Theoretical Formulation of Principal Components Analysis to Detect and Correct for Population Stratification

Jianzhong Ma*, Christopher I. Amos

Department of Epidemiology, The University of Texas M. D. Anderson Cancer Center, Houston, Texas, United States of America

Abstract

The Eigenstrat method, based on principal components analysis (PCA), is commonly used both to quantify population relationships in population genetics and to correct for population stratification in genome-wide association studies. However, it can be difficult to make appropriate inference about population relationships from the principal component (PC) scatter plot. Here, to better understand the working mechanism of the Eigenstrat method, we consider its theoretical or "population" formulation. The eigen-equation for samples from an arbitrary number of populations is reduced to that of a matrix of dimension $K$, the elements of which are determined by the variance-covariance matrix for the random vector of the $K$ allele frequencies. Solving the reduced eigen-equation, we investigate the within-population fluctuations around the axes of variation on the PC scatter plot for simulated datasets. Specifically, we show that there exists an asymptotically stable pattern of the PC plot for large sample size. Our results provide theoretical guidance for interpreting the pattern of PC plot in terms of population relationships. For applications in genetic association tests, we demonstrate that, as a method of correcting for population stratification, regressing out the theoretical PCs corresponding to the axes of variation is equivalent to simply removing the population mean of allele counts and works as well as or better than the Eigenstrat method.

The PCA method was first applied to detecting and characterizing population structure more than 30 years ago [1]. By taking allele frequencies at different loci as a random vector and using the first few principal components (PCs), Cavalli-Sforza and co-workers constructed synthetic maps in their study of the evolutionary history of human populations [1,2]. Recently, PCA has been applied to large-scale association studies using data for single-nucleotide polymorphisms (SNPs) in attempting to detect a few top axes of large genetic variation [10,11].

In 2006, Patterson and co-workers [8] developed a new approach that uses PCA to detect population structures from large-scale genotype data of a sample of individuals. Instead of treating different markers as components and constructing PCs to represent the main variations from all markers as was traditional ([10] e.g.), in this new approach, Patterson et al. [8] indexed the random vector by individuals, taking genotype data at different (genotype) loci as its realizations. In the resultant PC scatter plot using axes of the top PCs, individuals from different populations have different coordinates and thus have different locations. Price et al. [9] proposed a method of correcting for population stratification in association studies by regressing out the top PCs obtained by this new method from the genotype data. The method was implemented in the package EIGENSTRAT and is referred to as Eigenstrat method. The Eigenstrat method has been applied to...
quantifying fine structures and describing the relationships of many different populations, such as European American [12,13], European [14,15], and Japanese populations [16], and is now the gold standard for detecting and correcting for population stratification.

Although the Eigenstrat method is becoming more popular, appropriate inferences about the population relationships from the PC scatter plot remains a challenging task [17] e.g. In this paper, we address this issue by considering the theoretical or “population” formulation of the Eigenstrat PCA method. Here, the term “population” means that the PCA is formulated for a hypothetical marker with allele frequencies drawn from different distributions for different populations. In contrast, the term “sample” means that PCA is performed using markers observed on the sample. We establish an explicit connection between the pattern of the PC plot and the variance-covariance parameters of the random vector of allele frequencies. We propose that these parameters, independent of the relative sample sizes of the population, are more suitable than the patterns of PC plot for quantifying population divergence. Based on our theoretical formulation of PCA, we prove the existence of an asymptotic pattern of the PC plot when the population sizes become large, and derive the formula for numerically calculating this asymptotic pattern for given population parameters. We then illustrate how to apply our theory in quantifying population structures and relationships using HapMap [18] and simulated data. We also use the theoretical formulation to investigate the intra-population fluctuations on the PC scatter plot constructed using “sample” marker data.

Our theoretical formulation of PCA also applies to association studies. In the Eigenstrat method, the confounding effect of population stratification is controlled for by regressing out the first few top sample PCs obtained from genotype data. Here, we propose that population stratification can also be corrected for by regressing the theoretical or “population” PCs calculated from the estimates of the variance-covariance parameters using our formulation. This is not only an alternative to the Eigenstrat method in GWAS, but may also be applied in candidate gene association studies when fewer markers are available for study. It turns out, as we rigorously show, that this method is equivalent to subtracting the population mean of allele counts. The proposed method is tested and compared with other methods using simulations and found to have superior performance to the Eigenstrat method when the differentiation among populations was limited.

Results

Statistical model

Suppose we have genotype data for individuals sampled from $K$ populations with sample sizes $N_1, N_2, \ldots, N_K$, respectively. In Eigenstrat theory, the data is modeled in the following way: the components of the random vector are the genotypes of the sampled individuals, $\mathbf{C}^T = [\mathbf{C}_1^T, \mathbf{C}_2^T, \ldots, \mathbf{C}_K^T]$ with $\mathbf{C}_m^k$ the vector of the variant allele for individual $m$ from population $k$ for a random marker. Data for different markers are taken as different measurements (“samples”) of this random vector. In this statistical model, the randomness of a component, $C_m^k$, comes from two sources: the marker is randomly chosen, and the genotype is determined randomly conditional on the allele frequency of the chosen marker. The probability distribution of the allele frequency depends on which population the individual is from. The populations are characterized by the variance-covariance matrix of the random vector $\mathbf{P} = (p_1, p_2, \ldots, p_K)$ of allele frequencies

$$V_F = \begin{pmatrix} \Sigma_1^2 & \Sigma_1^2 & \cdots & \Sigma_1^K \\ \Sigma_1^2 & \Sigma_2^2 & \cdots & \Sigma_2^K \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_K^1 & \Sigma_K^2 & \cdots & \Sigma_K^K \end{pmatrix}.$$ 

In this model, the genetically independent individuals are not statistically independent. Individuals from the same population have stronger correlations than those from different populations. We denote the variance-covariance matrix of the random vector $\mathbf{C}$ as

$$V = \begin{pmatrix} V_1 & V_{12} & \cdots & V_{1K} \\ V_{12} & V_2 & \cdots & V_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ V_{1K} & V_{2K} & \cdots & V_K \end{pmatrix},$$

where

$$V_k = \text{COV}(\mathbf{C}_k^2) = \begin{pmatrix} \sigma_k^2 & \sigma_{k1} & \cdots & \sigma_{kK} \\ \sigma_{k1} & \sigma_1^2 & \cdots & \sigma_{1K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{kK} & \sigma_{1K} & \cdots & \sigma_K^2 \end{pmatrix} (k = 1, 2, \ldots, K) \quad (3)$$

and

$$V_{lk} = \text{COV}(\mathbf{C}_l^2, \mathbf{C}_k^2) = \begin{pmatrix} \sigma_{lk} & \cdots & \sigma_{lk} \\ \vdots & \ddots & \vdots \\ \sigma_{lk} & \cdots & \sigma_{lk} \end{pmatrix} \quad (l, k = 1, 2, \ldots, K; \ l \neq k), \quad (4)$$

where

$$\sigma_k^2 = \text{VAR}(C_i^k) \quad (i = 1, 2, \ldots, N_i) \quad (5)$$

$$\sigma_{lk} = \text{COV}(C_i^k, C_j^l) \quad (i, j = 1, 2, \ldots, N_i; \ i \neq j) \quad (6)$$

are the variance of $C_i^k$ for any individual $i$ in population $k$, the covariance of two different individuals $i$ and $j$ in population $k$, and the covariance of two individuals in populations $l$ and $k$, respectively. As shown in Text S1, $V$ is related to the variance-covariance matrix of $\mathbf{P}$ by

$$\sigma_k^2 = 2\Sigma_k^2 + 2p_k(1-p_k) \quad (8)$$

$$\sigma_{lk} = 4\Sigma_k^2 \quad (9)$$
\[ \sigma_{ik} = 4\Sigma_{ik}, \] (10)

where \( \bar{p}_k \) is the mean of \( p_k \), for \( l, k = 1, 2, \ldots, K \) and \( l \neq k \).

