Bioinformatics, 32(9), 2016, 1366–1372
doi: 10.1093/bioinformatics/btv752
Advance Access Publication Date: 31 December 2015
Original Paper

OXFORD

Genetics and population analysis

Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project

Dmitry Prokopenko[1,*], Julian Hecker[1], Edwin K. Silverman[2], Marcello Pagano[3], Markus M. Nöthen[4], Christian Dina[5,6,7,8], Christoph Lange[2,3] and Heide Loehlein Fier[1,3]

[1]Institute of Genomic Mathematics, University of Bonn, Bonn, Germany, [2]Channing Division of Network Medicine, Brigham and Women's Hospital, [3]Department of Biostatistics, Harvard School of Public Health, Boston, USA, [4]Institute of Human Genetics, University of Bonn, Bonn, Germany, [5]Institut National de la Santé et de la Recherche Médicale (INSERM) Unité Mixte de Recherche (UMR) 1087, l'institut du thorax, Nantes, France, [6]Centre National de la Recherche Scientifique (CNRS) UMR 6291, l'institut du thorax, Nantes, France, [7]Université de Nantes, l'institut du thorax, Nantes, France and [8]Centre Hospitalier Universitaire (CHU) de Nantes, l'institut du thorax, Service de Cardiologie, Nantes, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on 24 June 2015; revised on 22 November 2015; accepted on 19 December 2015

Abstract

Motivation: Population stratification is one of the major sources of confounding in genetic association studies, potentially causing false-positive and false-negative results. Here, we present a novel approach for the identification of population substructure in high-density genotyping data/next generation sequencing data. The approach exploits the co-appearances of rare genetic variants in individuals. The method can be applied to all available genetic loci and is computationally fast. Using sequencing data from the 1000 Genomes Project, the features of the approach are illustrated and compared to existing methodology (i.e. EIGENSTRAT). We examine the effects of different cut-offs for the minor allele frequency on the performance of the approach. We find that our approach works particularly well for genetic loci with very small minor allele frequencies. The results suggest that the inclusion of rare-variant data/sequencing data in our approach provides a much higher resolution picture of population substructure than it can be obtained with existing methodology. Furthermore, in simulation studies, we find scenarios where our method was able to control the type 1 error more precisely and showed higher power.

Availability and implementation:

Contact: dmitry.prokopenko@uni-bonn.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In genetic association studies, the importance of population substructure/admixture and its potentially negative impact on the analysis results have long been recognized. Numerous methods have been designed to detect population substructure and adjust genetic association analysis. So far, almost all methods implicitly assumed that the

analysis focus is on common variants, i.e. variants with common minor allele frequencies. Some of the most popular approaches, e.g. genomic control, structured association, EIGENSTRAT, STRATSCORE, multi-dimensional scaling, sparse principal component analysis (PCA) (Devlin *et al.*, 1999, 2004; Epstein *et al.*, 2007; Lee *et al.*, 2012; Li and Yu, 2008; Patterson *et al.*, 2006; Price *et al.*, 2006; Pritchard *et al.*, 2000; Reich and Goldstein, 2001; Satten *et al.*, 2001) extract the information about population substructure from the variance/covariance matrix of genotype data. The approaches have been shown to be efficient and are able to guard genetic association analysis under most circumstances.

With the arrival of next-generation sequencing data, rare variants, i.e. genetic loci with small minor allele frequencies, are moving to the center of the genetic association analysis. The hypothesis is that rare variant/sequencing data will enable the direct identification of disease susceptibility loci, which genome-wide association studies have not been very successful at. Although availability of rare variant/sequence analysis holds great promise, the effects of population substructure/admixture on the association analysis of rare variants are unclear and suitable adjustment methods have not been discussed widely in the literature (Babron *et al.*, 2012; Epstein *et al.*, 2012; Mathieson and McVean, 2012). The effectiveness of existing adjustment approaches which are mostly built on the estimation of the genetic variance/covariance matrix is unclear for rare variants, since the estimation of the allelic covariance based on rare variants might not provide sufficient power. This expected power loss can be attributed to the low variance of rare variant genotypes and thus the covariance matrix can be unreliable for small allele frequencies, e.g. it is suggested to use SNPs with common allele frequencies for the EIGENSTRAT approach. Moreover, the key feature of rare variant data in terms of population substructure is that such variants are genetically much 'younger' (Keinan and Clark, 2012; Kryukov *et al.*, 2009). The mutations that originated rare variants have occurred much more recently than those of common variants, allowing rare variant data to reflect even relatively new admixture/substructure. It is therefore unclear whether the application of adjustment methods, when they are computed based on common variants, is a suitable way to control for population substructure/admixture in rare-variant analysis.

