On the detection of genetic heterogeneity in whole-genome sequencing studies: A statistical test for the identification of “genetic outliers” due to population sub-structure or cryptic relationships

Daniel Schlauch[1] and Christoph Lange[1]

[1]Department of Biostatistics and Computational Biology,
Dana-Farber Cancer Institute and Department of Biostatistics,
Harvard TH Chan School of Public Health, Boston, MA 02115

June 6, 2016

Abstract

In order to minimize the effects of genetic confounding on the analysis of high-throughput genetic association studies, e.g. (whole-genome) sequencing studies, genome-wide association studies (GWAS), etc., we propose a general framework to assess and to test formally for genetic heterogeneity among study subjects. Even for relatively moderate sample sizes, the proposed testing framework is able to identify study subjects that are genetically too similar, e.g cryptic relationships, or that are genetically too different, e.g. population substructure. The approach is computationally fast, enabling the application to whole-genome sequencing data, and straightforward to implement. In an application to the 1,000 genome projects, our approach identifies studies subjects that are most likely related, but have passed so far standard qc-filters. Simulation studies illustrate the overall performance of our approach.

Introduction

The fundamental assumption in standard genetic association analysis is that the study subjects are independent and that, at each locus, the allele frequency is identical across study subjects. In the presence of population heterogeneity, e.g. population substructure or cryptic relatedness, these assumptions are violated. It can introduce confounding into the analysis and lead to biased results, e.g. false positive findings [6, 10, 12, 14]. Given the generality of the problem, it has been the focus of methodology research for a long time. For candidate gene studies and later genome-wide association studies (GWAS), genomic control was developed [1,4]. The approach adjusts the association test statistics at the loci of interest by an inflation factor that is estimated at a set of known null-loci. With the arrival of GWAS data, it became possible to estimate the genetic dependence between study subjects and the overall genetic variation for each study subject by computing the empirical genetic variance/covariance matrix between study subjects at a whole genome level. The genetic variance/covariance matrix can then be utilized in two ways to minimize the effects of population substructure on the association analysis.

The first method is to compute an eigenvalue decomposition of the matrix and to include the eigenvectors that explain the most variation as covariates in the association analysis [10, 11]. An alternative approach is to incorporate the estimated dependence structure of the study subjects directly into a generalized linear model and account so directly for the dependence at the model-level [7, 8, 15]. Both approaches have proven to work well in numerous applications. While the first approach is computationally fast and easy to implement, the direct modelling of the dependence structure between study subjects can be more efficient.

However, both approaches benefit if, prior to the analysis, study subjects whose genetic profile is very different from the other study subjects, e.g. "genetic outliers", are removed from the data set. The standard practice is currently to examine the Eigenvalue plots visually and to identify outliers by personal judgement on how far study subjects are from the "clouds" of study subjects. As typically up to 10 Eigenvectors have to be considered, this process of identifying outliers can become a complicated and subjective procedure.

In this communication, we propose a formal statistical test that assesses whether two study subjects come from the same population and whether they are unrelated. The test statistic is based on the an adaptation of the Jaccard Index which utilizes the idea that variants are differentially informative of relatedness based on their allele frequency. Furthermore, its distribution can be derived under the null-hypothesis which makes it computationally fast, enabling the application to whole-genome sequencing data. Our measure has clearly defined properties which can be used to test for homogeneity in a population and in particular identify individuals who are likely be related in a study population.

Methods

Exploiting the information in rare variants (RVs), such as one with minor allele fequency (MAF) $< 1\%$, is fundamental to our method. A RV tends to be less stable than common variants and thus is more likely to have arisen recently. Allele frequencies have been shown to differentially confound association studies in part due to the localized nature of RVs [9]. Motivated by this rationale, we developed a method that utilizes the differential informativeness of variants by allele frequency to increase the power to provide a high resolution picture of population structure. Our approach uses an intuitive, computationally straightforward approach towards identifying similarity between two study subjects.

Similarity measure among haploid genomes

Consider a matrix of n individuals ($2n$ haploid genomes), with N independent variants described by the genotype matrix $\mathbf{G}_{2n \times N}$. $\mathbf{G}$ is a binary matrix with value 1 indicating the presence of the minor allele and 0 indicating the major allele. We define the similarity index between two haploid genomes, $s_{i,j}$

$$s_{i,j} = \frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k} \mathbf{G}_{j,k}}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} \tag{1}$$

where

$$w_k = \begin{cases} \frac{\binom{2n}{2}}{\binom{\sum_{l=1}^{2n} \mathbf{G}_{l,k}}{2}} & \sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1 \\ 0 & \sum_{l=1}^{2n} \mathbf{G}_{l,k} \leq 1 \end{cases}$$

In the absence of population structure, i.e. homogeneous population we have

$$E\left(s_{i,j}\right) = 1$$

It therefore follows from the Central Limit Theorem that in the absence of populations structure, cryptic relatedness and dependence between loci (such as linkage disequilibrium) the distribution of the similarity index, $s_{i,j}$ is Gaussian.

$$s_{i,j} \sim N\left(1, \sigma_{i,j}^2\right)$$

Where the variance of s_{ij} can be estimated by

$$\hat{\sigma_{i,j}^2} = \hat{Var}\left(s_{i,j}\right) = \frac{\sum_{k=1}^{N} \left(w_k - 1\right)}{\left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2} \tag{2}$$

The similarity index $s_{i,j}$ provides an easily interpreted statistical test for evaluating possible relatedness between individuals in a purportedly homogeous dataset of unrelated individuals. Note that this formulation is independent of the samples i, j and depends only on the allele counts for each variant across the study group.

