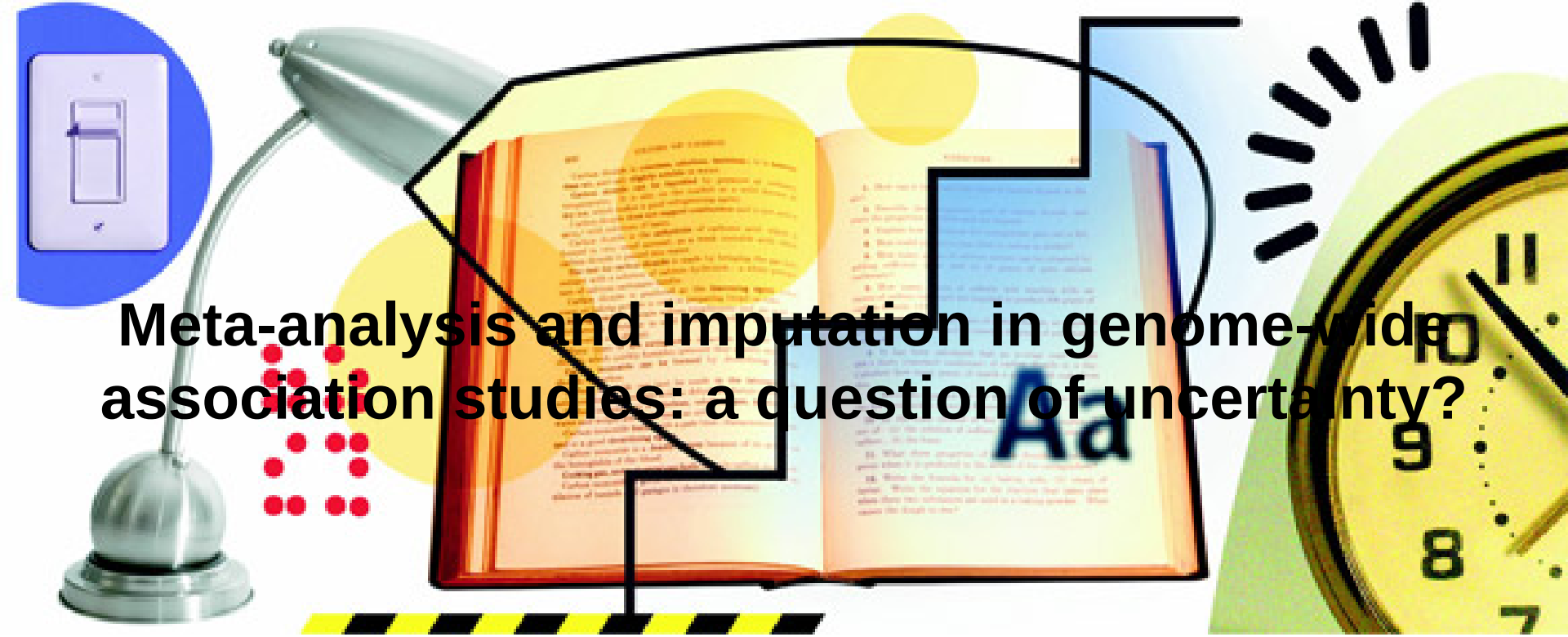


Meta-analysis and imputation in genome-wide association studies: a question of uncertainty?

Paul de Bakker

Assistant Professor of Medicine

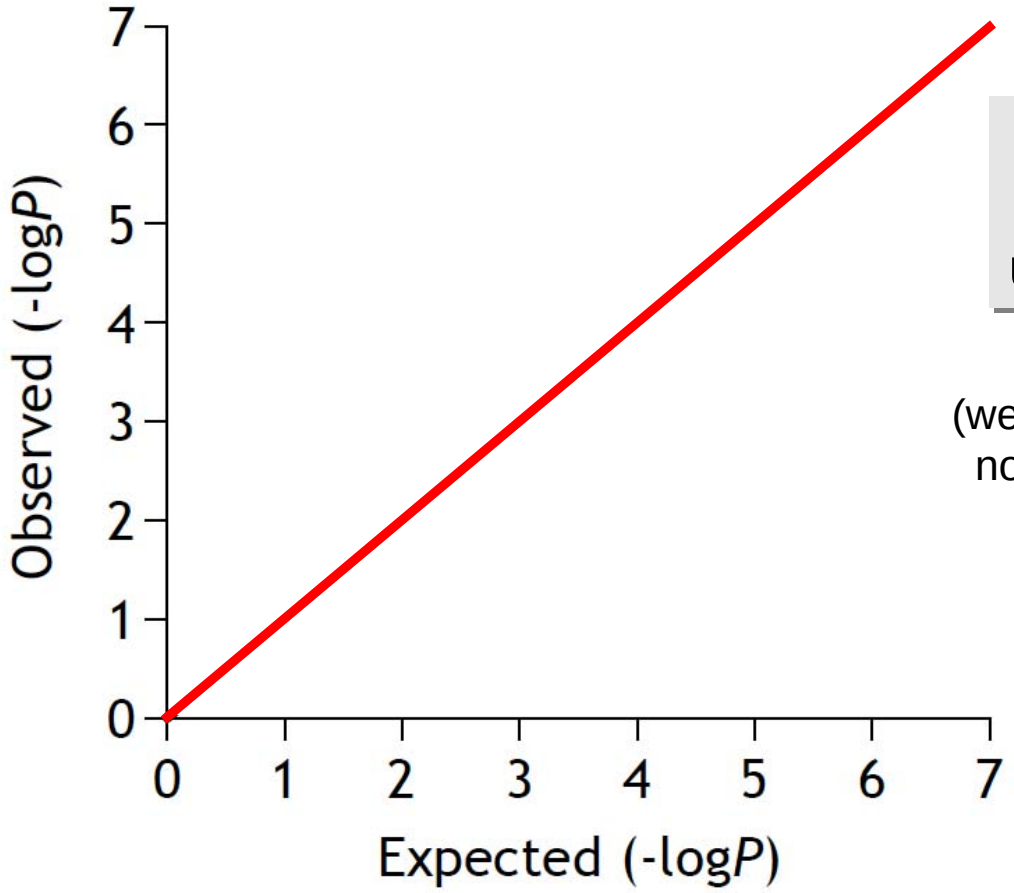
Brigham and Women's Hospital and Harvard Medical School

BRIGHAM AND WOMEN'S HOSPITAL

Combining multiple GWAS

- Rationale: more power
- Challenge is to achieve comparability between individuals studies
  - Need standardized distributions of test statistic
- Distortions can be due to:
  - Population stratification (sample ascertainment)
  - Technical artefacts (e.g. genotyping error, batch effects)
  - Statistical artefacts (e.g. overdispersion of test statistic, imputation)

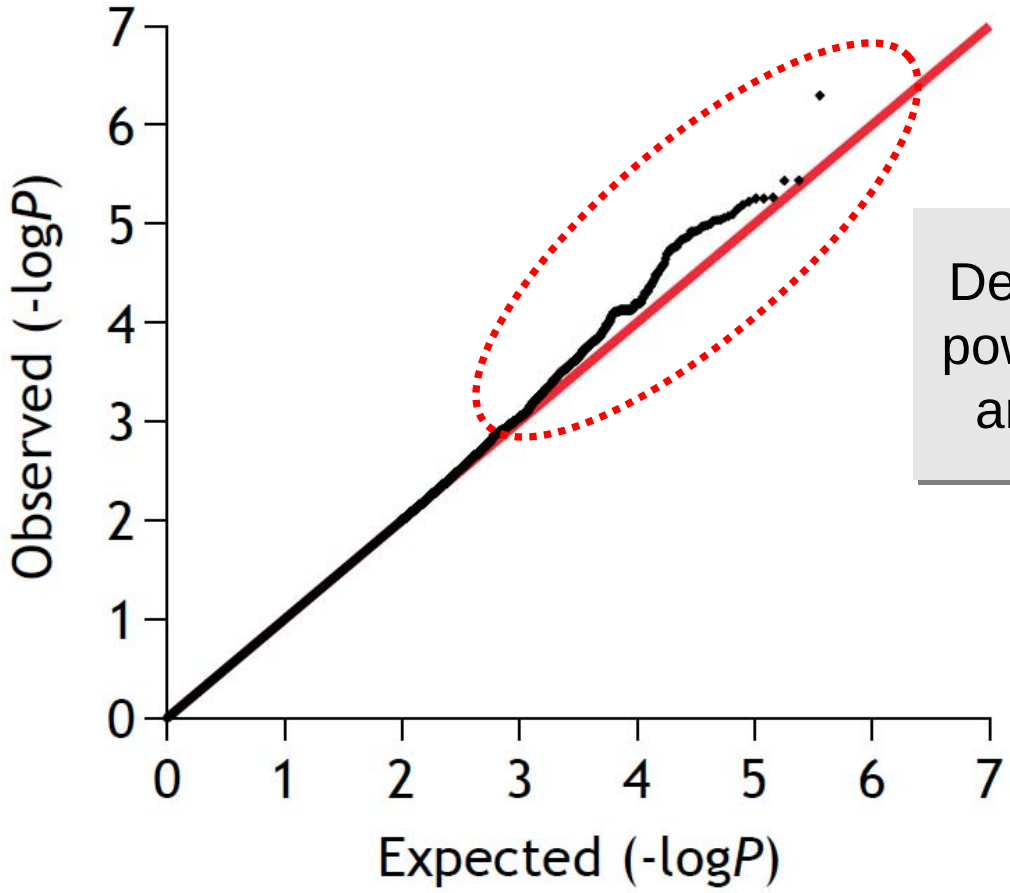
Q-Q plot of the test statistic: expected vs. observed



expected distribution under the null

(we expect most SNPs not to be associated)