One-population case

For PCA, we need to calculate the eigenvalues and eigenvectors of the variance-covariance matrix of the random vector of interest [19]. We begin with the simplest case, where all individuals are from the same population. Useful insights can be gained from this trivial situation. In this case, the variance-covariance matrix for the random vector \( \tilde{C} \) is given by Equation (3) with \( k = 1 \), for which the eigensolutions can be easily obtained ([19] pp. 469–470). The eigenvalues can be divided into two groups. The first group includes only one large eigenvalue, \( \lambda_1 = \sigma_1^2 + (N_1 - 1)\sigma_11 \), with the associated eigenvector \( \tilde{e}_1^T = [1/\sqrt{N_1}, 1/\sqrt{N_1}, \ldots, 1/\sqrt{N_1}] \). The second group includes all the other eigenvalues, which are smaller and are all equal: \( \lambda_2 = \cdots = \lambda_{N_1} = \sigma_1^2 - \sigma_11 \). The coordinate of the random vector \( \tilde{C} \) along the first PC is \( \tilde{e}_1^T \tilde{C} = \frac{1}{\sqrt{N_1}} \sum_{i=1}^{N_1} C_i \), proportional to the average of the allele counts over all samples. If the individuals within this population are not correlated to each other, \( \sigma_{11} = 0 \) and \( \lambda_1 = \sigma_1^2 \). That is, \( \lambda_1 \) reduces to the small eigenvalue. In contrast, if the individuals are completely correlated (e.g. if all individuals are monozygotic twins) \( \sigma_{11} = \sigma_1^2 \), \( \lambda_1 = N_1 \sigma_1^2 \), and \( \lambda_2 = \cdots = \lambda_{N_1} = 0 \). In general, the stronger the correlation between individuals, the larger \( \lambda_1 \) is and the smaller the small eigenvalues are.

This means that the {
only one PC here represents the co-variation of all individuals, whereas the small eigenvalues represent the variation between individuals.
}

Since there is only one population, the only PC with the large eigenvalue here does not reflect the variation caused by population structure. In Eigenstrat theory, therefore, one needs to perform the following mean adjustment:

\[ C_i' = C_i - \frac{1}{N} \sum_{j=1}^{N} C_j. \] (11)

For the mean-adjusted random vector \( \tilde{C}' \), the variance-covariance matrix has the same structure as that of \( \tilde{C} \), but with different diagonal and off-diagonal elements given by

\[ \sigma_i'^2 = \text{VAR}(C_i') = \sigma_i^2 - \sigma_{i1} \quad (i = 1, 2, \ldots, N_1) \] (12)

\[ \sigma_{i1}' = \text{COV}(C_i', C_j') = \sigma_{i1} - \sigma_{i1} \quad (i, j = 1, 2, \ldots, N_1; i \neq j) \] (13)

\[ \sigma_{11}' = \frac{1}{N_1} \left[ \sigma_1^2 + (N_1 - 1)\sigma_{11} \right]. \] (14)

It is easy to show that the large eigenvalue is now reduced to \( \lambda'_1 = \sigma_1'^2 - (N_1 - 1)\sigma_{11}' = 0 \) and the small eigenvalues remain unchanged:

\[ \sigma_i'^2 - \sigma_{i1}' = \sigma_i^2 - \sigma_{i1}. \] (15)

This shows how the mean adjustment in Eigenstrat theory removes the overall variance represented by the first PC that reflects the joint variation of all components because of their correlation instead of stratification. The same mean-adjustment will be performed for the general case where individuals are from two or more populations, for the same reason.

Two-population case

Now we turn to the first nontrivial case, where there are \( N_1 \) individuals in population 1 and \( N_2 \) individuals in population 2 in the samples. In this case, the variance-covariance matrix of \( \tilde{C} \) is

\[ V = \text{COV}(\tilde{C}) = \begin{pmatrix} V_1 & V_{12} \\ V_{12}^T & V_2 \end{pmatrix} \] (16)

and can be shown to have solutions as follows. The small eigenvalues of \( V \) are just those of \( V_1 \) and \( V_2 \), the same as in the one-population case. There are two large eigenvalues, each of which corresponds to an eigenvector whose coordinates are constant for individuals from the same population. However, these two large eigenvalues do not reflect only the variations caused by the population structure; as shown in the last subsection, they also represent the co-variation of the individuals. This can be easily seen if we assume \( \sigma_{12} = 0 \). In this case, the two large eigenvalues are simply the large eigenvalues of \( V_1 \) and \( V_2 \). If we were only interested in detecting population structure, the PCA for \( V \) would be sufficient. However, we are also interested in correcting for the population stratification, and therefore we hope to obtain PCs mainly representing variations due to population structure. This is why we need to investigate the mean-adjusted vector \( \tilde{C}' \) as defined in the last subsection. Here, again, the variance-covariance matrix of \( \tilde{C}' \) has the same structure as that of \( \tilde{C} \) and has the following eigensolution. The small eigenvalues of \( V' \) are still the same as those of \( V \). (Note that, as in the case of one population, \( \sigma_k^2 - \sigma_{kk} = \sigma_k^2 - \sigma_{kk}, k = 1, 2 \) ) The two large eigenvalues become

\[ \lambda'_1 = 0 \] with eigenvector \( \tilde{e}_1^T = \frac{1}{\sqrt{N_1 + N_2}} [1, 1, \ldots, 1] \) and

\[ \lambda'_2 = \frac{1}{N_1 + N_2} \left[ N_1\sigma_1^2 + N_2\sigma_2^2 + N_1(N_1 - 1)\sigma_{11} \right. \]

\[ + N_1(N_2 - 1)\sigma_{22} - 2N_1N_2\sigma_{12} \]

(17)

with eigenvector

\[ \tilde{e}_2^T = \frac{1}{\sqrt{N_1 + N_2}} \left[ -\frac{1}{N_1}, \ldots, -\frac{1}{N_1}, \frac{1}{N_2}, \ldots, \frac{1}{N_2} \right]. \] (18)

Note that Equation (18) is equivalent to Equations (13a) and (16b) in [17]. The first large eigenvalue \( \lambda'_1 = 0 \) reflects the fact that the mean of the vector is zero, a result of the mean adjustment, whereas the second large eigenvalue represents the variation caused by the population structure. The only nonzero large eigenvalue \( \lambda'_2 \) is very large compared with the small eigenvalues, for large \( N_1 \) and \( N_2 \). So, if there are only two populations, we would have only one eigenvector showing a clustering structure on a PC scatter plot using data of “samples” of markers. Any other eigenvectors would not have anything to do with stratification among populations.

