Centre for Molecular Medicine and Therapeutics



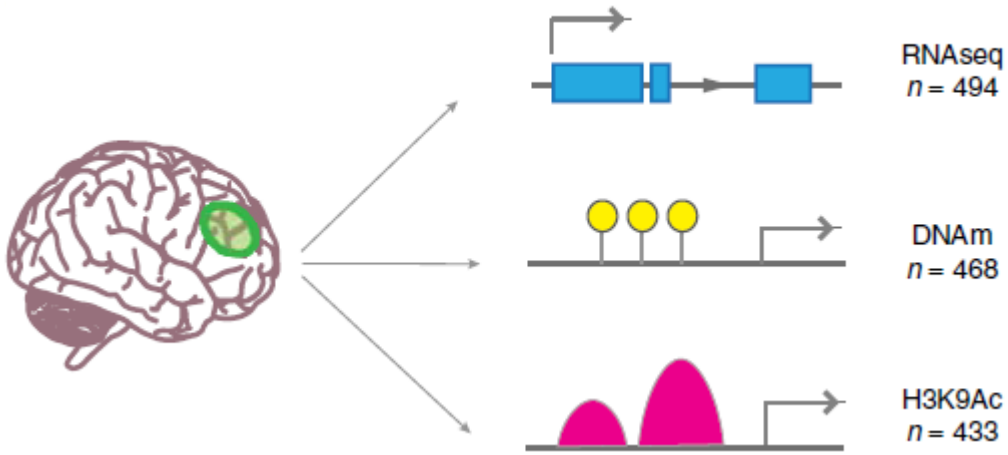
xQTL Analysis

Bernard Ng

Department of Statistics, UBC

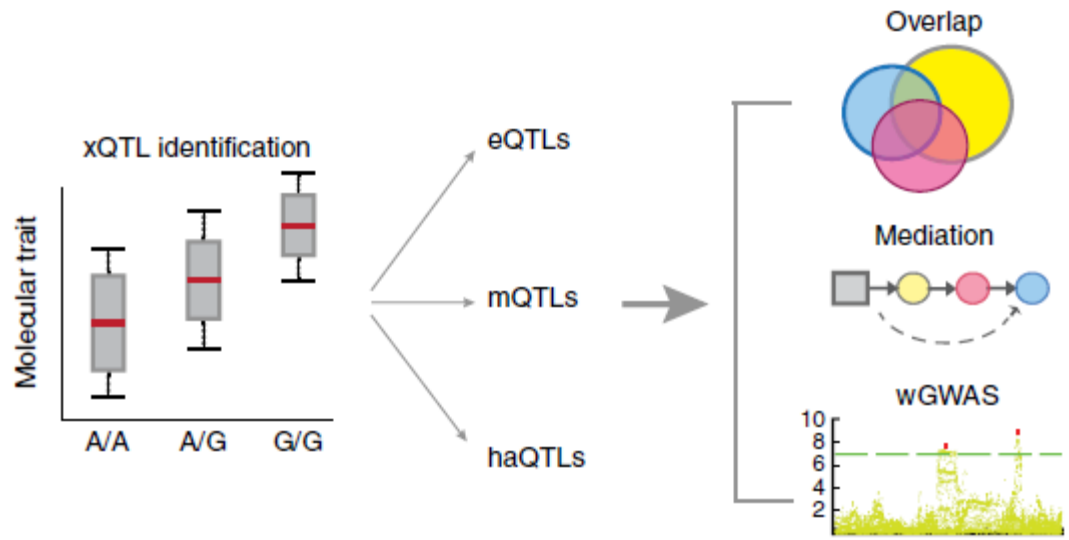
www.cmmt.ubc.ca

Outline



- Recap of GWAS
- xQTL Analysis
- Challenges

- Replication
- xQTL Sharing
- Mediation Analysis
- Cell Specific Analysis
- Weighted GWAS



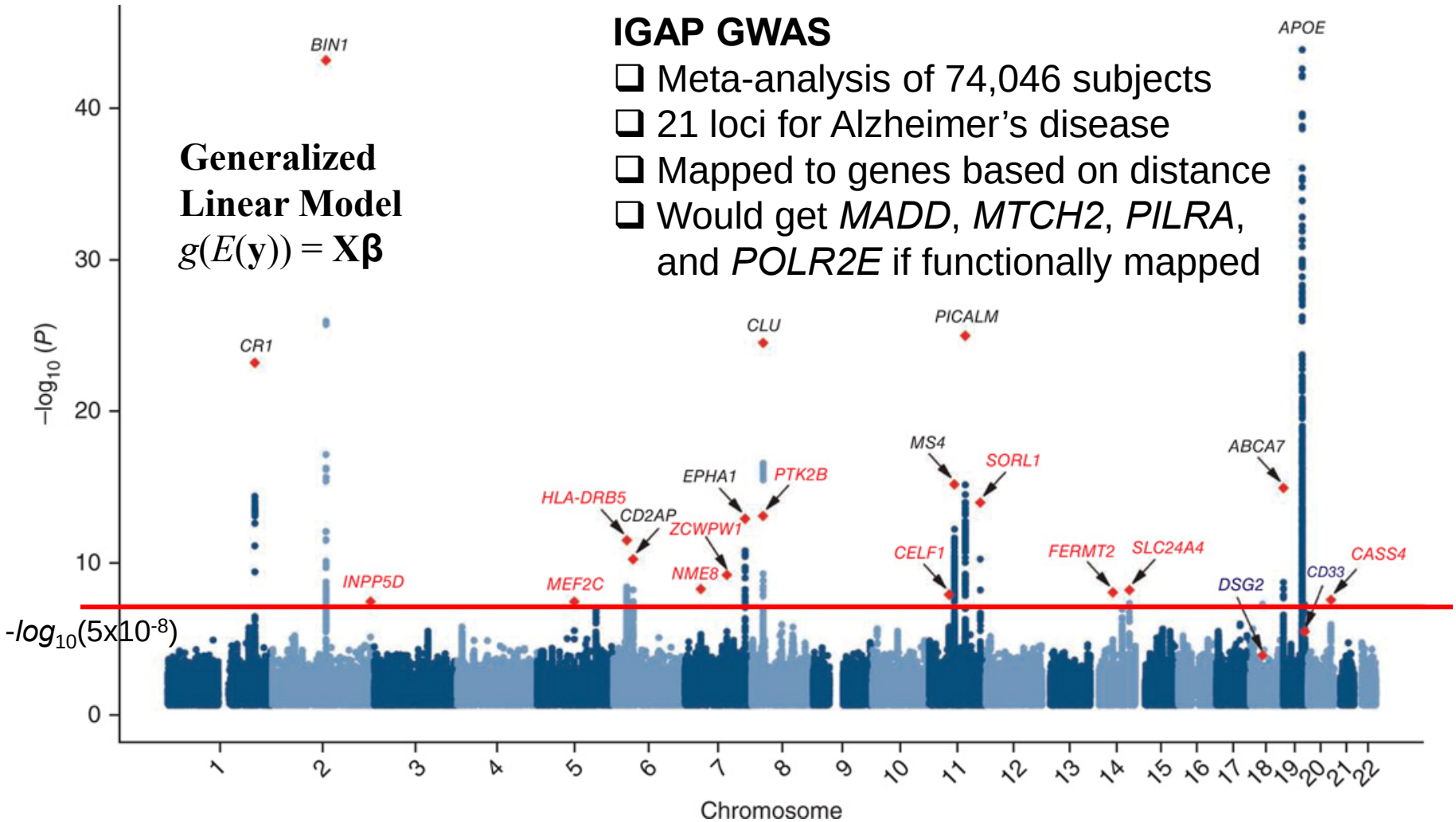
Motivation

Generalized Linear Model

$g(E(\mathbf{y})) = \mathbf{X\beta}$

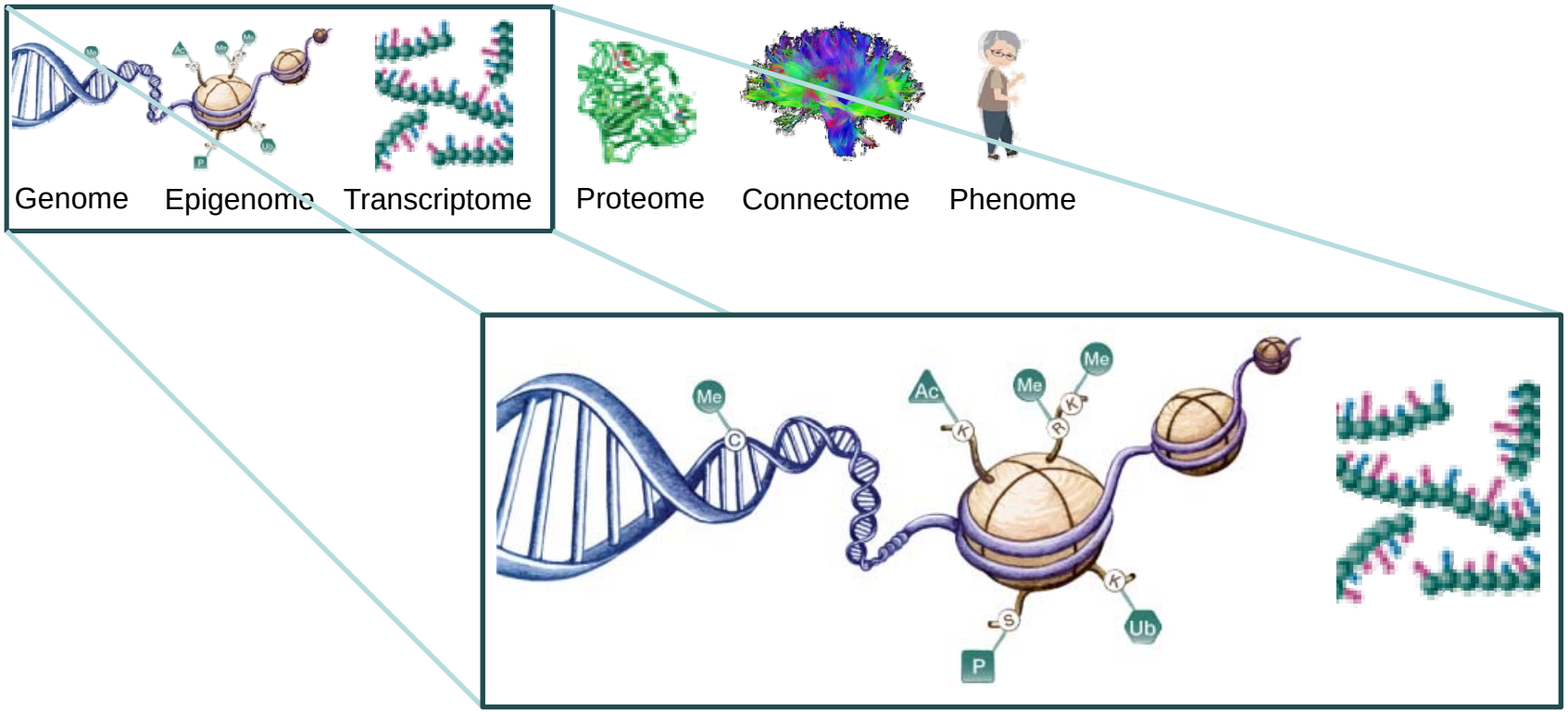
IGAP GWAS

- ❑ Meta-analysis of 74,046 subjects
- ❑ 21 loci for Alzheimer's disease
- ❑ Mapped to genes based on distance
- ❑ Would get *MADD*, *MTCH2*, *PILRA*, and *POLR2E* if functionally mapped



