noncoding SNPs, because, within the limits of our resolution (**Table 2**), they showed the same levels of population differentiation (with sufficient power, such differences can probably be detected[14]). The new markers also included a second set of SNPs chosen for their large allele frequency differences between west Africans and Europeans[15], which makes them particularly powerful for detecting stratification[16].

In this expanded data set we found significant evidence of stratification ($P < 0.0001$). When we restricted the analysis to 469 cases and 268 controls in whom all markers were successfully typed, the result was still significant ($P < 0.01$; **Table 1**). We then removed from the analysis 40 cases and 48 controls who reported that either they or their parents had some non-African American ancestry[2,5], because a small number of individuals with misclassified ancestry might disproportionately affect the result. The evidence for stratification was stronger in this subset ($P < 0.005$; **Table 1**). Notably, the Genomic Control estimate of stratification (removing SNPs that had been specifically chosen to have large differences in frequency across populations[17]) was $\lambda_{1000} = 1.5$, with a 95th percentile upper bound of 3.34. This indicates that an observation of $\chi^2 = 19.5$, expected only once by chance in a scan of 100,000 SNPs, would instead be seen 31 times (effective $\chi^2 = 19.5/1.5 = 13$) due to this level of stratification. At the 95% upper confidence limit of our estimate ($\lambda_{1000} = 3.34$), 1,568 false positives would be expected due to stratification.

The observation of population stratification in African Americans with prostate cancer is not entirely unexpected. People of west African descent are thought to have a higher genetic risk for prostate cancer than those of European descent[18], and hence African Americans with prostate cancer, who are known to have ancestry from both populations[15], might be expected to have more African ancestry, on average, than controls. Population stratification was also observed in a separate study of African Americans with prostate cancer[9]. Our analysis strengthens this result, in that our sample was prospectively collected in a population-based cohort[19], considered to be the optimal epidemiological design to minimize systematic differences between cases and controls (as opposed to the case-control design).

