Ma and Shi *BMC Genetics* (2020) 21:34
https://doi.org/10.1186/s12863-020-0833-x

BMC Genetics

RESEARCH ARTICLE Open Access

On rare variants in principal component analysis of population stratification

Shengqing Ma and Gang Shi[*]

Abstract

Background: Population stratification is a known confounder of genome-wide association studies, as it can lead to false positive results. Principal component analysis (PCA) method is widely applied in the analysis of population structure with common variants. However, it is still unclear about the analysis performance when rare variants are used.

Results: We derive a mathematical expectation of the genetic relationship matrix. Variance and covariance elements of the expected matrix depend explicitly on allele frequencies of the genetic markers used in the PCA analysis. We show that inter-population variance is solely contained in *K* principal components (PCs) and mostly in the largest *K*-1 PCs, where *K* is the number of populations in the samples. We propose F_{PC}, ratio of the inter-population variance to the intra-population variance in the *K* population informative PCs, and d^2, sum of squared distances among populations, as measures of population divergence. We show analytically that when allele frequencies become small, the ratio F_{PC} abates, the population distance d^2 decreases, and portion of variance explained by the *K* PCs diminishes. The results are validated in the analysis of the 1000 Genomes Project data. The ratio F_{PC} is 93.85, population distance d^2 is 444.38, and variance explained by the largest five PCs is 17.09% when using with common variants with allele frequencies between 0.4 and 0.5. However, the ratio, distance and percentage decrease to 1.83, 17.83 and 0.74%, respectively, with rare variants of frequencies between 0.0001 and 0.01.

Conclusions: The PCA of population stratification performs worse with rare variants than with common ones. It is necessary to restrict the selection to only the common variants when analyzing population stratification with sequencing data.

Keywords: Rare variant, Population stratification, Principal component analysis, Single nucleotide polymorphism

Background

Genome-wide association studies (GWAS) [1] have identified a considerable number of sequence variants, such as single nucleotide polymorphisms (SNPs), associated with human diseases or traits. Population stratification—allele frequencies of genetic markers of the studied samples having significant differences owing to systematic ancestry differences—can cause false positive results in case-control as well as cohort studies [2, 3]. Association mapping based on rare variants are much more susceptible to subtle effects of population stratification and therefore, more likely to yield false positive results [4]. From a population genetics point of view, exploring population structure is important for understanding the evolutionary history of populations. Many methods and software are available to study the population stratification, such as the principal component analysis (PCA) implemented in EIGENSOFT [5, 6], the multidimensional scaling analysis in PLINK [7], the clustering analysis in STRUCTURE [8, 9], and fastSTRUCTURE [10]. Recently, controlling population stratification in the

[*] Correspondence: gshi@xidian.edu.cn
State Key Laboratory of Integrated Services Networks, Xidian University, 2 South Taibai Road, Xi'an 710071, Shaanxi, China

association analysis using linear mixed models [11–14] was also suggested. Based on a large number of common variants whose minor allele frequencies (MAFs) are larger than 5%, the PCA of population structure is widely applied in GWAS.

With the advance of high-throughput sequencing technology, as well as the enormous reduction of the cost, it is capable and affordable in genetic studies to detect additional low-frequency and rare variants (MAF < 1%) [15]. Furthermore, existing sequencing data suggest that the vast majority of rare variants are population-specific. In the 1000 Genomes Project data [16, 17], there are a total of 77 million biallelic SNPs, among which 65 million are rare and 52 million are polymorphic in one of the five continental ancestry populations: East Asian (EAS), South Asian (SAS), African (AFR), European (EUR), American (AMR). It seems that rare variants are more informative in distinguishing population structure than common ones. However, the efficacy of using rare variants in population stratification analysis remains controversial [18–21].

A number of efforts have been made concerning the use of low-frequency and rare variants in population stratification analysis. Baye et al. illustrated that more fine substructure can be detected using rare variants and suggested that more SNPs are required to account for a similar level of population structure using rare variants compared to common ones [18]. Siu et al. showed that rare variants have a much higher power to identify population substructure than common variants [19]. In contrast, Zhang et al. demonstrated that PCAs based on common and less-frequency SNPs perform better than those based on rare ones in separating European and African samples [20]. The authors further concluded that there is little added value for PCA of population stratification with rare variants only [21]. All existing work was based on analysis of genotype data from the 1000 Genomes Project with known population structure.

In this work, we investigate how rare variants affect PCA of population stratification from a theoretical perspective. We derive mathematical expectation of the genetic relationship matrix (GRM) [22]. The GRM is commonly computed from the observed genotypes and eigen-decomposed in the analysis of population stratification. Elements of the expected genetic relationship matrix (EGRM), however, depend explicitly on the allele frequencies of the markers used. We show that inter-population variance is solely contained in K principal components (PCs) and mostly in the largest K-1 PCs, where K is the number of populations in the sample. We propose F_{PC}, ratio of the inter-population variance to the intra-population variance in the K population informative PCs, and d^2, sum of squared distances among populations, as measures of population divergence. We show analytically that when allele frequencies become small, the ratio F_{PC} abates, the population distance d^2 decreases, and portion of variance explained by the K PCs diminishes. Therefore, the PCA of population stratification performs worse with rare variants than with common ones. The results are further validated in the analysis of the 1000 Genomes Project data with 2504 individuals from five continental populations.