xQTL Analysis

xQTL Analysis



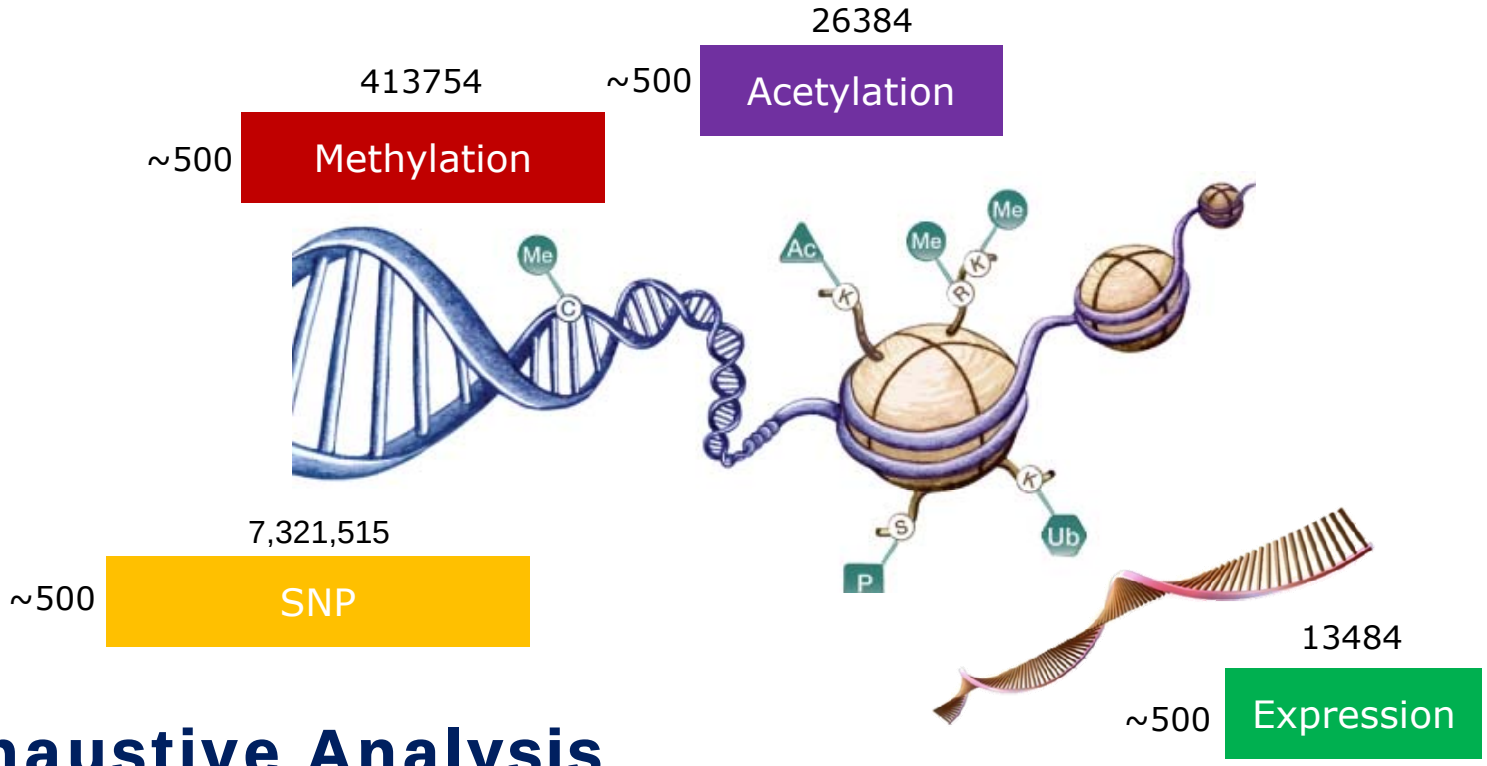
Linear Regression Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Spearman's correlation

$$corr(rank(\mathbf{y}^c), rank(\mathbf{X}_i))$$

Challenges: High Dimensionality

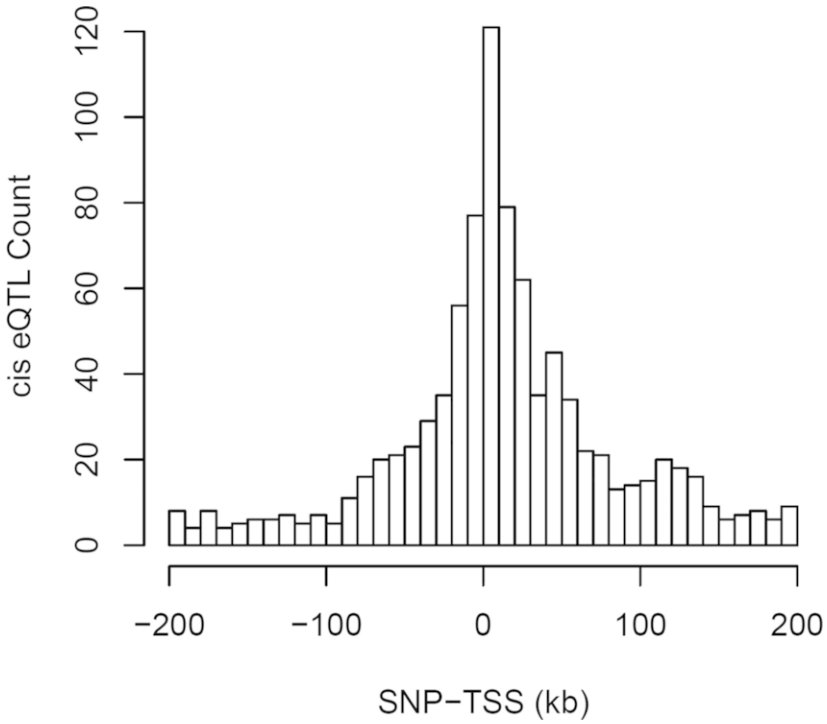


Exhaustive Analysis

- eQTL: 98,723,308,260
- mQTL: 3,029,306,117,310
- haQTL: 193,170,851,760

Trillion of tests =>
no statistical power

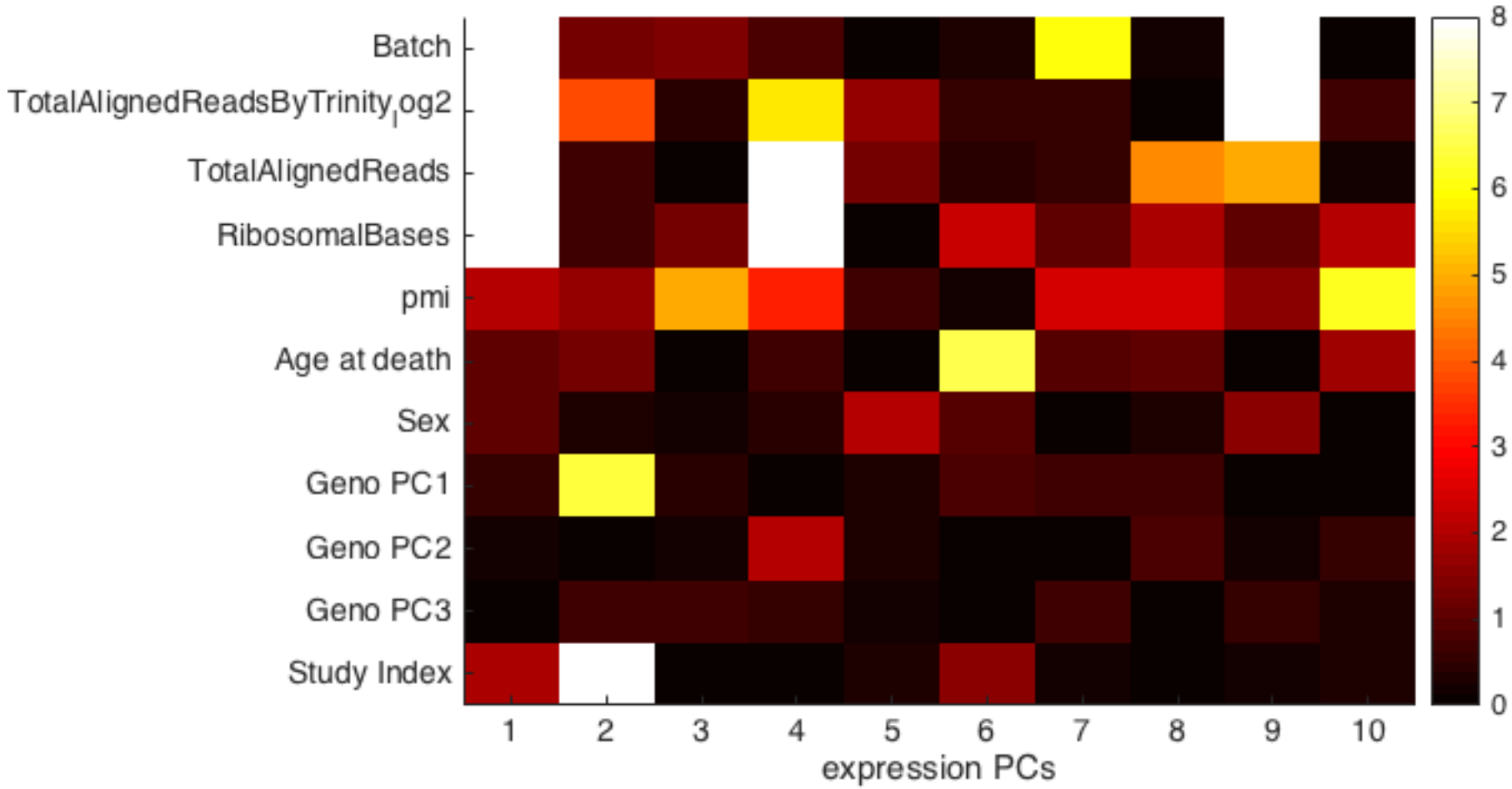
Challenges: High Dimensionality



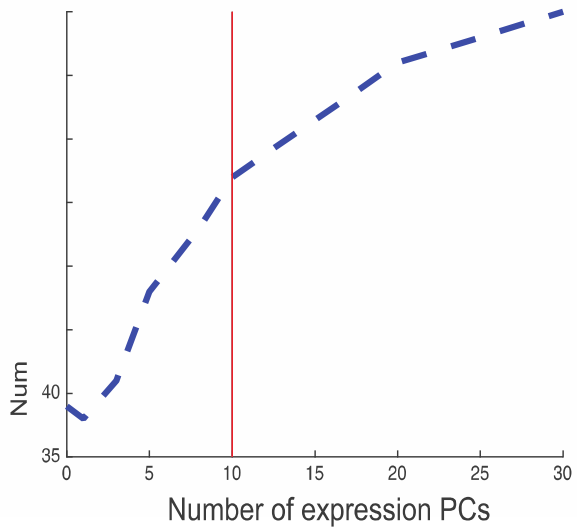
Windowing to reduce the number of tests

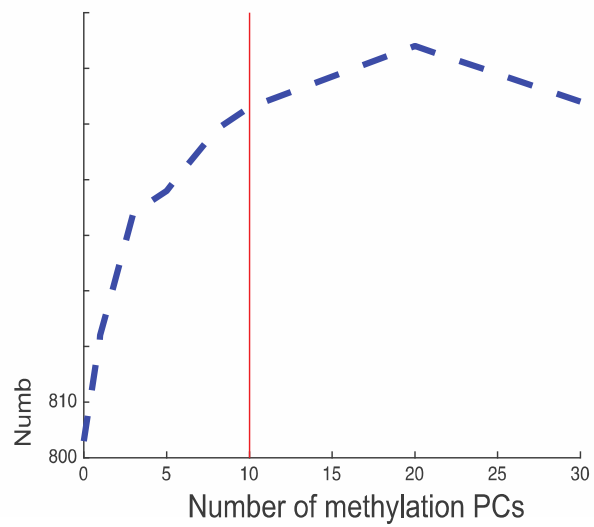
	#SNP-feature pairs	#Features	#SNPs
eQTL (1Mb)	60,456,556	12,979	6,442,864
mQTL (5Kb)	9,939,236	412,152	2,358,873
haQTL (1Mb)	125,100,450	25,720	6,756,597

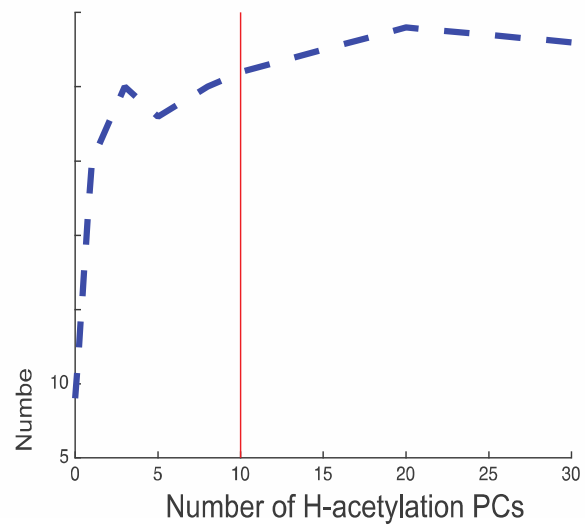
Challenges: Hidden Confounds



Challenges: Hidden Confounds

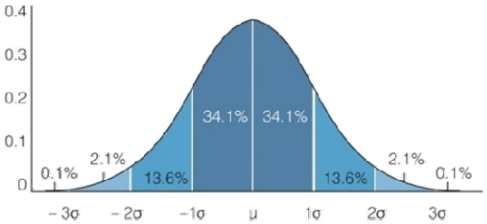






Challenges: Multiple Testing

- P(not rejecting 1 hypothesis) = 1- α
- P(not rejecting all *n* hypotheses) = $(1-\alpha)^n$
- $\alpha_{\text{FWER}} = 1-(1-\alpha)^n$
  - $\alpha = 0.05,\ n = 10$: $\alpha_{\text{FWER}} = 0.4013$
  - $\alpha = 0.05,\ n = 10^2$: $\alpha_{\text{FWER}} \approx 1$
- So if e.g. run 100 experiments, then $\alpha_{\text{FWER}} \cdot 100$ of them would have ≥1 hypothesis falsely rejected.
- Intuition is that the more we sample the variable space, the more "likely" we will get some "extreme" samples.



