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

July 21, 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. Simulation studies illustrate the overall performance of our approach. In an application to the 1000 genomes project, we outline an analysis/cleaning pipeline that utilizes our approach to formally assess whether study subjects are related and whether population substructure is present. In the analysis of 1000 Genomes Project, our approach revealed studies subjects that are most likely related, but have passed so far standard qc-filters. An implementation of our method, Similarity Test for Estimating Genetic Outliers (STEGO), is available in the R package **stego** from Github at https://github.com/dschlauch/stego.

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 [3, 19, 23]. 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 [9, 15, 18, 22]. 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 [2, 6]. 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 [15, 16]. 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 [10,11,24]. Both approaches have proven to work well in numerous applications. While the first approach is computationally fast and easy to implement, the direct modeling 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 judgment 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.

Many methods for exist for inferring relatedness which make the strong assumption of population homogeneity [3, 9, 19, 23]. These methods have been shown to be biased in the context of population heterogeneity [12]. More recently, methods have been developed recently which attempt to estimate relatedness with population structure [12, 21]. These developments improve the ability to detect existing pedigrees, which can aid in the removal of individuals who violate homogeneity. However, there is currently no quantitative measure of homogeneity which can be used to test a dataset prior to the application of GWAS.

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 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. Applications to the 1,000 Genome Project suggests that our approach is better suited to detect sub-populations than genetic variance/covariance approach. This is most likely attributable to the emphasis of our approach on small allele frequencies.

Methods

Exploiting the information in rare variants (RVs), such as one with minor allele frequency (MAF) $< 1\%$, is fundamental to our method, as our approach utilizes the features of RVs that they are typically more recent than common variants and that many of them are population/family specific. Since allele frequencies can differentially confound association studies [13], we developed a method that utilizes the differential informativeness of variants by allele frequency to obtain a high resolution picture of population structure and protect the association study against bias due to genetic confounding. Our approach uses an intuitive, computationally straightforward approach towards identifying similarity between two study subjects which is also directly linked to the kinship coefficient.

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

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

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 population 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

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

The similarity index $s_{i,j}$ provides an easily interpreted statistical test for evaluating possible relatedness between individuals in a purportedly homogeneous 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. See Supplemental Methods.

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 locus, 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

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

Test of Heterogeneity

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 family-wise error rate α, we use the Šidák procedure [20] 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 STEGO 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

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

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

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

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

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

For example, in an otherwise homogeneous study group of unrelated individuals a pair of cousins ($\phi = .0625$), with $MAF \sim Uniform\,(.02, .1)$ 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 outliers

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 $\Phi(x)$ is the cumulative distribution function for a standard normal random variable. Also note that this power is computed under the assumption of homogeneity among all unrelated individuals, which will yield a conservative estimate of probability of rejection. The presence of unknown population structure will necessarily increase the power of the test.

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

Simulations demonstrate power to detect heterogeneity and speed of method

We ran STEGO on simulated genotypes derived from a homogeneous dataset containing varying degrees of relatedness. A homogenized version of a real

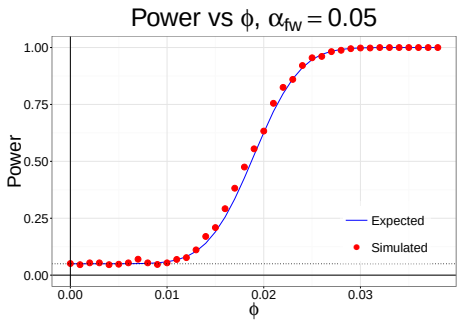

Figure 1: The probability of rejecting the null hypothesis given a simulated set of 301 homogeneous individuals containing a single related pair with coefficient of kinship, ϕ. The simulated power curve aligns with the analytically derived expectation demonstrating the clearly defined power of the method.

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

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

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 STEGO. 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 (Equation5).

Of additional interest is the computation time of STEGO 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 1,000,000 phased variants across n individuals and ran an R implementation of STEGO against the default implementation of correlation and Principal Components analysis in R, **cor()** and **princomp()** respectively.

n	stego	Correlation	PCA
250	2.009s	12.967s	6.014s
500	5.081s	49.852s	16.996s
1000	17.074s	202.049s	46.177s

Using a computer with Intel(R) Core(TM) i7-3630QM CPU @ 2.40GHz, and Microsoft R Open 3.2.5, we found a that our method ran substantially faster than correlation (**cor**) and PCA (**princomp**) in R.

