OXFORD

Briefings in Bioinformatics, 2023, **24(1)**, 1–11
https://doi.org/10.1093/bib/bbac611
Review

A comparison between similarity matrices for principal component analysis to assess population stratification in sequenced genetic data sets

Sanghun Lee, Georg Hahn, Julian Hecker, Sharon M. Lutz, Kristina Mullin, Alzheimer's Disease Neuroimaging Initiative (ADNI)†, Winston Hide, Lars Bertram, Dawn L. DeMeo, Rudolph E. Tanzi, Christoph Lange and Dmitry Prokopenko

Corresponding author. Sanghun Lee, Department of Medical Consilience, Division of Medicine, Graduate School, Dankook University, 152 Jukjeon-ro, Suji-gu, Yongin-si, Gyeonggi-do, 16890, Republic of Korea. E-mail: rehun@channing.harvard.edu

†Data used in preparation of this article were in part obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Abstract

Genetic similarity matrices are commonly used to assess population substructure (PS) in genetic studies. Through simulation studies and by the application to whole-genome sequencing (WGS) data, we evaluate the performance of three genetic similarity matrices: the unweighted and weighted Jaccard similarity matrices and the genetic relationship matrix. We describe different scenarios that can create numerical pitfalls and lead to incorrect conclusions in some instances. We consider scenarios in which PS is assessed based on loci that are located across the genome ('globally') and based on loci from a specific genomic region ('locally'). We also compare scenarios in which PS is evaluated based on loci from different minor allele frequency bins: common (>5%), low-frequency (5–0.5%) and rare (<0.5%) single-nucleotide variations (SNVs). Overall, we observe that all approaches provide the best clustering performance when computed based on rare SNVs. The performance of the similarity matrices is very similar for common and low-frequency variants, but for rare variants, the unweighted Jaccard matrix provides preferable clustering features. Based on visual inspection and in terms of standard clustering metrics, its clusters are the densest and the best separated in the principal component analysis of variants with rare SNVs compared with the other methods and different allele frequency cutoffs. In an application, we assessed the role of rare variants on local and global PS, using WGS data from multiethnic Alzheimer's disease data sets and European or East Asian populations from the 1000 Genome Project.

Keywords: population stratification, similarity matrix, Jaccard matrix, genetic relationship matrix, rare variant, principal component analysis

Dr. Sanghun Lee is an associate professor of Graduate School at Dankook University, a collaborator in the Channing Division of Network Medicine at Brigham and Women's Hospital, and a research scientist at the Harvard T.H. Chan School of Public Health.
Dr. Georg Hahn is a research scientist and instructor in the Biostatistics Department of the Harvard T.H. Chan School of Public Health. He holds an MSt in mathematics from Cambridge University and a PhD in statistics from Imperial College London.
Dr. Julian Hecker is an investigator in the Channing Division of Network Medicine at Brigham and Women's Hospital and Instructor in Medicine at Harvard Medical School. He holds a BSc in Mathematics, a BSc in Economics, a Master's degree in Mathematics, and a PhD in Epidemiology.
Dr. Sharon Lutz is an assistant professor in the Department of Population Medicine at Harvard Medical School and Harvard Pilgrim Health Care Institute. She holds a BA and MA in Mathematics and a MA and PhD in Biostatistics.
Kristina Mullin is a senior laboratory manager and researcher in the Genetics and Aging Research Unit at The Massachusetts General Hospital. Her research interests are in genetic factors leading to Alzheimer's disease. She holds a BSc in Biology.
Dr. Winston Hide is an associate professor at Harvard Medical School and Beth Israel Deaconess Medical Center (BIDMC). He is also a co-Director of the Non-Coding RNA Precision Diagnostics and Therapeutics Core Facility. His group develops and performs data-driven systems approaches to disease causal discovery.
Dr. Lars Bertram is a professor for Genome Analytics at the University of Lübeck in Germany and the head of the Lübeck Interdisciplinary Platform for Genome Analytics (LIGA). The main focus of his research is on the genetic and epigenetic foundations of genetically complex human traits and diseases, in particular those related to aging.
Dr. Dawn DeMeo is an associate professor and associate physician in the Division of Network Medicine, and Pulmonary and Critical Care Division in the Department of Medicine at Brigham and Women's Hospital and Harvard Medical School. She holds a BSc in Biology, a MPH and is a Doctor of Medicine (MD).
Dr. Rudolph Tanzi is the director of the Genetics and Aging Research Unit, Co-Director of the McCance Center for Brain Health, Co-Director of the Mass General Institute for Neurodegenerative Disease, and Vice-Chair of Neurology (Research), at Massachusetts General Hospital, and the Joseph P. and Rose F. Kennedy Professor of Neurology at Harvard Medical School. His main research focus is on Alzheimer's disease, its genetics, treatment and prevention. Dr. Tanzi co-discovered the first three Alzheimer's disease genes, including APP and directs the Cure Alzheimer's Fund Alzheimer's Genome Project, which identified the first neuroinflammation-related Alzheimer's gene, CD33.
Dr. Christoph Lange is a professor of Biostatistics at the Harvard T.H. Chan School of Public Health. His research interests are at the intersection of biostatistical methodology, numerical analysis and computer science.
Dr. Dmitry Prokopenko is an instructor at the Genetics and Aging Unit and McCance Center for Brain Health, Massachusetts General Hospital and Harvard Medical School. His research interests focus on Alzheimer's disease genetics and statistical and computational problems posed by omics data.

Received: September 13, 2022. **Revised:** December 7, 2022. **Accepted:** December 11, 2022

Introduction

Population stratification/substructure (PS), meaning systematic differences of minor allele frequencies (MAF) between populations, can cause spurious associations in population-based genome-wide association studies (GWASs) if not accounted for [1, 2]. To guard against biases introduced by PS, the most widely used method is to include principal components (PCs), calculated based on the genetic relationship matrix (GRM), as fixed covariates in a regression model [3, 4]. Another popular method is to include the GRM in mixed models as it captures both PS and cryptic relatedness [5, 6]. In the context of population genetics, genetic similarity matrices are utilized to compare different populations and their relative distance to each other.

Historically, single-nucleotide variations (SNVs) with a MAF > 5% are almost exclusively used for the computation of genetic similarity matrices because principal component analysis (PCA) of PS was reported to perform worse with rare variants than with common ones [7–9]. When over 80% of the SNVs are low frequency (0.5% < MAF ≤ 5%) or rare (MAF ≤ 0.5%) in the era of next-generation sequencing data [10, 11], these SNVs helped to identify additional PS and better control for Type 1 error in association studies of rare variants [12, 13]. We and others showed that using rare variants and alternative similarity matrices can provide a finer scale/higher resolution of PS [14–16]. Specifically, we have shown that the standard (i.e. unweighted) Jaccard index could be used to reveal finer scale population structure [15]. Subsequently, the weighted Jaccard index was introduced providing a more formal connection to the kinship coefficient and a similarity test for genetic outliers [17].

Local PS (i.e. PS within a small defined genomic locus/region on a chromosome) has been shown to differ from global PS along the genome [18, 19] and is well recognized in local ancestry studies [20, 21]. Therefore, capturing local PS could be helpful in the GWASs when small chromosomal regions can show stratification patterns that differ from the global stratification pattern. Here, we systematically compare three similarity matrices used in PCA: (1) the GRM, (2) the unweighted and (3) the weighted Jaccard similarity matrix. Using simulations, we assessed the ability of similarity matrices to capture PS under different MAF scenarios globally and locally. Furthermore, we validated our findings in whole-genome sequencing (WGS) data from a disease-focused [Alzheimer's disease (AD)] cohort and a population-based cohort from the 1000 Genome Project [22].

