Fine-scale population stratification with rare alleles

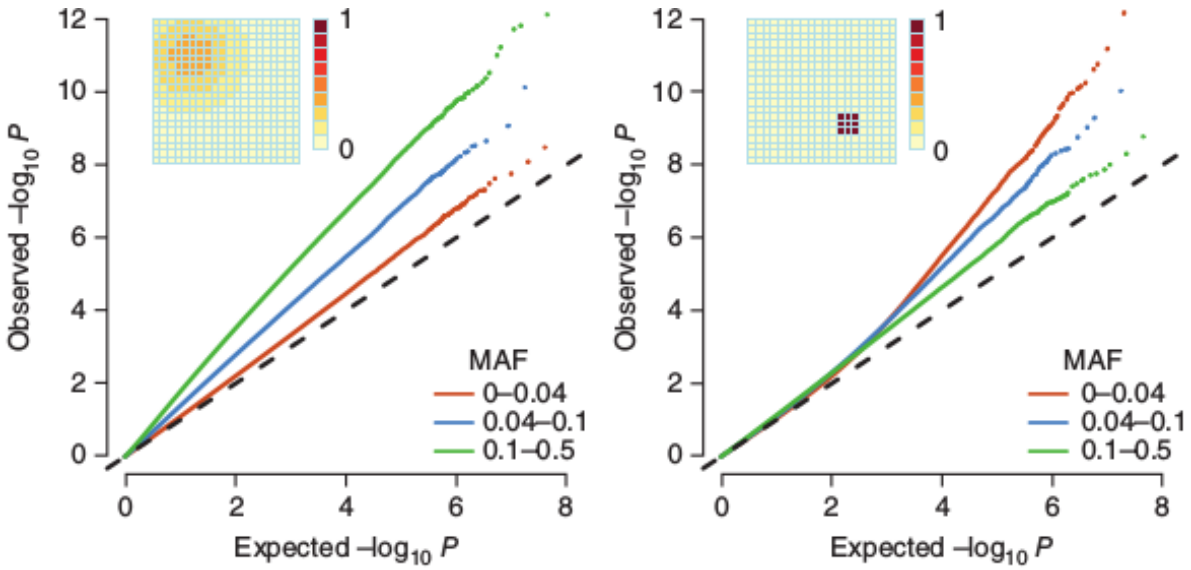
Dan Schlauch, PhD Candidate

Department of Biostatistics
Harvard School of Public Health

November 16, 2015

Outline

Rare allele association inflation



QQ plots of p-values separated by allele frequency. Comparing two types of non-genetic risk distributions. **Mathieson, Nature Genetics 2012**

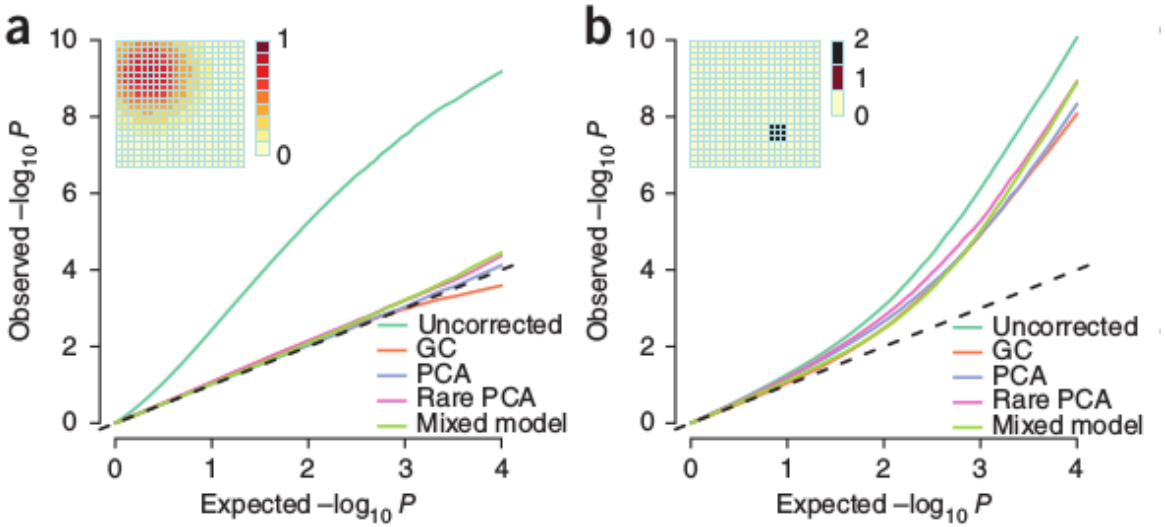
Rare allele association inflation

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 use linear functions to define risk and do not distinguish a tree-like ancestry particularly well.
   - 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 overdispersion.
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:

01110001010**01**101011 - .5
11100010111**01**100110 - .5
00000010101**10**001000 - .25
01010010101**00**000000 - .25
00010100010**01**000010 - .25
00000000000**00**000000 - 0
00000000000**11**000000 - .1
01100000000**00**001010 - .2
00000100010**10**001000 - .2

