nature genetics

Principal components analysis corrects for stratification in genome-wide association studies

Alkes L Price[1,2], Nick J Patterson[2], Robert M Plenge[2,3], Michael E Weinblatt[3], Nancy A Shadick[3] & David Reich[1,2]

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies. We describe a method that enables explicit detection and correction of population stratification on a genome-wide scale. Our method uses principal components analysis to explicitly model ancestry differences between cases and controls. The resulting correction is specific to a candidate marker's variation in frequency across ancestral populations, minimizing spurious associations while maximizing power to detect true associations. Our simple, efficient approach can easily be applied to disease studies with hundreds of thousands of markers.

Population stratification—allele frequency differences between cases and controls due to systematic ancestry differences—can cause spurious associations in disease studies[1–8]. Because the effects of stratification vary in proportion to the number of samples[9], stratification will be an increasing problem in the large-scale association studies of the future, which will analyze thousands of samples in an effort to detect common genetic variants of weak effect.

The two prevailing methods for dealing with stratification are genomic control and structured association[9–14]. Although genomic control and structured association have proven useful in a variety of contexts, they have limitations. Genomic control corrects for stratification by adjusting association statistics at each marker by a uniform overall inflation factor. However, some markers differ in their allele frequencies across ancestral populations more than others. Thus, the uniform adjustment applied by genomic control may be insufficient at markers having unusually strong differentiation across ancestral populations and may be superfluous at markers devoid of such differentiation, leading to a loss in power. Structured association uses a program such as STRUCTURE[15] to assign the samples to discrete subpopulation clusters and then aggregates evidence of association within each cluster. If fractional membership in more than one cluster is allowed, the method cannot currently be applied to genome-wide association studies because of its intensive computational cost on large data sets. Furthermore, assignments of individuals to clusters are highly sensitive to the number of clusters, which is not well defined[14,16].

We propose a method to detect and correct for population stratification that addresses these limitations. Our method, EIGENSTRAT, consists of three steps (**Fig. 1**). First, we apply principal components analysis[17] to genotype data to infer continuous axes of genetic variation. Intuitively, the axes of variation reduce the data to a small number of dimensions, describing as much variability as possible; they are defined as the top eigenvectors of a covariance matrix between samples (see Methods). In data sets with ancestry differences between samples, axes of variation often have a geographic interpretation: for example, an axis describing a northwest-southeast cline in Europe would have values that gradually range from positive for samples from northwest Europe, to near zero in central Europe, to negative in southeast Europe. Second, we continuously adjust genotypes and phenotypes by amounts attributable to ancestry along each axis, via computing residuals of linear regressions; intuitively, this creates a virtual set of matched cases and controls. Third, we compute association statistics using ancestry-adjusted genotypes and phenotypes.

The EIGENSTRAT method has arisen out of our systematic exploration of the use of principal components analysis in a more general population genetic context. Principal components analysis was originally applied to genetic data to infer worldwide axes of human genetic variation from the allele frequencies of various populations[18,19]. We have further developed this approach in a parallel paper (N.J.P., A.L.P. and D.R., unpublished data), focusing instead on individual genotype data and placing the method on a firm statistical footing by rigorously assigning statistical significance to each axis of variation[20–22]. EIGENSTRAT applies this toolkit to analyze population structure in the context of disease studies.

Correcting for stratification using continuous axes of variation has several advantages. Continuous axes provide the most useful description of within-continent genetic variation, according to recent studies[23]. Because our continuous axes are constructed to be orthogonal, results are insensitive to the number of axes inferred, as we verify

[1]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [2]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA. [3]Division of Rheumatology, Immunology and Allergy, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. Correspondence should be addressed to A.L.P. (aprice@broad.mit.edu).

Received 23 March; accepted 21 June; published online 23 July 2006; doi:10.1038/ng1847

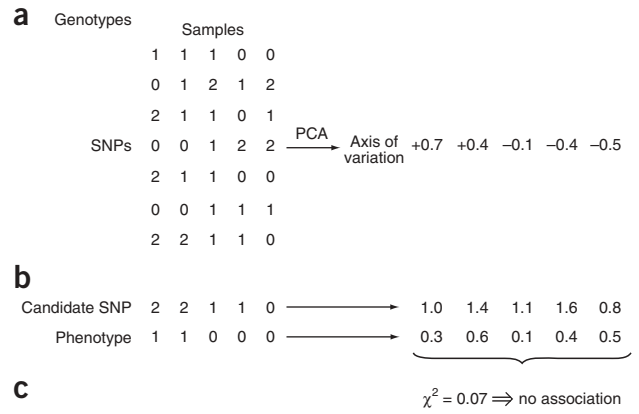

Figure 1 The EIGENSTRAT algorithm, illustrated on simulated data. (**a**) Principal components analysis is applied to genotype data to infer continuous axes of genetic variation; a single axis of variation is illustrated here. (**b**) Genotype at a candidate SNP and phenotype are continuously adjusted by amounts attributable to ancestry along each axis, removing all correlations to ancestry. (**c**) After ancestry adjustment, an association statistic between genotype at the candidate SNP and phenotype shows no significant association.

empirically. In addition, our approach is computationally tractable on a genome-wide scale. Here, we test EIGENSTRAT on simulated genome-wide disease studies and apply it to a real data set of 488 European American samples genotyped at 116,204 SNPs.