It should be noted that the eigenvector corresponding to the only large eigenvalue and reflecting the population structure (Equation (18)) depends only on the ratio of sample sizes of the two populations, \( N_1/N_2 \), not on the other parameters \( \sigma_1, \sigma_2, \sigma_{11}, \sigma_{22}, \) or \( \sigma_{12} \). This is true only in the two-population case and is not a generic property, as will be clear soon.

K-population case

In the general case of \( K \) populations, the eigensolutions of the variance-covariance matrix Equation (2) are as follows. The small eigenvalues are the same as those for the individual populations.
There are $K$ large eigenvalues, each corresponding to an eigenvector whose coordinates are constant for individuals from the same population. These large eigenvalues, again, reflect not only the variation caused by population stratification but also the overall co-variation of the individuals. After the mean adjustment, the variance-covariance matrix $V'$ has the same structure as $V$ in Equation (2), and its submatrices $V'_1$ and $V'_K$ have the same structure as the corresponding $V_1$ and $V_K$ in Equations (3) and (4), respectively. The elements of $V'_1$ and $V'_K$ are given by

$$\sigma'_{1k} = \sigma_{1k} - 2\sigma_{0k} + \sigma'_c,$$

(19)

$$\sigma'_{kk} = \sigma_{kk} - 2\sigma_{0k} + \sigma'_c,$$

(20)

$$\sigma'_{lk} = \sigma_{lk} - \sigma_{0l} + \sigma'_c,$$

(21)

where

$$\sigma_{0k} = \frac{1}{\sum_{m=1}^{M} N_m} \left[ \sigma_{0m}^{2} + (N_{mk} - 1)\sigma_{kk} + \sum_{m}(\sigma_{mk}) \right]$$

(22)

$$\sigma'_c = \frac{1}{(\sum_{m} N_m)^2} \left[ \sum_{m} N_m \sigma_{m}^{2} + \sum_{m} N_m(N_m - 1)\sigma_{mm} 
+ 2 \sum_{m>n} N_m N_n \sigma_{mn} \right].$$

(23)

It is not difficult to show that the eigensolutions with small eigenvalues are still the same as those in the one-population and two-population cases (see also [8]). To find the eigensolutions with large eigenvalues of $V'$, which describe the population distinctions, we define

$$\tilde{X}^{T} = [\tilde{X}_1^{T}, \tilde{X}_2^{T}, \ldots, \tilde{X}_K^{T}],$$

(24)

and for each of the $\tilde{X}_k$ with dimension $N_k$,

$$\tilde{X}_k^{T} = [x_k, x_k, \ldots, x_k].$$

(25)

Then the eigenequation

$$V'\tilde{X} = \lambda'\tilde{X}$$

(26)

is reduced to

$$\begin{pmatrix}
\sigma'_{11} + (N_1 - 1)\sigma'_{11} & N_2\sigma'_{12} & \cdots & N_K\sigma'_{1K} \\
N_1\sigma'_{12} & \sigma'_{22} + (N_2 - 1)\sigma'_{22} & \cdots & N_K\sigma'_{2K} \\
\cdots & \cdots & \cdots & \cdots \\
N_1\sigma'_{1K} & N_2\sigma'_{2K} & \cdots & \sigma'_{kk} + (N_K - 1)\sigma'_{kk}
\end{pmatrix} \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_K
\end{pmatrix} = \lambda \begin{pmatrix}
x_1 \\
x_2 \\
\vdots \\
x_K
\end{pmatrix}.$$
covariance between each pair of populations are the same, as shown in Table 1, the three representative points in the PC scatter plot are distributed unevenly because of the unbalanced sample sizes, as seen in Figure 1. The representative points of populations with small sample sizes are around the borders, whereas the points for populations with large sample sizes are located near the zero point, as can be explained by Equation (30).

In the second example, we have five populations, the first three of which are close to one another (in the sense that the corresponding covariances are large) and are far away from the other two distant populations (see Table 2). Figure 2 shows that the first three populations can hardly be distinguished from the two-dimensional PC plot using the first two eigenvectors. In eigenvector 1, P4 and P5 are contrasted with the three closely related populations (P1, P2 and P3), while in eigenvector 2 P4 is contrasted with P5, and the other three are in the middle. The three closely related populations are distinguished in eigenvectors 3 and 4. This example suggests that one has to examine a large enough number of eigenvectors in order to find all the significant population differences. The first two eigenvectors are the most important, but the others are also needed if the samples are from more than three populations. However, if there are only two populations, a two-dimensional PC plot is not needed; only the first eigenvector shows the population structure.

The representative points depend on the sample sizes as well as the variance-covariance parameters, as pointed out in [8]. We note that, even if equal sample sizes are used, the representative points cannot replace the variance-covariance parameters in characterizing the population relationships, because their values depend on the presence of one another in the analysis.

Estimation of variance-covariance parameters and axes of variation

In practice, the variance-covariance parameters in Equations (32) and (33) are unknown and can only be estimated from genotype data of a large number of markers. The estimates of these parameters can be used to calculate the estimates of axes of variation using Equation (27). If the population memberships of the samples are known, estimation of the variance-covariance parameters is straightforward. For any one of the parameters in Equations (32) and (33), we can simply take the average of the corresponding elements of the sample variance-covariance matrix as its estimate. However, in reality, the information on population membership is usually unavailable and needs to be inferred using the PCA or other methods such as STRUCTURE. For inference of population structure from a PC plot, a generic clustering algorithm may be appropriate [20].

