Identifying disease-relevant cell types from GWAS data

June 1, 2018

Symposium on Advances in Genomics, Epidemiology, and Statistics

Hilary Finucane

Schmidt Fellow, Broad Institute

Acknowledgements

Harvard T.H. Chan School of Public Health:

- **Alkes Price**
- Steven Gazal
- Alexander Gusev
- Sara Lindstrom
- Po-Ru Loh
- Samuela Pollack
- Yakir Reshef

Harvard Medical School:

- Steve McCarroll
- Soumya Raychaudhuri
- Arpiar Saunders
- Kamil Slowikowski

University of Cambridge:

- Felix Day
- John Perry

Broad Institute:

- **Ben Neale**
- **Brendan Bulik-Sullivan**
- Verneri Anttila
- Andrea Byrnes
- Mark Daly
- Kyle Farh
- Giulio Genovese
- Evan Macosko
- Nick Patterson

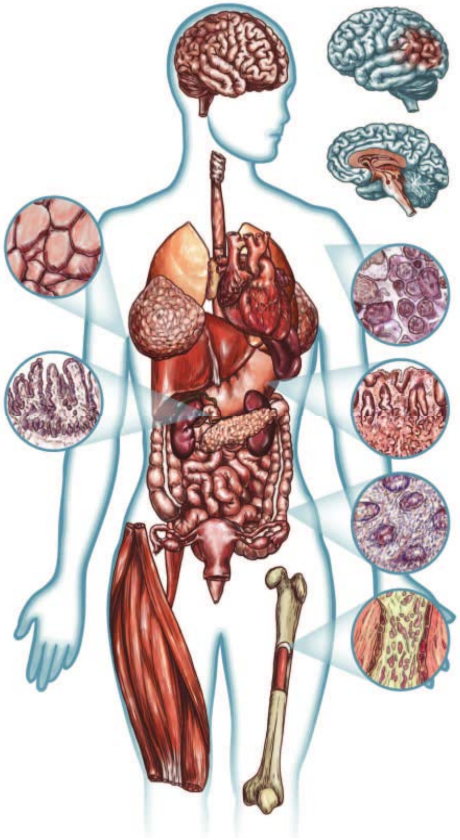
Consortia:

- Brainstorm Consortium
- GTEx Consortium
- Psychiatric Genomics Consortium
- RACI Consortium
- ReproGen Consortium
- UK Biobank

NIH National Institutes of Health

the Hertz FOUNDATION *freedom to innovate*

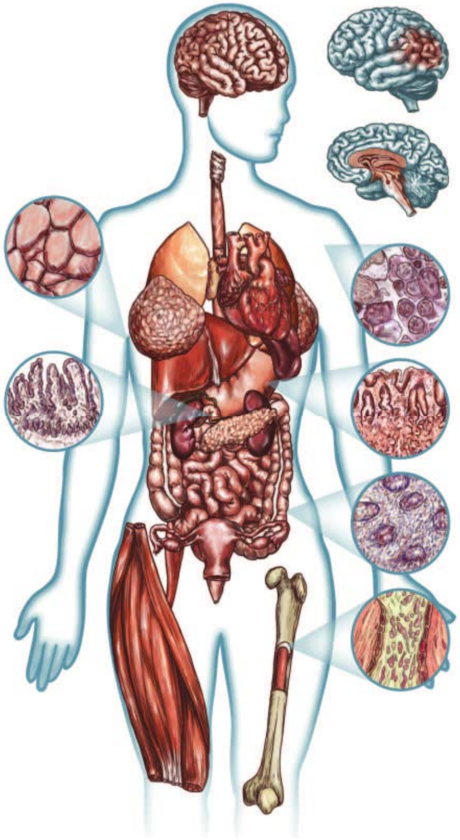
What cell types should we be studying?



Anterior caudate

Roadmap Epigenomics Consortium 2015 *Nature*

What cell types should we be studying?



Anterior caudate

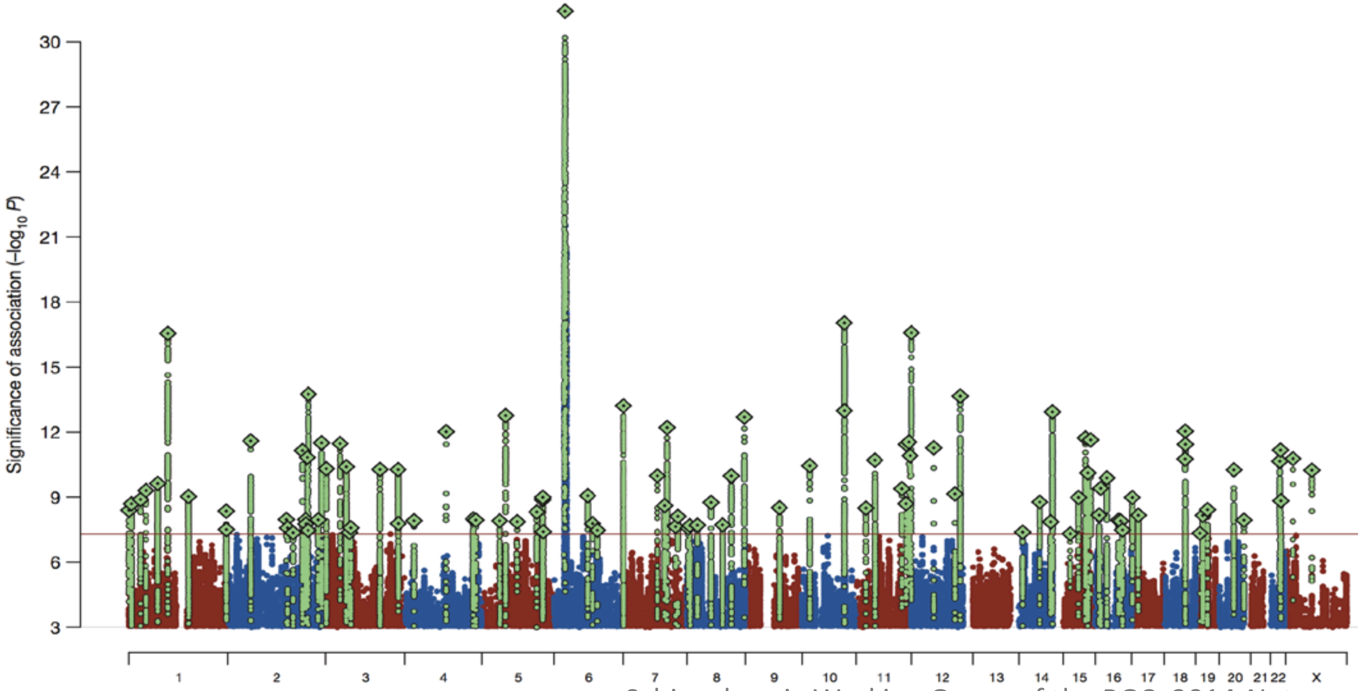
GWAS + external data

↓

phenotype-relevant cell types and tissues

Roadmap Epigenomics Consortium 2015 *Nature*

How much of the genetic signal falls in this genome annotation?

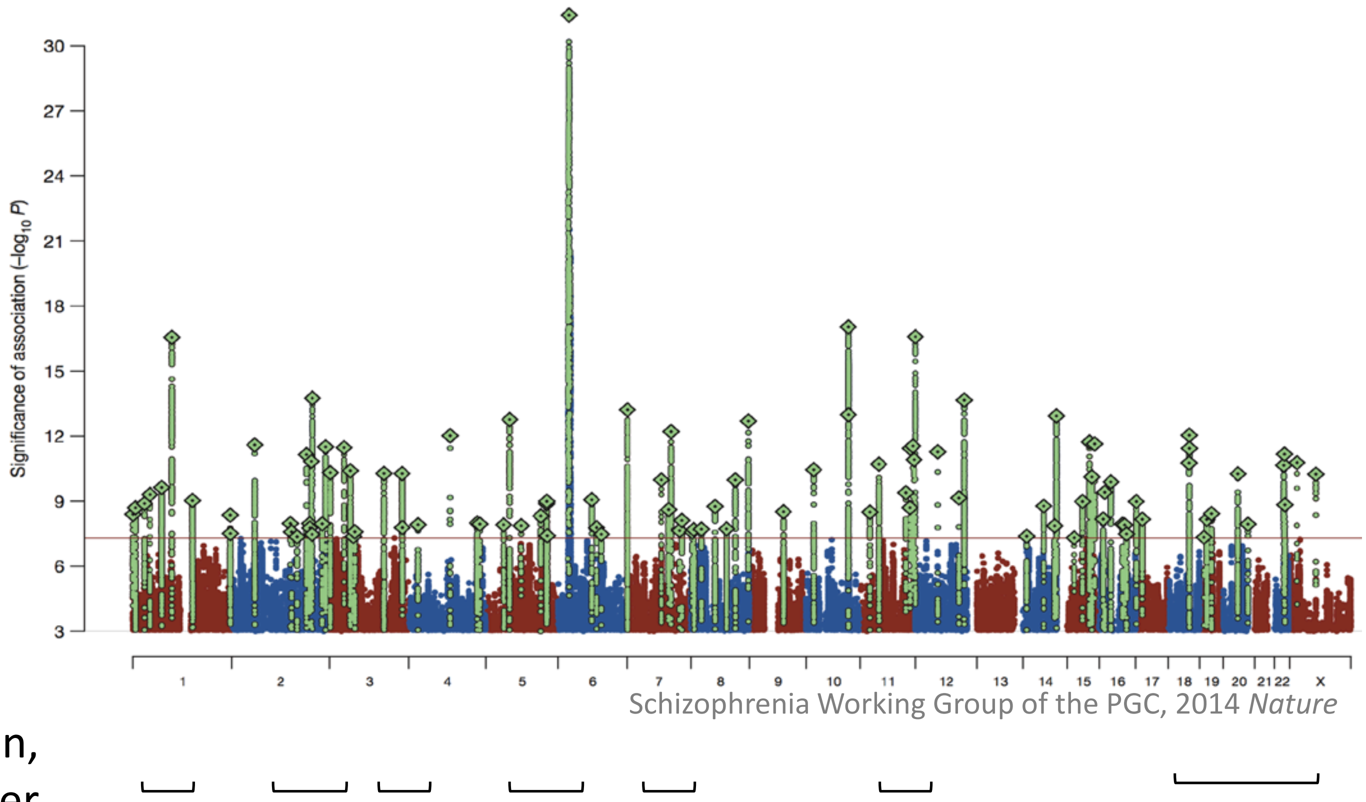


Schizophrenia Working Group of the PGC, 2014 *Nature*

Genome annotation,
e.g. H3K27ac in Cortex

Common approach: e.g., Hu et al. 2011 AJHG, Maurano et al. 2012 *Science,* Trynka et al. 2013, Pickrell 2014 *AJHG*, Kichaev et al. 2014 *PLoS Gen,* Gusev et al. 2014 *AJHG,* Pers et al. 2015 *Nat Commun*, Marbach et al. 2016 *Nat Methods,* Shooshtari et al. 2016 *bioRxiv*, Sarkar et al. 2016 *bioRxiv*, Lu et al. 2016 *bioRxiv*, Iotchkova et al. 2016 *bioRxiv*

How much of the genetic signal falls in this genome annotation?