Similarity measure among diploid genomes

This approach is easily generalized to the diploid scenario. A diploid similarity score, $s_{diploid}$, is obtained by averaging each of the four pairwise haploid $s_{haploid}$ scores between each person's two haploid genotypes. For n individuals, $2n$ genotypes per loci, the similarity between individuals i and j is defined as

$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^{N} \left[w_k \mathbf{G}_{i_1,k} \mathbf{G}_{j_1,k} + w_k \mathbf{G}_{i_1,k} \mathbf{G}_{j_2,k} + w_k \mathbf{G}_{i_2,k} \mathbf{G}_{j_1,k} + w_k \mathbf{G}_{i_2,k} \mathbf{G}_{j_2,k} \right] /4}{\sum_{k=1}^{N} I \left[\left(\sum_{l=1}^{n} \left[\mathbf{G}_{l_1,k} + \mathbf{G}_{l_2,k} \right] \right) > 1 \right]}$$

where $\mathbf{G}_{i_2,k}$ refers to the 2^{nd} genotype of individual i at locus k.

Here it becomes clear that the method can be applied to phased and unphased data alike. For an unphased data matrix $\mathbf{H}_{n \times N}$, where $\mathbf{H}$ contains the number of minor alleles, $\{0, 1, 2\}$, for a subject at a particular variant.

$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^{N} \left[w_k \mathbf{H}_{i,k} \mathbf{H}_{j,k} \right] /4}{\sum_{k=1}^{N} I \left[\left(\sum_{l=1}^{n} \mathbf{H}_{l,k} \right) > 1 \right]}$$

This formulation will have the same mean

$$E \left[s_{i,j}^{(diploid)} \right] = 1$$

and assuming independence of each individual's haploid genomes, such as in the absence of inbreeding,

$$\hat{Var} \left(s_{i,j}^{(diploid)} \right) = \frac{\hat{Var} \left(s_{i,j}^{(haploid)} \right)}{4} = \hat{\sigma^2}_{i,j}$$

Which yields the asymptotic result

$$s_{i,j} \sim N \left(\mu_{i,j}, \hat{\sigma^2}_{i,j} \right)$$

We can test the null hypothesis that population structure does not exist and all subjects are unrelated, with respect to the alternative that at least one pair of individuals is related.

$$H_0 : \mu_{i,j} = 1 \forall i, j \in 1 \dots n$$

$$H_A : \exists i, j \in 1 \dots n | \mu_{i,j} \neq 1$$

In a homogeneous dataset lacking relatedness, we consider each of the $\binom{n}{2}$ comparisons to be independent. To achieve a familywise error rate α, we use the Šidák procedure [13] or the approximately equivalent Bonferroni procedure. We reject the null at the α level when we obtain similarity scores in the rejection region

$$R : max \left(s_{i,j} \right) > 1 - probit \left(\frac{\alpha}{\binom{n}{2}} \right)$$

Estimating cryptic relatedness

Furthermore, the measure is particularly powerful for measuring relatedness. Intuitively, we can imagine two subjects which have a kinship coefficient, ϕ, indicating a probability of a randomly chosen allele in each person being identical by descent (IBD). For an allele which belongs to the one person, the probability of it belonging to a related person with kinship coefficient ϕ is $\phi + (1-\phi) \times p$, where p is the allele frequency in the population. We can clearly see that for rare alleles, such that p is small compared to ϕ, there will be a much larger relative difference in the probability of shared alleles among related individuals ($\phi > 0$) compared to unrelated individuals ($\phi = 0$). Given that our method weights more highly these rarer alleles, there is increased sensitivity to detection of relatedness.

Consider a coefficient of kinship between two individuals i, j, $\phi_{i,j} > 0$ with no other population structure present in the data. For an individual variant, k, with sufficient allele frequency, the expected contribution to the statistic for an allele from each individual, s_{i_1,j_1} is

$$E\left(s_{i_1,j_1,k}|\phi_{i,j}\right) = (1-\phi_{i,j}) + \phi_{i,j}\left[p_k \frac{\binom{2n}{2}}{\binom{p_k(2n-2)+2}{2}}\right]$$

and the expectation for the similarity score between those haploid genomes is

$$E\left(s_{i_1,j_1}|\phi_{i,j}\right) = \frac{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\left[(1-\phi_{i,j}) + \phi_{i,j}\left[p_k \frac{2n(2n-1)}{(p_k(2n-2)+2)(p_k(2n-2)+1)}\right]\right]}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} \tag{3}$$

It can be seen that in the presence of cryptic relatedness, $\phi_{i,j} > 0$,

$$E\left(s_{i_1,j_1}|\phi_{i,j} > 0\right) > 1$$

With $\sum_{i=1}^{2n} \mathbf{G}_{i,k}$ as the maximum likelihood estimator for $p_k n$, by the invariance principle, w_k is a consistent estimator for $\frac{\binom{2n}{2}}{\binom{p_k(2n-2)+2}{2}}$.

This yields a maximum likelihood estimate of this kinship defined as

$$\hat{\phi}_{i,j} = \frac{s_{i,j} - 1}{\left[\frac{\sum_{k=1}^{N} \hat{p_k} w_k}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} - 1\right]} \tag{4}$$

with

$$\hat{Var}\left(\hat{\phi}_{i,j}\right) = \frac{\sigma_{i,j}^{\hat{2}}}{\left[\frac{\sum_{k=1}^{N} \hat{p_k} w_k}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} - 1\right]^2}$$

In some instances we may prefer to assume a discrete distribution of the relatedness parameter, such as the identification of first cousins $\left(\phi = \frac{1}{16}\right)$ or closer in th absence of inbreeding. In this scenario, we may consider $\phi \in \left\{0, \frac{1}{16}, \frac{1}{8}, \frac{1}{4}, 1\right\}$ and obtain the maximum likelihood estimate among this set. We use the likelihood function

$$\mathcal{L}\left(s_{i,j}|\phi\right) =$$

For example, in an otherwise homogeneous study group of unrelated individuals a pair of cousins ($\phi = .0625$), with $MAF \sim Uniform\left(.02, .1\right)$ we can directly calculate the expectation of their similarity statistic, $s_{i,j}$

$$E\left(s_{i,j} | \phi = .0625, \text{No other structure}\right) \approx 2.19$$

Statistical power to detect cryptic relatedness

The properties of this similarity measure lend themselves toward straightforward power calculations. It is often of interest to consider some coefficient of relatedness, γ that is acceptable for a study. Setting a $\phi \geq \gamma$ allows for the calculation of the probability of obtaining a pair of samples inside the rejection region given two unacceptably closely related individuals.

$$P\left(Reject\, H_0 | \phi_{i,j} = \gamma\right) = \alpha + (1 - \alpha)\left(1 - \Phi\left(\frac{\mu_{i,j} - 1}{\sqrt{\hat{\sigma^2}_{i,j}}}\right)\right) \tag{5}$$

Where $\Phi\left(x\right)$ is the cumulative distribution function for a standard normal random variable.

