A genetic similarity measure for identifying fine-scale population stratification and cryptic relatedness

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

March 17, 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. 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. 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. 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. 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 perform 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 Jaccard Index and its distribution can be derived under the null-hypothesis which makes 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 relative value of rare alleles is fundamental to our method, which uses an intuitive, computationally straightforward approach towards identfying similarity between two individuals. Effectively, we give a larger weight to a genotype which is common to two individuals if the allele frequency is low among the rest of the population.

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 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

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

And in a homogeneous population we have

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

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

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

Where σ^2 is estimated by the maximum likelihood estimator

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

This provides an easily interpreted statistical test for evaluating possible relatedness between individuals in a purportedly homogeous dataset of unrelated individuals.

Furthermore, this measure is particularly sensitive 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 the related person 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

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

It can easily be shown 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

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

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, $s_{i,j}$

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

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

$$s_{diploid,i,j} = \frac{\sum_{k=1}^{N} \left[w_k \mathbf{G}_{i_1,k} \mathbf{G}_{j_1,k} + w_k \mathbf{G}_{i_1,k} \mathbf{G}_{j_2,k} + w_k \mathbf{G}_{i_2,k} \mathbf{G}_{j_1,k} + w_k \mathbf{G}_{i_2,k} \mathbf{G}_{j_2,k}\right] / 4}{\sum_{k=1}^{2N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]}$$

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

This formulation will have the same mean

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

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.

$$H_0 : \mu_{i,j} = 1 \forall i, j \in 1 \ldots n$$
$$H_A : \exists i, j \in 1 \ldots n | \mu_{i,j} \neq 1$$

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 [6] or the approximately equivalent Bonferroni procedure. We reject the null at the α level when we obtain similarity scores in the rejection region

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

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.

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

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 qunatitatively 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.

Identification of relatedness in 1000GP data

We applied our method to data from the 1000 Genomes Project [1,2], a consortium...[].

These populations were not identified to have cryptic relatedness or had 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 [3].

Phase 3 of the 1000 Genomes Project contains approximately 2504 individuals with a combined total of 88 million variants. To test our method, we sampled 80,000 variants uniformly spaced in the dataset **[This is a methodological weakness, IMO]** in order to limit the impact of linkage disequilibrium and ensure independence of variants. Our method was then run on each of the 26 populations in the study which were derived from 5 super populations sampled across the globe.

We discovered that there was great variation in the presence of cryptic relatedness and population structure across the 26 populations of the study. Under

