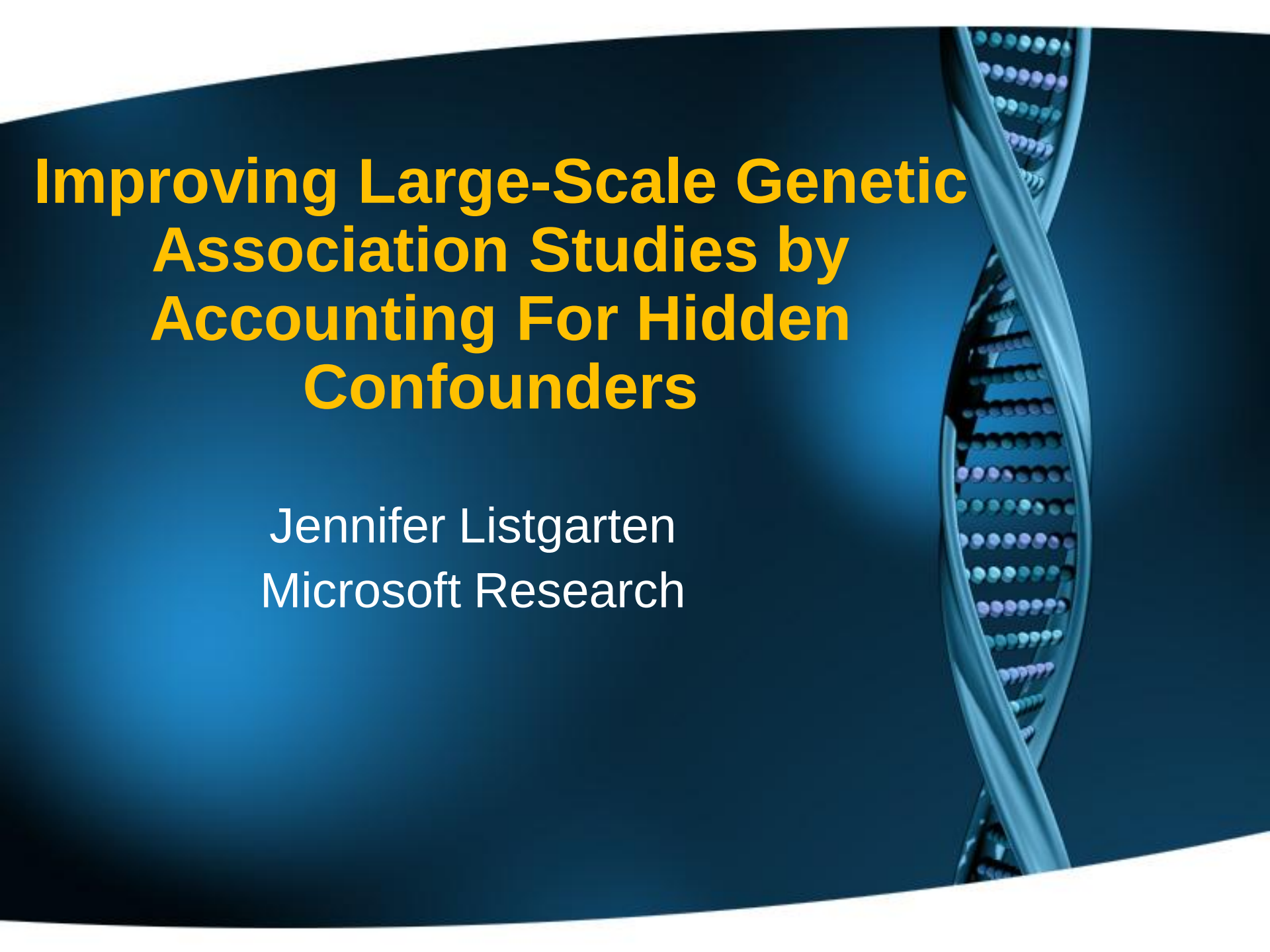


Improving Large-Scale Genetic Association Studies by Accounting For Hidden Confounders

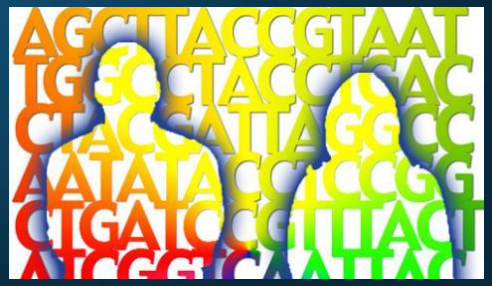
Jennifer Listgarten
Microsoft Research

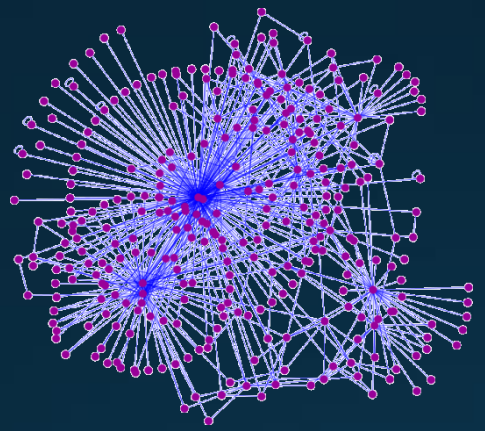
Talk Outline

1. Introduction
2. GWAS and Problem of Hidden Structure
3. New Statistical Method for eQTL Analysis

Importance of Genetic Studies

1. Adapting treatments to a person's genetic make-up.
2. Early risk identification
3. Refining disease definitions (*i.e.*, sub-classification for more targeted treatment)
4. Basic biological insight





Focus on GWAS (Genome Wide Association Studies)

Input:

- A set of people with/without a disease
- Measure a large set of genetic markers for each person (*e.g.*, SNPs).

Desired output:

- A list of genetic markers underlying the disease.

DNA from set of individuals

list of genetic markers e.g. SNPs

Hidden Structure?

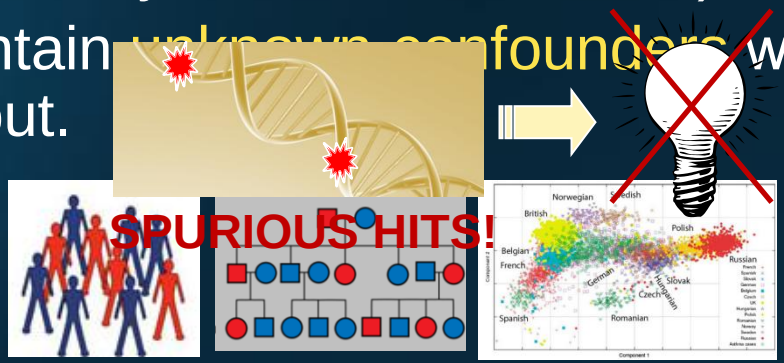
Fundamental assumption in most models is that the subjects are sampled *identically and independently* from the same distribution.

