The Omnigenic model for complex traits and disease

How is phenotypic variation encoded in the human genome?

Yang I Li
Assistant Professor of Medicine
University of Chicago

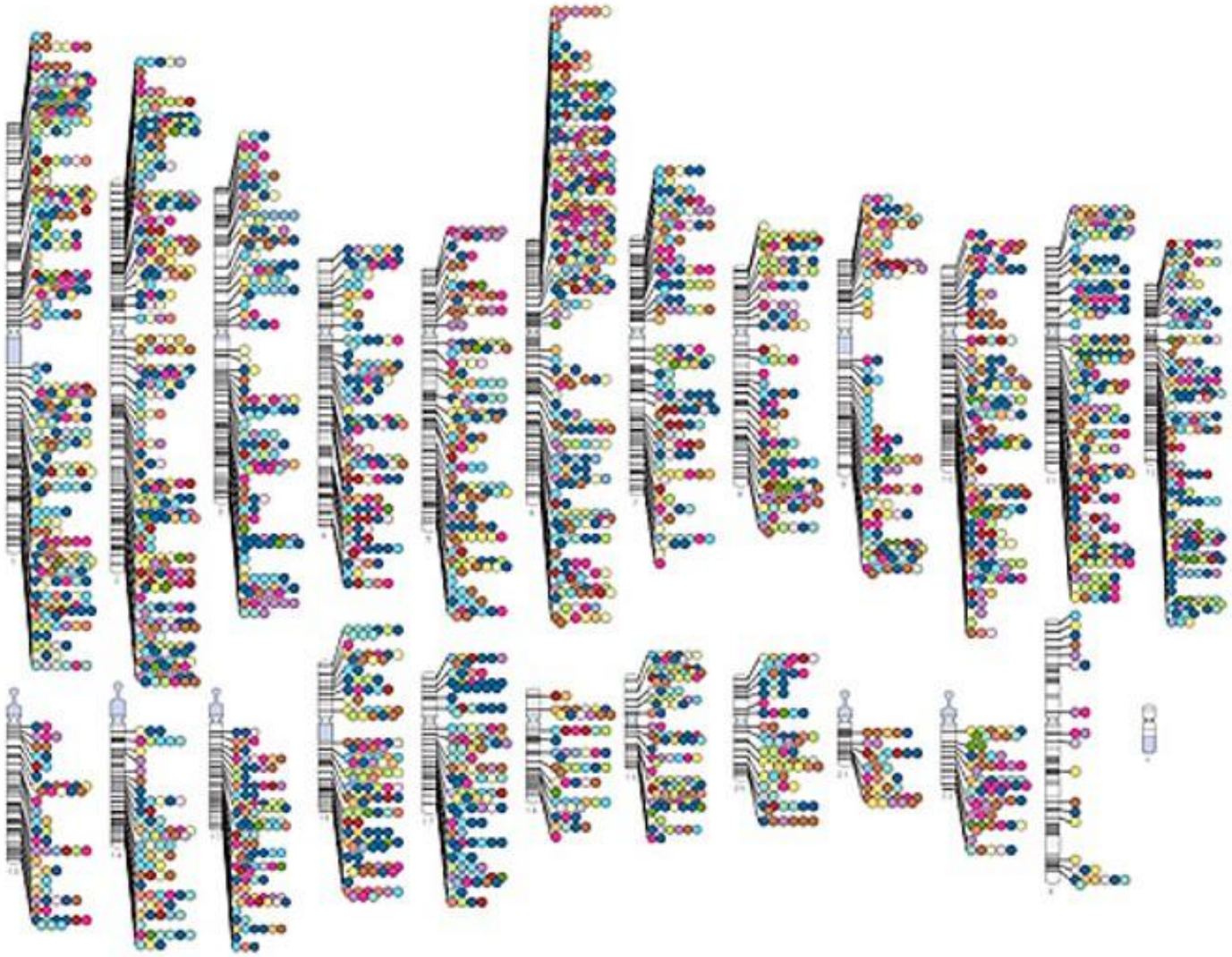
Evan Boyle

Jonathan Pritchard

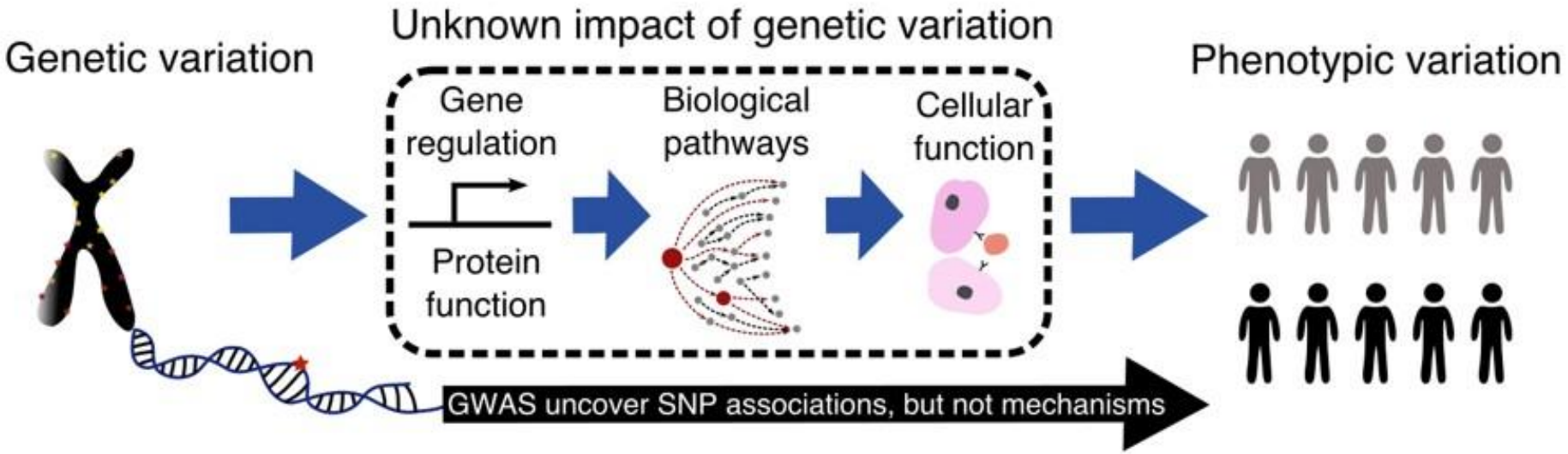
THE UNIVERSITY OF CHICAGO

Stanford University

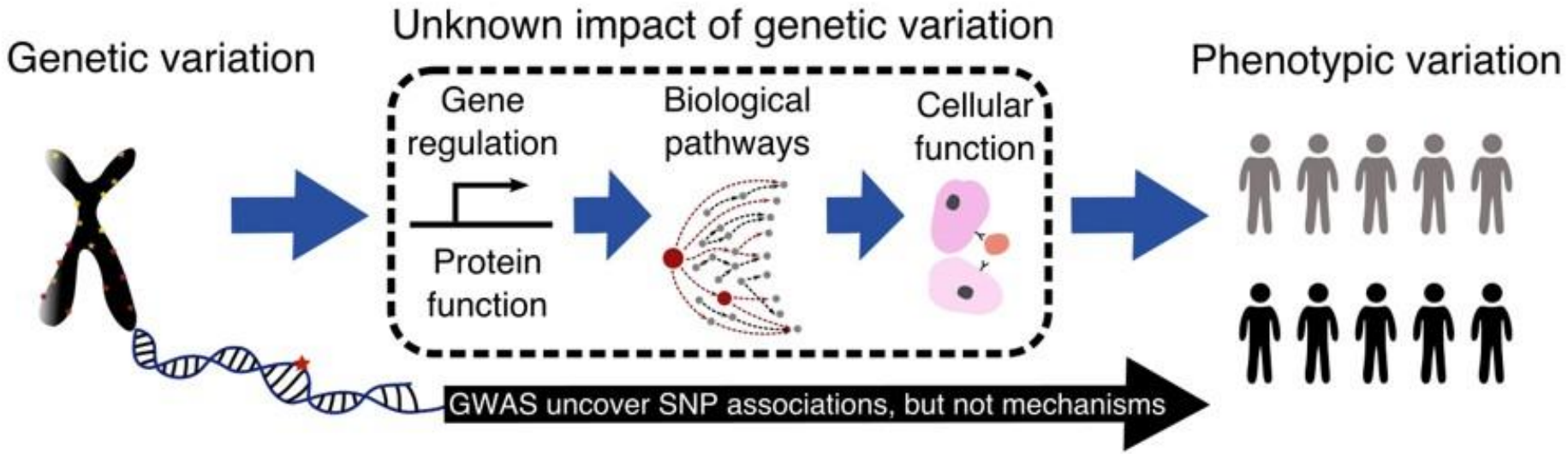
GWAS have been hugely successful



Fundamental question: which genes, pathways, and function do trait associated variants affect?



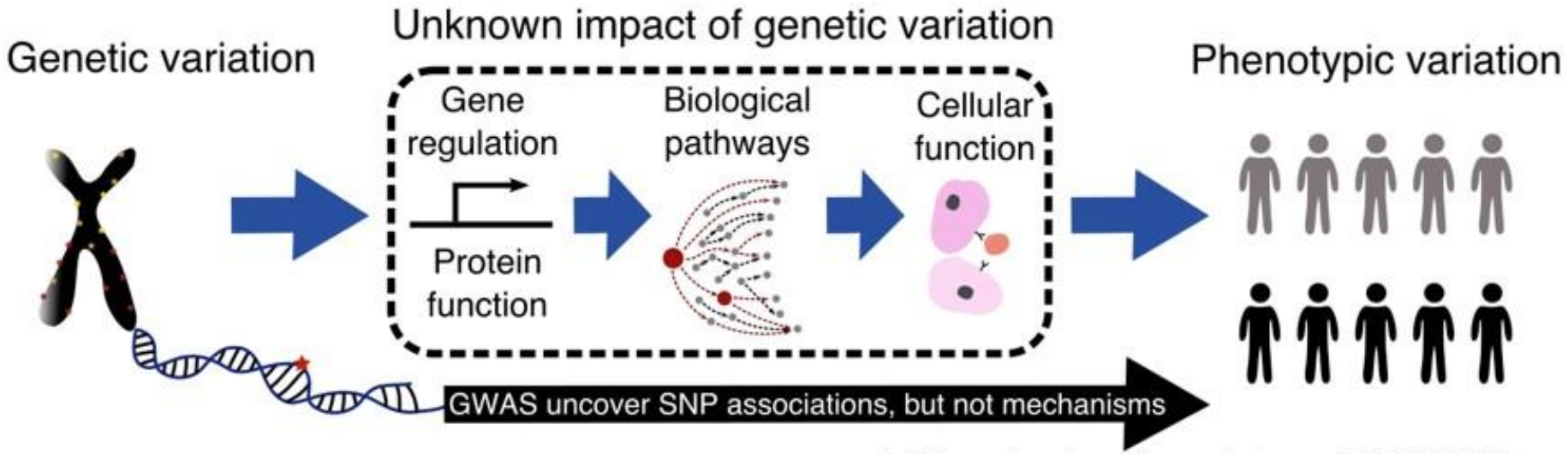
Fundamental question: which genes, pathways, and function do trait associated variants affect?



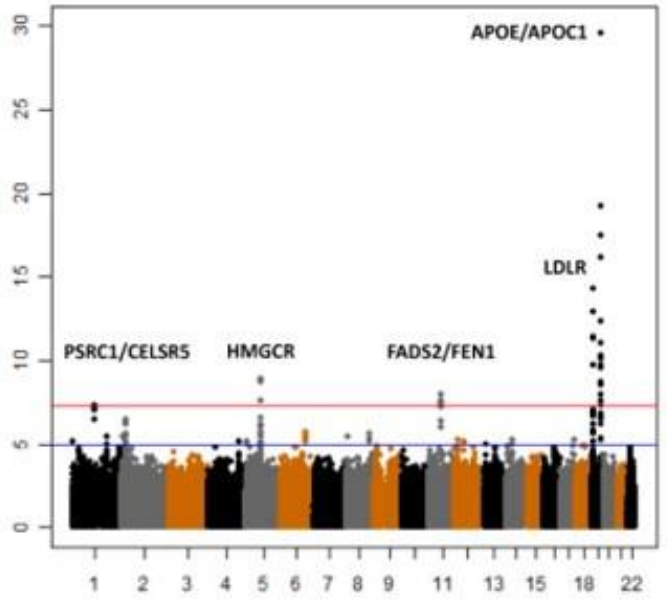
Classical model:

Disease-associated variants affect genes that are directly involved in disease-causing pathways

Fundamental question: which genes, pathways, and function do trait associated variants affect?



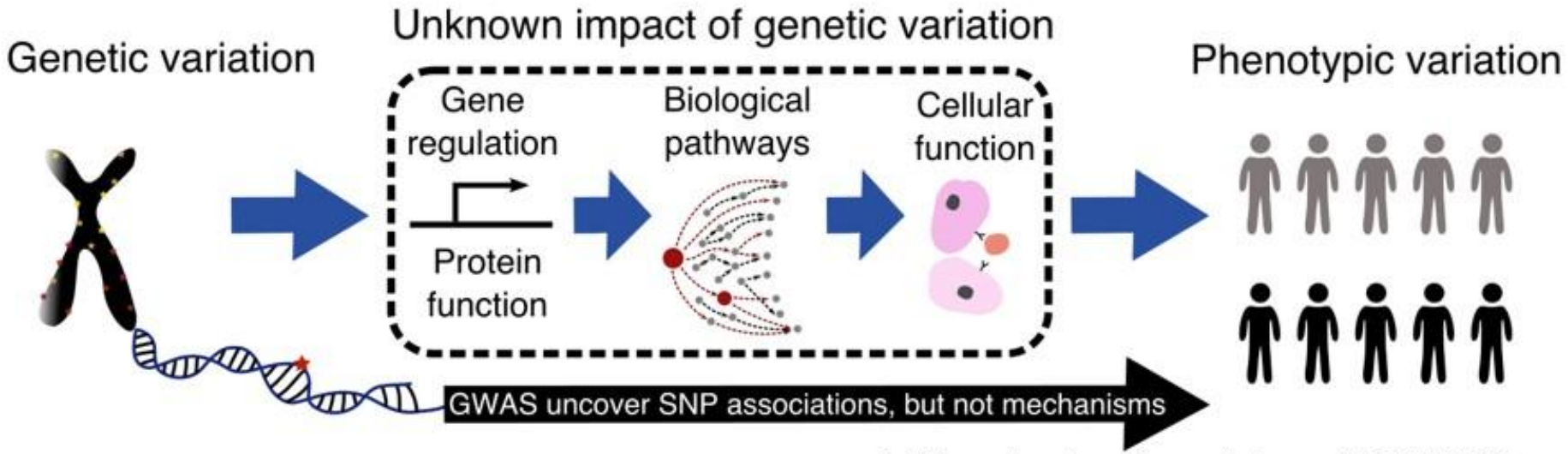
LDL cholesterol level GWAS



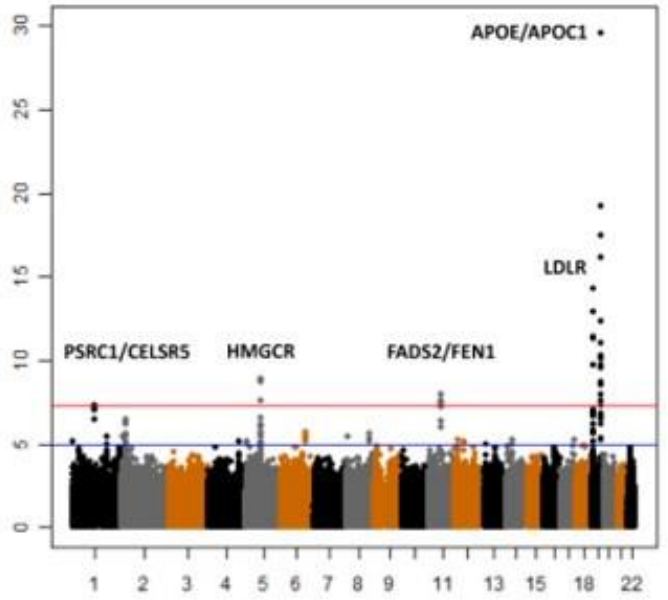
Classical model:

Disease-associated variants affect genes that are directly involved in disease-causing pathways

Fundamental question: which genes, pathways, and function do trait associated variants affect?



LDL cholesterol level GWAS



Classical model:

Disease-associated variants affect genes that are directly involved in disease-causing pathways

All with have to do is to perform bigger GWAS and find which genes are in associated loci?