Methods

Genetic relationship matrix

In the scenario where genotype data of individuals is sampled from K populations, there are N_k individuals in population k and the number of individuals in the total population is $N = N_1 + N_2 + \cdots + N_K$. We have M SNPs, whose frequencies of their coded alleles in population k are $[f_{k1}, f_{k2}, \cdots, f_{kM}]$. Let $\boldsymbol{X}$ be the genotype matrix of dimension $N \times M$. The genotypic value $\boldsymbol{X}(n,m)$ is the number of the coded allele of SNP m for individual n, where $n = 1, 2, \cdots, N$ and $m = 1, 2, \cdots, M$. Typically, the number of individuals is less than the number of markers, i.e. $N < M$. We assume that all SNPs are under the Hardy-Weinberg equilibrium in each population. The GRM can be calculated as

$$\boldsymbol{Z} = \frac{1}{M}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}, \tag{1}$$

where each entry of $\boldsymbol{Y}$ is a normalized version of the coded genotype in $\boldsymbol{X}$, i.e.

$$\boldsymbol{Y}(n,m) = \frac{\boldsymbol{X}(n,m) - \mu_m}{\sigma_m} \tag{2}$$

for $n = 1, 2, \cdots, N$ and $m = 1, 2, \cdots, M$. Here, μ_m and σ_m denote the genotypic mean and standard deviation of SNP m in the total population, respectively. It can be shown that μ_m and σ_m relate to the population structure and allele frequencies as follows (Supplemental Text S1)

$$\mu_m = 2\frac{\sum_{k=1}^{K} N_k f_{km}}{N}, \tag{3}$$

$$\sigma_m^2 = 2\frac{\sum_{k=1}^{K} N_k f_{km}(1-f_{km})}{N} + 4\frac{\sum_{k=1}^{K}\sum_{\substack{l=1\\ k\neq l}}^{K} N_k N_l}{N^2}(f_{km}-f_{lm})^2. \tag{4}$$

The coded-allele frequency of SNP m in the total population can be found as $f_m = \mu_m/2$, where $m = 1, 2, \cdots, M$. The GRM is of dimension $N \times N$, whose diagonal elements are genotypic variance of individuals and off-diagonal elements are genotypic covariance between two individuals. It should be noted that genotypes follow mixed binomial distributions, and elements of $\boldsymbol{Z}$ are sample variances and covariance computed from the genotype data. The PCA analysis of population stratification is based on the eigen-analysis of the observed GRM $\boldsymbol{Z}$.

In practice, μ_m and σ_m are unknown, and therefore sample mean $\hat{\mu}_m$ and sample standard deviation $\hat{\sigma}_m$ or some other quantities similar are used instead. Usually, $\hat{\mu}_m = 2\hat{f}_m$ is used for the centralization in (2). In EIGENSOFT, $\sqrt{\hat{f}_m(1-\hat{f}_m)}$ is adopted for the normalization in (2), while $\sqrt{2\hat{f}_m(1-\hat{f}_m)}$ is employed in GCTA [22]. Different estimates of the allele frequency f_m were suggested as well [5, 6]. In the following theoretical derivations, we assume that μ_m and σ_m are known for the sake of simplicity. This will bring about clear connections between variance and covariance elements of the EGRM and allele frequencies of the SNPs used in the analysis. The connections further provide clues and insights for understanding the effect of rare variants on the PCA of population stratification.

Expected genetic relationship matrix

We assume that all individuals are unrelated. When the number of markers M goes large, the sample variance and covariance elements in the GRM will converge to their mathematical expectations in probability due to the law of large numbers. We denote the EGRM as Z, which is the expectation of the GRM $\boldsymbol{Z}$. Without loss of generality, we assume that individuals are ordered according to their population memberships. As such, the first N_1 rows and columns correspond to individuals from population 1, the next N_2 rows and columns are from population 2, and so on. Thus, the EGRM can be partitioned into a block matrix consisting of $K \times K$ submatrices

$$Z = \begin{pmatrix} Z_{11} & Z_{12} & \dots & Z_{1K} \\ Z_{12}^{\mathrm{T}} & Z_{22} & \dots & Z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{1K}^{\mathrm{T}} & Z_{2K}^{\mathrm{T}} & \dots & Z_{KK} \end{pmatrix}. \tag{5}$$

Diagonal sub-matrices of the EGRM Z have the following structure

$$Z_{kk} = \begin{pmatrix} z^k & z^{kk} & \dots & z^{kk} \\ z^{kk} & z^k & \dots & z^{kk} \\ \vdots & \vdots & \ddots & \vdots \\ z^{kk} & z^{kk} & \dots & z^k \end{pmatrix}, k = 1, 2, ..., K. \tag{6}$$

Here, diagonal elements of the submatrix Z_{kk} are of the mathematical form

$$z^k = \frac{1}{M}\sum_{m=1}^{M} \frac{2f_{km}(1-f_{km}) + (2f_{km}-\mu_m)^2}{\sigma_m^2} \tag{7}$$

which is the genotypic variance of individuals from population k. All off-diagonal elements share the form

$$z^{kk} = \frac{1}{M}\sum_{m=1}^{M} \frac{(2f_{km}-\mu_m)^2}{\sigma_m^2} \tag{8}$$

which is the genotypic covariance between two individuals from population k.

The off-diagonal sub-matrices of the EGRM Z are structured as follows

$$Z_{kl} = \begin{pmatrix} z^{kl} & z^{kl} & \dots & z^{kl} \\ z^{kl} & z^{kl} & \dots & z^{kl} \\ \vdots & \vdots & \ddots & \vdots \\ z^{kl} & z^{kl} & \dots & z^{kl} \end{pmatrix}, k \neq l. \tag{9}$$

Elements of Z_{kl} share the value

$$z^{kl} = \frac{1}{M}\sum_{m=1}^{M} \frac{(2f_{km}-\mu_m)(2f_{lm}-\mu_m)}{\sigma_m^2} \tag{10}$$

which is the genotypic covariance between one individual from population k and one from population l. Details of the derivations are presented in Supplemental Text S2.

The EGRM Z, the mathematical expectation of GRM $\boldsymbol{Z}$, depends only on the population sizes N_1, N_2, $\cdots$, N_K and allele frequencies of the M SNPs in K populations $[f_{k1}, f_{k2}, \cdots, f_{kM}]$, $k = 1, 2, \cdots, K$. Here, we treat the SNP number M and the allele frequencies as fixed numbers. A theoretical formulation of the PCA that considers genotypic values as a random vector and allele frequencies in different populations being random was proposed in Ma and Amos, 2010 [23]. Based on different assumptions, a genotypic variance-covariance matrix of the same structure was attained; nevertheless, elements of the EGRM are different from those of the variance-covariance matrix in [23].