In this article, we propose a method that is designed to detect population substructure/admixture both in rare variant data and in common variant data. Instead of utilizing the genetic variance/covariance matrix, the Jaccard similarity index (Jaccard, 1908) is used. Among existing similarity measures, the Jaccard similarity index is most suitable for the similarity calculation based on sparse binary data, since it emphasizes mutual presences of minor alleles. In comparison to identity-by-state similarity measure the Jaccard index has the advantage that it doesn't account for SNPs where a pair of individuals shares a 0 genotype, which is very efficient in a sparse dataset with only rare variants. We recommend, when available, to exclude potentially associated variants from the population structure estimation. If no such information is available, we recommend to include all rare variants in the analysis. In fact, rare variants typically do not show strong linkage disequilibrium with other rare or common variants. This is particularly due to their low minor allele frequencies (Pritchard, 2001; Pritchard and Cox, 2002). If the sample size is marginal and the sample contains only variants with very small allele frequency, excluding some long-range LD regions can reduce the resolution of population structure. However, if the Jaccard index is applied to common variants, we recommend to apply LD-pruning on the common variants prior to analysis.

The approach has the advantage that it can be applied to all available genetic loci and it is computationally fast. We have implemented a C program, which takes the genotype matrix and the position file as input and produces the Jaccard similarity matrix between individuals. The application to the complete dataset of the 1000 Genomes Project (2504 individuals, 31 202 510 variants with MAF $\leq$ 0.01) on one processor with 2 GB RAM running under Linux took only 8 h.

First, we assess the performance of our approach by application to the 1000 Genomes Project phase 3 data. A comparison of the results with existing methodology suggests that our approach provides better discrimination between subpopulations than current methodology. The approach seems to work particularly well, when we restricted the method to rare variant loci. In simulation studies, we evaluate the type-1 error of our approach and compare it to EIGENSTRAT, which uses PCA based on the genetic covariance matrix. For completeness we include in the simulation studies a model-based method ADMIXTURE (Alexander *et al.*, 2009), which uses a likelihood model to estimate population structure. Under different scenarios, the proposed methodology preserves the type-1 error much better.

2 Methods

Babron *et al.* (2012) have shown for the UK population that the population stratification pattern for rare variants is different from that of more common variants. Moreover, several Consortia (The 1000 Genomes Consortium, 2012, 2015; The UK10K Consortium, 2015) showed that rare variants are found more often within the same population, especially those which are present only twice through the sample.

Based on these findings, we aim to develop a measure, which is suitable to capture the most recent differences between individuals based on rare variant data. This information would represent a finer resolution of population structure in rare variant data. In order to keep the notation and the derivations simple, we outline the proposed method for one chromosome/genomic region. We assume that one defined genetic region or one entire chromosome has been sequenced in N individuals. For the purposes of this article, we do not distinguish whether one or two copies of the minor alleles are observed. Given the fact that we deal with rare variants it is rather unlikely to observe homozygous variants. We only differentiate between the absence and the presence of the minor allele at a variant.

Let G represent our modified genotype matrix with N individuals and M markers:

$$G = \begin{pmatrix} g_{11} & \cdots & g_{1M} \\ \vdots & \ddots & \vdots \\ g_{N1} & \cdots & g_{NM} \end{pmatrix}, g_{ik} \in \{0, 1\}$$

For two sample sets the Jaccard similarity coefficient is defined as the proportion of intersection between samples A and B divided by the proportion of their union:

$$\mathrm{Jac}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

When the two samples A and B represent binary vectors or modified genotypes, as in our case, the Jaccard similarity for individual i and j is defined as following. Let a_{ij} represent the number of markers shared between individual i and j, that is when $g_{ik} = 1 \wedge g_{jk} = 1$, b_{ij} represents the total number of markers, which satisfy $g_{ik} = 1 \wedge g_{jk} = 0$, c_{ij} is the total number of markers, which

satisfy $g_{ik} = 0 \wedge g_{jk} = 1$. Then, the Jaccard similarity coefficient is equal to:

$$\mathrm{Jac}_{ij} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Based on the measures Jac_{ij} we construct a symmetric similarity matrix with N rows and N columns for our study subjects, where each entry corresponds to a pair of study subjects i and j:

$$\mathrm{SM} = \begin{pmatrix} 1 & \mathrm{Jac}_{12} & \cdots & \mathrm{Jac}_{1N} \\ \mathrm{Jac}_{21} & 1 & \cdots & \mathrm{Jac}_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Jac}_{N1} & \mathrm{Jac}_{N2} & \cdots & 1 \end{pmatrix}$$