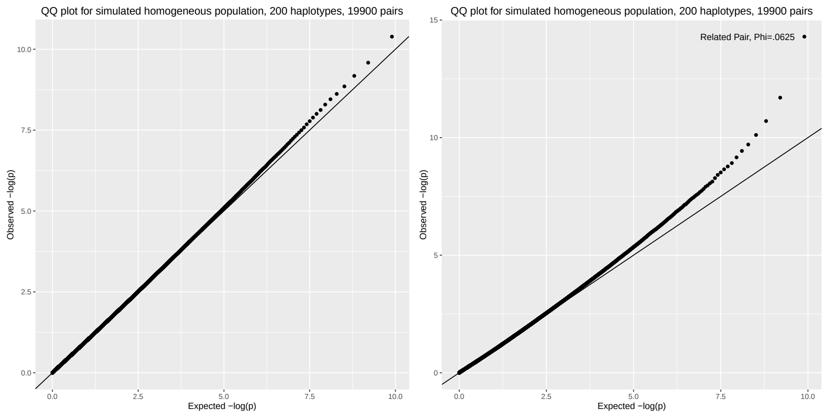
It would be of interest in any study seeking to quantitatively demonstrate the homogeneity of participants to produce this statistic which can demonstrate that a lack of homogeneity would have been found with low probability given the presence of some specified degree of relatedness, γ.

Simulations demonstrate power to detect heterogeneity and speed of method

We ran our method on simulated genotypes derived from a homogeneous dataset containing varying degrees of relatedness. A homogenized version of a real dataset was generated by randomly resampling each variant across all samples. This eliminates correlations between individuals and variants, preserving only the allele frequency distribution. To test the power of our method to identify relatedness we generated an additional sample, S_{N+1} which was related to an arbitrarily chosen individual, S_N, in the homogenized dataset. The genotype for S_{N+1} was generated by assigning one of their values for each allele to be the same as one of the alleles of S_N with probability 4ϕ and assigning the other to be a randomly chosen allele across all samples. With probability $1 - 4\phi$, both haplotypes for S_{N+1} were selected randomly from the homogenized data.

For variant i, allele j, the genotype at $S_{N+1,i,j}$ is given as

$$S_{N+1,i,j} = \begin{cases} S_{N,i,1} & \text{with probability } \phi + \frac{1-2\phi}{2N} \\ S_{N,i,2} & \text{with probability } \phi + \frac{1-2\phi}{2N} \\ S_{1,i,1} & \text{with probability } \frac{1-2\phi}{2N} \\ \vdots & \vdots \\ S_{N-1,i,2} & \text{with probability } \frac{1-2\phi}{2N} \end{cases}$$



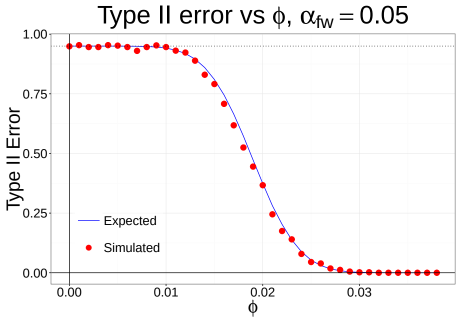


Figure 1: Simulated QQ plots for homogeneous populations in the absence of cryptic relatedness (left) and in the presence of a single $\phi = .0625$ pairing (right)

For each coefficient of kinship we simulated 1,000 studies containing 301 individuals across 100,000 variants in the above manner to evaluate the power of our method. Each simulated study contained only a single related pair with relatedness, ϕ, among an otherwise homogeneous dataset. We demonstrate that under the null hypothesis, $H_0 : \phi = 0$, the family-wise type I error rate, $\alpha = .05$ is preserved. We then compared the proportion of simulated studies which were found to have significantly related pairs to the analytically derived probability of type II error 5.

Of additional interest is the computation time of our method in comparison to other similarity metrics. Commonly, in PCA, a decomposition of the correlation matrix is used. We compared our method in terms of computation time of generating a correlation matrix. We simulated a study of 100,000 variants across n individuals and ran an R implementation of our method against the default implementation of correlation in R, **cor()**.

n	Our Method	Correlation
500	7.514s	13.267s
1000	20.011s	51.460s
2000	73.566s	202.640s

Using a computer with Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz, and R 3.2.3, we found a that our method ran faster than **cor** in R.

Identification of relatedness and structure in 1000GP data

We applied our method to data from the 1000 Genomes Project (TGP) [2, 3], an international consortium which has sequenced individuals from 26 distinct populations sampled from around the globe.

These populations were not identified by the TGP to have cryptic relatedness or had known cryptic relatedness removed [**citation difficult (pptx file posted online KGP website) ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/cryptic_relation_analysis/Nemesh_crypticrelatedness_20120213.pptx**]. However, subsequent analyses have discovered numerous inferred relationships closer than first cousins [5].

Phase 3 of the 1000 Genomes Project contains approximately 2504 individuals with a combined total of approximately 88 million variants. To test our method, we divided the data into blocks of 400 consecutive variants and selected only one loci from each block in order to limit the impact of linkage disequilibrium and promote the independence of measurements. The selected variant within each block was chosen based on the smallest minor allele frequency observed which was larger than our cutoff of 2%. This yielded approximately 200,000 variants for each of the 26 populations in the TGP. Our method was then run on each of these populations separately to test for heterogeneity and relatedness within population groups.