Table 1. Parameters for the three populations in Figure 1.

|       | P1 | P2 | P3 |
|-------|----|----|----|
| Sample size* | 5  | 10 | 100 |
| Variance  | 1  | 1  | 1  |
| Covariance|    |    |    |
| P1       | 0.8| 0.1| 0.1 |
| P2       | 0.8| 0.1|    |
| P3       |    | 0.8|    |

*Sample size: The number of individuals of each population.

doi:10.1371/journal.pone.0012510.t001

Table 2. Parameters for the five populations in Figure 2.

|       | P1 | P2 | P3 | P4 | P5 |
|-------|----|----|----|----|----|
| Sample size* | 100| 110| 120| 130| 140|
| Variance  | 1  | 1  | 1  | 1  | 1  |
| Covariance|    |    |    |    |    |
| P1       | 0.8| 0.7| 0.7| 0.1| 0.11|
| P2       | 0.73| 0.7| 0.13| 0.14|
| P3       | 0.75| 0.14| 0.09|
| P4       | 0.9| 0.17|    |
| P5       |    | 0.91|    |

*Sample size: The number of individuals of each population.

doi:10.1371/journal.pone.0012510.t002

In contrast, PCA in practice is based on the “sample” of markers. Namely, eigenvectors are calculated using genotypes of a large number of markers. These eigenvectors are therefore referred to as “sample” eigenvectors. The points on the PC scatter plot using the sample eigenvectors, called the sample points, are scattered to some extent because of sampling fluctuation. Our representative points using the estimated axes of variation should be located in the middle of the corresponding clusters of the sample points. We performed simulations to show how the axes of variation can be evaluated and to compare them with the sample eigenvectors used in Eigenstrat theory. Figure 3 shows an example of our simulation, using the simulating parameter values given in Table 3. P1 and P2 were simulated...
using an $F_{st}$ of 0.01 and thus were closer to each other than to P3 or P4, for which $F_{st}$ was much larger (0.43). The distance between P3 and P4 was even larger than their distances to P1 or P2. The representative points obtained from the estimated parameters listed in Table 3 were right in the centers of the corresponding sample points. Also listed in Table 3 are the estimated correlation coefficients $r_{kl}$ ($k, l = 1, 2, \ldots$). Compared with the covariances $s_{kl}$, the corresponding correlations $r_{kl} \approx s_{kl}/\sigma_k \sigma_l$ seem to be more suitable for representing the population distances.

Fluctuations in sample eigenvectors within populations and asymptotic PC plot patterns

In practice, within-population fluctuations of the PC scatter plot using sample eigenvectors may be so strong that closely related populations have overlapping clusters and hence cannot be distinguished. Here, we first investigate the factors that affect the within-population fluctuations for a given population divergence: the sample sizes of the populations and the number of markers.

Table 3. Simulating values for $F_{st}$ used for the four populations P1, P2, P3, and P4 and the estimated parameters.

|          | P1   | P2   | P3   | P4   |
|----------|------|------|------|------|
| Sample size* | 20   | 20   | 20   | 20   |
| $F_{st}$   | 0.01 | 0.01 | 0.43 | 0.43 |
| Variance   | 0.565 | 0.562 | 0.750 | 0.747 |
| Covariance |      |      |      |      |
| P1         | 0.128 | 0.119 | 0.114 | 0.116 |
| P2         | 0.127 | 0.112 | 0.115 |      |
| P3         | 0.500 | 0.113 |      |      |
| P4         |      | 0.493 |      |      |
| Correlation ($r$) |      |      |      |      |
| P1         | 0.227 | 0.211 | 0.174 | 0.178 |
| P2         | 0.225 | 0.172 | 0.178 |      |
| P3         | 0.667 | 0.151 |      |      |
| P4         |      | 0.660 |      |      |

*Sample size: The number of individuals of each population.

doi:10.1371/journal.pone.0012510.g003
We then study the asymptotic behavior of the patterns of the PC scatter plot as the sample size becomes large.

A remark is in order. The fluctuations of the sample points on the PC scatter plot within a population should not be confused with the random variation of marker data between individuals within a population. Recall from our theoretical consideration in Section 2 that the between-individual variation within a pure population exists even theoretically and is represented by the small eigenvalues and the corresponding eigenvectors. However, the within-population fluctuation of the sample points is mainly due to the limited “sample size” of markers and should be decreased as more markers are included in the analysis. The observed fluctuations of sample points around the representative points may also reflect effects from subtle population structure and cannot be reduced by increasing marker numbers.

We performed simulations to demonstrate the effect of increasing population size (N) and number of markers (M) on the within-population fluctuations of sample points on the PC scatter plot. In our simulations, we first generated three populations (P1, P2, and P4) each with the same $F_{st}$ value of 0.003; the three populations should therefore be equidistant from one another. In order to mimic a subtle subpopulation structure, we created another population (P3) based on P2 by using the allele frequency vector of P2 added to a random vector, each of the elements of which was independently and uniformly distributed within $\pm 0.01$. This resulted in three distinct populations, P1, P2+P3 and P4; P2+P3 has a subtle structure. Figure 4 shows the PC scatter plot using the first two eigenvectors for various values of M and N. If the subtle structure within P2+P3 is ignored, this plot should be all that is needed to distinguish these three populations. As we see from this figure, as the number of markers (M) increased, the fluctuations within each population gradually decreased, but the distance between P2 and P3 remained unchanged, indicating that fluctuations due to limited “sample size” can be reduced by increasing it, whereas fluctuations reflecting subtle structure cannot. Since there were actually four populations, we also plotted the first and third eigenvectors in Figure 5. Here, we see that after mainly addressing the difference between P1+P4 and P2+P3 in eigenvector 1, and the difference between P1 and P4 in eigenvector 2, the difference between P2 and P3 was further addressed in eigenvector 3.

![Figure 4. Effects of sample size and marker number on within-population fluctuations of PC scatter plot: eigenvector 2 vs. eigenvector 1.](image)

We plot the first two axes of variation (top row) and sample eigenvectors (other rows) for simulated data sets with different values for the number of markers (M) and the number of individuals (N) in the samples. In the simulation, we first generated three populations, P1, P2 and P4, each with the same $F_{st}$ value of 0.003. These three populations should therefore be equidistant from one another. Then another population, P3, was so generated that it was very close to P2 (see main text for details). P2+P3 can be viewed as a single population with a subtle structure. As M or N increased, the fluctuations within each population gradually decreased, while the distance between P2 and P3 remained unchanged.

doi:10.1371/journal.pone.0012510.g004
Figures 4 and 5 also show the effect of increasing the number of individuals in each population. As \( N \) increased, the within-population fluctuations became smaller and the distinctions among populations became clearer. The effect of increasing \( N \) was much stronger than that of increasing \( M \), in agreement with what was found in [8]. Here, we give an explanation for this phenomenon as follows. For a finite number of markers, each individual in PCA actually acts as a population. When the number of individuals (\( N \)) is small, the variation between individuals from the same population is comparable to that caused by population difference, so PCA tends to address these variations in the first few eigenvectors. As \( N \) increases, the variation due to population differences becomes overwhelming, so PCA addresses only this variation in the first few eigenvectors and leaves the trivial ones to other eigenvectors with small eigenvalues.

In Figures 4 and 5, we also plotted the theoretical patterns calculated using the population sizes (\( N_1 = \cdots = N_4 = N \)) and the estimated variance-covariance parameters from the case with the largest \( N \) (500) and largest \( M \) (100,000). The absolute distances between the representative points became smaller as \( N \) increased, because of the normalization equation (31), but the relative pattern remained almost the same, especially for large \( N \). The ratio of distance between \( P_2 \) and \( P_3 \) to that between \( P_2+P_3 \) and \( P_3 \) on the PC plot approached a constant value, implying that the pattern on the PC plot reached an asymptotic shape as \( N \) became large. This kind of asymptotic behavior can be derived from our theoretical considerations as follows.