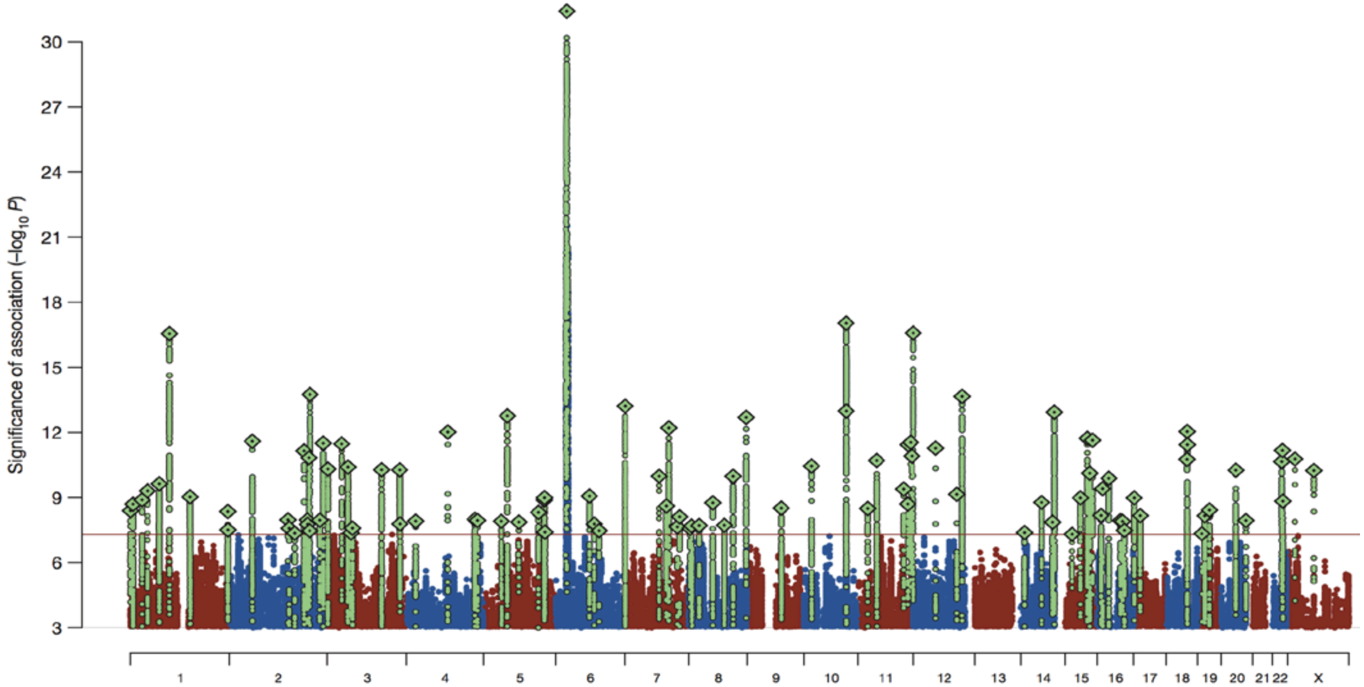


Schizophrenia Working Group of the PGC, 2014 *Nature*

Genome annotation,
e.g. H3K27ac in Liver

Common problems: polygenicity & LD

How much of the genetic signal falls in this genome annotation?



Schizophrenia Working Group of the PGC, 2014 *Nature*

Genome annotation,
e.g. H3K27ac in Liver

Our approach: *leverage polygenicity & LD* by fitting a random effects model from summary statistics.

Random effects models for GWAS leverage polygenicity

- Common approach: Identify causal SNPs, look for patterns
- Random effects: Model SNP effects as random, look at properties of the distribution

Random effects models for GWAS leverage polygenicity

- Common approach: Identify causal SNPs, look for patterns
- Random effects: Model SNP effects as random, look at properties of the distribution
  - Variance (SNP-heritability)
  - Correlation across two traits (genetic correlation)
  - Category-specific variance (partitioning SNP-h^2)

Yang et al. 2010 Nat Genet
Yang et al. 2011 Nat Genet
Lee et al. 2012 Bioinformatics
Lee et al. 2012 Nat Genet
Vattikuti et al. 2012 PLOS Gen
Davis et al. 2013 PLOS Genet
CDG-PGC 2013 Nat Genet
Chen et al. 2014 Hum Mol Gen
Gusev et al. 2014 AJHG

Random effects model for genetics

$$\boxed{Y} = \sum_j X_j \beta_j + \epsilon$$

Quantitative phenotype

Random effects model for genetics

$$\boxed{Y} = \sum_j \boxed{X_j}\beta_j + \epsilon$$

Quantitative phenotype

Genotype at SNP j. (0/1/2 valued, standardized to mean 0, variance 1.)

Random effects model for genetics

$$\boxed{Y} = \sum_j \boxed{X_j} \beta_j + \boxed{\epsilon}$$

Quantitative phenotype

Genotype at SNP j.

Noise and environmental factors. Random, mean-0, independent across individuals.

Random effects model for genetics

$$\boxed{Y} = \sum_j \boxed{X_j}\boxed{\beta_j} + \boxed{\epsilon}$$

Quantitative phenotype

Genotype at SNP j.

Noise and environmental factors.

Effect size of SNP j. Random with mean 0.

Random effects model for genetics

$$\boxed{Y} = \sum_j \boxed{X_j}\boxed{\beta_j} + \boxed{\epsilon}$$

Quantitative phenotype

Genotype at SNP j.

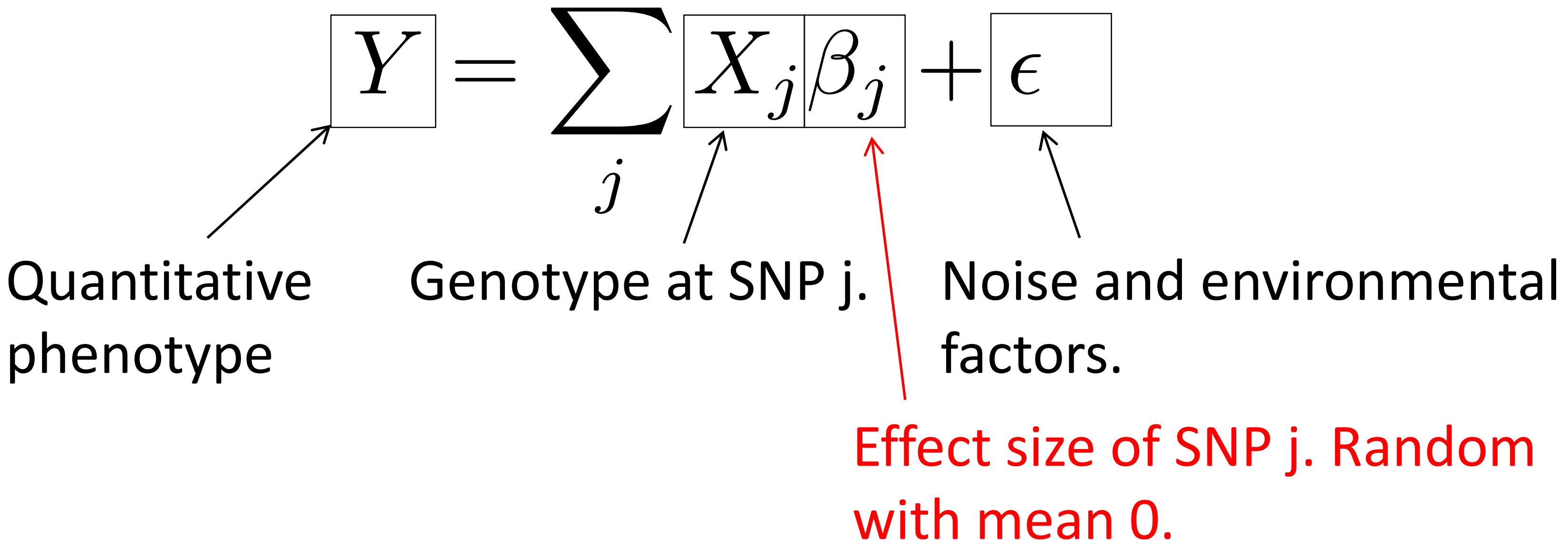
Noise and environmental factors.

Effect size of SNP j. Random with mean 0.

Possibilities for variance:

- One variance for all SNPs
- One variance for all SNPs, correlation across traits
- Variance for SNP j depends on category of SNP j

Random effects model for genetics



$$\boxed{Y} = \sum_j \boxed{X_j}\boxed{\beta_j} + \boxed{\epsilon}$$

Quantitative phenotype

Genotype at SNP j.

Noise and environmental factors.

Effect size of SNP j. Random with mean 0.

