Exploiting high dimensionality in big data

David Heckerman

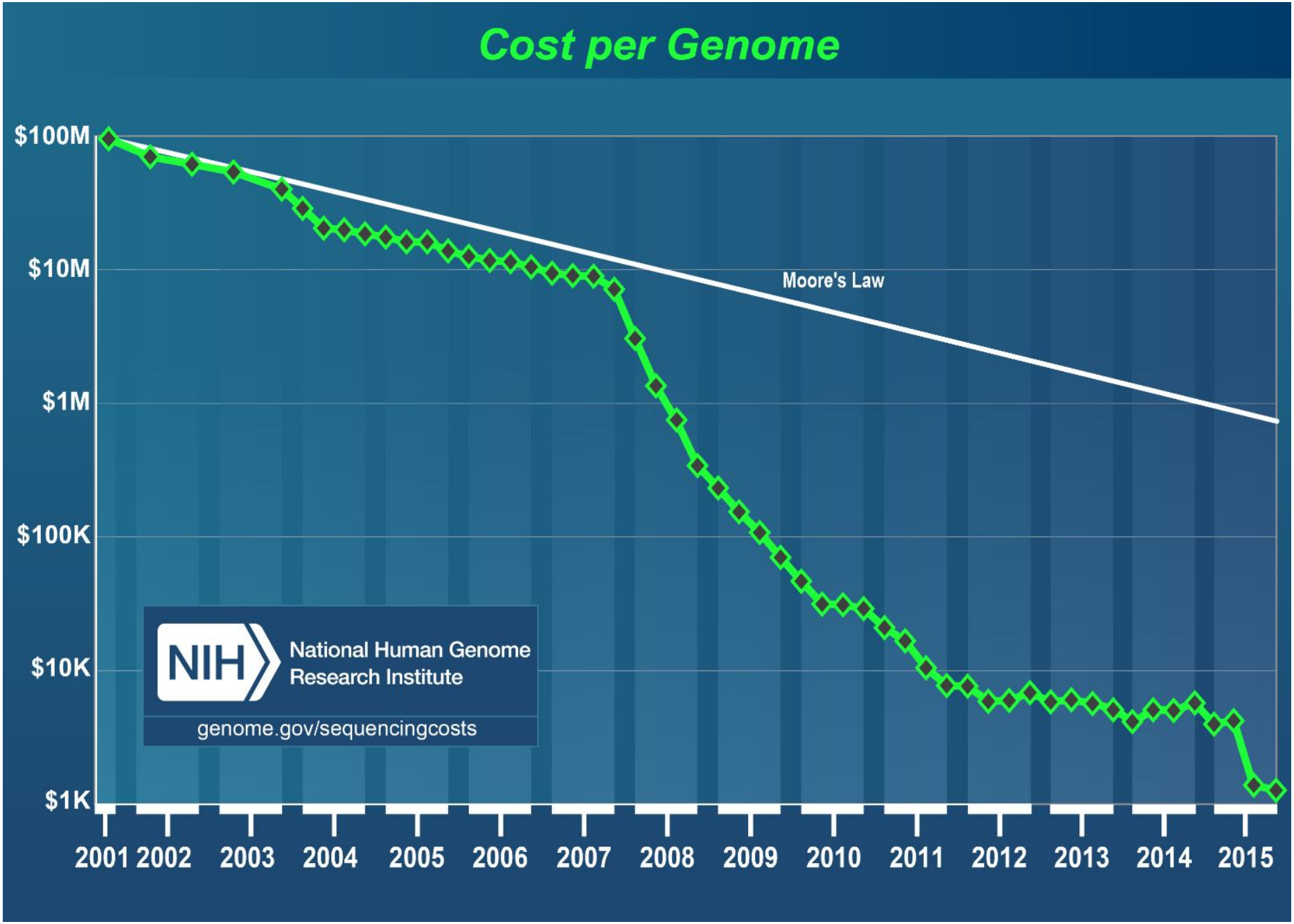
(while at Microsoft Research)

Theme #1: High dimensionality

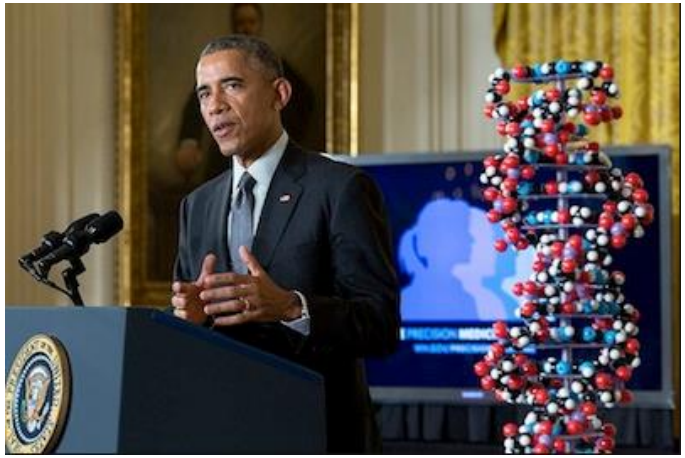
- Data can be big in two ways:
  sample size and dimensionality
- Large samples are a blessing
- High dimensionality is a “curse”
- This talk: blessings of high dimensionality
- Example: causal discovery on genomic data
  sample size and dimension ~1M

Theme #2: Causal discovery

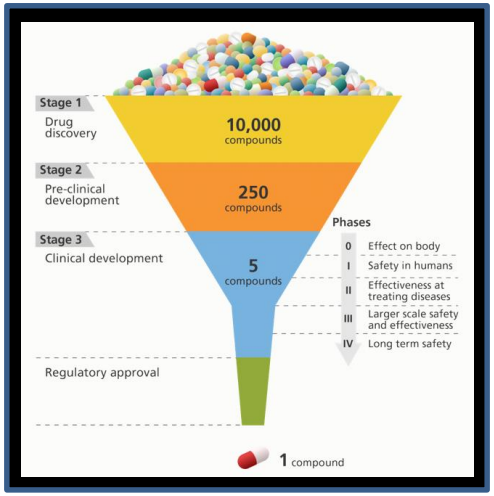
- Knowledge of cause and effect helps us predict in the face of intervention
- Classical causal discovery is done through intervention (e.g., randomized trials)
- Can we discover cause and effect without interventions?
- Very tricky, as we shall see


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

Genomics applications



https://www.whitehouse.gov/precision-medicine



yourgenome.org

https://www.entm.purdue.edu/conference/

http://blog.illumina.com/blog

http://www.genengnews.com/gen-news-highlights

Genomics 101

- Human genome: 3 billion base pairs (A/C/T/G) x 2
- Only about 0.1% difference in genome between individuals
- Only millions of differences
  - SNPs (single nucleotide polymorphisms)
    - 0, 1, or 2 copies of lest frequent allele
  - Insertions
  - Deletions
  - Copy number variations

An exercise in causal discovery without intervention: Genome-wide association studies (GWAS)

Which genetic markers (e.g., SNPs) cause

- Disease
- Favorable reaction to a drug
- Bad reaction to a drug
- Various physical abilities

An exercise in causal discovery without intervention: Genome-wide association studies (GWAS)

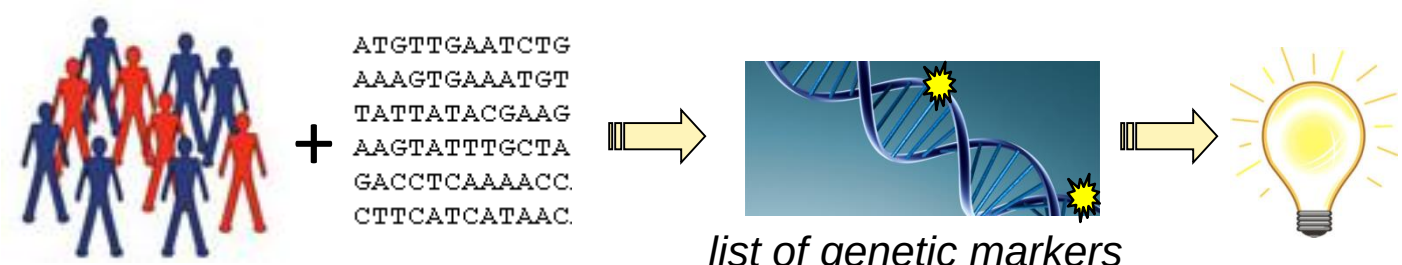
<u>Input</u>:

A set of individuals each with measurements of

- A trait (e.g., have disease X?)
- Set of genetic markers (e.g., SNPs)

