6.874, 6.802, 20.390, 20.490, HST.506
Computational Systems Biology
Deep Learning in the Life Sciences

Lecture 13 – GWAS mechanism
Epigenomic Enrichments, eQTLs, Mediation, Causality

Prof. Manolis Kellis
Guest lecture: Yongjin Park

Massachusetts Institute of Technology

http://mit6874.github.io

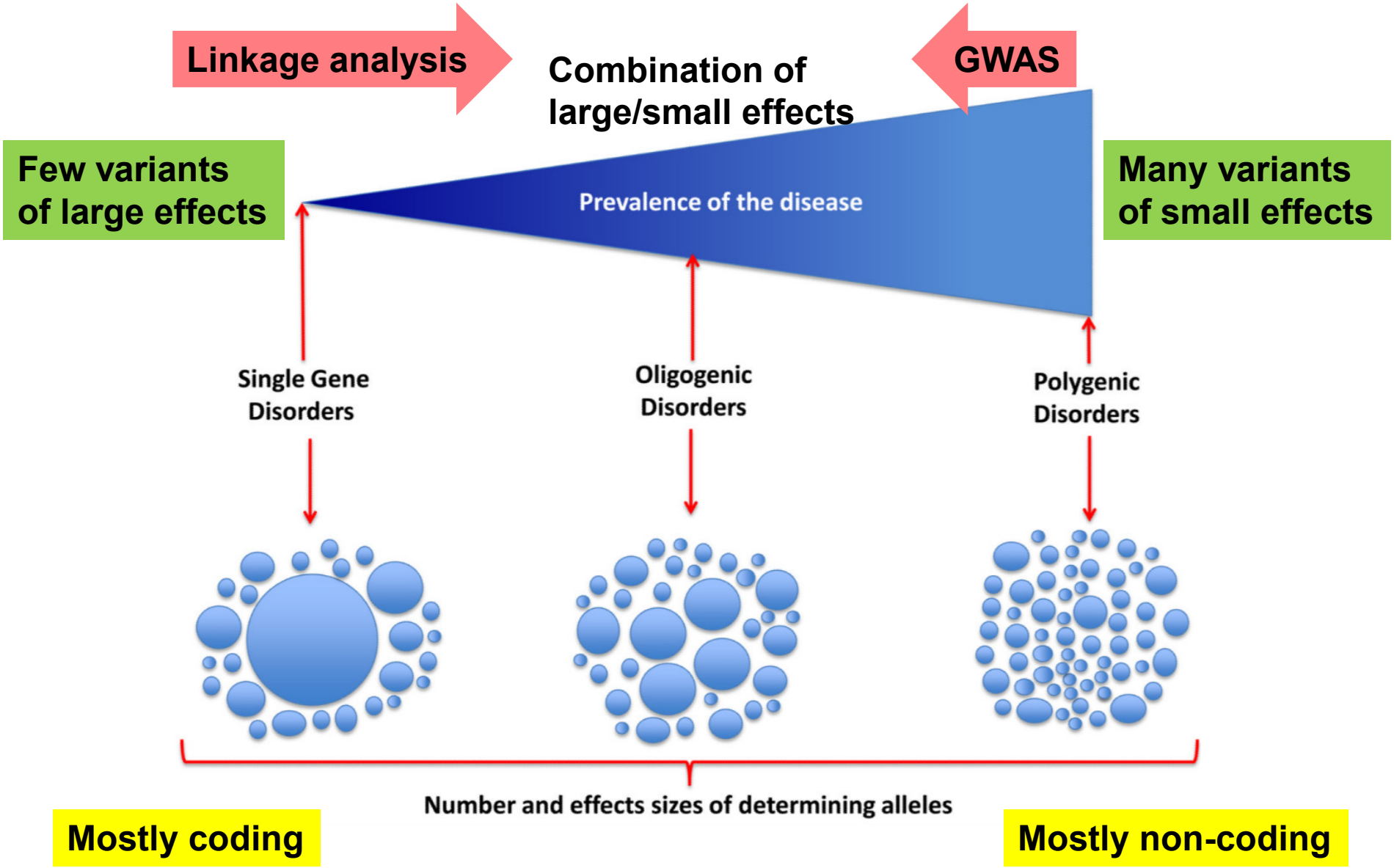
Slides credit: Yongjin Park, Abhishek Sarkar, Mark Daly, David Gifford, et al

GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

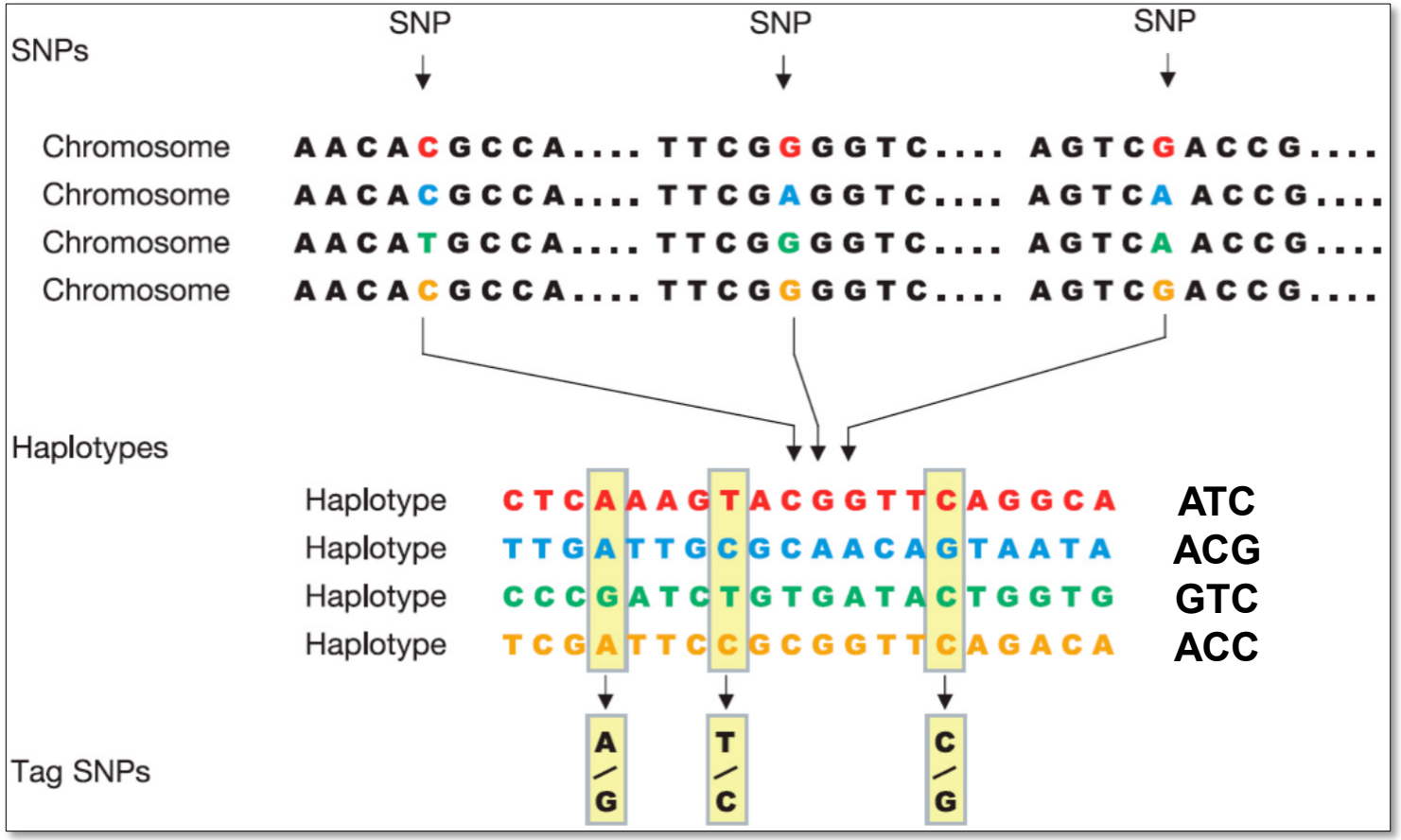
1. Review: GWAS, fine-mapping, Bayesian methods for variant prioritization

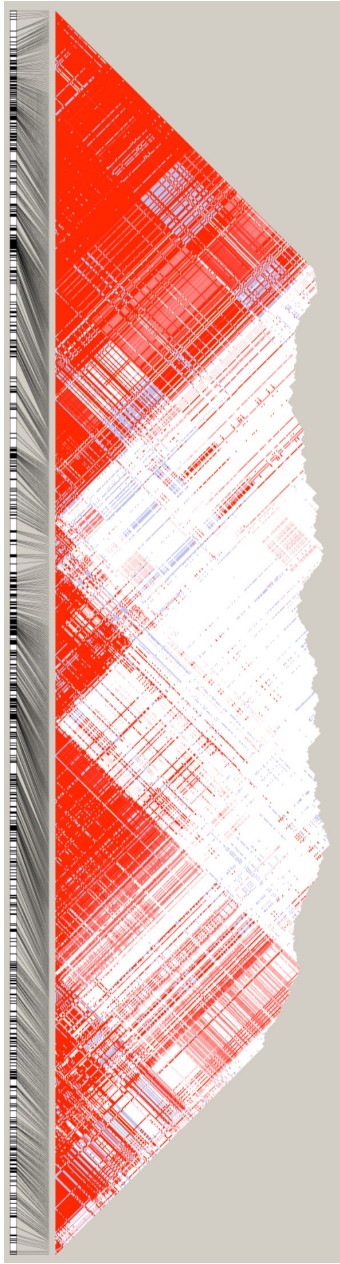
Monogenic vs. oligogenic vs. polygenic disorders



Molecular genetic studies of complex phenotypes. Marian - Translational Research, 2012

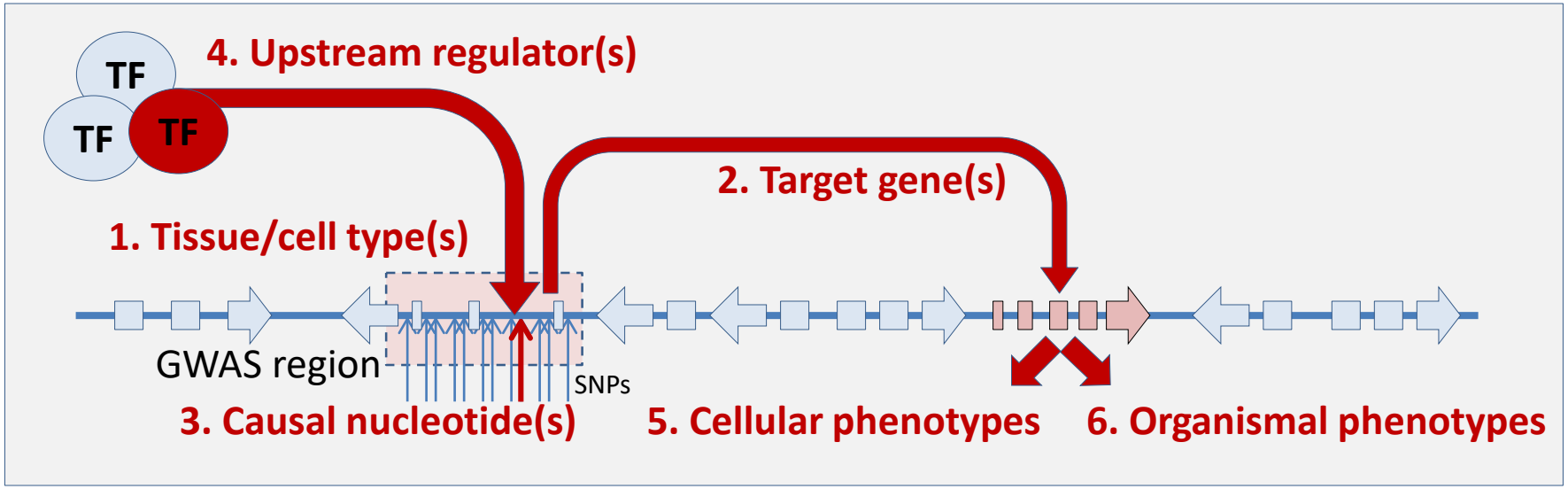
Common variants (SNPs) live in Haplotypes





- Common SNPs only once every 1000 nucleotides or so
- These are co-inherited, so only need to profile a subset
- Markers selected for haplotype profiling are “tag” SNPs

Dissecting non-coding genetic associations

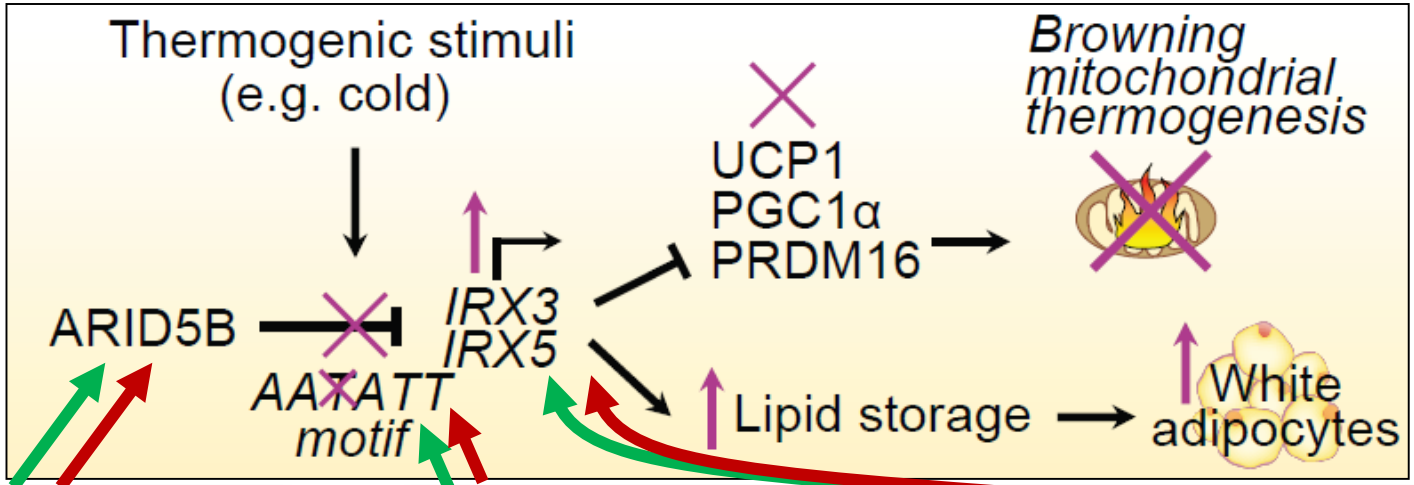


1. Establish relevant **tissue/cell type**
2. Establish downstream **target** gene(s)
3. Establishing **causal** nucleotide variant
4. Establish upstream **regulator** causality
5. Establish **cellular** phenotypic consequences
6. Establish **organismal** phenotypic consequences

<u>Goal:</u>

Apply these to the FTO locus in obesity

Manipulate circuitry ➔ reverse disease phenotypes



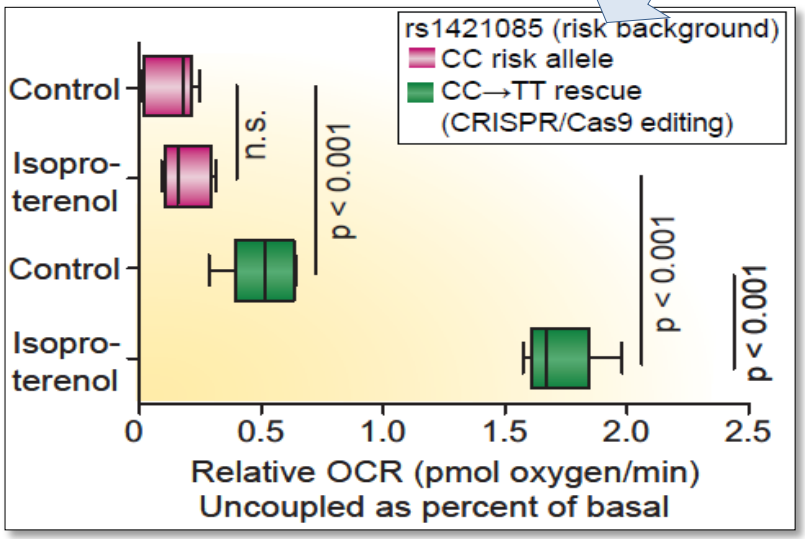
Lean

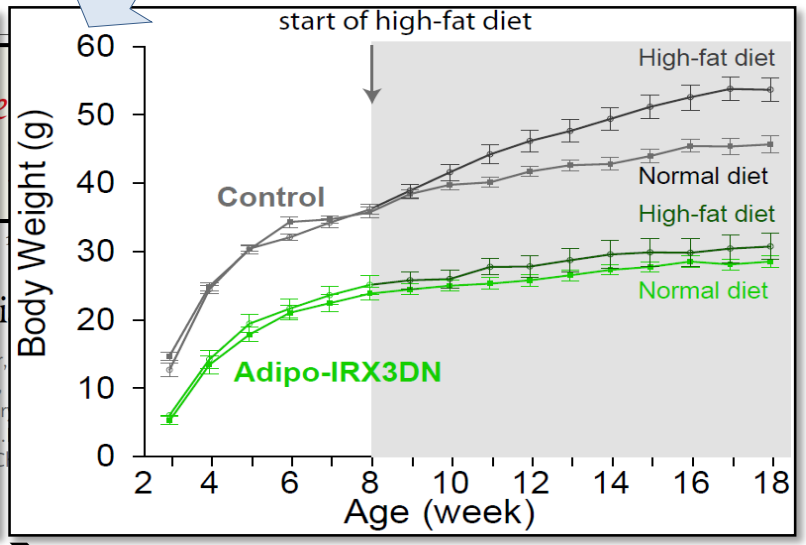
Obese

Incr. ARID5B ➔ Lean
Decr ARID5B ➔ Obese

C-to-T ➔ Lean
T-to-C ➔ Obese

Decrease IRX3, IRX5 ➔ Lean
Increase IRX3, IRX5 ➔ Obese





CRISPR-edit human fat cells
➔ able to burn calories again

IRX3 KD ➔ Burn calories in their sleep
➔ 54% weight loss. Can't gain weight

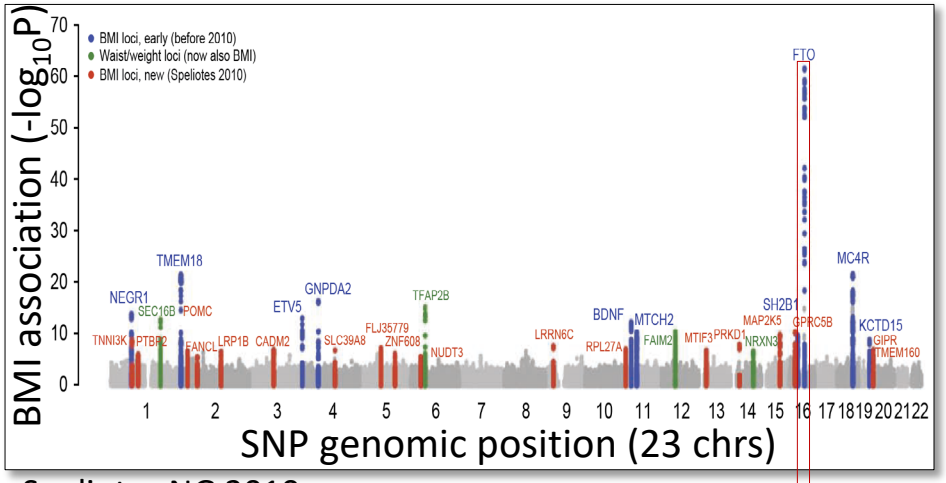
GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

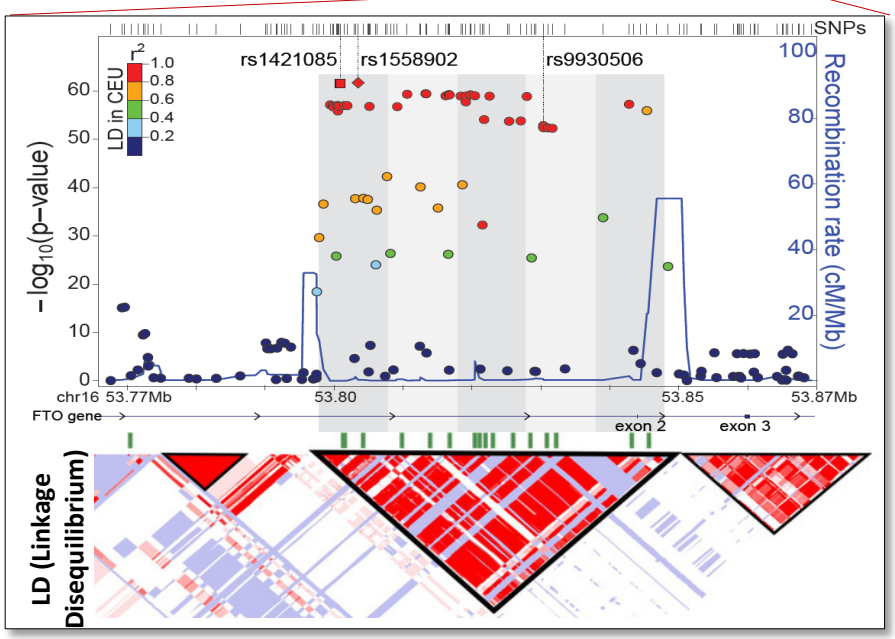
2. Global enrichment analyses:
Predicting disease-relevant Tissues, Regulators, Cell Types, Target Genes

Genomic medicine today: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

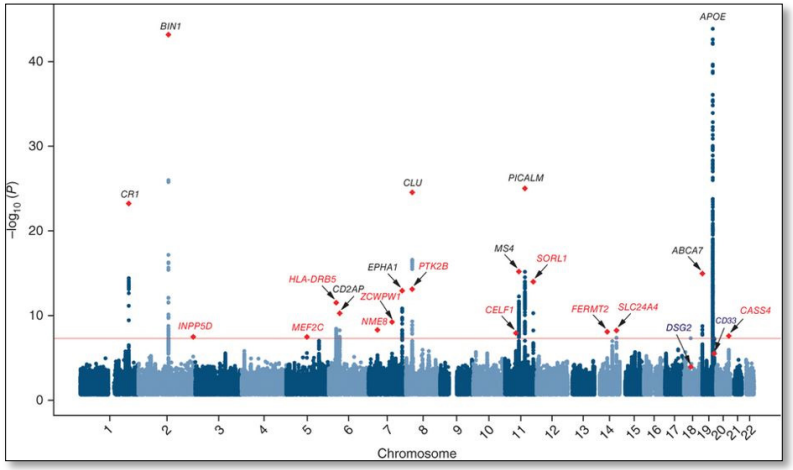
The promise of genetics

- Unbiased, Causal, Uncorrected
- New disease mechanisms
- New target genes
- New therapeutics
- Personalized medicine

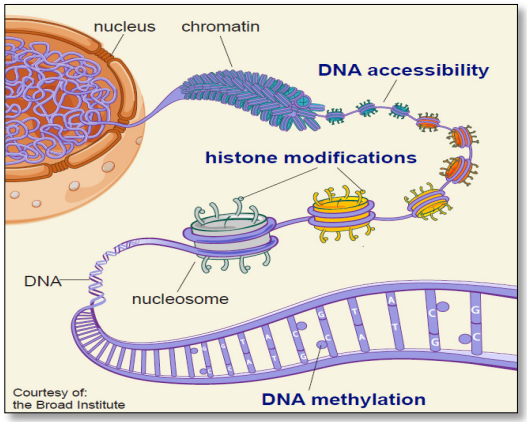
The challenge of mechanism

- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

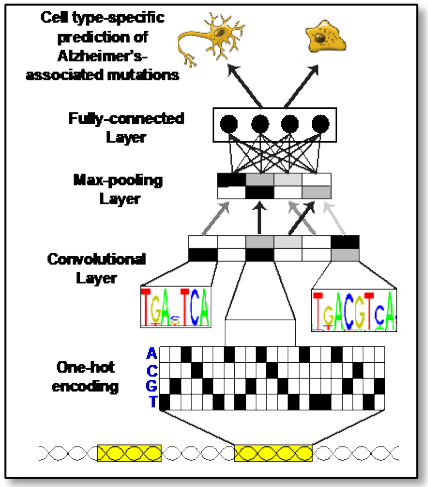
Dissect mechanisms of disease-associated regions



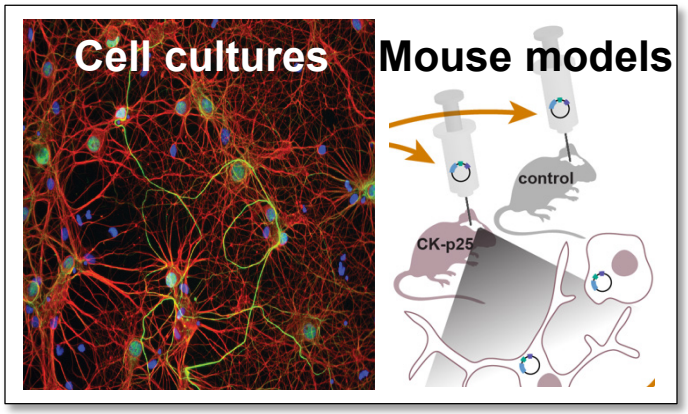
1. Disease genetics reveals common + rare variants/regions



2. Profile RNA + Epigenome in healthy + disease samples



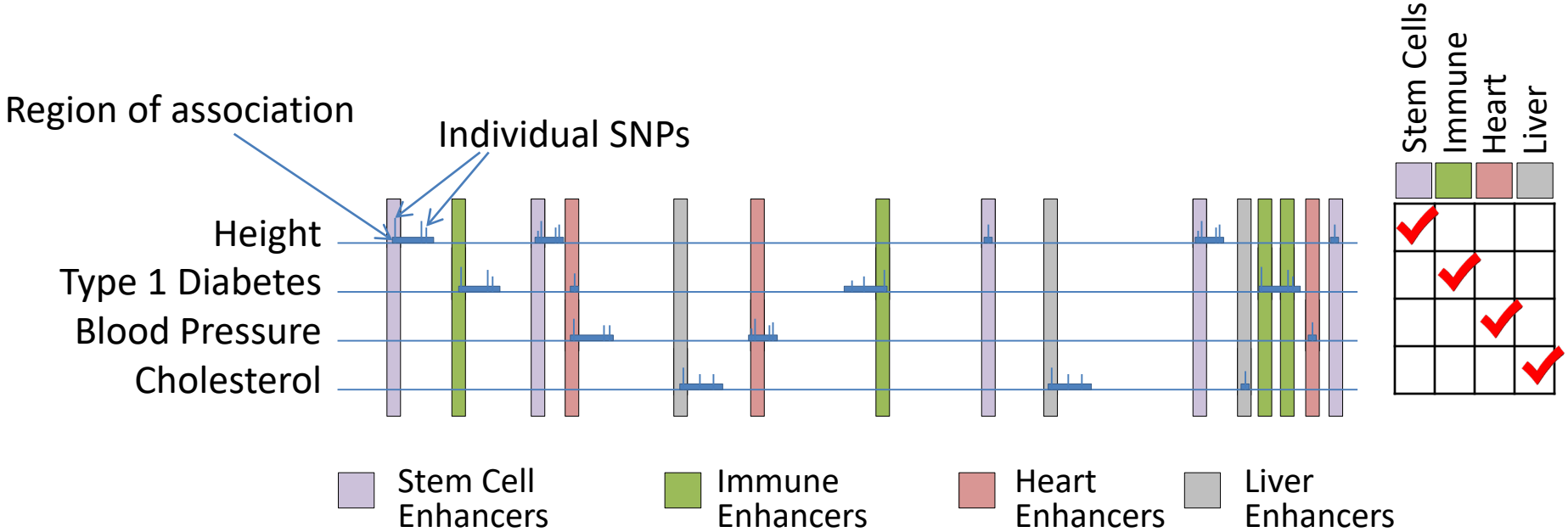
3. Integrate data to predict driver genes, regions, cell types



4. Validate predictions in human cells + mouse models

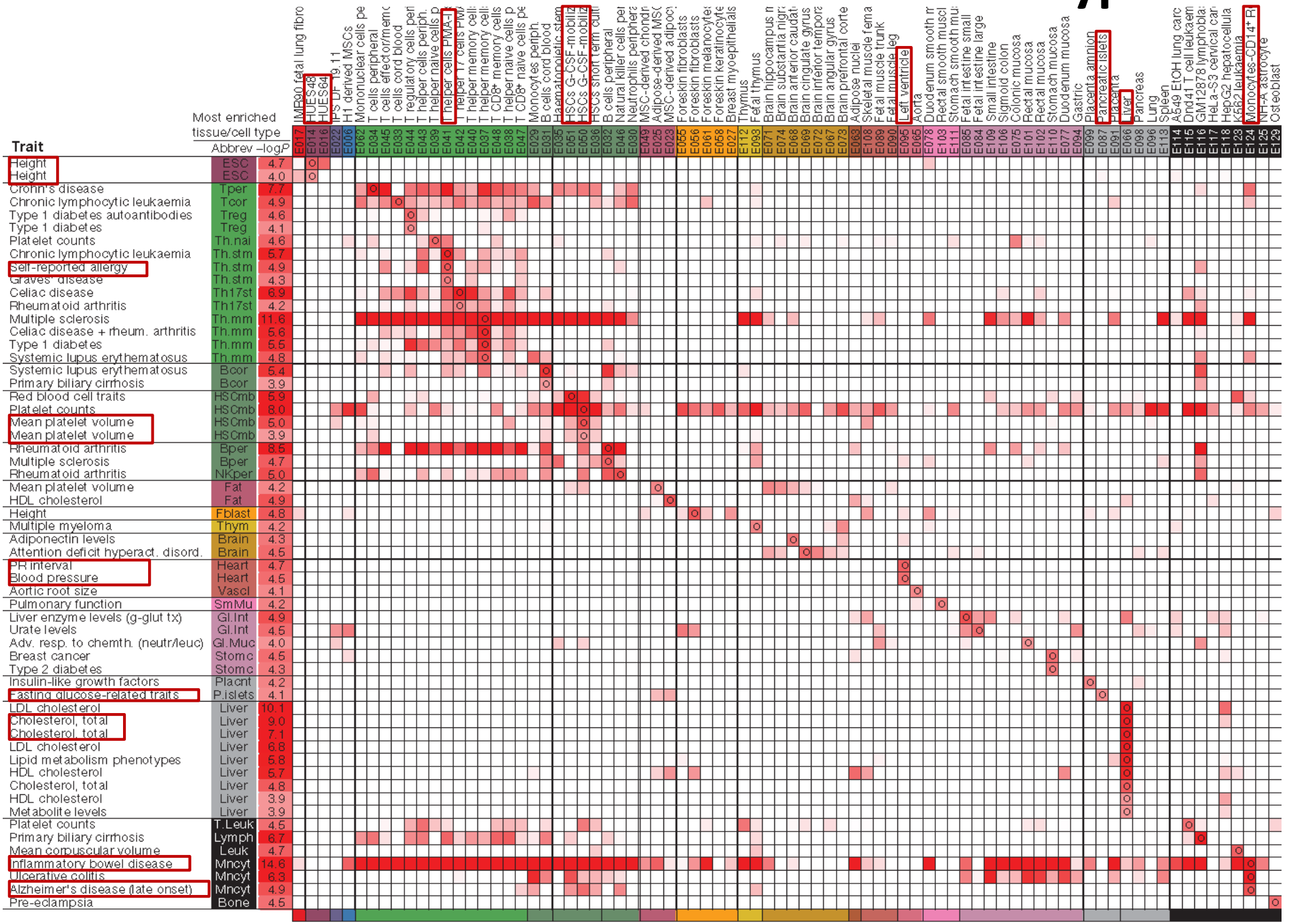
5. Disseminate results

Identifying disease-relevant cell types

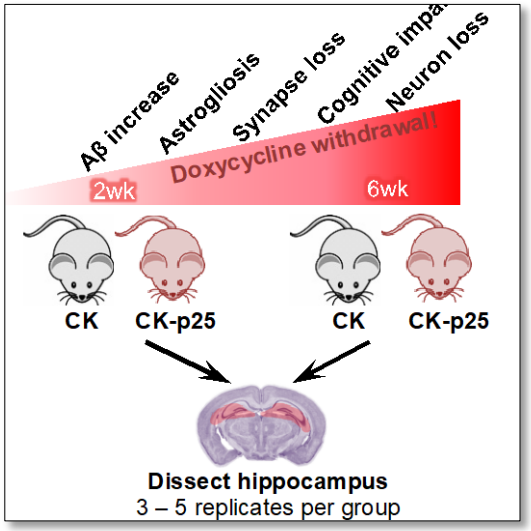


- For every trait in the GWAS catalog:
  - Identify all associated regions at P-value threshold
  - Consider all SNPs in credible interval ($R^2 \geq .8$)
  - Evaluate overlap with tissue-specific enhancers
  - Keep tissues showing significant enrichment ($P<0.001$)
- Repeat for all traits (rows) and all cell types (columns)

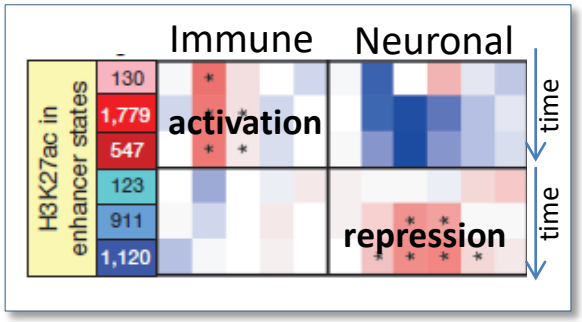
GWAS hits in enhancers of relevant cell types



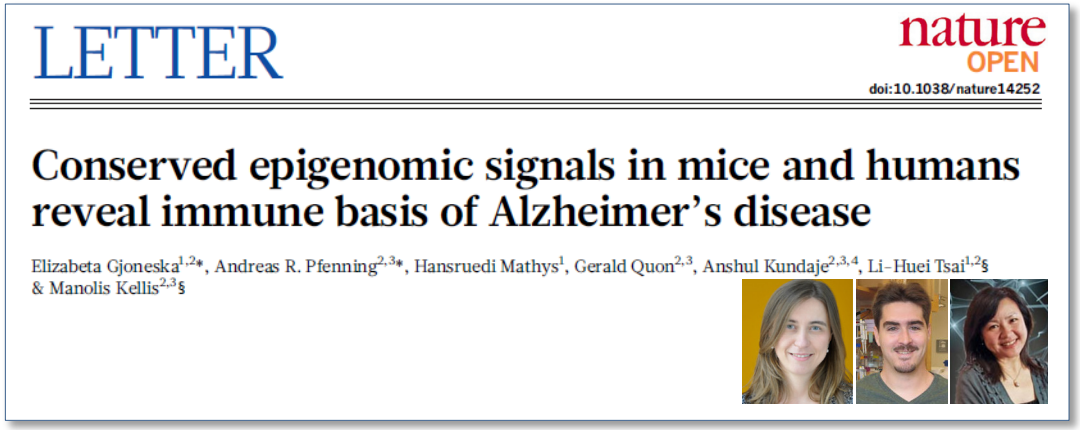
Immune activation + neural repression in human + mouse