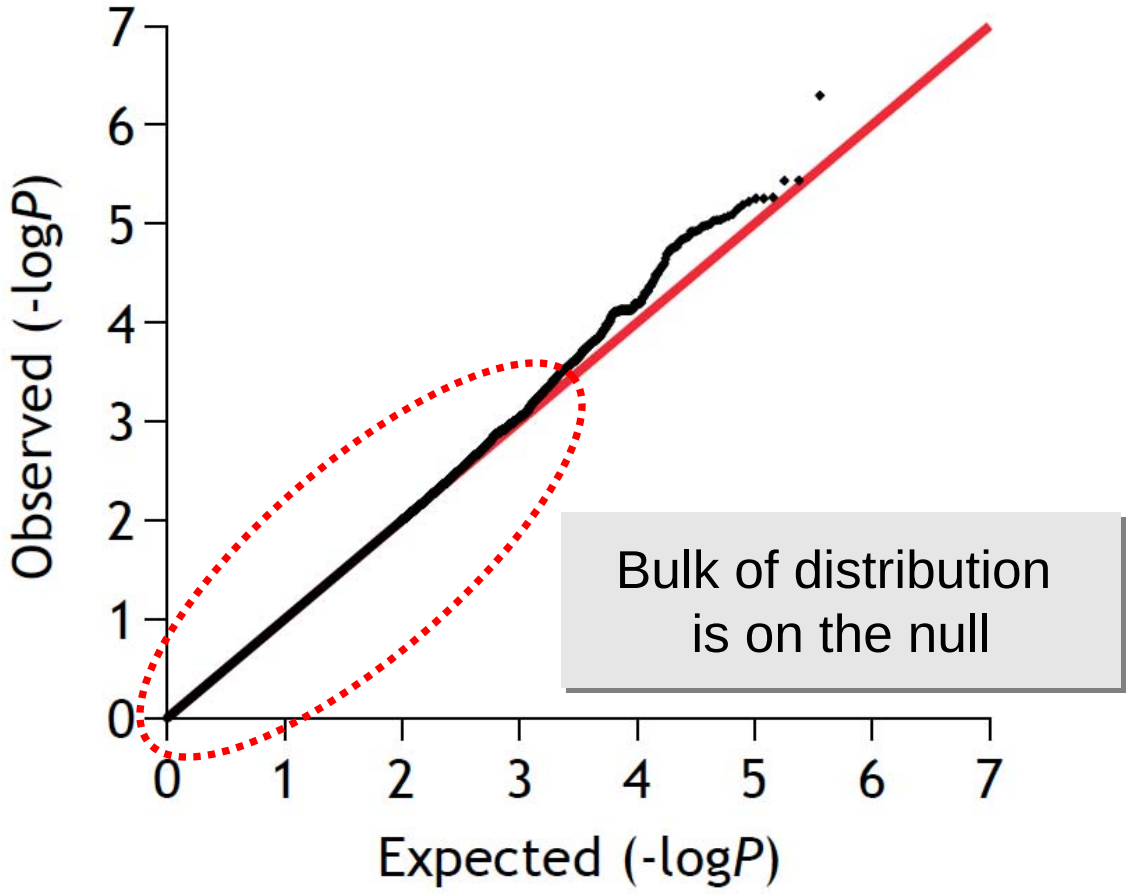
Q-Q plot of the test statistic: expected vs. observed



Depending on study power, true positives are enriched in tail

$\lambda_{GC} = 1.05$

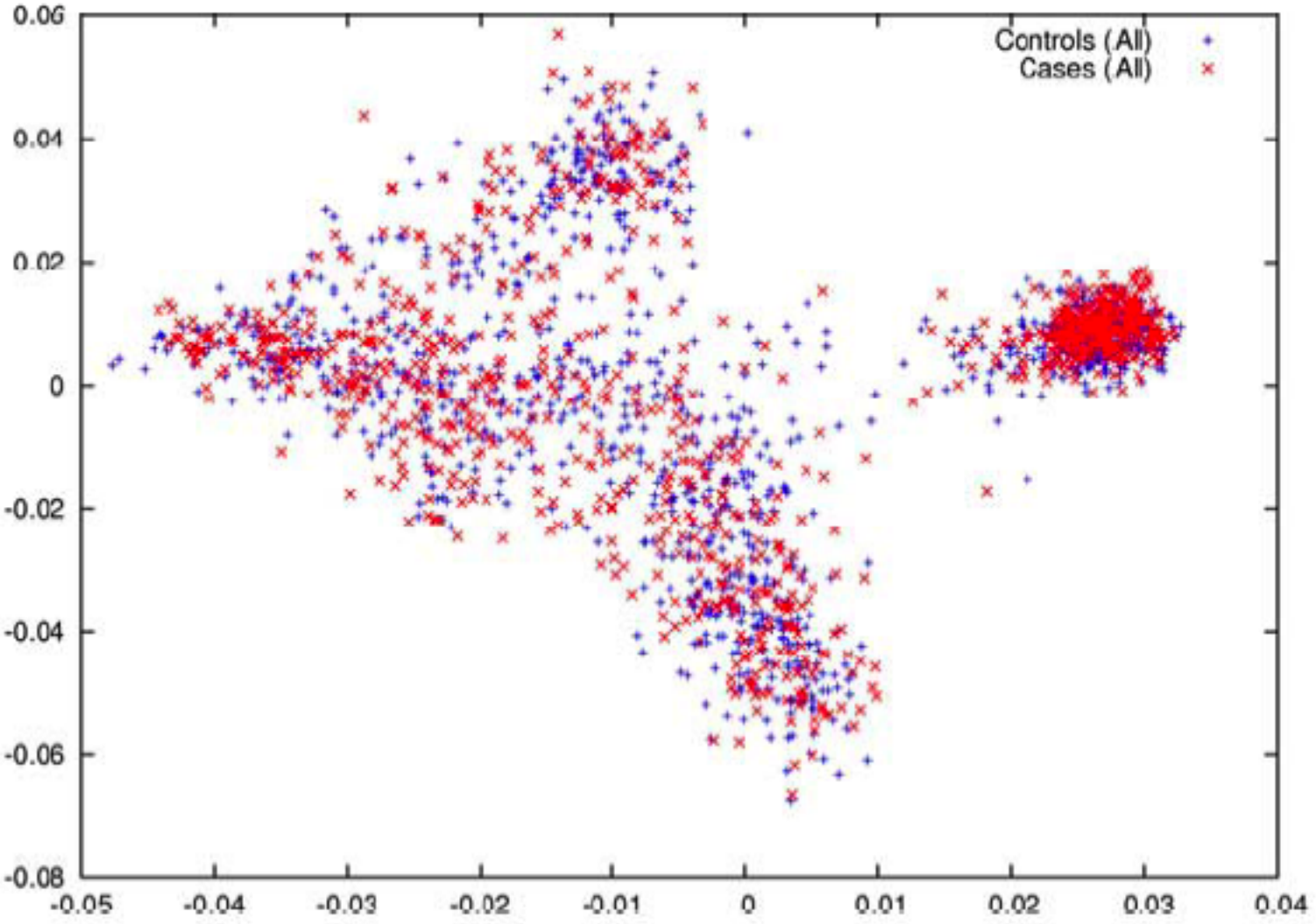
Q-Q plot of the test statistic: expected vs. observed



