Identifying fine-scale population stratification with rare alleles

Dan Schlauch, PhD Candidate

Department of Biostatistics
Harvard School of Public Health

November 18, 2015

Outline

1. Population stratification
   - What is population stratification?
   - Differential confounding by allele frequency
   - Problems with stratification correction
2. Estimating population structure
   - Addressing fine-scale stratification
   - Identifying structure in 1000 Genomes Project
   - Controlling confounding in 1000 Genomes Project
3. Corrected association test statistic
   - Bias of stratification adjusted tests
   - Applying corrected estimator to 1000GP data

What is population stratification?

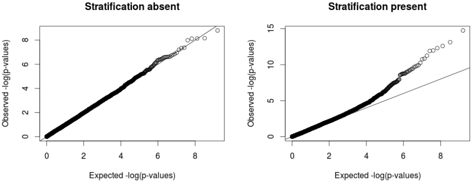
- Population stratification is the existence of allele frequency differences across population groups due to differing ancestries.
- PS is typically caused by geographical isolation, leading to non-random mating patterns.
- Direct associations between a genetic variant and a phenotype may be confounded by ancestry.

Confounding due to population stratification

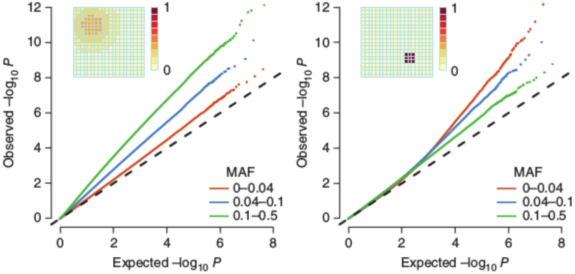
For example, in a case-control GWAS we are typically interested in how genetic variants contribute to disease:
If ancestry is associated with disease and population stratification exists

$$cor\left(Disease, \mathbf{G}\right) \neq cor\left(Disease, \mathbf{G}|\mathbf{C}\right)$$

An uncontrolled test will lead to spurious associations and inflation of type I error.



Rare allele association inflation



Mathieson, Nature Genetics 2012

QQ plots of p-values separated by allele frequency. Comparing two types of non-genetic risk distributions.

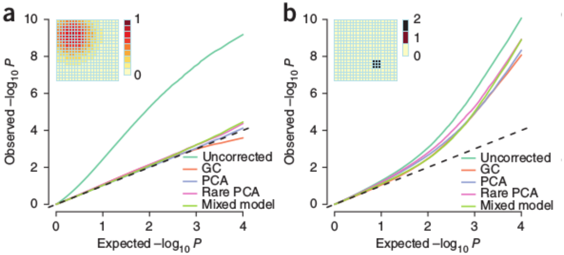
Rare allele association inflation

Differential confounding by allele frequency

The magnitude of confounding due to stratification is a function of allele frequency and phenotypic distribution

- For a gradual phenotypic distribution.
  - Greater inflation of **common** alleles
- For a sharp phenotypic distribution
  - Greater inflation of **rare** alleles

Rare allele association inflation



Mathieson, Nature Genetics 2012

Existing methods for correction for population stratification do not work for sharp phenotypes (and are particularly ineffective for rare variants).

Why do we observe inflation?

There are at least 3 problems

1. Common stratification correction methods inadequately distinguish the tree-like ancestry.
    - Need better estimates of genetic relatedness.
2. Differential genotype/phenotype variances lead to scaling of null test statistic distribution.
    - Need better estimates of test statistic distribution.
3. Finite sample sizes lead to overdispersion of the association test statistic.
    - Need improvements over use of asymptotic distributions.

Addressing fine-scale stratification

Common stratification correction approach

- Build a variance-covariance matrix between all samples using all variants and identify top axes of variation via PCA. (Eigenstrat)
- Apply correction using the top PCs.

Limitations

- Assumes populations are linearly structured in space.
- Inherently relies on common variants relative to rare variants.
  - Unable to clearly separate closely related populations, such as Europeans from Spain vs Italy

Addressing fine-scale stratification

Consider the following haplotype matrix, with columns as samples and rows as variants:

01110001010101101011 - .5
11100010111001100110 - .5
00000001010110001000 - .25
01010010101000000000 - .25
00010100010101000010 - .25
00000000000000000000 - 0
00000000000011000000 - .1
01100000000000001010 - .2
00000100010010001000 - .2