Bonferroni Correction

Procedures

- Recall $\alpha_{\text{FWER}} = 1\text{-}(1\text{-}\alpha)^n$
- Set $\alpha = 1\text{-}(1\text{-}\ \alpha_{\text{FWER}})^{1/n} \approx 1\text{-}(1\text{-}\alpha_{\text{FWER}}/n) = \alpha_{\text{FWER}}/n$

Examples

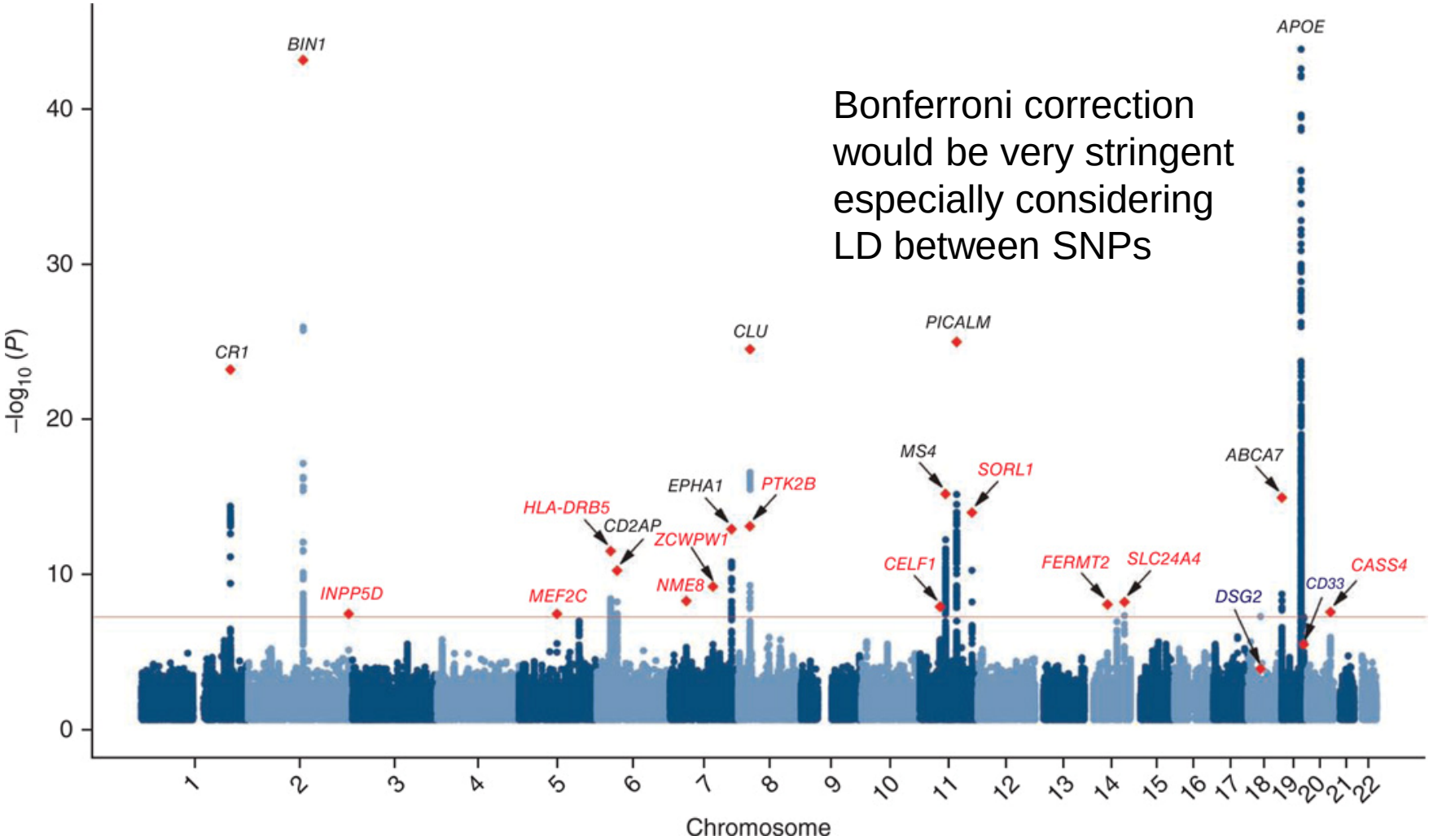
- $\alpha_{\text{FWER}} = 0.05$ and $n = 10$, needs $\alpha = 0.05/10 = 0.005$
- $\alpha_{\text{FWER}} = 0.05$ and $n = 10^6$, needs $\alpha = 0.05/10^6 = 5\times10^{-8}$

Properties

- Ctrls FWER = P(V≥1) in *strong* sense.
- Can handle correlated hypotheses.
- Very stringent

		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n\text{-}n_0$
		n-R	R	n

Challenges: Multiple Testing



Bonferroni correction would be very stringent especially considering LD between SNPs

False Discovery Rate

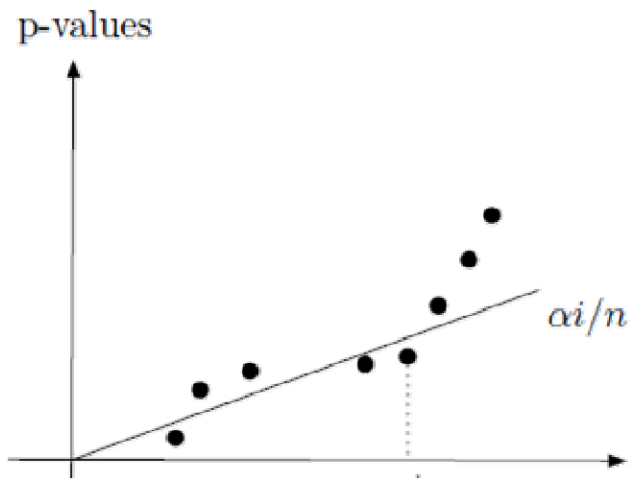
		Predicted: True	Predicted: False	
Ground Truth	True	U	V	n_0
	False	T	S	$n-n_0$
		n-R	R	n

Idea

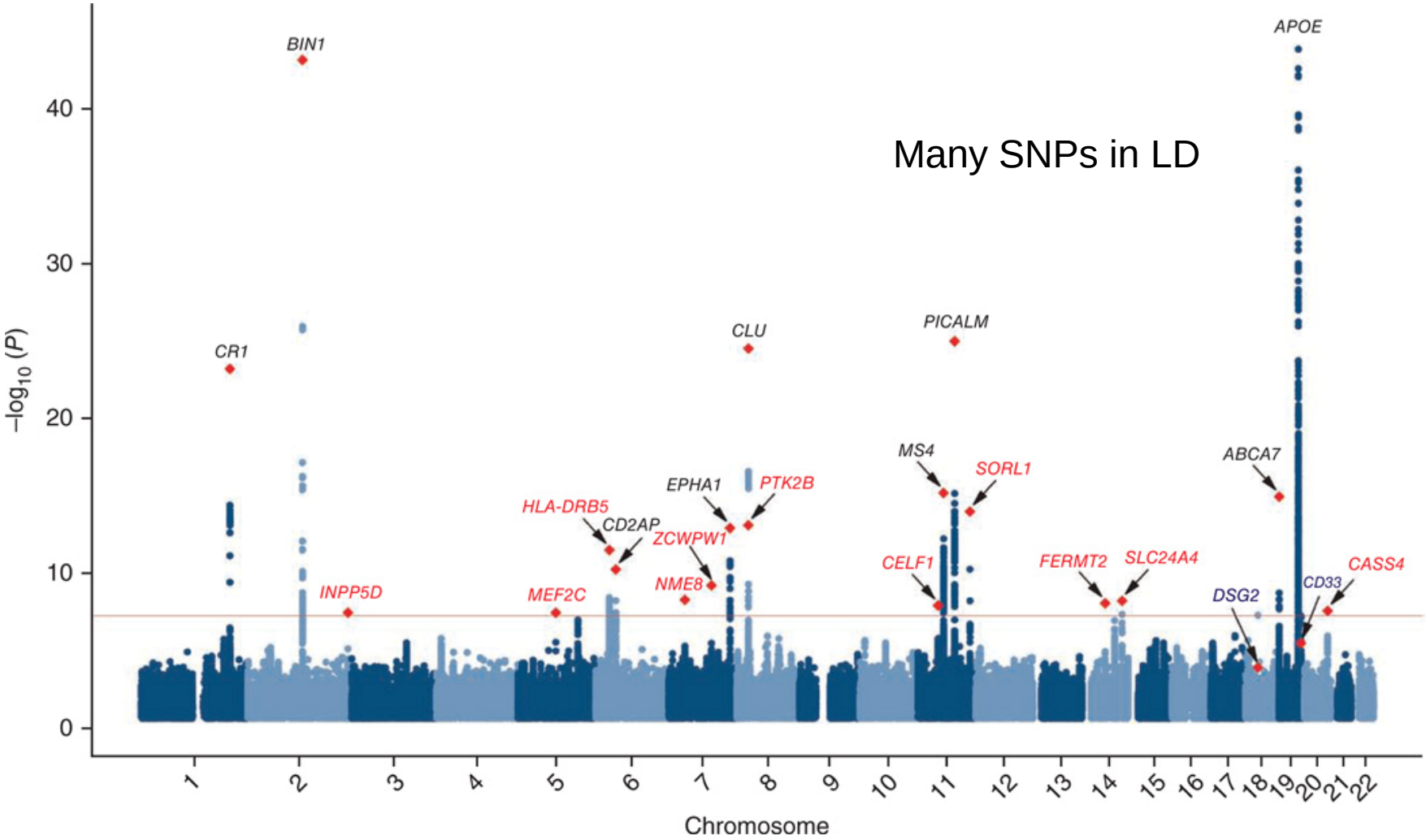
- Benjamini & Hochberg, 1995
- Recall FWER = $P(V \geq 1)$
- Fdp = $V/\max(R,1)$
- But V unobserved, so: FDR = E(Fdp)

Procedures

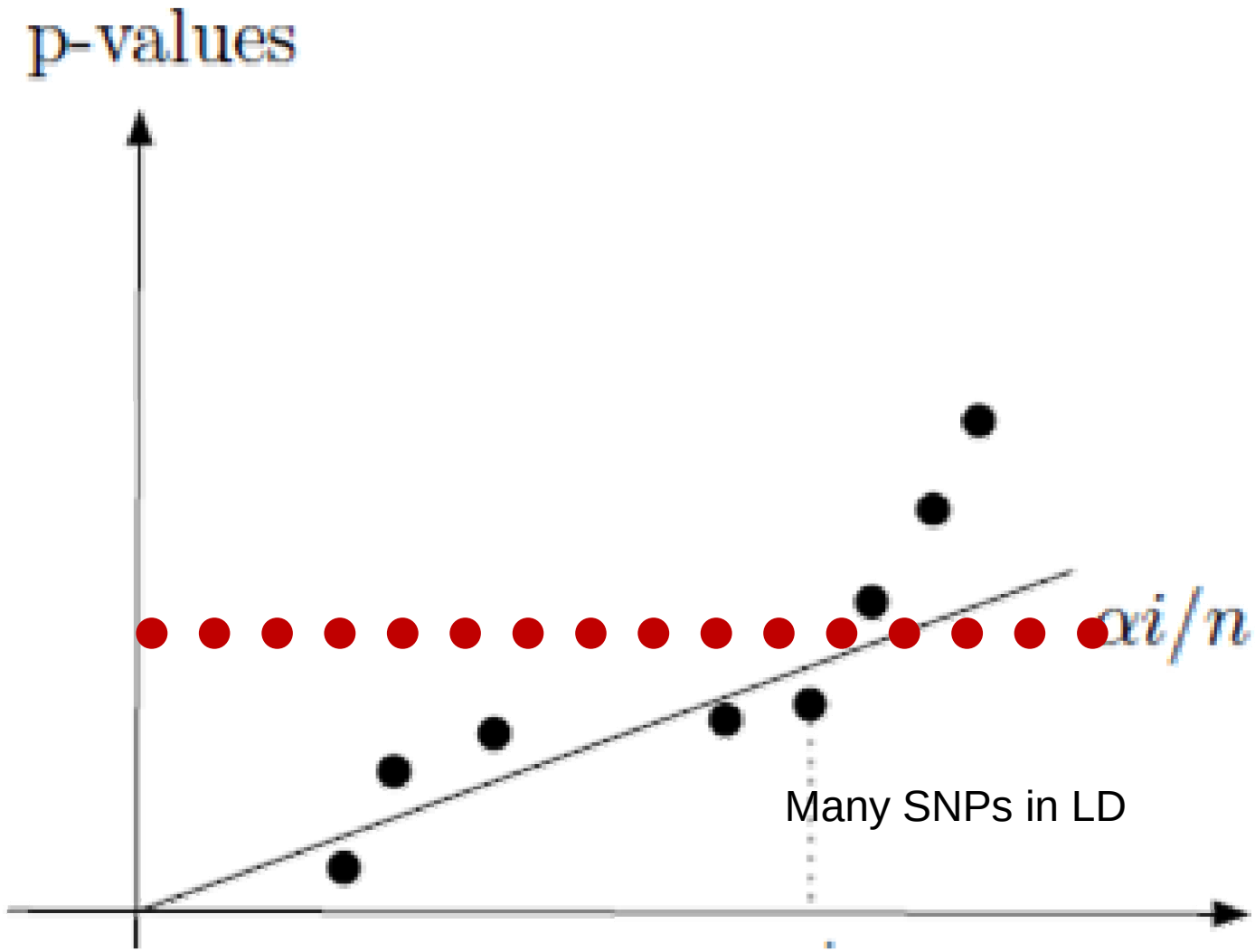
- Sort p in ascending order.
- Find $i_0 = \max i$ s.t. $p(i) \leq i \cdot q/n$