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

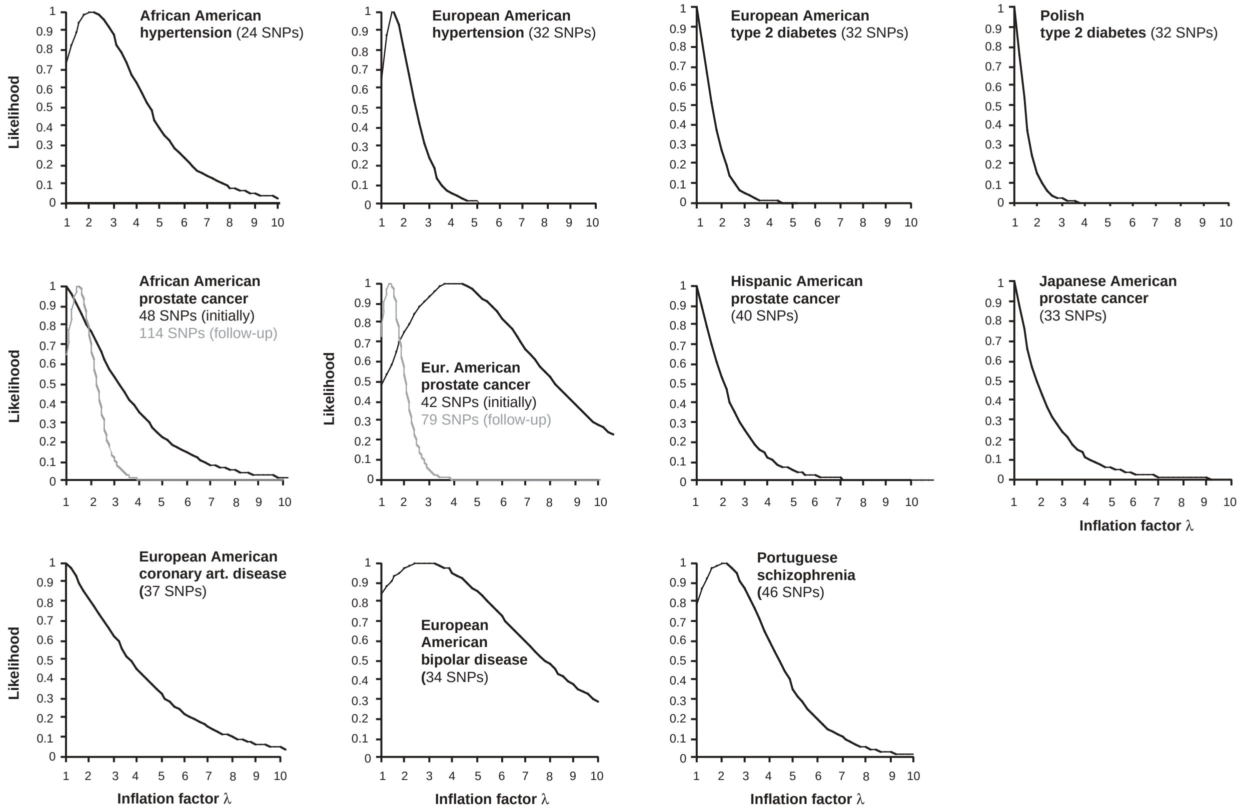

Figure 2 Likelihood surfaces for stratification for the 11 studies, assuming 1,000 cases and 1,000 controls (we provide results for λ_{1000}, but likelihood surfaces for other numbers of cases and controls could be obtained simply by rescaling the axis using the equation in Methods). The upper bound on the level of stratification can be obtained from the figures as the point where the likelihood drops to 4.5% of its maximum, which is a log-likelihood criterion for a 95th percentile upper bound (one-sided test). With the handful of SNPs genotyped initially (24–48), the likelihood distribution is broad. Although no studies show significant stratification, all are consistent with levels of stratification that could produce notable numbers of false positives. Increasing the number of SNPs and samples can tighten the estimate of population stratification. This is shown for the African American and European American prostate cancer studies, for which we provide both an initial estimate of stratification based on the noncoding SNPs and a more precise estimate based on an expanded sample size and inclusion of missense SNPs.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Table 2 Comparison of levels of stratification in missense versus randomly chosen SNPs

Comparison	SNPs used in calculation (noncoding/missense)	Mean χ^2 for noncoding SNPs	Mean χ^2 for missense SNPs	Mann-Whitney significance
African American (n = 88) versus Asian American (n = 70)	40/68	12.5	17.2	$P < 0.36$
African American (n = 88) versus European American (n = 156)	41/70	12.5	14.8	$P < 0.11$
Asian American (n = 70) versus European American (n = 156)	35/66	8.3	15.3	$P < 0.04$

Application of the Genomic Control approach relies on the assumption that the distribution of noncoding, unlinked SNPs across populations is similar to that for the missense SNPs typically tested in association studies. We assessed this by genotyping missense SNPs in 50 African American, 88 European American and 42 Asian American population samples and comparing the χ^2 values to the noncoding SNPs. The differentiation between populations is not significantly greater among missense SNPs than noncoding SNPs (accounting for the fact that three hypotheses were tested). This suggests that noncoding SNPs can be used to assess stratification in a way that is roughly applicable to missense SNPs as well.

We also followed up with a study of European Americans with prostate cancer (approximately doubling the number of SNPs to 79 and quadrupling the number of samples to 391 cases and 456 cohort controls). In this study, we did not find statistically significant evidence for stratification ($P < 0.10$). The 95th percentile upper bound on stratification from Genomic Control, however, was similar to that in the study of African Americans ($\lambda_{1000} = 3.03$; **Fig. 2**). Much more data will be needed from many studies before it is possible to assess whether matching cases and controls solely on the basis of their self-reported ancestry, in a population such as European Americans without recent mixture, is adequate to take into account population stratification.