In the EIGENSTRAT approach (Patterson *et al.*, 2006; Price *et al.*, 2006) PCA is applied to the genetic variance/covariance matrix to visually detect population structures. The principal components could also be incorporated into the genetic association analysis as covariates in order to adjust for population substructure. PCA could also directly be applied to the similarity matrix SM. This corresponds to kernel PCA (Schölkopf *et al.*, 1998), where the Jaccard index is considered as a positive definite kernel (Gower, 1971). In this article the kernel (Jaccard) matrix is centered, as this allows us to relate the eigenvalues of the Jaccard matrix to the variance explained by a given component in the feature space. Subpopulation structure can be visualized by plotting the first principal components, and association analysis results can be adjusted for population substructure by including the principal components of SM as covariates.

The number of covariates to use in an association analysis strongly depends on the dataset, which is analyzed. It is important to note that there exists no generally accepted rule to determine the optimal number of eigenvectors sufficient to correct for population structure in association analysis. However, there exist several mostly heuristic methods in order to choose the number of components to use. The Kaiser-Guttman criterion and its modifications (Guttman, 1954; Lambert *et al.*, 1990) are based on the average value of the eigenvalues from the data matrix. Another popular method is to visually inspect the scree plot, which plots the eigenvalues in a decreasing order against the number of principal components. Usually one chooses only those PCs, which explain the highest amount of variance in the data, i.e. until the slope is still steep. Another modification of this approach is the eigengap heuristic (Lee *et al.*, 2010), which is based on differences between consecutive eigenvalues.

We provide the code that we used to calculate the Jaccard similarity matrix. It is provided on request from authors. The further calculation, analysis and simulations were performed with *R*.

To assess the properties of the Jaccard similarity matrix SM and compare them to similarity matrices that are built on the genetic variance/covariance matrix, the matrix SM will be computed for the 1000 Genomes Project data based on loci with common minor allele frequency (i.e. MAF $>$ 1%) and based on rare variant data, i.e. minor allele frequencies of $<$1%. Then, we will apply PCA to visualize the information about population substructure that is contained in the Jaccard similarity matrix. The results are benchmarked with the EIGENSTRAT approach on real data from the 1000 Genomes Project and further assessed in a simulation study.

3 Results

3.1 Application to 1000 genomes data

The data of the 1000 Genomes Project Consortium (2012) is an ideal instrument to assess the capability of the proposed methodology to detect population substructure in sequence data and to benchmark its results with the existing standard approach (i.e. PCA of the genetic covariance matrix). The recent phase 3 release of the project includes now 2504 indivduals from 26 populations. For the preliminary analysis of the 1000 Genomes Project data, we created two datasets. For dataset 1, we selected sequence data that is available for the autosomal chromosomes of 2504 unrelated subjects from 26 different populations (Table 1). For dataset 2, we created a subset with only European individuals (504 subjects). We assessed how well PCA based on the Jaccard similarity matrix SM is able to differentiate between the known populations. We compared the results of our proposed method to PCA based on the genetic covariance matrix as proposed by EIGENSTRAT (Price *et al.*, 2006). Each approach was applied twice to the 1000 Genomes Data, once to data for common variants (MAF $>$ 1%) and then to rare variant data (MAF$\leq$1%, singletons excluded).

3.1.1 Results for common variants

We created a setting that is commonly used in practice. We selected all common SNPs which remained after linkage disequilibrium (LD) pruning in order to reduce the LD structure between the variants. In addition, known, long-range LD regions were excluded from the analysis (Price *et al.*, 2008). This resulted in a total of 1 934 536 common variants for dataset 1 and 744 520 common variants for dataset 2. Then, based on the selected loci, we calculated the genetic covariance matrix and the Jaccard similarity matrix SM, and applied PCA to both matrices. We used all available variants, although, when information about potentially associated variants is provided, they should be excluded from population structure estimation. The

Table 1. Description of the population structure of the 1000 Genomes Project data

