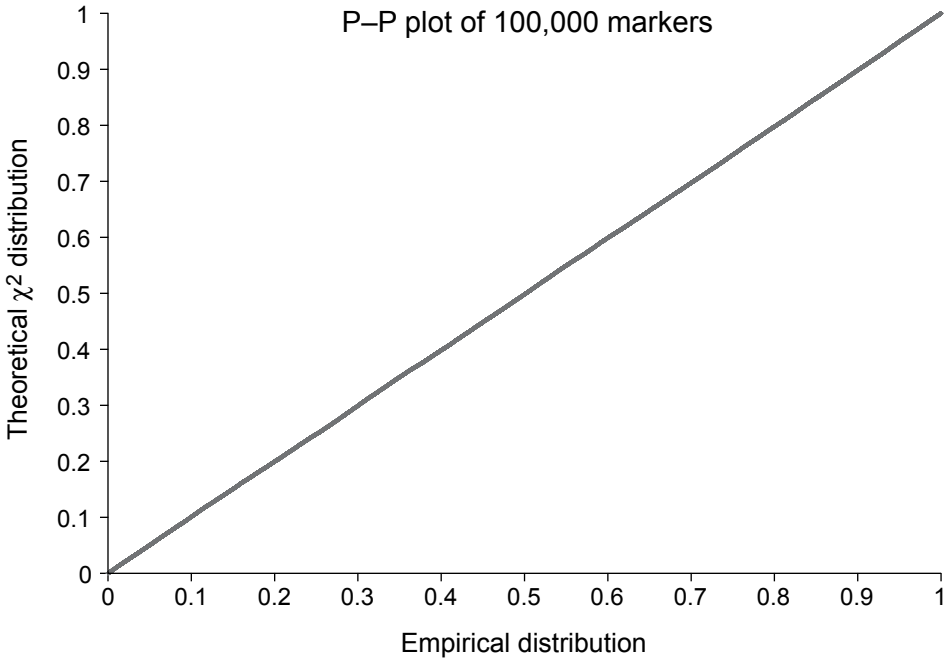

Supplementary Figure 1. P–P plot of EIGENSTRAT test statistics. The empirical distribution of EIGENSTRAT test statistics closely matches a theoretical χ^2 distribution.

Supplementary Table 1: Simulations using *K* axes of variation

	$K = 1$	$K = 2$	$K = 5$	$K = 10$
Random SNPs	0.0001	0.0001	0.0001	0.0001
Differentiated SNPs	0.0001	0.0001	0.0001	0.0001
Causal SNPs	0.4923	0.4916	0.4891	0.4860

Proportion of associations reported as significant by EIGENSTRAT adjusting along the top *K* axes of variation, for various values of *K*.

Supplementary Table 2: Simulations using *M* SNPs

M	False positive rate	Correlation of top axis
100	0.0826	68.4%
200	0.0079	80.9%
500	0.0016	90.8%
1,000	0.0007	94.8%
2,000	0.0002	97.4%
5,000	0.0001	99.0%
10,000	0.0001	99.5%
20,000	0.0001	99.7%
50,000	0.0001	99.9%
100,000	0.0001	99.9%

Proportion of associations falsely reported as significant by EIGENSTRAT at highly differentiated candidate SNPs for various values of *M*, the number of random SNPs used to infer population structure. The correlation between the top axis of variation and population membership across samples is also reported.

Supplementary Table 3: Simulations of Pritchard and Donnelly

(a)

M	χ^2	GC	SA	EIGENSTRAT
50	0.016	0.008	0.012	0.010
200	0.016	0.008	0.009	0.008
1,000	0.016	0.007	0.009	0.008

(b)

M	χ^2	GC	SA	EIGENSTRAT
50	0.449	0.334	0.362	0.432
200	0.449	0.351	0.285	0.430
1,000	0.449	0.355	0.281	0.433

Proportion of associations reported as significant by Armitage trend χ^2 statistic, Genomic Control (GC), Structured Association (SA) and EIGENSTRAT for various values of *M*, the number of random SNPs used to infer population structure. For each value of *M*, we report the proportion of SNPs at which each method reports a causal association with P-value less than 0.01. Results are given for (a) random candidate SNPs and (b) causal candidate SNPs.

