Inflation issues

Distribution of **non-genetic**, **non-stratified** armitage trend test is not chi-sq(1) for rare alleles.

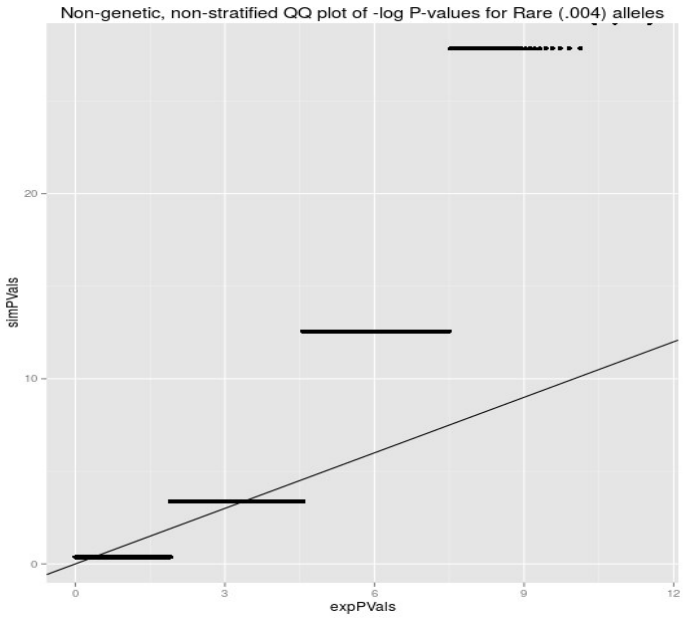
E(ATT)=1

Var(ATT)!=2

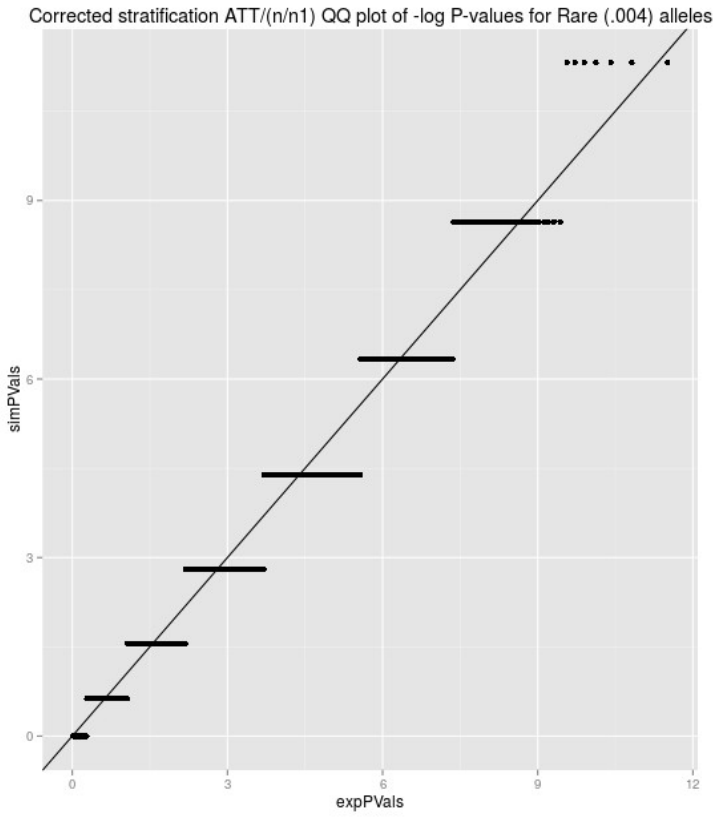
For example, with no population stratification:
MAF=.004, PhenoFreq=.01

mean(ATT)=.997

Var(ATT)=3.1



Additionally, even after a perfect correction for stratification, for a genotype and phenotype that appear in only one subpopulation, for a non-genetic risk, ATT~n/n1*chi-sq(1)



Individual SNP variance inflation

So we have a variance inflation factor of 1/n1 for that particular SNP.

