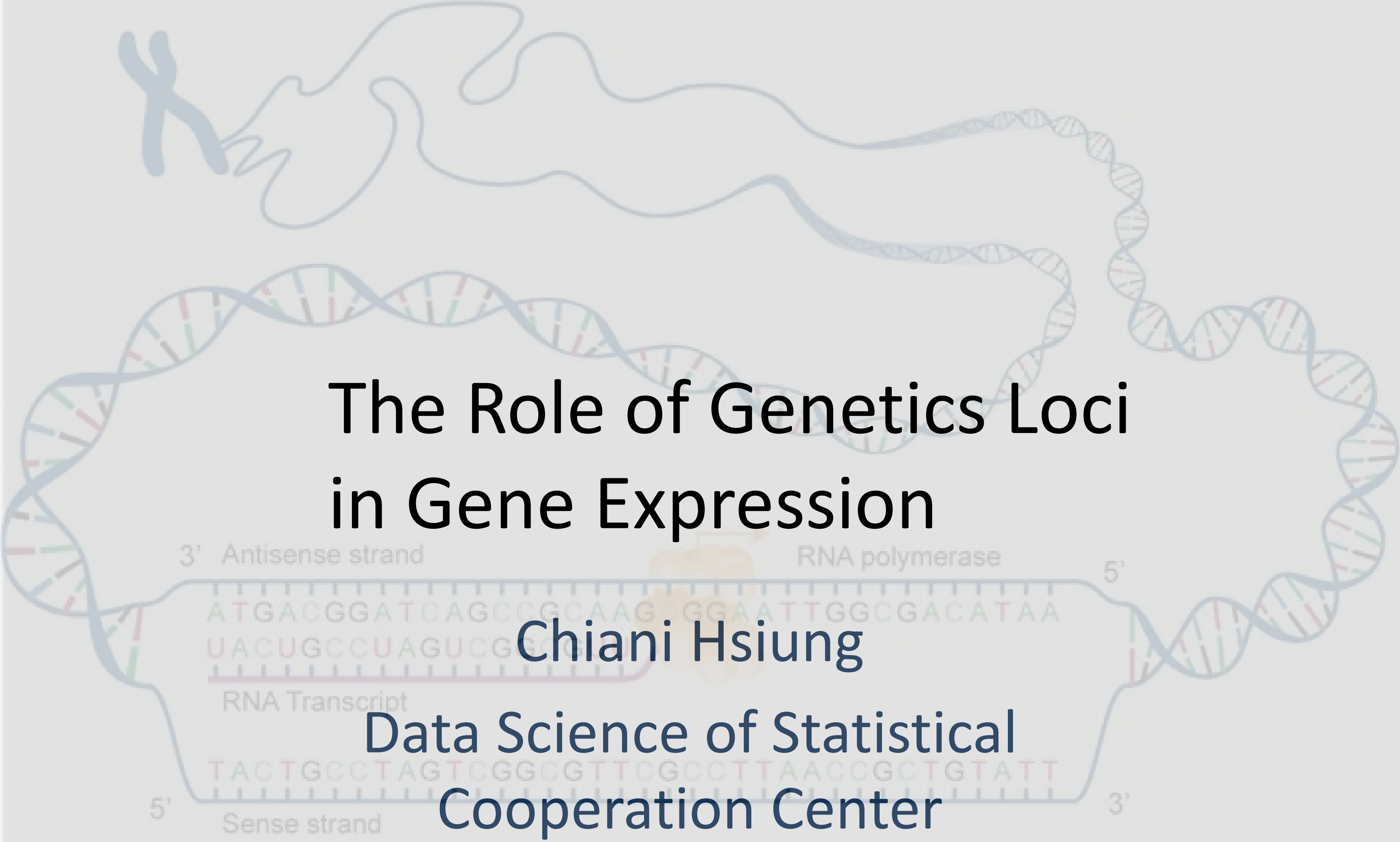

3'
Antisense strand
RNA polymerase
5'
ATGACGGATCAGCCGCAAGCGGAATTGGCGACATAA
RNA Transcript
TACTGCCTAGTCGGCGTTCGCCTTAACCGCTGTATT
5'
Sense strand
3'

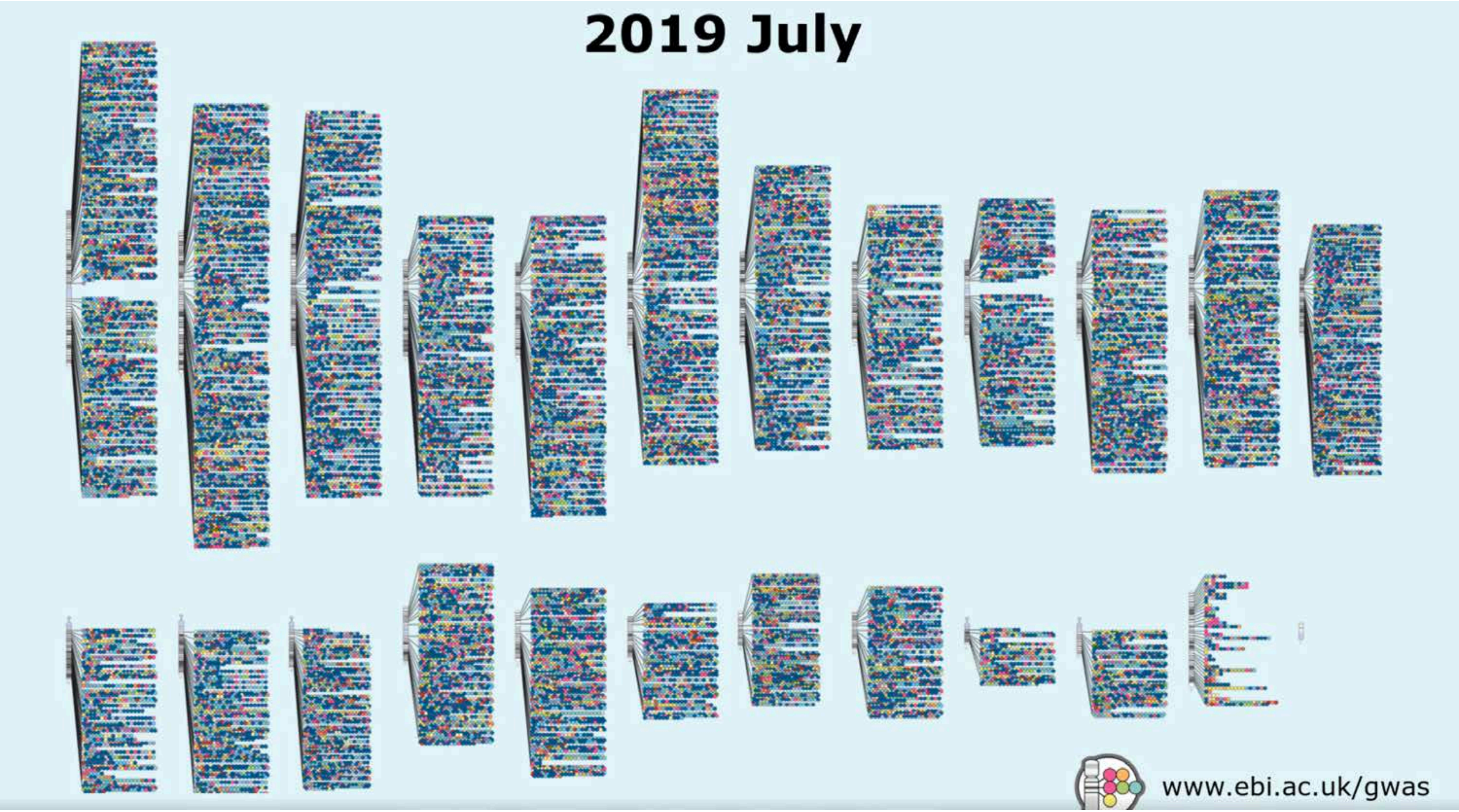
The Role of Genetics Loci in Gene Expression

Chiani Hsiung

Data Science of Statistical Cooperation Center

1 GENETICS OF GENE EXPRESSION

Chromosome map of disease-associated regions



“GWAS have so far identified only a small fraction of the heritability of common diseases, so the ability to make meaningful predictions is still quite limited”

Francis Collins, Director of the NIH, *Nature*, April 2010

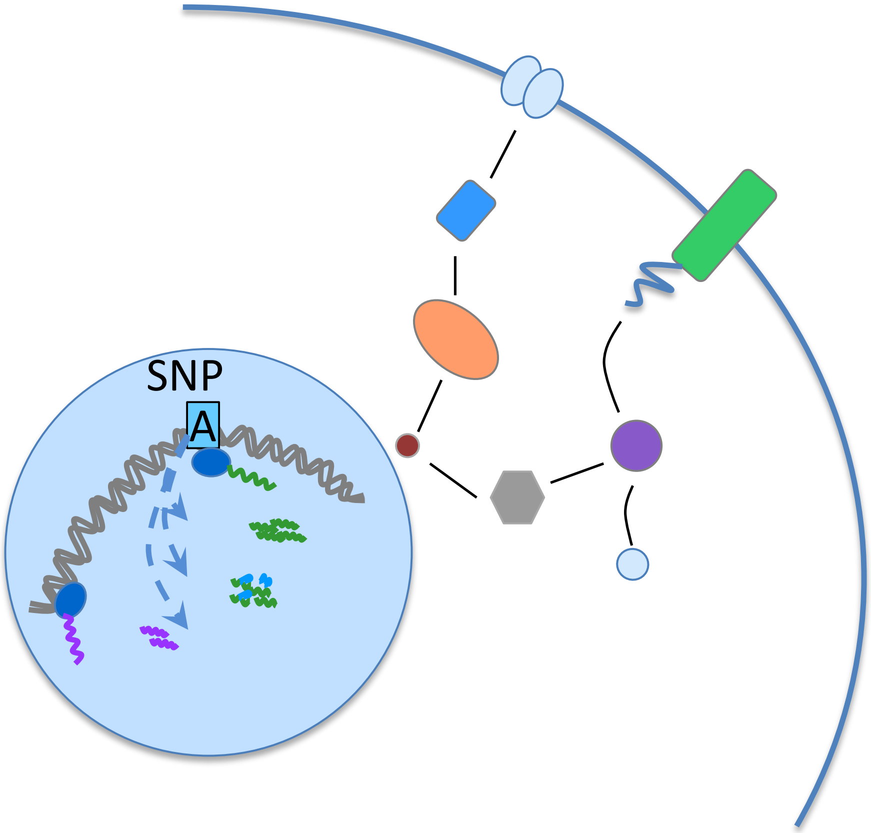
Trait	Heritability (Family base)	Individuals studied	Heritability explained
Coronary artery disease	40%	86995	10%
Type 2 Diabetes	40%	47117	10%
BMI	50%	249796	3%
Blood pressure	50%	34433	1%
Circulating lipids	50%	100000	25%
Height	80%	183727	12.5%

Motivation

How can we use gene expression and epigenetics to help us understand complex trait genetics?

Majority of trait-associated variation is non-coding.

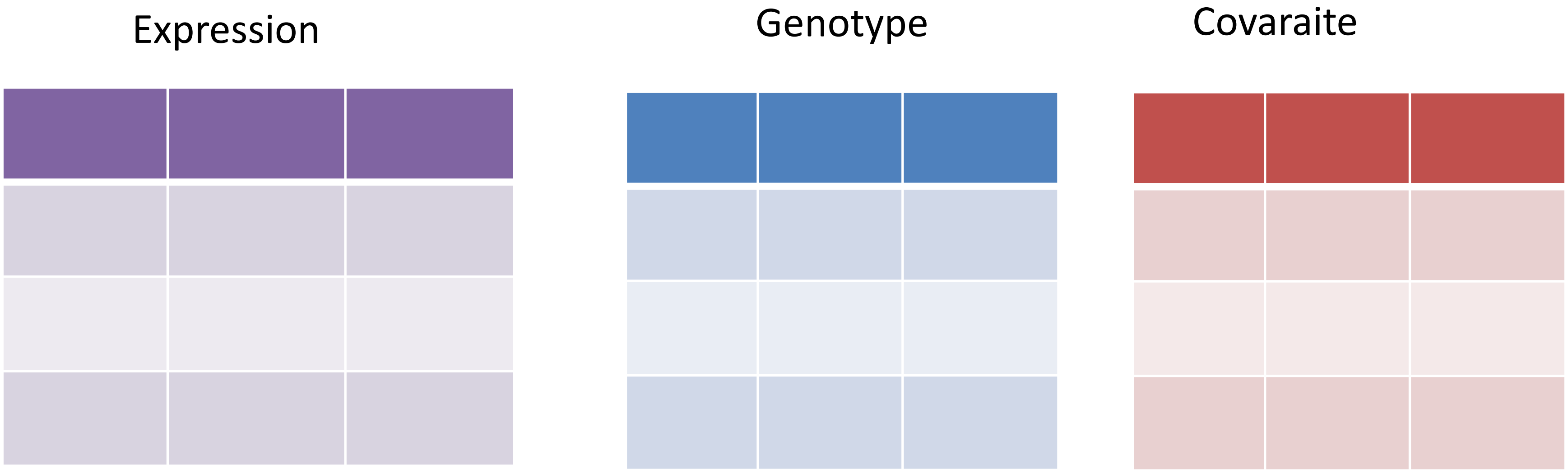
Common hypothesis is that most of these function by altering gene expression.



