

# UQ Genetics and Genomics Winter School 2023

## Systems Genomics and Pharmacogenomics Module 6

# 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

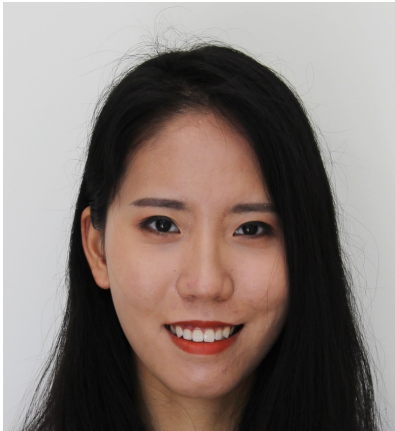
To maximize your learning experience, we will be working with genuine human genetic data, during this module.

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

Please email [pctgadmin@imb.com.au](mailto:pctgadmin@imb.com.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.”

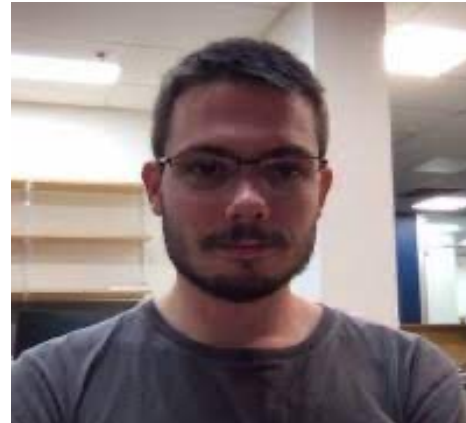
# Systems Genomics and Pharmacogenomics Module



Clara  
Jiang



Gagandeep  
Singh



Solal  
Chauquet



Sonia  
Shah



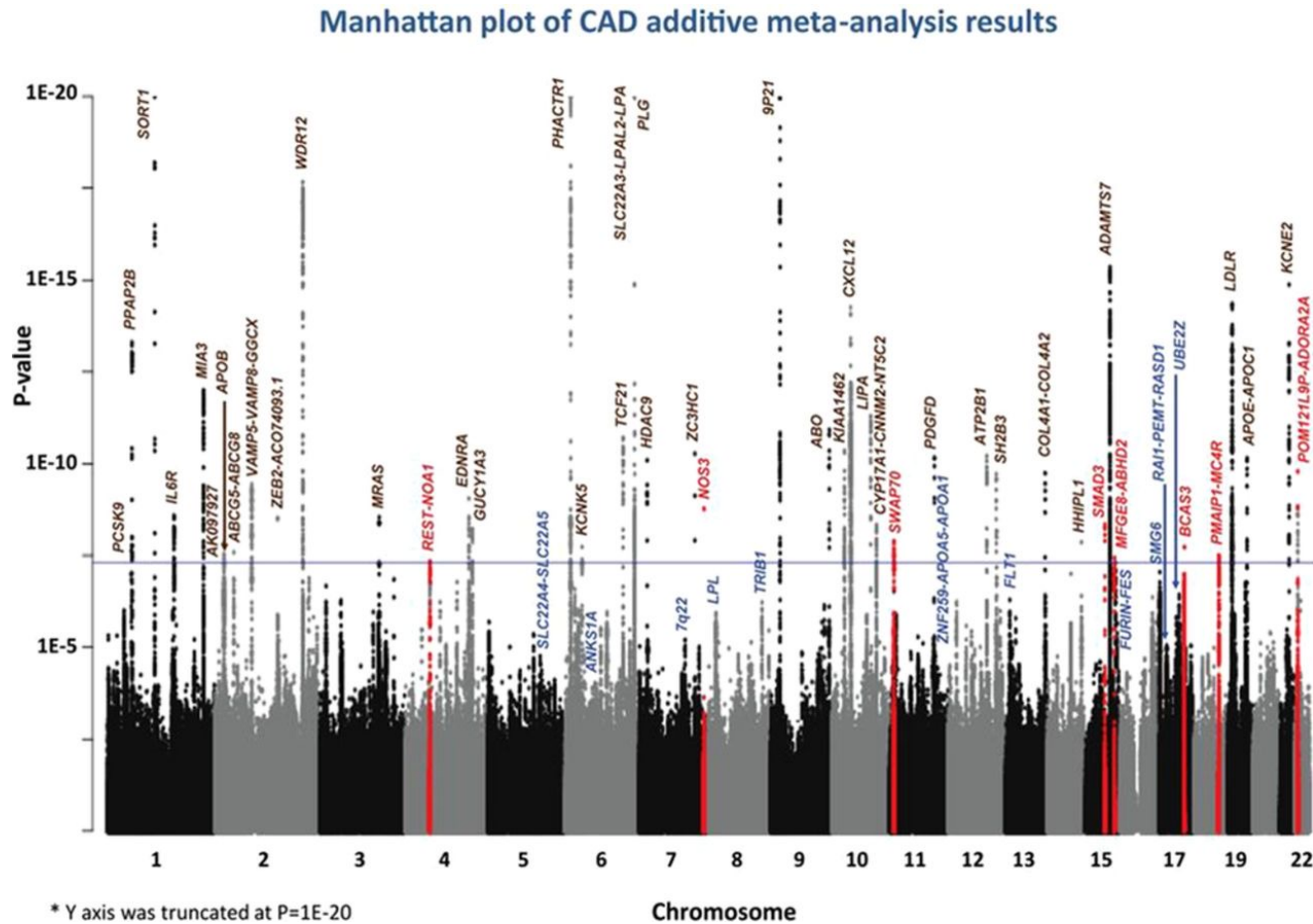
Zhihong  
Zhu

# eQTL Mapping Lecture

# Lecture overview

1. Translation of GWAS results
2. What is an eQTL?
3. Performing eQTL mapping
4. eQTL data resources
5. Dynamic QTLs
6. Splice QTLs

# Challenge: Linking GWAS SNPs to mechanism




**GWAS Catalog**  
The NHGRI-EBI Catalog of human genome-wide association studies

Search the catalog

Examples: breast carcinoma, rs7329174, Yao, 2q37.1, HBS1L, 6:16000000-25000000

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

**As of 2023-06-03**  
6401 publications  
529,481 top associations  
60,071 full summary statistics



# Challenge: Linking GWAS SNPs to mechanism

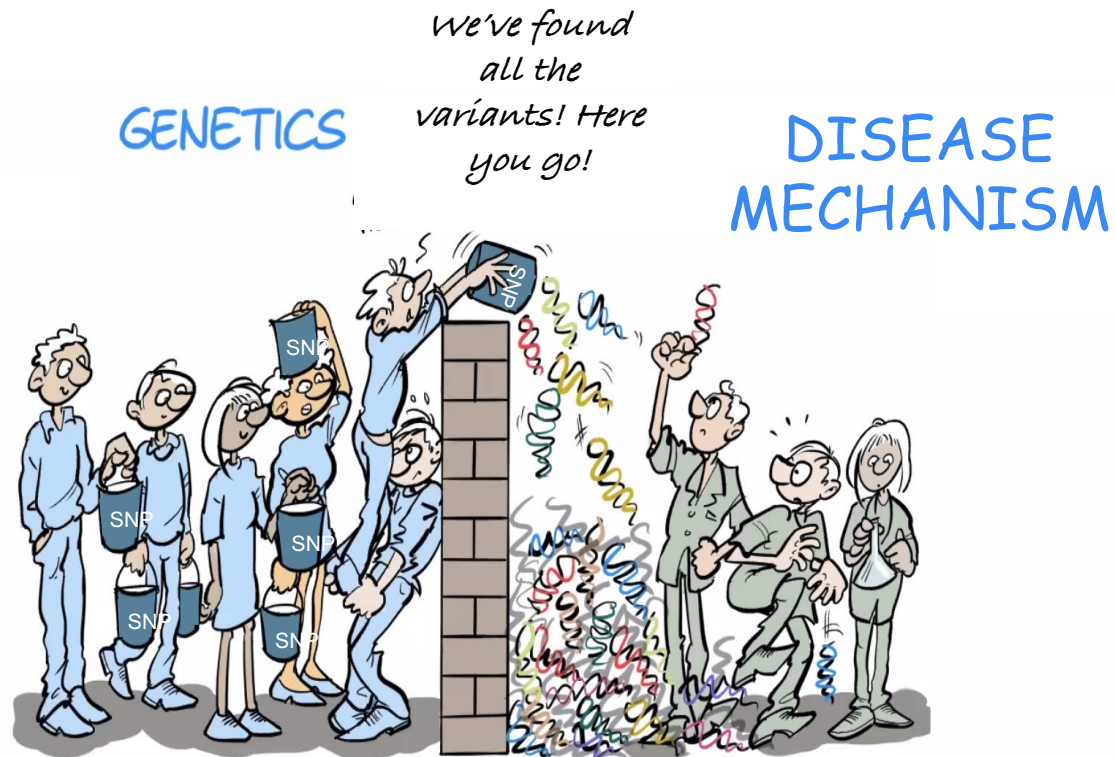
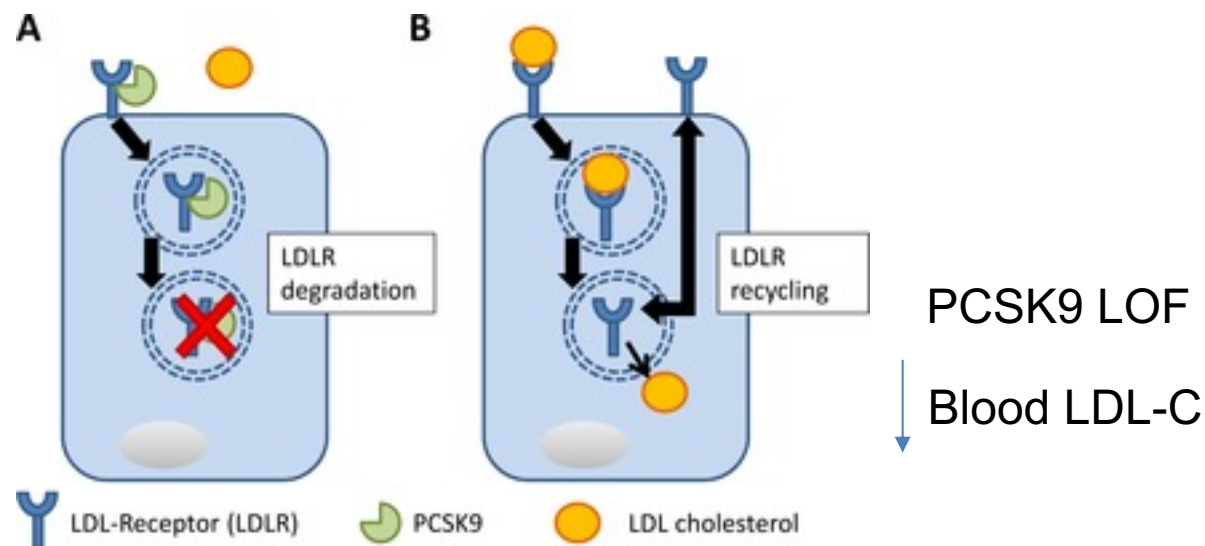