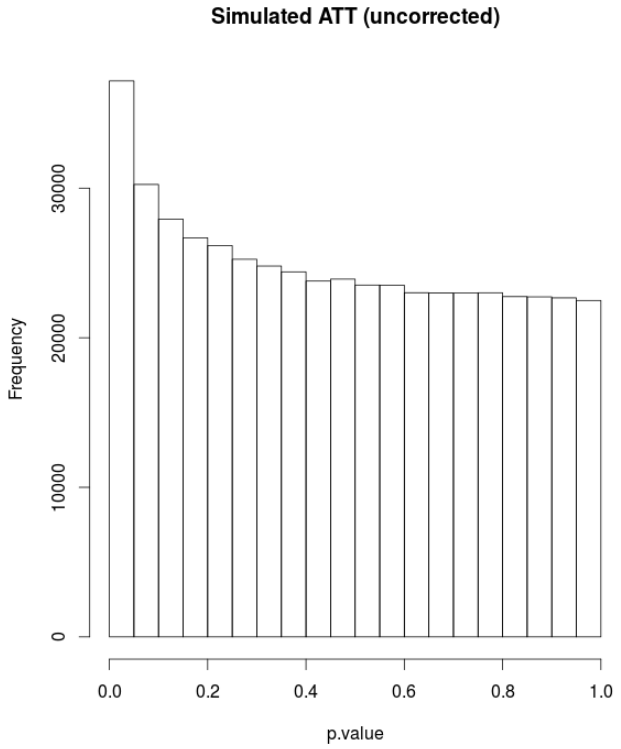
Generalizing this we have

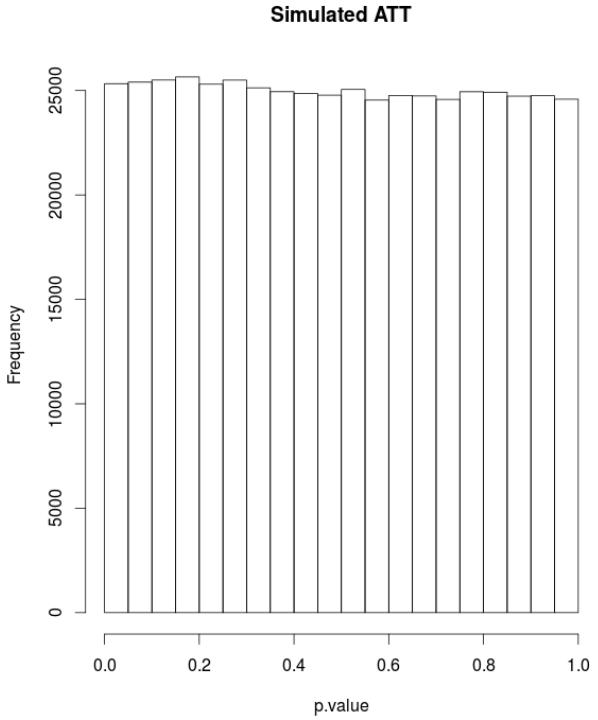
$$
\begin{aligned}
VIF_i &= \frac{\left[\sum_{k=1}^{N} Leverage_{i,k}\right]^2}{\sum_{j=1}^{N}\left[Leverage_{i,j}^2\right]} \\
Leverage_{i,j} &= Var\left(geno_{i,j}\right) \times Var\left(pheno_{i,j}\right)
\end{aligned}
$$

Where the leverage is the product of variance of genotype(i,j) and the variance of phenotype(i,j).
The variances are found by p(1-p) where p is the fitted value from the population correction.

Simulated variance inflation correction

```
z <- (1:100)/200
zvar <- z*(1-z)
sumviSq <- sum(zvar)^2
sumSqvi <- sum(zvar^2)
vFactor <- 1/(sumSqvi/sumviSq)
hist(replicate(500000, {
    x <- rbinom(100,1,prob=z)-z
    y <- rbinom(100,1,prob=z)-z
    1-pchisq(vFactor*cor(x,y)^2, 1)
    vFactor*cor(x,y)^2
}))
```





Phenotypes

- Binary
- 3 types
  - Uniformly random phenotype
  - Gradual phenotype risk determined by superpop and subpop membership
    - X~Uniform(0,.5) for each superpop
    - Y~Uniform(0,.2) for each subpop
    - Risk=X+Y
  - Sharp phenotype risk
    - Risk=I[randomly chosen subgroup]*.2

Corrections

- 4 types
  - Uncorrected
  - Superpop
    - Take residuals of regression of genotype and phenotype on superpopulation matrix
  - Superpop
    - Take residuals of regression of genotype and phenotype on subpopulation matrix
  - Jaccard
    - Take residuals of regression of genotype and phenotype on top X eigenvectors of jaccard matrix