$$cov\left(\mathbf{G}_i, \mathbf{G}_j\right) = .28$$

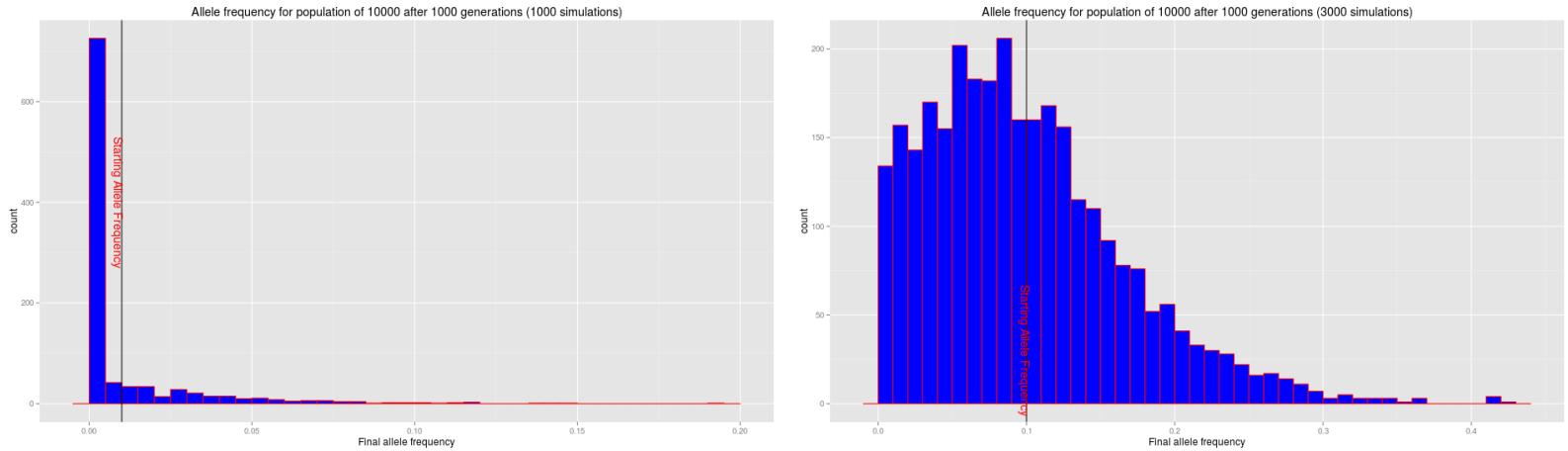
$$cov\left(\mathbf{G}_i^{(-6)}, \mathbf{G}_j^{(-6)}\right) = 0$$

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, $d_{i,j}$

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

where

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

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

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

Genetic Similarity Measure

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

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

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

$$> 1$$

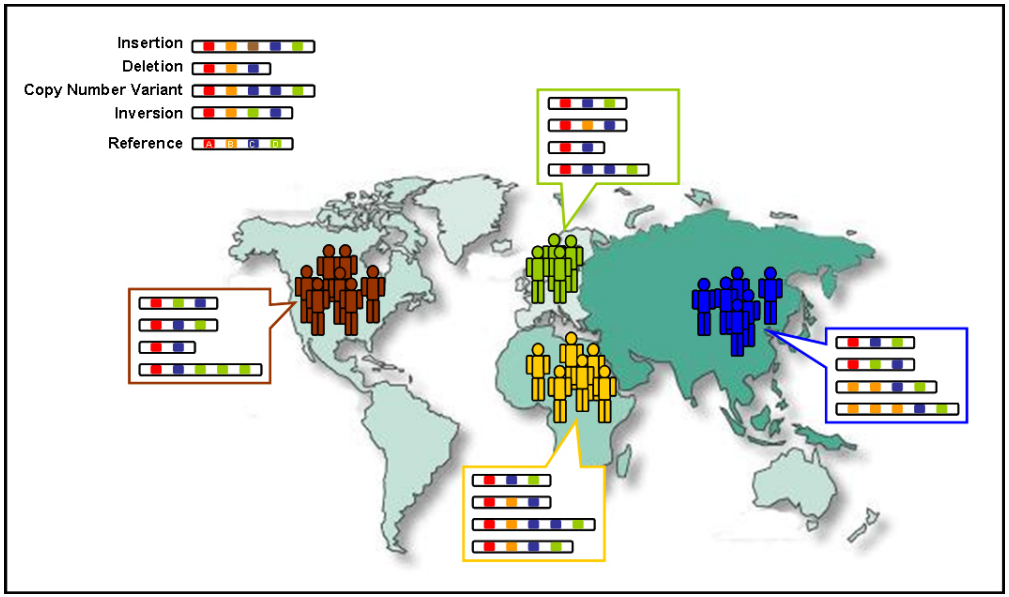
e.g.