Supplementary Table 4: Simulations with no stratification and *n* subpopulations

n	χ^2	GC	SA	EIGENSTRAT
3	0.446	0.435	0.279	0.449
5	0.443	0.435	0.225	0.452
10	0.450	0.440	0.166	0.448

Proportion of associations reported as significant by Armitage trend χ^2 statistic, Genomic Control (GC), Structured Association (SA) and EIGENSTRAT for causal candidate SNPs, assuming no systematic ancestry differences between cases and controls, and using 1,000 random SNPs to infer population structure. For each value of *n*, the number of subpopulations used to generate data, we report the proportion of causal candidate SNPs at which each method reports a causal association with P-value less than 0.01.

Supplementary Table 5: Stratification correction at rs10511418 using *M* SNPs

M	EIGENSTRAT	Correlation of top axis
1,000	29.85	78.6%
2,000	20.45	89.5%
5,000	16.10	95.7%
10,000	18.14	98.0%
20,000	14.45	99.0%
50,000	11.22	99.7%
100,000	12.27	99.95%
116,204	11.61	100.00%

Association statistics reported by EIGENSTRAT at the candidate SNP rs10511418 for various values of *M*, the number of random SNPs used to infer population structure. The correlation between the top axis of variation and the axis inferred using all SNPs is also reported.

Supplementary Note

1. Distribution of EIGENSTRAT statistics

As described in the main text, we computed EIGENSTRAT statistics for random candidate markers which were simulated under a model with systematic ancestry differences between cases and controls. We analyzed EIGENSTRAT statistics at 100,000 random candidate markers to verify that they follow a χ^2 distribution with 1 degree of freedom.

We first produced a P-P plot of our empirical distribution against a theoretical χ^2 distribution (**Supplementary Fig. 1** online). The P-P plot indicates the empirical versus theoretical proportion of values exceeding any given threshold. The visual fit seems entirely satisfactory.

We checked the tail of the distribution. Under a theoretical χ^2 distribution, we would expect 5% of values to exceed the threshold $X = 3.841$. In our empirical distribution, 5,116 of 100,000 values exceed this threshold. Under binomial sampling with frequency 5%, the two-sided P-value of this event is 0.09, which is not significant. We further checked the 5,116 points greater than $X = 3.841$ for a fit to the theoretical χ^2 distribution. A one-sample Kolmogorov-Smirnov test gives a P-value of 0.2, which is not significant.

2. Simulations of Pritchard and Donnelly

We compared EIGENSTRAT, Genomic Control and Structured Association by duplicating the simulations of Pritchard and Donnelly[1]. These simulations were based on 200 cases and 200 controls from three subpopulations, with the numbers of cases from each subpopulation

fixed at 50, 50, 100 and the number of controls at 66, 67, 67. Subpopulation allele frequencies were drawn from beta distributions with mean p and variances $0.01p(1-p)$, $0.02p(1-p)$ and $0.04p(1-p)$ respectively, where the ancestral allele frequency p is uniform on [0.1,0.9]. Results were averaged over 20 simulated data sets of M random SNPs (M = 50, 200 or 1,000) used to infer population structure before testing for association at 500 candidate loci each; the results for Genomic Control represent average rates after simulating 10,000 sets of M loci to estimate an inflation factor. For each method, a significant association is reported if a P-value less than 0.01 is obtained.

We first tested what proportion of random candidate SNPs generate a spurious association. Results for the χ^2 statistic, Genomic Control, Structured Association and EIGENSTRAT are reported in **Supplementary Table 3a** online, which shows that each method is effective in correcting for stratification at random candidate SNPs. These results are generally consistent with Pritchard and Donnelly[1], except that the false positive rate for the uncorrected χ^2 statistic is slightly lower; much of this can be attributed to our use of the Armitage trend χ^2 statistic[2] recommended by Devlin and Roeder[3] in lieu of an allelic 2×2 χ^2 statistic.

We next tested what proportion of causal candidate SNPs generate a true association, assuming a multiplicative risk model with a disease risk factor of 1.5 for the causal allele. Results for the uncorrected χ^2 statistic, Genomic Control, Structured Association and EIGENSTRAT are reported in **Supplementary Table 3b** online, which shows that EIGENSTRAT achieves superior power. Results for Genomic Control and Structured Association are virtually identical to those reported by Pritchard and Donnelly[1].