<u>Possibilities for variance</u>:

- One variance for all SNPs
- One variance for all SNPs, correlation across traits
- **Variance for SNP j depends on category of SNP j**

Random effects model for genetics

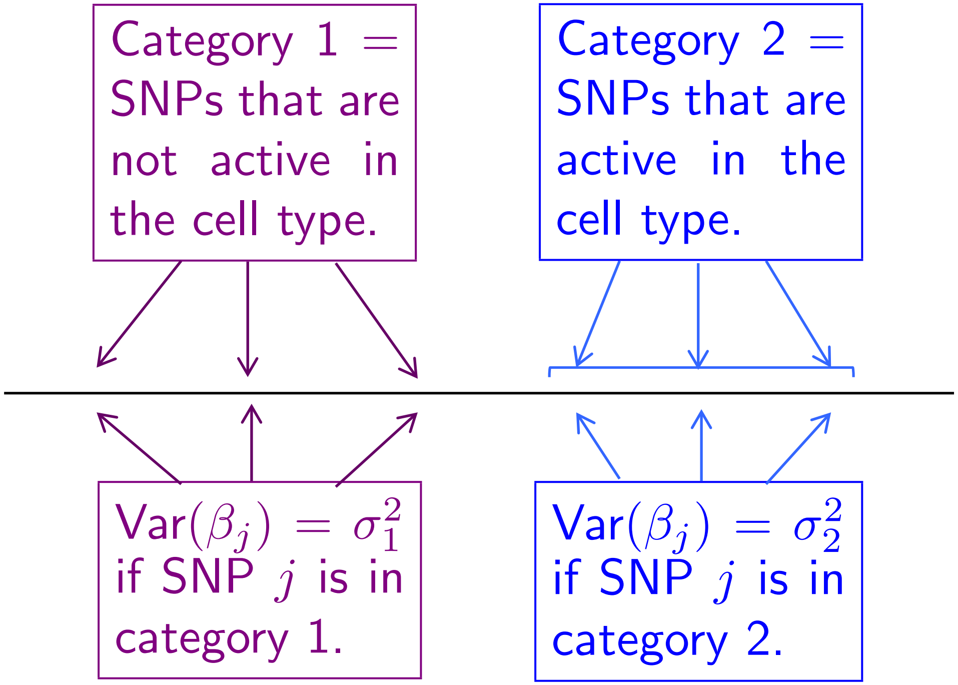
$$\boxed{Y} = \sum_j \boxed{X_j}\boxed{\beta_j} + \boxed{\epsilon\quad}$$

Quantitative phenotype

Genotype at SNP j.

Noise and environmental factors.

Effect size of SNP j. Random with mean 0. Variance depends on which category it belongs to.



Category 1 = SNPs that are not active in the cell type.

Category 2 = SNPs that are active in the cell type.

$\text{Var}(\beta_j) = \sigma_1^2$ if SNP j is in category 1.

$\text{Var}(\beta_j) = \sigma_2^2$ if SNP j is in category 2.

Random effects model for genetics

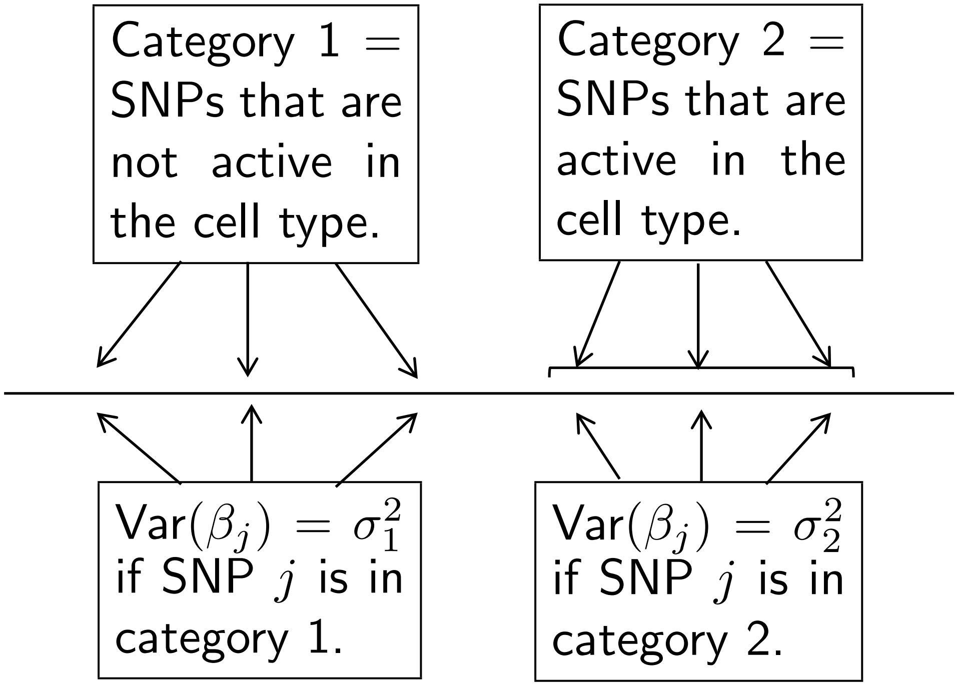
$$\boxed{Y} = \sum_j \boxed{X_j}\boxed{\beta_j} + \boxed{\epsilon}$$

Quantitative phenotype

Genotype at SNP j.

Noise and environmental factors.

Effect size of SNP j. Random with mean 0. Variance depends on which category it belongs to.



Category 1 = SNPs that are not active in the cell type.

Category 2 = SNPs that are active in the cell type.

$\mathrm{Var}(\beta_j) = \sigma_1^2$ if SNP j is in category 1.

$\mathrm{Var}(\beta_j) = \sigma_2^2$ if SNP j is in category 2.

The goal: infer σ_1^2 and σ_2^2.

This model has been used to identify cell types previously

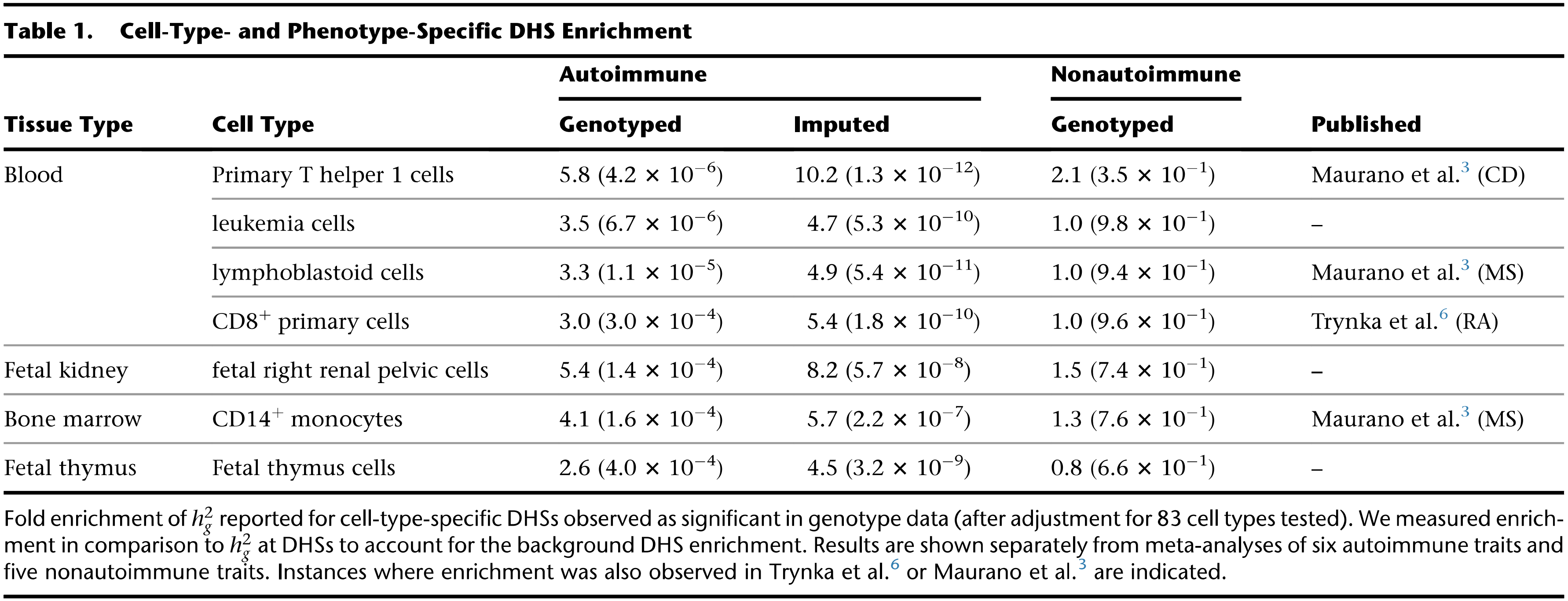


Table 1. Cell-Type- and Phenotype-Specific DHS Enrichment

		Autoimmune		Nonautoimmune	
Tissue Type	**Cell Type**	**Genotyped**	**Imputed**	**Genotyped**	**Published**
Blood	Primary T helper 1 cells	5.8 (4.2×10^{-6})	10.2 (1.3×10^{-12})	2.1 (3.5×10^{-1})	Maurano et al.[3] (CD)
	leukemia cells	3.5 (6.7×10^{-6})	4.7 (5.3×10^{-10})	1.0 (9.8×10^{-1})	–
	lymphoblastoid cells	3.3 (1.1×10^{-5})	4.9 (5.4×10^{-11})	1.0 (9.4×10^{-1})	Maurano et al.[3] (MS)
	$CD8^+$ primary cells	3.0 (3.0×10^{-4})	5.4 (1.8×10^{-10})	1.0 (9.6×10^{-1})	Trynka et al.[6] (RA)
Fetal kidney	fetal right renal pelvic cells	5.4 (1.4×10^{-4})	8.2 (5.7×10^{-8})	1.5 (7.4×10^{-1})	–
Bone marrow	$CD14^+$ monocytes	4.1 (1.6×10^{-4})	5.7 (2.2×10^{-7})	1.3 (7.6×10^{-1})	Maurano et al.[3] (MS)
Fetal thymus	Fetal thymus cells	2.6 (4.0×10^{-4})	4.5 (3.2×10^{-9})	0.8 (6.6×10^{-1})	–

Fold enrichment of h_g^2 reported for cell-type-specific DHSs observed as significant in genotype data (after adjustment for 83 cell types tested). We measured enrichment in comparison to h_g^2 at DHSs to account for the background DHS enrichment. Results are shown separately from meta-analyses of six autoimmune traits and five nonautoimmune traits. Instances where enrichment was also observed in Trynka et al.[6] or Maurano et al.[3] are indicated.