Continent	Population	Number of samples
AFR	ACB (African Caribbean in Barbados)	96
AFR	ASW (African Ancestry in Southwest US)	61
SAS	BEB (Bengali in Bangladesh)	86
EAS	CDX (Chinese Dai in Xishuangbanna, China)	93
EUR	CEU (Utah residents (CEPH) with Northern and Western European ancestry)	99
EAS	CHB (Han Chinese in Beijing, China)	103
EAS	CHS (Han Chinese South)	105
AMR	CLM (Colombian in Medellin, Colombia)	94
AFR	ESN (Esan in Nigeria)	99
EUR	FIN (Finnish from Finland)	99
EUR	GBR (British from England and Scotland)	91
SAS	GIH (Gujarati Indian in Houston, TX)	103
AFR	GWD (Gambian in Western Division, The Gambia)	113
EUR	IBS (Iberian populations in Spain)	107
SAS	ITU (Indian Telugu in the UK)	102
EAS	JPT (Japanese in Toyko, Japan)	104
EAS	KHW (Kinh in Ho Chi Minh City, Vietnam)	99
AFR	LWK (Luhya in Webuye, Kenya)	99
AFR	MSL (Mende in Sierra Leone)	85
AMR	MXL (Mexican Ancestry in Los Angeles, CA)	64
AMR	PEL (Peruvian in Lima, Peru)	85
SAS	PJL (Punjabi in Lahore, Pakistan)	96
AMR	PUR (Puerto Rican in Puerto Rico)	104
SAS	STU (Sri Lankan Tamil in the UK)	102
EUR	TSI (Toscani in Italia)	107
AFR	YRI (Yoruba in Ibadan, Nigeria)	108

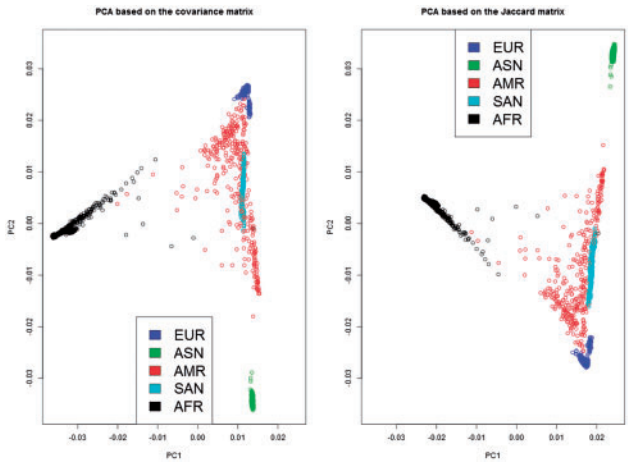

Fig. 1. Visualization of the first two principal components of the PCA based on the covariance matrix (left plot) and on the Jaccard matrix (right plot) for *common* variants, all populations

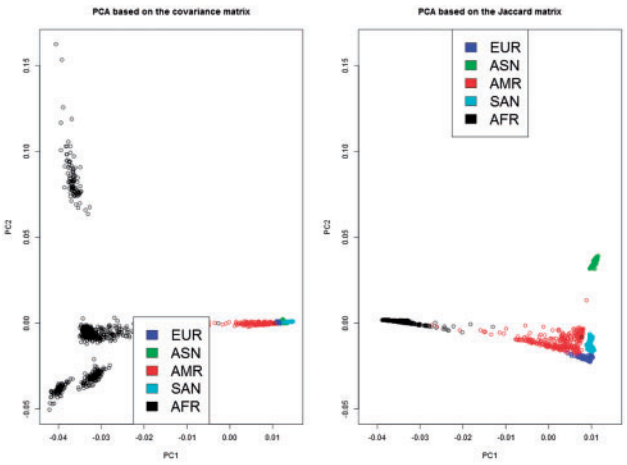

Fig. 3. Visualization of the first two principal components of the PCA based on the covariance matrix (left plot) and on the Jaccard matrix (right plot) for *rare* variants, all populations

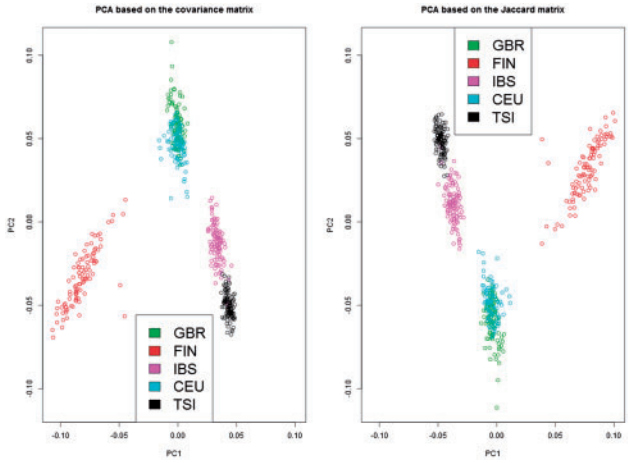

Fig. 2. Visualization of the first two principal components of the PCA based on the covariance matrix (left plot) and on the Jaccard matrix (right plot) for *common* variants, only European populations

