

# UQ Genetics and Genomics Winter School 2025


Systems Genomics and  
Pharmacogenomics  
Module 6 Day 1

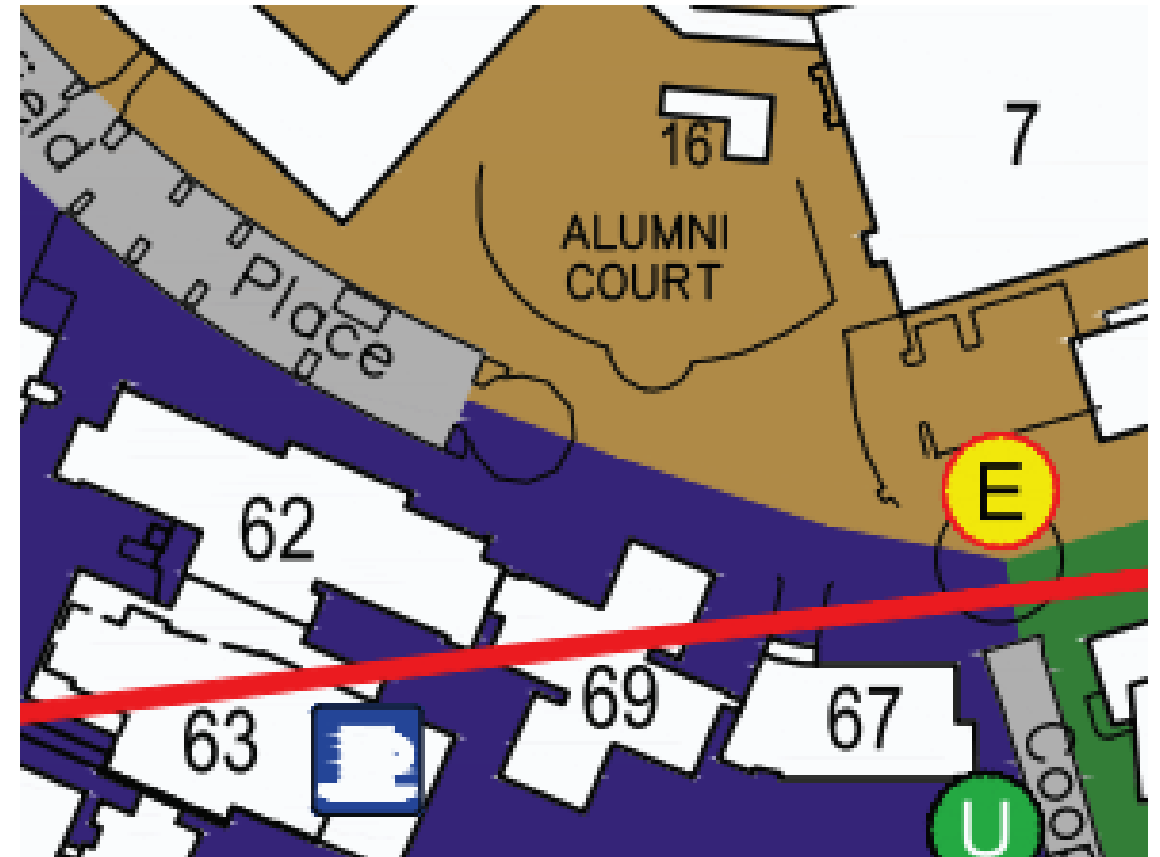
# Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.



# General Information

- We are currently located in Building 69
- Emergency evacuation point 
- Food court and bathrooms are located in Building 63
- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module



# Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

If you haven't done so, please email <ctr-pdg-admin@imb.uq.edu.au> with your name and the below statement to confirm that you agree with the following:

“I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

# Learning materials

Instructions to access WiFi/desktop/server:

<https://cnsgenomics.com/data/teaching/GNGWS25/module0/>

The winter school server is available until **25<sup>th</sup> July 2025** (2 weeks after the course)

Slides and practical notes for this module:

<https://cnsgenomics.com/data/teaching/GNGWS25/module6/>



# Tutors



Lauren  
Barker



Solal  
Chauquet



Sonia  
Shah



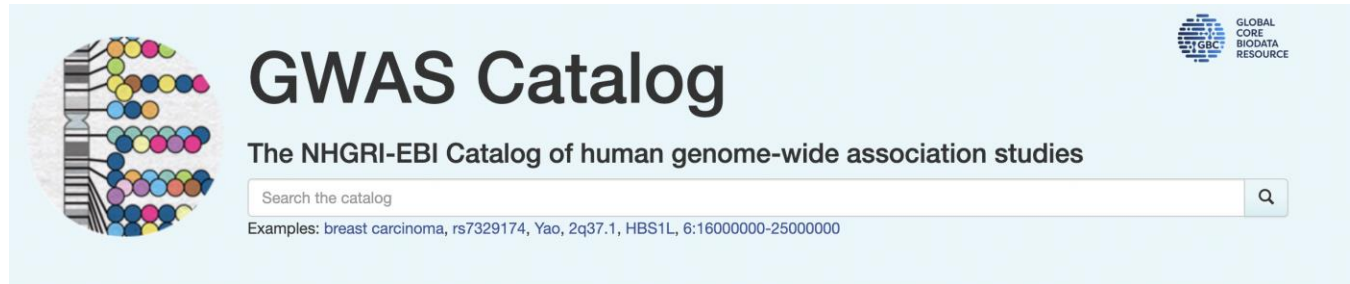
Zhihong  
Zhu

# Why do we do GWAS?

- GWAS identifies variants that are associated with disease
- Identifying genes and biological pathways involved in disease
- Identifying which cell types and tissues most relevant to disease
- Identifying new avenues for drug development
- Risk stratification (polygenic risk scores)
- Understanding causal risk factors (genetic epidemiology)

**GOAL: Progress from genetic maps to mechanism to medicine**

# We're great at generating genetic maps



a freely accessible curated collection of all human genome-wide association studies

**As of 2025-06-07**  
7286 publications  
891,200 top SNP associations  
136,189 full summary statistics available



BUT...moving from genetic maps to mechanism and medicine is challenging

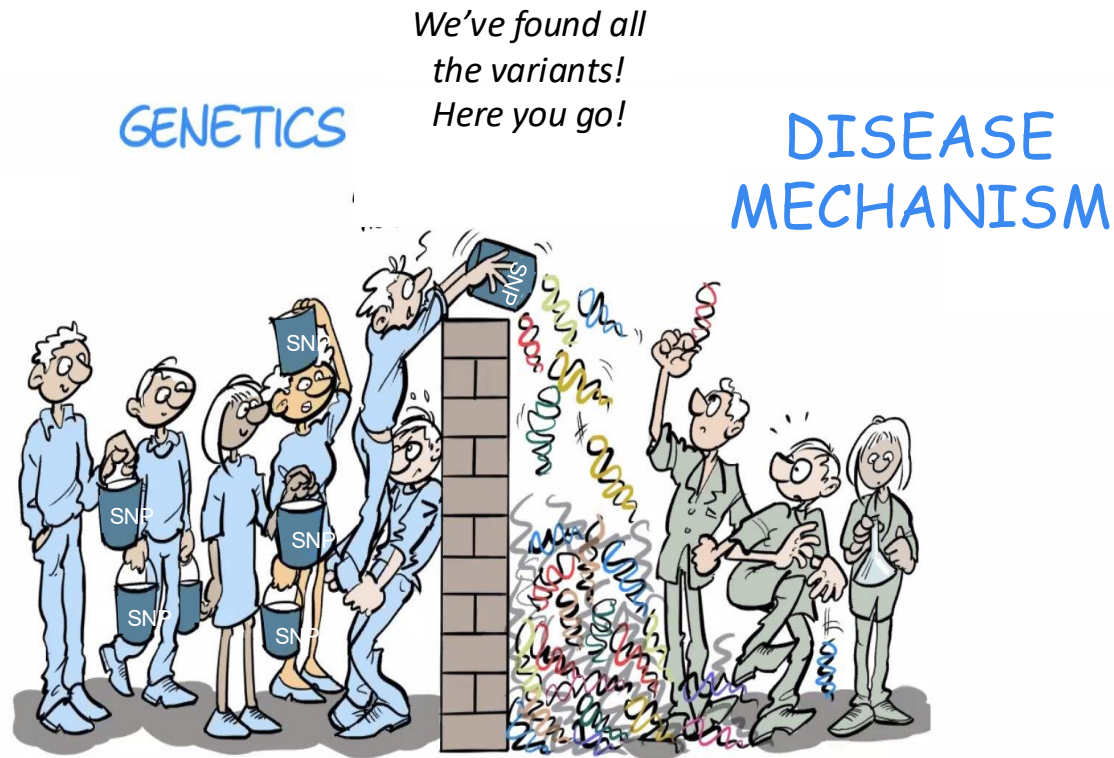
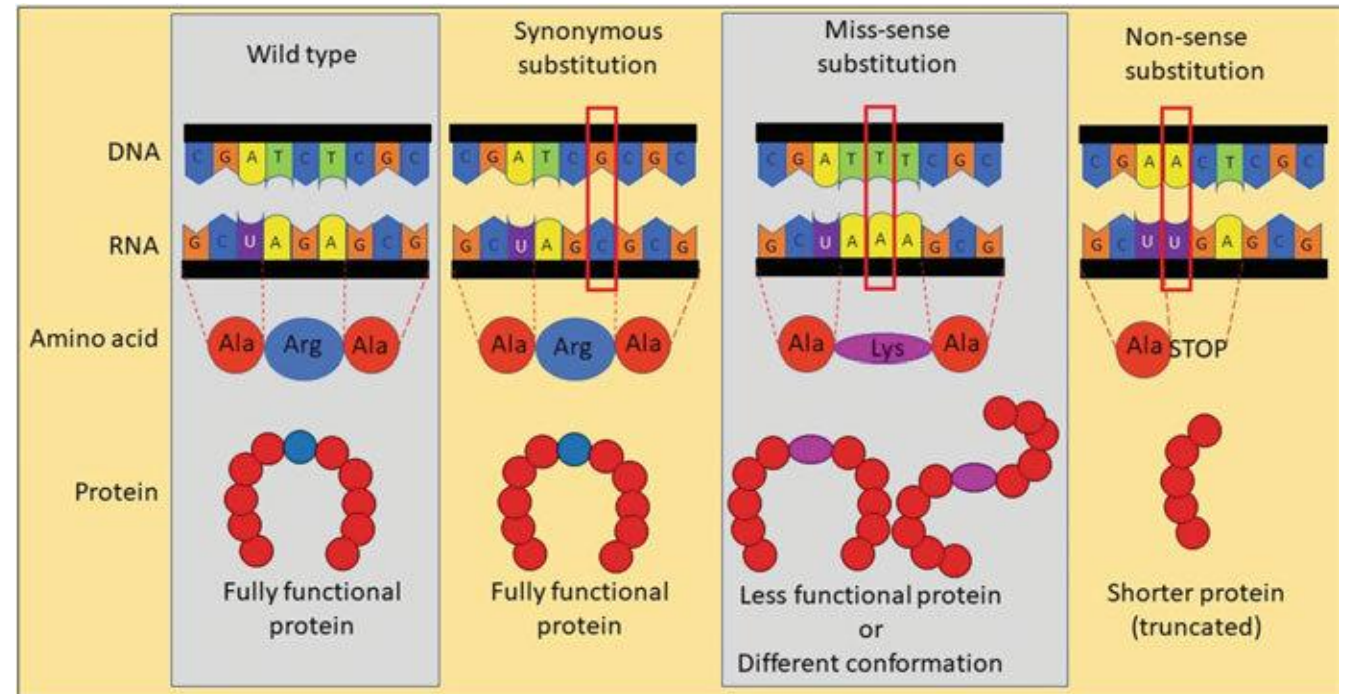
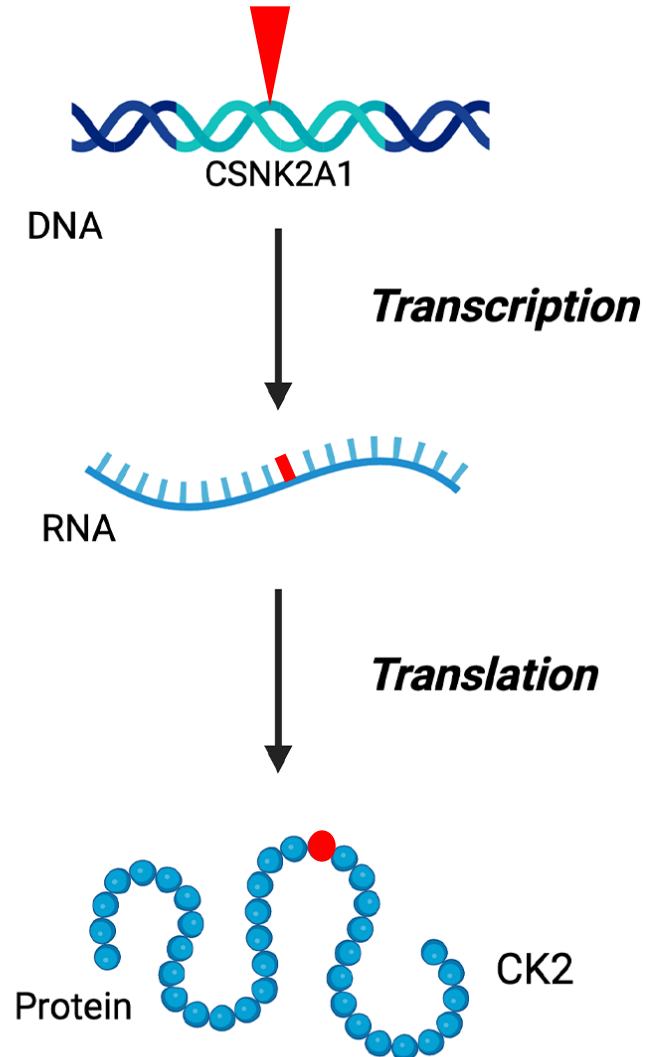


IMAGE CREDIT: [BRAINSCAPES](#)

- What are the causal variants?
- What are the causal genes?
- What are the relevant cell types and tissues?
- How does the variant and gene affect the disease?

