# Supplementary Materials for

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[1,2], Heide Fier[3], and Christoph Lange[1,2]

[1]Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, MA 02115
[2]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02115
[3]Institute of Genomic Mathematics, University of Bonn, Bonn, Germany

August 19, 2016

## 1 Supplementary Methods

### 1.1 Variance of $\hat{s}$

The variance of $\hat{s}_{ij}$ can be estimated by

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

This formulation is independent of the samples $i, j$ and depends only on the allele counts for each variant across the study group.

$$
\begin{aligned}
Var\left(s_{i,j}\right) &= Var\left(\frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2 Var\left(\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2 \sum_{k=1}^{N} Var\left(w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) && \text{Independence of variants} \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2 \sum_{k=1}^{N} w_k^2 Var\left(\mathbf{G}_{i,k}\mathbf{G}_{j,k}\right) \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2 \sum_{k=1}^{N} w_k^2 P\left[\mathbf{G}_{i,k}\mathbf{G}_{j,k} = 1\right]\left(1 - P\left[\mathbf{G}_{i,k}\mathbf{G}_{j,k} = 1\right]\right) && \text{Variance of Bernouli RV} \\
&= \left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2 \sum_{k=1}^{N} w_k^2 \frac{1}{w_k}\left(1 - \frac{1}{w_k}\right) \\
&= \frac{\sum_{k=1}^{N}\left(w_k - 1\right)}{\left(\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]\right)^2}
\end{aligned}
$$

### 1.2 Linkage Disequilibrium Pruning

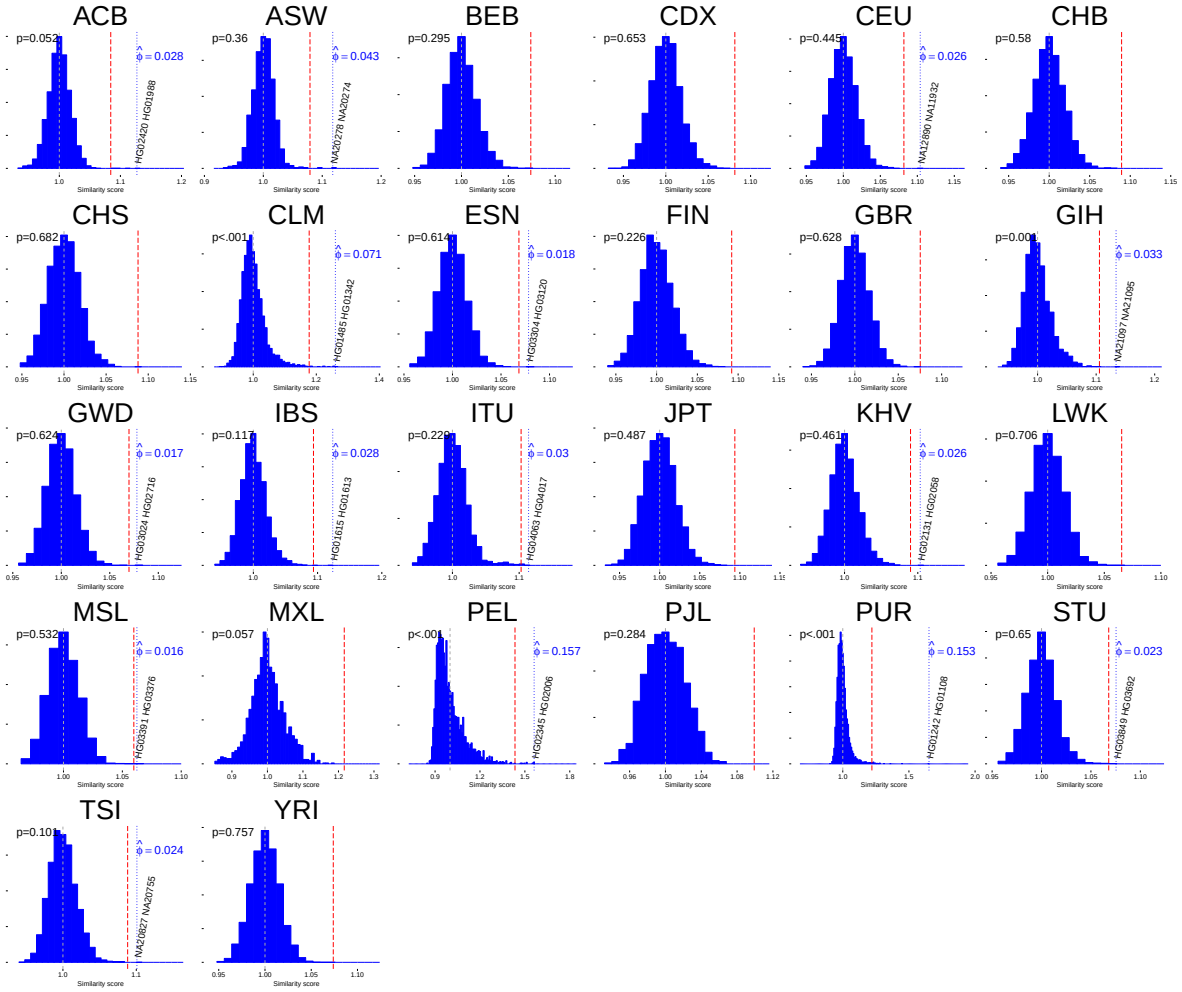
Phase 3 of the 1000 Genomes Project contains 2504 individuals with a combined total of over 80 million variants. Assumptions of STEGO include the independence of variants, which may be violated in the presence of Linkage Disequilibrium (LD). Our method focuses variants with low minor allele frequency, which are less susceptible to high $R^2$ between loci. However, to help reduce the impact of correlated variants, we filtered the data such that the impact of LD was limited. Prior to analysis, we divided the data into blocks of 800 consecutive variants and selected only one locus from each block. The selected variant within each block was chosen based on the smallest minor allele frequency observed which was larger than our cutoff of 1%. We chose this somewhat arbitrary cutoff as a compromise between our desire to focus on rare alleles and the recognition that QC concerns may become an increasingly valid concern at the lowest allele frequencies. We recommend the use of a cutoff which balances the value of rare variants with the confidence in the technology used to obtain the data.