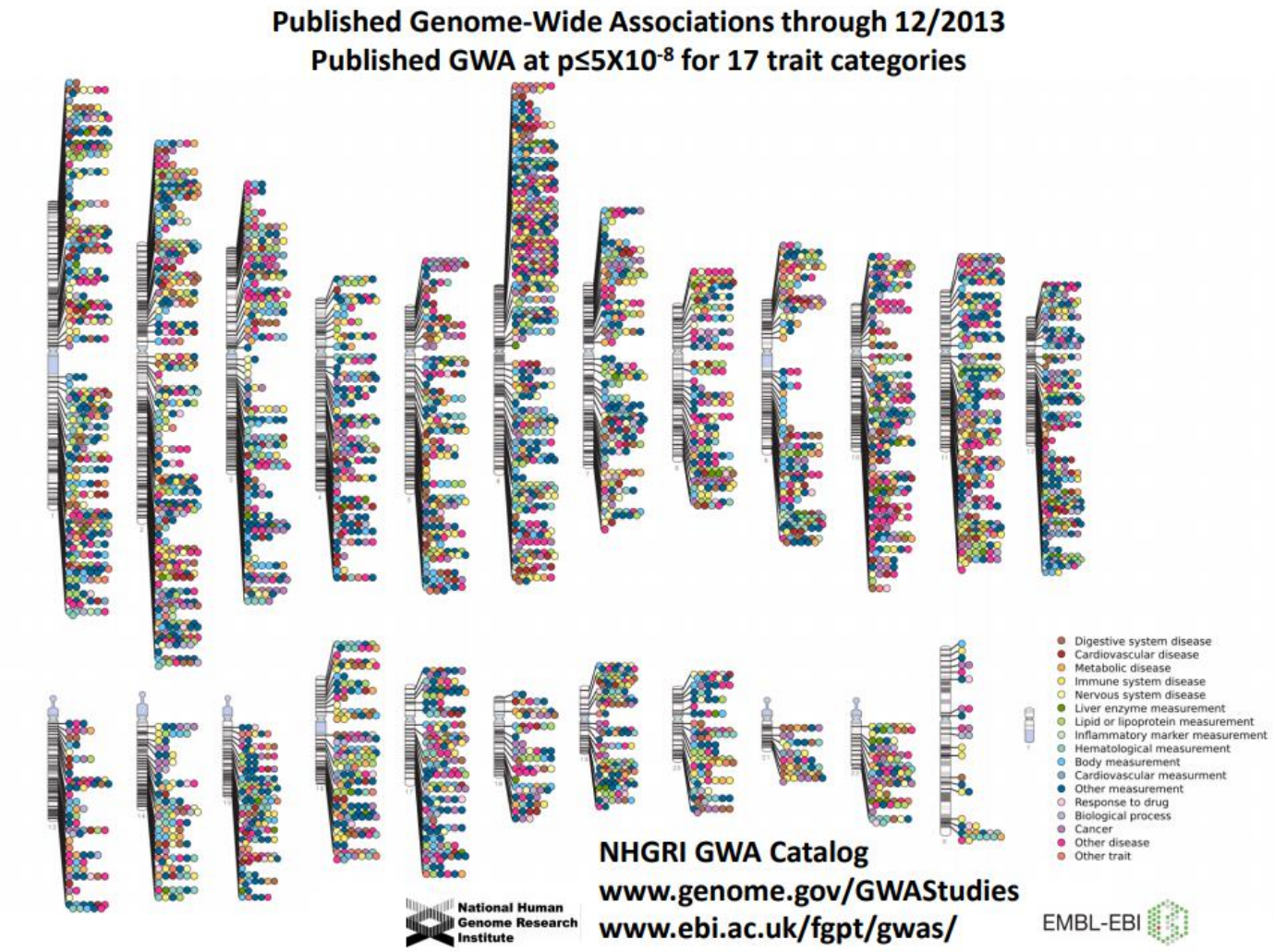
<u>Output:</u>

A list of genetic markers that <u>cause</u> the trait.



DNA from set of individuals

GWAS findings



Large sample sizes are essential for GWAS

- Many interesting traits are highly polygenic, with each SNP/variant playing only a small role

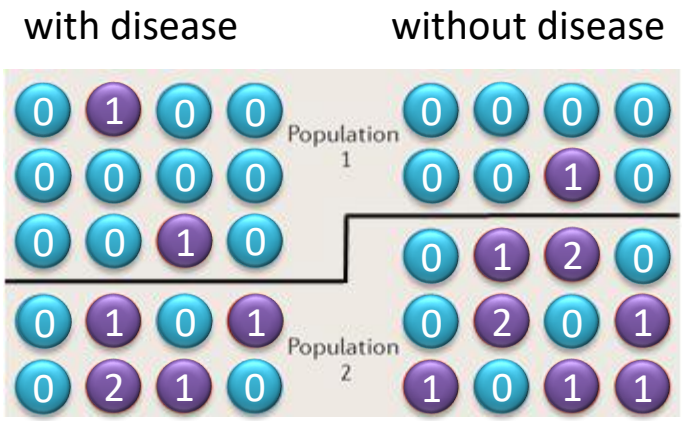
- Sample sizes in the hundreds-of-thousands to millions are needed to uncover a comprehensive picture of genetic influence

Large sample sizes are essential for GWAS

- Many interesting traits are highly polygenic, with each SNP/variant playing only a small role
- Sample sizes in the hundreds-of-thousands to millions are needed to uncover a comprehensive picture of genetic influence
- Large samples tend to be confounded by population structure and family relatedness

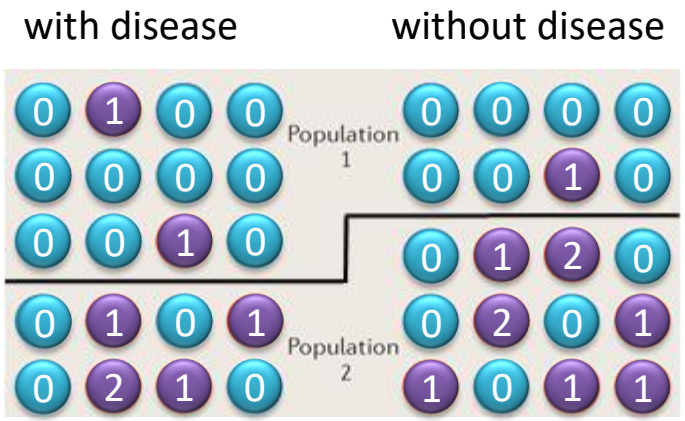
Example confounder: Population structure

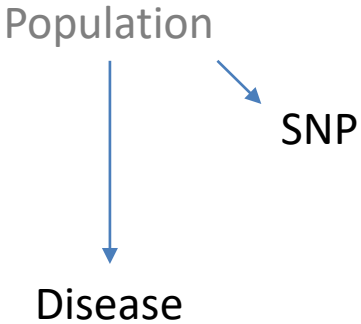
- Consider a single SNP with values 0, 1, and 2 copies of less frequent allele
- Suppose SNP does NOT cause the disease
- Suppose there are two populations that differ in frequency of the SNP
- Suppose there are more individuals from population 1 with the disease
- Ignoring population, there is a correlation between SNP and trait – the correlation is non-causal or "spurious"



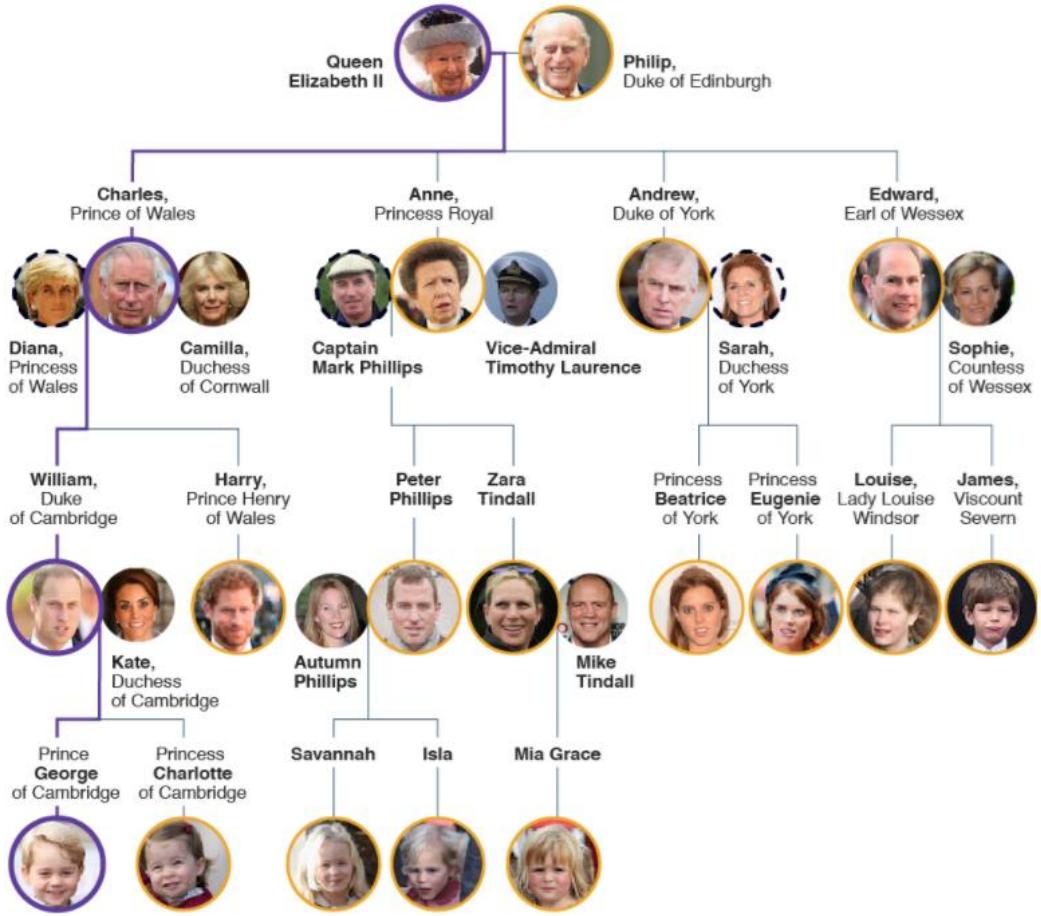
Example confounder: Population structure

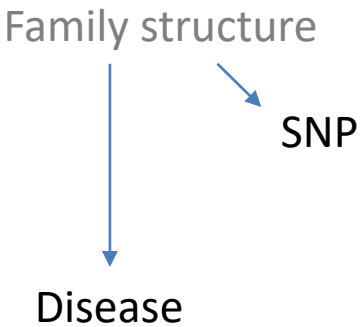
- Consider a single SNP with values 0, 1, and 2 copies of less frequent allele
- Suppose SNP has no effect on a disease
- Suppose two populations differ in frequency of the SNP
- Suppose there are more individuals from population 1 with the disease
- Ignoring population, there is a correlation between SNP and trait – the correlation is “spurious”





Another confounder: Family structure





Blessings of high dimension (lots of SNPs)