# SNP to mechanism – protein-coding variants (low hanging fruit)



## SNP to mechanism – protein-coding variants

- *PCSK9* c.426C>G (p.Tyr142X) associated with lower risk of coronary artery disease
- Premature termination codon and truncation of the encoded protein or absence of the protein due to nonsense mediated decay i.e. Loss-of-function (LOF) variant
- LOF carriers have substantially lower plasma LDL-C
- Predicting functional consequence of coding variants – SIFT, PolyPhen

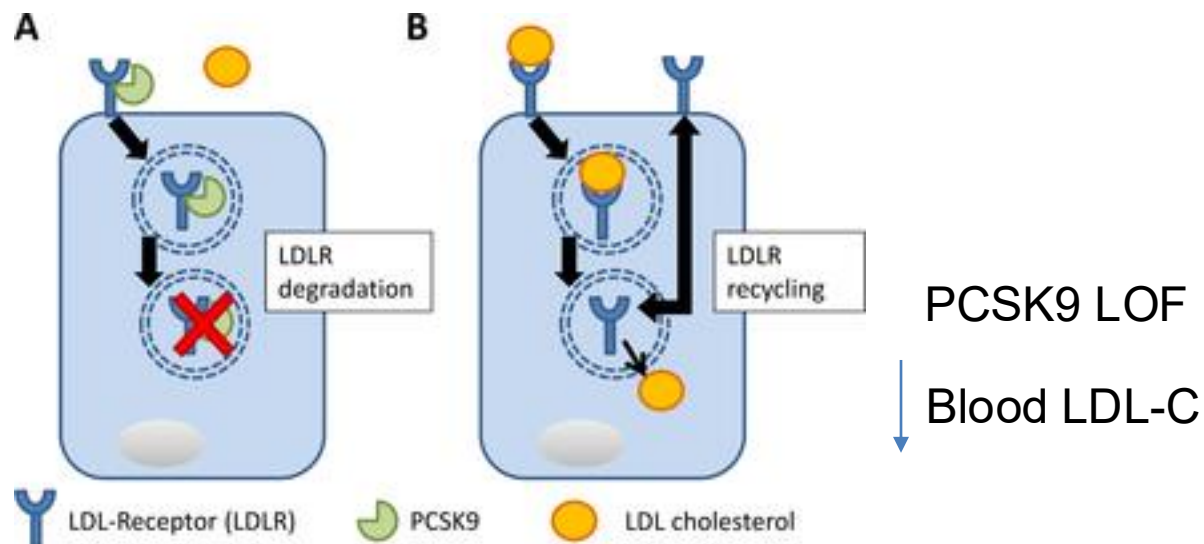


Image source: <https://www.biovendor.com/>

# SNP to mechanism – protein-coding variants



The screenshot shows the top of a Science journal article page. At the top left is the 'Science' logo in red. To its right are navigation links: 'Current Issue', 'First release papers', 'Archive', and 'About' with a dropdown arrow. Further right is a 'Submit manuscript' button. Below these is a breadcrumb trail: 'HOME > SCIENCE > VOL. 381, NO. 6664 > ACCURATE PROTEOME-WIDE MISSENSE VARIANT EFFECT PREDICTION WITH ALPHAMISSENSE'. Below the breadcrumb is a row of tags: a lock icon, 'RESEARCH ARTICLE', and 'MACHINE LEARNING'. To the right of these are social media icons for Facebook, X, Twitter, LinkedIn, YouTube, and others. The main title 'Accurate proteome-wide missense variant effect prediction with AlphaMissense' is prominently displayed. Below the title, the authors are listed: 'JUN CHENG', 'GUIDO NOVATI', 'JOSHUA PAN', 'CLARE BYCROFT', '[...]', and 'ŽIGA AVSEC', each with an ORCID icon. To the right of the authors is a '+11 authors' button and a link to 'Authors Info & Affiliations'. At the bottom left of the article preview, it says 'SCIENCE • 19 Sep 2023 • Vol 381, Issue 6664 • DOI: 10.1126/science.adg7492'.

Science

Current Issue First release papers Archive About ▼ Submit manuscript

HOME > SCIENCE > VOL. 381, NO. 6664 > ACCURATE PROTEOME-WIDE MISSENSE VARIANT EFFECT PREDICTION WITH ALPHAMISSENSE

RESEARCH ARTICLE | MACHINE LEARNING

Accurate proteome-wide missense variant effect prediction with AlphaMissense

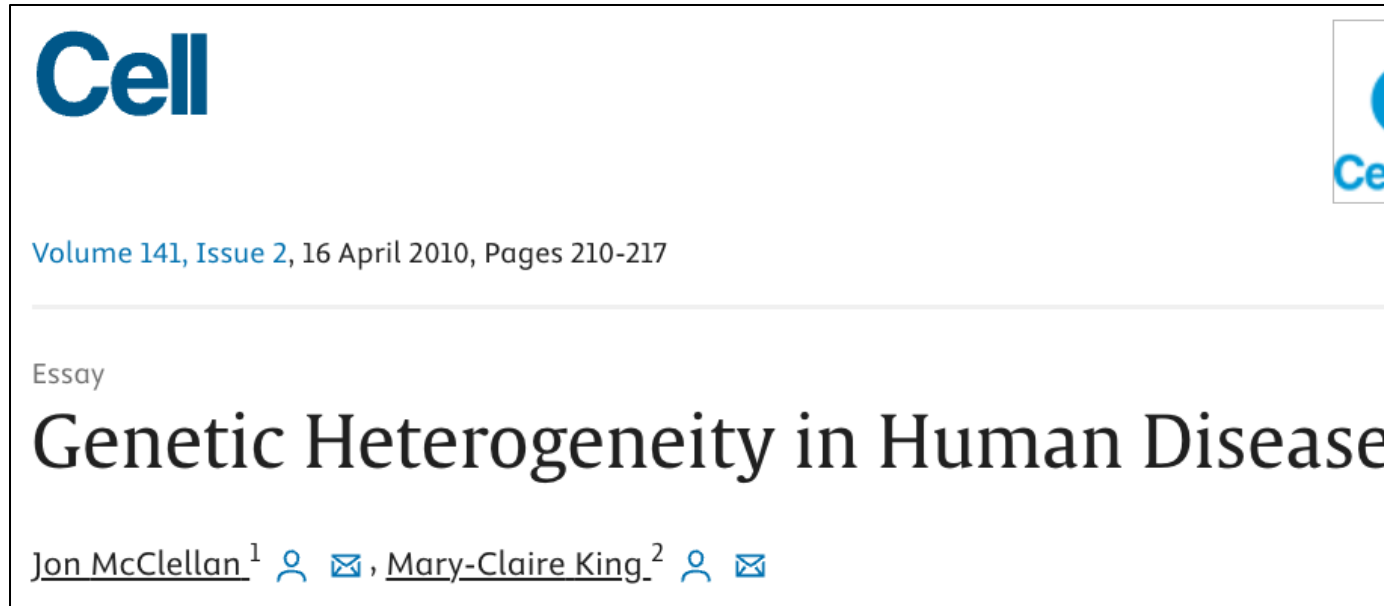
JUN CHENG , GUIDO NOVATI, JOSHUA PAN, CLARE BYCROFT , [...], AND ŽIGA AVSEC  +11 authors [Authors Info & Affiliations](#)

SCIENCE • 19 Sep 2023 • Vol 381, Issue 6664 • DOI: 10.1126/science.adg7492

- (i) unsupervised protein language modelling to learn amino acid distributions conditioned on sequence context
- (ii) incorporates structural context
- (iii) incorporates population frequency data, thereby avoiding bias from human-curated annotations.

Provide a database of predictions for all possible single amino acid substitutions in the human proteome. We classify 32% of all missense variants as likely pathogenic and 57% as likely benign

<2% of the human genome is protein-coding



*“To date, GWAS have published hundreds of common variants...  
However, the vast majority of such variants  
have no biological relevance to disease or clinical utility...”*



# 98% of the genome is not junk!



Image source [www.biocomicals.com](http://www.biocomicals.com)

## The Encyclopedia of DNA Elements (ENCODE)

Goal: Build a comprehensive list of functional elements in the human genome, including elements that act at the protein and RNA levels, and regulatory elements that control cells and circumstances in which a gene is active.

ENCODE
Data
Encyclopedia
Materials & Methods
Help

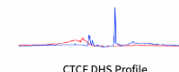
Search...

### Ground Level Annotations

**Open chromatin (DNase-seq, ATAC-seq)**

DNase I hypersensitive sites (DHSs) computed from DNase-seq experiments, and ATAC-seq peaks (enriched genomic regions).

[\[Open chromatin regions\]](#)




CTCF DHS Profile

**Histone mark enrichment (ChIP-seq)**

Peaks (enriched genomic regions) of a variety of histone marks computed from ChIP-seq experiments.

[\[Histone mark peaks\]](#)




H3K27ac from mouse e11.5 hindbrain

**Transcription factor binding (TF ChIP-seq)**

Peaks (enriched genomic regions) of TFs computed from ChIP-seq experiments. Visualize sequence motifs and other information on Factorbook.

[\[TF peaks\]](#) [Factorbook](#)

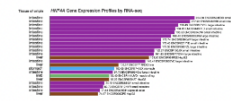


CTCF Motif from [Factorbook](#)

**Gene expression (RNA-seq)**

Expression levels of genes and transcripts annotated by GENCODE, which can be visualized on SCREEN.

[\[Expression levels\]](#) [SCREEN](#)

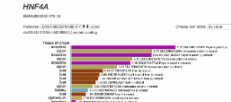


HNF4A Gene Expression

**Transcription start site (TSS) activity profiling (RAMPAGE)**

Identification of transcription start sites (TSSs) and quantification of transcript expression, which can be visualized on SCREEN.

[\[RAMPAGE peaks\]](#) [SCREEN](#)

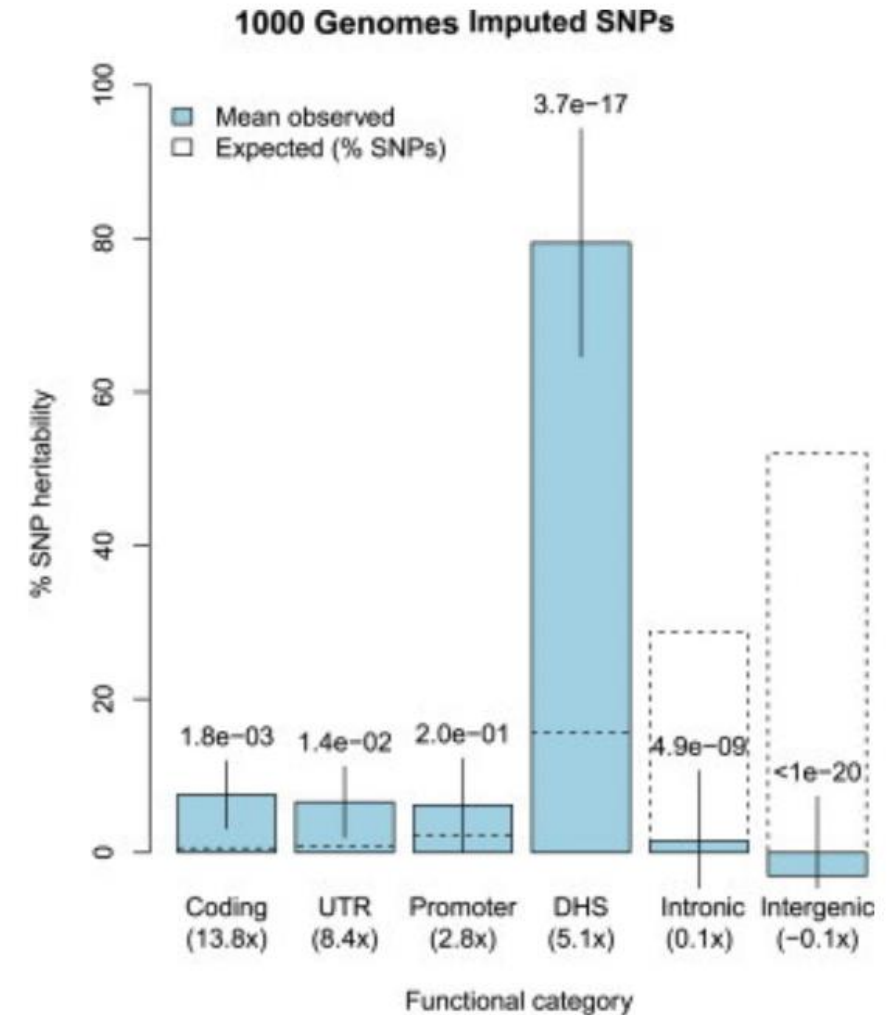


HNF4A Transcript Expression

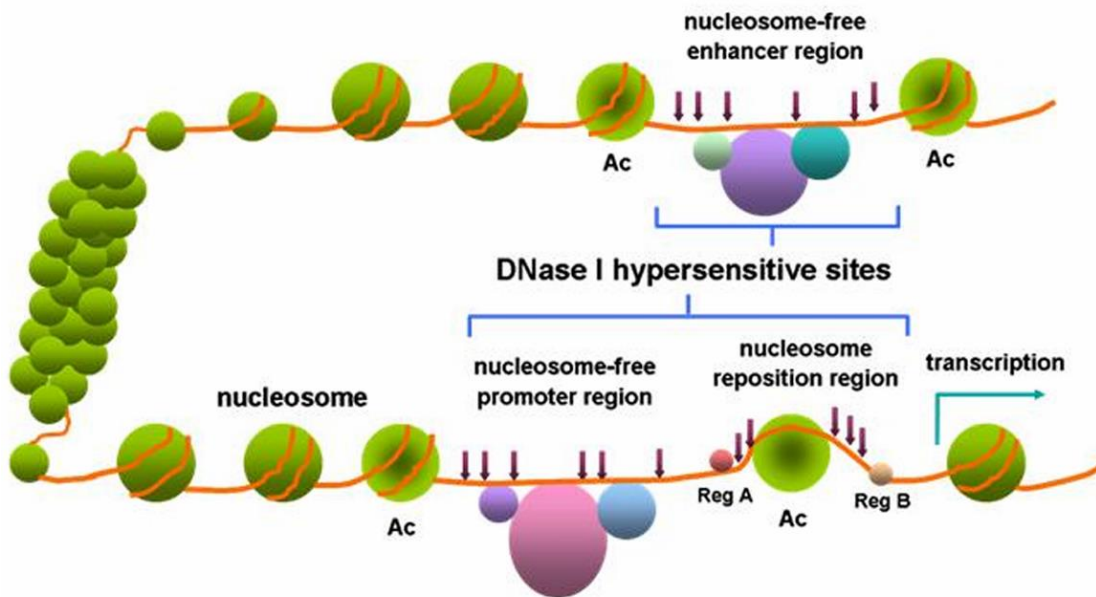


# >90% GWAS significant SNPs are in non-coding regions

- SNP heritability  $h^2$  partitioned by functional category across 11 common diseases (psychiatric, metabolic, immune and cardiovascular)
- **Coding variants:** explain 8% of SNP  $h^2$
- **DHSs** from 217 cell types explain ~80% of SNP  $h^2$