Methods

Simulation studies for similarity matrices

Simulation studies for similarity matrices were adopted from the R-package 'jacpop' ver. 0.6 (https://cran.r-project.org/web/packages/jacpop) and followed the design as described in Price *et al.* [4]. Briefly, we used the Balding–Nichols model [23] with a fixation index (a measure of population differentiation, F_{st}) of 0.1% (on the order of within-country differences) [24] to simulate 3000 subjects from three populations of equal sizes where the number of markers in each data set was roughly 10 000. Ancestral allele frequencies p for each SNP were drawn from a uniform distribution and allele frequencies for each population were drawn from a beta distribution with the following parameters: $p(1 - F_{st})/F_{st}$ and $(1 - p)(1 - F_{st})/F_{st}$. We divided the data set into three MAF bins (Table 1): common (>5%), low-frequency (5–0.5%) and rare SNVs (<0.5%) without singletons. We did not simulate linkage disequilibrium (LD) in order to make a fair comparison between different similarity matrices and MAF thresholds.

The similarity matrices were calculated with the locStra R-package ver. 1.9 [18]: (1) the GRM defined in Yang *et al.* [25], (2) the unweighted and (3) weighted Jaccard similarity matrices defined in Prokopenko *et al.* [15] and Schlauch *et al.* [17], respectively, where the weight was larger for the rarer variants because it was computed to be the inverse of the probability that two alleles selected belong to the set of minor alleles without replacement.

$$\mathrm{GRM}_{jk} = \frac{1}{N}\sum_{i=1}^{N}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - 2p_i)}$$

where x_{ij} or x_{ik} are the number of copies of the reference allele for the ith SNP of the individual (j or k, respectively) with N independent variants, and p_i is the frequency of the reference allele.

$$\text{Unweighted Jaccard}_{jk} = \frac{\sum_{i=1}^{N} G_{i,j}G_{i,k}}{\sum_{i=1}^{N} I\left[\sum_{l=1}^{2n} G_{i,l} > 1\right]}$$

where G is a genotype matrix with n individuals ($2n$ haploid genomes), with N independent variants, and $I[\cdot]$ is an indicator function, evaluating to 1 if true and 0 if false.

$$\text{Weighted Jaccard}_{jk} = \frac{\sum_{i=1}^{N} w_k G_{i,j}G_{i,k}}{\sum_{i=1}^{N} I\left[\sum_{l=1}^{2n} G_{i,l} > 1\right]}$$

where

$$w_k = \begin{cases} \dfrac{\binom{2n}{2}}{\binom{\sum_{l=1}^{2n} G_{i,l}}{2}} & \text{if } \sum_{l=1}^{2n} G_{i,l} > 1 \\ 0 & \text{if } \sum_{l=1}^{2n} G_{i,l} \le 1 \end{cases}$$

All analyses were carried out using R ver. 4.0.3, a free software from the R Foundation for Statistical Computing. Our simulation process was displayed in a flow diagram (Supplementary Figure 1).

The decomposition method for PCA analysis with similarity matrices

To extract PCs, we performed eigenvalue decomposition of the corresponding similarity matrices. While the GRM is column centered by construction, we have centered the other similarity matrices by subtracting the column means from their corresponding columns. Several R packages for the computation of eigenvalue or singular value decompositions are available (RSpectra ver. 0.16-1 or Matrix ver. 1.4-1): 'eigs_sym', 'eigs', 'eigen' and 'svd'. The 'eigs' function in RSpectra has shown the most stable outcome for all similarity matrices (Supplementary Figure 2). Therefore, we used the 'eigs' function for the comparison of clustering populations among the similarity matrices.

Population clustering in each similarity matrix under common, low-frequency and rare MAF simulations

We have evaluated how well the simulated populations are clustered based on the first and second eigenvectors for each matrix and respective simulation scenario. As an objective measure for comparing similarity matrices, a goodness of fit test based on the within sum of squares in clustering the three populations was employed. All results are based on a total of 100 independent simulations for each matrix with the same F_{st} of 0.1% and 3000

Table 1. The outlined simulation scenarios with three populations

Data set	Variants' condition	Sample numbers in each population	After removing minimum variants or singletons	
			Number of SNVs	MAF range
Data set 1	All SNVs MAC ≥ 2	1000 + 1000 + 1000	9904	0.033–50.0%
Data set 2	Common SNVs MAC ≥ 300	1000 + 1000 + 1000	9862	5.0–50.0%
Data set 3	Low-frequency SNVs MAC ≥ 30	1000 + 1000 + 1000	9897	0.5–5.0%
Data set 4	Rare SNVs MAC ≥ 2	1000 + 1000 + 1000	9920	0.033–0.5%

subjects in three populations having the category of MAFs as described above.

Canonical correlation analysis (CCA) between global and local substructure in each similarity matrix under common, low-frequency and rare MAF simulations

Local substructure can differ from global substructure, especially in the admixed populations [26], which may lead to false positives or reduce power in association studies. In order to examine each similarity matrix's ability to capture the local substructure compared with the global one, we generated a data set having PS with the fixation index (F_{st} = 0.1%) and 1000 subjects. SNVs were simulated either coming from two discrete populations (outer regions containing 50 000 + 50 000 SNVs, respectively), or from five discrete populations (mid-region containing 20 000 SNVs) using the same criteria of MAF thresholds as described previously. Next, we calculated global PCs based on all 120K SNVs (outer plus mid-regions) and local PCs were calculated using a non-overlapping sliding window of 2000 SNVs for 60 consecutive regions. We evaluated in 100 simulations the canonical correlation between global 10 PCs and local 10 PCs for each consecutive region based on each MAF data set using either GRM or Jaccard matrix. For CCA, we used the CCA R-package ver. 1.2.1 [27].

In addition, we generated confidence intervals of the expected difference in canonical correlation. For each matrix, we used bootstrapping with 1000 replicates, where we calculated the canonical correlation between global PCs and local PCs on a randomly selected coherent window of the same size (2000 SNVs).

WGS AD data sets

We have used two WGS data sets with AD cases and controls from the National Institute of Mental Health (NIMH) and the National Institute of Aging Alzheimer's Disease Sequencing Project (NIA ADSP; Supplementary Table 1). Briefly, the former sequencing and quality control is described elsewhere [28, 29]. Vcf files for the NIA ADSP cohort were obtained from the National Institute on Aging Genetics of Alzheimer's Disease Data Storage Site (NIAGADS) under the accession number NG00067. NIA ADSP was divided into three subpopulations based on self-reported population: non-Hispanic white (NHW), African American (AA) and Hispanic (HISP). We have verified the population assignment based on PCA using the Jaccard matrix and removed outliers that were more than 5 SD away from the mean based on each of the first ten PCs. CCA was performed similarly to simulations. Briefly, we have selected a 5 MB region at each of the reported GWAS hits in [30] centered at the corresponding hit. Using the same three MAF bins and two similarity matrices (Jaccard and GRM), we calculated global PCs based on a genome-wide independent set of 100 000 variants and local PCs based on 1 MB regions for each similarity matrix. We assessed the canonical correlation based on 10 PCs using only AD cases (for the NIMH family-based data set, we have extracted one case per family). Confidence intervals were calculated using bootstrapping with 1000 replicates, as described above.