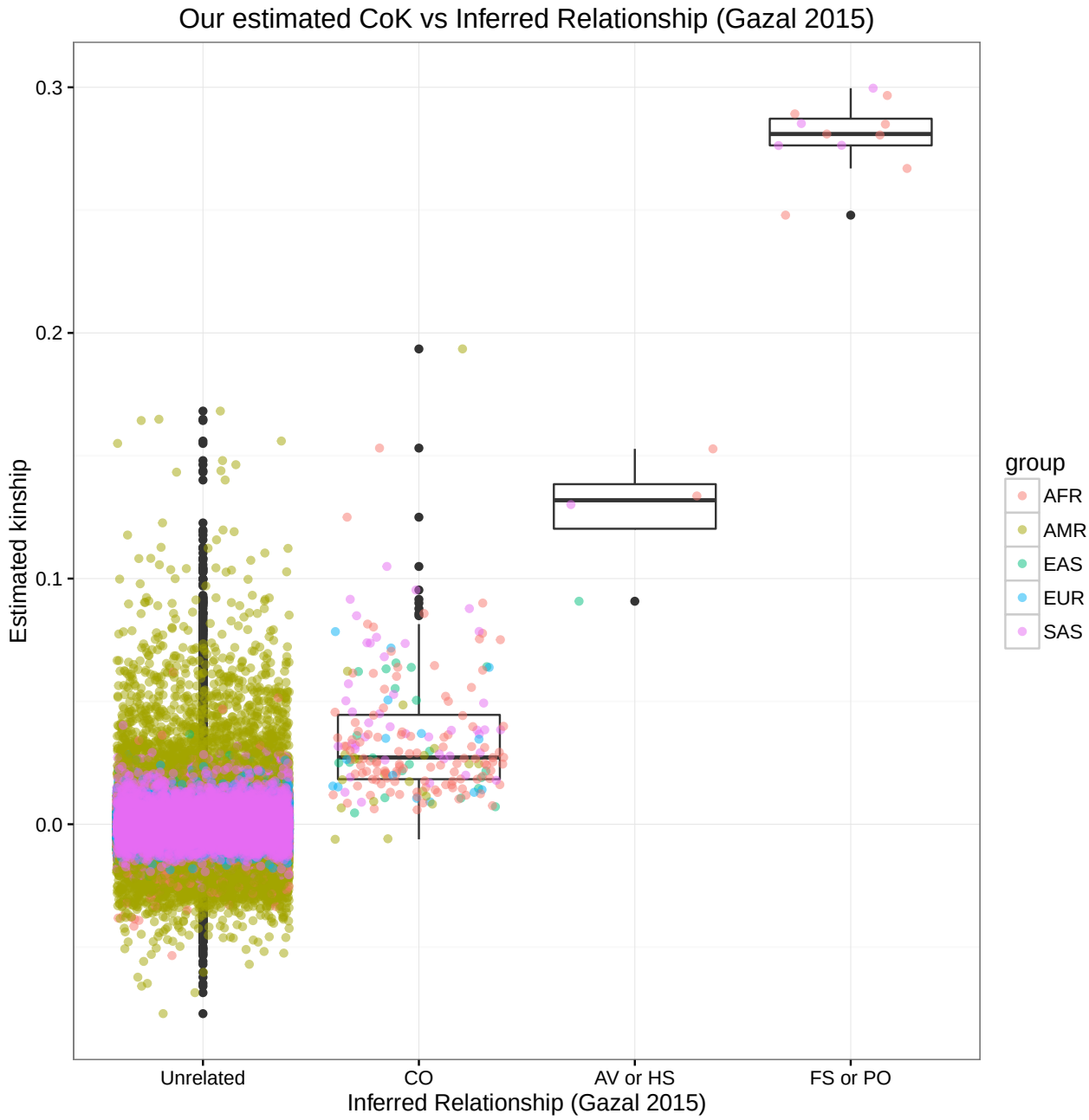
the assumptions that each study contained a homogeneous population of unrelated individuals, only a handful of groups contained neither large outliers nor heavily inflated numbers of significant results.

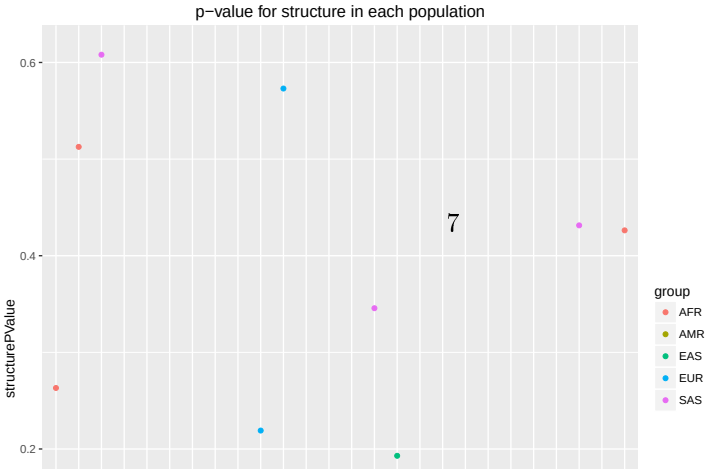
We defined the presence of population structure as applying to those populations which had a greater than expected number of similarity scores below the mean. We performed a standard one-sided binomial test with $p = 0.5$ using α cutoffs of .01 and .0011. Using this criterion, eight of the 26 populations 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.

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 6 of the 26 populations using this method.

The overlap in these two groups may be due to the fact that the variance in similarity is inflated in the presence of population structure. So it is not necessarily accurate to identify cryptic relatedness in this manner in popultations which contain 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) had 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 ($p = .0002$). 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, we found a pair of individuals HG03998 from the STU population and HG03873 from the ITU population which exhibited strikingly high relatedness. The plot below2 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 expected 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 full siblings 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.


Our estimated CoK vs Inferred Relationship (Gazal 2015)
Estimated kinship
0.3
0.2
0.1
0.0
Unrelated
CO
AV or HS
FS or PO
Inferred Relationship (Gazal 2015)
group
AFR
AMR
EAS
EUR
SAS


p−value for structure in each population
structurePValue
0.6
0.4
0.2
group
AFR
AMR
EAS
EUR
SAS

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 ($\alpha = .01$) at after multiple testing correction.
*- $p < .001$
$^{+}$- $p < .001$

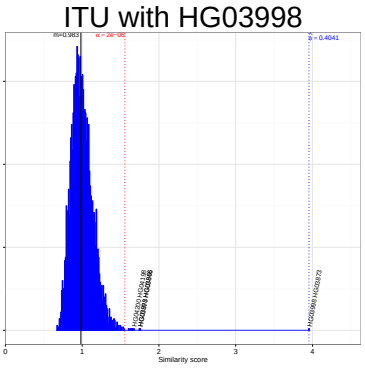


Figure 2: Distribution of s statistic for population Indian Telugu from the UK (ITU) with individual HG03998 added, who is believed to be related to HG03873, despite being labeled in the Sri Lankan Tamil from the UK (STU) population.