We first computed the genetic similarity matrix across all 5 super populations and compared it to the variance-covariance matrix commonly used in PCA 9. Additionally, we performed an eigenvalue decomposition of the GSMs and plotted the samples on the first two components, showing highly similar results for our method compared with standard principal components analysis. We found that despite the focus on less frequent alleles that is inherent to our approach, we were able to easily partition the data by super population, suggesting that our metric may be effectively used in computing a genetic similarity matrix even for distantly related populations which do not require the prioritization of less frequent alleles.

We discovered that there was great variation in the presence of cryptic relatedness and population structure across the 26 populations of the study. Under the assumptions that each study contained a homogeneous population of unrelated individuals, only a handful of groups contained neither large outliers nor heavily inflated numbers of significant results.1

We defined the presence of population structure as applying to those populations which had a greater than expected number of similarity scores below the mean. By definition, the mean of each study is $\mu = 1$. We performed a standard one-sided binomial test with $p = 0.5$ using α cutoffs of .05 and .01 1. Using this criterion, eight of the 26 populations met this threshold. Using the stricter cutoff- CLM (Colombians from Medellin, Colombia) $\left(p = 4 \times 10^{-6}\right)$, PUR (Puerto Ricans from Puerto Rico) $\left(p = 1 \times 10^{-11}\right)$, and PEL (Peruvians from Lima, Peru) $\left(p = 7 \times 10^{-34}\right)$. Each of these populations are part of the Ad Mixed American super population and represent "new world" groups which have undergone extensive admixture in recent centuries. It is therefore reassuring that these groups of individuals would exhibit the greatest amount of structure among the populations surveyed.

[**Note: another way of defining structure is via the ratio of sample variance to calculated variance which yields an arguably better list of populations. But may be difficult to avoid having to make some arbitrary cutoff (or calculate the variance of that ratio)**]

We defined the presence of cryptic relatedness as those individual pairs which exceed the cutoff for a family-wise error rate of $\alpha = .01$ and were estimated to have a coefficient of relatedness $\phi > .0625$, which corresponds to first cousins. By this measure, cryptic relatedness was discovered in all but six of the 26 populations using this method. Eleven pairs of first order (parent-offspring or full sibling) relationships were detected among individuals within the same population group, $\left(.2 < \hat{\phi_{i,j}} < .3\right)$, a set of pairings which corresponds identically with the conclusions of Gazal et al.

Inference on our kinship estimate is made under the assumption of homogeneity of the background study population. Identified significant relatedness may be due to the fact that the variance is inflated in the presence of population structure. So it is incomplete to identify cryptic relatedness in this manner in popultations which contain identified structure. However, in populations which do not exhibit detectable structure, we still find many instances of related individuals in this study. For example, two individuals from the ACB population (African Caribbeans in Barbados) produced a $s_{i,j}$ score of 2.6 $\left(p < 10^{-30}\right)$, whereas no other pairing exceeded the family-wise cutoff of 1.3. Using the formula above, we estimate this relationship to be $\hat{\phi} = .27$, suggesting that those individuals are first degree relatives. Two pairs of individuals in the STU population- (HG03899/HG03733 and HG03754/HG03750) were both estimated to have a kinship coefficient $\hat{\phi} \approx .25$, similarly indicating a relatedness of the first degree.

Interestingly, not all related pairs belonged to the same population groups. We additionally discovered a pair of individuals HG03998 from the STU population and HG03873 from the ITU population which exhibited strikingly high relatedness. The plot below4 was generated by placing HG03998 into the ITU population and running our analysis on that population. Given an individual who belongs to a separate population from all others in a dataset would be ex-

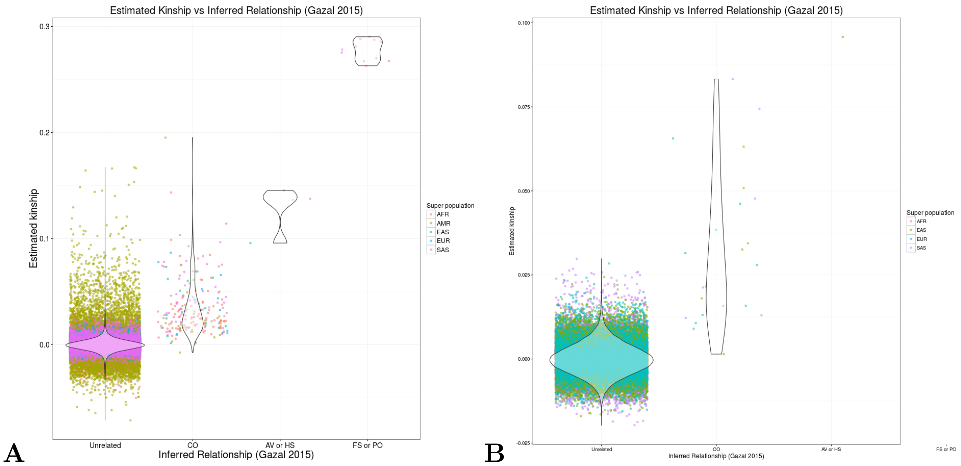


Figure 2: Violin plots for observed similarity scores separated by inferred relationship by Gazal et al. We see inflated estimation of cryptic relatedness using all populations, of which many include population structure (**A**). Examining only populations which have low or undetected structure, we observe a strong separation of individuals based on relatedness (**B**).

pected to produce similarity scores less than 1. However, the similarity between HG03998 and HG03873 was found to be $s = 3.9$ significant at $p < 10^{-30}$ with an estimated relatedness $\hat{\phi} > .25$, suggesting that these individuals are first order relatives despite belonging to different population groups. Both populations were sampled from locations in the United Kingdom, increasing the possibility that one these individuals was mislabeled in the data.

Gazal et al propose a subset of the TGP which removes individuals from 227 related pairs such that no two individuals are as related as cousins or closer. This results in a reduced set of 2261 individuals which are assumed to be no more closely related than cousins ($\phi = .0625$). We applied this filter and analyzed each of the 26 populations again to test for heterogeneity and cryptic relatedness.