WGS data sets in European and east Asian populations based on the 1000 genome project

We checked how well each similarity matrix could identify subpopulations in European and East Asian populations from the 1000 Genome Project WGS data. The 503 European-ancestry individuals consisted of 99 Utah residents (CEPH) with northern and western European ancestry (CEU), 107 individuals from Iberian populations in Spain (IBS), 91 British in England and Scotland (GBR), 99 Finnish in Finland (FIN) and 107 Toscani in Italia (TSI), whereas the 504 East Asian-ancestry individuals consisted of 104 Japanese in Tokyo, Japan (JPT), 103 Han Chinese in Beijing, China (CHB), 105 Southern Han Chinese, China (CHS), 93 Chinese Dai in Xishuangbanna, China (CDX) and 99 Kinh in Ho Chi Minh City, Vietnam (KHV). The same criteria of MAFs were applied: common, low-frequency and rare SNVs without singletons. The SNV quality criteria were 0.0% genotyping missing rate and no deviations from Hardy–Weinberg proportions (P-value < 10^{-6}). LD pruning was performed to select independent variants and reduce the computational burden. The following clustering metrics assuming five population clusters in each PCA plot were used to assess clustering performance based on different similarity matrices: (1) within sum of squares, (2) Davies–Bouldin index, (3) Fowlkes–Mallows index and (4) average silhouette width [31–33]. Specifically, the Davies–Bouldin index is based on a ratio of within-cluster and between-cluster distances [31]. The optimal clustering solution has the smallest Davies–Bouldin index value. Fowlkes–Mallows index is a performance metric to evaluate and compare a cluster label set with a true label set [32]. A higher value for the Fowlkes–Mallows index indicates a higher similarity between the clusters. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared with other clusters (separation) [33]. The value of the silhouette ranges between [−1, 1], where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. In our analysis, the Fowlkes–Mallows index was based on K-medoids, whereas the rest indexes were computed using the Euclidean distance measure. K-medoids clustering uses the most centrally located object of a cluster instead of using the mean point as the center of a cluster in K-means. Therefore, it is more robust to noises and outliers compared with K-means. We would like to point out

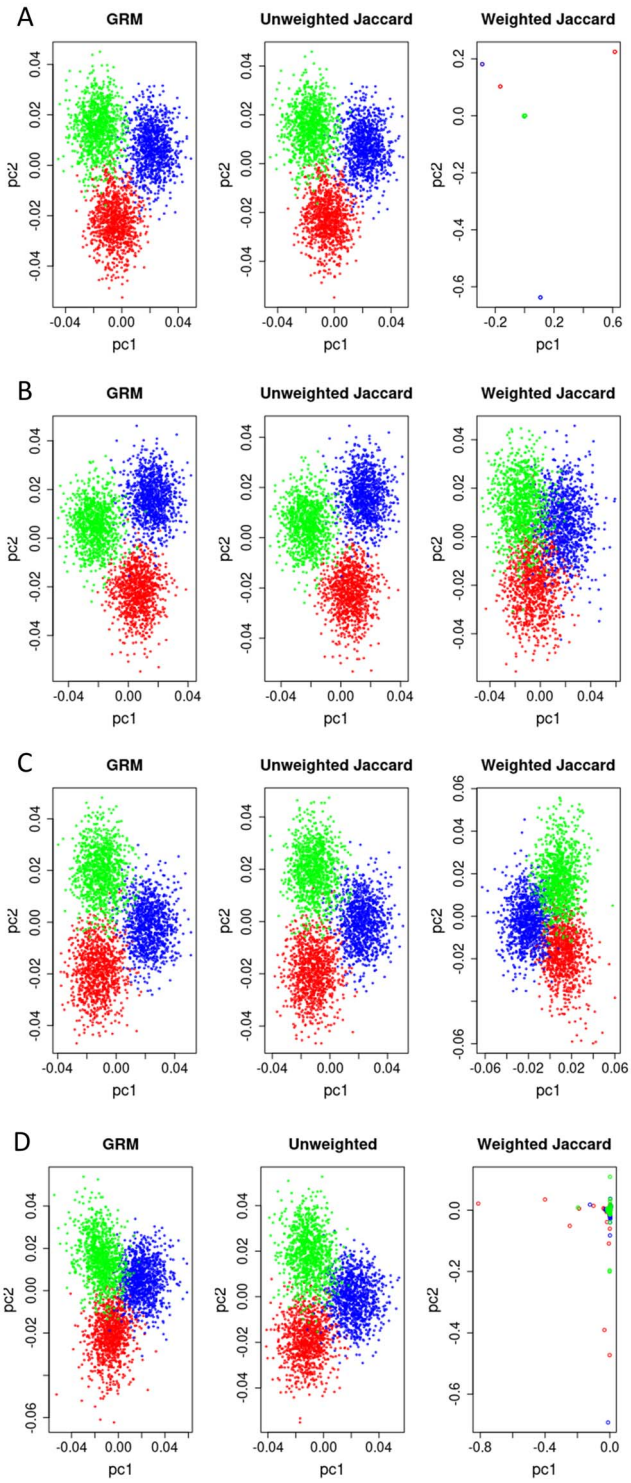

Figure 1. PCA plots according to similarity matrices in all (**A**: Data set 1), common (**B**: Data set 2), low-frequency (**C**: Data set 3) and rare SNVs (**D**: Data set 4). In simulation data sets, three populations were assigned to green, red or blue colors. In simulation scenarios that include rare variants (A and D), weighted Jaccard matrix shows very poor clustering because of extreme weights for the rarest variants. All simulation data (A–C) showed little difference between GRM and unweighted Jaccard matrix except the rare SNVs (D) where red population is more clustered in the unweighted Jaccard matrix.

that such a clustering evaluation is best applicable to discrete subpopulations, separation of which can be evaluated in the PC space of corresponding PCA method.