eQTL analysis Statistics

- Regression: find the coefficients for the effect of expression on genotype when conditioned on the covariates in a linear model and test if they are significant diffetent than 0

$$gene\ expression = \beta_0 + \beta_1 genotype + \beta_2 covaraites$$



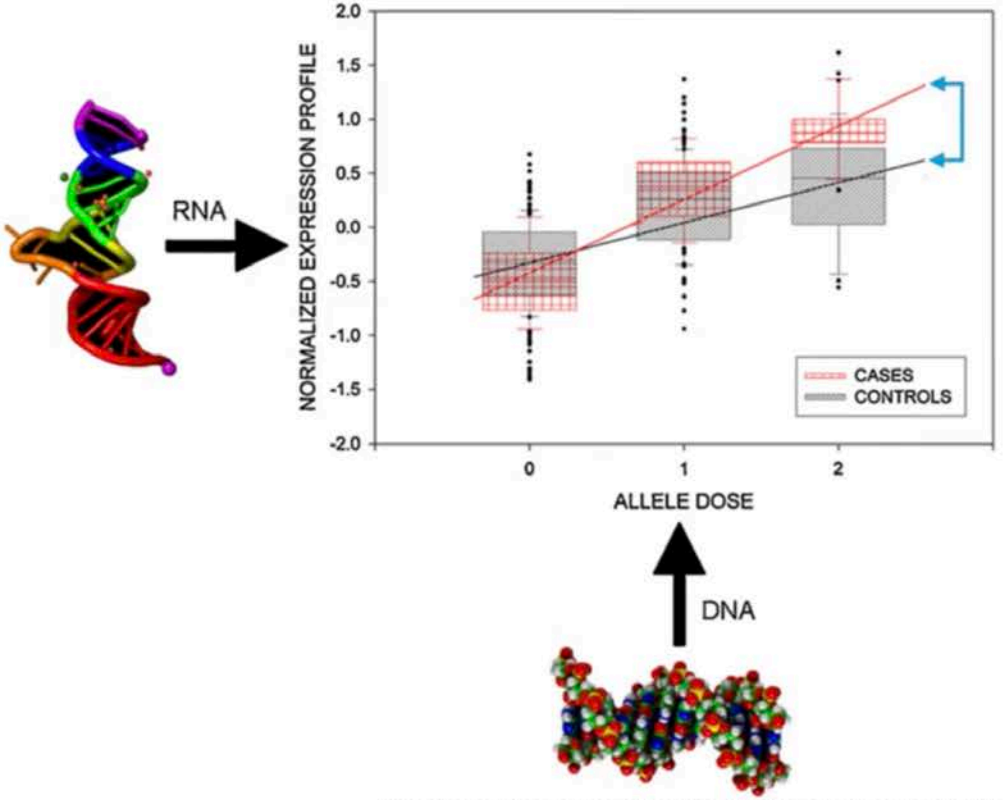
Expression

Genotype

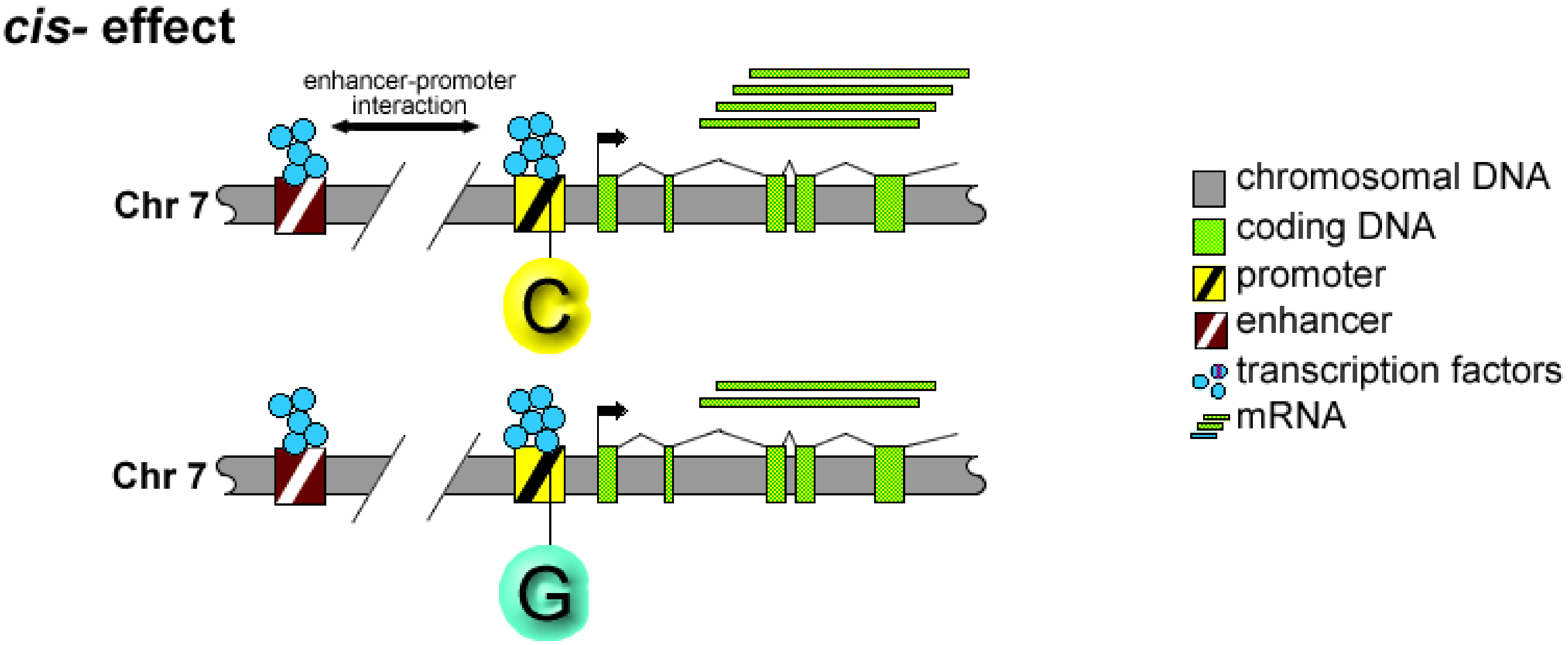
Covaraite

Mapping expression (e)QTL

- RNA expression levels can be treated like any other quantitative trait in QTL mapping.
- 30,000 genes by 10,000 SNPs = 300,000,000 comparisons!
- eQTL studies are sometimes called genetical genomics



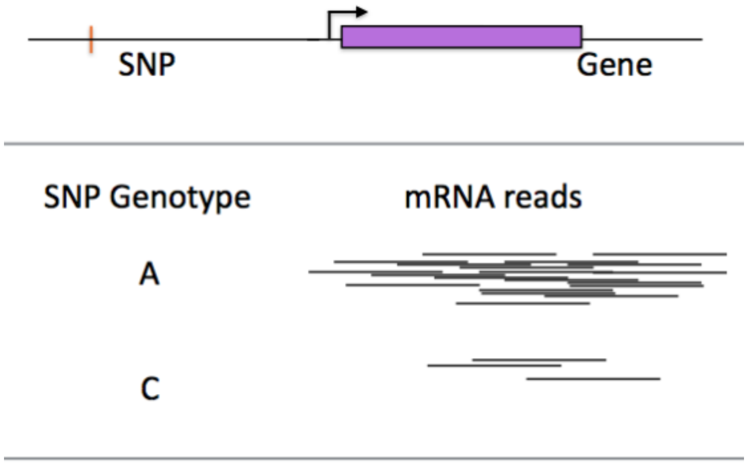
Myers, AJ. The age of the "ome": Genome, transcriptome and proteome data set collection and analysis. Brain Research Bulletin Volume 88, Issue 4 2012 294 - 301

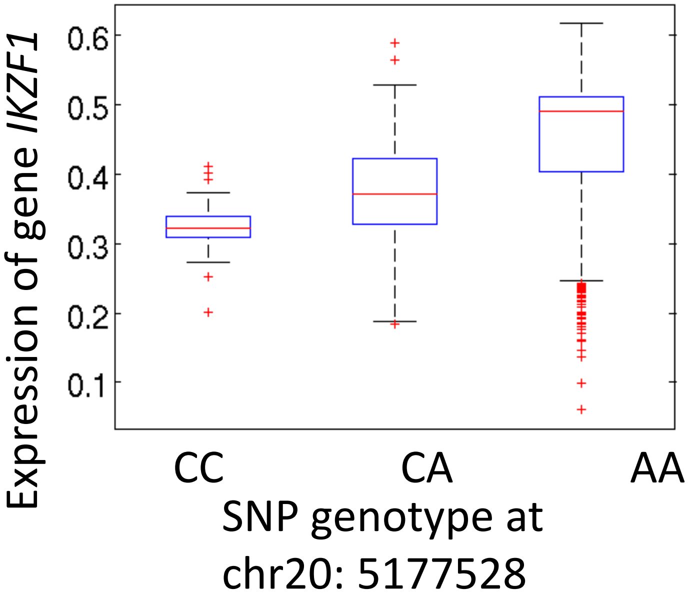


Canonical model

Genetic variants affect gene expression

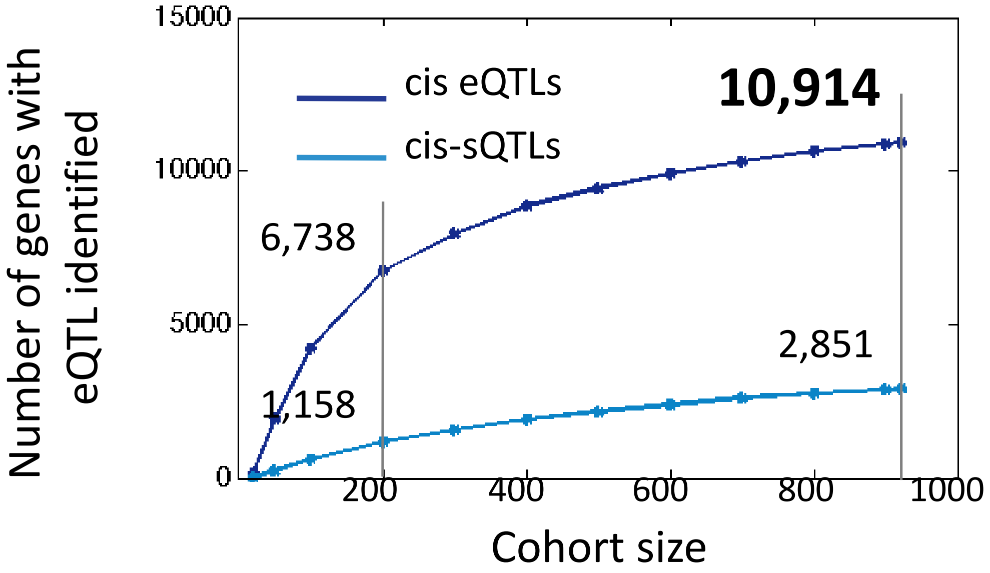
<u>eQTL (expression Quantitative Trait Locus) analysis</u>:
Association between genotype and RNA expression levels





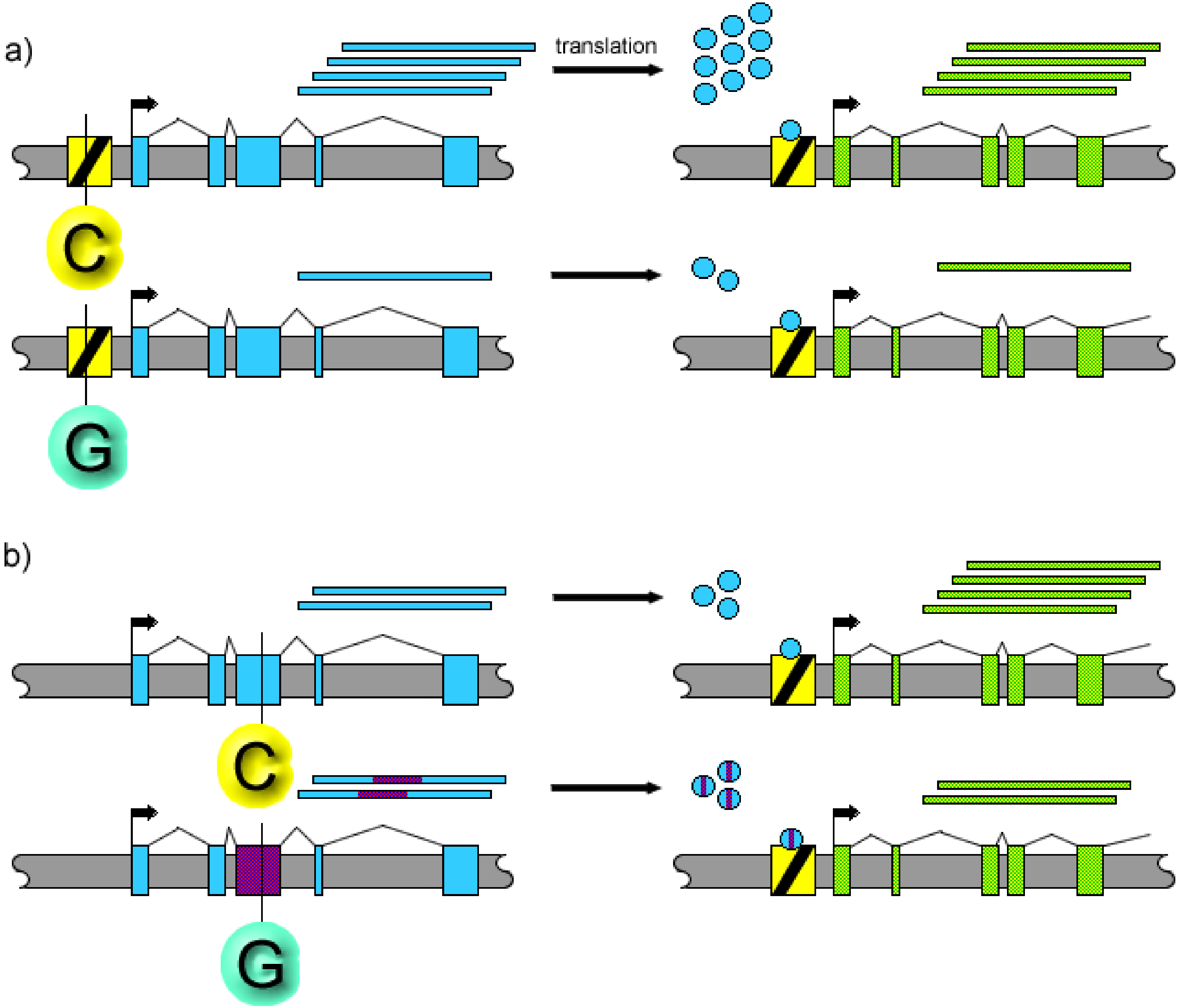
Prevalence of eQTLs

Cis-eQTLs have now been identified for nearly every human gene, with numerous large studies available