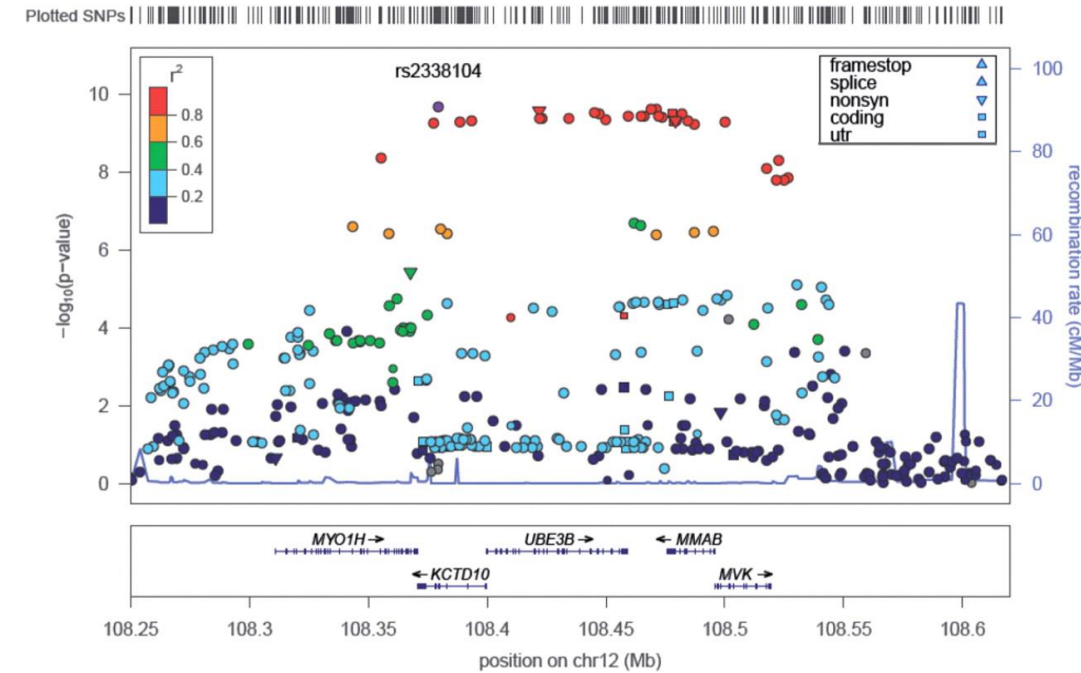
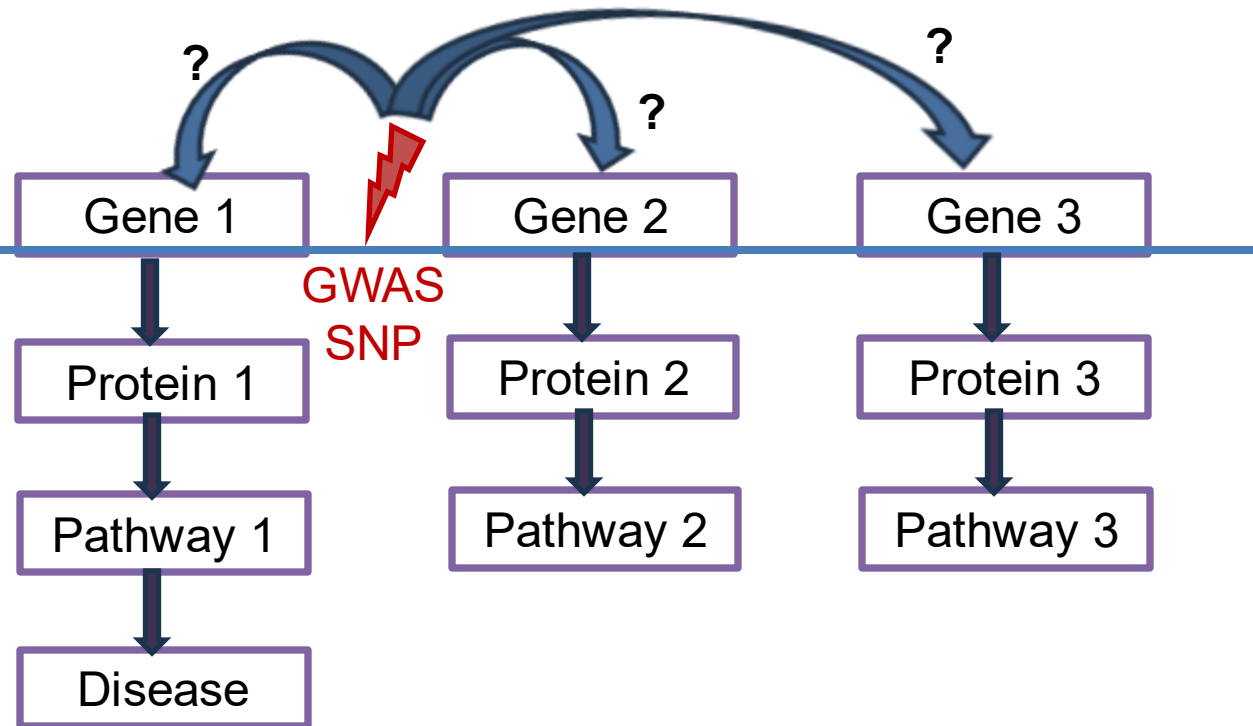


# DNase I hypersensitive sites



- Regions that are sensitive to cleavage by the DNase I enzyme
- Associated with open chromatin, and thus accessibility to regulatory elements and transcriptional activity.
- Maps to regulatory elements (promoters, enhancers, insulators, silencers)
- Accessibility of regulatory regions is highly cell type- and state-selective

# Do these SNPs affect gene expression?

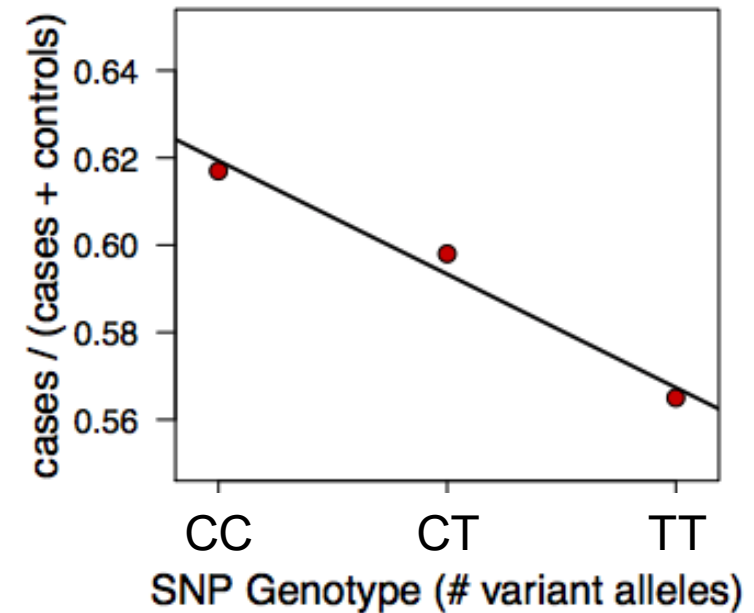
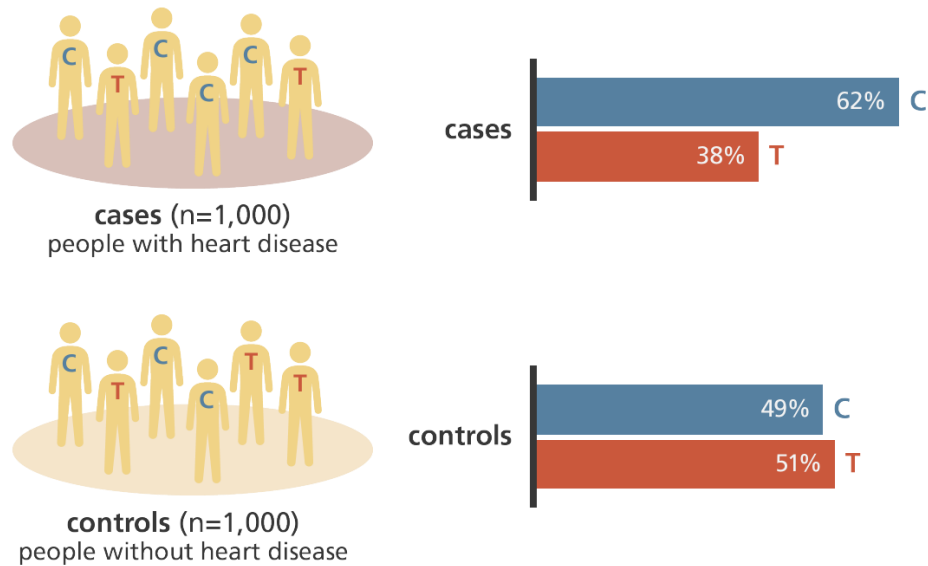


HDL cholesterol-associated region  
(Kathiresan et al Nature Genetics 2009)

# eQTL Mapping

QTLs are genetic variants that are associated with gene expression (eQTLs) or protein expression (pQTLs)

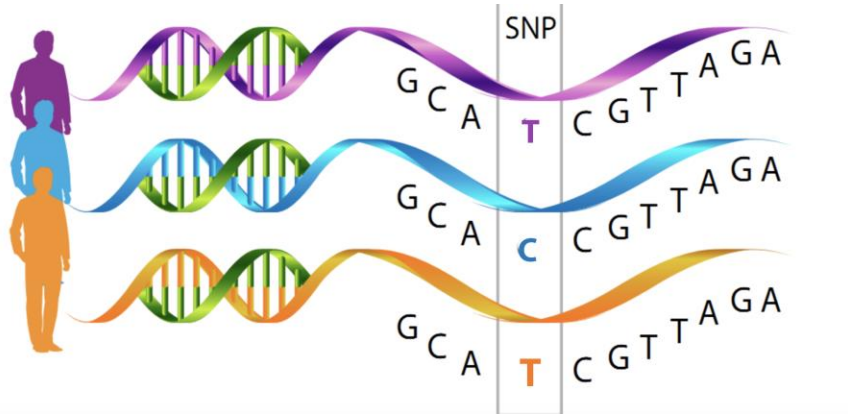
# Genetic association – binary trait



OR = increase in odds of being a case for each additional C allele



# Genetic association – quantitative trait



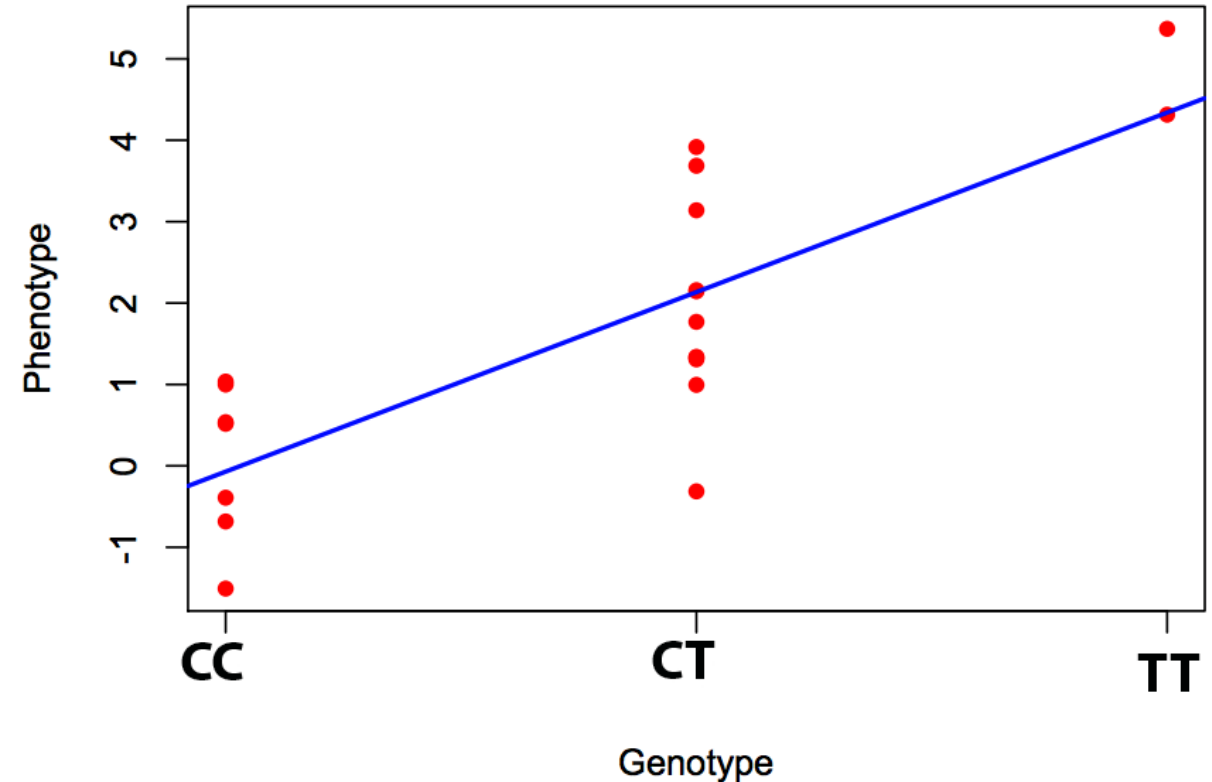
$$Y = b_0 + b_1X + e$$

**Y** phenotype e.g. LDL-cholesterol levels

***b*0** intercept

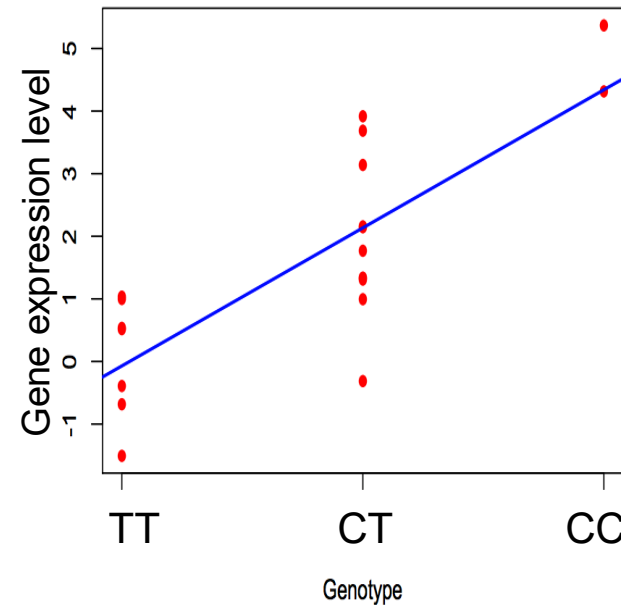
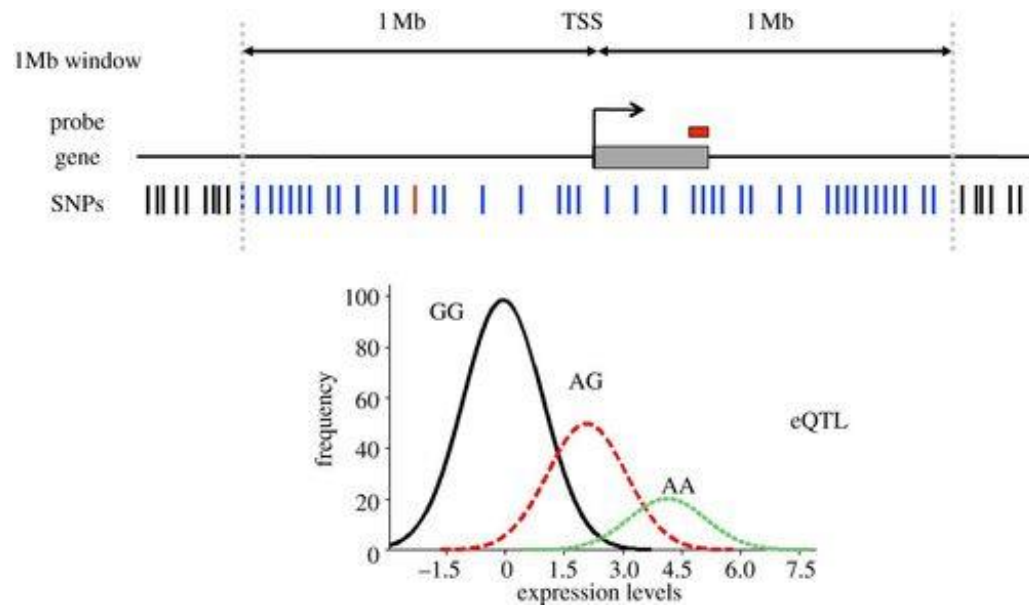
***b*1** effect of each copy of the risk allele on the mean phenotype

**e** noise or the part of y that is not explained by the SNP (e.g., environmental effect)



# Expression quantitative trait locus (eQTL)

- Gene expression is a complex phenotype with both genetic and environmental determinants
- eQTL - variant that contributes to inter-individual variation in gene expression