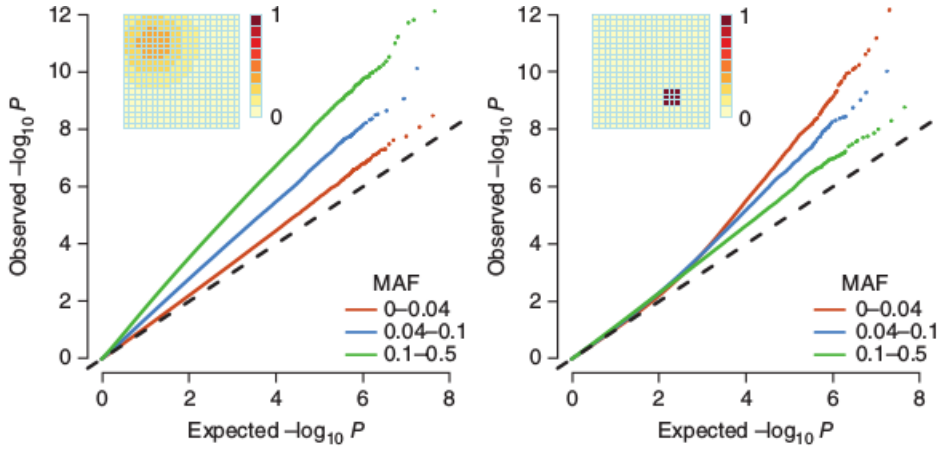
Simulations

- Simulated 100 of each type of phenotype (300 total
- Used each of the correction methods to get corrected phenotypes for each of the 300.
- Read in each SNP and corrected it.
- Calculated “variance-factor” for each phenotype-genotype pair.
- Calculated R^2 for each phenotype-genotype pair.
- Determined test statistic for each phenotype-genotype pair.
- Plotted observed distribution against expected: chisq(1)

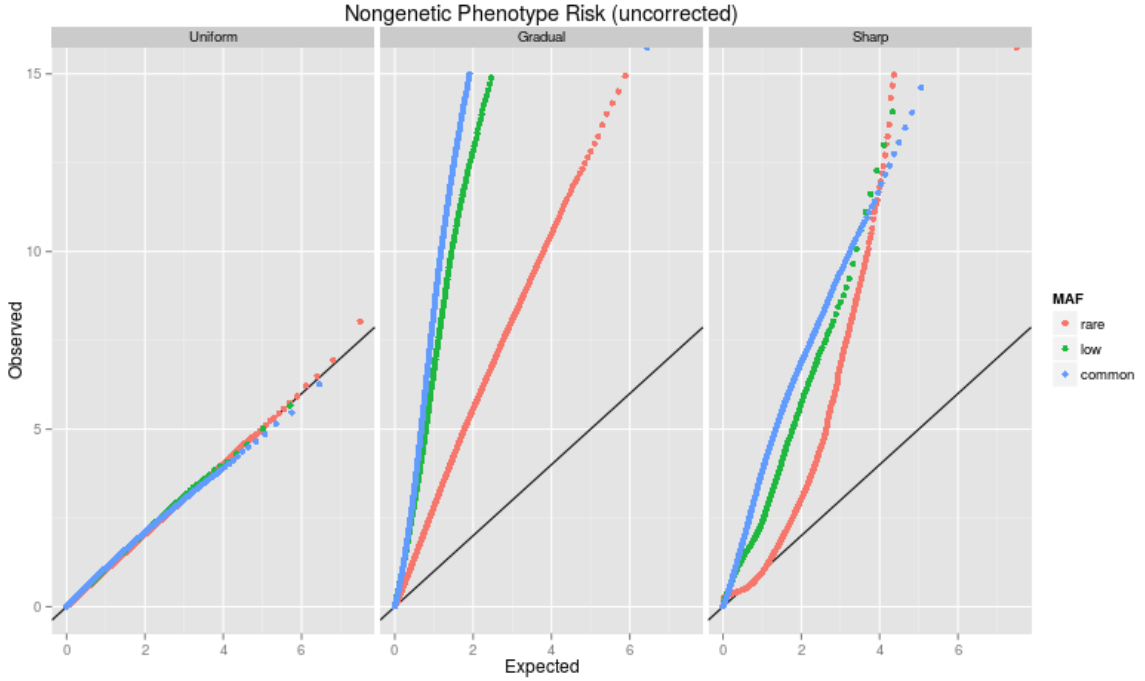
  Note: All simulations were run on subsets of ~100k-400k loci on a single chromosome in the interest of time. Takes about 10 minutes per 200k loci. May take a long time to do whole genome.