Epigenomics of AD progression



Immune activation precedes neuronal repression



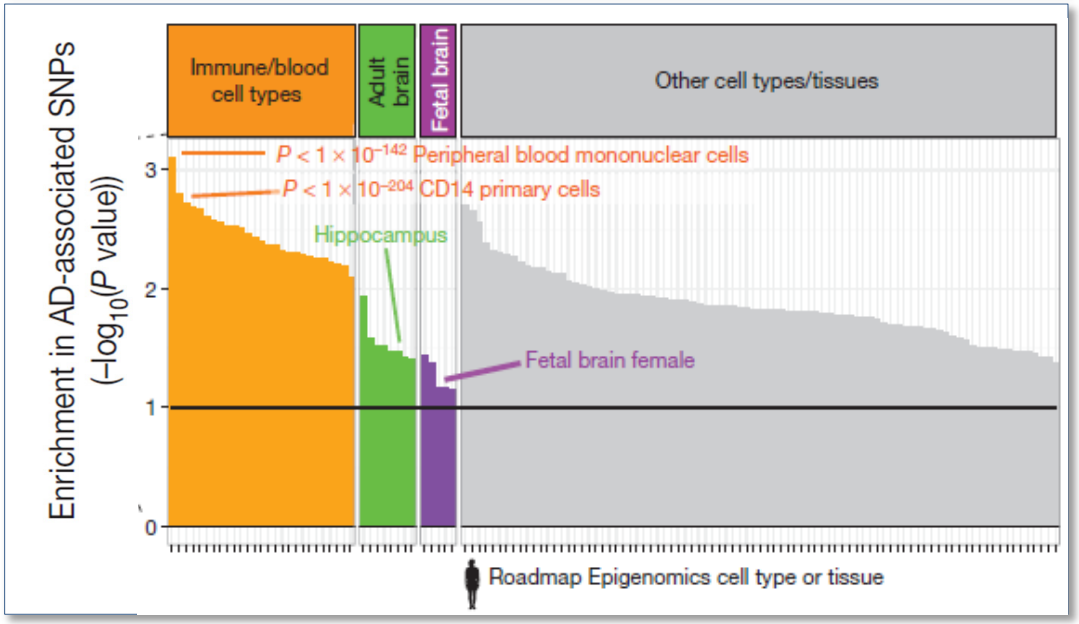
LETTER

nature OPEN

doi:10.1038/nature14252

Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease

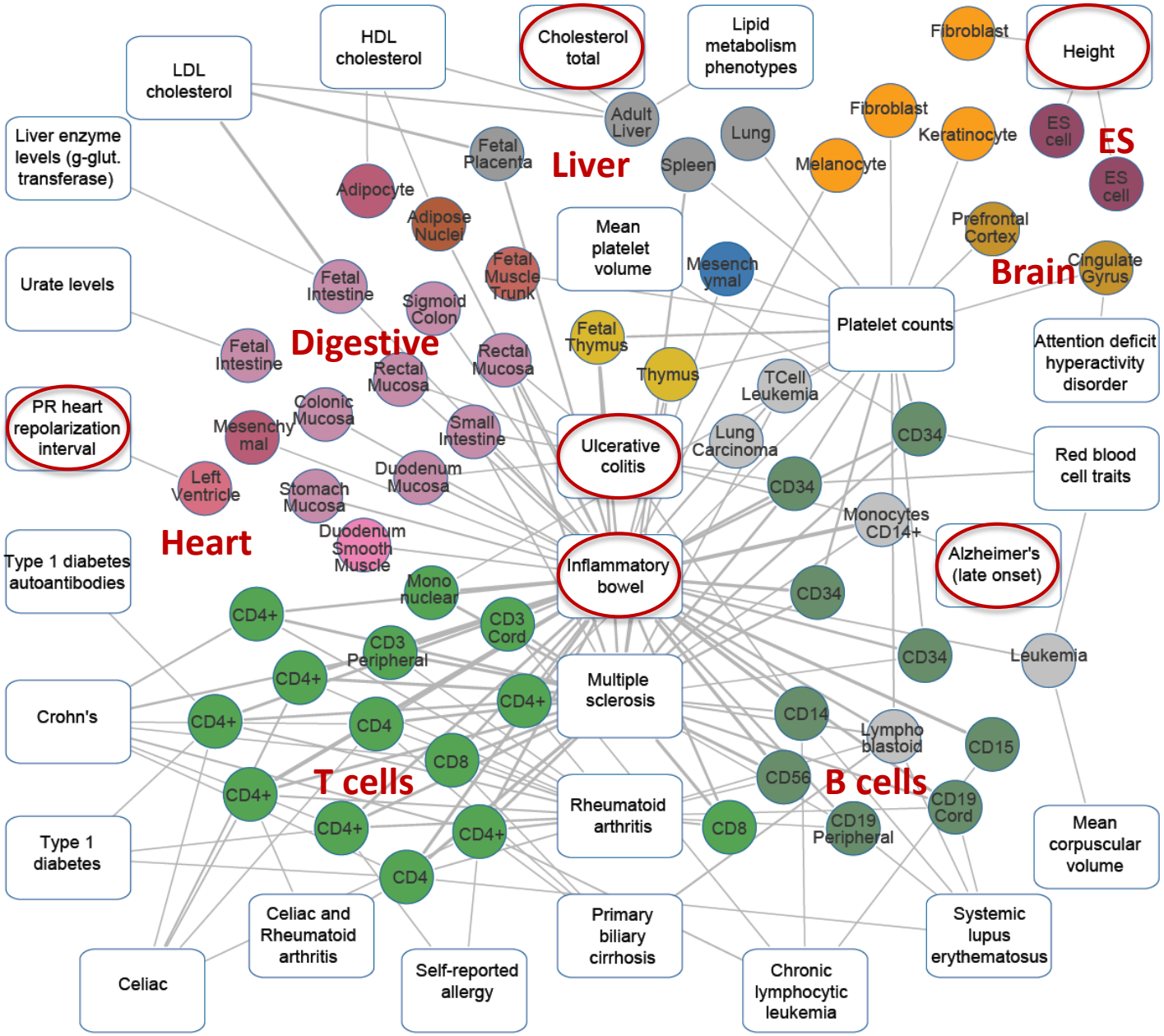
Elizabeta Gjoneska[1,2]*, Andreas R. Pfenning[2,3]*, Hansruedi Mathys[1], Gerald Quon[2,3], Anshul Kundaje[2,3,4], Li-Huei Tsai[1,2]§ & Manolis Kellis[2,3]§



AD variants localize in immune cells, not neuronal

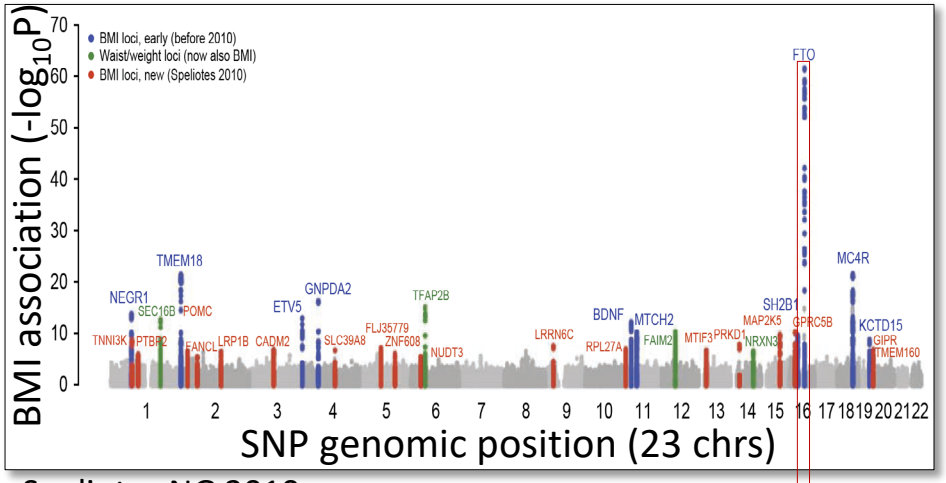
Inflammation as the causal component of Alzheimer's disease

Linking traits to their relevant cell/tissue types

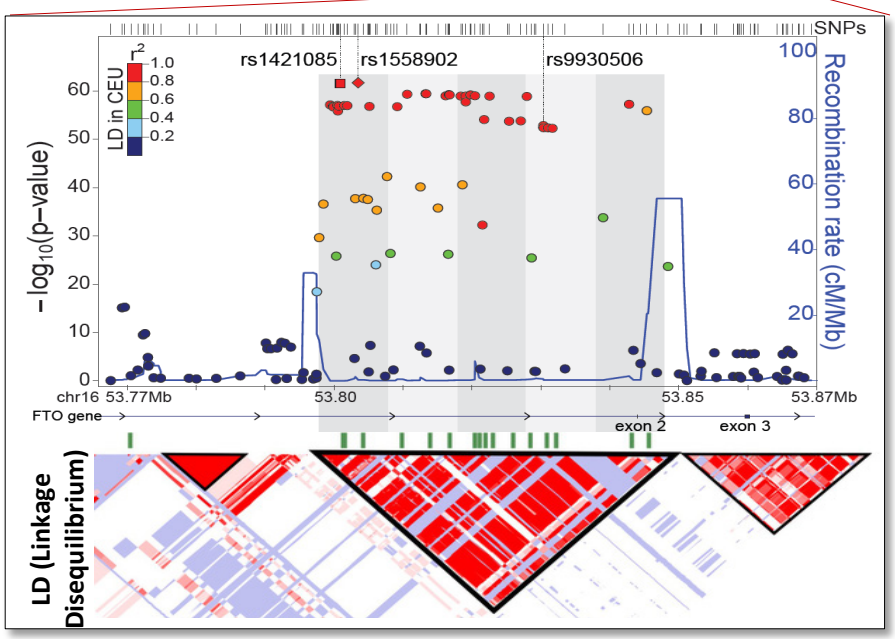


Genomic medicine: challenge and promises

GWAS Manhattan Plot: simple χ^2 statistical test



Speliotes NG 2010



Dina NG 2007, Frayling Science 2007, Claussnitzer NEJM 2015

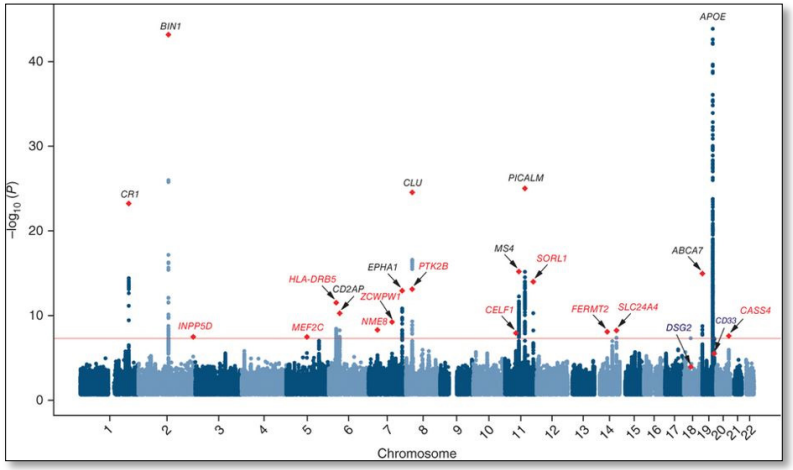
The promise of genetics

- Disease mechanism
- New target genes
- New therapeutics
- Personalized medicine

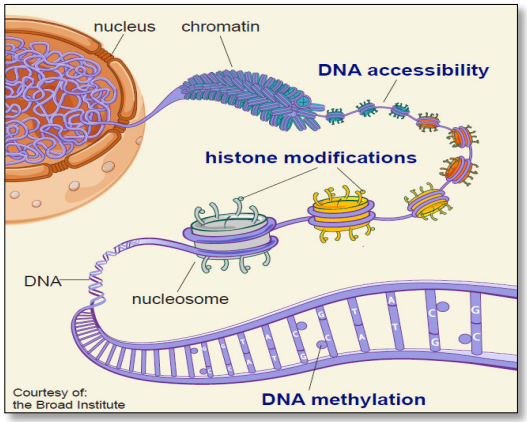
The challenge of mechanism

- **90+% disease hits non-coding**
- Target gene not known
- Causal variant not known
- Cell type of action not known
- Relevant pathways not known
- Mechanism not known

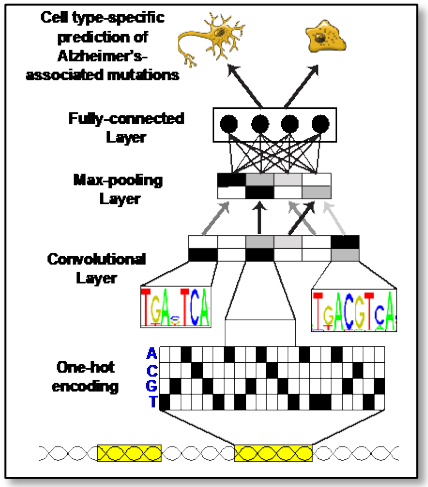
Summary: Dissect circuitry of disease-associated regions



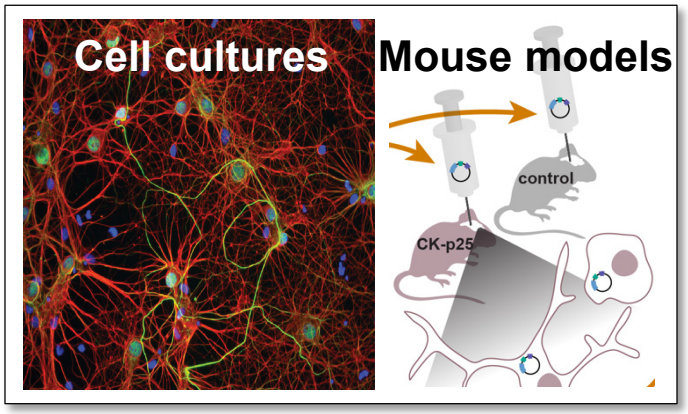
1. Disease genetics reveals common + rare variants/regions



2. Profile RNA + Epigenome in healthy + disease samples



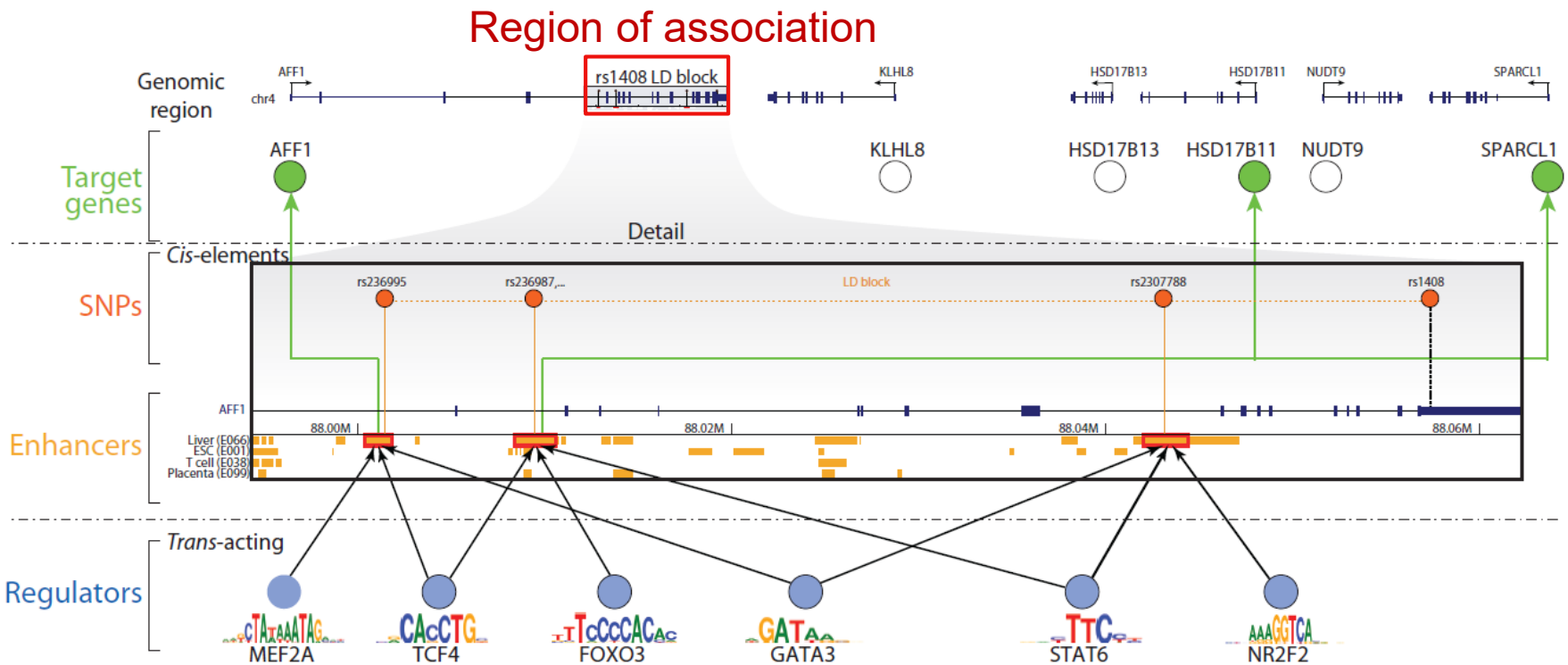
3. Integrate data to predict driver genes, regions, cell types



4. Validate predictions in human cells + mouse models

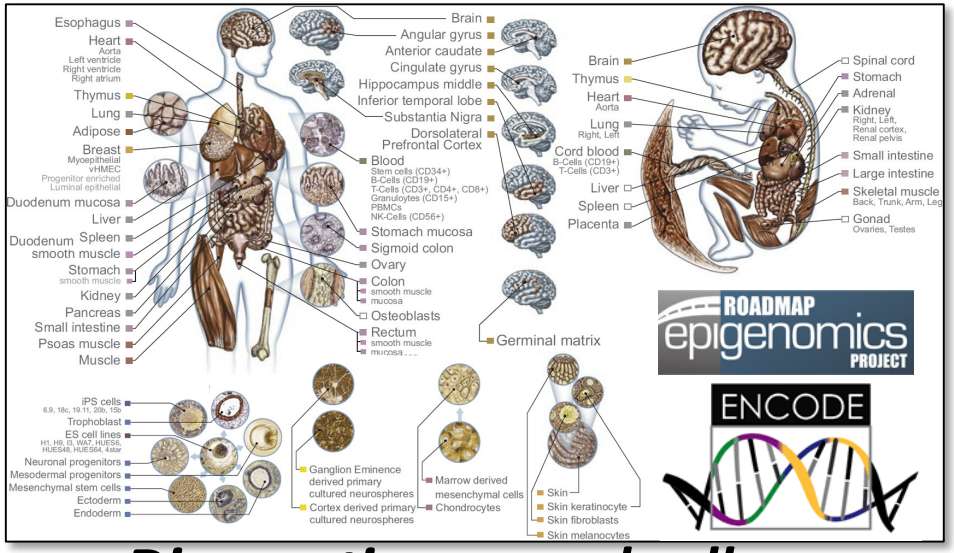
5. Disseminate results

Regulatory circuitry of GWAS loci



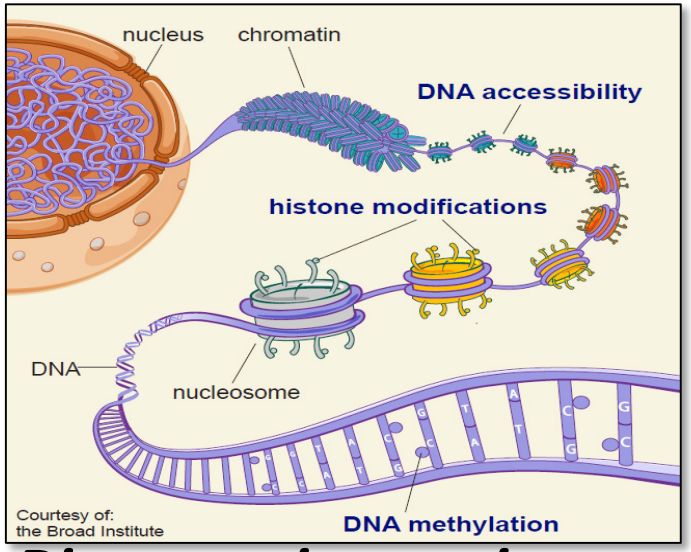
- Expand each GWAS locus using SNP linkage disequilibrium (LD)
  - Recognize **relevant cell types**: tissue-specific enhancer enrichment
  - Recognize **driver TFs**: enriched motifs in multiple GWAS loci
  - Recognize **target genes**: linked to causal enhancers

Epigenomic mapping across 800+ tissues/cell types

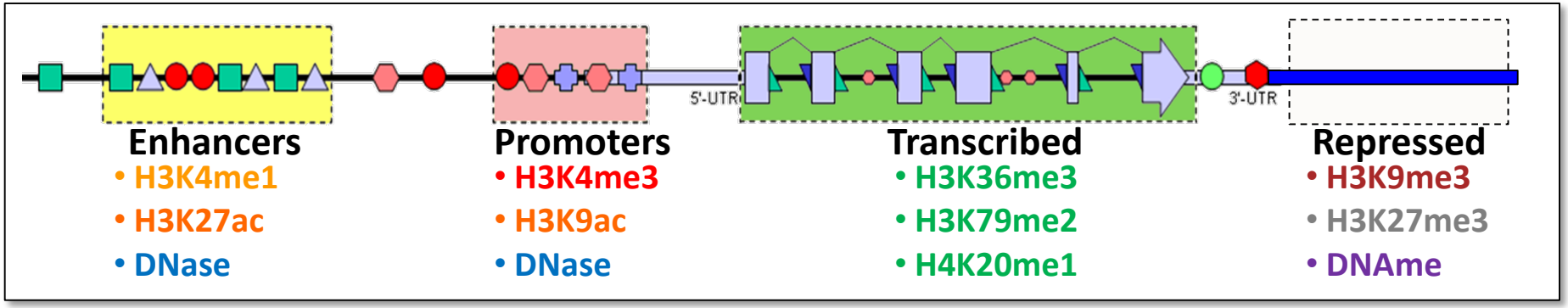


Diverse tissues and cells

X

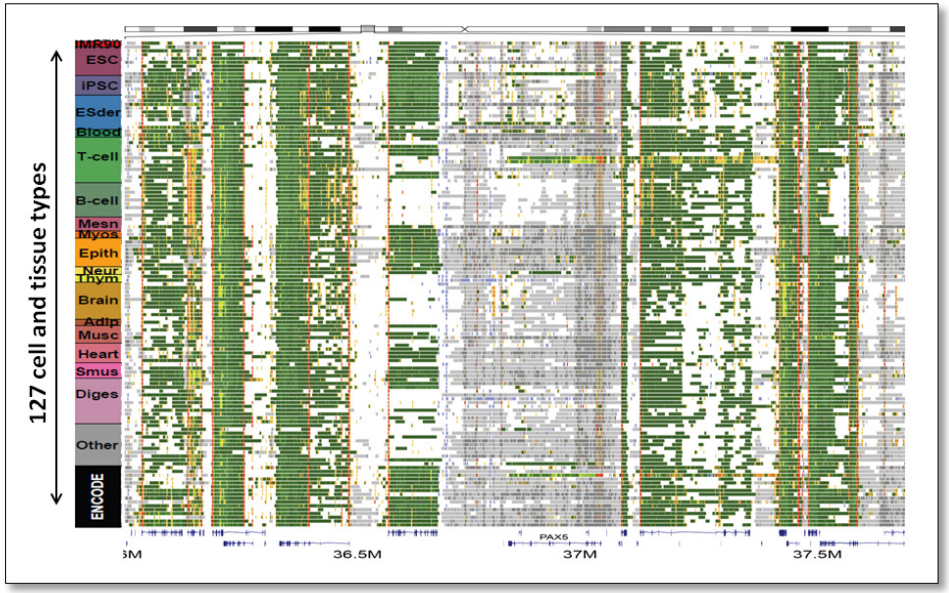


Diverse epigenomic assays

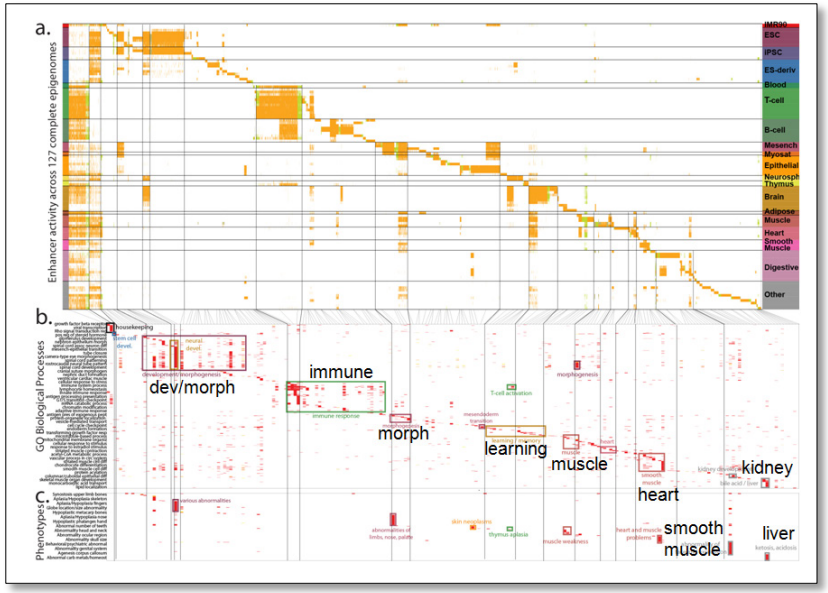


*Their **combinations** define diverse classes of elements*

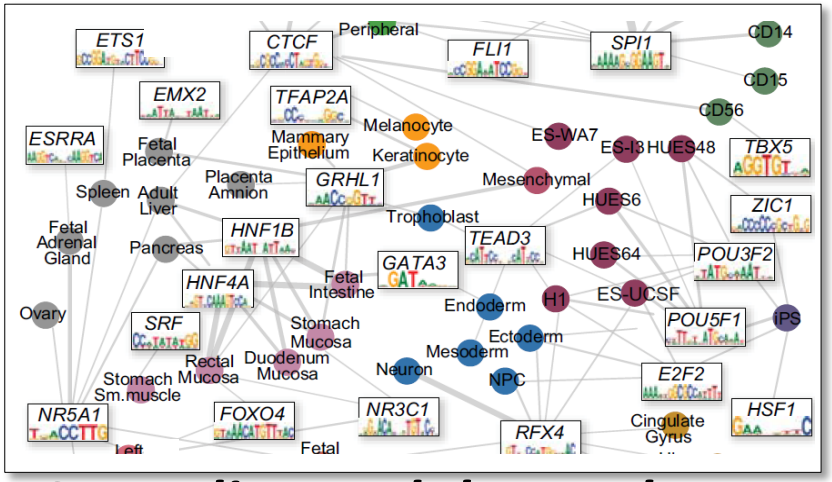
Enhancer modules, regulators, and target genes



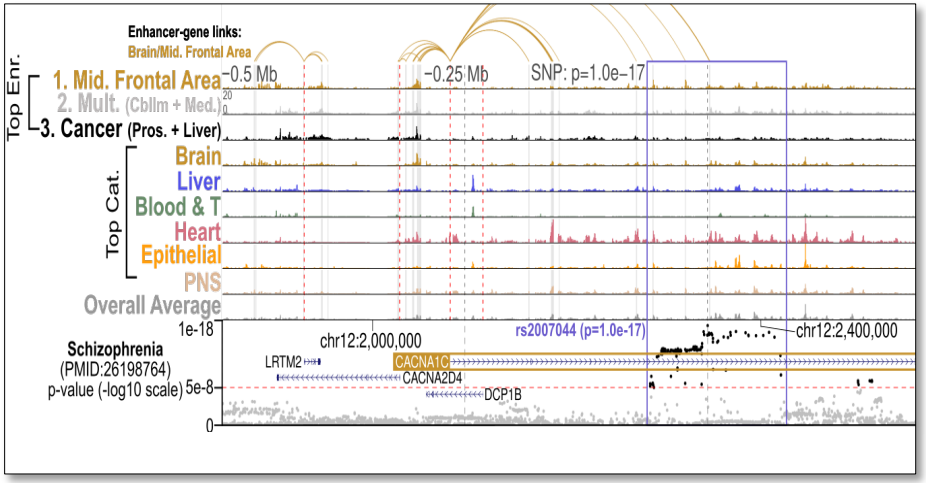
1. *Map chromatin states across 127 tissue/cells*



2. *Group enhancers into modules of common function*

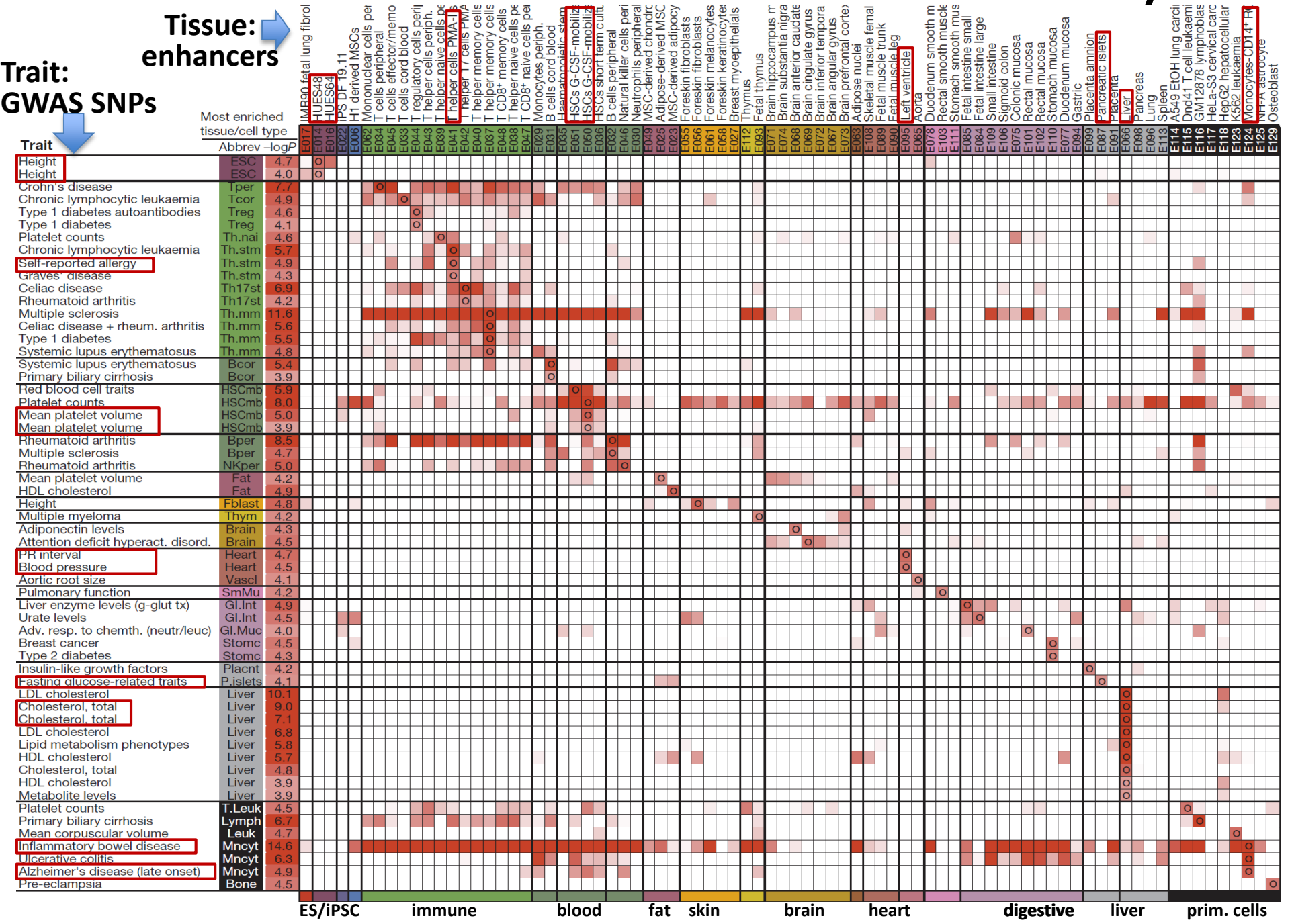


3. *Predict module regulators using motif enrichment*

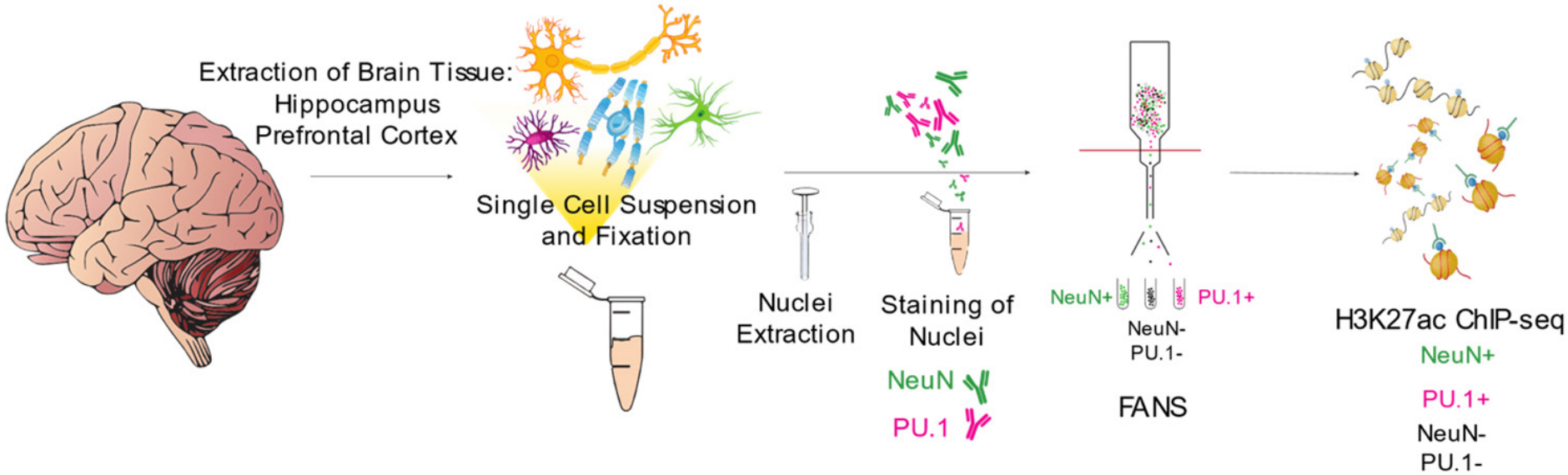


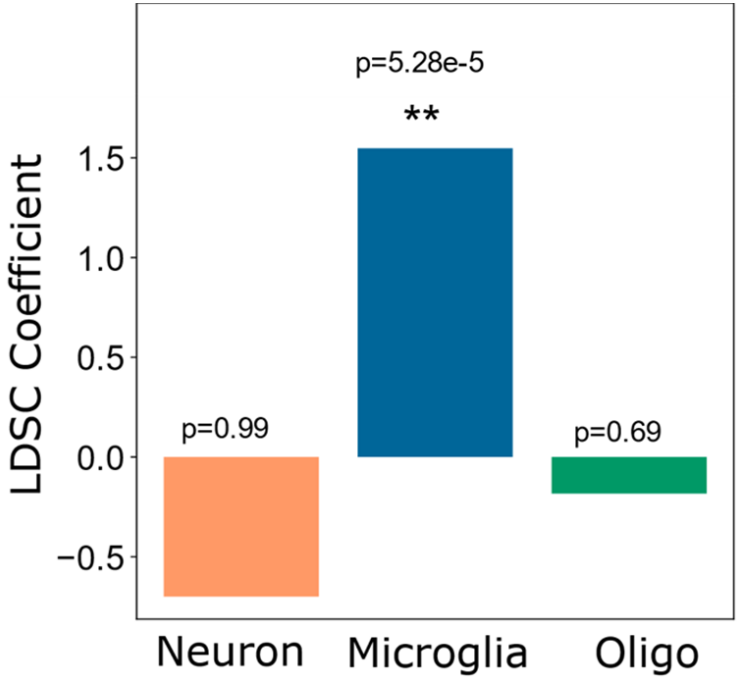
4. *Predict target genes using activity correlation, Hi-C, eQTLs*

Enhancer enrichment reveals trait-relevant tissues/cells



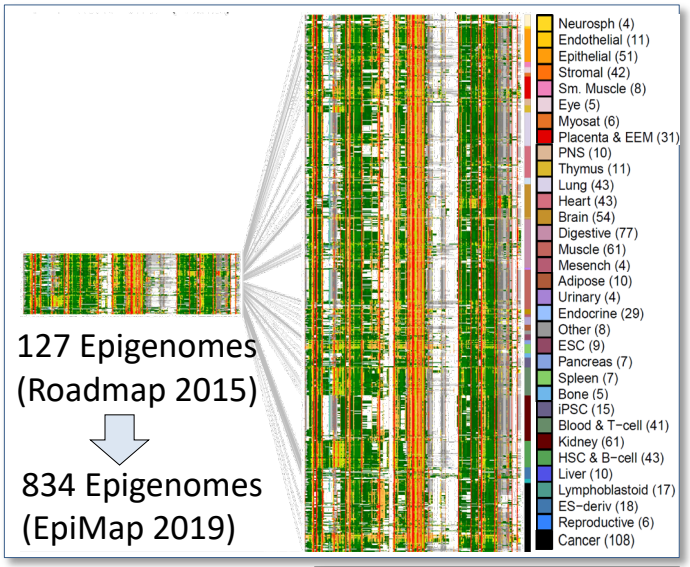
Cell-sorted H3K27ac ➔ AD variants in microglia, not neurons





- **No enrichment found in whole-brain samples**
- **Cell-sorted H3K27ac shows strong enrichment for AD variants in microglia**
- **No enrichment found in neurons or oligodendrocyte H3K27ac for AD variants**