Challenges: Multiple Testing



False Discovery Rate



False Discovery Rate

Properties

- If hypotheses are independent, then FDR < q for *all* configurations of hypotheses.
- If hypotheses are correlated,

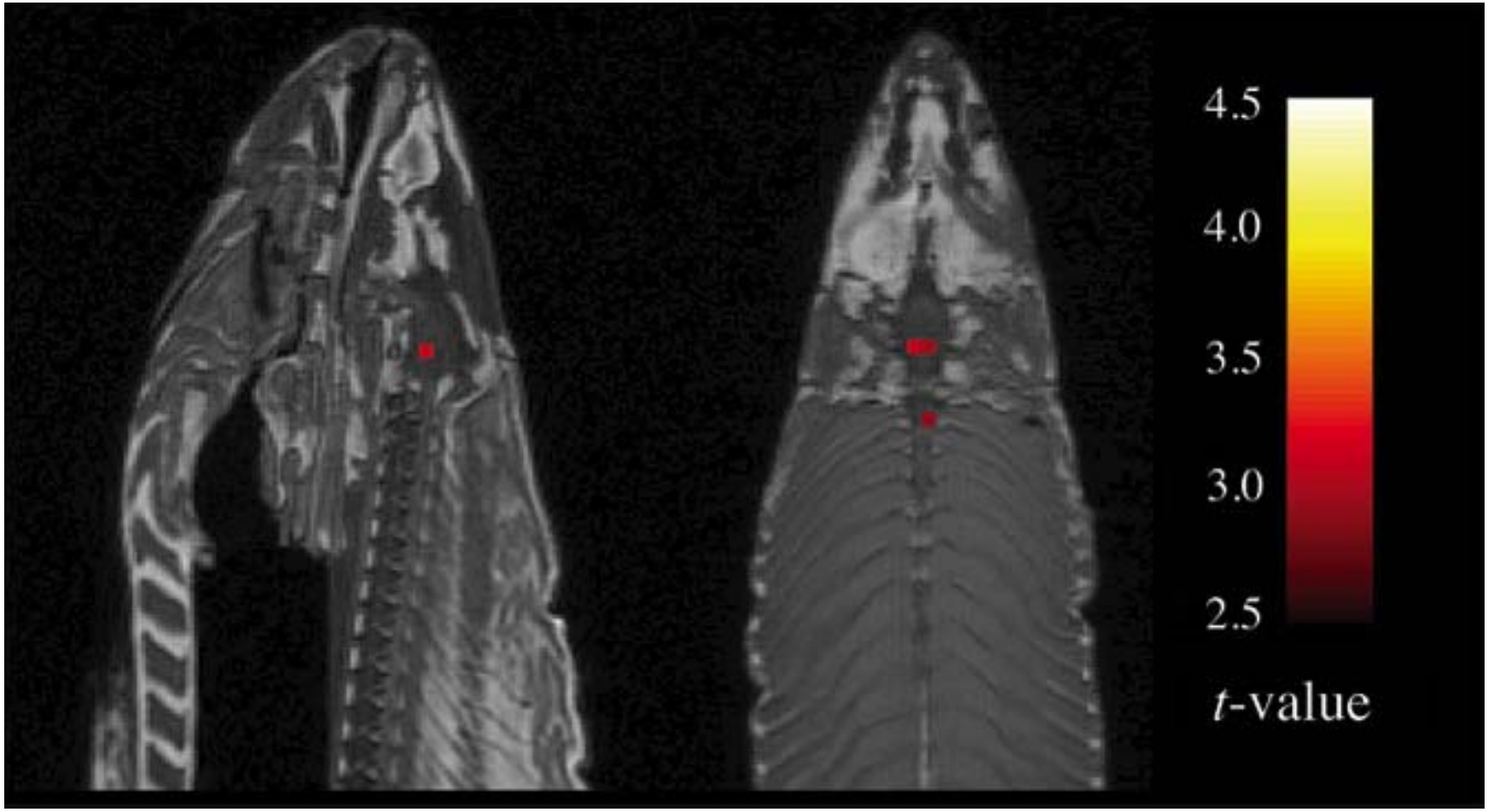
  FDR < q·(log(n)+0.577)

  => p(i) < i·q/n / (log(n)+0.577)

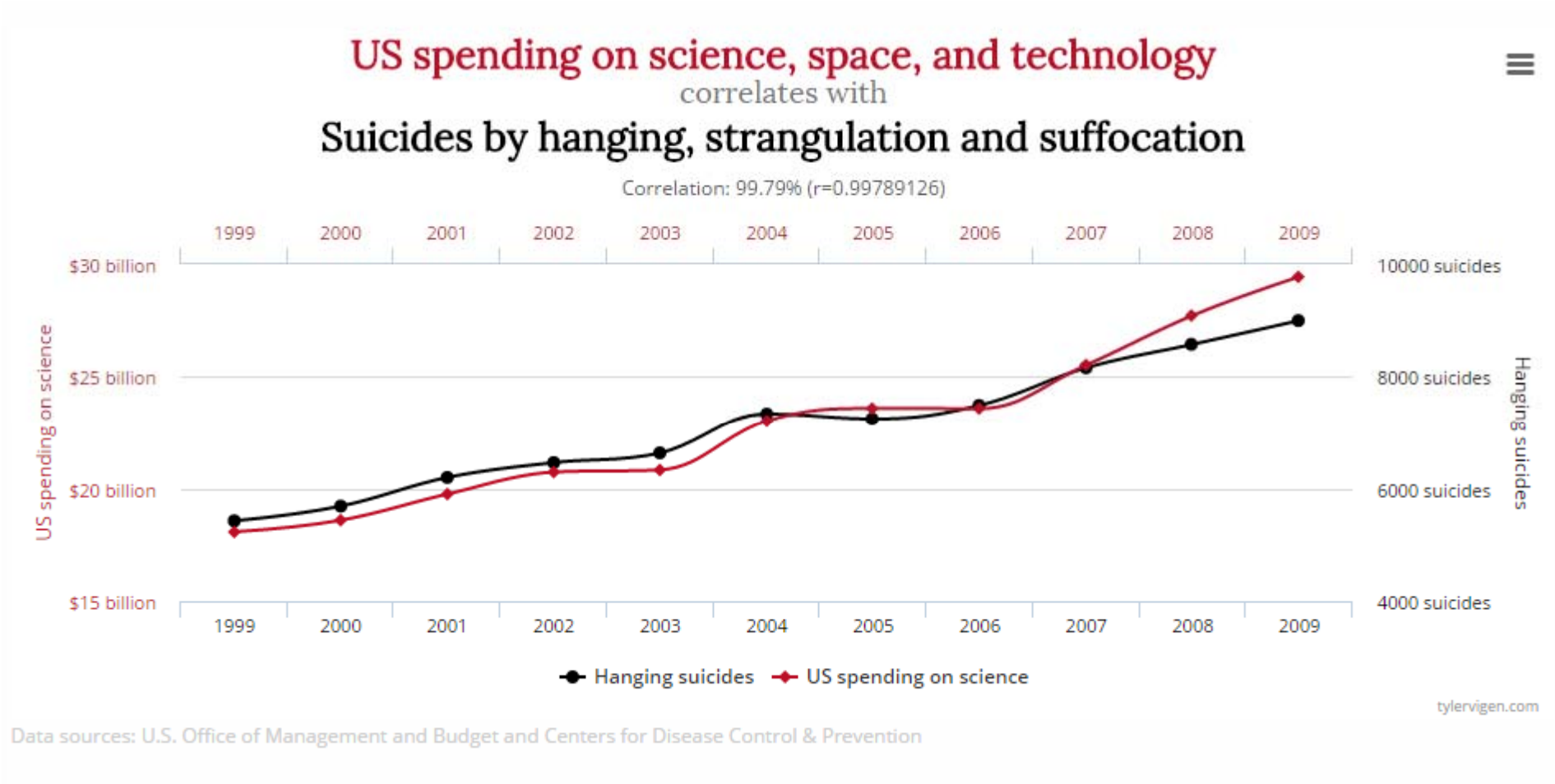
  BUT i = 1, p(i) < q/n / (log(n)+0.577) < q/n

		Predicted		
		True	**False**	
Ground Truth	**True**	U	V	n_0
	False	T	S	$n-n_0$
		n-R	R	n

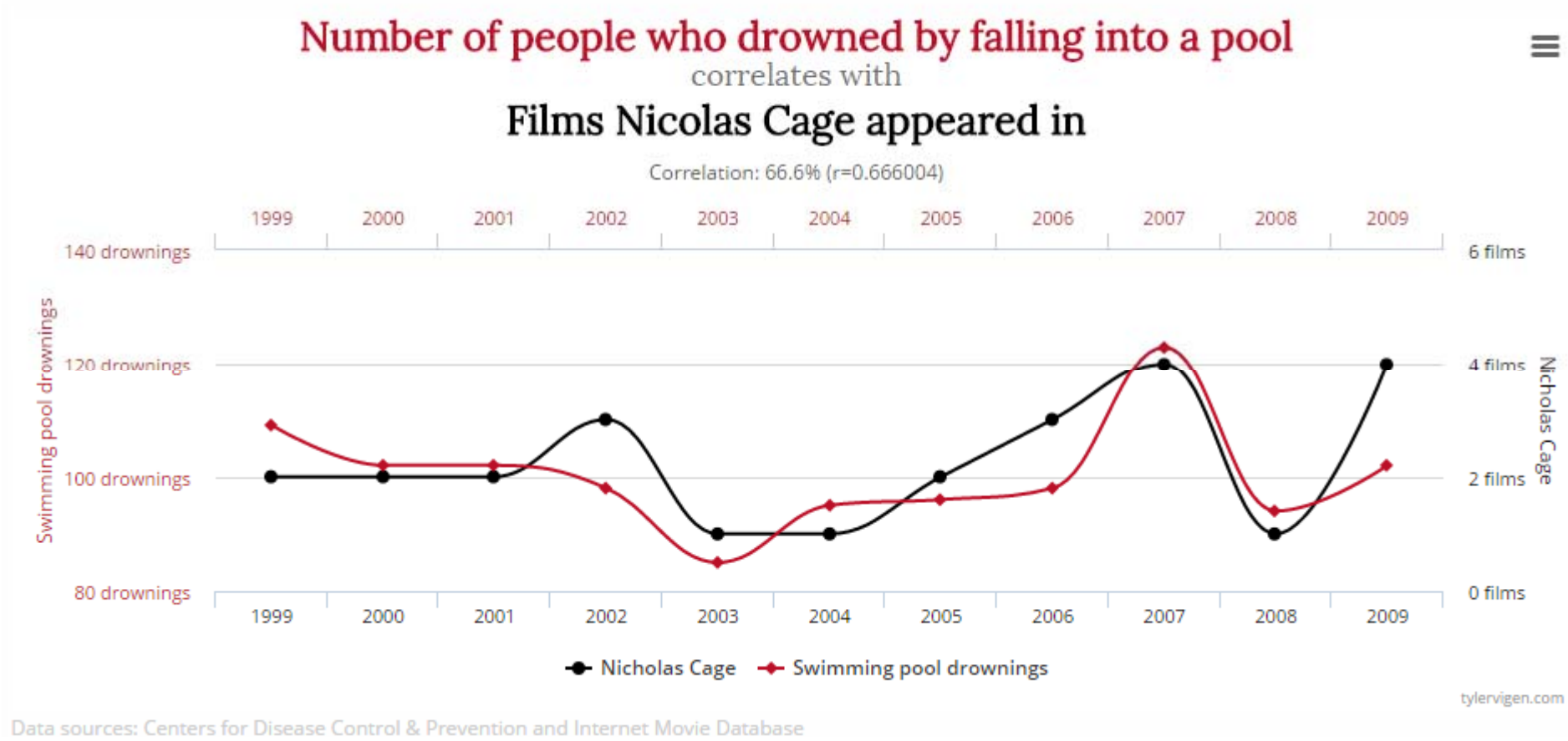
Why Important?