Today's presentation:

- For typical complex traits, most of the heritability is mediated through genes that are not directly related to the trait

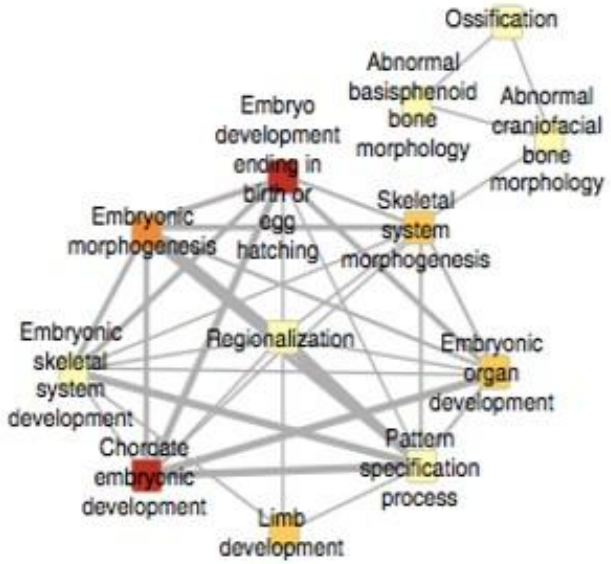
Today's presentation:

- For typical complex traits, most of the heritability is mediated through genes that are not directly related to the trait

- We need new conceptual models for thinking about the molecular processes that link genetic variation to complex phenotypes

Observation #1: For many traits, there is a huge number of causal variants across the genome

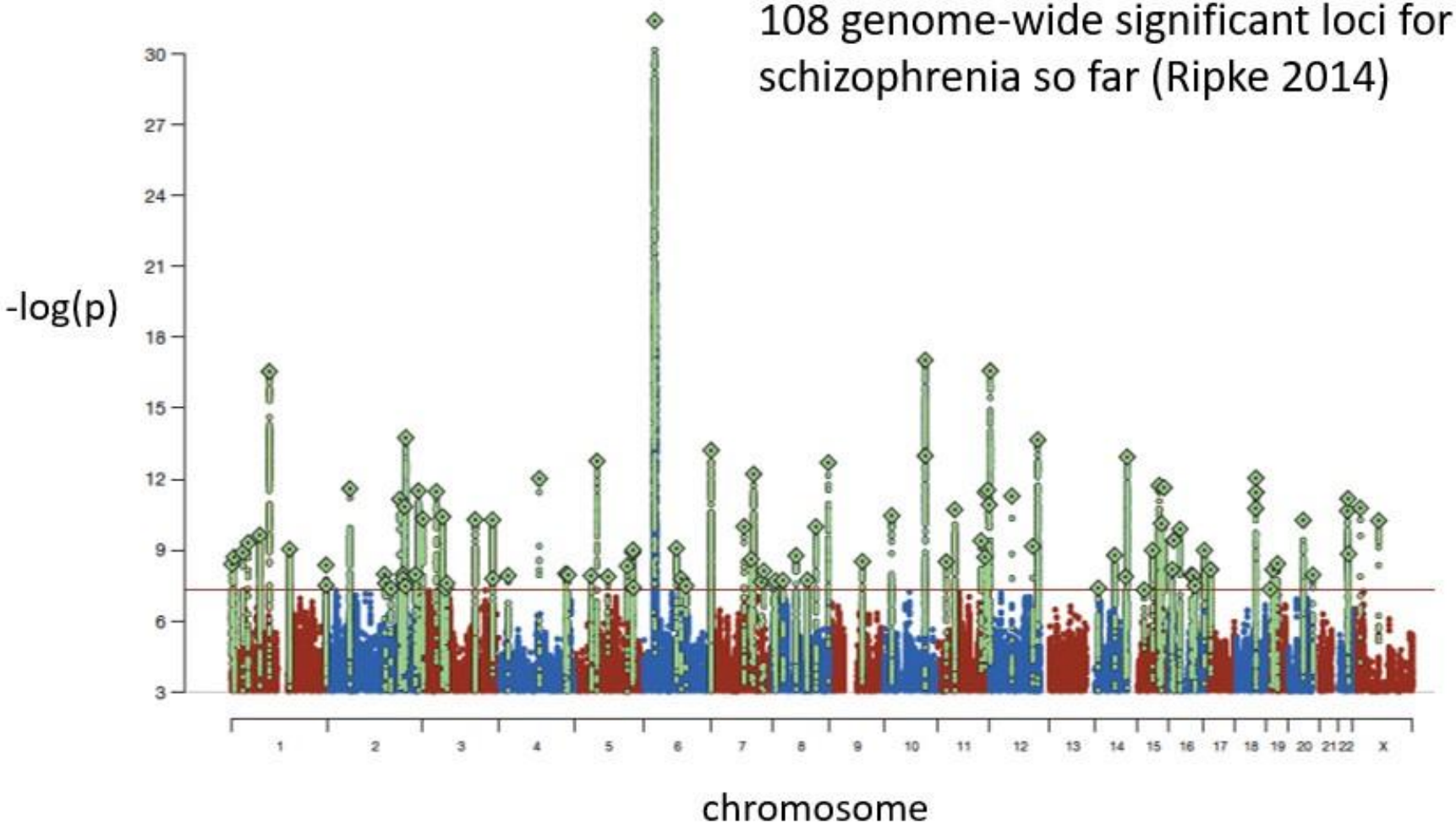
Observation #1: For many traits, there is a huge number of causal variants across the genome



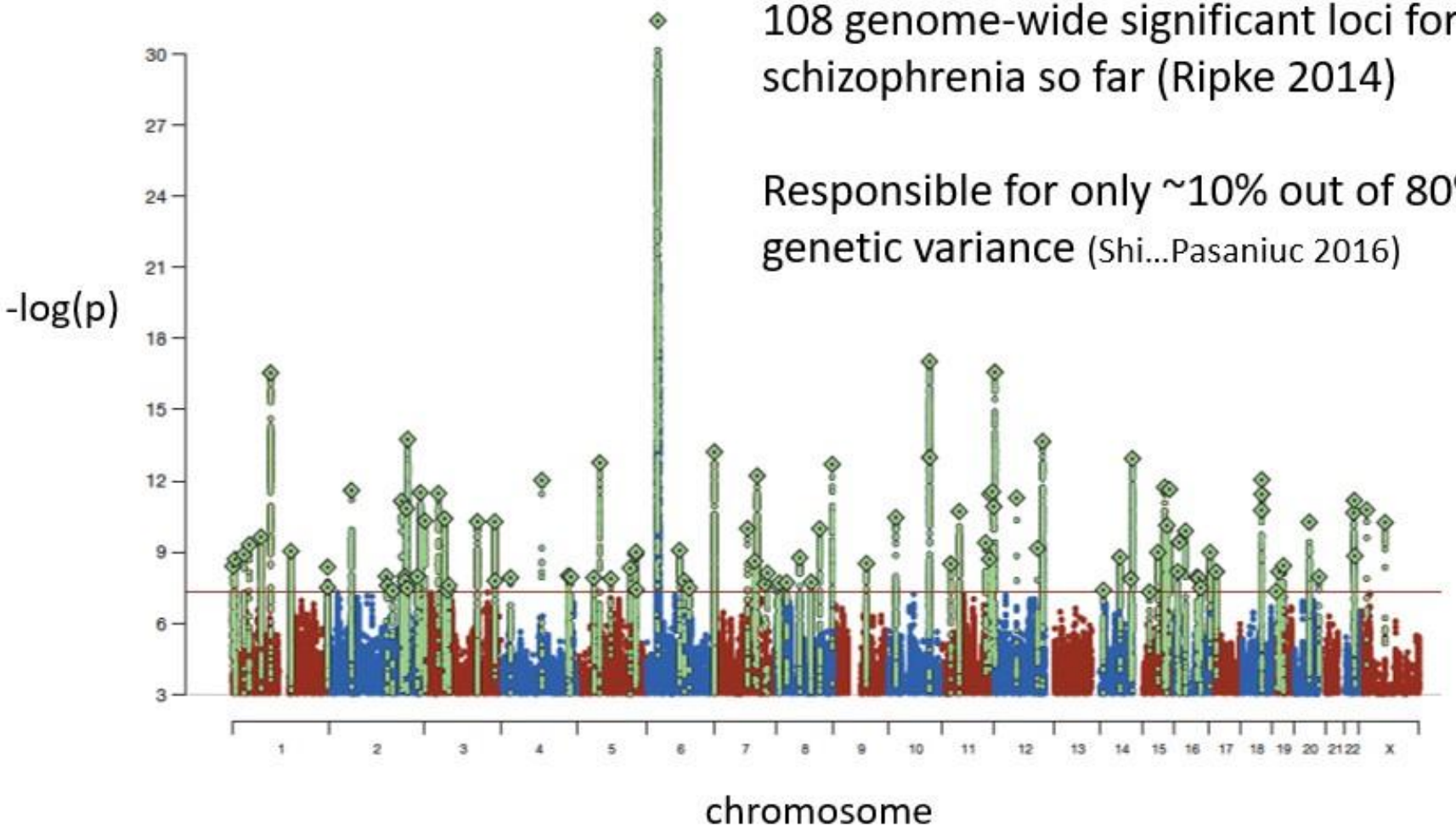
(From Wood 2014)

This suggests trouble for the expectation that heritability is "funneled" into key pathways

Most complex traits are polygenic.



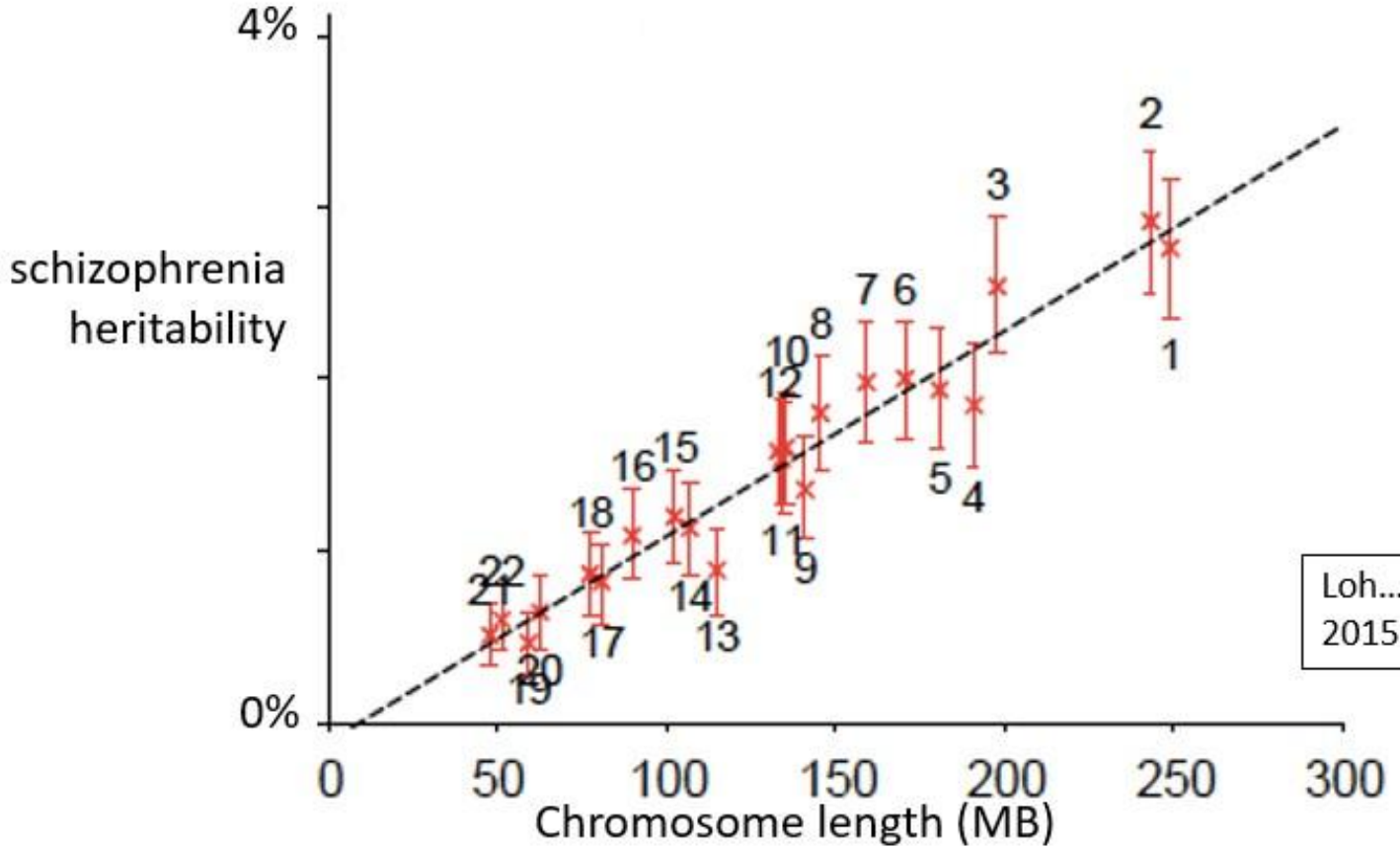
Most complex traits are polygenic.



108 genome-wide significant loci for schizophrenia so far (Ripke 2014)

Responsible for only ~10% out of 80% of genetic variance (Shi…Pasaniuc 2016)

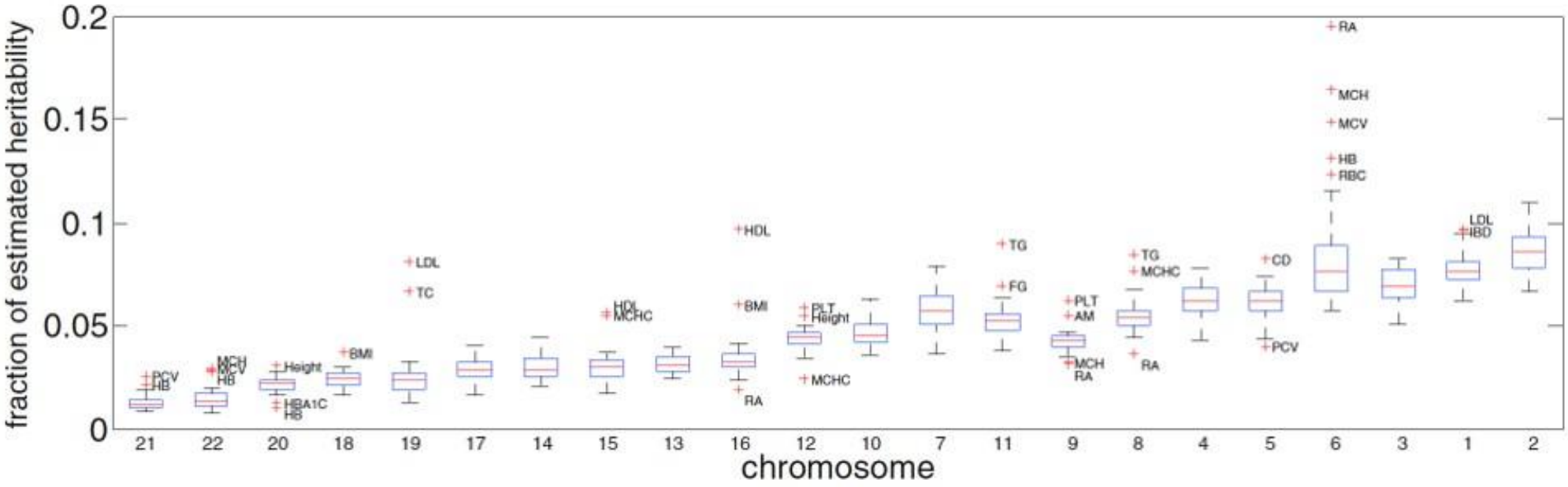
Schizophrenia is hugely polygenic.



Loh…Price, Nature Genetics 2015.

- At a broad scale, causal SNPs are spread widely across the genome
- Loh et al: >70% of MB windows in the genome contribute to schizophrenia heritability

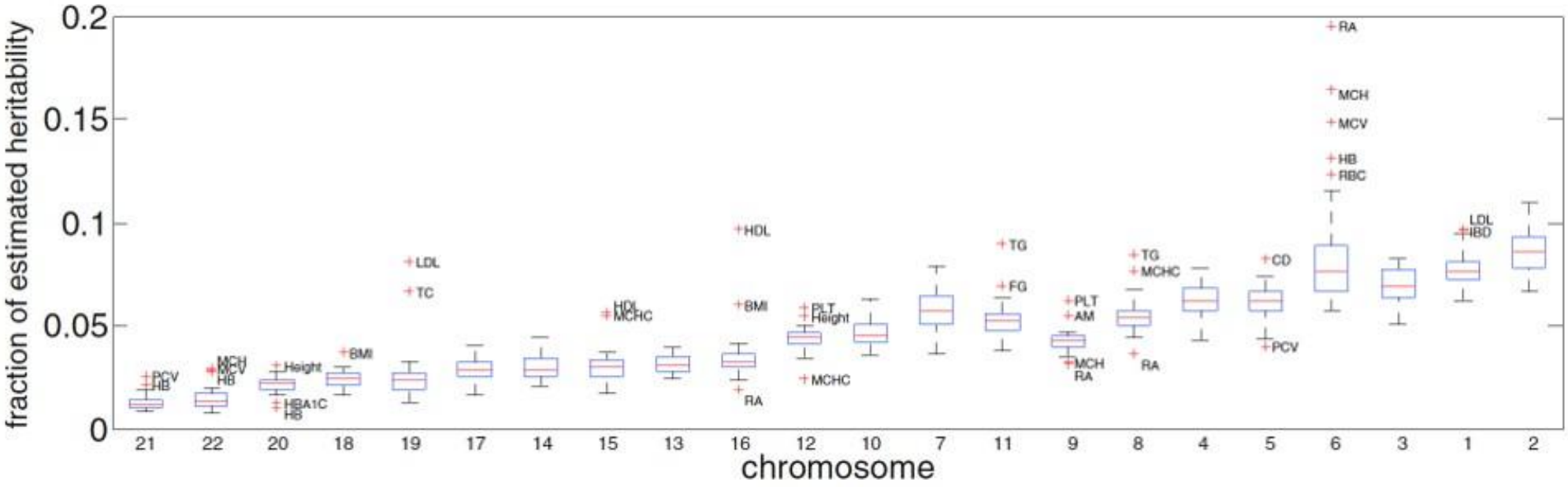
Nearly all complex traits show a strong polygenic signature at a broad scale



- 30 traits were considered ranging from autoimmune diseases, to anthropomorphic traits, to metabolic traits, etc...