- Helps identify confounding
- Helps correct for confounding

Identifying confounding

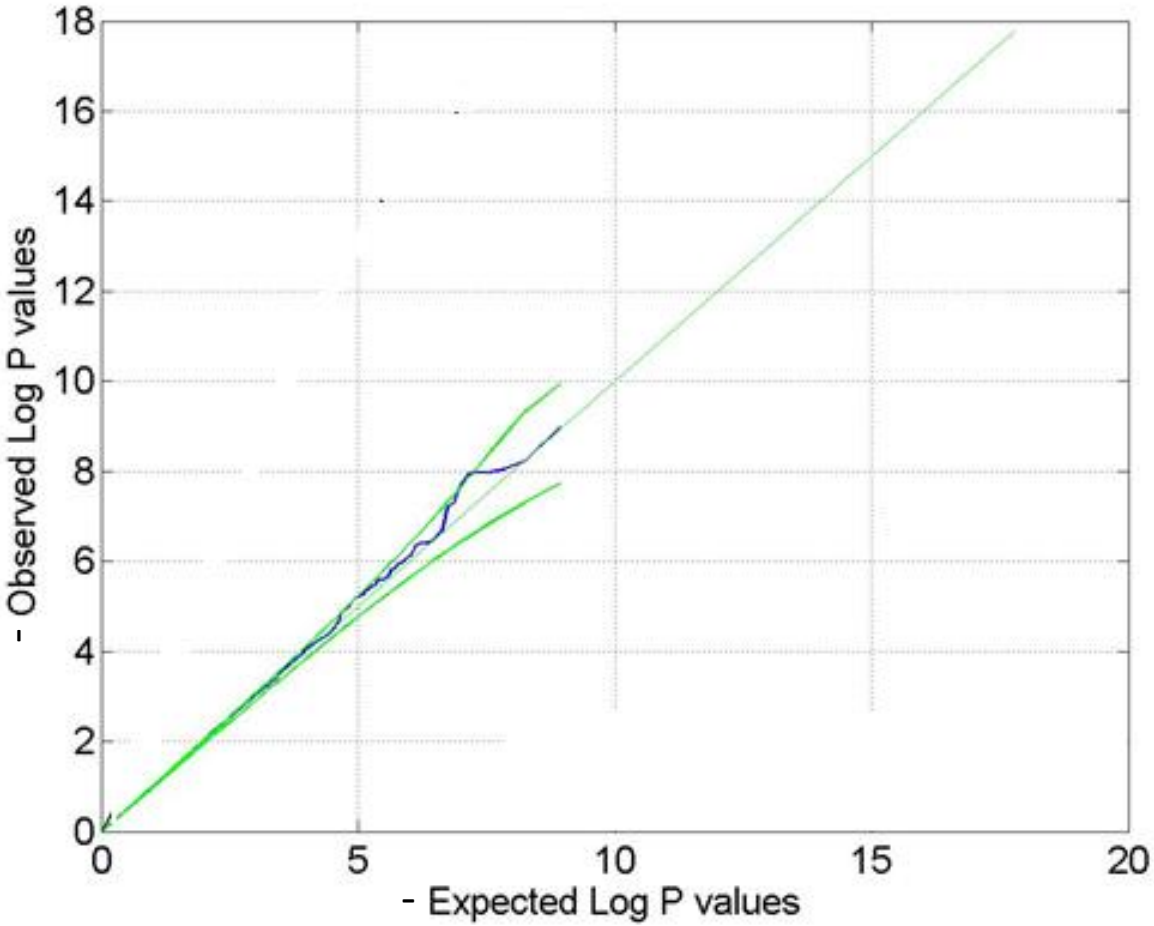
<u>QQ plot:</u>

- For every SNP, compute a P value for the association between the SNP and the trait
- Sort these P values
- Plot observed P values vs expected P values

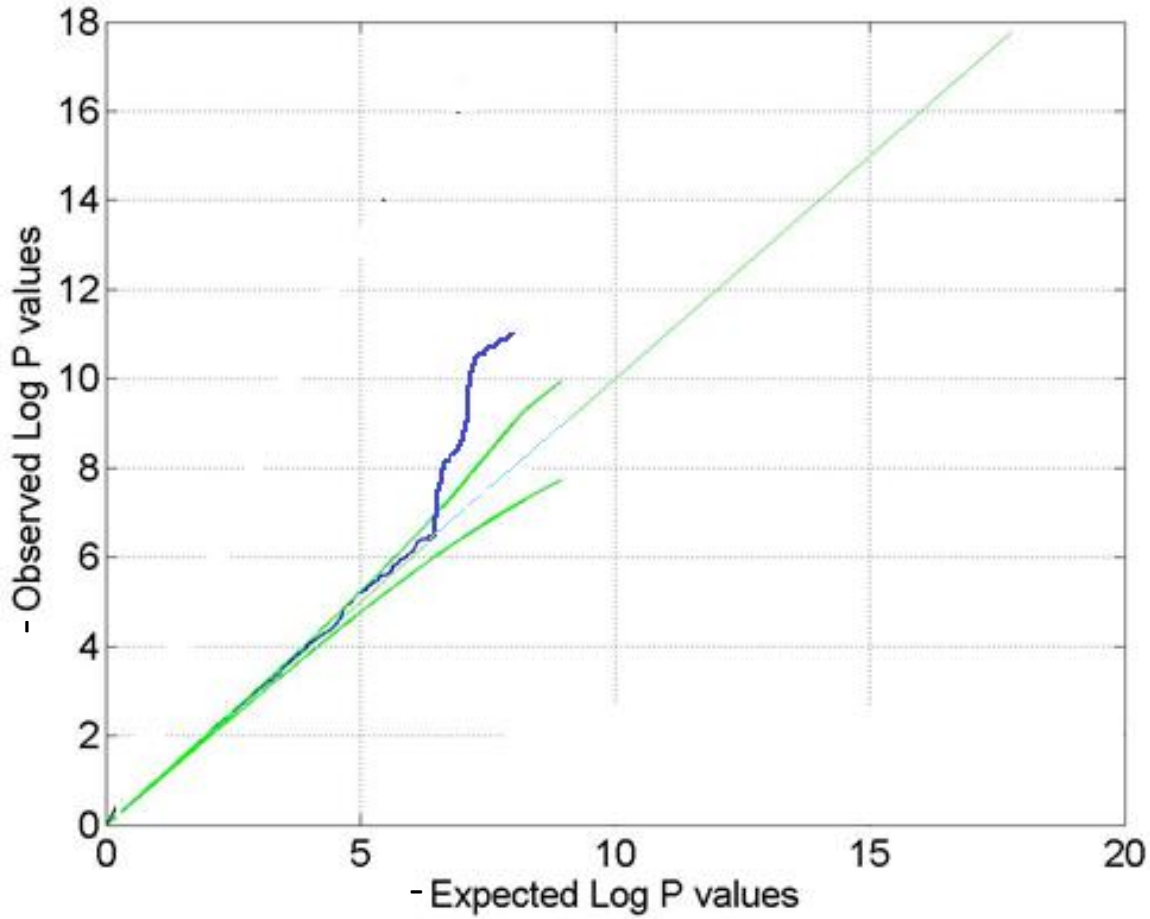
P-value: The probability that, when the null hypothesis is true, the test statistic would be greater than the statistic on the observed data.

When testing null hypotheses, P values should be uniformly distributed.