Test for an association between genotype group and **mean** gene expression

$$Y = b_0 + b_1X + e$$

**Y** gene expression

**b<sub>0</sub>** intercept

**b<sub>1</sub>** effect of risk allele on mean expression

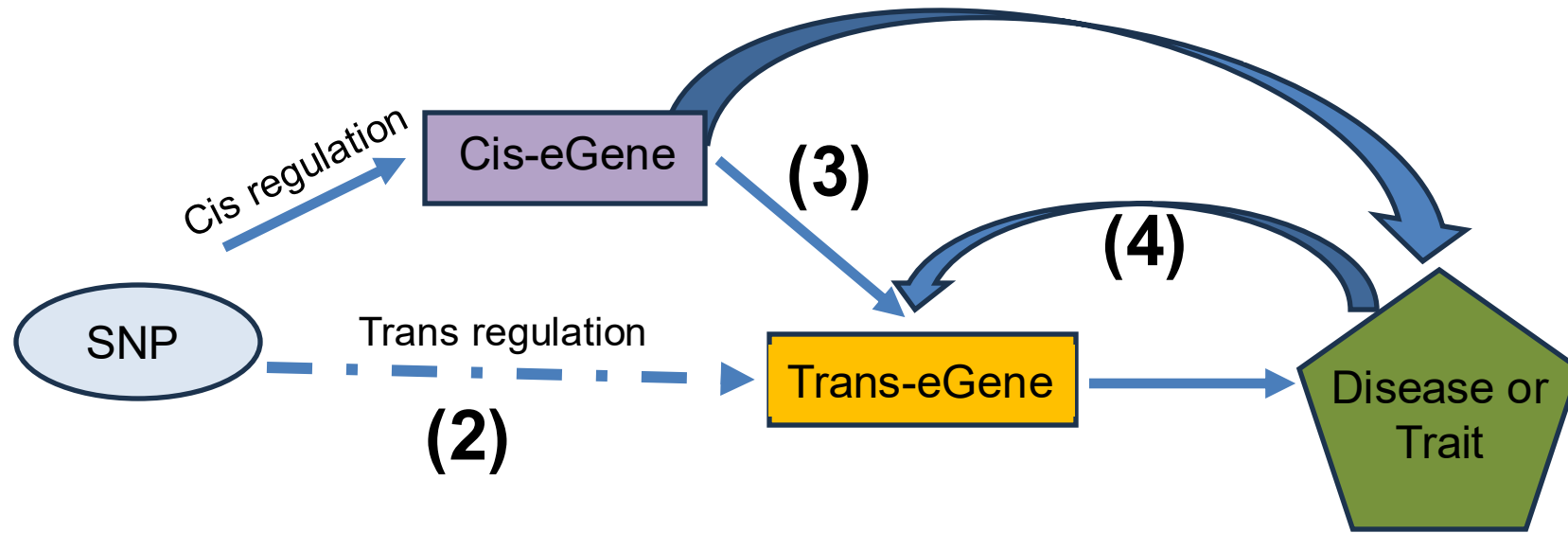
**e** noise or the part of y that is not explained by the SNP x (e.g. environmental, batch effect)

# Cis vs trans QTLs

Cis-eQTL: SNP affects a gene located  $< 1\text{Mb}$  away

Trans-eQTL: SNP affects a gene located  $> 1\text{Mb}$  away (could be on a different chromosome)

# Trans QTLs – how does a SNP affect a gene located far away?



Yao et al 2017 AJHG  
Most trans-eQTLs are on the same chromosome - likely mediated by mechanism 3

## Mediation Mechanisms of eQTLs (Yao et al 2017 AJHG)

- (1) non-coding SNP affects expression of nearby gene (*cis-eQTL*)
- (2) non-coding SNP affects remote gene expression directly (*trans-eQTL*) (e.g. SNP in an enhancer region)
- (3) *cis-eGene* mediation of the *trans-eGene* (e.g. if *cis-eGene* is a TF for *trans-eGene*);
- (4) reverse causality (trait has feedback effect on gene expression).

# Performing eQTL mapping

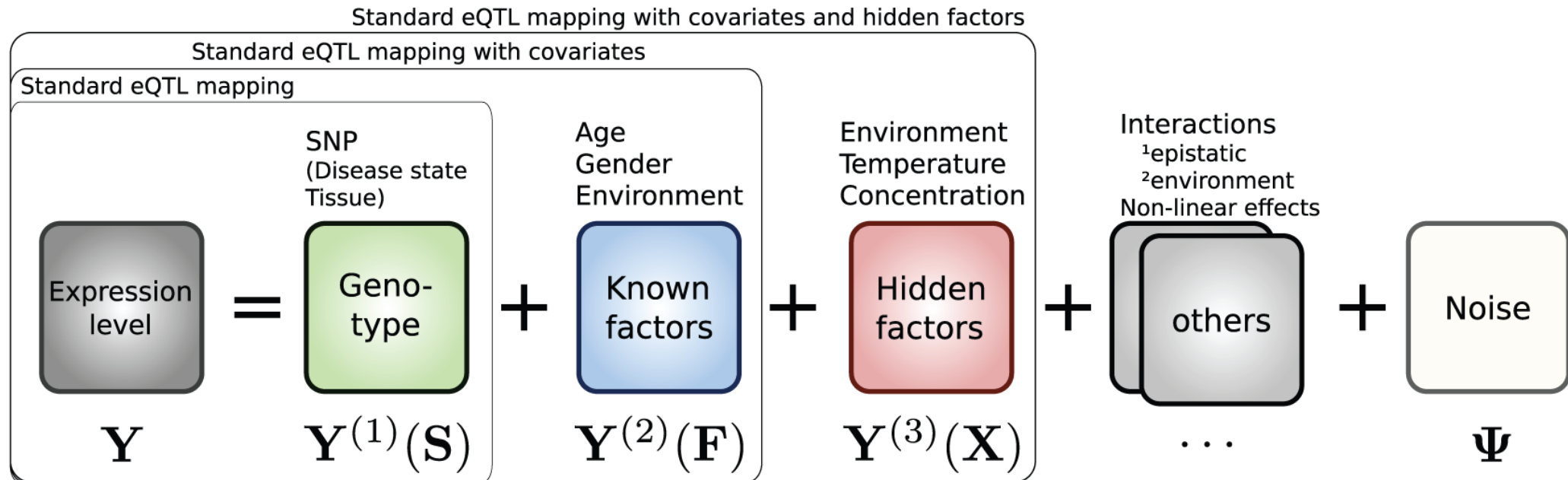
# Performing genome-wide QTL mapping

1. Need to measure gene expression (bulk tissue or single cell) and generate genotype data in the same individuals
2. Separate processing of genotype and expression data.  
(Expression data requires quality control, normalisation and correcting for batch factors and unmeasured confounders)



# eQTL Mapping – Covariate adjustment

- Covariate adjustment

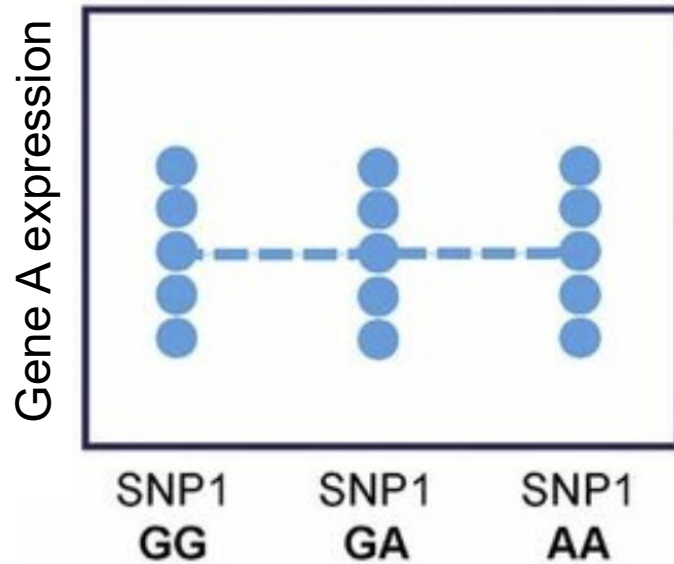


## Performing genome-wide QTL mapping

1. Need to measure gene expression (bulk tissue or single cell) and generate genotype data in the same individuals
2. Separate processing of genotype and expression data.
3. Conduct a GWAS for every single gene, where gene expression levels are your phenotype

# Genome-wide QTL mapping software

SNP 1 is not an eQTL  
for Gene A



Test every single SNP with every single gene – A LOT of tests (requires multiple testing correction) and computation power required

# Genome-wide QTL mapping software

**Plink** – most-commonly used software for GWAS - legacy approach

- default software to manipulate genetic files and run genetic association analysis

**matrix eQTL** (2012)

[http://www.bios.unc.edu/research/genomic\\_software/Matrix\\_eQTL/](http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/)

- Computationally efficient
- fast performance is achieved by special data pre-processing and using matrix operations
- Calculates FDR only for gene-SNP pairs that pass a user-defined significance

**fastQTL** (2016) <https://hpc.nih.gov/apps/FastQTL.html>

- faster processing time (16× faster than matrix QTL).
- Different permutation schemes available for multiple testing correction

**Table 1.** FastQTL and Matrix eQTL running times

Number of permutations	Matrix eQTL	FastQTL		
		1000	500	100
GTEX_AS	337.4	19	10.8	3.5
GTEX_AT	330.3	21	10.8	3.6
GTEX_HLV	312.4	15	8	3
GTEX_L	364.9	25.6	13.1	4
GTEX_MS	335.9	23.6	12.6	3.9
GTEX_NT	343.8	18.4	9.5	3.4
GTEX_SSEL	349.7	20.7	10.8	3.6
GTEX_T	358.1	22.3	11.8	3.9
GTEX_WB	340.5	25.5	13.7	4.1
ALL	3073	191.1	101.1	33

Table 1 shows the running times in CPU hours to produce the results shown in Figure 2e; nine GTEx datasets (column 1) processed with 1000 Matrix eQTL permutations (column 2) and FastQTL with 1000 (column 3), 500 (column 4) and 100 permutations (column 5). Total running times for all nine datasets together are shown in the last row.

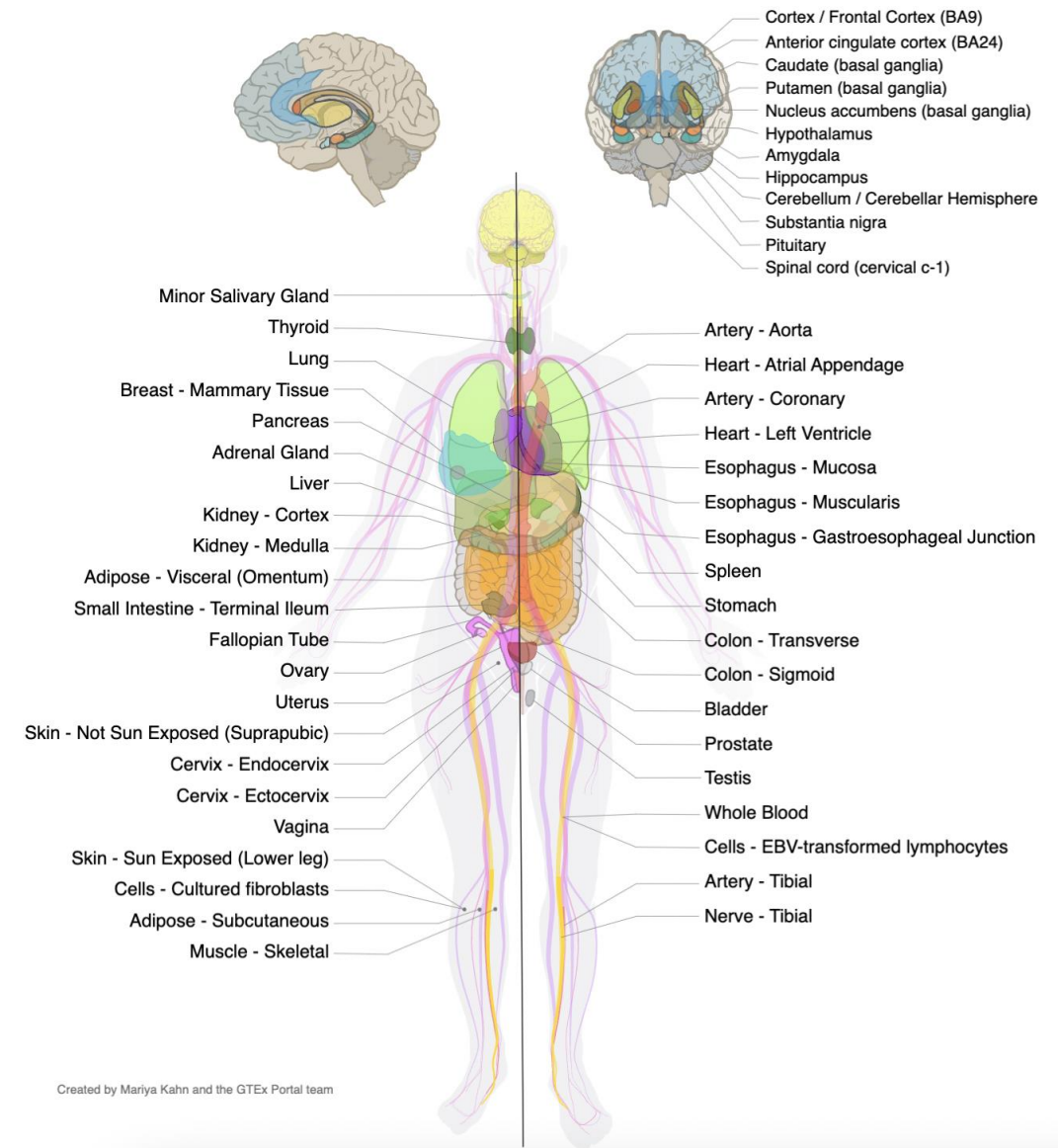
# eQTL data resources

# The GTEx Project

- Launched 2010
- Now at GTEx v10
- Catalogue of genetic effects on gene expression across a large number of human tissues
- eQTL data from 943 donors and 19,466 samples in 50 tissues
- Gene expression (RNA-seq) and genotype data (WGS data)

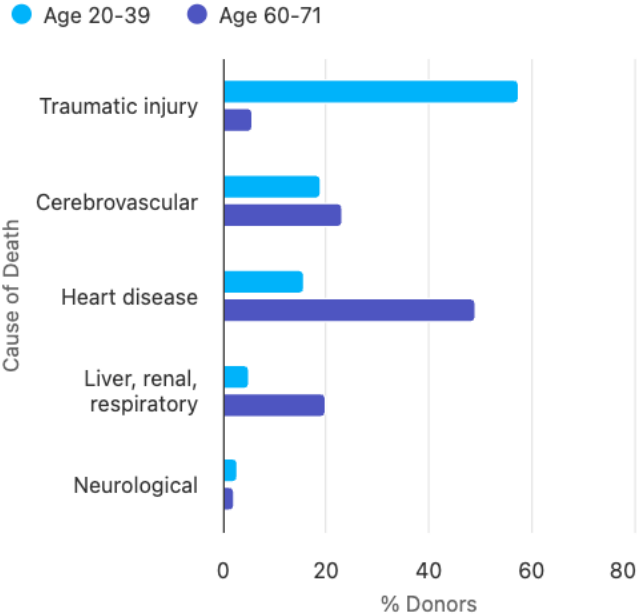
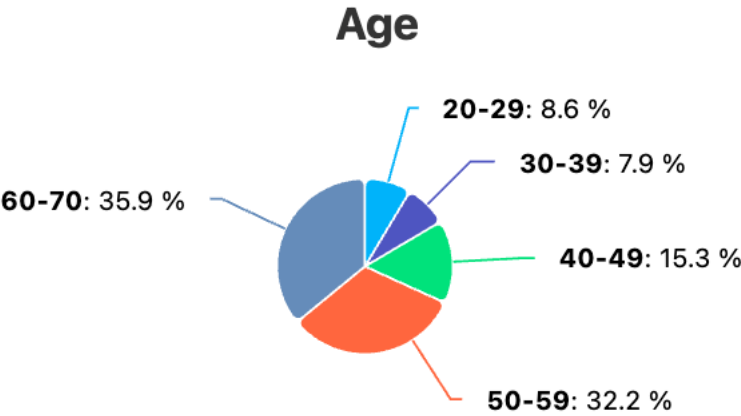
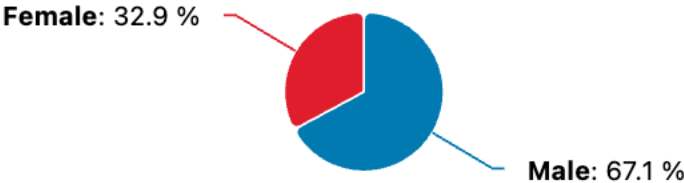
**The GTEx Consortium atlas of genetic regulatory effects across human tissues**