Let \( n_k = N_k/N \) (\( k = 1, 2, \cdots, K \)) denote the relative sample size (or proportion) of population \( k \) in the samples of interest. It is shown in Text S2 that the asymptotic form of the eigen-equation (27) is

\[
N \begin{pmatrix}
- \sum_{l \neq 1} n_l \sigma_{1l} & n_2 \sigma_{12} & \cdots & n_K \sigma_{1K} \\
 0 & - \sum_{l \neq 2} n_l \sigma_{2l} & \cdots & n_K \sigma_{2K} \\
 \vdots & \vdots & \ddots & \vdots \\
 n_1 \sigma_{1K} & n_2 \sigma_{2K} & \cdots & - \sum_{l \neq K} n_l \sigma_{lK}
\end{pmatrix}
\]

(34)
with
\[
\hat{\sigma}_{kk'} = \sigma_{kk'} + \sum_{m,l=1}^{K} n_m n_l \sigma_{ml} - \sum_{m=1}^{K} n_m (\sigma_{km} + \sigma_{k'm})
\]  
(35)

for \( k, k' = 1, 2, \cdots, K \) and \( k \neq k' \). The small part being neglected for large \( N \) is
\[
\sigma_{kk'} = \frac{1}{N} \left[ \sum_{m=1}^{K} n_m \sigma_m^2 - \sigma_k^2 - \sigma_{k'}^2 \right] \sim \frac{1}{N},
\]  
(36)

where \( \sigma_m^2 = \sigma_m^2 - \sigma_{nn} \) is the small eigenvalue for population \( m \).

From Equation (34), we see that asymptotically the eigenvectors are independent of \( N \) for given proportions of populations and variance-covariance parameters listed in Equations (32) and (33), whereas the large eigenvalues increase linearly with \( N \). Figure 6 shows how the theoretical predictions of the dimensions of the pattern on the PC plot vary with sample size for the simulated datasets plotted in Figures 4 and 5. The dimensions shown are: \( d_1 \), the distance between P2 and P3 on eigenvector 1; \( d_2 \), the distance between P1 and P3 on eigenvector 1; \( d_3 \), the distance between P1 and P4 on eigenvector 2; and \( d_4 \), the distance between P2 and P3 on eigenvector 3. Here, the estimated variance-covariance parameters from the case with the largest number of markers (\( M = 100,000 \)) were used.

Note that the asymptotic form of the eigen-equation, and hence the asymptotic form of the PC plot patterns, do not depend on the values of the variances \( \sigma_k^2 (k = 1, 2, \cdots, K) \). They are determined only by the intra- and inter-population covariances, and the relative sample sizes.

For \( \hat{\sigma}_{kk'} \) to be neglectable compared to \( \hat{\sigma}_{kk} \), we need to have an \( N \) such that
\[
N \gg \frac{\sum_{m=1}^{K} n_m \sigma_m^2 - \sigma_k^2 - \sigma_{k'}^2}{\sigma_{kk'} + \sum_{m=1}^{K} n_m (\sigma_{mm} - \sigma_{km})}
\]  
(37)

In the simplest case, where all populations have the same variance (\( \sigma_1^2 \)) and the same intra-population covariance (\( \sigma_{11} \)) and each pair of populations has the same inter-population covariance (\( \sigma_{12} \)), we have a simple expression for the critical size \( N \) as
\[
N \gg \frac{\sigma_1^2 - \sigma_{11}}{\sigma_{11} - \sigma_{12}},
\]  
(38)

Figure 6. Approach to the asymptotic form. We plot the theoretical predictions for the dimensions of the PC plot pattern a function of the sample size for the simulated data sets used in Figures 4 and 5. Here, \( d_1 \) is the distance between P2 and P3 on eigenvector 1; \( d_2 \) is the distance between P1 and P3 on eigenvector 1; \( d_3 \) is the distance between P1 and P4 on eigenvector 2; and \( d_4 \) is the distance between P2 and P3 on eigenvector 3. When \( N \approx 30 \), the pattern of the PC plot reached its asymptotic form.

doi:10.1371/journal.pone.0012510.g006
indicating that the closer the populations \((\sigma_{11} - \sigma_{12} \rightarrow 0)\) the larger the sample size needed for the asymptotic pattern to be approached. Note, however, that this is true only for cases where there are more than two populations in the sample. If there are only two populations, as shown in the previous section, the pattern in eigenvector 1 is determined only by the relative sample sizes (see Equation (18)).

### Application to HapMap data

We estimated the variance-covariance parameters and the axes of variation for some of the HapMap [18] populations. An example is given in Figure 7, where the first three sample eigenvectors are plotted from the analysis of the four populations: Chinese in Denver (CHD), Gujarati Indians in Houston (GIH), Japanese in Tokyo (JPT), and Tuscan Italians (TSI) using markers on chromosome 1. The estimates of the variance-covariance parameters for these four populations are given in Table 4. The axes of variation calculated from these estimates are also shown in Figure 7 and were in consistent with the sample eigenvectors calculated directly from the raw variance-covariance matrix. It can be seen from Figure 7 that the two genetically very close populations, CHD and JPT, are contrasted only on the third eigenvector; the first two eigenvectors are used to address the difference between PHD+JPT vs GIH and TSI, and the difference between GIH vs TSI. We note that CHD and JPT can be distinguished not only by PCA together with other populations but also by PCA by themselves (data not shown). As shown in Table 4, the genetic diversity within GIH or TSI is so large that the average covariance between two random individuals both from GIH or from TSI is larger than the covariance between an individual from CHD and an individual from JPT. This explains why on eigenvector 3, where CHD and JPT are contrasted, the clusters of GIH and TSI are also elongated, showing a within-population structure.

### Application to genetic epidemiology

**Correcting for population stratification using axes of variation.** Now we turn to the issue of correcting for population stratification in genetic association studies. In the Eigenstrat method, the correction for stratification is performed by regressing out the variation caused by population structures [9]. In the theory of PCA, the random vector \(\tilde{C}^t\) can be expanded as a linear combination of all PCs:

\[
\tilde{C}^t = y_1 \tilde{X}_1 + y_2 \tilde{X}_2 + \cdots + y_N \tilde{X}_N,
\]

(39)

where \(N\) is the total number of individuals and

\[
y_i = \tilde{X}_i^T \tilde{C}^t
\]

(40)

is the \(i\)th PC. To correct for population stratification, we subtract the first \(K - 1\) terms from this expansion, which are the variations due to the differences between the \(K\) populations. The sum of the remaining terms, describing the variations between individuals,

\[
\tilde{C}^0 = \tilde{C}^t - \sum_{k=1}^{K-1} y_k \tilde{X}_k,
\]

(41)

is then used for disease association test.