IMAGE CREDIT: [BRAINSCAPES](#)

# SNP to mechanism – protein-coding variants

- *PCSK9* c.426C>G (p.Tyr142X) identified through targeted sequencing in a population cohort (50% African American) (Cohen et al 2005 Nat Gen)
- Individuals with this variant had lower levels of plasma LDL-C
- Premature termination codon and truncation of the encoded protein or absence of the protein due to nonsense mediated decay.
- Predicting functional consequence of coding variants – SIFT, PolyPhen, CADD



PCSK9 inhibitors for lowering cholesterol



# Protein-coding regions make up only ~2% of the human genome



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).  
[\[Open chromatin regions\]](#)
- Histone mark enrichment (ChIP-seq)**  
Peaks (enriched genomic regions) of a variety of histone marks computed from ChIP-seq experiments.  
[\[Histone mark peaks\]](#)
- 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](#)
- Gene expression (RNA-seq)**  
Expression levels of genes and transcripts annotated by GENCODE, which can be visualized on SCREEN.  
[\[Expression levels\]](#) [SCREEN](#)
- 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](#)

CTCF DHS Profile

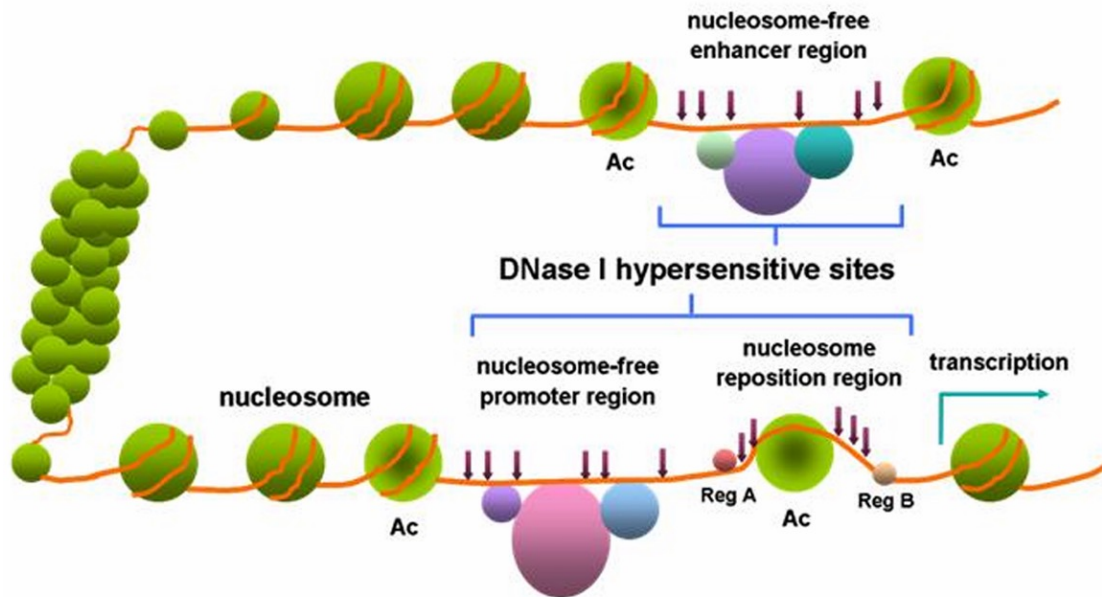
H3K27ac from mouse e11.5 hindbrain

CTCF Motif from Factorbook

HNF4A Gene Expression

HNF4A Transcript Expression

# DNase I hypersensitive sites

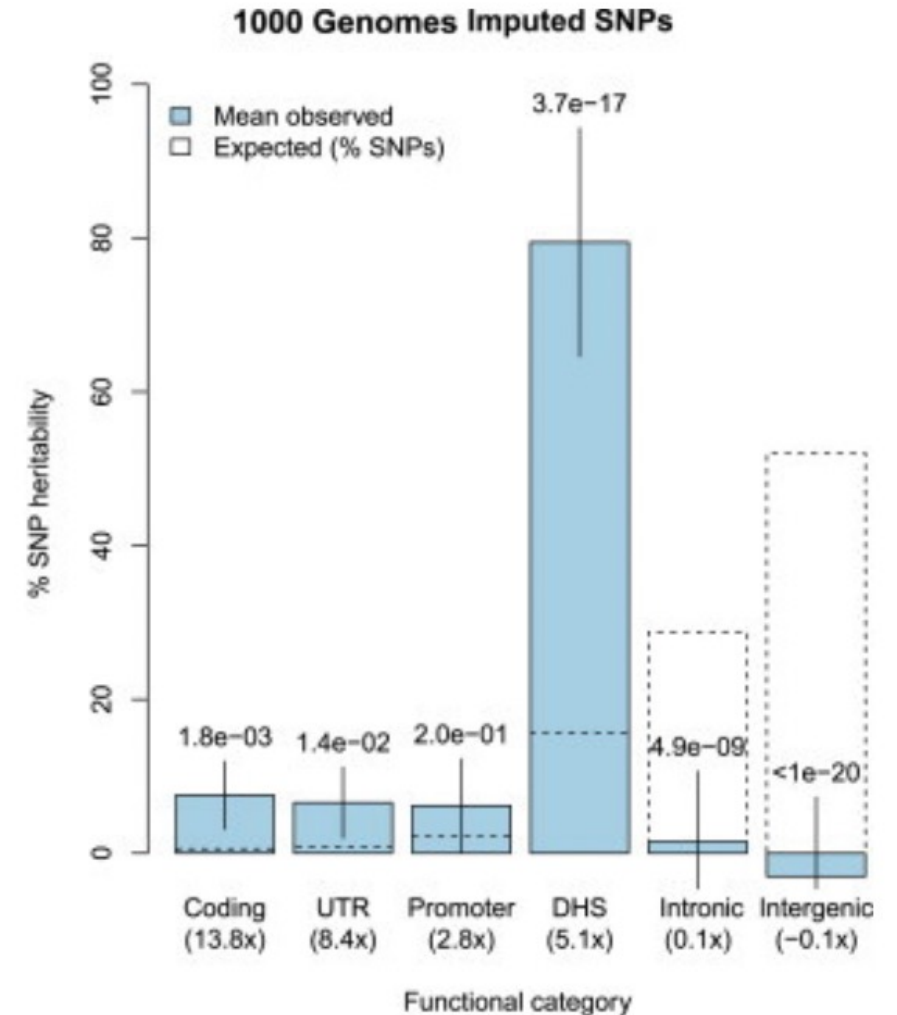


- Regions that are sensitive to cleavage by the DNase I enzyme
- Associated with open chromatin and therefore transcriptional activity.
- Map to regulatory elements (promoters, enhancers, insulators, silencers)

image source: Wikipedia

# Linking SNPs to genes – Regulatory variants

- SNP heritability  $h^2$  partitioned by functional category across 11 common diseases
- $h^2$  enrichment = GWAS significant SNP  $h^2$  compared to  $h^2$  for a random set of SNPs from the same functional category
- Coding variants: 13.8 fold enrichment but accounted for only 8% of SNP  $h^2$
- DHSs from 217 cell types spanned 16% of imputed SNPs, explained ~80% of SNP  $h^2$  (5-fold enrichment)