Shi et al. AJHG 2016

Nearly all complex traits show a strong polygenic signature at a broad scale



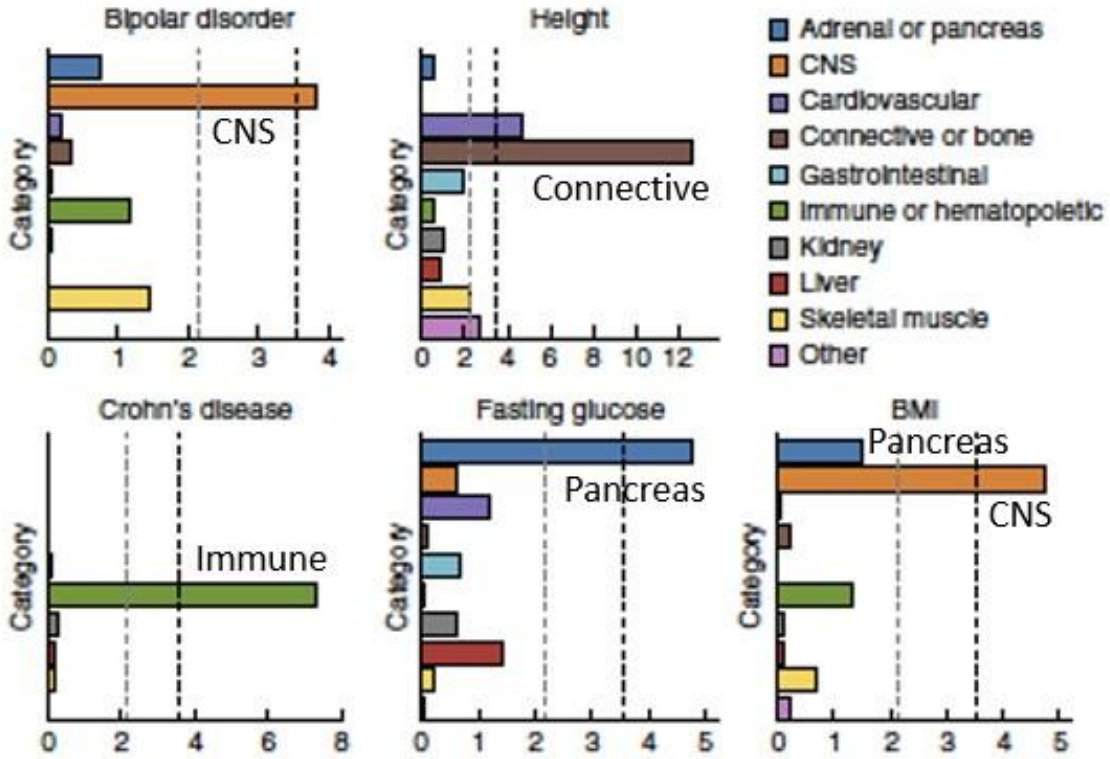
- 30 traits were considered ranging from autoimmune diseases, to anthropomorphic traits, to metabolic traits, etc...
- Polygenic architecture inconsistent with the view that trait-variants are all located near genes with direct effects on trait/disease.

Shi et al. AJHG 2016

Observation #2: Many groups have shown that GWAS signals are enriched in chromatin that is active in cell-types that "make sense"…

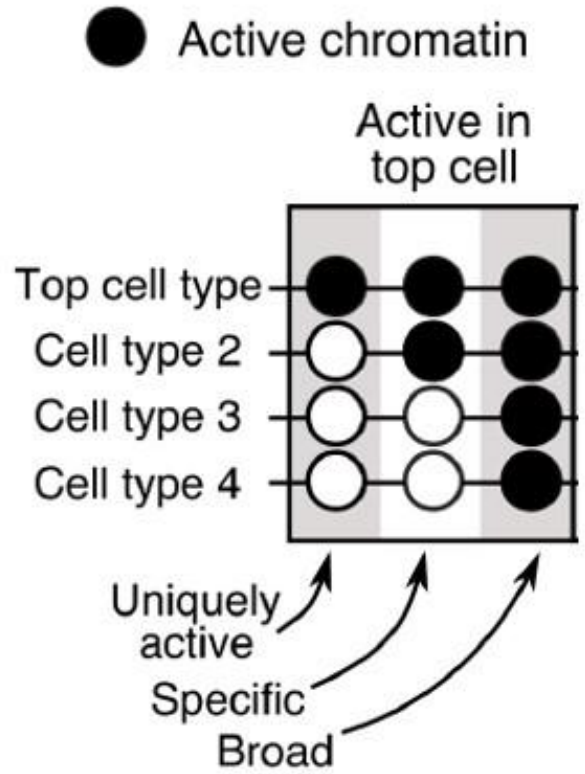
…but surprisingly, it doesn't matter much whether the chromatin is broadly active, or active only in relevant cell types

GWAS signals are enriched in chromatin that is active in cell-types that "make sense".



Analysis by Stratified LD Score regression
[Finucane…Alkes Price (2015) NG]

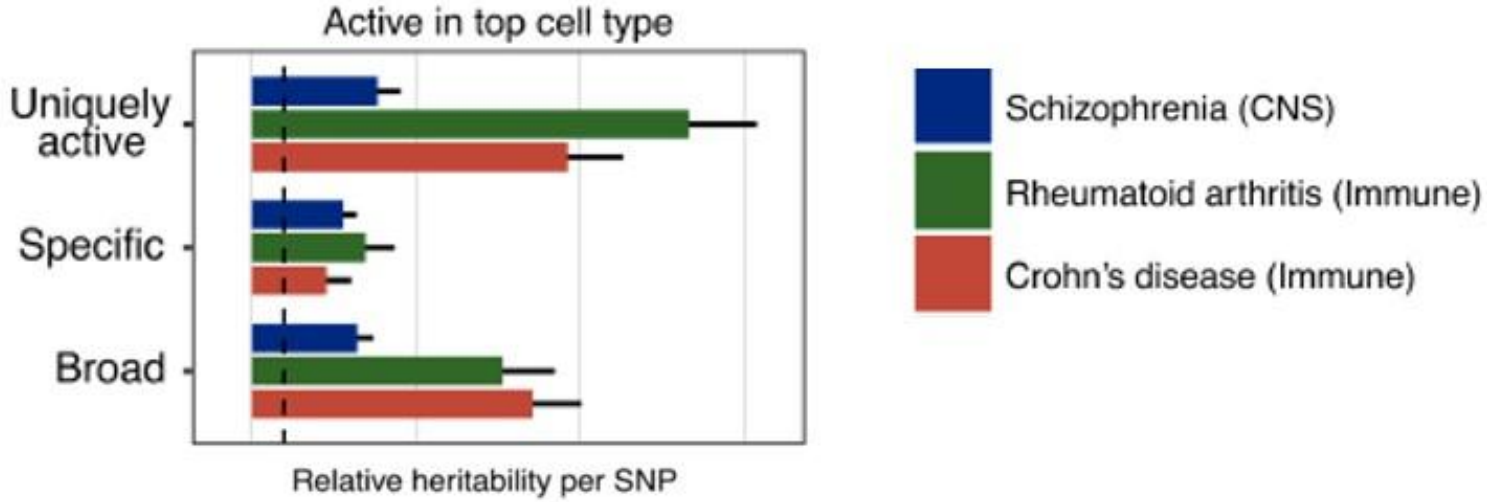
- GWAS SNPs generally affect cell-type-specific processes (classical view).



Three types of chromatin regions:

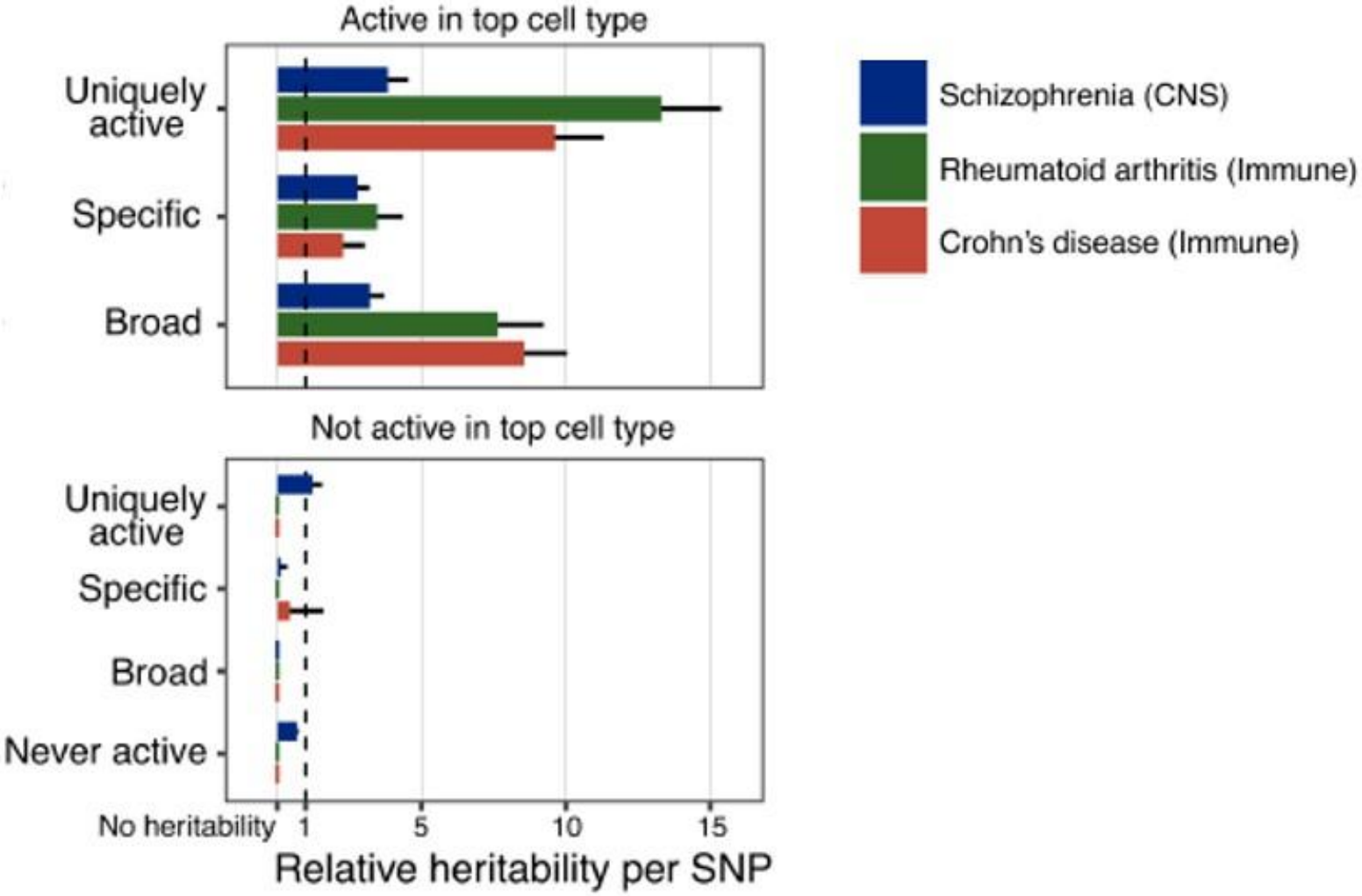
In principle we might expect cell type-specific open chromatin to be most enriched in disease heritability

Per-SNP heritability is similar in cell type-specific vs broadly active chromatin



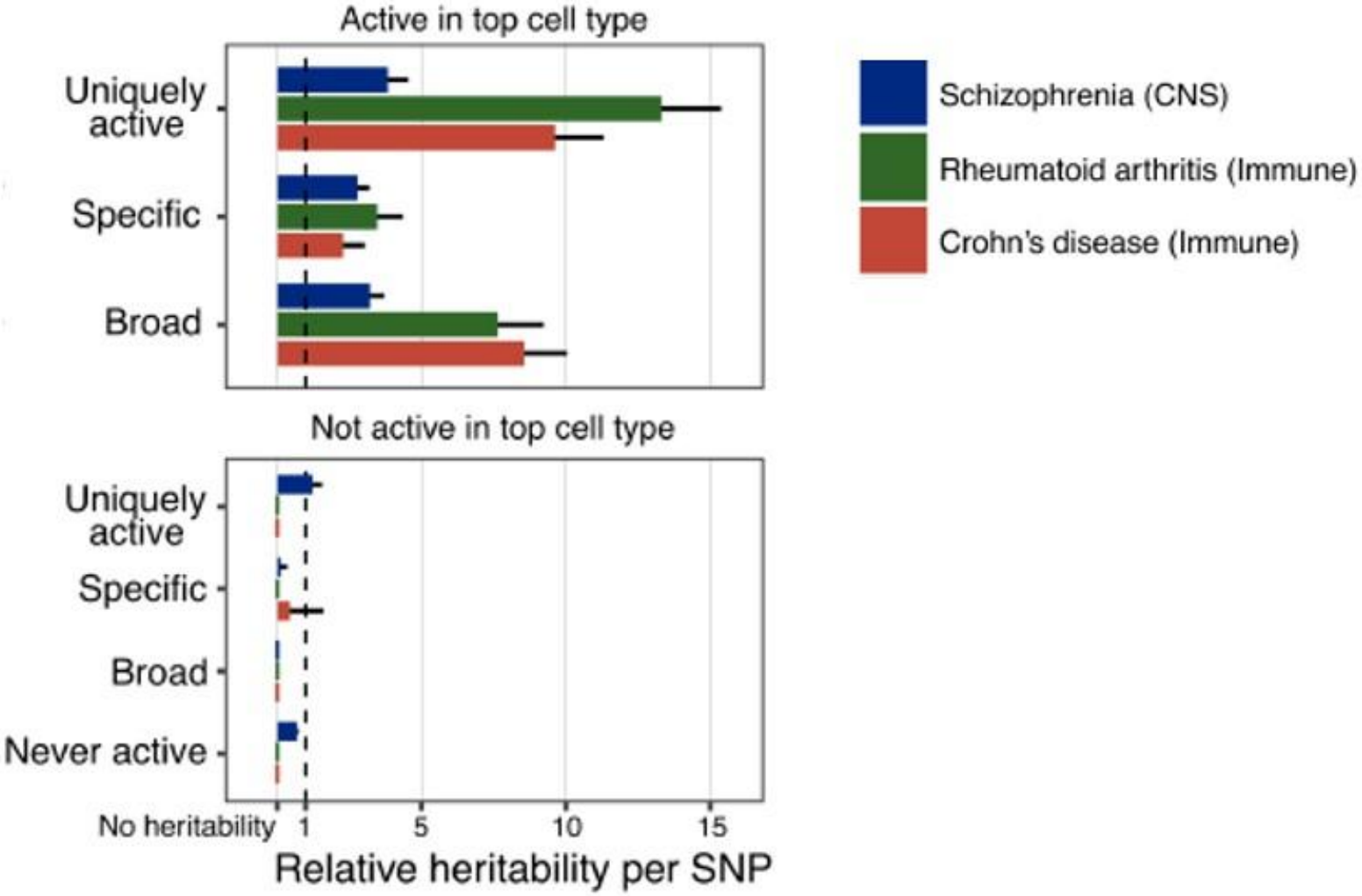
Analysis using stratified LD Score regression and annotations from Finucane et al

Per-SNP heritability is similar in cell type-specific vs broadly active chromatin



Analysis using stratified LD Score regression and annotations from Finucane et al

Per-SNP heritability is similar in cell type-specific vs broadly active chromatin



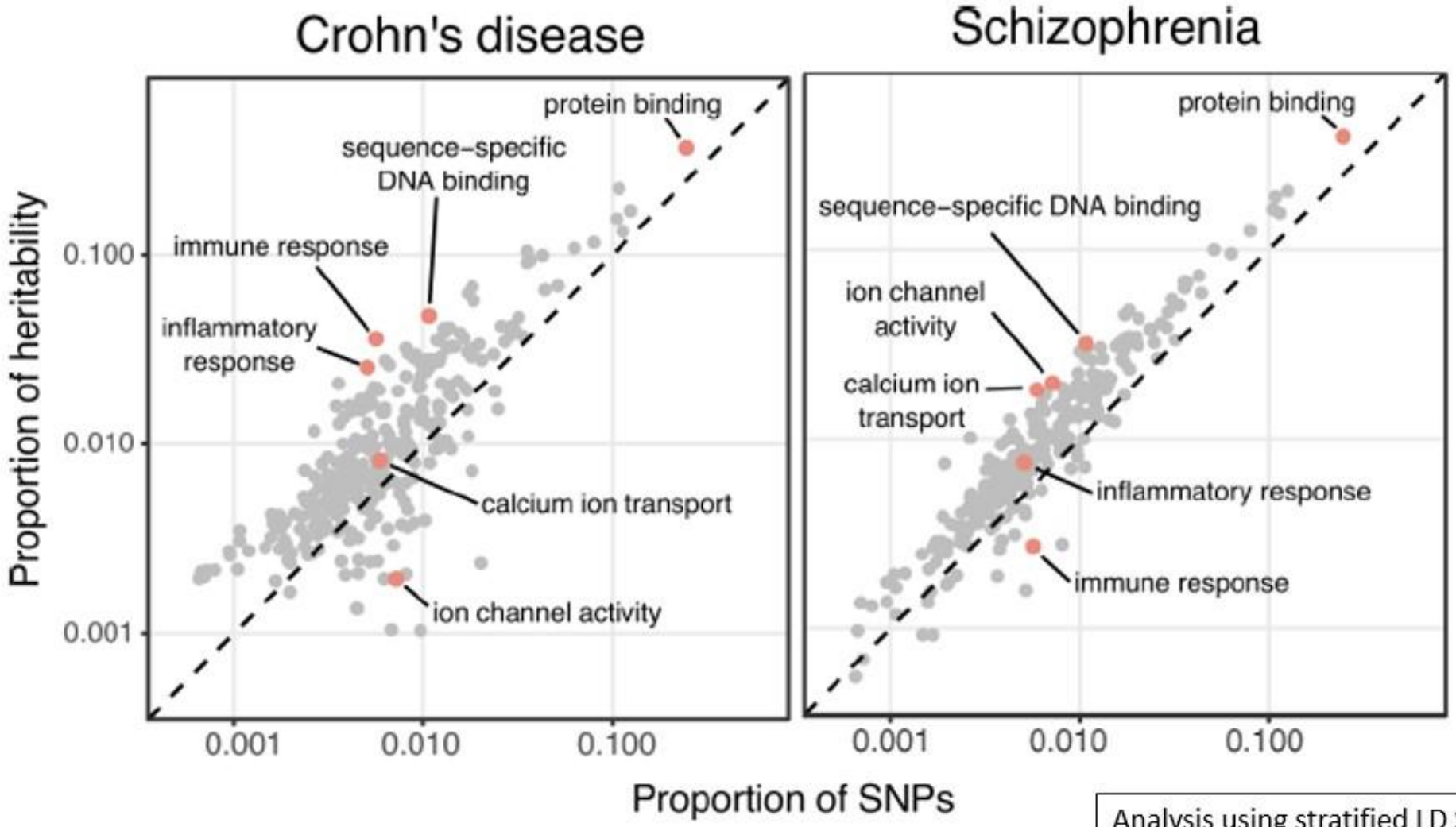
Not all genetic effects are mediated through cell-type-specific function?