We repeated the analysis of causal candidate SNPs with both the number of cases and the number of controls from each subpopulation fixed at 66, 67, 67, eliminating the systematic ancestry differences between cases and controls. For M = 1,000, we see that Structured Association loses power in this scenario, while Genomic Control and EIGENSTRAT do not (**Supplementary Table 4** online). Thus, Genomic Control loses power only when stratification is present, whereas Structured Association loses power even in the absence of stratification. We repeated the experiment using 200/n cases and 200/n controls from each of n = 5 or n = 10 subpopulations, with subpopulation allele frequencies generated using mean p and variance equal to $cp(1–p)$ where $c = 0.05/n, 2\times0.05/n, \ldots, 0.05$ for the respective subpopulations. We found that Structured Association, but not Genomic Control or EIGENSTRAT, suffers a further loss in power as the number of dimensions of underlying population structure increases (**Supplementary Table 4** online).

3. Level of population structure in European American data set

In the European American data set, the ratio between the top eigenvalue of the covariance matrix χ and the average of all eigenvalues of χ is equal to 2.61. Letting N be the number of samples, this quantity is roughly equal to $1 + N F_{ST}$ in the case of two discrete subpopulations each of size $N/2$, assuming that F_{ST} is small and the number of SNPs is large (N.J.P., A.L.P. and D.R., unpublished data). Setting N = 449 (excluding outliers) and solving for F_{ST}, we obtain $F_{ST} = 0.0036$. Alternately, if we assume that each sample is an admixture of two populations, with admixture proportions uniform on [0,1], then the mean square difference in ancestry across all pairs of individuals decreases by a factor of 3, thus requiring a value of F_{ST}

= 0.0108 between the underlying populations. **Figure 2** suggests that the truth is somewhere in between these two scenarios.

4. Application to Campbell et al. data set

We applied our method to a data set of European American samples discordant for the height phenotype. Campbell et al.[4] demonstrated that the association between the Lactase (*LCT*) SNP and the height phenotype in this data set is spuriously due to stratification. After genotyping a subpanel of 368 samples at 178 markers, they further observed that Genomic Control and Structured Association find no evidence of any population structure in the subpanel, and thus cannot correct for this stratification.

Our usual definition of outliers as individuals whose ancestry was at least 6 standard deviations from the mean on one of the top 10 inferred axes of variation yielded one outlier individual who, strikingly, explained more than half of the variance along the top axis of variation. We observed that the outlier individual is an FY*O homozygote at the Duffy marker; this allele is fixed in sub-Saharan Africans but virtually absent elsewhere[5], suggesting that the outlier may have substantial African ancestry. Indeed, using 43 Campbell et al. markers which were also typed in our African American admixture map[6], we determined that the outlier has 85 +– 5% African ancestry. (In detail, we inferred the posterior distribution of the genomewide African vs. European ancestry of the individual using a model that computes the joint likelihood of the individual's genomewide ancestry θ and the true unobserved African and European allele frequencies A_i and E_i at each SNP i. The joint likelihood incorporates, at each SNP, both the probability of the individual's genotype and the probability of binomially

sampling the counts in the admixture map, conditional on θ, A_i and E_i. Inference was performed via MCMC[7].)

We computed top axes of variation with the genetic outlier removed from the data. We checked to see if any of the top 10 axes of variation inferred by EIGENSTRAT is correlated to grandparent-ancestry labels (northwest European, southeast European, or European American of unknown ancestry) collected by Campbell et al. as part of their study. We found that the second axis is 35% correlated to northwest European vs. southeast European ancestry in the 147 samples of known ancestry (P-value = 2×10^{-4} after correction for multiple hypothesis testing), and 28% correlated to southeast European ancestry vs. northwest European or unknown ancestry in all 368 samples (P-value = 8×10^{-7}). Although the correlation to true ancestry is highly significant, it is only a partial correlation. Thus, EIGENSTRAT fails to correct for stratification, reporting a chisq association statistic virtually equal to the uncorrected chisq statistic, yielding a P-value of 0.003 on the subpanel of 368 samples. Based on our simulations, EIGENSTRAT's inability to infer an accurate axis and correct for stratification using only 178 random background markers is not surprising; nevertheless, the method's ability to detect a previously undetected African outlier and to detect within-Europe population structure with partial accuracy is encouraging.