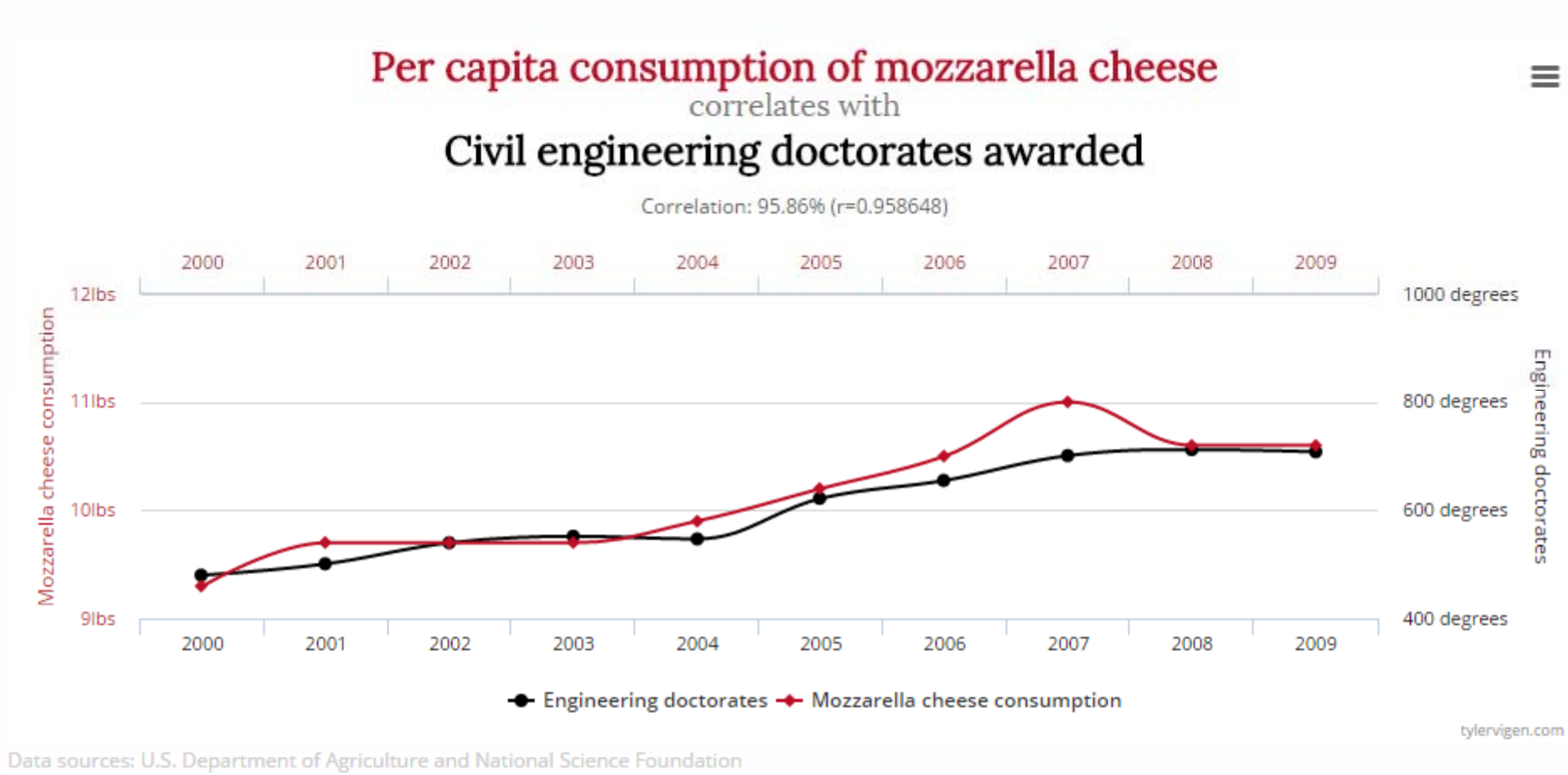
Why Important?



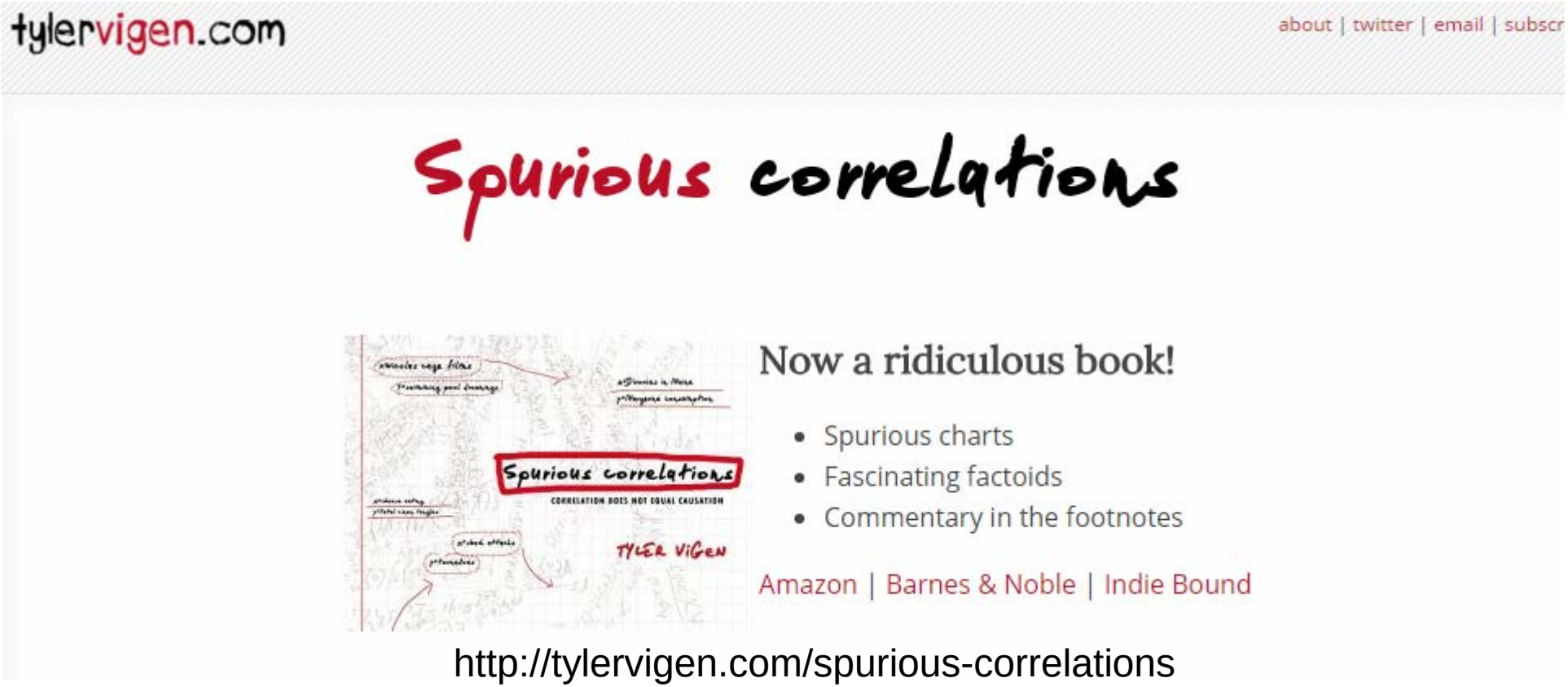
Why Important?



Why Important?



Why Important?



http://tylervigen.com/spurious-correlations

Why Important?

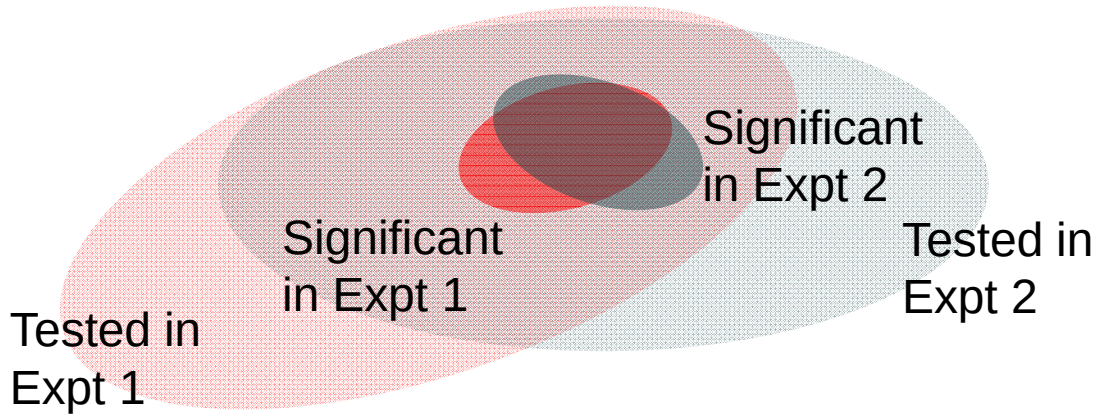
Problems with scientific research

How science goes wrong

Scientific research has changed the world. Now it needs to change itself

Last year researchers at one biotech firm, Amgen, found they could reproduce just **six of 53** "landmark" studies in cancer research. Earlier, a group at Bayer, a drug company, managed to repeat just a **quarter of 67** similarly important papers. A leading computer scientist frets that **three-quarters** of papers in his subfield are bunk. In 2000-10 roughly **80,000 patients** took part in clinical trials based on research that was **later retracted** because of mistakes or improprieties.

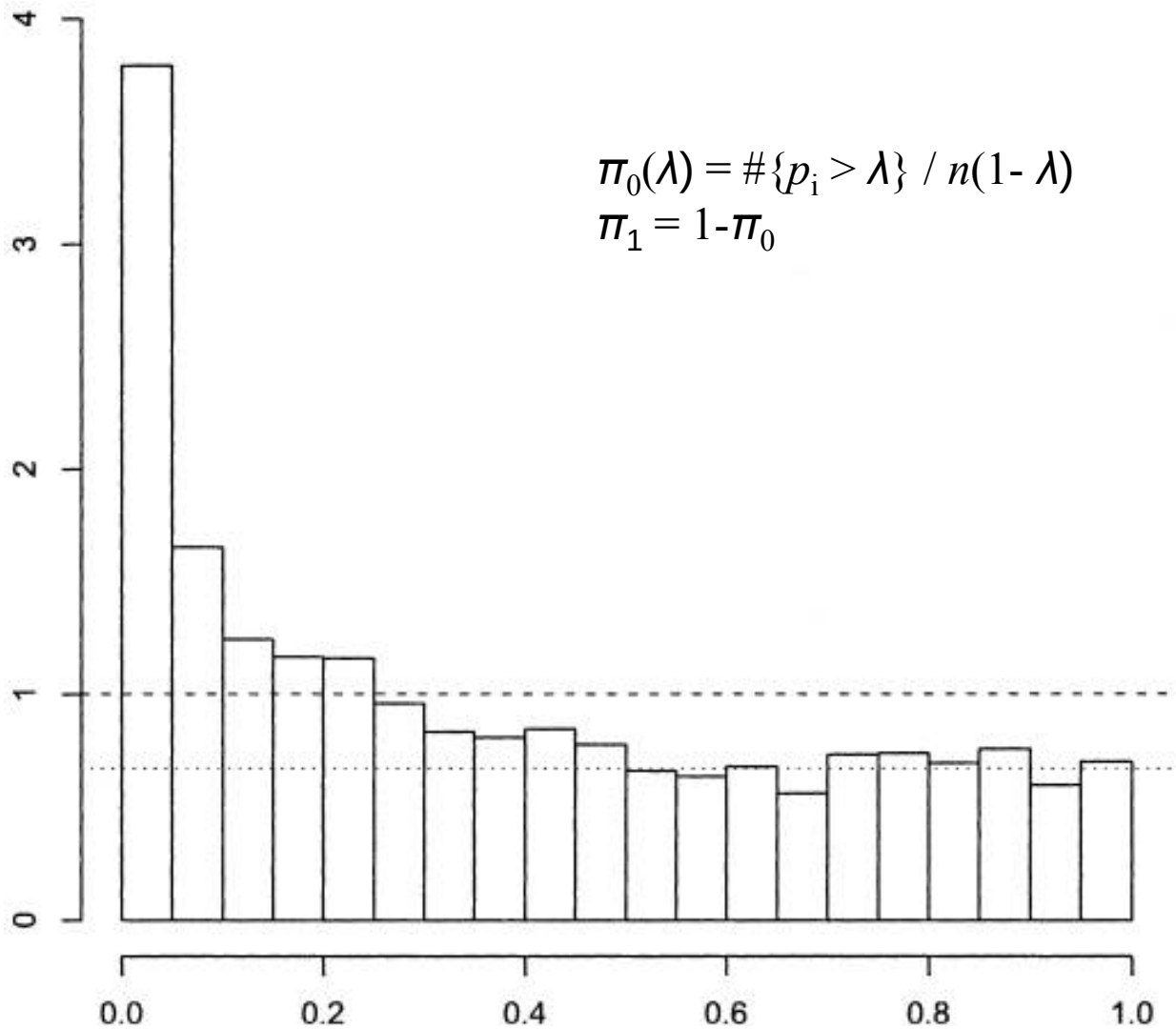
Replication



Strategies

- Lenient: Significant at *a/n* in Expt 1 and at α in Expt 2
- Stringent: Significant at *a/n* in Expt 1 and Expt 2
- In between: π_1 statistics = proportion of non-nulls