Analysis using stratified LD Score regression and annotations from Finucane et al

Observation #3: Enrichment analysis of rare variants or strongest GWAS hits often identifies key genes and pathways….

….but most of the heritability is broadly distributed across genes with diverse functions.

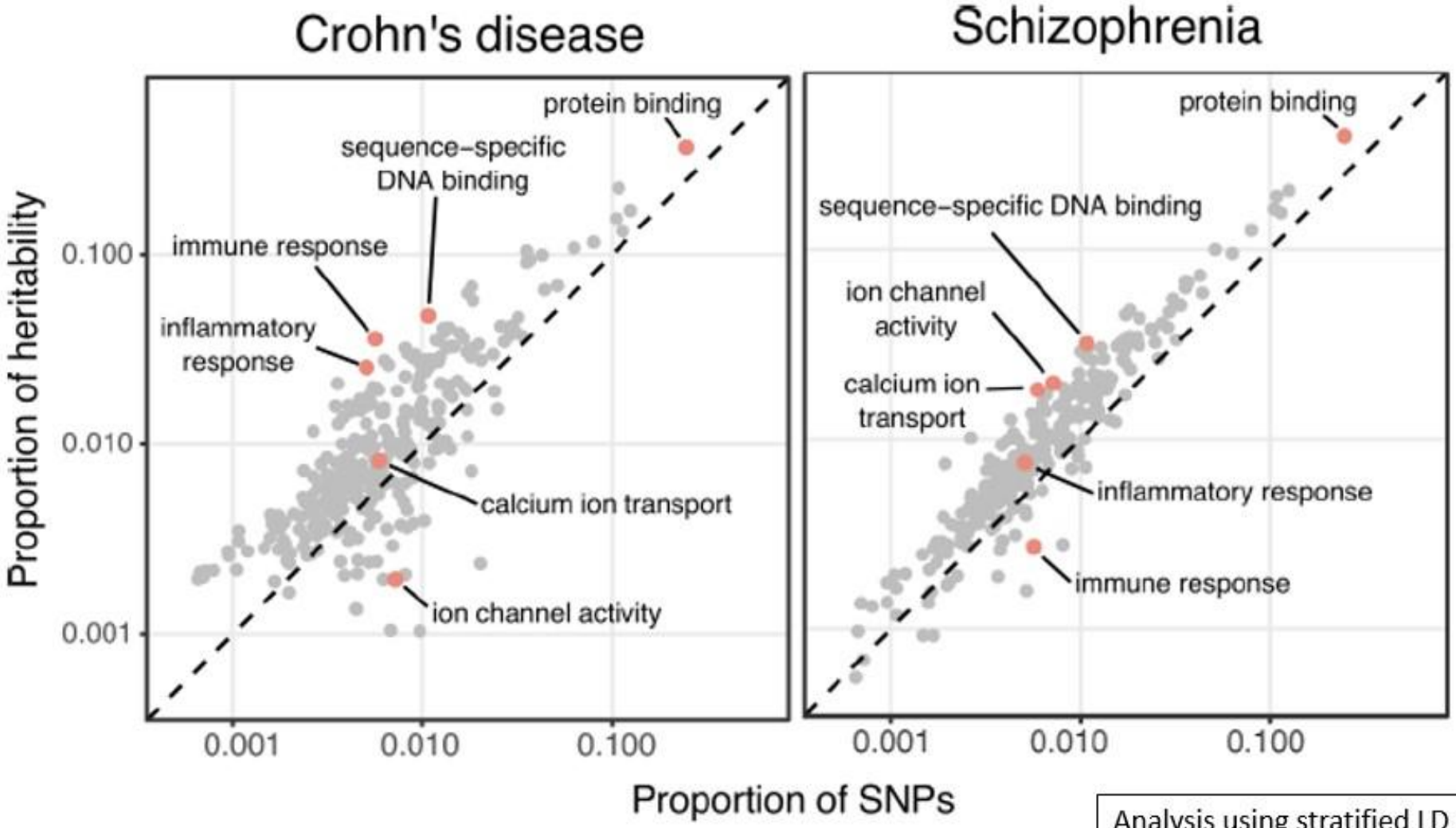
Heritability: Gene Ontology enrichments



- Relevant functional categories are enriched in heritability.

Analysis using stratified LD Score
Finucane et al

Heritability: Gene Ontology enrichments



Analysis using stratified LD Score
Finucane et al

- Relevant functional categories are enriched in heritability.
- For both diseases the category that explained the most heritability was simply the largest category, i.e. “protein binding”.

Our model to describe the data:
The "*omnigenic*" model

Our model to describe the data: The "*omnigenic*" model

3 types of genes:

- Tier 1: Core genes: direct roles in disease

Our model to describe the data:
The "*omnigenic*" model

3 types of genes:

- Tier 1: Core genes: direct roles in disease
- Tier 2: Peripheral genes: all other expressed genes may affect the regulation of core genes

Our model to describe the data: The "*omnigenic*" model

3 types of genes:

- Tier 1: Core genes: direct roles in disease
- Tier 2: Peripheral genes: all other expressed genes may affect the regulation of core genes
- Tier 3: Genes not expressed in the "right" cell types do not contribute to heritability

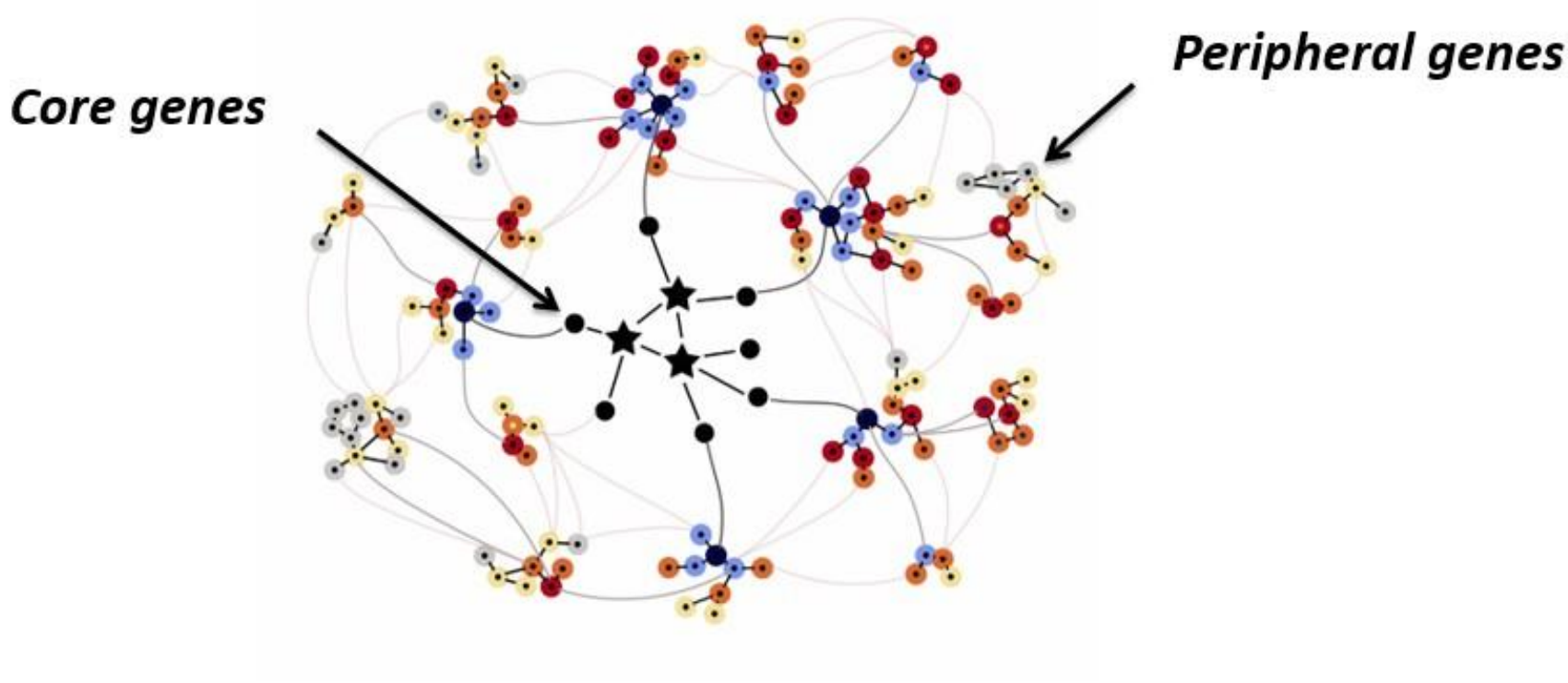
Our model to describe the data: The "*omnigenic*" model

3 types of genes:

- Tier 1: Core genes: direct roles in disease
- Tier 2: Peripheral genes: all other expressed genes may affect the regulation of core genes
- Tier 3: Genes not expressed in the "right" cell types do not contribute to heritability

Can explain why most phenotypic variance could be due to genetic variants affecting peripheral genes

Hypothetical scenario: Peripheral genes outnumber core genes by ~100:1, and so they can dominate the phenotypic variance.

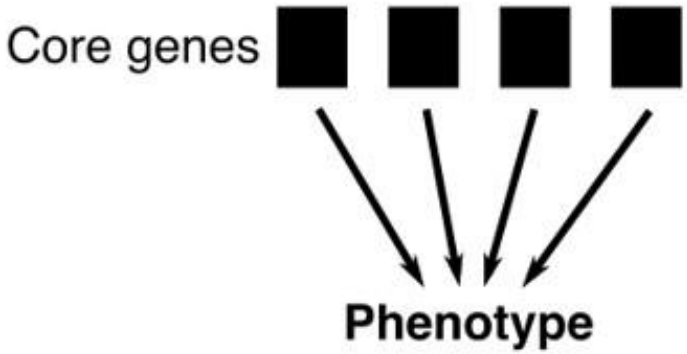


Proposed mechanism: Possibly through weak effects rippling through gene or PPI networks?

(Boyle, Li, and Pritchard, 2017)

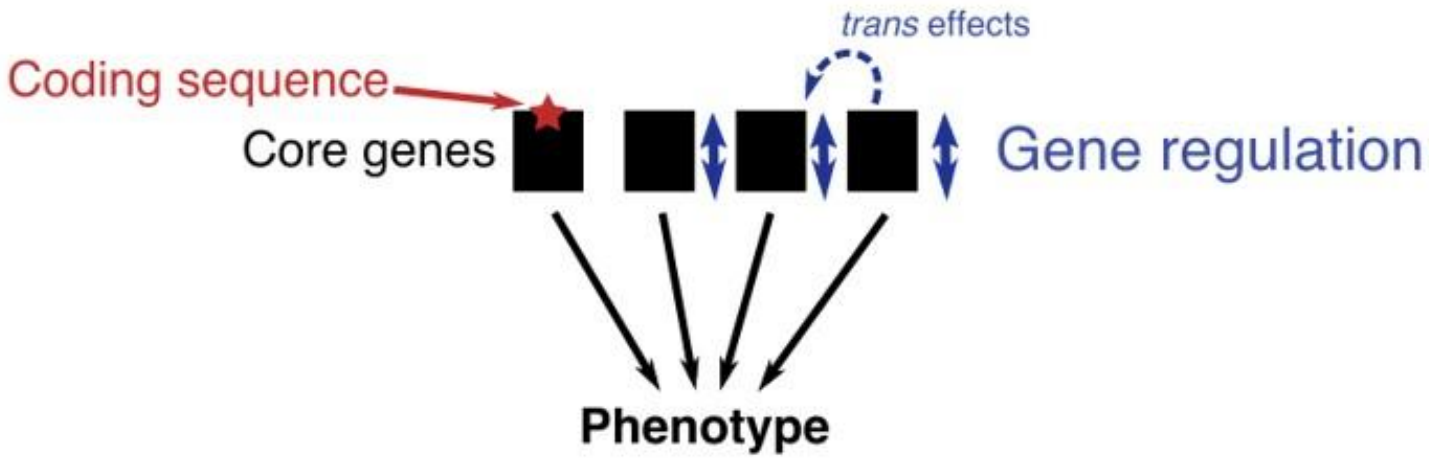
Starting point of a quantitative phenotype model

1) Genes with direct effect on phenotypes are defined to be core genes



Starting point of a quantitative phenotype model

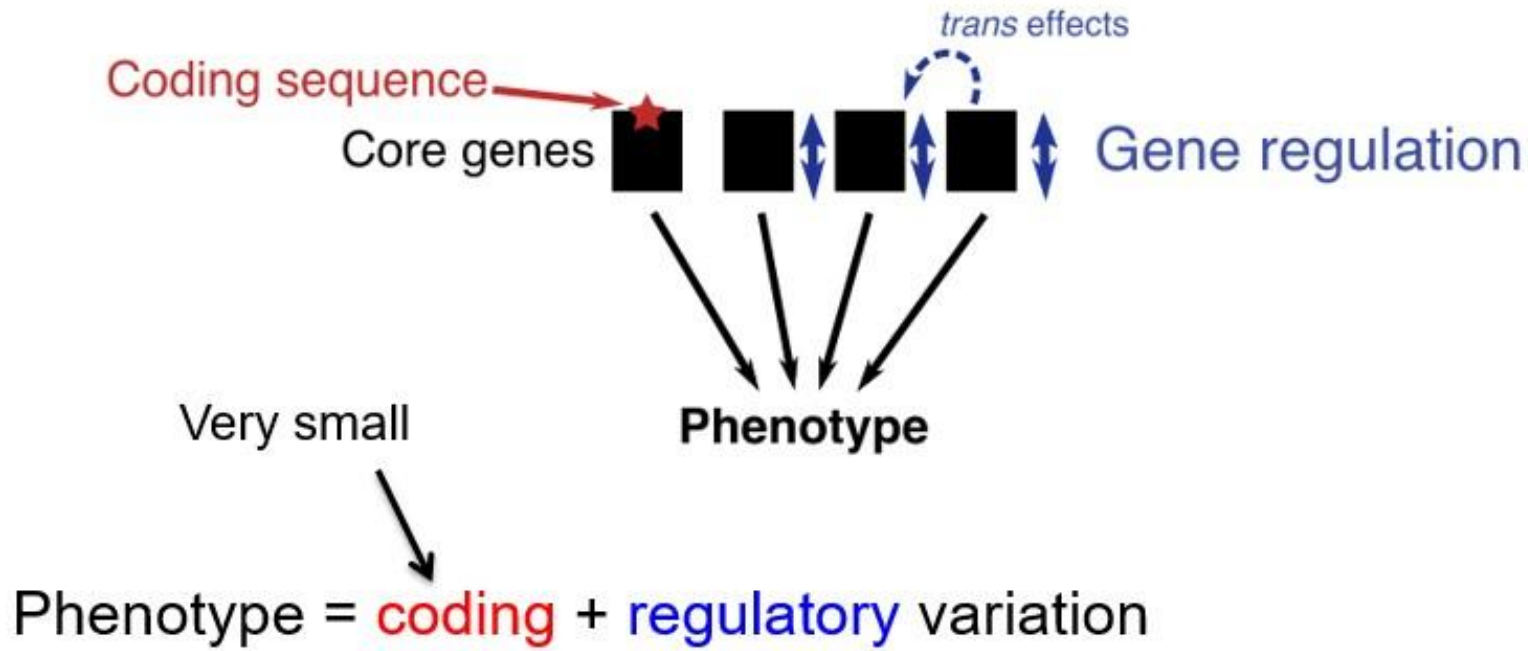
1) Genes with direct effect on phenotypes are defined to be core genes



Phenotype = coding + regulatory variation

Starting point of a quantitative phenotype model

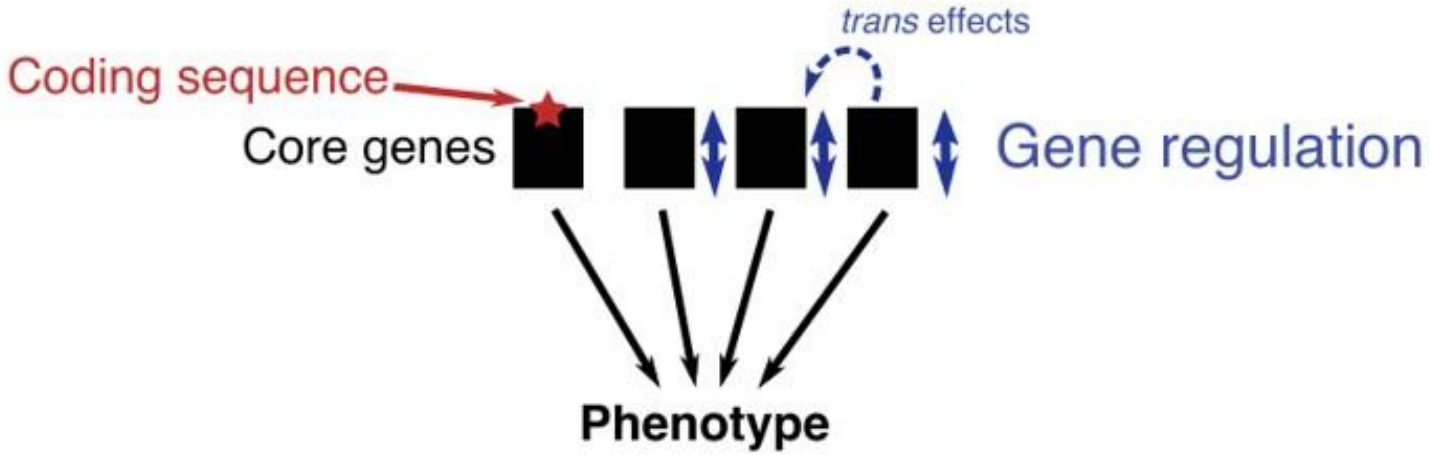
1) Genes with direct effect on phenotypes are defined to be core genes
2) Coding variants contribute very little to phenotypic variation.