RESULTS

Simulated disease studies

Our simulated disease studies are similar to those carried out in ref. 24 in a comparison between structured association and genomic control methods, except that our studies are many orders of magnitude larger. We generated data at 100,000 random SNPs for 500 cases and 500 controls, with 60% of the cases and 40% of the controls sampled from population 1 and the remaining cases and controls sampled from population 2. Allele frequencies for population 1 and population 2 were generated using the Balding-Nichols model[25] with $F_{ST} = 0.01$ (see Methods) ($F_{ST} = 0.01$ is typical of differentiation between divergent European populations[26,27] and leads to allele frequency differences under 0.10 for typical common SNPs).

We used principal components analysis to infer orthogonal axes of continuous variation for each of ten independent data sets of 100,000 random SNPs (see Methods). Because the data contain only one added dimension of population structure, we expect only the top axis of variation to be of interest. However, in order to test the sensitivity of EIGENSTRAT to the number of axes of variation used, we actually inferred the top ten axes of variation. We checked whether the top axis of variation accurately tracks membership in population 1 versus population 2, and we found that the coordinate along the top axis is 99.9% correlated to population membership across samples; the remaining axes are artifacts of sampling.

We simulated three categories of candidate SNPs to compare the effectiveness of different stratification correction methods. For the first category (random SNPs with no association to disease), we again used the Balding-Nichols model with $F_{ST} = 0.01$ (see Methods). For the second category (differentiated SNPs with no association), we assumed population allele frequencies of 0.80 for population 1 and 0.20 for population 2. This category was motivated by previous studies of a SNP in the lactase (*LCT*) gene, which was shown to be spuriously associated to the height phenotype in European Americans because of stratification[6]; this SNP varies in frequency from 0.20 to 0.80 in European populations because of positive selection[28]. For the third category (causal SNPs), we used the Balding-Nichols model with $F_{ST} = 0.01$ and a multiplicative disease risk model with a relative risk of 1.5 for the causal allele (see Methods). For each of the ten data sets of 100,000 random SNPs, we simulated 1,000,000 candidate SNPs in each of the three categories, and computed association statistics using three methods: (i) Armitage trend χ^2 statistic[29] with no stratification correction, (ii) genomic control using random SNPs to infer an inflation factor, and (iii) EIGENSTRAT using random SNPs to infer ten axes of variation (see Methods). Association statistics producing a P value < 0.0001 were reported as significant.

Our simulations show that by explicitly modeling the ancestry of cases and controls, EIGENSTRAT achieves an equal or lower rate of false positive associations and achieves superior sensitivity to detect true associations, relative to genomic control. For random candidate SNPs, both methods reduce the false positive rate to the expected value for a $P < 0.0001$ cutoff (**Table 1**, top); we have verified that the EIGENSTRAT statistic is χ^2 distributed with 1 degree of freedom (**Supplementary Note** and **Supplementary Fig. 1** online). For highly differentiated candidate SNPs, genomic control is likely to produce false positives, whereas EIGENSTRAT still perfectly corrects for stratification (**Table 1**, top). On the other hand, for causal disease SNPs, EIGENSTRAT has higher power (nearly equivalent to the statistic uncorrected for stratification): 49% versus 30% for genomic control (**Table 1**, top). These results confirm the hypothesis that the uniform adjustment of genomic control is insufficient at markers that show unusually strong differentiation across ancestral populations and is superfluous at markers devoid of such differentiation, leading to a loss in power.

A possible concern with our EIGENSTRAT simulations is that we adjusted for ancestry along the top ten axes of variation. As this data set contains only one added dimension of population structure, however, we expect only the top axis of variation to be of interest. We tested the sensitivity of EIGENSTRAT to the number (K) of axes of variation used (**Supplementary Table 1** online). We see that in each SNP category, results are virtually identical for $K = 1$, 2, 5 or 10. This implies that EIGENSTRAT results are not sensitive to the number of axes of variation used, as long as there is a sufficient number of axes to capture true population structure effects. This is a natural consequence of the fact that the axes are orthogonal by definition; for example, allowing $K > 1$ has no effect whatsoever on the top axis. In practice, we have chosen $K = 10$ as a default value for running EIGENSTRAT; a more rigorous approach is to set K equal to the number of statistically significant axes of variation (N.J.P., A.L.P. and D.R., unpublished data).