BUT IF subjects:

- Are closely/distantly related to each other.
- Comprise different ethnicities.
- Samples batch effects (processed slightly differently, and not at random).
- Contain unknown confounders we don't yet know about.

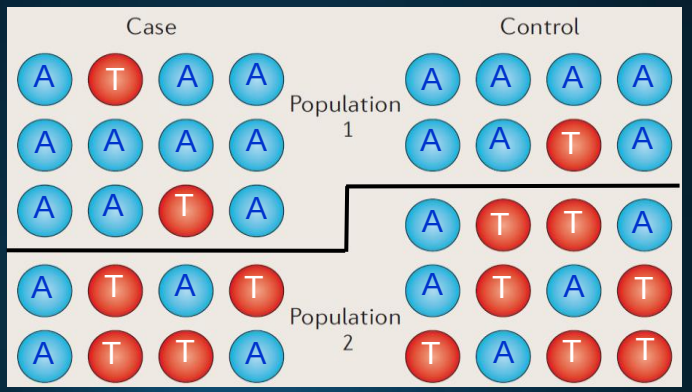
THEN…

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



Intuition of Hidden Ethnicity Structure

- 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.



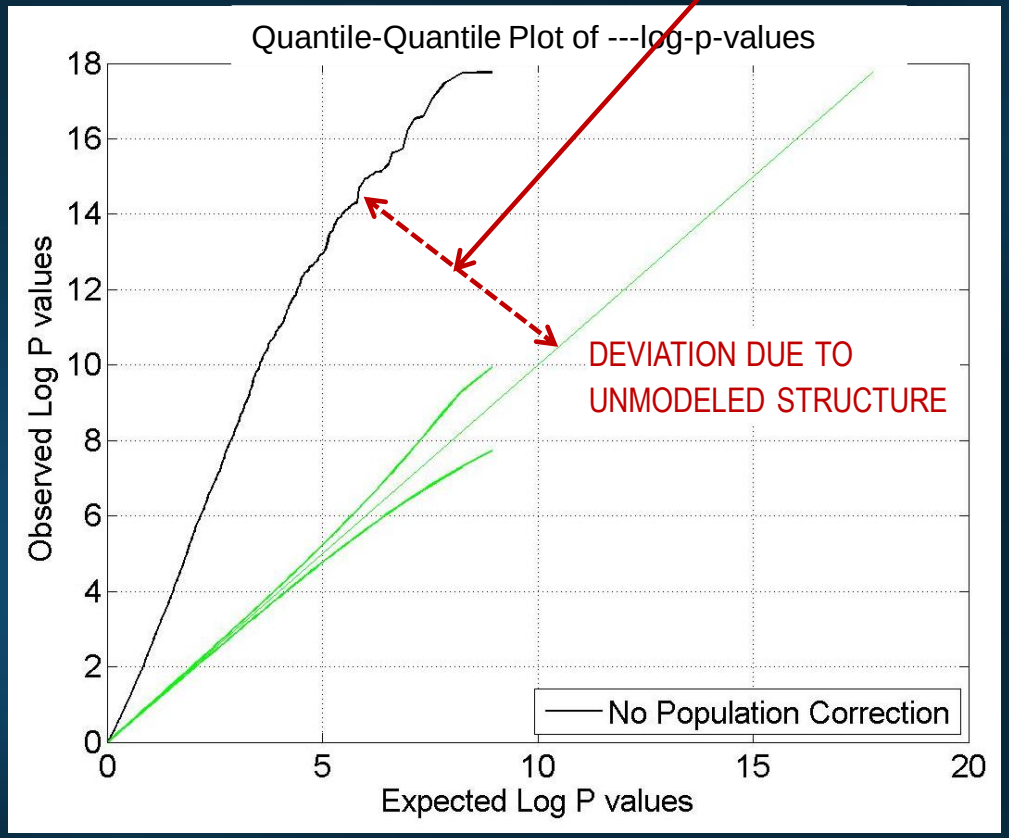
[Balding, *Nat Rev Genet*. 2006]

- These spurious associations can swamp out the true signal of interest, or induce spurious associations.

- 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.

Diagnosing Analyses w P-value Distributions

Evidence of problematic analysis

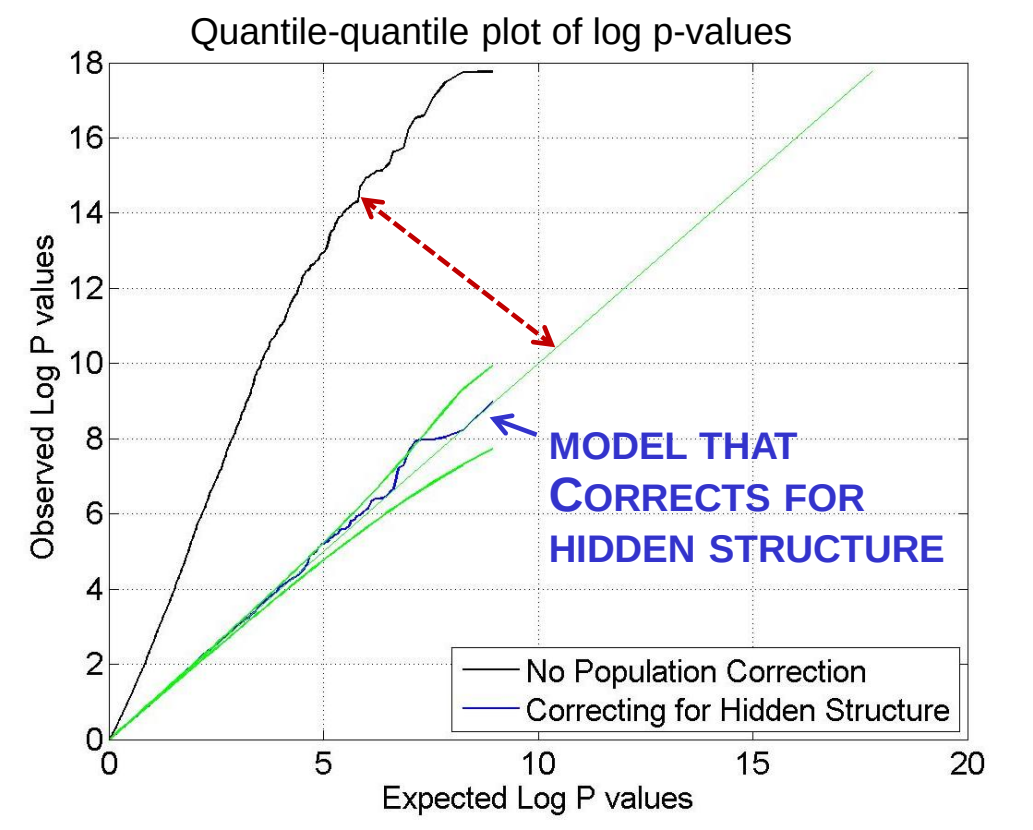


~7500 SNPs, ~1000 people,
contains multiple ethnicities and families

- Expect very few markers to truly be associated with disease.
- Distribution of p-values should be close to a uniform (diagonal on this plot).