Starting point of a quantitative phenotype model

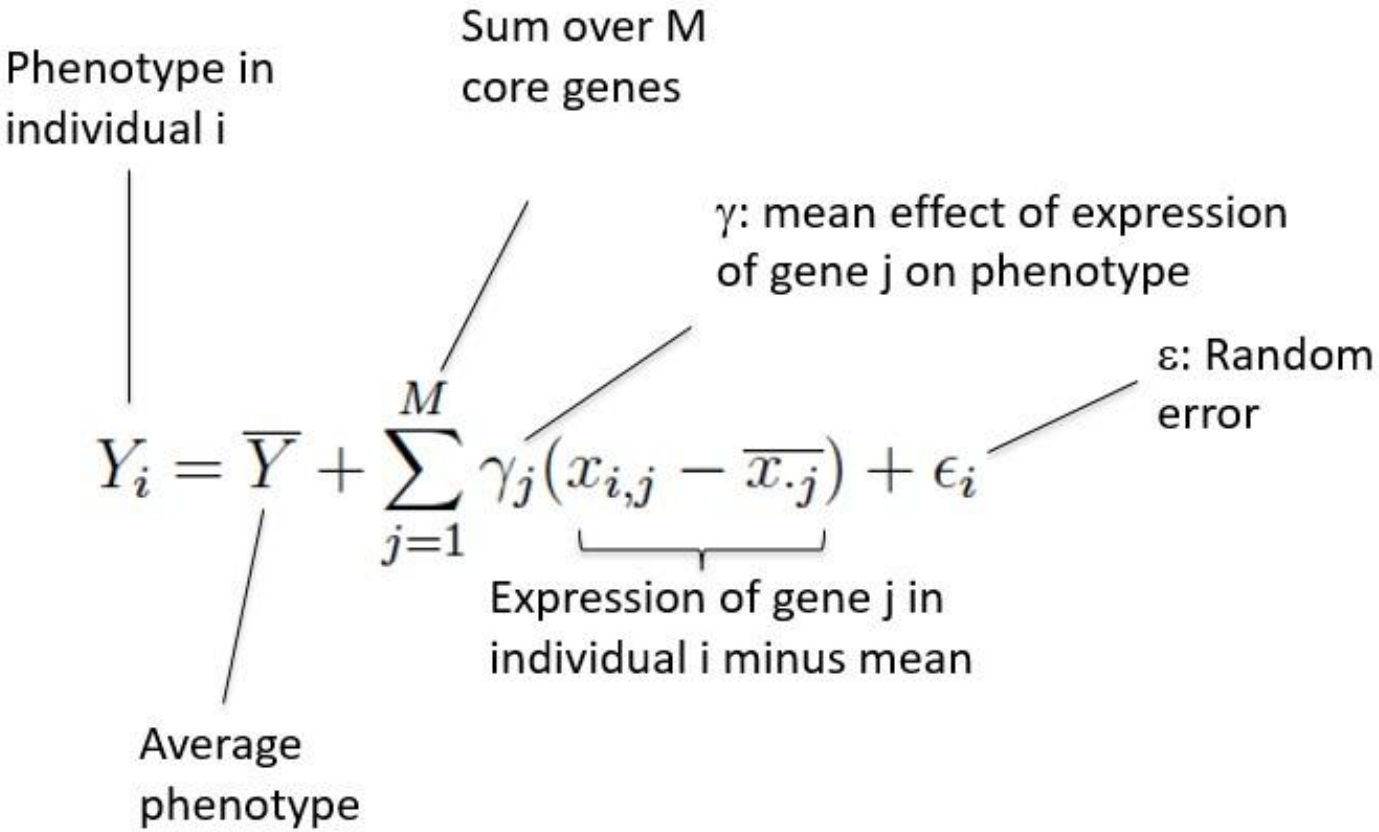
1) Genes with direct effect on phenotypes are defined to be core genes
2) Coding variants contribute very little to phenotypic variation.

Bulk of phenotypic variation is explained by variation in core gene regulation.



Phenotype ≈ regulatory variation

A simple phenotype model based on expression of core genes



$$Y_i = \overline{Y} + \sum_{j=1}^{M} \gamma_j (x_{i,j} - \overline{x_{\cdot j}}) + \epsilon_i$$

Expression variance

M of these terms

Phenotypic variance

$$\overbrace{\mathrm{Var}(\mathrm{Y_i})} = \sum_{\mathrm{j=1}}^{\mathrm{M}} \gamma_{\mathrm{j}}^2 \overbrace{\mathrm{Var}(\mathrm{x_{\cdot j}})} + \sum_{\mathrm{j \neq k}}^{\mathrm{M}} \gamma_{\mathrm{j}} \gamma_{\mathrm{k}} \underbrace{\mathrm{Cov}(\mathrm{x_{\cdot j}}, \mathrm{x_{\cdot k}})} + \mathrm{Var}(\epsilon)$$

Expression covariance

Nearly M^2 of these terms

$$Y_i = \overline{Y} + \sum_{j=1}^{M} \gamma_j (x_{i,j} - \overline{x_{\cdot j}}) + \epsilon_i$$

Expression variance

M of these terms

Phenotypic variance

$$\overbrace{\mathrm{Var}(\mathrm{Y_i})} = \sum_{\mathrm{j=1}}^{\mathrm{M}} \gamma_{\mathrm{j}}^2 \overbrace{\mathrm{Var}(\mathrm{x_{\cdot j}})} + \sum_{\mathrm{j \neq k}}^{\mathrm{M}} \gamma_{\mathrm{j}} \gamma_{\mathrm{k}} \underbrace{\mathrm{Cov}(\mathrm{x_{\cdot j}}, \mathrm{x_{\cdot k}})} + \mathrm{Var}(\epsilon)$$

What contribute to this gene expression variance?

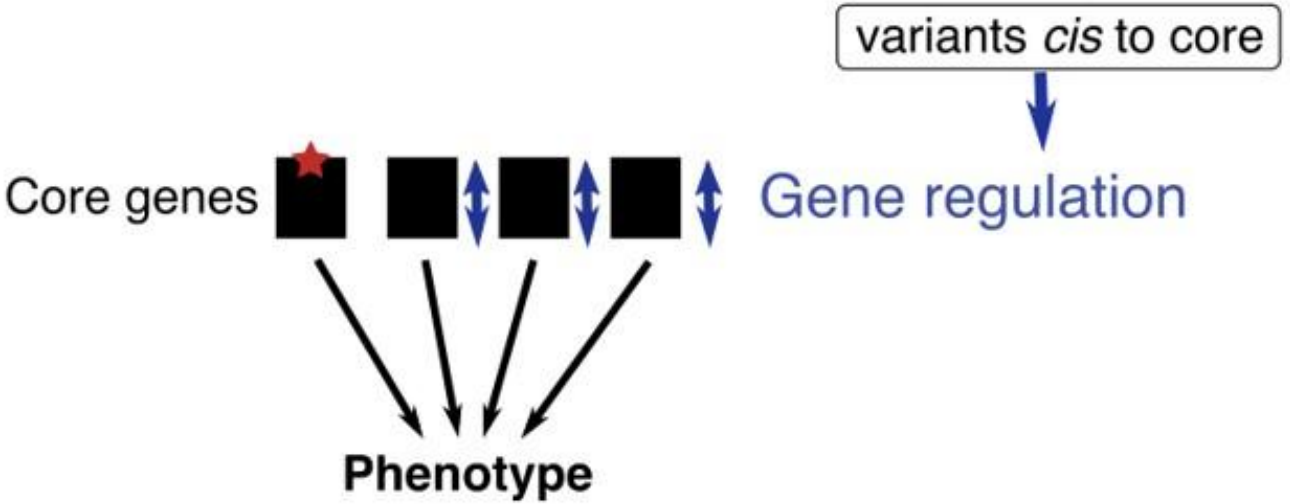
Expression covariance

Nearly M^2 of these terms

Core gene expression variance partitions into:

$$\sum_{j=1}^{M} \gamma_j^2 \overbrace{\mathrm{Var}(x_{\cdot j})}$$

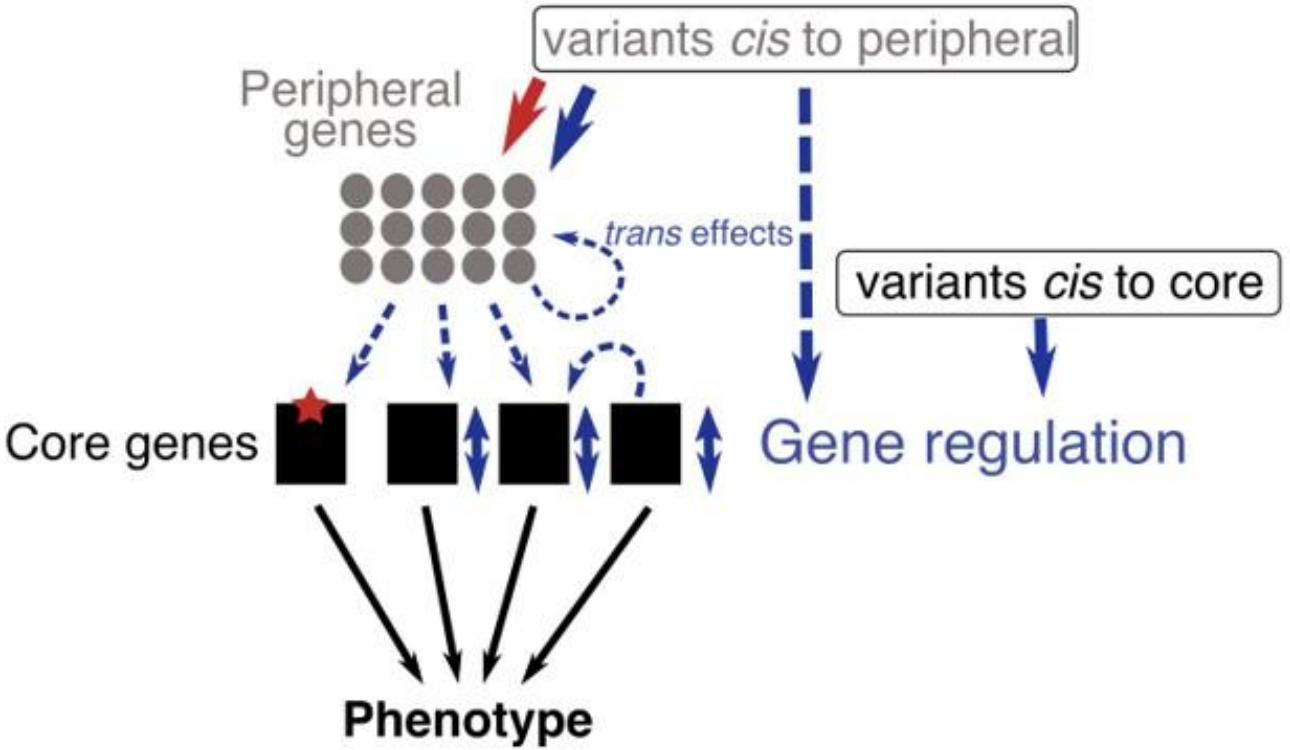
1) Variance from *cis* (nearby) core genes.



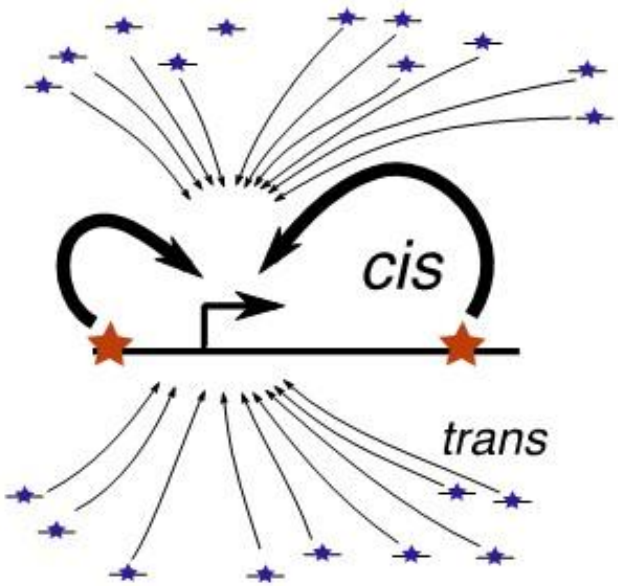
Core gene expression variance partitions into:

$$\sum_{j=1}^{M} \gamma_j^2 \mathrm{Var}(x_{\cdot j})$$

1) Variance from *cis* (nearby) core genes.
2) Variance from *trans* (far from) core genes (**likely *cis* to peripheral**)



How much of expression variance is due to *cis* vs *trans* effects?



Literature review: genetic variance in gene expression

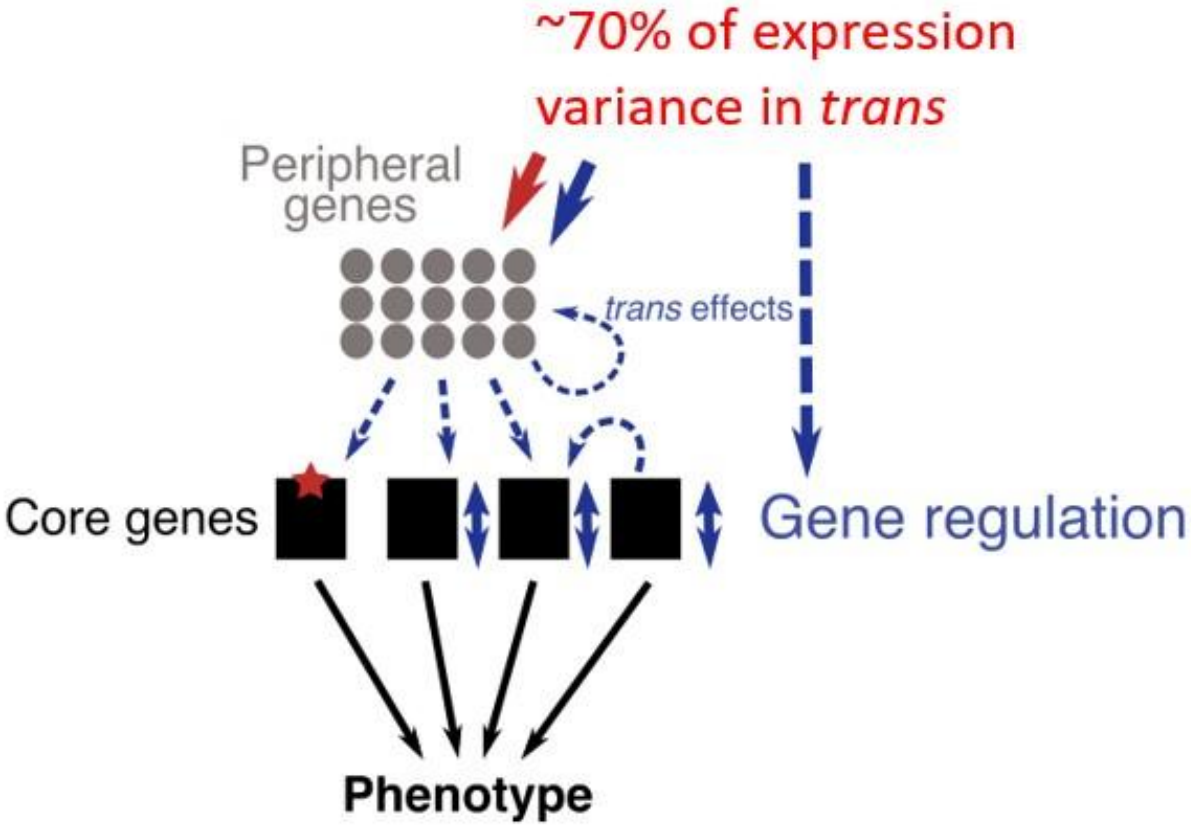
Percent h^2 in trans	Tissue/ organism	Platform	Sample size	Method
88%	LCL from admixed inds	Affymetrix Array	89	African-European ancestry
76%, 61%	Drosophila, whole body	RNA-seq	multi-fly pools	fly hybrids
76%, 63%	adipose, blood	custom array	638, 687	cis/trans IBD in families
70%, 65%, 64%	adipose, LCL, skin	Illumina Array	856	twin design
77%, 69%	peripheral blood	Affymetrix Array	2,752	twin design, LD Score
72%	yeast segregants	RNA-seq	1012	cis vs. trans eQTLs
62%	mouse liver	RNA-seq	192	GCTA
72%	mouse liver (proteins)	Mass Spec	192	GCTA
78%	human plasma (proteins)	protein aptamers	3301	LD Score Regression