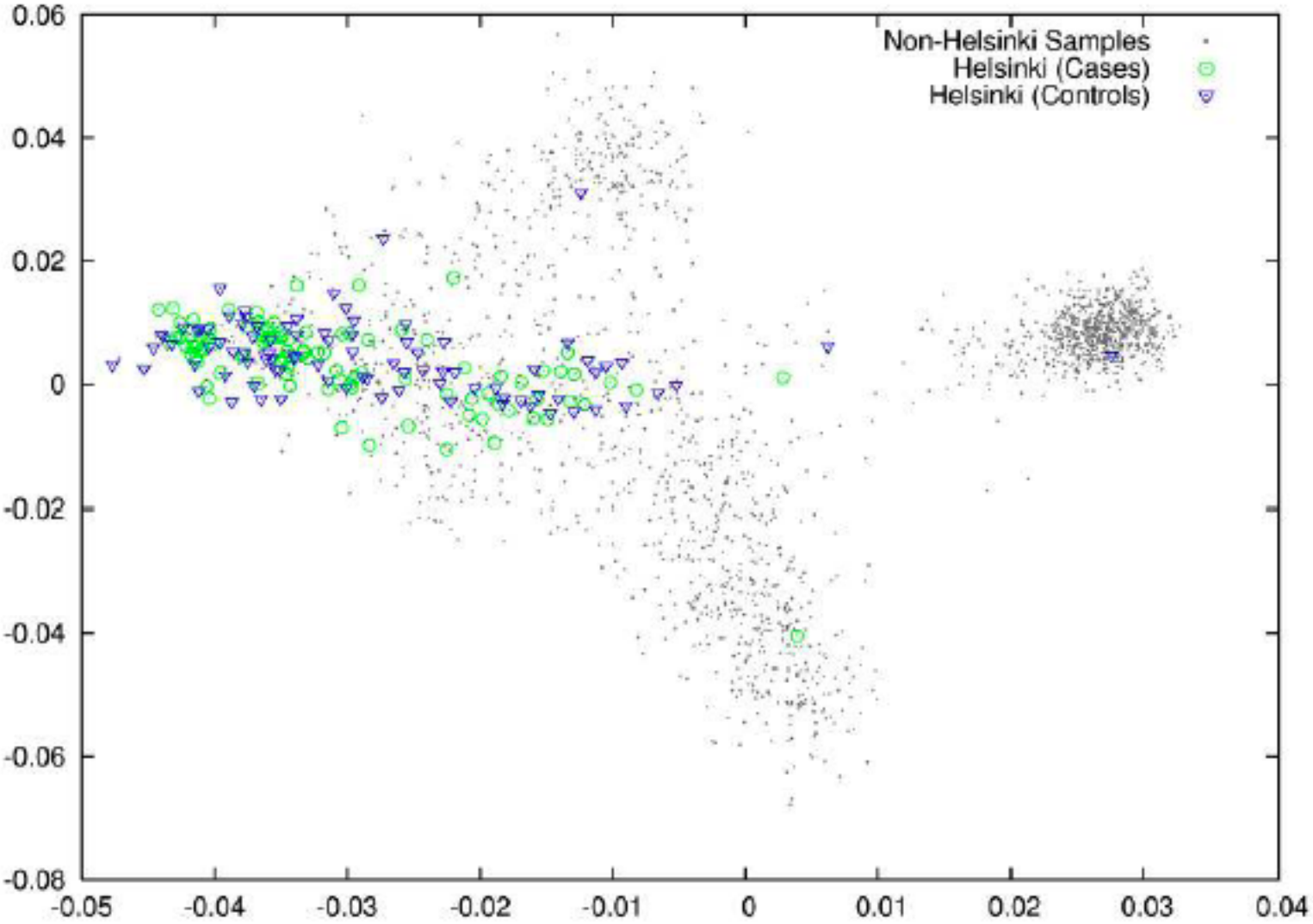
$\lambda_{GC} = 1.05$

population stratification

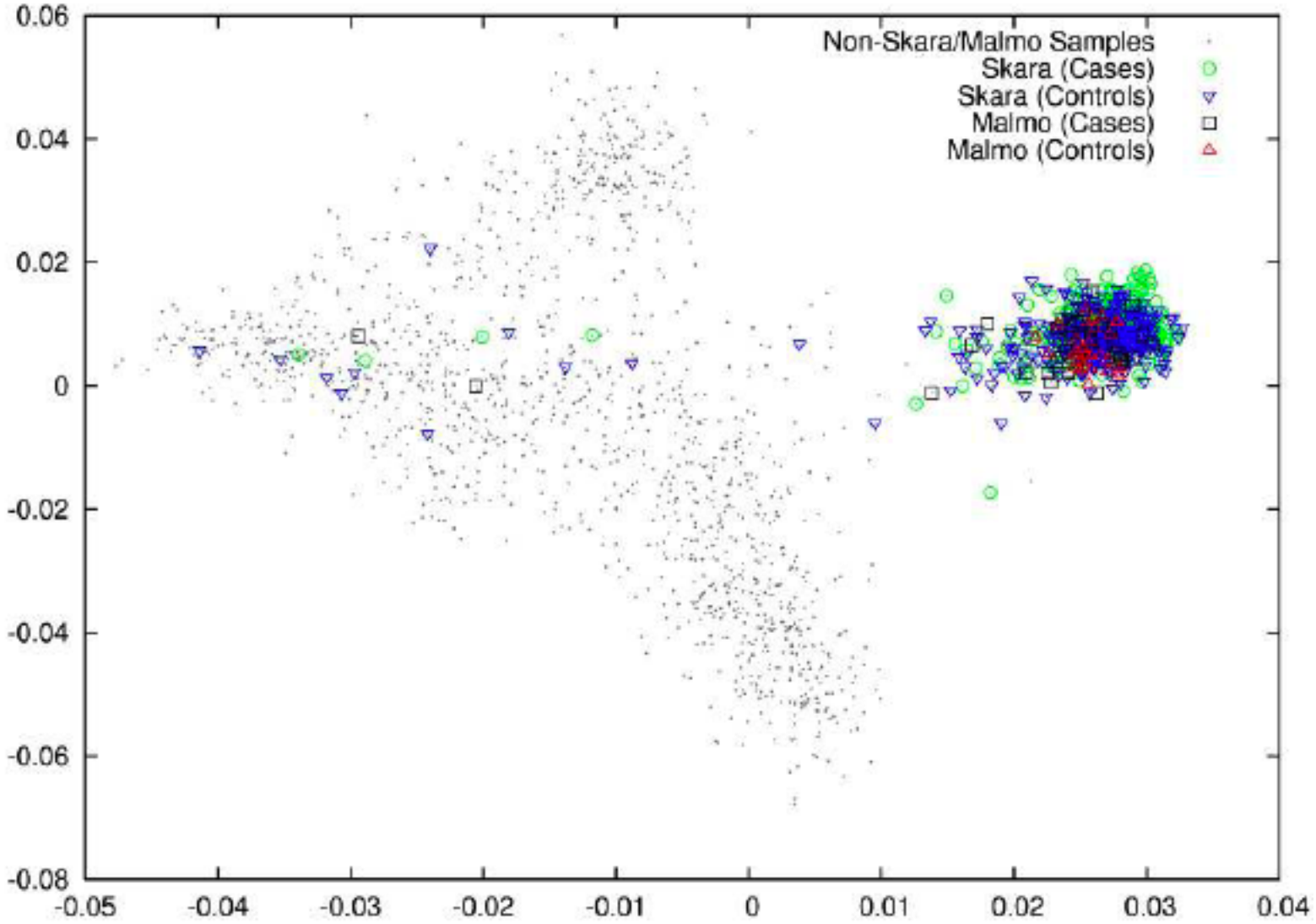
Principal components analysis (PCA) to test for differences between cases and controls



