Microsoft® Research

Faculty Summit

10

YEAR ANNIVERSARY

Microsoft Research

Improving Detection in GWAS by Discovering and Accounting for Race, Relatedness, and Other Hidden Relationships

Jennifer Listgarten
Researcher
Microsoft Research

joint work with:
David Heckerman & Carl Kadie,
Microsoft Research
Hyun Min Kang, UCLA

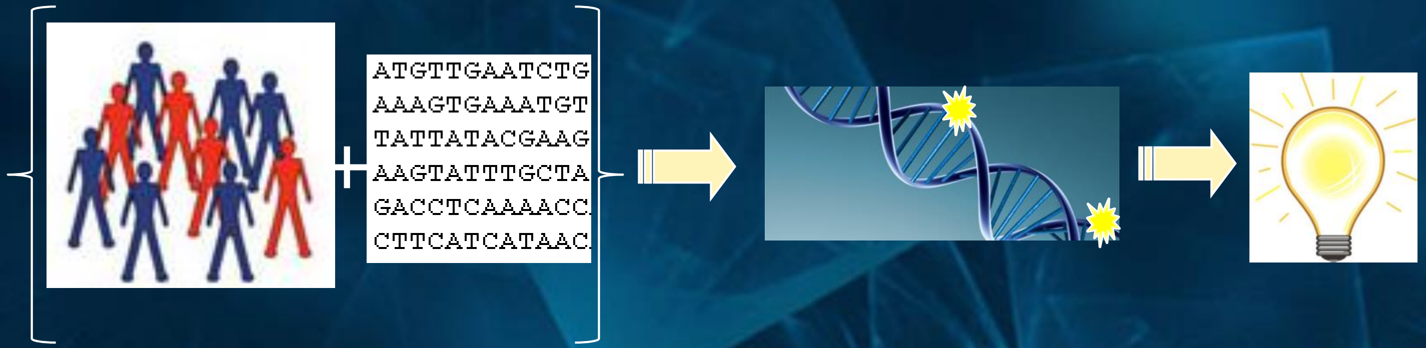
GWAS Overview

Input:

- A set of people with/without a disease (*e.g.*, cancer)
- Measure a large set of genetic markers for each person (*e.g*., measurement of DNA at various points)

Desired output:

- A list of genetic markers causing the disease



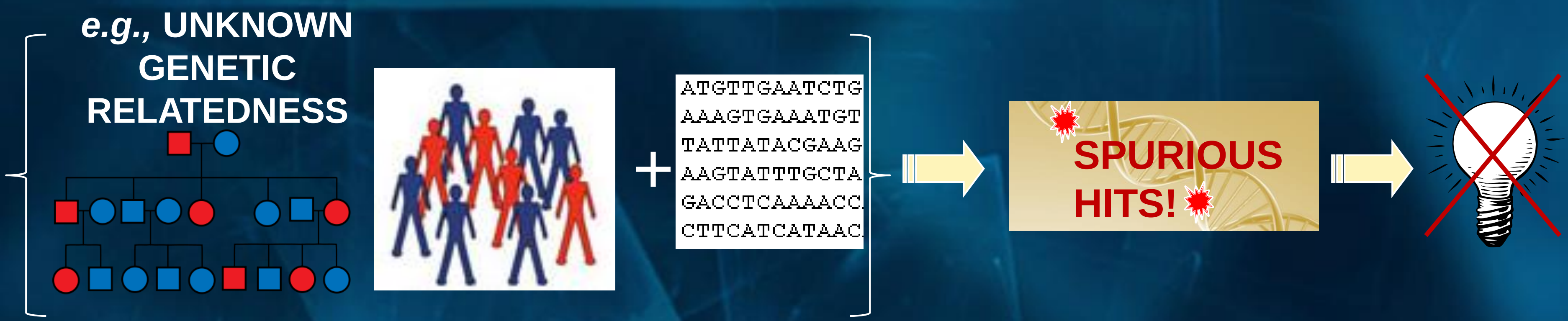
```
ATGTTGAATCTG
AAAGTGAAATGT
TATTATACGAAG
AAGTATTTGCTA
GACCTCAAAACC
CTTCATCATAAC
```

-personalized medicine
-new drug targets
-screening & preventative measures
-genetic counseling
-disease mechanism understanding

Major Statistical Modeling Challenge

Hidden structure in the data leads to:

1. Loss of power to detect signal of interest
2. Spurious hits (*i.e.*, false positives) due to unaccounted confounding signal

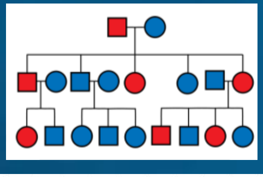


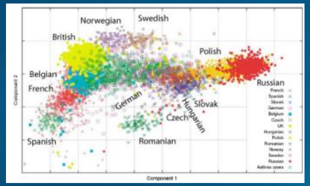
Hidden Structure?