Identification of relatedness and structure in 1000GP data

We applied our method to data from the 1000 Genomes Project (TGP) [4, 5], 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 [14]. However, subsequent analyses have discovered numerous inferred relationships closer than first cousins [1,7,8].

Phase 3 of the 1000 Genomes Project contains approximately 2504 individuals with a combined total of over 80 million variants. To test STEGO, we divided the data into blocks of 800 consecutive variants and selected only one locus 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 1%. This yielded approximately 100,000 variants for each of the 26 populations in the TGP. STEGO was then run on each of these populations separately to test for heterogeneity and relatedness within population groups. (Figure 4)

Our investigation revealed a great deal of 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.(Table 1)

We defined the presence of population structure as applying to those populations which deviated from the normal distribution defined under the null model. From Equation 2, we have the expected distribution under H_0 which we tested for in each of the populations using a standard Kolmogorov-Smirnov test. Using a significance cutoff of $\alpha = .01$, 15 of the 26 populations were found to have violated population homogeneity.

In addition to investigating population structure, we examined the presence of cryptic relatedness in the study. We defined relatedness as those individ-

ual pairs which exceed the cutoff for a family-wise error rate of $\alpha = .01$ and were estimated to have a coefficient of relatedness $\hat{\phi} > \frac{1}{32}$, which approximately corresponds to half first cousins. By this measure, cryptic relatedness was observed 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 [8].

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 of the similarity score is inflated in the presence of population structure. So it is incomplete to identify cryptic relatedness in this manner in populations 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 (Figure 4). Using the formula above, the estimated coefficient of kinship is $\hat{\phi} = .27$, suggesting that those individuals are first degree relatives. Additionally, 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 below (Figure 3) was generated by placing HG03998 into the ITU population and running STEGO on that population. 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 first order relatives despite belonging to different population groups. Both populations were sampled from locations in the United Kingdom, further supporting the evidence that these individuals are related.

With strong evidence of relatedness in the data, we sought to test our method on pruned data which contained no known related pairs. 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 half first cousins ($\phi = .0312$) [8]. We applied this filter and re-analyzed each of the 26 populations again to test for heterogeneity and cryptic relatedness.(Figure 2, Table 1, Supplemental Figure S1)

Eleven 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, four populations, including each of the ad-mixed American groups, continued to violate homogeneity even after the

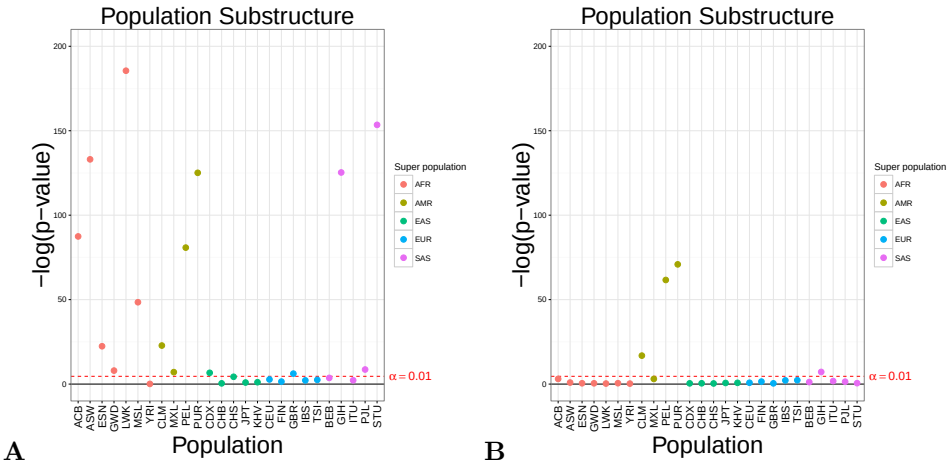

Figure 2: **Significance of population heterogeneity in 26 populations of the TGP**. Detection of population structure was found at $p < .01$ in 15 of the 26 populations using the full dataset (A). Upon removal of suspected related individuals, four populations (CLM, PEL, PUR and GIH) violated homogeneity in the relatedness-removed populations (B).