Nine populations which had been identified as violating homogeneity ($\alpha = .01$) in the full TGP dataset were no longer identified as violating homogeneity after removal of suspected related pairs. However, seven populations, including each of the ad-mixed American groups, continued to violate homogeneity even after the attempts to limit the impact of cryptic relatedness.3

Population structure detection in 1000 Genomes Project

There are many methods for detecting population structure. Most commonly, Principal Components Analysis [10,11]is applied for identifying the components

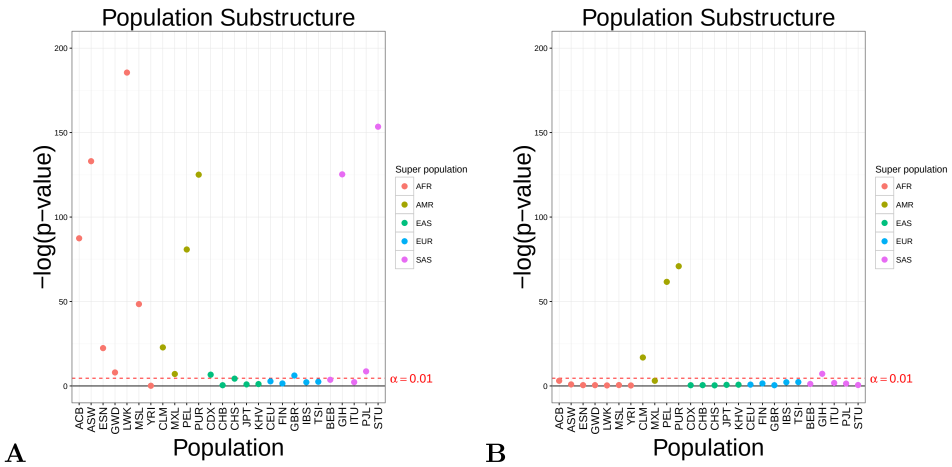


Figure 3: Detection of population structure before (**A**) and after (**B**) removal of related individuals.

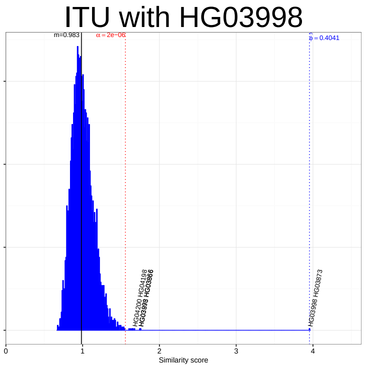


Figure 4: Distribution of s statistic for population Indian Telugu from the UK (ITU) with individual HG03998 added, who is believed to be related to HG03873, despite being labeled in the Sri Lankan Tamil from the UK (STU) population.

Distribution of similarity statistic within population subgroups from 1000 Genomes Project

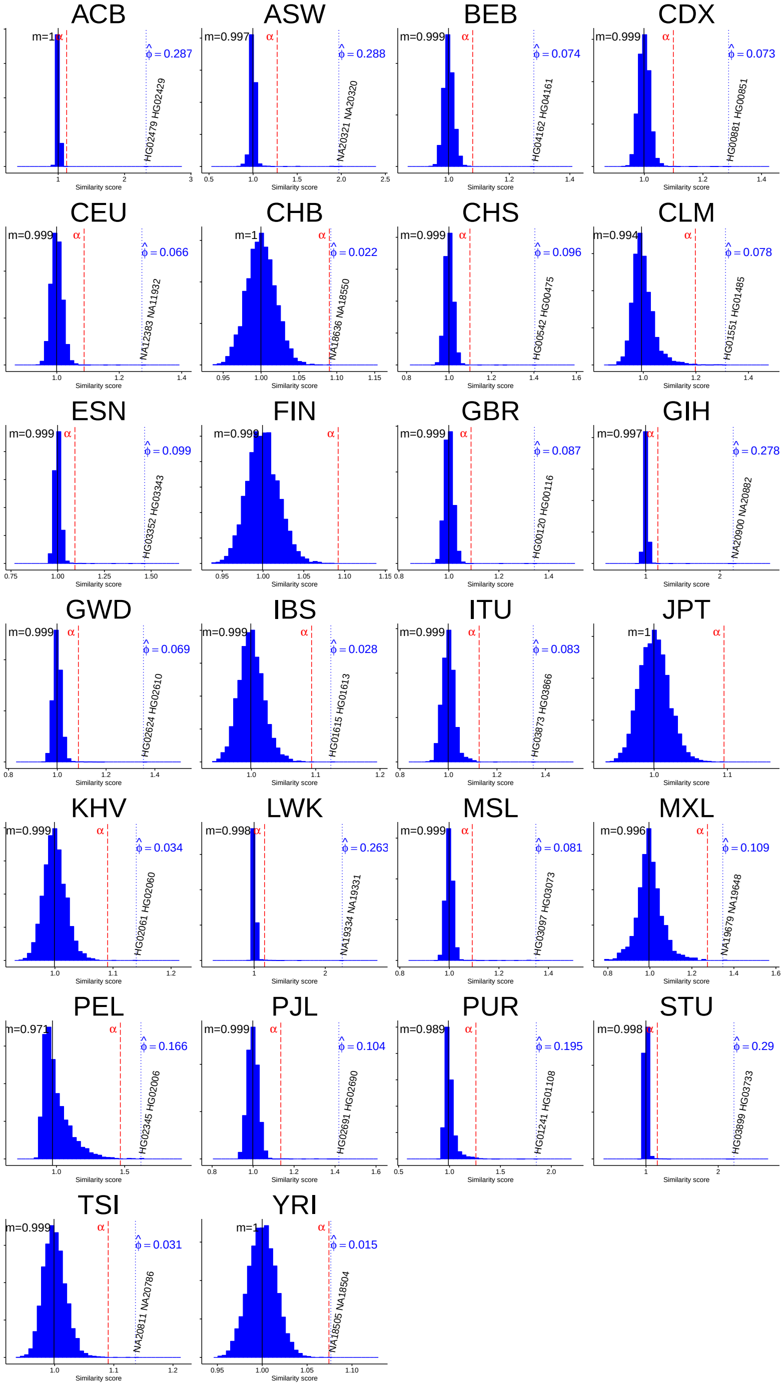


Figure 5: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise$\alpha = .05$ level cutoff for $\binom{n}{2}$ comparisons. Many of the population groups do demonstrate the null behavior (e.g. JPT, KHV, FIN)- however, a number of populations show the presence of extreme outliers (e.g. STU, PUR) or systematic right skew (e.g. MXL, PEL)