To explore how EIGENSTRAT performs under more extreme mismatching of cases and controls, we next simulated a disease study in which some of the cases have no matching control: we sampled 50% of the cases and 0% of the controls from population 1 and the remaining cases and all controls from population 2. This simulates a situation in which, for example, European American cases are compared with controls from a single European country. For each of ten data sets of 100,000 random SNPs used to infer population structure, we computed association statistics at 1,000,000 candidate SNPs in each of three categories using the Armitage trend χ^2 statistic, genomic control and EIGENSTRAT (**Table 1**, center). For random candidate SNPs, stratification is more severe than before but is still perfectly corrected for by genomic control and EIGENSTRAT. For

Table 1 Proportion of associations reported as significant by Armitage trend χ^2 statistic, genomic control and EIGENSTRAT

	χ^2	Genomic control	EIGENSTRAT
Discrete subpopulations with moderate ancestry differences between cases and controls			
Random SNPs	0.0008	0.0001	0.0001
Differentiated SNPs	0.8520	0.5007	0.0001
Causal SNPs	0.5117	0.2980	0.4860
Discrete subpopulations with more extreme ancestry differences between cases and controls			
Random SNPs	0.0365	0.0001	0.0001
Differentiated SNPs	1.0000	1.0000	0.0001
Causal SNPs	0.5073	0.0342	0.2666
Admixed population with ancestry differences between cases and controls based on ancestry risk *r*			
$r = 2$			
Random SNPs	0.0002	0.0001	0.0001
Differentiated SNPs	0.1600	0.1004	0.0001
Causal SNPs	0.5180	0.4367	0.4863
$r = 3$			
Random SNPs	0.0007	0.0001	0.0001
Differentiated SNPs	0.7757	0.5553	0.0001
Causal SNPs	0.5158	0.3328	0.4442

We report the proportion of candidate SNPs in each category at which each method reports a significant association with $P < 0.0001$. Each row of the table reflects an average across 1,000,000 candidate SNPs in each of ten independent simulations. Results are given for three types of stratification (see text): (i) discrete subpopulations with moderate ancestry differences between cases and controls, (ii) discrete subpopulations with more extreme ancestry differences between cases and controls and (iii) admixed population with ancestry differences between cases and controls based on ancestry risk *r*.

highly differentiated SNPs, stratification is now guaranteed to generate a false positive association that genomic control cannot correct for, whereas EIGENSTRAT again achieves perfect stratification correction. For causal SNPs, genomic control loses nearly all power, whereas EIGENSTRAT suffers a partial power loss. These results again confirm the hypothesis that the uniform adjustment of genomic control is insufficient at markers showing unusually strong differentiation across ancestral populations and is superfluous at markers devoid of such differentiation, leading to a loss in power. We further examined the power attained by EIGENSTRAT and determined that it is identical to the power achieved by computing the uncorrected χ^2 statistic using only the 250 cases and 500 controls from population 2. Intuitively, the ancestry adjustment of EIGENSTRAT effectively removes the 250 cases from population 1 from the study, which is exactly what is supposed to happen given the lack of matching controls. Thus, given a specified set of cases and controls with no prior knowledge of ancestry, EIGENSTRAT will implicitly and automatically match cases and controls to extract the maximum possible amount of power from the data while avoiding false positives due to stratification. We caution that this does not obviate the need to carefully match cases and controls when designing a disease study: in the current example, a more closely matched set of 500 cases and 500 controls would have achieved superior power to detect true associations.

We next explored how EIGENSTRAT performs in an admixed population. We sampled individuals with ancestry proportions a from population 1 and $(1 - a)$ from population 2, with a uniformly distributed on [0,1] and case/control status simulated using disease risk proportional to r^a, based on ancestry risk r (see Methods). For each of ten data sets of 100,000 random SNPs used to infer population structure, we computed association statistics at 1,000,000 candidate SNPs in each of three categories using the Armitage trend χ^2 statistic, genomic control and EIGENSTRAT. Results for $r = 2$ and $r = 3$ are reported in **Table 1** (bottom). Once again, EIGENSTRAT is far more effective than genomic control in correcting for stratification at highly differentiated SNPs and achieves higher power at causal SNPs. However, even EIGENSTRAT incurs a slight power loss at causal SNPs: because its ancestry inference is extremely accurate—the top axis is 99.8% correlated to true ancestry for either value of r—we hypothesize that this power loss may be an unavoidable consequence of imperfect matching of cases and controls, analogous to the unavoidable power loss of **Table 1**, center.