Our data indicate that genotyping a few dozen markers cannot rule out modest levels of population stratification that could generate false positives in an association study designed to detect alleles of weak effect—even in the setting of a prospectively collected cohort study. Stratification is probably most problematic in populations whose ancestors recently mixed due to intercontinental migrations and for diseases that have different prevalence rates across these ancestral populations[11,13] (such as hypertension, obesity, diabetes and autoimmunity). Because the importance of stratification grows with sample size[12,13], however, it seems possible that, even for diseases whose incidence rates are not currently known to vary across populations, stratification could exist. Thus, our study argues that stratification cannot be excluded based on either first principles or published empirical data. We suggest instead that investigators continue to monitor for stratification. In addition to presenting nominal P values, investigators should also report the range of values consistent with the Genomic Control estimate of stratification in the samples based on genotyping unlinked markers. Alternatively, investigators could present a P value corrected for the full range of possible values of λ_{1000}, using the full Bayesian approach to Genomic Control[12].

Table 3 Number of SNPs necessary to ensure an association is not due to stratification

Number of markers evaluated	Maximum factor by which λ can exceed the best estimate of stratification
5	1.847
10	1.599
15	1.487
20	1.421
25	1.375
30	1.342
40	1.295
50	1.263
60	1.240
80	1.207
100	1.185
125	1.166
150	1.151
200	1.130
250	1.116
300	1.106
350	1.098
400	1.092
450	1.086
500	1.082
1,000	1.058

Number of markers than must be genotyped to be 95% confident that the upper bound on stratification is within a particular factor of the best estimate. If we observe a χ^2 value of x and the genome-wide threshold of significance is y, then ruling out stratification as an explanation for the positive association means genotyping enough markers so that the second column in the table is less than x/y. See ref. 13 for a related table.

Our data show that stratification cannot be excluded as a possibility in real case-control studies, but that there is no need to abandon case-control and case-cohort studies in favor of family-based designs (such as transmission disequilibrium tests). Two powerful approaches are available to detect and correct for stratification[20]. The first clusters samples based on multilocus genotypes (*e.g.*, STRUCTURE[21]) to identify individuals with different ancestries. This provides a way to adjust for ancestry as a covariate in the association analysis[7,21]. Genomic Control, on the other hand, makes a quantitative estimate of the degree of stratification and uses it to adjust for any stratification that might be present. The two methods are not mutually exclusive: STRUCTURE can be used first to identify and eliminate samples that contribute unduly to stratification, and a smaller Genomic Control correction can then be made in the final study.

How many SNPs need to be used in an assessment of stratification? This question must be viewed in relation to the magnitude of genetic effects under study. Given a substantial magnitude of effect and a highly significant P value, only a few dozen markers probably need to be genotyped to rule out gross stratification as an explanation for the positive association[11,22] (**Table 3**). In contrast, if the results point to more modest influences on disease, such as the risk due to variation in *CTLA4* on autoimmune thyroid disease and type 1 diabetes[23], it may be necessary to genotype a larger number of markers to rule out modest amounts of stratification. Genotyping more than 340 markers can bring the conservative 95th percentile upper bound on the level of stratification to within 10% of the true value (**Table 3**). Fortunately, as the number of SNPs tested in association studies grows larger (to survey the genome for risk-associated alleles of increasingly modest effect), the bounds on the estimate of stratification should become increasingly precise with no additional effort, as all the markers in a study can be used to assess and adjust for stratification[12,13].

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

METHODS

Clinical samples. We obtained all samples for the new data collections with permission of the principal investigators and with approval of the Institutional Review Boards of the Massachusetts General Hospital, the Cleveland Clinic, SUNY/Upstate Medical University, the University of Hawaii and the University of Southern California. Informed consent was obtained from all subjects by the institutions responsible for the collections. Citations provide additional detail on the ascertainment of cases and controls.

GeneQuest coronary artery disease study[24]. We randomly selected 83 cases and 80 controls, all European Americans. The cases were from Cleveland, and controls were identified by random digit phone dialing in Atlanta, Georgia, USA.