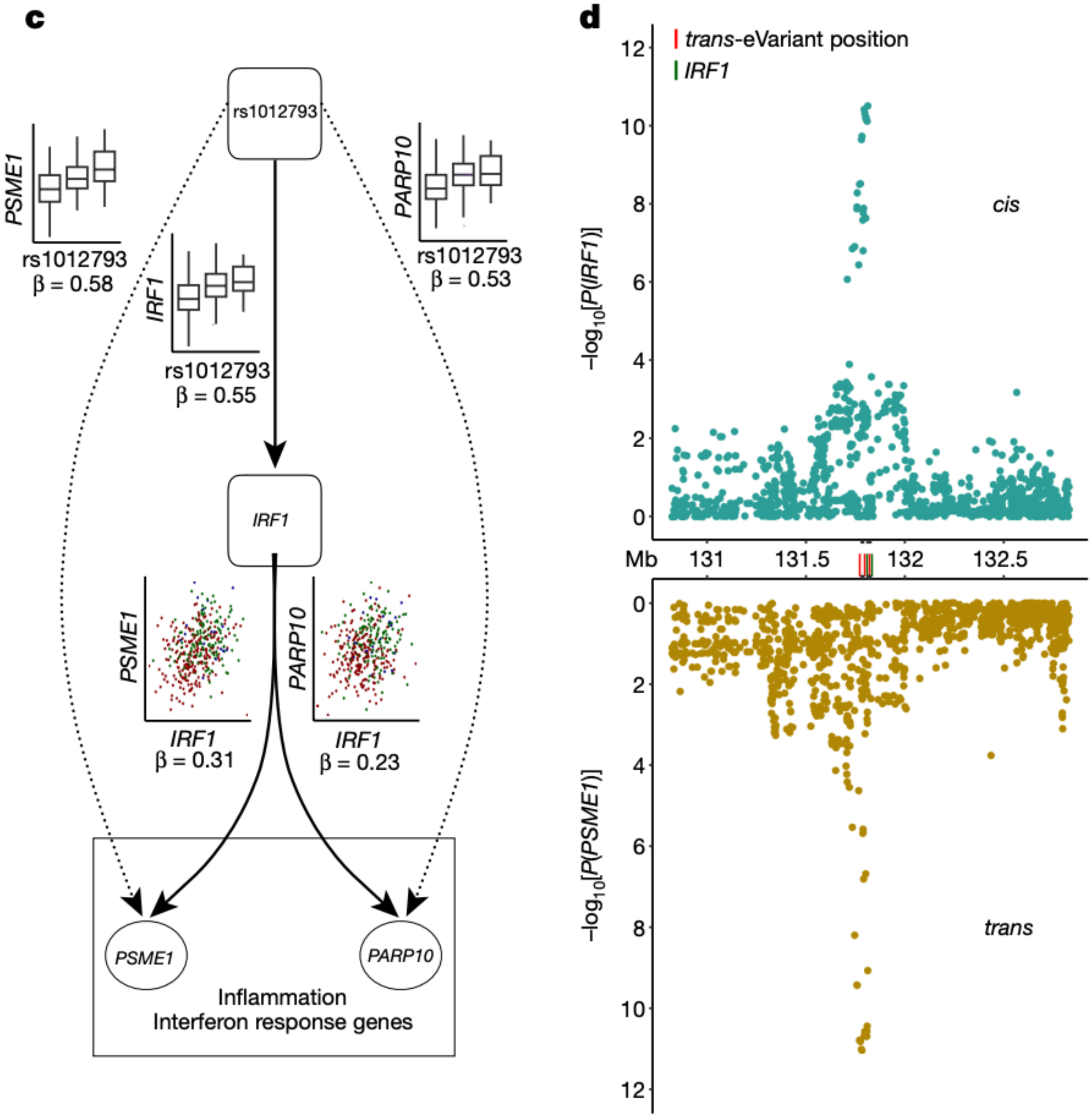


Battle, Genome Research, 2014

trans- effect



Trans-eQTL example

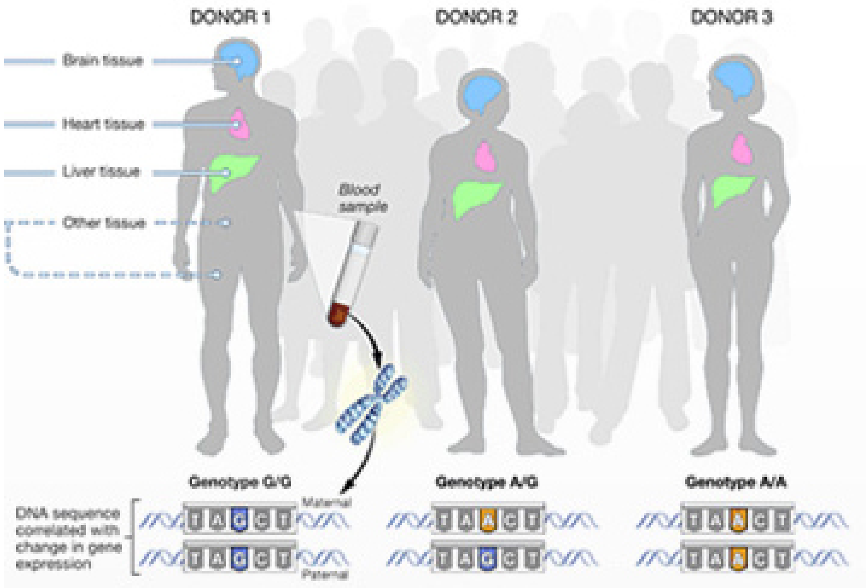


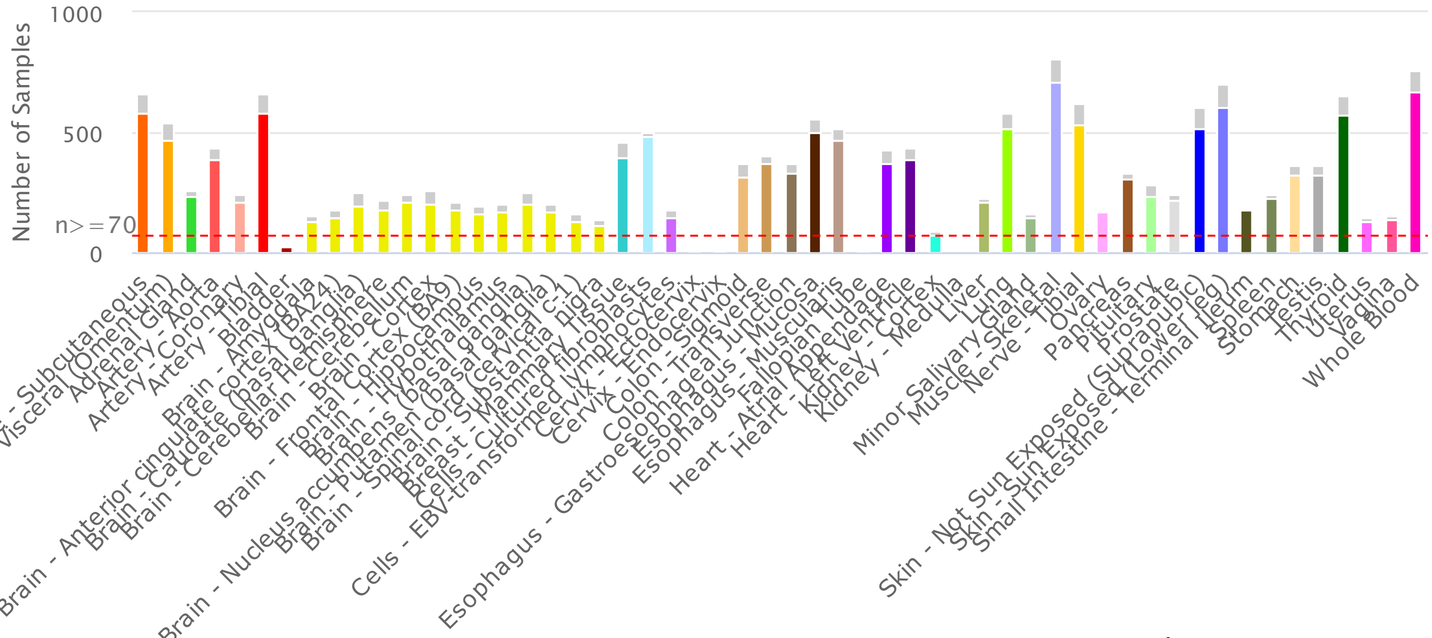
The GTEx Consortium, 2017

GTEx Project

GTEx Consortium v8 data

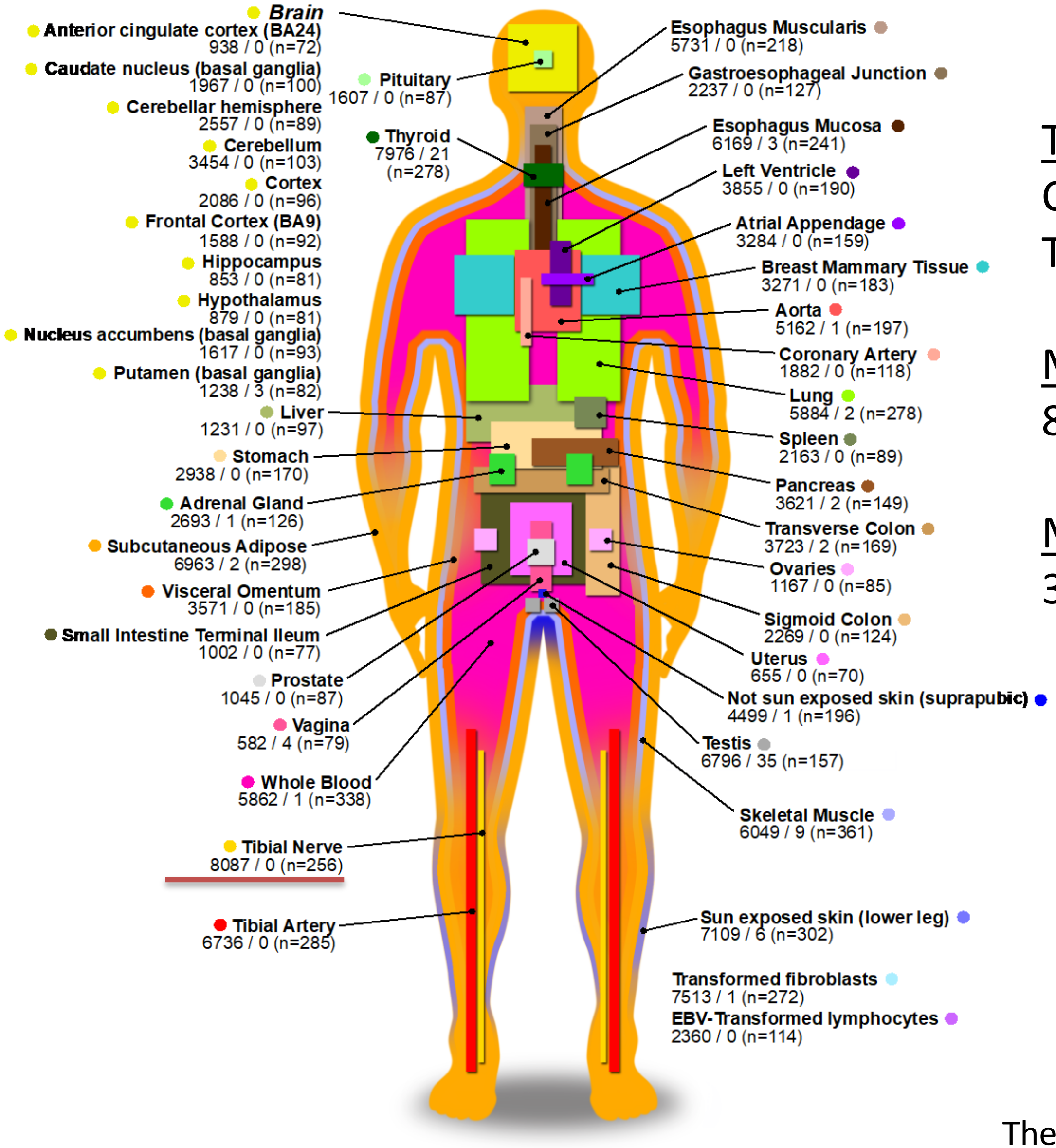
- 838 genotyped donors
- 17832 gene expression samples





The GTEx Consortium

Genetic effects across human tissues



Total unique eQTL genes:

Cis: 19,725 (FDR 5%)

Trans: 93 (FDR 10%)

Most cis per tissue:

8,087 Tibial nerve (N=256)

（脛神經）

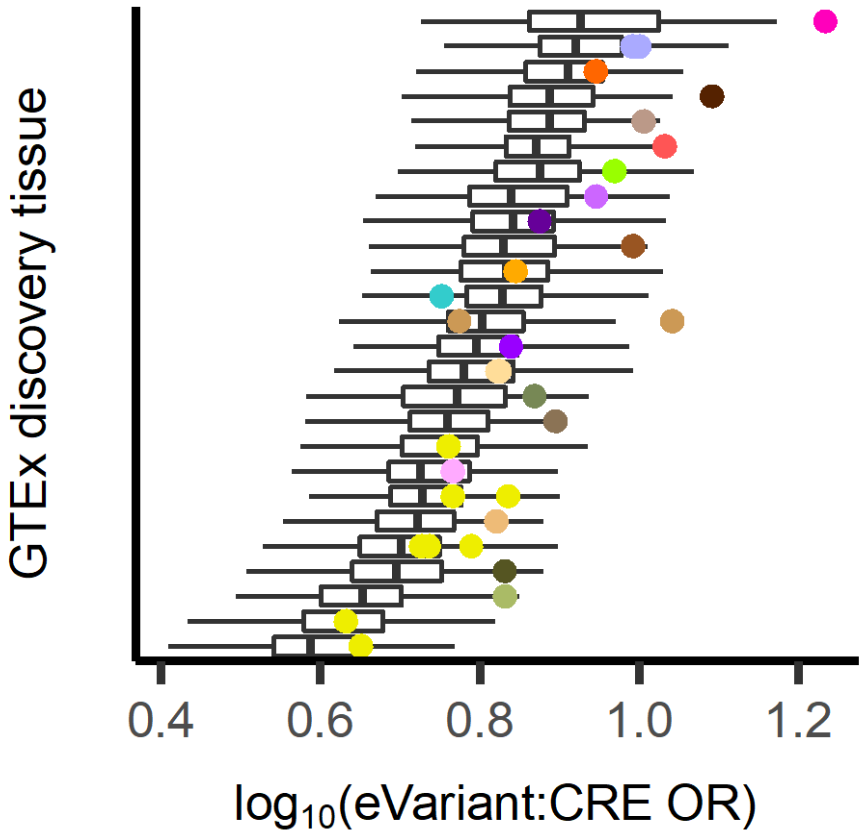
Most trans per tissue:

35 Testis (N=157)

The GTEx Consortium, Nature 2017

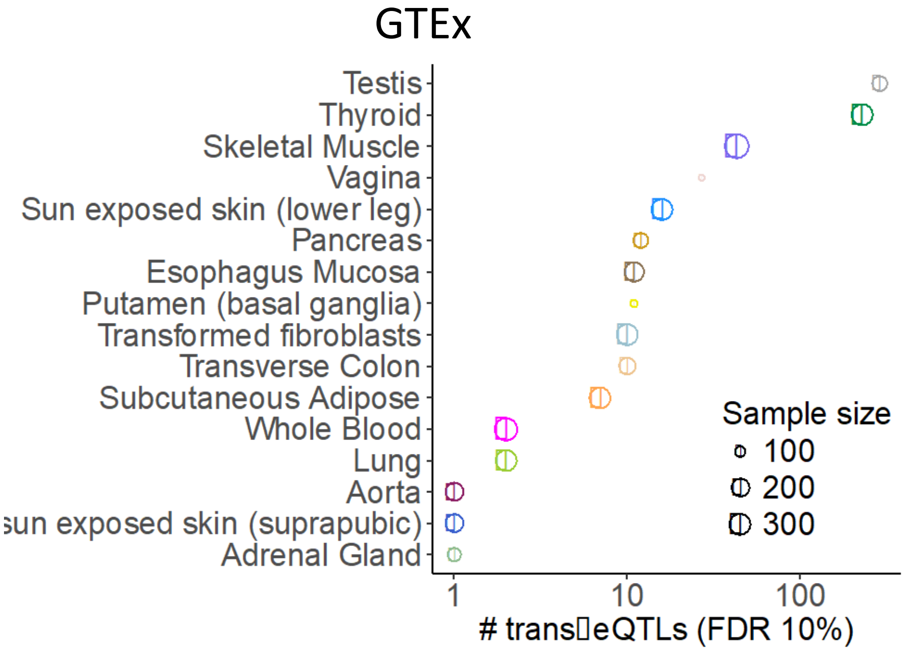
Characterizing eQTLs across tissues

- Cis-eQTL variants fall in tissue-specific regulatory elements (from Roadmap Epigenomics)