EpiMap: 834 tissue/cell types ➔ 30k GWAS SNPs in 534 traits



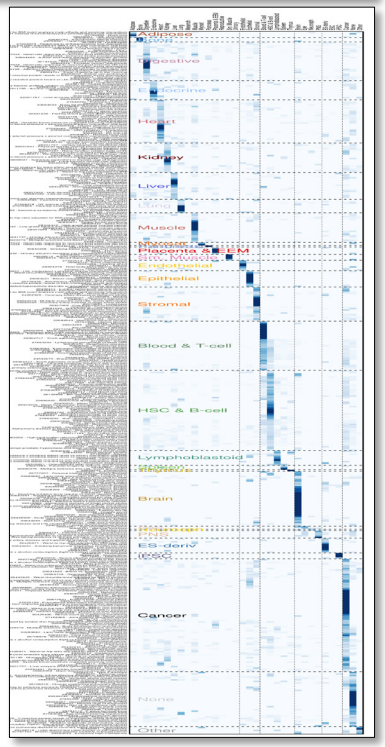
127 Epigenomes (Roadmap 2015)

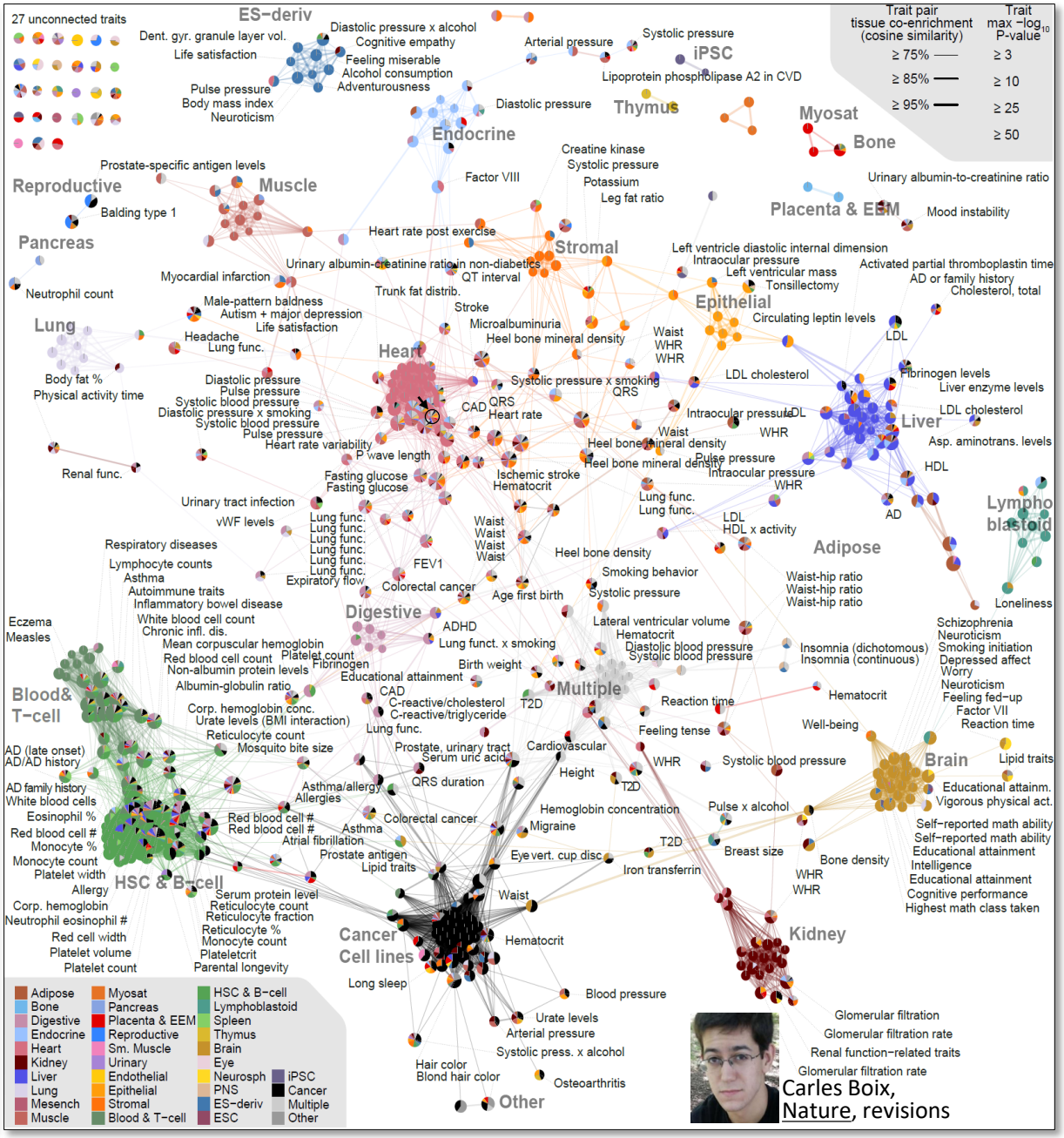
834 Epigenomes (EpiMap 2019)

54 enriched GWAS traits (2015)

534 enriched traits

30,247 SNPs in enriched enhancers ➔ Highly-specific associations Emerge ➔ Precise biological hypotheses on mechanistic basis



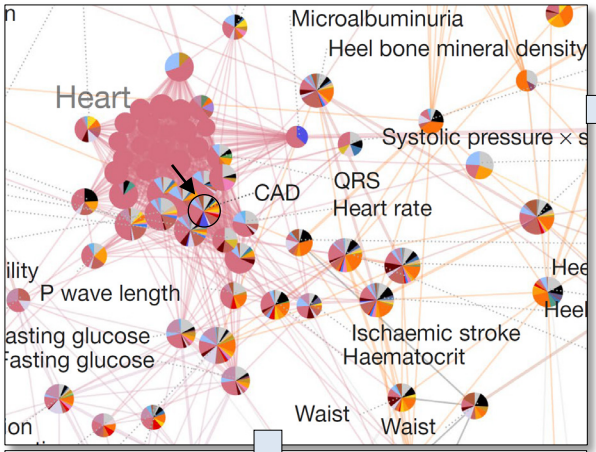


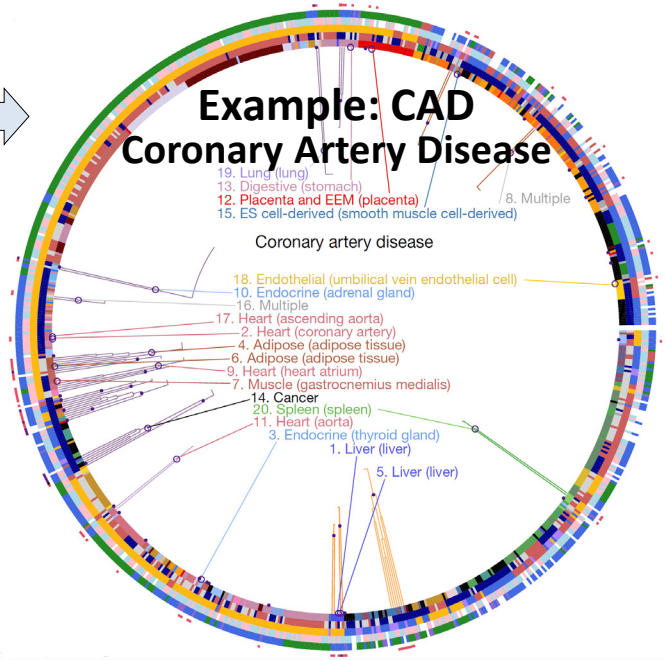
Carles Boix, Nature, revisions

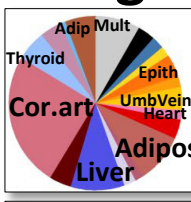
➔ http://compbio.mit.edu/epimap

Tissue enrich/co-enrichments ➔ trait clustering, trait-tissue network

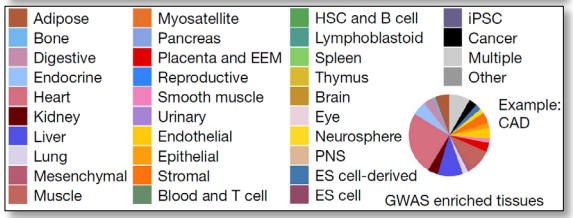
Dissect circuitry of 30,000 GWAS loci: TF➔Enh➔SNP➔gene➔pathways



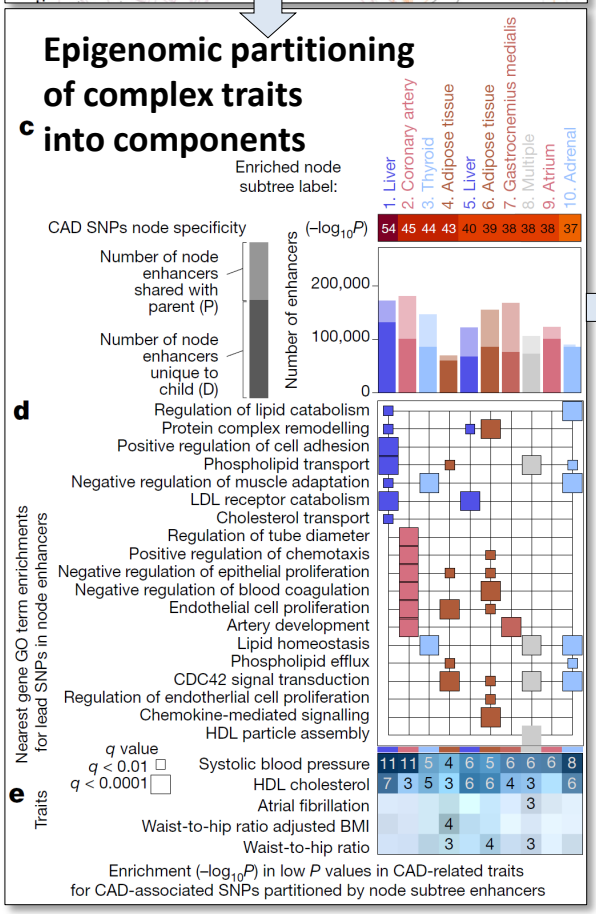


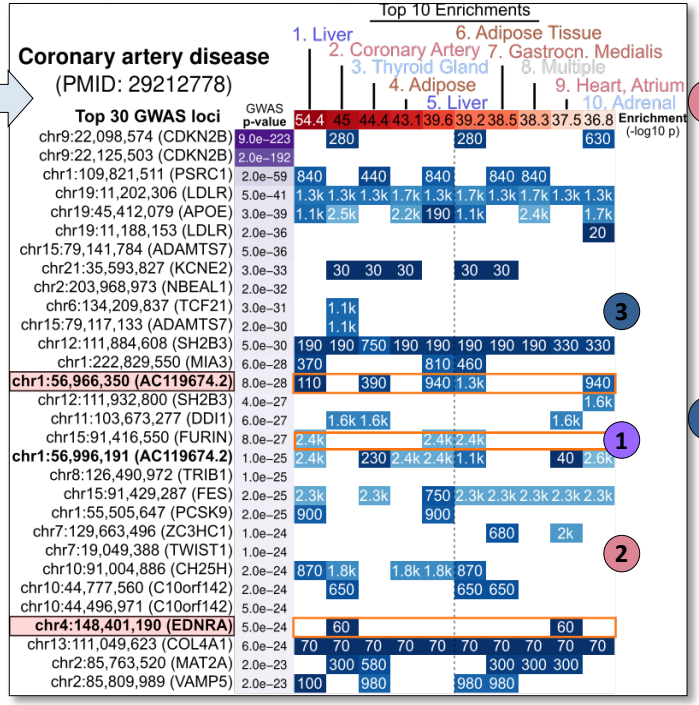


Example:
CAD
Coronary Artery Disease

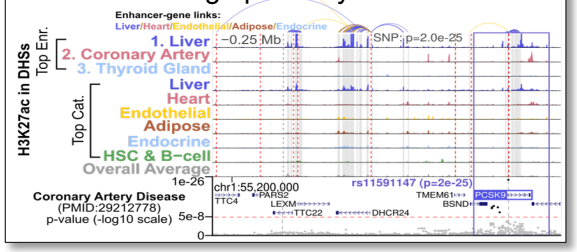


Epigenomic partitioning of complex traits into components

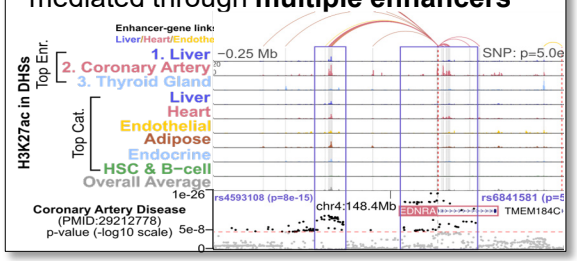




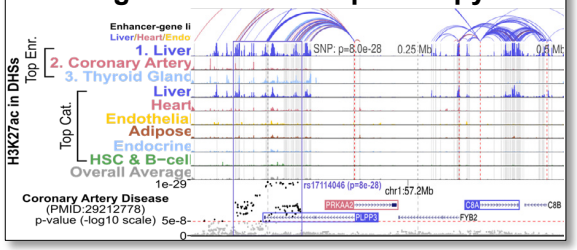
1 PCSK9: Liver-only mechanism, mediated through primarily one variant



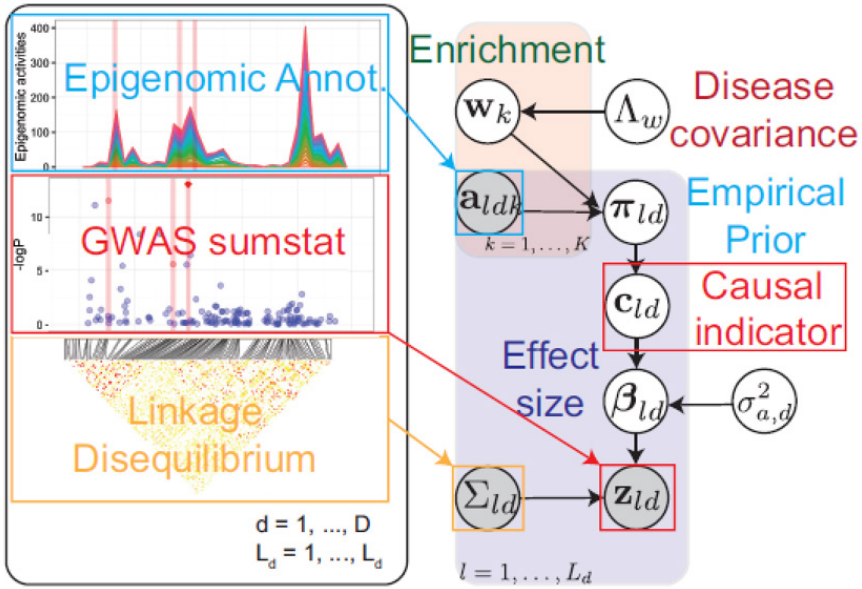
2 EDNRA Heart/vasculature-only, mediated through multiple enhancers



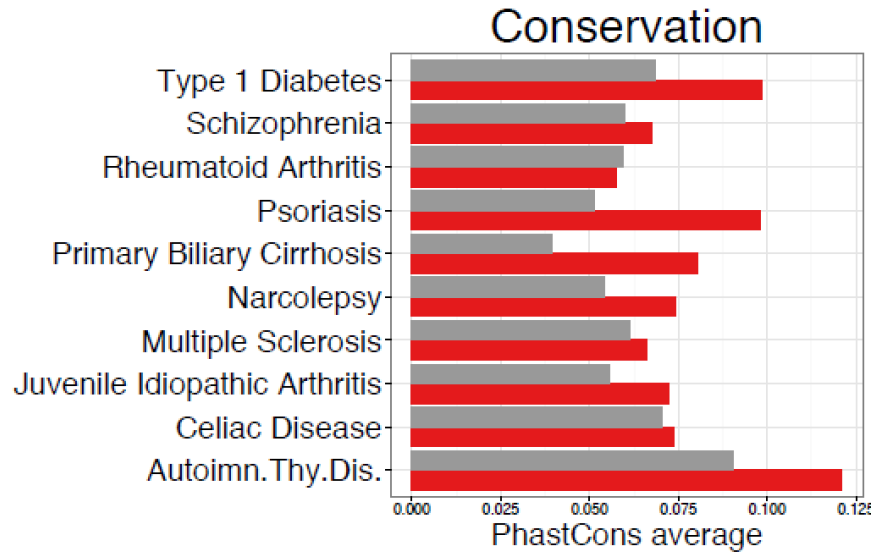
3 PLPP3: Both liver and coronary artery: multi-gene/multi-tissue pleiotropy



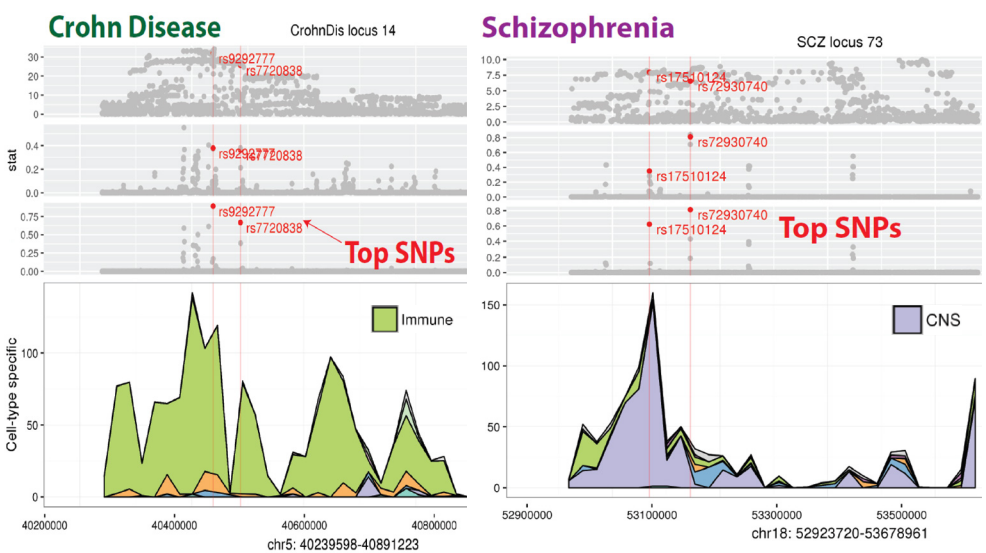
Bayesian fine-mapping: Predict causal variant and cell type



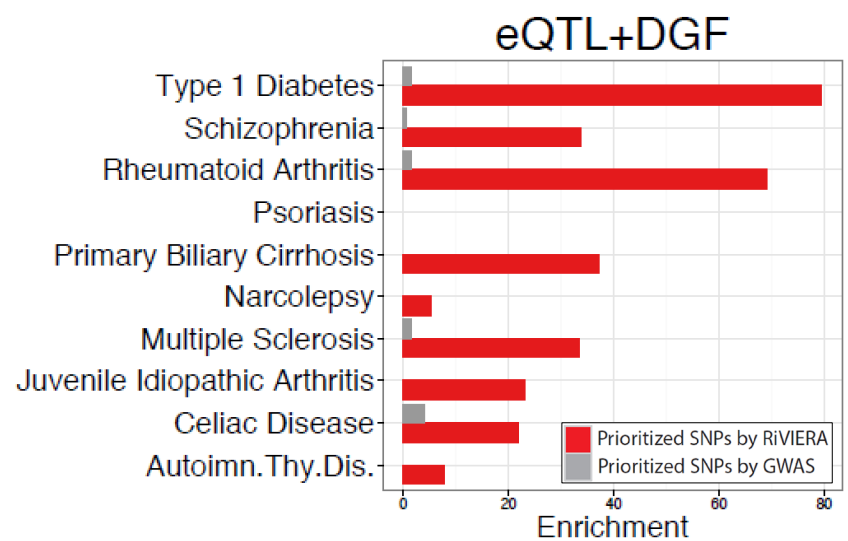
RiVIERA: multi-trait GWAS integration



Capture conserved elements

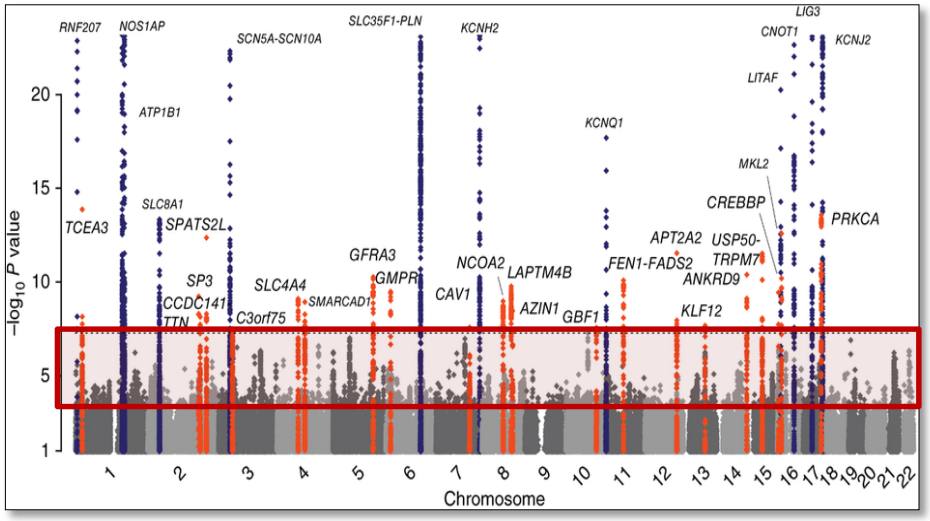


Predict causal variants and cell types



Capture eQTLs from GTEx

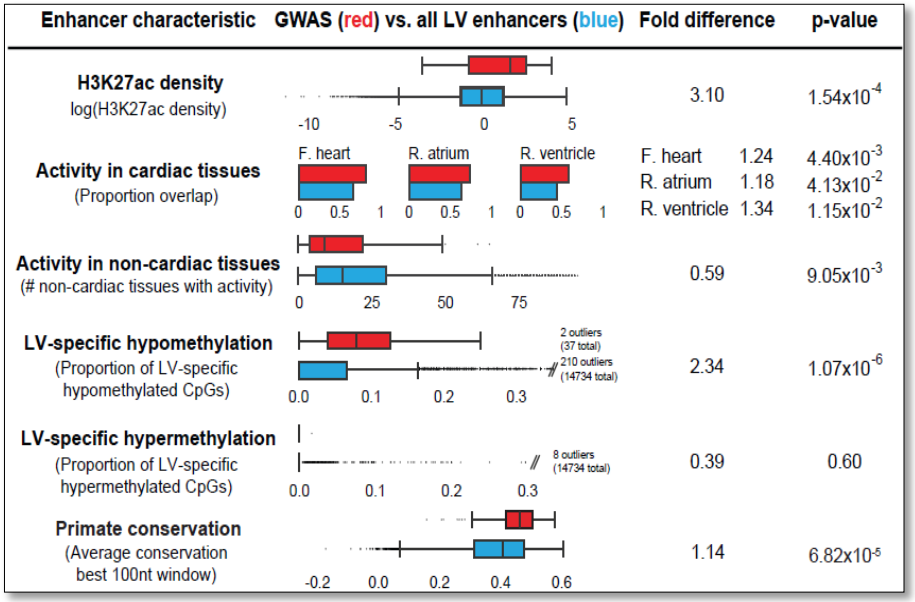
Combine GWAS+Epig to find new target genes/SNPs

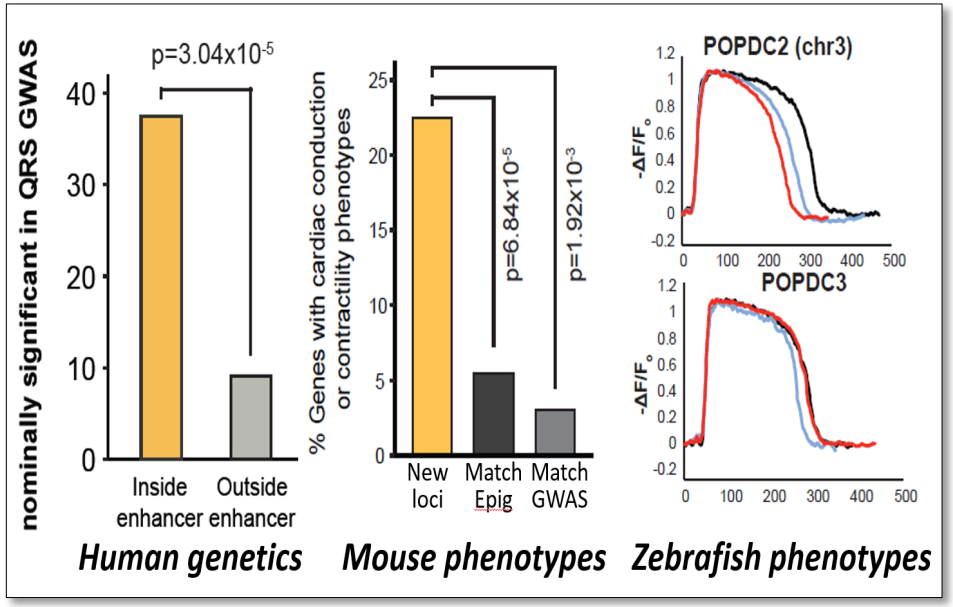


Lead SNP	p-value	Enhancer	1. Luciferase reporter	2. 4C-seq interactions
rs1886512	4.30×10^{-8}	chr13:74,520,000-74,520,400	0.015	No interactions
rs1044503	5.13×10^{-7}	chr14:102,965,400-102,972,000	4.70×10^{-9}	CINP, RCOR1
rs10030238	6.21×10^{-7}	chr4:141,807,800-141,809,600	1.35×10^{-14}	RNF150
		chr4:141,900,800-141,908,000	-	RNF150
rs6565060	1.52×10^{-5}	chr16:82,746,400-82,750,800	5.00×10^{-3}	No interactions
rs3772570	1.73×10^{-5}	chr3:148,733,200-148,738,600	0.67	-
rs3734637	2.23×10^{-5}	chr6:126,081,200-126,081,800	1.06×10^{-4}	HDDC2
rs1743292	6.48×10^{-5}	chr6:105,706,600-105,710,200	3.20×10^{-4}	BVES, POPDC3
		chr6:105,720,200-105,723,000	-	BVES, POPDC3
rs11263841	6.87×10^{-5}	chr1:35,307,600-35,312,200	0.22	GJA4, DLGAP3
rs11119843	7.14×10^{-5}	chr1:212,247,600-212,248,600	0.031	-
rs6750499	7.37×10^{-5}	chr2:11,559,600-11,563,000 (split into two 2kb fragments)	0.54	ROCK2
			3.26×10^{-7}	
rs17779853	7.73×10^{-5}	chr17:30,063,800-30,066,800	4.33×10^{-3}	No interactions

Prioritize sub-threshold loci ($<10^{-4}$)

Validate new enhancers: allelic activity, enh-prom looping





Machine learning predictive features

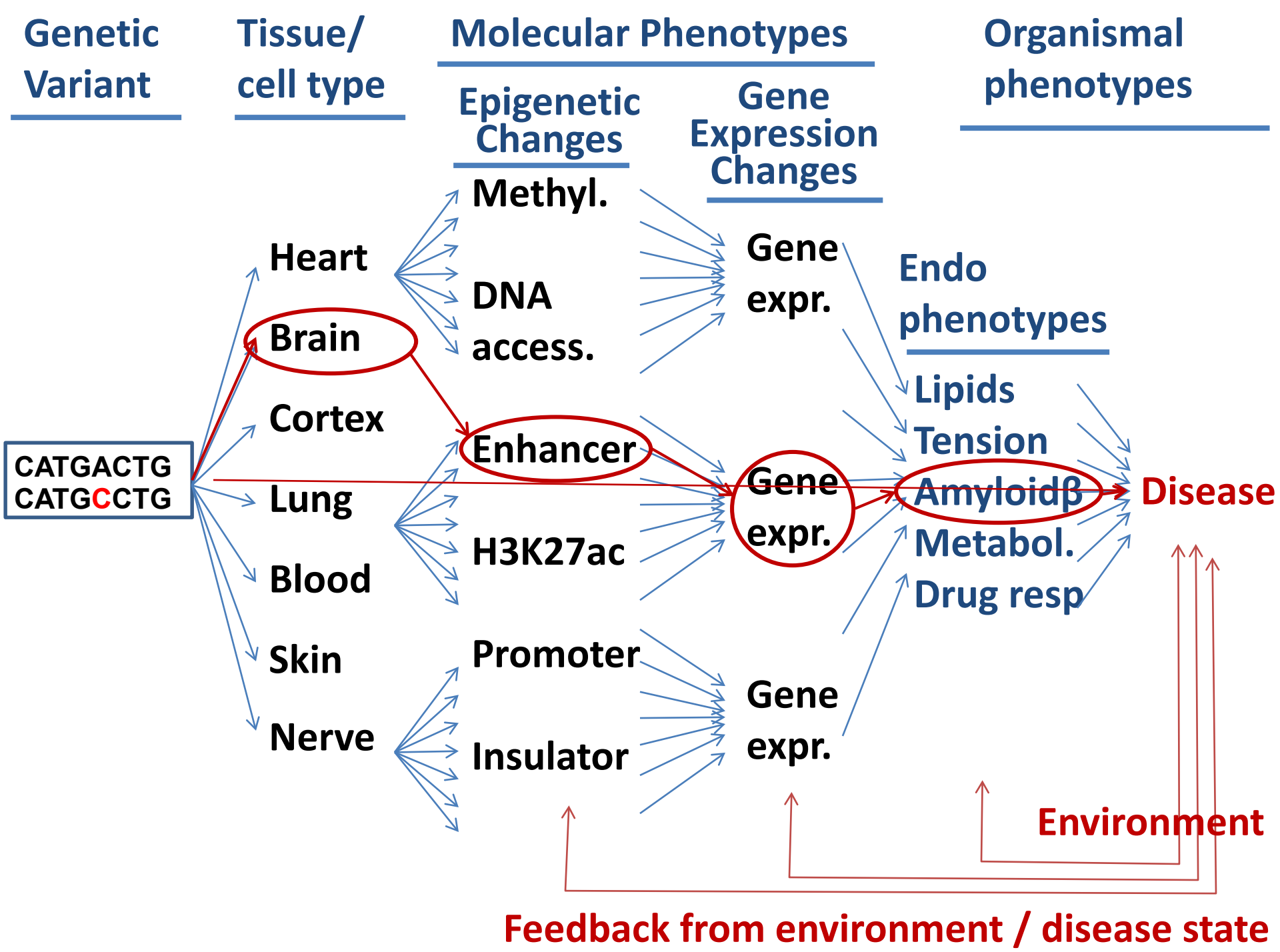
Validate new genes in hum/mou/zb

GWAS mechanism: epigenomics, eQTLs, Causality

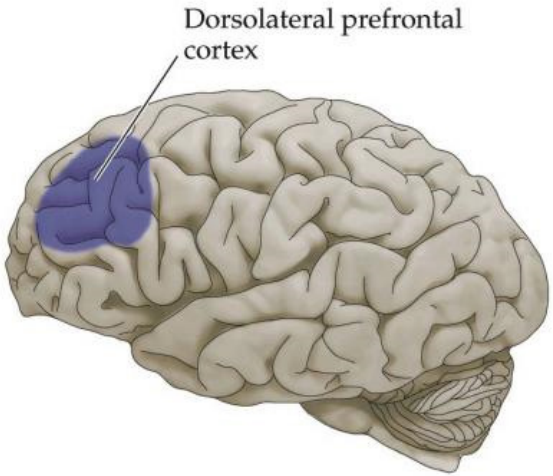
1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

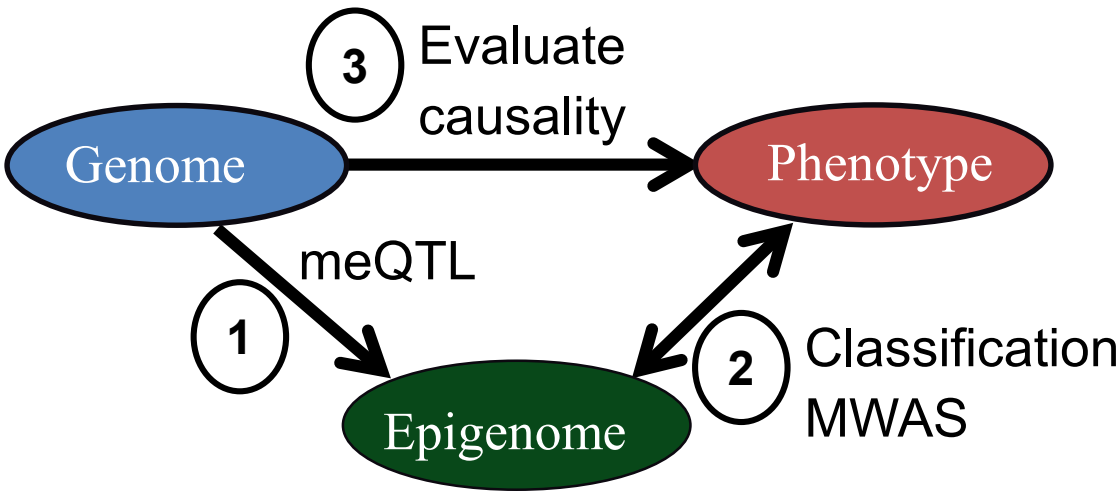
3. eQTLs and Mediation Analysis

Intermediate molecular phenotypes to disease


Genetic Variant
Tissue/ cell type
Molecular Phenotypes
Epigenetic Changes
Gene Expression Changes
Organismal phenotypes
Methyl.
DNA access.
Enhancer
H3K27ac
Promoter
Insulator
Heart
Brain
Cortex
Lung
Blood
Skin
Nerve
CATGACTG
CATGCCTG
Gene expr.
Gene expr.
Gene expr.
Endo phenotypes
Lipids
Tension
Amyloidβ
Metabol.
Drug resp
Disease
Environment
Feedback from environment / disease state