**Majority of GWAS significant SNPs in non-coding regions**

# Linking SNPs to genes

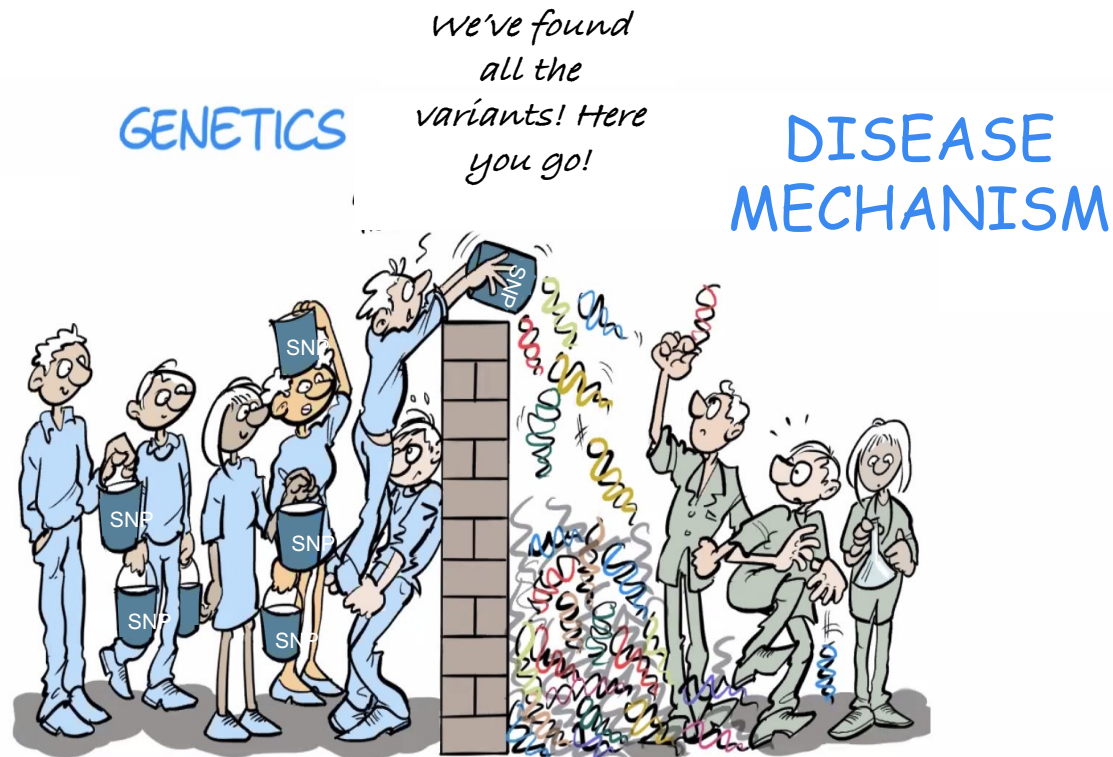


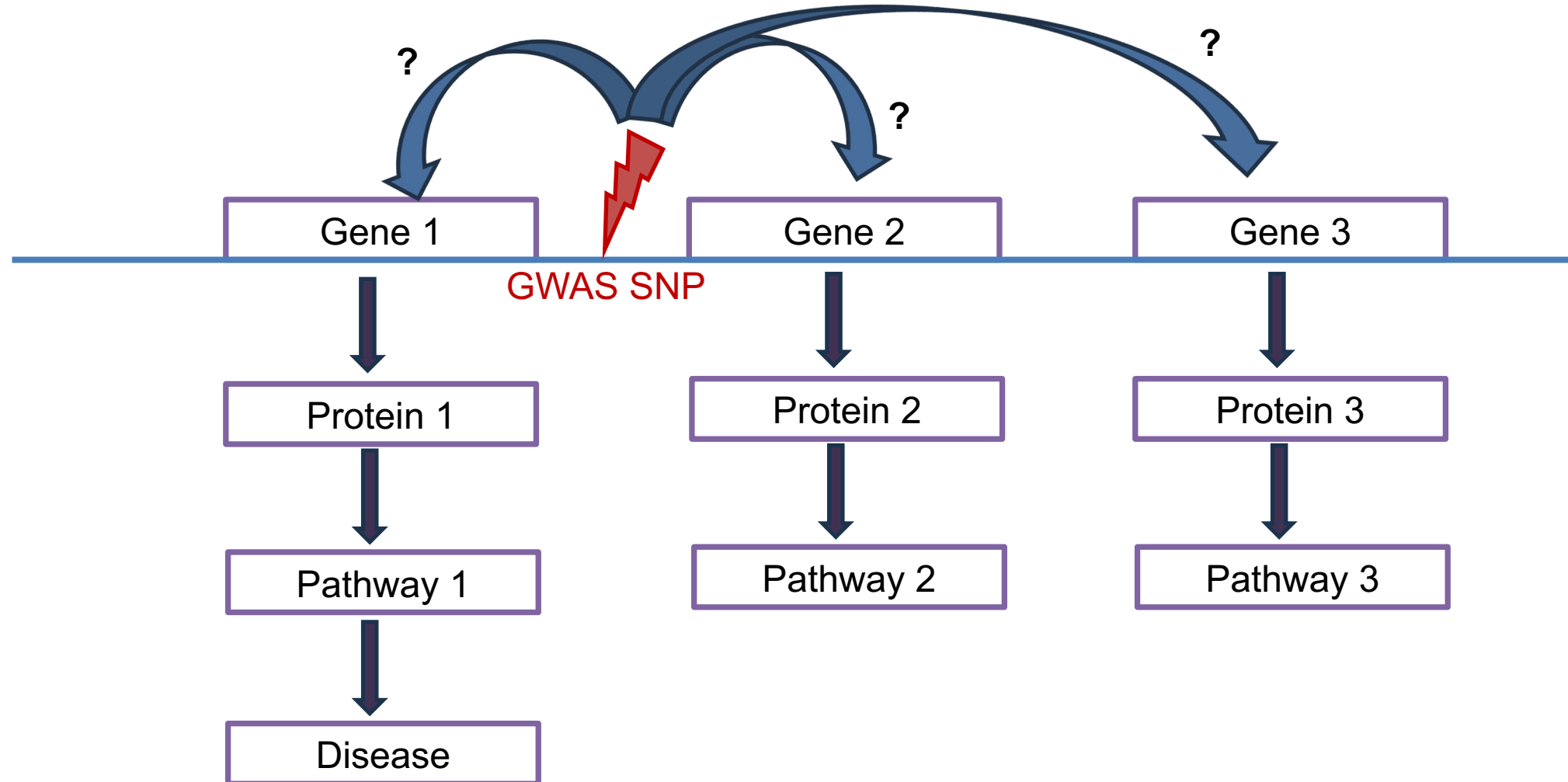
IMAGE CREDIT: [BRAINSCAPES](#)

## Challenges:

1. GWAS hits are in non-coding regions making it harder to determine the causal gene
2. Identifying causal variants made difficult by correlation (Linkage Disequilibrium) between neighbouring SNPs.
3. Unclear what cell types are relevant to disease

Overlaying cell regulatory region annotation can help narrow down putative causal variants – used in fine-mapping strategies

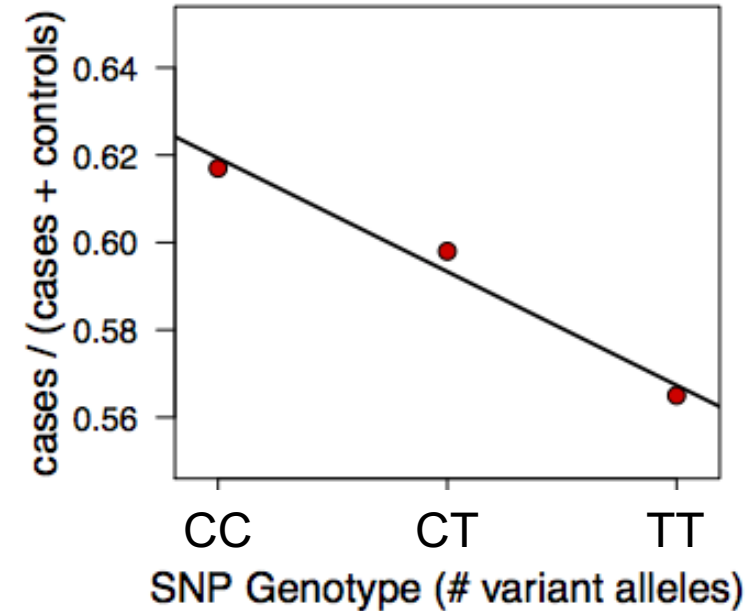
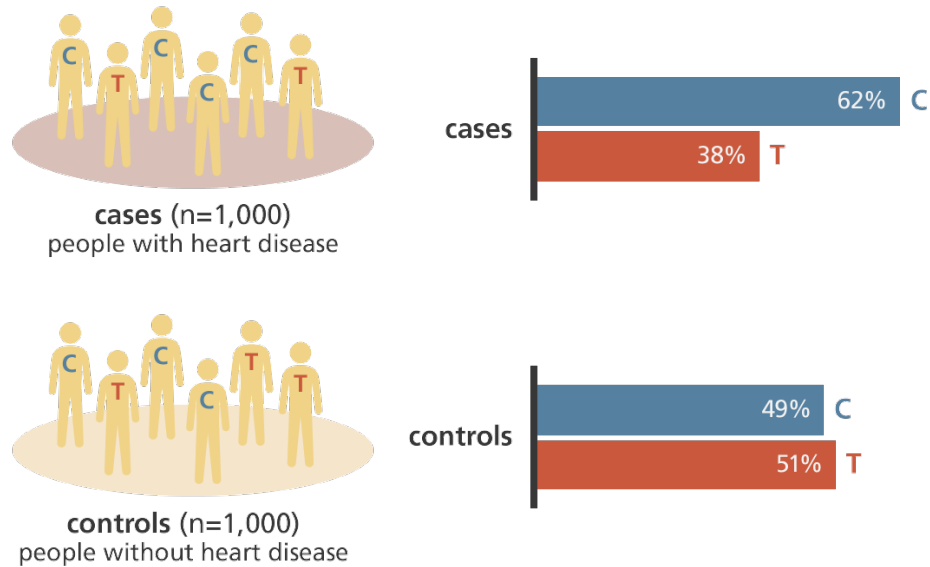
# Linking SNPs to genes – eQTL mapping