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

1000 Genomes Project

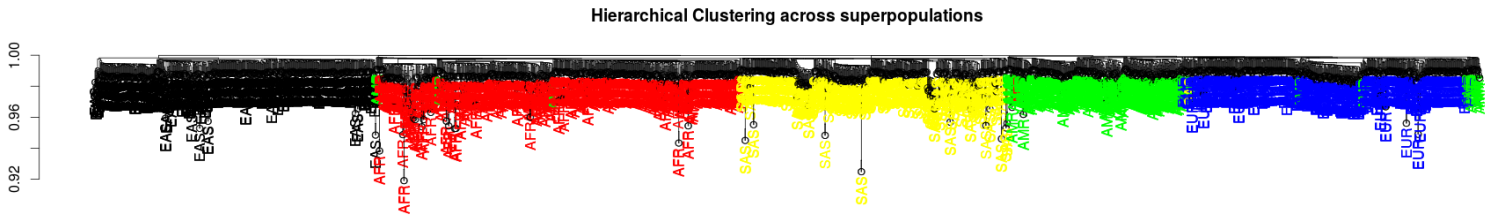
1000 Genomes Project dataset

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

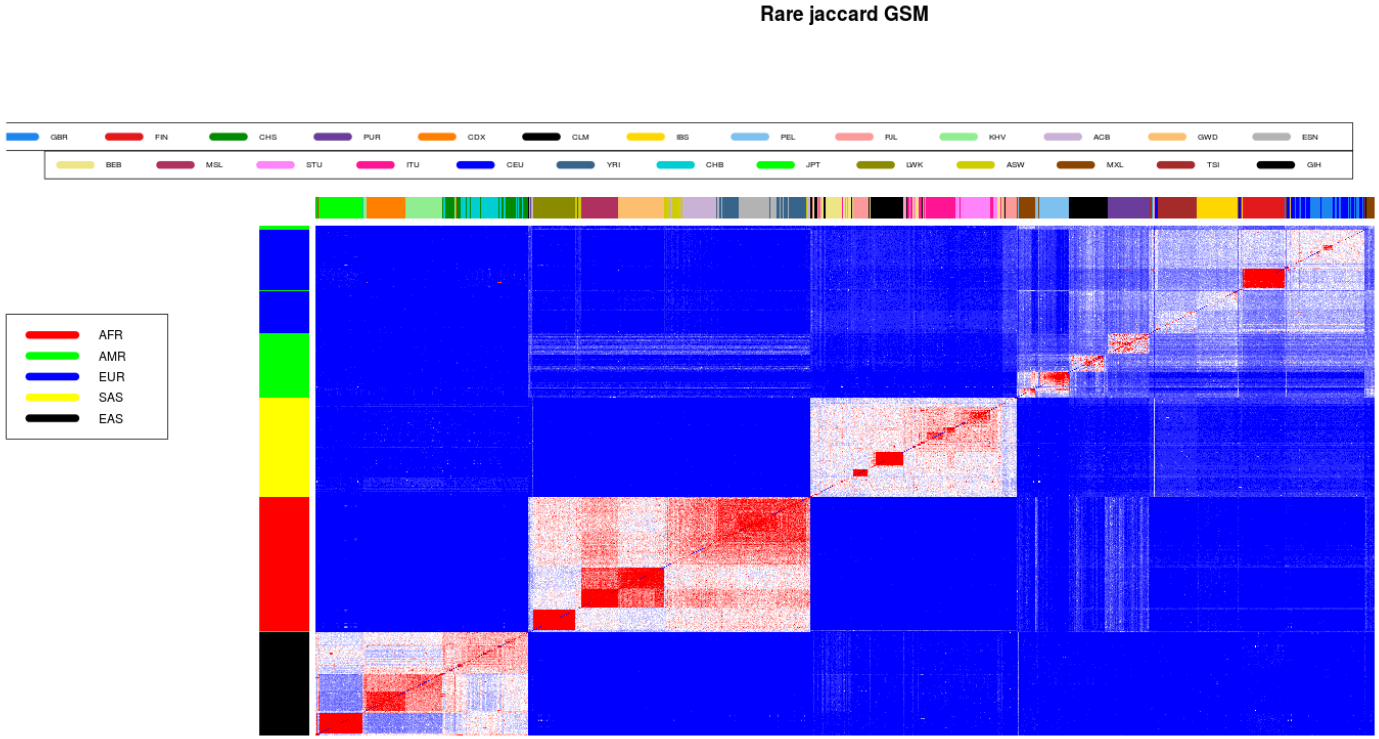


Separation of all individuals

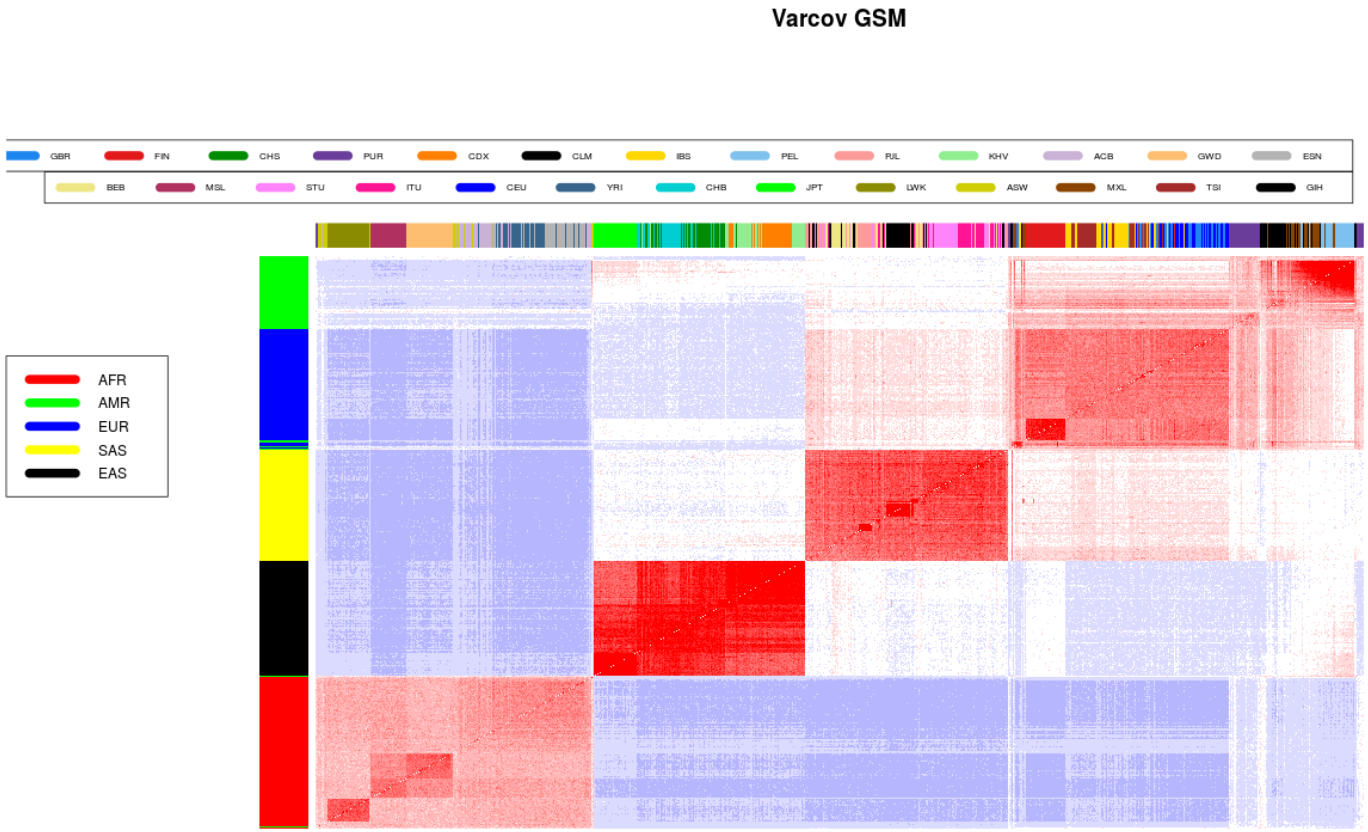
Hierachical clustering of all 2504 samples by jaccard similarity vs covariance



