A similarity measure for detecting genetic outliers

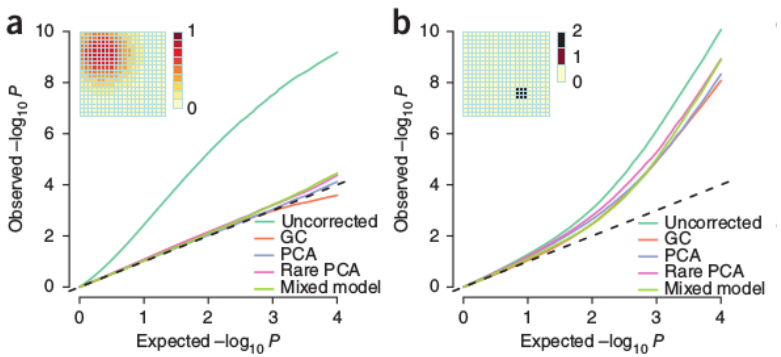
Dan Schlauch, PhD Candidate

Department of Biostatistics, Harvard School of Public Health

June 10, 2016

Background

- Individuals may be too similar (due to cryptic relatedness) or too different (due to population structure).
- Both features may lead to spurious results, inflation of type I error.
- Many methods exist for addressing some of these concerns (e.g. PCA, LMM).
- Limitations exist, such as with rare alleles and sharp localized effects, or with the assumption of linear or discrete population structure.



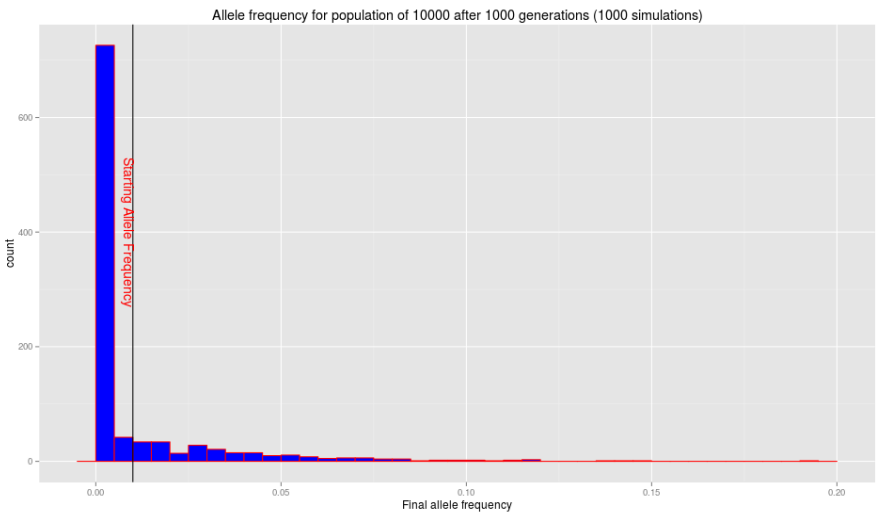
Mathieson, Nature Genetics 2012

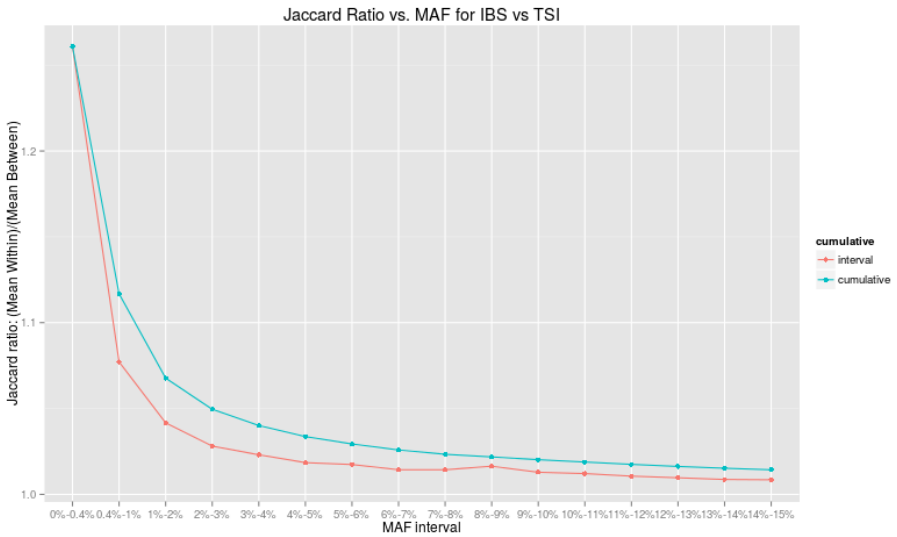
We want to create a similarity measure that...