Population structure detection in 1000 Genomes Project

There are many methods for detecting population structure. Most commonly, Principal Components Analysis [4, 5]is applied for identifying the components 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 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 **[Adjust for filtered]** variants, we generated the two GSMs and plotted the similarity matrices, ordered by hierarchical clustering with average linkage (Figure 5).

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

Distribution of similarity statistic within population subgroups from 1000 Genomes Project

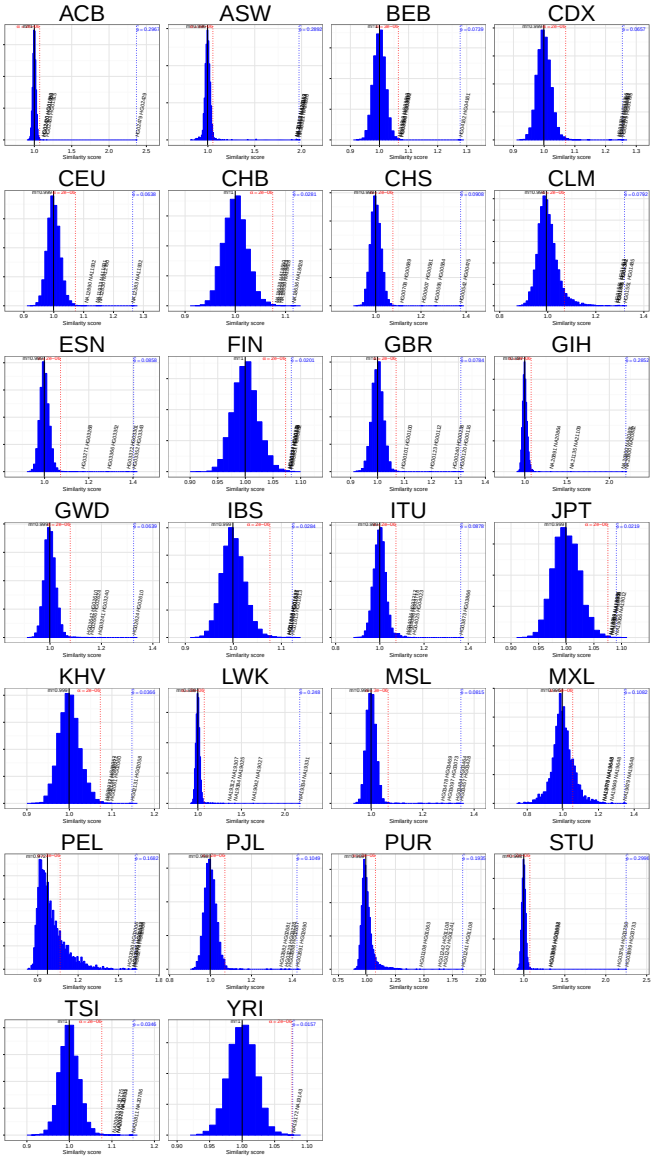


Figure 3: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise$\alpha = .05$ 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)