Addressing fine-scale stratification

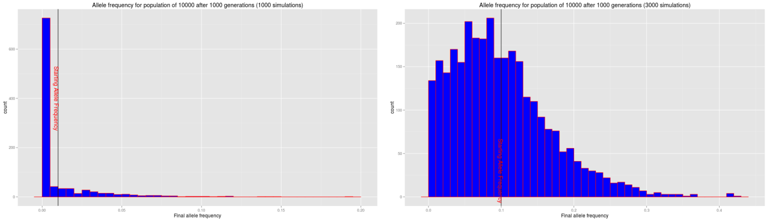
Consider the following haplotype matrix, with columns as samples and rows as variants:

011100010101**01**101011 - .5
111000101110**01**100110 - .5
000000010101**10**001000 - .25
010100101010**00**000000 - .25
000101000101**01**000010 - .25
000000000000**00**000000 - 0
000000000000**11**000000 - .1
011000000000**00**001010 - .2
000001000100**10**001000 - .2

Addressing fine-scale stratification

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

Starting MAF: .01 vs .1



P[Fixation]=.678 vs P[Fixation]=.017

Addressing fine-scale stratification

A simple proposed approach

- Utilize the intuition that rare variants are more informative than common variants.
- Build a genetic similarity matrix based on a weighted variation of the Jaccard Index and perform eigendecomposition.
- Apply correction using the top PCs.

Jaccard Index:

$$J = \frac{|A \cap B|}{|A \cup B|}$$

Genetic Similarity Measure

For a matrix of n individuals ($2n$ haploid genomes), with N variants described by the genotype matrix $\mathbf{G}_{2n\times N}$, we define the weighted Jaccard similarity between two haploid genomes, $s_{i,j}$

$$s_{i,j} = \frac{\sum_{k=1}^{N} w_k \mathbf{G}_{i,k}\mathbf{G}_{j,k}}{\sum_{k=1}^{N} \mathbf{G}_{i,k} + \sum_{k=1}^{N} \mathbf{G}_{j,k} - \sum_{k=1}^{N} \mathbf{G}_{i,k}\mathbf{G}_{j,k}}$$

where

$$w_{k,i,j} = \begin{cases} \frac{2(2n-1)}{\sum_{l=1}^{2n} \mathbf{G}_{l,k} - 1} - 1 & \sum_{l=1}^{2n} \mathbf{G}_{l,k} > 1 \\ 0 & \sum_{l=1}^{2n} \mathbf{G}_{l,k} \leq 1 \end{cases}$$

$$E\left(s_{i,j} | \text{No structure}\right) = 1$$

$$\hat{Var}\left(s_{i,j} | \text{No structure}\right) \approx \frac{\sum_{k=1}^{N} \hat{p}_k^2 \left(1 - \hat{p}_k^2\right) w_{k,i,j}^2}{\left(\sum_{k=1}^{N} \mathbf{G}_{i,k} + \sum_{k=1}^{N} \mathbf{G}_{j,k} - \sum_{k=1}^{N} \mathbf{G}_{i,k}\mathbf{G}_{j,k}\right)^2}$$

Genetic Similarity Measure

This measure is particularly sensitive for measuring kinship. Given a Coefficient of relatedness, $r > 0$,

$$E\left(s_{i,j} | r, \text{No other structure}\right) =$$

$$= \frac{(1-r)\sum_{i=1}^{N}\left(2p_i - p_i^2\right) + rN}{(1-r)\sum_{i=1}^{N}\left(2p_i - p_i^2\right) + r\sum_{i=1}^{N} p_i}$$

$$> 1$$

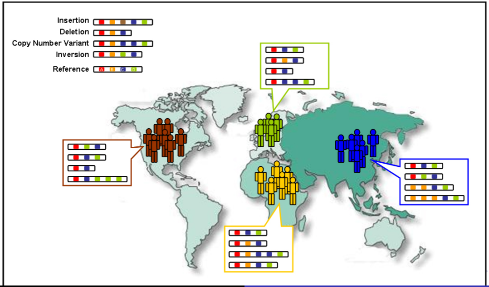
e.g. with $MAF \sim Uniform\,(.01, .1)$

$$E\left(s_{i,j} | r = .125, \text{No other structure}\right) \approx 2.9$$

1000 Genomes Project

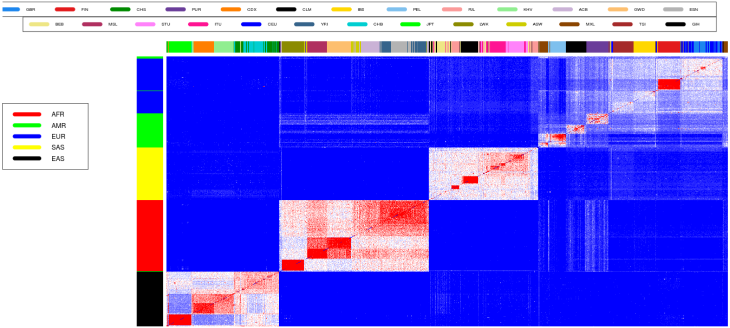
1000 Genomes Project dataset

- 2504 individuals
- 6 superpopulations (African, Ad-Mixed American, East Asian, European, South Asian)
- 26 populations
- 60 million variants