A challenge: sample size

Gusev et al. 2014 AJHG; see also Lee et al. 2012 Nat Genet and Davis et al. 2013 PLoS Genet

Stratified LD score regression fits this model from summary statistics

Why?

- For meta-analyses, no one has all of the genotypes.
- Lots of publicly available summary statistics.
- Existing methods are computationally expensive.

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Fitting the model from summary statistics: What are summary statistics?

- In a GWAS, test for positive *marginal correlation.*

$$\chi_j^2 \approx (\vec{X}_j \cdot \vec{Y})^2/N$$

- Reflects causal effects of SNP j **and SNPs in LD with SNP j**

Fitting the model from summary statistics: What is $E[\chi_j^2]$?

Without LD:

$$E[\chi_j^2|\beta] = 1 + N\beta_j^2$$

$$E[\chi_j^2] = 1 + NE[\beta_j^2]$$

$E[\chi_j^2] = 1 + N\sigma_C^2$ for $j \in C$.

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Fitting the model from summary statistics: We can use regression!

$$E[\chi_j^2] \approx 1 + N \sum_C \sigma_C^2 \ell(j, C)$$

To estimate σ_C^2:

- Estimate LD Scores from a reference panel.
- Regress chi-square statistics on LD Scores.

Details:

- Significance via jackknife
- Weighted regression

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Chi-square is linear in LD score

With only one category:

$$E[\chi_j^2] \approx 1 + N\ell(j, \text{all SNPs})$$

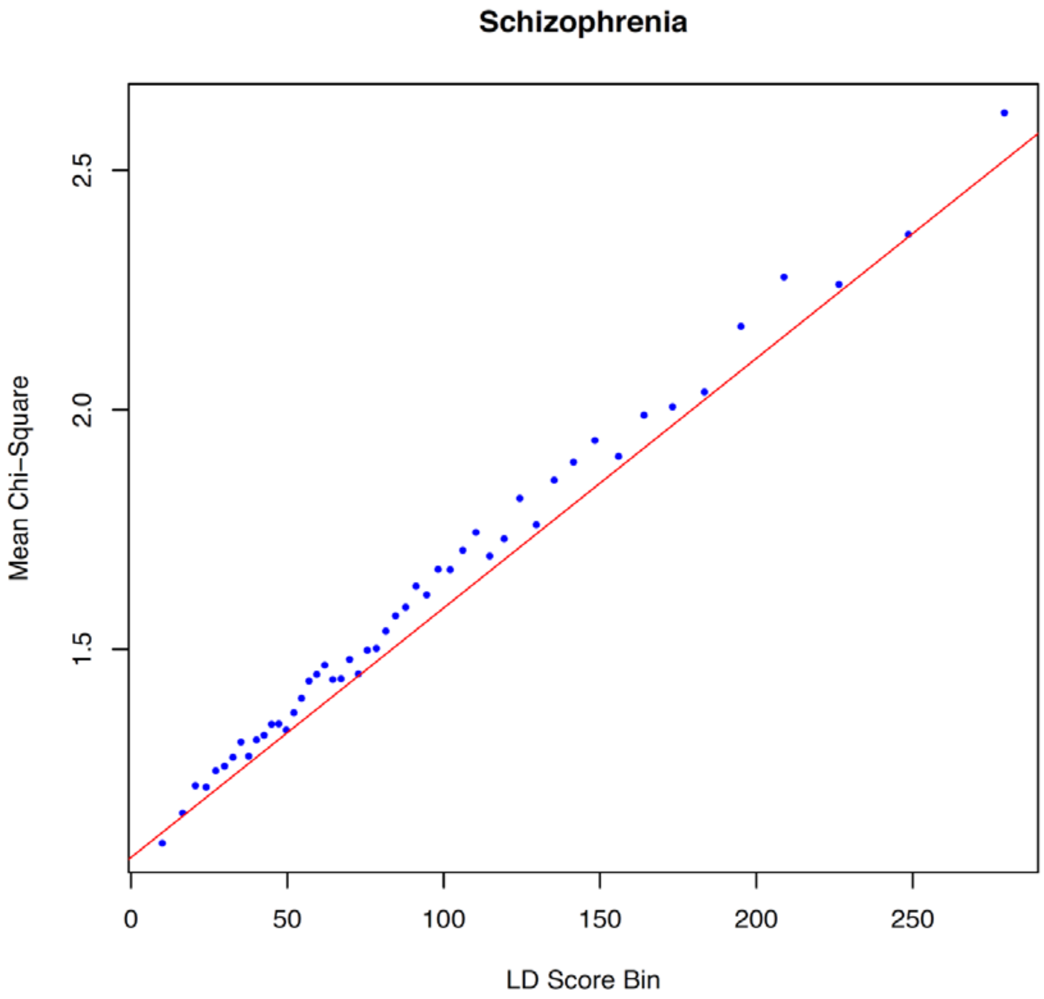
[Bulik-Sullivan et al. 2015 Nat Genet]

Chi-square is linear in LD score

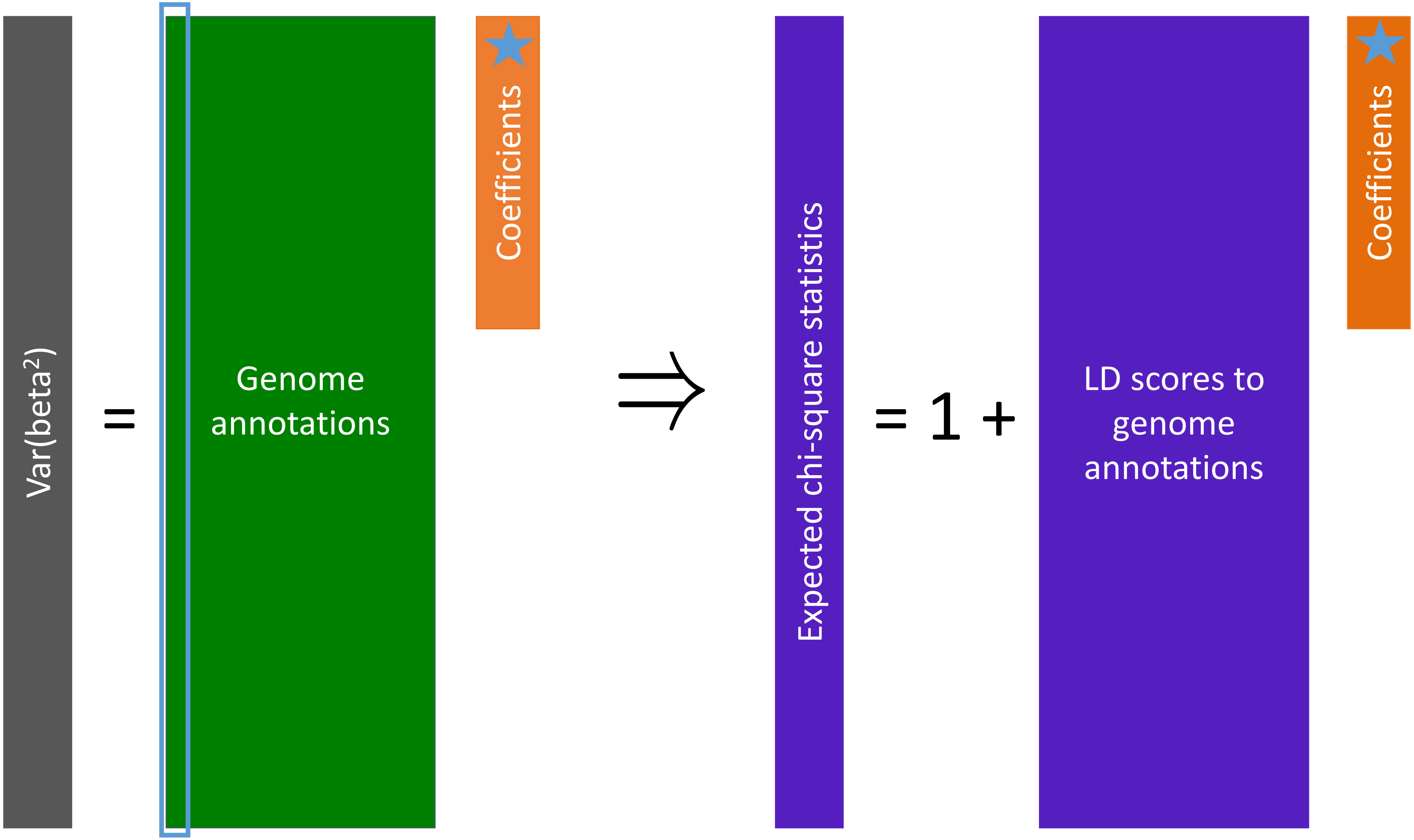
With only one category:

$$E[\chi_j^2] \approx 1 + N\ell(j, \text{all SNPs})$$

[Bulik-Sullivan et al. 2015 Nat Genet]

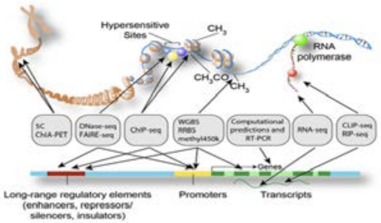


SCZ working group of the PGC 2014 Nature (data)
Bulik-Sullivan et al 2015 Nat Genet (LD Score plot)

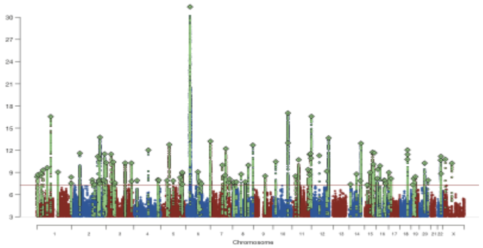


Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Genome annotation of interest
(e.g., DHS peaks in liver)