**Science 2020 [DOI: 10.1126/science.aaz1776](https://doi.org/10.1126/science.aaz1776)**



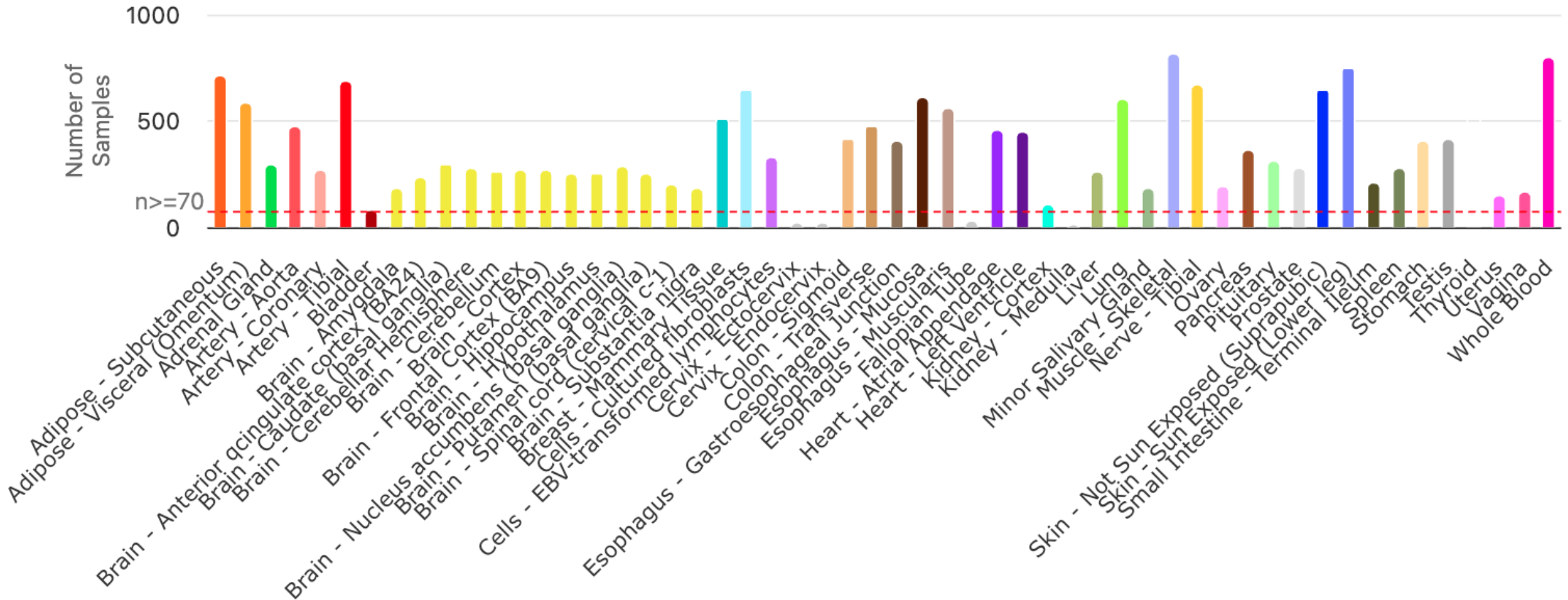


# The GTEx Project

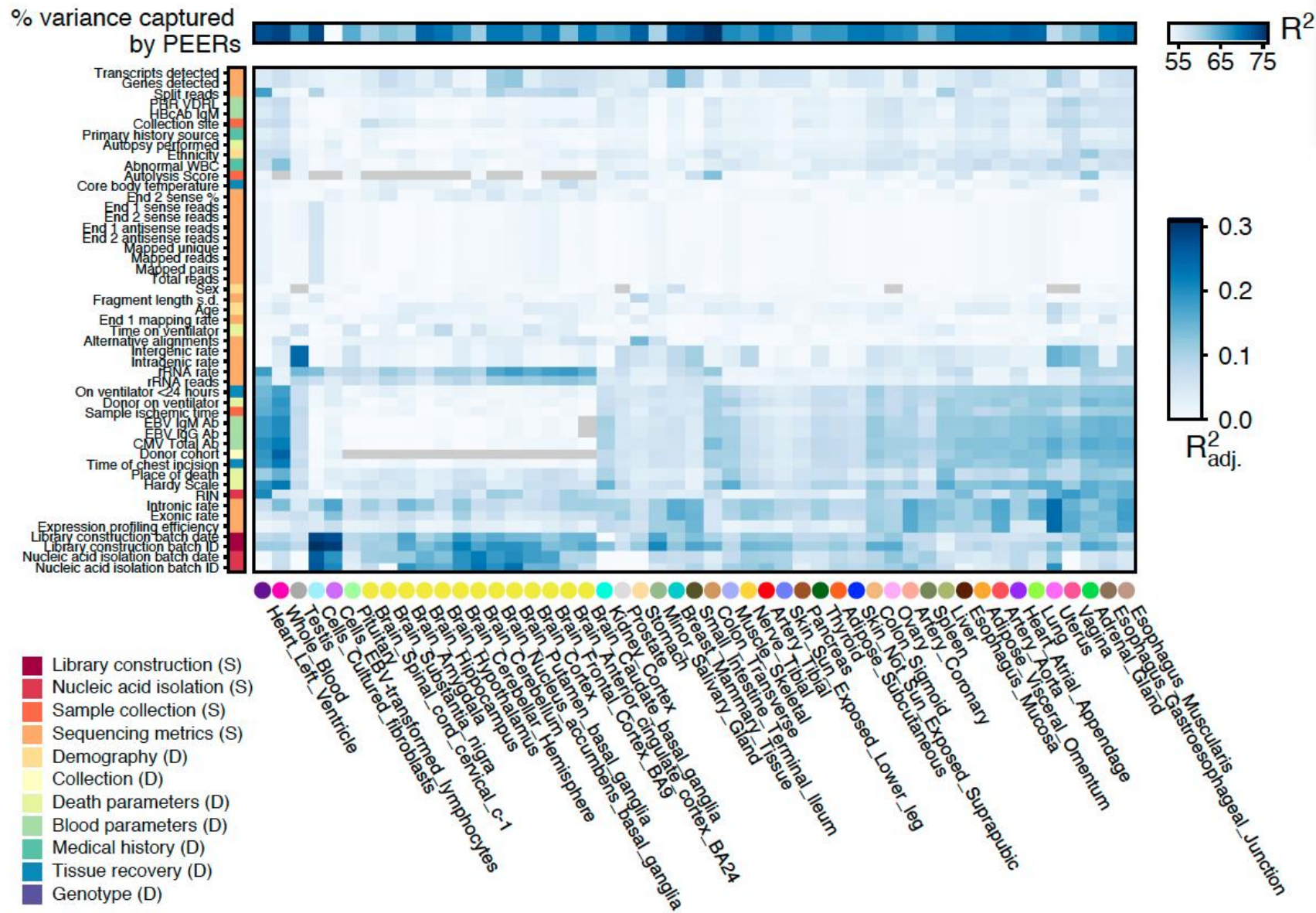


Cause of Death	Age 20 - 39	Age 60 - 71
Traumatic injury	57.4%	5.6%
Cerebrovascular	19.1%	23.2%
Heart disease	15.6%	49.2%
Liver, renal, respiratory	5.0%	19.8%
Neurological	2.8%	2.2%

# Tissue sample size



# PEER analysis – accounting for batch effects in expression data



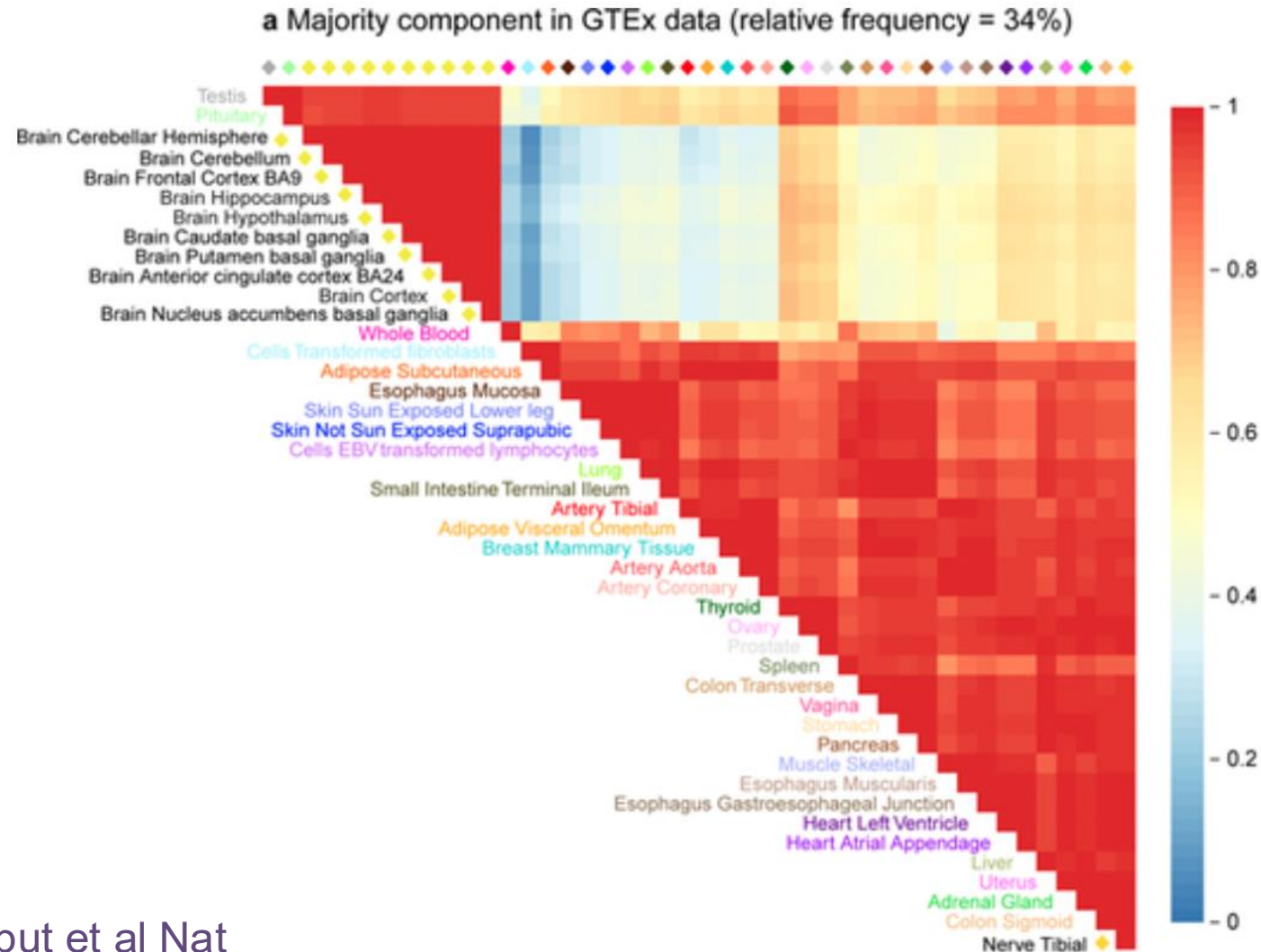
Covariates most consistently associated with PEER factors include factors related to donor death, ischemic time, sequencing quality control metrics, and nucleic acid isolation and library construction batches.

# The GTEx Project

- cis-eQTLs identified (at 5% FDR per tissue) for 94.7% of all protein-coding and 67.3% of all lincRNA genes detected in at least one tissue
- most cis-eQTLs had small effect sizes: an average of 22% of cis-eQTLs had allelic Fold Change > twofold
- Genes lacking a cis-eQTL enriched for those not expressed in the tissues analysed, including genes involved in early development
  - Bulk RNAseq – may lose cell-specific effects
- Interchromosomal trans-eQTLs for 143 trans eGenes



# Many eQTLs are shared across tissues

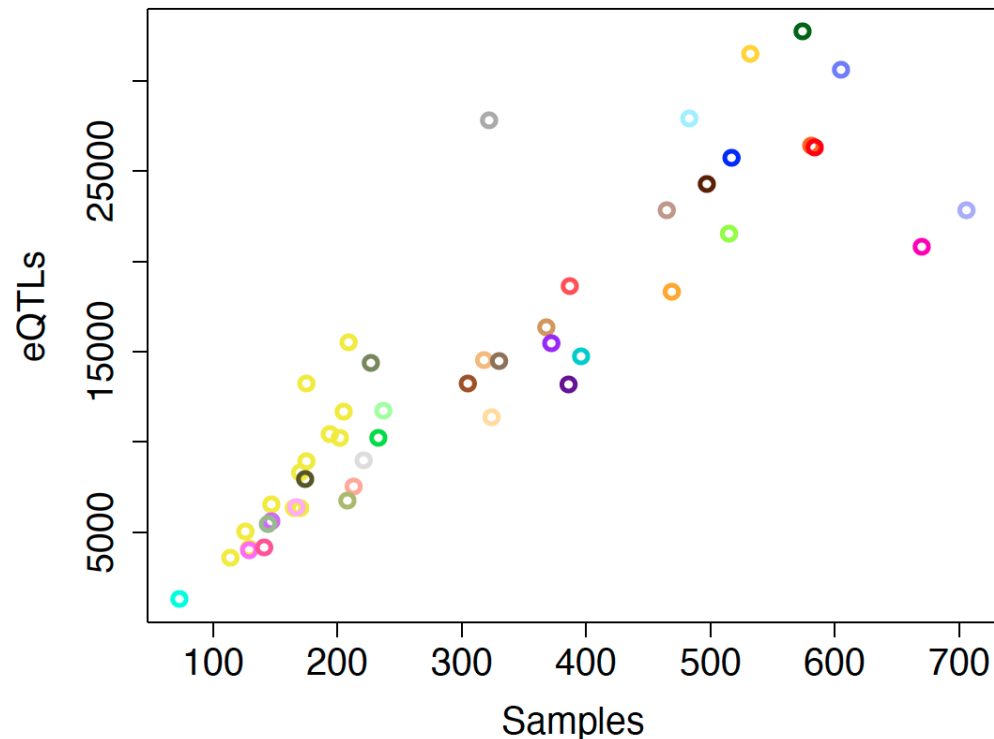


Correlation of eQTL effect estimates for 16,069 (genes expressed and have effect estimates in all 44 tissues)

- (1) effects are **positively** correlated among all tissues;
- (2) the brain tissues—and, to a lesser extent, testis and pituitary—are particularly strongly correlated with one another, and less correlated with other tissues;
- (3) effects in whole blood less well correlated with other tissues

# Power to detect an eQTL

- eQTL discovery related to sample size
- Increased discovery in larger samples driven by increased power to detect small effects
- Papers that identify tissue-specific genes use p-value threshold for detecting eqtls
- Power is an important consideration when trying to identify tissue-specific eQTLs
  - Down-sampling to check tissue-specificity.