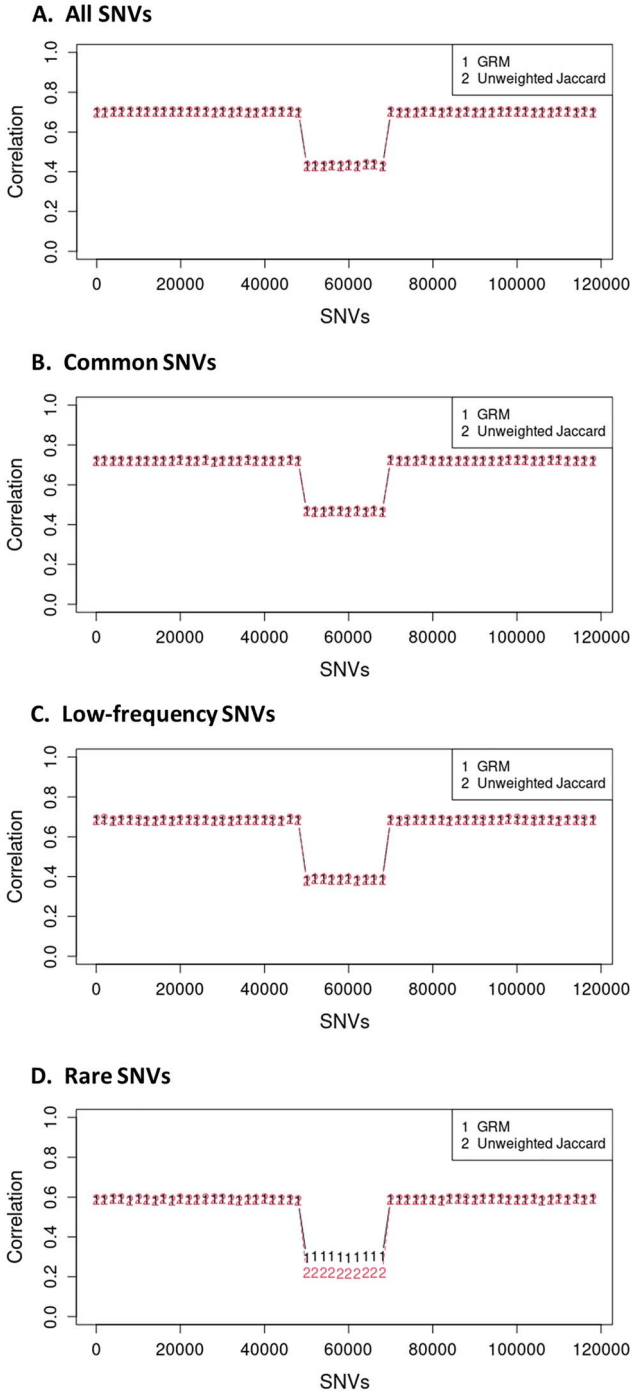

Figure 2. The highest correlation between global and local PCs in each similarity matrix reflected the local substructure of all (**A**), common (**B**), low-frequency (**C**) and rare SNVs (**D**), as illustrated in the plots with an average of 100 data sets (mid-region: five populations versus outer-region: two populations). The correlation in mid-region is lower as MAF decreases. The unweighted Jaccard matrix produces the lowest correlation in the case of rare variants.

Results

Comparison of clustering performance using each similarity matrix under various MAF conditions in a simulation study

We have simulated three populations with an F_{st} of 0.1% and used different MAF bins as described in Table 1, Supplementary Figure 3 and Methods. We performed PCA based on each similarity

Table 2. The goodness of fit test based on the within sum of squares in PC1 and PC2 for population stratification in each similarity matrix under the common, low-frequency and rare MAF simulations (total sum of squares = 2.0)

MAF	Simulation	Number of SNVs	GRM	Unweighted Jaccard matrix	Weighted Jaccard matrix	P-value*
All SNVs	Data set 1	9904	0.4702	0.4711	1.998	–
	Repeated simulation ($n = 100$)	9899.97 ± 9.66	0.4709 ± 0.0126	0.4718 ± 0.0126	1.998 ± 0.00096	0.6079
Common SNVs	Data set 2	9862	0.4468	0.4472	0.8178	–
	Repeated simulation ($n = 100$)	9856.03 ± 11.35	0.4371 ± 0.0105	0.4380 ± 0.0105	0.8022 ± 0.0274	0.5707
Low-frequency SNVs	Data set 3	9897	0.5051	0.5025	0.7504	–
	Repeated simulation ($n = 100$)	9918.47 ± 28.05	0.4834 ± 0.0118	0.4801 ± 0.0118	0.7175 ± 0.0200	0.0542
Rare SNVs	Data set 4	9920	0.6288	0.5776	1.996	–
	Repeated simulation ($n = 100$)	9931.64 ± 36.48	0.6120 ± 0.0143	0.5608 ± 0.0132	1.995 ± 0.0022	<2.2E−16

*P-value was calculated with t-test between GRM and unweighted Jaccard matrix based on 100 simulations.

Table 3. The average correlation between global and local PCs in each similarity matrix with 100 data sets (mid-region: five populations versus outer-region: two populations)

MAF	Simulation	GRM				Unweighted Jaccard matrix			
		Outer-region	Mid-region	Difference	P-value*	Outer-region	Mid-region	Difference	P-value*
All SNVs	Repeated simulation ($n = 100$)	0.7017 ± 0.0013	0.4332 ± 0.0031	0.2685	–	0.7011 ± 0.0013	0.4320 ± 0.0031	0.2691	–
Common SNVs	Repeated simulation ($n = 100$)	0.7221 ± 0.0016	0.4684 ± 0.0024	0.2538	1.53×10^{-09}	0.7216 ± 0.0016	0.4672 ± 0.0026	0.2544	1.77×10^{-09}
Low-frequency SNVs	Repeated simulation ($n = 100$)	0.6837 ± 0.0017	0.3905 ± 0.0032	0.2931	1.17×10^{-12}	0.6844 ± 0.0016	0.3849 ± 0.0029	0.2995	3.09×10^{-14}
Rare SNVs	Repeated simulation ($n = 100$)	0.5897 ± 0.0024	0.3023 ± 0.0030	0.2874	5.59×10^{-11}	0.5888 ± 0.0027	0.2205 ± 0.0023	0.3683	$<2.2 \times 10^{-16}$

*P-value was calculated with t-test where the differences were compared with the case of all SNVs in each GRM and unweighted Jaccard matrix.

matrix and assessed the clustering visually and using different goodness of fit tests. Most of the variance is explained by the first two PCs for each similarity matrix (Supplementary Figure 4). We did not find substantial differences in the PCA plots among three similarity matrices, except for the weighted Jaccard matrix (PC1 versus PC2, Figure 1A–D). When rare variants are included in the data set, PCA based on the weighted Jaccard matrix leads to extreme outliers as compared with other matrices (Data sets 1 and 4 shown in Figure 1A and D). This is a natural consequence of the weight definition in Schlauch *et al*. [17] that up and down weighs all columns of the similarity matrix by an inverse binomial weight vector computed using the number of variants, thus making the weighted Jaccard matrix extremely sensitive to the rarest SNVs (Supplementary Figure 5).

Next, we compared how well the first two PCs based on different similarity matrices can separate the three artificial populations. As a goodness of fit, we used the total within sum of squares criterion with the original population assignment as a baseline. Well-defined dense clusters provide a small value, whereas dispersed clusters or clusters with outliers increase the total within sum of squares. Among the similarity matrices, the weighted Jaccard matrix had the largest values driven by the sensitivity of the weight definition, which in the extreme case of rare variants generates outliers as shown before (Table 2). However, we observed that the unweighted Jaccard matrix had a much smaller within sum of squares than both: the weighted Jaccard matrix and the GRM based on rare variants (Figure 1D). Furthermore, the repeated simulations ($n = 100$) confirmed the advantages of the unweighted Jaccard matrix for clustering subjects with rare variants (Table 2 and Supplementary Figure 6). The results in simulations are consistent regardless of larger or smaller numbers of SNVs (around 20 000 or 5000 SNVs, Supplementary Table 2).

The performance among similarity matrices and the role of the rare variants capturing the local substructure based on CCA

Next, we compared the performance of similarity matrices to capture local PS. The weighted Jaccard matrix was not included in the comparison because of its performance in the clustering analysis. To create local PS, we simulated a genetic region where the outer parts contained two artificial populations (50 000 + 50 000 SNVs) and the inner part contained five artificial populations (mid-region: 20 000 SNVs). Using the same MAF bins as in the previous section, we calculated global PCs based on the whole region and local PCs based on evenly divided regions ($n = 60$ and window size = 2000 SNVs) for each similarity matrix. Finally, we compared the canonical correlation between 10 local PCs and 10 global PCs. In our repeated simulations ($n = 100$), the PS was detected, indicated by the drop in the correlation between global and local PCs in the CCA (Figure 2). In the CCA with common and low-frequency SNVs, the local substructure was captured but there was no difference among similarity matrices (Figure 2B and C). However, we observed that the correlation between global and local PCs in the mid-region, especially with rare SNVs, was lower in the unweighted Jaccard matrix than in the GRM (Figure 2D). Moreover, the gap in correlation between mid- and outer-regions of the unweighted Jaccard matrix is significantly larger in the case of rare SNVs compared with low-frequency or common SNVs (0.3693, 0.2995, 0.2544, respectively, in Table 3), which suggests that local substructure

could be captured better using rare variants. Therefore, the CCA simulations suggested that the unweighted Jaccard matrix can capture a finer local substructure using rare SNVs.