Rare variants on the eigenvalues

Carrying out eigen-decomposition on the EGRM, it can be shown that there are $N_k - 1$ eigenvalues of value $z^k - z^{kk}$, for $k = 1, 2, \cdots, K$. Here,

$$z^k - z^{kk} = \frac{1}{M}\sum\nolimits_{m=1}^{M} \frac{2f_{km}(1-f_{km})}{\sigma_m^2}.$$

The sum of the $N - K$ eigenvalues is

$$\sum\nolimits_{k=1}^{K}(N_k-1)\left(z^k-z^{kk}\right)=\frac{1}{M}\sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K}\frac{2(N_k-1)f_{km}(1-f_{km})}{\sigma_m^2} \tag{11}$$

Clearly, variations in the $N-K$ PCs are entirely intra-population variance of the K populations. The sum of the other K eigenvalues is

$$\begin{aligned}\sum_{k=1}^{K}\lambda_k&=\sum\nolimits_{k=1}^{K}N_kz^{kk}+\sum\nolimits_{k=1}^{K}\left(z^k-z^{kk}\right)\\&=\sigma_{\mathrm{B}}^2+\sigma_{\mathrm{W}}^2,\end{aligned} \tag{12}$$

where

$$\sigma_{\mathrm{B}}^2=\sum\nolimits_{k=1}^{K}N_kz^{kk}=\frac{1}{M}\sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K}\frac{N_k\left(2f_{km}-\mu_m\right)^2}{\sigma_m^2}$$

represents the inter-population variance component and

$$\begin{aligned}\sigma_{\mathrm{W}}^2&=\sum\nolimits_{k=1}^{K}\left(z^k-z^{kk}\right)\\&=\frac{1}{M}\sum\nolimits_{m=1}^{M}\sum\nolimits_{k=1}^{K}\frac{2f_{km}\left(1-f_{km}\right)}{\sigma_m^2}\end{aligned}$$

stands for the intra-population variance component. Here, the intra-population covariance z^{kk} of the EGRM factor in the K PCs as the inter-population variance after the eigen-decomposition. Note that all inter-population variance is solely contained in the K PCs. Hence, it is sufficient to conduct the population stratification analysis based on the K PCs alone.

Given a set of SNPs, the divergence among the K populations can be measured by the ratio of the two variance components, i.e.

$$\mathrm{F}_{\mathrm{PC}}=\frac{\sigma_{\mathrm{B}}^2}{\sigma_{\mathrm{W}}^2}. \tag{13}$$

Its normalized version can be defined as

$$\mathrm{F}_{\mathrm{PC}}^{*}=\frac{\sigma_{\mathrm{B}}^2}{\sigma_{\mathrm{B}}^2+\sigma_{\mathrm{W}}^2}, \tag{14}$$

which measures the portion of inter-population variance in the K population informative PCs and takes a value between 0 and 1. The larger the F_{PC} and $\mathrm{F}_{\mathrm{PC}}^{*}$ are, the more divergence among the populations.

Note that $\mu_m=2f_m=\frac{2}{N}\sum_{k=1}^{K}N_kf_{km}$, terms in σ_{B}^2 are quadratic functions of f_{km}, $k=1, 2, \cdots, K$, $m=1, 2, \cdots, M$. However, terms in σ_{W}^2 are linear and quadratic functions of the frequencies. Therefore, F_{PC} will decrease when frequencies of the coded alleles become smaller, see Supplemental Text S3 for more details. As a result, instead of improving the population stratification analysis, using rare variants will deteriorate the analysis performance. Meanwhile, since σ_{B}^2 decreases faster than σ_{W}^2, the K eigenvalues will be closer to the other $N-K$ eigenvalues when frequencies of the coded alleles become smaller. Thus, the percent of variance explained by the K PCs will be smaller.

It can be shown that among the K eigenvalues, $K-1$ are of large values and one small. When intra-population variance z^k-z^{kk} of the K populations are equal, all inter-population variance is contained in the largest $K-1$ eigenvalues. In addition, when the sample size is large and the portions of populations remain, inter-population variance contained in the small eigenvalue is negligible, almost all information on the population structure is contained in the largest $K-1$ PCs.

For cases with two populations, it can be shown that the two eigenvalues are

$$\lambda_1=\frac{N_1z^{11}}{2}+\frac{N_2z^{22}}{2}+\frac{z^1-z^{11}}{2}+\frac{z^2-z^{22}}{2}+\frac{\sqrt{a}}{2},$$

$$\lambda_2=\frac{N_1z^{11}}{2}+\frac{N_2z^{22}}{2}+\frac{z^1-z^{11}}{2}+\frac{z^2-z^{22}}{2}-\frac{\sqrt{a}}{2},$$

where

$$\begin{aligned}a=&\left[\left(z^1-z^{11}\right)-\left(z^2-z^{22}\right)+N_1z^{11}-N_2z^{22}\right]^2\\&+4N_1N_2\left(z^{12}\right)^2.\end{aligned}$$

When inter-population variance of the two populations are equal, i.e. $z^1-z^{11}=z^2-z^{22}=\sigma_{\mathrm{W}}^2/2$, we have

$$\lambda_1=N_1z^{11}+N_2z^{22}+\frac{\sigma_{\mathrm{W}}^2}{2},$$

$$\lambda_2=\frac{\sigma_{\mathrm{W}}^2}{2}.$$

That is, all information on the population structure is contained in the largest PC. All proofs are presented in Supplemental Text S4.