# GTEx web browser

## Explore GTEx



### Browse

Browse and search all data by gene

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 catalogue

<https://www.ebi.ac.uk/eql/>

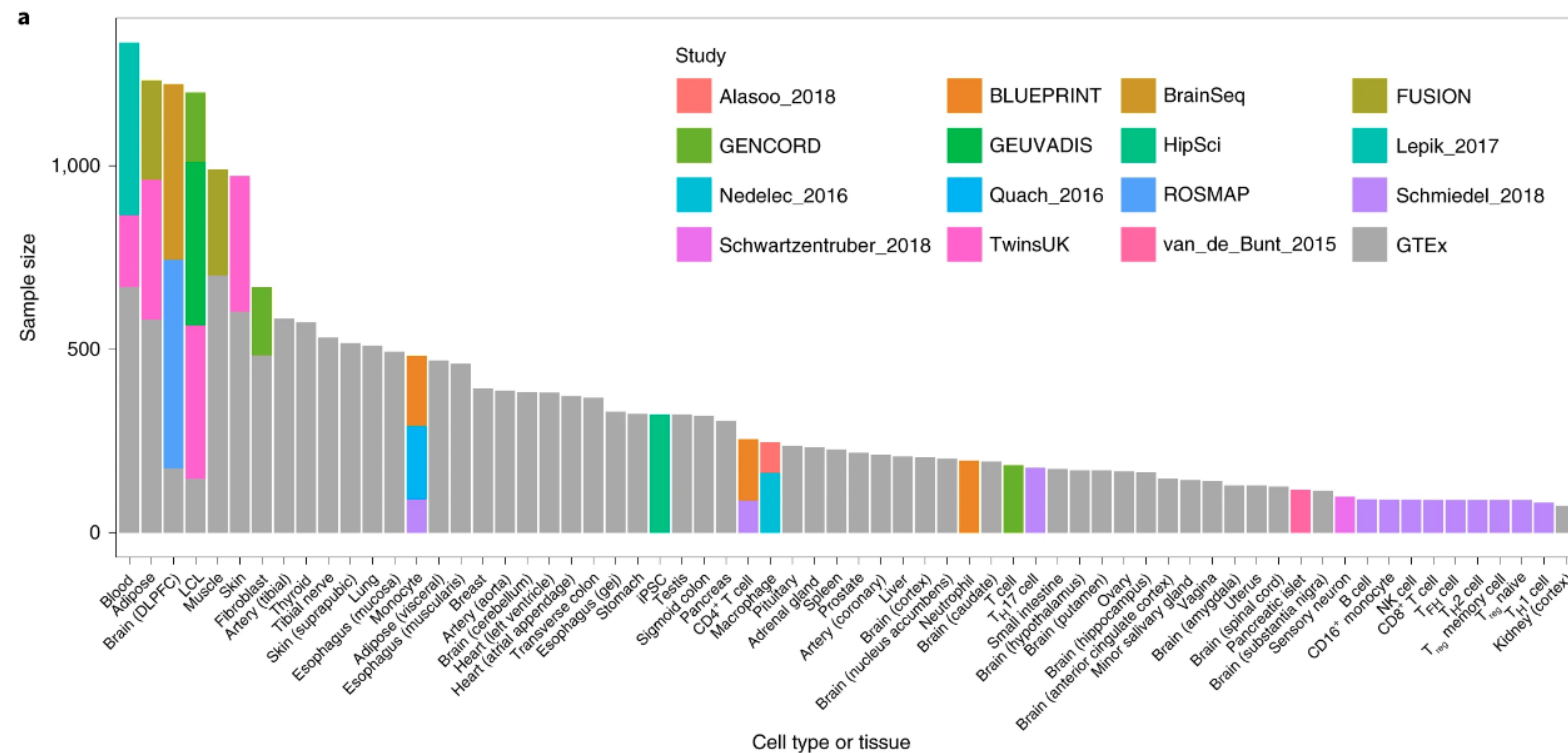
Provides uniformly processed cis-eQTLs and sQTLs from all available public studies on human.

Article | [Open Access](#) | Published: 06 September 2021

## A compendium of uniformly processed human gene expression and splicing quantitative trait loci

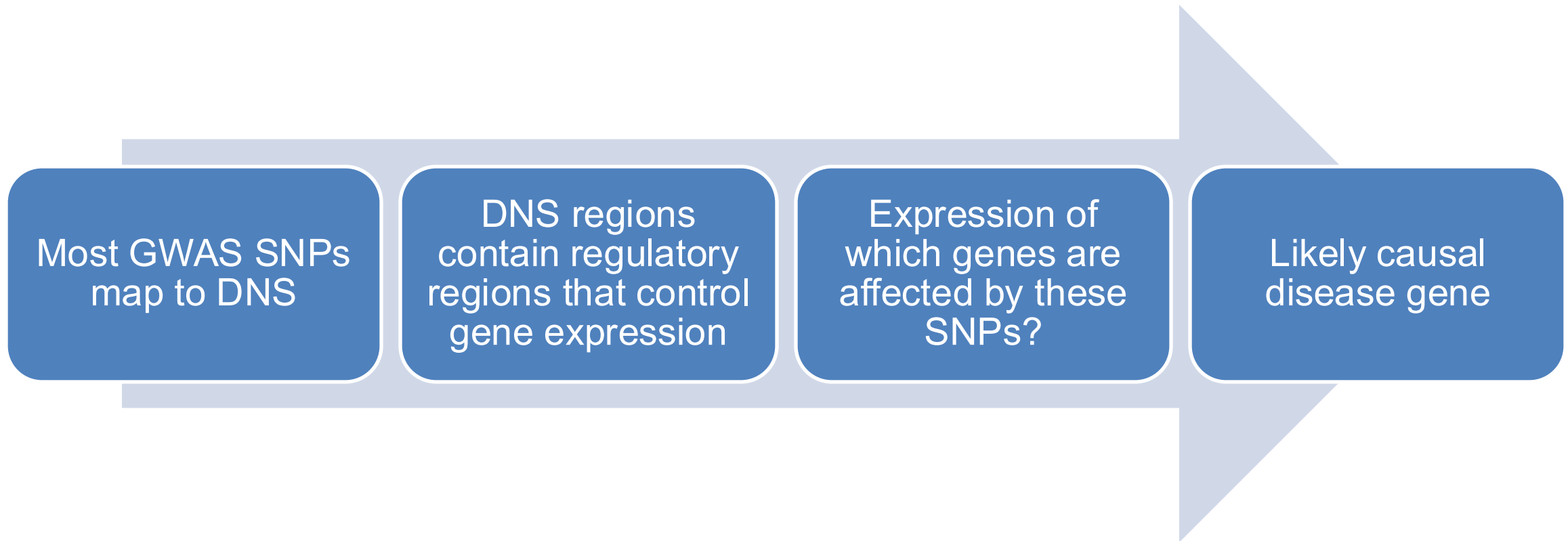
[Nurlan Kerimov](#), [James D. Hayhurst](#), [Kateryna Peikova](#), [Jonathan R. Manning](#), [Peter Walter](#), [Liis Kolberg](#), [Marija Samoviča](#), [Manoj Pandian Saktivel](#), [Ivan Kuzmin](#), [Stephen J. Trevanion](#), [Tony Burdett](#), [Simon Jupp](#), [Helen Parkinson](#), [Irene Papatheodorou](#), [Andrew D. Yates](#), [Daniel R. Zerbino](#) & [Kaur Alasoo](#)

*Nature Genetics* **53**, 1290–1299 (2021) | [Cite this article](#)





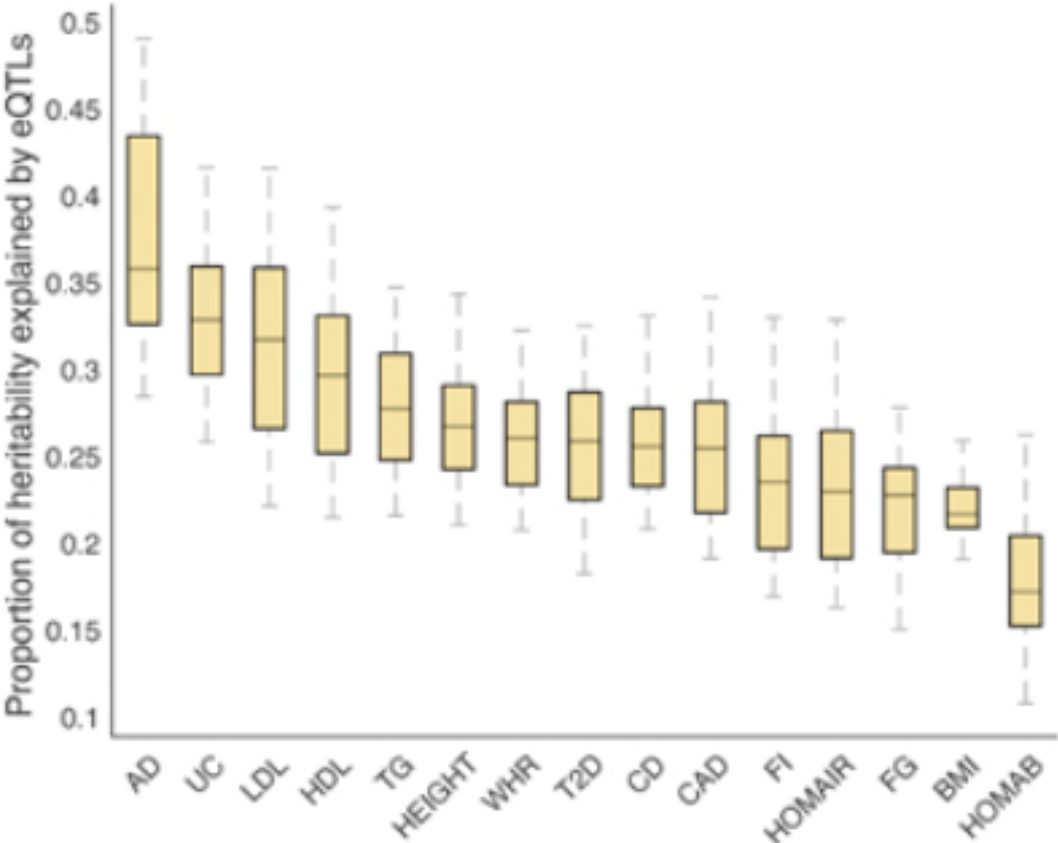
# eQTL mapping could give clues to causal genes



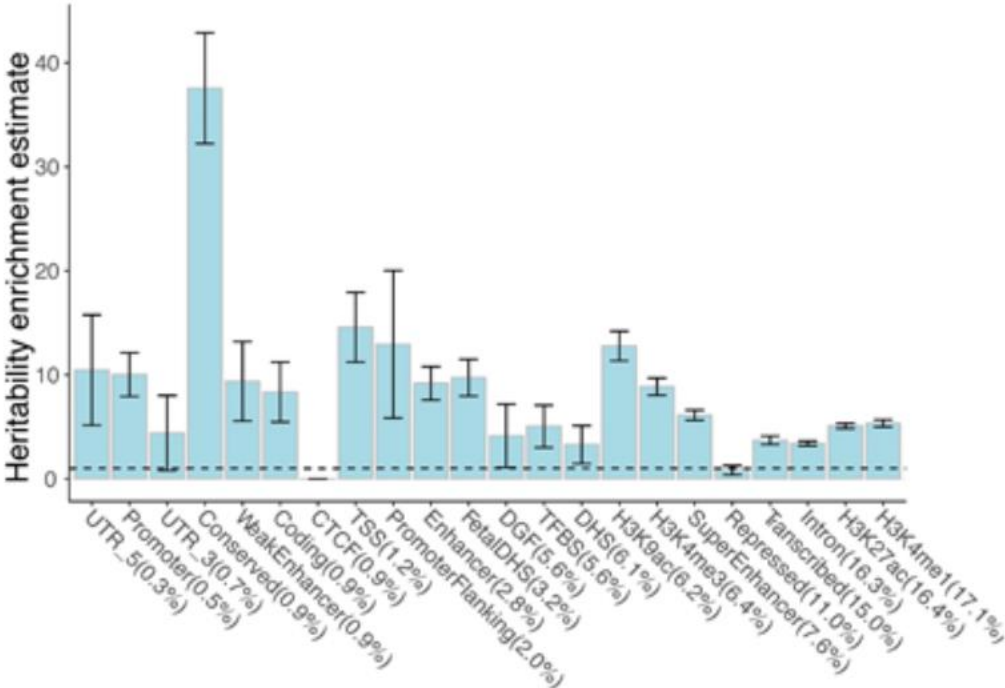
Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation

Eric R Gamazon<sup>1,2,§,\*</sup>, Ayellet V Segre<sup>3,4,§,\*</sup>, Martijn van de Bunt<sup>5,6,§</sup>, Xiaoquan Wen<sup>7</sup>, Hualin S Xi<sup>8</sup>, Farhad

Distribution of proportion of heritability of 15 traits explained by eQTLs in 44 tissues (GTEx data)



Heritability enrichment estimate computed for subsets of eQTLs that fall in different genomic features



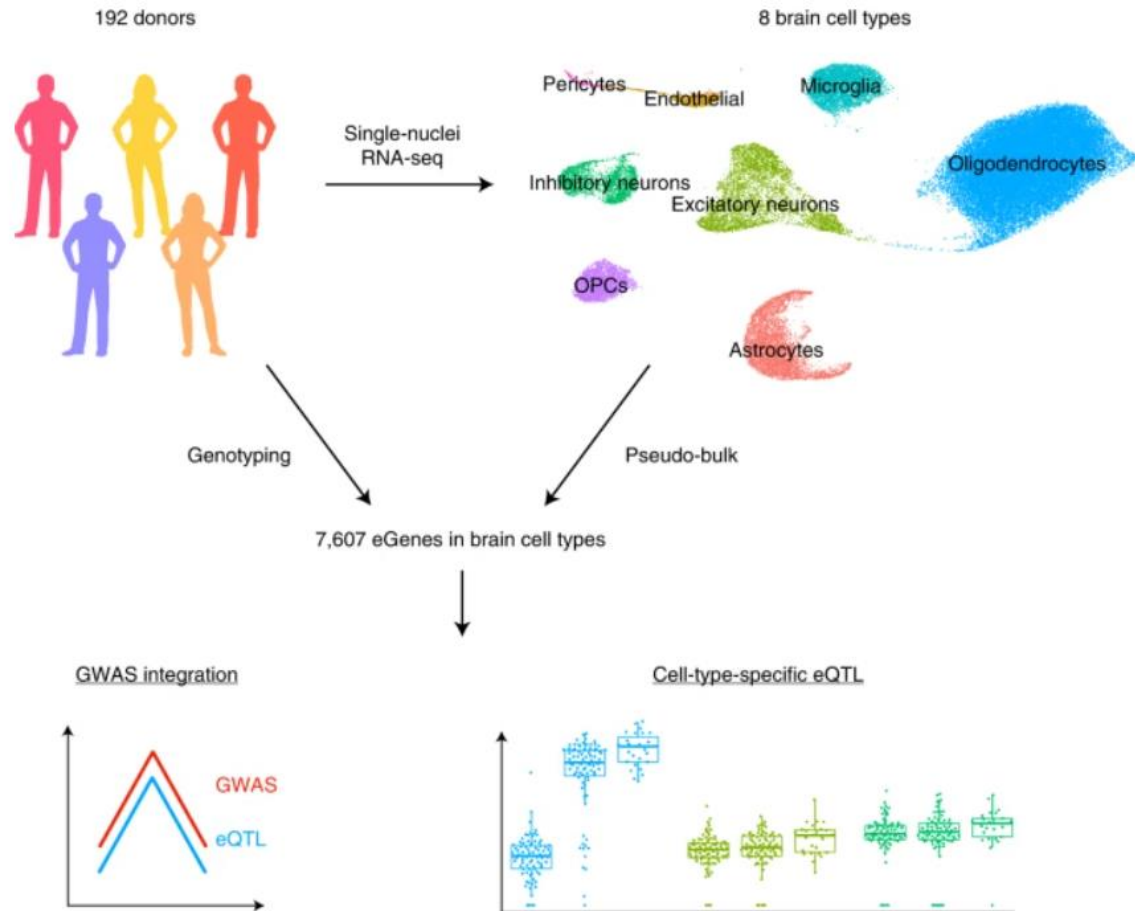
eQTLs can be context-dependent

# Context-specific eQTLs

- eQTLs can be time-dependent or environment-specific
  - Cell-type-specific expression
  - Response to treatment
  - Developmental stage
- Most eQTL studies measure gene expression
  - At a **single timepoint**
  - In adult tissues
  - Using bulk RNA seq methods (no information on cell type differences)

# Cell-type-specific QTLs

**Fig. 1: Study summary.**



We performed single-nuclei RNA-seq on brain samples from 192 genotyped donors. We mapped *cis*-eQTLs for eight major brain cell types and identified a total of 7,607 *cis*-eQTL genes. We identified cell-type-specific genetic effects and leveraged our results to identify risk genes for brain disorders.