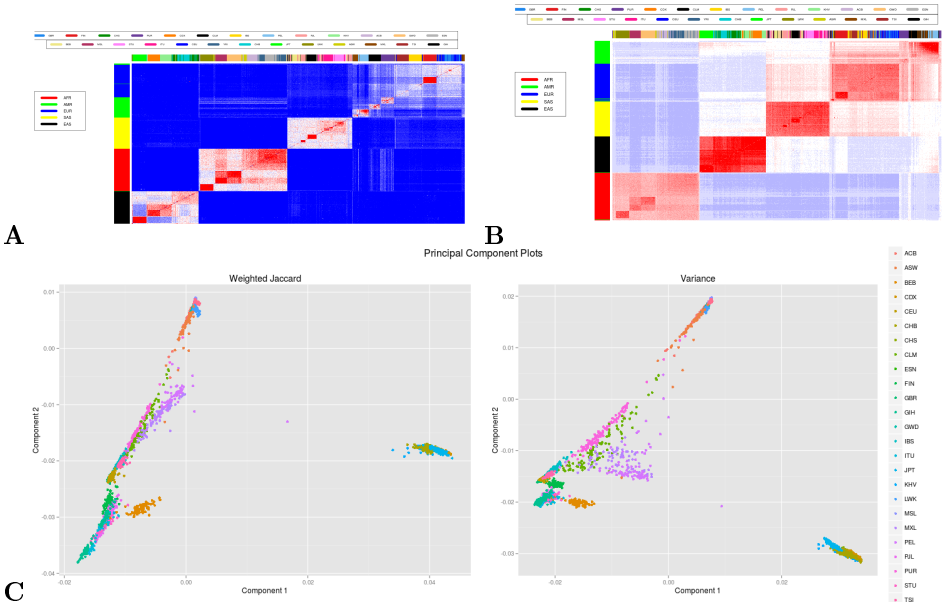


Figure 4: Heatmap of GSM generated by our method (A) and variance-covariance (B) using 80,000 uniformly spaced variants. Samples have been ordered by hierarchical clustering (dendrogram no shown). The vertical colorbar indicates membership in one of the five superpopulations, while the horizontal colorbar 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

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.

Interestingly, we found a strong case for cryptic relatedness between three pairs of individuals, one pair of which (HG03998 and HG03873) spanned the two populations. Considering that both groups were sampled in the UK, this suggests a distant genetic relationship is possible from members of different population groups.

Rarer alleles are more informative for recent ancestry

More figure and text, not sure where this goes

Separation of recent shared ancestries

Example: Indian Telugu from the UK (ITU) Sri Lankan Tamil from the UK (STU)

Separation of recent shared ancestries

Example: Iberian Population in Spain (IBS) Toscani in Italia (TSI)

Separation of recent shared ancestries

Ratio of within-group mean distance to out-of group mean distance:

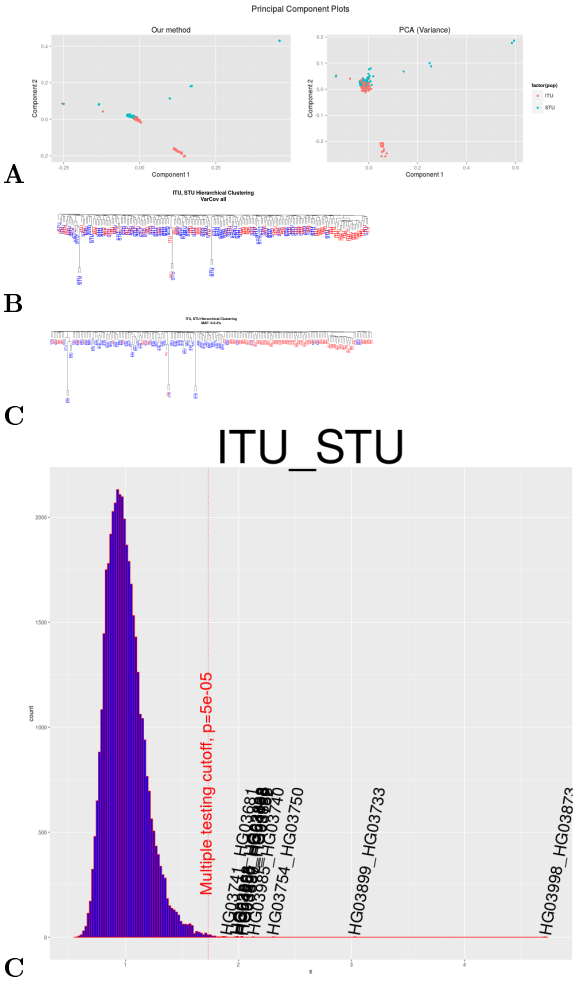
Populations	Our method	PCA
TSI-IBS	.417	.504
BEB-PJL	.748	.794
ITU-STU	.836	.889
ITU-BEB	.905	.951
CHB-CHS	.605	.681
LWK-ESN	.178	.197
GIH-ITU	.513	.552
CEU-YRI	.025	.022

Our method outperformed standard PCA in differentiating groups for *every* same-continent subpopulation pairing across all continents. (≈ 50 comparisons)

References

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

[2] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.



[Rerun this figure on whole dataset]

Figure 5: **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. Hierarchical clustering using the GSM as a similarity matrix (**B**) was unable to clearly visually distinguish between ITU and STU, but use of our method provided much clearer results (**C**).