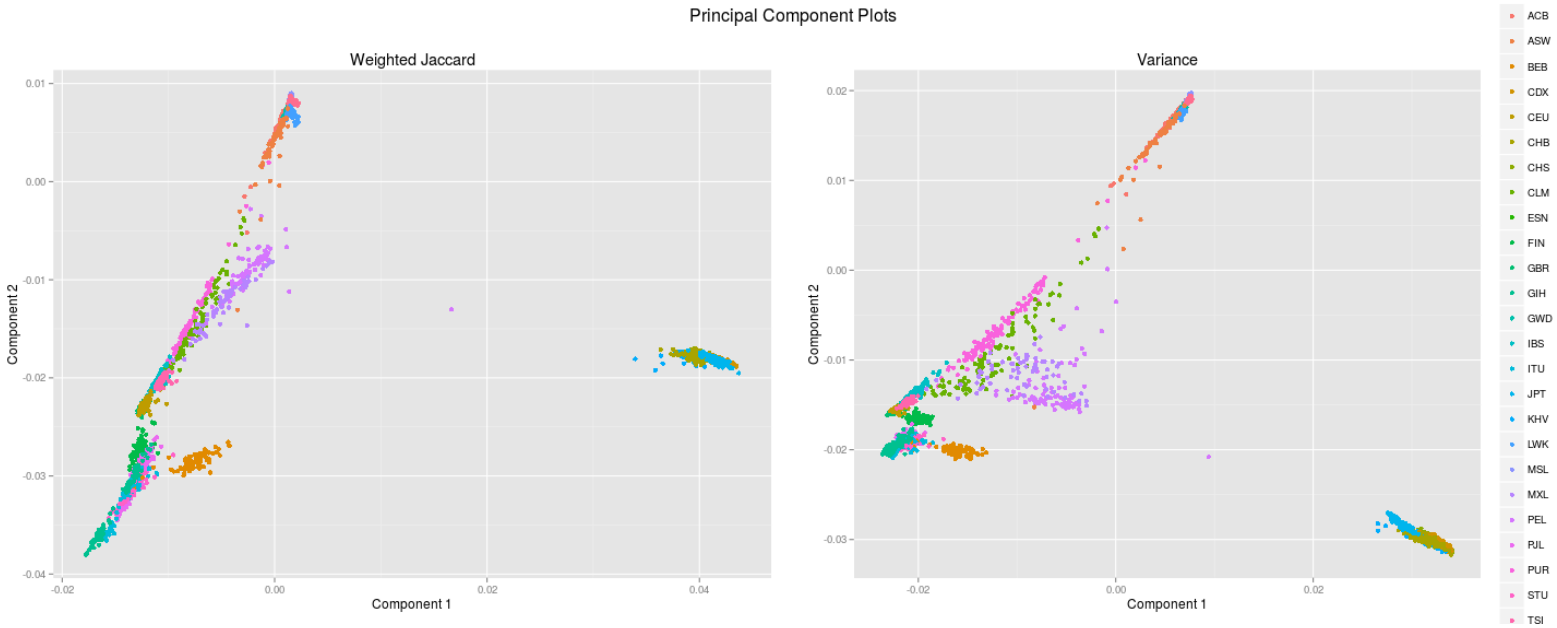
Separation of all individuals



Separation of all individuals

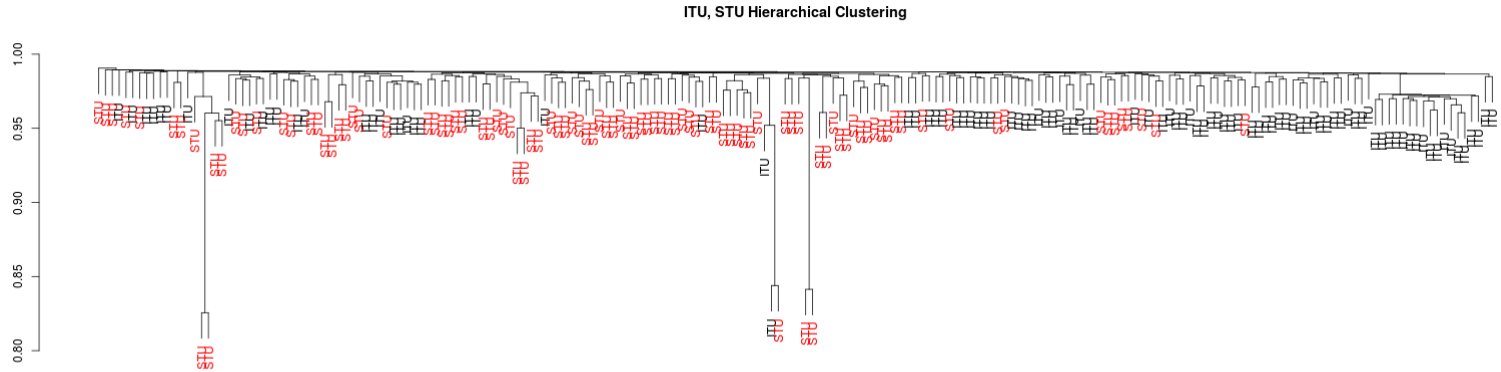


Separation of all individuals



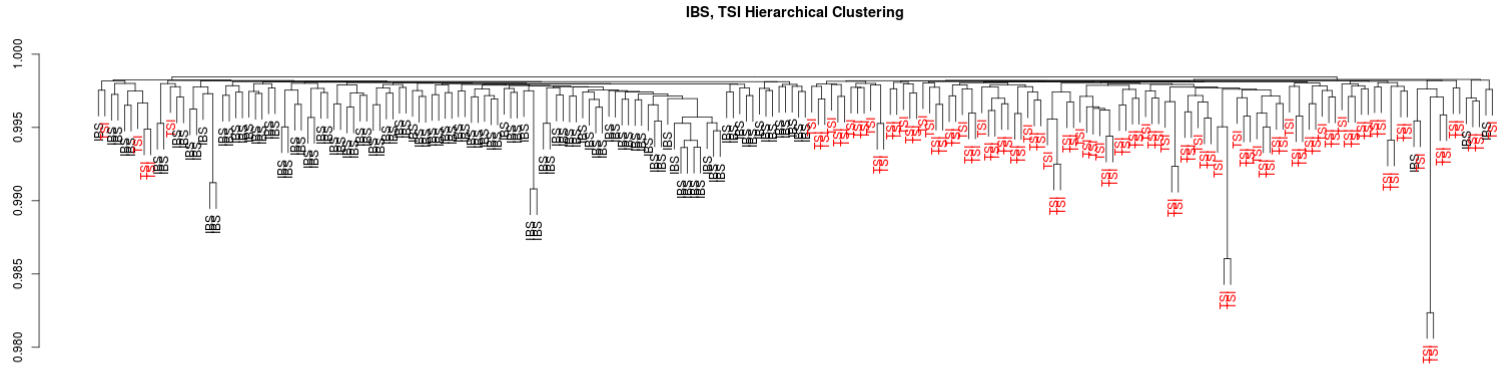