- is more sensitive to fine scale population stratification
- can be used as a formal test for cryptic relatedness
- can be used as a formal test for population structure

Basis for measure

- Rare variants are recent variants.
- In the absense of selection, rare variants become fixed at 0% with high probability over a relatively short timeframe.





P[Fixation|n=10000,g=1000,maf=.01]=.678

- Key Idea: **Less** frequent variants are **more** informative of ancestry.

Test Statistic

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

where

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

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

It therefore follows from the CLT 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 Normal.

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

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

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

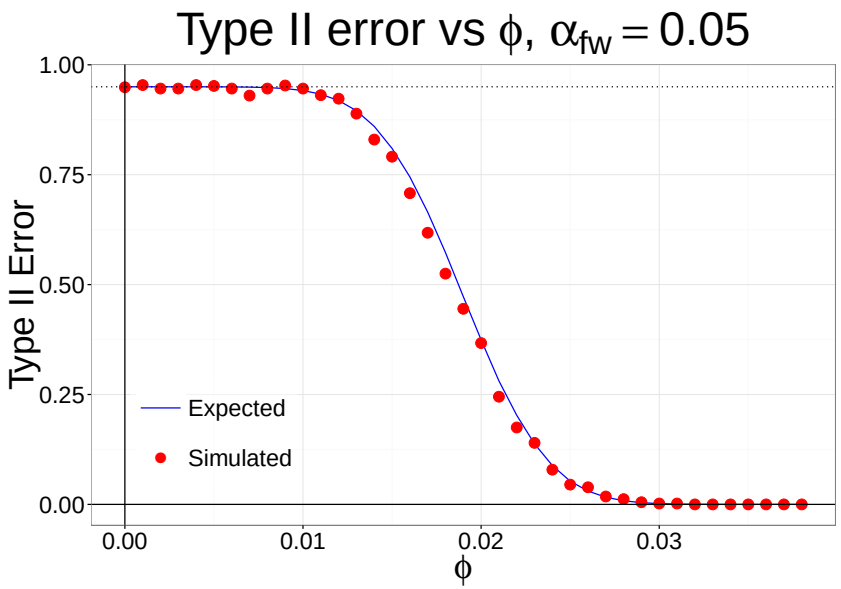
$$s_{i,j}^{(diploid)} = \frac{\sum_{k=1}^{N} \left[w_k \mathbf{H}_{i,k} \mathbf{H}_{j,k}\right] / 4}{\sum_{k=1}^{N} I\left[\left(\sum_{l=1}^{n} \mathbf{H}_{l,k}\right) > 1\right]}$$

Properties of test statistic

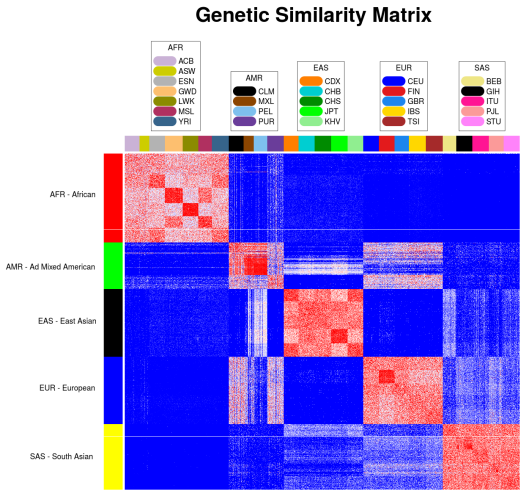
$$\hat{\phi}_{i,j} = \frac{s_{i,j} - 1}{\left[\frac{\sum_{k=1}^{N} \hat{p_k} w_k}{\sum_{k=1}^{N} I\left[\sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1\right]} - 1\right]}$$

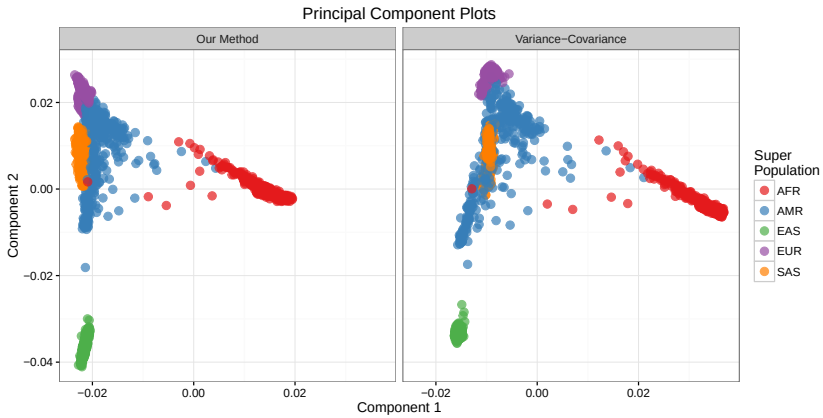
$$R : max\left(s_{i,j}\right) > 1 - probit\left(\frac{\alpha}{\binom{n}{2}}\right)$$