attempts to limit the impact of related individuals. The three most significant populations are all part of the Ad Mixed American super population and represent "new world" groups which have undergone extensive admixture in recent centuries. - CLM (Colombians from Medellin, Colombia)$\left(p = 7 \times 10^{-8}\right)$, PUR (Puerto Ricans from Puerto Rico) $\left(p = 3 \times 10^{-31}\right)$, and PEL (Peruvians from Lima, Peru) $\left(p = 2 \times 10^{-27}\right)$. It is therefore reassuring that these groups of individuals would exhibit the greatest amount of structure among the populations surveyed.

Population differentiation in 1000 Genomes Project

There are many methods for detecting population structure. Most commonly, Principal Components Analysis [15,16] 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, failing to fully utilize the fact that the overall allele frequency is informative of the value of each variant. Recently, the use of the Jaccard Index has been used to estimate genetic similarity [17]. This approach provides a higher resolution picture of the genetic landscape by exploiting the co-occurrence of rare-variants in sequencing data. STEGO di-

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

Table 1: **Presence of population structure and cryptic relatedness detected in each of the 26 populations in the 1000 Genomes Project.** STEGO was run separately on each population group following the removal of suspected related individuals. Population structure was defined as a significant ($p < .01$) Kolmogorov-Smirnov statistic comparing the observed test statistic distribution to that expected under the assumption of homogeneity. Cryptic relatedness was defined as those populations containing at least one pair of individuals with estimated kinship $\hat{\phi} > \frac{1}{32}$ and statistically significant ($p < .01$) kinship after multiple testing correction.

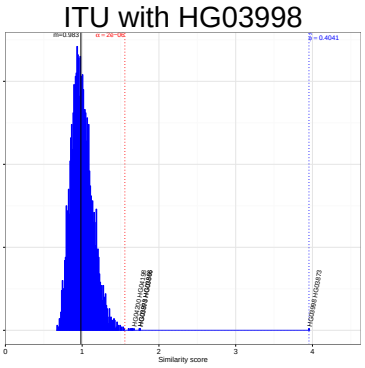

Figure 3: Distribution of all pairwise s statistics for population Indian Telugu from the UK (ITU) with individual HG03998 included. HG03998 is now believed to be related to HG03873, despite being labeled in the Sri Lankan Tamil from the UK (STU) population. The family-wise $\alpha = .01$ cutoff is indicated by the dotted red vertical line and the s statistic for HG03998 and HG03873 is seen as an extreme outlier at 3.97.

rectly utilizes this the differential value of alleles based on minor allele frequency by weighting variants by how unlikely such a co-occurrence would have been in a homogeneous population.

We evaluated the effectiveness of our similarity measure to differentiate populations in the TGP in both global and localized contexts. For the global scenario we used data from all 26 populations in a single analysis. In the localized scenarios, we ran 57 separate analyses corresponding to all possible pairs of populations within each of the five superpopulations. In each analysis, STEGO was used to compute the GSM containing all pairwise similarity scores. An eigendecomposition of the GSM was performed and each individual in the study was plotted against the top two eigenvectors.

In comparing our results with those of PCA, we achieve highly similar results on the global scale depicting the two dimensional linear migrations of ancient human history. However, despite a focus on separating recently related populations, STEGO is effective at partitioning samples of more distant common ancestry as well (Figure 5).

Despite no loss of performance on the global scale, the approach we describe here outperforms standard PCA when the task involves classifying individuals of recent ancestry. Focusing only on populations belonging to the same continental super-population, every possible pair was merged following the removal of suspected related pairs. STEGO and standard PCA were then run on each merged dataset and the two methods were compared by computing the ratio of mean within-population variance to total variance across the first three principal components.