Methylation in 750 Alzheimer patients/controls



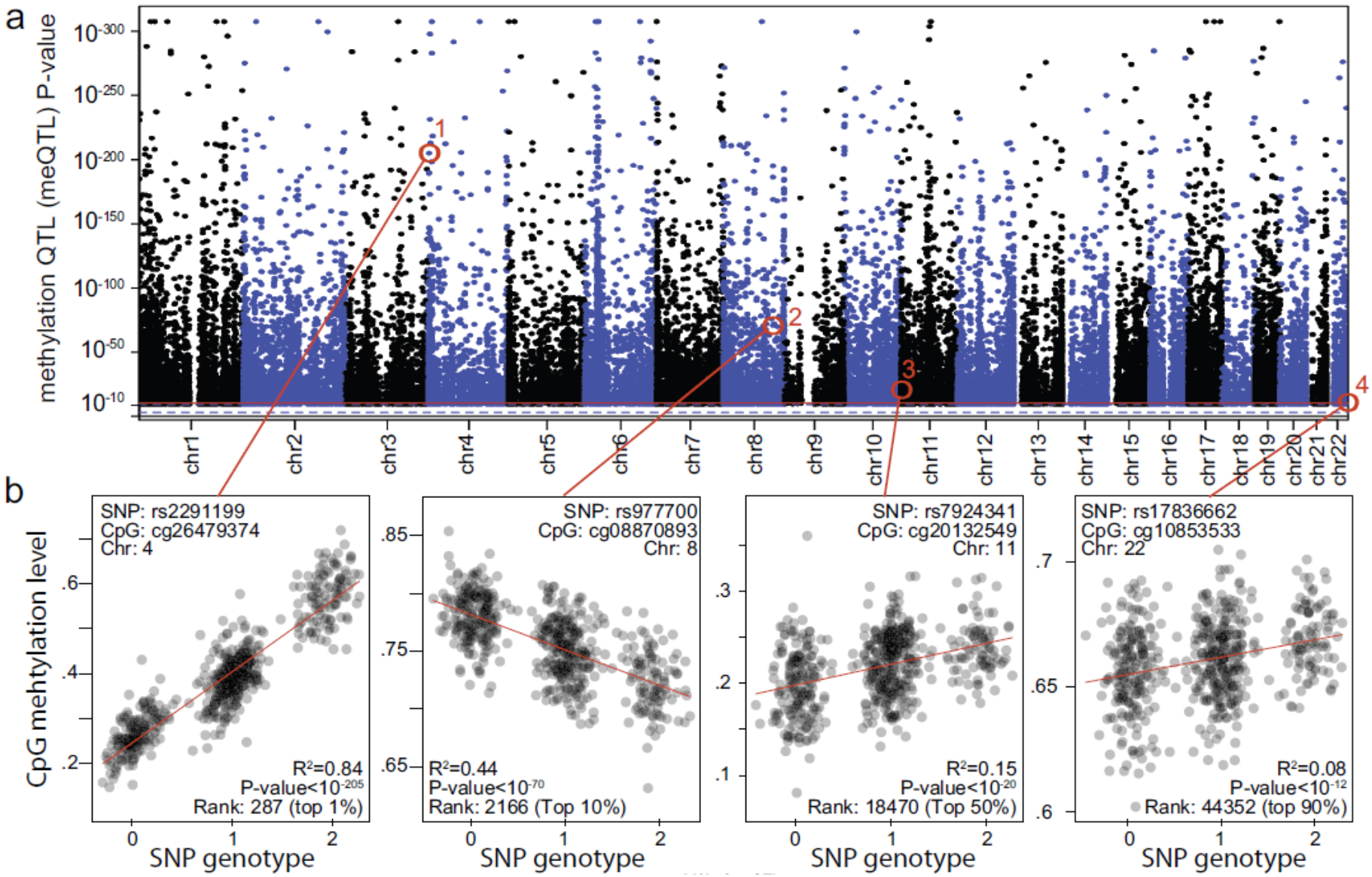


Methylation variation in 723 individuals

Relate to genotype and AD variation

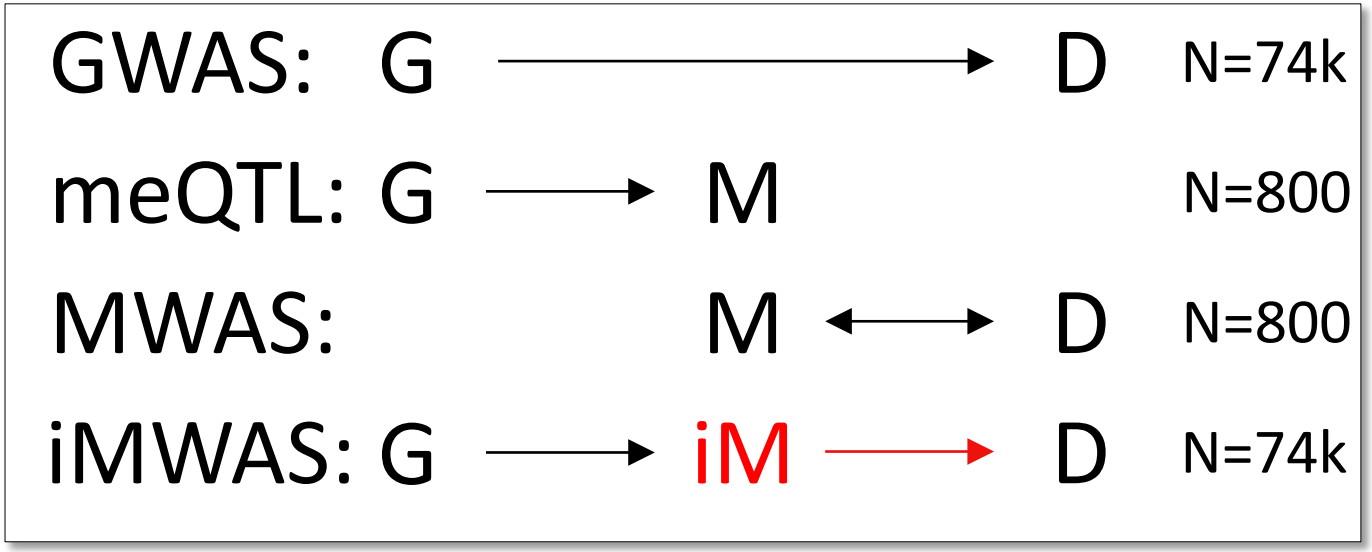
- ***ROS-MAP cohort (RUSH: David Bennett, HMS: Phil De Jager)***
  - ***Patients followed for 10+ years with cognitive evaluations***
  - ***Brain samples donated post-mortem methylation/genotype***
- ***Seek predictive features: SNPs, QTLs, mQTLs, regulation***

50,000 significant meQTLs after Bonferroni



- Strong effects across entire range of discovery values

Imputed MWAS: increased power, genetic component



Learn G→D directly (complex phenotype)

Learn G→M (simpler phenotype)

M⇔D (no causality)

Apply G→M to get iM
iM→D (causality)

<u>Key Idea:</u>

- Learn G→M model (ROSMAP n=800) Fewer indiv. Simpler phenotype
- Impute methylation iM for GWAS cohort (n=74k)
- iMWAS between <u>genotype-driven</u> M and AD phenotype (n=47k)

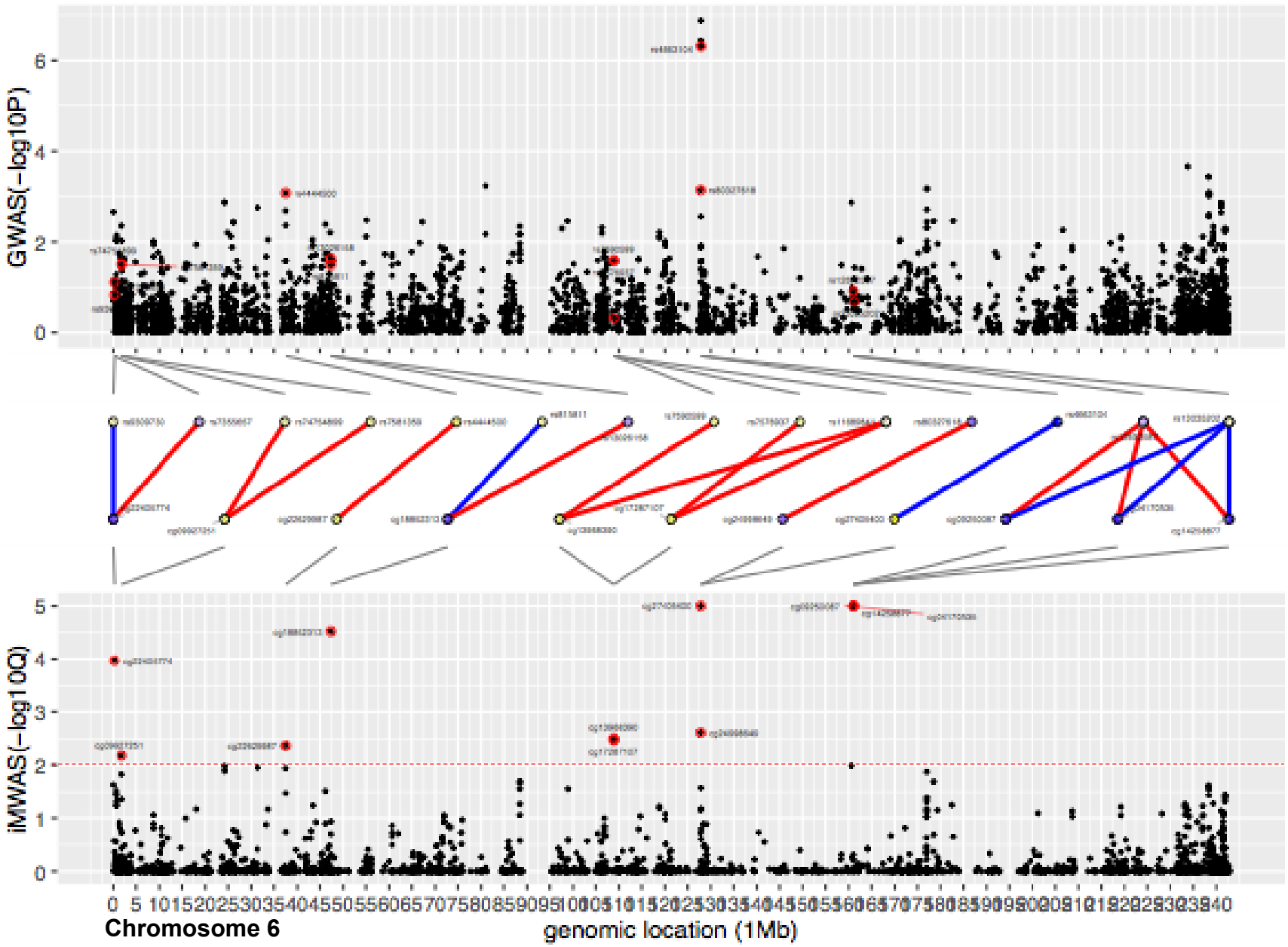
<u>Advantage:</u>

- Much larger GWAS cohorts (>>MWAS): increased power
- Genetic component of methyl. variation

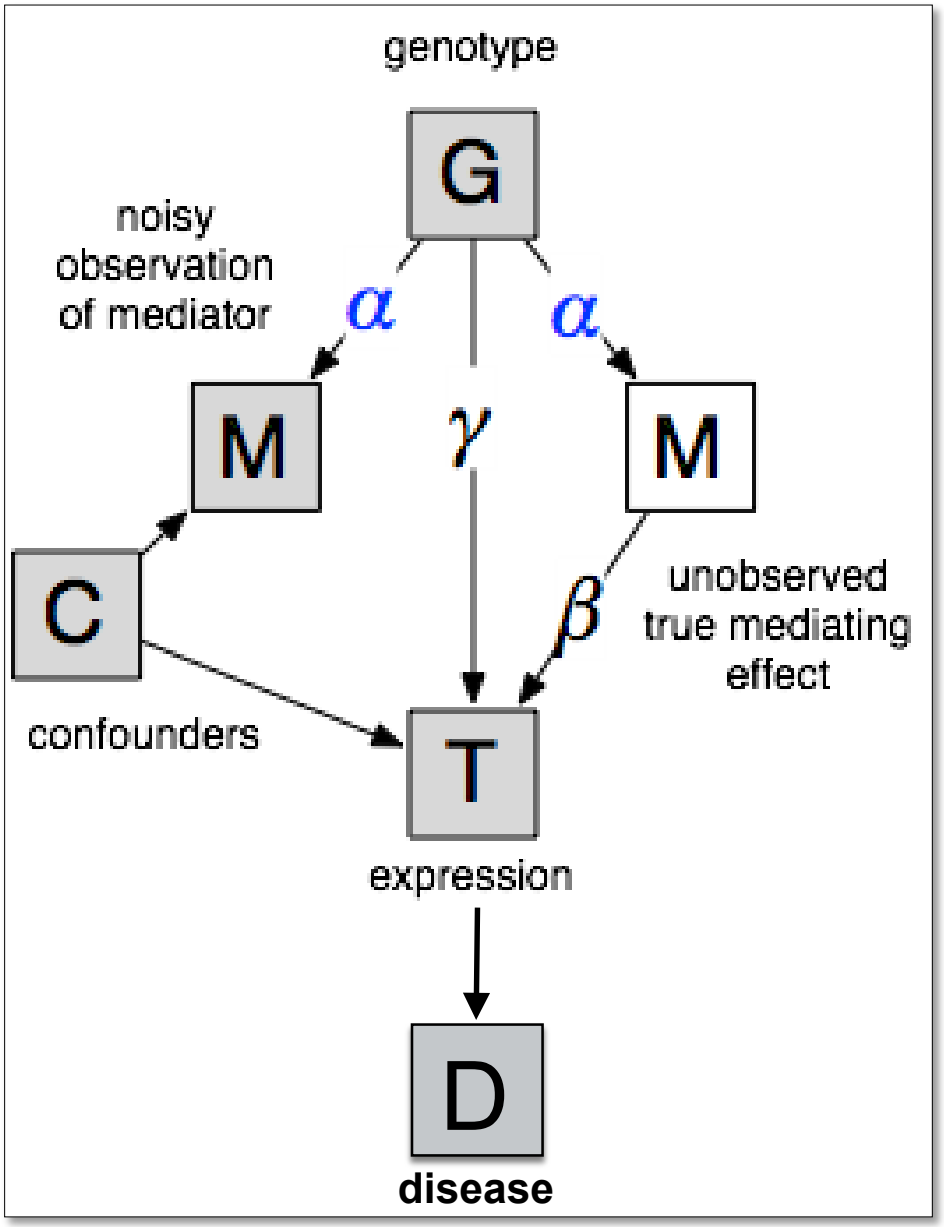
<u>Logistical challenge:</u>

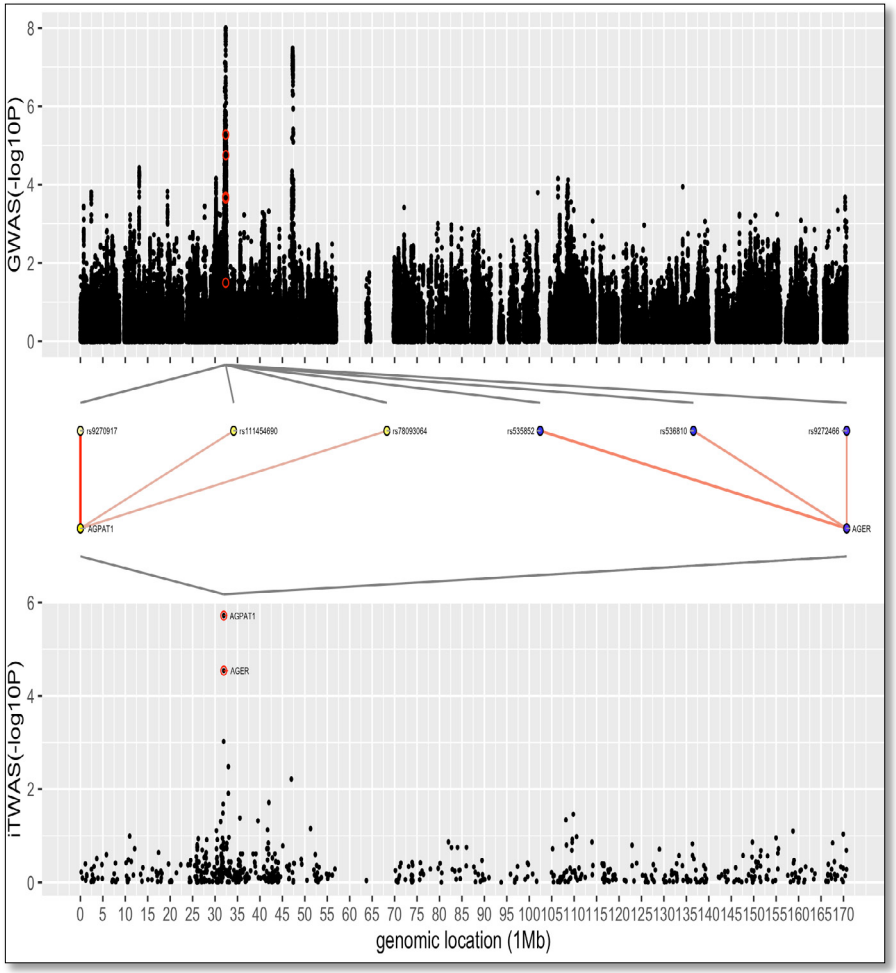
- Summary stats, not full genotypes ➔ Linear model, impute stats direct

iMWAS results: new loci, multiple contributing SNPs



iMTWAS: Imputation across multiple intermediate variables





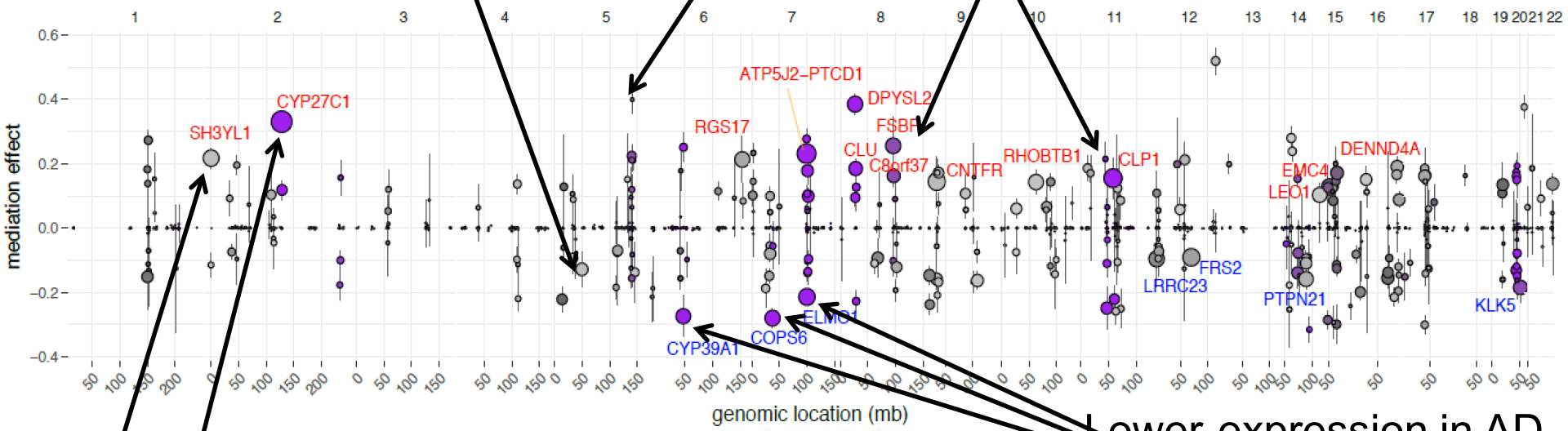
Model multiple mediator variables
SNP → Methylation → Expression → Disease
Predict new loci, increased power
Predict regulatory regions & target genes

CaMMEL: 206 significant mediating genes in AD

Small expression change (short),
large variance explained (big circle)

Large expression change (tall),
little variance explained (small circle)

Higher-expression in AD
(risk increasing)

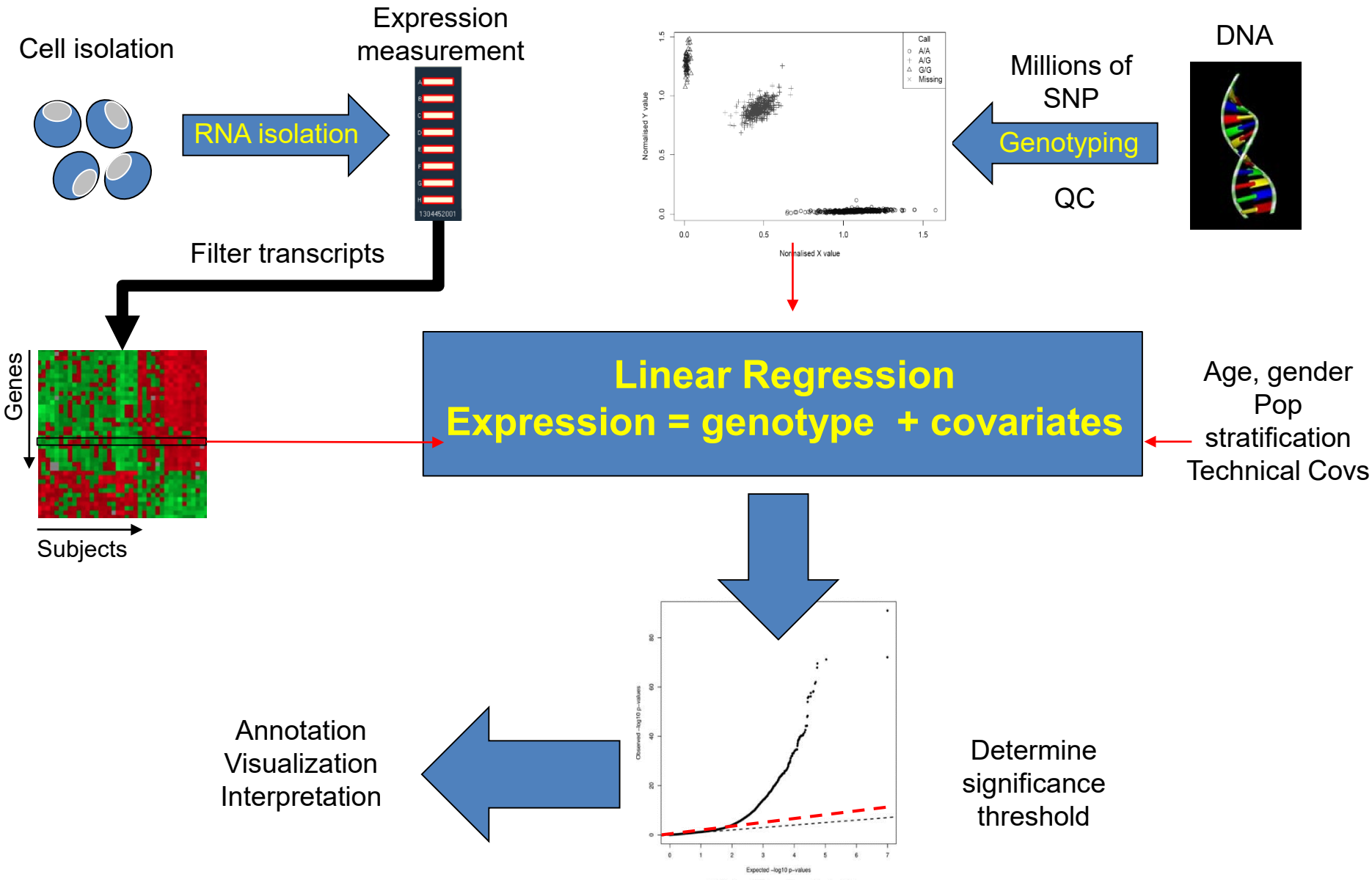


Lower-expression in AD
(protective)

Genome-wide significant locus (purple)

Sub-threshold locus (grey)

The nuts and bolts of an eQTL study



Expanded eQTL models

$$Y_{ij} = \alpha + \beta_{ijs}\text{genotype} + \varepsilon$$

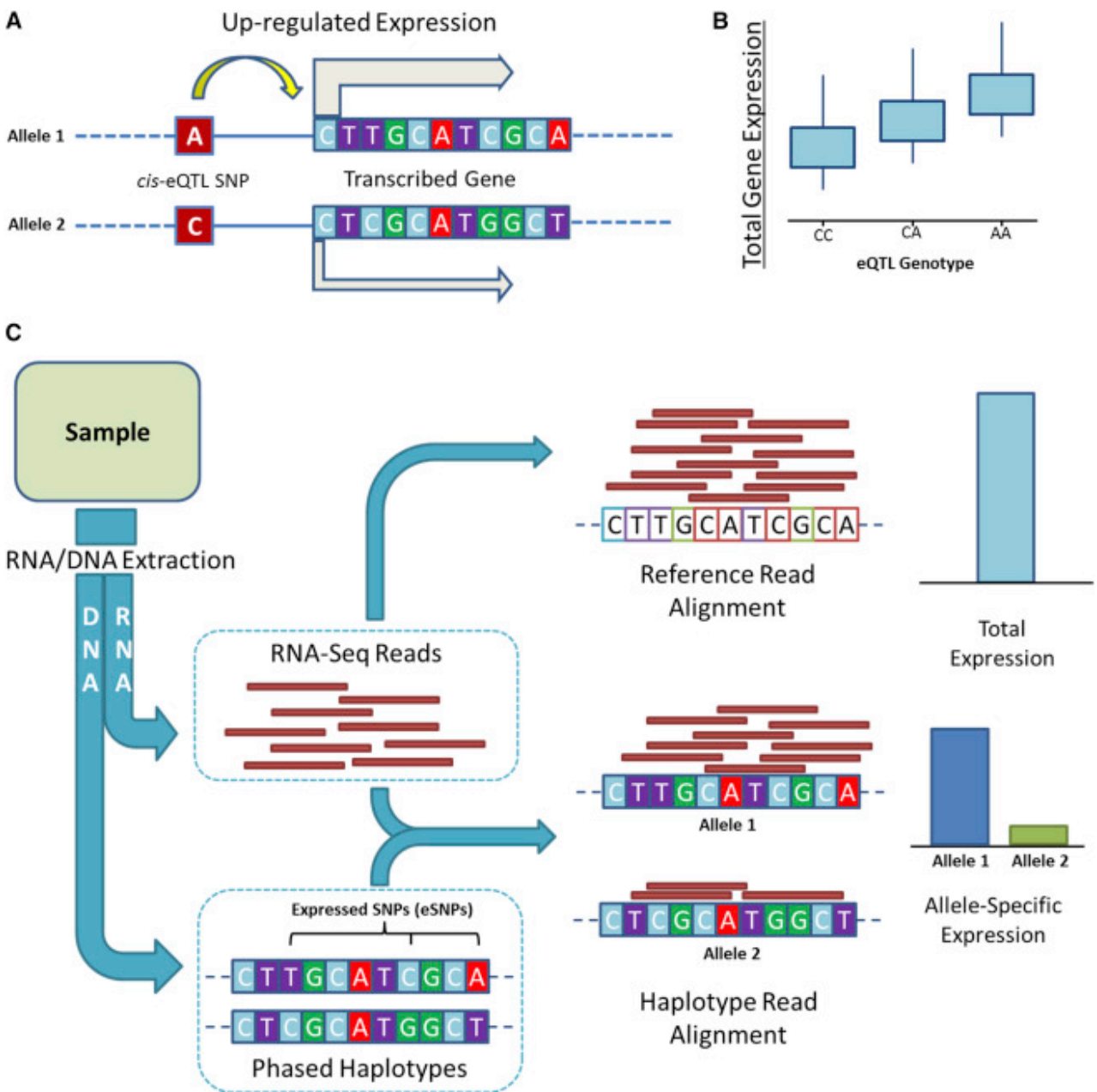
$$Y_{ij} = \alpha + \beta 1_{ijs}\text{genotype} + \beta 2_{i}\text{gender} + \beta 3_{i}\text{age} +$$

$$\beta 4_{i}\text{gPC1} + \beta 5_{i}\text{gPC2} + \beta 6_{i}\text{gPC3} + \beta 7_{i}\text{gPC4} +$$ Genotype PCs

$$\beta 8_{i}\text{ePC1} + \beta 9_{i}\text{ePC2} + \beta 10_{i}\text{ePC3} + \beta 11_{i}\text{ePC4} + \beta 12_{i}\text{ePC5} + \beta 13_{i}\text{ePC6} + \beta 14_{i}\text{ePC7}$$ Expression PCs

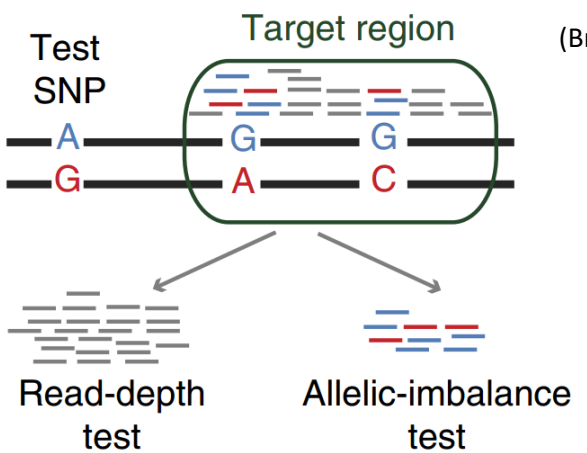
$$+ \varepsilon$$

Allelic analysis complements eQTLs



Distinguish reads **within the same heterozygous individual**

Combined Haplotype Test



(Bryce van de Geijn, et.al *Nature Method* 2015)

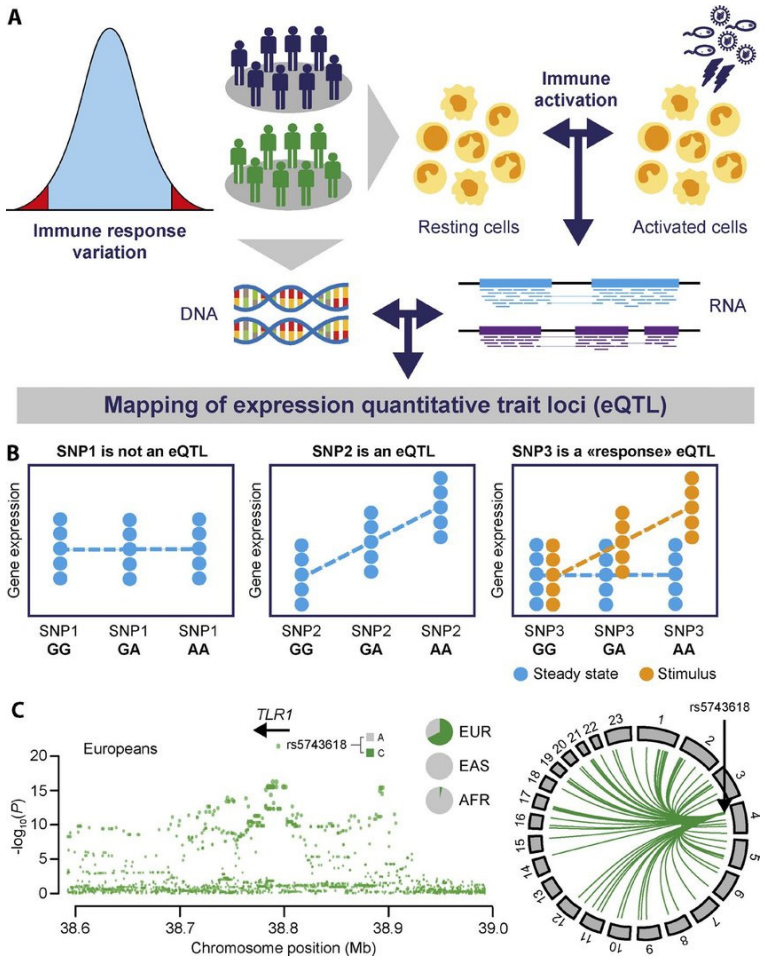
Maximize likelihood of two observed components:

Total Read-depth Allelic imbalance

$$\mathrm{L}\left(\alpha_h, \beta_h, \phi_j \,|D\right) = \prod_i \left[\Pr_{\mathrm{BNB}}\left(X = x_{ij} \,|\lambda_{hi}, \Omega_i, \phi_j\right) \prod_k \Pr_{\mathrm{BB-mix}}\left(Y = y_{ik} \,|p_h, n_{ik}, \Upsilon_i\right) \right]$$

Beta-Negative-Binomial Beta-Binomial

“Response eQTLs”: Trait-conditional eQTLs



GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

4. Linear Mixed Models (LMMs)

for GWAS and for eQTL calling

Formal definition of a linear model

n individuals

p SNPs

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, X = \begin{pmatrix} X_{11} & \cdots & X_{1p} \\ X_{21} & \cdots & X_{2p} \\ \vdots & \cdots & \vdots \\ X_{n1} & \cdots & X_{np} \end{pmatrix}, \boldsymbol{\theta} = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_p \end{pmatrix}$$

In matrix notation, phenotype y as a factor of genetic information x

$$\mathbf{y} = X\boldsymbol{\theta} + \epsilon,\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I).$$

θ = effect size (can be itself sampled from a normal prior)

What are we missing in the previous multivariate model?

$$\mathbf{y} = X\boldsymbol{\theta} + \epsilon, \boxed{\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I).}$$

Assume IID individuals.
This may not be true.

$$\mathbf{y} = X\boldsymbol{\theta} + \boxed{\boldsymbol{u}} + \epsilon.$$

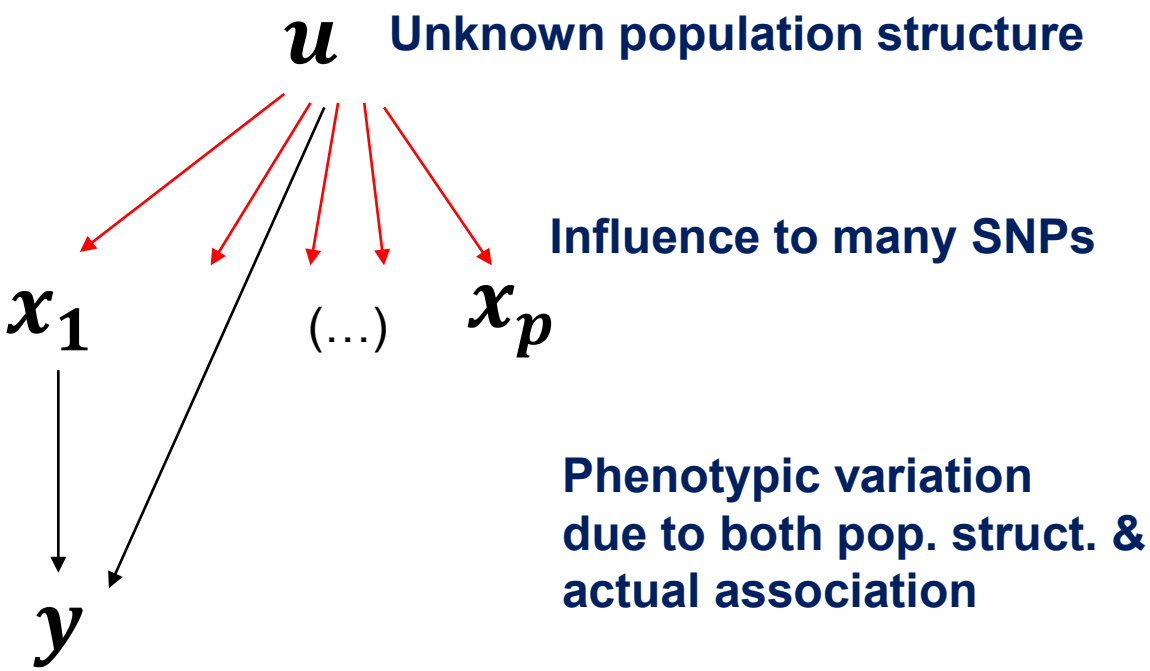
Add random effects to account for the unknown

$$\boxed{\boldsymbol{u}} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

We assume this random effect can be captured by Kinship covariance.

In GWAS problems, the most influential/spurious random effect stems from population structure.

Why do we need a random effect?