Rare variants on the population distance

Suppose that $\boldsymbol{x}_k$, $k=1, 2, \cdots, K$ are the eigenvectors associated with the K eigenvalues containing inter-population variance. We can represent each individual as a point in the K-dimension space. Vector $\sqrt{\lambda_k}\boldsymbol{x}_k$ consists of coordinates of N individuals in the k-th dimension. Average value $\sqrt{\lambda_k}\boldsymbol{x}_k^{\mathrm{T}}\mathbf{1}_N/N$ represents center of the total population in the k-th dimension, where $\mathbf{1}_N$ is a column vector of dimension N and with each element as 1. Due to the structure of Z, individuals from the same population share the same coordinates in the K-dimension space, and the common points denote the representative points of the populations, or centers of the populations [23]. We define

$$
\begin{aligned}
d^2 &= \sum\nolimits_{k=1}^{K} \left[\sqrt{\lambda_k}\boldsymbol{x}_k - \left(\sqrt{\lambda_k}\boldsymbol{x}_k^{\mathrm{T}}\mathbf{1}_N/N\right)\mathbf{1}_N\right]^{\mathrm{T}} \\
&\qquad \left[\sqrt{\lambda_k}\boldsymbol{x}_k - \left(\sqrt{\lambda_k}\boldsymbol{x}_k^{\mathrm{T}}\mathbf{1}_N/N\right)\mathbf{1}_N\right] \\
&= \sum\nolimits_{k=1}^{K}\lambda_k - \frac{1}{N}\mathbf{1}_N^{\mathrm{T}} Z \mathbf{1}_N \\
&= \sum\nolimits_{k=1}^{K}\lambda_k - \frac{1}{N}\sum\nolimits_{k=1}^{K} N_k\left(z^k - z^{kk}\right) \qquad (15)
\end{aligned}
$$

which measures the population divergence as the sum of squared distances among populations. The proof is shown in Supplemental Text S5. Here, the second term in (15) is an average intra-population variance. As explained earlier that when allele frequencies become smaller, the K eigenvalues decrease. Hence, the population distance d^2 is smaller when using rare SNPs compared to common ones.

The 1000 genomes project data

We used genotype data from the 1000 Genomes Project to validate our theoretical results. Genotype data used in this work was obtained from Phase 3 version 5a of the 1000 Genomes Project [16, 17], which contains 84.4 million genetic markers and 2504 individuals from EUR, EAS, SAS, AFR and AMR. We extracted biallelic SNPs that are polymorphic in the total population. In summary, there are 77,279,863 SNPs; 5,261,820 are common ($0.1 < \mathrm{MAF} \le 0.5$), 6,770,457 are low-frequency ($0.01 < \mathrm{MAF} \le 0.1$), and 65,247,586 are rare ($0.0001 < \mathrm{MAF} \le 0.01$). Genotype data were converted to PLINK format with VCFtools [24]. The SNPs were divided into six frequency bins according to their MAFs, as shown in Table 1. For each bin, we randomly sampled approximately 100,000 SNPs using PLINK for the population stratification analyses. Here, MAFs of the SNPs in the total population were used, hence their frequencies in the five populations may be different and may not be in the same bins as in the total population. For SNPs in bin 5 and 6, they are polymorphic in the total population and may not be polymorphic in all of the five populations. PCAs were carried out, with GRMs computed by EIGENSOFT and PCAs on EGRMs conducted using GCTA. Default parameters were used when analyzing with EIGENSOFT, which excluded 68 and 116 outliers in the analyses of the data from frequency bin 5 and 6, respectively.

Results

Theoretical and empirical EGRMs

To calculate the theoretical results (5)–(10), we computed MAFs of the SNPs with PLINK. Values of variance z^k and covariance z^{kk}, z^{kl}, $k, l = 1, 2, \cdots K$, were calculated as in (7), (8), and (10), respectively, in which μ_m was computed with (3) and $\sigma_m^2 = 2f_m(1-f_m)$, $m = 1, 2, \cdots M$. Values of z^k and z^{kk} for the five populations with SNPs from the six frequency bins are presented in Table 2. Absolute values of the inter-population covariance z^{kl} are much smaller and the results are shown in Supplemental Tables S1–6.

To obtain the empirical values of variance z^k, as well as covariance z^{kk} and z^{kl}, we first computed GRMs with SNPs from the six bins using EIGENSOFT. Each GRM included $N(N+1)/2$ variance and covariance terms of N individuals based on the observed genotype data. Empirical value of z^k was computed as the average variance of the N_k individuals from population k. The empirical value of z^{kk} is the average covariance of $N_k(N_k-1)/2$ pairs of individuals from population k. Lastly, the value of z^{kl} is the average covariance of N_kN_l pairs of individuals, one from population k and one from population l. The results of z^k and z^{kk} are shown in Table 2, and those of z^{kl} are presented in Supplemental Tables S1–6.

We can see that across the six frequency bins, theoretical values of z^k, z^{kk}, and z^{kl} predicted by (7), (8), and (10), respectively, are close to their empirical values. When MAFs of the SNPs become smaller, intra-population covariance z^{kk} decreases. For example, z^{kk} was 0.2 for EAS with SNPs whose MAFs are between 0.4 and 0.5, which reduced to 0.003 in

Table 1 Summary of SNPs from the 1000 Genomes Project data

	Common SNP				Low-frequency SNP	Rare SNP
MAF	(0.4,0.5]	(0.3,0.4]	(0.2,0.3]	(0.1,0.2]	(0.01,0.1]	(0. 0001,0.01]
Pop (N)	Bin 1	Bin 2	Bin 3	Bin 4	Bin 5	Bin 6
EUR (503)	995,352	1,048,669	1,190,239	1,581,788	3,717,490	13,531,139
EAS (504)	970,359	1,010,549	1,130,961	1,440,178	2,982,582	14,189,976
AMR (347)	1,004,970	1,068,395	1,234,095	1,613,443	4,827,083	16,092,172
SAS (489)	1,001,330	1,077,620	1,239,727	1,626,183	3,989,855	15,562,799
AFR (661)	981,929	1,097,944	1,436,771	2,403,901	8,852,607	24,044,176
Total (2504)	1,023,570	1,105,365	1,308,728	1,824,157	6,770,457	65,247,586

MAF minor allele frequency, *Pop* population, *EUR* European, *EAS* East Asian, *AMR* American, *SAS* South Asian, *AFR* African