Separation of all individuals

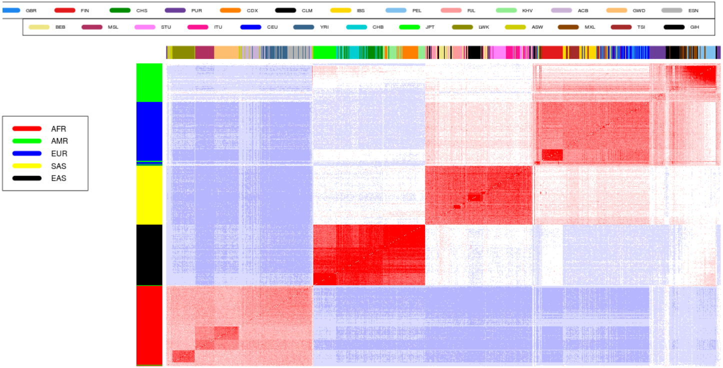
Clustered heatmap of GSM based on our method.



Clustered heatmap of genetic similarity using our method.

Separation of all individuals

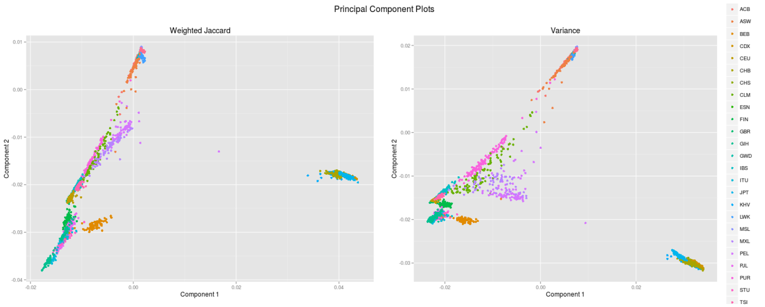
Clustered heatmap of GSM based on variance-covariance matrix.



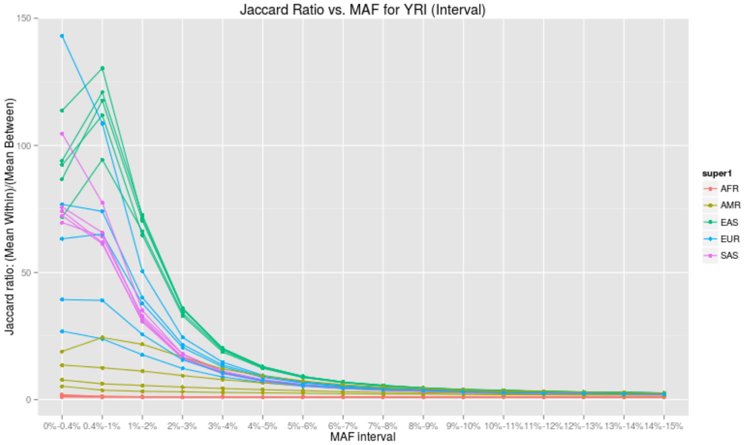
Clustered heatmap of genetic similarity using PCA.

Separation of all individuals

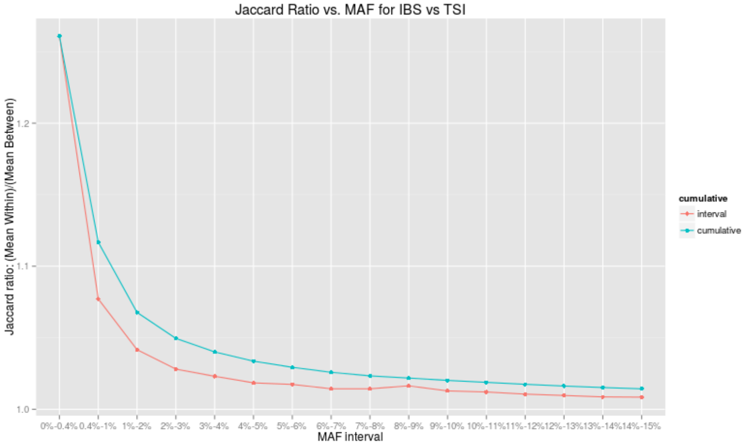
First two principal components using our method vs Var-Cov yield very similar results.
Continental level population structure is not meaningfully affected.