# What is an eQTL?

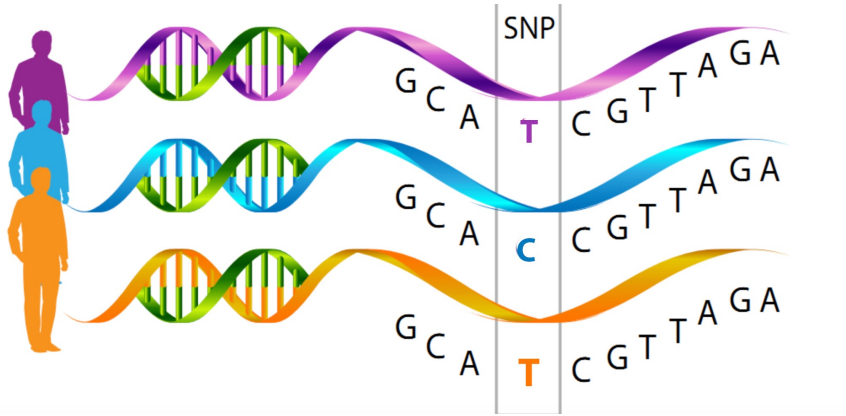


# Genetic association – binary trait



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

# Genetic association – quantitative trait



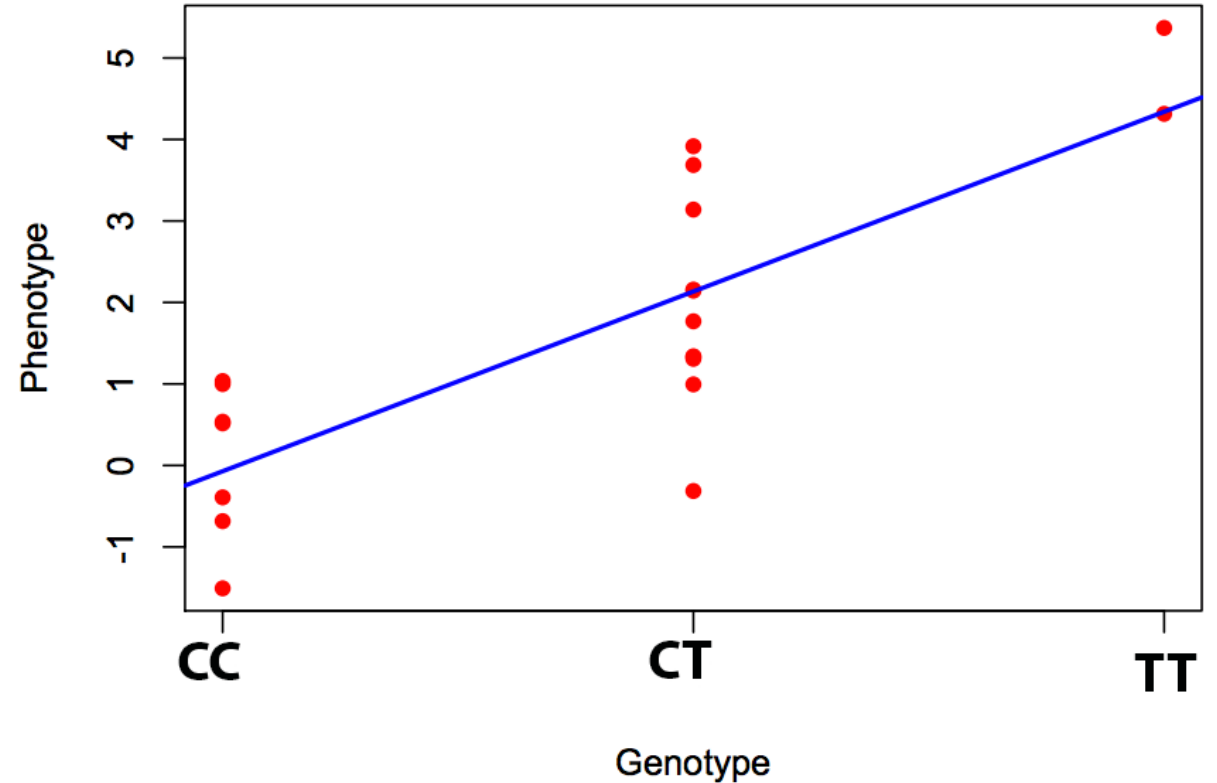
$$Y = b_0 + b_1X + e$$

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

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

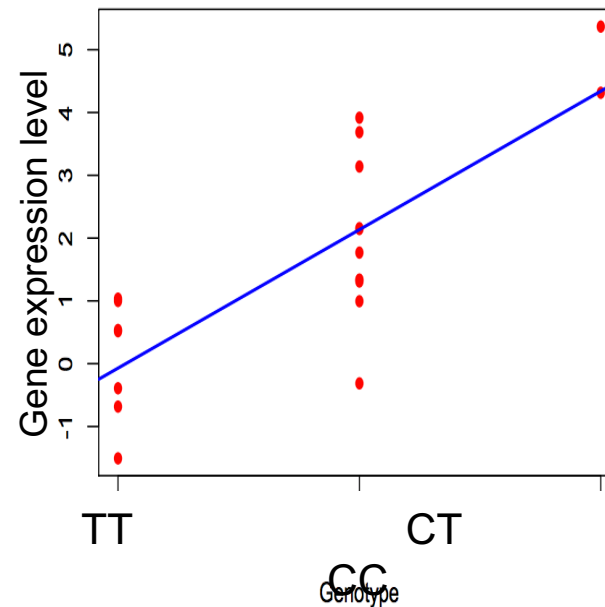
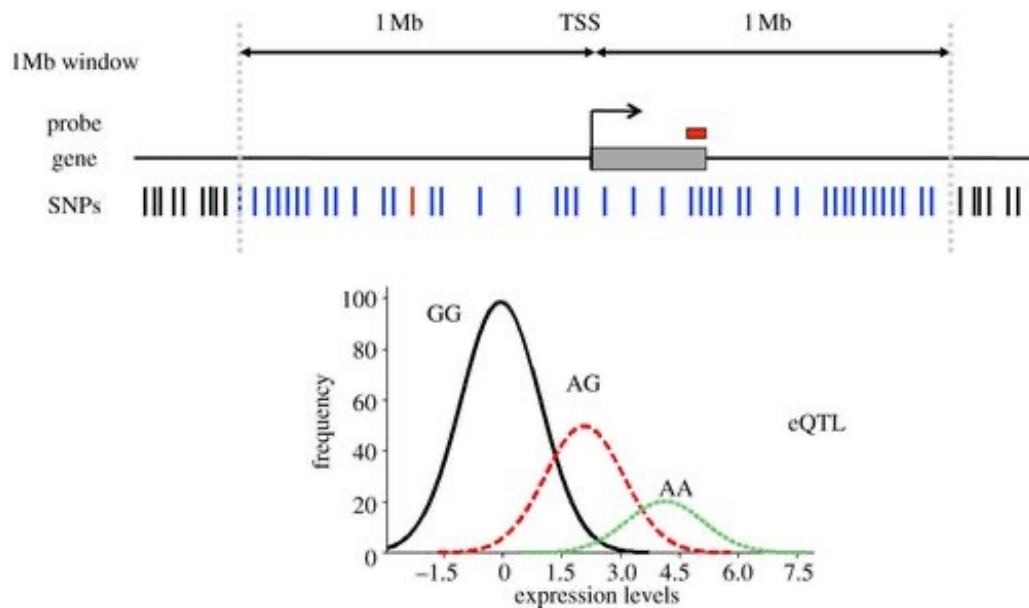
**b<sub>1</sub>** effect of each copy of the risk allele on the mean phenotype

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



# Expression quantitative trait locus (eQTL)

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



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)

# Performing eQTL mapping

# 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
- not parallelised

**matrix eQTL** (2012) [http://www.bios.unc.edu/research/genomic\\_software/Matrix\\_eQTL/s](http://www.bios.unc.edu/research/genomic_software/Matrix_eQTL/s)

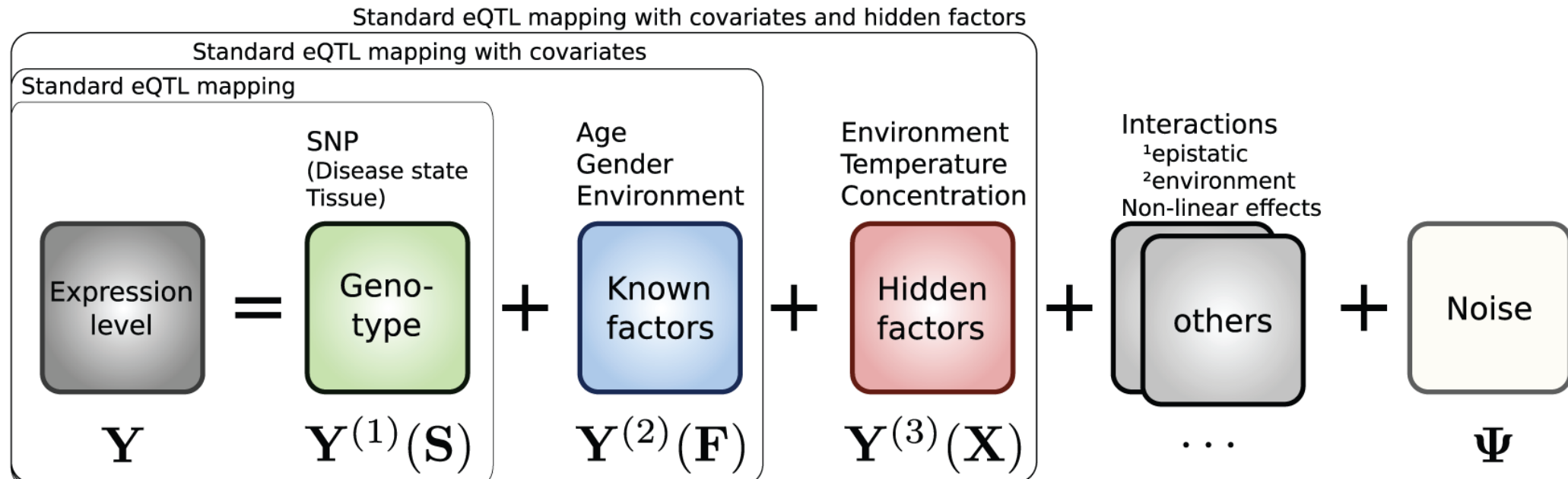
- Computationally efficient
- fast performance is achieved by special data pre-processing and using matrix operations
- no efficient built-in permutation scheme for multiple testing correction

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

- faster processing time (16× faster than matrix QTL).

# eQTL Mapping – things to consider

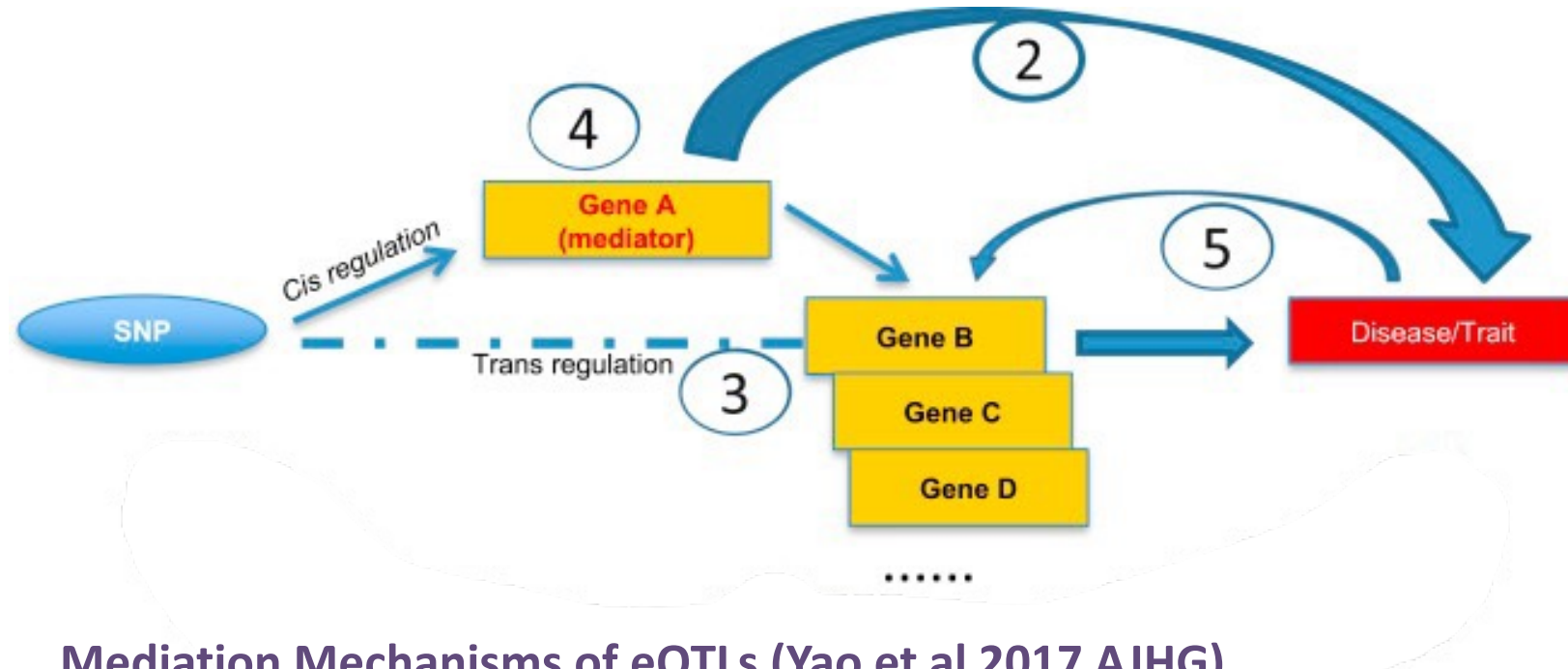
- Data normalisation - dependent on transcriptomics technology (gene arrays vs RNAseq)
- Removal of batch-effect e.g. sample preparation differences, tools such as ComBat, SVA, PEER
- Covariate adjustment
- P-value of association – high multiple-testing burden