Separation of recent shared ancestries

Example:
Indian Telugu from the UK
Sri Lankan Tamil from the UK



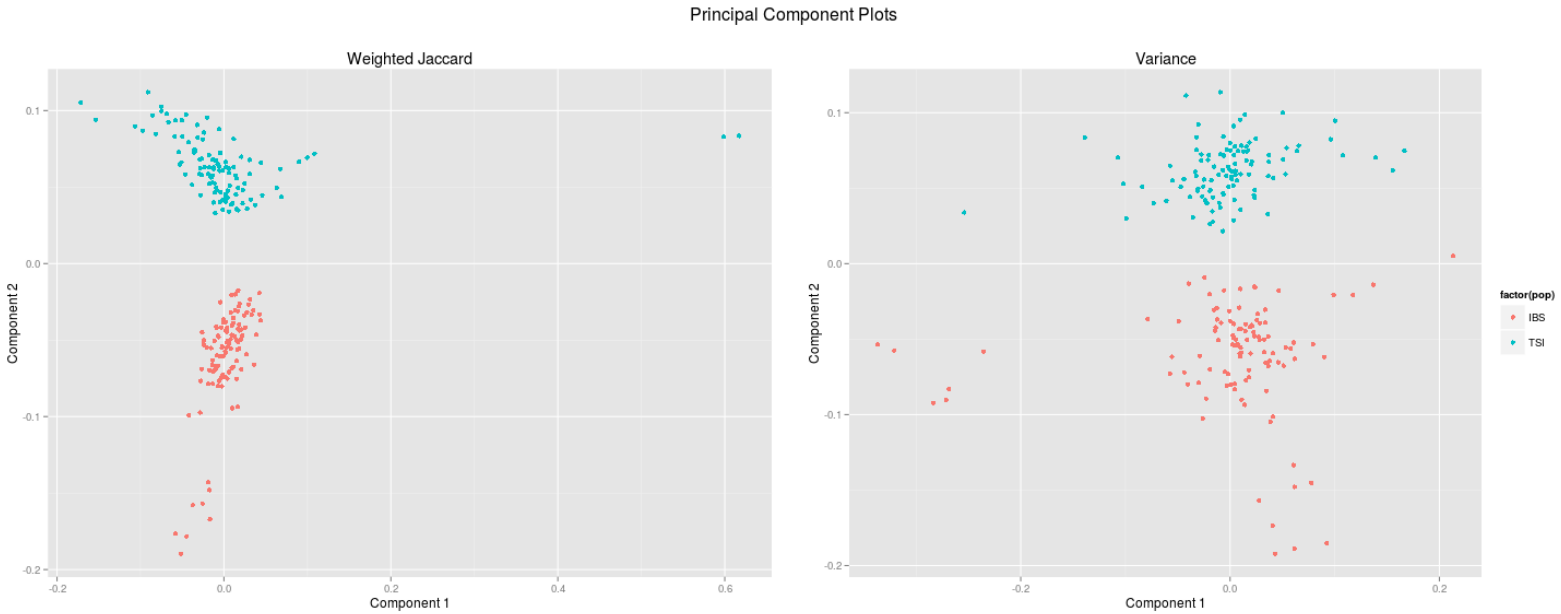
Separation of recent shared ancestries

Example:
Iberian Population in Spain