Diagnosing Analyses w P-value Distributions

Evidence of improved analysis

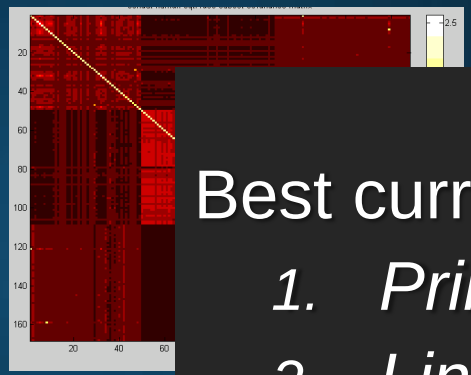


Data informs us of confounding structure

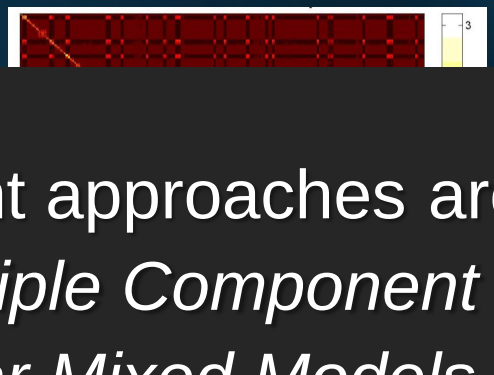
Use the large scale of the genetic markers themselves, in aggregate, to see how ‘similar’ every two people are, and incorporate this into the analysis.

“genetic similarity matrix ” (IBD/IBS/covariance)

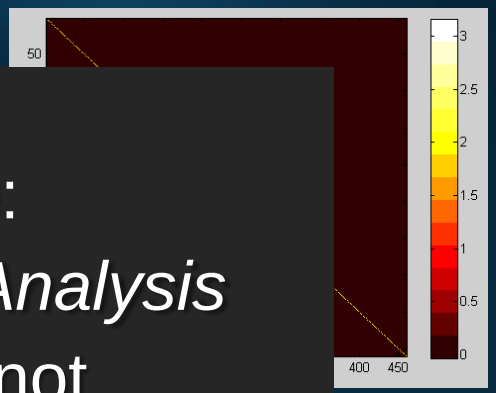
three ethnicities



families



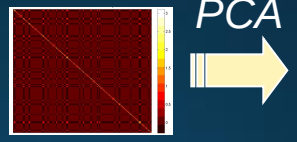
no structure



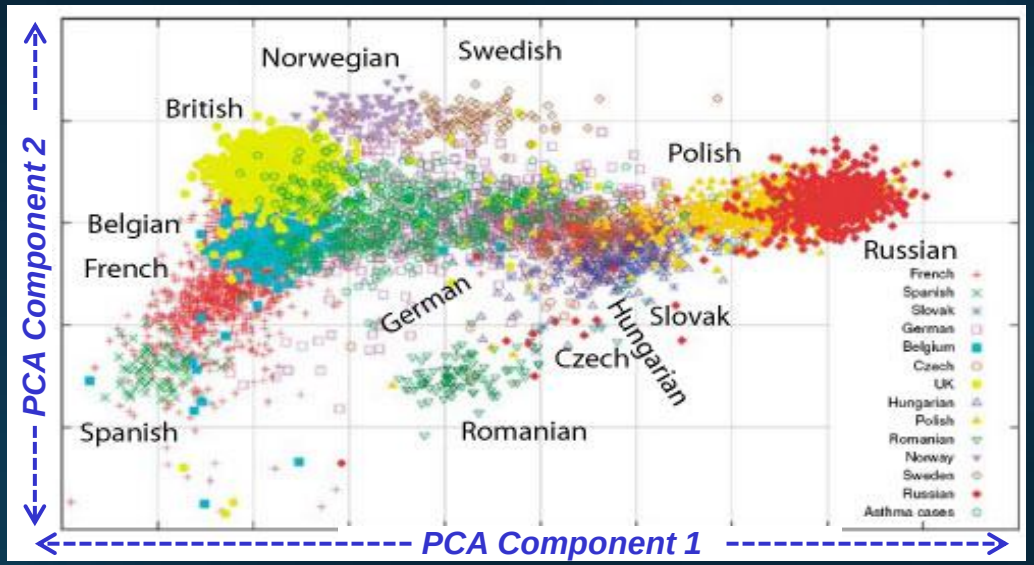
Best current approaches are:

1. *Principle Component Analysis*
2. *Linear Mixed Models* (not mixture models!)

Richness of Genetic Similarity Matrix



PCA

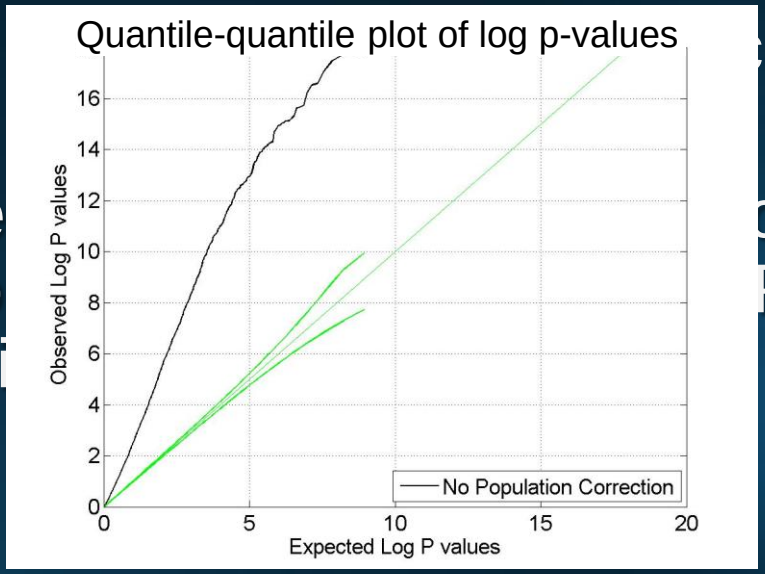


[Novembre et al, *Nature*, 2008]

Principle components correspond well to geographic map.