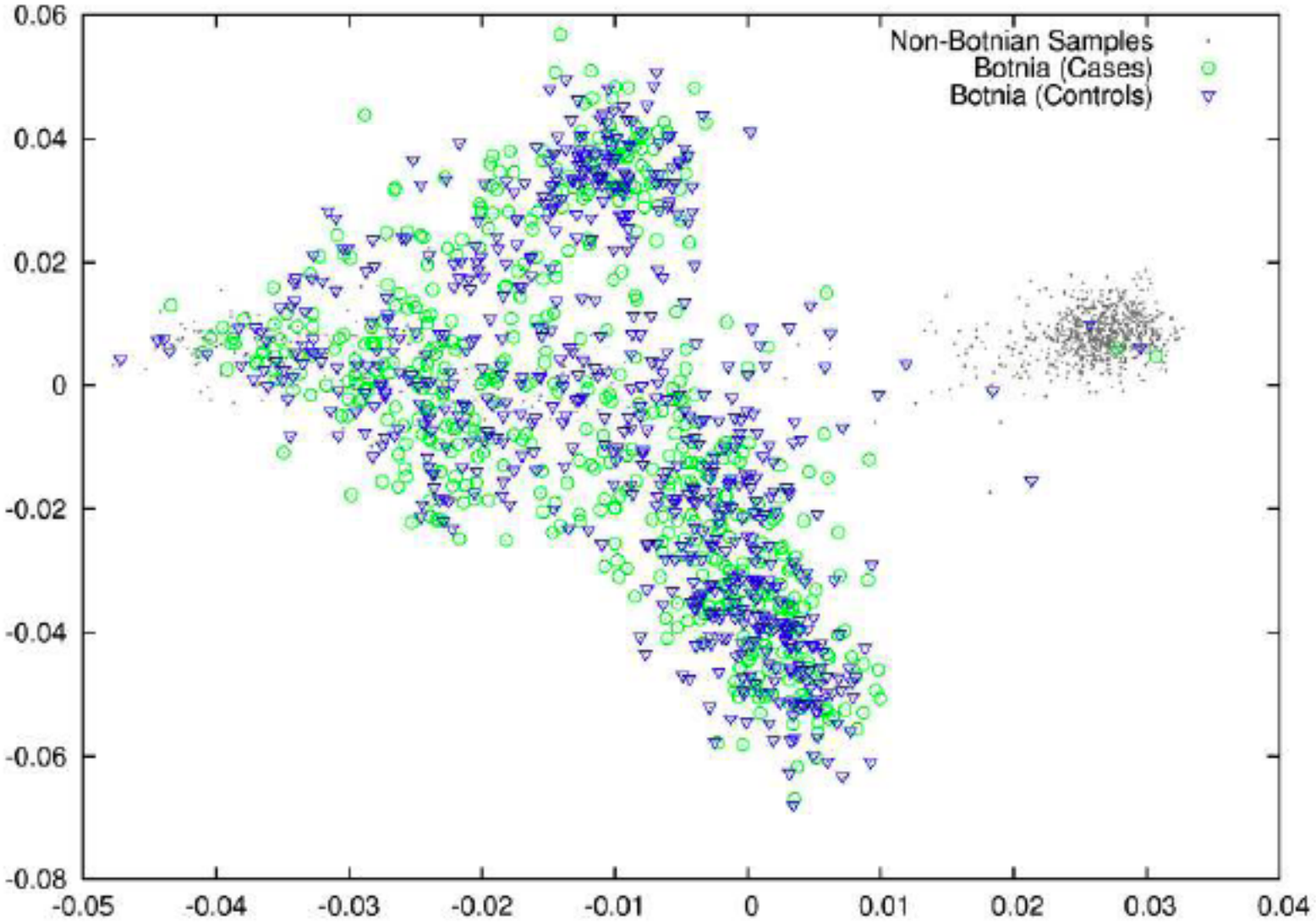
Helsinki


Non-Helsinki Samples
Helsinki (Cases)
Helsinki (Controls)
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04

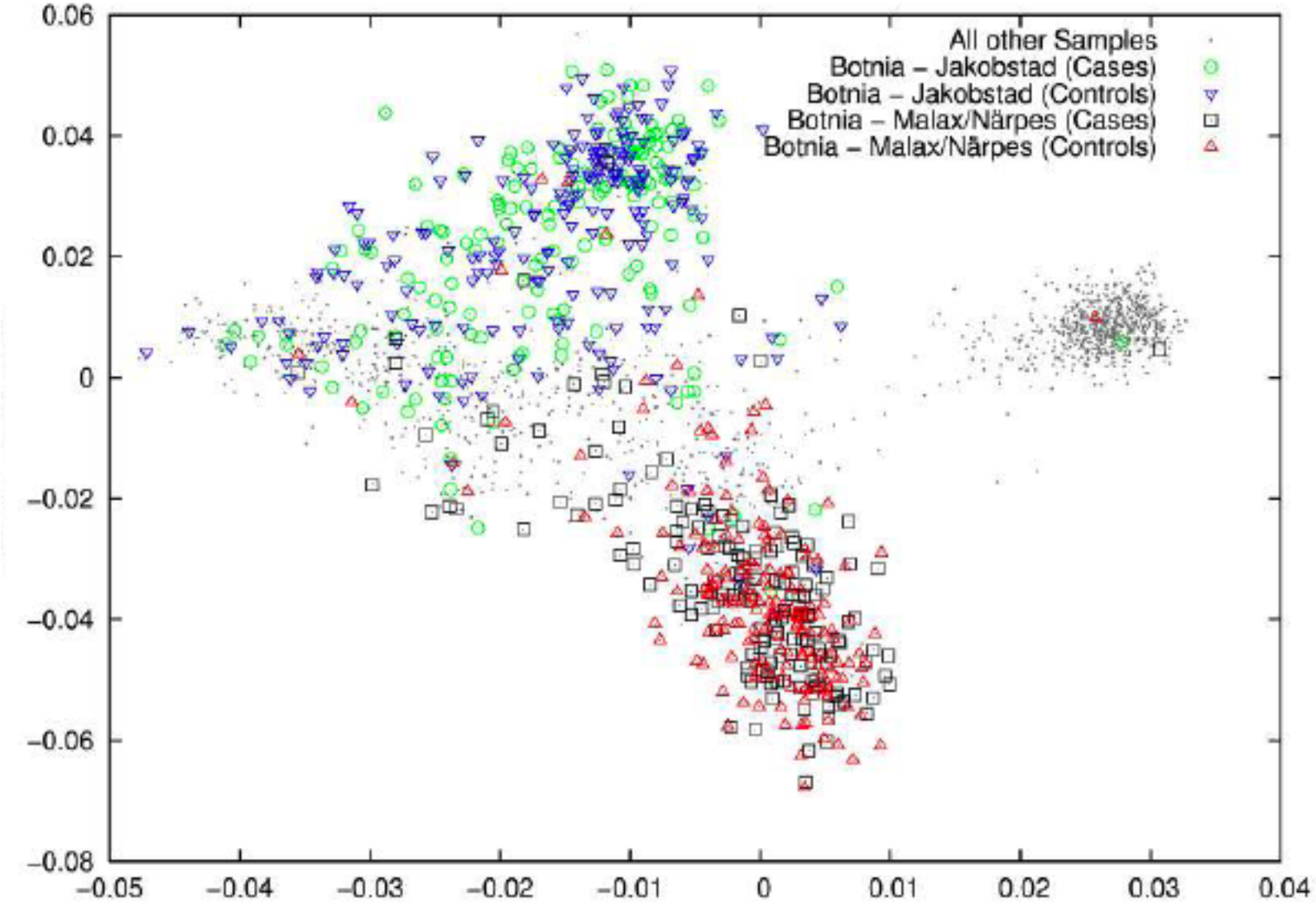
Skara and Malmö



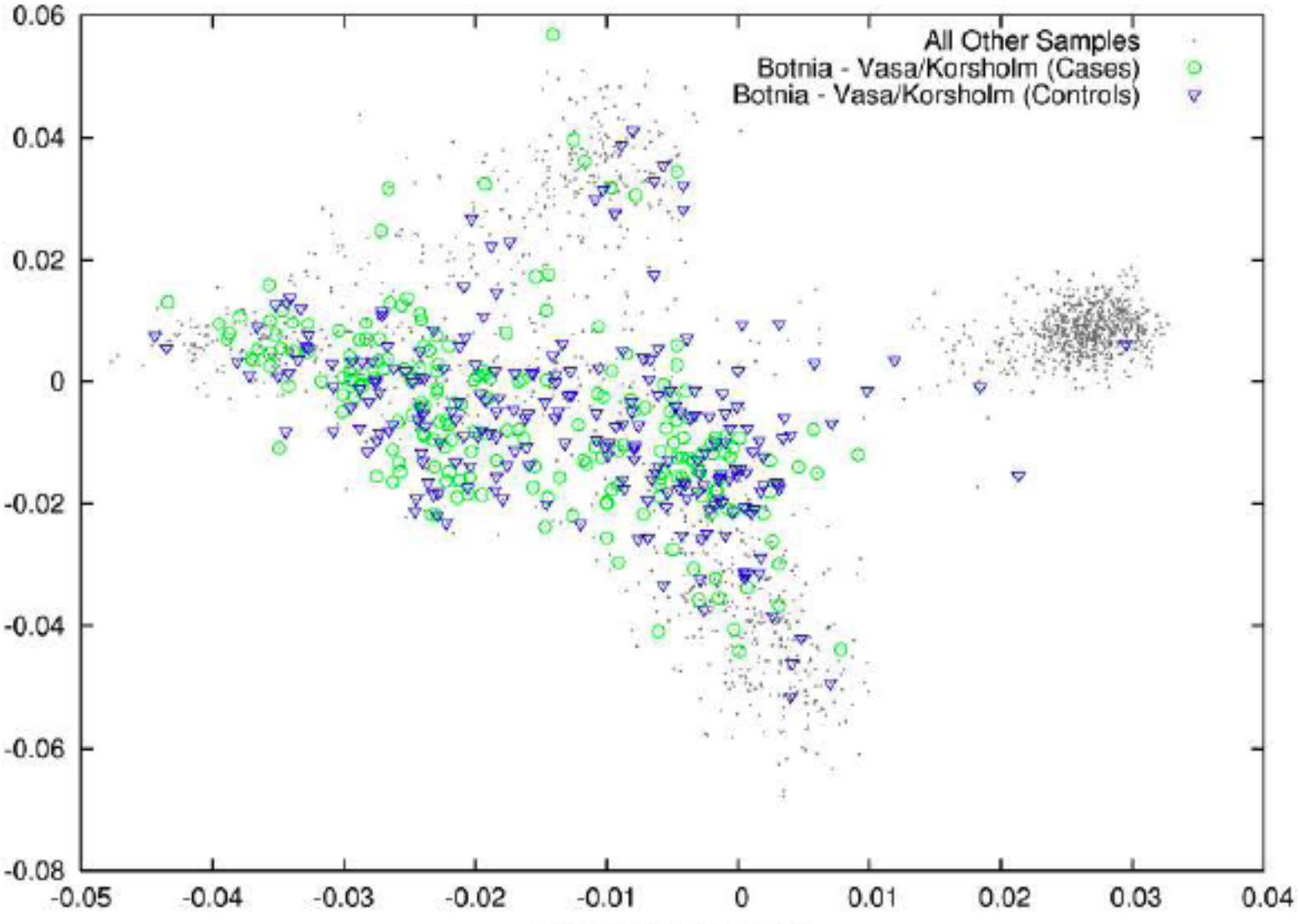
Botnia


Non-Botnian Samples
Botnia (Cases)
Botnia (Controls)
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.05
-0.04
-0.03
-0.02
-0.01
0
0.01
0.02
0.03
0.04