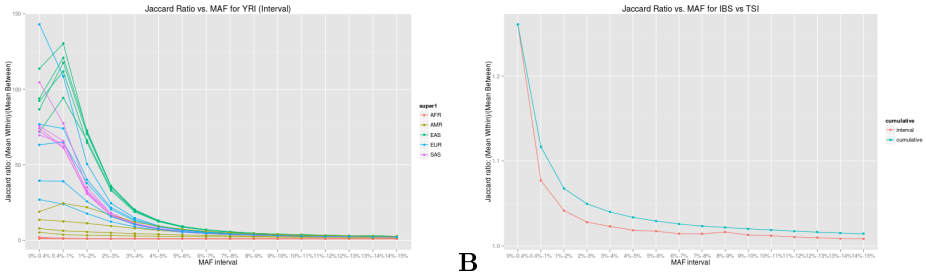


Figure 6: (**A**) Ratio for mean within-population jaccard ratio vs mean out-of-population jaccard ratio for Yoruba in Ibadan, Nigeria (YRI) compared to all other populations. We find that in comparising YRI to all other populations, the within-population vs out-of-population ratio increases as the allele frequency decreases. This trend held true for all 26 populations. For the smallest allele frequency bin, 0%-0.4%, the trend is not as clear suggesting that this is the point at which quality control is of concern when considering rare variants. (**B**) Ratio for mean within-population jaccard ratio vs mean out-of-population jaccard ratio for two populations with a relatively recent common ancestry- Iberian Population in Spain (IBS) and Toscani in Italia (TSI). Although the ratio is unsurprisingly on a smaller scale, we find that the most informative variants are those with the smallest allele frequency.

[3] 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.

[4] 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.

[5] 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.

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

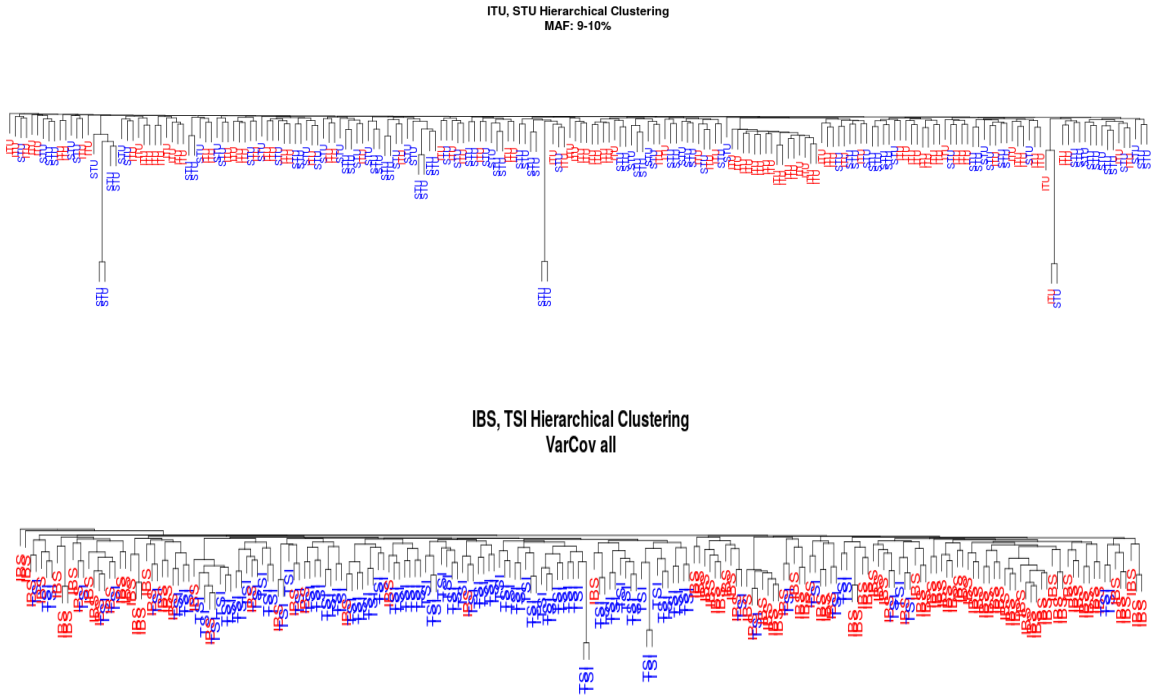


Figure 7: Figures...