Naïve Approach

- Regress
- Evaluate



h genetic marker.

comparing this model to P (*e.g.* use LRT stati

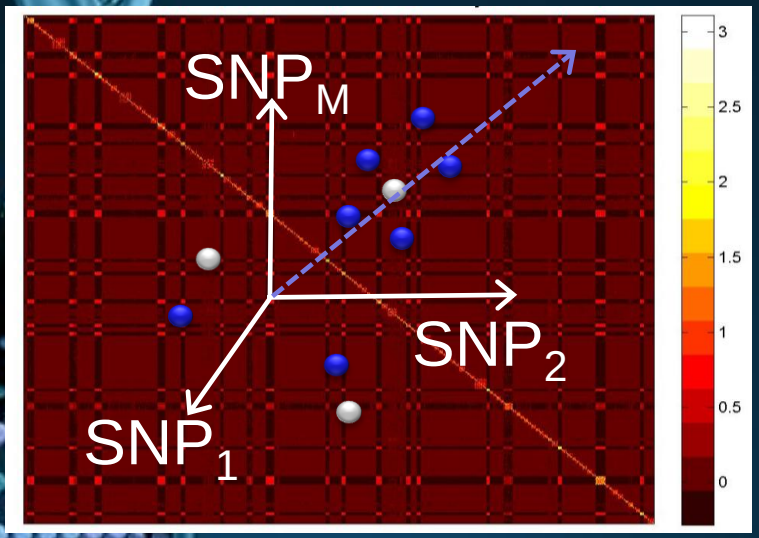
$$p(y_i) = N(X_i\beta, \sigma^2)$$ linear regression

phenotype (e.g. blood pressure level)

SNP, covariates and offset

learned regression weight (importance of SNP to blood pressure)

Principle Components Analysis Approach (aka *Eigenstrat* [Price *et al Nature Genetics 2006*])



genetic similarity between every two people

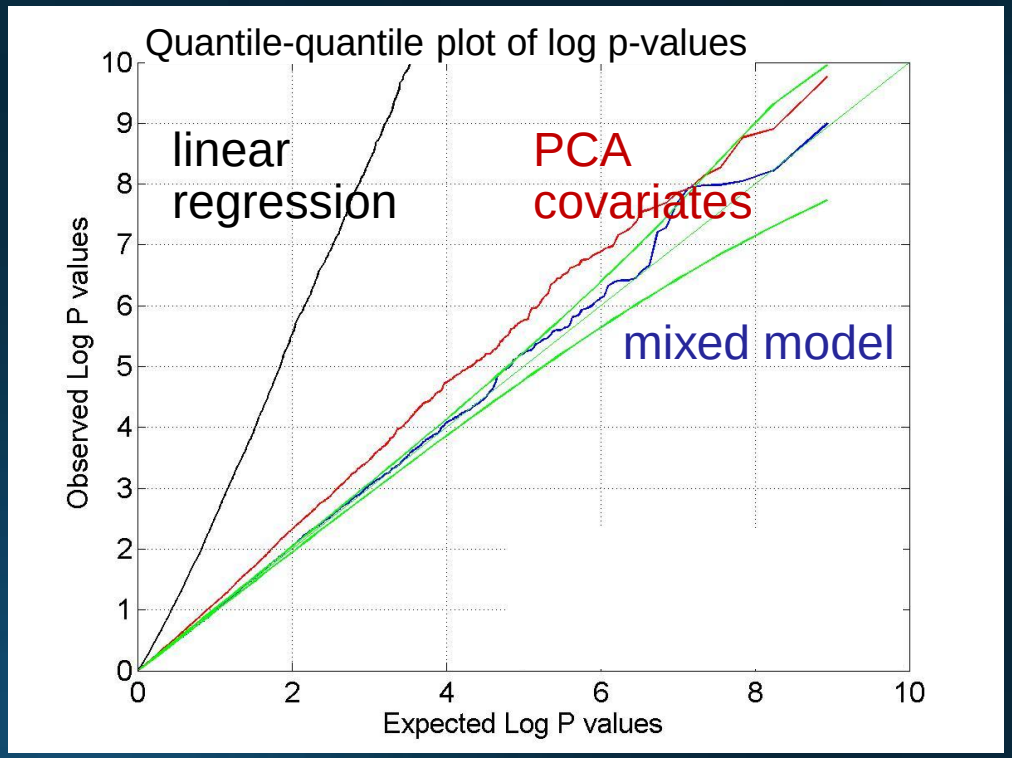
- Project each person's markers into the low dimensional space capturing confounders.
- Add projections as covariates to standard linear regression.

$$p(y_i) = N(X_i\beta + \sum_t d_i^t C_t, \sigma^2)$$

projection of i^{th} person onto t^{th} eigenvector

regression weight for t^{th} Principle Direciton

Performance on Real Data



Mixed model works better than PCA approach here.

GAW14 data set

~7500 SNPs,

~1300 people: variety of ethnicities (~900 Caucasian, ~200 Hispanic + ~200 African American) and families within these

Mixed Model Approach

- Do not reduce space to a set of directions: use the full structure in this space.
- Like standard linear regression, but with latent "ancestry" variables generated from a Gaussian in the "genetic similarity" matrix:

$$p(\vec{u}) = N(\vec{0}, \blacksquare)$$

"amount of confounding genetic signal"