Table 2 Theoretical and empirical values of the variance and covariance elements of EGRMs

	MAF	EUR	EAS	AMR	SAS	AFR
z^k	(0.4,0.5]	1.06/0.97	1.11/1.02	1.04/0.97	1.04/0.96	1.15/1.05
	(0.3,0.4]	1.06/0.97	1.11/1.02	1.04/0.97	1.04/0.96	1.14/1.05
	(0.2,0.3]	1.05/0.96	1.09/1.00	1.03/0.97	1.03/0.96	1.17/1.07
	(0.1,0.2]	0.99/0.91	1.01/0.93	0.98/0.93	0.99/0.92	1.32/1.22
	(0.01,0.1]	0.61/0.57	0.50/0.47	0.73/0.70	0.56/0.53	2.50/2.37
	(0.0001,0.01]	0.71/0.71	0.94/0.94	0.82/0.82	0.98/0.98	1.46/1.46
z^{kk}	(0.4,0.5]	0.13/0.11	0.22/0.20	0.08/0.07	0.08/0.07	0.30/0.28
	(0.3,0.4]	0.13/0.11	0.22/0.20	0.07/0.06	0.08/0.07	0.29/0.26
	(0.2,0.3]	0.12/0.11	0.21/0.19	0.07/0.06	0.08/0.07	0.27/0.24
	(0.1,0.2]	0.11/0.10	0.18/0.17	0.07/0.06	0.08/0.07	0.27/0.25
	(0.01,0.1]	0.06/0.05	0.08/0.07	0.04/0.03	0.05/0.04	0.25/0.23
	(0.0001,0.01]	0.004/0.002	0.005/0.003	0.004/0.002	0.005/0.003	0.011/0.008

The first values are theoretical values of the variance and covariance, and second values are empirical values

the sixth bin that included rare SNPs only. A similar pattern can be observed for the other four populations. F_{PC} was estimated by (13) for the six bins, where empirical values of z^k and z^{kk} were used. The F_{PC} decreases from 93.85 in bin 1 to 55.01 in bin 5, and further to 1.83 in bin 6. Thus the divergence among the populations is much larger when measured by common SNPs than by rare ones.

PCAs of the 1000 genomes project data

With genotypes of SNPs from each frequency bin, we carried out PCAs of population stratification by EIGENSOFT, which was essentially based on the eigen-analysis of the observed GRMs. Scatter plots of the largest three PCs are shown in Figs. 1 and 2, where eigenvectors were scaled by square roots of their corresponding eigenvalues.

From Figs. 1 and 2, we can see patterns of population structure computed with common and less-frequency SNPs. For example, Figs. 1a-e and 2a-e displayed similar patterns, whereas the scatter plots based on rare SNPs differed significantly. For example, AMR and SAS are separated mostly by the third PC with common SNPs, while they are distinguished by the second PC with rare ones. The third PC from rare SNPs reveals mostly substructure of AFR, likely because more rare SNPs are polymorphic in AFR than in other populations. Portions of variance explained by the largest five PCs decrease from 17.09% in bin 1 to 10.41% in bin 5, and it falls dramatically to 0.74% with rare SNPs only. As a result, the five populations are more closely distributed around the origin in Figs. 1f and 2f, compared with those in Figs. 1a-e and 2a-e. Clearly, common variants show much better performance in dissecting the population structure than rare variants do.

PCAs of EGRMs

For each frequency bin, we also constructed a EGRM with structure as described in (5), (6), and (9), whose variance and covariance elements were their theoretical values calculated by (7), (8), and (10), respectively. We conducted PCAs of the EGRMs using GCTA, and scatter plots of the largest three PCs shown in Figs. 1 and 2. Large symbols in black are representative points or centers of the five continental populations from eigen-analyses of the EGRMs. Similarly, coordinates were scaled by square roots of their eigenvalues.

Upon comparing the representative points in Figs. 1 and 2, we can see that distances between populations decrease as the SNPs change from common to rare. Sum of the squared distance d^2 was calculated for the six frequency bins by (15), where λ_k, $k = 1, 2, \cdots K$ were the eigenvalues of the EGRM Z and z^k, z^{kk}, $k = 1, 2, \cdots K$ were their theoretical values. The d^2 decreases from 444.38 in bin 1 to 254.10 in bin 5, and further to 17.83 in bin 6.

In addition, when portions of variance explained by the PCs become small, deviations between the representative points of the populations and true centers of the populations can be observed. This is particularly evident in the scatter plots with rare SNPs. In the PCAs of a single population, such deviations are more obvious when percents of variance explained by the largest PCs are much smaller.

Discussion

We showed that all information about the population structure is contained in K PCs. Genotypic variance explained by the K PCs can be further decomposed