The GTEx Consortium, Nature 2017

Trans-eQTLs



Large studies:

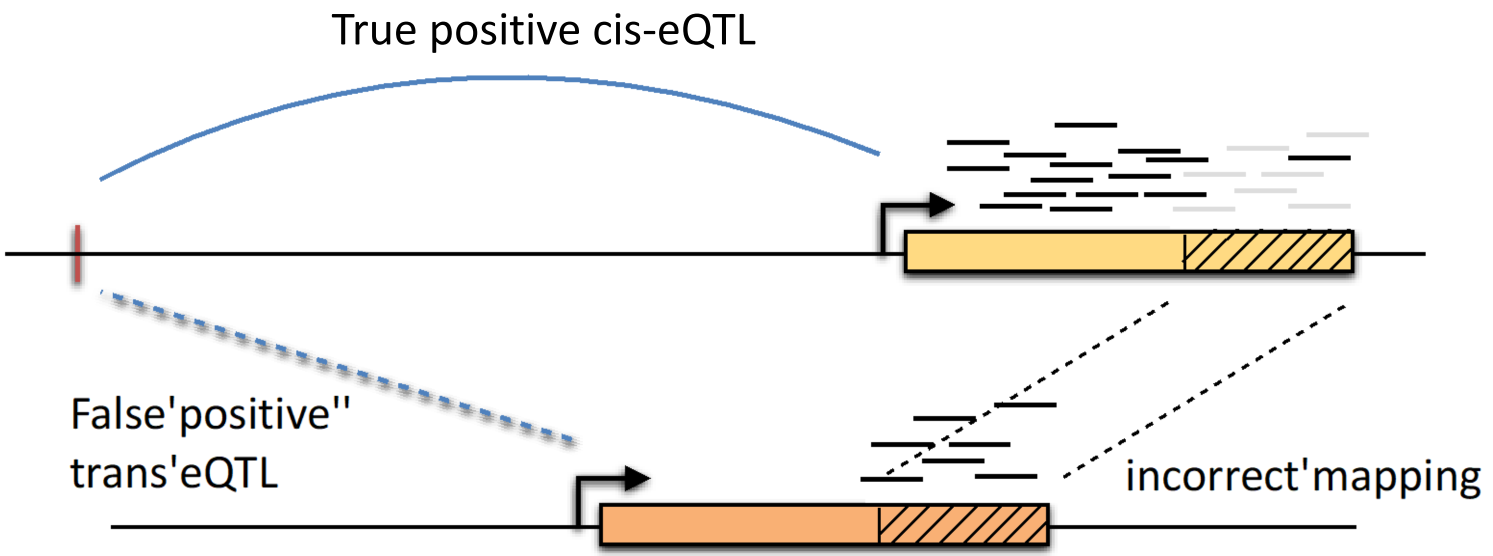
Westra et al (N=5,311, using GWAS variants only)
ALSPAC (N=869)
MUTHER (N=850)
DGN (N=922)
Framingham (N=5257)

Studies report wildly different # hits (10s–10000s)

Replication and validation remains poor

Challenges for trans-eQTL detection

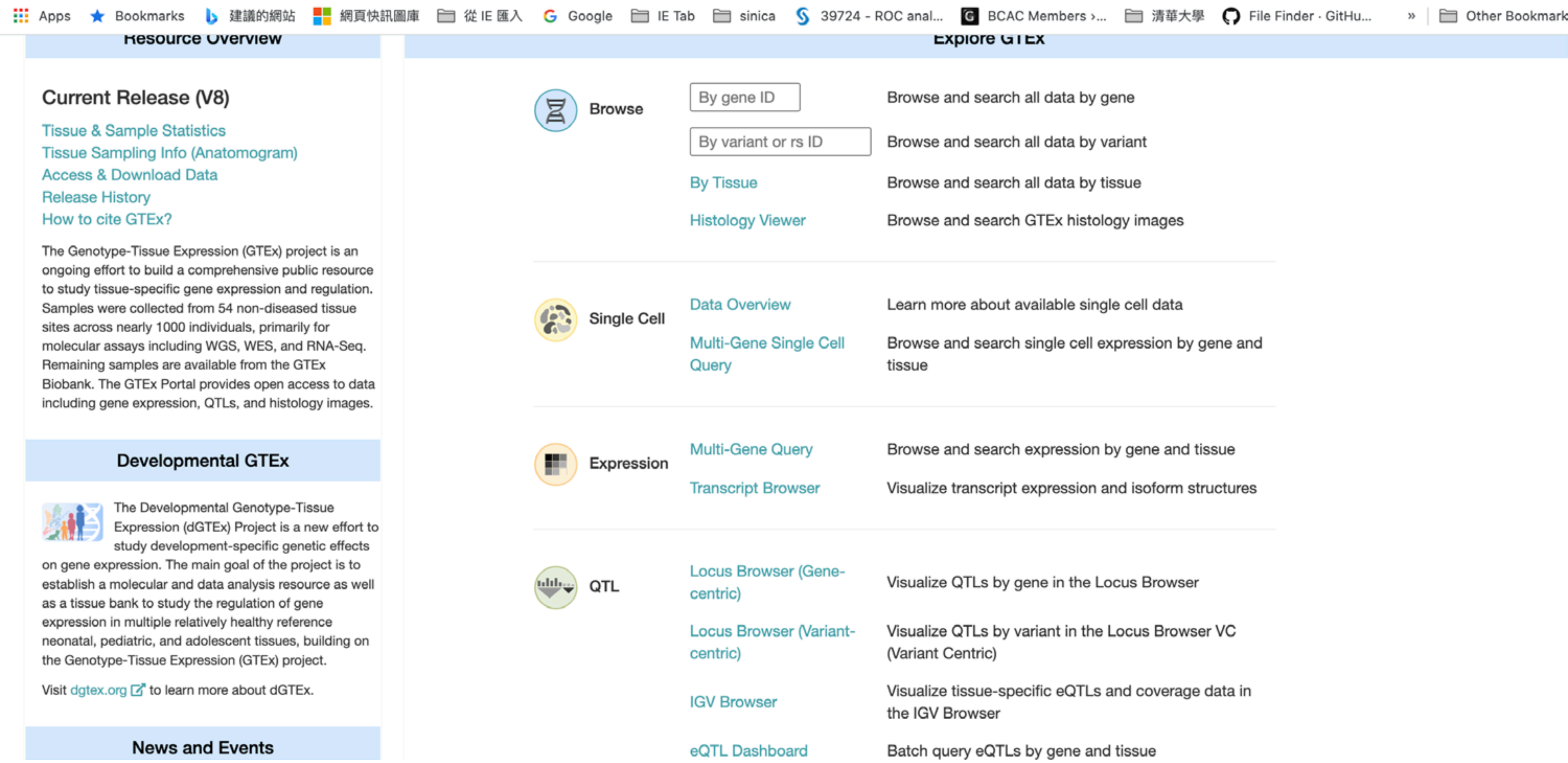
- Power
- False positives from many sources e.g. over and under correcting confounders (Dahl et al, 2017)
- Mapping error (similar to probe cross-hybrid.)



The GTEx Consortium, Nature 2017

Introduction of GTEx protal

- https://gtexportal.org/home/



Resource Overview

Current Release (V8)

Tissue & Sample Statistics
Tissue Sampling Info (Anatomogram)
Access & Download Data
Release History
How to cite GTEx?

The Genotype-Tissue Expression (GTEx) project is an ongoing effort to build a comprehensive public resource to study tissue-specific gene expression and regulation. Samples were collected from 54 non-diseased tissue sites across nearly 1000 individuals, primarily for molecular assays including WGS, WES, and RNA-Seq. Remaining samples are available from the GTEx Biobank. The GTEx Portal provides open access to data including gene expression, QTLs, and histology images.

Developmental GTEx

The Developmental Genotype-Tissue Expression (dGTEx) Project is a new effort to study development-specific genetic effects on gene expression. The main goal of the project is to establish a molecular and data analysis resource as well as a tissue bank to study the regulation of gene expression in multiple relatively healthy reference neonatal, pediatric, and adolescent tissues, building on the Genotype-Tissue Expression (GTEx) project.

Visit dgtex.org to learn more about dGTEx.

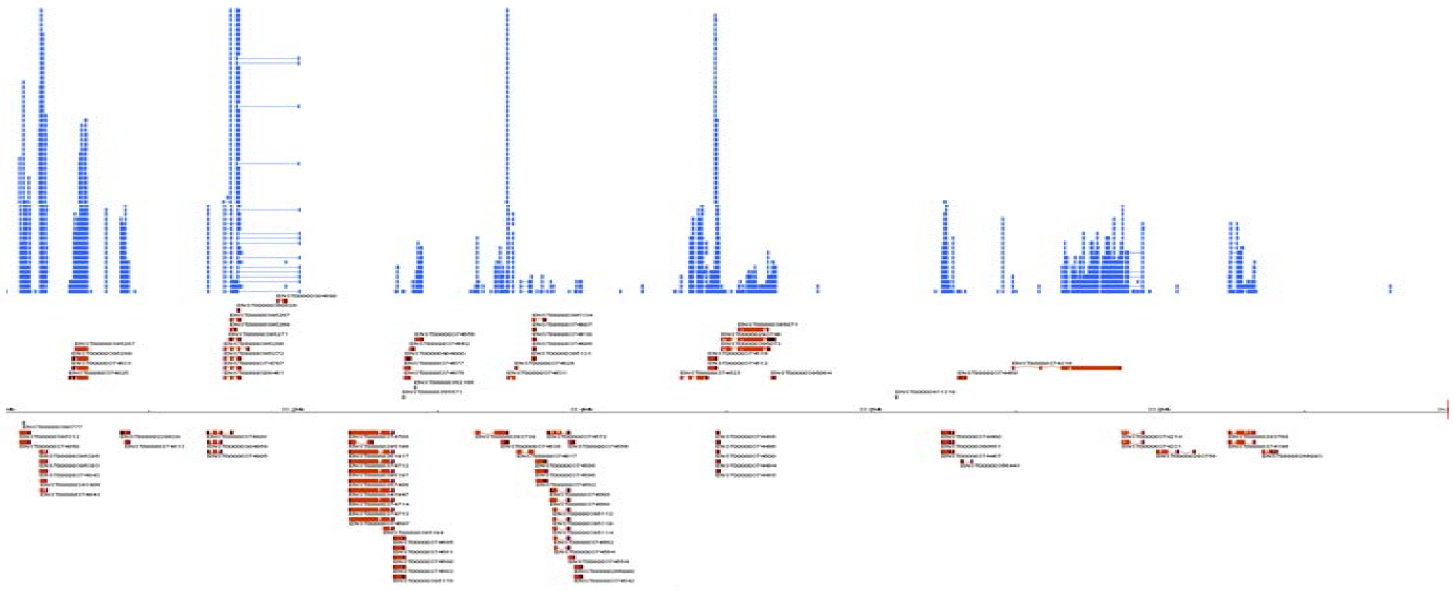
News and Events

Explore GTEx

Browse	By gene ID	Browse and search all data by gene
	By variant or rs ID	Browse and search all data by variant
	By Tissue	Browse and search all data by tissue
	Histology Viewer	Browse and search GTEx histology images
Single Cell	Data Overview	Learn more about available single cell data
	Multi-Gene Single Cell Query	Browse and search single cell expression by gene and tissue
Expression	Multi-Gene Query	Browse and search expression by gene and tissue
	Transcript Browser	Visualize transcript expression and isoform structures
QTL	Locus Browser (Gene-centric)	Visualize QTLs by gene in the Locus Browser
	Locus Browser (Variant-centric)	Visualize QTLs by variant in the Locus Browser VC (Variant Centric)
	IGV Browser	Visualize tissue-specific eQTLs and coverage data in the IGV Browser
	eQTL Dashboard	Batch query eQTLs by gene and tissue

2 MANY TYPES OF QTLS

Next generation sequencing has increased our ability to survey the transcriptome.

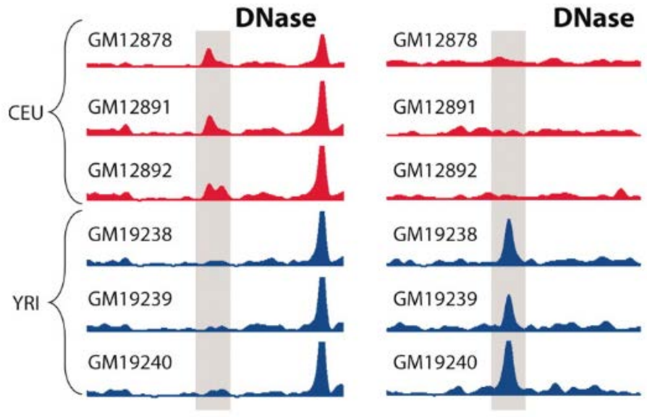


RNA-Seq

Montgomery, Nature 2010
Pickrell, Nature 2010

ChIP-Seq

McDaniell, Science 2010

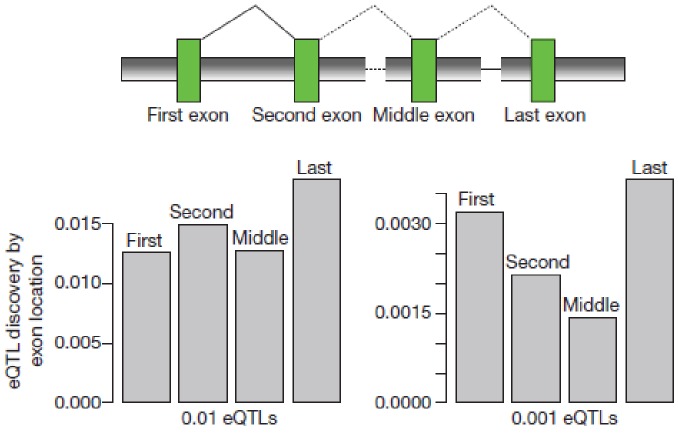


RNA-seq provides resolution of more QTLs

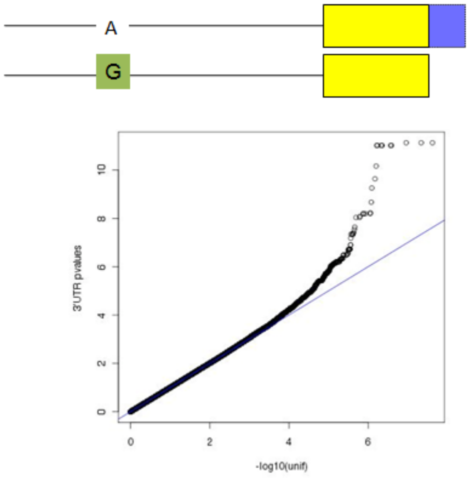
RNA-sequencing in 60 Europeans (HapMap genotypes; LCLs)

Found 2x more expression Quantitative Trait Loci (eQTLs) and...