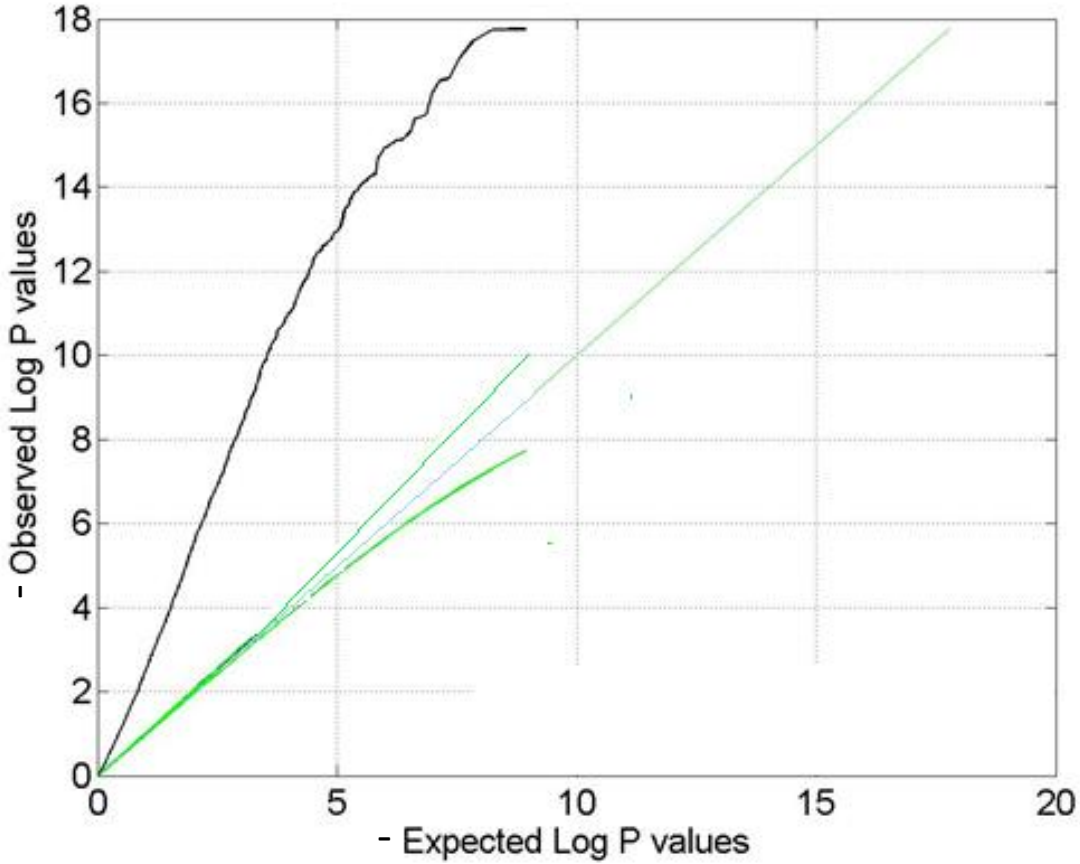
QQ plot; “null” data, no confounding



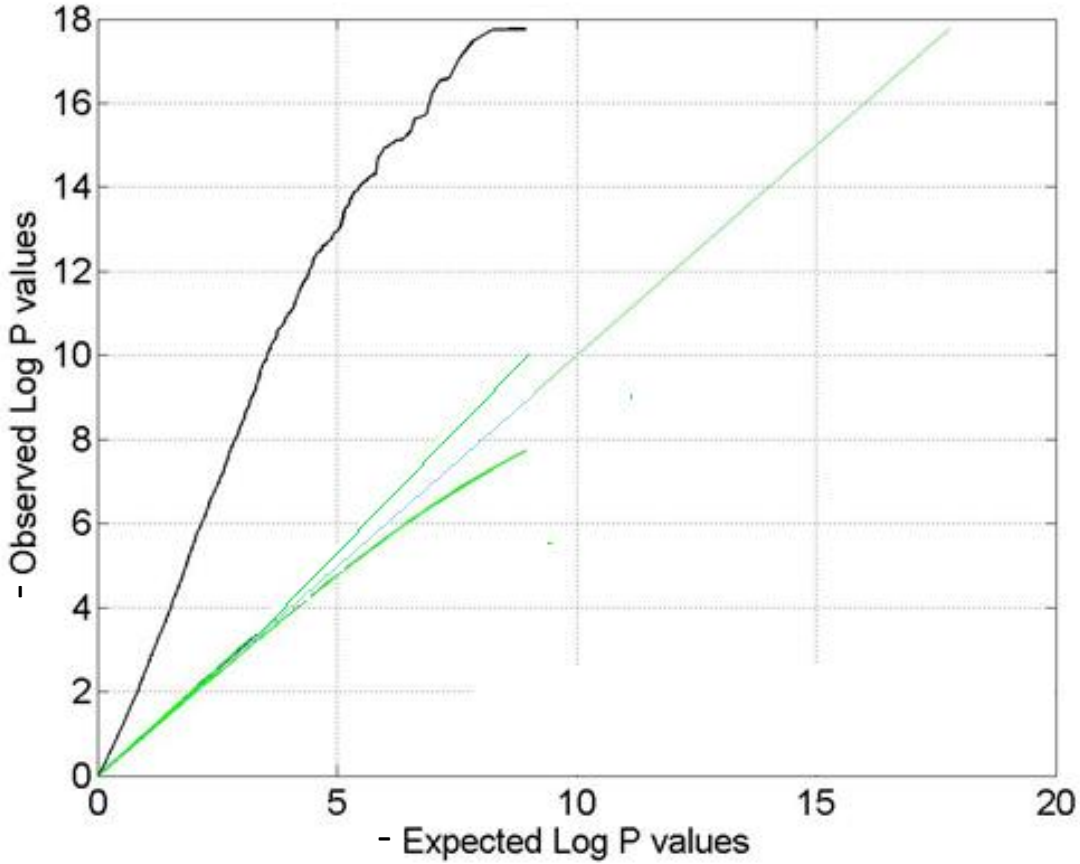
QQ plot; some causal SNPs, no confounding



QQ plot; confounding



QQ plot; confounding

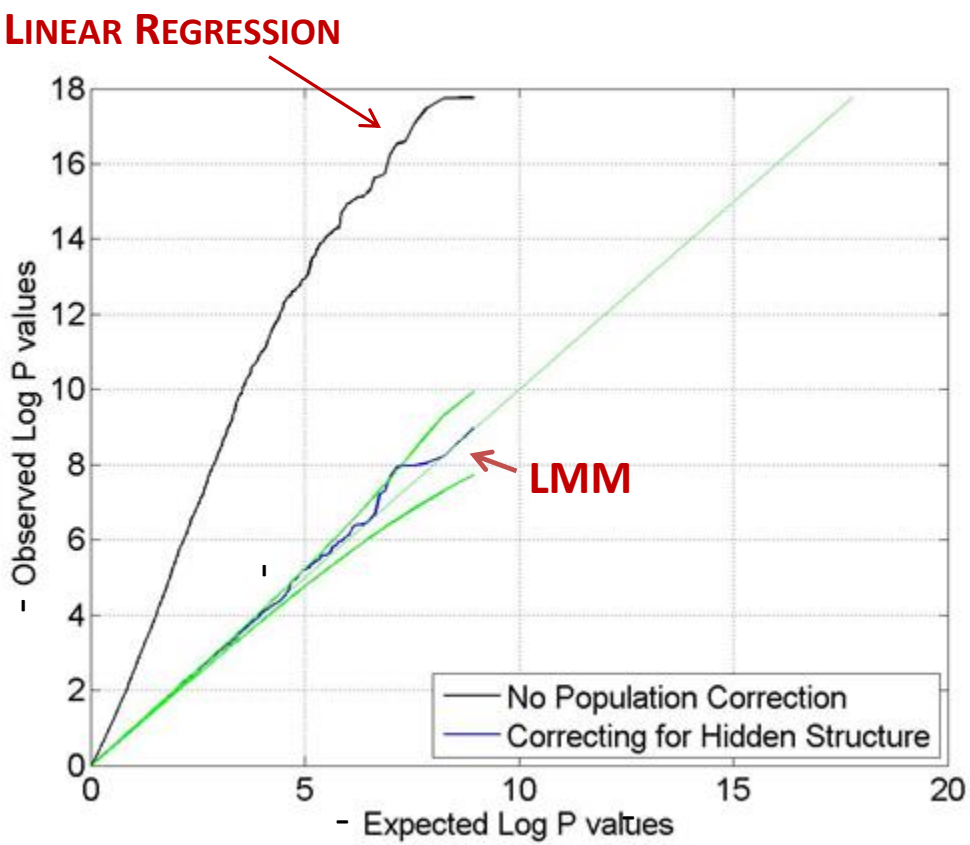


Works well when there are lots of SNPs & P values

Blessings of high dimension (lots of SNPs)

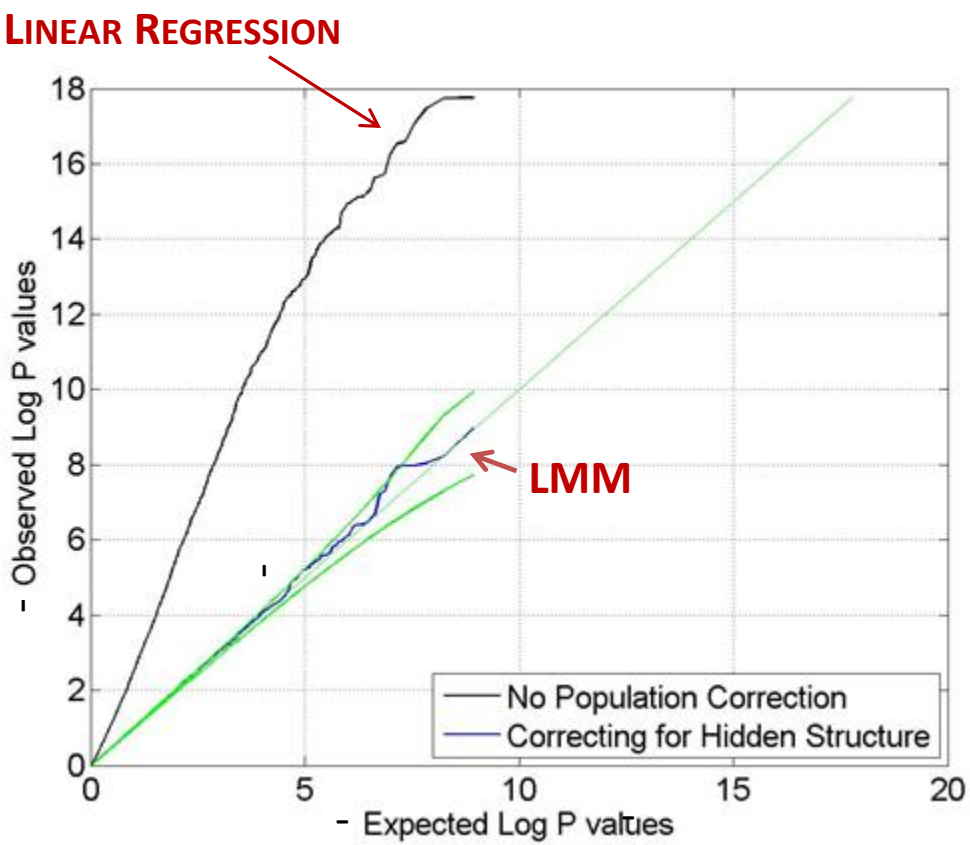
- Helps identify confounding
- <u>Helps correct for confounding</u>

Correcting for confounding: Linear Mixed Models



- Data: From Genetic Analysis Workshop 14
  - About 1300 individuals
  - About 8000 SNPs
  - Multiple ethnicities, related individuals

Correcting for confounding: Linear Mixed Models



- Data: From Genetic Analysis Workshop 14
  - About 1300 individuals
  - About 8000 SNPs
  - Multiple ethnicities, related individuals

Works well when there are lots of SNPS & P values

What is an LMM and why does it correct so well?

ACM Transactions on
Intelligent Systems and Technology (TIST)

publishes the highest quality papers on intelligent systems, applicable algorithms and technology with a multi-disciplinary perspective

Towards accounting for hidden common causes when inferring cause and effect from observational data