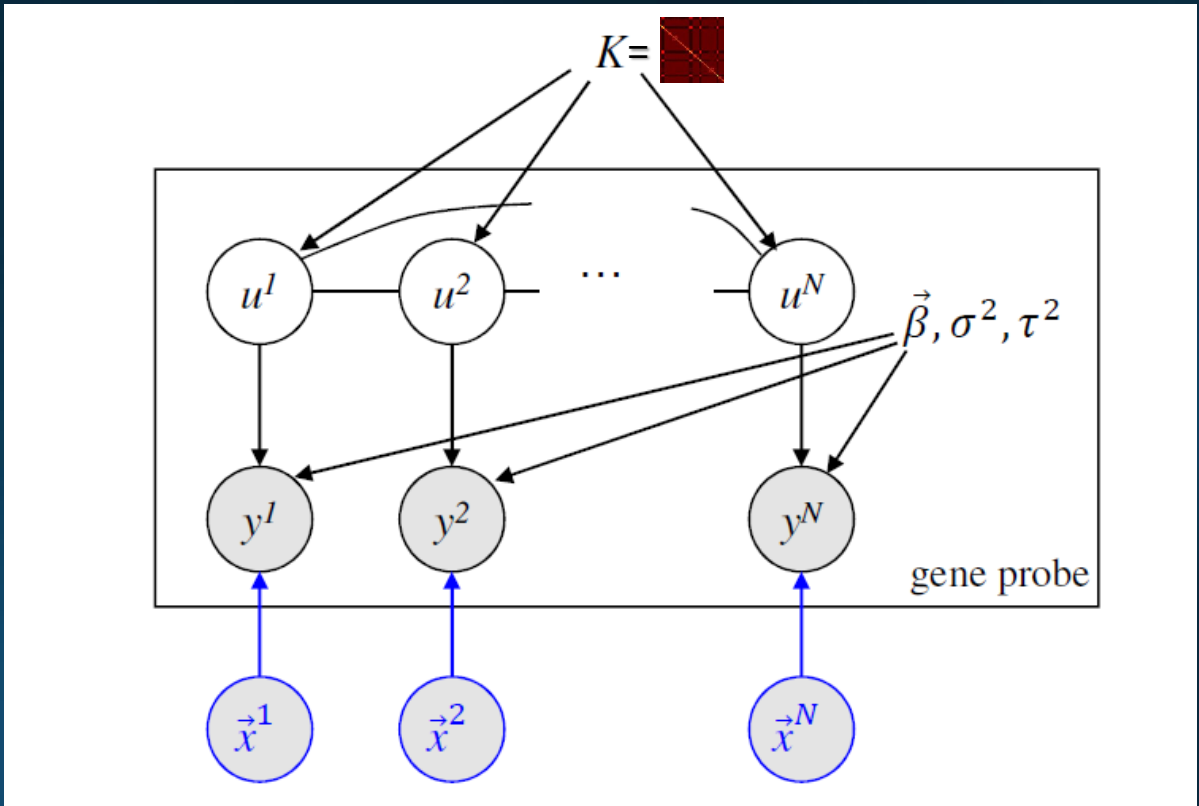
$$p(\vec{y}|\vec{u}) = N(X\beta + \eta\vec{u}, I\sigma^2)$$

LIKELIHOOD

$$p(\vec{y}) = \int p(\vec{y}|\vec{u})p(\vec{u})d\vec{u}$$

$$= N(X\beta, \eta\blacksquare + I\sigma^2)$$

Mixed Model Approach



$$p(\vec{u}) = N(\vec{0}, K)$$

$$p(\vec{y}|\vec{u}) = N(X\beta + \eta\vec{u}, I\sigma^2)$$

Efficient Parameter Learning in LMM

[Emma: Kang *et al Genetics* 2008]

Likelihood:

$$L = p(\vec{y}) = N(X\beta, \eta K + I\sigma^2), \quad K=$$ 

1. Re-parameterize L in terms of $\delta = \frac{\sigma}{\eta}, \beta, \eta$.
2. Plug in ML estimates $\hat{\beta} = \hat{\beta}(\delta)$, $\vec{\eta} = \vec{\eta}(\delta)$, and $\hat{\sigma}^2 = \hat{\sigma}^2(\delta)$.
3. Now ML is reduced to (non-convex) problem of finding ML value of δ.
4. Use SVD tricks so that $L = L(\delta)$ can be

Our recent paper on eQTL analysis

PNAS PNAS PNAS

Correction for hidden confounders in the genetic analysis of gene expression

Jennifer Listgarten[a,1], Carl Kadie[b], Eric E. Schadt[c], and David Heckerman[a,1]

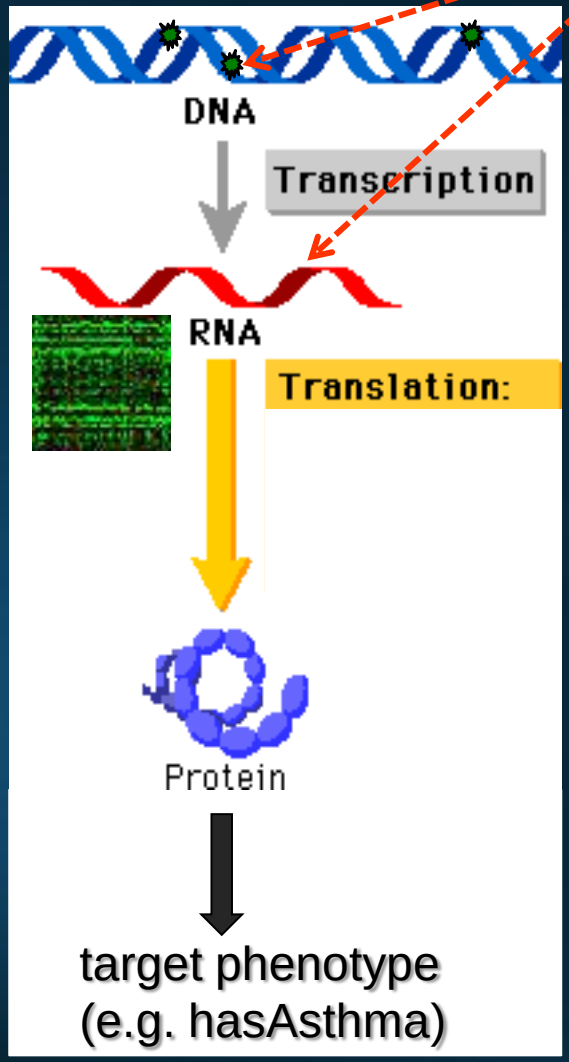
[a]Microsoft Research, 1100 Glendon Avenue, Suite PH1, Los Angeles, CA; [b]Microsoft Research, 1 Microsoft Way, Redmond, WA; and [c]Pacific Biosciences, 1505 Adams Drive, Menlo Park, CA

Edited* by David Haussler, University of California, Santa Cruz, CA, and approved July 21, 2010 (received for review February 26, 2010)

Joint work with:

- David Heckerman & Carl Kadie (Microsoft Research)
- Eric Schadt (Pacific Biosciences)

eQTL (expression Quantitative Trait Loci)



- Searching to find genetic markers (SNPs) associated with changes in gene expression (mRNA).

  {SNP} x {gene expression}

- *e.g.,* ask which of 1 million SNPs are correlated with which of 20K gene expression levels (1 million x 20K hypotheses)

eQTL Analysis *{SNP} x {gene expression}*

Can think of eQTL analysis in two ways:

1. One GWAS analysis for each gene probe.
2. One Gene Expression analysis for each SNP.

➡ Need to properly deal with:

i. Confounding by Population Structure and Family Relatedness (as required in GWAS)

ii. Confounding by e*xpression heterogeneity*: technical, environmental, demographic, or genetic factors [Leek & Storey *PLoS Genetics* 2007; Kang *Genetics* 2008.].