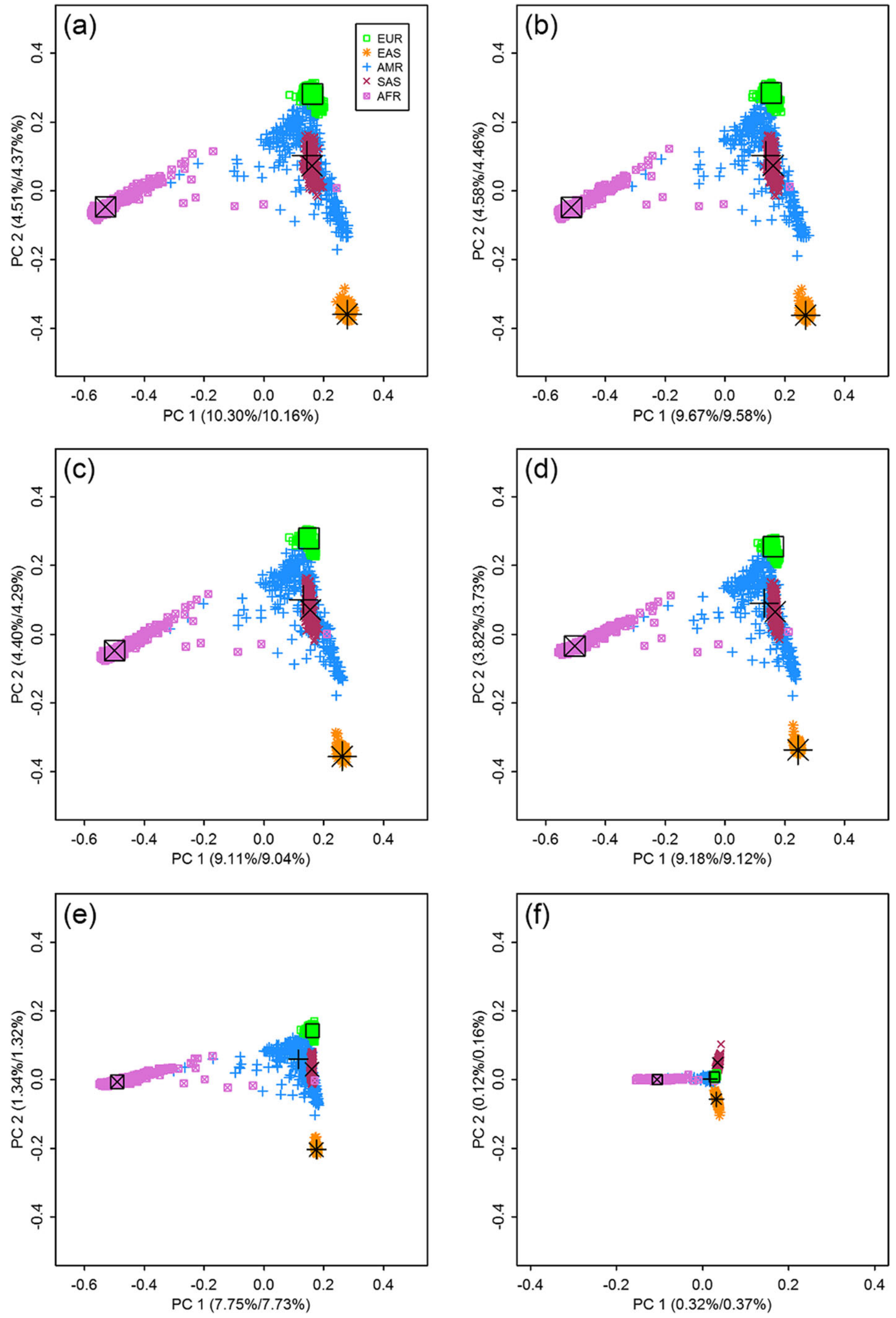

Fig. 1 Scatter plots and representative points with SNPs from six MAF bins, PC 1 vs. PC 2. (**a**) 0.4 < MAF ≤ 0.5 (**b**) 0.3 < MAF ≤ 0.4 (**c**) 0.2 < MAF ≤ 0.3 (**d**) 0.1 < MAF ≤ 0.2 (**e**) 0.01 < MAF ≤ 0.1 (**f**) 0.0001 < MAF ≤ 0.01. EUR: European, EAS: East Asian, AMR: American, SAS: South Asian, AFR: African. The first values in brackets are the percentages of variance explained from the PCAs of GRMs; and the second values are from the PCAs of EGRMs. Large symbols in black are the representative points of the five populations

into the intra-population variance σ_W^2 and inter-population variance σ_B^2. Using more SNPs will improve convergence of the realized GRM to its mathematical expectation, i.e. the EGRM. As a result, individuals belonging to the same population will be more closely distributed around the representative point or center of the population on the PC-PC plots. On the other hand, note that σ_B^2 is the average inter-population variance contributed from M SNPs. When rare variants are used, adding more SNPs will not