A Bayesian approach to account for the random effect u

Likelihood model:

$$\mathbf{y} = X\boldsymbol{\theta} + \boxed{\boldsymbol{u}} + \epsilon.$$

(Empirical) prior knowledge:

$$\boxed{\boldsymbol{u}} \sim \mathcal{N}(\mathbf{0}, \mathrm{K})$$

A Bayesian method ≈ Address/remove uncertainty by averaging out

$$p(\boldsymbol{y}|X\theta) = \int p(\boldsymbol{y}|X\theta, \boldsymbol{u})p(\boldsymbol{u})d\boldsymbol{u}$$

A Linear mixed effect model:

two components in covariance matrix

$$\mathbf{y} = X\boldsymbol{\theta} + \tilde{\epsilon} \quad \text{with} \quad \tilde{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boxed{\sigma^2 I} + \boxed{\tau^2 \mathrm{K}})$$

IID error **Kinship components**

Linear mixed models

$$p \sim N(0, h^2 G + (1 - h^2) I)$$
$$G = XX' / p$$

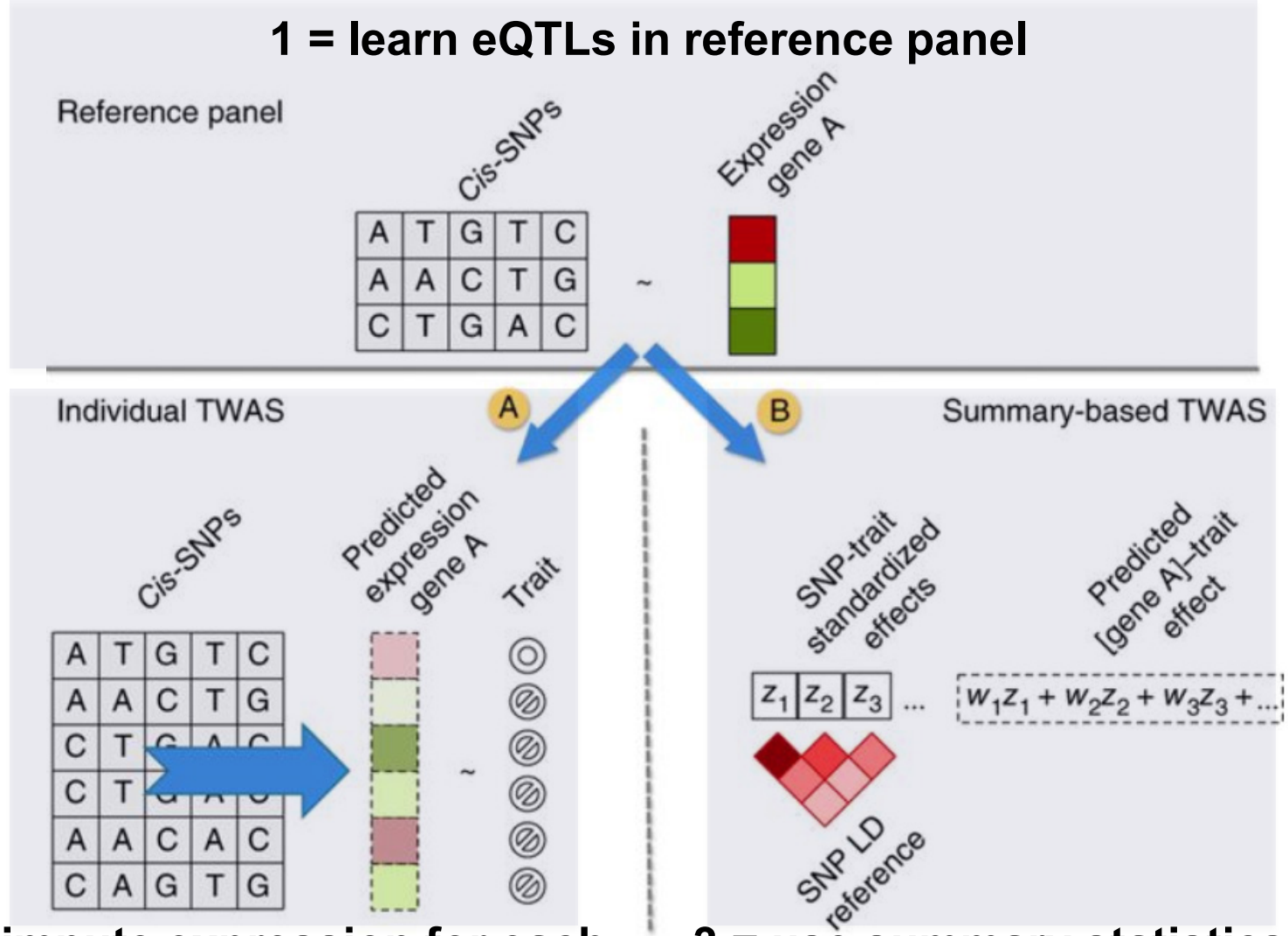
- Joint model of all SNPs explains more heritability (Yang 2010)
- Idea: under suitable assumptions, $V[a] = \Sigma\beta_j^2$
- Under the infinitesimal assumption $\beta_j \sim N(0, h^2/p)$, we can estimate V[a] without estimating individual β_j using residual maximum likelihood (REML)
- REML avoids using ML fit of parameters, instead uses transformed data so that nuisance parameters have no effect.
- In variance components analysis (random effects model), transformation focuses on differences, sum of variances
- **This works despite not knowing the causal variants**
- Example (height): ; $h^2_{GWAS} = 0.16$, $h^2 = 0.73$, $h^2_g = 0.5$

Linear mixed models

$$p \sim N(0, h^2 G - (1 - h^2) I)$$
$$G = XX' / p$$
$$E[p_i p_j] = h^2 G_{ij}$$

- We can generalize Haseman-Elston regression to estimate heritability for unrelated individuals using LMM
- Intuition: genetic relationship matrix G captures identity by state in unrelated individuals
- This is again the probability of sharing the same allele at the causal variants
- This is called **PCGC regression** (Golan 2015) (phenotype correlation – genotype correlation regression

Imputation-based association

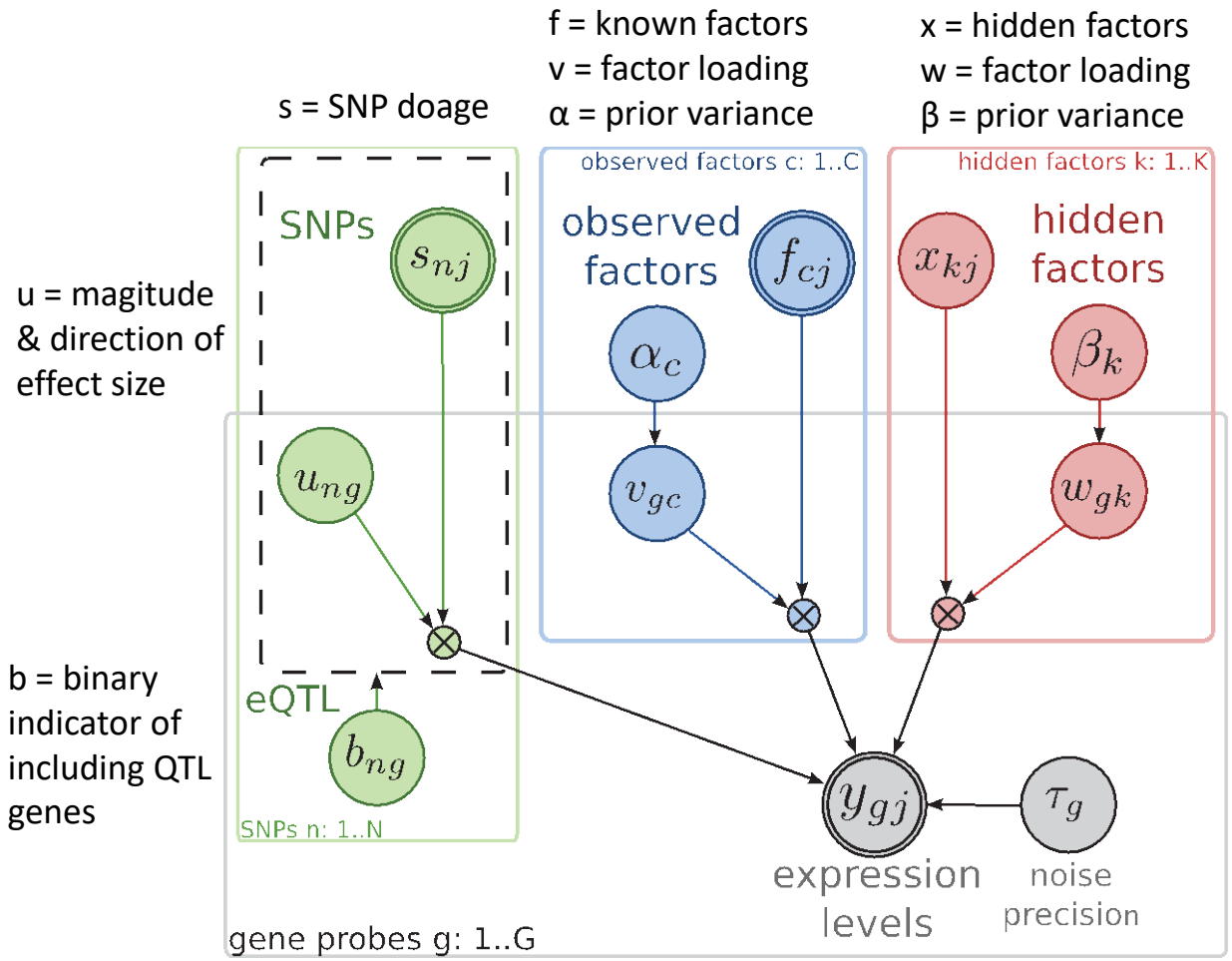


2 = impute expression for each person in a genotyped cohort

3 = use summary statistics to get to associations directly

Gusev et al. “Integrative approaches for large-scale transcriptome-wide association studies” 2016 *Nature Genetics*

Bayesian linear regression for eQTL modeling



Stegle *et al.* PLoS Genetics (2010)

Bayesian extension to ordinary regression models

1. Spike-slab prior to select relevant variables
2. Random effect models
3. Bayesian sparse linear mixed effect model
4. Fine mapping causal variants in LD correlation

Extension 1: spike-slab prior on θ

$p(\theta|z=1) \sim N(0, 1/\tau)$ ← Fat Gaussian for true effects (slab; magnitude and direction)

$p(\theta|z=0) = \delta(\theta)$ ← Completely set to zero if not selected

$z = 1 \sim \text{Bernoulli}(\pi)$ ← π determines prior prob. of including variables (usually < .1; spike; prescribed or optimized)

$p(\theta) \sim \exp(-\lambda|\theta|)$

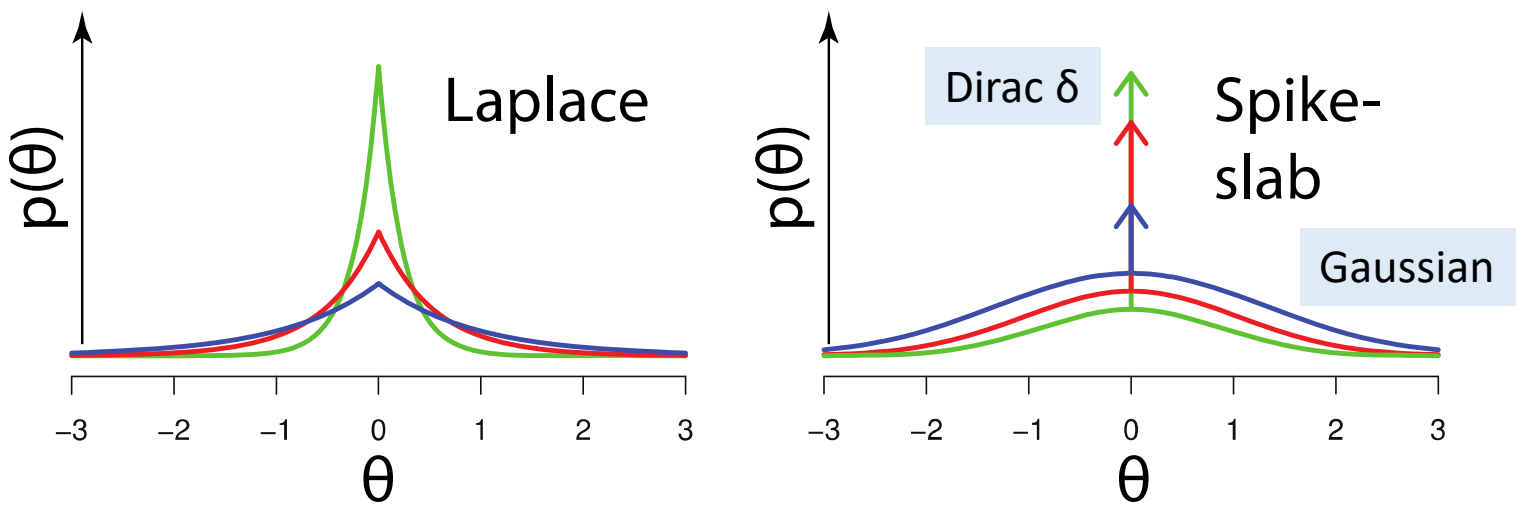
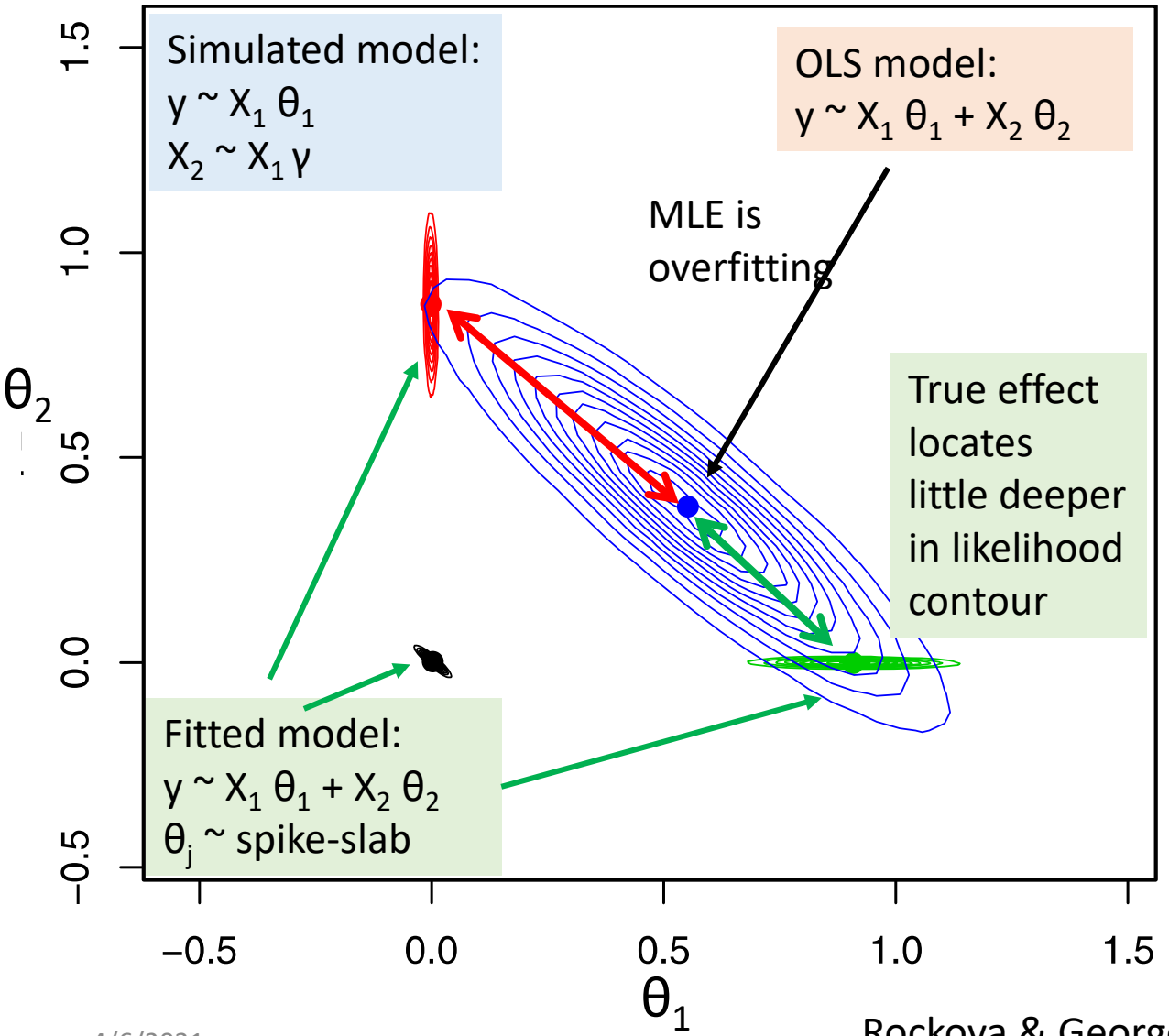
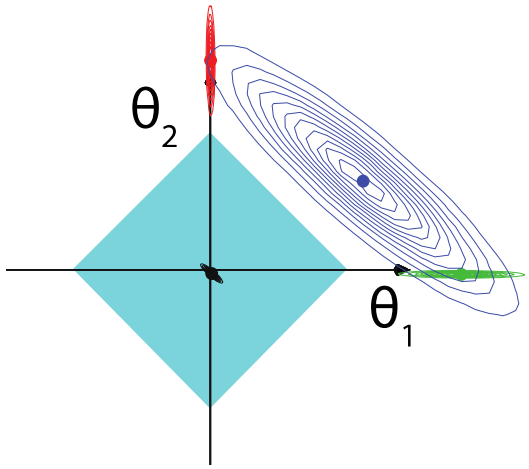


Figure: Hernandez-Lobato (2014)

Spike-slab prior model effectively avoid colinearity



Can L1-regularized one handle this?



If correlation between $X_1 \sim X_2$ is strong, probably not … (best solution within the box is still non-zero for both vars).

Rockova & George, *Metron* (2014)

Ext 2: random-effect for pop. stratification

Additive effect of random vector u (n × 1):

$$\mathbf{y} = X\boldsymbol{\theta} + \boxed{\mathbf{u}} + \boldsymbol{\epsilon}$$

The random effect captures population structure K (kinship matrix):

$$\boxed{\mathbf{u}} \sim \mathcal{N}(0, \tau^2 \boxed{K})$$

n × n covar. (~PCs)

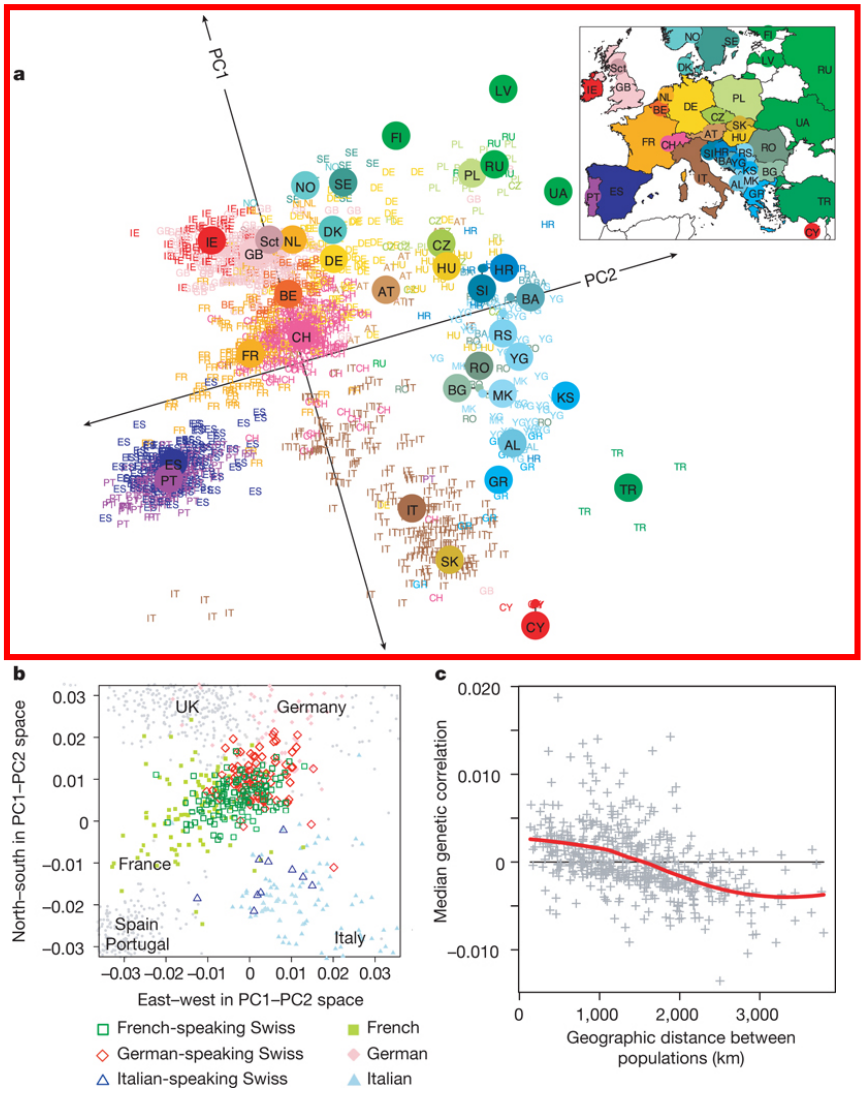
Integrate out uncertain random effect u:

$$\int \mathbf{p}(\mathbf{y}|X, \boldsymbol{\theta}, \mathbf{u})\mathbf{p}(\mathbf{u}|\boldsymbol{\tau}, \mathbf{K})\mathbf{du}$$
$$= \mathcal{N}(\mathbf{y}|\mathrm{X}\theta, \tau^2 K + \sigma^2 I)$$

population structure

random noise

Linear Gaussian model with two variance components.



J Novembre *et al. Nature* **000**, 1-4 (2008)

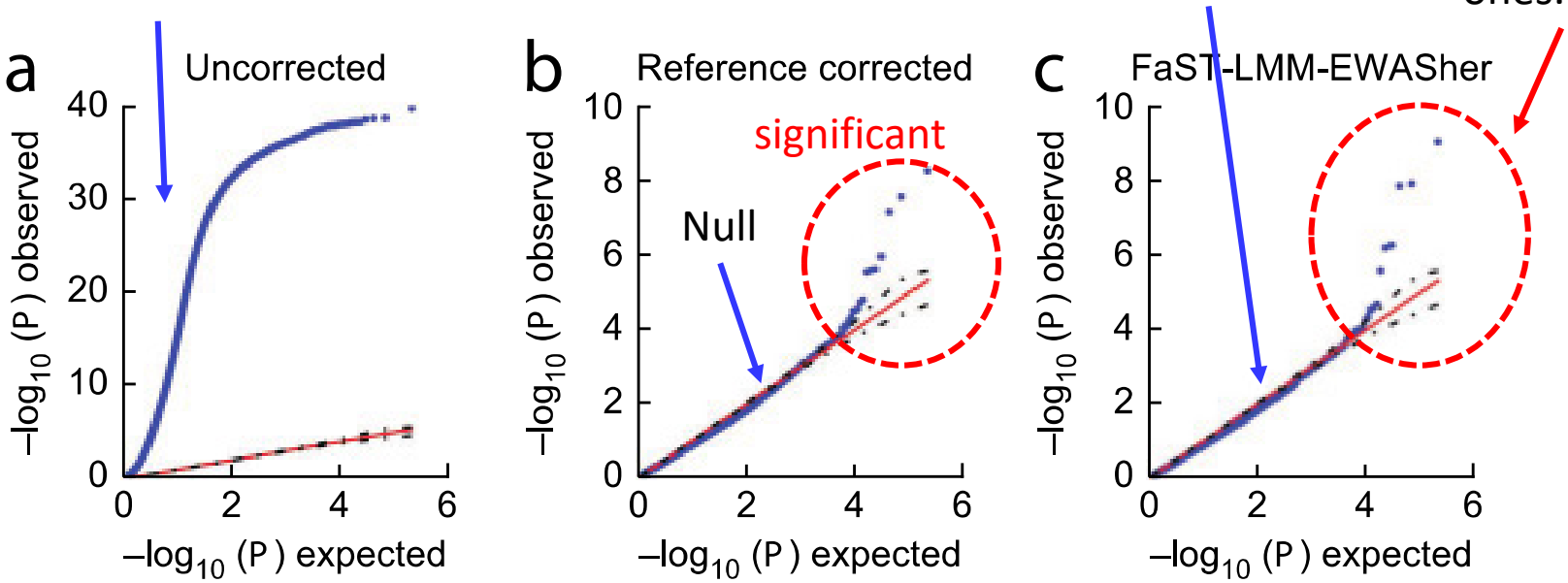
Extension 2: random effect model

Inflated statistics due to unknown population structure (almost all loci are significant)

Adjusted GWAS qq-plot with correct structure

Linear mixed-effect calibrated the null distrib.

LMM can correctly capture significant ones.



Zou .. Listergarten, *Nat. Methods* (2014)

Extension 3: Bayesian sparse linear mixed effect model

Random effect

$$\mathbf{y} = X\boldsymbol{\theta} + \mathbf{u} + \boldsymbol{\epsilon},$$

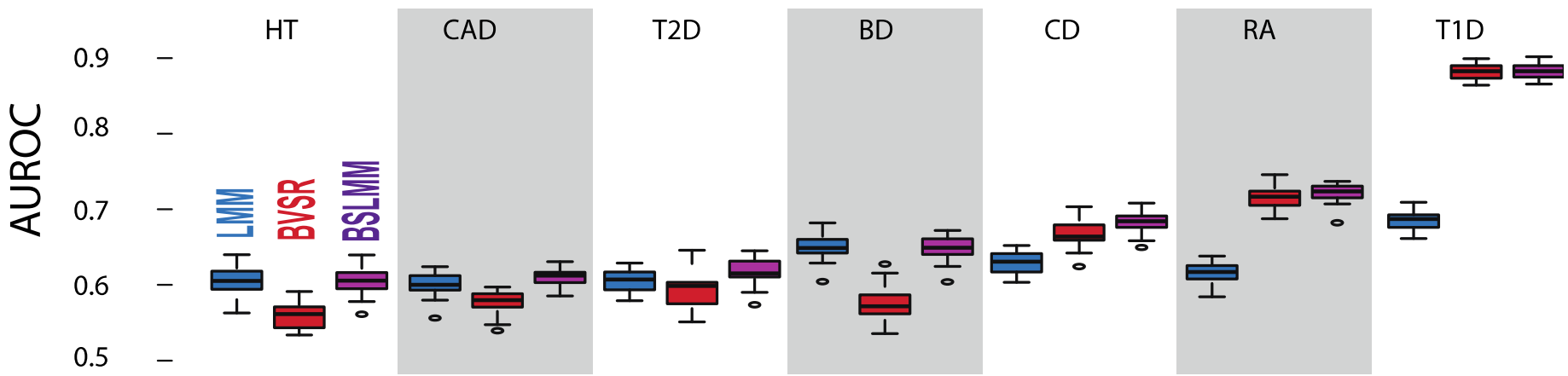
$$\mathbf{u} \sim \mathcal{N}(0, K),$$

A sort of spike-slab (two mixture model)

$$\theta_j \sim \pi\mathcal{N}(0, \tau_1^2) + (1-\pi)\mathcal{N}(0, \tau_2^2)$$

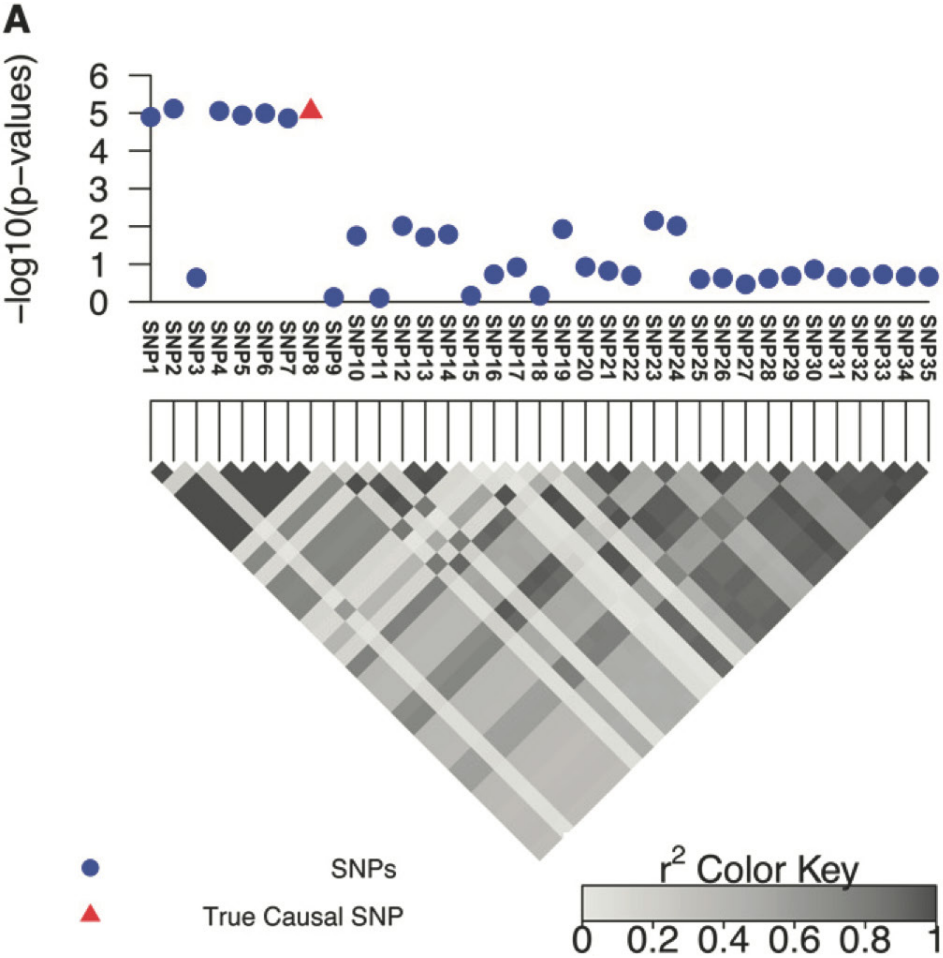
causal effect

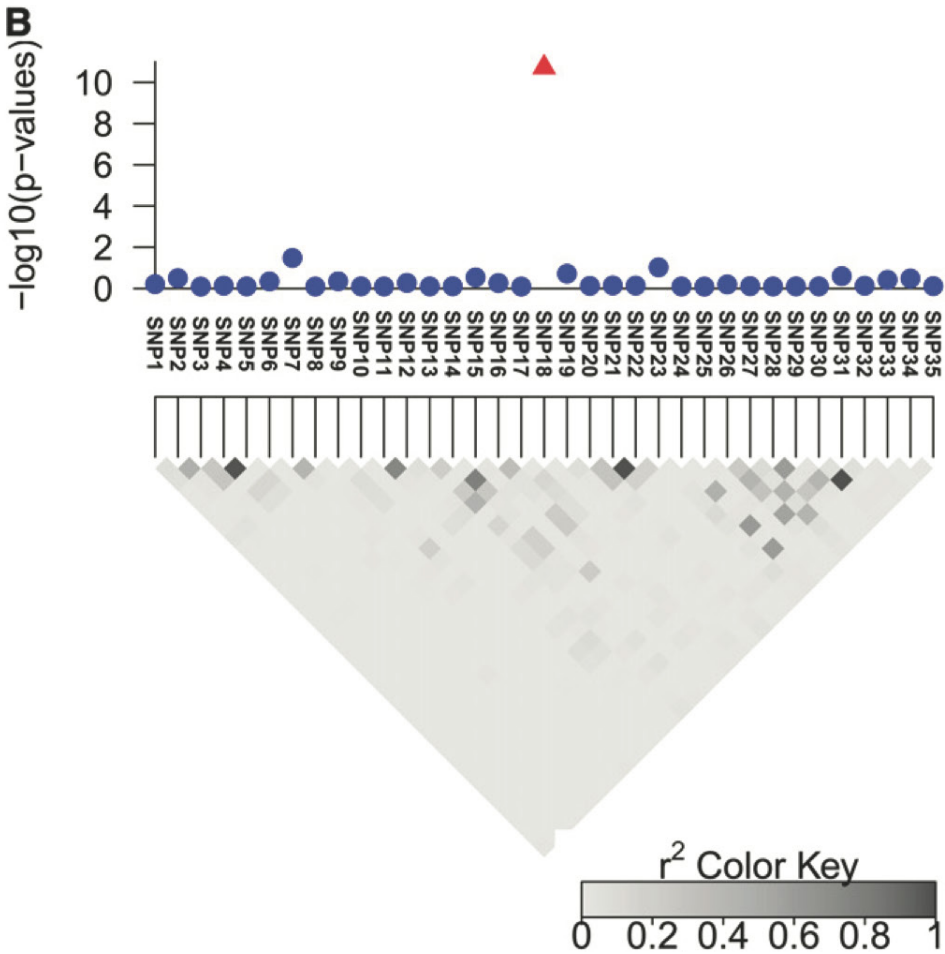
infinitesimal background effect



Zhou, Carbonetto, Stephens, *PLoS Gen.* (2013)

Extension 4: Fine-mapping causal variants





Hormozdiari *et al.* (2014)

Extension 4: Fine-mapping under the hood

unknown genotype

unkonwn phenotype y vector

summary z-score obs.

$$\mathbf{z} \approx X^\top \mathbf{y}/\sqrt{n}\sigma$$

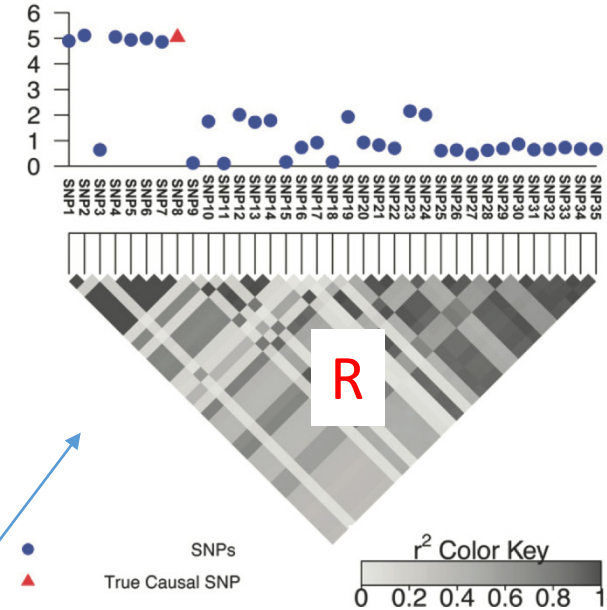
We assume phenotype vector were generated by

$$\mathbf{y} \sim \mathcal{N}(X\boldsymbol{\theta}, \sigma^2 I).$$

Therefore $p \times 1$ vector follows

$$\mathbf{z} \sim \mathcal{N}\left(\frac{X^\top X\boldsymbol{\theta}}{\sqrt{n}\sigma}, \frac{X^\top X}{n}\right) \approx \mathcal{N}(\lambda R\boldsymbol{\theta}, R).$$

where LD matrix $R = n^{-1}X^\top X$ and $\lambda = (n\sigma^2)^{-1/2}$ absorbs all scaling factors.



(a) Considering potential colinearity embedded in the R matrix, θ desperately needs spike-slab prior.

(b) For computational efficiency, previously developed algorithms restrict number of causal variants (e.g., at most 3).

Hormozdiari *et al.* (2014)

Bayesian inference algorithms

	Exact inference	Markov Chain Monte Carlo	Variational Bayes
Accuracy	correct	approximate, stochastic	approximate, deterministic
Convergence	sure	Global optima at equilibrium	Local optima in finite time
Flexibility	very limited	high	high
Examples	HMM's forward-backward, Dynamic programming	Importance sampling, Metropolis-Hastings, Gibbs, Hamiltonian MC, Elliptical slice sampling	Laplace, Mean-field approx., Belief propagation, Expectation propagation

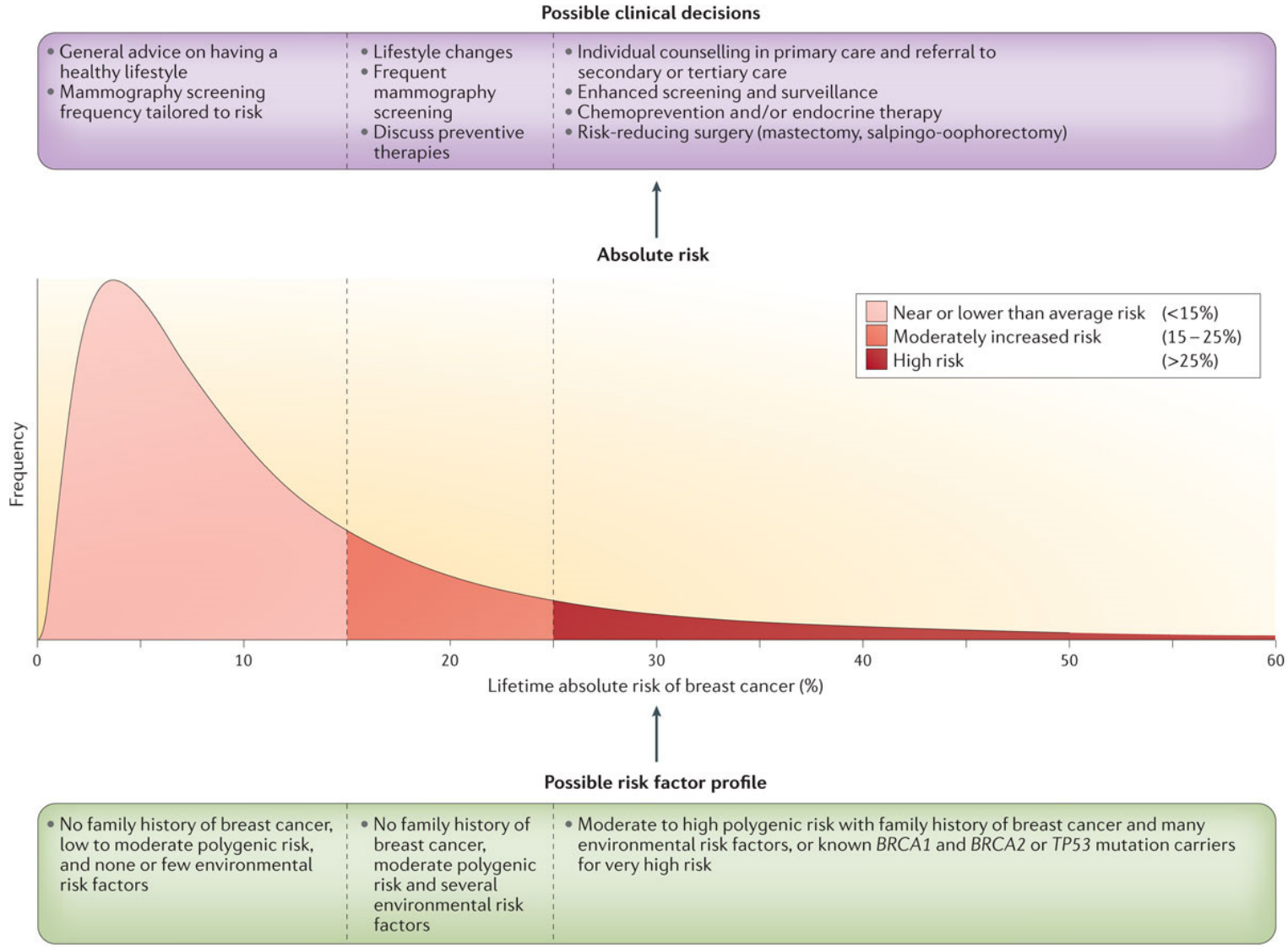
GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

5. Polygenic Risk Scores (PRS):

Summing over all variants (and more)

Estimate absolute risk combining genetic and environmental risk factors



Chatterjee *et al.* Nature Reviews Genetics (2016)

Nature Reviews | Genetics

How do we estimate polygenic risk score?

Univariate GWAS statistics teach us:

β_j = log(odds ratio of SNP j)

g_j = genotype (dosage)

Predict overall risk by combining many, many variants!

$$PRS = \Sigma_{j \in \{SNPs\}} \beta_j g_j$$

Can we just combine all the SNPs? Why not?

- Is correlation between g_1 and g_2 zero?
- Can we trust the estimate β of all the SNPs?
- Can we just select GWAS significant SNPs?