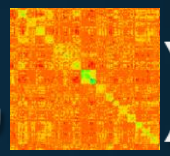
Double Correction Model: PS+EH

1. *SNP-space confounder coefficients*

$$p(\vec{u}) = N(\vec{0}, \quad)$$

2. *Gene-expression-space confounder coefficients*

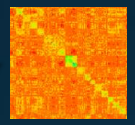
$$p(\vec{v}) = N(\vec{0}, \quad)$$



$$p(\vec{y}|\vec{u},\vec{v}) = N(X\beta + \eta_{PS}\vec{u} + \eta_{EH}\vec{v}, I\sigma^2)$$

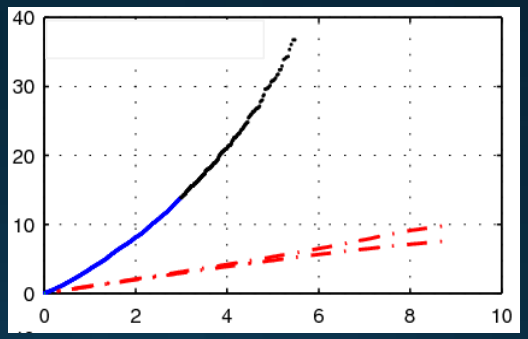
LIKELIHOOD

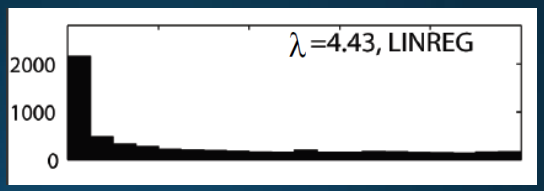
$$p(\vec{y}) = N(X\beta, \eta_{PS} \quad + \eta_{EH} \quad , I\sigma^2)$$



Results on Mouse Data

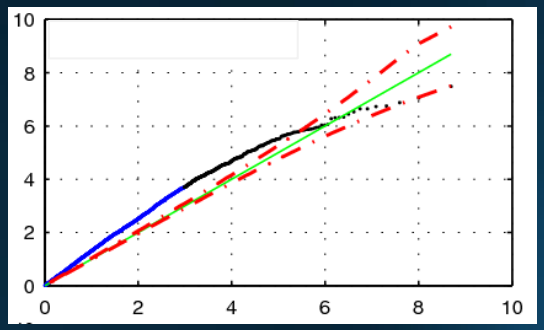
LINEAR REGRESSION

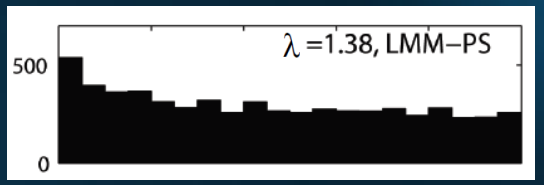




Serious inflation!

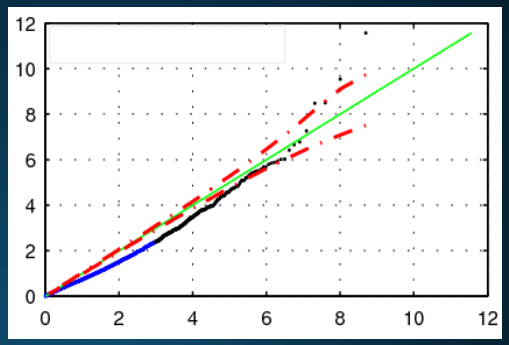
LMM-PS

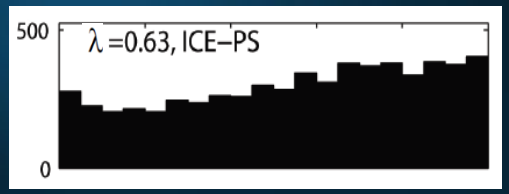




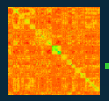
Still inflation.

LMM-ICE-PS





Deflation?!

What's going on? Generative model for gene expression with inconsistent estimator for model parameter  .

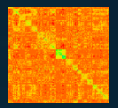
LIKELIHOOD

$$p(\vec{y}) = N(X\beta, \eta_{PS}\ \eta_{EH}\ , I\sigma^2)$$

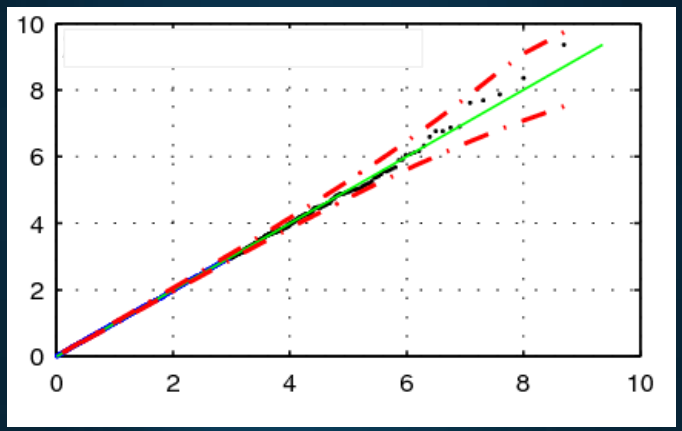
gene expression data

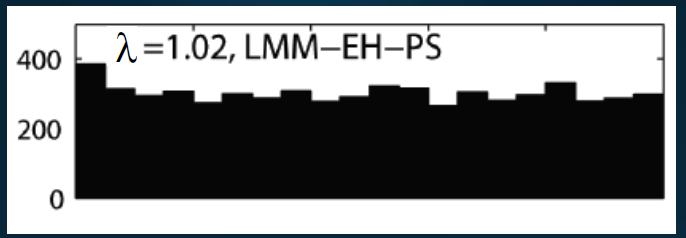
based on gene expression data

How to fix the ICE+PS model?

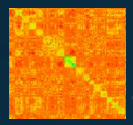
Instead of computing  ahead of time as a pre-processing step, learn it jointly with the other parameters of the model so that we have a *consistent* estimate of it.