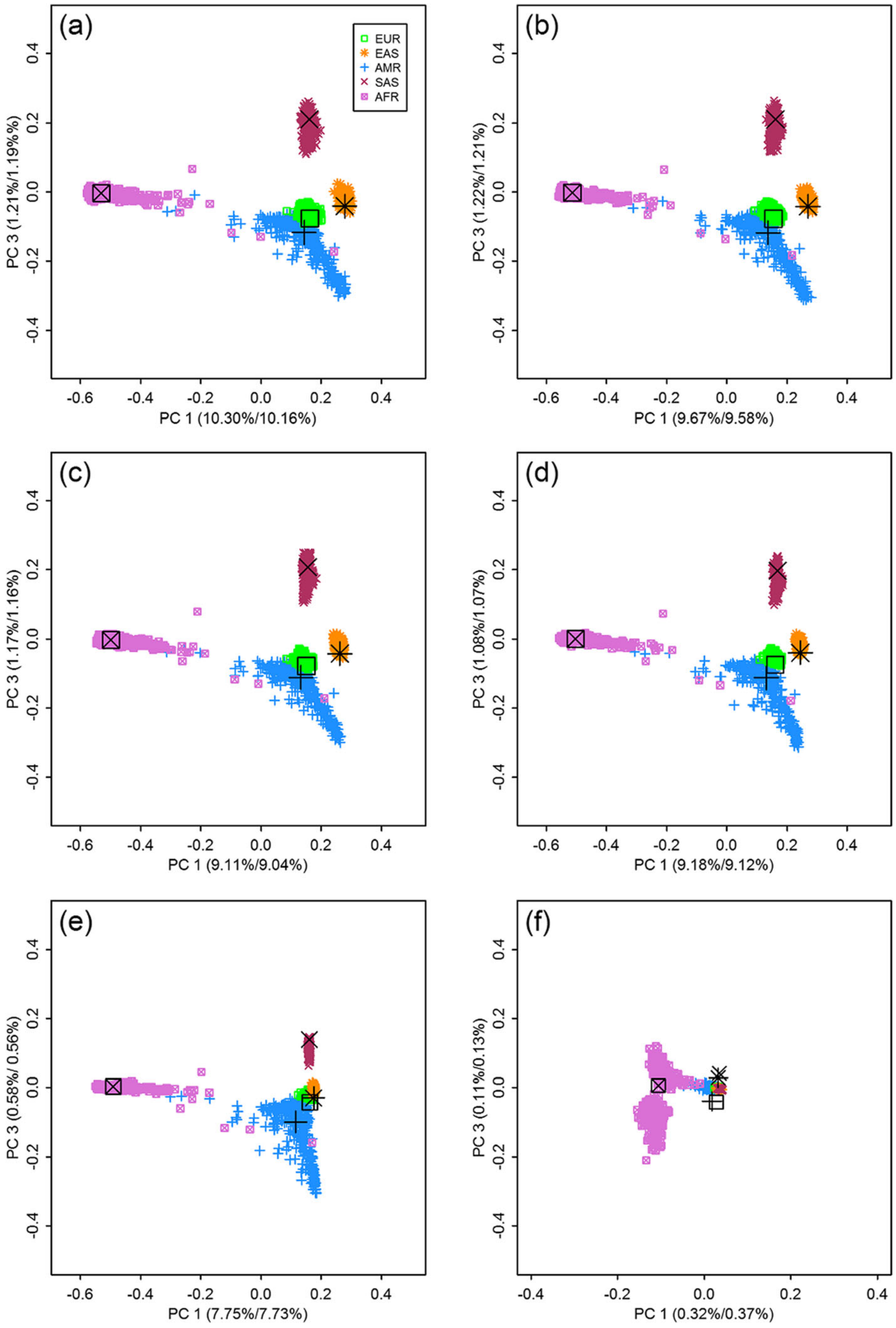

Fig. 2 Scatter plots and representative points with SNPs from six MAF bins, PC 1 vs. PC 3. (**a**) 0.4 < MAF ≤ 0.5 (**b**) 0.3 < MAF ≤ 0.4 (**c**) 0.2 < MAF ≤ 0.3 (**d**) 0.1 < MAF ≤ 0.2 (**e**) 0.01 < MAF ≤ 0.1 (**f**) 0.0001 < MAF ≤ 0.01. EUR: European, EAS: East Asian, AMR: American, SAS: South Asian, AFR: African. The first values in brackets are the percentages of variance explained from the PCAs of GRMs; and the second values are from the PCAs of EGRMs. Large symbols in black are the representative points of the five populations

increase the average level of σ_B^2, hence neither the ratio F_{PC} nor the sum of squared distances d^2 will improve. For same reason, using a combination of common and rare SNPs will result in lower F_{PC} and d^2 compared with using common SNPs only and therefore result in worse performance.

In the case where there is one SNP, our F_{PC} and F_{PC}^* can be further reduced to

$$\mathrm{F_{PC}} = \frac{\mathrm{F^*_{PC}}}{1-\mathrm{F^*_{PC}}} = 2\frac{\sum_{k=1}^{K} N_k\left(f_k-f\right)^2}{\sum_{k=1}^{K} f_k\left(1-f_k\right)},$$

where f_k is the allele frequency in the population k, and f the frequency in the total population. The classical Wright's fixation index $\mathrm{F_{ST}}$ is widely used to gauge population stratification [25], which measures the deviation from Hardy-Weinberg equilibrium at the total population level. In this case, it can be shown that

$$\frac{\mathrm{F_{ST}}}{1-\mathrm{F_{ST}}} = \frac{\sum_{k=1}^{K} N_k\left(f_k-f\right)^2}{\sum_{k=1}^{K} N_k f_k\left(1-f_k\right)}.$$

Therefore, our $\mathrm{F^*_{PC}}$ is much larger than $\mathrm{F_{ST.}}$ It is worth pointing out that we assign numeric values to genotypes as numbers of the coded alleles, hence our results are dependent on such coding scheme. $\mathrm{F_{ST}}$, however, does not involve in the numeric coding of genotypes. Note also that $\mathrm{F^*_{PC}}$ measures the portion of inter-population variance in the K population informative PCs. After the eigen-decomposition, most of the intra-population variance associated with the other N-K PCs was excluded. If our $\mathrm{F_{PC}}$ were defined as the ratio of the inter-population variance to the intra-population variance in the N PCs, it would be related to $\mathrm{F_{ST}}$ as $\mathrm{F_{PC}} = 2\mathrm{F_{ST}}/(1 - \mathrm{F_{ST}})$.

In GWAS, it is a common practice to conduct population stratification analyses using a large number of random markers [26], which usually yields satisfactory results. As shown in this work, the capability of dissecting population structure depends on the allele frequencies of markers used in the analyses, and common variants perform much better than rare ones. This is not much of a concern for GWAS because SNP panels implemented in the genotyping platforms are mostly of common ones. In sequencing studies, however, the majority of the called variants are rare, and selecting SNPs randomly will yield a large portion of rare SNPs, which will deteriorate the analysis performance. Therefore, it is necessary to restrict the selection to only the common SNPs when analyzing population stratification with sequencing data. This would also be true for controlling population stratification based on the linear mixed models [11–14].

In this work, we assumed that μ_m and σ_m are known constants in (2) in order to simplify the theoretical derivations. Our results are approximations of those when estimates of the two quantities are used. When sample size N is large, variations associated with $\hat{\mu}_m$ and $\hat{\sigma}_m$ are much smaller than those with the genotype data. Therefore, the mathematical expectations are largely taken with respect to the genotypes and difference between the two sets of results would be small. As shown in Table 2, the predicted values of the EGRM are close to their empirical values in the 1000 Genome Project data. We carried out additional simulation studies to evaluate the effect of lacking knowledge on μ_m and σ_m. We randomly chose one SNP from each of the six frequency bins (Supplemental Table S7). Based on their MAFs observed in the five populations of the 1000 Genomes Project, we simulated genotypes of five populations each with 500 individuals. Values of μ_m and σ_m were computed with the assumption of known population structure and MAF information, and theoretical values of z^k and z^{kk} were then calculated. For comparison, we first estimated $\hat{\mu}_m$ and $\hat{\sigma}_m$ from the simulated genotype data. $\boldsymbol{Y}(n, m)$ were normalized with $\hat{\mu}_m$ and $\hat{\sigma}_m$, and z^k and z^{kk} were obtained as averages of sample variance and covariance from 1000 replicates. The two sets of results are presented in Supplemental Table S8 and the differences between the two sets of results are negligible except for small differences in the results with the rare SNP.