A common practice of PRS estimation

Univariate GWAS statistics:

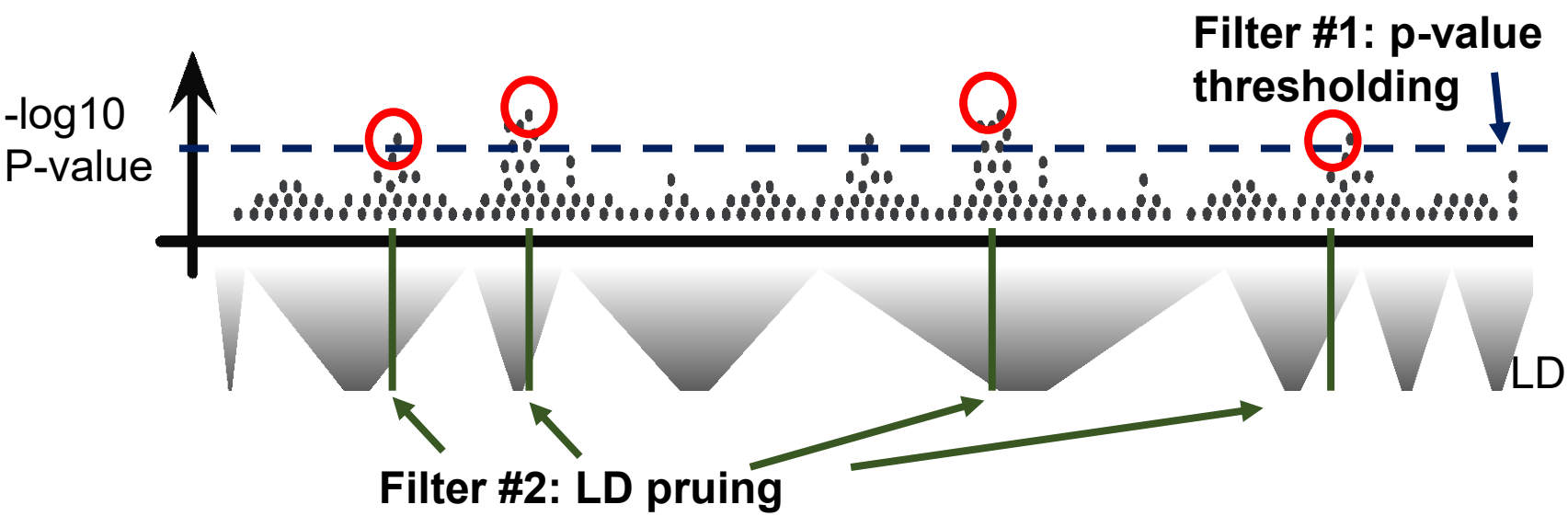
β_j = log(OR of SNP j)

g_j = genotype (dosage)

PRS model:

$$PRS[i] = \Sigma_{j \in \{SNPs\}} \beta_j \, g_j[i]$$

Goal: Tuning this parameter



A common practice of PRS estimation: Cross-validation with observed phenotype

Univariate GWAS statistics:

$\beta_j = \log(\text{OR of SNP } j)$

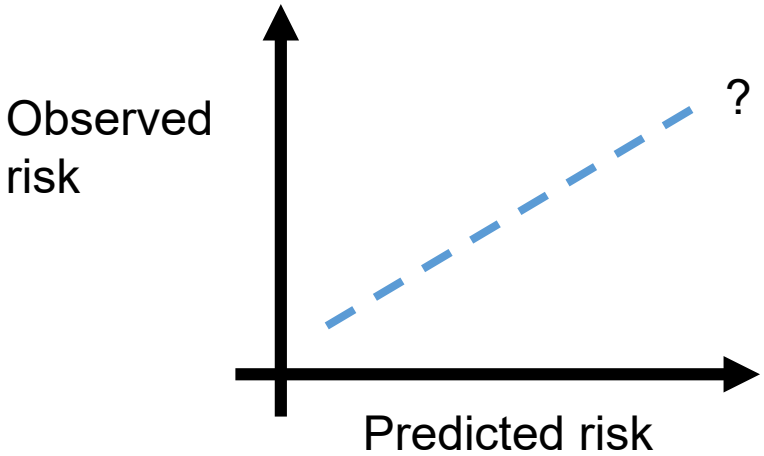
g_j = genotype (dosage)

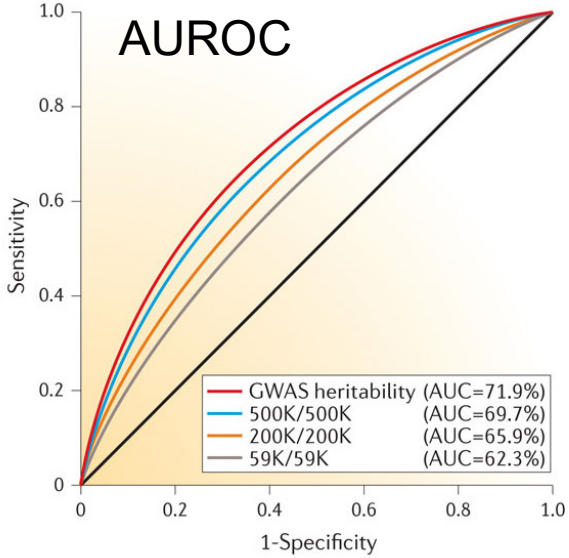
PRS model:

$$PRS[i] = \Sigma_{j \in \{SNPs\}} \beta_j \, g_j[i]$$

Goal: Tuning this parameter

How do we know the selected SNPs are good?





An alternative method for estimating PRS (and a simpler and more powerful way)

Univariate GWAS statistics:

β_j = log(OR of SNP *j*)

g_j = genotype (dosage)

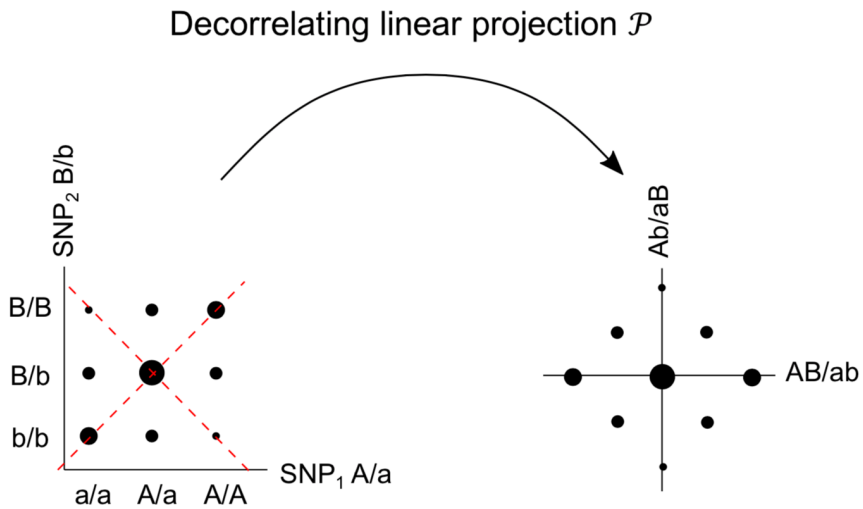
PRS model:

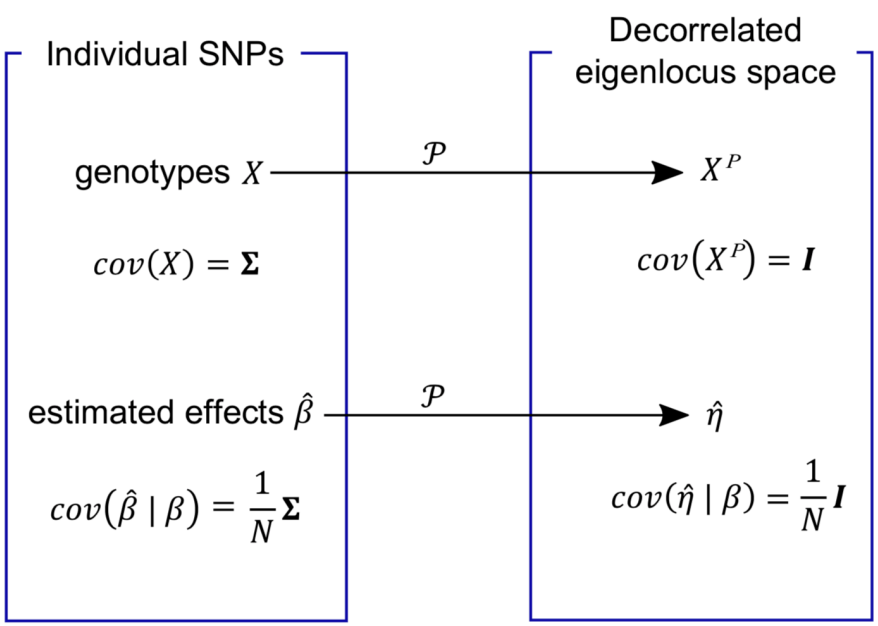
$$PRS[i] = \Sigma_{j \in \{SNPs\}} \beta_j \, g_j[i]$$

What's wrong with using all the SNPs? LD between them. Adjust spurious weak effects.

Chun .. Sunyeav, BioRxiv (2019)
Baker *et al.,* Genetic Epidemiology (2017)

Idea: Decorrelate LD structure

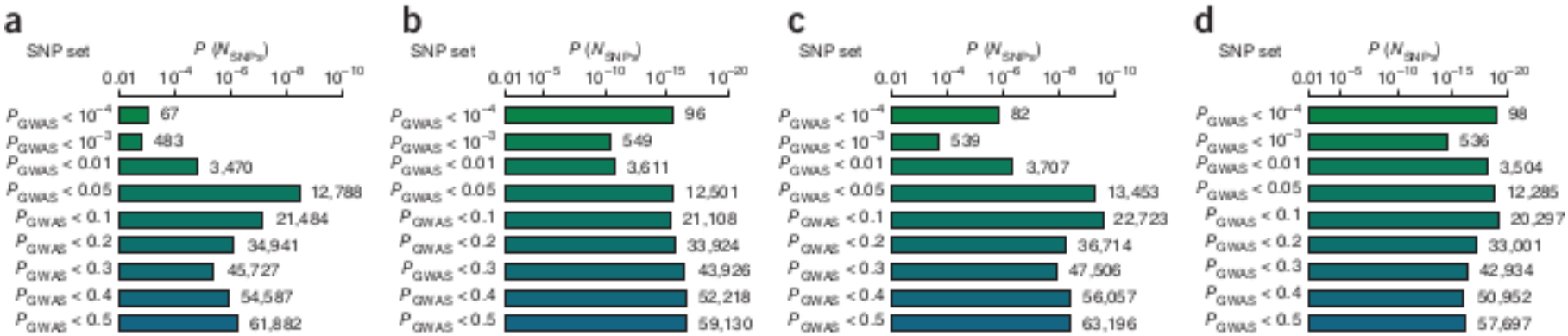




- Transform SNP space to multi-SNP space (SVD)
- Select independent & orthogonal factors.
- Or regularize eigen-values to smooth out spurious associations.
- We don't need much tuning with regularization.

Chun .. Sunyeav, BioRxiv (2019)
Baker *et al.,* Genetic Epidemiology (2017)

Polygenic risk scores



Rheumatoid Arthritis | Celiac | Myocardial infarction | Coronary artery disease

- Aggregate burden of sub-threshold SNPs to improve prediction performance (Stahl 2012)
- As we include more SNPs in the risk score, the association with RA, celiac disease, MI, CAD gets stronger
- In practice, requires tuning of p-value threshold, LD pruning threshold

Phasing diploid genomes is hard

- Humans are **diploid** organisms
- Each individual carries two **homologous** copies of each chromosome
- Therefore, they carry two copies of each variant (called the **maternal/paternal allele**)
- Variants co-occur in **haplotypes** which are inherited as a unit
- Experimentally possible, but currently infeasible, to directly measure haplotypes over the whole genome
- Cheaper and more efficient to measure **genotypes** (counts of minor allele)
- Genotyping loses information, which we need algorithms and statistical models to recover (**phasing, imputation**)

Haplotypes

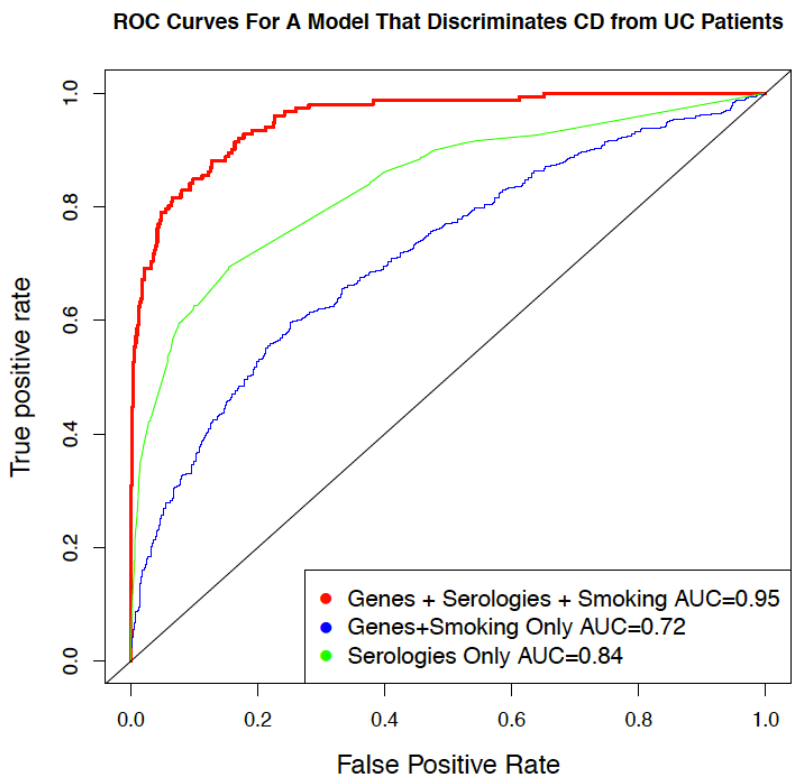
`0 0 1 0 1 1 0` (maternal)

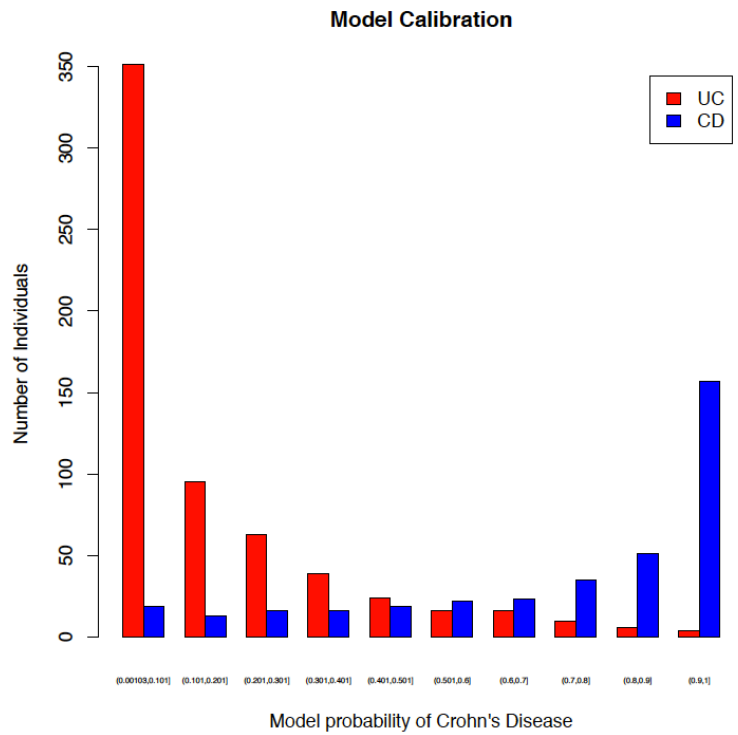
`0 1 1 0 0 1 0` (paternal)

Genotypes

`0 1 2 0 1 2 0`

Molecular diagnostics in IBD



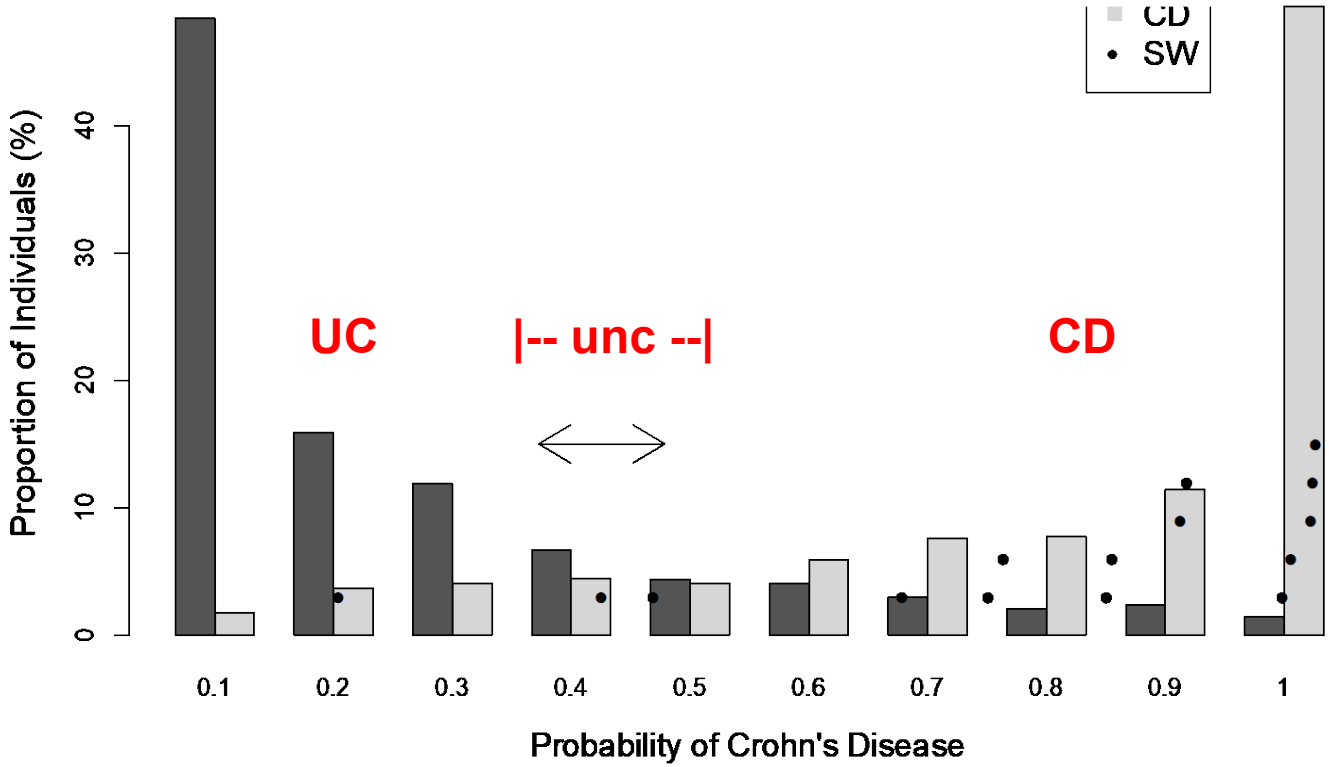


'Molecular' diagnosis (based on GWAS SNPs & serologic biomarkers) concordant with GI dx: CD & UC patients can be distinguished accurately

>90% of patients correctly classified with >90% reliability

Jonah Essers (MGH/CHB), Dermot McGovern (CSMC)

Molecular diagnostics flag patients with worst outcome



Black dots represent patients diagnosed with UC who later underwent colectomy and then developed full-blown Crohn's disease

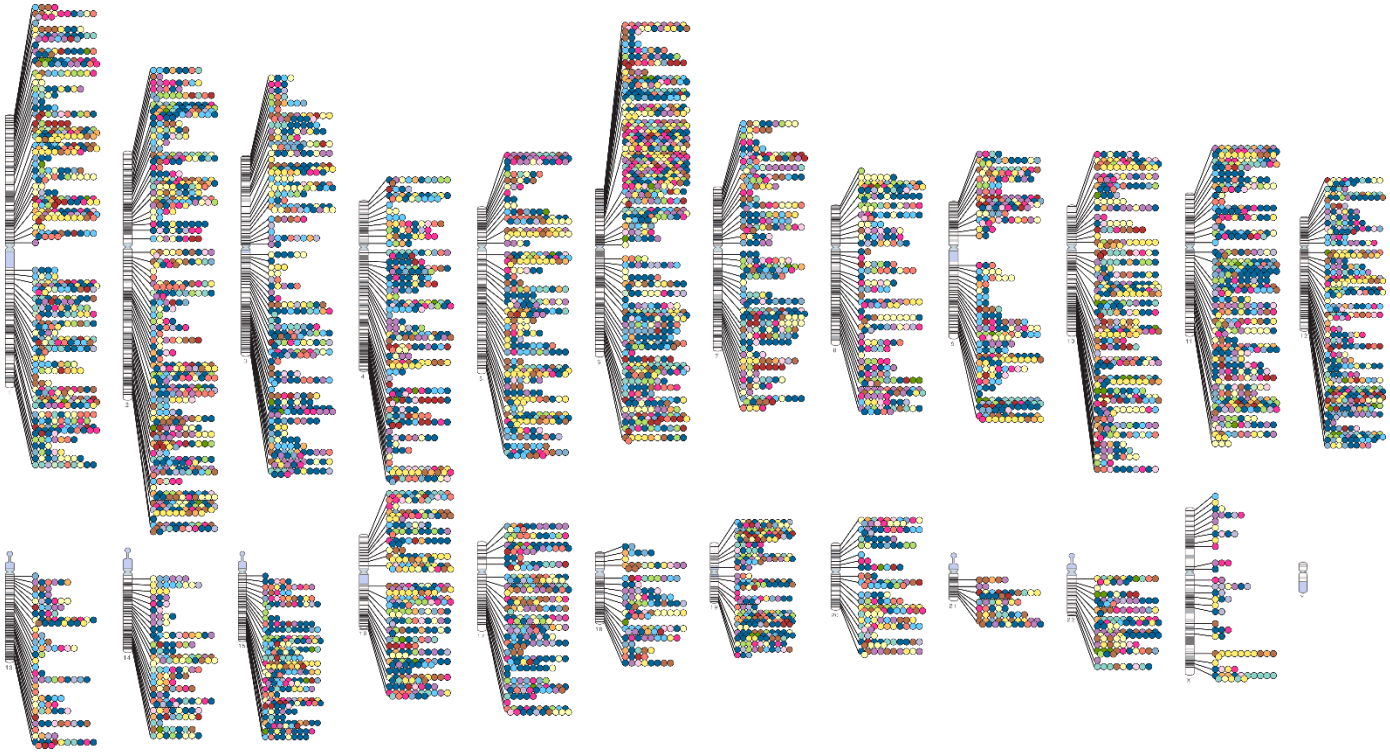
GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

6. Heritability:

Definition, Missing Heritability, Partitioning

Lessons of GWAS



1. **We haven't found all causal loci:** known loci explain little phenotypic variance
2. **Most loci affect transcriptional regulation:** they don't tag coding variation

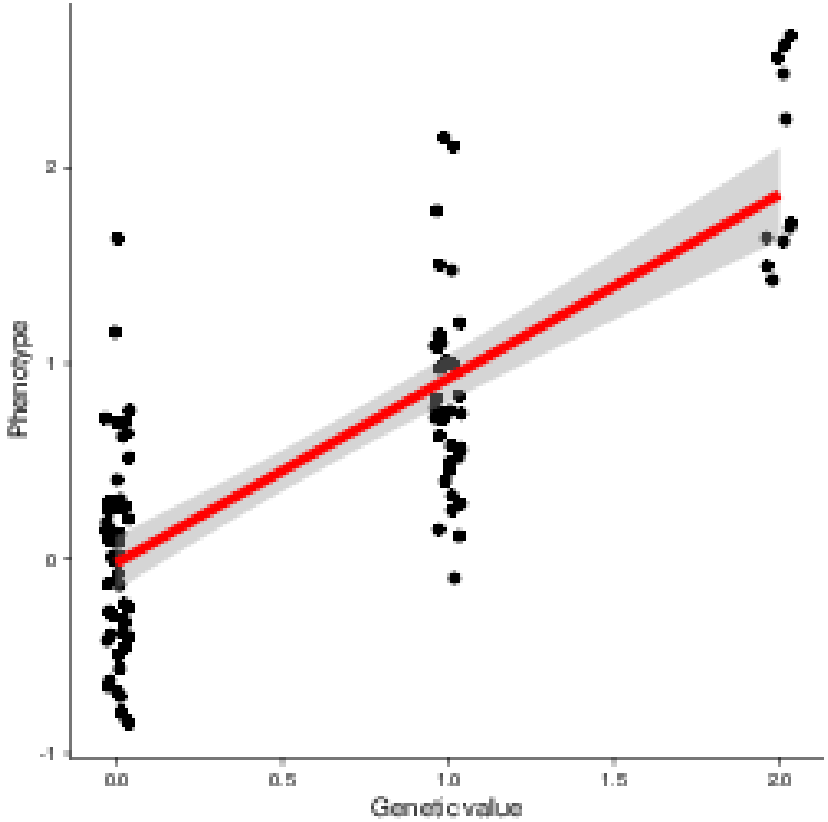
NHGRI GWAS catalog

Components of phenotypic variance

- Assume p (phenotype) = g (genetic) + e (environment)
- Then, V[p] = V[g] + V[e] + ~~2Cov(G,E)~~
  (assume no gene-environment interactions)

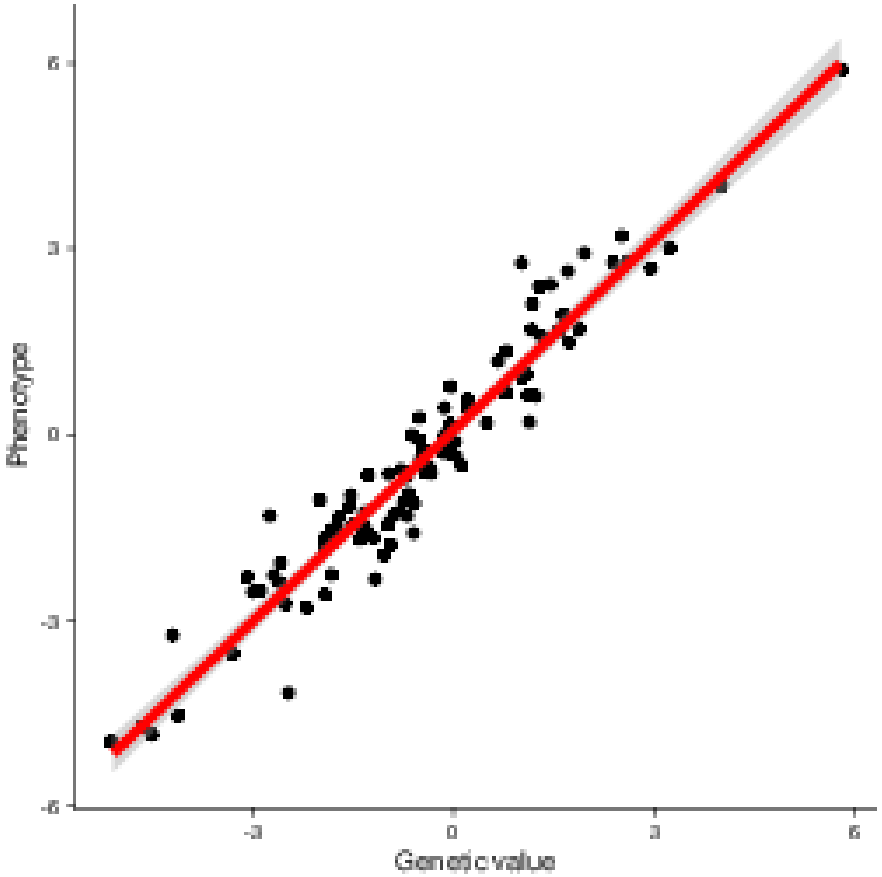
V[phenotype]	
V[genetics]	V[environment]

- Example: one causal variant
- Three possible **genetic values** in the population
- Intuition: V[g] is the variance of mean phenotype across different genetic values
- V[e] is the variance of phenotype for the same genetic value



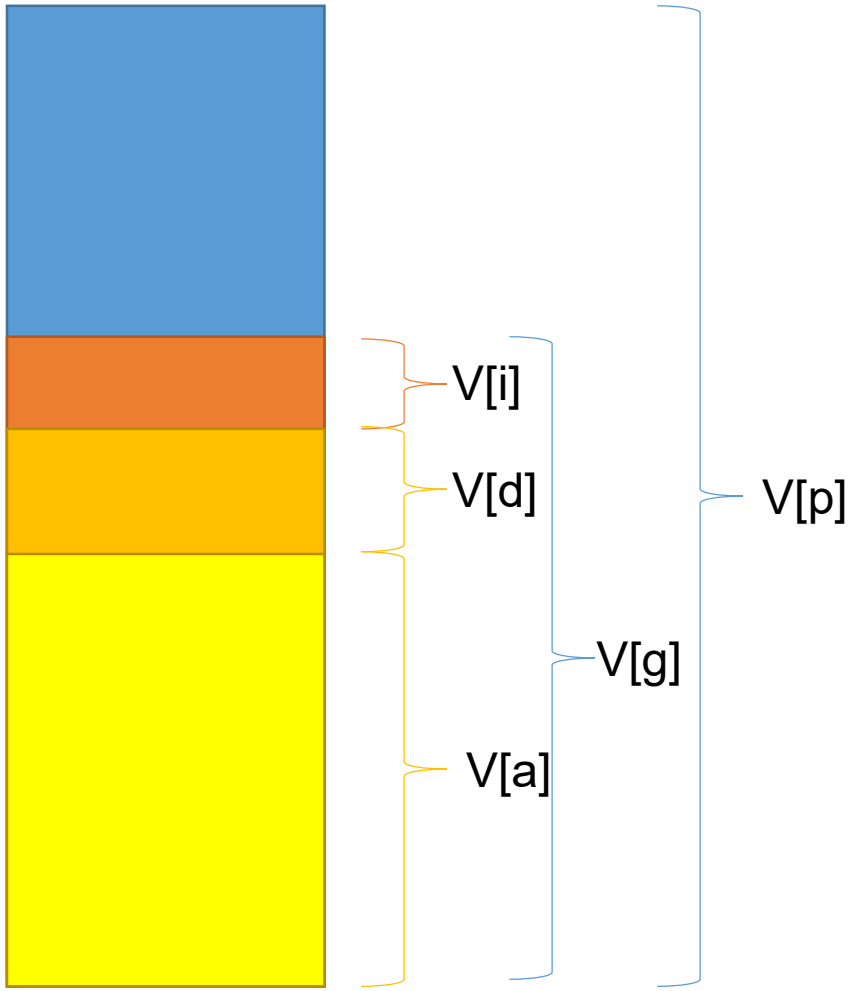
Components of genetic variance

- Assume V[g] = V[a] (additive) + V[d] (dominance) + V[i] (interactions)
- The additive component corresponds to a linear model
- As we add more causal variants, phenotypes become closer to Gaussian
- We could further decompose interactions
- We could include variance due to *de novo* mutations



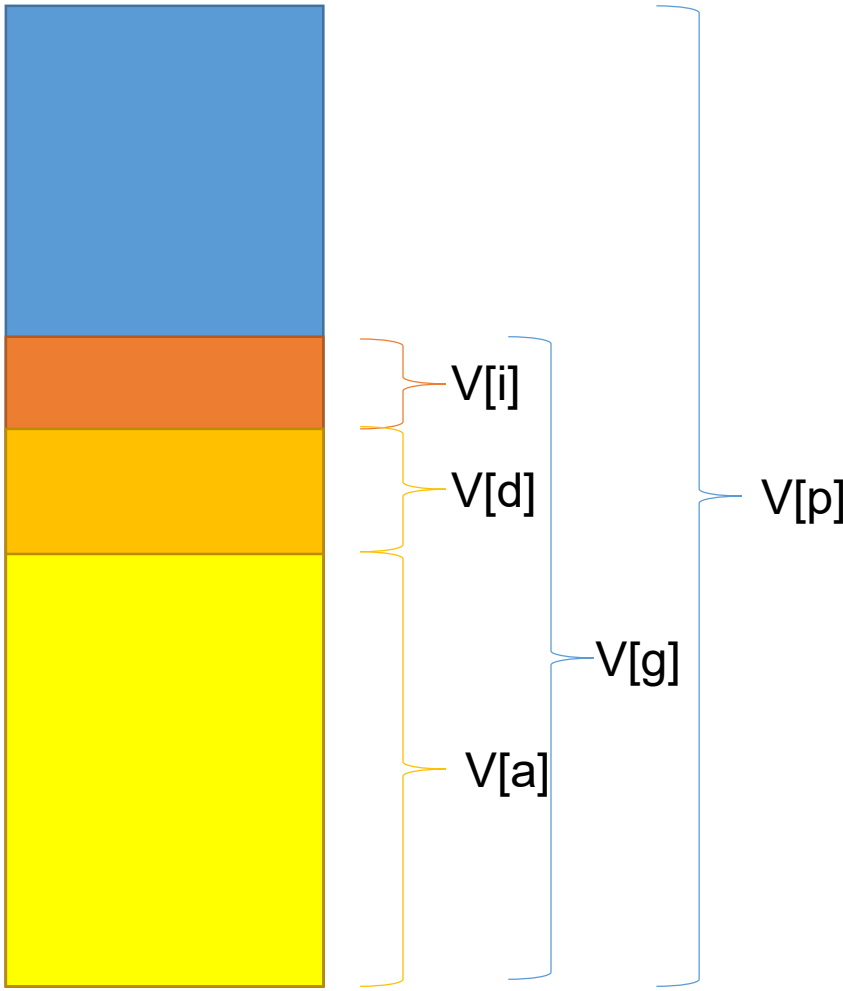
Heritability is a ratio of variances

- V[p] = V[g] + V[e]
- V[g] = V[a] + V[d] + V[i]
- **Broad sense heritability**
  H^2 = V[g] / V[p]
- Broad sense captures all genetic factors
- **Narrow sense heritability**
  h^2 = V[a] / V[p]
- Narrow sense captures only additive effects
- Ongoing debate about the relative importance of additive vs. other effects in disease, selection, etc.



Why study heritability?