5. Practical Concerns

Linkage disequilibrium and choice of markers. Genome-wide data sets containing hundreds of thousands of markers are likely to exhibit substantial linkage disequilibrium (LD) between markers, even in the case of markers chosen to optimally tag all variation in the genome. Thus,

inferring population structure from the set of all markers has two potential problems. The first problem is that, due to varying levels of LD, some regions will have more redundant markers than others and will thus be overrepresented. The second problem is that strong LD at a given locus which affects many markers could result in an axis of variation which corresponds to genetic variation specifically at that locus, rather than to genome-wide ancestry. Nonetheless, we recommend inferring population structure using all markers. This recommendation is based on an analysis of HapMap[8] data which suggests that these potential problems will not affect results in practice, even on a data set with over 3 million markers. Specifically, we computed principal components using data from 45 Chinese and 45 Japanese individuals at 3,351,221 markers from Phase II HapMap. We repeated the computation keeping only 1 of every s markers, based on order along each chromosome, for various values of s (s = 1, 2, 5, 10). Correlation of the top axis of variation to population label (Chinese or Japanese) was 0.983, 0.983, 0.982 and 0.982 respectively in the four runs. Correlation between the top axis of variation from one run and the top axis of variation from a different run was greater than 0.999 for any pair of runs. Thus, different weighting of different regions due to different levels of LD does not significantly affect results. In each of the four runs, only one statistically significant axis of variation (N.J.P., A.L.P. and D.R., unpublished data) was observed. Thus, statistically significant axes arising from strong LD at a specific locus do not occur.

In theory, it is appropriate to exclude a candidate marker from the set of markers used to infer population structure and correct for possible stratification at the candidate marker. However, in genome-wide association studies involving hundreds of thousands of markers, it is not practical to separately infer population structure using all markers except the candidate marker,

for each choice of candidate marker. It is possible to alleviate this problem by inferring population structure for several different subsets of markers which each contain a majority of markers, with each candidate marker absent (along with nearby linked markers) from at least one of the subsets. However, in large data sets with > 100,000 markers, we instead recommend simply using all markers to infer population structure. Our results suggest that in data sets with > 100,000 markers, axes of variation and the resulting stratification correction are robust to inclusion or exclusion of candidate markers (see Results).

Assay effects. It is advisable to check for correlations between each axis of variation and assay variables such as amount of missing data per sample or plate membership. Large correlations are indicative of assay effects, which often occur in real data (see below).

Removal of outliers. We describe an outlier removal procedure which we believe to be reasonable (see Methods), but other outlier removal methods may be used instead. As data sets grow very large, sensitivity to detect outlier effects will increase. Because nearly all individuals are likely to have at least a small fraction of "unusual" ancestry, investigators will need to carefully weigh the desire to avoid samples with unusual ancestry against the desire to maximize power.

Cryptic relatedness. If cryptic relatedness is a concern, investigators may wish to preprocess the data by applying existing methods to detect cryptically related individuals[9] and selecting a maximal subset of unrelated individuals for subsequent analysis.

Residual confounding. If investigators are concerned as to whether residual confounding could remain after applying the EIGENSTRAT correction, a conservative and careful approach would be to test this by applying Genomic Control to the results of EIGENSTRAT. We note that applying EIGENSTRAT first to correct for population stratification will generally retain higher power and reduce spurious associations at highly differentiated SNPs, relative to the exclusive application of Genomic Control (see Results).

6. Extensions of our approach

Quantitative traits. Though we have focused on case/control phenotypes, extension of our method to quantitative traits is straightforward. One simply starts with phenotypes p_j which represent continuous-valued quantitative traits (instead of 0 or 1), then performs the adjustment for ancestry as described previously (see Methods).

Non-multiplicative disease models. Though we have focused on multiplicative disease models, extension of our method to other disease models is straightforward. Given possible genotypes aa:aA:AA, one can evaluate the dominant (respectively, recessive) model for the A allele by using genotype values 0:1:1 (respectively, 0:0:1) instead of the standard genotypes values 0:1:2. The adjustment for ancestry then follows as described previously (see Methods).