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

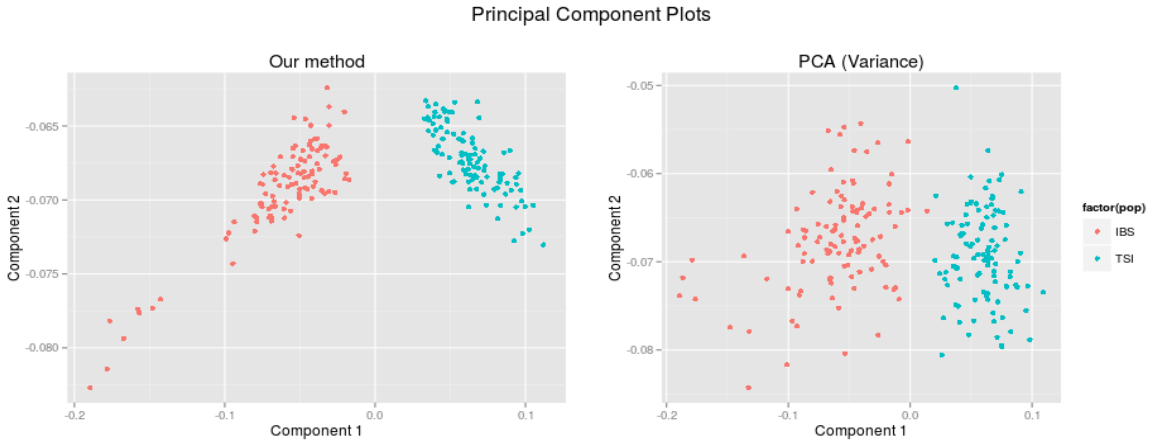


Results: Application to 1000 Genomes Project



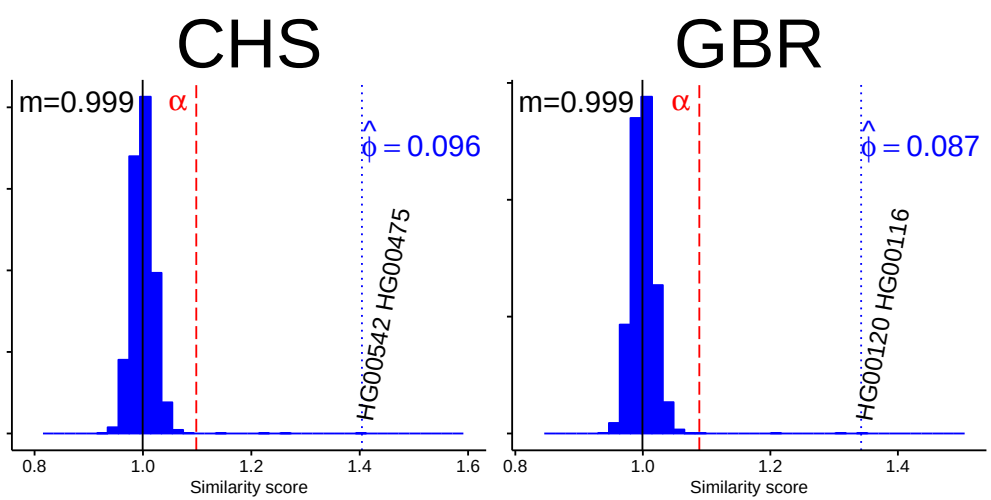


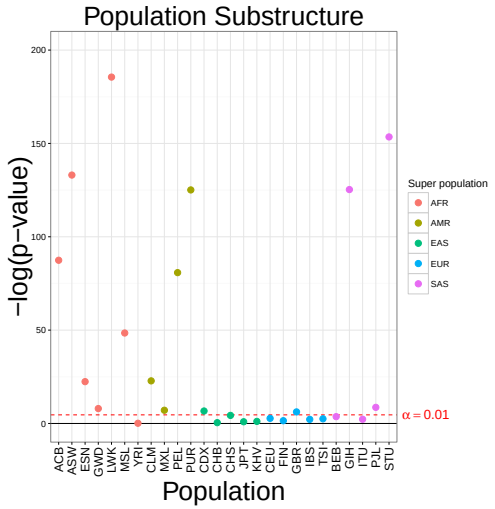
Our method is comparable to PCA when applied on a global scale.



But produces superior separation for recently related populations.

Application to 1000 Genomes Project





After removal of related pairs