Jakobstad and Malax/Närpes



Vasa/Korsholm

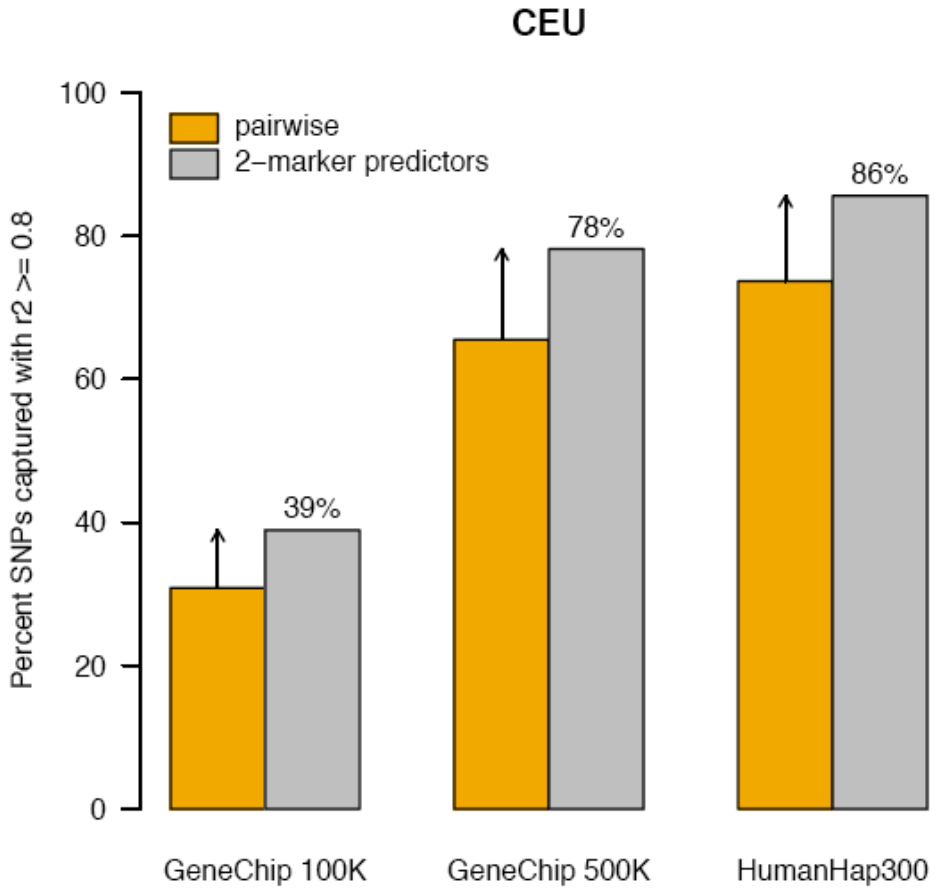


Got stratification?

- Analytical methods to optimize matching between cases and controls
  - EIGENSTRAT (PCA)
  - PLINK (clustering based on identity-by-state)
- For meta-analysis: distributions must be corrected for (e.g. λ_{GC})
- But can't save data if cases and controls are severely differentiated
  - Other control data available? (data sharing)

statistical artifacts due to imputation

Coverage of common SNPs by genome-wide genotyping platforms



Barrett and Cardon; Pe'er, de Bakker *et al*., Nat Genet, 2006

Increasing coverage and power by genome-wide imputation

- Genotyping platforms have partially overlapping SNP sets
  - Roughly 50K SNPs between Affy 500K and Illumina 317K
- Imputation (prediction) of “missing” SNPs
  - Majority of SNPs are highly correlated to genotyped SNPs
  - Minority of SNPs are difficult to impute $\rightarrow$ uncertainty
- Questions:
  - How does this affect the test statistic?
  - What can we do about it?
  - Example: Diabetes Genetics Initiative (DGI) and MACH imputations

1,022 diabetics and 1,075 euglycemic controls matched by age, sex, BMI, location

after QC: 370,847 SNPs

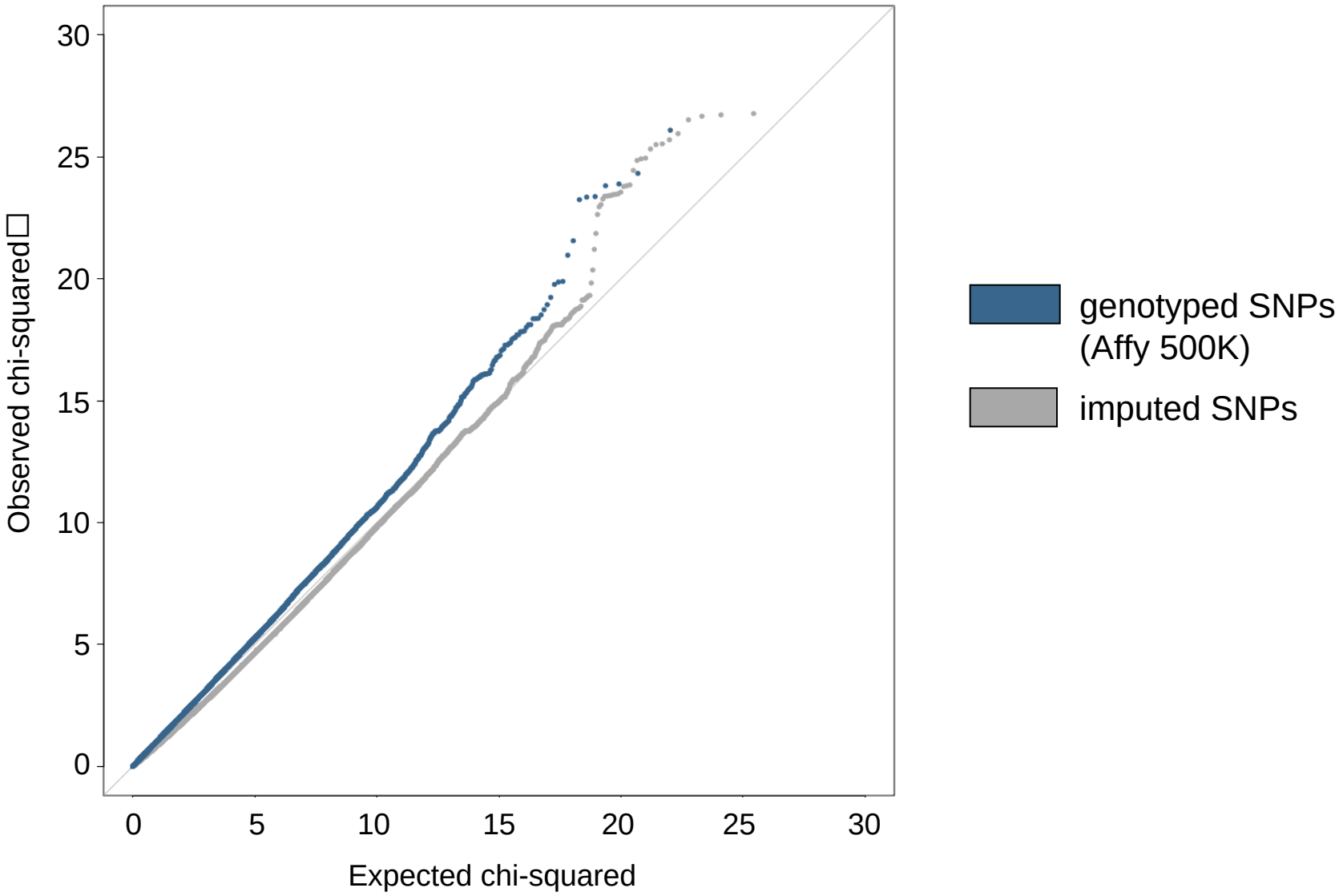
phased haplotypes

MACH

2.55 million SNPs
("dosage vector" in all 2,097 individuals)