Association for Computing Machinery

SCIENTIFIC REPORTS

OPEN

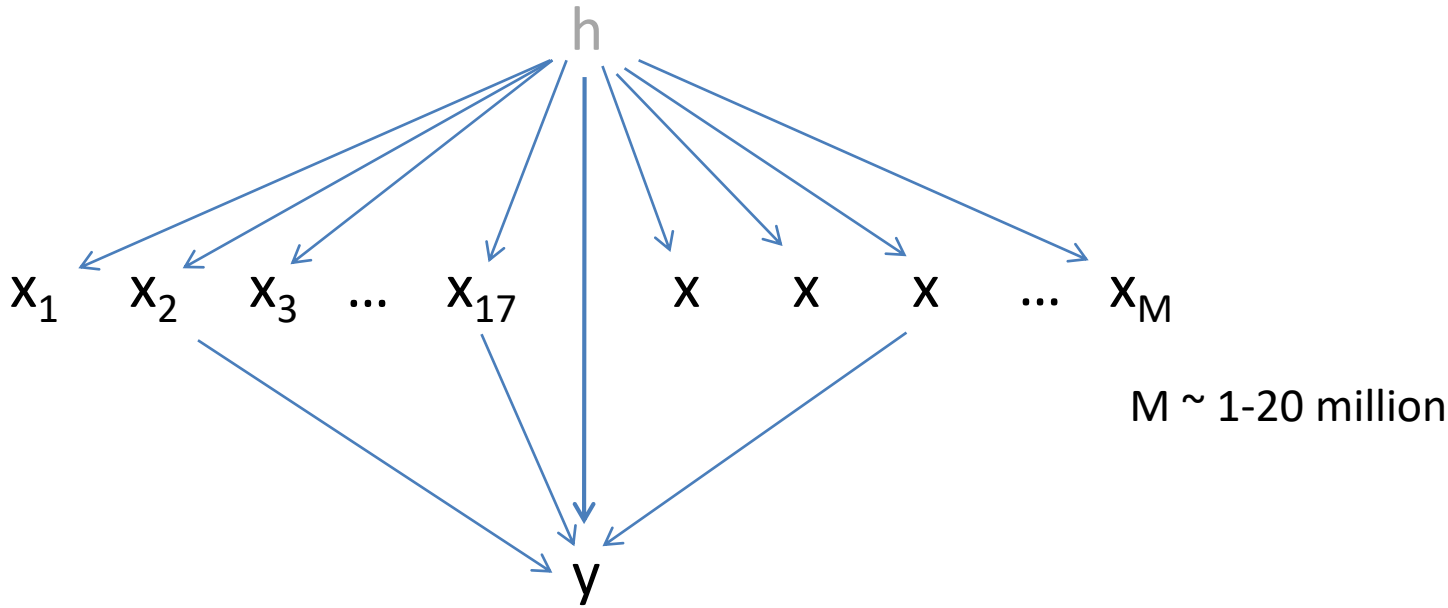
SUBJECT AREAS:
MACHINE LEARNING
GENOME INFORMATICS

Received
9 August 2013

Further Improvements to Linear Mixed Models for Genome-Wide Association Studies

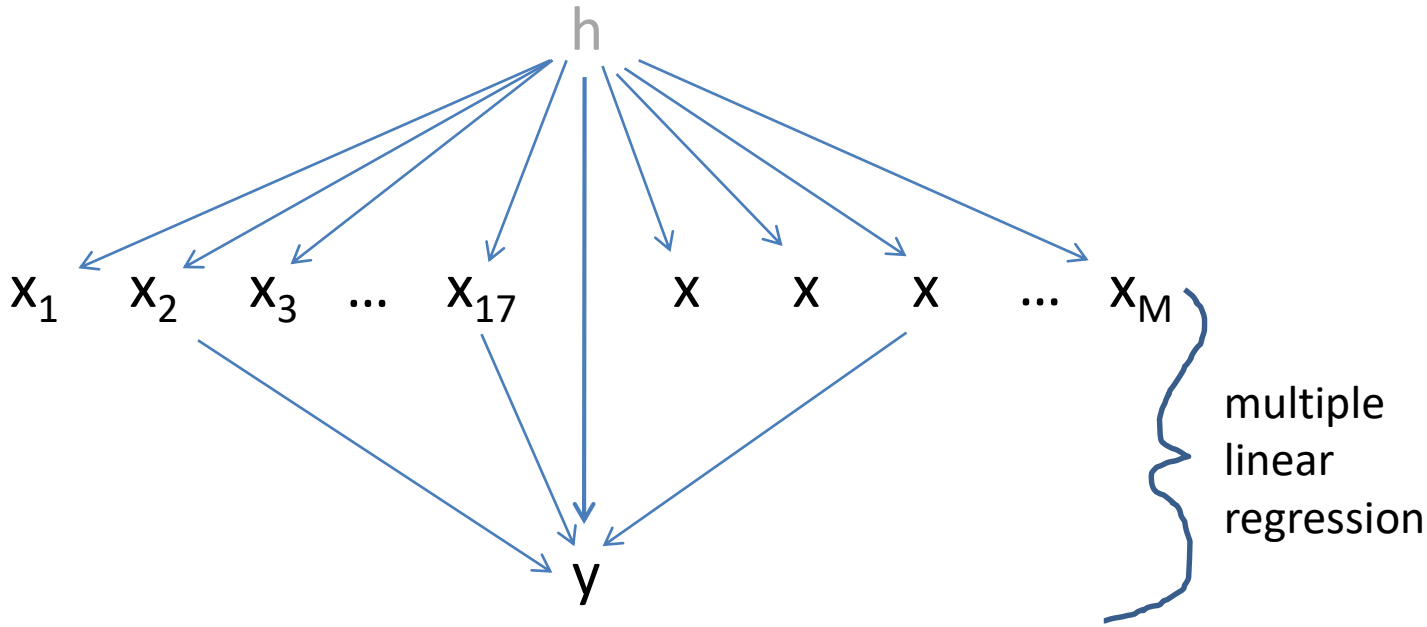
Christian Widmer[1]*, Christoph Lippert[1]*, Omer Weissbrod[2], Nicolo Fusi[1], Carl Kadie[3], Robert Davidson[3], Jennifer Listgarten[1] & David Heckerman[1]*

A graphical causal model for GWAS

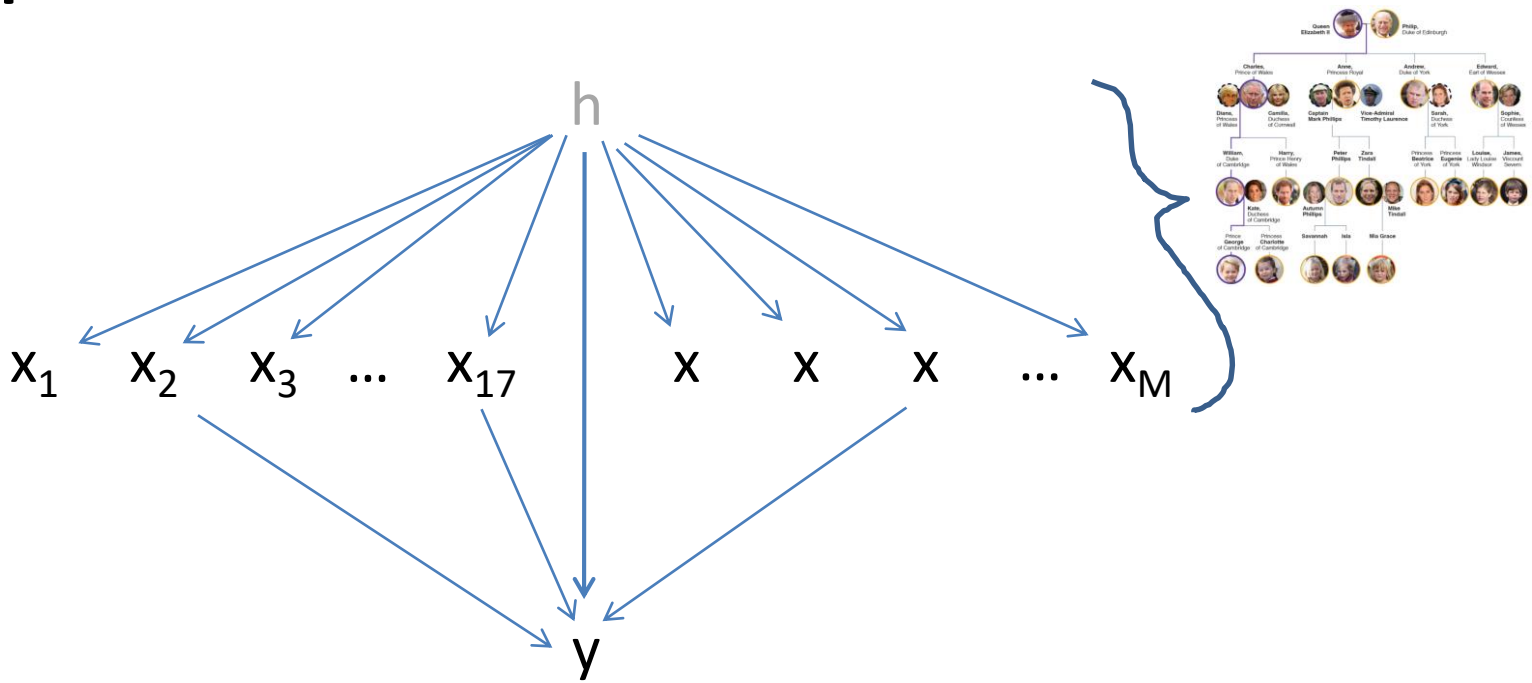


y is some trait of interest (e.g., disease?)
h represents hidden common causes of the SNPs