Multiethnic Cohort prostate cancer study. The Multiethnic Cohort[19] is an ongoing (n = 215,251) study focusing on the effects of diet, genes and environment on the risk of cancer. The cohort samples include four main ethnic groups in Los Angeles and Hawaii. For European Americans, we randomly selected 110 incident cases and 97 cohort controls; for African Americans, we selected 90 incident cases and 69 cohort controls; for Japanese Americans, we selected 121 incident cases and 106 cohort controls; and for Hispanic Americans, we selected 142 incident cases and 124 cohort controls. We followed up in the study of African American prostate samples by genotyping all the missense and ancestry-informative SNPs described below. We genotyped an expanded sample of 469 African-American incident cases and 268 cohort controls from the cohort for all the SNPs and genotyped an additional 5 cases and 208 cohort controls for 31 of the SNPs that had high allele frequency differences across populations before the DNA for these samples ran out.

Bipolar disorder in European Americans. We obtained 93 DNA samples from Massachusetts General Hospital from individuals with diagnoses of bipolar disorder 1 or bipolar disorder 2. As controls, we used GeneQuest samples (both this and the coronary artery disease study are examples where cases are matched to controls only using self-reported ancestry.)

Schizophrenia. We obtained samples from 149 cases diagnosed with schizophrenia according to the criteria of the Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition, and 152 matched controls as part of a study of schizophrenia in the Portuguese population. Samples were descended from continental Portugal (83% of cases, 87% of controls), the Azore islands (13% of cases, 3% of controls) or the Madeira islands (3% of cases, 10% of controls). Some individuals were from Fall River, Rhode Island, but in each of these cases, all four grandparents were from the Azore islands.

For comparison of noncoding to missense SNPs, we studied 50 African American, 88 European American and 42 Asian American population samples. These were identical to those previously studied[25], except that the 88 European American samples were replaced by the parents of the 44 samples resequenced previously[26].

Choice of markers. The physical and genetic map positions, along with flanking sequences, of all SNPs used in this study are available from the authors on request.

We obtained noncoding SNPs (67) from the SNP Consortium website. They were identified by comparing a single sequencing read from a diverse panel of individuals with the publicly available genome sequence[27]. The SNPs were evenly spaced throughout the autosomes, each at least 20 Mb from the others. In practice, only 34–48 of these SNPs genotyped successfully and were of high enough frequency in any study to use in our analysis (the expected number of reference and variant alleles based on the allele frequency and sample size was ≥5).

We identified missense SNPs (100) from a database of SNPs in coding regions of genes[28], obtained as part of an effort to catalog SNPs in genes of interest for disease. We used only genes that were not designated in the database or in a published meta-analysis[10] as having any relationship with prostate cancer, coronary artery disease, asthma or atopy. We excluded from the study those SNPs with a minor allele frequency <10% in a multiethnic screening panel. SNPs were chosen to be at least 1 Mb away from each other and from all the noncoding SNPs.

We obtained ancestry-informative SNPs (101) with high allele frequency differences comparing European and African Americans by combining data from ref. 15 with unpublished data from our own laboratory. These SNPs were all chosen to be at least 20 Mb from each other. The average frequency difference comparing west Africans and Europeans was 67%.

Genotyping. The genotypes collected for this study are available from the authors to the extent that is consistent with the informed consent provided by the study participants. We used matrix-associated laser desorption ionization–time of flight mass spectrometry (MALDI-TOF)[29] with 5 ng of DNA per multiplex genotyping reaction to genotype most SNPs in this study. The PCR protocol is described elsewhere[25]. Error rates with the Sequenom MassARRAY system have been estimated to be ~0.4% at our laboratory[25], although the discrepancy rate in the present data set suggests closer to 0.25% (215 conflicts out of 42,766 genotypes, each done at least in duplicate).

Elimination of poorly performing SNPs. We removed all SNPs from our analysis that showed Hardy Weinberg P values of <0.01 in at least two of the three diversity samples (CEPH, East Asian and African American). We also excluded SNPs from the analysis if the combined Hardy-Weinberg P value, over all populations excluding African Americans and Hispanic Americans, was <0.01. To calculate the P value, we summed the χ^2 values for the Hardy-Weinberg test over all n populations for which the statistic could be calculated and assessed significance using a χ^2 distribution with n degrees of freedom. (We excluded African Americans and Hispanic Americans from the Hardy-Weinberg assessment because different levels of population mixture across individuals in these groups can produce a deficiency of heterozygotes, even with accurate genotyping.) We also excluded from analysis SNPs for those studies in which the genotyping success rates were <75% in either cases or controls[25]. We also eliminated from analysis SNPs that showed discrepancy rates of >3% in duplicate genotypes.