Fundamental assumption in most statistical tests is that the subjects are sampled independently from the same distribution

BUT…IF subjects:

- Are closely/distantly related to each othe.
- Comprise different ethnicities
- Have samples that contain batch effects (processed slightly differently, and not at random)
- *etc*. (unknown confounders we don't yet know about)



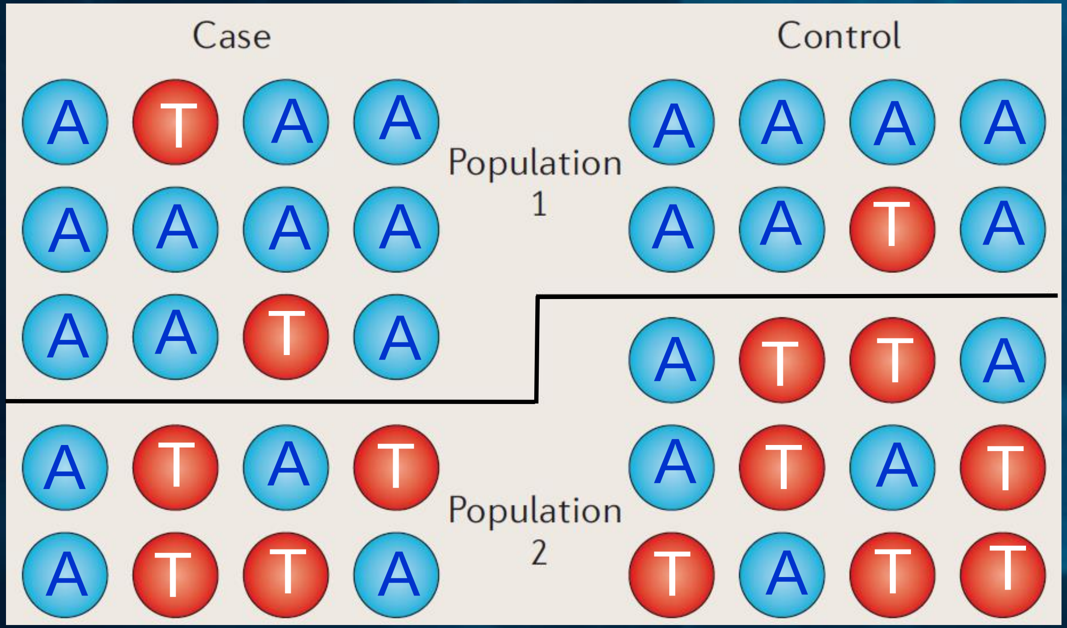


THEN…

- Spurious correlations induced giving spurious hits
- True signal swamped out, reducing power to detect true associations

SPURIOUS HITS!