1000GP data -Uncorrected variance inflation

Mathieson

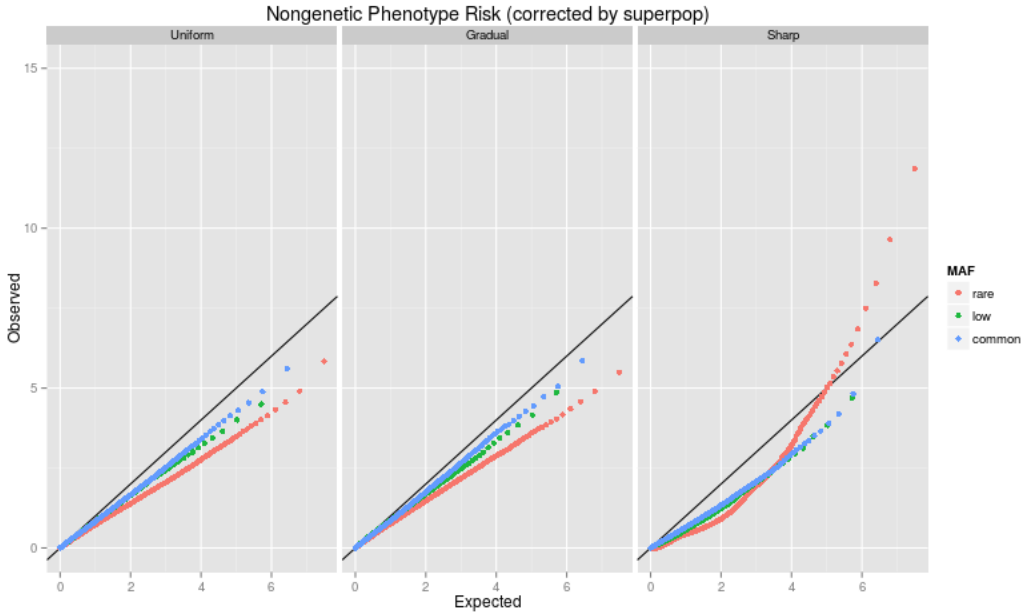


1000GP

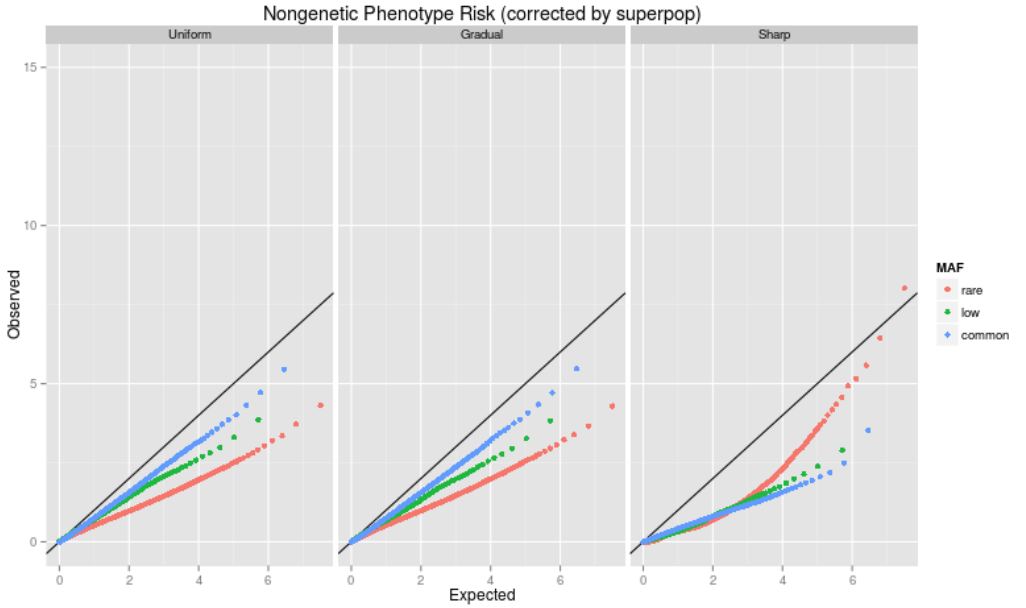