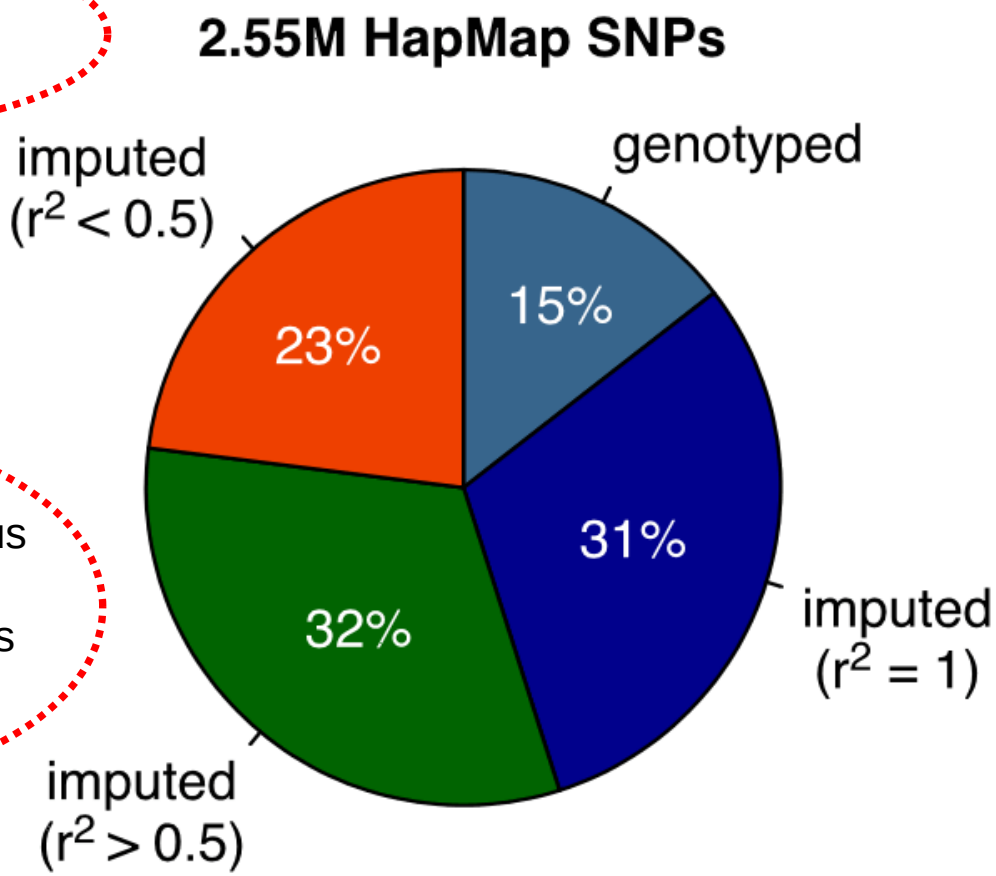
association testing

Q-Q plot: genotyped vs. imputed SNPs



Parsing all imputed SNPs by their correlation (r^2) to the genotyped SNPs

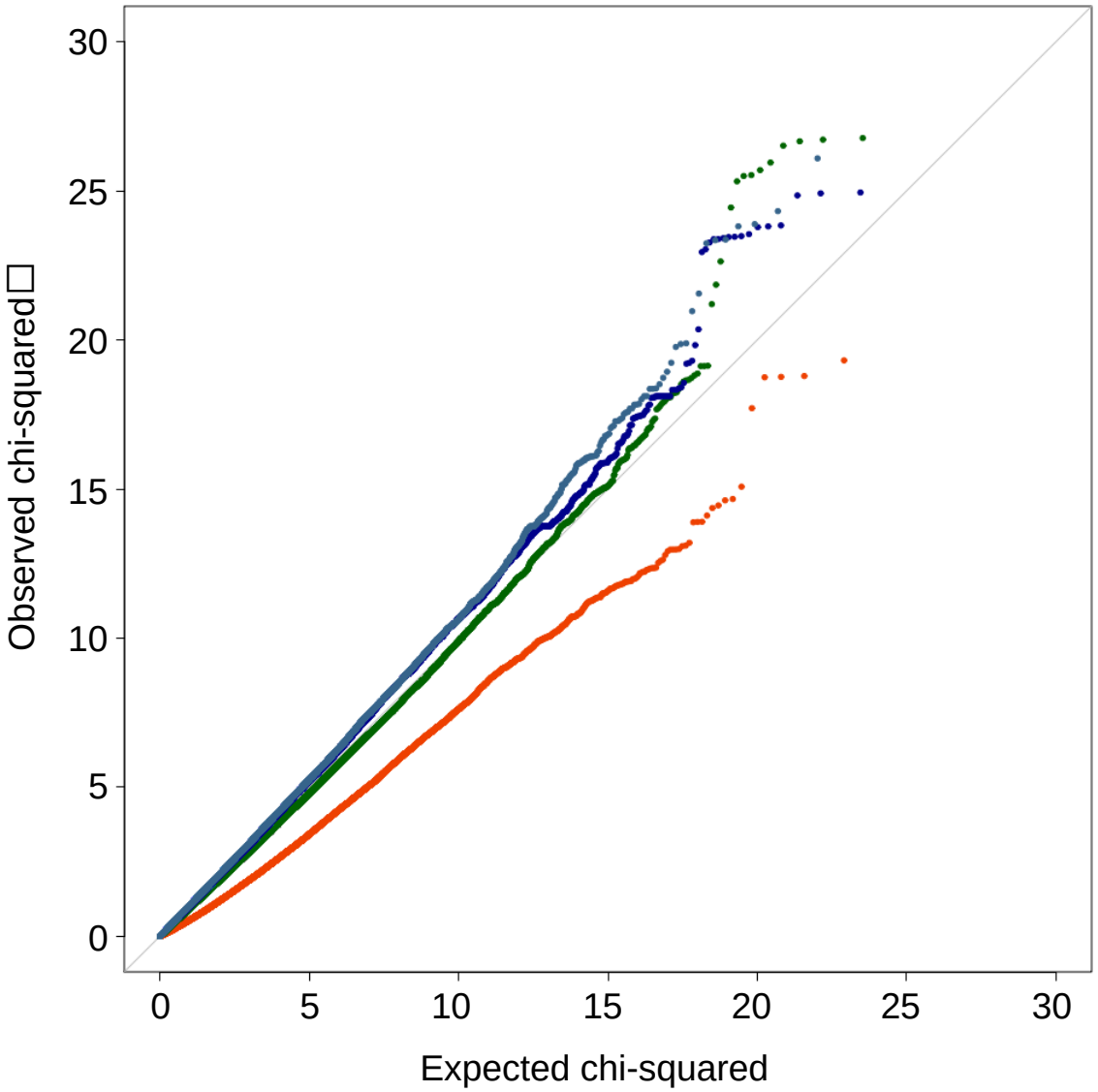
we have not "seen" these SNPs before

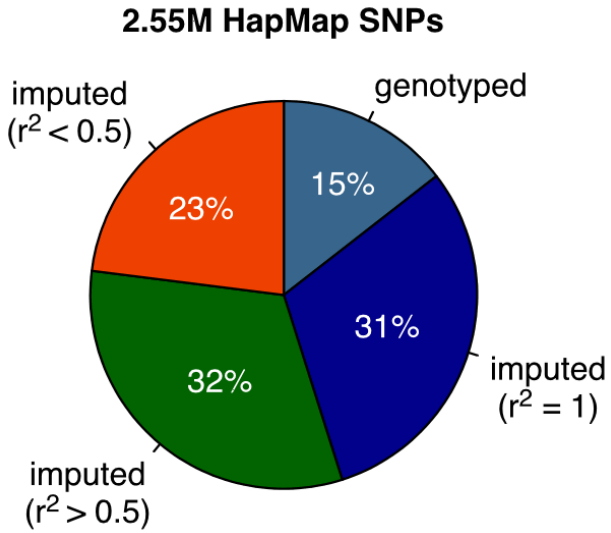


consistent with previous coverage estimates: 66% of common SNPs captured with r^2>0.8

these come for "free"

Serious deflation observed for imputed SNPs that are in poor (pairwise) LD to genotyped SNPs