# Cis vs trans QTLs

Cis-eQTL: SNP affects a gene located < 1Mb away

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



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

- (2) non-coding SNP affects expression of nearby gene (*cis*; < 1Mb away)
- (3) non-coding SNP affects remote (*trans*; >1Mb away) gene expression directly or by
- (4) *cis*-eGene mediation of the *trans*-eQTL-*trans*-eGene association; or
- (5) reverse causality (trait has feedback effect on gene expression).

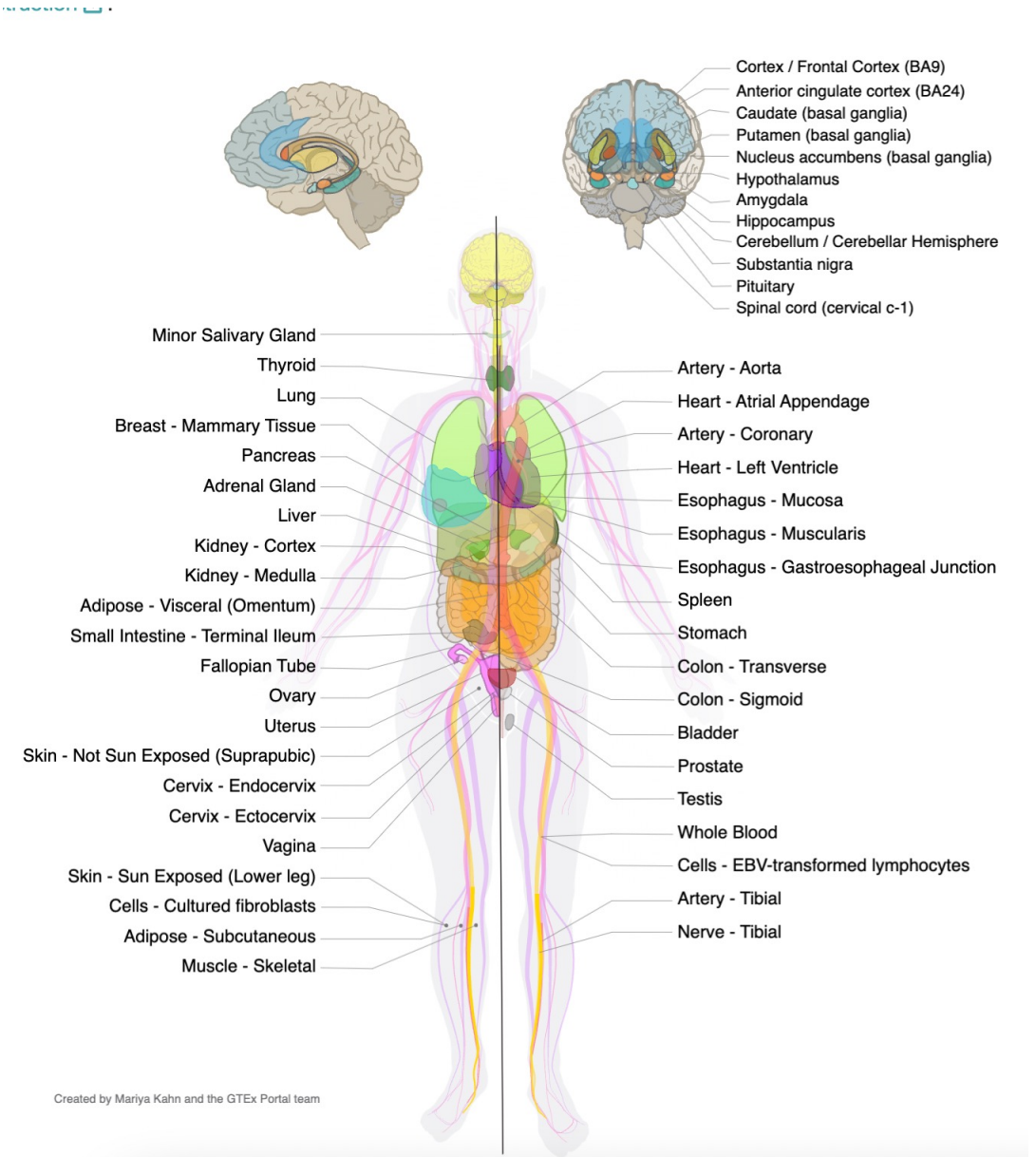
# eQTL data resources



# The GTEx Project

- Launched 2010
- Catalogue of genetic effects on gene expression across a large number of human tissues
- 838 donors and 17,382 samples from 50 tissues
- Gene expression (RNA-seq) and genotype data (WGS data)

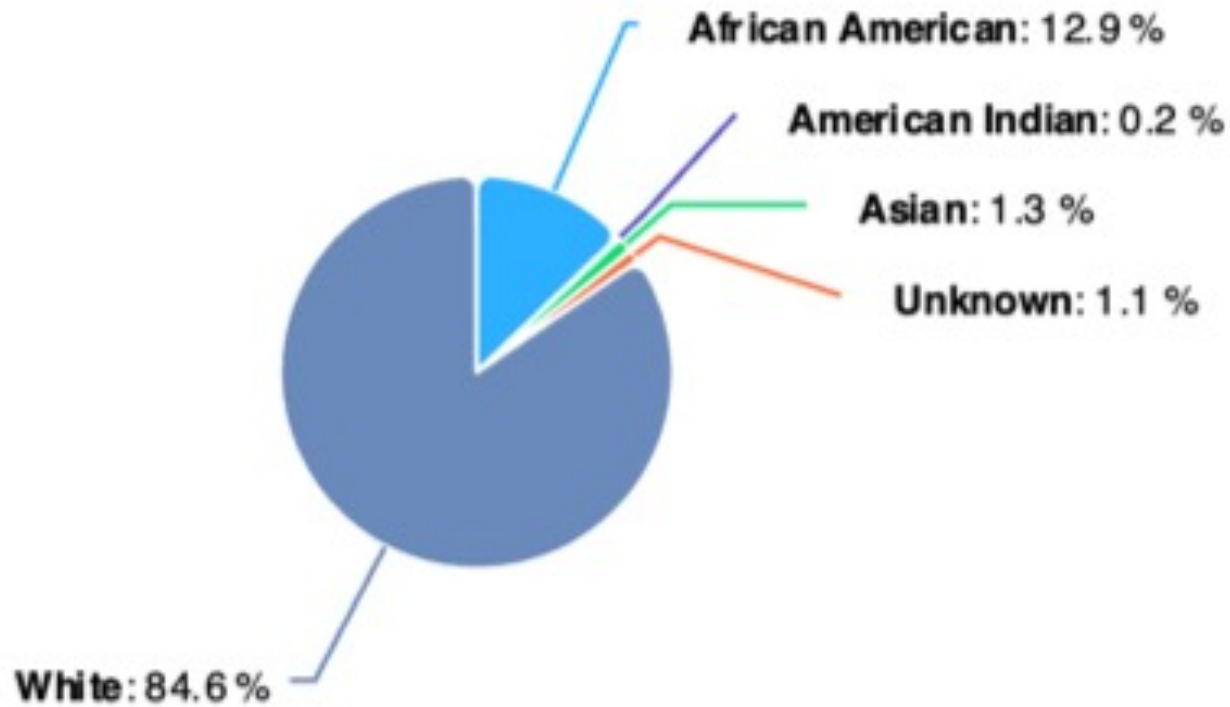
**The GTEx Consortium atlas of genetic regulatory effects across human tissues**  
**Science 2020 [DOI: 10.1126/science.aaz1776](https://doi.org/10.1126/science.aaz1776)**



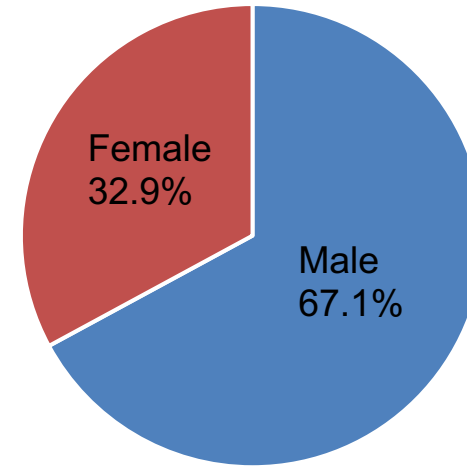
Created by Mariya Kahn and the GTEx Portal team