1000GP data - Corrected by Superpop label

ATT

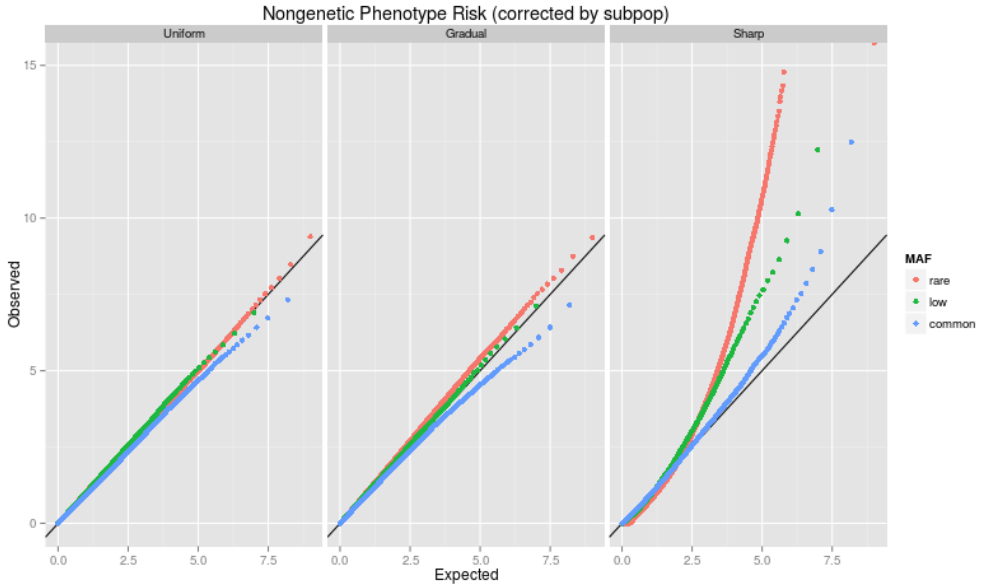


Individual VIF

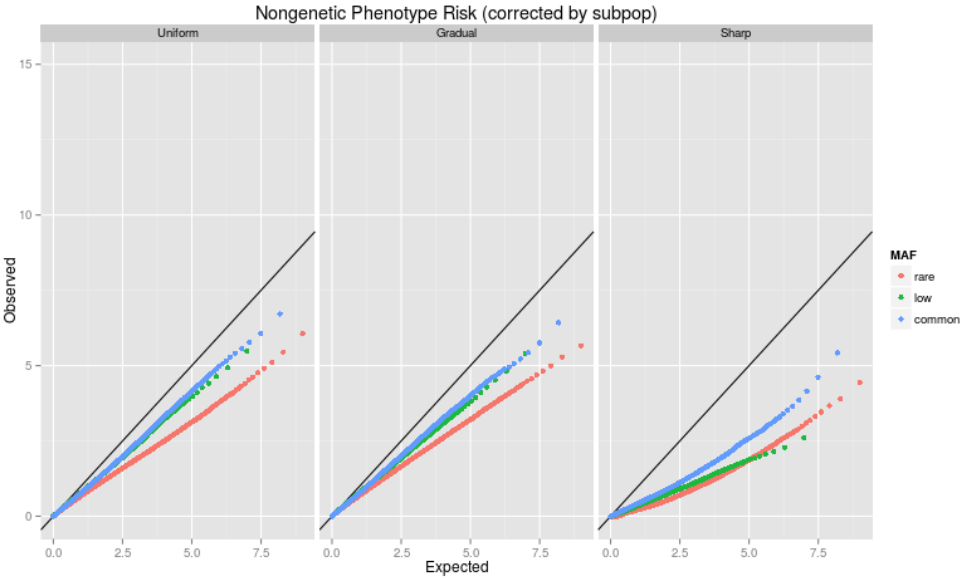


1000GP data - Corrected by Subpop label