Furthermore, evidence supports the presence of long-range LD, long stretches along chromosomes which have relatively highly correlated variants. Twenty-four such features have been documented in Price et al [18] and to address this we simply excluded these regions from our analysis.

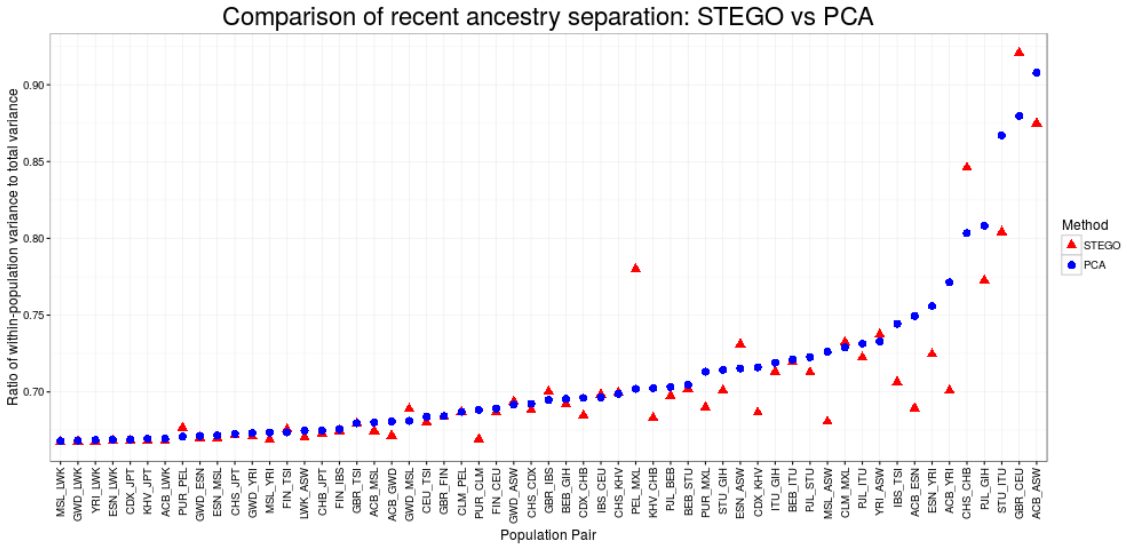
This filtering yielded approximately 100,000 variants for each of the 26 populations in the TGP.