[nature](#) > [nature neuroscience](#) > [resources](#) > [article](#)

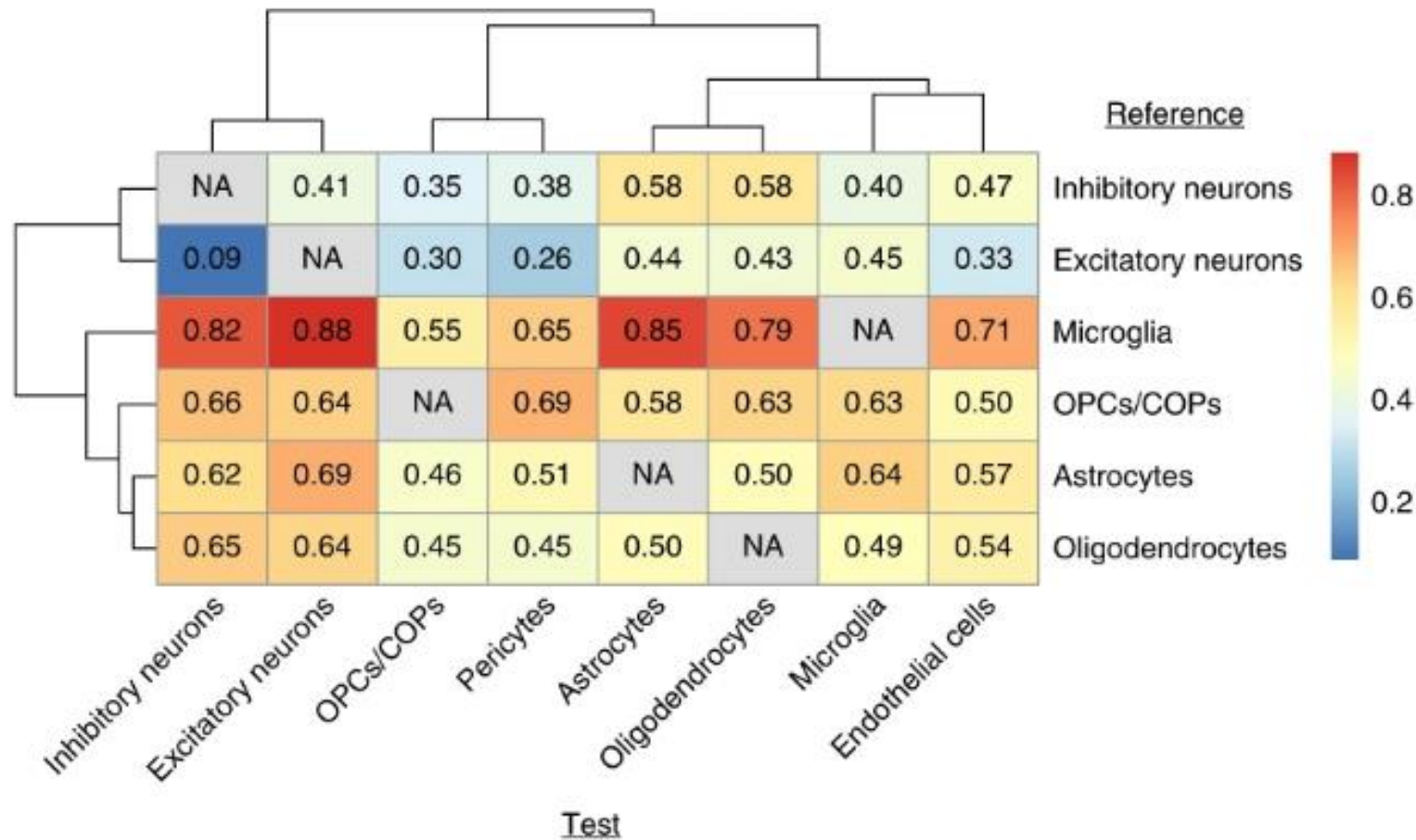
Resource | Published: 01 August 2022

## Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders

[Julien Bryois](#) , [Daniela Calini](#), [Will Macnair](#), [Lynette Foo](#), [Eduard Urich](#), [Ward Ortmann](#), [Victor Alejandro Iglesias](#), [Suresh Selvaraj](#), [Erik Nutma](#), [Manuel Marzin](#), [Sandra Amor](#), [Anna Williams](#), [Gonçalo Castelo-Branco](#), [Vilas Menon](#), [Philip De Jager](#) & [Dheeraj Malhotra](#) 

[Nature Neuroscience](#) **25**, 1104–1112 (2022) | [Cite this article](#)

# Cell-type-specific QTLs



Estimates of the proportions of cis-eQTLs that have a different genetic effect in another cell type.

[nature](#) > [nature neuroscience](#) > [resources](#) > [article](#)

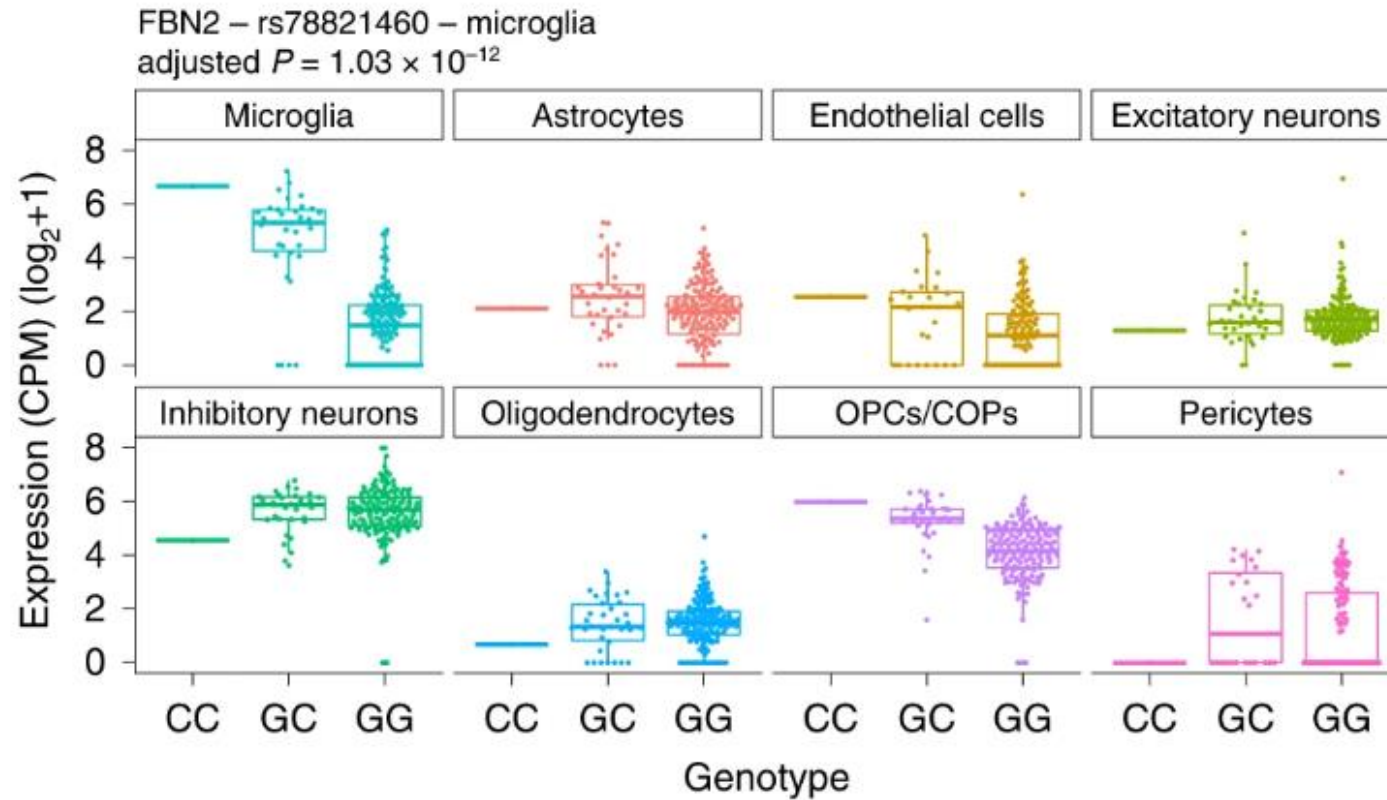
Resource | Published: 01 August 2022

**Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders**

[Julien Bryois](#) , [Daniela Calini](#), [Will Macnair](#), [Lynette Foo](#), [Eduard Urich](#), [Ward Ortmann](#), [Victor Alejandro Iglesias](#), [Suresh Selvaraj](#), [Erik Nutma](#), [Manuel Marzin](#), [Sandra Amor](#), [Anna Williams](#), [Gonçalo Castelo-Branco](#), [Vilas Menon](#), [Philip De Jager](#) & [Dheeraj Malhotra](#) 

[Nature Neuroscience](#) **25**, 1104–1112 (2022) | [Cite this article](#)

# Cell-type-specific QTLs



[nature](#) > [nature neuroscience](#) > [resources](#) > [article](#)

Resource | Published: 01 August 2022

**Cell-type-specific *cis*-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders**

[Julien Bryois](#) , [Daniela Calini](#), [Will Macnair](#), [Lynette Foo](#), [Eduard Urich](#), [Ward Ortmann](#), [Victor Alejandro Iglesias](#), [Suresh Selvaraj](#), [Erik Nutma](#), [Manuel Marzin](#), [Sandra Amor](#), [Anna Williams](#), [Gonçalo Castelo-Branco](#), [Vilas Menon](#), [Philip De Jager](#) & [Dheeraj Malhotra](#) 

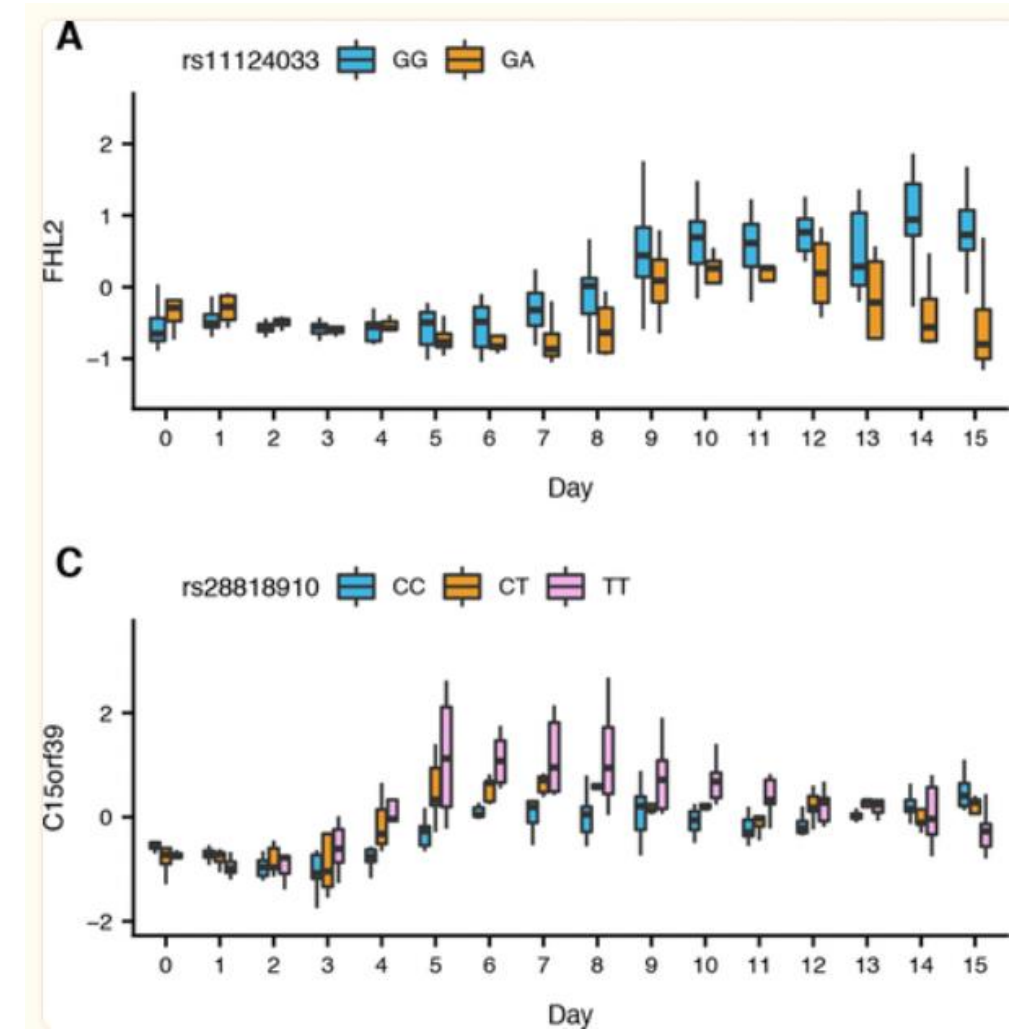
[Nature Neuroscience](#) **25**, 1104–1112 (2022) | [Cite this article](#)



# Developmental stage-specific QTLs



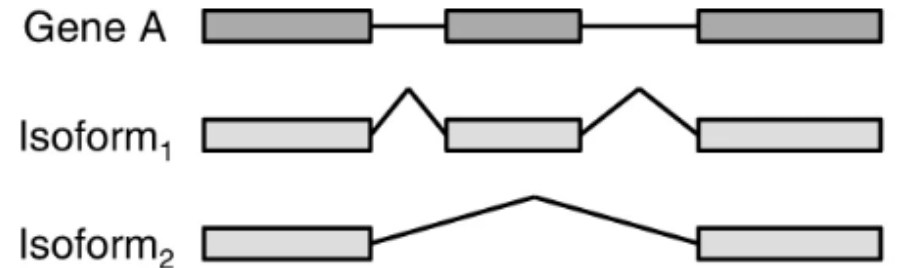
- iPSC differentiation into cardiomyocytes.
- eQTL analysis at 16 time points in 19 human cell lines