Detection of population stratification. We calculated χ^2 association statistics for all k SNPs in a study, including only those for which the expected number of allele counts (based on the combined frequency in the two population samples) was at least 5. We then summed the values and assessed significance using a χ^2 distribution with k degrees of freedom[11].

Quantitative assessment of population stratification. For each SNP in each study for which at least 40% of the cases and controls had been successfully genotyped, we calculated χ^2 values for all SNPs for which the expected number of allele counts (based on the combined frequency in the two population samples) was at least 5.

We carried out a likelihood analysis to estimate the level of stratification consistent with the data in each study. Defining c_j as the association statistic observed at marker j genotyped in n_j cases and m_j controls and f as the χ^2 distribution with 1 degree of freedom, the likelihood of a given inflation factor due to stratification is simply

$$L_j = f\left(c_j/\lambda_{n_j,m_j}\right)/\lambda_{n_j,m_j}$$

a consequence of the fact that the χ^2 distribution scales with the inflation factor[12]. The likelihood at all K markers is then

$$L = \prod_{j=1}^{K} \frac{f\left(c_j/\lambda_{n_j,m_j}\right)}{\lambda_{n_j,m_j}} .$$

To estimate a likelihood distribution for the level of stratification, we define a reference sample size (we use n_{ref} = 1,000 cases and m_{ref} = 1,000 controls). We then use an equation derived in ref. 12 and confirmed by simulation as in ref. 13 to relate this to the inflation factor applicable to n_j cases and m_j controls. The inflation factor should be different from marker to marker because it scales with sample size:

$$\lambda_{n_j,m_j} = 1 + \left(\lambda_{n_{ref},m_{ref}} - 1\right)\left(\frac{1}{n_{ref}} + \frac{1}{m_{ref}}\right) \Big/ \left(\frac{1}{n_j} + \frac{1}{m_j}\right)$$

In this paper we abbreviate $\lambda_{1000,1000}$ as λ_{1000}.

© 2004 Nature Publishing Group http://www.nature.com/naturegenetics

Substituting equation 3 into equation 2 allows us to obtain a likelihood distribution for λ_{1000}. The maximum likelihood estimate for λ_{1000} is simply the value for which L is maximized, with the requirement that $\lambda_{1000} \geq 1$. We obtained the likelihood surfaces shown in **Figure 2** by plotting the values of L for different λ_{1000}, normalizing by the maximum likelihood (set equal to 1 in **Fig. 2**). We obtained the upper bound on λ_{1000} by picking the value such that the likelihood ratio $2\log_{10}(L_{max}/L) = 2.7$; that is, the point for which the likelihood was 4.5% of the maximum, corresponding roughly to a $P < 0.05$ cutoff (one-sided test).

To test for a difference in the distribution of χ^2 values between missense and noncoding SNPs (**Table 2**), we compared the random African American, European American and Asian American population samples in our study. For each SNP for which at least 70% both sample sets had been successfully genotyped, we randomly dropped samples until we had the same number at all sites. We then calculated χ^2 values and used a Mann-Whitney U test to assess whether the empirical distributions of statistics at missense and noncoding SNPs were distinguishable.

URL. The SNP Consortium website is available at http://snp.cshl.org.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We thank A. Villapakkam for assistance in genotyping and data checking and K. Ardlie, D. Goldstein and C. Haiman for discussions. M.L.F. is supported by a Department of Defense Health Disparity training grant and a Postdoctoral Fellowship for Physicians from the Howard Hughes Medical Institute. D.A. is a Clinical Scholar in Translational Research from the Burroughs Wellcome Fund, as well as a Charles E. Culpeper Medical Scholar. J.N.H. and D.R. are recipients of Career Development Awards from the Burroughs Welcome Fund. T.L.P. is supported by a Canadian Institutes of Health Research Postdoctoral Fellowship and is a NARSAD Young Investigator. This work was supported in part by a grant from the Functional Genomics Program at the Whitehead Institute/MIT Center for Genome Research (supported by Affymetrix, Millennium Pharmaceuticals and Bristol Myers Squibb) and by a grant from the US National Institutes of Health to B.H. and L.K.