ATT

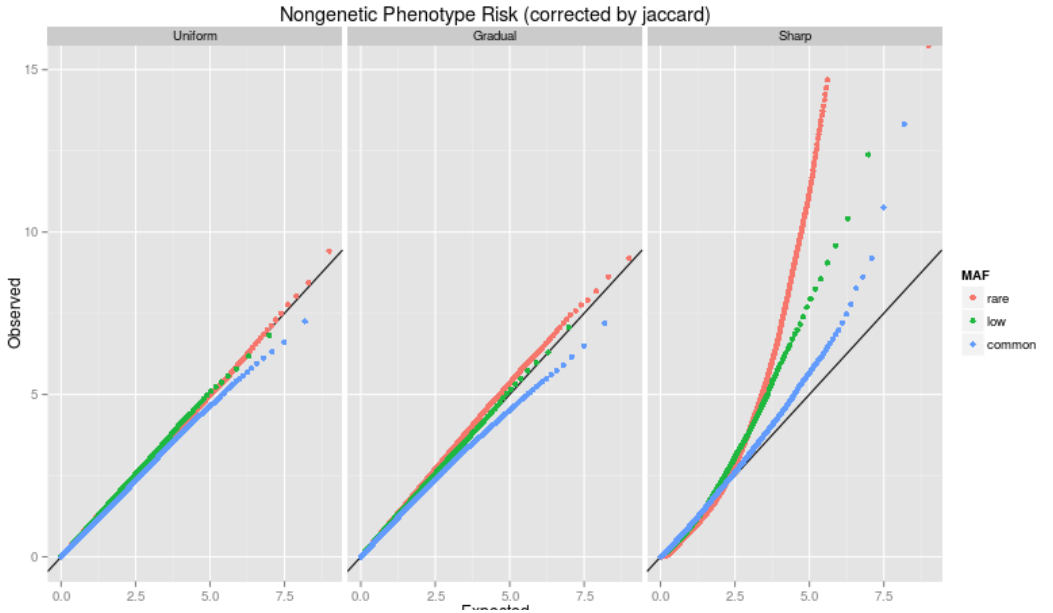


Individual VIF

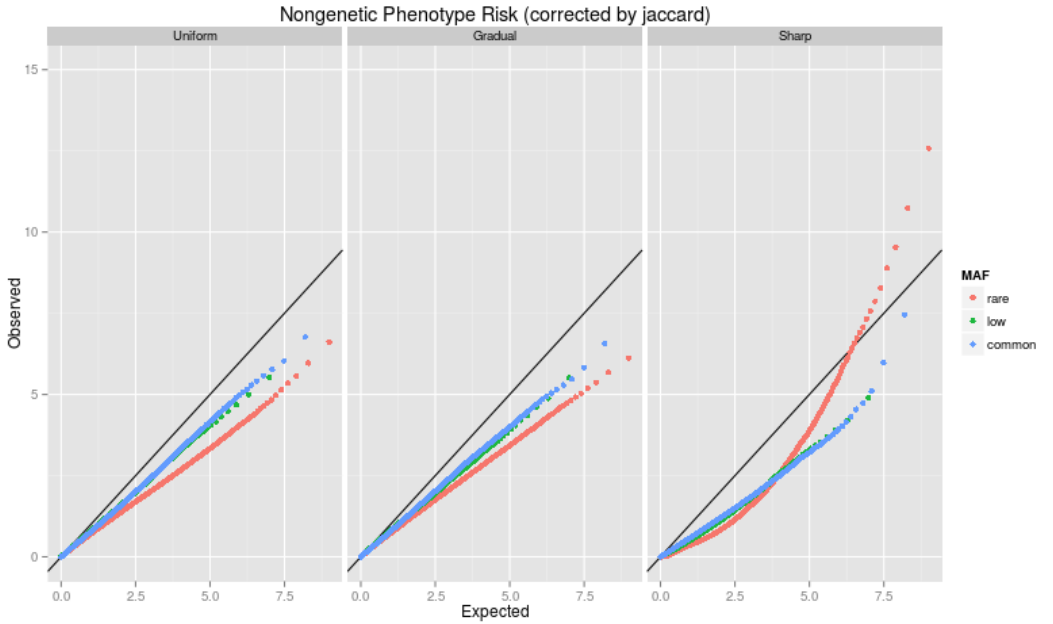


1000GP data - Corrected by Jaccard eigenvectors(10)