Separation as a function of allele frequency

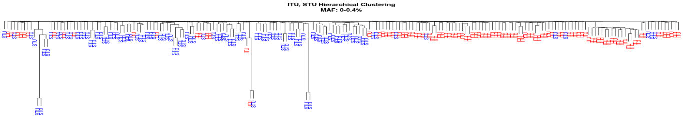


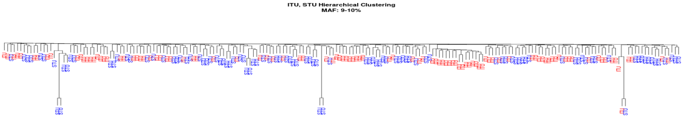
Separation as a function of allele frequency

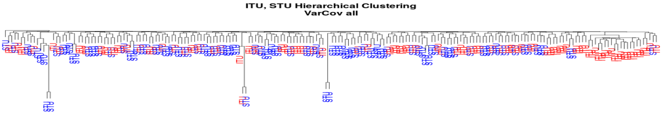


Separation of recent shared ancestries

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

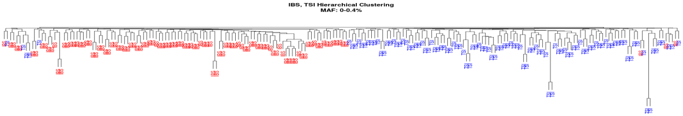

ITU, STU Hierarchical Clustering
MAF: 0-0.4%

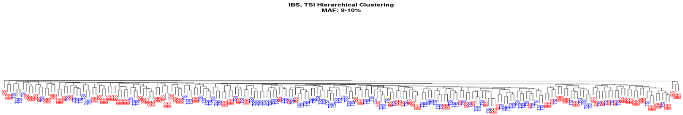

ITU, STU Hierarchical Clustering
MAF: 9-10%

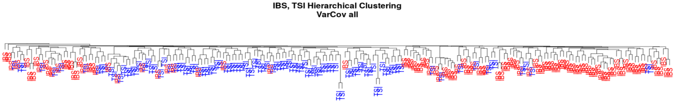

ITU, STU Hierarchical Clustering
VarCov all

Separation of recent shared ancestries

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

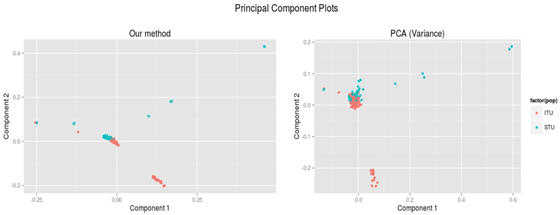




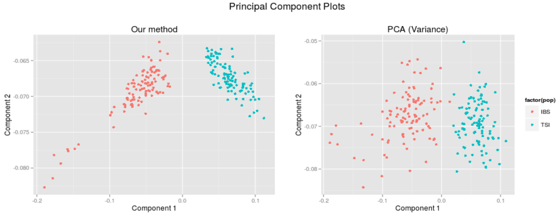


Separation of recent shared ancestries

Example: ITU vs STU



Example: IBS vs 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. ($\approx$ 50 comparisons)

Simulated non-genetic phenotypes with 1000GP genotypes

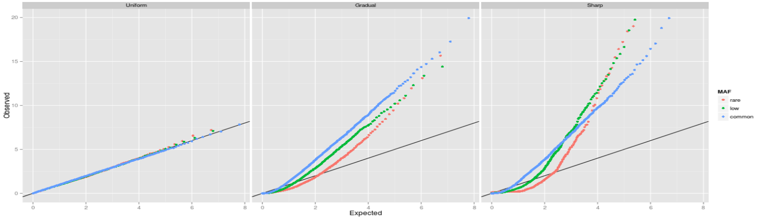
Phenotype simulation

- Phenotypes were simulated as Bernouli or exponential RV
- Risk was assigned based on 3 separate risk models:
  - Uniform risk
  - Super and sub-population differentiated risk (gradual risk)
  - Sub-population alone differentiated risk (sharp risk)
- 100 phenotypes generated per model (300 total per distribution type)
- GWAS performed on each phenotype for "LD sampled" set of 100k variants
- Mean rank-ordered p-value taken for each simulated phenotype.