Finally, we explored how much data is needed to accurately infer population structure and correct for stratification. (We note that this is greater than the amount of data needed to merely detect the existence of population structure). There are many variables of interest, but we restricted our attention to the number of samples (N), the number of random SNPs (M) used to infer population structure, and F_{ST}. All other variables were fixed as in our original simulations; in particular, 60% of cases and 40% of controls were sampled from population 1, with the remainder from population 2. First, we tried altering the number N of samples and found that simulations at $N = 100$, 200, 500 or 1,000 each yielded a top axis of variation that is 99.9% correlated to population membership across samples, with EIGENSTRAT effectively correcting for stratification in each case, even at highly differentiated candidate SNPs (data not shown). Thus, effective stratification correction is insensitive to the number of samples. We then fixed the sample size N at 1,000 and ran simulations at various values of M, the number of random SNPs used to infer population structure. We focused our attention on highly differentiated candidate SNPs, which are particularly likely to produce false positive associations, as demonstrated above. False positive rates for these SNPs, along with the correlation between the top axis of variation and population membership across samples, are reported in **Supplementary Table 2** online. We see that EIGENSTRAT has difficulty inferring a perfectly accurate axis of variation when $M < 5{,}000$, leading to incomplete stratification correction. On the

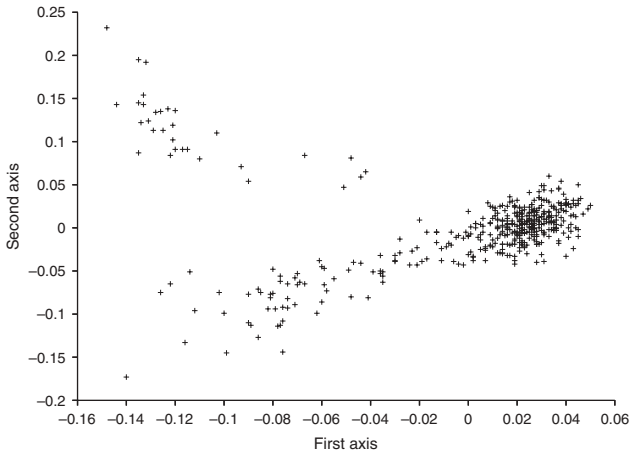

Figure 2 The top two axes of variation of European American samples. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, with a fraction of the samples showing southeast European ancestry (first axis < 0; see text). It follows that the second axis separates two southeast European subpopulations.

other hand, stratification correction at random candidate SNPs is effective for $M \geq 200$ (data not shown), even when inference of axes of variation is slightly inaccurate. We repeated this analysis for values of $F_{ST} < 0.01$ and found that fully correcting for stratification at highly differentiated SNPs requires 20,000 SNPs for $F_{ST} = 0.005$; 50,000 SNPs for $F_{ST} = 0.002$ and 100,000 SNPs for $F_{ST} = 0.001$. Thus, genome scans on European Americans with hundreds of thousands of SNPs will be able to detect and correct for stratification across closely related European populations, even in the case of highly differentiated SNPs.

We were unable to include the structured association method in the above comparisons because of its intensive computational cost. For example, execution of the STRAT program[12] on a data set with 1,000 samples and 110 markers requires 72 hours of computation time[14]; our simulations are many orders of magnitude larger than this. (We note that EIGENSTRAT runs on a data set with 1,000 samples and 100,000 markers in less than 15 min.) Thus, we compared EIGENSTRAT, genomic control and structured association by duplicating the much smaller simulations of ref. 24. We observed that the three methods achieve similar success in correcting for stratification at random candidate SNPs, and that EIGENSTRAT achieves superior power to genomic control or structured association in detecting true associations at causal SNPs (**Supplementary Note** and **Supplementary Tables 3** and **4** online).

European American data set

We applied our method to a data set of 488 European Americans genotyped on an Affymetrix platform containing 116,204 SNPs as part of an ongoing disease study (see Methods). We removed outlier individuals from all analyses (see Methods). We used principal components analysis to infer the top axes of variation. Statistical methods that we have developed (N.J.P., A.L.P. and D.R., unpublished data) indicate that there are ten statistically significant axes ($P < 0.01$ for each). The top two axes ($P < 10^{-12}$ for each) are shown in **Figure 2**. Interestingly, we observed both continuous and discrete genetic effects. We hypothesize that the first axis reflects genetic variation between northwest and southeast Europe, based on its correlation with lactase persistence (see below). The sign of the correlation implies that a fraction of the samples provide the bulk of southeast European ancestry, and the second axis separates two southeast European subpopulations (**Fig. 2**).