# Splice QTLs

# splice QTLs (sQTLs)

- Alternative splicing (AS) produces multiple transcript isoforms from a single gene  
tissue-, cell type-, or condition-specific
- sQTLs - genetic variants that regulate AS
- sQTLs may change:
  - UTRs, affecting RNA stability or translational efficacy
  - Coding sequence by skipping or inclusion of coding exons, affecting protein structure and function



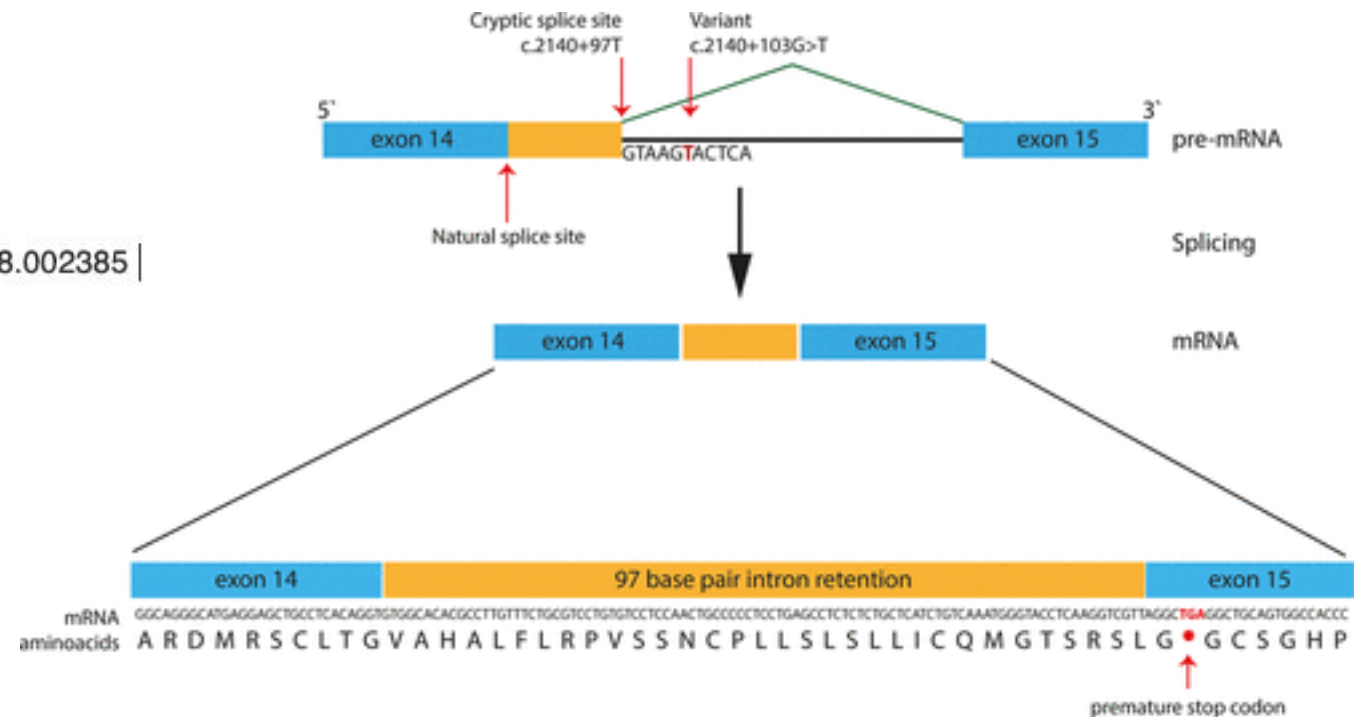
# Splice variants

## A Deep Intronic Variant in *LDLR* in Familial Hypercholesterolemia

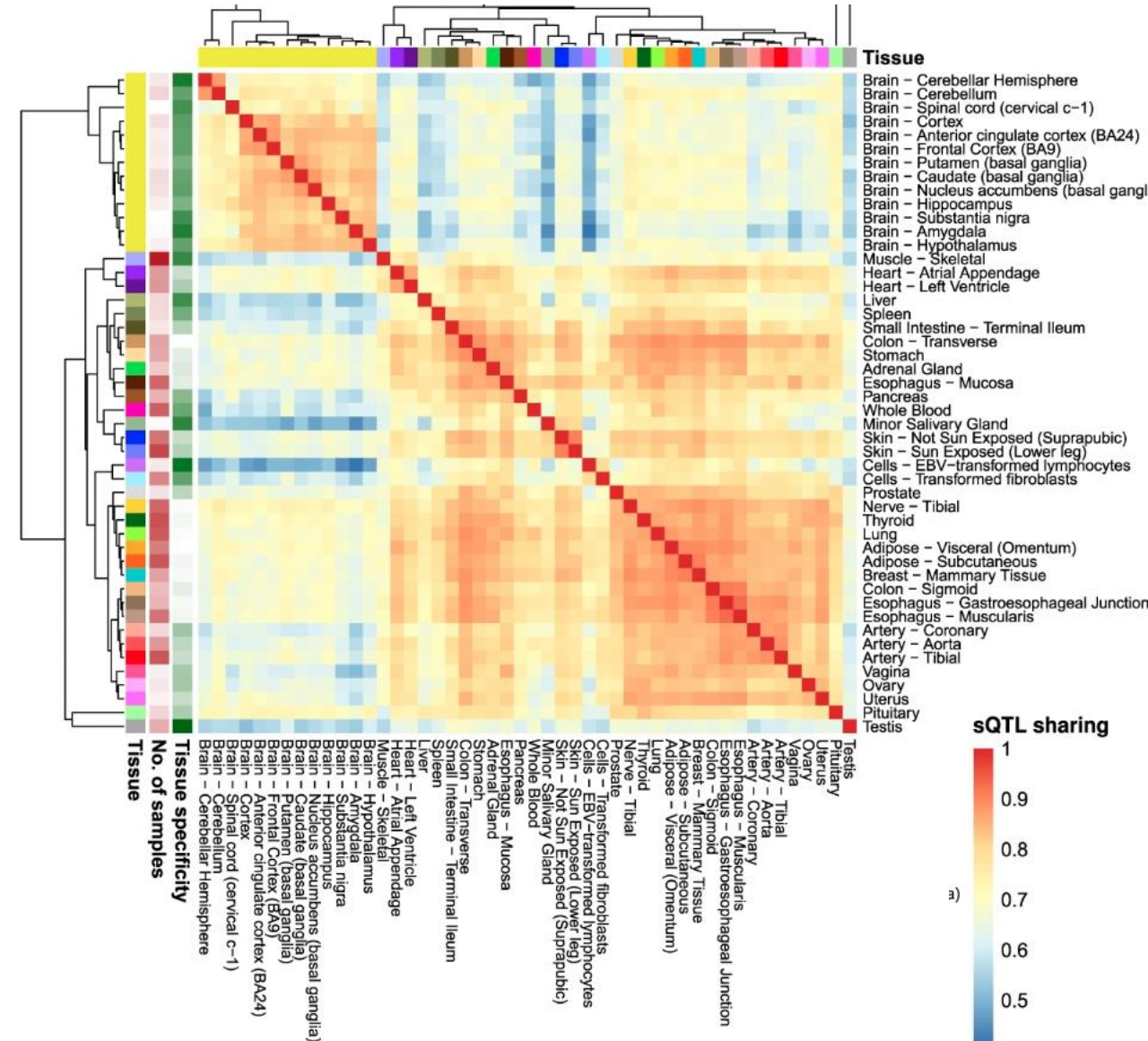
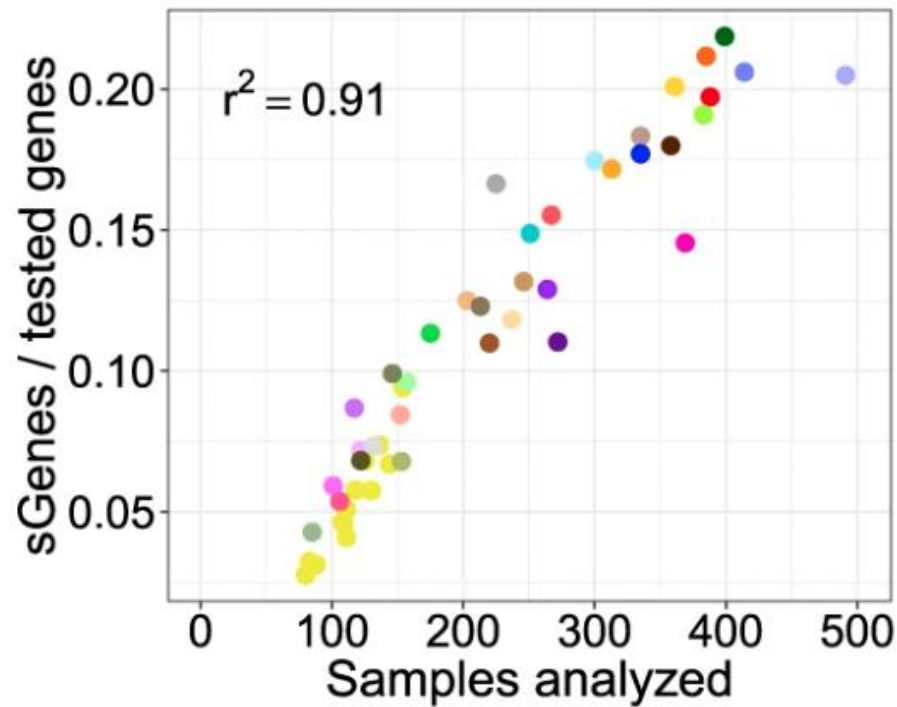
### Time to Widen the Scope?

Laurens F. Reeskamp, Merel L. Hartgers, Jorge Peter, Geesje M. Dallinga-Thie, Linda Zuurbier, Joep C. Defesche, Aldo Grefhorst and G. Kees Hovingh ✉

Originally published 11 Dec 2018 | <https://doi.org/10.1161/CIRCGEN.118.002385> | Circulation: Genomic and Precision Medicine. 2018;11:e002385

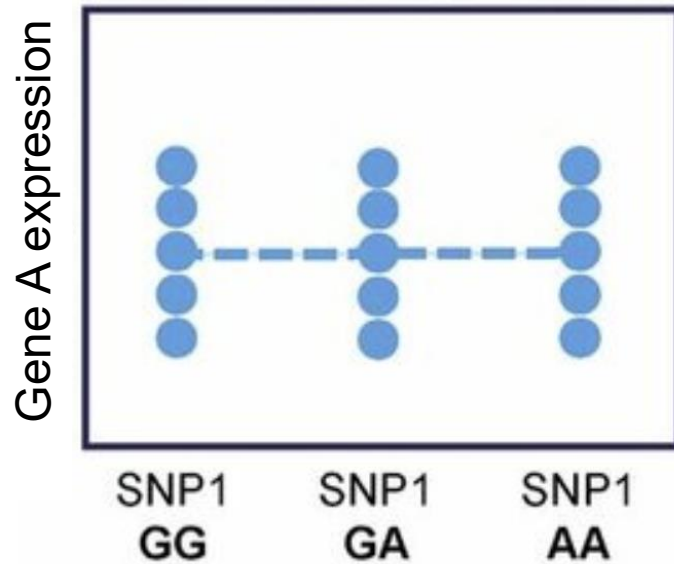


# sQTL sharing across tissues

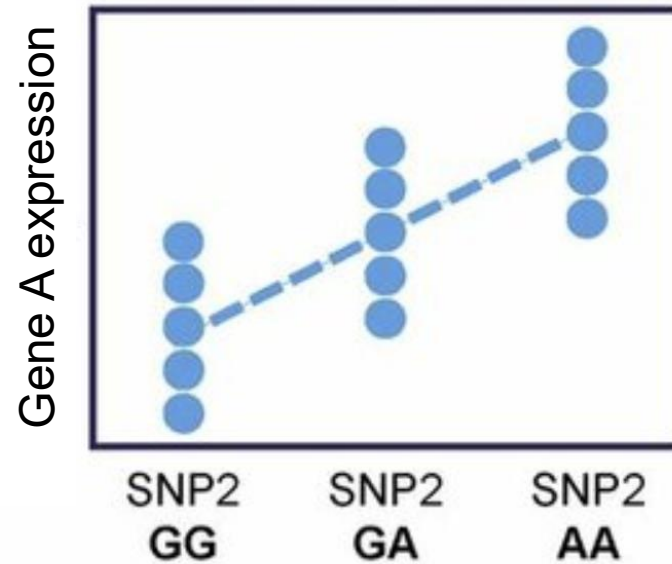


# Stimulus-dependent eQTL

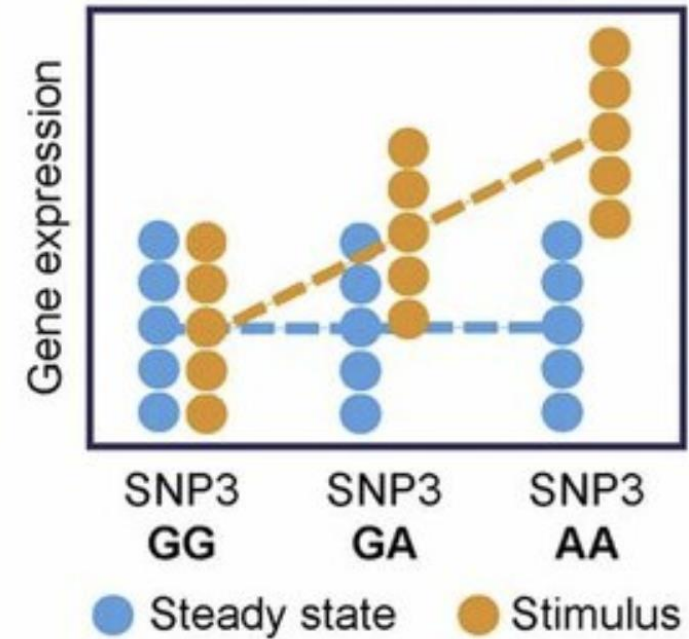
SNP 1 is not an eQTL  
for Gene A



SNP 2 is an eQTL for  
Gene A



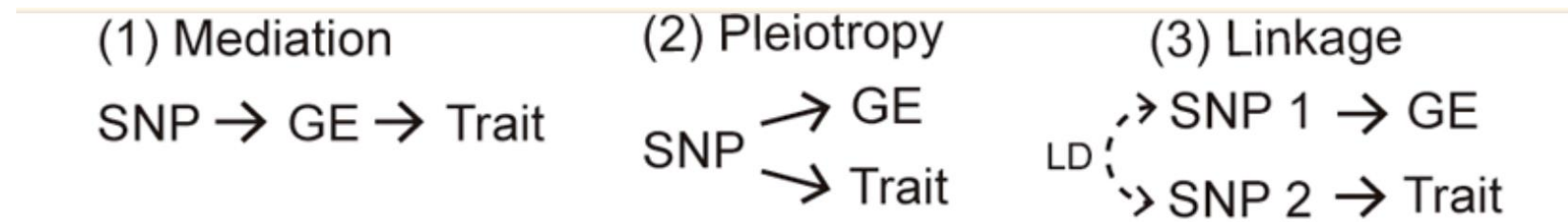
SNP3 is a «response» eQTL





# Association does not mean causality

Several different causal scenarios can result in similar patterns of heritability enrichment or overlap between GWAS loci and eQTLs



- Co-regulation of nearby genes – multiple eGenes for the same SNPs
- Unclear which is the causal gene just based on overlapping association
- Several statistical methods to determine if a variant impacts phenotype through gene expression change (SMR, coloc)

Changes in gene expression doesn't necessarily translate to changes in protein levels

Integration of genetic and omic data for greater understanding of variant consequence