The PCs removed for correcting population stratification in Equation (39) are those obtained using the “sample” markers in the Eigenstrat method and is referred to as sample PCs. Here, we propose to use the same strategy for correcting population stratification but with the PCs defined using the axes of variation (referred to as representative PCs or “population” PCs). Our motivation is as follows. When the Eigenstrat method is used, some of the genotype variations corresponding to within-population fluctuations in the sample eigenvectors are removed, in addition to those corresponding to population stratification. As shown in previous sections, these within-population fluctuations are mainly

![Figure 7. Four HapMap populations.](image)

The first three eigenvectors for the four populations CHD, GIH, JPT and TSI using the HapMap data for chromosome 1. As in Figure 3, the small symbols represent sample eigenvectors from the Eigenstrat analysis, whereas the large ones are the estimates of the axes of variation based on the estimated variance-covariance parameters of the four populations, listed in Table 4. The CHD and JPT populations are so close to each other that they can be distinguished only on eigenvector 3.

doi:10.1371/journal.pone.0012510.g007

Theoretical Formulation of PCA

### Table 4. Estimates and the corresponding standard errors for the variance-covariance matrix of the four populations CHD, GIH, JPT, and TSI using HapMap data for chromosome 1.

| Sample size | CHD | GIH | JPT | TSI |
|-------------|-----|-----|-----|-----|
| Variance    | 0.680(0.006) | 0.661(0.008) | 0.680(0.007) | 0.665(0.008) |
| Covariance  |     |     |     |     |
| CHD         | 0.360(0.008) | 0.279(0.008) | 0.354(0.008) | 0.260(0.007) |
| GIH         | 0.314(0.012) | 0.280(0.008) | 0.298(0.008) |         |
| JPT         | 0.357(0.008) | 0.260(0.007) |         |         |
| TSI         | 0.330(0.008) |         |         |         |

*Sample size: The number of individuals of each population.

doi:10.1371/journal.pone.0012510.t004
due to the finite “sample size” (i.e. limited number of markers) and thus are irrelevant to the issue of population stratification. Indeed, in practice, a proportion of the within-population fluctuations may be due to a subpopulation structure. However, as shown in previous sections, the fluctuations in the first few eigenvectors only partially represent this kind of subpopulation structure. Consider the example given in Figures 4 and 5. When the samples are thought of as coming from three distinct populations P1, P2, and P4, only the first two eigenvectors should be used to correct for population stratification. The variation caused by the difference between P2 and P3 would then be partially removed; although the fluctuation described by d1 on eigenvector 1 would be taken into account, that described by d4 on eigenvector 3 would remain in the residual. So the variation caused by the difference between P2 and P3 would only be altered and not completely removed. Only partially removing the variations caused by a subpopulation structure may not be really helpful for reducing false-positive rates in a case-control study. In this specific example, where the subpopulation structure is the simplest, we could remove the corresponding variation by simply adding the third PC in the sum in Equation (41). In reality, however, subpopulation structures are far more complex and are hence represented by many PCs. Regressing out too many PCs in Equation (41) would remove too much inter-individual variation within a pure population and hence would significantly reduce the power of association tests. We therefore prefer to use the representative PCs in Equation (41) for removing the main variations caused by major population stratification, while keeping the variations due to subtle subpopulation structures unchanged.

Since the “population” PCs are used here, the proposed method is referred to as popu-Eigenstrat.

Now let us derive the theoretical expression of the residuals in Equation (41). Since we use the representative PCs, for a given \( \hat{C} \), the vector \( \hat{R} = \sum_{k=1}^{K} \gamma_k \hat{X}_k \), which is subtracted from \( \hat{C} \), has a structure like

\[
\hat{R}^T = \begin{bmatrix} \hat{R}_1^T, \hat{R}_2^T, \ldots, \hat{R}_K^T \end{bmatrix},
\]

and for each of the \( \hat{R}_k \),

\[
\hat{R}_k^T = [r_k, r_k, \ldots, r_k].
\]

In Text S3, it is shown that

\[
r_k = \sum_{l=1}^{K} C_{l}^k N_l \sum_{l=1}^{K-1} x_{l}^k x_{l}^k = C_{k}^k,
\]

where

\[
C_{k}^k = \frac{1}{N_k} \sum_{l=1}^{N_k} C_{l}^k
\]

is the mean of \( C^* \) over individuals from population \( k \). So it turns out that our representative PC-based correction is simply equivalent to subtracting the population group means for each individual. Our method does not even use any information about the variance-covariance parameters.

Results of simulations

We conducted simulations for comparing the performance of the popu-Eigenstrat method with that of the original Eigenstrat method as well as with that of the covariate-adjustment method, which used population labels as a covariate. See Method section for details of the simulations. Table 5 shows the results of our simulations. For \( F_{st} = 0.01 \), which is typical of differentiation between divergent European populations, the proposed method, popu-Eigenstrat, using the representative PCs, achieved almost the same rates of false-positive associations and comparable power as the original EIGENSTRAT method, which uses the sample PCs. Compared with the covariate-adjustment method, our method has slightly lower power, but also a lower rate of false-positive associations.

It is interesting to compare the results for the two different allele frequency sets for the Causal-Specific SNPs (see Methods for definitions of these simulated SNPs). When the allele frequency ratio was the inverse of the sample size ratio, the power was zero if no correction was performed. The reason was that in this case, even though a very high proportion of tests had very low p-values, the disease allele was incorrectly identified as the wild allele. The power increased to \( 41\% - 47\% \) after the methods of stratification correction were applied. In contrast, when the allele frequency ratio was the same as the sample size ratio, the power without correction was as high as \( >99\% \). Without population stratification, we could not have achieved such an extremely high power for the given sample size, allele frequencies and relative risk. In this case, the population with a high disease allele frequency was over-sampled and thus the difference of allele frequency between the case and control groups was enlarged. After the methods of stratification correction were applied, the power was reduced to its “normal” level (\( 37\% - 43\% \)). This observation indicates that special attention has to be paid to the population’s allele frequency spectra when powers for stratification correction strategies are compared. For the causal-StruInfo SNPs in our simulation (see Methods for definitions of these simulated SNPs), some of the SNPs for which the allele frequency ratio was the inverse of the sample size ratio may have contributed to an increase in power after stratification correction was applied. However, more SNPs had allele frequency ratio with the same trend as the sample size ratio and hence contributed more to a reduction in power when stratification correction was applied.

For a smaller \( F_{st} = 0.003 \), we found that the performance of the original Eigenstrat method was poorer. The rate of false positives was reduced less by Eigenstrat method than by the other methods. The power was increased less by Eigenstrat method than it was by the other methods in the case when the ratio of disease-causing allele frequencies was the inverse of the ratio of sample sizes. Only when the two ratios were the same did the Eigenstrat method improve the power to significantly higher than the others. The other methods of stratification correction worked similarly as in the case of larger \( F_{st} \). This can be explained as follows. As \( F_{st} \) decreased, the distances between the populations decreased, and hence the fluctuations in the sample eigenvectors increased. This in turn made these sample PCs less representative of the population structure, resulting in a poorer performance in correcting for stratification.

We also examined how power was affected by regressing out too many PCs using Eigenstrat method. In the case of \( F_{st} = 0.003 \), for the Causal-StruInfo SNPs, the power was reduced from \( 53\% \) (see Table 5) obtained using only two PCs to \( 48\% \), \( 38\% \) and \( 28\% \) when using 30, 90 and 150 PCs, respectively.