Summary statistics for trait of interest



Other annotations to control for:

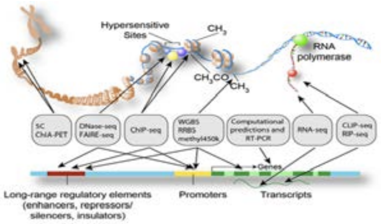
- Exons
- Promoters
- Repressed regions
- Conserved regions
- Etc…

Stratified LD score regression

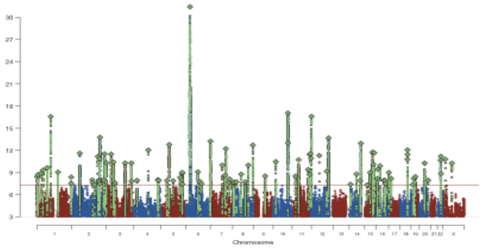
Heritability enrichment of the **annotation of interest**, controlling for **other annotations**

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Genome annotation for
Central Nervous System



Summary statistics for
Schizophrenia



Other annotations to
control for:

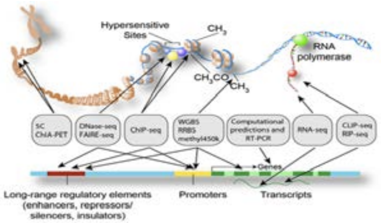
- Exons
- Promoters
- Repressed regions
- Conserved regions
- Etc…

Stratified LD score regression

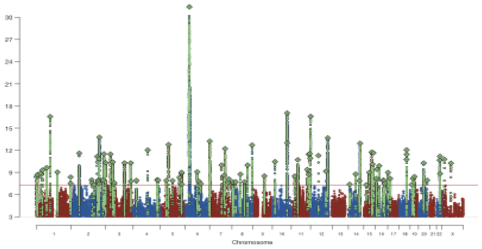
P-value for **Central Nervous System**
enrichment for **Schizophrenia**

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Genome annotation for **Cardiovascular System**



Summary statistics for **Schizophrenia**



Other annotations to control for:

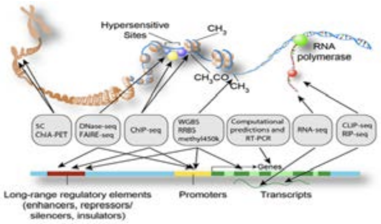
- Exons
- Promoters
- Repressed regions
- Conserved regions
- Etc…

Stratified LD score regression

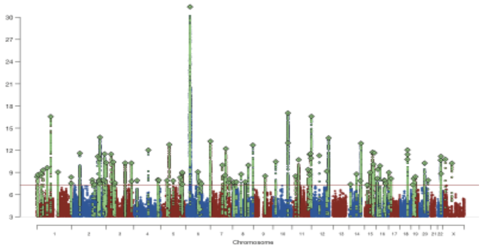
P-value for **Cardiovascular system** enrichment for **Schizophrenia**

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

Genome annotation for **Liver**



Summary statistics for **Schizophrenia**



Other annotations to control for:

- Exons
- Promoters
- Repressed regions
- Conserved regions
- Etc…

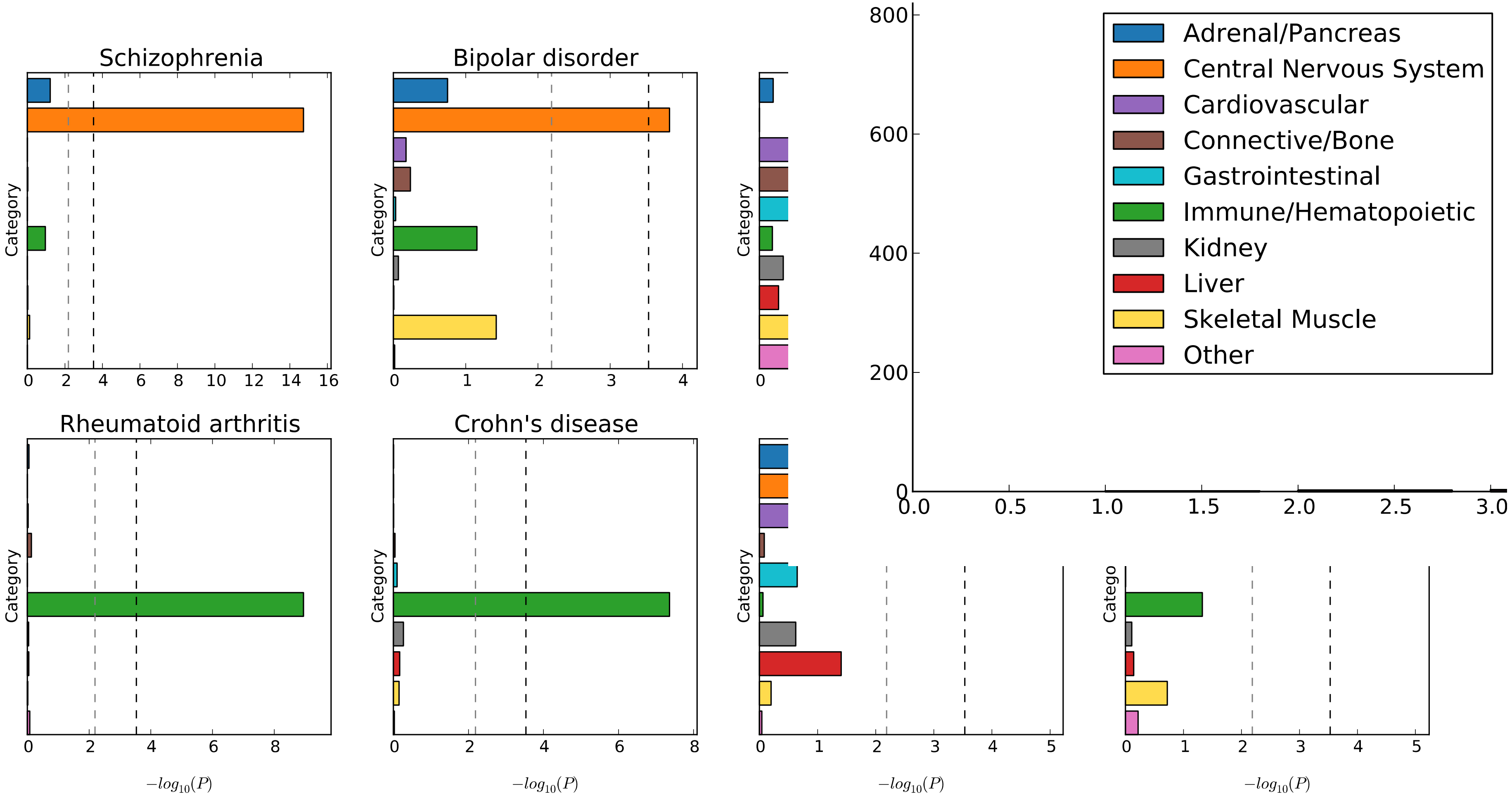
Stratified LD score regression

P-value for **Liver** enrichment for **Schizophrenia**

Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

S-LDSC identifies relevant tissues

- 10 annotations using histone marks from ENCODE/Roadmap
- 17 phenotypes with publicly available GWAS summary statistics



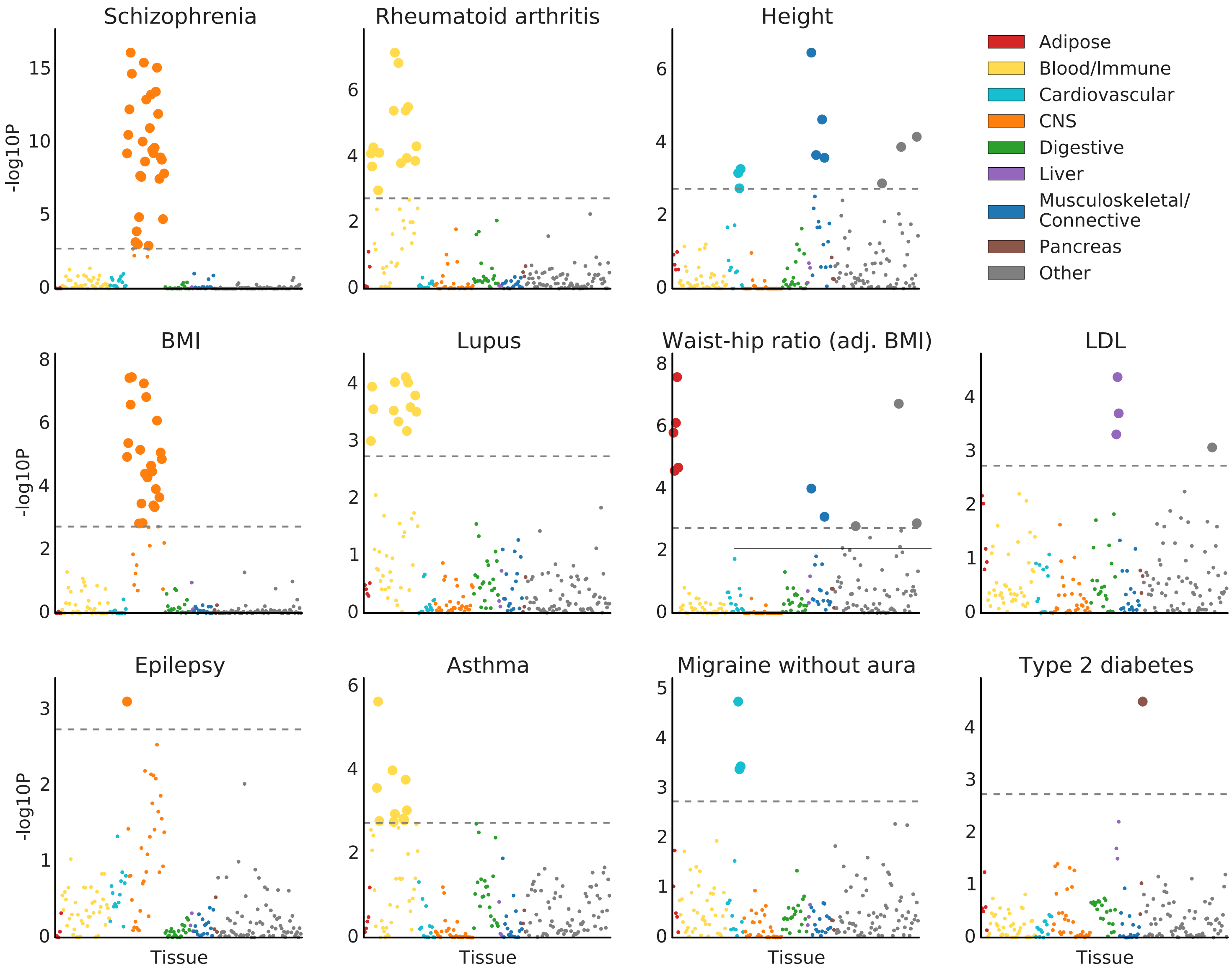
Finucane*, Bulik-Sullivan*, et al. 2015 *Nature Genetics*