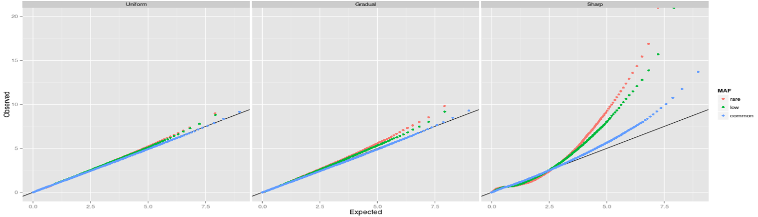
Does use our method preserve type I error?

QQ-plots of p-values for variant association with a non-genetic binary phenotype

Uncorrected



Our method (20 components) association Q-Q plot



Differential genotype/phenotype variances

Consider the stratification-adjusted test statistic

$$t_k = n \times \left(\frac{\mathbf{r}_P \mathbf{r}_G^T}{\sqrt{\sum_i^n \mathbf{r}_{P,k,i}^2 \sum_i^n \mathbf{r}_{G,k,i}^2}} \right)^2 \sim \chi_1^2$$

Where $\mathbf{r}_P$ and $\mathbf{r}_G$ are the residuals for the phenotype and genotype, respectively.

Now consider P, a binary phenotype with $p_{P,i}$ and $p_{G,i}$ as the mean phenotypes and genotypes given stratification, and k be the index of a variant

t_k is biased

$$E\left(t_k | H_0\right) = n \frac{\sum_{i=1}^n \left[2p_{G,k,i}\left(1 - p_{G,k,i}\right) p_{P,i}\left(1 - p_{P,i}\right)\right]}{\sum_{i=1}^n 2p_{G,k,i}\left(1 - p_{G,k,i}\right) \sum_{i=1}^n p_{P,i}\left(1 - p_{P,i}\right)}$$
$$\neq 1$$

Differential genotype/phenotype variances

The residual correlation estimate of association is biased when there is a mean-variance relationship for phenotype (there is always a mean-variance relationship for genotype)

$$cor\left(var\left(\mathbf{r}_{G,k}\right), var\left(\mathbf{r}_P\right)\right) > 0 \rightarrow E\left(t_k\right) > 1$$
$$cor\left(var\left(\mathbf{r}_{G,k}\right), var\left(\mathbf{r}_P\right)\right) < 0 \rightarrow E\left(t_k\right) < 1$$

Correcting biased estimator

Consistent estimator for stratification adjusted association with mean-variance:

$$s_k = w \times n \times \left(\frac{\mathbf{r}_P \mathbf{r}_G^T}{\sqrt{\sum_i^n \mathbf{r}_{P,k,i}^2 \sum_i^n \mathbf{r}_{G,k,i}^2}} \right)^2 \sim \chi_1^2$$

Where

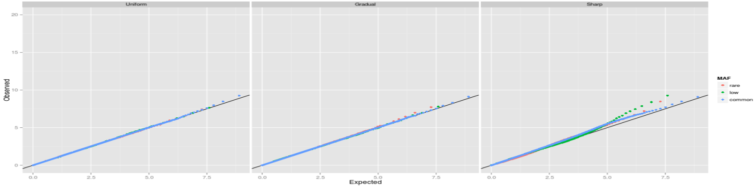
$$w = \left(\frac{\sum_{i=1}^n \hat{\sigma}_{P,k,i}^2 \sum_{i=1}^n \hat{\sigma}_{G,k,i}^2}{\sum_{i=1}^n \left[\hat{\sigma}_{P,k,i}^2 \hat{\sigma}_{G,k,i}^2 \right]} \right)$$

$$E\left[s_k \middle| H_0 \right] = 1$$

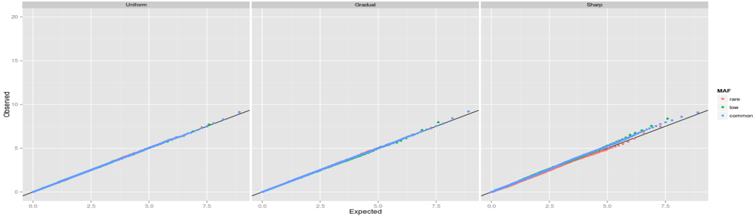
Applying corrected estimator to 1000GP data

QQ-plots of p-values for variant association with a non-genetic binary phenotype across TSI-IBS

PCA (2 components)



Our method (2 components)



Acknowledgements

Thanks to:

- Prof. Christoph Lange
- Matt Goodman