ATT

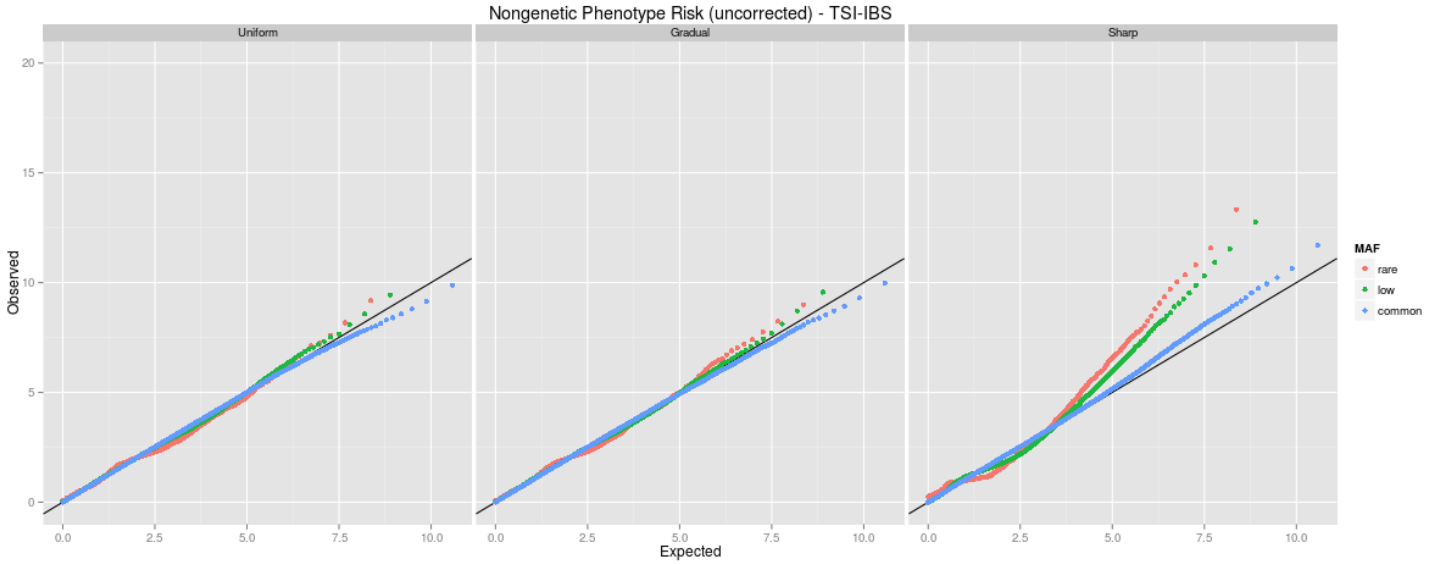


Individual VIF



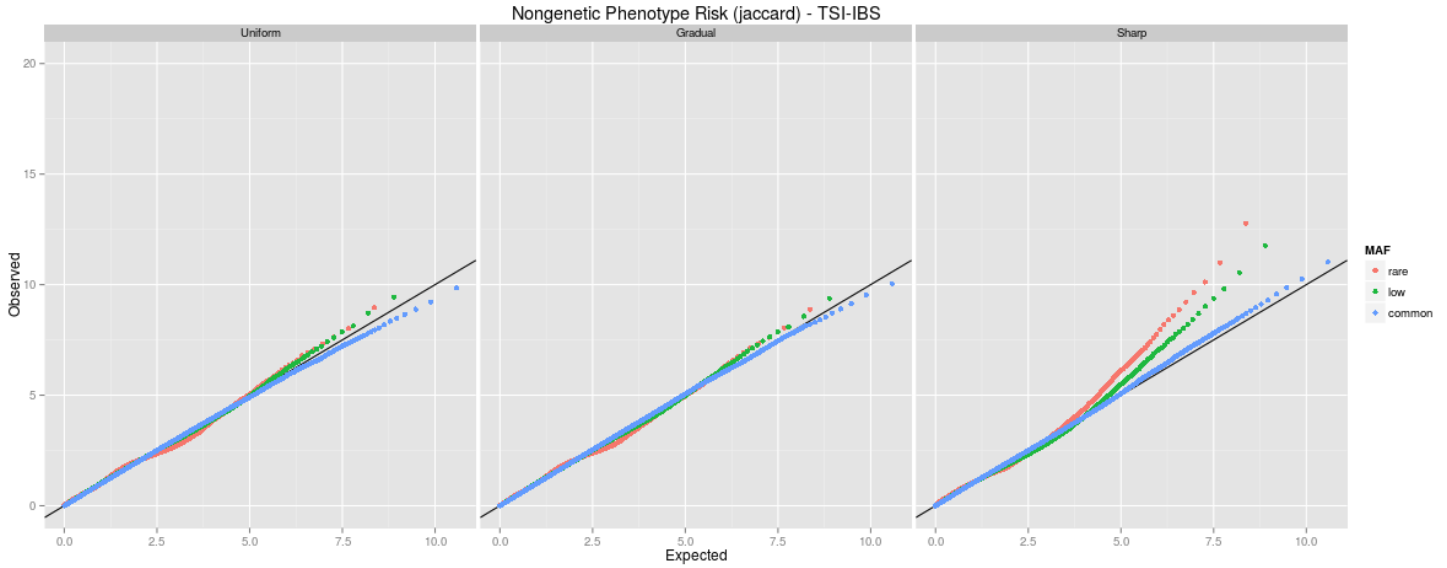
TSI vs IBS uncorrected

ATT

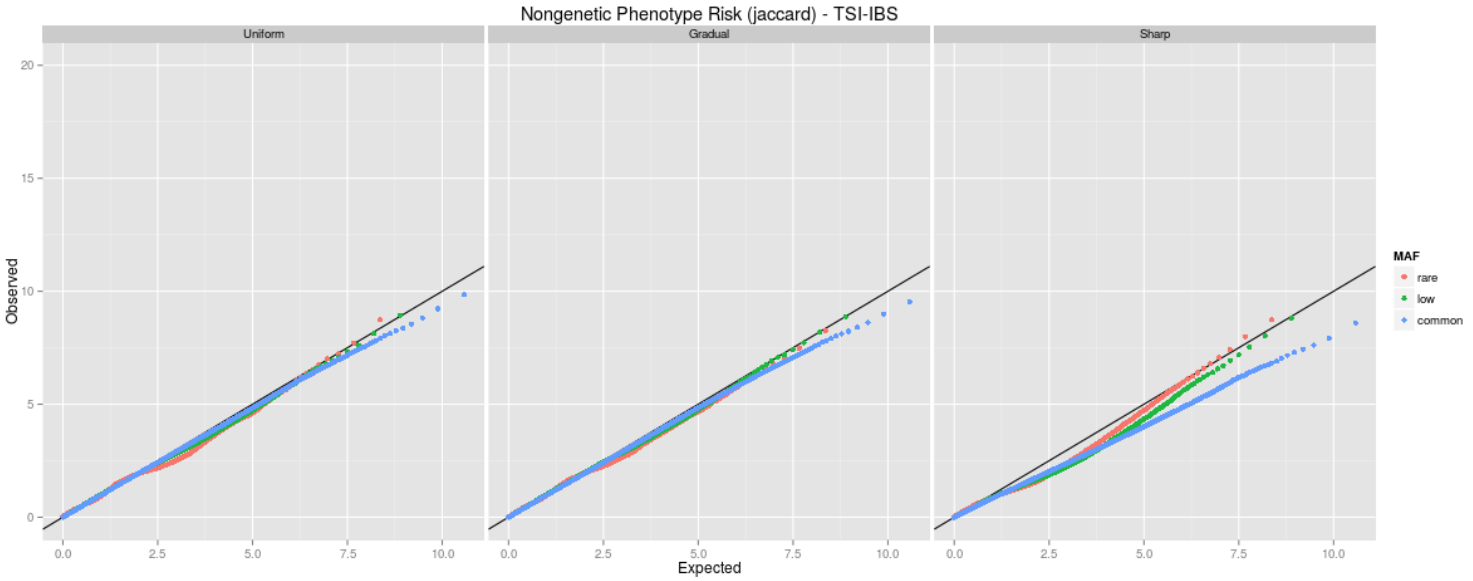


TSI vs IBS corrected by Jaccard eigenvectors(2)

ATT



Individual VIF



What we need

- Comparison to PCA
  - This will take awhile unless there is a pre-run result somewhere online (couldn't find one)
- Demonstrate ability to find causal SNPs
  - Next step in simulations