- Quantify the importance of genetics vs. environment in traits of interest
- Learn about *genetic architecture*: how many causal variants, effect sizes, allele frequencies
- Narrow sense heritability is the fundamental parameter needed for phenotype prediction (and is the theoretical best possible prediction performance with a linear model)



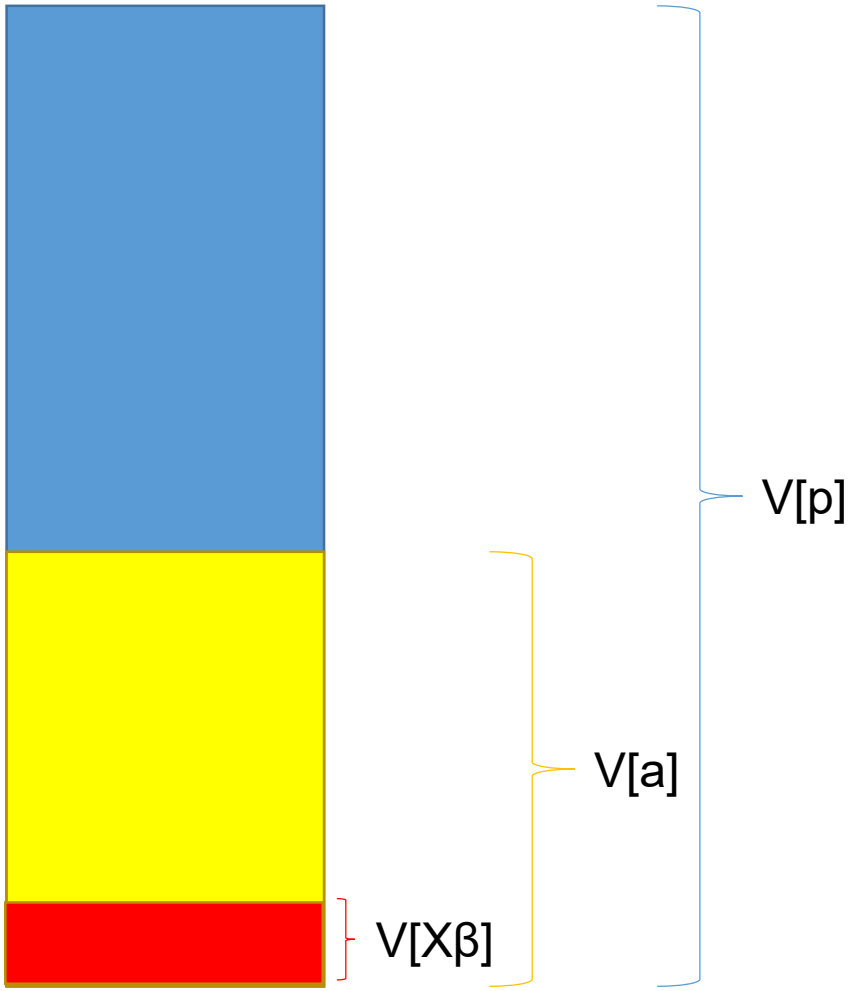
Estimating heritability in relatives

$$p = g + e$$
$$E[p_i\, p_j] = h^2\, E[g_i\, g_j]$$

- Intuition: heritability relates phenotypic correlations to genotypic correlations
- If two individuals have the same allele at each of the causal variants, they will have the same phenotype
- **Haseman-Elston regression:** fit linear regression of phenotypic correlations against genotypic correlations
- Derive genotypic correlation from family relationships: monozygotic twins share 100% of genome, siblings share 50%, etc.
- Example (height): $h^2 = 0.73$

Estimating heritability from GWAS

- Linear model $g = X\beta$
- We can estimate SNP effect sizes β from GWAS
- The variance explained by each SNP depends on effect size and MAF
- $V[X_j \beta_j] = 2 f_j (1 - f_j) \beta_j^2$
- If we do this with genome-wide significant SNPs, we usually $h^2_{GWAS} < h^2$
- Example (height): 253,288 samples; 697 genome-wide significant loci; $h^2_{GWAS}=0.16$, $h^2 = 0.73$
- Known as the **missing heritability problem**

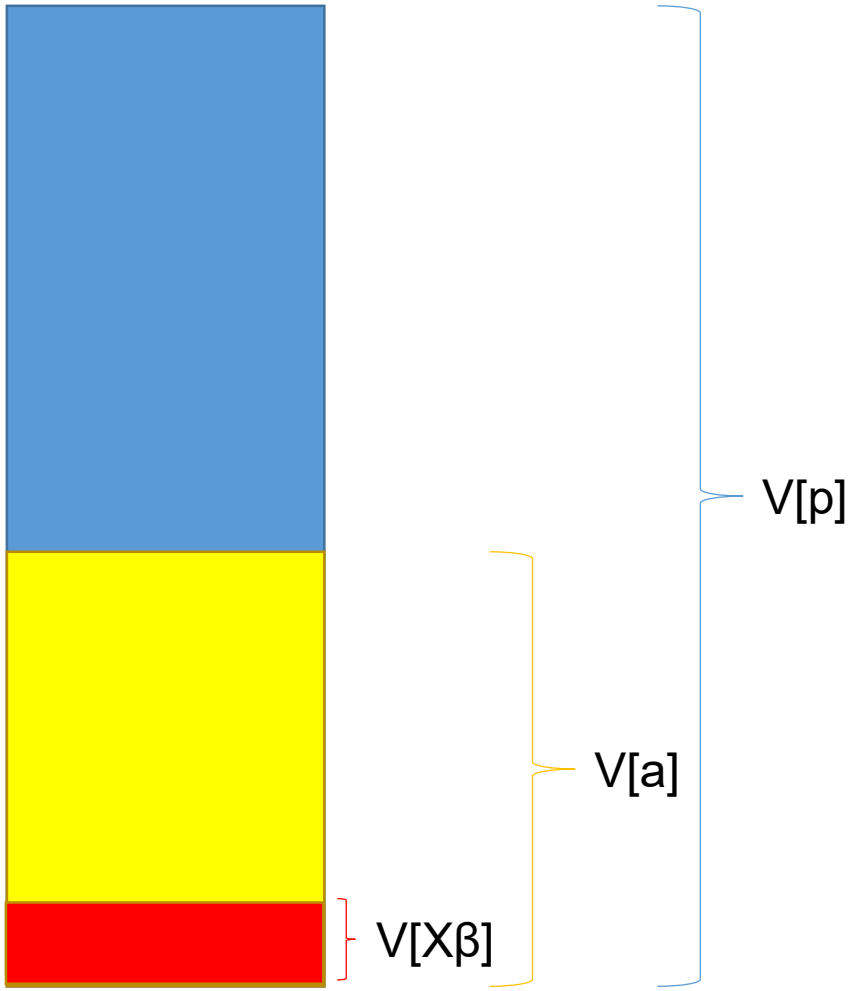


Sources of missing heritability

Ongoing debate about several possible explanations for the missing heritability problem.

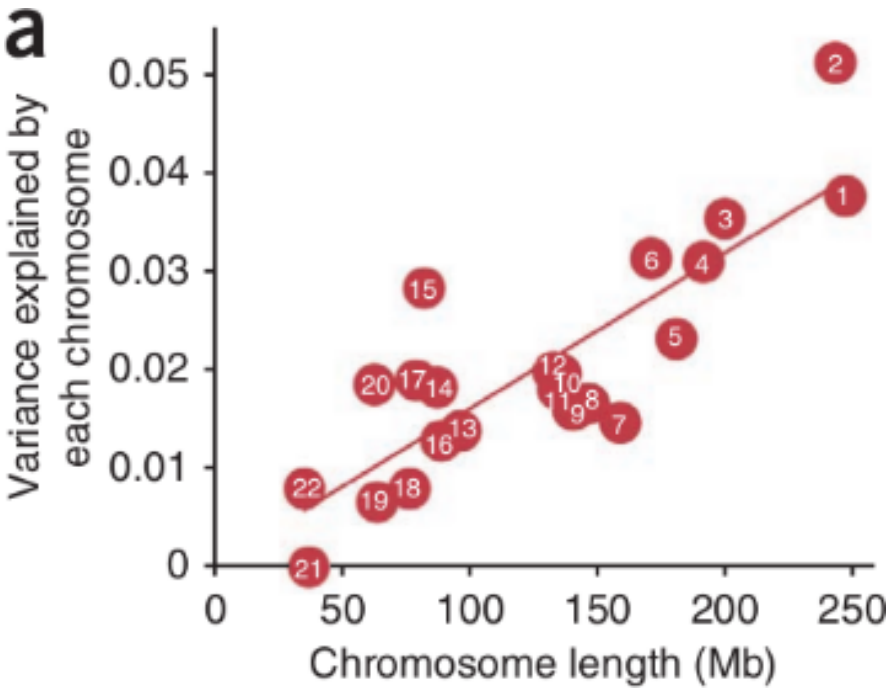
1. Many common variants, small effects
2. Unobserved rare variants, large effects
3. Wrong model assumptions

Each has very different implications for the future of human genetics studies.

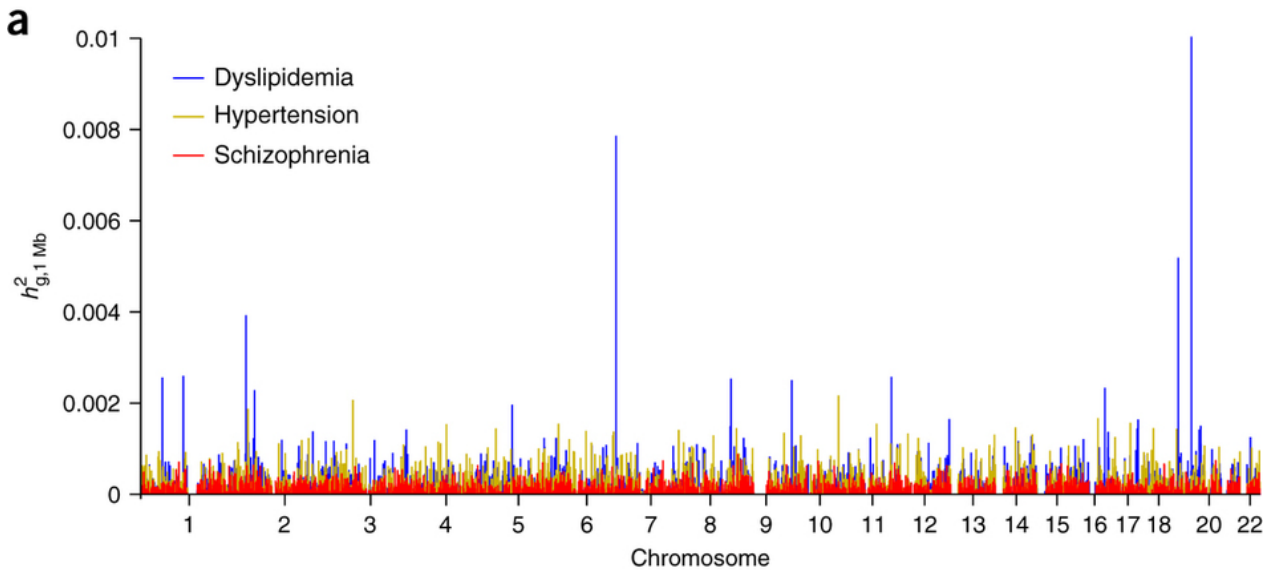


Partitioning heritability

- Extend the model so chromosomes can explain different proportions of variance
- Intuition: add more variance parameters for each partition of SNPs
- Each partition induces a different genetic relationship matrix
- Longer chromosomes explain more heritability
- Suggests causal variants are spread uniformly through the genome

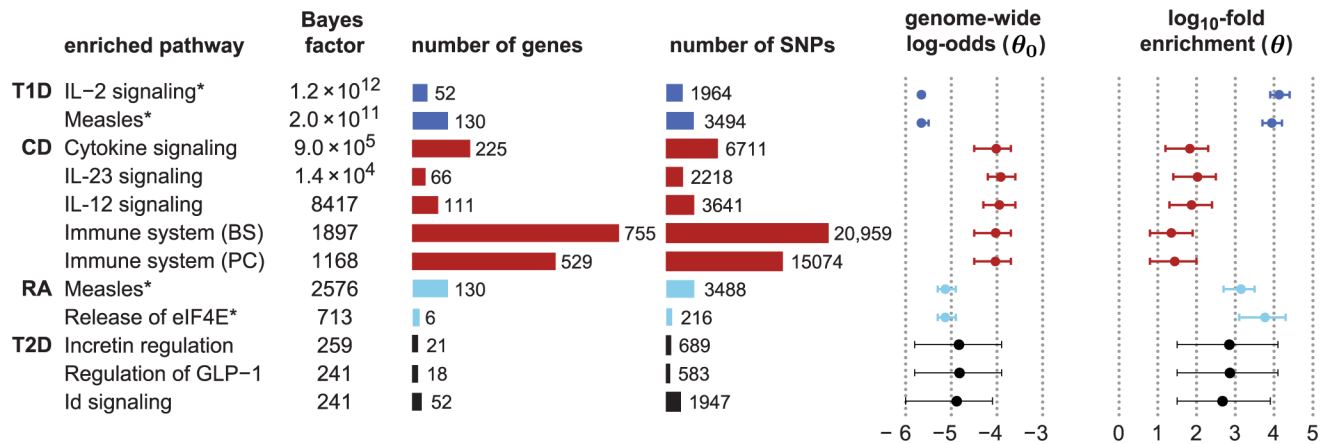


Partitioning heritability



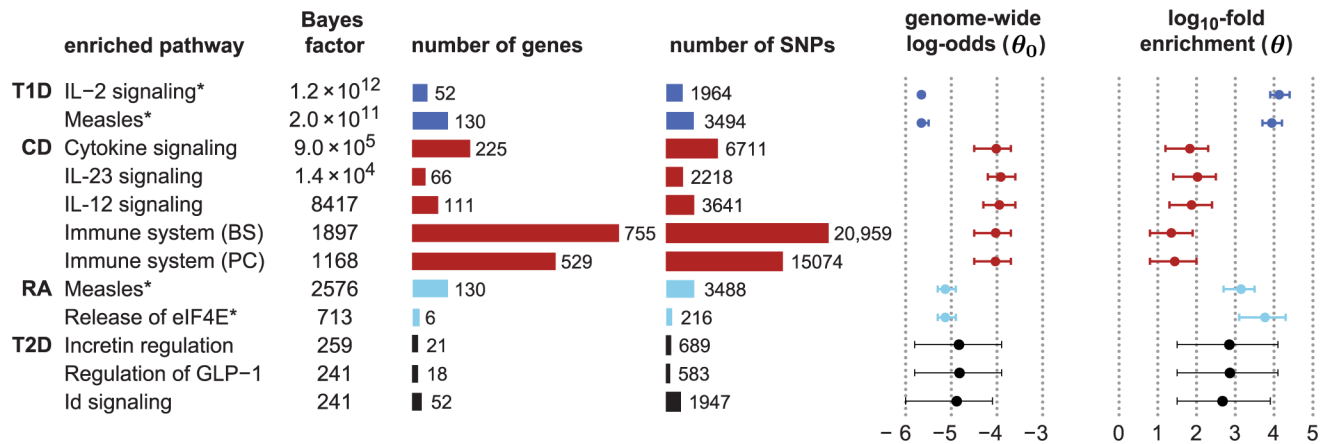
- Fit a model with one component per 1MB window (Loh 2015)
- Bound cumulative heritability explained to estimate number of regions
- Most of the genome explains non-zero heritability

Bayesian variable selection



- Directly fitting the underlying linear model is ill-posed: we have n < p so there are infinitely many solutions
- Idea: use **spike and slab** prior to force many effects to be exactly 0 and regularize the problem (one solution)
- Inference goal: estimate the effect sizes and the level of sparsity (Carbonetto 2013)

Pathways-informed prior from enrichments

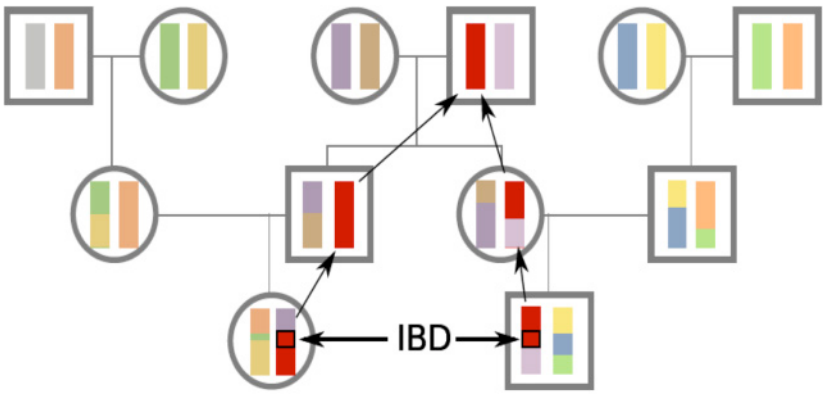


- Extension: some pathways contain more causal variants than the rest of the genome
- Incorporate into the prior
- Identifies relevant immune signaling pathways which are not found using existing methods
- Identifies tens of thousands of SNPs which could be affecting those pathways

Evidence for other explanations

- Incorporating Identity by Descent (IBD) in unrelated individuals
- Partitioning SNPs by MAF, LD
- Assumptions do not hold in real data

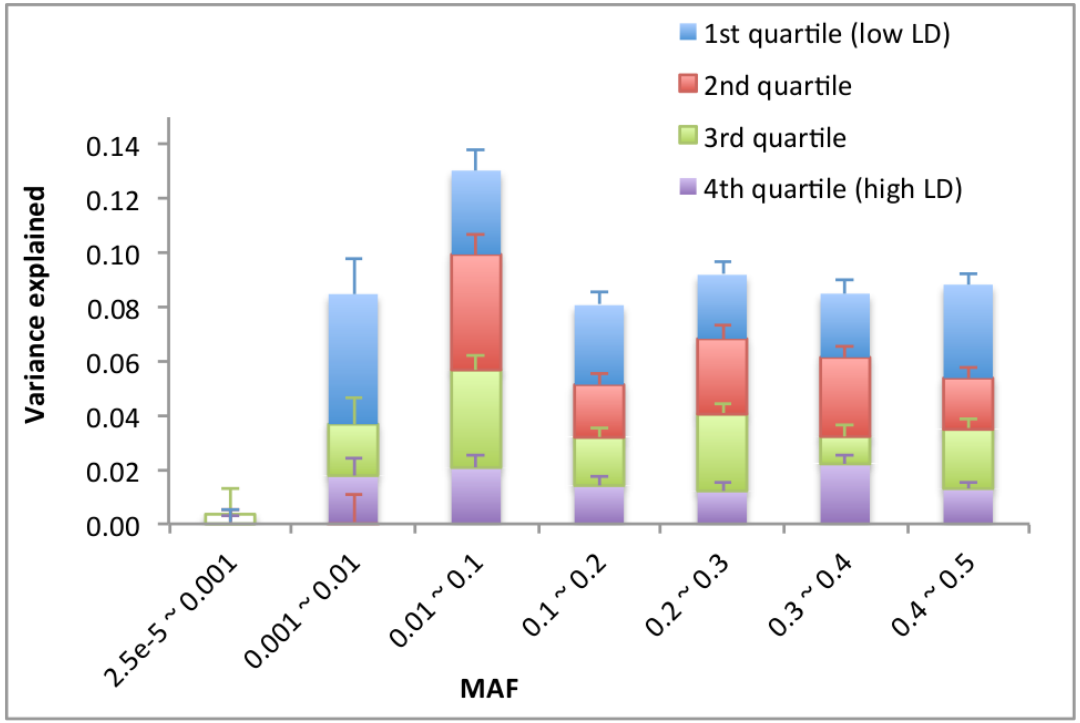
Estimating heritability: shared haplotypes



- Shared haplotypes explain more heritability than tag SNPs
- There is a still a discrepancy between h^2_g and h^2
- If two individual share a chromosomal segment, unobserved variants should also be shared (Bhatia 2015)
- Idea: Identify IBD segments by quickly scanning SNPs and finding stretches of identical alleles
- Inferring shared segments captures rarer variants more effectively than LD

Image credit: http://gcbias.org/european-genealogy-faq/

Partitioning SNPs by MAF/LD



- Low frequency/low LD variants are poorly tagged by observed/imputed variants, so estimate variance for them separately (Yang 2015)
- Partitioning appears to explain all of the heritability of height using only common/low frequency variants!

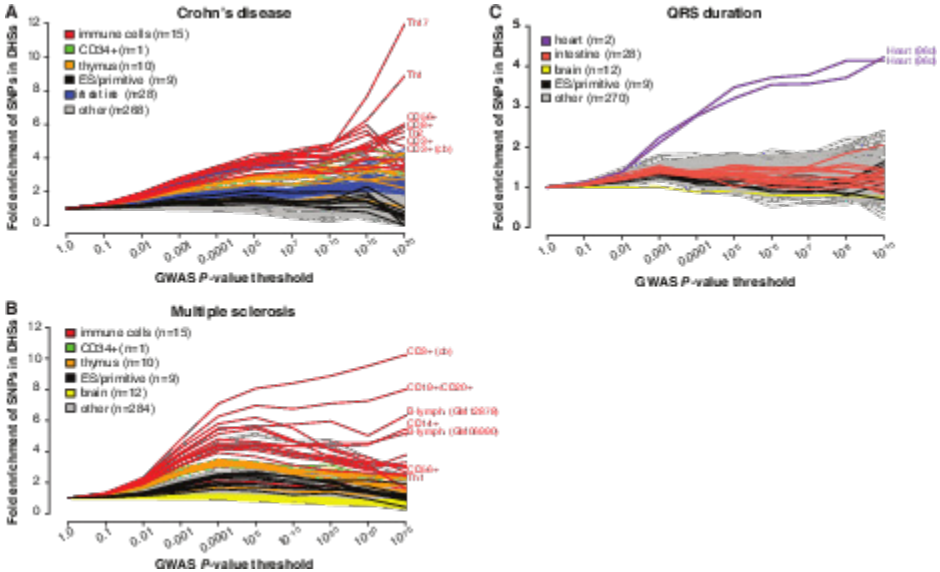
Examining model assumptions

- Phenotypes might not be Gaussian
- GWAS samples are not independent and identically distributed
- SNPs are not independent
- Not all SNPs have an effect
- Not all causal SNPs have equal effects
- There are gene-environment interactions
- There are gene-gene interactions

Limitations of heritability

- Explaining all of the heritability of complex traits is not enough
- As sample size goes to infinity, will the entire genome be associated with all traits? (Goldstein 2009)
- **Goal:** Find biological pathways recurrently disrupted by non-coding variation

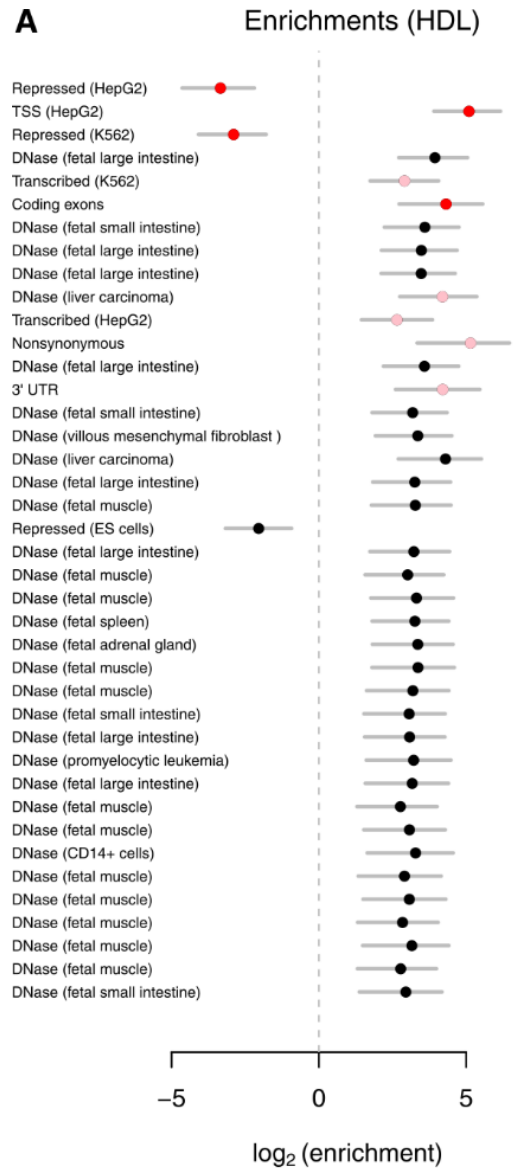
Regulatory enrichments



- Weakly associated variants overlap accessible chromatin more often than expected by chance (Maurano 2012)
- Same trend observed in other predicted regulatory elements: histone peaks, ChromHMM segments, super enhancer clusters

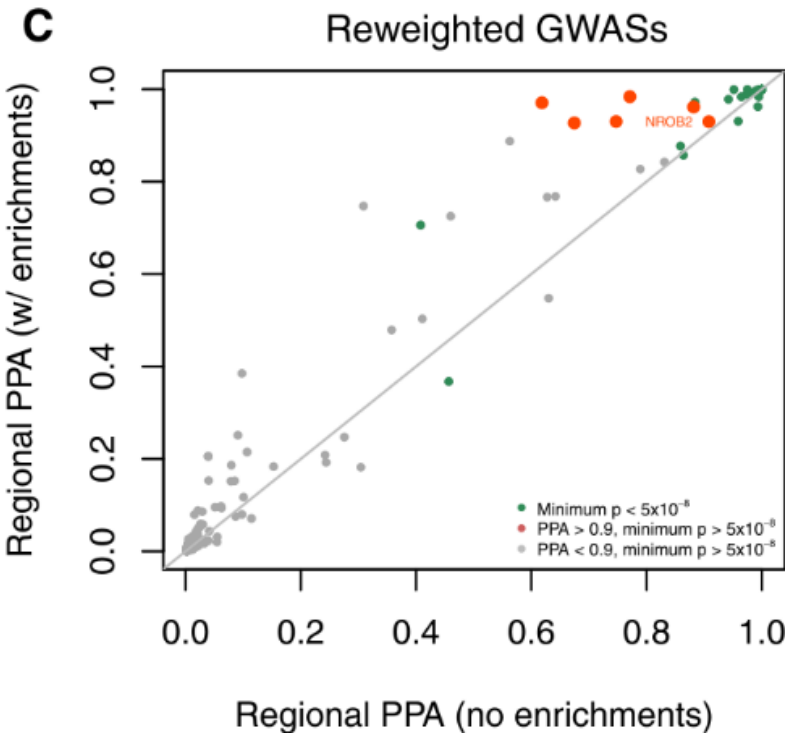
Joint model of SNPs and annotations

- Use **penalized stepwise regression** to pick relevant annotations (Pickrell 2014)
- Use approximate Bayes factors to compute posterior probability of association
- Forward steps: add annotations to the model until they don't explain enough variance
- Backward steps: remove annotations from the fitted model until variance explained drops too much



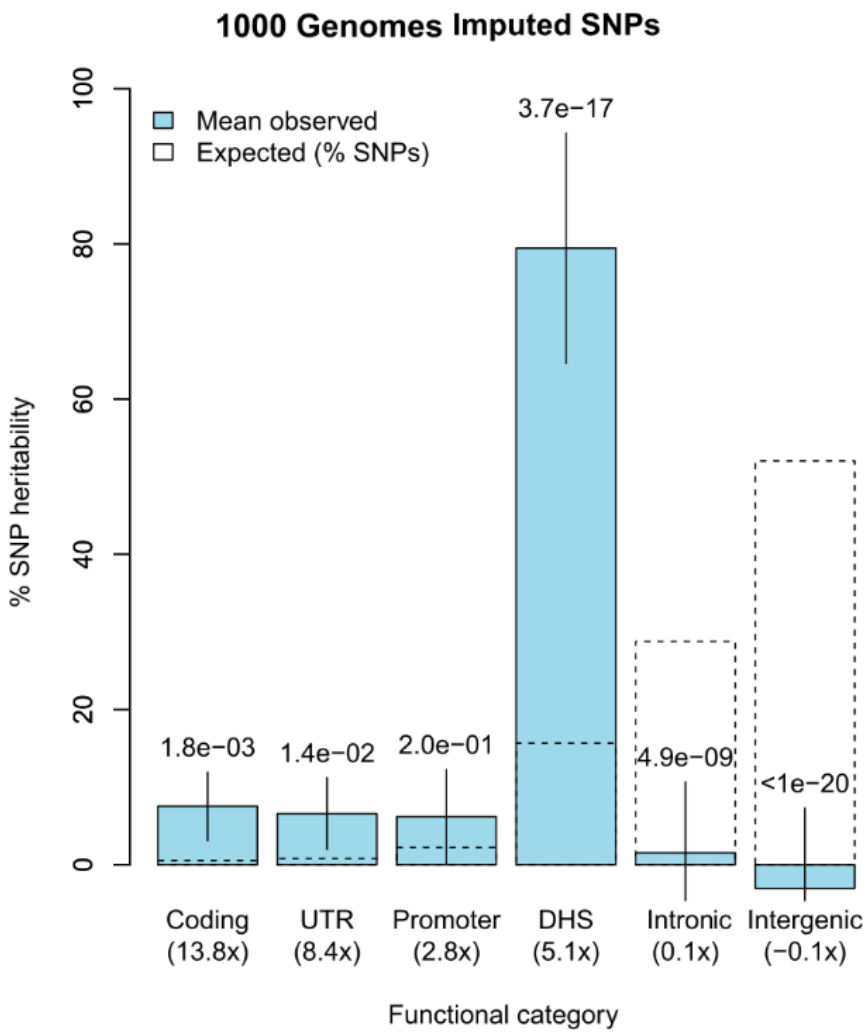
Joint model of SNPs and annotations

- Use approximate Bayes factors to compute posterior probability of association
- Posterior probability of association re-prioritizes new GWAS loci



Partitioning heritability by annotation

- Accessible chromatin explains more heritability
- Combine DHS in >100 cell types: 70% of genome is accessible in some cell type, but only 16% is accessible in multiple cell types
- Implies non-coding SNPs explain more variance per SNP than coding SNPs



GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

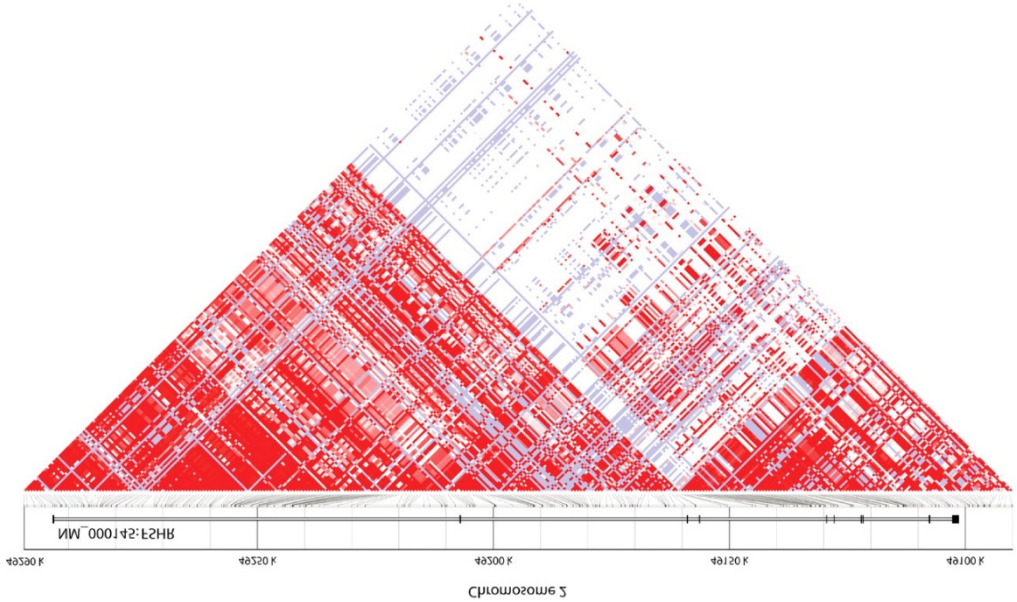
7. LD SCore regression (LDSC):

Computing and partitioning* heritability quickly

(* with stratified LD SCore regression)

LD *SCore* regression (LDSC)

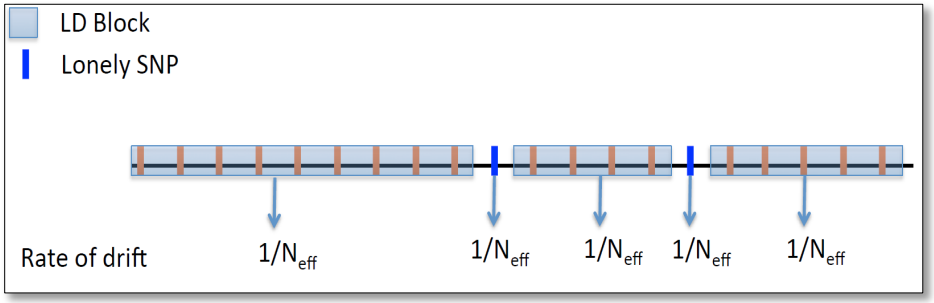
$$E[z_j^2] = N\, l_j\, h^2 / M$$



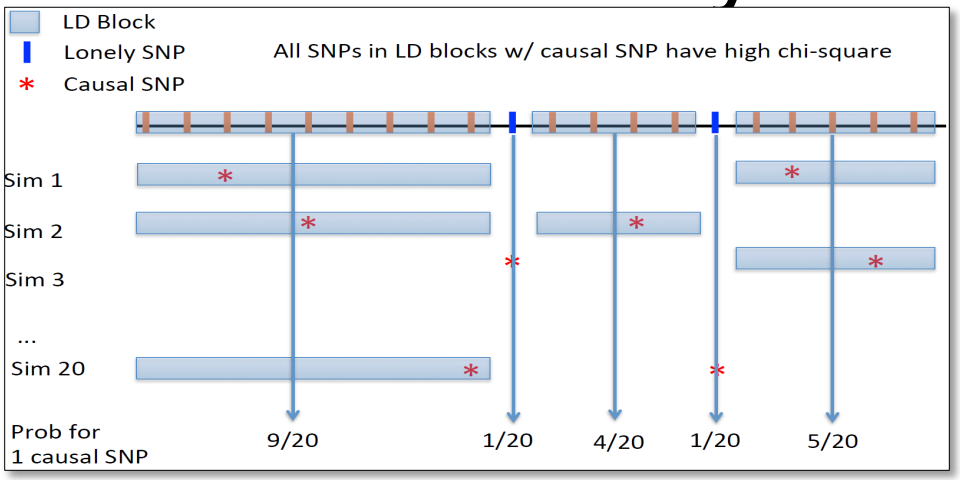
- Intuition: Causal variants drawn uniformly at random from the genome are more likely to come from larger LD blocks (Bulik-Sullivan 2014)
- Linear regression of summary statistics against LD score gives h^2 without access to individual-level genotype matrix

Image credit: Simoni 2008

Intuition: LD score ⇔ heritability



Under pure drift, LD is uncorrelated to magnitude of allele frequency differences between populations

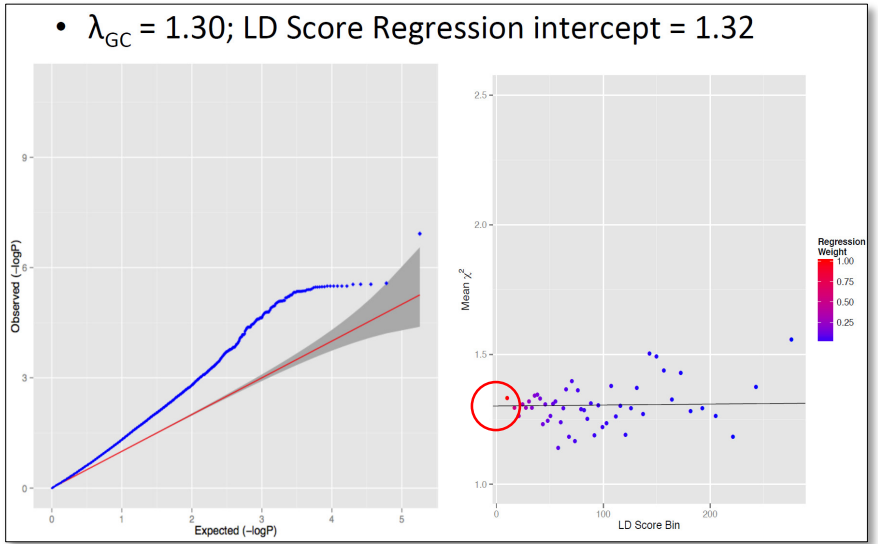


Assuming *i.i.d.* (standardized) effect sizes, more LD yields higher chi-square (on average)

More tags ➔ more causal SNPs.
More shots ➔ more shots on goal

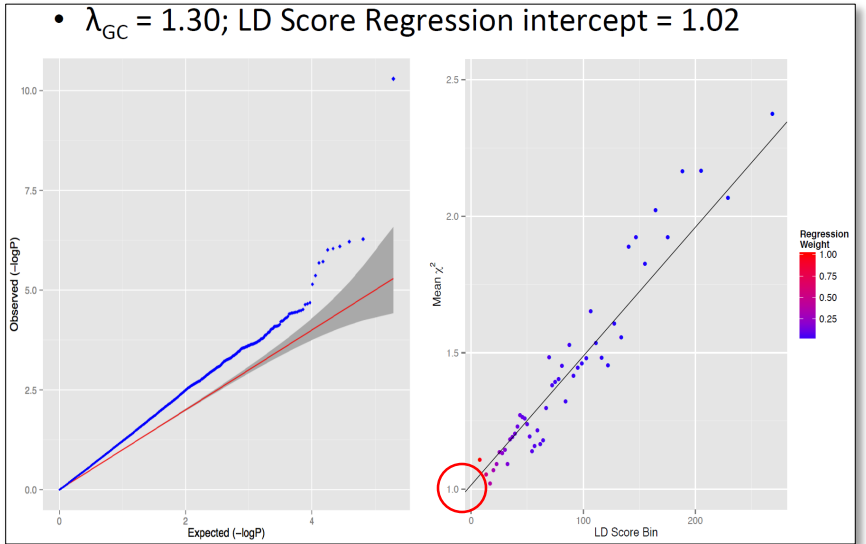
Simulation under stratification

- λ_{GC} = 1.30; LD Score Regression intercept = 1.32



Simulation under association

- λ_{GC} = 1.30; LD Score Regression intercept = 1.02



Linkage disequilibrium: D and D'

- Genetic variants do not segregate independently
- D = coeff. of linkage disequilibrium between alleles A and B at loci L1 and L2
  - $D_{AB}=P_{11}P_{00}-P_{10}P_{01}=0.07$
  - Property of the specific **alleles**. Different alleles at these loci will have diff D_{AB}
- If independent, then $D_{AB}=0$ ($P_{11}P_{00}=P_{10}P_{01}$)
- Linkage disequilibrium measures the degree of departure from Mendel's laws of independent assortment

Haplotype AB	Marginal allele frequency
0*	0.54
1*	0.46
*0	0.30
*1	0.60

How to interpret actual values?