COMPETING INTERESTS STATEMENT

The authors declare competing financial interests (see the *Nature Genetics* website for details).

Received 15 October 2003; accepted 23 February 2004
Published online at http://www.nature.com/naturegenetics/

1. Thomas, D.C. & Witte, J.S. Point: population stratification: a problem for case-control studies of candidate-gene associations? *Cancer Epidemiol. Biomarkers Prev.* **11**, 505–512 (2002).
2. Wacholder, S., Rothman, N. & Caporaso, N. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiol. Biomarkers Prev.* **11**, 513–520 (2002).
3. Ziv, E. & Burchard, E.G. Human population structure and genetic association studies. *Pharmacogenomics* **4**, 431–441 (2003).
4. Cardon, L.R. & Palmer, L.J. Population stratification and spurious allelic association. *Lancet* **361**, 598–604 (2003).
5. Ardlie, K.G., Lunetta, K.L. & Seielstad, M. Testing for population subdivision and association in four case-control studies. *Am. J. Hum. Genet.* **71**, 304–311 (2002).
6. Schork, N.J. *et al.* The future of genetic case-control studies. *Adv. Genet.* **42**, 191–212 (2001).
7. Hoggart, C.J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
8. Knowler, W.C., Williams, R.C., Pettitt, D.J. & Steinberg, A.G.. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.* **43**, 520–526 (1988).
9. Kittles, R.A. *et al.* CYP3A4-V and prostate cancer in African Americans: causal or confounding association because of population stratification? *Hum. Genet.* **110**, 553–560 (2002).
10. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. & Hirschhorn, J.N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
11. 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).
12. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
13. Reich, D.E. & Goldstein, D.B. Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.* **20**, 4–16 (2001).
14. Akey, J.M., Zhang, G., Zhang, K., Jin, L. & Shriver, M.D. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res.* **12**, 1805–1814 (2002).
15. Parra, E.J. *et al.* Estimating African American admixture proportions by use of population-specific alleles. *Am. J. Hum. Genet.* **63**, 1839–1851 (1998).
16. Pfaff, C.L., Kittles, R.A. & Shriver, M.D. Adjusting for population structure in admixed populations. *Genet. Epidemiol.* **22**, 196–201 (2002).
17. Reich, D.E. & Goldstein, D.B. Response to Pfaff *et al.*: Adjusting for population structure in admixed populations. *Genet. Epidemiol.* **22**, 196–201 (2002).
18. Bunker, C.H. *et al.* High prevalence of screening-detected prostate cancer among Afro-Caribbeans: the Tobago prostate cancer survey. *Cancer Epidemiol. Biomarkers Prev.* **11**, 726–729 (2002).
19. Kolonel, L.N. *et al.* A multiethnic cohort in Hawaii and Los Angeles: baseline characteristics. *Am. J. Epidemiol.* **151**, 346–357 (2000).
20. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
21. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
22. Siddiqui, A. *et al.* Association of multidrug resistance in epilepsy with a polymorphism in the drug-transporter gene ABCB1. *N. Engl. J. Med.* **348**, 1442–1448 (2003).
23. Ueda, H. *et al.* Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
24. Topol, E.J. *et al.* Single nucleotide polymorphisms in multiple novel thrombospondin genes may be associated with familial premature myocardial infarction. *Circulation* **104**, 2641–2644 (2001).
25. Gabriel, S.B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
26. Reich, D.E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
27. Sachidanandam, R. *et al.* A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933 (2001).
28. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**, 231–238 (1999).
29. Tang, K. *et al.* Chip-based genotyping by mass spectrometry. *Proc. Natl. Acad. Sci. USA* **96**, 10016–10020 (1999).