Toscani in Italia



Separation of recent shared ancestries

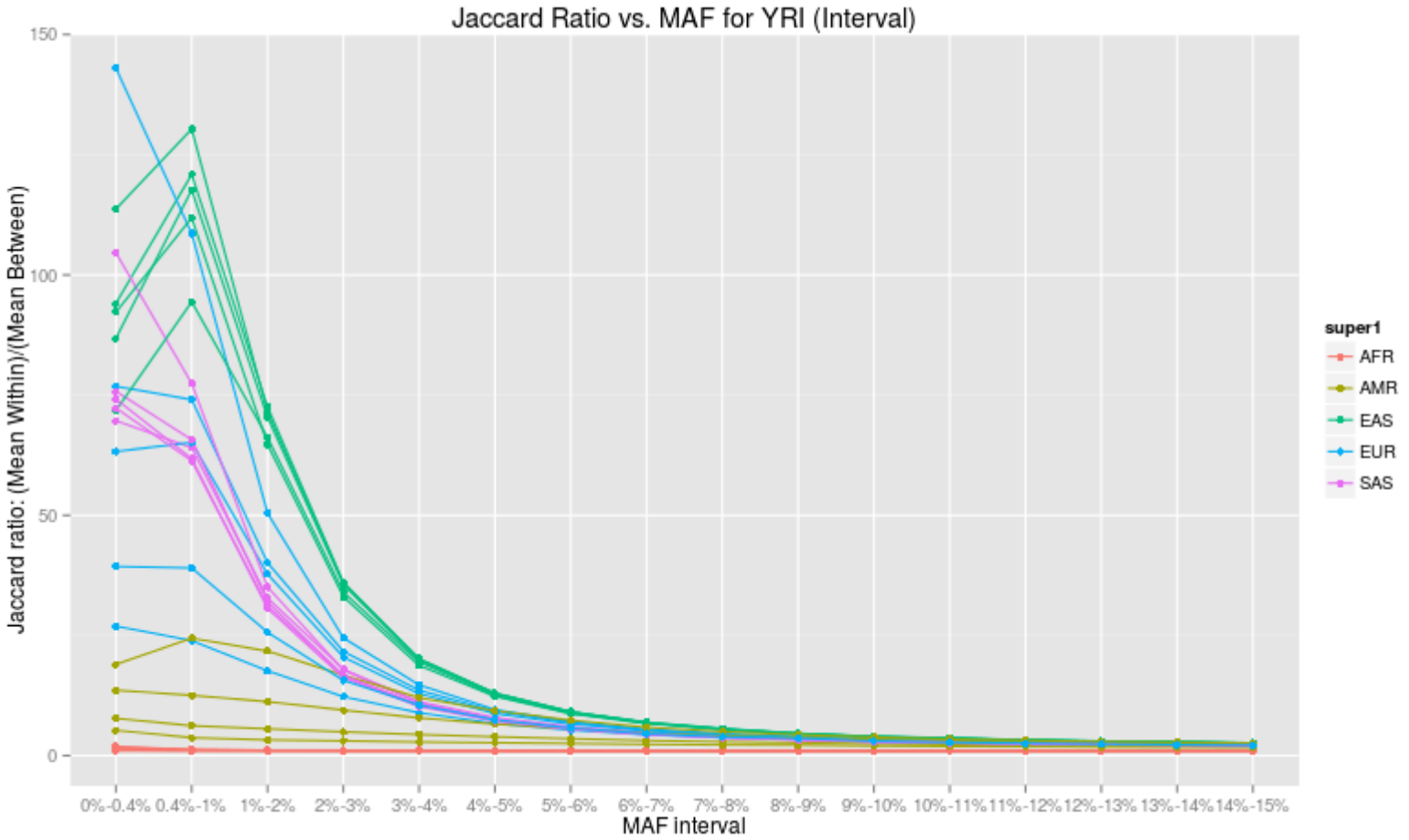
Example: ITU vs STU



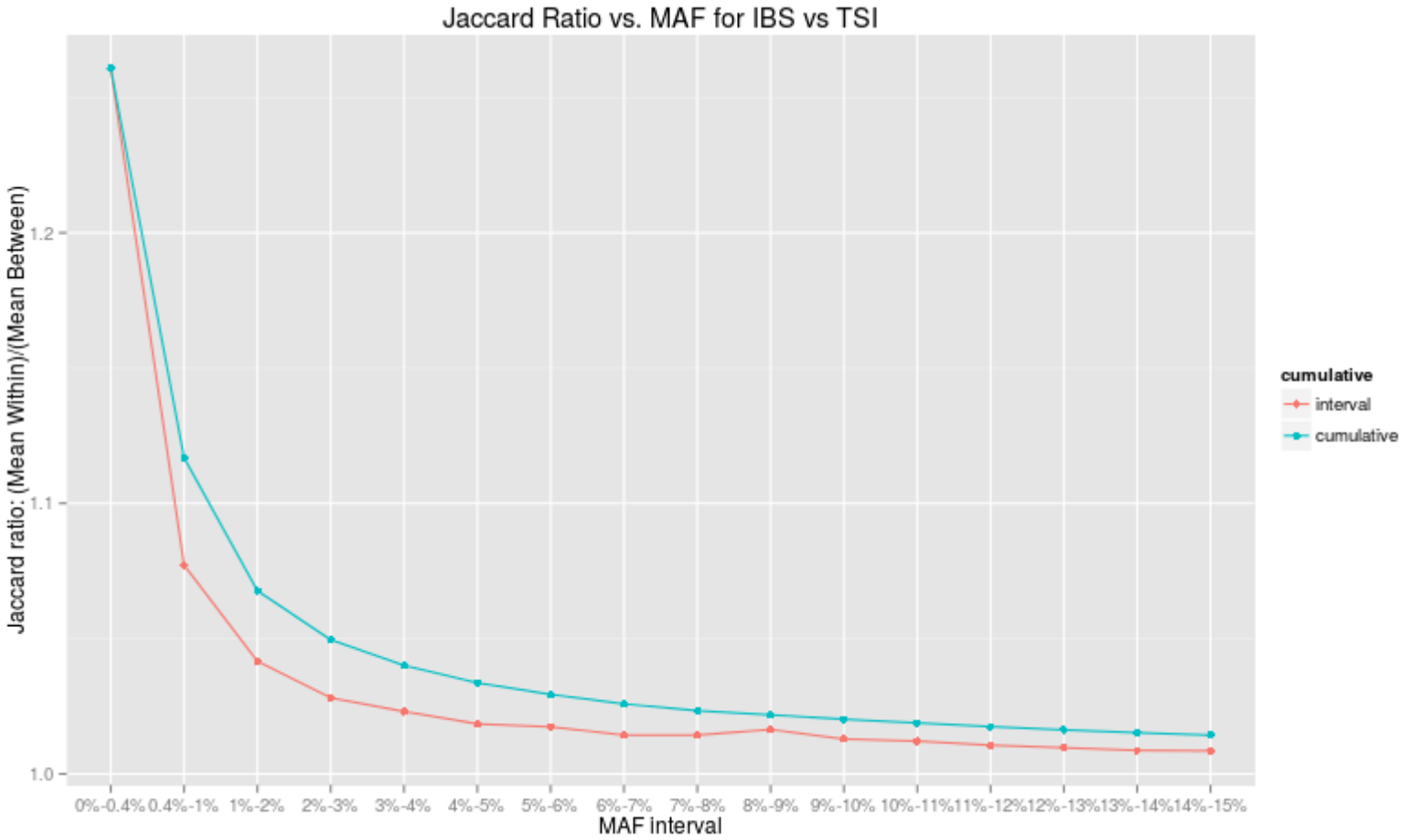
Ratio of within-group mean distance to out-of group mean distance:

	TSI-IBS	ITU-STU	CHB-CHS	LWK-ESN	GIH-ITU
wJaccard	.417	.836	.605	.178	.513
Variance	.504	.889	.681	.197	.552

Separation as a function of allele frequency



Separation as a function of allele frequency



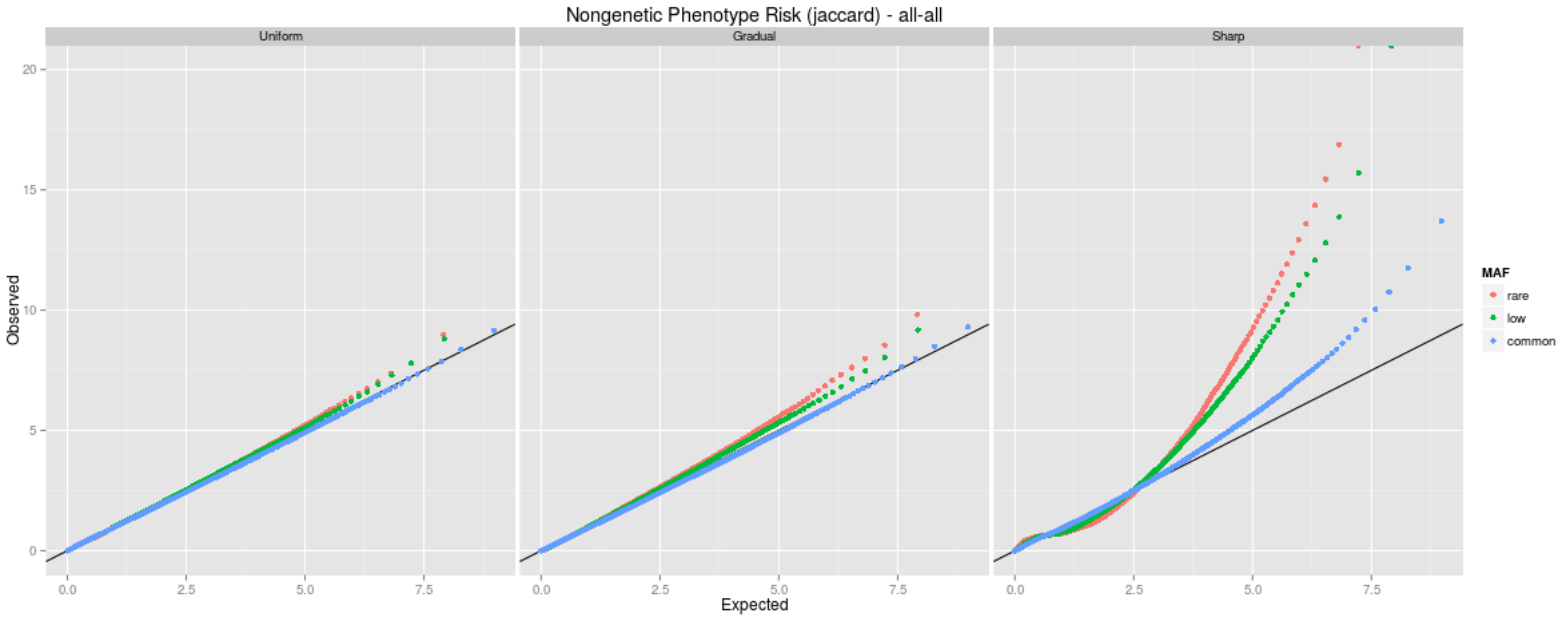
Simulated non-genetic phenotypes with 1000GP genotypes

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 of weighted Jaccard preserve type I error?