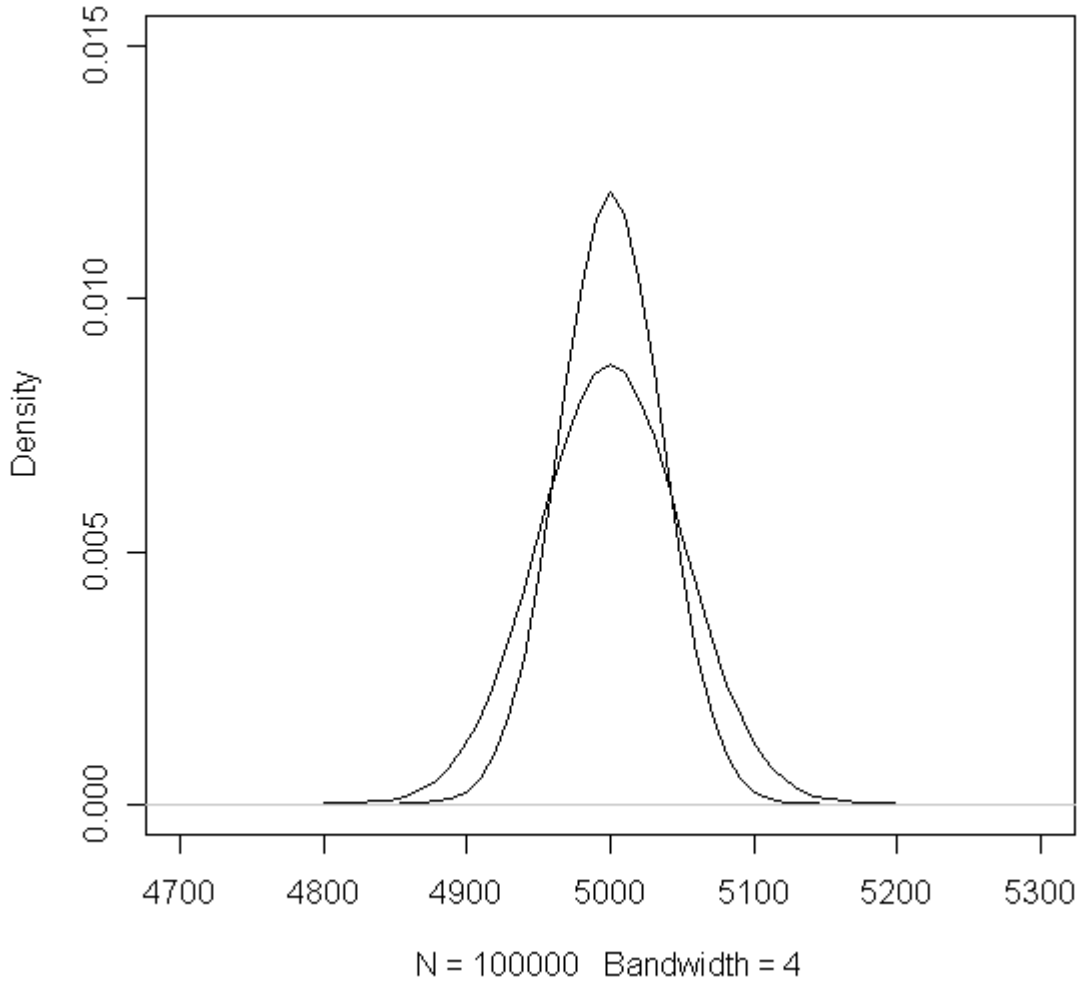




$$\chi^2 = \frac{(p_{case} - p_{control})^2}{\left(\frac{1}{n_{case}} + \frac{1}{n_{control}}\right)(p(1-p))}$$

binomial variance

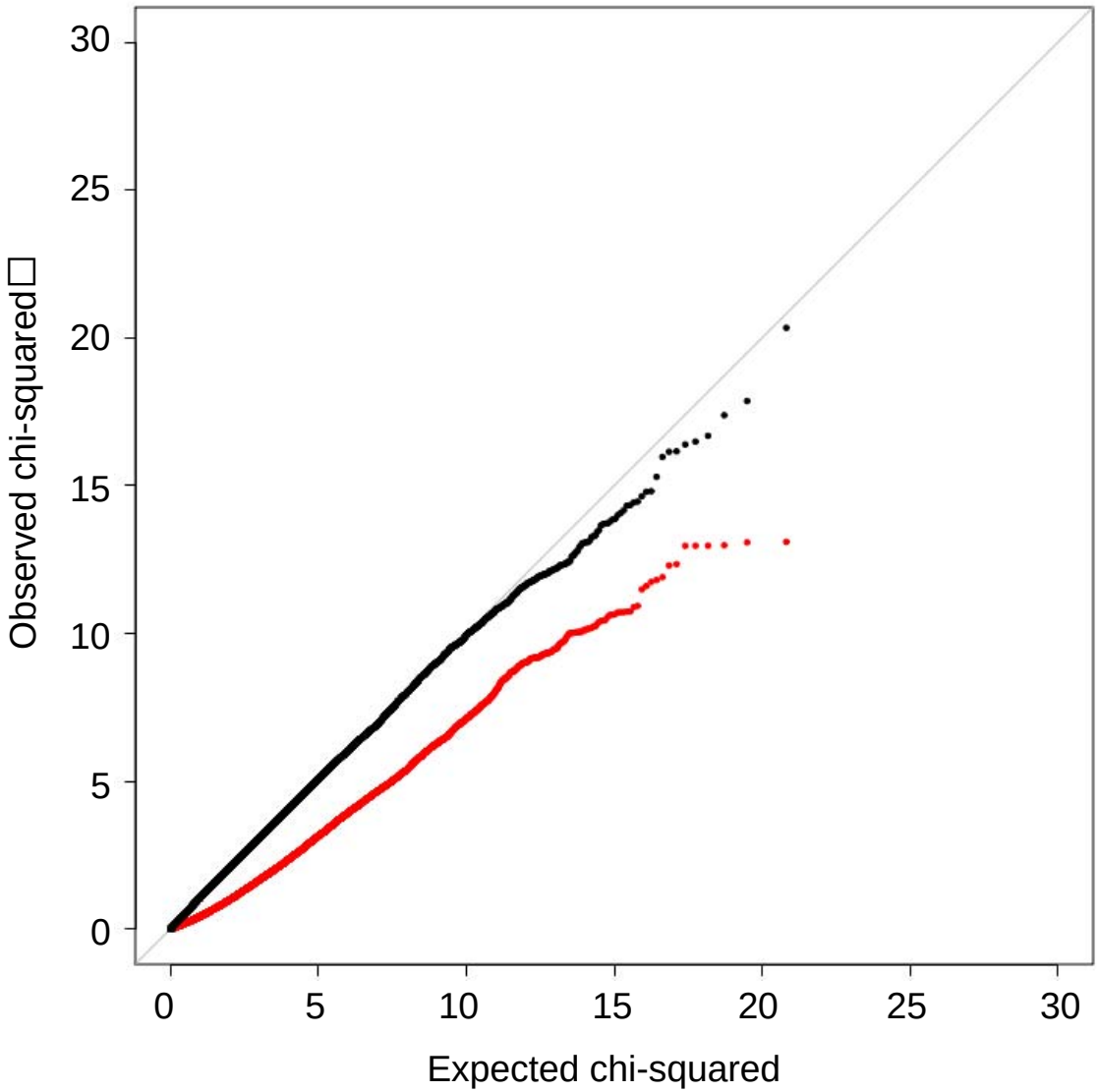
Lack of information (uncertainty) leads to decreased variance of dosage

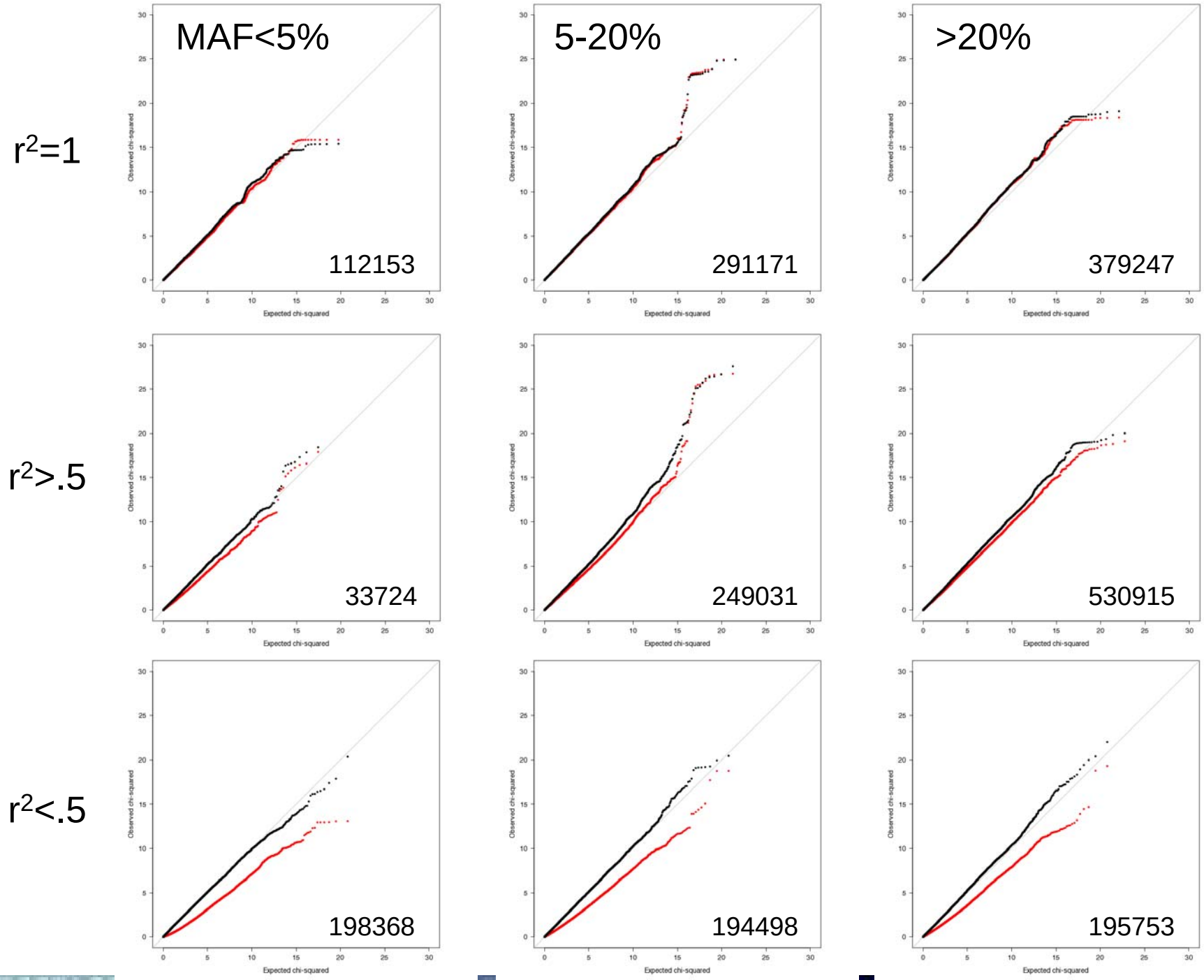


$$\chi^2 = \frac{(p_{case} - p_{control})^2}{\left(\frac{1}{n_{case}} + \frac{1}{n_{control}}\right)(p(1-p))}$$