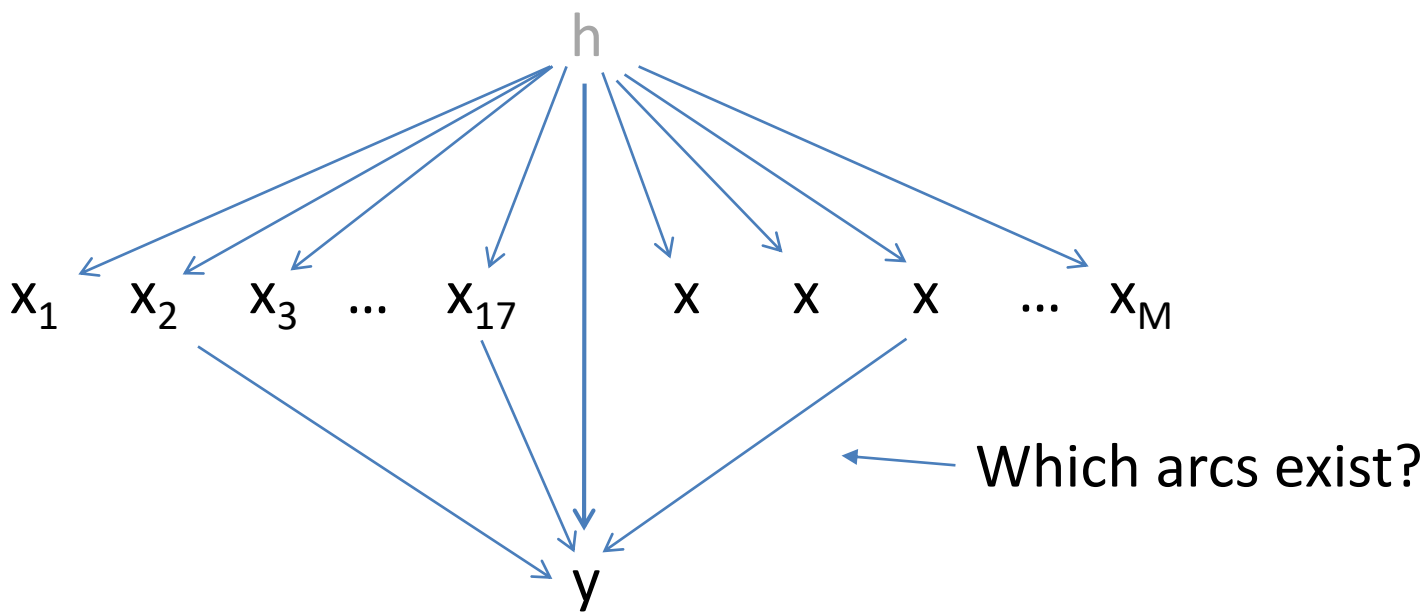
A graphical causal model for GWAS



A graphical causal model for GWAS



Causal discovery without intervention

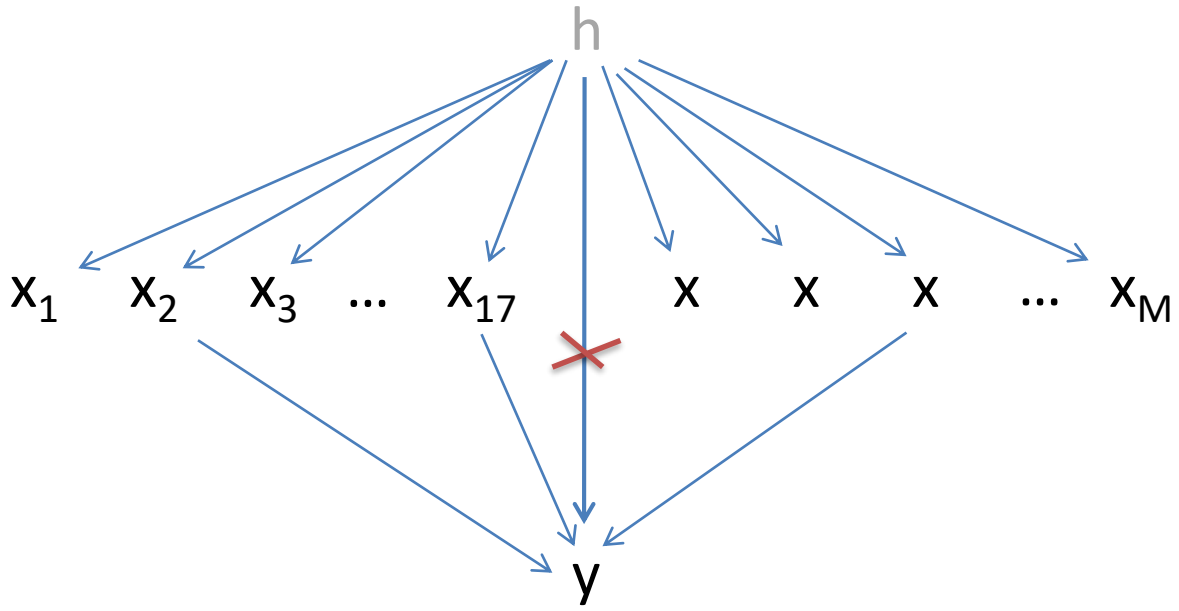


Acknowledge that false positives (type 1 errors) will exist

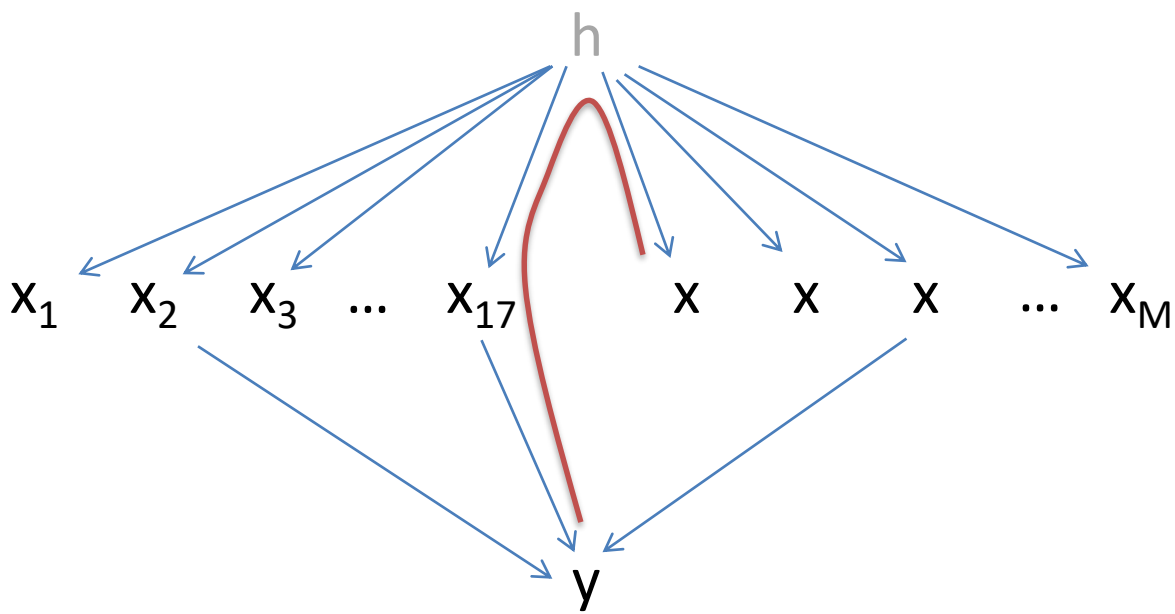
Manage with P values

- Null hypothesis: there is no arc from x->y
- Reject the null when P value < threshold α

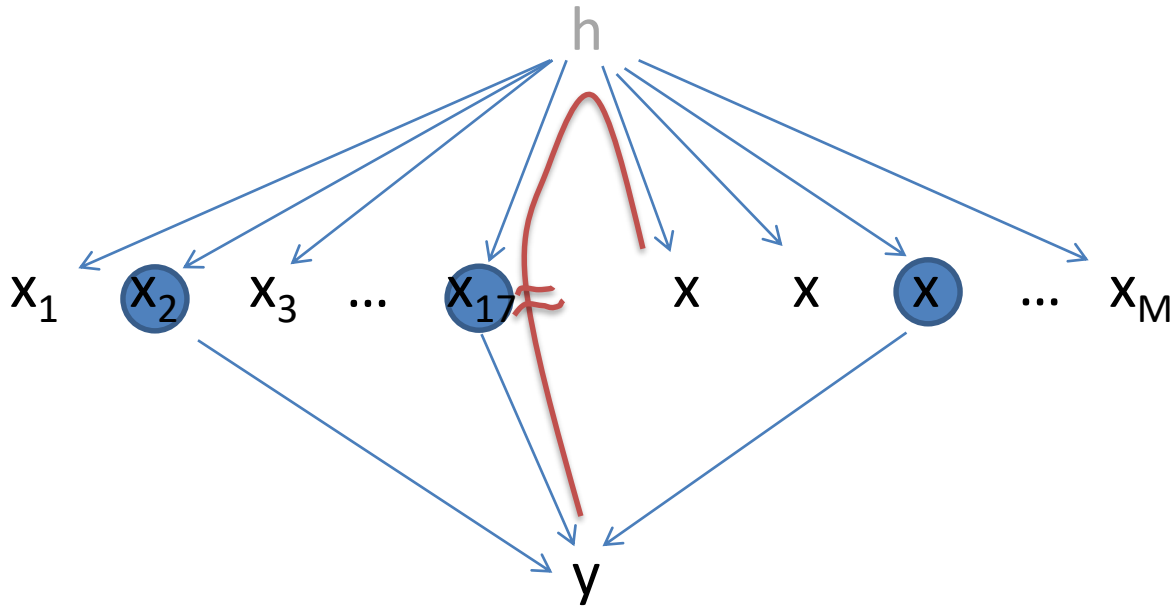
First, consider a simpler case



Spurious associations



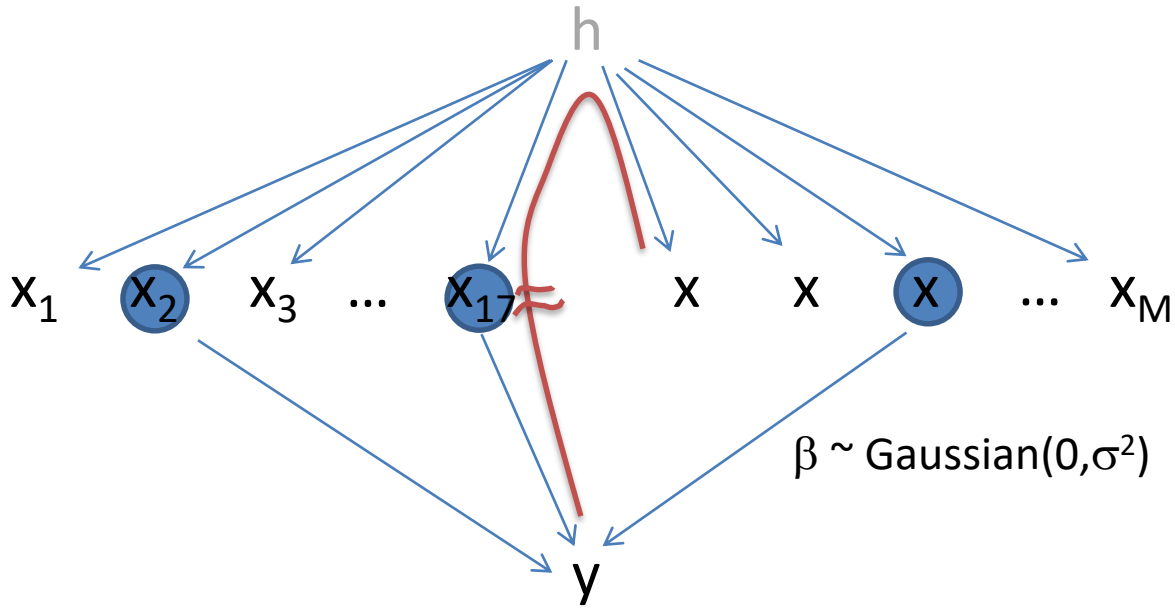
Can eliminate spurious associations by conditioning on causal SNPs



Unfortunately, we don't know which SNPs are causal.

So instead, condition on all SNPs (except the one being tested).

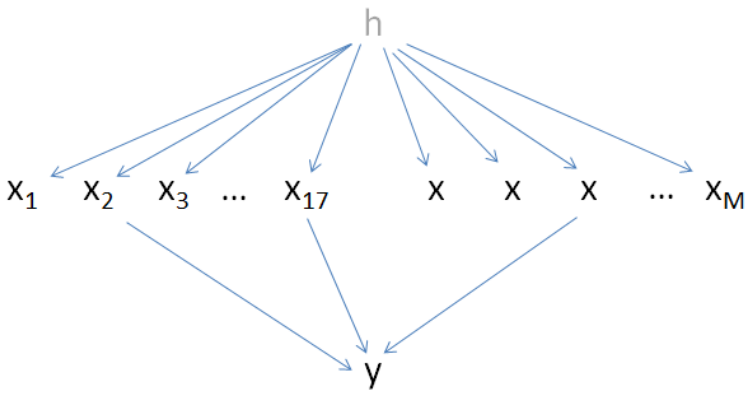
Can eliminate spurious associations by conditioning on causal SNPs



Unfortunately, we don't know which SNPs are causal.