We can also use gene expression

- **Genome annotation of interest**:
    - Rank genes by specific expression
    - Take top 10% of genes
    - Add 100kb window
- **Annotation data**:
    - GTEx project
    - Public dataset from Franke lab
- **GWAS data**: 48 GWAS, avg N =86,850
    - Public data
    - Brainstorm consortium
    - UK Biobank

Finucane et al. 2018 *Nature Genetics*

We can also use gene expression data

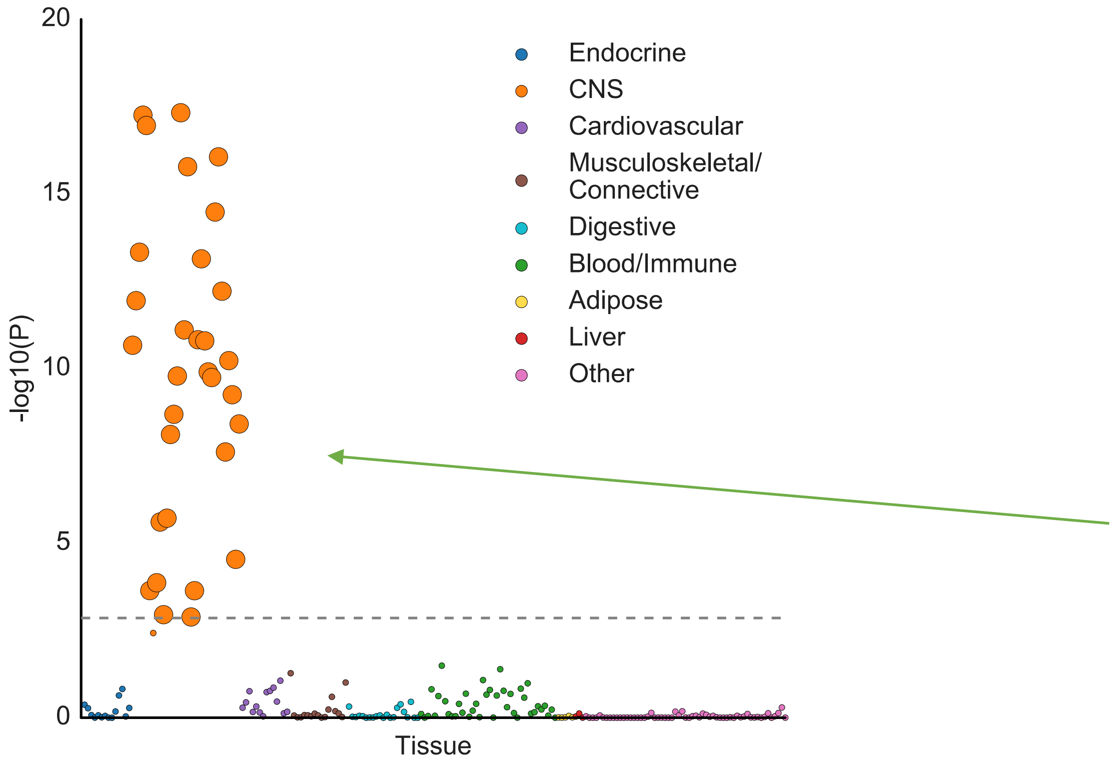


(Large dot = FDR < 5%)

Finucane et al. 2018 *Nature Genetics*

Zooming in Part 1: the brain

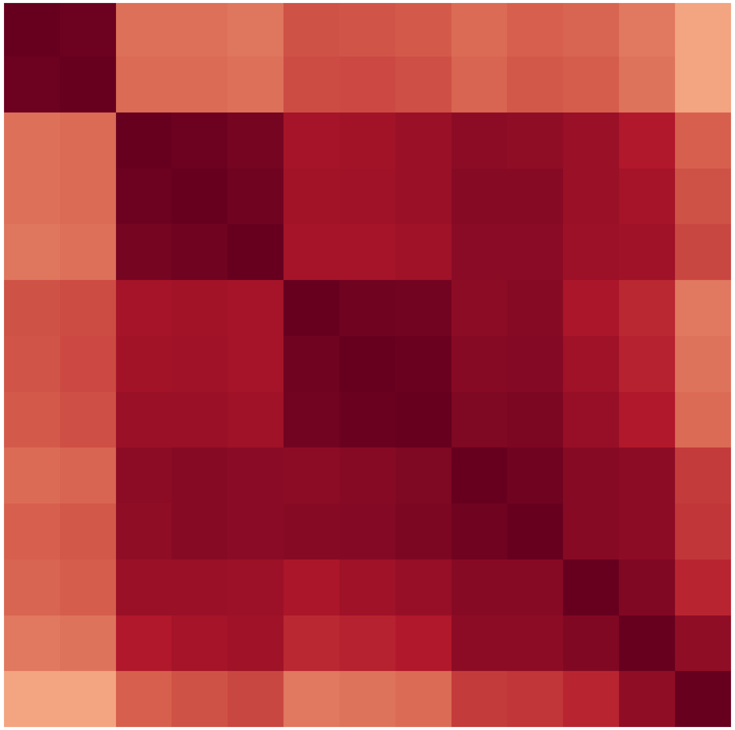
Schizophrenia, multi-tissue analysis



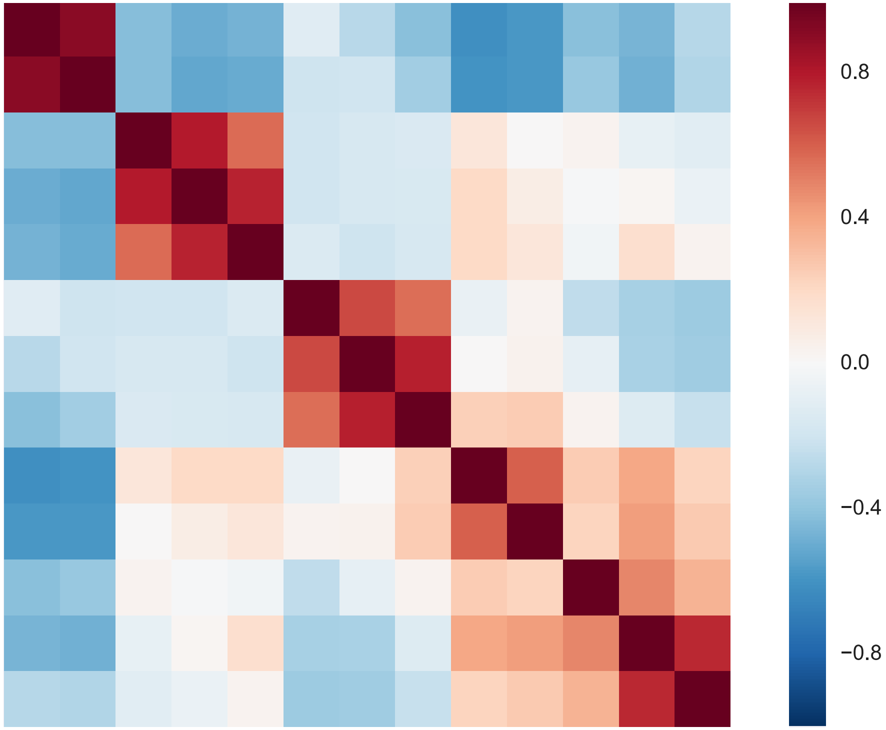
Almost every CNS annotation passes FDR < 5%.

Finucane et al. 2018 *Nature Genetics*

Differential expression within brain differentiates brain regions



Correlations among brain region LD scores: **Multi-tissue analysis**



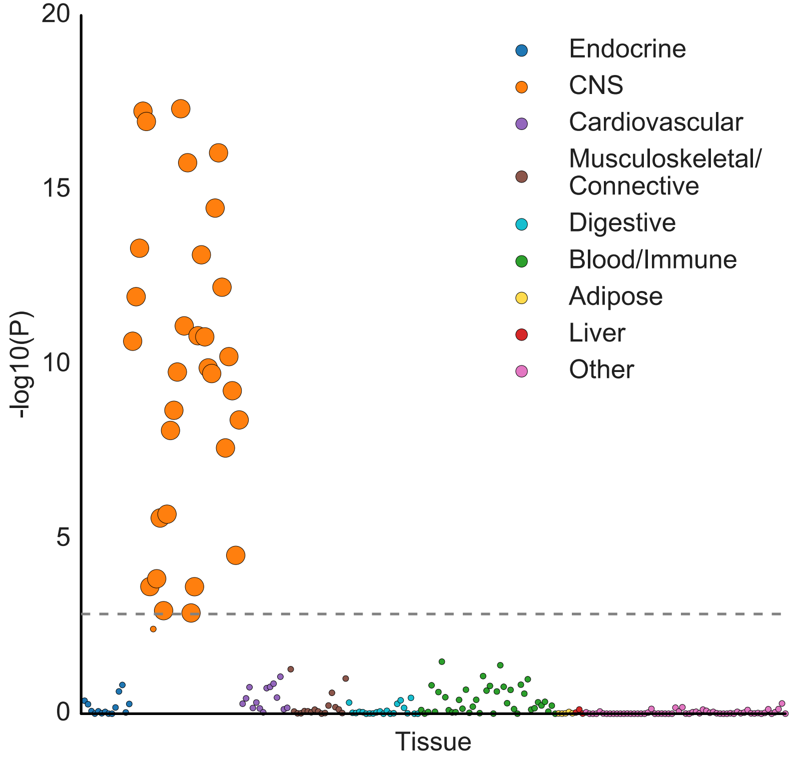
Correlations among brain region LD scores: **Within-brain analysis**