We conducted an association study for the lactase persistence phenotype. We chose this phenotype because it correlates with within-Europe ancestry[28] and can be inferred from the data, as it is 100% associated to genotype at the *LCT* gene[30]. Although the SNP perfectly associated to this phenotype was not one of the 116,204 SNPs genotyped, the nearby SNP rs3769005 was genotyped and is 90% correlated to the perfectly associated SNP in European samples from HapMap[31]; thus, the lactase persistence phenotype can be inferred with reasonable accuracy from the genotype at this SNP. We computed association statistics using the Armitage trend χ^2 statistic[29] (see Methods), correcting P values for the number of SNPs tested. As expected, a large number of SNPs on chromosome 2 showed a highly significant association, reflecting the strong selective sweep that occurred at the *LCT* gene[28]. We thus restricted our subsequent analysis to SNPs outside chromosome 2. Four SNPs showed significant associations (**Table 2**); we hypothesized that these might be due to stratification.

We first attempted to correct for stratification using genomic control[10]. We computed a genome-wide inflation factor of $\lambda = 1.43$ (see Methods). After dividing the uncorrected χ^2 statistics by this quantity, the top associated SNP remained significant (**Table 2**). We then ran EIGENSTRAT, which did not report any significant association either at any of the originally associated SNPs (**Table 2**) or at any other SNP. Notably, the top axis of variation inferred by EIGENSTRAT is strongly correlated both to the four originally associated SNPs and to the lactase persistence phenotype—presumably because lactase persistence varies between northwest and southeast Europe[6,28]. Thus, correcting along this axis addresses the spurious associations.

In the above analysis, the set of SNPs used by EIGENSTRAT to infer axes of variation included the candidate SNPs of interest. This raises the question of whether the axes of variation could be biased by the inclusion of these SNPs. To address this question, we reran EIGENSTRAT with the four candidate SNPs, the *LCT* region SNP rs3769005, and all SNPs within 5 Mb of each of those SNPs excluded when inferring axes of variation. Results were essentially unchanged. This suggests that the method is robust to inclusion or exclusion of candidate SNPs when inferring axes of variation in large data sets.

To test how many random SNPs are needed to effectively correct for stratification, we reran EIGENSTRAT using a subset of M SNPs, for various values of M. The association reported by EIGENSTRAT at the candidate SNP rs10511418 is reported in **Supplementary Table 5** online, together with the correlation of the top axis of variation to the top axis inferred using all SNPs. We see that EIGENSTRAT has

Table 2 SNPs outside chromosome 2 that are spuriously associated to the lactase persistence phenotype

SNP	χ^2	Genomic control	EIGENSTRAT
rs10511418	**45.11 (0.0000022)**	**31.55 (0.0023)**	11.57 (1.00)
rs2493880	**26.12 (0.037)**	18.27 (0.89)	8.17 (1.00)
rs4306808	**26.04 (0.039)**	18.21 (0.90)	8.83 (1.00)
rs2243133	**25.60 (0.049)**	17.90 (0.93)	5.88 (1.00)

We list association statistics (P values in parentheses, corrected for 116,204 SNPs tested) for three methods: Armitage trend χ^2 statistic, genomic control and EIGENSTRAT. Significant associations ($P < 0.05$) are listed in bold.

difficulty inferring a perfectly accurate axis of variation when $M <$ 20,000, leading to incomplete stratification correction. Thus, the number of SNPs needed is larger than we had determined in simulation for $F_{ST} = 0.01$ (**Supplementary Table 2**). We hypothesized that because European Americans are an admixed population, the effective F_{ST} might be smaller than 0.01. Indeed, the level of population structure is similar to what would be observed in the case of two discrete subpopulations with $F_{ST} = 0.004$ (**Supplementary Note** online). Given these results, it is not surprising that EIGENSTRAT fails to correct for stratification in the data set from ref. 6 of 368 European Americans typed at 178 markers (**Supplementary Note**).

DISCUSSION

We have described a new method to detect and correct for population stratification that explicitly models ancestry differences between cases and controls along continuous axes of variation. Practical issues such as linkage disequilibrium between markers and extending to quantitative traits are discussed in the **Supplementary Note**. The EIGENSTRAT method outperforms prevailing methods on simulated and real data sets and can easily be applied to disease studies with hundreds of thousands of markers. The method should be particularly valuable in disease studies involving European Americans, as genetic risk has already been reported to vary across Europe for numerous diseases[32–36].