Literature review: genetic variance in gene expression

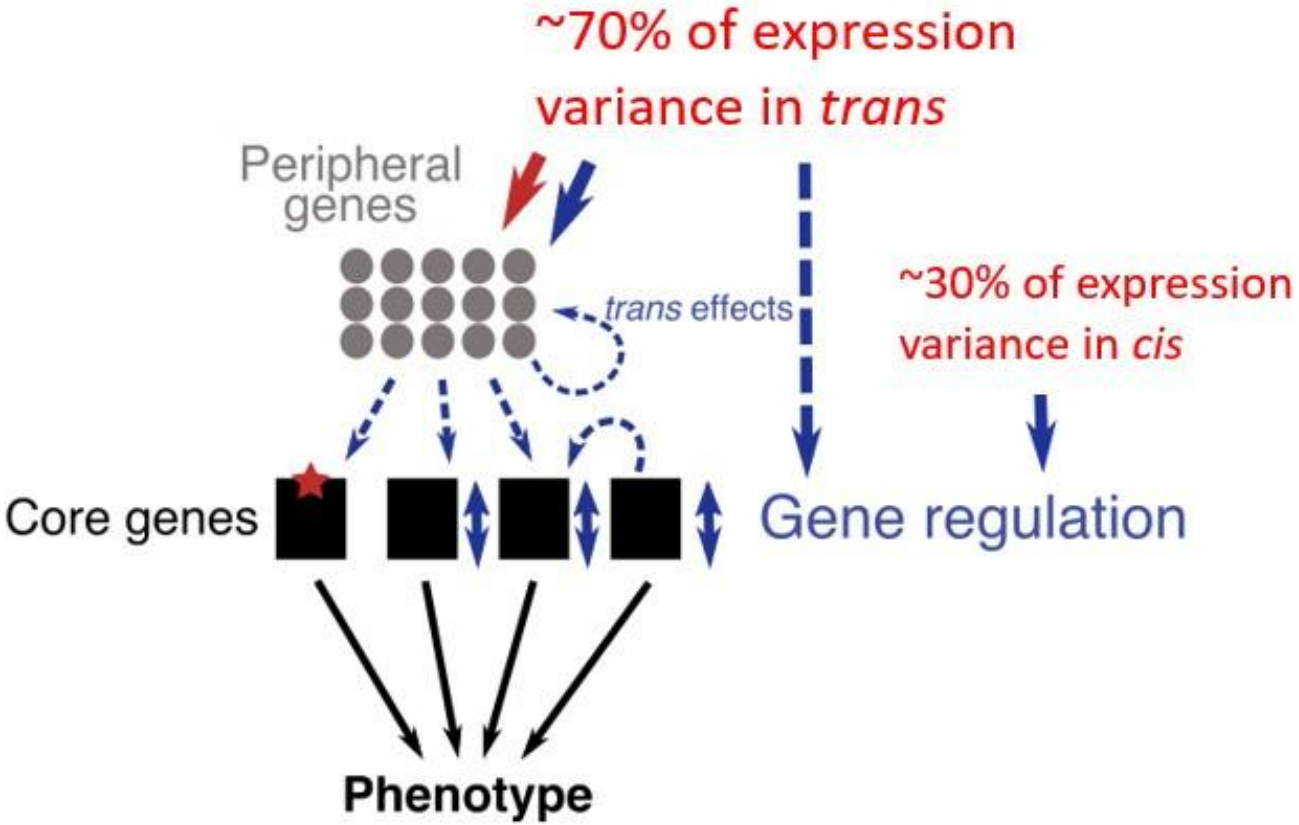
Percent h^2 in trans	Tissue/ organism	Platform	Sample size	Method
88%	LCL from admixed inds	Affymetrix Array	89	African-European ancestry
76%, 61%	Drosophila, whole body	RNA-seq	multi-fly pools	fly hybrids
76%, 63%	adipose, blood	custom array	638, 687	cis/trans IBD in families
70%, 65%, 64%	adipose, LCL, skin	Illumina Array	856	twin design
77%, 69%	peripheral blood	Affymetrix Array	2,752	twin design, LD Score
72%	yeast segregants	RNA-seq	1012	cis vs. trans eQTLs
62%	mouse liver	RNA-seq	192	GCTA
72%	mouse liver (proteins)	Mass Spec	192	GCTA
78%	human plasma (proteins)	protein aptamers	3301	LD Score Regression

~70% in trans

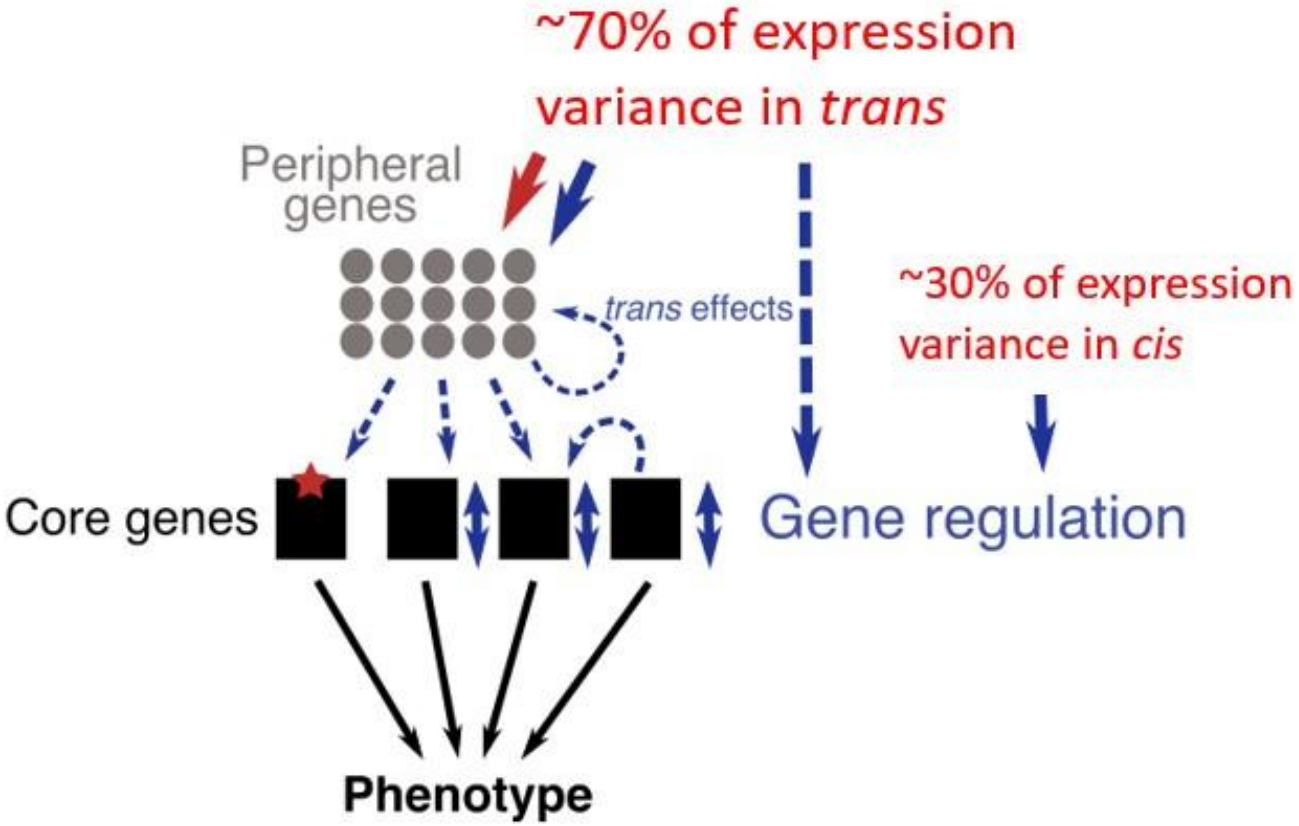
Only a minority of the trait heritability is expected to come from variants *cis* to core genes



Only a minority of the trait heritability is expected to come from variants *cis* to core genes

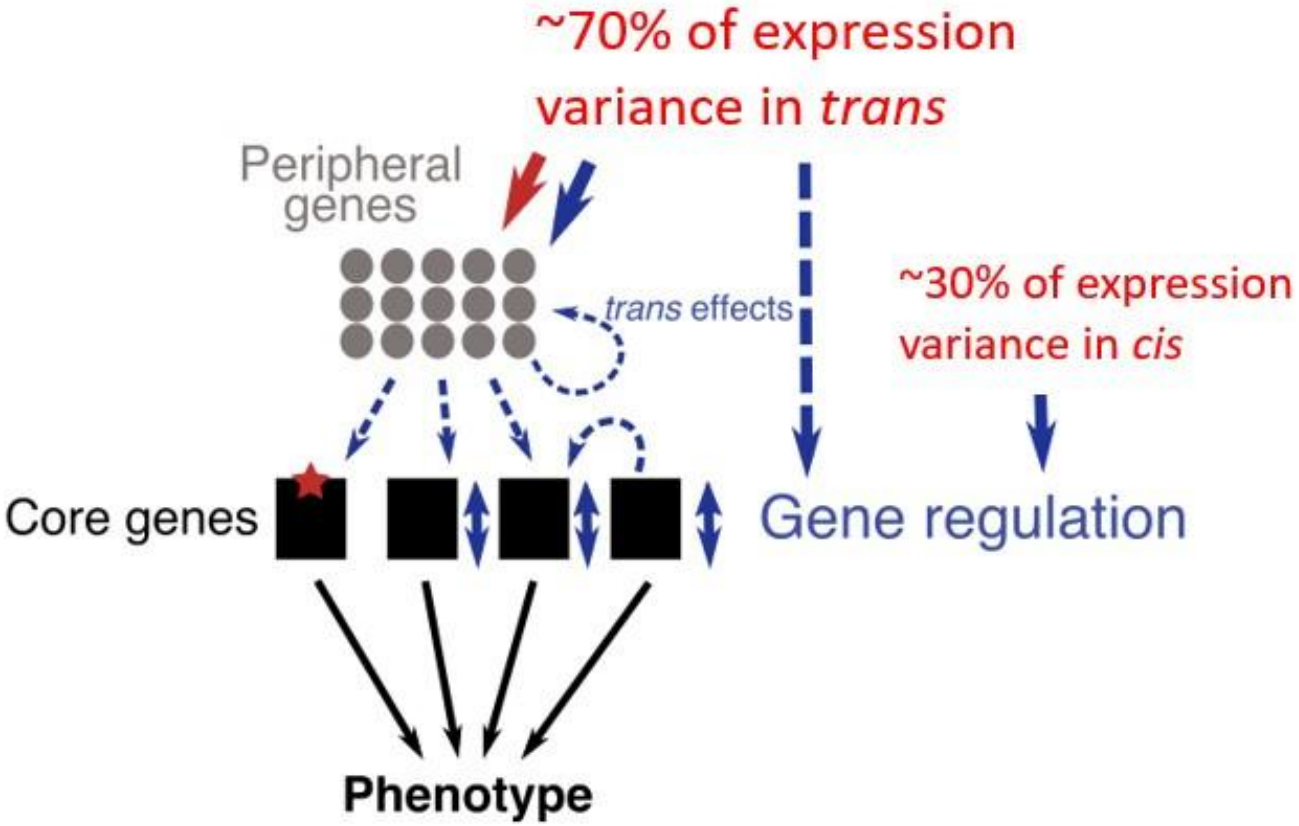


Only a minority of the trait heritability is expected to come from variants *cis* to core genes



~30% of heritability *cis* to core genes

Only a minority of the trait heritability is expected to come from variants *cis* to core genes



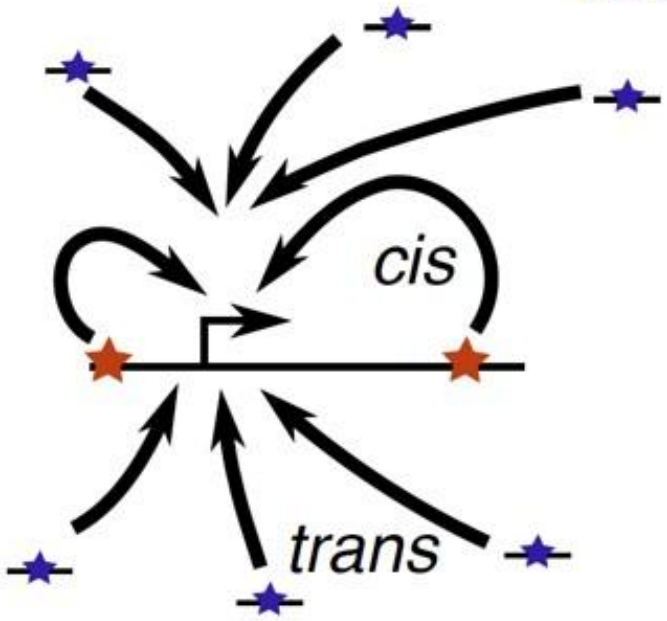
~30% of heritability *cis* to core genes

If a trait depends on the expression level of a single core gene, most genetic variance for this trait would be explained by variants located far away.

How many *trans* effects?

Model A:
Few trans-acting variants,
some with big effects

~70% of variance in *trans*

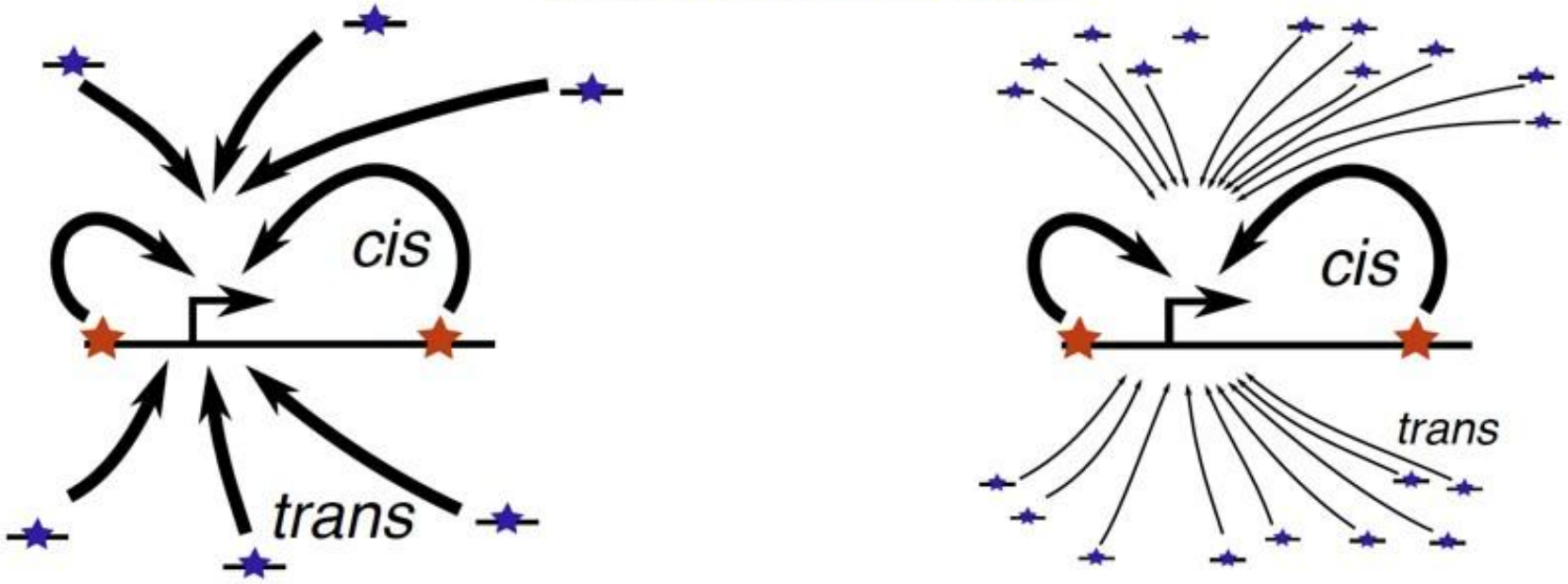


How many *trans* effects?