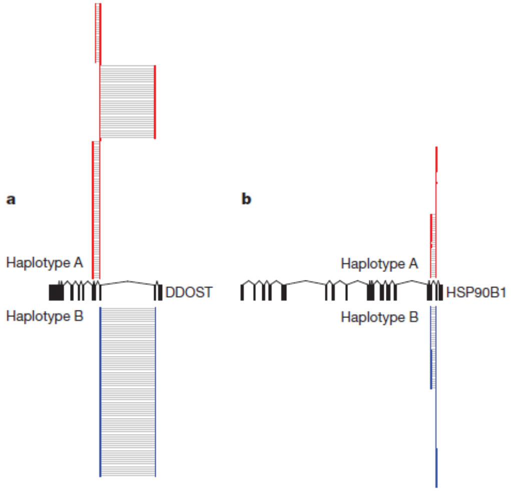
Exon-eQTLs



UTR Length-QTLs



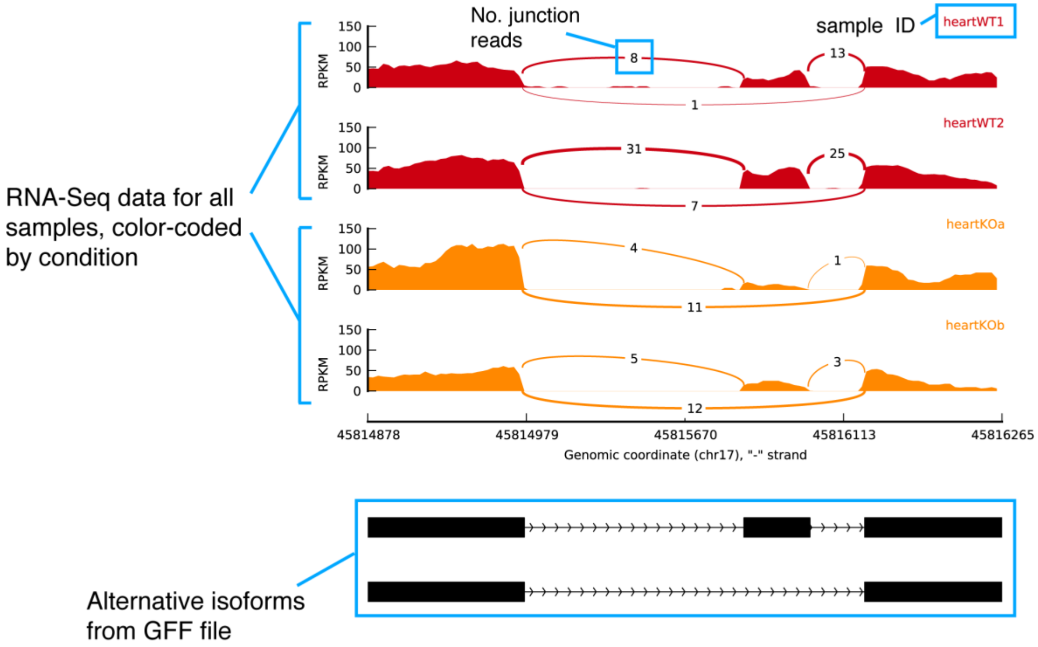
Splicing eQTLs



Rare eQTLs with allele specific expression-based approaches

Splicing eQTL

Can investigate relative transcript ratios or reads across junctions.

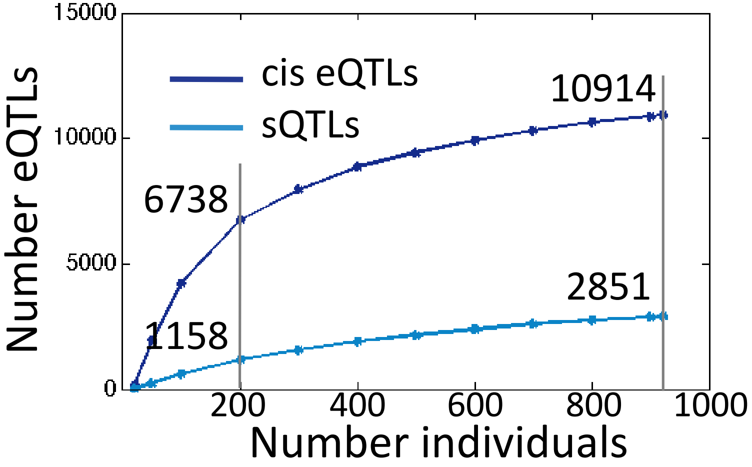


RNA-Seq data for all samples, color-coded by condition

Alternative isoforms from GFF file

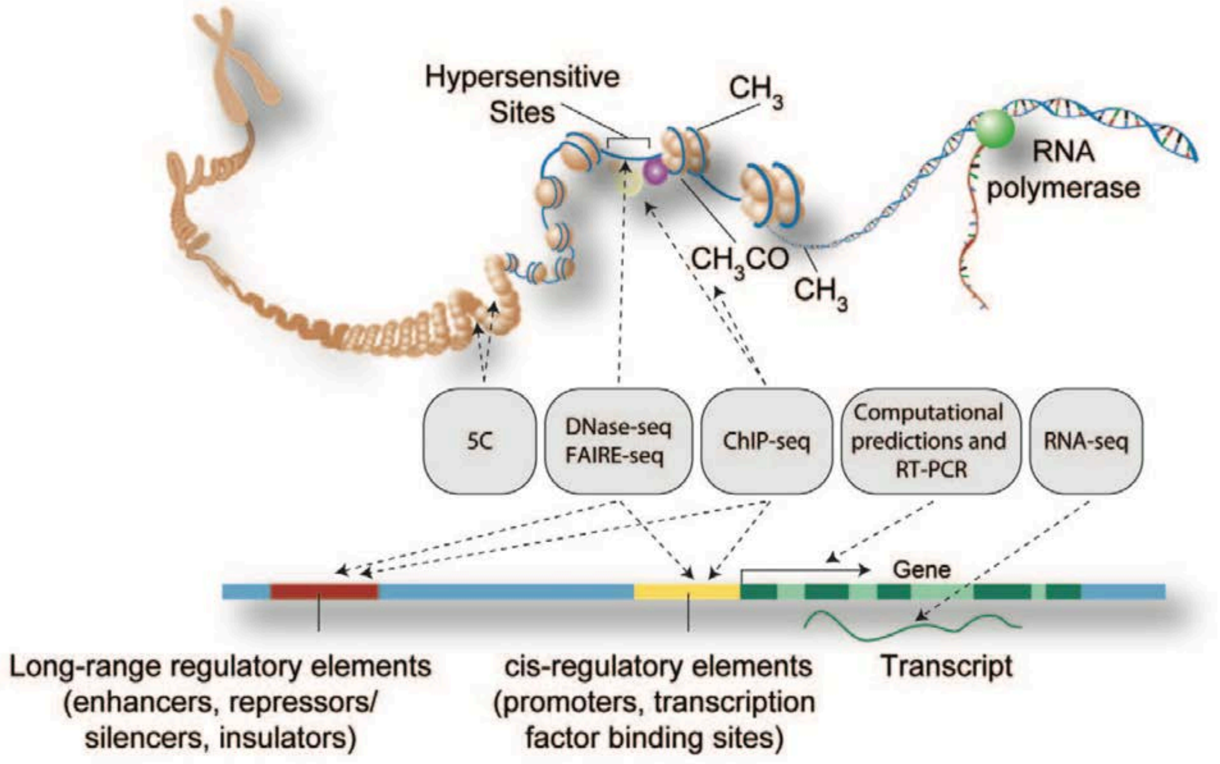
Katz et al, Nature Methods, 2010

- Splicing also affected for many genes



Battle et al, Genome Research, 2014

Epigenetic data



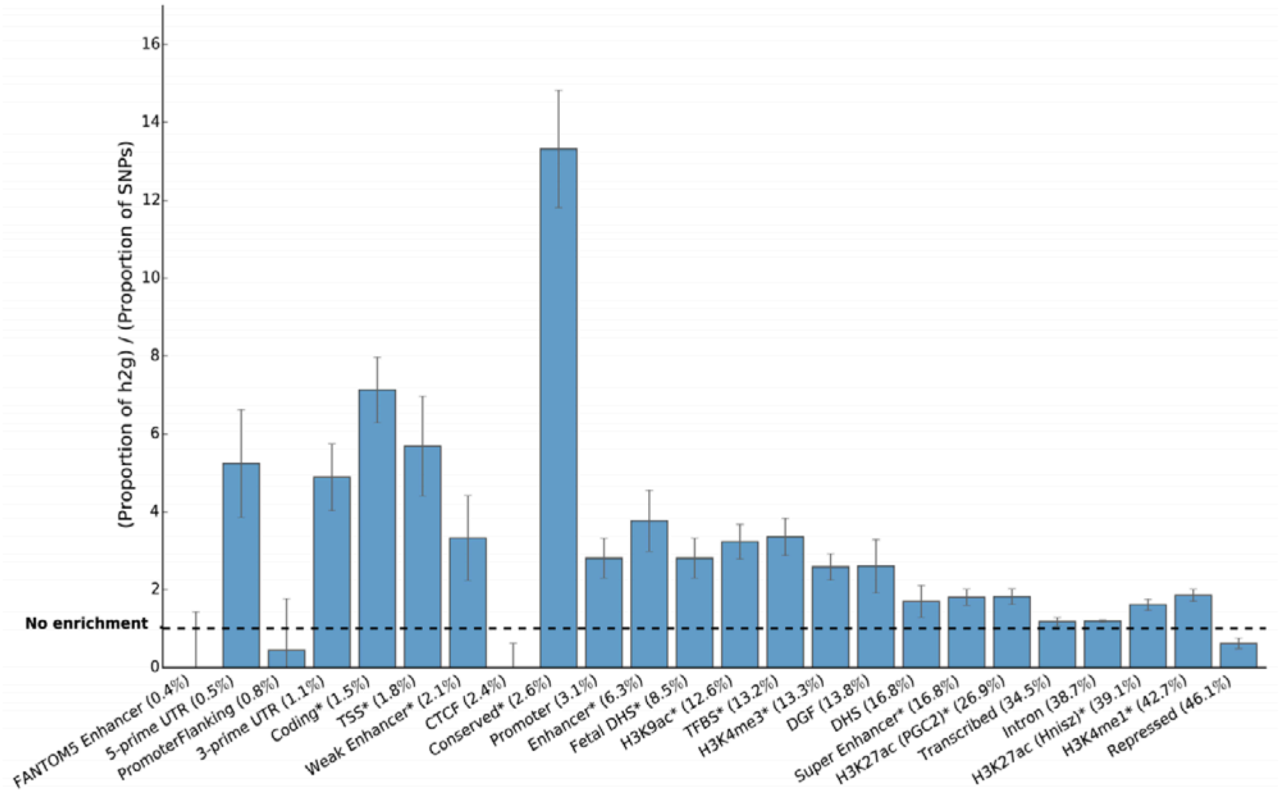
Hypersensitive Sites
CH_3
RNA polymerase
CH_3CO
CH_3
5C
DNase-seq FAIRE-seq
ChIP-seq
Computational predictions and RT-PCR
RNA-seq
Gene
Long-range regulatory elements (enhancers, repressors/ silencers, insulators)
cis-regulatory elements (promoters, transcription factor binding sites)
Transcript

- ENCODE, Roadmap Epigenomics
- Regulatory elements: promoters, enhancers
- Transcription factor binding sites
- CpG sites
- ChromHMM

ENCODE Project Consortium. Plos Biology 2011.

Epigenetic data informs heritability

LD score regression, related approaches partition *h^2*

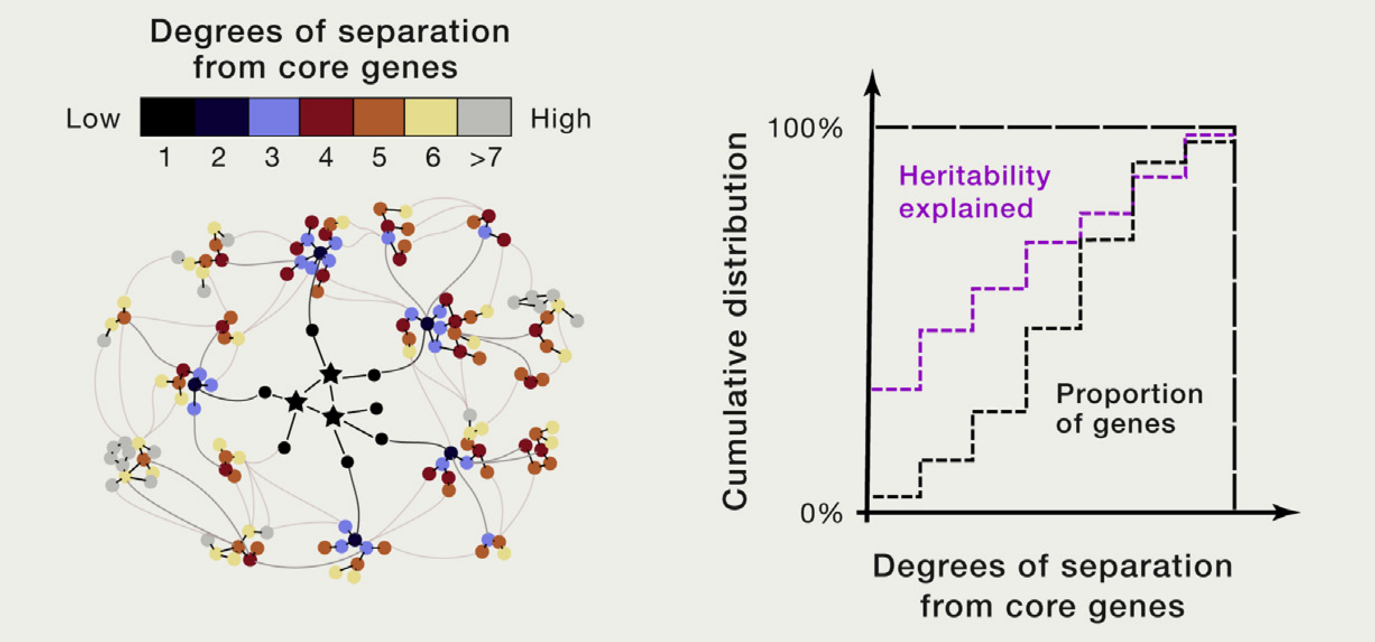


Large scale epigenetic data (Roadmap, ENCODE) enable analysis, indicate contribution of gene regulation