Although EIGENSTRAT is a robust and powerful method for correcting for stratification, it is not a panacea, and researchers should adhere to the principles of careful experimental design, matching the ancestry and laboratory treatment of cases and controls to the fullest extent possible. If violation of these principles leads to a strong bias between cases and controls, EIGENSTRAT is likely to detect the bias; however, a loss in power will inevitably result, because any putative disease association will resemble an unusually strong instance of the bias. Though our focus here has been on ancestry effects, a recent study has suggested that differences in laboratory treatment among samples is a pervasive issue that will often outweigh the effects of population stratification[37]. These effects are so common that it is not surprising if assay artifacts are detected by our methods, especially in a large study where our sensitivity is high. Indeed, in the European American data set described here, the top two axes of variation describe ancestry effects, but subtle evidence of differences in laboratory treatment among samples is detected in the third axis (**Supplementary Note**). EIGENSTRAT's ability to explicitly address such subtle effects is an encouraging prospect.

METHODS

Simulated disease studies. Following ref. 24, simulated data for populations 1 and 2 with a specified value of F_{ST} were generated using the Balding-Nichols model[25]. For each SNP, an ancestral population allele frequency p was drawn from the uniform distribution on [0.1,0.9]. The allele frequencies for populations 1 and 2 were each drawn from a beta distribution with parameters $p(1 - F_{ST})/F_{ST}$ and $(1 - p)(1 - F_{ST})/F_{ST}$. This distribution has mean p and variance $F_{ST}\,p(1 - p)$. It follows that the quantity F_{ST} agrees with its usual measure of genetic distance between two populations[26,38]. The risk model with a relative risk of R for the causal allele was implemented as follows: for individuals from population l with population allele frequency p_l, control individuals were assigned genotype 0, 1 or 2 with probabilities $(1 - p_l)^2$, $2p_l(1 - p_l)$, or p_l^2, respectively, and case individuals were assigned genotype 0, 1 or 2 with relative probabilities $(1 - p_l)^2$, $2Rp_l(1 - p_l)$, or $R^2p_l^2$, respectively, each scaled by $(1 - p_l)^2 + 2Rp_l(1 - p_l) + R^2p_l^2$.

Simulated disease studies in an admixed population. Case/control status for individuals with ancestry proportions a from population 1 and $(1 - a)$ from population 2 were simulated using disease risk proportional to r^a, based on ancestry risk r. To insure an average value of 0.5 across possible values of a, the probability of disease was set to 0.5 $\log(r)\, r^a/(r - 1)$. The risk model with a relative risk of R for the causal allele was implemented as above, replacing p_l with $ap_1 + (1 - a)p_2$, the allele frequency conditional on an individual's ancestry proportion a.

Inference of axes of variation. Let g_{ij} be a matrix of genotypes for SNP i and individual j, where $i = 1$ to M and $j = 1$ to N. We subtract the row mean $\mu_i = (\Sigma_j g_{ij})/N$ from each entry in row i to obtain a matrix with row sums equal to 0; missing entries are excluded from the computation of μ_i and are subsequently set to 0. We then normalize row i by dividing each entry by $\sqrt{p_i(1 - p_i)}$, where p_i is a posterior estimate of the unobserved underlying allele frequency of SNP i defined by $p_i = (1 + \Sigma_j g_{ij})/(2 + 2N)$, with missing entries excluded from the computation. We denote the resulting matrix X. We compute an $N \times N$ covariance matrix Ψ of individuals, where $\Psi_{jj'}$ is defined to be the covariance of column j and column j' of X. We define the kth axis of variation to be the kth eigenvector of Ψ (that is, the eigenvector with kth largest eigenvalue). Thus, the ancestry a_{jk} of individual j along the kth axis of variation equals coordinate j of the kth eigenvector. We note that eigenvectors are orthonormal by definition; thus, $\Sigma_j a_{jk} = 0$, $\Sigma_j a_{jk}^2 = 1$ and $\Sigma_j a_{jk} a_{jk'} = 0$ for distinct axes k and k'. In particular, the ancestry values a_{jk} can be either positive or negative and should not be interpreted as percentages. Each axis is invariant to multiplying by a factor of –1, which does not change its interpretation.

The above procedure is motivated by the decomposition $X = USV^T$, where U is an $M \times N$ matrix whose kth column contains coordinates of each SNP along the kth principal component, S is a diagonal matrix of singular values and V is an $N \times N$ matrix whose kth column contains ancestries a_{jk} of each individual j along the kth principal component. It follows that $X^TX = VS^2V^T$; thus, the columns of V are the eigenvectors of the matrix X^TX. The matrix X^TX is equivalent up to a constant to the covariance matrix Ψ, and the matrix S^2 of squared singular values is equivalent up to a constant to the diagonal matrix of eigenvalues of Ψ.