In Figure 2, we observed that the similarity matrices clearly capture the mid-region containing rare SNVs. A natural question in this context pertains to the significance of certain correlations observed in such local substructure plots: how would one determine that an observed area of local substructure differs from what is expected? We attempted to answer this question using bootstrapping. For a given data set of genetic regions such as the one depicted in Figure 2, we applied the following procedure. We calculated global PCs based on the whole data set and local PCs based on randomly selected coherent windows of the given window size (instead of evenly divided regions of a given window size). We then computed the canonical correlation between the 10 local PCs per window and the 10 global PCs for the whole data set. This is repeated for R random windows (we use R = 1000 in our experiments), thus allowing us to obtain an empirical bootstrap distribution of correlations for random windows. We reported the mean and the 5% and 95% quantiles of this bootstrap distribution. Using the simulation setting of the Methods section for CCA, Supplementary Figure 7 (with local PS) and Supplementary Figure 8 (without PS) show the results of this experiment. The mean of the bootstrap distribution is given as a solid line, with the 5% and 95% confidence bands as dashed lines. The bootstrapping method based on the Jaccard matrix is displayed in blue, whereas results based on the GRM are shown in purple. We observed that this metric allows one to quantify to some extent the expected range of correlations for a given data set. Observed correlations outside of the confidence band could be flagged for special attention. We observed in this example that using the Jaccard similarity matrix results in lower correlation with smaller confidence bands than using the GRM.

Assessment of local population structure in a WGS data set of AD

Motivated by simulation results in the previous section, we evaluated local PS around the 'APOE' region, a well-known risk factor for AD, in two WGS data sets of AD (Methods). In the 'APOE' locus, we see that the correlation between global and local structure is higher in NIMH AD cases with a smaller sample size ($n = 418$, Figure 3A) than in NIA ADSP NHW AD cases ($n = 4515$, Figure 3B). Also, in NIMH, we see a large difference in performance between the unweighted Jaccard matrix and GRM. This could be because of lower minor allele counts (MAC) in low frequency and rare bins due to sample size. In data sets with predominantly NHW individuals, we see that the correlation between the global and local PCs is consistently lower when using the Jaccard matrix as compared with the GRM matrix for low-frequency and especially, rare variants. Finally, although none of the methods shows a significant (i.e. outside of confidence bands) correlation difference in the region that surrounds rs429358 (middle region), we do see such difference in the neighboring regions among NHW and AA for low-frequency variants and among HISP for rare variants (Figure 3B–D). This could imply that local population structure using low-frequency/rare variants is different from global structure in the 'APOE' region, which might indicate some allelic heterogeneity in this region. This observation would be supported by the varying effect size estimates of the 'APOE $\varepsilon 4$' allele across ethnicities [34–36].

Next, we extended this analysis to the reported loci associated with AD (38 SNPs, respectively) in the recent large GWASs

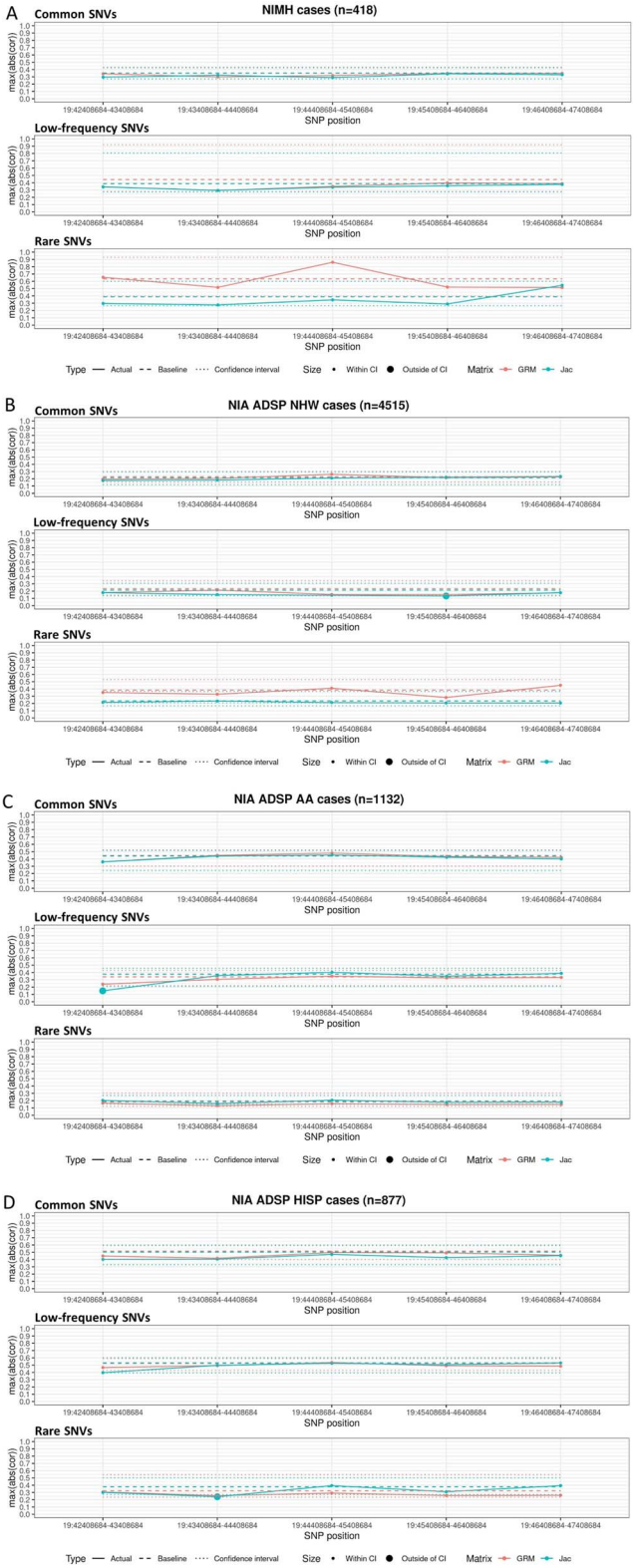

Figure 3. Maximum canonical correlation between local and global PCs in unrelated cases from NIMH (**A**), NIA ADSP NHW (**B**), NIA ADSP AA (**C**), NIA ADSP HISP population data sets (**D**) for the 'APOE' region centered around rs429358. Confidence bands were calculated using bootstrapping with 1000 replicates as described in Methods (red line: GRM versus blue line: unweighted Jaccard matrix).