So instead, condition on all SNPs (except the one being tested).
- Go Bayesian to help with high dimensionality.
- Equivalent to linear mixed model!
- Also equivalent to a Gaussian process with a linear kernel.

Generative model for data:



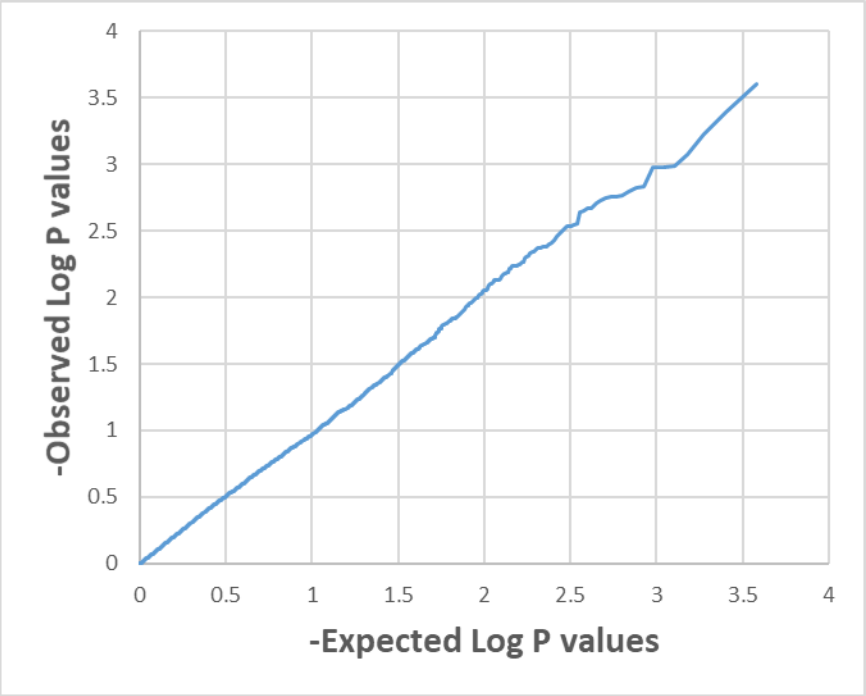
Generate SNPs

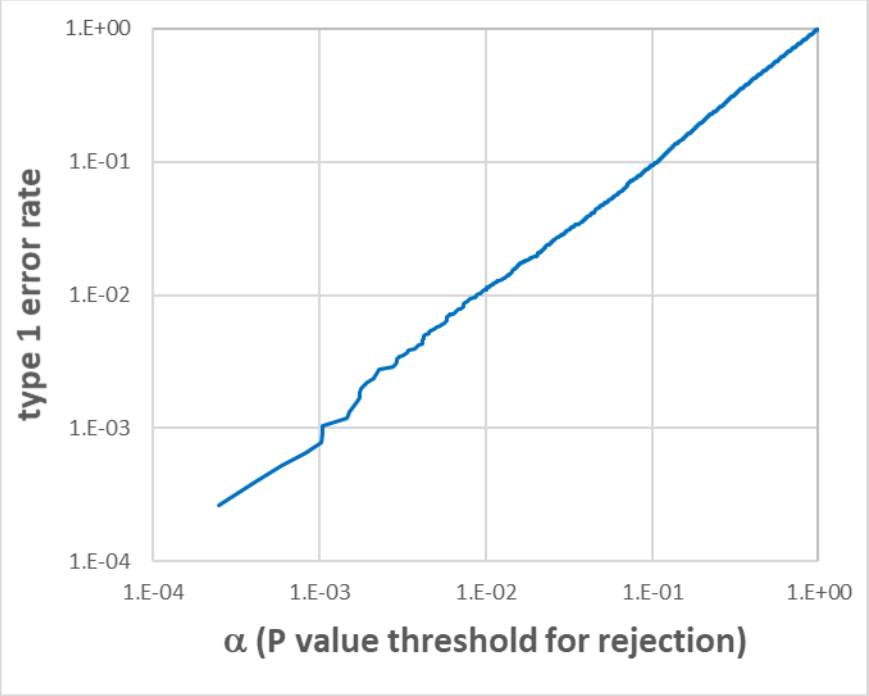
- M=50,000 SNPs and N= 4,000 individuals
- Family relatedness: 10 offspring per parent pair
- Randomly choose a subset of C SNPs to be causal
- Generate trait from causal SNPs with noise $N(0, \sigma_e^2)$
- Linear weights drawn from $N(0, \sigma_g^2/C)$

Generate data sets of each of the following:

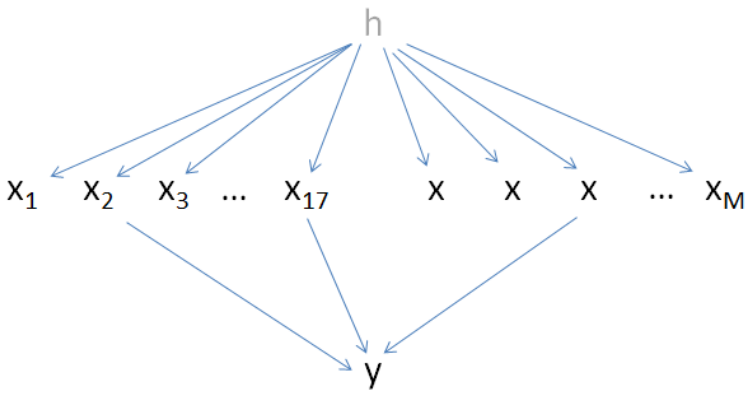
- Number of causal SNPs: 10, 50, 100, 500, 1000
- $\sigma_g^2 \ / \ (\sigma_g^2 + \sigma_e^2)$: 0.1, 0.2, 0.3, 0.4, 0.5, 0.6
- Fraction of individuals belonging to a family: 0.5, 0.6, 0.7, 0.8, 0.9

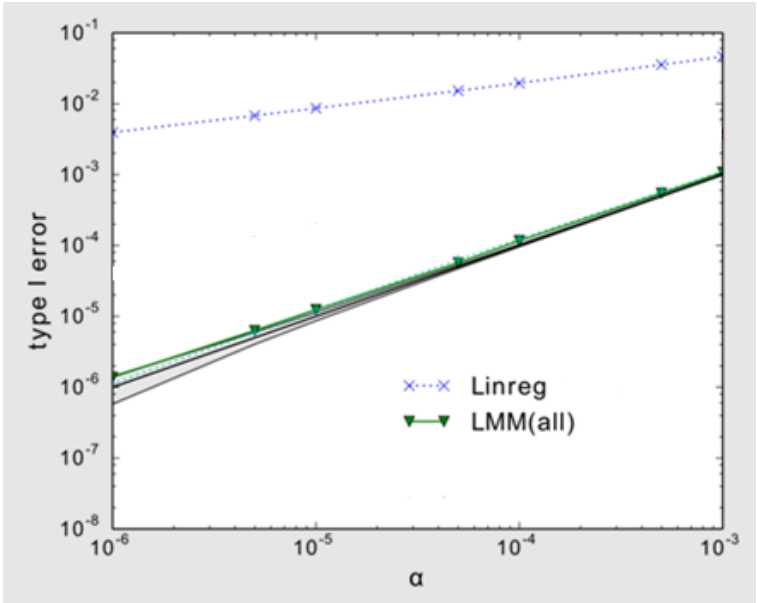
Another type of plot to check for confounding, useful for synthetic data