e.g. of How Hidden Structure Can Hurt

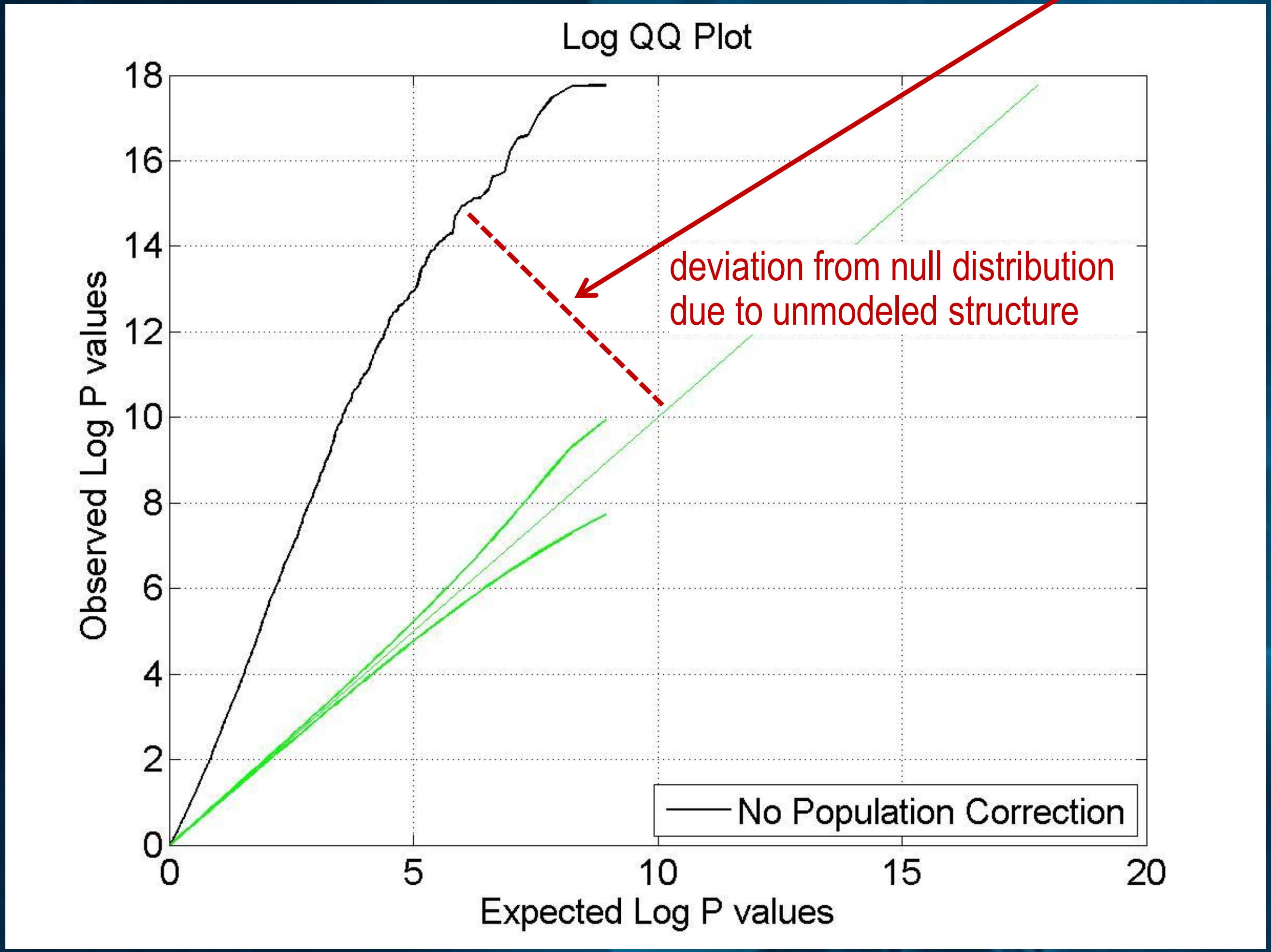


(Balding, Nat Rev Genet. 2006)

- Suppose the set of *cases* has a different proportion of ethnicity X from *control*
- Then genetic markers that differ between X and other ethnicities in the study, Y, will appear artificially to be associated with disease
- Furthermore, these (often numerous and strong) spurious associations can swamp out the true signal of interest

- Also, the larger the study (# people), the worse the problem, since the power to detect ‘spurious’ signal increases
- But large studies are needed to detect markers with weak effect