Figure from Finucane, NG, 2015

Ommigenic model

- Most/all expressed genes in disease-relevant cell types affect trait

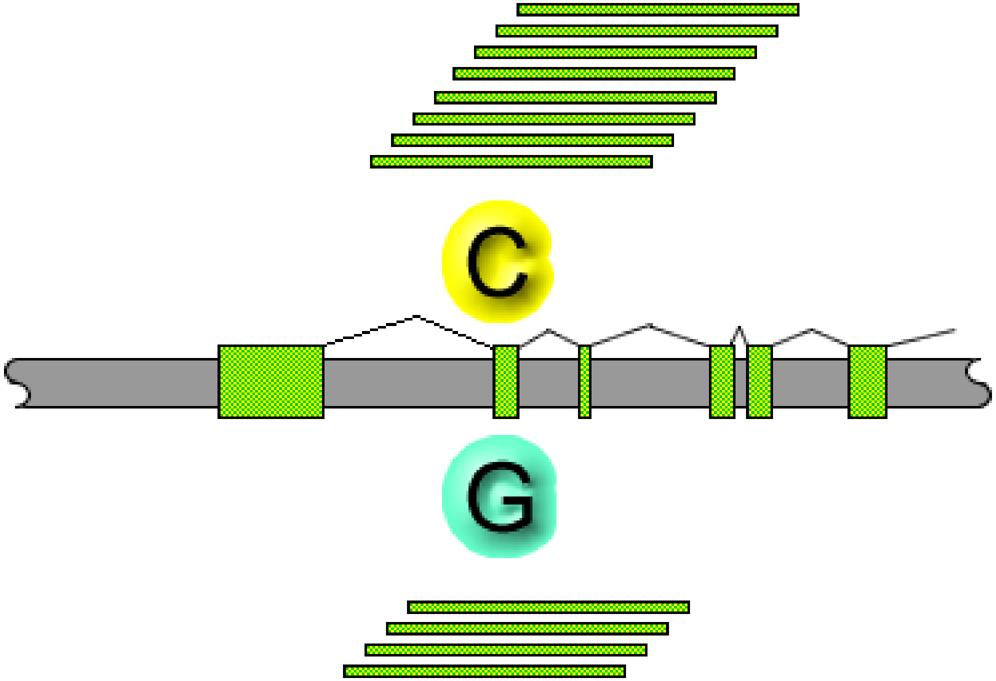


- Highlights potential role of eQTLs, trans effects

Boyle et al., Cell, 2017

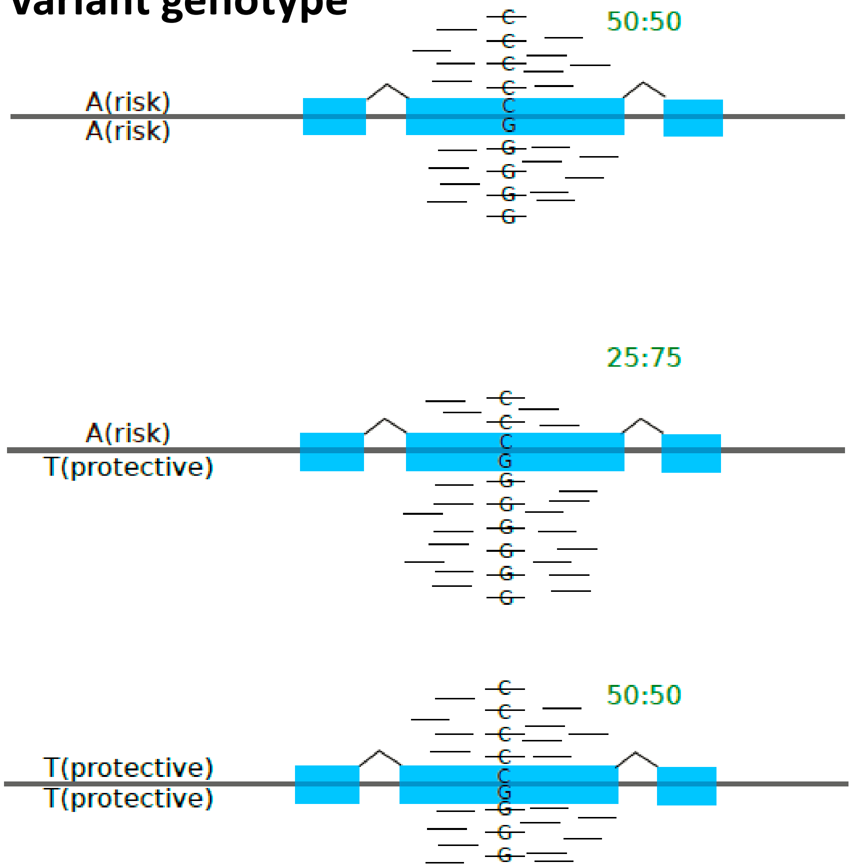
Advantages of ASE

- Test within an individual allelic imbalance, given one has sufficient reads.



Using ASE to detect GWAS signals driven by <u>multiple</u> causal variants

GWAS variant genotype



No ASE ⇨ LACK OF ASE FOR HOMS

ASE ⇨ ABUNDANT ASE FOR HETS

No ASE ⇨ LACK OF ASE FOR HOMS

Tests functional differences between alleles in population

Lucia Conde et al, *AJHG*, 2013

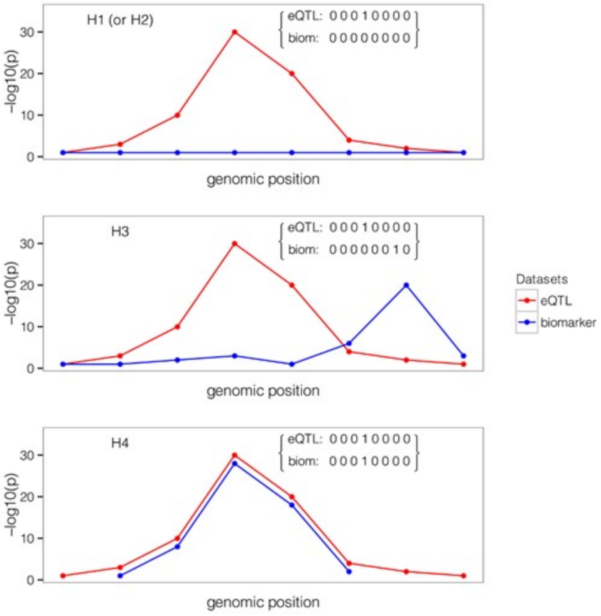
Coloc: A Bayesian test for colocalization of pairs of association signals

H1 is the hypothesis that there is only an eQTL signal at a locus

H2 is the hypothesis that there is only a GWAS signal at a locus.

H3 is the hypothesis that there are two independent eQTL and GWAS signals in linkage.

H4 is the strong hypothesis that the same SNP (not just the locus) is responsible for both the GWAS and eQTL.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

GWAS eQTL colocalization

- **Coloc**
- **eQTpLot**
- hypercoloc

Examples of H3 and H4

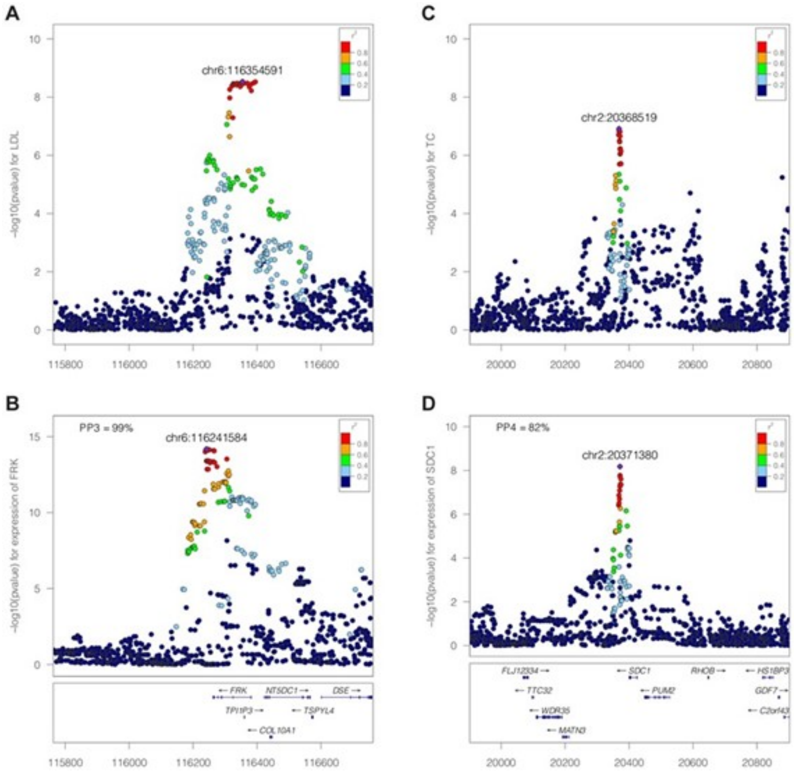
On the left, the profile of association at the FRK locus with LDL (top) is very different from that with *FRK* expression.

H3 is the supported hypothesis.

On the right, even though there are two different peak SNPs, they are in the same strong LD region and the profiles are almost the same for Total Cholesterol and *Soc1* expression.

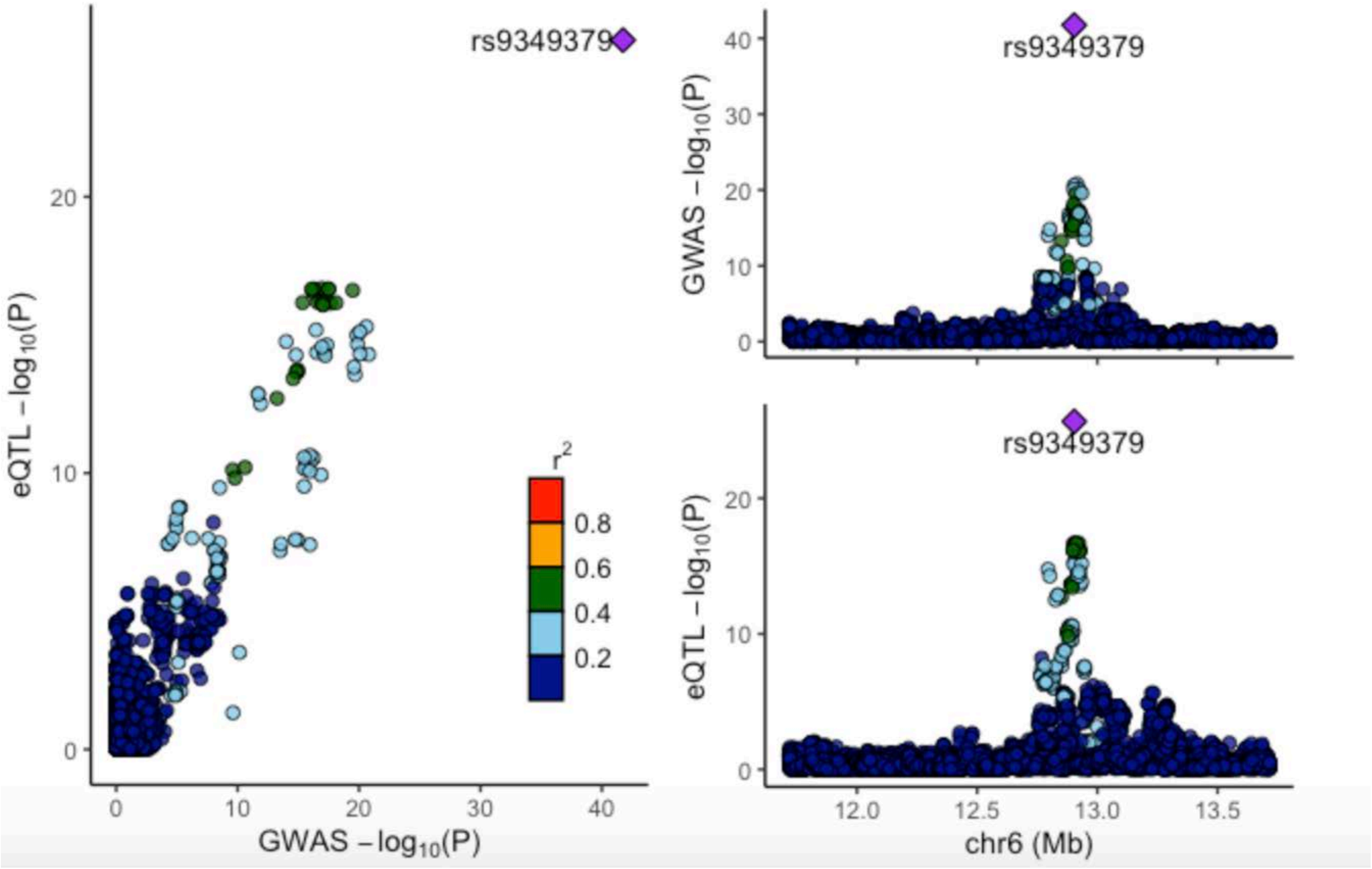
H4 is the supported hypothesis.

Bayesian analysis evaluate each H relative to the other four and generates a confidence level for the most likely one.



Giambartolomei et al (2014) *PLoS Genetics* **10(5)**: e1004383

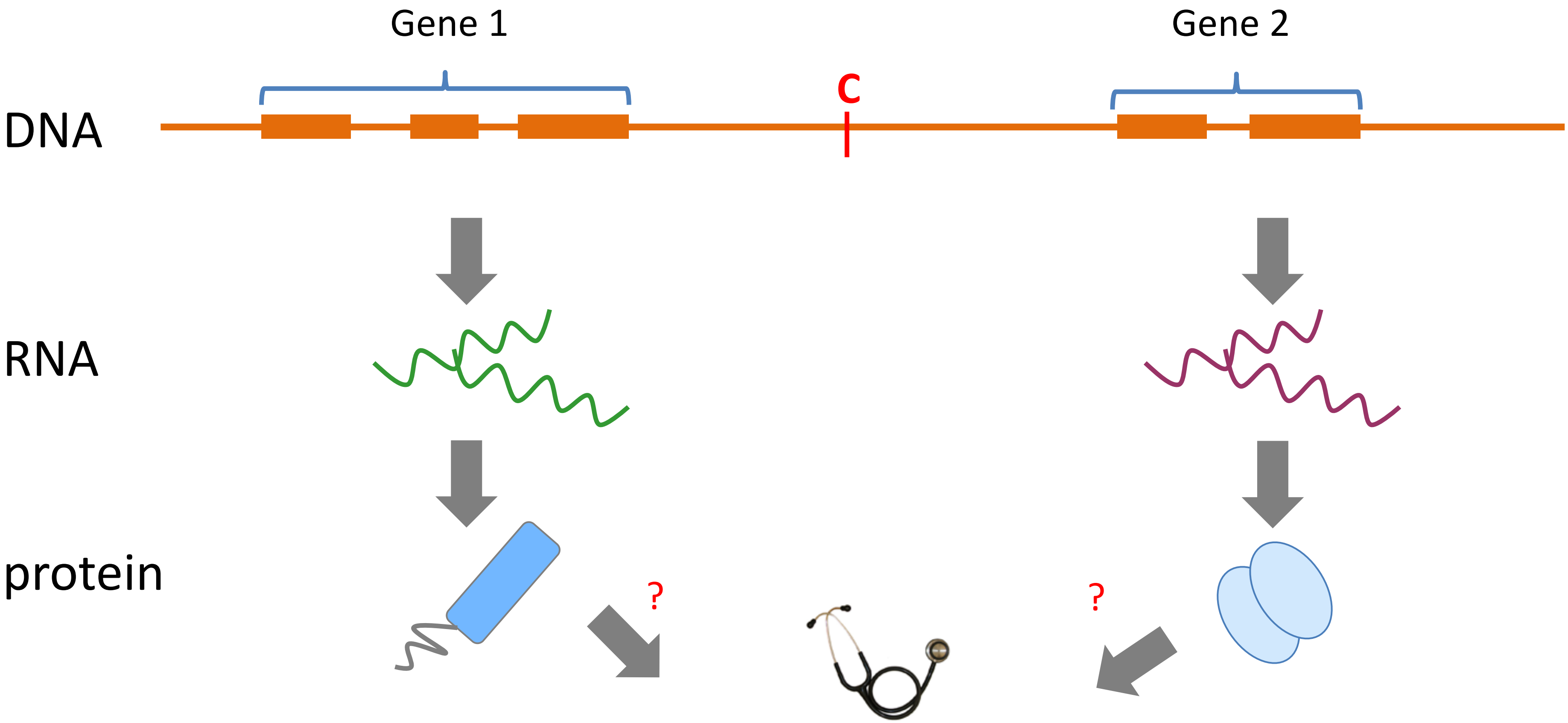
Coloc results



eQTLs and complex disease genetics

Help interpret GWAS variants (especially non-coding):