Inferring population structure based on a large number of genome-wide markers are likely to include markers in linkage disequilibrium (LD). Practical concerns on the LD and choice of markers were extensively discussed in [5]. It is worth noting that each marker contributes to the elements in GRM additively, see eqs. S1–3 in Supplemental Text S2. Because of the linearity of expectation, our EGRM formulae as well as the eigen-analysis on the EGRM still hold when LD exists among markers. When the number of markers goes large, convergence of the GRM to EGRM will be slower with LD among markers, compared with the case that independent markers are used. Since there are always limited number of markers in the PCA practice, our EGRM and the eigen-analysis on it represent asymptotic results of the real PCA analysis.

Despite the fact that the vast majority of rare variants are population-specific, we showed that performance of the PCA of population stratification is better when based on common SNPs rather than rare ones. On the other hand, the PCA results with rare SNPs do reveal a population structure that differs from that of common SNPs. Existing methods may not exploit ancestry information embedded in the rare variants efficiently, and different approaches from those applied to common variants should be developed [26].

Conclusions

To quantify population divergence as a function of allele frequencies of genetic markers used in the PCA analysis, we derived the expected genetic relationship matrix. We proposed $\mathrm{F_{PC}}$, ratio of the inter-population variance to the intra-population variance, and population distance

d^2 as measures of population divergence. Our theoretical results as well as the analyses of the 1000 Genomes Project data showed that employing rare variants yielded smaller F_{PC} in the K population informative PCs, smaller d^2, and smaller portion of variance explained by the K PCs than those using common variants. Therefore, the PCA of population stratification performs worse with rare variants than with common ones. When analyzing population stratification with sequencing data, it is necessary to restrict the selection to only the common variants.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12863-020-0833-x.

Additional file 1 Text S1. Proof of eq. (4). **Text S2**. Proofs of eqs. (5–10). **Text S3**. F_{PC} as a function of allele frequencies. **Text S4**. Proofs of rare variants on the eigenvalues. **Text S5**. Proof of eq. (15). **Table S1**. Theoretical and empirical values of the inter-population covariance of EGRM (0.4 < MAF ≤ 0.5). **Table S2**. Theoretical and empirical values of the inter-population covariance of EGRM (0.3 < MAF ≤ 0.4). **Table S3**. Theoretical and empirical values of the inter-population covariance of EGRM (0.2 < MAF ≤ 0.3). **Table S4**. Theoretical and empirical values of the inter-population covariance of EGRM (0.1 < MAF ≤ 0.2). **Table S5**. Theoretical and empirical values of the inter-population covariance of EGRM (0.01 < MAF ≤ 0.1). **Table S6**. Theoretical and empirical values of the inter-population covariance of EGRM (0.0001 < MAF ≤ 0.01). **Table S7**. MAFs of the six SNPs used in the simulations. **Table S8**. Expected variance and covariance with and without the knowledge of μ_m and σ_m.

Abbreviations

GWAS: Genome-wide association study; LD: Linkage disequilibrium; MAF: Minor allele frequency; PC: Principal component; PCA: Principal component analysis; SNP: Single nucleotide polymorphism

Authors' contributions

SM: conceived the concept, conducted the analyses, and drafted the manuscript. GS: conceived the concept, supervised the work, reviewed and revised the manuscript. All authors have read and approved the manuscript.

Funding

This work was supported by the national Thousand Youth Talents Plan. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The datasets analyzed during the current study are available at https://www.internationalgenome.org. The accession number at https://www.ebi.ac.uk/ena is PRJNA262923.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 7 October 2019 Accepted: 25 February 2020
Published online: 17 March 2020

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet. 2008;9(5):356–69.
2. Freedman ML, Reich D, Penney KL, McDonald GJ, Mignault AA, Patterson N, Gabriel SB, Topol EJ, Smoller JW, Pato CN, Pato MT, Petryshen TL, Kolonel LN, Lander ES, Sklar P, Henderson B, Hirschhorn JN, Altshuler D. Assessing the impact of population stratification on genetic association studies. Nat Genet. 2004;36(4):388–93.
3. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. Hum Mol Genet. 2008;17(R2):R143–50.
4. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44(3):243–6.
5. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
6. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet. 2006;2(12):e190.
7. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81(3):559–75.
8. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155(2):945–59.
9. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164(4):1567–87.
10. Raj A, Stephens M, Pritchard JK. fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics. 2014;197(2):573–89.
11. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S, Buckler ES. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38(2):203–8.
12. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42(4):355–60.
13. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42(4):348–54.
14. Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. Nat Methods. 2014;11(4):407–9.
15. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet. 2016;17(6):333–51.
16. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, Gibbs RA, Hurles ME, McVean GA. A map of human genome variation from population-scale sequencing. Nature. 2010;467(7319):1061–73.
17. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56–65.
18. Baye TM, He H, Ding L, Kurowski BG, Zhang X, Martin LJ. Population structure analysis using rare and common functional variants. BMC Proc. 2011;5(Suppl 9):S8.
19. Siu H, Jin L, Xiong M. Manifold learning for human population structure studies. PLoS One. 2012;7(1):e29901.
20. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. Genet Epidemiol. 2013;37(1):99–109.
21. Zhang Y, Shen X, Pan W. Adjusting for population stratification in a fine scale with principal components and sequencing data. Genet Epidemiol. 2013;37(8):787–801.
22. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88(1):76–82.
23. Ma J, Amos CI. Theoretical formulation of principal components analysis to detect and correct for population stratification. PLoS One. 2010;5(9):e12510.

24. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R1000 Genomes Project Analysis Group. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8.
25. Wright S. The genetical structure of populations. Ann Eugenics. 1951;15: 323–45.
26. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11(7): 459–63.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.