Distribution of similarity statistic within population subgroups from 1000 Genomes Project after removal of related individuals

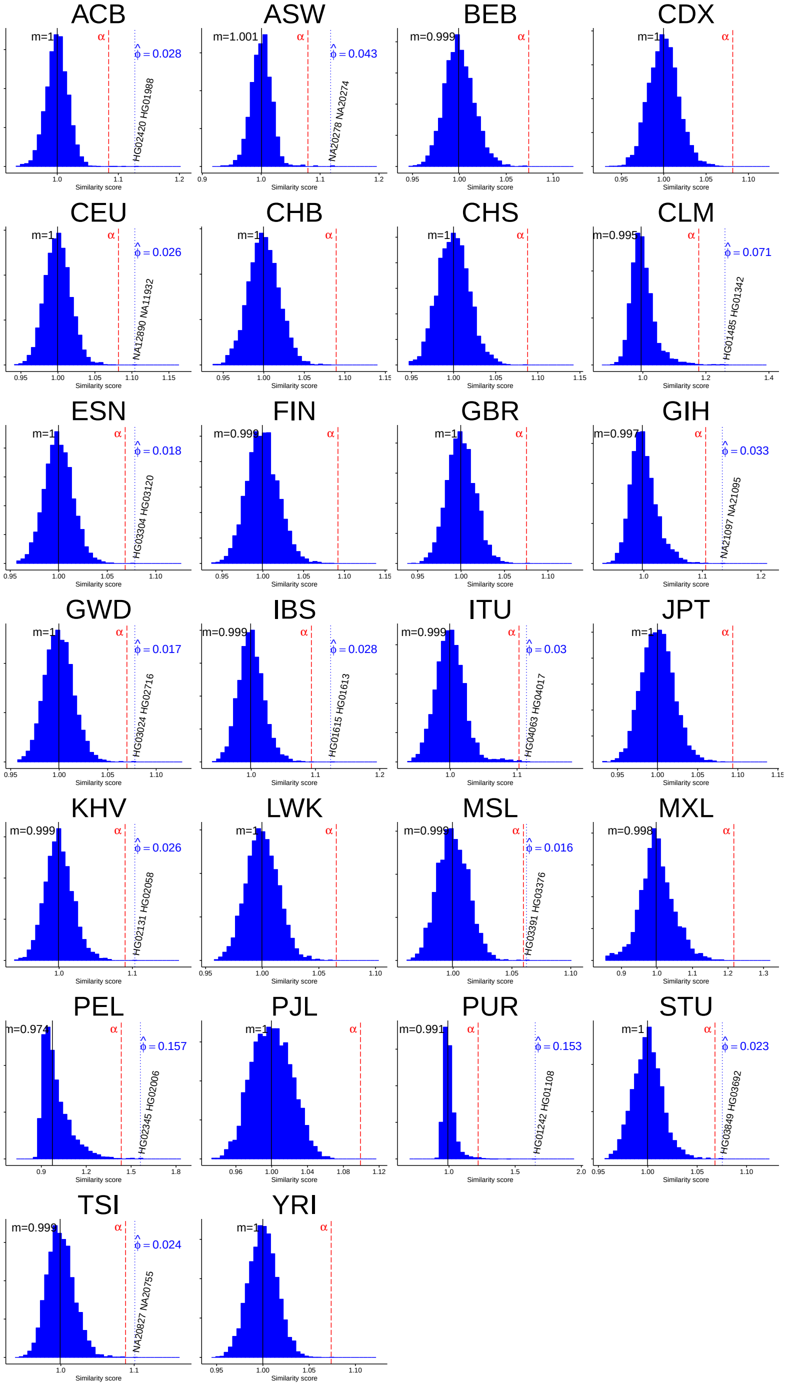


Figure 6: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise$\alpha = .01$ level cutoff for $\binom{n}{2}$ comparisons. Many of the population groups do demonstrate the null behavior (e.g. JPT, KHV, FIN)- however, a number of populations show the presence of extreme outliers (e.g. STU, PUR) or systematic right skew (e.g. MXL, PEL)

Population	Super Population	Structure	Cryptic Relatedness
CDX	EAS - East Asian	No	No
CHB		No	No
CHS		No	No
JPT		No	No
KHV		No	No
ACB	AFR - African	No	**Yes**[+]
ASW		No	**Yes**[+]
ESN		No	No
GWD		**Yes**	No
LWK		**Yes**	**Yes**
MSL		No	No
YRI		No	No
BEB	SAS - South Asian	No	No
GIH		**Yes**	**Yes**[+]
ITU		**Yes**	No
PJL		No	No
STU		**Yes**	**Yes**[+]
CEU	EUR - European	No	No
FIN		No	No
GBR		No	No
IBS		No	No
TSI		No	No
CLM	AMR - Ad Mixed American	**Yes**[*]	No
MXL		No	**Yes**
PEL		**Yes**[*]	No
PUR		**Yes**[*]	**Yes**

Table 1: **Presence of population structure and cryptic relatedness detected in each of the 26 populations in the 1000 Genomes Project.** Population structure was defined as a significant ($\alpha = .01$) number of pairs below the mean. Cryptic relatedness was defined as those populations containing at least one pair of individuals with an estimated kinship coefficient greater than .1 and statistically significant ($\alpha = .01$) after multiple testing correction. [**Note: this table needs to be re-updated with latest runs and definitions**]
[*]- $p < .001$
[+]- $p < .001$