replace with empirically observed variance

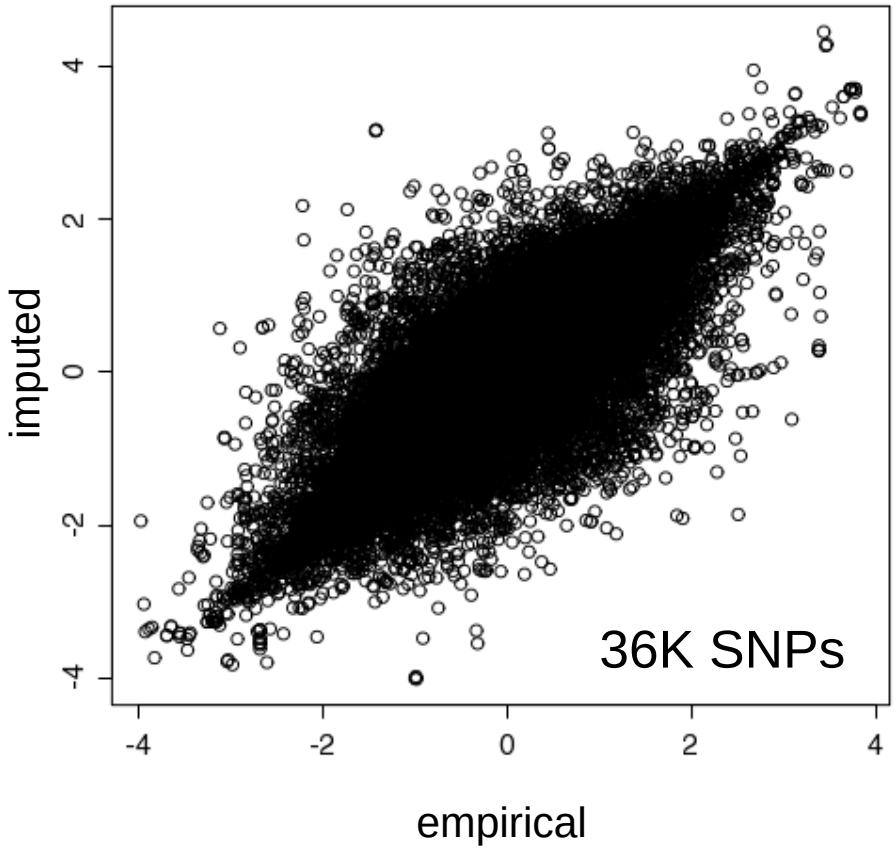
This correction re-inflates the distribution



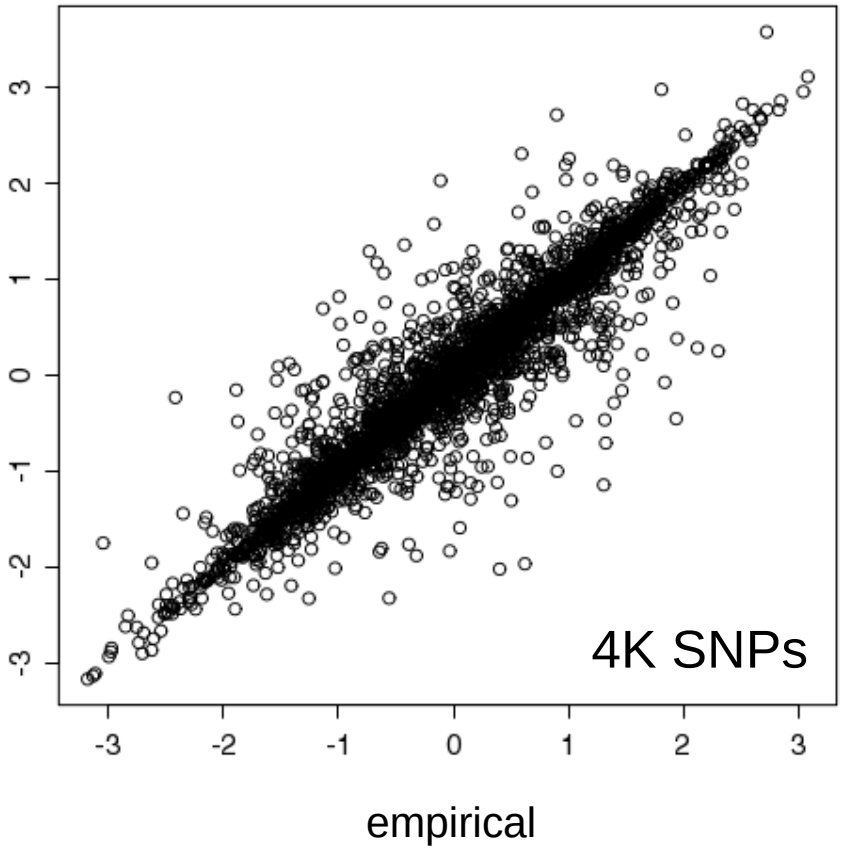

MAF<5%
5-20%
>20%
r2=1
r2>.5
r2<.5
112153
291171
379247
33724
249031
530915
198368
194498
195753
Observed chi-squared
Expected chi-squared

Correlation in test statistic for rare and common SNPs: genotyped vs. imputed data

r^2=0.68 MAF<5%

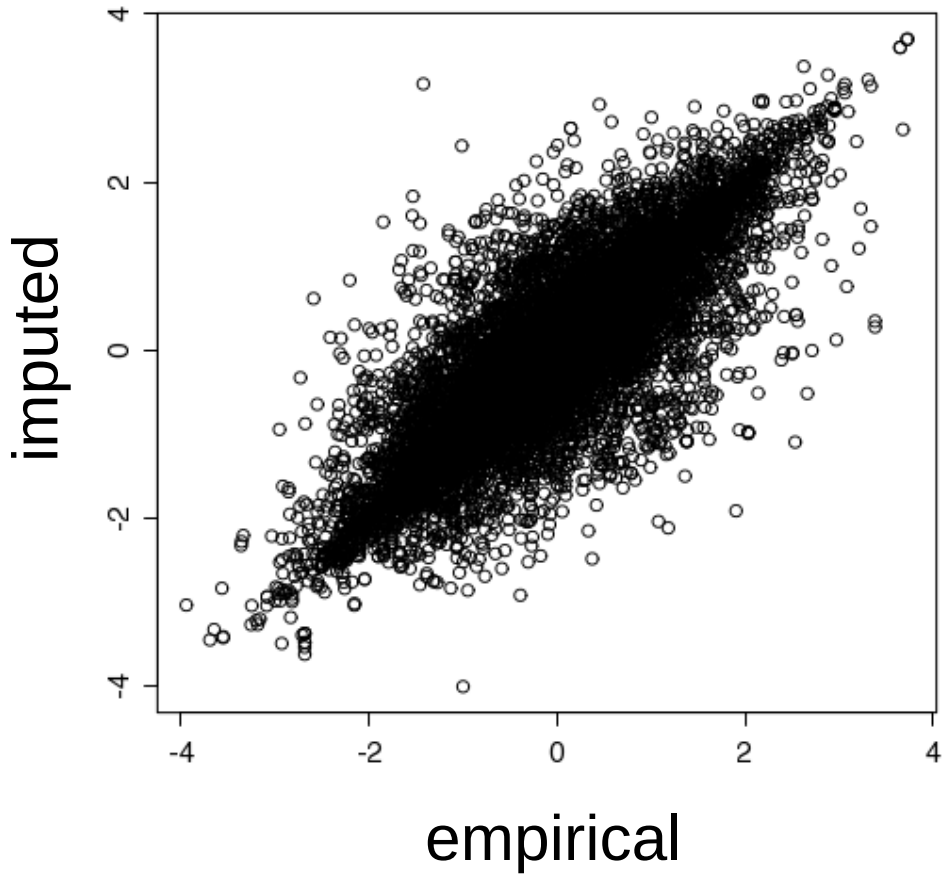


r^2=0.88 MAF>5%



Same effect observed in ultra-clean set of rare SNPs

(missingness <0.1% and HWE p-val>0.1)



r^2=0.69

Conclusions

- Imputation methods available and user-friendly
- Word of caution for subset of SNPs that show deflated test statistics
  - Simple correction is proposed
- Some SNPs (mostly rare) would benefit from a larger HapMap

Acknowledgements

Benjamin Neale and Mark Daly

Diabetes Genetics Initiative

Richa Saxena, Benjamin Voight, Noel Burtt, Valeriya Lyssenko, Leif Groop, David Altshuler

WTCCC/UKT2D

Eleftheria Zeggini, Jonathan Marchini, Mark McCarthy, Andrew Hattersley

FUSION

Laura Scott, Yun Li, Gonçalo Abecasis, Francis Collins, Mike Boehnke