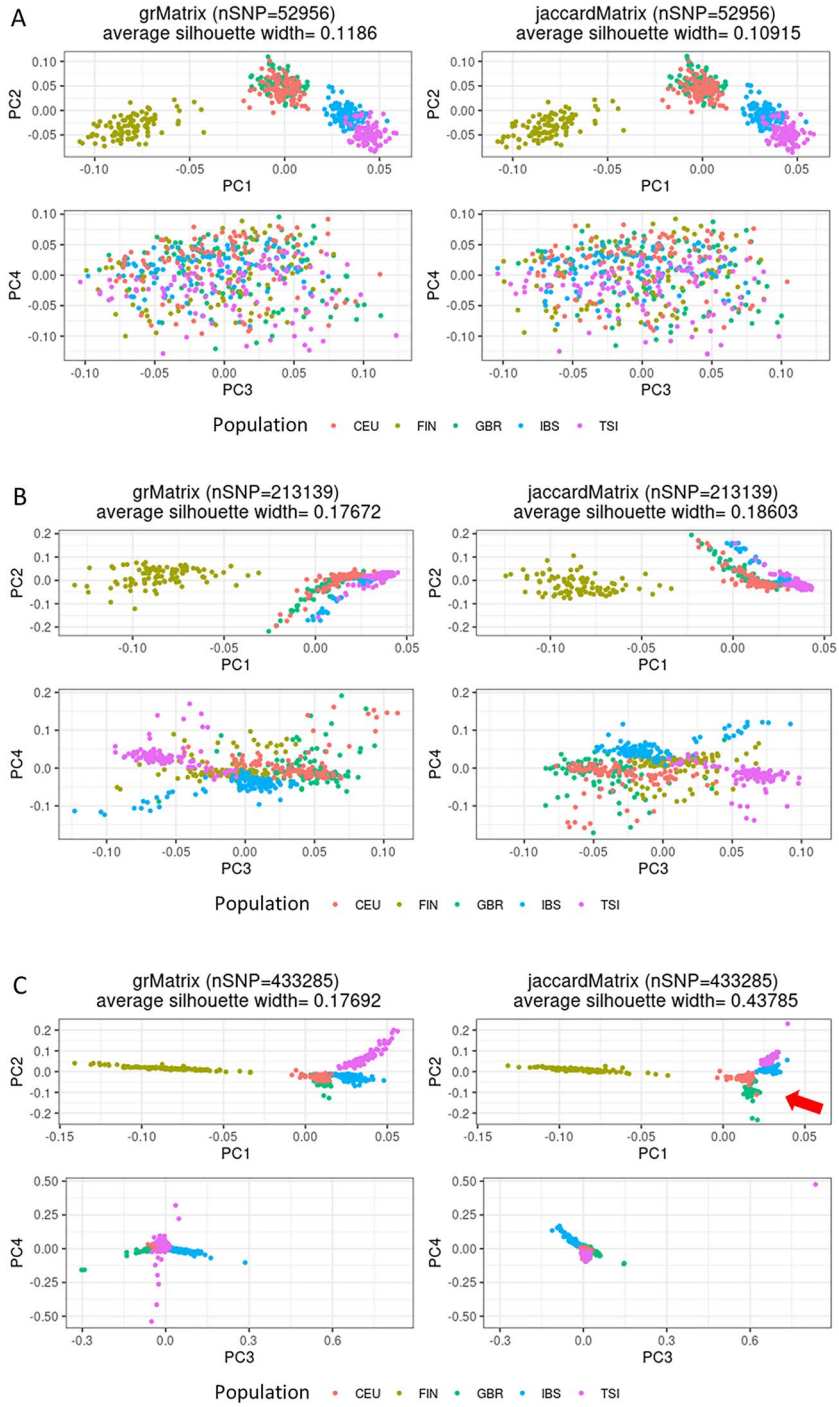

Figure 4. The PCA plots according to the genome-wide data sets in the European populations from 1000 Genome Project data (**A**: common SNVs, **B**: low-frequency SNVs and **C**: rare SNVs; left panel: GRM and right panel: unweighted Jaccard matrix). The SNV quality criteria were 0.0% genotyping missing rate or deviations from Hardy–Weinberg proportions (P-value $< 10^{-6}$). The red arrow (C: rare SNVs) indicates that the GBR (green color) population is better separated from CEU (orange color) population in the unweighted Jaccard matrix.

[30]. Using bootstrapping with 1000 random replicates for each region, we have quantified the number of significant (i.e. outside of the CI) differences between local and global structures around these GWAS loci across all MAF bins. Surprisingly, we saw significant differences in more than half of the loci depending on the data set, population, similarity matrix and MAF cutoff (Supplementary Figure 9). In particular, we saw more significant differences when using the Jaccard matrix for the rare variants except for the HISP population.

The performance among similarity matrices in the European and East Asian populations from the 1000 Genome Project

Finally, using a real population-based data set from the 1000 Genomes Project, we assessed how well we could cluster sub-populations using three different MAF bins (common, low frequency and rare) and two different similarity matrices (GRM and unweighted Jaccard). We selected two continental populations for our analysis: Europeans and East Asians. In the European cohort

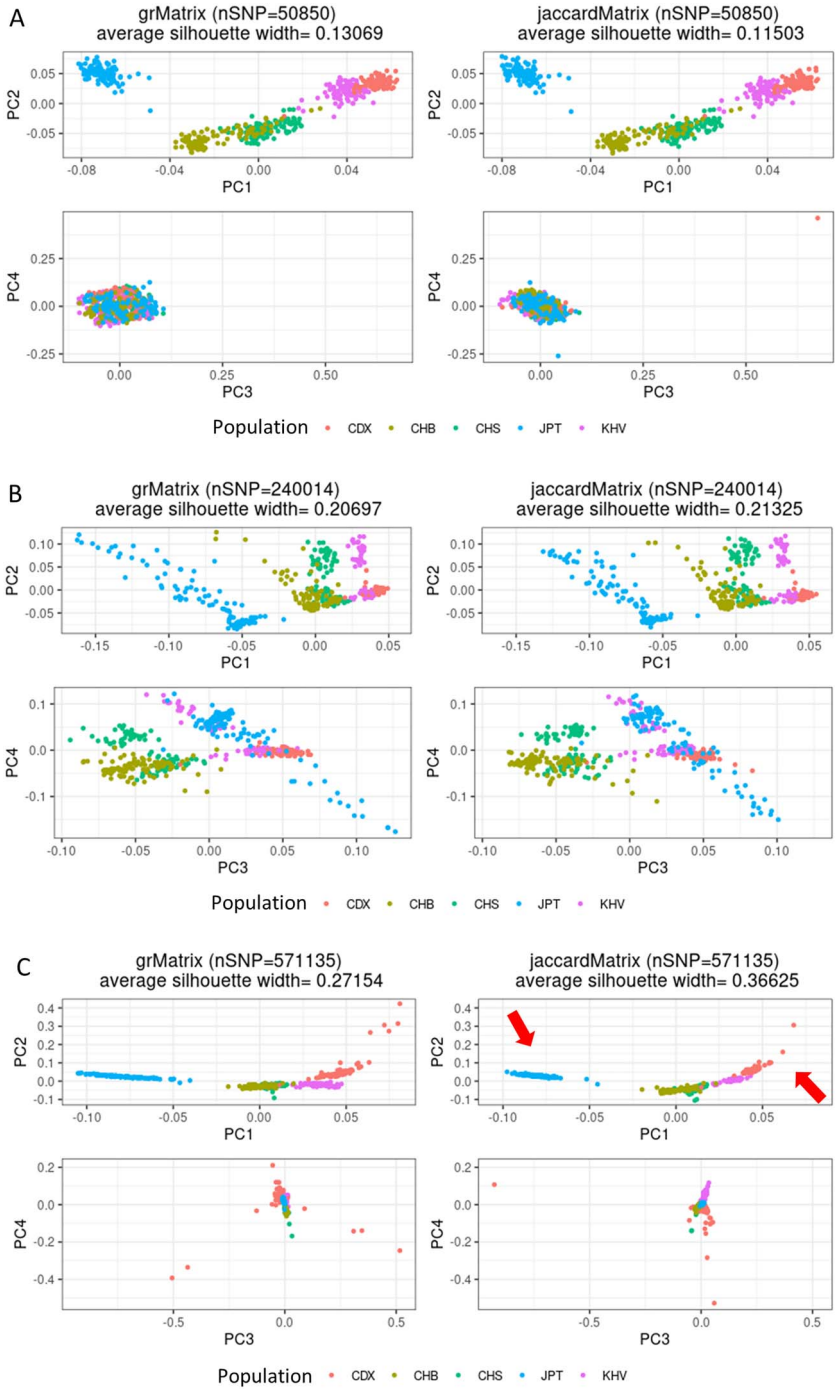

Figure 5. The PCA plots according to the genome-wide data sets in the East Asian populations from 1000 Genome Project data (**A**: common SNVs, **B**: low-frequency SNVs and **C**: rare SNVs; left panel: GRM and right panel: unweighted Jaccard matrix). The SNV quality criteria were 0.0% genotyping missing rate or deviations from Hardy–Weinberg proportions (P-value $< 10^{-6}$). The red arrows (C: rare SNVs) indicate that the CDX (orange color) and the JPT (blue color) populations are more densely clustered in the unweighted Jaccard matrix.