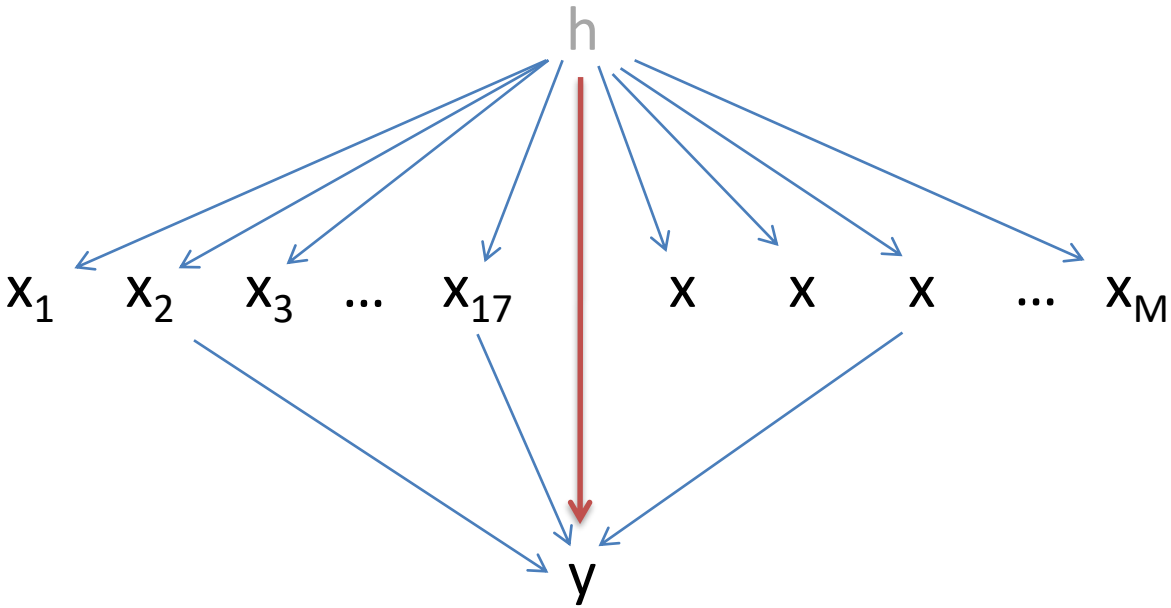


LMM corrects for confounding

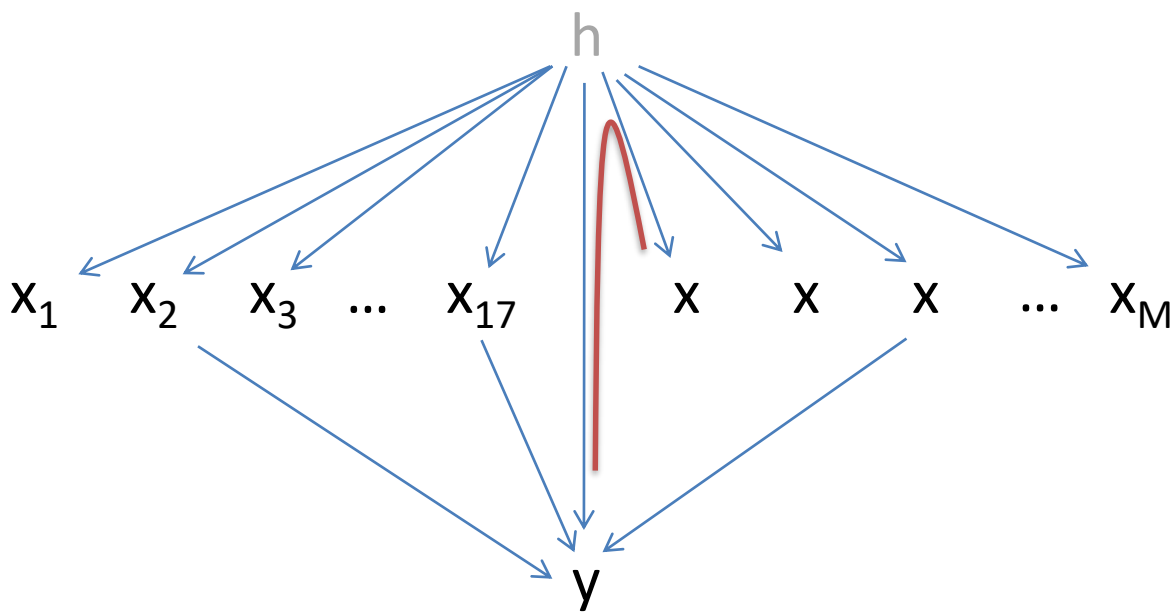




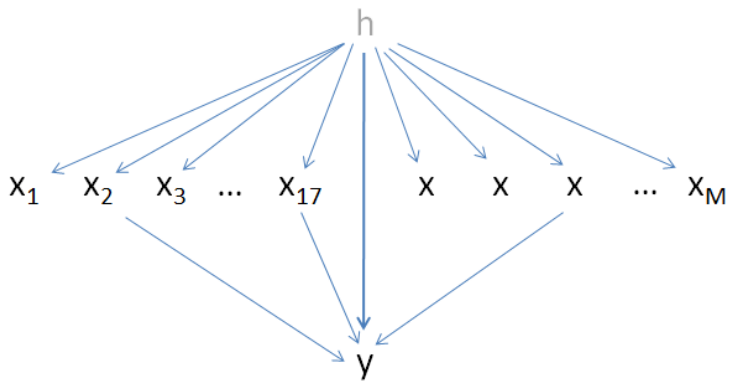
Back to the more realistic case



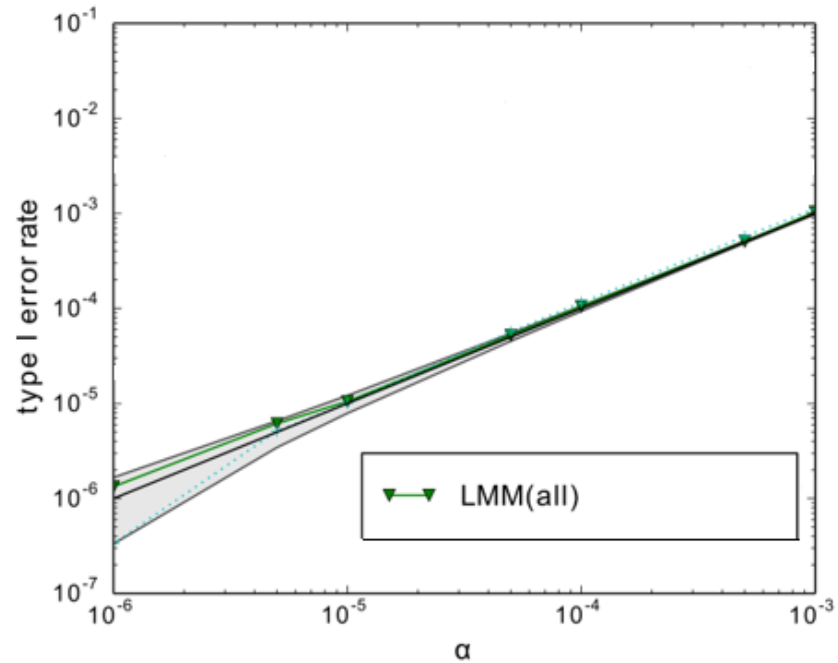
Still confounded after conditioning on all SNPs



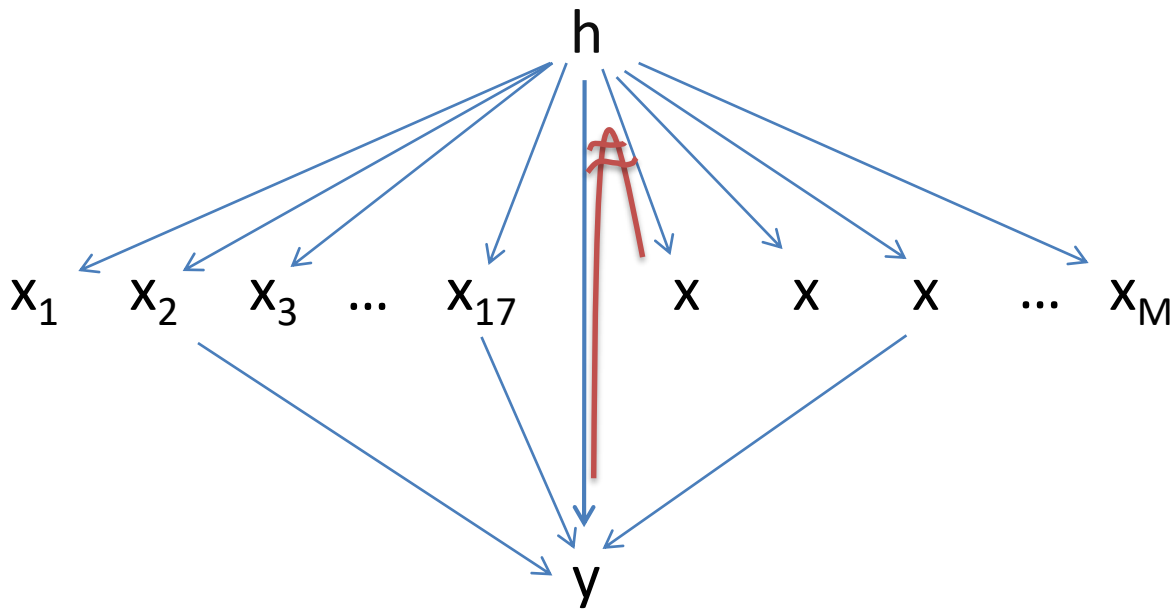
But conditioning on all SNPs still corrects well



- h->y implemented via 100 hidden causal SNPs
- 30% of the variance in y due to h->y



What's going on?
h is "inferentially" observed

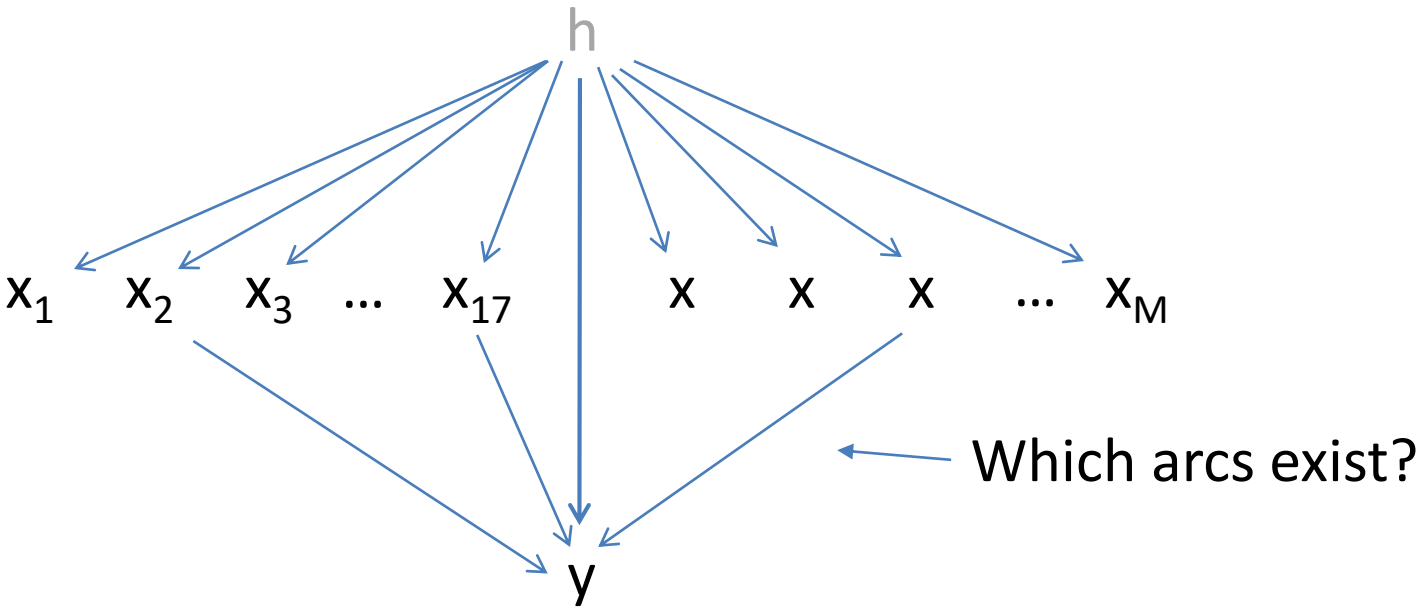


High dimensionality is important

Benefits of high dimensionality

- Diagnose the presence of confounders
- Correct for their presence

Causal discovery without intervention



- Causal discovery in GWAS <u>seems</u> simple
- Surprising amount of complexity

Acknowledgments

Christoph Lippert

Chris Widmer

Carl Kadie

Bob Davidson