Leveraging Scale Of GWAS to Find Evidence of Hidden Structure



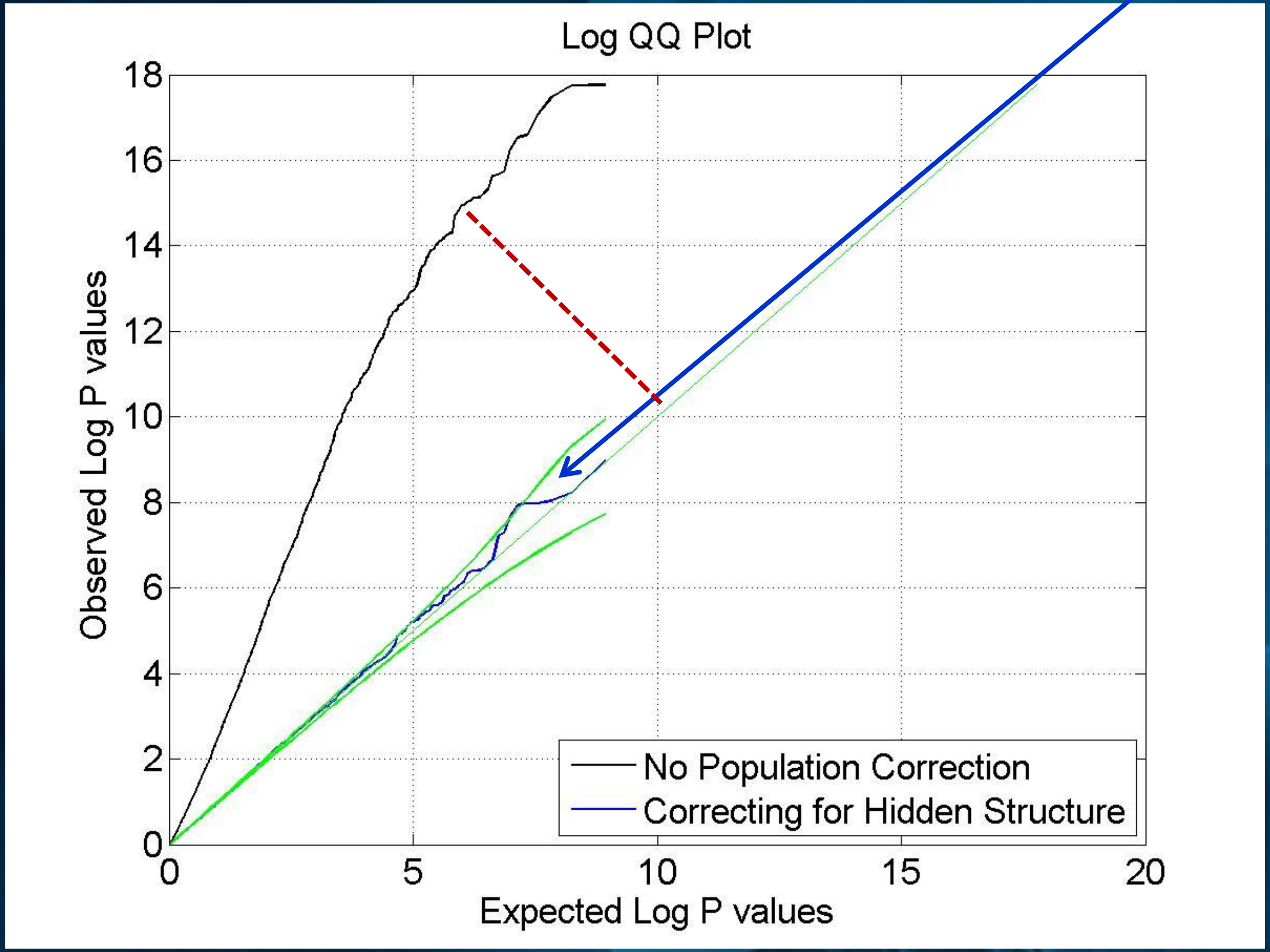
deviation from null distribution due to unmodeled structure

~7500 SNPs, ~1000 people, contains multiple ethnicities and families)

- When testing thousands of genetic markers for association with a disease, we expect very few of them to truly be associated with disease
- Key insight: the resulting distribution of test statistics should be close to a uniform p-value distribution

Leveraging Scale Of GWAS to Correct For Hidden Structure

Microsoft Research

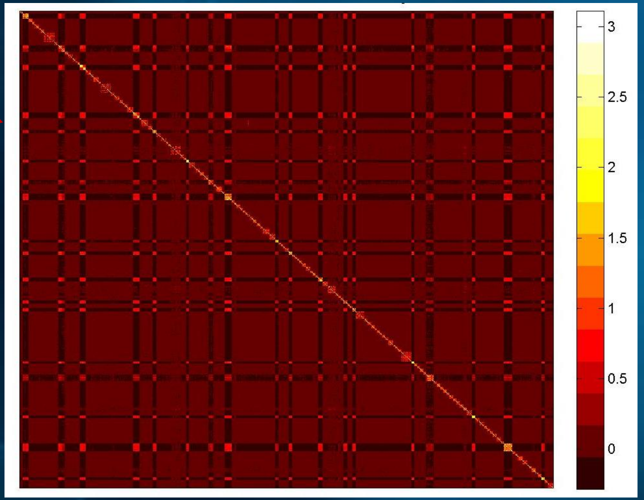


~7500 SNPs, ~1000 people, contains multiple ethnicities and families)