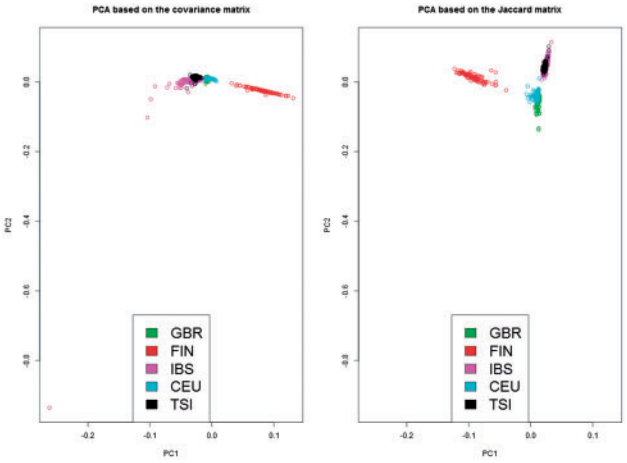

Fig. 4. Visualization of the first two principal components of the PCA based on the covariance matrix (left plot) and on the Jaccard matrix (right plot) for *rare* variants, only European populations

first two principal components of both matrices are shown in Figures 1 and 2. We used common axis ranges for both methods for comparison reasons. Both approaches provide quite similar results for common variants.

3.1.2 Results for rare variants

Because our approach was designed to identify population substructure based on rare variants, we re-calculated the two matrices, the genetic covariance matrix and the Jaccard similarity matrix SM, based on the rare variants of the prior utilized 1000 Genomes Project data, i.e. 31 202 510 loci for dataset 1 and 4 828 940 for dataset 2 with a MAF of <1% and excluded singletons. Then, PCA was again performed for both matrices.

The results are given in Figures 3–5. We used common axis ranges for both methods for comparison reasons. For PCA based on the genetic covariance matrix (EIGENSTRAT), the ability to distinguish between the different populations has been substantially diminished. Many of the populations are clumped together and are not identifiable. The left panel in Figure 3 depicts that the first two

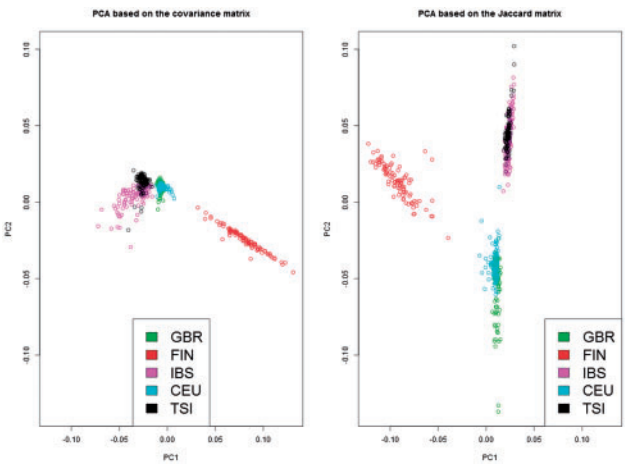

Fig. 5. Visualization of the first two principal components of the PCA based on the covariance matrix (left plot) and on the Jaccard matrix (right plot) for *rare* variants, only European populations, without outliers

PCs are driven by the large amount of variance within African populations which is captured by the covariance matrix on rare variants. The left panel in Figure 4 illustrates that the second PC is driven by individual with the largest variance in the sparse genotype matrix. This creates 'outliers' in the principal components and reduces the resolution of the picture. In Figure 5, we removed the Iberian outlier for visual inspection for both methods, but we can still see that only the Finnish population is separated well in the left panel (PCA based on the covariance matrix).

However, for the PCA based on the Jaccard similarity matrix SM, we observe the opposite effect. The plot for all 26 worldwide populations is similar to the one based on common variants and the first two principal components separate the main continents quite well. For the European subpopulations the first principal component separates the Finnish subpopulation and the second PC separates the two southern populations (IBS and TSI) from the two north-western populations (GBR and CEU).