of the 1000 Genomes Project, the remaining SNVs in each data set (common, low frequency and rare SNVs) were 52 956, 213 139 and 433 285, respectively, after SNV quality control and LD-pruning steps. In the East Asian cohort, they were 50 850, 240 014 and 571 135 in each data set (common, low frequency and rare SNVs), respectively.

In cases of common and low-frequency variants, the PCA plots (PC1 versus PC2 and PC3 versus PC4) showed no substantial difference in clustering of European or East Asian subpopulations between GRM and the unweighted Jaccard matrix (Figures 4A and B, 5A and B). In the case of rare variants, however, the unweighted Jaccard matrix outperformed the GRM in both European and East Asian populations (Figures 4C and 5C). In detail, GBR among the European populations is clearly separated when using the unweighted Jaccard matrix with rare variants compared with the GRM with the same variants (Figure 4C, green population). Among the East Asian populations, CDX and JPT were more densely clustered in the unweighted Jaccard matrix

Table 4. The goodness of fit tests based on clustering indexes for population stratification in each similarity matrix in the European and East Asian populations in the 1000 genome project. Better clustering between and within subpopulations provides a larger silhouette width or Fowlkes–Mallows index and a smaller Davies–Bouldin index

Index	Matrix used in European populations	Common SNVs ($n = 52\,956$)	Low-frequency SNVs ($n = 213\,139$)	Rare SNVs ($n = 433\,285$)
Within sum of squares	GRM	7.33387	7.15871	7.41454
	Unweighted Jaccard matrix	7.33805	7.10288	7.12859
Average silhouette width	GRM	0.11860	0.17672	0.17692
	Unweighted Jaccard matrix	0.10915	0.18603	0.43785
Fowlkes–Mallows index	GRM	0.54780	0.76288	0.83010
	Unweighted Jaccard matrix	0.60959	0.75301	0.84687
Davies–Bouldin index	GRM	3.34277	2.41018	1.81481
	Unweighted Jaccard matrix	3.24087	2.56822	1.79457
Index	**Matrix used in East Asian populations**	**Common SNVs ($n = 50\,850$)**	**Low-frequency SNVs ($n = 240\,014$)**	**Rare SNVs ($n = 571\,135$)**
Within sum of squares	GRM	7.12564	7.22644	7.25937
	Unweighted Jaccard matrix	7.26501	7.10025	7.11338
Average Silhouette width	GRM	0.13069	0.20697	0.27154
	Unweighted Jaccard matrix	0.11503	0.21325	0.36625
Fowlkes–Mallows index	GRM	0.65764	0.65277	0.79487
	Unweighted Jaccard matrix	0.70543	0.61634	0.79797
Davies–Bouldin index	GRM	2.53257	2.03141	1.68264
	Unweighted Jaccard matrix	2.36855	1.99933	1.62926

with rare variants (Figure 5C, orange and blue populations, respectively) compared with the other cases. As a measure of the goodness of fit, we used the total within sum of squares criterion as well as clustering indices such as Davies–Bouldin index, Fowlkes–Mallows index and average silhouette width. Better clustering between and within subpopulations is indicated with a smaller Davies–Bouldin index and a larger silhouette width or Fowlkes–Mallows index.

Within sum of squares and the average silhouette width among clustering indices clearly showed the outperformance of the unweighted Jaccard matrix with the rare variants as shown in PCA plots (Figures 4 and 5 and Table 4). The two other measures show a better clustering when using the unweighted Jaccard matrix as well; however, the difference is less pronounced. Therefore, as shown here and in the simulation study, we can reveal finer PS with the unweighted Jaccard matrix based on the rare variants.

Discussion

Despite the vast majority of rare variants being population specific, the utility of rare variants in PS analyses remains controversial. Some reports mention that using rare variants for PCA provides little benefit when handling PS in the admixed genomic analysis compared with common ones [7–9]. Others report that rare variants can be used to detect a finer PS [13, 37]. In this study, we examined the use of rare variants when performing PCA. We show that variants with a low MAC can indeed inform on PS in a PCA analysis at least as well as common variants and provide better clustering/fine-scale resolution in some cases. Our study suggests some new insights for local PS capturing using rare variants and shows an outlier issue when using the weighted Jaccard matrix.

Through simulation studies and an application to real data, we showed that PCA on the unweighted Jaccard matrix is more sensitive than other similarity matrices (GRM) when applied to rare variants (MAF < 0.5%) and shows better visual and clustering performance as measured by several clustering indices. We have not observed significant differences between similarity matrices in other categories of MAF bins (common: >5%, or low frequency: 0.5–5%). Considering that the vast majority of human genomic variation is rare (MAF < 0.5%) and population specific in WGS [16], we suggested using the unweighted Jaccard similarity matrix with the rarest variants when studying population structure instead of GRM on common variants. In addition, we have asked the question of whether rare variants could be used to capture local PS along a chromosome. In simulation studies, we showed that, indeed, the difference in canonical correlation between the outer region (less PS) and inner region (more PS) is more pronounced when using rare variants.

Furthermore, we assessed local PS in the 'APOE' region (a known risk factor for AD) in two AD WGS data sets using only AD cases. In particular, we saw a different local PS pattern in rare variants among the NHW population, which could indicate some allelic heterogeneity in this complex region. Additionally, we proposed a bootstrapping method to assess the significance of deviation of canonical correlation between global PS and local PS for a particular region. Using this technique, we evaluated 38 AD-associated loci from a recent large GWAS paper and reported that we see more significant differences between local and global PS when focusing on low-frequency and rare variants.

This work leaves scope for a variety of avenues for future work. First, outliers in the PCs seem to be commonplace, especially for the weighted Jaccard matrix. This behavior is induced by SNVs with the lowest MAF, causing the inverse weights in the weighted Jaccard matrix to be (disproportionately) large. This made the weighted Jaccard matrix somewhat less suited for PS analyses in our studies. Outliers can be avoided by changing the weighting scheme in the weighted Jaccard matrix. However, this remains for future research. Second, calculating PCs on the (unweighted) Jaccard matrix showed some instabilities when using the function 'eigen' in R (Supplementary Figure 2). Other functions are available, such as the function 'eigs' in the R-package 'RSpectra', which might be more stable numerically. Overall, some care needs to be taken to ensure all computations are numerically stable. Lastly, we demonstrated in CCA that error bars can be computed for local

to global PS. These error bands can potentially help in identifying correlations that are of interest. The precise interpretation of the local to global stratification plots remains for future research. We also recommend exercising extreme caution in the interpretation of PCA results, as those results can be sensitive to sample size, number of used variants and population composition [38]. In the era of WGS, we hope that our work encourages the use of the Jaccard matrix with rare variants to inform PS in GWASs.