Computation of Armitage trend χ^2 statistic. As discussed in ref. 10, the Armitage trend χ^2 statistic[29] is more appropriate than a χ^2 statistic obtained from a 2×2 allelic or 2×3 genotypic χ^2 table. The Armitage trend χ^2 statistic is equal to N times the squared correlation between genotype (0, 1 or 2) and phenotype (0 or 1), where N is the number of samples. Though we believe that $(N - 1)$ times the squared correlation is a more appropriate statistic, we used the original definition of Armitage in all of our calculations.

Computation of genome-wide χ^2 inflation factor for genomic control. As described in ref. 10, a robust genome-wide inflation factor λ is computed as the median χ^2 statistic divided by 0.456, the predicted median χ^2 if there is no inflation.

Adjustment of genotypes and phenotypes using axes of variation. Let g_{ij} be the genotype of individual j ($g_{ij} = 0$, 1 or 2) at SNP i, and let a_j be the ancestry of individual j along a given axis of variation. We define $g_{ij,\ adjusted} = g_{ij} - \gamma_i a_j$, where $\gamma_i = \Sigma_j a_j g_{ij}/\Sigma_j a_j^2$ is a regression coefficient for ancestry predicting genotype across individuals j with valid genotypes at SNP i. (If there are no missing genotypes at SNP i, then $\Sigma_j a_j^2 = 1$ by definition, and thus $\gamma_i = \Sigma_j a_j g_{ij}$.) A similar adjustment is performed for each axis of variation. The adjustment of phenotype p_j is analogous. We note that the procedure we have described is equivalent to using the axes of variation as covariates in a multilinear regression, but is simpler because the axes of variation are orthogonal, and thus the adjustments can be performed independently for each axis of variation.

Computation of χ^2 statistic using ancestry-adjusted genotypes and phenotypes. Our χ^2 statistic is equal to $(N - K - 1)$ times the squared correlation between ancestry-adjusted genotype and ancestry-adjusted phenotype, where N is the number of samples and K is the number of axes of variation used to adjust for ancestry. This is a generalization of the Armitage trend χ^2 statistic[29] for discrete genotypes and phenotypes (see above). The idea is to test for correlation between two vectors which have been projected into a space of reduced dimension, namely the space orthogonal to the K axes of variation.

We note that using more axes of variation than necessary will in theory lead to a loss in power; however, for $K << N$ the effect will be minimal.

European American data set. The data set consisted of 488 individuals chosen from the Brigham Rheumatoid Arthritis Sequential Study (BRASS), an ongoing single-center cohort of subjects seen at the Brigham & Women's Hospital Arthritis Center. The individuals chosen were unrelated and self-described as white, suggesting European ancestry. The data set exclusively contained individuals with rheumatoid arthritis (as diagnosed by a board-certified rheumatologist), as cohort specimens collected at enrollment as controls were not yet genotyped. The cohort is predominantly female (82%) with a mean age of 57 years and an average disease duration of 15 years[39]. Sample collection was approved by the Human Research Committee of Brigham and Women's Hospital, and informed consent was obtained from all subjects. Genotyping was performed using the Affymetrix GeneChip Mapping Array containing 116,204 SNPs. Samples were processed in a 96-well plate format using Biomek FX robotics according to the manufacturer's protocol. Individual samples with <90% genotype call rates or with more than two Hind-Xba discrepancies were excluded from the data set.

We identified 39 outlier individuals and removed them from all analyses, keeping 449 individuals. Although EIGENSTRAT is designed to automatically and implicitly match cases and controls, we view the removal of outliers as a prudent step, as outliers may skew other axes of interest because of the orthogonality property, obfuscating their interpretation and potentially hindering their detection in the case of subtle effects. Outliers were defined as individuals whose ancestry was at least 6 standard deviations from the mean on one of the top ten inferred axes of variation. This trimming step was iteratively applied, removing 39 outliers in five iterations. After correcting for the total of 23,800 hypotheses tested, each outlier was still highly statistically significant ($P < 5 \times 10^{-5}$). We note that the alternative single-iteration approach of removing individuals whose ancestry is at least 6 standard deviations from the mean on any statistically significant axis of variation (N.J.P., A.L.P. and D.R., unpublished data) produces similar results.

URLs. Software for running EIGENSTRAT on a Linux platform is available at http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm.

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

ACKNOWLEDGMENTS

The authors are grateful to B. Blumenstiel, M. DeFelice, M. Parkin, R. Barry, W. Winslow, C. Healy and S. Gabriel for generation of the Affymetrix genotype data. We are grateful to the BRASS study participants, the BRASS study team, and our rheumatology colleagues at the Brigham and Women's Hospital Arthritis Center. We thank C. Campbell and J. Hirschhorn for helpful comments and sharing data from their paper[6]. The BRASS study was supported by a grant from Millennium Pharmaceuticals. D.R. is supported in part by a Burroughs Wellcome Career Development Award in the Biomedical Sciences.