LMM-EH-PS





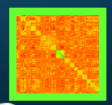
LIKELIHOOD

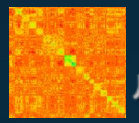
$$p(\vec{y}) = N(X\beta, \eta_{PS}$$  $$+ \eta_{EH}$$  $$, I\sigma^2)$$

Algorithm for Learning LMM Parameters

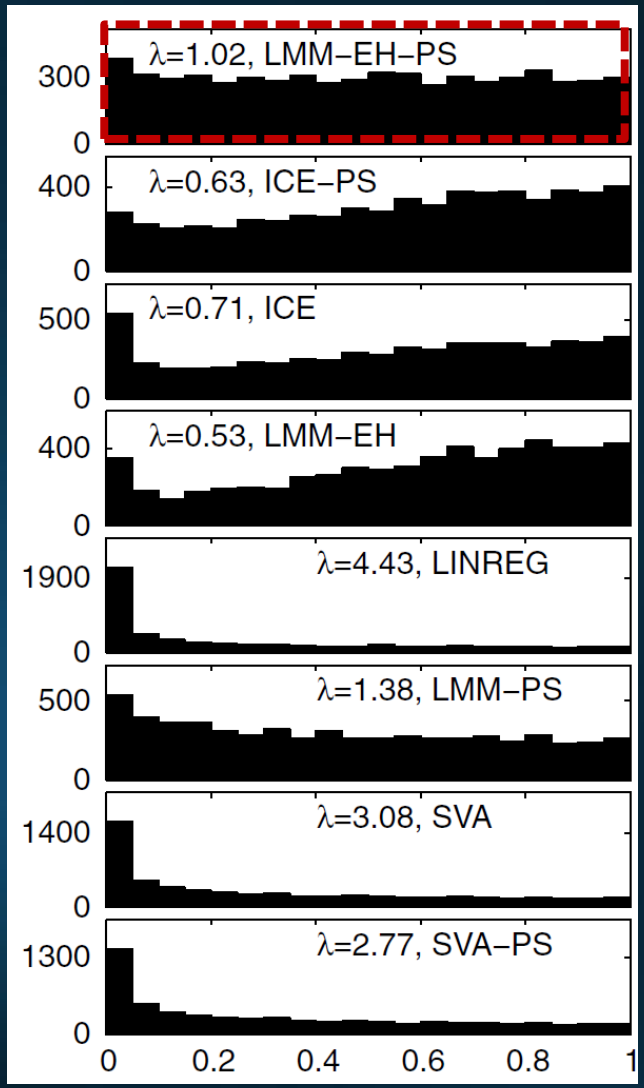
Coordinate Ascent:

a) Find Maximum Likelihood "standard" LMM parameters $(\beta, \eta_{PS}, \eta_{EH}, \sigma^2)$ by numerical optimization, assuming known  .

b) Find Maximum Likelihood value for  assuming known values for $(\beta, \eta_{PS}, \eta_{EH}, \sigma^2)$. Uses Expectation Maximization.

$$p(\vec{y}) = N(X\beta, \eta_{PS}$$  $$+ \eta_{EH}$$  $$, I\sigma^2)$$

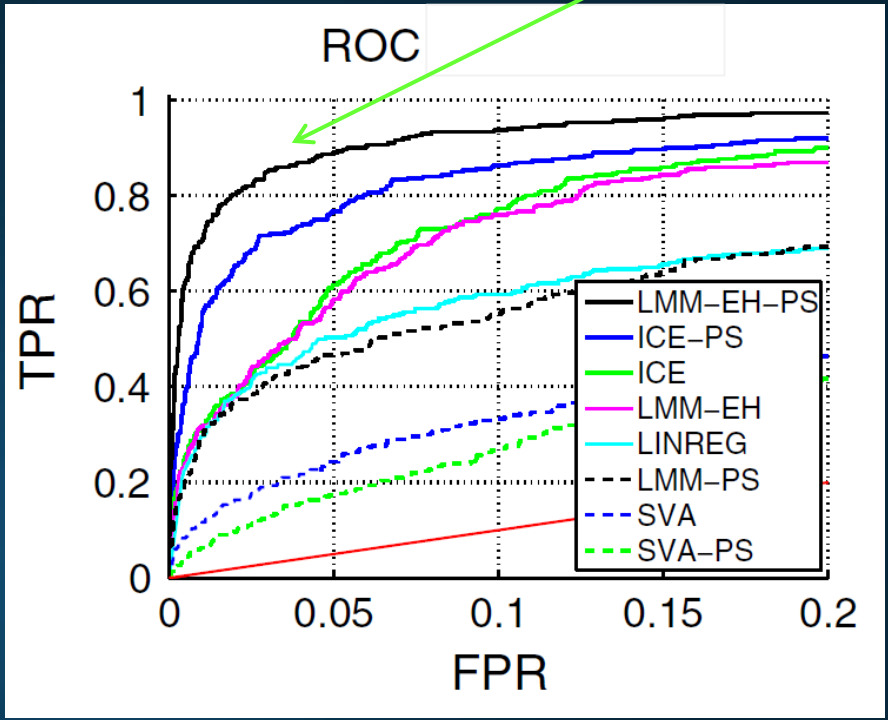
Comprehensive Comparison



Only our model has a reasonable p-value distribution on the real data.

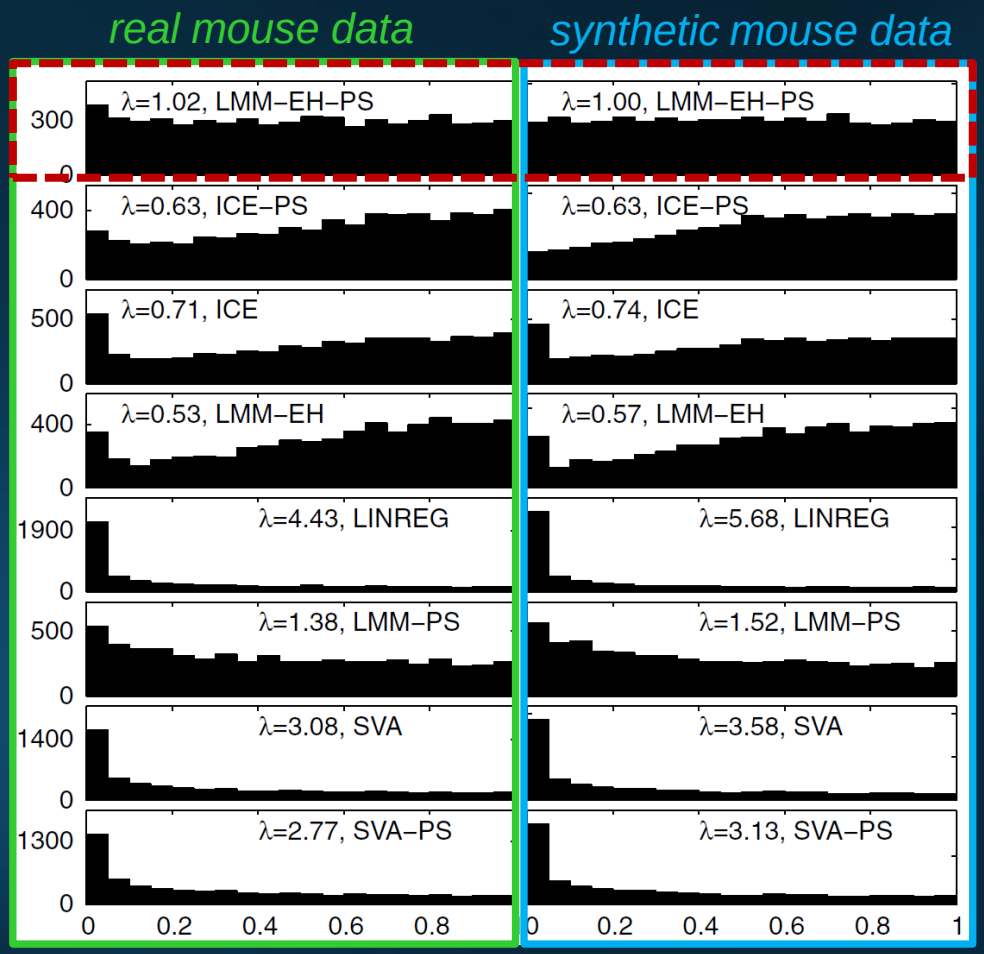
What about power?