- When testing thousands of genetic markers for association with a disease, we expect very few of them to truly be associated with disease
- Key insight: the resulting distribution of test statistics should be close to a uniform p-value distribution

Leveraging Scale Of GWAS to Correct For Hidden Structure

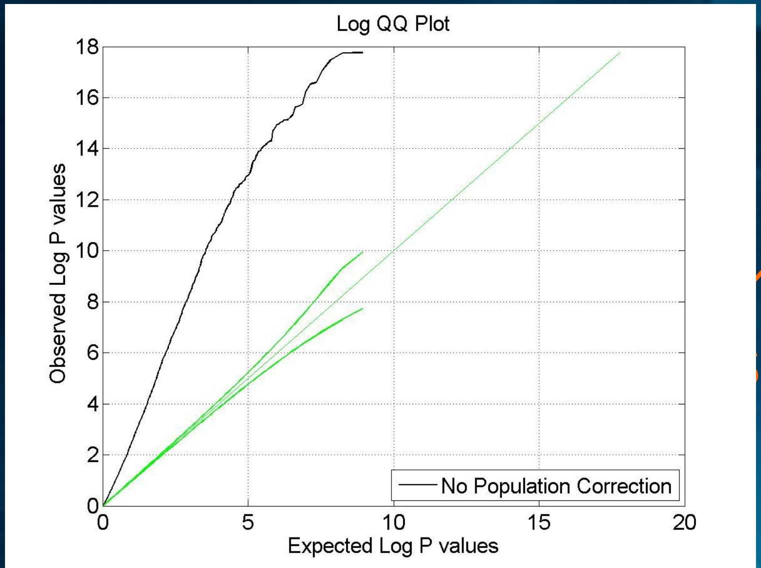
- Use the large scale of the data set itself to infer hidden population structure
- *i.e.*, Use the genetic markers themselves, in aggregate, to see how 'similar' every two people are, and incorporate this into the analysis
- Best current approaches are:
  1. *Principle Component Analysis* –based
  2. *Linear Mixed Models*



genetic 'similarity' matrix

Digression: Naïve Approach→ Linear Regression

- Regress target phenotype on each genetic marker
- *e.g.*, regress blood pressure level on SNP (and do for each SNP)
- Evaluate SNP for association by comparing this model to one that ignores the SNP (*e.g.* use LRT statistical test)