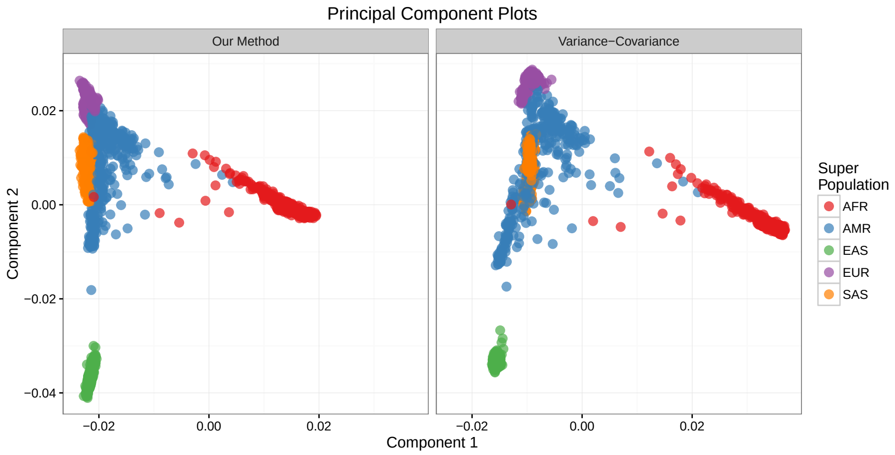


Figure 7: **Principal Components Analysis of all samples in 1000 Genomes Project** using GSM computed using our method (left) and variance-covariance (right) yield similar projections which elucidate the migratory patterns of early humans.

of largest variation which ideally corresponds to the population structure. This procedure first involves the calculation of a genetic similarity matrix (GSM) via the correlation between all samples, which is commonly followed by an eigendecomposition of that matrix. There are a number of limitations to this straightforward approach, one of which is that the calculation of a variance-covariance matrix equally weights the impact of all loci [**unless standardized by rows?**], failing to fully utilize the fact that the overall allele frequency is informative of the value of each variant.

We can similarly use the GSM defined as the pairwise similarity between all samples to feed the eigendecomposition and achieve highly similar results 7 depicting the two dimensional linear migrations of ancient human history.

Despite a focus on separating recently related populations, our method is effective at partitioning samples of more distant common ancestry as well. However, the approach we describe here outperforms PCA when the task involves separating individuals of recent ancestry.

We choose a set of closely related populations from the 1000 Genomes Project in order to demonstrate this performance. The population pairs STU and ITU, and IBS and TSI we among the most closely related populations and represent two pairs of groups which are geographically near and have common ancestry dating back fewer than X years [citation].

We used the 1000 Genomes Project data to compare the GSMs obtained via the conventional variance-covariance matrix step and the use of our method. We evaluated the ability of each method to separate the same quantity of data into the 5 superpopulations and 26 populations. Using approximately 80,000 variants, we generated the two GSMs and plotted the similarity matrices, ordered by hierarchical clustering with average linkage (Figure 9).

Both methods performed well at separating the five superpopulations, but the our method outperformed variance covariance in separating populations of the same superpopulation. As expected, the lack of focus on less frequent alleles, which are more important for distinguishing recent ancestry allowed variance-covariance to adequately separate continental origins, but failed to sufficiently partition the samples according to subgroups.

The first two eigenvectors of the GSMs generated using our method vs variance-covariance yield very similar results. Both methods provide a sufficient separation of coarse-scale population structure. But closer examination of fine-scale population structure reveals our method to be an improvement over variance-covariance. We were able to provide a stronger separation of all 26 populations, particularly those of recent ancestry. As an example, we explored two populations- Sri Lankan Tamil and Indian Telugu, which have relatively small geographical separation.

The origins of two populations are geographically near and can trace their cultural divergence back several thousand years [Citation]. We found that the initially, the greatest axes of variation separated only a handful of the individuals from the rest of the group, strongly suggesting cryptic relatedness and further highlighting the need for identification of related pairs, especially in recently diverged populations. Figure 9A compares our method with that of a correlation based Principal Components Analysis. Unsurprisingly, it is clear that our method assigns greater weight for separating the suspected related pairs from the rest of the pack. Still, the separation of the two populations is clearer in our method compared to PCA.

Next, we removed individuals found to be related, including HG03998 from the STU population and HG03873 from the ITU population. We found that removal of these genetic outliers improved the ability both our method and PCA to separate the two populations. However, the separation remains clearer with the use of our method (Figure 9B).

The reasoning behind the superior performance in fine scale population stratification is due to the focus on rarer alleles. Rare alleles tend to be less stable over generations and are therefore more likely to have arisen recently. It stands to reason that these alleles would therefore be the most informative of recently related populations.

Discussion

The ability to identify genetic outliers has well-established utility in genome-wide association studies. Many existing methods for identification of genetic associations are predicated on the assumptions that population homogeneity holds in the study. Checking for violations of these assumptions typically involves a qualitative assessment without any specific concern for effect size and power. Our method provides an approach for quantitatively assess homogeneity