Targeted disease association studies. Though we have focused on genome-wide association studies in which a large number of SNPs are used to infer accurate axes of variation, there is also a desire to correct for stratification in targeted disease association studies in which a much smaller number of markers are genotyped. A possible plan in such studies is to genotype

samples at a preselected set of ancestry informative markers (in addition to candidate disease markers), then infer axes of variation using the ancestry informative markers. The choice of ancestry informative markers will depend on the population being studied. Our methods will likely prove useful in identifying ancestry informative markers for use in future targeted disease studies, however the identification and validation of ancestry informative markers requires caution, as a marker's observed informativeness in a particular sample (488 European Americans) may exceed true informativeness in the entire population (all European Americans) since sample sizes of only hundreds of individuals may be dominated by sampling error in the case of very subtle ancestry effects.

Nonlinear methods. Though we have focused on linear methods, axes of variation inferred by EIGENSTRAT could be incorporated as covariates in the context of nonlinear methods such as logistic regression.

Additional covariates. Additional covariates (such as age, gender or environmental factors) can be incorporated by computing residuals of (linear or nonlinear) regressions corresponding to the covariate(s). If the covariates are correlated to each other or to ancestry, multivariate regressions should be used, in lieu of the univariate regressions we have described.

Geographical information. If geographic information on the samples is available (for example, geographic information on within-Europe ancestry), analyzing the correspondence between axes of variation and geographic labels may facilitate a geographic interpretation of the axes of variation, and will indicate whether the geographic labels contain information

which is not already encoded in the axes of variation. If so, the geographic labels can be included as additional covariates (see above). We note that, in addition to ancestry, geographic labels could be correlated to environmental factors.

Ancestry-dependent risk. Though we have focused on the case where the risk conferred by a disease allele is independent of ancestry, EIGENSTRAT can be extended to model ancestry-dependent risk[1] by incorporating as covariates one or more ancestry-modulation terms, equal to the product of ancestry-adjusted genotype and ancestry along a given axis of variation.

7. Assay effects

For each of the top 10 axes of variation and for each subset of the set of 6 plates genotyped using the Affymetrix 100K array, we computed the correlation across samples between that axis and plate membership in that subset. We also computed the correlation across samples between each axis and the proportion of missing data. We observed that the third axis of variation is 58% correlated with membership in plates 1,2,3,6 versus 4,5 (P-value $< 10^{-12}$ after correcting for 320 hypotheses tested) and 38% correlated with the proportion of missing data (P-value $< 10^{-12}$ after correcting for 10 hypotheses tested). We determined that this axis is the result of a large number of SNPs that have both a higher rate of heterozygotes and a higher rate of missing data on plates 4,5 versus plates 1,2,3,6. These findings are suggestive of laboratory effects; possible explanations include differences in sample collection and preparation or differences in genotyping procedure. An important question is whether to remove from the data set the large number of SNPs, characterized by higher than average rate of missing data, which give rise to this axis; this risks discarding valuable data and missing true positive

findings[10]. We recommend this step only in the event of an axis attributable to laboratory effects that leads to a strong bias between cases and controls, risking an even greater power loss when correcting for this bias.

The existence of an axis of variation strongly correlated to the proportion of missing data highlights EIGENSTRAT's treatment of missing data: although axes of variation arising from missing data effects should be regarded as spurious if the underlying goal is to detect true population structure, correcting along such axes is entirely appropriate in the context of disease studies, in which the goal is to prevent spurious associations due either to population structure or to laboratory effects.

8. Supplemental References

[1] Pritchard, J.K. & Donnelly P. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227-37 (2001).
[2] Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375-86 (1955).
[3] Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
[4] Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **37**, 868-72 (2005).
[5] Hamblin, M., Thompson, E.E. & Di Rienzo, A. Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* **70**, 369-83 (2002).
[6] Smith, M.W. *et al*. A high-density admixture map for disease gene discovery in African Americans. *Am. J. Hum. Genet.* **74**, 1001-13 (2004).
[7] Chen, M., Shao, Q. & Ibrahim J.G. *Monte Carlo Methods in Bayesian Computation* (Springer, New York, NY, 2000).
[8] The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-320 (2005).
[9] Epstein, M.P., Duren, W.L. & Boehnke, M. Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* **67**, 1219-31 (2000).
[10] Clayton, D.G. *et al*. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243-6 (2005).