COMPETING FINANCIAL INTERESTS

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

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Lander, E.S. & Schork, N.J. Genetic dissection of complex traits. *Science* **265**, 2037–2048 (1994).
2. Lohmueller, K. *et al.* Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.* **33**, 177–182 (2003).
3. Freedman, M. *et al.* Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* **36**, 388–393 (2004).
4. Marchini, J. *et al.* The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).
5. Helgason, A. *et al.* An Icelandic example of the impact of population structure on association studies. *Nat. Genet.* **37**, 90–95 (2005).
6. Campbell, C.D. *et al.* Demonstrating stratification in a European American population. *Nat. Genet.* **37**, 868–872 (2005).
7. Hirschhorn, J.N. & Daly, M.J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
8. Thomas, D.C. *et al.* Recent developments in genomewide association scans: a workshop summary and review. *Am. J. Hum. Genet.* **77**, 337–345 (2005).
9. Reich, D. & Goldstein, D. Detecting association in a case-control study while allowing for population stratification. *Genet. Epidemiol.* **20**, 4–16 (2001).
10. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
11. Devlin, B. *et al.* Genomic control to the extreme. *Nat. Genet.* **36**, 1129–1130 (2004).
12. Pritchard, J.K. *et al.* Association mapping in structured populations. *Am. J. Hum. Genet.* **67**, 170–181 (2000).
13. Satten, G. *et al.* Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* **68**, 466–477 (2001).
14. Setakis, E., Stirnadel, H. & Balding, D.J. Logistic regression protects against population structure in genetic association studies. *Genome Res.* **16**, 290–296 (2006).
15. Pritchard, J.K. *et al.* Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
16. Serre, D. & Paabo, S. Evidence for gradients of human genetic diversity within and among continents. *Genome Res.* **14**, 1679–1685 (2004).
17. Jackson, J.E. *A User's Guide to Principal Components* (John Wiley & Sons, New York, 2003).
18. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
19. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. Demic expansions and human evolution. *Science* **259**, 639–646 (1993).
20. Johnstone, I. On the distribution of the largest eigenvalue in principal components analysis. *Ann. Stat.* **29**, 295–327 (2001).
21. Soshnikov, A. A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices. *J. Stat. Phys.* **108**, 1033–1056 (2002).
22. Baik, J., Ben Arous, G. & Peche, S. Phase transition of the largest eigenvalue for non-null complex sample covariance matrices. *Ann. Probab.* **33**, 1643–1697 (2005).
23. Rosenberg, N.A. *et al.* Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* **1**, 660–671 (2005).
24. Pritchard, J.K. & Donnelly, P. Case-control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
25. Balding, D.J. & Nichols, R.A. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identify and paternity. *Genetica* **96**, 3–12 (1995).
26. Cavalli-Sforza, L.L., Menozzi, P. & Piazza, A. *The History and Geography of Human Genes* (Princeton Univ. Press, Princeton, New Jersey, 1994).
27. Nicholson, G. *et al.* Assessing population differentiation and isolation from single-nucleotide polymorphism data. *J. R. Statist. Soc. (B)* **64**, 695–715 (2002).
28. Bersaglieri, T. *et al.* Genetic signatures of strong recent positive selection at the lactase gene. *Am. J. Hum. Genet.* **74**, 1111–1120 (2004).
29. Armitage, P. Tests for linear trends in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).
30. Enattah, N.S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
31. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
32. Cimmino, M.A. *et al.* Prevalence of rheumatoid arthritis in Italy: the Chiavari study. *Ann. Rheum. Dis.* **57**, 315–318 (1998).
33. Rosati, G. The prevalence of multiple sclerosis in the world: an update. *Neurol. Sci.* **22**, 117–139 (2001).
34. Panza, F. *et al.* Shifts in angiotensin I converting enzyme insertion allele frequency across Europe: implications for Alzheimer's disease risk. *J. Neurol. Neurosurg. Psychiatry* **74**, 1159–1161 (2003).
35. Bernardi, F. *et al.* Contribution of factor VII genotype to activated FVII levels. Differences in genotype frequencies between northern and southern European populations. *Arterioscler. Thromb. Vasc. Biol.* **17**, 2548–2553 (1997).
36. Angastiniotis, M. & Modell, B. Global epidemiology of hemoglobin disorders. *Ann. NY Acad. Sci.* **850**, 251–269 (1998).
37. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
38. Wright, S. The genetical structure of populations. *Ann. Eugen.* **15**, 323–354 (1951).
39. Benito-Garcia, E. *et al.* Dietary caffeine does not affect methotrexate efficacy in rheumatoid arthritis patients. *J. Rheumatol.* (in the press).