The results show that STEGO outperforms PCA by this measure in 41 of the 57 possible comparisons (binomial test $p < .001$). We chose a pair of closely

Distribution of similarity statistic within population subgroups from 1000 Genomes Project

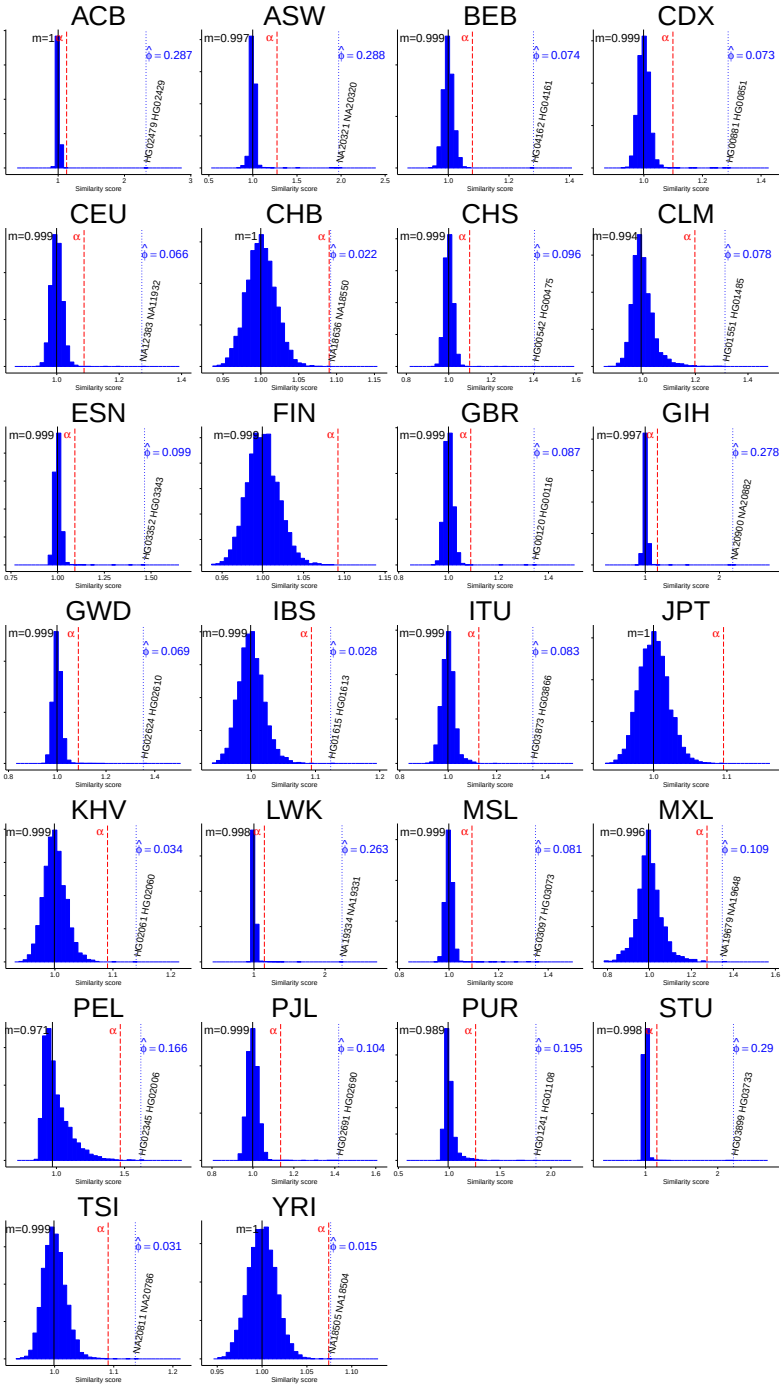

Figure 4: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. The red dotted vertical line on each plot indicates the family-wise$\alpha = .05$ 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)