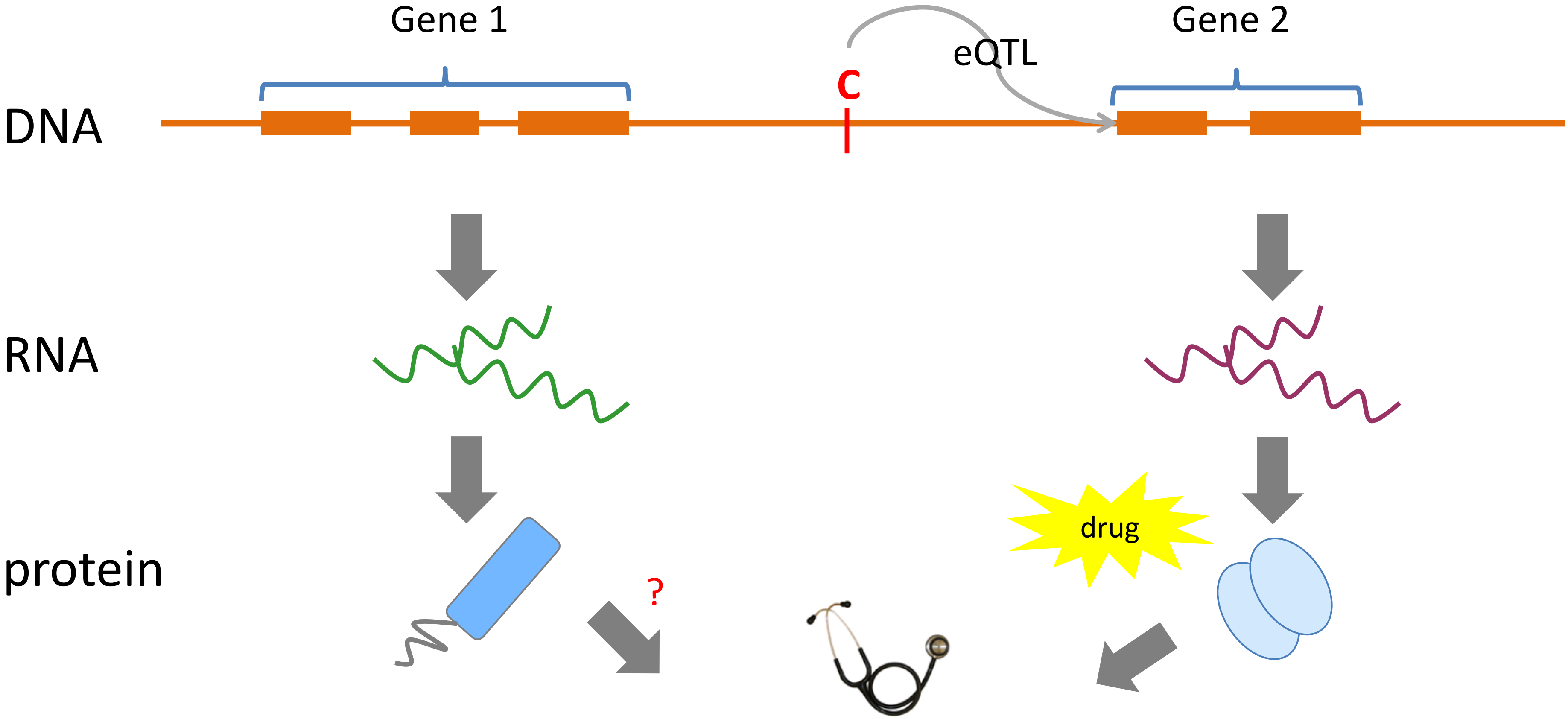
- understand mechanism
- guide interventions



eQTLs and complex disease genetics

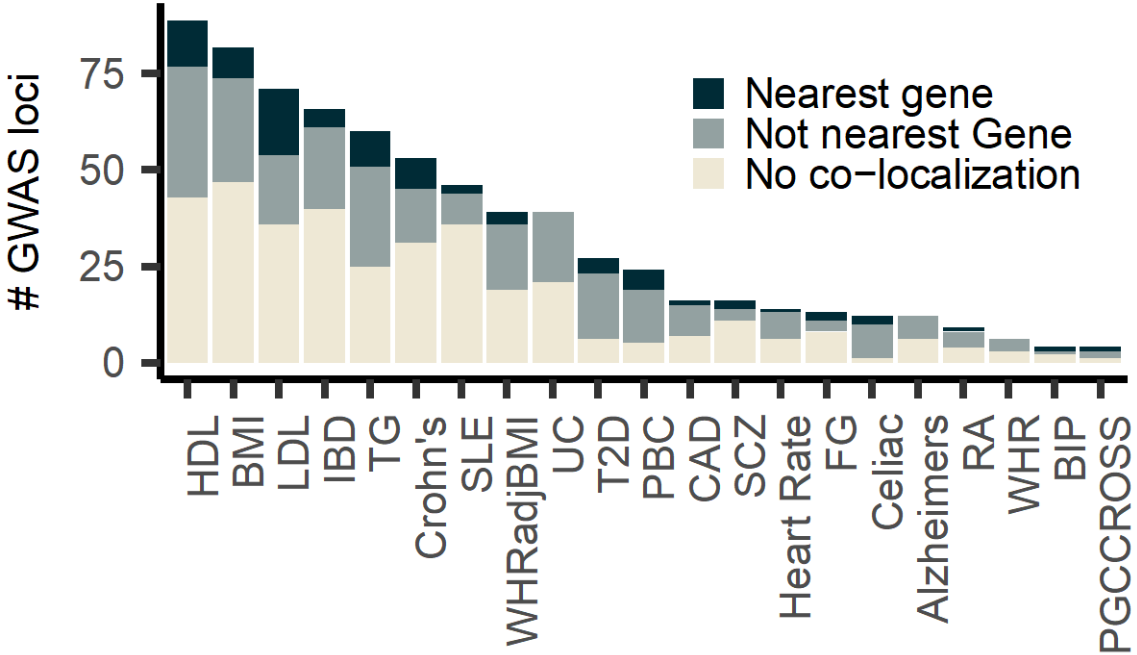
Help interpret GWAS variants (especially non-coding):

- understand mechanism
- guide interventions



eQTLs and complex disease genetics

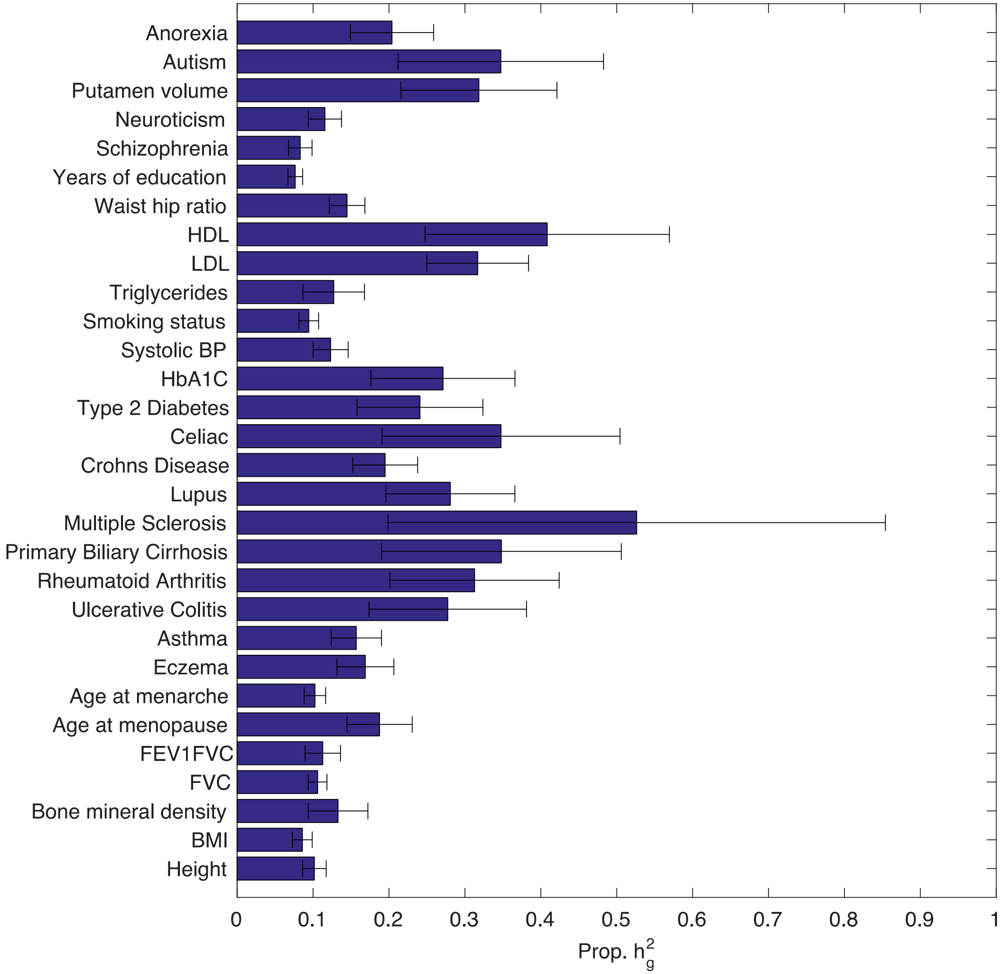
52% of genetic variants associated with human disease co-localize with an eQTL



The GTEx Consortium, Nature 2017

eQTL data informs heritability

GE co-score regression indicates cis-eQTLs explain mean 21% of h^2 across a set of complex traits



O'Connor et al. bioRxiv, 2017

3. Complex effects of genetic variation on gene expression

What are we missing?

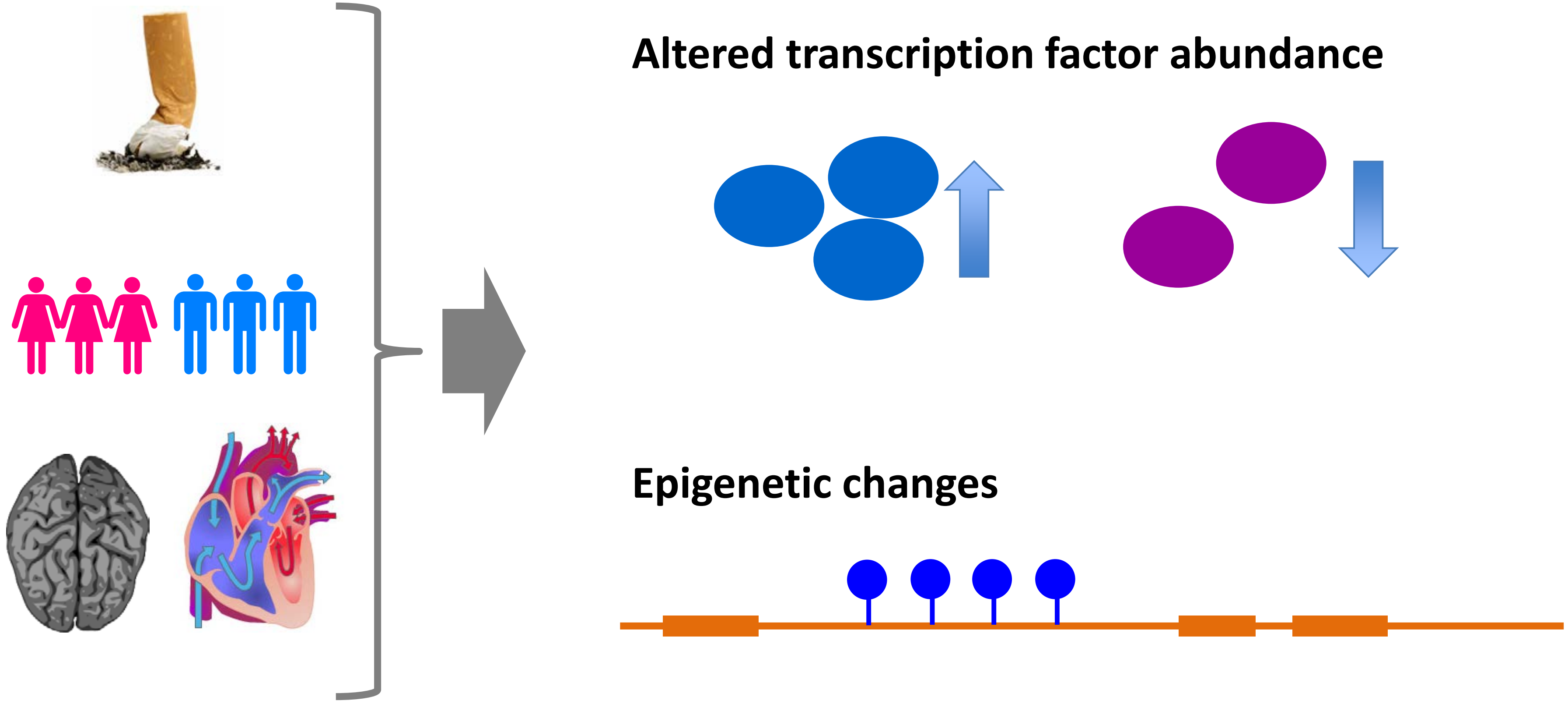
- Most studies are done on steady-state total expression measurements at a single adult or post-mortem time point

- Disease-relevant states include different developmental stages, environmental exposures, cell types