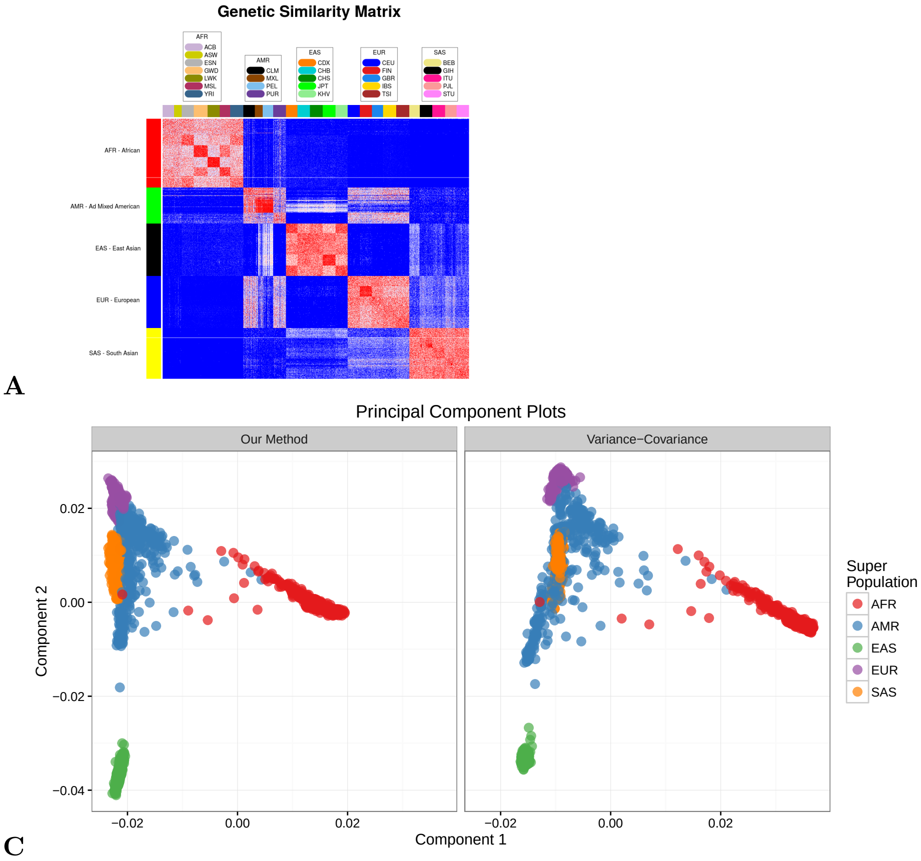


Figure 8: Heatmap of GSM generated by our method (A) and variance-covariance (B) using 80,000 uniformly spaced variants. Samples have been ordered by hierarchical clustering (dendrogram not shown). The vertical colorbar indicates membership in one of the five superpopulations, while the horizontal colorbar indicates membership in one of the 26 populations. (C) Projecting each individual onto the top two eigenvectors resulted in a similar 2-dimensional distribution of global ancestry

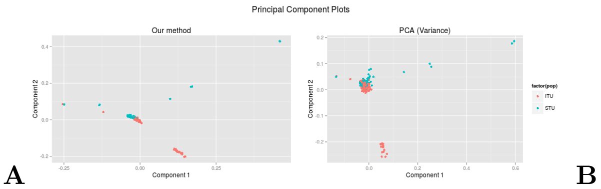


Figure 9: **Example: ITU vs STU**. Two populations of Southern Asian origin, Indian Telugu from the UK (ITU) and Sri Lankan Tamil from the UK (STU) show poor separation using the variance-covariance approach. When using our method, we see improved separation of populations when individuals are projected onto the top two eigenvectors (**A**) despite the fact that our method appears to have expended a greater proportion of the variance explanation on identification of related individuals. After removal of individuals determined to be related (**B**), we see further improvement in the subpopulation separation.

and a formal test for the identification of cryptic relatedness and population stratification.

Several limitations exist with our approach. First, the method assumes that the variants are independent. We satisfy this assumption by performing LD sampling, but in doing so limit the number of informative markers to less than 100k, potentially omitting much of our data and reducing our power to detect heterogeneity. Furthermore, one's choice of LD sampling method will necessarily impact the performance of the method. Additionally, with respect to the detection of population structure, we cannot design a uniformly most powerful test for structure due to the complex nature in which structure can exist.

In spite of these limitations, our method provides formal, interpretable tool which is computationally fast and straightforward to implement which can test a critical assumption in a wide range of genome wide association studies.

References

[1] Silviu-Alin Bacanu, Bernie Devlin, and Kathryn Roeder. Association studies for quantitative traits in structured populations. *Genetic epidemiology*, 22(1):78–93, 2002.

[2] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[3] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[4] B Devlin, Kathryn Roeder, and Larry Wasserman. Genomic control, a new approach to genetic-based association studies. *Theoretical population biology*, 60(3):155–166, 2001.

[5] Steven Gazal, Mourad Sahbatou, Marie-Claude Babron, Emmanuelle Génin, and Anne-Louise Leutenegger. High level of inbreeding in final phase of 1000 genomes project. *Scientific reports*, 5, 2015.

[6] Hyun Min Kang, Jae Hoon Sul, Susan K Service, Noah A Zaitlen, Sit-yee Kong, Nelson B Freimer, Chiara Sabatti, Eleazar Eskin, et al. Variance component model to account for sample structure in genome-wide association studies. *Nature genetics*, 42(4):348–354, 2010.

[7] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M Kadie, Robert I Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nature methods*, 8(10):833–835, 2011.

[8] Jennifer Listgarten, Christoph Lippert, Carl M Kadie, Robert I Davidson, Eleazar Eskin, and David Heckerman. Improved linear mixed models for genome-wide association studies. *Nature methods*, 9(6):525–526, 2012.

[9] Iain Mathieson and Gil McVean. Differential confounding of rare and common variants in spatially structured populations. *Nature genetics*, 44(3):243–246, 2012.

[10] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[11] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[12] Susan E Ptak and Molly Przeworski. Evidence for population growth in humans is confounded by fine-scale population structure. *Trends in Genetics*, 18(11):559–563, 2002.

[13] Zbyněk Šidák. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[14] Benjamin F Voight and Jonathan K Pritchard. Confounding from cryptic relatedness in case-control association studies. *PLoS Genet*, 1(3):e32, 2005.

[15] Zhiwu Zhang, Elhan Ersoz, Chao-Qiang Lai, Rory J Todhunter, Hemant K Tiwari, Michael A Gore, Peter J Bradbury, Jianming Yu, Donna K Arnett, Jose M Ordovas, et al. Mixed linear model approach adapted for genome-wide association studies. *Nature genetics*, 42(4):355–360, 2010.