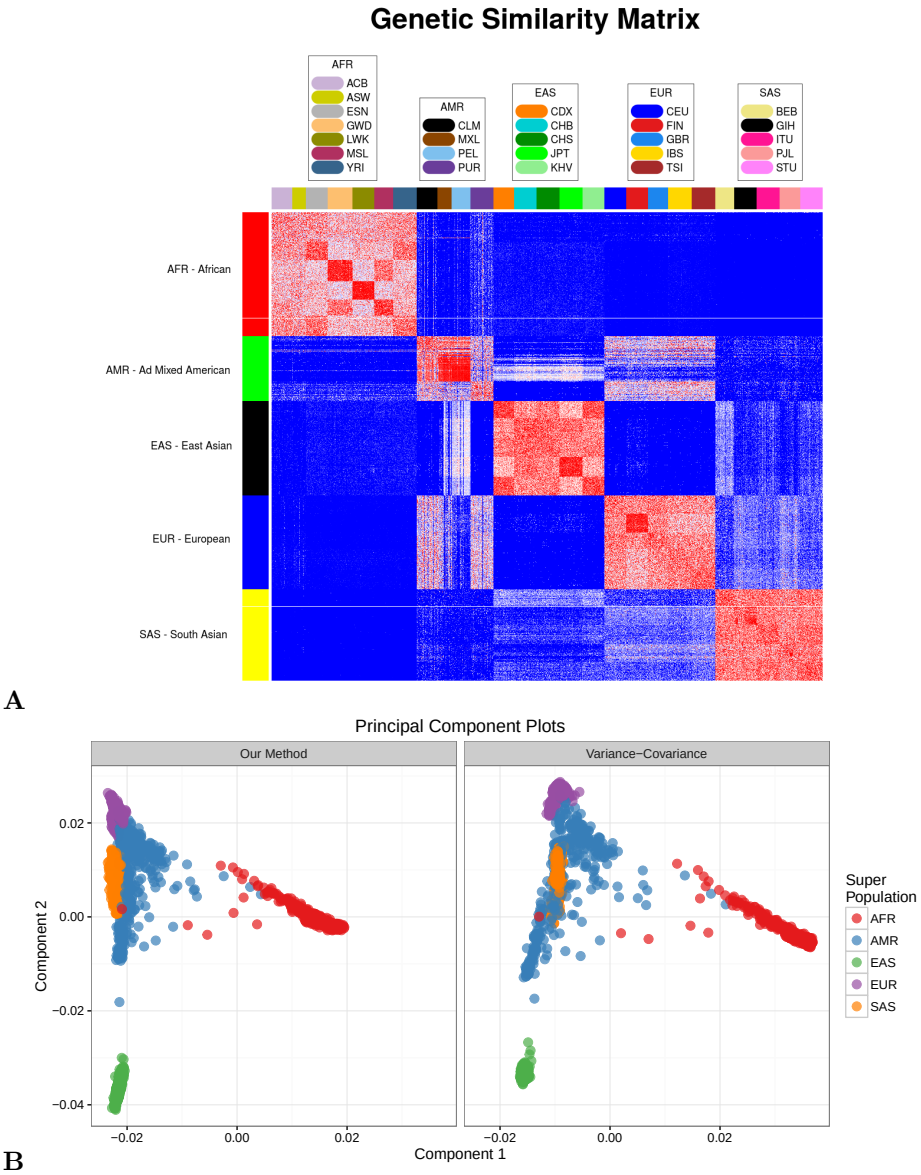

Figure 5: Population structure in 2504 samples from 1000 Genomes Project. (**A**) Heatmap of the GSM generated by STEGO using 80,000 LD-sampled variants. The vertical colorbar indicates membership in one of the five superpopulations, while the horizontal colorbar indicates membership in one of the 26 populations. (**B**) Projecting each individual onto the top two eigenvectors resulted in a similar 2-dimensional distribution of global ancestry. Both STEGO and PCA show similar projections which elucidate the migratory patterns of early humans.