Data: GTEx

Finucane et al. 2018 *Nature Genetics*

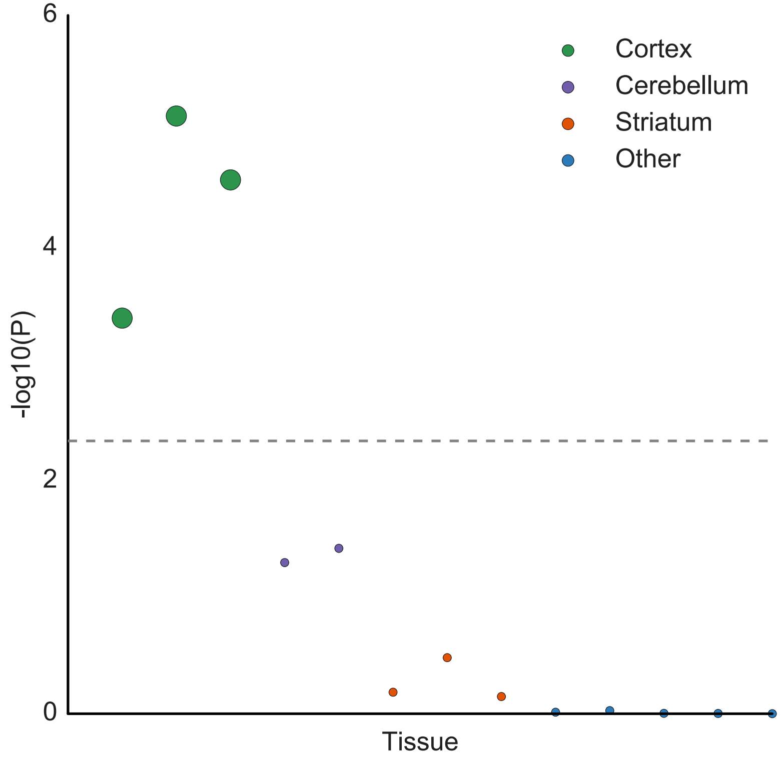
Differential expression within brain differentiates brain regions

Schizophrenia, multi-tissue analysis



Data: GTEx, Pers et al. 2015 *Nat Commun*

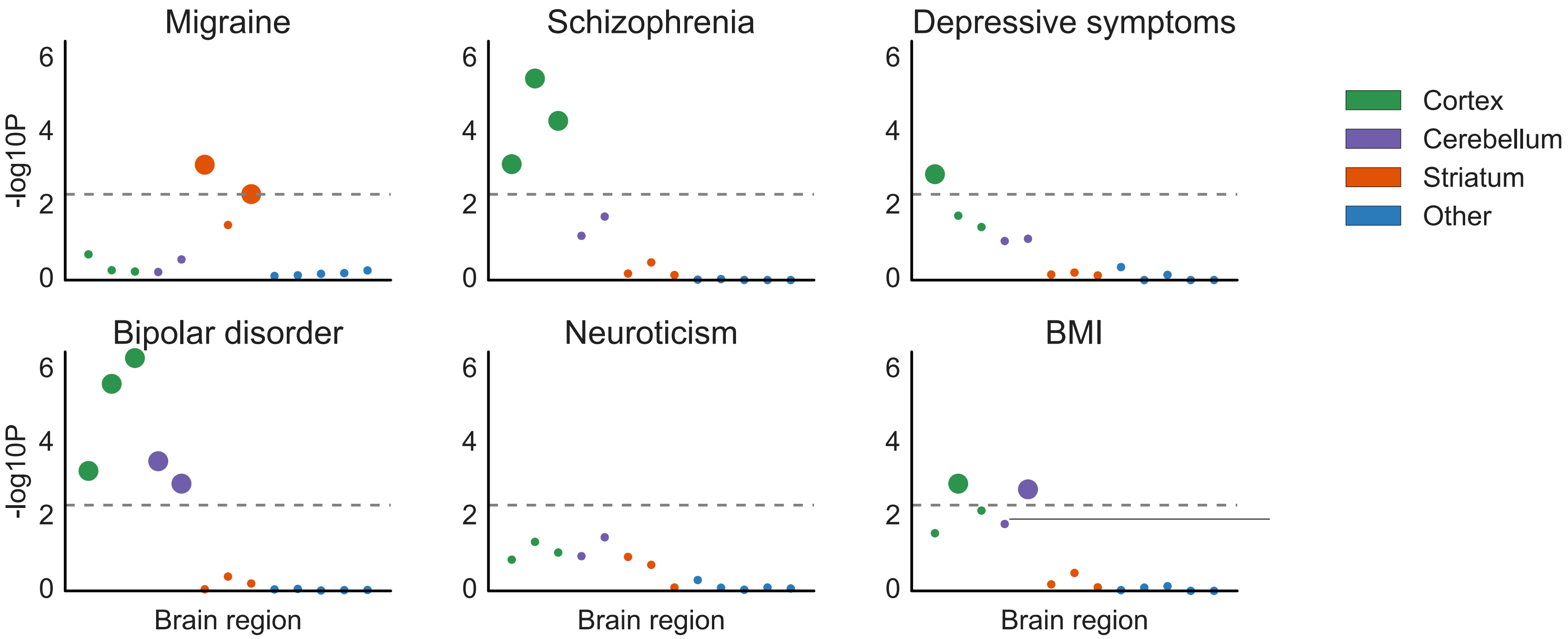
Schizophrenia, GTEx brain only



Data: GTEx

Finucane et al. 2018 *Nature Genetics*

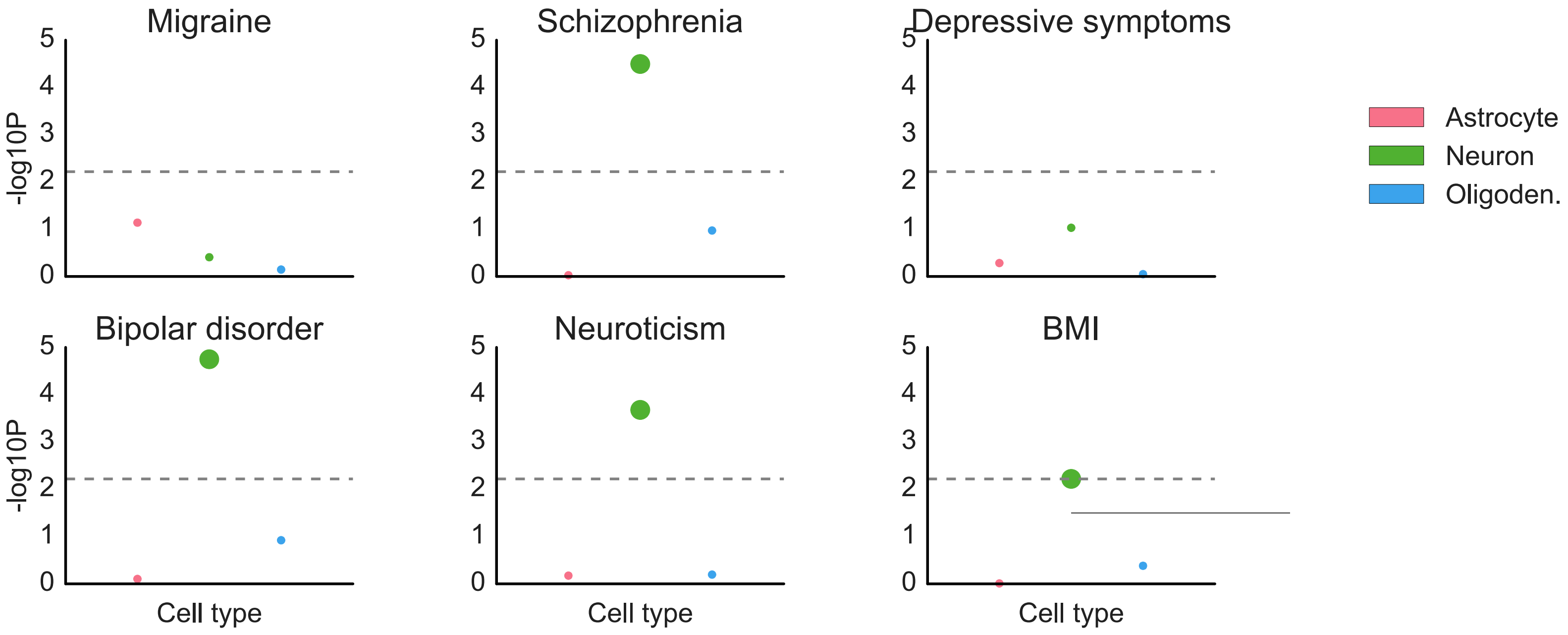
Differential expression within brain differentiates brain regions