# The GTEx Project

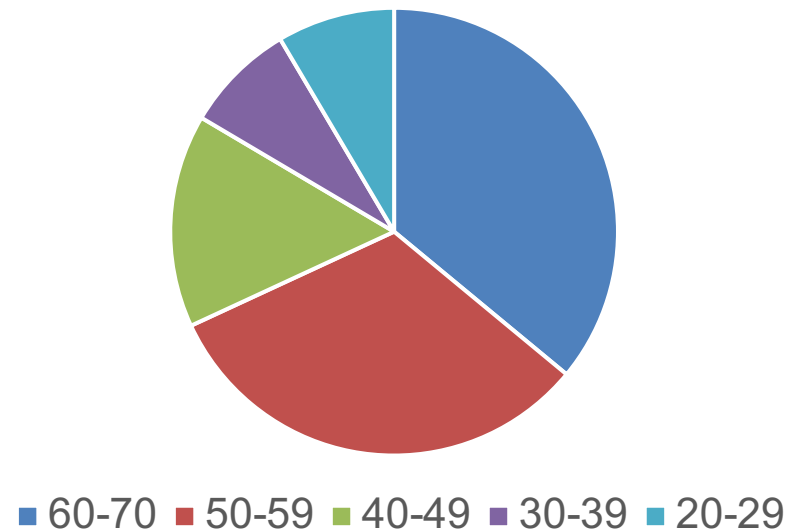
## Race



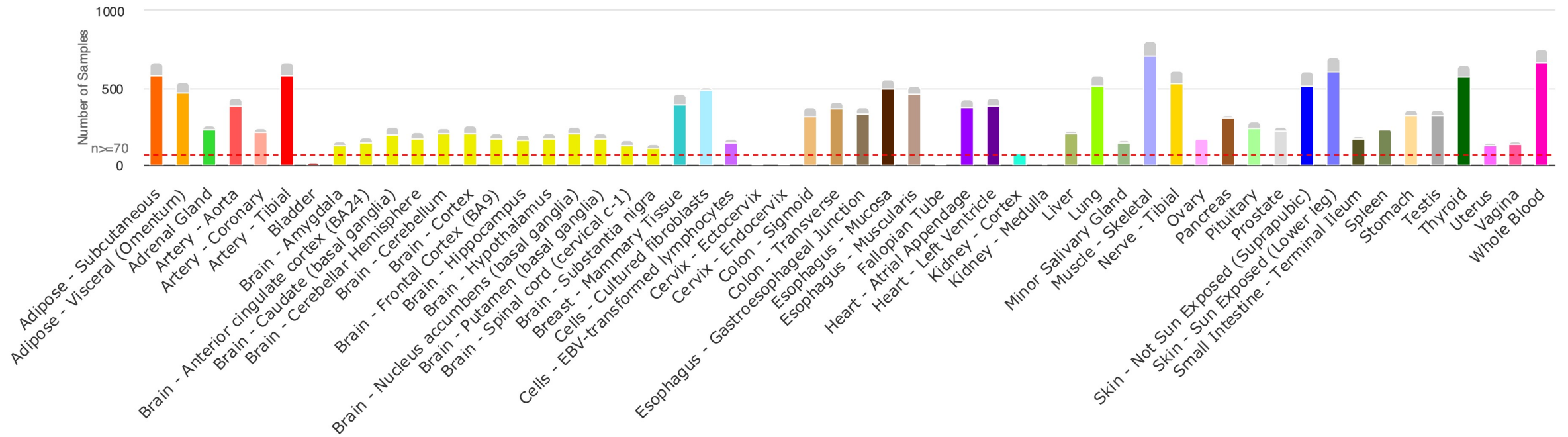
## Sex



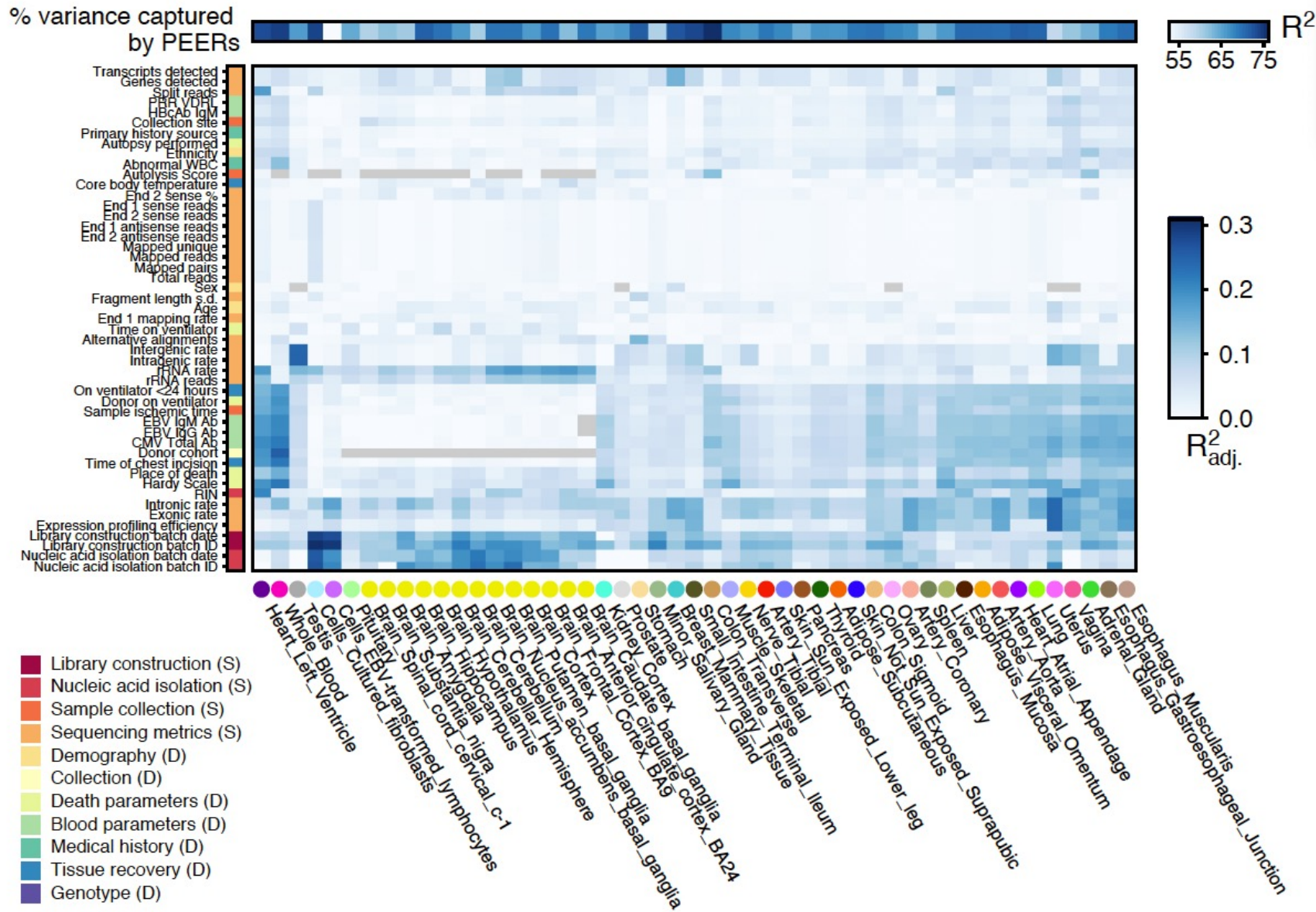
## Age



# Tissue sample size



# PEER analysis – accounting for batch effects

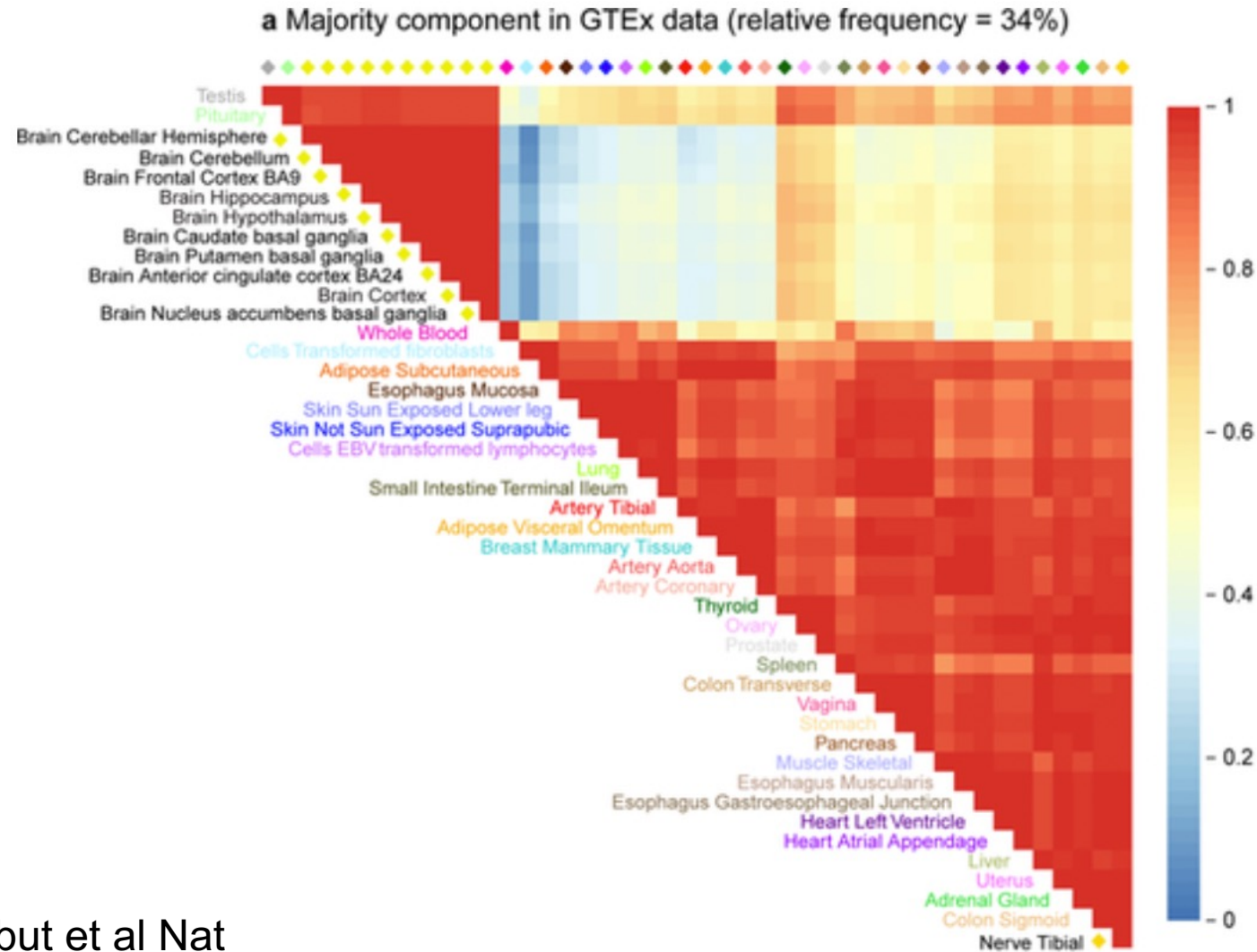


Covariates most consistently associated with PEER factors include factors related to donor death, ischemic time, sequencing quality control metrics, and nucleic acid isolation and library construction batches.