Weighted-Jaccard-corrected association Q-Q plot

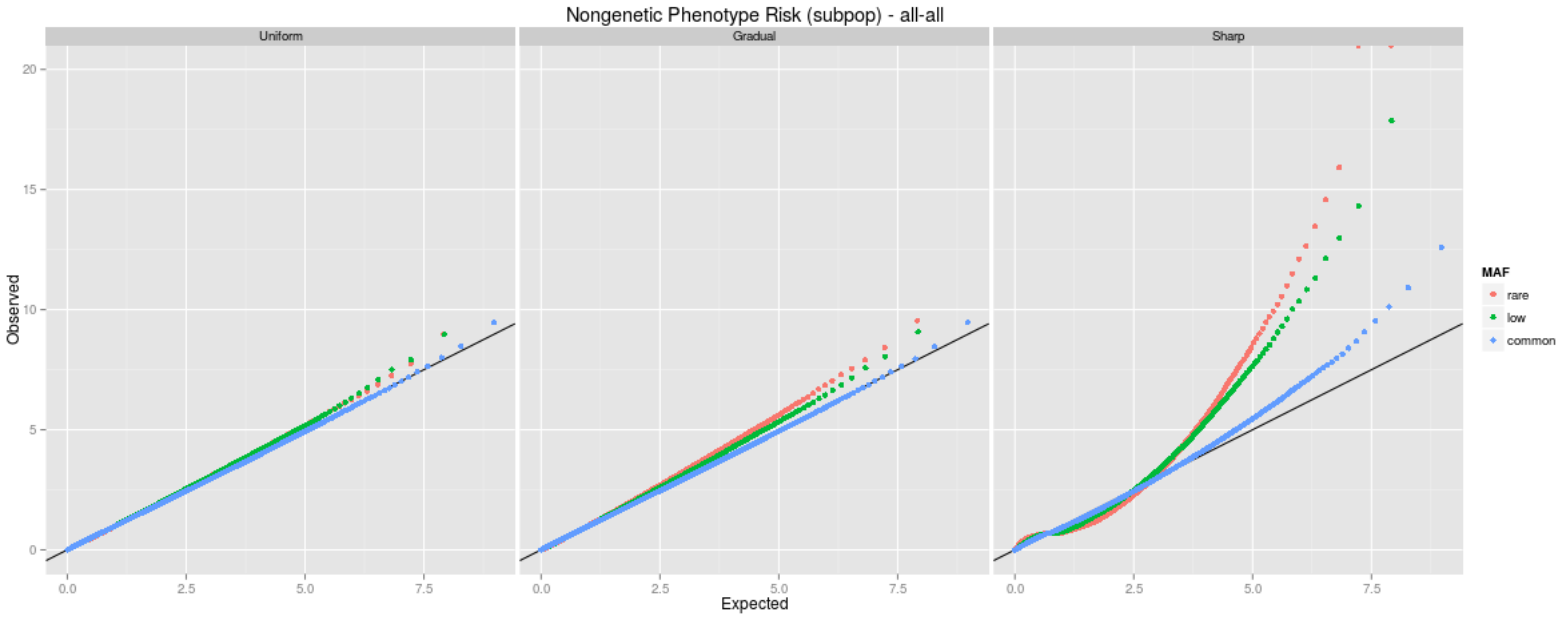


No, inflation is still present.

Does perfect population stratification identification control the type I error?

Use of population labels to control for stratification

Subpopulation-corrected association Q-Q plot



Type I error inflation *still* occurs when controlling for groups which define disease risk.

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

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

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:

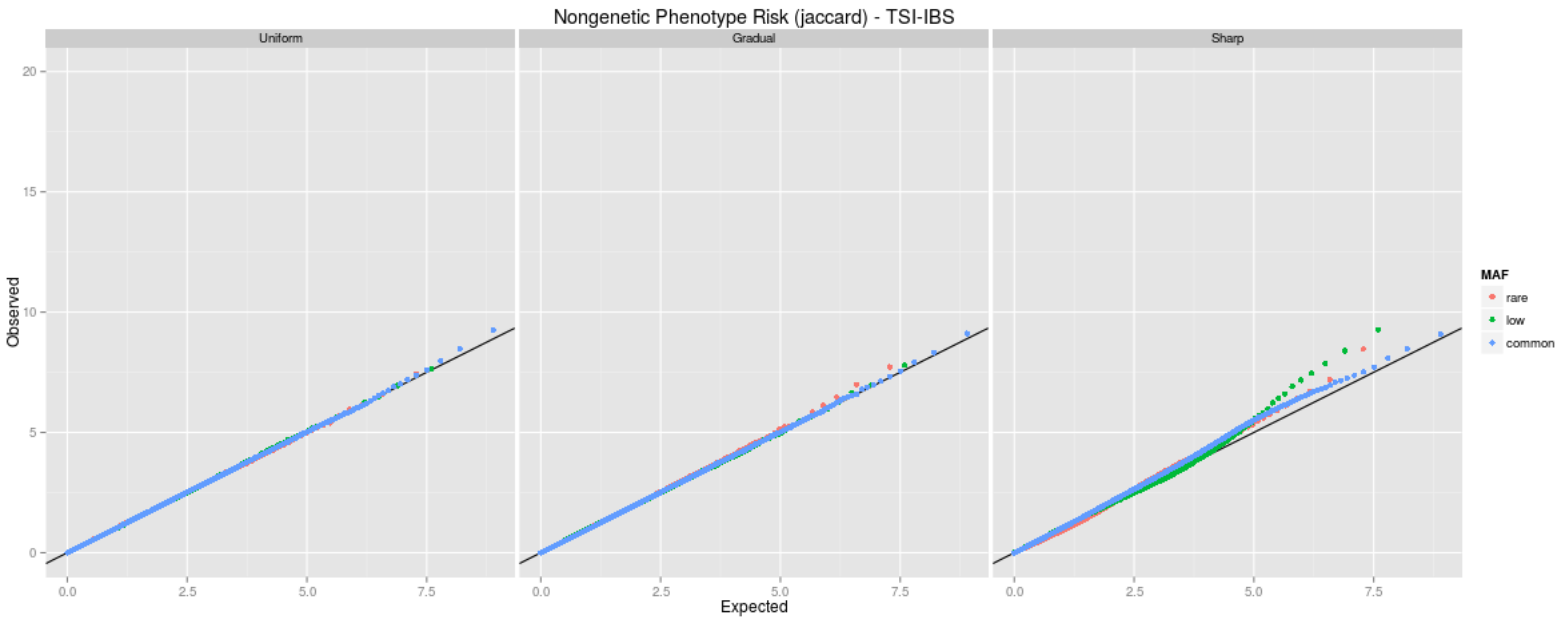
$$s_k = w \times n \times \left(\frac{\mathbf{r}_P \mathbf{r}_G^T}{\sqrt{\sum_i^n \mathbf{r}_{P,k,i}^2 \sum_i^n \mathbf{r}_{G,k,i}^2}} \right)^2 \sim \chi_1^2$$

Where

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

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

Applying corrected estimator to 1000GP data



Acknowledgements

Thanks to:

- Christoph Lange
- Matt Goodman