3.2 Simulation study

We performed a simulation study for rare variants in order to evaluate the type 1 error and power of our approach. The goal of the simulations was to evaluate the performance of standard association analysis adjusted for population substructure, using PCA based on the Jaccard similarity matrix and using PCA based on the genetic covariance matrix. For completeness we also evaluated the performance of a model-based approach, which is not utilizing PCA. We modified the simulation design described in Price *et al* (2006) and utilized it. We used three different minor allele frequency cutoffs: 0.001, 0.005 and 0.01. A small diagram displaying the simulation workflow can be found in Supplementary Figure S1.

For each MAF setting we extracted the allele frequencies corresponding to this cutoff from the 1000 Genomes four European populations: CEU, GBR, IBS and TSI. We excluded the Finnish population, which is known to be a population isolate. Due to a restricted sample size (404 individuals), which is currently available at 1000 Genomes for given populations, we decided to keep the overall proportion of the allelic counts during our simulations. That means we generated data based on allelic counts and not allele frequencies which are sample size dependent. This allowed us to simulate larger sample sizes without loss of information about variants with really small minor allele frequencies. For the association analysis, we used the multi-loci rare variant test SNP-set Kernel Association Test (SKAT) (Wu *et al.* 2011). This is a flexible regression approach, which is able to incorporate population specific covariates. As covariates, we included principal components obtained either from genetic covariance matrix or from Jaccard similarity matrix or ancestry estimates obtained from ADMIXTURE. We expect the top three principal components to distinguish between the four subpopulations, so we used three principal components or ancestry estimates as covariates. As recommended by Price *et al.* we also included a setting with 10 principal components as covariates based on the genetic covariance matrix or the Jaccard similarity matrix.

In each replication, we considered three categories of SNPs: training SNPs, null SNPs and causal SNPs. Category 1 and 2 were simulated without any disease model, i.e. without genetic effect on the phenotype, whereas for category 3 we considered a multiplicative disease model with a relative risk of 2 for every SNP. Within each category, we randomly generated 100 000 SNPs for a given MAF cutoff from the drawn allele frequencies. In total that resulted in 300 000 independent SNPs. The proportions of simulated individuals coming from four populations, which mimic the four European populations were kept as following: (0.4, 0.1, 0.1 and 0.4) for cases and (0.1, 0.4, 0.4 and 0.1) for controls. Category 1 was used for population structure identification and the other 2 categories were used for association analysis, where we divided the 100 000 SNPs into 1000 groups, which are supposed to represent genes or regions of interest. We then performed the variance-component test SKAT on each group and assessed the type 1 error and power with the nominal significance level of 0.01.

We performed 500 replications with different minor allele cutoffs (0.001, 0.005, 0.01) and different sample sizes (500 and 2000

Table 2. Type I errors for the outlined simulation scenarios averaged over 500 replications with standard errors in brackets

MAF cutoff	Naive	3 PCs EIG	10 PCs EIG	3 PCs Jaccard	10 PCs Jaccard	ADM
2000 individuals						
0.001	0.564 (1.63e-02)	0.090 (8.99e-03)	0.052 (2.96e-02)	0.008 (2.86e-03)	0.008 (2.84e-03)	0.523 (1.78e-02)
0.005	0.420 (1.51e-02)	0.008 (2.80e-03)	0.008 (2.75e-03)	0.008 (2.85e-03)	0.008 (2.80e-03)	0.009 (2.92e-03)
0.01	0.360 (1.43e-02)	0.008 (2.71e-03)	0.008 (2.71e-03)	0.008 (2.58e-03)	0.008 (2.71e-03)	0.009 (2.76e-03)
500 individuals						
0.005	0.638 (1.54e-02)	0.004 (2.06e-03)	0.005 (2.15e-03)	0.005 (2.15e-03)	0.005 (2.16e-03)	0.595 (2.08e-02)
0.01	0.655 (1.44e-02)	0.004 (2.09e-03)	0.005 (2.25e-03)	0.004 (2.14e-03)	0.005 (2.10e-03)	0.005 (2.07e-03)

The nominal significance level was set to 0.01. EIG, EIGENSTRAT; ADM, ADMIXTURE.