# 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

# Co-regulation 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

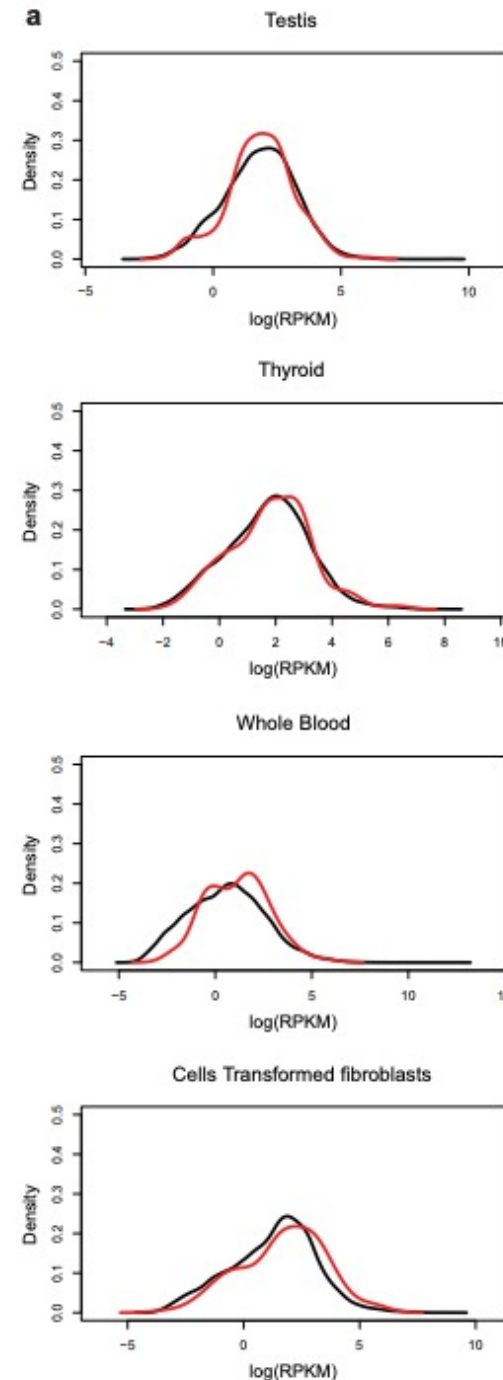
# Tissue-specific eQTLs

Urbut et al 2019 Nature Genetics

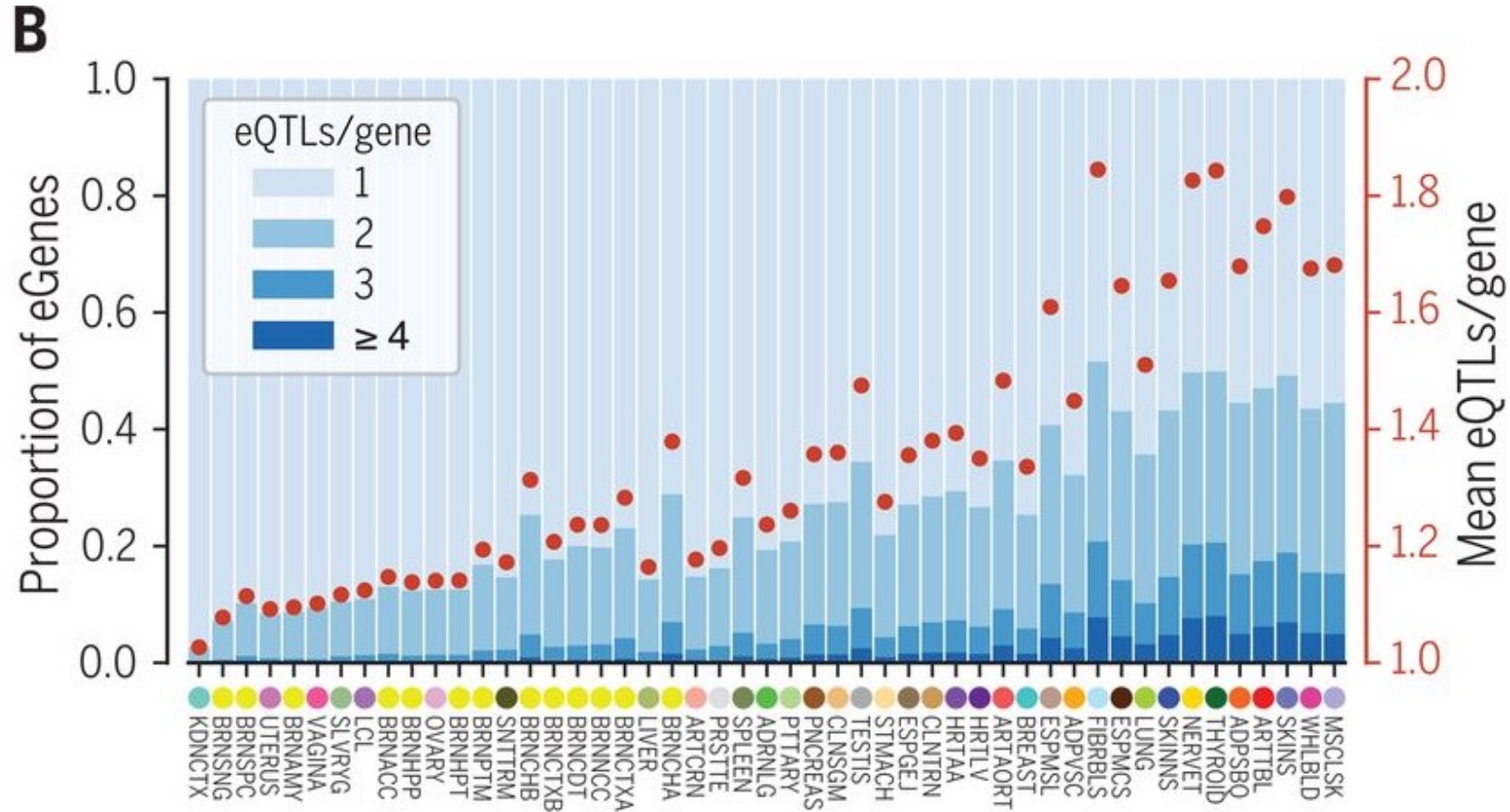
Subset of eQTLs that have a much stronger effect in one tissue than in any other (i.e. “tissue-specific”)

most tissue-specific eQTLs identified here do not solely reflect tissue-specific expression.

May reflect power due to different sample size



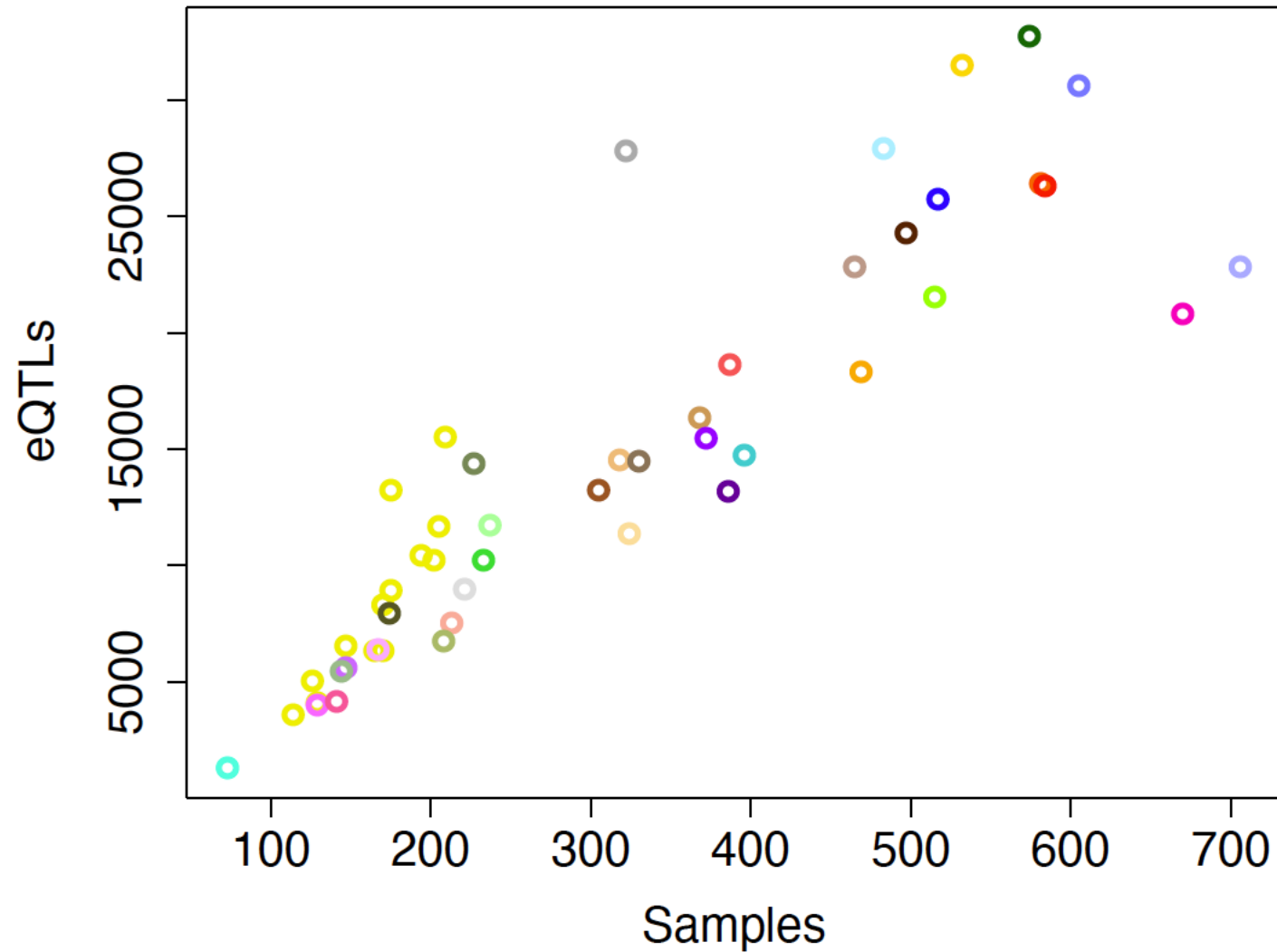
# The GTEx Project



up to 50% of eGenes having more than one independent cis-eQTL in the tissues with the largest sample sizes