$$y = X\beta + \varepsilon$$

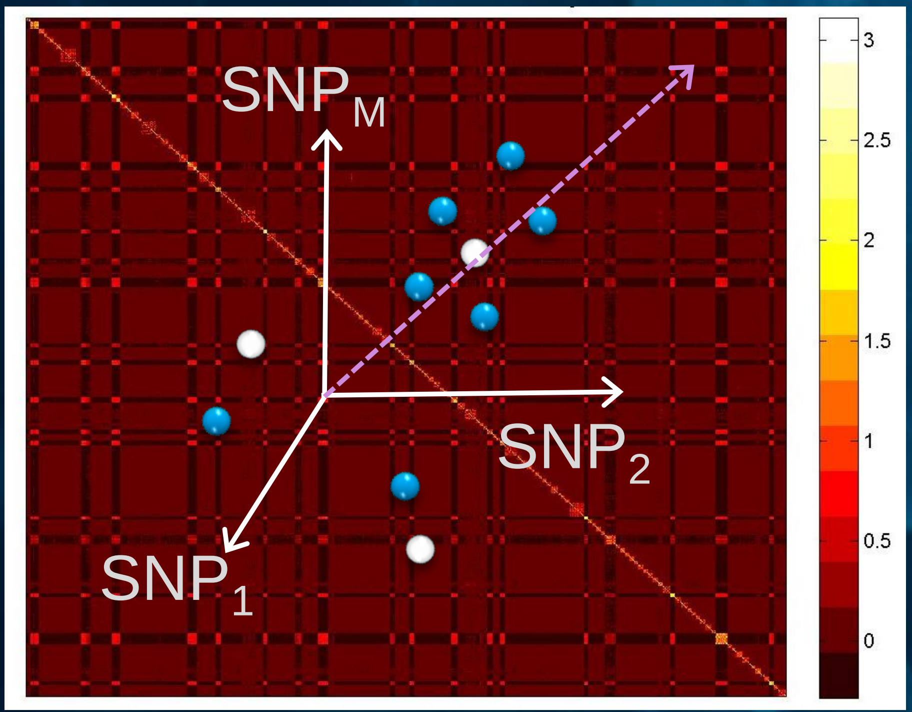
gaussian noise

ure

SNP

learned regression weight
(importance of SNP to blood pressure)

Principle Components Analysis Approach



genetic similarity between every two people

- Find major ‘*axes of variation*’ of the high dimensional space (# markers)
- Project each person’s markers into the low dimensional space captured by the top few axes
- Add projections as covariates in a standard regression analysis that looks for associations between marker and phenotype

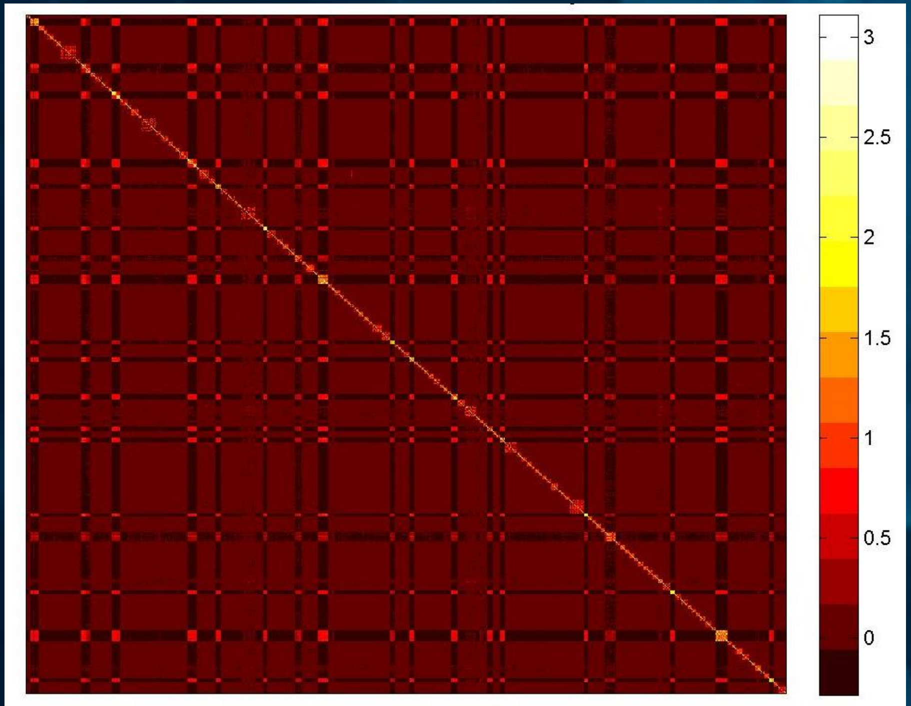
- ➢ *Works well to capture broad structure*
- ➢ *Sensitive to outliers (bad!)*
- ➢ *Cannot capture fine-grained structure (bad!)*
- ➢ *Fast computations (good!)*