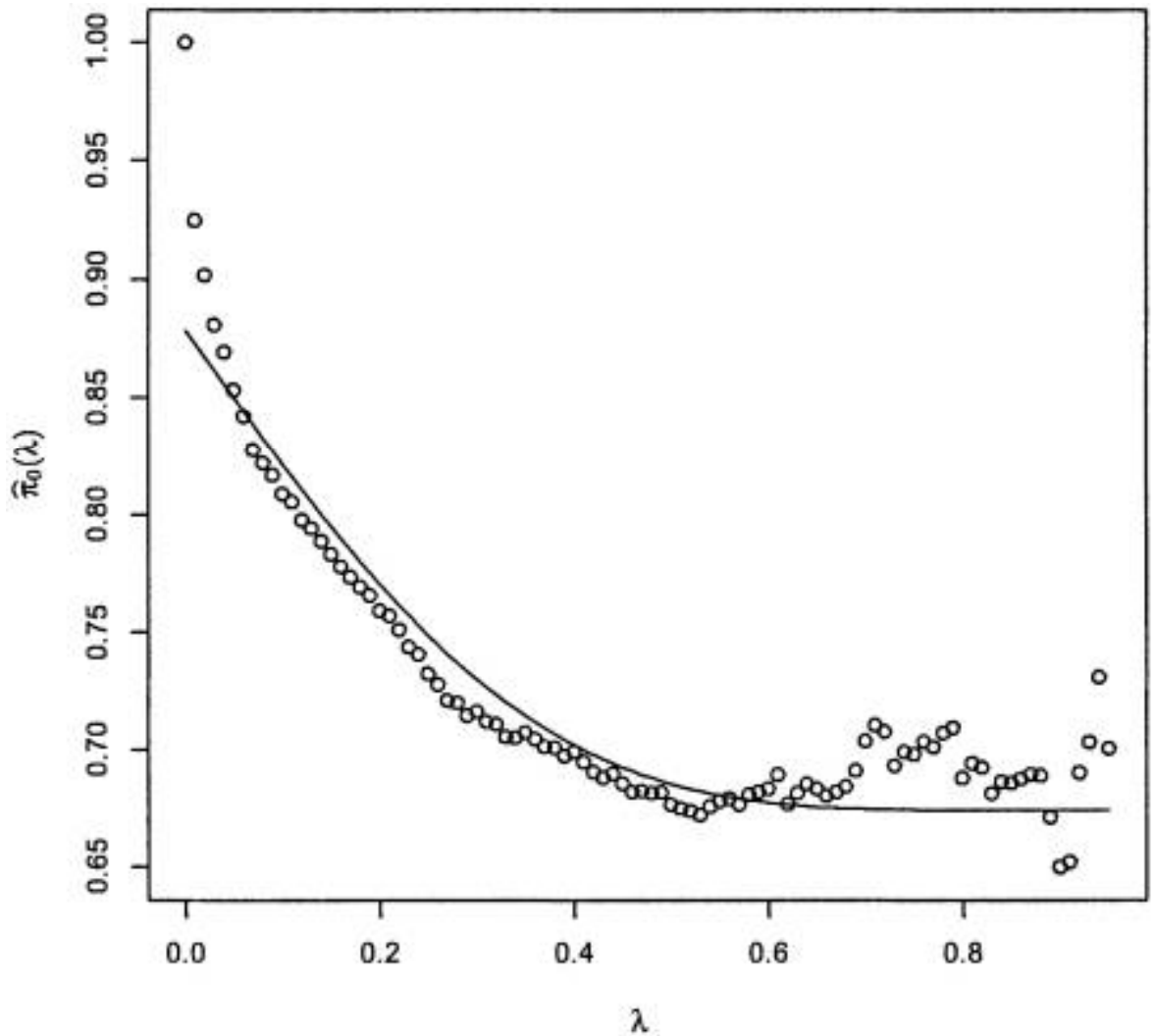
π_1 Statistics



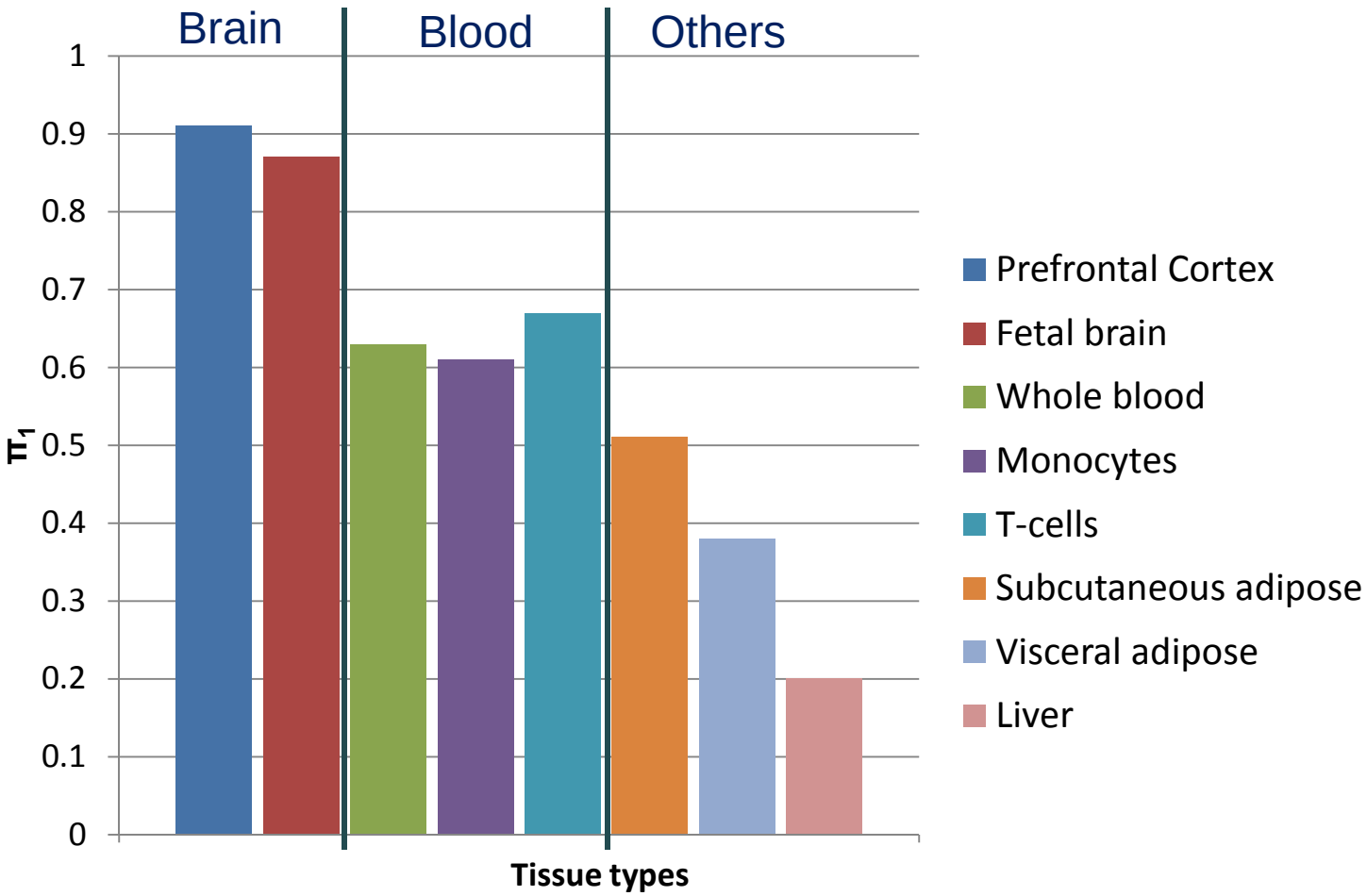
$\pi_0(\lambda) = \#\{p_i > \lambda\} \ / \ n(1- \lambda)$