Finally, the results on our comparison between Eigenstrat and popu-Eigenstrat should be taken with caution, because additional information (i.e. population memberships) was given to popu-Eigenstrat.
very subtle subpopulation structures are entangled with the normal
subpopulation structure. These fluctuations corresponding to the
then mainly determined by the number of markers and the
The within-population fluctuations in the asymptotic pattern are
investigated the factors that affect the within-population fluctua-
tions of the sample eigenvectors (as obtained from the Eigenstrat
methods) around the axes of variation. As the sample size becomes
become large. We have also shown that the asymptotic pattern can
exists an asymptotic pattern on the PC plot as the sample size
plot. Using the developed formulation, we have shown that there
correctly infer population structures from the pattern of the PC
practice, our formulation provides theoretical guidance on how to
Accordingly, we have also shown that the proposed method can also
association tests. The proposed method outperformed the method
of simply using the population label as a covariate in reducing
false-positive rates, and it had slightly lower power. Our proposed
method can also be used in candidate gene association studies or
replication studies as long as the population memberships are
known and a trend test is preferred.

Discussion

The Eigenstrat method is a powerful tool to detect and correct
for population stratification by treating genotype data as
“samples” of markers. In this work, we have provided a framework
in which the large eigenvalues and the corresponding eigenvectors
necessary for differentiating the population structure are theoret-
ically connected to the variance-covariance parameters of the
random vector of the allele frequencies of the populations. These
variance-covariance parameters can serve as an alternative to the
traditional $F_{st}$ statistic for quantifying population relationships. In
practice, our formulation provides theoretical guidance on how to
correctly infer population structures from the pattern of the PC
plot. Using the developed formulation, we have shown that there
exists an asymptotic pattern on the PC plot as the sample size
become large. We have also shown that the asymptotic pattern can
be easily obtained by numerically solving the asymptotic form of the
reduced eigen-equation for given covariance parameters and the
relative sample sizes.

Based on our theoretical consideration and simulations, we have
investigated the factors that affect the within-population fluctua-
tions of the sample eigenvectors (as obtained from the Eigenstrat
method) around the axes of variation. As the sample size becomes
large, the overall asymptotic pattern of the PC plot quickly forms.
The within-population fluctuations in the asymptotic pattern are
then mainly determined by the number of markers and the
subpopulation structure. These fluctuations corresponding to the
very subtle subpopulation structures are entangled with the normal
inter-individual variations within a pure population and hence can
hardly be adjusted without significantly affecting the power of
association tests.

These conclusions led us to a novel method of correcting for
population stratification: We can regroup out the representative
PGs, instead of the sample PCs as done in Eigenstrat theory. We
theoretically showed that this method is equivalent to simply
removing the population mean of the allele counts. Therefore,
implementation of the proposed method becomes trivial, whence
the samples’ population memberships are known (either self-
reported or identified using the Eigenstrat method or any other
methods, such as STRUCTURE). Our simulation studies showed
that the proposed method worked as well as the Eigenstrat method
for reducing false-positive rates and for maintaining the power of
association tests. The proposed method outperformed the method
of simply using the population label as a covariate in reducing
false-positive rates, and it had slightly lower power. Our proposed
method can also be used in candidate gene association studies or
replication studies as long as the population memberships are
known and a trend test is preferred.

In the present work, we have not considered admixture of
populations. As shown in [8], PCA carried out on samples that
include admixed individuals produces an interesting pattern on the
PC scatter plot: the admixed samples are lying along a line
between the two source populations. Similar patterns have been
observed for other populations [13,16]. Work is currently in
progress to extend our theoretical formulation to the situation of
admixture in order to explain the observed patterns.

Table 5. Proportion of associations reported as significant ($P < 0.001$) using different methods of stratification correction for simulated data.

| $F_{st}$ | Without correction | Eigenstrat | popu-Eigenstrat | Covariate-adjustment |
|---------|-------------------|------------|----------------|---------------------|
| 0.01    |                   |            |                |                     |
| StruInfo SNPs* | Causal* | 0.6002 | 0.5129 | 0.5077 | 0.5651 |
| Specific SNPs* | Null* | 0.0281 | 0.0057 | 0.0052 | 0.0090 |
| 0.75:0.30:0.15 | Causal* | 0.0000 | 0.4106 | 0.4218 | 0.4690 |
| Specific SNPs* | Null* | 0.9938 | 0.0060 | 0.0059 | 0.0090 |
| 0.15:0.30:0.75 | Causal* | 1.0000 | 0.4105 | 0.3712 | 0.4253 |
| Specific SNPs* | Null* | 0.9730 | 0.0074 | 0.0073 | 0.0113 |
| 0.003   |                   |            |                |                     |
| StruInfo SNPs* | Causal* | 0.6939 | 0.5263 | 0.5168 | 0.5738 |
| Specific SNPs* | Null* | 0.0144 | 0.0062 | 0.0068 | 0.0102 |
| 0.75:0.30:0.15 | Causal* | 0.0000 | 0.2175 | 0.4180 | 0.4674 |
| Specific SNPs* | Null* | 0.9947 | 0.0143 | 0.0058 | 0.0092 |
| 0.15:0.30:0.75 | Causal* | 1.0000 | 0.5537 | 0.3572 | 0.4112 |
| Specific SNPs* | Null* | 0.9716 | 0.0130 | 0.0064 | 0.0094 |

*StruInfo SNPs: Allele frequencies were simulated based on the Balding-Nichols model and the SNPs were population structure informative.

Causal: Genotypes were generated according to the proportion: $(1-p)^2 \cdot 2Rp(1-p) \cdot R^2 p^2$ where $p$ is the causal allele frequency and $R = 1.5$ is the relative risk.

Null: Genotypes were generated according to the Hardy-Weinberg proportion: $(1-p)^2 \cdot 2p(1-p) \cdot p^2$. 10,000 SNPs were simulated for each of the categories.

doi:10.1371/journal.pone.0012510.t005
Methods

Simulations of population structure
Following [4] and [9], we simulated genotype data for a specified number of populations with specified values of $F_{st}$ using the Balding-Nichols model [21]. The ancestral allele frequency $p_k$ was first generated from the uniform distribution on $(0,0.5,0.9)$ for each locus $l$. The allele frequencies in population $k$ were then drawn from a beta distribution with parameters $(1-F_{st})p_k/F_k$ and $(1-F_{st})(1-p_k)/F_k$, where $F_k$ is the $F_{st}$ for population $k$. No linkage disequilibrium was considered here. Distances between a pair of population was determined by the populations’ $F_{st}$s with the ancestral population. Only when $F_k$ is chosen to be the same for all populations to be simulated does it become an estimate of $F_{st}$ for all the populations.