Our model achieves maximum power.



- This evaluation is based on synthetic data generated from our model, using same parameters as on the real data (so realistic).
- No gold standards exist, so must rely on synthetic. Bronze standard on real data to follow.

How representative is our synthetic data?



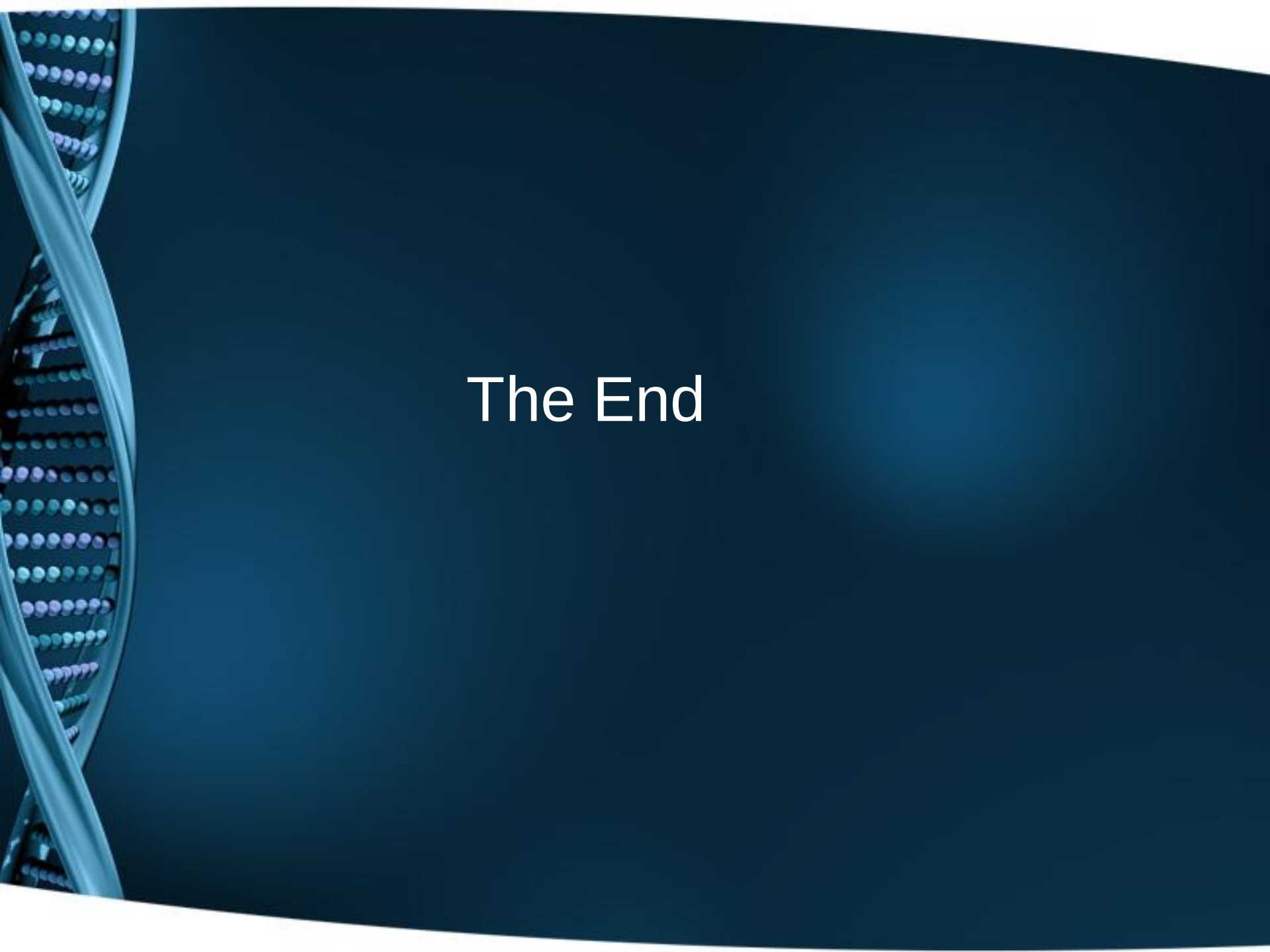
- Striking similarity between histogram of p-values on real data and synthetic data.
- Suggests synthetic data is representative of the real data.

Bronze Standard Power Evaluation

- Assume that “cis” eQTL should be picked up by the model more strongly than “trans” eQTL.
- Test all pairs of models to see which model is better than each other model:

Table 2. Mouse cis-enrichment, a starred *p*-value indicates that the starred model outperforms the nonstarred model

mouse	ICE-PS	LMM-EH	ICE	LMM-PS	LINREG	SVA	PS-SVA
LMM-EH-PS*	1.80e−014*	<1e−16*	<1e−16*	1.89e−008*	2.22e−015*	<1e−16*	<1e−16*
ICE-PS*		4.02e−14*	<1e−16*	0.163*	4.97e−5*	<1e−16*	<1e−16*
LMM-EH*			0.497	0.403*	0.1959*	<1e−16*	<1e−16*
ICE*				0.106	0.390*	<1e−16*	2.22e−16*
LMM-PS*					8.88e−5*	<1e−16*	<1e−16*
LINREG*						<1e−16*	2.44e−14*
SVA*							<1e−16


The End