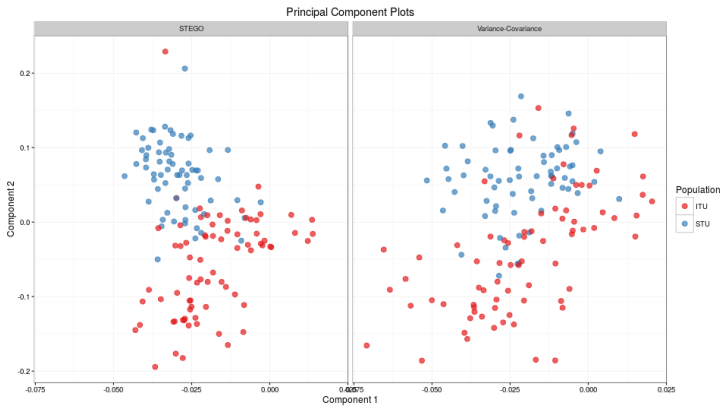

Figure 6: **Example: ITU vs STU**. Two populations of Southern Asian origin, Indian Telugu from the UK (ITU) and Sri Lankan Tamil from the UK (STU). A genetic similarity matrix was computed using STEGO and standard correlation. An eigendecomposition of each matrix was performed. These plots show the set of unrelated individuals projected on to the first two eigenvectors. We see clearer clustering by population (colored) in our method (Left) compared to standard PCA (right). This performance boost is attributed to the value added by preferentially considering genetic agreement in less frequent alleles.

related populations from the 1000 Genomes Project in order to demonstrate this performance. The populations Sri Lankan Tamil (STU) and Indian Telugu (ITU) have relatively small geographical separation and recent common ancestry relative to other populations in the TGP. We used a subset of each population determined to be homogeneous and lacking cryptic relatedness as described above. We demonstrate the clearer separation in (Figure 6) comparing our method with that of standard Principal Components Analysis.

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 become fixed at 0% with high probability. Therefore, rare alleles that are observed are more likely to have arisen recently. It stands to reason that these alleles would therefore be the most informative of recently related populations. By appropriately recognizing the increased information contained in the co-occurrences of less frequent alleles, we achieve superior separation 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 in-

volves a qualitative assessment without any specific concern for effect size and power. STEGO provides an analytical approach for quantitatively assess homogeneity 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, STEGO provides a formal, interpretable tool which is directly linked to the kinship coefficient. It provides a formal statistical test for population substructure, identifying study subjects which are related and subjects which are genetic outliers in their assigned population.

References

[1] Ahmed Al-Khudhair, Shuhao Qiu, Meghan Wyse, Shilpi Chowdhury, Xi Cheng, Dulat Bekbolsynov, Arnab Saha-Mandal, Rajib Dutta, Larisa Fedorova, and Alexei Fedorov. Inference of distant genetic relations in humans using '1000 genomes'. *Genome biology and evolution*, 7(2):481–492, 2015.

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

[3] Yoonha Choi, Ellen M Wijsman, and Bruce S Weir. Case-control association testing in the presence of unknown relationships. *Genetic epidemiology*, 33(8):668–678, 2009.

[4] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[5] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[6] 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.

[7] Larisa Fedorova, Shuhao Qiu, Rajib Dutta, and Alexei Fedorov. Atlas of cryptic genetic relatedness among 1000 human genomes. *Genome biology and evolution*, 8(3):777–790, 2016.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] Ani Manichaikul, Josyf C Mychaleckyj, Stephen S Rich, Kathy Daly, Michèle Sale, and Wei-Min Chen. Robust relationship inference in genome-wide association studies. *Bioinformatics*, 26(22):2867–2873, 2010.

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

[14] James Nemesh and Steve McCarroll. Addressing cryptic relatedness in candidate samples for 1kg. ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/cryptic_relation_analysis. Accessed: 2016-06-06.

[15] 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.

[16] 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.

[17] Dmitry Prokopenko, Julian Hecker, Edwin K Silverman, Marcello Pagano, Markus M Nöthen, Christian Dina, Christoph Lange, and Heide Loehlein Fier. Utilizing the jaccard index to reveal population stratification in sequencing data: a simulation study and an application to the 1000 genomes project. *Bioinformatics*, 32(9):1366–1372, 2016.

[18] 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.

[19] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

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

[21] Timothy Thornton, Hua Tang, Thomas J Hoffmann, Heather M Ochs-Balcom, Bette J Caan, and Neil Risch. Estimating kinship in admixed populations. *The American Journal of Human Genetics*, 91(1):122–138, 2012.

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

[23] Jian Yang, Beben Benyamin, Brian P McEvoy, Scott Gordon, Anjali K Henders, Dale R Nyholt, Pamela A Madden, Andrew C Heath, Nicholas G Martin, Grant W Montgomery, et al. Common snps explain a large proportion of the heritability for human height. *Nature genetics*, 42(7):565–569, 2010.

[24] 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.