$\pi_1 = 1-\pi_0$

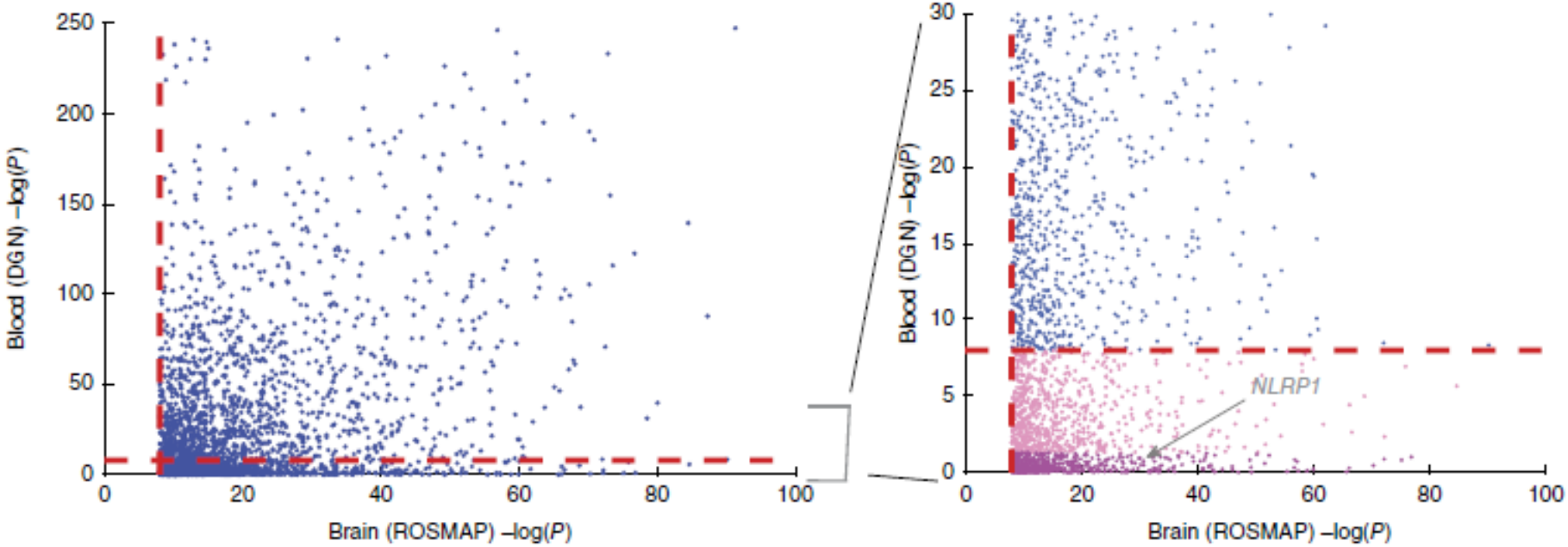
Π_1 Statistics



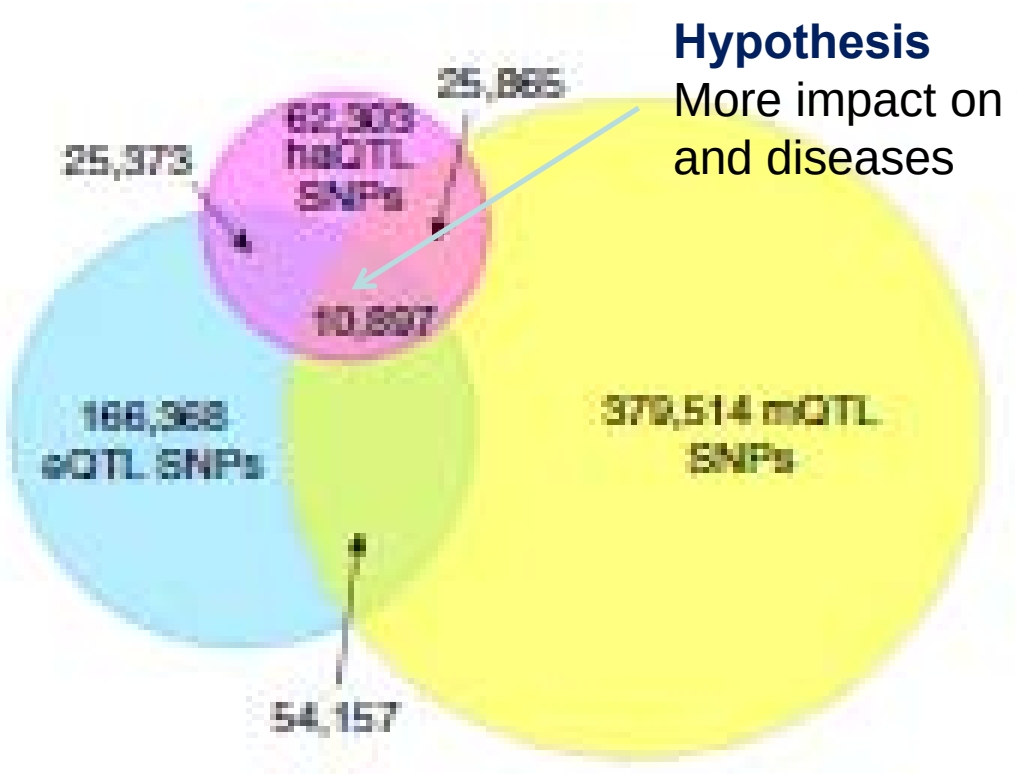
Replication



Replication

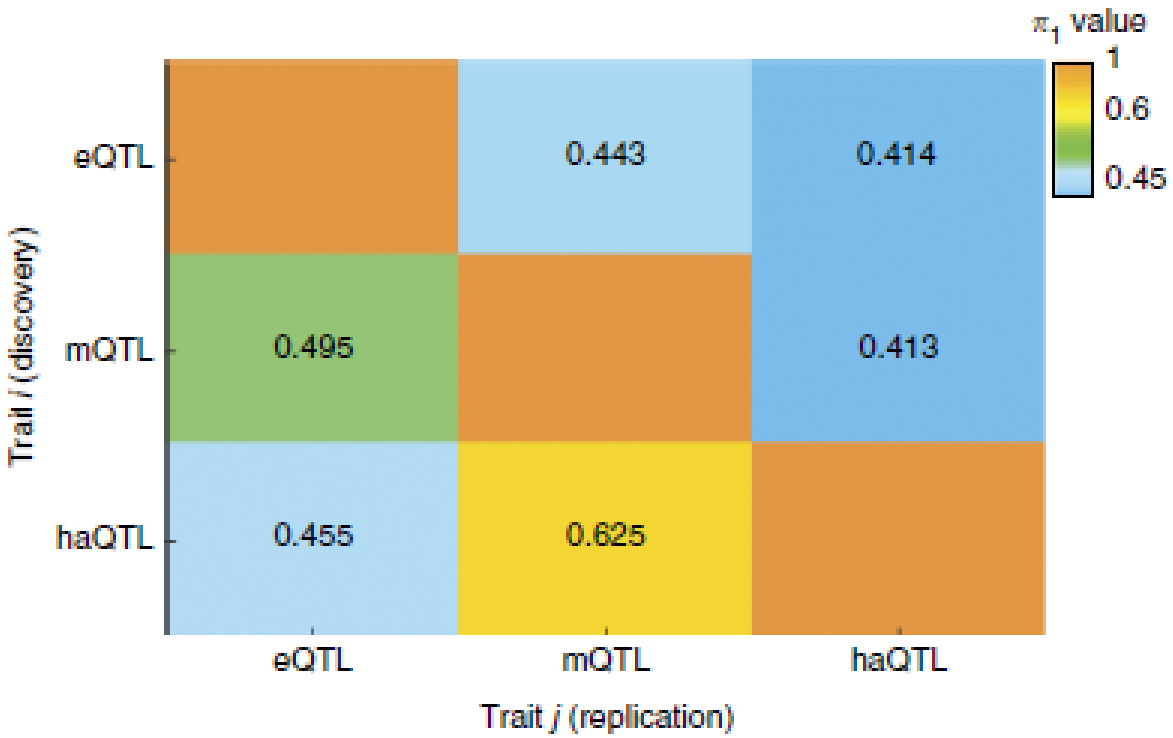


xQTL Sharing

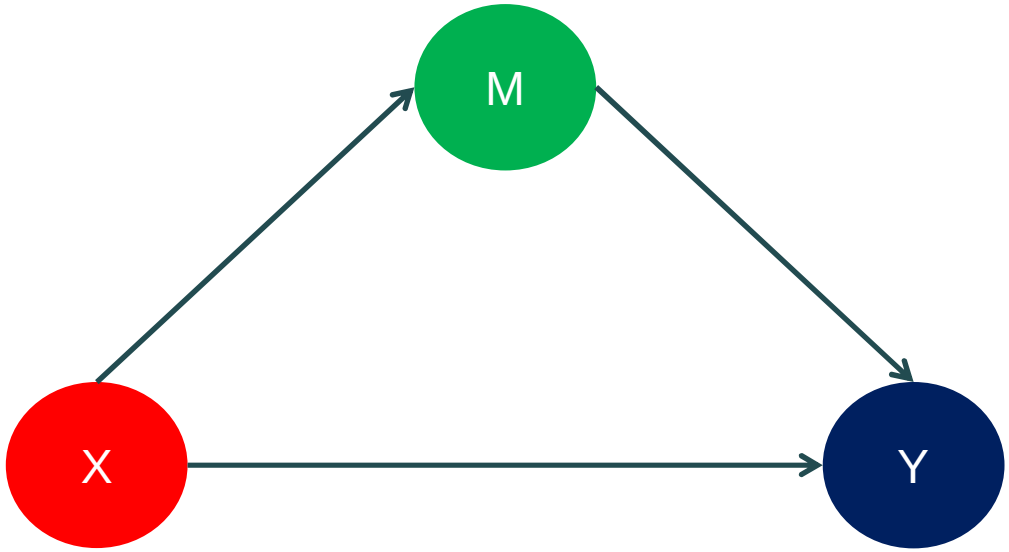


Hypothesis
More impact on traits and diseases

xQTL Sharing



Mediation Analysis

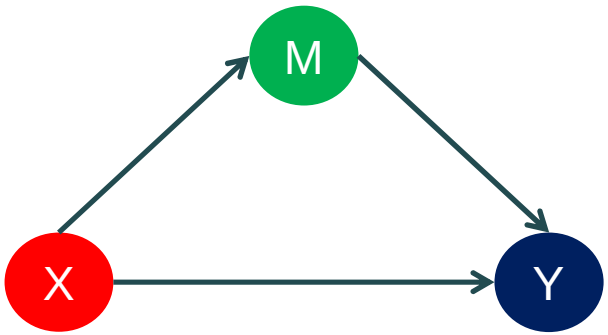


Casual Inference Test

- X & Y associated: $Y = X\beta + \varepsilon$ with H_1: $\beta \neq 0$
- X & M|Y associated: $M = X\beta + Y\gamma + \varepsilon$ with H_1: $\beta \neq 0$
- M & Y|X associated: $Y = X\beta + M\alpha + \varepsilon$ with H_1: $\alpha \neq 0$
- $X \perp Y|M$: $Y = X\beta + M\alpha + \varepsilon$ with H_1: $\beta = 0$

Requires equivalence test

Mediation Analysis



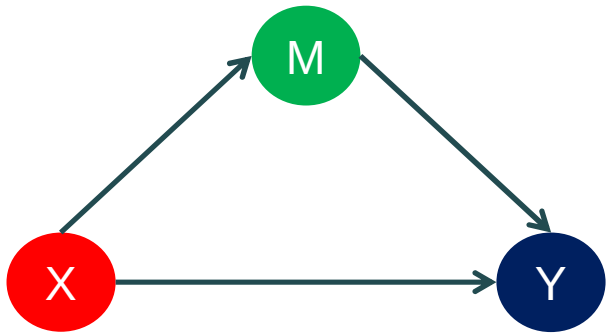
Casual Inference Test

- X & Y associated: $Y = X\beta + \varepsilon$ with H_1: $\beta \neq 0$
- X & M|Y associated: $M = X\beta + Y\gamma + \varepsilon$ with H_1: $\beta \neq 0$
- M & Y|X associated: $Y = X\beta + M\alpha + \varepsilon$ with H_1: $\alpha \neq 0$
- $X \perp Y|M$: $Y = X\beta + M\alpha + \varepsilon$ with H_1: $\beta = 0$

Equivalent Test for H_1: $\beta = 0$ under independence model

- $M = X\beta + \varepsilon$
- $M^p = X\beta + \varepsilon^p$
- $Y = X\beta^p + M^p\alpha^p + \varepsilon$
- $\beta^p \text{ -> } F^p$
- Generate *m* F^p, then p-value = $\#\{F > F^p\}/m$

Mediation Analysis



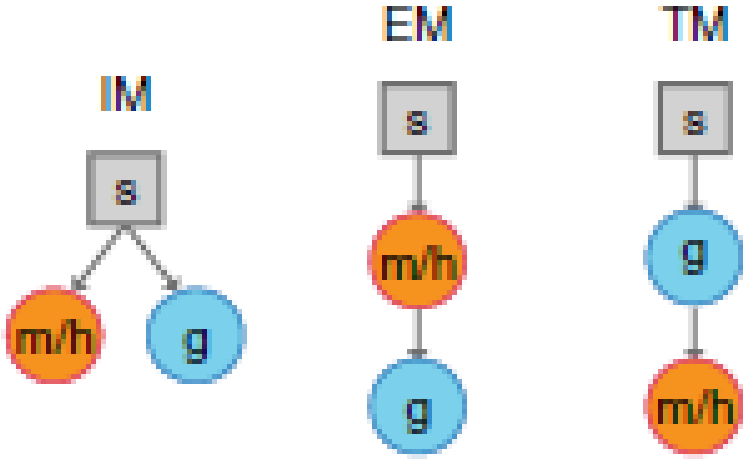
Casual Inference Test

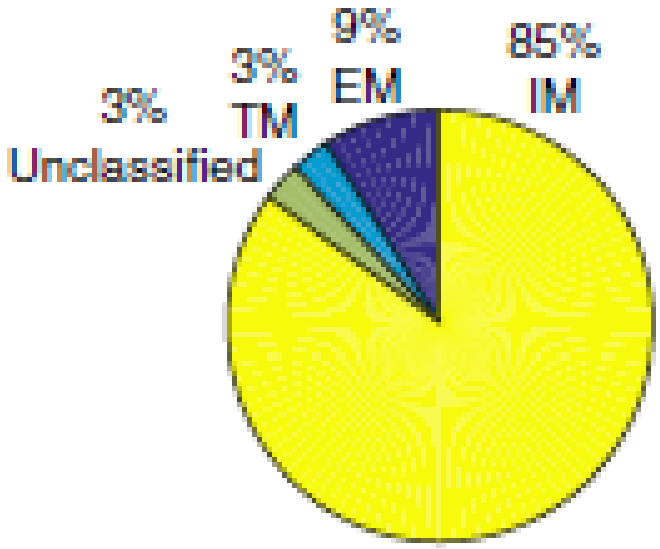
- X & Y associated: p_1
- X & M|Y associated: p_2
- M & Y|X associated: p_3
- $X \perp Y|M$: p_4

Multiple Testing

- $p = \max(p_1,p_2,p_3,p_4)$ based on intersection-union test, i.e. weakest link
- Correct for number of mediation tested