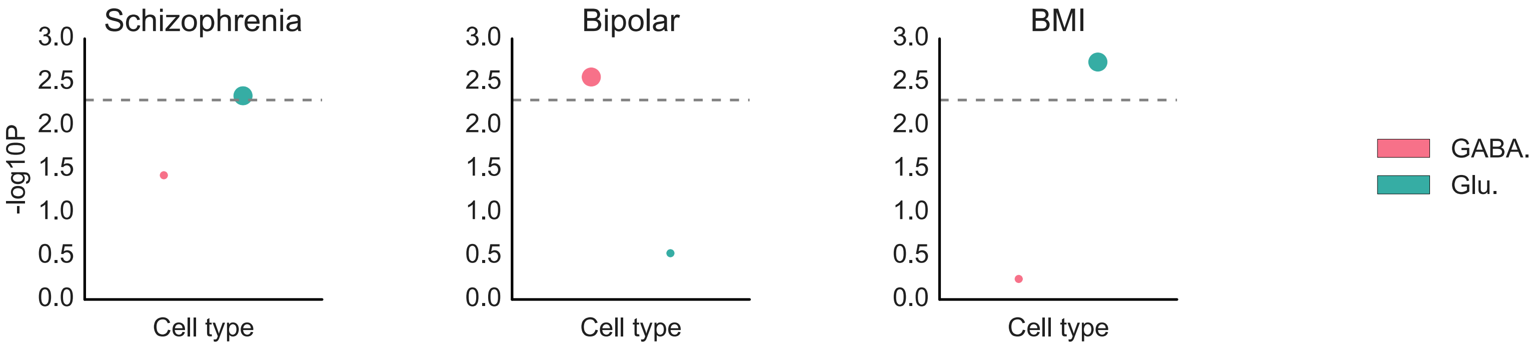


Potential confounder: cell type composition

Finucane et al. 2018 *Nature Genetics*

Differential expression within brain differentiates brain cell types

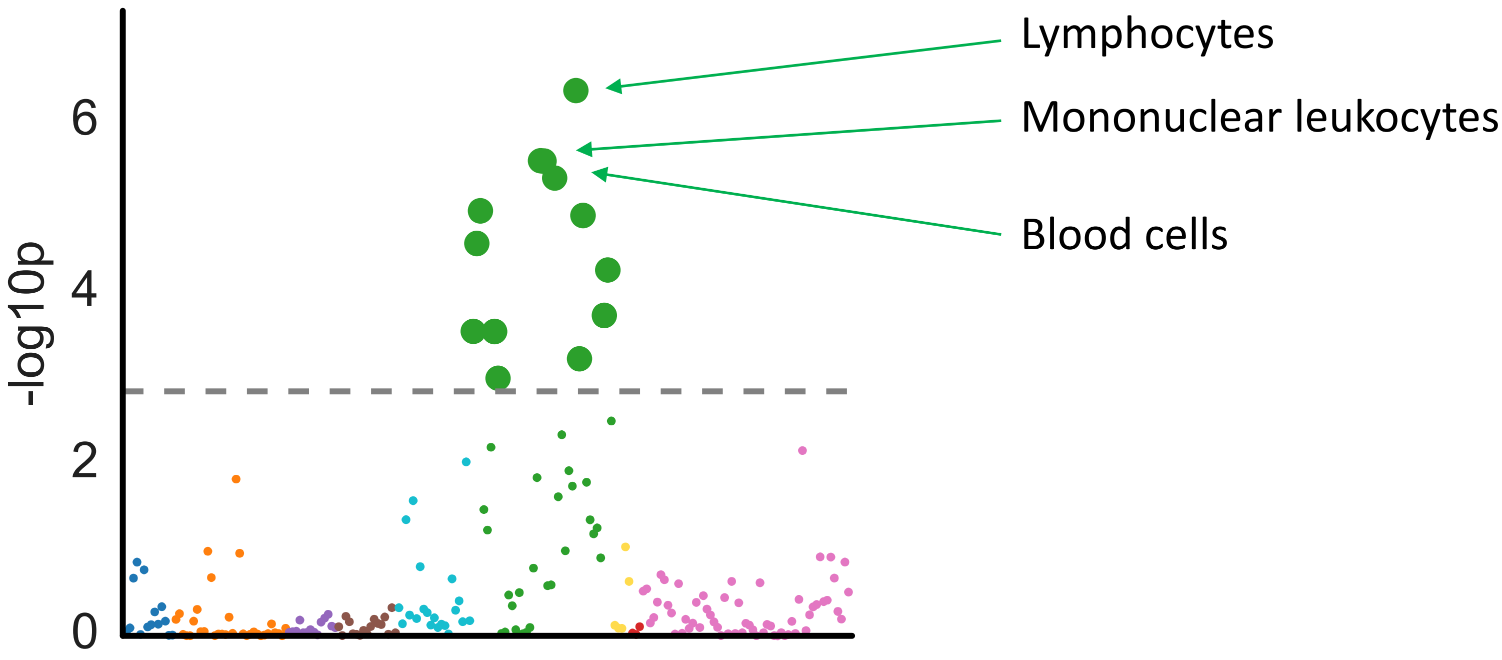




Finucane et al. 2018 *Nature Genetics*

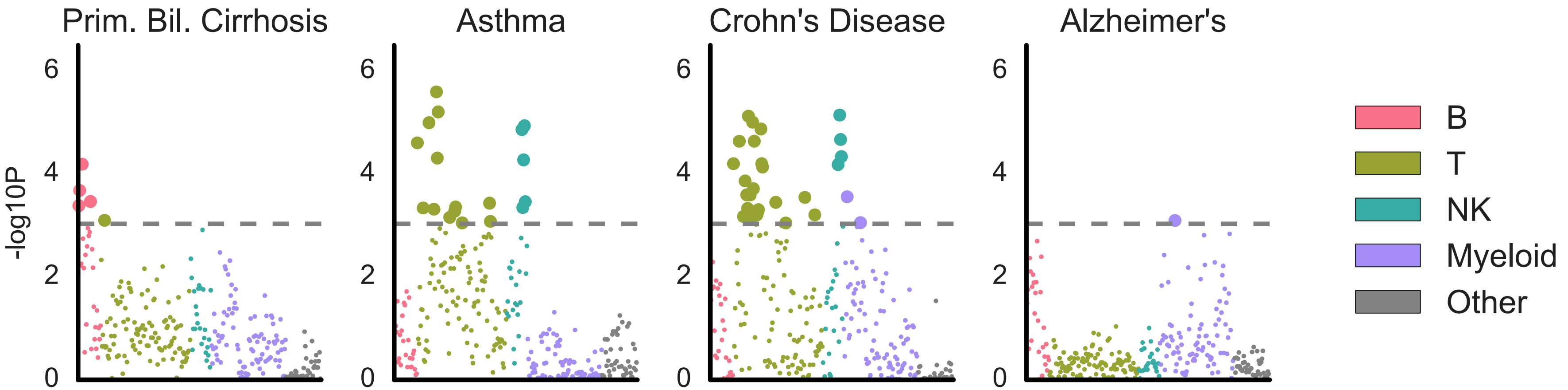
Zooming in Part 2: Blood/Immune

Rheumatoid arthritis, multi-tissue analysis

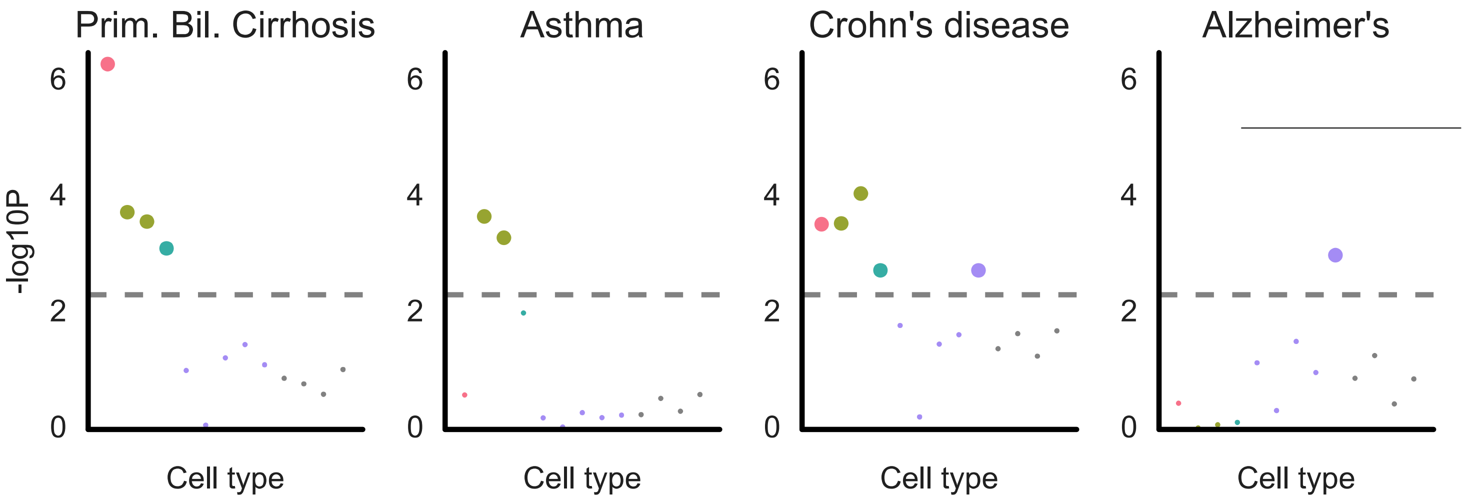


Finucane et al. 2018 *Nature Genetics*

Mouse microarray, 292 immune cell types [Data: ImmGen Consortium]



Human ATACseq, 13 cell types spanning hematopoiesis [Data: Corces et al.]



Finucane et al. 2018 *Nature Genetics*

Acknowledgements

Harvard T.H. Chan School of Public Health:

- **Alkes Price**
- Steven Gazal
- Alexander Gusev
- Sara Lindstrom
- Po-Ru Loh
- Samuela Pollack
- Yakir Reshef

Harvard Medical School:

- Steve McCarroll
- Soumya Raychaudhuri
- Arpiar Saunders
- Kamil Slowikowski

University of Cambridge:

- Felix Day
- John Perry

Broad Institute:

- **Ben Neale**
- **Brendan Bulik-Sullivan**
- Verneri Anttila
- Andrea Byrnes
- Mark Daly
- Kyle Farh
- Giulio Genovese
- Evan Macosko
- Nick Patterson

Consortia:

- Brainstorm Consortium
- GTEx Consortium
- Psychiatric Genomics Consortium
- RACI Consortium
- ReproGen Consortium
- UK Biobank

NIH
National Institutes
of Health

the Hertz
FOUNDATION
freedom to innovate