## 2 Supplementary Figures

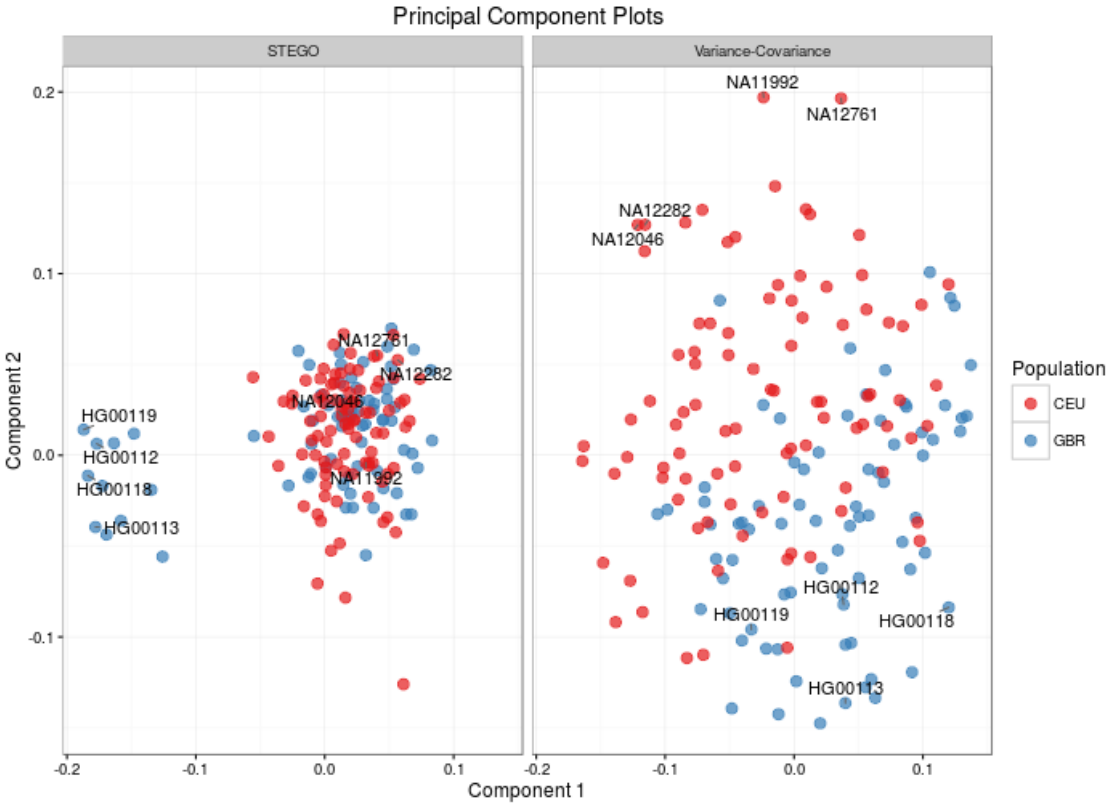
**Distribution of similarity statistic within population subgroups from 1000 Genomes Project after removal of related individuals**




Supplemental Figure 1: Distribution of similarity coefficients for each of the 26 populations in the 1000 Genomes Project after the removal of suspected related individuals. Homogeneous populations lacking cryptic relatedness should be expected to exhibit distributions centered around 1 with no outliers. A heterogeneous population is expected to exhibit a normal distribution centered around 1. Non-normal distributions such as right-skewed (e.g. PUR, PEL, CLM) or bimodal are indicative of population structure. The red dotted vertical line on each plot indicates the family-wise$\alpha = .01$ level cutoff for $\binom{n}{2}$ comparisons. The most significant related pair is labeled for each population with the estimated kinship for that pairing indicated in blue. The p-value for the KS test for homogeneity is reported for each population. Outliers in the absence of non-normally distributed statistics are an indication of relatedness among pairs of individuals.




Supplemental Figure 2: The eigendecomposition of the STEGO matrix separates individuals of populations belonging to the same continent with greater efficiency than PCA. In 43 of 57 possible within-continent population pairs, STEGO had super separation of populations. Separation was measured as the ratio of within population variance to total variance along the first 3 eigenvectors for STEGO and PCA.




Supplemental Figure 3: Despite a clear trend of superior performance with STEGO, notable exceptions occur. For example, by this measure, the populations GBR and CEU were more clearly divided by PCA (Right, ) than by STEGO (Left). Closer inspection revealed that the first eigenvector from STEGO isolates 11 samples exclusively from the GBR population. It is not readily apparent what features of the data are being captured here or the relative value of those features (this may be a result of population structure, relatedness, batch effect, etc.). But it is notable that all 11 samples came from the same population in the 1000 Genomes Project. It is reasonable to infer that this subset of samples is scientifically relevant. It most likely contains a disproportionate number of co-occurences of rare variants, which were not observed separately by PCA.