Model A:
Few trans-acting variants, some with big effects

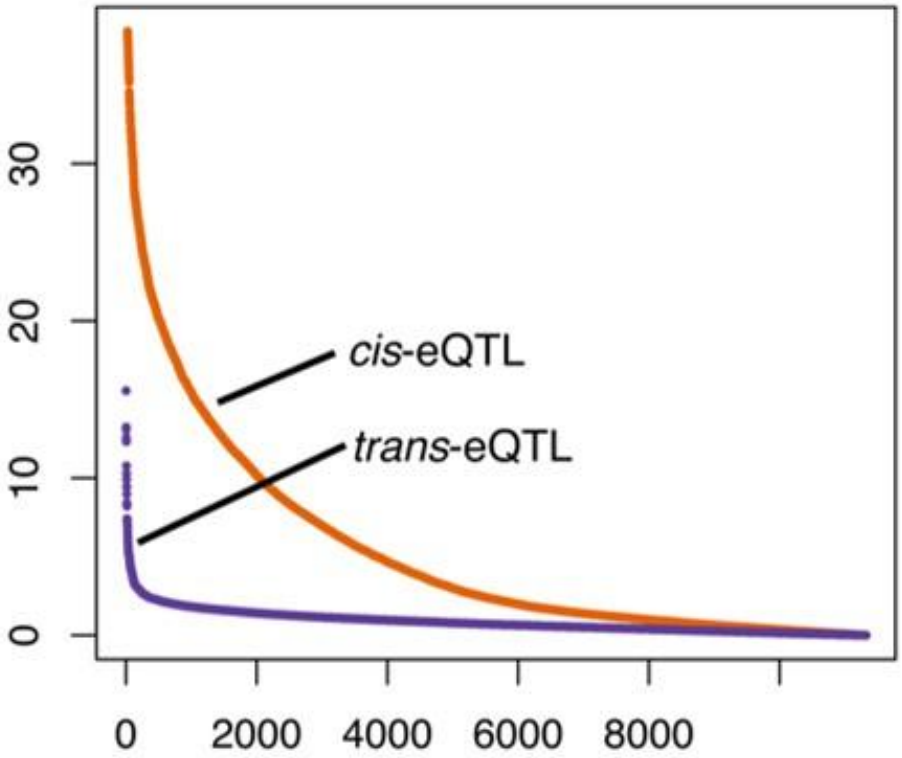
Model B:
Many trans-acting variants, all with small effects

~70% of variance in *trans*



Trans eQTLs have very small effect sizes compared to *cis*

Effect sizes of SNPs on expression (|Z| scores)



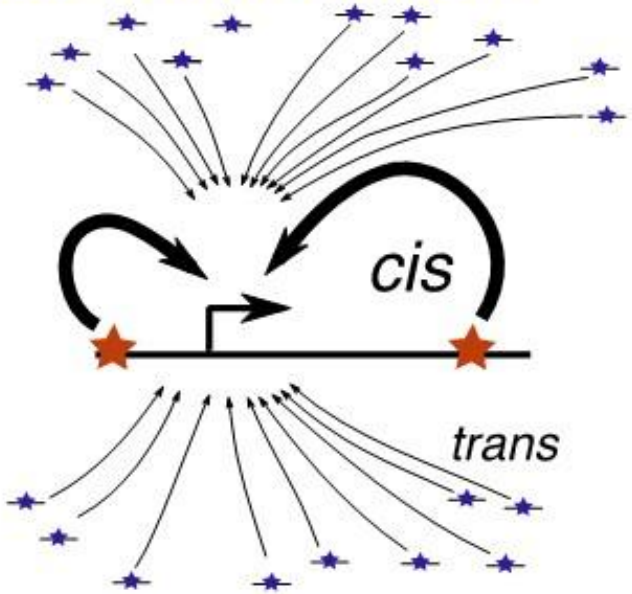
Distribution of effect sizes for top hits, cis and trans

This difference is even more dramatic for (effect size)^2

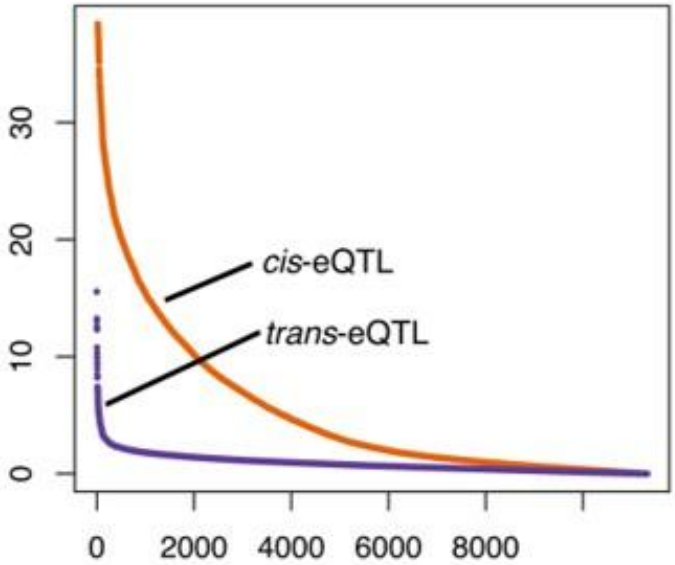
(Liu, Li, and Pritchard, bioRxiv, 2018)
Plot shows effect sizes of strongest cis and trans signals from Wright et al. 2014 replicated in DGN

Together these observations imply that a typical gene must have huge numbers of weak *trans*-regulators

~70% of variance in *trans*



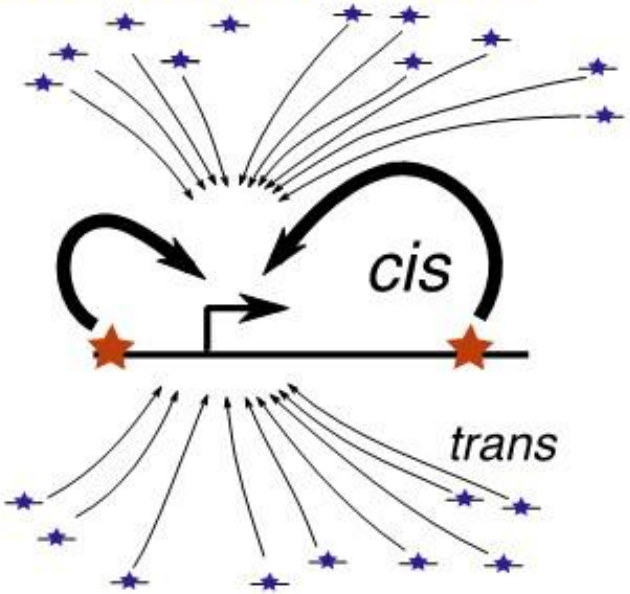
Trans effect much smaller than cis



(Liu, Li, and Pritchard, bioRxiv, 2018)

Together these observations imply that a typical gene must have huge numbers of weak *trans*-regulators

~70% of variance in *trans*



So assuming > tens of core genes, this model explains why such a large fraction of the genome can contribute to any given complex trait

(Liu, Li, and Pritchard, bioRxiv, 2018)

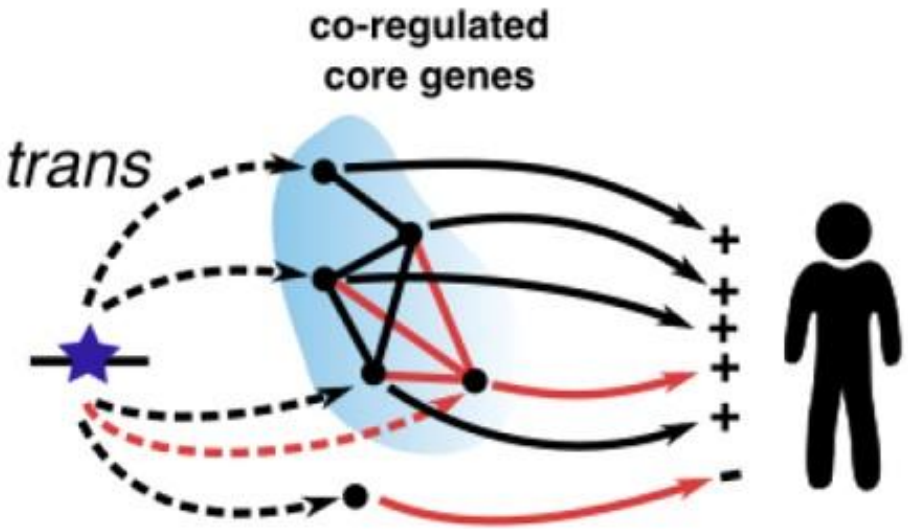
Co-regulation of core gene expression can amplify the contribution of trans-variation/peripheral genes.

$$\sum_{j=1}^{M} \gamma_j^2 \mathrm{Var}(x_{\cdot j}) + \sum_{j \neq k}^{M} \gamma_j \gamma_k \mathrm{Cov}(x_{\cdot j}, x_{\cdot k}) + \mathrm{Var}(\epsilon)$$

(Liu, Li, and Pritchard, bioRxiv, 2018)

Co-regulation of core gene expression can amplify the contribution of trans-variation/peripheral genes.

$$\sum_{j=1}^{M} \gamma_j^2 \mathrm{Var}(x_{\cdot j}) + \sum_{j \neq k}^{M} \gamma_j \gamma_k \mathrm{Cov}(x_{\cdot j}, x_{\cdot k}) + \mathrm{Var}(\epsilon)$$

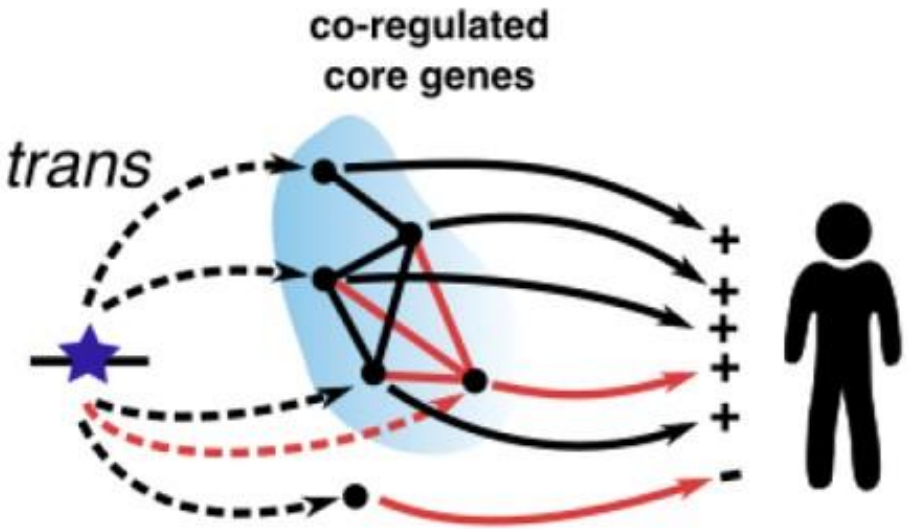


If (1) core gene expression levels are positively correlated, and (2) core genes have coordinated effects, then covariance term starts to contribute to phenotypic variance.

(Liu, Li, and Pritchard, bioRxiv, 2018)

Co-regulation of core gene expression can amplify the contribution of trans-variation/peripheral genes.

$$\sum_{j=1}^{M} \gamma_j^2 \mathrm{Var}(x_{\cdot j}) + \sum_{j \neq k}^{M} \gamma_j \gamma_k \mathrm{Cov}(x_{\cdot j}, x_{\cdot k}) + \mathrm{Var}(\epsilon)$$



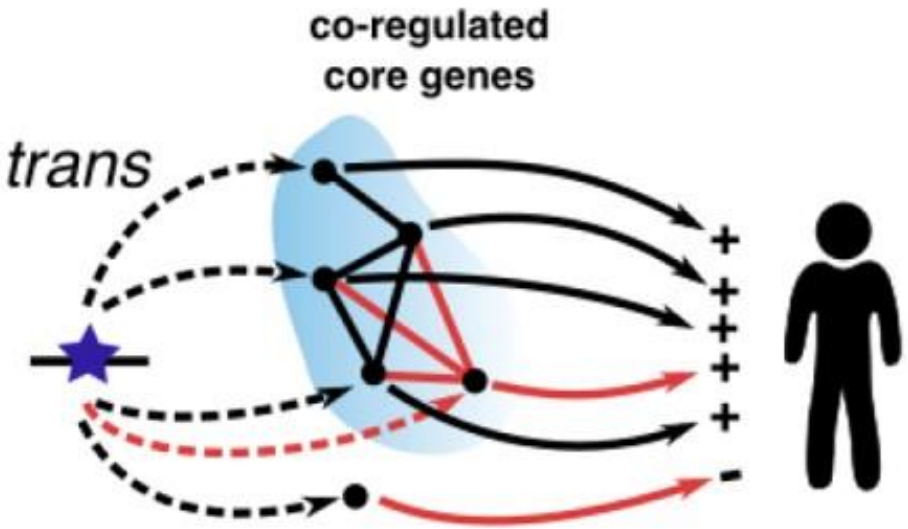
If (1) core gene expression levels are positively correlated, and (2) core genes have coordinated effects, then covariance term starts to contribute to phenotypic variance.

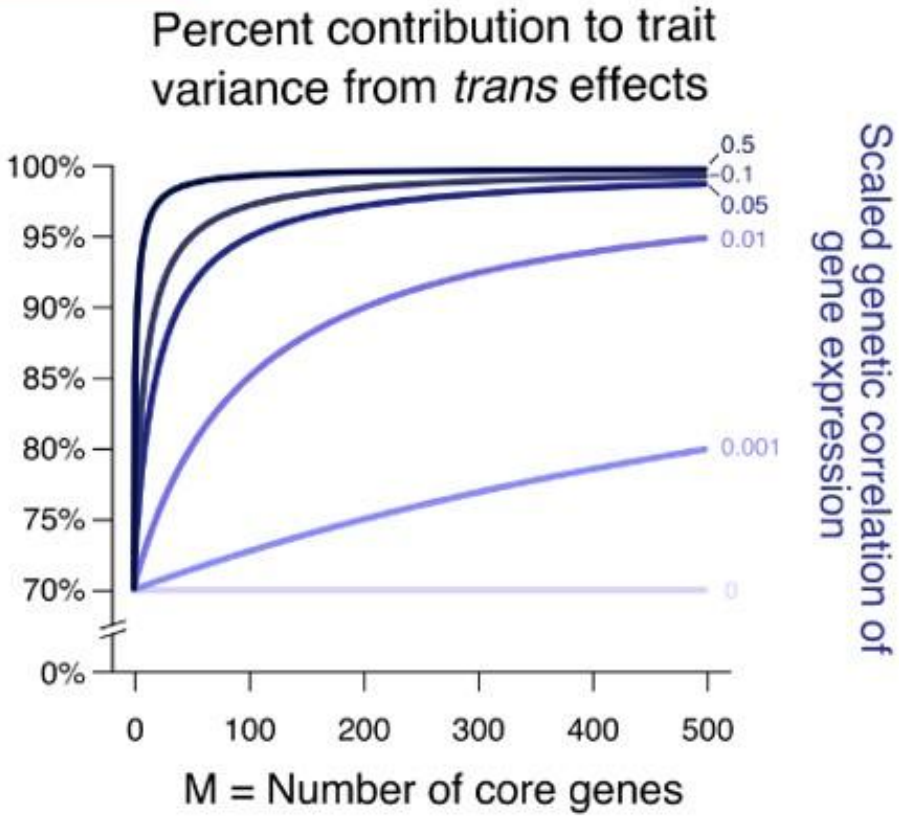
Genetic covariance likely to come from trans-effects.

(Liu, Li, and Pritchard, bioRxiv, 2018)

Co-regulation of core gene expression can amplify the contribution of trans-variation/peripheral genes.

$$\sum_{j=1}^{M} \gamma_j^2 \mathrm{Var}(x_{\cdot j}) + \sum_{j \neq k}^{M} \gamma_j \gamma_k \mathrm{Cov}(x_{\cdot j}, x_{\cdot k}) + \mathrm{Var}(\epsilon)$$





Estimated genetic correlation from real data = 0.12 (data from Lukowski et al., 2017)

(Liu, Li, and Pritchard, bioRxiv, 2018)

Summary: Part 1

Three main observations suggest that:

- Most heritability is due to SNPs outside core genes/pathways.
- All genes expressed in disease-relevant cells may affect the regulation of core genes. Thus, most effects are "peripheral".

We refer to this hypothesis as an ''omnigenic'' model.

(Boyle, Li, and Pritchard, 2017)

Summary: Part 2

The omnigenic model is consistent with known properties of *cis*- and *trans*-eQTLs (and likely other types of molecular QTLs)

- Variation in *trans* is responsible for ~70% of gene expression heritability (implying that the bulk of complex trait heritability are likely to be explained by trans-effects).

- *Trans* effect sizes are nearly all tiny, which implies that there are a huge number of contributing genetic loci/peripheral gene.

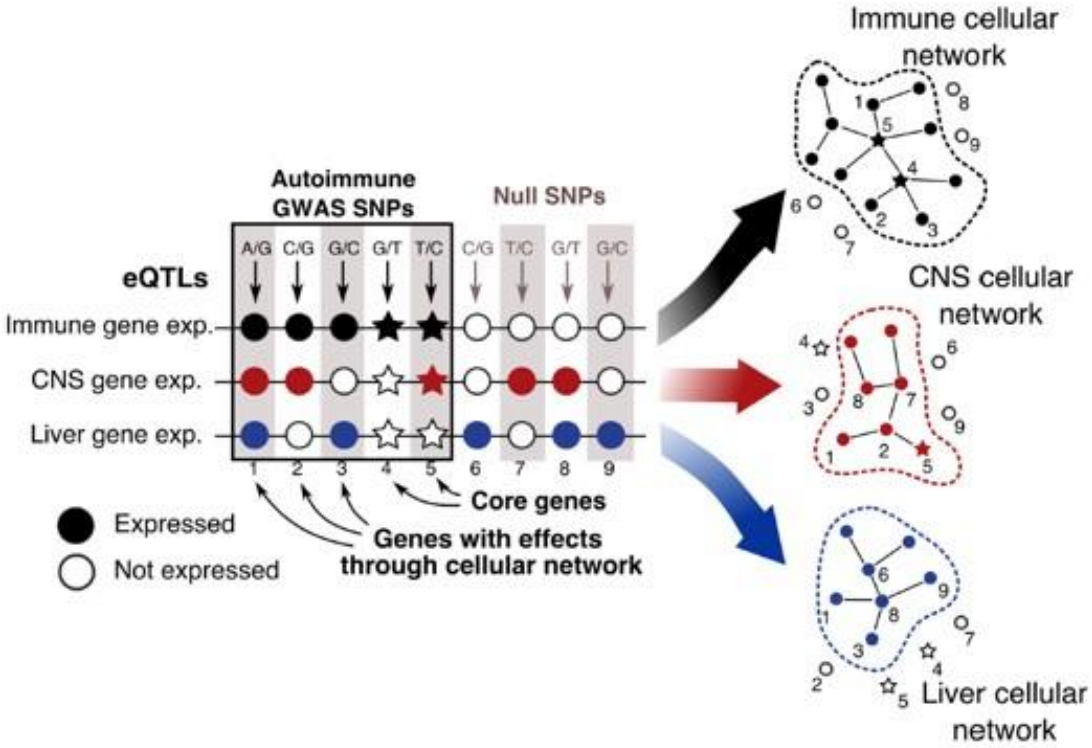
- Co-regulation of core gene expression can amplify the contribution of trans-variation/peripheral genes.