- Relative to D_{ABmax}, which depends on frequencies of individual alleles at A, B
- $D_{ABmax}=P_{0*}P_{*1}-P_{1*}P_{*0}=0.138$
- $D'=D/D_{max}=0.51$

➔ 51% of max possible disequilibrium

Haplotype	Expected	Observed
00	0.162	**0.24****
01	0.324	0.31
10	0.138	**0.07****
11	0.276	**0.39****

Linkage disequilibrium: r^2

- Define
- $r^2 = \frac{D^2}{P(A=0)P(B=0)P(A=1)P(B=1)} = 0.37$
- This really is the squared Pearson correlation of the two SNPs
- In practice, Pearson correlation is efficiently computed for all SNPs in windows as $\mathrm{X}'\mathrm{X}/n$
- This is a fundamental quantity for modeling GWAS *z*-scores

Haplotype AB	Marginal allele frequency
0*	0.54
1*	0.46
*0	0.30
*1	0.60

Haplotype	Expected	Observed
00	0.162	0.24
01	0.324	0.31
10	0.138	0.07
11	0.276	0.39

Key property: r^2 correlation for individual SNPs is exactly the r^2 of the GWAS association summary statistics of these SNPs

LD score regression estimates heritability from summary data

A multivariate model for phenotype variation

phenotype indiv. *i*

$$y_i = \sum_j X_{ij}\beta_j + \varepsilon_i$$

non-genetic for indiv. *i*

multivar. effect on SNP *j*

Assuming $E[X_j]=0$ and $V[X_j] = 1$,
heritability= $V[X\beta] \approx \Sigma X^2\beta^2 \approx \Sigma\beta^2$

$$h^2 = \sum_j \beta_j^2$$

Heritability by partitioning
(restricting on a set C):

$$h^2(C) = \sum_{j \in C} \beta_j^2$$

Finucane *et al.* (2015)

LD score regression estimates heritability from summary data

A multivariate model

$$y_i = \sum_j X_{ij}\beta_j + \varepsilon_i$$

Summary statistics data

$$\chi_j^2$$

(1) X-square tests statistic for all SNP *j*

$$r_{jk}^2$$

and (2) LD matrix (or correlation between SNP *j* and *k*)

Assuming $E[X_j]=0$ and $V[X_j] = 1$,
heritability= $V[X\beta] \approx \Sigma X^2\beta^2 \approx \Sigma\beta^2$

$$h^2 = \sum_j \beta_j^2$$

Heritability by partitioning (restricting on a set C):

$$h^2(C) = \sum_{j \in C} \beta_j^2$$

Finucane *et al.* (2015)

Idea: Reverse-engineer summary data to find multivar. parameters

A univariate effect (GWAS)

$$\hat{\beta}_j = \frac{1}{N} X_j^T \left(X\beta + \epsilon \right)$$

$$= \sum_k \boxed{\hat{r}_{jk}} \beta_k + \epsilon'_j$$

LD between SNP *j* and *k*

A univariate chi-square (GWAS)

$$\chi_j^2 = N \hat{\beta}_j^2$$

$$\mathrm{E}[\chi_j^2] = N \mathrm{E} \left(\sum_k \hat{r}_{jk} \beta_k + \epsilon'_j \right)^2$$

Finucane *et al.* (2015)

Idea: Reverse-engineer summary data to find multivar. parameters

A univariate effect (GWAS)

$$\hat{\beta}_j = \frac{1}{N} X_j^T (X\beta + \epsilon)$$

$$= \sum_k \boxed{\hat{r}_{jk}} \beta_k + \epsilon'_j$$

LD between SNP *j* and *k*

A univariate chi-square (GWAS)

$$\chi_j^2 = N\hat{\beta}_j^2$$

$$\mathrm{E}[\chi_j^2] = N\mathrm{E}\left(\sum_k \hat{r}_{jk}\beta_k + \epsilon'_j\right)^2$$

$$= N\sum_k \hat{r}_{jk}^2 \boxed{\mathrm{E}[\beta_k^2]} + N\mathrm{E}[\epsilon'^2_j]$$

Per SNP variance (heritability)

$$\mathrm{Var}(\beta_j) = \sum_{c:j\in\mathcal{C}_c} \tau_c$$

= $E[\beta_j^2]$ (assuming $E[\beta_j] \approx 0$)

Finucane *et al.* (2015)

Idea: Reverse-engineer summary data to find multivar. parameters

A univariate effect (GWAS)

$$\hat{\beta}_j = \frac{1}{N} X_j^T (X\beta + \epsilon)$$

$$= \sum_k \boxed{\hat{r}_{jk}} \beta_k + \epsilon'_j$$

LD between SNP *j* and *k*

A univariate chi-square (GWAS)

$$\chi_j^2 = N\hat{\beta}_j^2$$

$$\mathrm{E}[\chi_j^2] = N\mathrm{E}\left(\sum_k \hat{r}_{jk}\beta_k + \epsilon'_j\right)^2$$

$$= N\sum_k \hat{r}_{jk}^2 \boxed{\mathrm{E}[\beta_k^2]} + N\mathrm{E}[\epsilon'^2_j]$$

Per SNP variance (heritability)

$$\mathrm{Var}(\beta_j) = \sum_{c:j\in\mathcal{C}_c} \tau_c$$

= $E[\beta_j^2]$ (assuming $E[\beta_j] \approx 0$)

$$\boxed{\mathrm{E}[\chi_j^2] = N\sum_c \tau_c \sum_{k\in\mathcal{C}_c} \hat{r}_{jk}^2 + \sigma_e^2}$$

Finucane *et al.* (2015)

Regression of chi-square statistics on LD scores

$$\mathrm{E}[\chi_j^2] = N \sum_c \tau_c \sum_{k \in \mathcal{C}_c} \hat{r}_{jk}^2 + \sigma_e^2$$

$$E\left[\chi_j^2\right] = N \sum_c \tau_c \ell(j,c) + 1$$

$$\ell(j,c) := \sum_{k \in C_c} r_{jk}^2$$

LD-scores between SNP ***j*** and other SNP ***k*** in annotation ***c***

Intuition: Remove unwanted "double-counting" of annotation enrichment due to LD

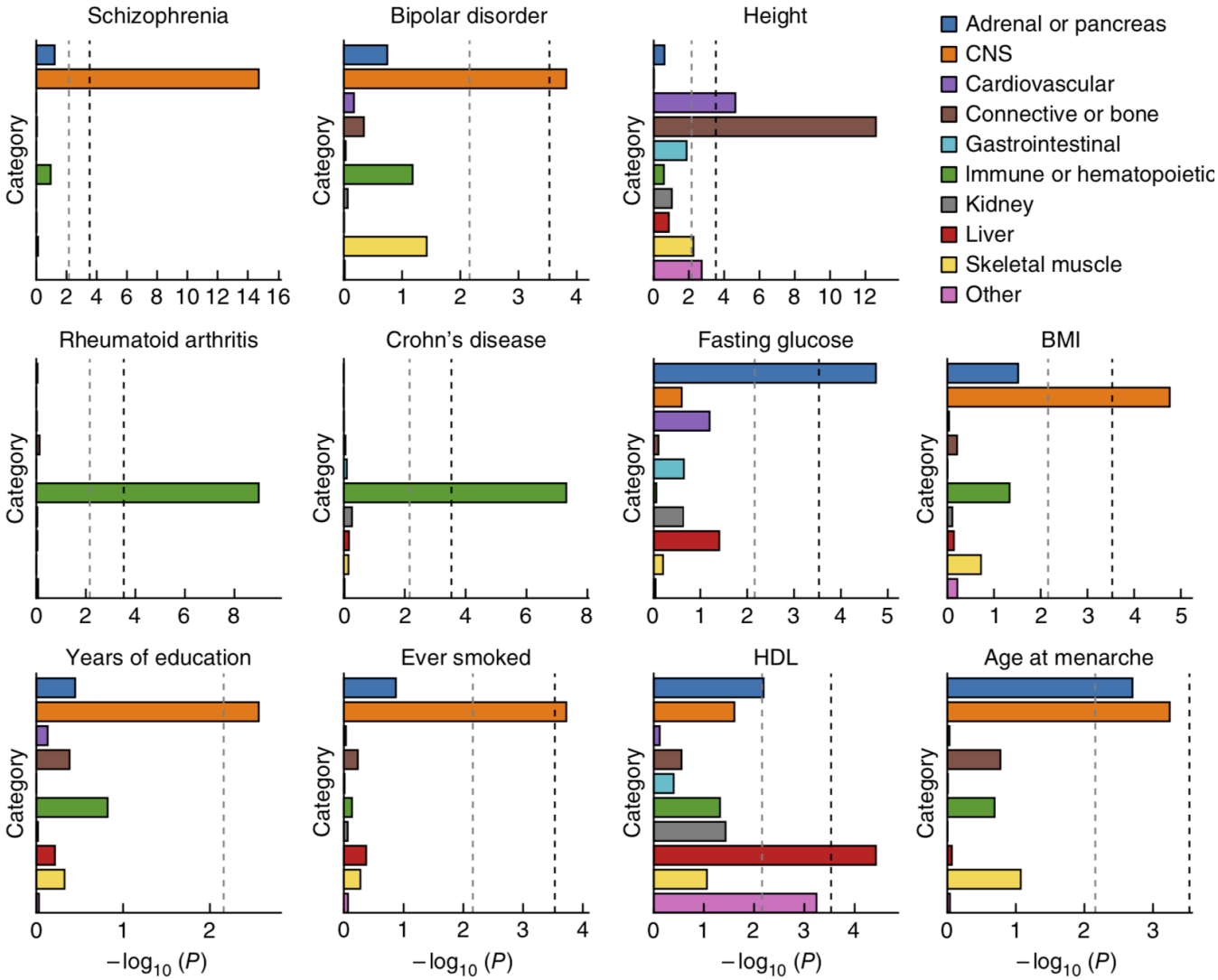
Finucane *et al.* (2015)

Regression to estimate τ_c**:**

$$\begin{pmatrix} \chi^2_1 \\ \chi^2_2 \\ (\ldots) \\ \chi^2_{p-1} \\ \chi^2_p \end{pmatrix} \sim \sum_c \begin{pmatrix} l(1,c) \\ l(2,c) \\ (\ldots) \\ l(p-1,c) \\ l(p,c) \end{pmatrix} \tau_c$$

p SNPs = p observations

Stratified LDSC partitions heritability of complex trait GWAS summary



Finucane *et al.* (2015)

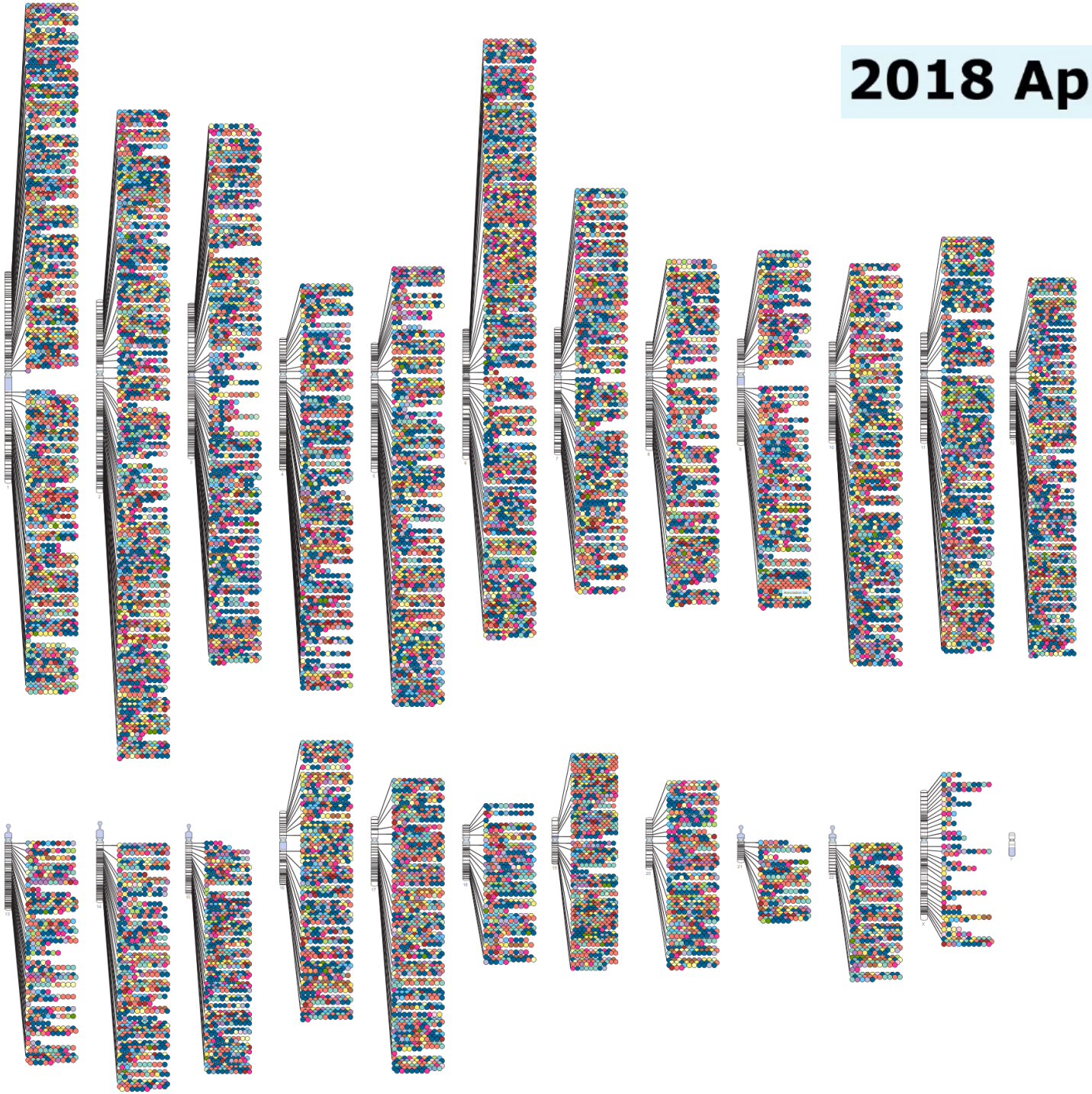
GWAS mechanism: epigenomics, eQTLs, Causality

1. Review: GWAS, fine-mapping, locus mechanistic dissection
2. Global enrichment analyses: epigenomics, Tissues, Regulators, Cell types, target genes
3. eQTLs and mediation analysis: intermediate molecular phenotypes
4. Linear Mixed Models for GWAS and for eQTL calling
5. Polygenic Risk Scores (PRS): Summing over all variants (and more)
6. Heritability: Definition, Missing Heritability, Partitioning Heritability
7. LD Score Regression (LDSC): Computing and partitioning heritability
8. Polygenic and Omnigenic models of disease
9. Guest Lecture: Yongjin Park (UBC) on Causality

8. Polygenic → Omnigenic models of disease

Recognizing “core” vs. “periphery” pathways

Schizophrenia GWAS: Number of significant loci



2018 Apr

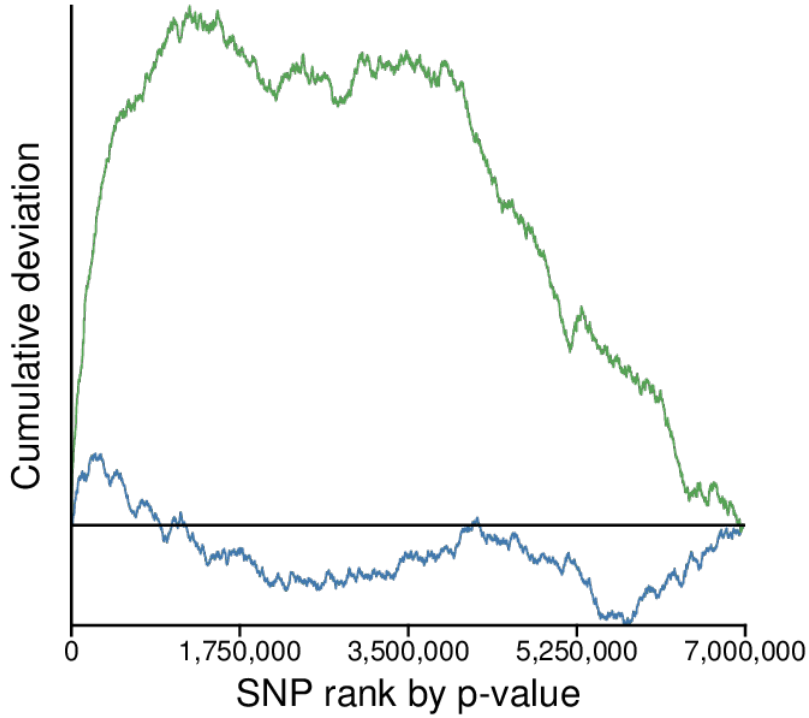
Associations: 69,885

Studies: 5,152

Papers: 3,378

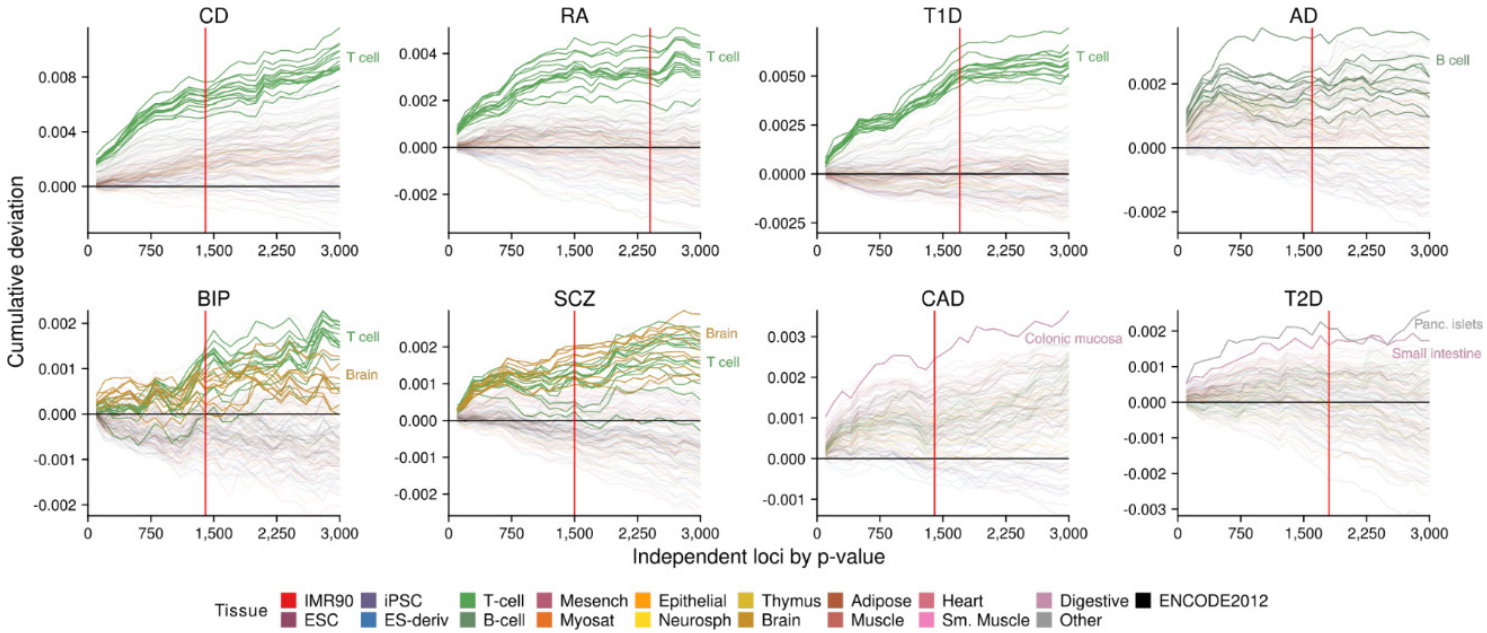
Digestive system disease	968
Cardiovascular disease	674
Metabolic disease	226
Immune system disease	1201
Nervous system disease	1065
Liver enzyme measurement	154
Lipid or lipoprotein measurement	464
Inflammatory marker measurement	326
Hematological measurement	2249
Body weights and measures	1158
Cardiovascular measurement	679
Other measurement	5044
Response to drug	275
Biological process	631
Cancer	979
Other disease	1160
Other trait	3871

How far down the SNP list does enrichment go?



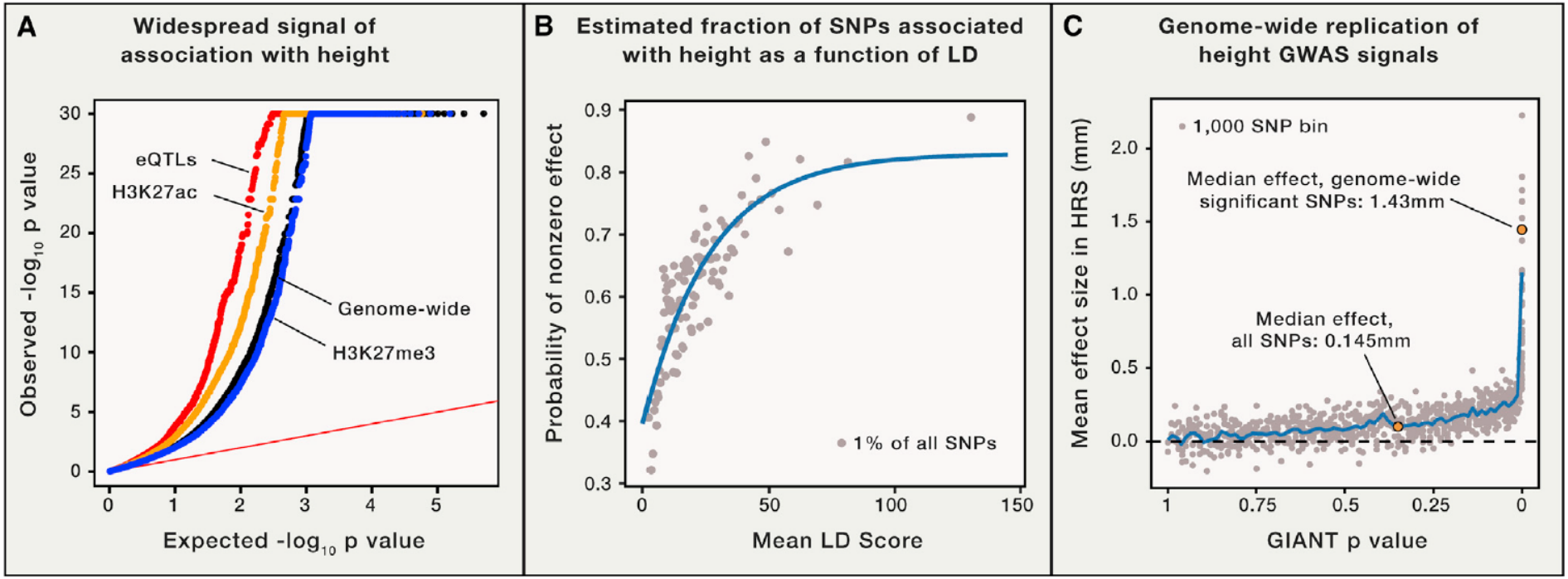
- Use functional enrichment to gain insight into genetic architecture (Sarkar 2016)
- Idea: as we consider more SNPs beyond genome-wide significance, relevant regulatory regions will be disrupted more often than irrelevant regions

Long tails of enrichment for 8 diseases



- Use functional enrichment to gain insight into genetic architecture (Sarkar 2016)
- Idea: as we consider more SNPs beyond genome-wide significance, relevant regulatory regions will be disrupted more often than irrelevant regions

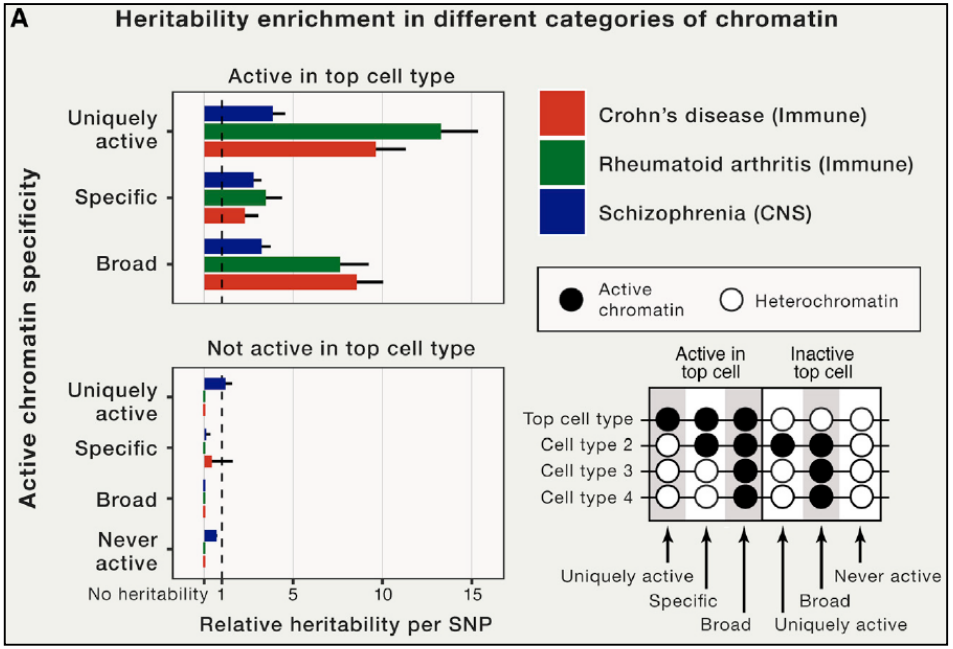
Omnigenic model of heritability

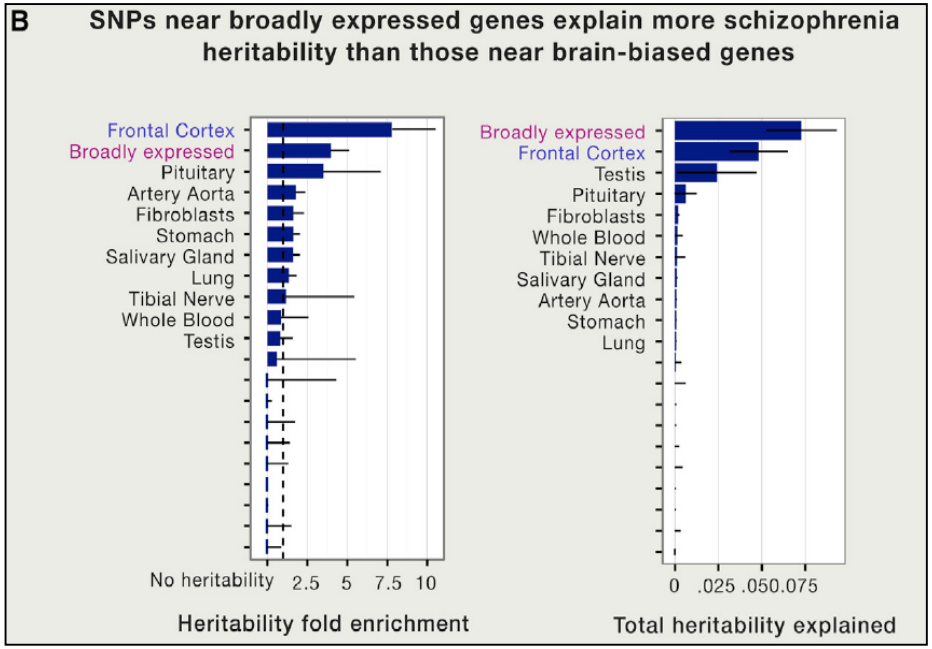


- (A) Genome-wide inflation of small p values from the GWAS for height, with particular enrichment among expression quantitative trait loci and single-nucleotide polymorphisms (SNPs) in active chromatin (H3K27ac).
- (B) Estimated fraction of SNPs associated with non-zero effects on height (Stephens, 2017) as a function of linkage disequilibrium score (i.e., the effective number of SNPs tagged by each SNP; Bulik-Sullivan et al., 2015b). Each dot represents a bin of 1% of all SNPs, sorted by LD score. Overall, we estimate that 62% of all SNPs are associated with a non-zero effect on height. The best-fit line estimates that 3.8% of SNPs have causal effects.
- (C) Estimated mean effect size for SNPs, sorted by GIANT p value with the direction (sign) of effect ascertained by GIANT. Replication effect sizes were estimated using data from the Health and Retirement Study (HRS). The points show averages of 1,000 consecutive SNPS in the p-value-sorted list. The effect size on the median SNP in the genome is about 10% of that for genome-wide significant hits.

Boyle, Li, Pritchard, Cell, 2017

More heritability in broad classes



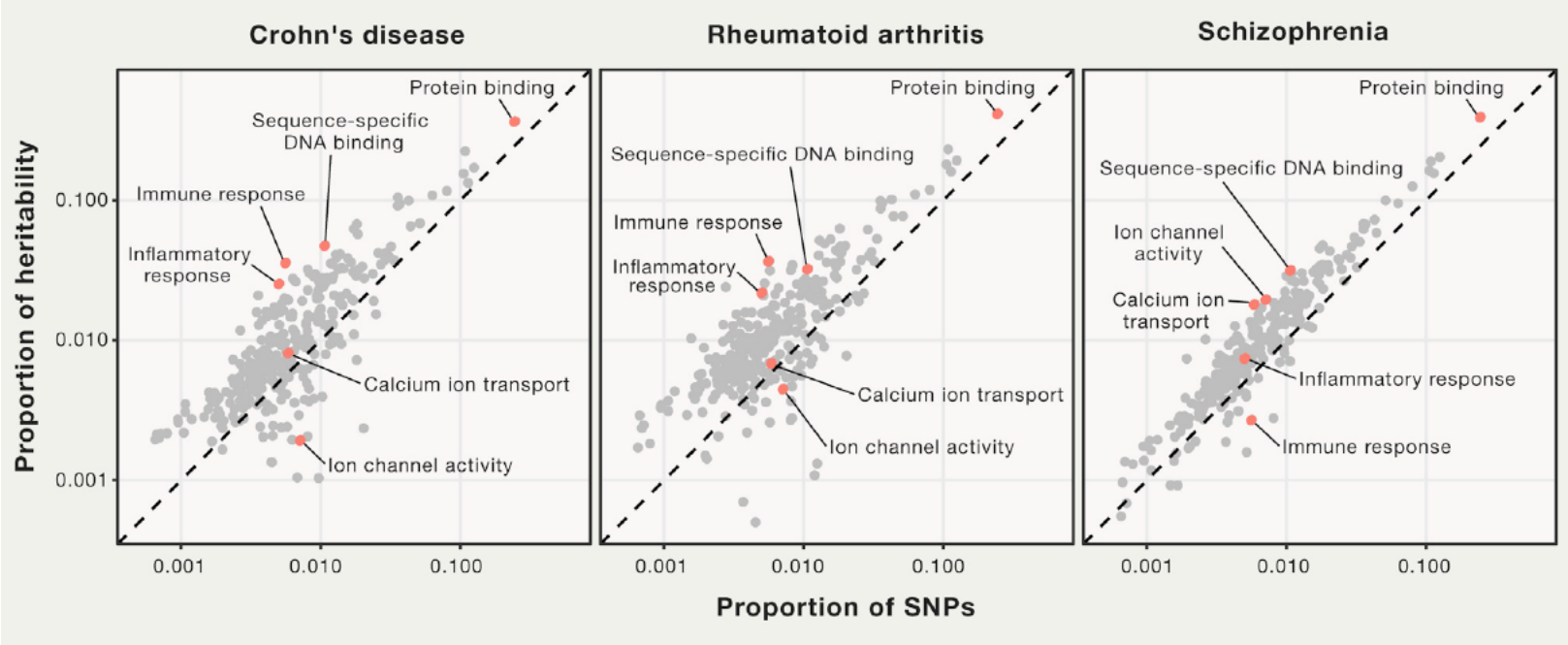


- Contributions to heritability (relative to random SNPs) as a function of chromatin context. There is enrichment for signal among SNPs that are in chromatin active in the relevant tissue, regardless of the overall tissue breadth of activity

- Genes with brain-specific expression show the strongest enrichment of schizophrenia signal (left), but broadly expressed genes contribute more to total heritability due to their greater number (right)

Boyle, Li, Pritchard, Cell, 2017

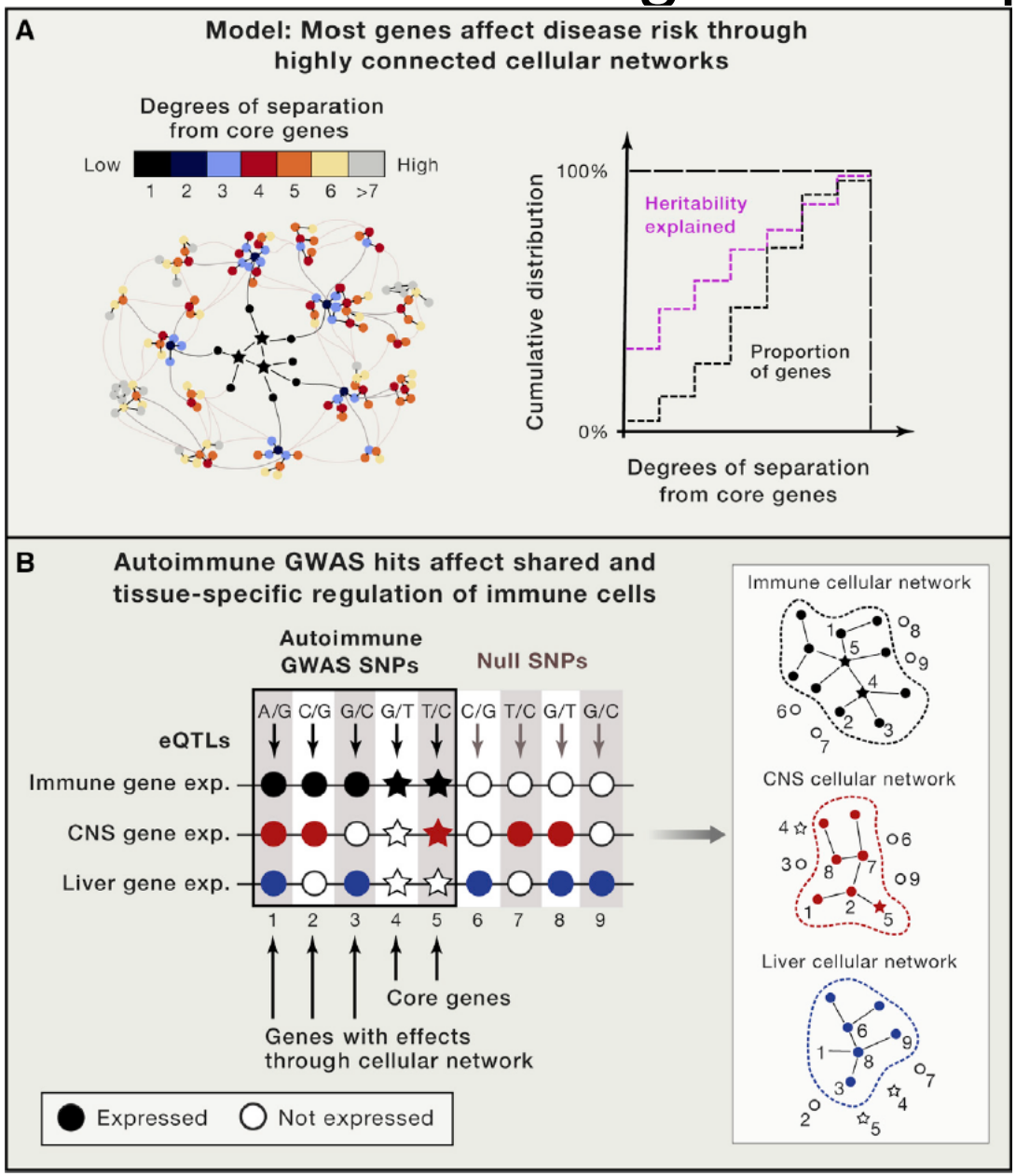
Most GO categories are enriched



- Gene Ontology Enrichments for Three Diseases, with Categories of Particular Interest Labeled. The x axis indicates the fraction of SNPs in each category; the y axis shows the fraction of heritability assigned to each category as a fraction of the heritability assigned to all SNPs. Note that the diagonal indicates the genome-wide average across all SNPs; most GO categories lie above the line due to the general enrichment of signal in and around genes. Analysis by stratified LD score regression

Boyle, Li, Pritchard, Cell, 2017

Core genes vs. periphery



- Omnigenic Model of Complex Traits
- (A) For any given disease phenotype, a limited number of genes have direct effects on disease risk. However, by the small world property of networks, most expressed genes are only a few steps from the nearest core gene and thus may have non-zero effects on disease. Since core genes only constitute a tiny fraction of all genes, most heritability comes from genes with indirect effects.
- (B) Diseases are generally associated with dysfunction of specific tissues; genetic variants are only relevant if they perturb gene expression (and hence network state) in those tissues. For traits that are mediated through multiple cell types or tissues, the overall effect size of any given SNP would be a weighted average of its effects in each cell type.

Boyle, Li, Pritchard, Cell, 2017