Key Points

- Rare variants (<0.5%), which are the vast majority of the human genome, can provide better clustering/fine-scale resolution when analyzing population structure than common variants.
- Similarity matrices such as GRM and Jaccard matrix make no difference in the categories of minor allele frequency (common: >5%, or low frequency: 0.5–5%).
- The unweighted Jaccard matrix provides a better visual and clustering performance when applied to rare variants in population structure analysis.
- Canonical correlation between global and local population structure is different across all MAF spectra, and the significance of this deviation can be evaluated via bootstrapping.

Supplementary Data

Supplementary data are available online at https://academic.oup.com/bib.

Data availability

The NIMH data set analyzed during the current study is available from the authors on reasonable request. The family component and the case-control component of the NIA ADSP WGS data set are available from DSS NIAGADS under accession number: NG00067. The 1000 Genomes Project data are available from https://www.internationalgenome.org/.

Author contributions

C.L. conceptualized and designed the project. S.L., G.H. and D.P. performed statistical analyses and interpretation, and drafted the manuscript. J.H. assisted the analyses and manuscript preparation. All authors contributed to the critical revision of the manuscript. All authors contributed to the relevant sections and approved the final manuscript.

Acknowledgements

The computations in this paper were run in part on the FASRC Cannon cluster supported by the FAS Division of Science Research Computing Group at Harvard University. The funding body has no role in the design of the study and collection, analysis and interpretation of data and in writing the manuscript. Please refer to the Supplementary Note for full acknowledgements.

Funding

The National Human Genome Research Institute (R01HG008976, 2U01HG008685); the National Heart, Lung, and Blood Institute (P01 HL132825, U01HL089856, U01HL089897, P01HL120839); National Institute of Mental Health (R01MH129337); Cure Alzheimer's Fund.

References

1. Campbell CD, Ogburn EL, Lunetta KL, *et al.* Demonstrating stratification in a European American population. *Nat Genet* 2005;**37**:868–72.
2. Tian C, Gregersen PK, Seldin MF. Accounting for ancestry: population substructure and genome-wide association studies. *Hum Mol Genet* 2008;**17**:R143–50.
3. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.
4. Price AL, Patterson NJ, Plenge RM, *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 2006;**38**:904–9.
5. Kang HM, Sul JH, Service SK, *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 2010;**42**:348–54.
6. Listgarten J, Lippert C, Kadie CM, *et al.* Improved linear mixed models for genome-wide association studies. *Nat Methods* 2012;**9**:525–6.
7. Ma S, Shi G. On rare variants in principal component analysis of population stratification. *BMC Genet* 2020;**21**:34.
8. Zhang Y, Guan W, Pan W. Adjustment for population stratification via principal components in association analysis of rare variants. *Genet Epidemiol* 2013;**37**:99–109.
9. Zhang Y, Shen X, Pan W. Adjusting for population stratification in a fine scale with principal components and sequencing data. *Genet Epidemiol* 2013;**37**:787–801.
10. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* 2016;**17**:333–51.
11. Keinan A, Clark AG. Recent explosive human population growth has resulted in an excess of rare genetic variants. *Science* 2012;**336**:740–3.
12. Persyn E, Redon R, Bellanger L, *et al.* The impact of a fine-scale population stratification on rare variant association test results. *PLoS One* 2018;**13**:e0207677.
13. Siu H, Jin L, Xiong M. Manifold learning for human population structure studies. *PLoS One* 2012;**7**:e29901.
14. Mathieson I, McVean G. Demography and the age of rare variants. *PLoS Genet* 2014;**10**:e1004528.
15. Prokopenko D, Hecker J, Silverman EK, *et al.* Utilizing the Jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 Genomes Project. *Bioinformatics* 2016;**32**:1366–72.
16. Taliun D, Harris DN, Kessler MD, *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* 2021;**590**:290–9.
17. Schlauch D, Fier H, Lange C. Identification of genetic outliers due to sub-structure and cryptic relationships. *Bioinformatics* 2017;**33**:1972–9.
18. Hahn G, Lutz SM, Hecker J, *et al.* locStra: fast analysis of regional/global stratification in whole-genome sequencing studies. *Genet Epidemiol* 2021;**45**:82–98.
19. Li H, Ralph P. Local PCA shows how the effect of population structure differs along the genome. *Genetics* 2019;**211**:289–304.
20. Tekola-Ayele F, Ouidir M, Shrestha D, *et al.* Admixture mapping identifies African and Amerindigenous local ancestry loci associated with fetal growth. *Hum Genet* 2021;**140**:985–97.

21. Atkinson EG, Maihofer AX, Kanai M, *et al.* Tractor uses local ancestry to enable the inclusion of admixed individuals in GWAS and to boost power. *Nat Genet* 2021;**53**:195–204.
22. Genomes Project C, Abecasis GR, Auton A, *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 2012;**491**:56–65.
23. Balding DJ, Nichols RA. A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 1995;**96**: 3–12.
24. Nelis M, Esko T, Magi R, *et al.* Genetic structure of Europeans: a view from the north-east. *PLoS One* 2009;**4**:e5472.
25. Yang J, Lee SH, Goddard ME, *et al.* GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 2011;**88**:76–82.
26. Zhou Q, Zhao L, Guan Y. Strong selection at MHC in Mexicans since admixture. *PLoS Genet* 2016;**12**:e1005847.
27. González I, Déjean S, Martin PGP, *et al.* CCA: an R package to extend canonical correlation analysis. *J Stat Softw* 2008;**23**: 1–14.
28. Prokopenko D, Hecker J, Kirchner R, *et al.* Identification of novel Alzheimer's disease loci using sex-specific family-based association analysis of whole-genome sequence data. *Sci Rep* 2020;**10**:5029.
29. Prokopenko D, Morgan SL, Mullin K, *et al.* Whole-genome sequencing reveals new Alzheimer's disease-associated rare variants in loci related to synaptic function and neuronal development. *Alzheimers Dement* 2021;**17**:1509–27.
30. Wightman DP, Jansen IE, Savage JE, *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* 2021;**53**:1276–82.
31. Davies DL, Bouldin DW. A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell* 1979;**PAMI-1**:224–7.
32. Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. *J Am Stat Assoc* 1983;**78**:553–69.
33. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 1987;**20**: 53–65.
34. Rajabli F, Feliciano BE, Celis K, *et al.* Ancestral origin of ApoE epsilon4 Alzheimer disease risk in Puerto Rican and African American populations. *PLoS Genet* 2018;**14**:e1007791.
35. Granot-Hershkovitz E, Tarraf W, Kurniansyah N, *et al.* APOE alleles' association with cognitive function differs across Hispanic/Latino groups and genetic ancestry in the study of Latinos-investigation of neurocognitive aging (HCHS/SOL). *Alzheimers Dement* 2021;**17**:466–74.
36. Blue EE, Horimoto A, Mukherjee S, *et al.* Local ancestry at APOE modifies Alzheimer's disease risk in Caribbean Hispanics. *Alzheimers Dement* 2019;**15**:1524–32.
37. Baye TM, He H, Ding L, *et al.* Population structure analysis using rare and common functional variants. *BMC Proc* 2011; **5**:S8.
38. Elhaik E. Principal component analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated. *Sci Rep* 2022;**12**:14683.