- Other variant classes and regulatory effects

Context-specificity

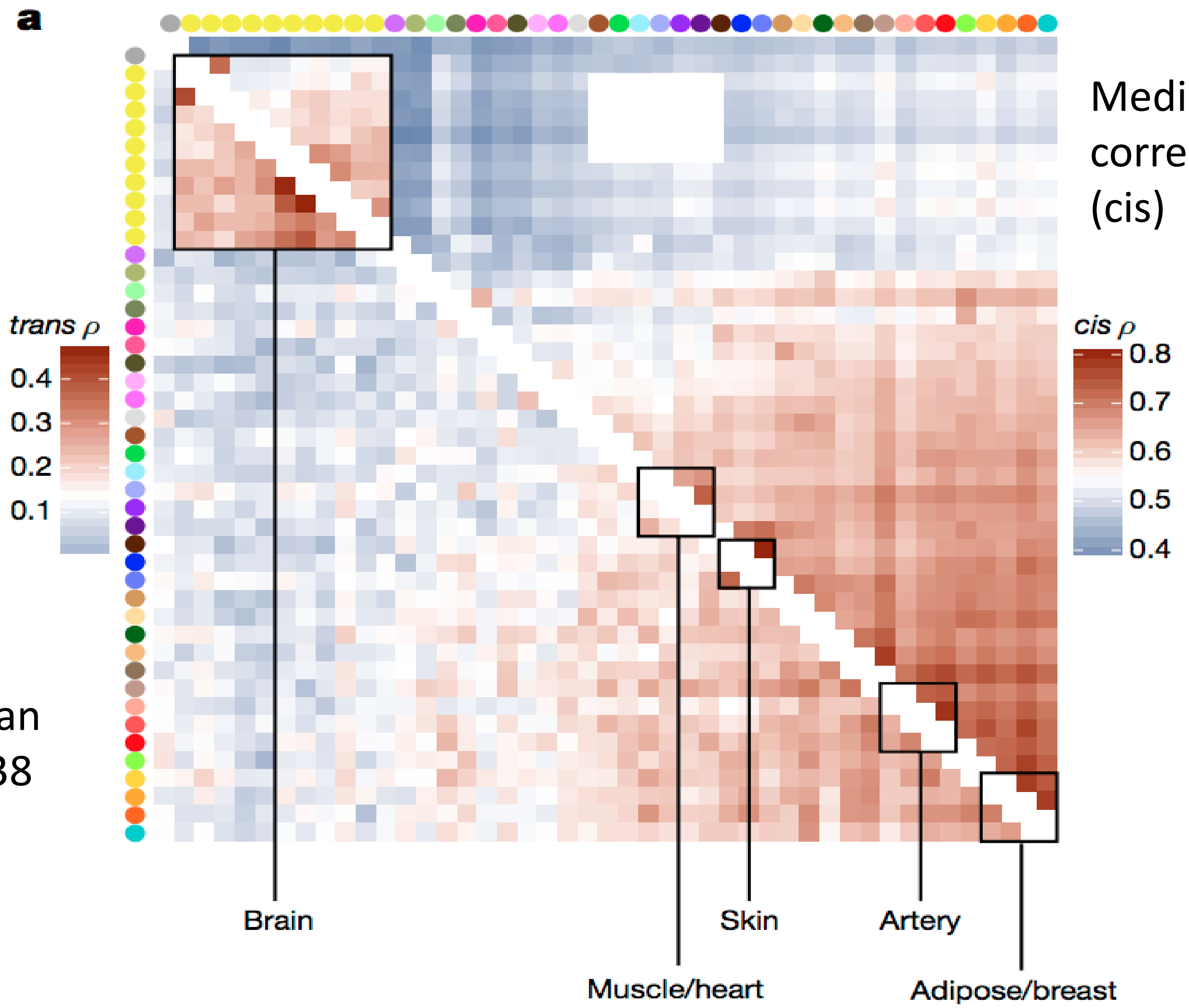
Many factors can *modulate* regulatory effects



Altered transcription factor abundance

Epigenetic changes

GTEx tissue-specificity of cis and trans



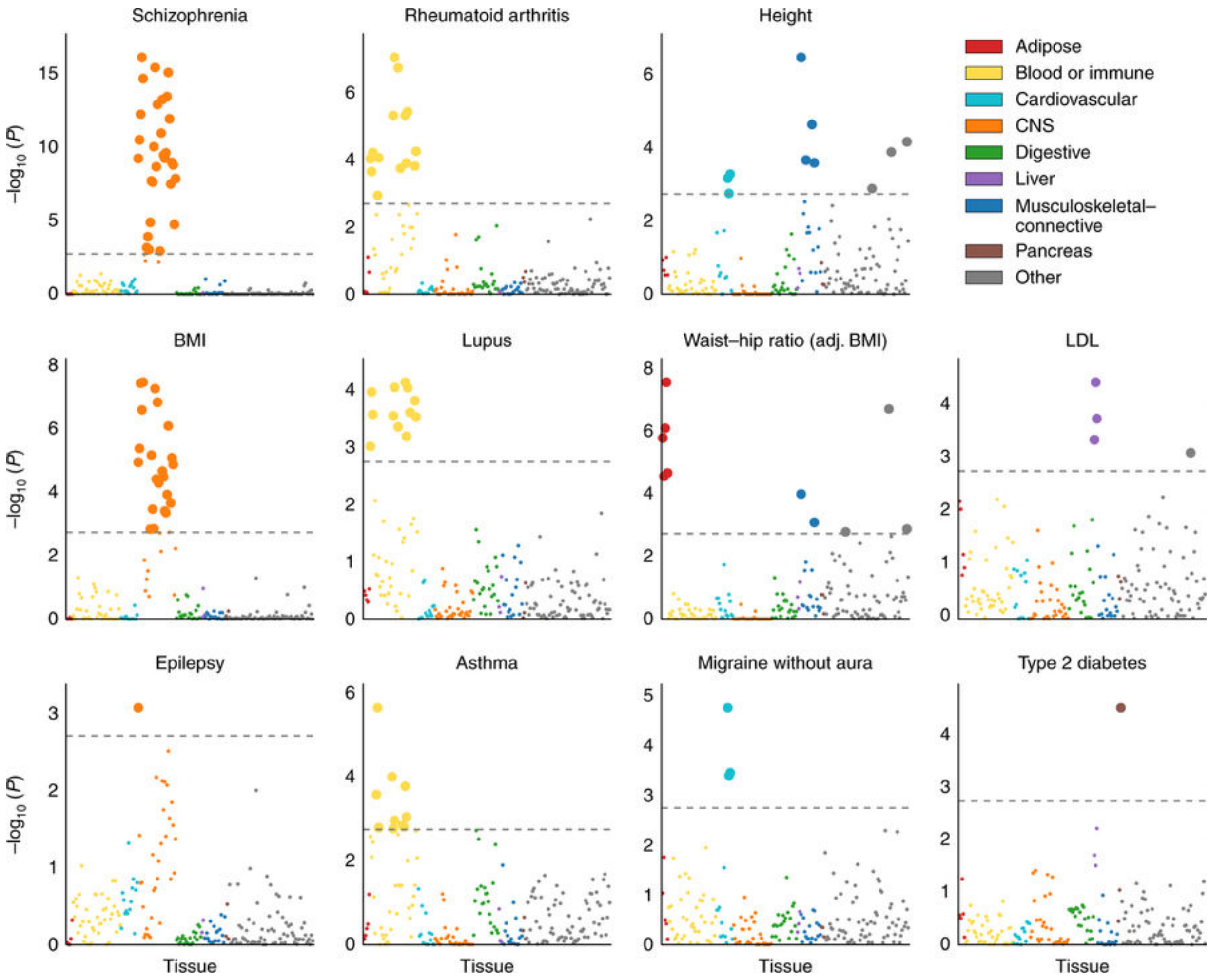
Median spearman correlation=0.547 (cis)

Median spearman correlation=0.138 (trans)

Trans eQTLs appear more highly tissue-specific than cis-eQTLs

The GTEx Consortium, Nature 2017

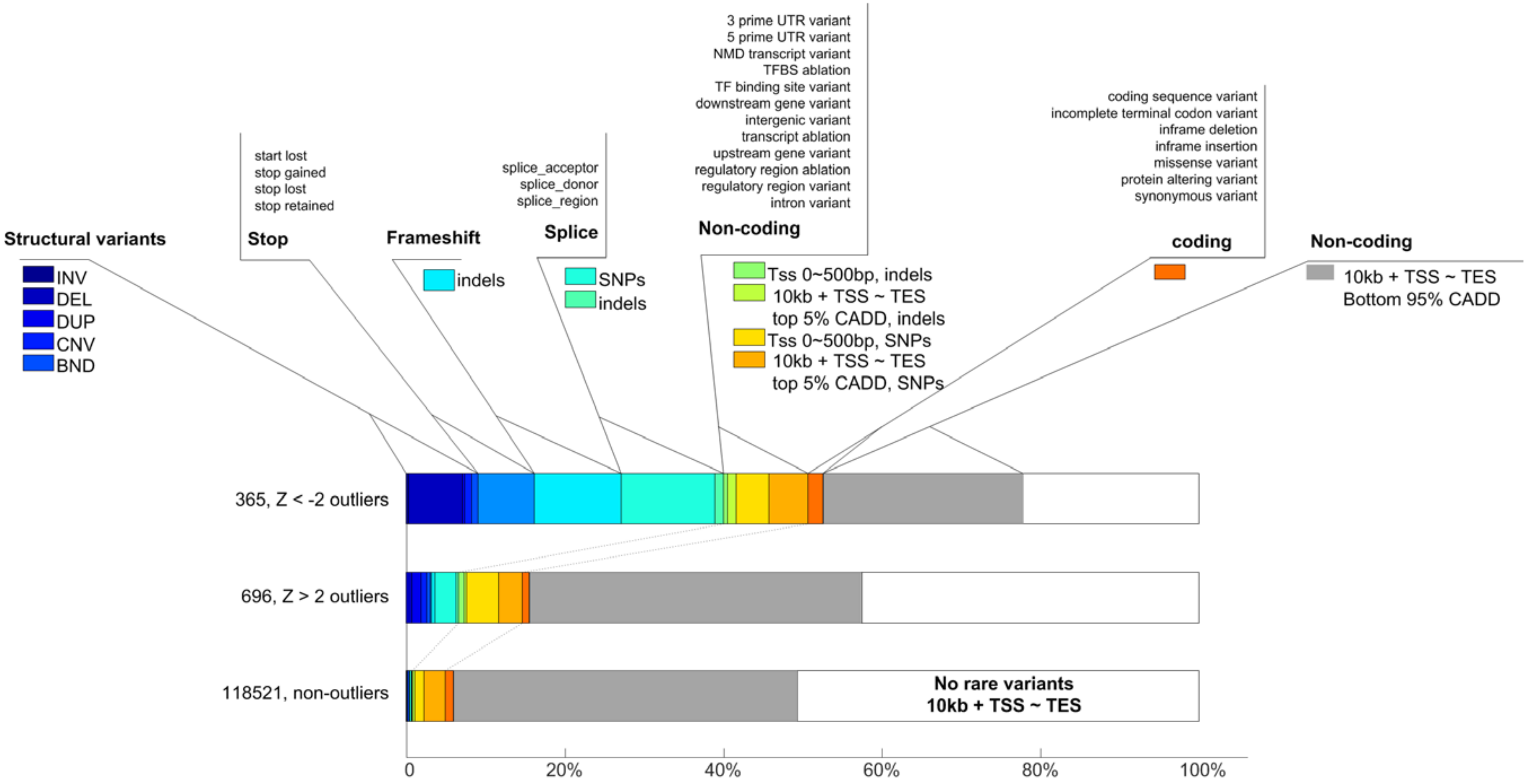
Tissue specificity and heritability



From Finucane et al, NG, 2018

Rare variants

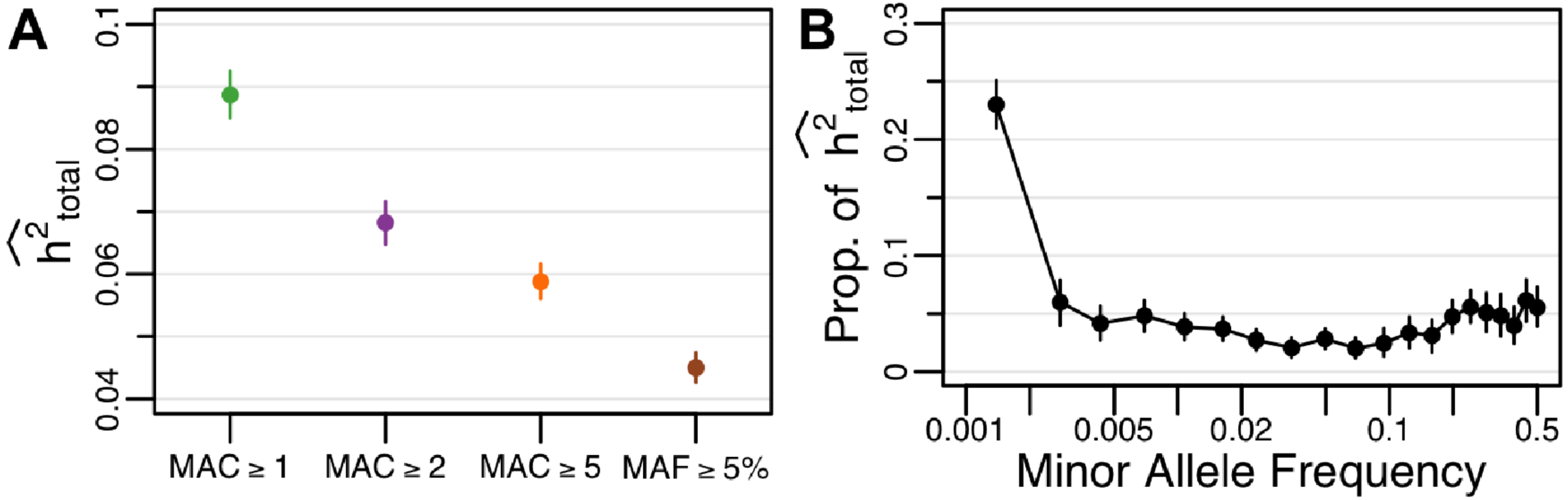
Recent work emphasizes importance of rare variation in driving extreme expression levels



Li et al, Nature, 2017

Rare variants

Preprint (Hernandez et al 2017) suggests rare variants explain a large fraction of heritability of gene expression



4. Conclusions

Why delve deeper into expression?

- Help determine when and how much to invest in WGS, expression, epigenetic data

- To continue understanding implicated
  - Genes
  - Tissue and cell types
  - Epigenetic and other regulatory mechanisms

- Challenges and caveats
  - Ambiguity: many variants affect multiple genes
  - Interpretability: missing relevant cell types
  - Power: trans-eQTLs also require large sample sizes

5. eQTL tool demo

matrixEQTL

Statistical Clinics

1. It is a special service for free for researchers
   in Academia Sinica.

2. The service is offered at 14:00-16:00, Thursday,
   in Institute of Statistical Science Building,
   Room 401.

ACADEMIA SINICA DATA SCIENCE STATISTICAL COOPERATION CENTER

資料科學
統計合作社

……什麼都問

統計診所
Statistical Clinics

什麼都不奇怪

統計診所是統計合作社特別為院內同仁開放的一段免費統計問題討論的時間。統計診所提供的服務不論是實驗設計、資料分析、資料圖表呈現，或是統計觀念的釐清、文章投稿的統計問題等等，只要是統計相關，不論大小問題都歡迎到統計診所提出討論。

您正在苦惱資料該如何收集嗎？
您有樣本數該收幾個的問題嗎？
資料該用什麼統計方法分析呢？
統計結果該如何解讀呢？
來統計診所找專家準沒錯！

研究路途不卡關，
統計診所陪您一起渡過。

時　間
每週四　14:00 - 16:00

地　點
統計所大樓4樓401室

如何取得服務
1. 直接現場提問
2. 網路預約

FREE
線上預約

統計暨資料科學諮詢服務 中央研究院 統計科學研究所

3. You are welcome to apply for the service.
http://disc.stat.sinica.edu.tw/statistical-clinic-service-appointment/

Thank you