Table 3. Power for the outlined simulation scenarios averaged over 500 replications with SEs in brackets

MAF cutoff	Naive	3 PCs EIG	10 PCs EIG	3 PCs Jaccard	10 PCs Jaccard	ADM
2000 individuals						
0.001	0.813 (1.28e-02)	0.414 (1.78e-02)	0.302 (9.63e-02)	0.131 (1.19e-02)	0.130 (1.17e-02)	0.785 (1.33e-02)
0.005	0.998 (1.29e-03)	0.684 (1.48e-02)	0.677 (1.57e-02)	0.683 (1.48e-02)	0.681 (1.56e-02)	0.678 (1.34e-02)
0.01	0.999 (1.53e-04)	0.972 (5.34e-03)	0.971 (5.52e-03)	0.974 (5.19e-03)	0.973 (5.31e-03)	0.970 (5.36e-03)
500 individuals						
0.005	0.869 (1.03e-02)	0.140 (1.32e-02)	0.112 (1.84e-02)	0.140 (1.14e-02)	0.136 (1.45e-02)	0.850 (1.24e-02)
0.01	0.988 (3.37e-03)	0.250 (1.51e-02)	0.214 (2.39e-02)	0.250 (1.51e-02)	0.244 (1.79e-02)	0.250 (1.37e-02)

The nominal significance level was set to 0.01. EIG, EIGENSTRAT; ADM, ADMIXTURE. Values in grey represent scenarios with inflated type 1 error.

individuals). We report the averaged over 500 replications type 1 error and power in Tables 2 and 3 with SEs in brackets.

The outlined correction for population structure for rare variants based on Jaccard similarity matrix performed better compared to EIGENSTRAT and ADMIXTURE. Our method showed great performance in the scenario with the smallest minor allele frequency cutoff, it was the only method, which was able to control the type 1 error in this scenario. All three methods were able to maintain the type 1 error under the nominal level in the other scenarios, except ADMIXTURE in the scenario with 500 individuals and 0.005 minor allele frequency cutoff. All methods showed similar power, when 3 covariates were included in the analysis. When increasing the number of covariates to 10, we found that our proposed correction method did not lose the achieved power and showed the best performance, whereas the values for EIGENSTRAT dropped down. Overall, our proposed correction method showed the best performance ratio between type one error and power.

4 Conclusion

Next generation sequencing data is becoming highly available and numerous methods have been developed to study the association of variants with low minor allele frequencies with complex traits (Ionita-Laza *et al.*, 2011; Li and Leal, 2008; Madsen *et al.*, 2009; Price *et al.*, 2010b; Wu *et al.*, 2011). However, the fact that rare variants might represent different patterns of population stratification is currently underestimated. Recently, several Consortia (1000 Genomes, UK10K) showed that variants with a very low minor allele frequency often cluster within one population. Those variants might represent 'younger' population-specific mutations and, hence, can be utilized for population structure correction in rare variant tests.

We propose an approach for the detection of population substructure in sequencing data that utilizes the information about co-appearances of the minor alleles between individuals. The attractive features of the approach are its computational speed and its capability to expose population substructure with higher resolution within sequence data than standard methodology. It requires no removal of long-range LD regions, when the dataset contains only rare variants. The approach can be seen as a modification of the EIGENSTRAT approach in which the genetic covariance matrix is replaced by the Jaccard similarity matrix. Consequently, existing methodology can be used to guard association tests against confounding due to population substructure. For example, the principal components based on the Jaccard similarity matrix can be included into the association analysis as covariates (Price *et al.*, 2010a, Wu *et al.*, 2011), or into the conditional sampling approach for rare variant analysis (Epstein *et al.*, 2012).

Several studies (Baye *et al.*, 2011; Zhang *et al.* 2013) assessed the population structure in rare variants by applying standard PCA, based on genetic covariance data. Zhang *et al.* showed that including multiple classical PCs based on variants with low minor allele frequency might lead to power loss (overestimation) in association analysis. They also showed that classical PCA, when applied to variants with MAF $< 1\%$ is sensitive to outliers and hence isn't effective in separating subpopulations. We confirm this in Figure 4, left panel. Here the second PC is driven by the largest variance of the sparse individual genotypes based on rare variants.

Our method based on Jaccard similarity matrix was specifically designed to detect population stratification based on variants with small minor allele frequencies. We can see that due to sparsity of datasets with rare variants standard covariance-based methods fail to detect population structure. Given the imminent arrival of large-scale whole genome sequencing (WGS) studies, there is a urgent need for methods that account for population structure in datasets with rare variants. We evaluated our approach by simulation studies and by application to the 1000 Genomes Project data. Our empirical evidence based on the application to the 1000 Genomes Project Data and simulation studies suggests that the proposed methodology provides a finer resolution picture of population substructure than existing methodology. The simulation studies show that our method is able to achieve equal or higher power in most of the simulated scenarios, while maintaining the type 1 error under the nominal level. Overall, our method did not lose the power, when increasing the number of principal components, included in the analysis, in comparison the classical PCA. This is an important feature for real data analysis with big sample sizes.