Simulations of association tests
Our simulations for association tests were similar to those reported in [9]. One of the differences between our simulations and those in [9] is that we considered three populations rather than two populations. In our simulations, we assumed that the prior probability of sampling individuals from each of these three populations was the same, 1/3. We assumed that the ratio of disease prevalences in the three populations was 1:2:5, and that these prevalences were very small. The numbers of cases and controls simulated for each population were their expected values, namely, 30, 60, 150 for cases and 80, 80, 80 for controls.

For each individual in each population, we simulated four different categories of SNPs. The first and second categories were generated using allele frequencies based on the Balding-Nichols model [21] with $F_{st}$ = 0.01 or 0.003 for all populations, and the SNPs were thus population-structure-informative (StruInfo SNPs). The genotype data were generated differently for the first and second categories. For the first category, data for both cases and controls were simulated using allele frequencies based on the Balding-Nichols model [21] with $F_{st}$ = 0.01 or 0.003 for all populations, and the SNPs were thus population-structure-informative (StruInfo SNPs). The genotype data were generated differently for the first and second categories. For the first category, data for both cases and controls were generated in the same way, by using Hardy-Weinberg equilibrium, and hence were not associated with the disease. SNPs in the first category were thus referred to as Null-StruInfo SNPs. We simulated genotypes of 10,000 Null-StruInfo SNPs. They were first used to infer the variance-covariance matrix of the three populations and the sample PCs and then served as replicates for estimating the type I error rate. For the second category of SNPs, referred to as Causal-StruInfo SNPs, the genotypes were simulated differently for cases and controls. For controls, the simulation of genotypes was the same as for the Null-StruInfo SNPs, whereas for cases, we used a risk model with a relative risk $R = 1.5$ for the causal allele. A case individual was assigned genotype 0, 1, or 2 with probabilities $(1-p)^2$, $2Rp(1-p)$, or $R^2p^2$, respectively, where $p$ is the causal allele frequency.

We also simulated 10,000 SNPs for each of the third and fourth categories. The third category of SNPs, referred to as Null-Specific SNPs, were disease-independent, like the Null-StruInfo SNPs, but had a fixed allele frequency set for the three populations. For the fourth category, Causal-Specific SNPs, the allele frequencies for the three populations were also fixed, but the cases and controls were simulated differently, as for the Causal-StruInfo SNPs, using the same relative risk, $R = 1.5$. We used two different allele frequency sets for the specific SNPs: 0.15, 0.30, and 0.75; and 0.75, 0.30 and 0.15. In the first set, the ratio of the allele frequencies of the populations was the same as that of the sample sizes. In the second set, the ratio of the allele frequencies was the inverse of that of the sample sizes. The Null-Specific SNPs were intended for estimating type I error rate and the Causal-Specific SNPs for estimating power for a specific allele frequency set in the populations.

Following [8], we used the Armitage trend $p^2$ statistic for association tests without stratification correction (without-correction), and we used the generalized Armitage trend $p^2$ statistic when stratification was corrected for by regressing out the sample PCs (Eigenstrat) or the representative PCs (popu-Eigenstrat). For a fourth test, we used the population label as a covariate (covariate-adjustment). Association statistics producing a P value $< 0.001$ were reported as significant.

Programs and scripts used in this work are available at https://cge.mandelson.org/dma/User/ProgramsScripts/popuPCA/: (a) VarCov, a C++ program for calculating the sample variance-covariance matrix from genotype data; (b) EstimateSigma, a C++ program for estimating the variance-covariance parameters from the sample variance-covariance matrix; (c) ReducedMat.awk, an awk script for calculating the reduced matrix from the variance-covariance parameters and the sample sizes; and (d) An R script for calculating the eigenvalues and eigenvectors of the reduced matrix.

Supporting Information

Text S1 Derivation of Equations (8-10) Found at: doi:10.1371/journal.pone.0012510.s001 (0.03 MB PDF)

Text S2 Derivation of Equation (38) and Equation (34) Found at: doi:10.1371/journal.pone.0012510.s002 (0.03 MB PDF)

Text S3 Derivation of Equation (44) Found at: doi:10.1371/journal.pone.0012510.s003 (0.02 MB PDF)

Author Contributions
Conceived and designed the experiments: JM CIA. Performed the experiments: JM CIA. Analyzed the data: JM. Contributed reagents/materials/analysis tools: JM CIA. Wrote the paper: JM CIA.

References
1. Menozzi P, Piazza A, Cavalli-Sforza L (1978) Synthetic maps of human gene frequencies in europeans. Science 201: 786–792.
2. Cavalli-Sforza L, Menozzi P, Piazza A (1994) The History and Geography of Human Genes Princeton University Press.
3. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997–1004.
4. Pritchard J, Donnelly P (2003) Case-control studies of association in structured or admixed populations. Theor Popul Biol 63: 227–237.
5. Pritchard J., Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959.
6. Falush D, Stephens M, Pritchard J (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164: 1567–1587.
7. Tang H, Coram M, Wang P, Zhu X, Risch N (2006) Reconstructing genetic ancestry blocks in admixed individuals. Am J Hum Genet 79: 1–12.
8. Patterson N, Price A, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 4: 2074–2093.
9. Price A, Patterson N, Plenge R, Weinblatt M, Shadick N, et al. (2006) Principal components analysiscorrects for stratification in genome-wide association studies. Nat Genet 38: 904–909.
10. Zhu X, Zhang S, Zhao H, Cooper RS (2002) Association mapping, using a mixture model for complex traits. Genet Epidemiol 23: 104–196.
11. Yu K, Wang Z, Li Q, Washolder S, Hunter D, et al. (2008) Population substructure and control selection in genome-wide association studies. PLoS ONE 3: e2551.
12. Price A, Butler J, Patterson N, Capelli C, Pascali VL, et al. (2008) Discerning the ancessty of european americans in genetic association studies. PLoS Genet 4: 1300–0017.
13. Tian C, Plenge RM, Ransom M, Lee A, Villoslada P, et al. (2008) Analysis and application of european genetic substructure using 300k snp information. PLoS Genet 4: e4.
14. Heath SC, Gut IG, Brennan P, McKay JD, Bencko V, et al. (2008) Investigation of the fine structure of European populations with applications to disease association studies. Eur J Hum Genet 16: 1413–1429.
15. Nelis M, T E, Magi R, Zamprich F, Zamprich A, et al. (2009) Genetic structure of Europeans: A view from the north-east. PLoS One 4: e5472.
16. Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, et al. (2008) Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. Am J Hum Genet 83: 445–456.
17. McVean G (2009) A genealogical interpretation of principal components analysis. PLoS Genet 5: e1000696.
18. Gibbs R, Belmont J, Hardenbol P, Willis T, Yu F, et al. (2003) The international hapmap project. Nature 426: 789–796.
19. Johnson R, Wichern D (1982) Applied Multivariate Statistical Analysis. Prentice-Hall, Englewood Cliffs, NJ.
20. Lee C, Abdool A, Huang C (2009) PCA-based population structure inference with generic clustering algorithms. BMC Bioinformatics 10(suppl 1): S73.
21. Balding D, Nichols R (1995) A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. Genetica 96: 3–12.