(Liu, Li, and Pritchard, bioRxiv, 2018)

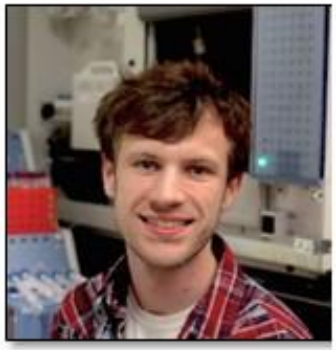
Happy to take questions!

Acknowledgments

Arbel Harpak (Columbia)
William Greenleaf (Stanford)
Alex Marson (UCSF)
Joe Pickrell (Gencove)
Guy Sella (Columbia)
Molly Przeworski (Columbia)



Stanford collaborators:

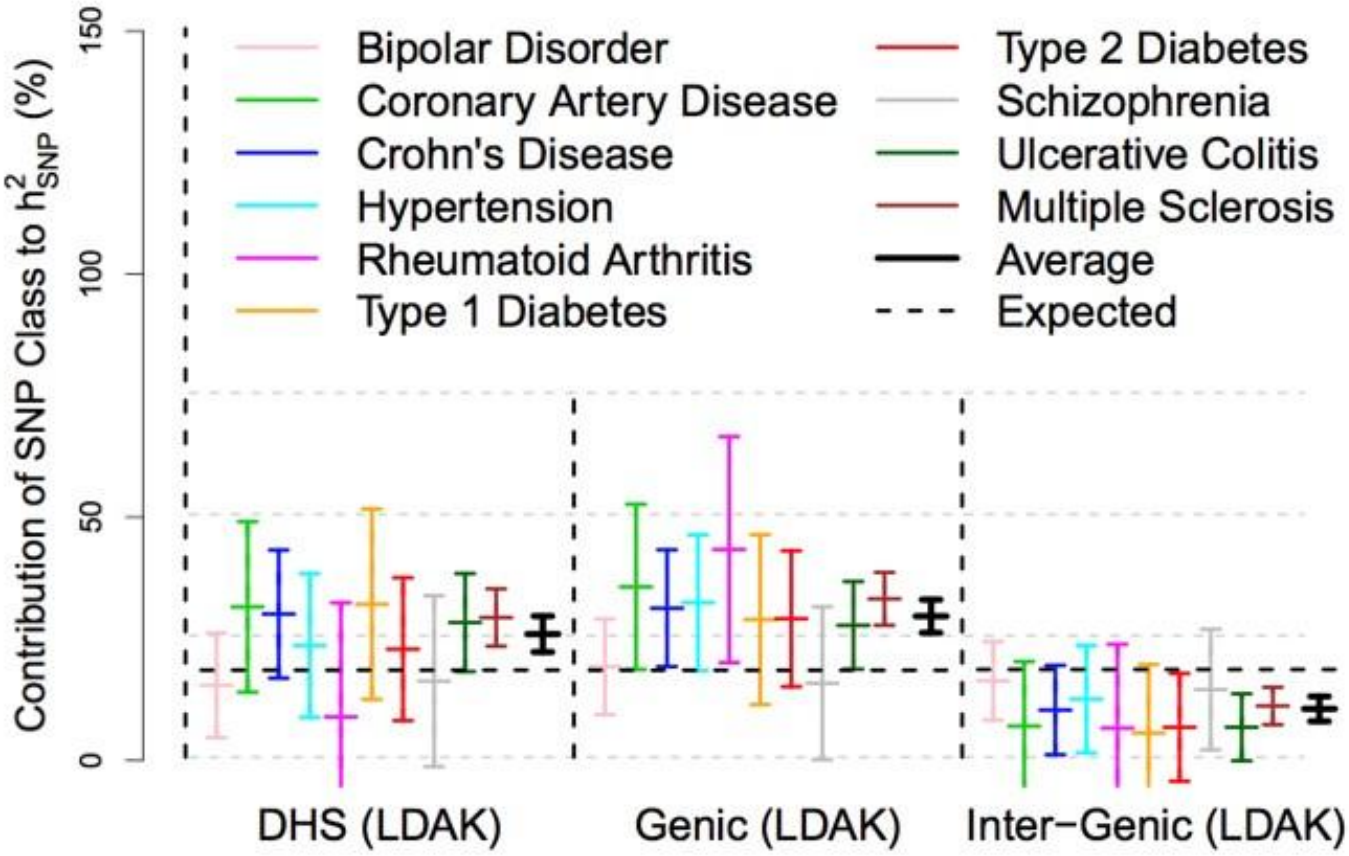


Evan Boyle

Jonathan Pritchard

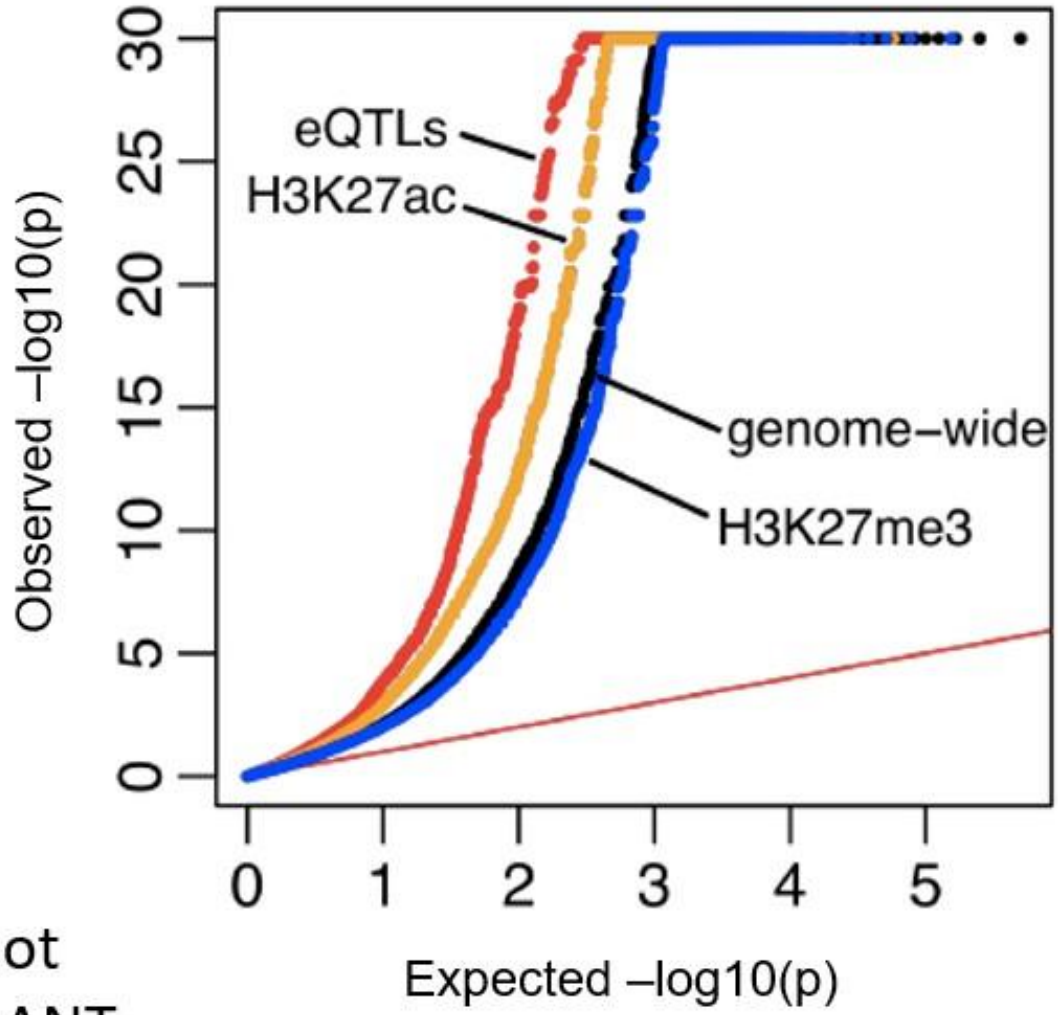
Xuanyao Liu

24% of heritability is explained by SNPs in DNase-I sites



Speed et al., Nat. Genet., 2016

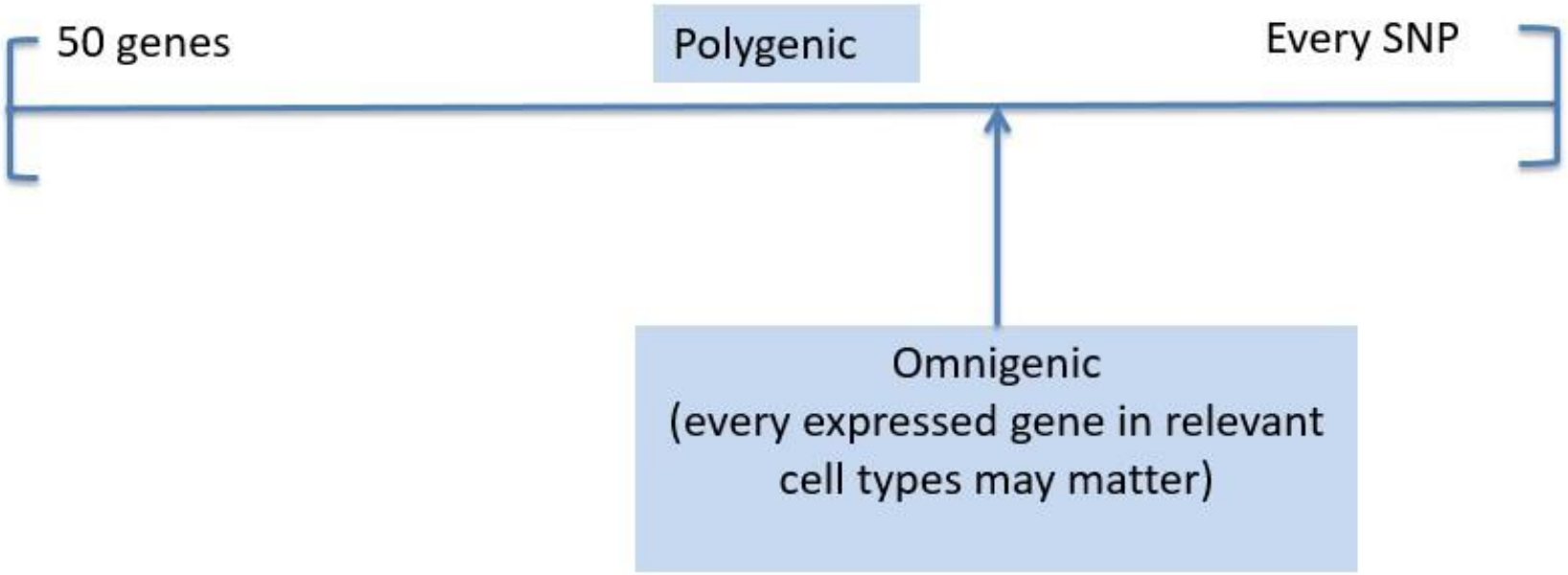
GWAS hits are enriched among functional SNPs



QQ-plot
Data: GIANT
(height)

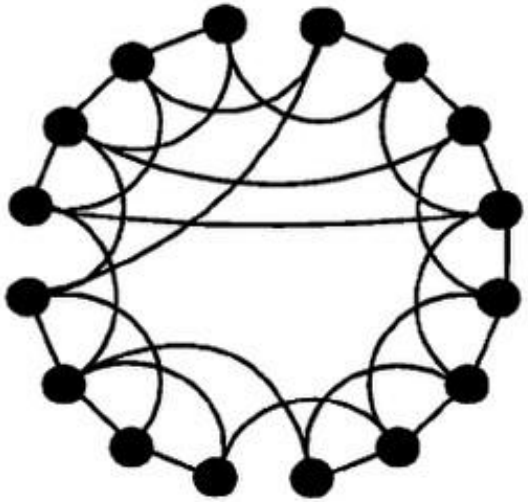
Li et al., Science, 2016
Boyle, Li, and Pritchard, Cell, 2017

We think of “Omnigenic” as having a more precise meaning than “polygenic”

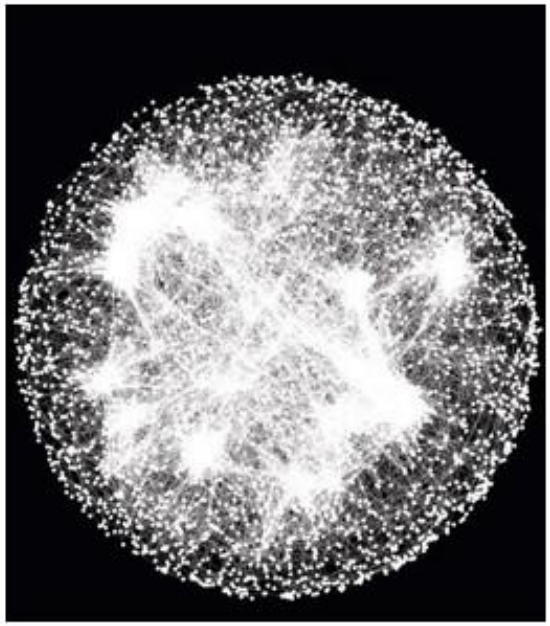


Small world property of networks: most nodes can be reached from every other node by a small number of steps

"Small world" property:
(Watts & Strogatz, 1998, Strogatz 2001)

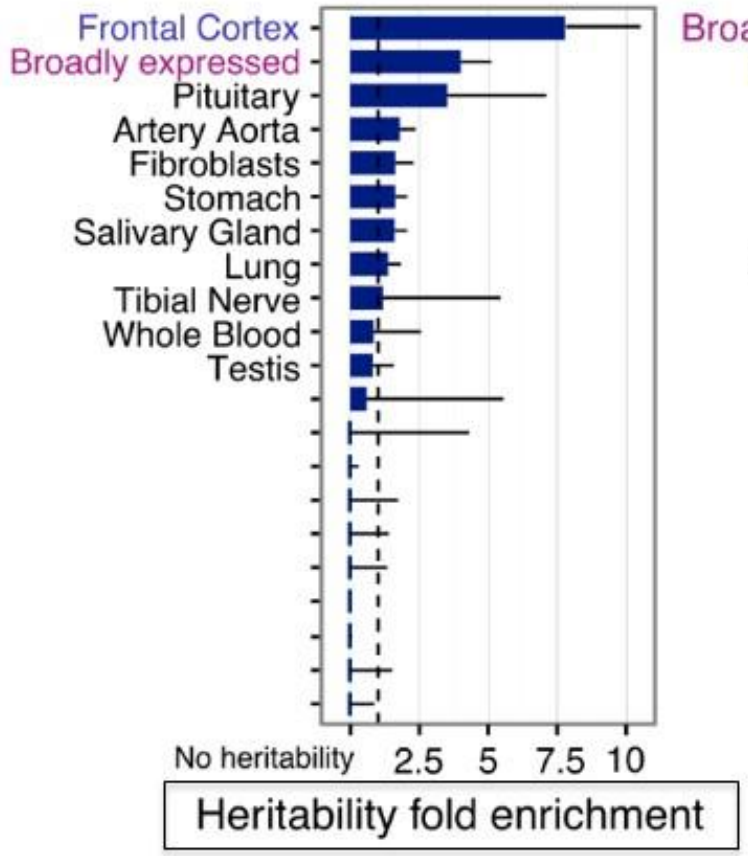


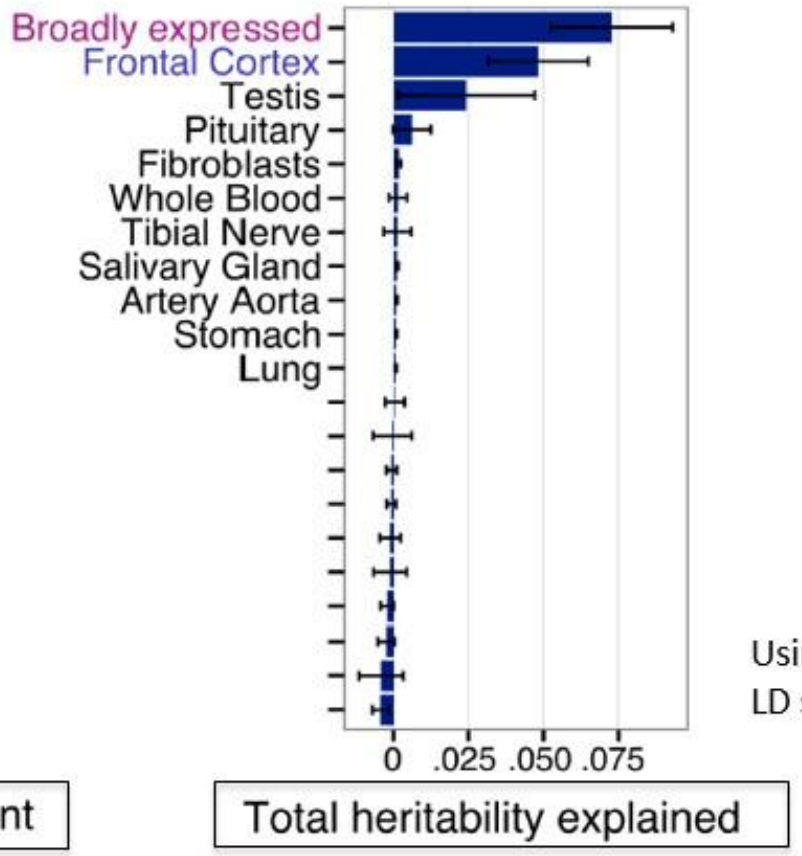
All possible pairwise genetic interactions
in yeast (Costanzo et al., 2016)



- Suggests that peripheral genes may be "close" to core genes.
- Network could be transcriptional, PPI, etc…

But SNPs in broadly expressed genes explain more total heritability!





Using stratified LD score regression

- *Genes that do not have a direct function in disease pathways might play a large role in disease*