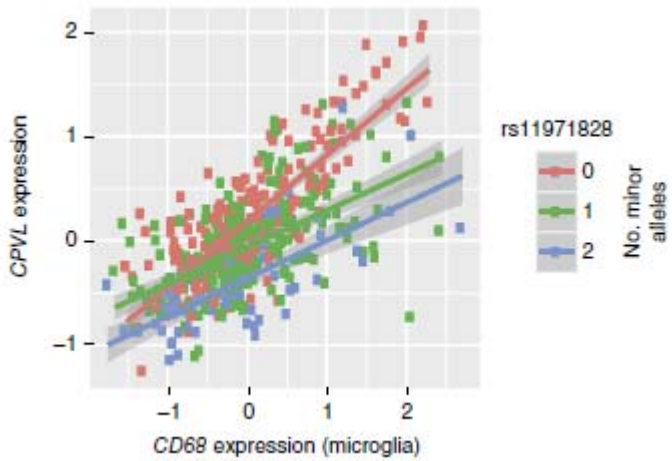
Mediation Analysis





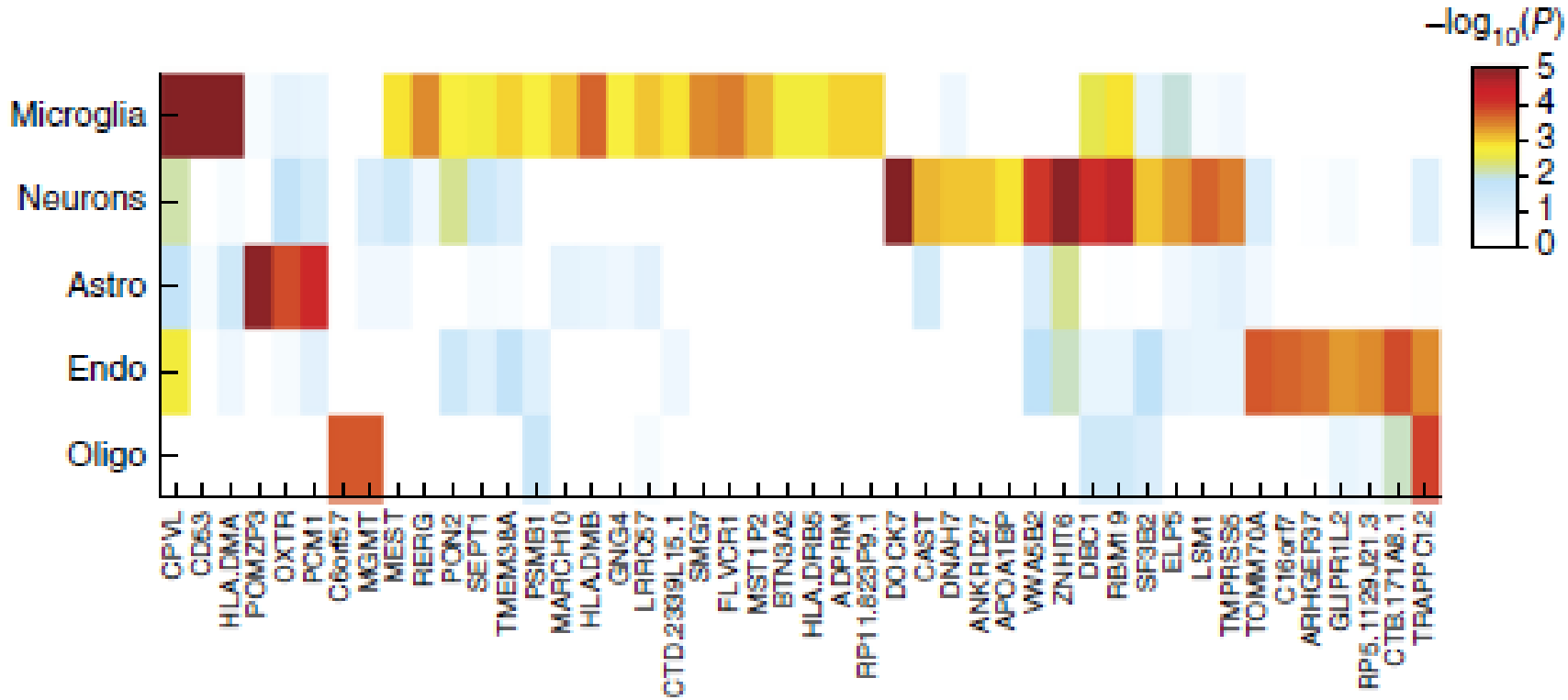
Cell Type Specific Analysis

- $E_i = S_j\alpha + C\beta + S{\cdot}C\gamma + \varepsilon$
  - E_i = Expression of gene *i*
  - S_j = Genotype values of SNP *j*
  - C = cell type proportion
  - $S{\cdot}C$ = element-wise product



- C estimated based on expression markers
  - ENO2 for Neuron
  - OLIG2 for Oligodendrocyte
  - GFAP for Astrocyte
  - CD68 for Microglia
  - CD34 for Endothelial

Cell Type Specific Analysis



- 46 genes at liberal FDR of 0.2
- This type of analysis falls under the general area of GxE interaction, which is nontrivial to detect

Weighted GWAS

Theory

- Given $\{p_i\}$, $i=1,\ldots,n$
- If $\{w_i p_i\}$, $w_i \geq 0$ and $\sum_i w_i = 1$, then FWER controlled

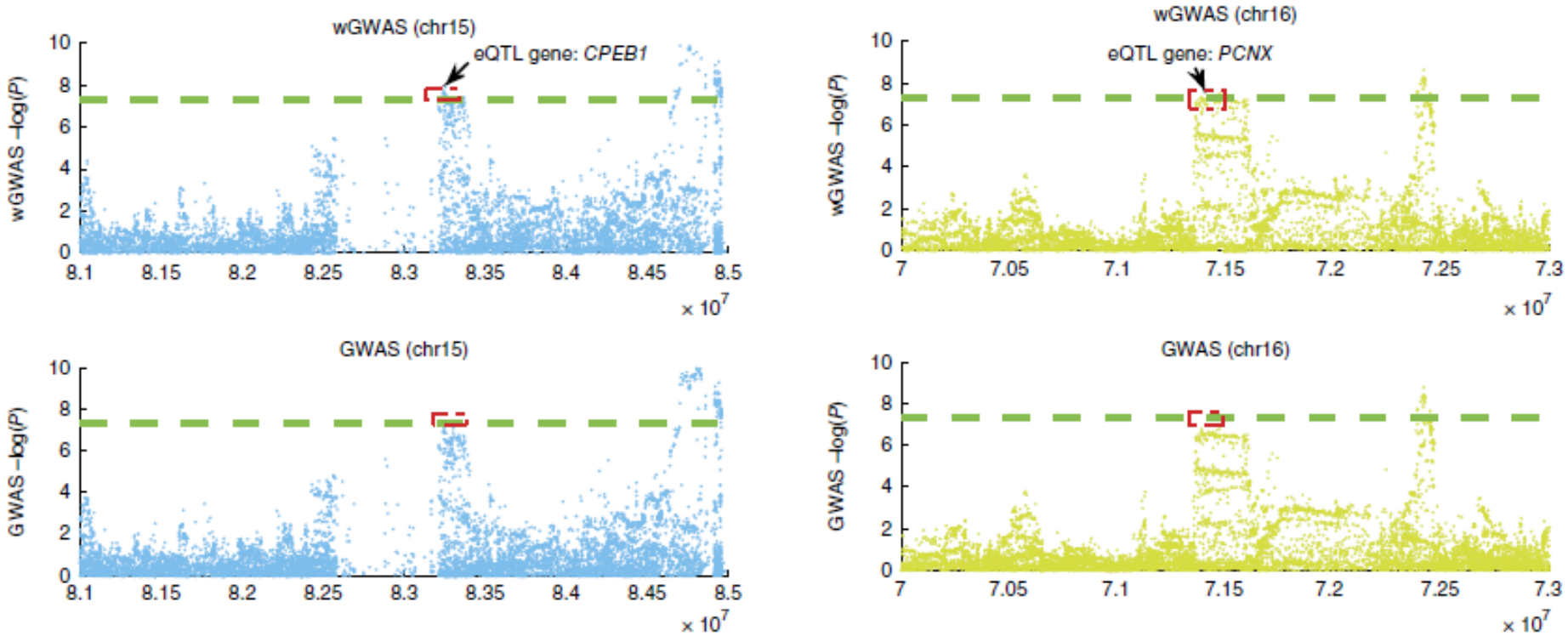
Binary Strategy

- $w_1 = s/[1 + (s - 1)n_1/n]$
- $w_0 = 1/[1 + (s - 1)n_1/n]$
- $s = w_1/w_0$ ranging from 1 to 100
- n_1 = #xQTL SNPs, n = #SNPs

s Selection

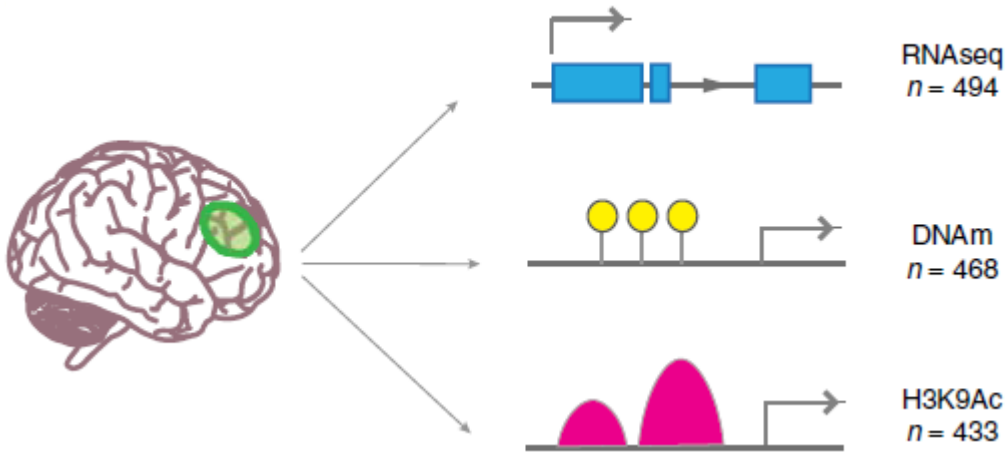
- Randomly split the SNPs into 2 halves
- $J(s) = (D^1(s)/\pi_1^1 + D^2(s)/\pi_1^2) / |D^1(s)/\pi_1^1 - D^2(s)/\pi_1^2|$

Weighted GWAS



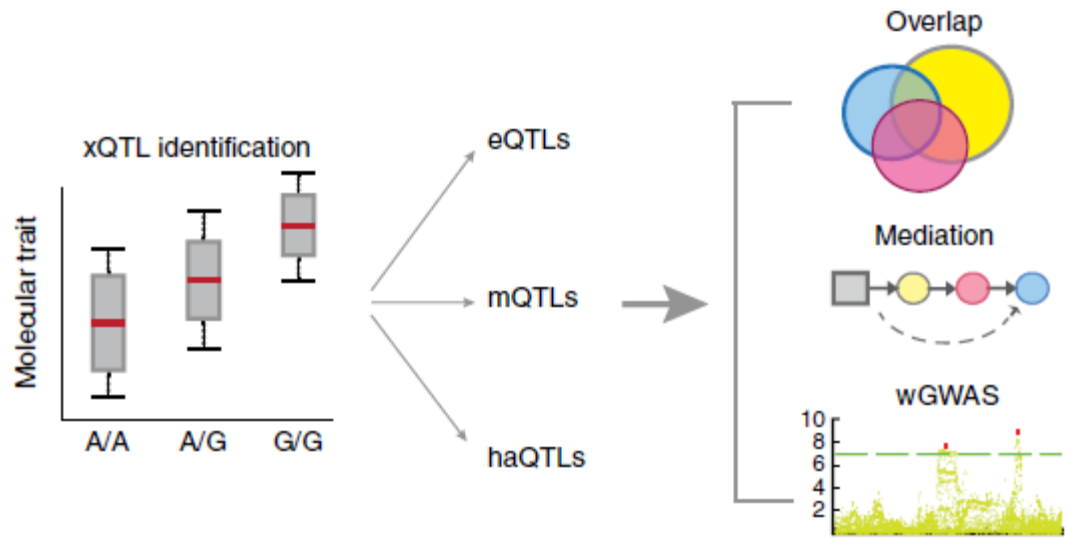
Applied on the largest Schizophrenia data set that found 108 loci, we found **18** additional loci that met genome-wide significance.

Summary



- Recap of GWAS
- xQTL Analysis
- Challenges

- Replication
- xQTL Sharing
- Mediation Analysis
- Cell Specific Analysis
- Weighted GWAS



xQTL Serve

RESOURCE

nature neuroscience

An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome

Bernard Ng[1,2], Charles C White[3], Hans-Ulrich Klein[3,4], Solveig K Sieberts[5], Cristin McCabe[3], Ellis Patrick[3], Jishu Xu[3], Lei Yu[6], Chris Gaiteri[6], David A Bennett[6], Sara Mostafavi[1,2,7,8] & Philip L De Jager[3,4,8]

http://mostafavilab.stat.ubc.ca/xQTLServe/