Funding

The project described was supported by Cure Alzheimer's fund, Award Number (R01MH081862, R01MH087590) from the National Institute of Mental Health and Award Number (R01HL089856, R01HL089897) from the National Heart, Lung and Blood Institute, the BONFOR Programme of the University of Bonn and Integrated Network IntegraMent. The research of C.L. was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A2A2028559).

Conflict of Interest: none declared.

References

Alexander,D.H. *et al.* (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.

Babron,M.C. *et al.* (2012) Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One*, **7**, e46519.

Baye,T.M. *et al.* (2011) Population structure analysis using rare and common functional variants. *BMC Proc.*, **5**, S8

Devlin,B. *et al.* (2004) Genomic control to the extreme. *Nat. Genet.*, **36**, 1129–1130.

Devlin,B. and Roeder,K. (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Epstein,M.P. *et al.* (2007) A simple and improved correction for population stratification in case-control studies. *Am. J. Hum. Genet.*, **80**, 921–930.

Epstein,M.P. *et al.* (2012) A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am. J. Hum. Genet.*, **91**, 215–223.

Gower,J.C. (1971) A general coefficient of similarity and some of its properties. *Biometrics*, **27**, 857.

Guttman,L. (1954) Some necessary conditions for common factor analysis. *Psychometrika*, **19**, 149–161.

Ionita-Laza,I. *et al.* (2011) A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet.*, **7**, e1001289.

Jaccard,P. (1908) Nouvelles recherches sur la distribution florale. *Bull. Soc. Vaud. Des. Sci. Nat.*, **44**, 223–270.

Keinan,A. and Clark,A.G. (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science*, **336**, 740–743.

Kryukov,G.V. *et al.* (2009) Power of deep, all-exon resequencing for discovery of human trait genes. *Proc. Natl. Acad. Sci. USA*, **106**, 3871–3876.

Lambert,Z.V. *et al.* (1990) Assessing sampling variation relative to number-of-factors criteria. *Educ. Psychol. Meas.*, **50**, 33–49.

Lee,A.B. *et al.* (2010) Discovering genetic ancestry using spectral graph theory. *Genet. Epidemiol.*, **34**, 51–59.

Lee,S. *et al.* (2012) Sparse principal component analysis for identifying ancestry-informative markers in genome-wide association studies. *Genet. Epidemiol.*, **36**, 293–302.

Li,B. and Leal,S.M. (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.*, **83**, 311–321.

Li,Q. and Yu,K. (2008) Improved correction for population stratification in genomewide association studies by identifying hidden population structures. *Genet. Epidemiol.*, **32**, 215–226.

Madsen,B.E. and Browning,S.R. (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.*, **5**, e1000384.

Mathieson,I. and McVean,G. (2012) Differential confounding of rare and common variants in spatially structured populations. *Nat. Genet.*, **44**, 243–246.

Patterson,N.J. *et al.* (2006) Population structure and eigenanalysis. *PLoS Genet.*, **2**, e190.

Price,A.L. *et al.* (2010a) Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. *Am. J. Hum. Genet.*, **86**, 832–838.

Price,A.L. *et al.* (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

Price,A.L. *et al.* (2008) Long-range LD can confound genome scans in admixed populations. *Am. J. Hum. Genet.*, **83**, 132–135.

Price,A.L. *et al.* (2010b) New approaches to population stratification in genome-wide association studies. *Nat. Rev. Genet.*, **11**, 459–463.

Pritchard,J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.

Pritchard,J.K. and Cox,N.J. (2002) The allelic architecture of human disease genes: common disease-common variant . . . or not? *Hum. Mol. Genet.*, **11**, 2417–2423.

Pritchard,J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170–181.

Reich,D. and Goldstein,D. (2001) Detecting association in a case-control study while correcting for population stratification. *Genet. Epidemiol.*, **20**, 4–16.

Satten,G.A. *et al.* (2001) Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.*, **68**, 466–477.

Schölkopf,B. *et al.* (1998) Nonlinear component analysis as a kernel eigenvalue problem. *Neural Comput.*, **10**, 1299–1319.

The 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.

The 1000 Genomes Project Consortium. (2015) A global reference for human genetic variation. *Nature*, **526**, 68–74.

The UK10K Consortium. (2015) The UK10K project identifies rare variants in health and disease. *Nature*, **526**, 82–90.

Wu,M.C. *et al.* (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, **89**, 82–93.

Zhang,Y. *et al.* (2013) Adjustment for population stratification via principal components in association analysis of rare variants. *Genet. Epidemiol.*, **37**, 99–109.