$$y = X\beta_1 + P\beta_2 + \varepsilon$$

projection in low-dim space

learned regression weight

Linear Mixed Model Approach



genetic similarity between every two people

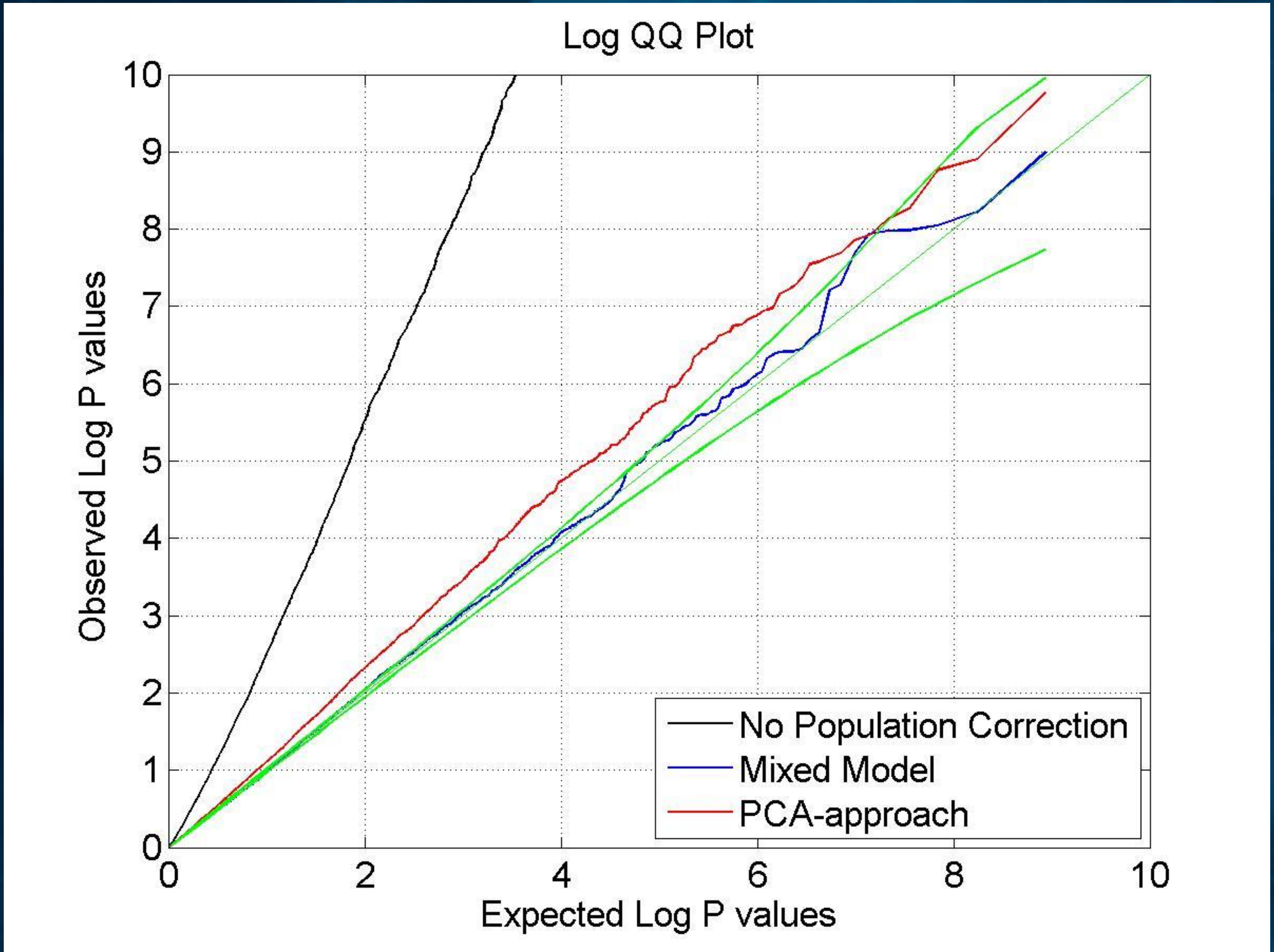
- Do *not* reduce space to a set of directions Use it in its entirety!
- Use similarity as a (Bayesian) *prior over* hidden regression coefficients that are integrated out within a standard regression analysis

$$\vec{u} \leftarrow Normal(\vec{0}, \quad)$$

$$\vec{y} = X\beta_1 + \left(\int \vec{u}\beta_2 d\vec{u} \right) + \vec{\varepsilon}$$

- ➢*Captures multiple levels of similarity: broad and fine (good!)*
- ➢*Not sensitive to outliers (good!)*
- ➢*Computationally expensive (bad!)*

PCA-based Approach vs. Mixed Model



Mixed model works better than PCA approach here

~7500 SNPs, 1000 people, variety of ethnicities + people that are related

Our Contributions to Mixed Model Approach

- Learning similarity matrixes from the data and showing them to be better than prior known structure usually used (e.g. pedigree)
- Combining heterogeneous sources of ‘similarity’ to gain power and reduce spurious association
- Using approximation tricks to make the models as fast as Principle Components Analysis approaches

Microsoft® Research

Microsoft®

Your potential. Our passion.™