# Power to detect an eQTL



# eQTL catalogue

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

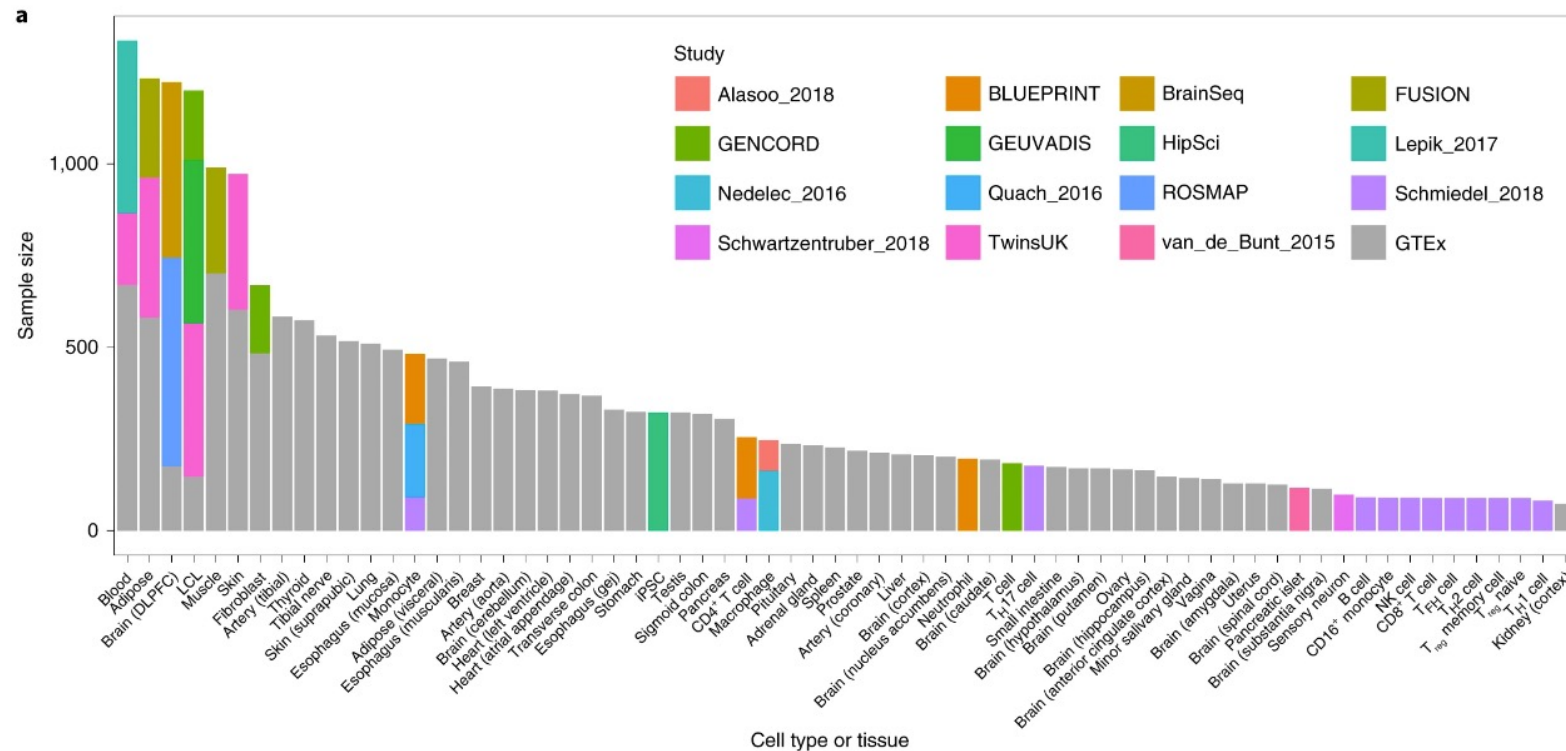
Provides uniformly processed cis-eQTLs and sQTLs from all available public studies on human.

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

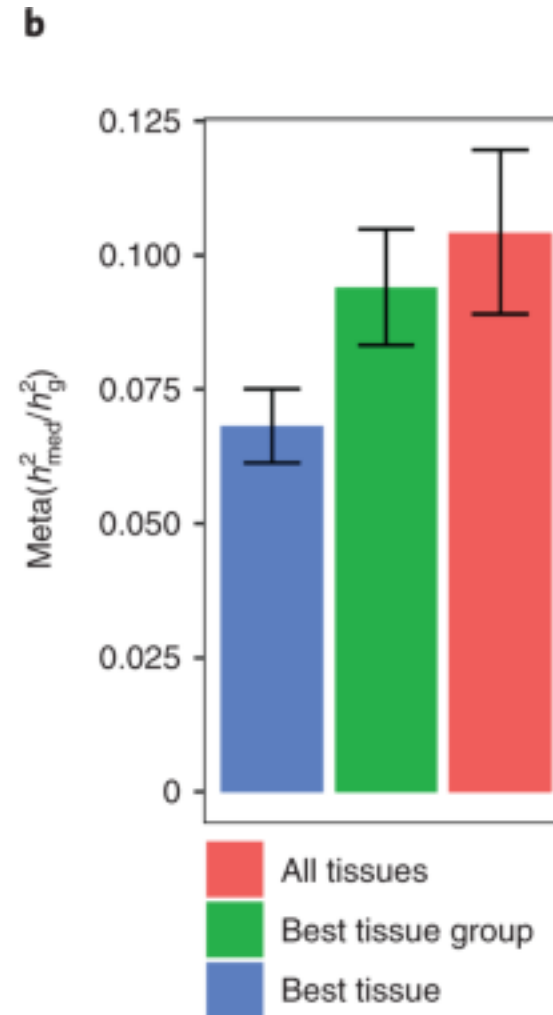
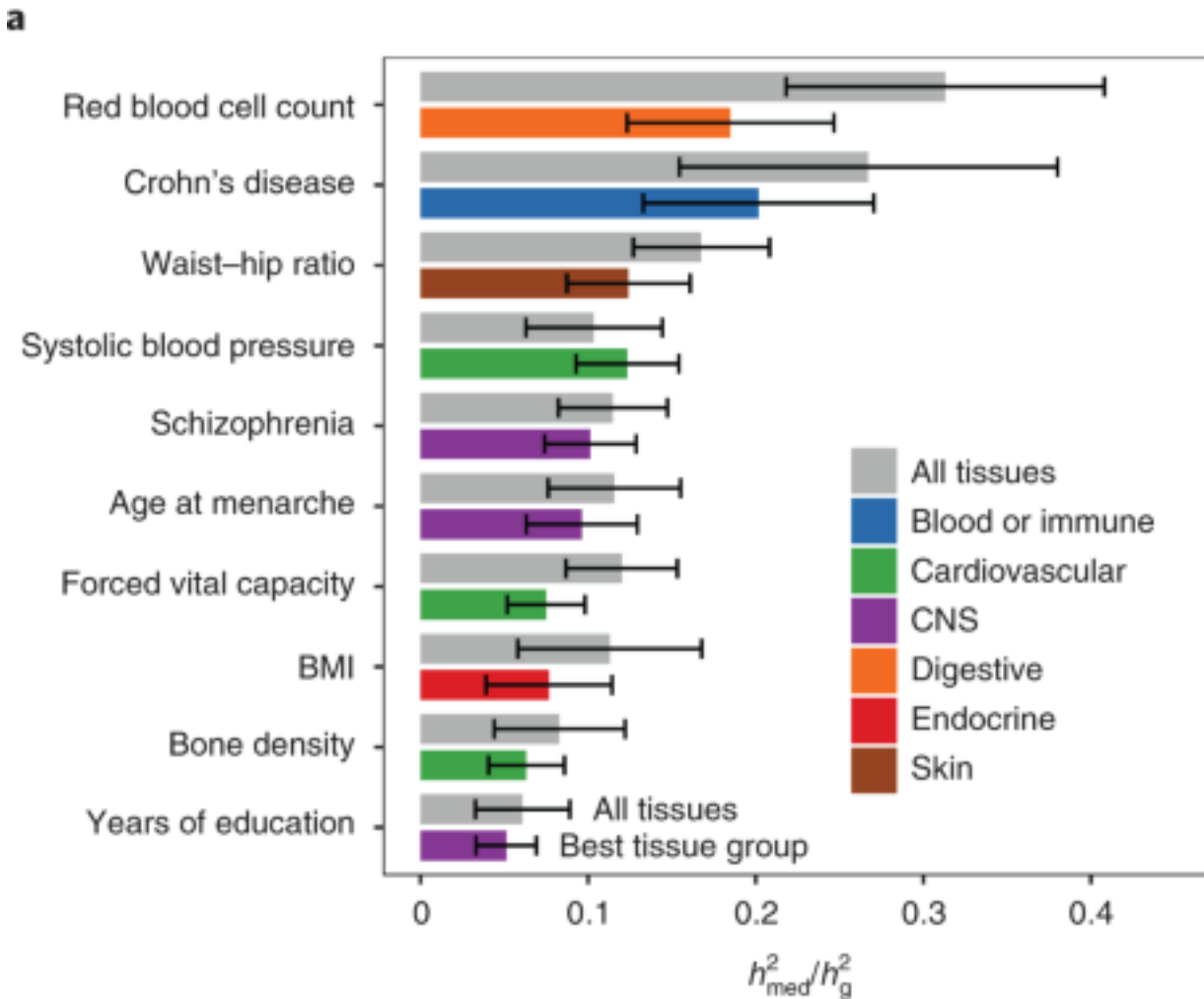
## A compendium of uniformly processed human gene expression and splicing quantitative trait loci

[Nurlan Kerimov](#), [James D. Hayhurst](#), [Kateryna Peikova](#), [Jonathan R. Manning](#), [Peter Walter](#), [Liis Kolberg](#), [Marija Samoviča](#), [Manoj Pandian Sakhivel](#), [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](#)



# Significant but only modest proportion of complex trait heritability mediated by the cis-eQTLs

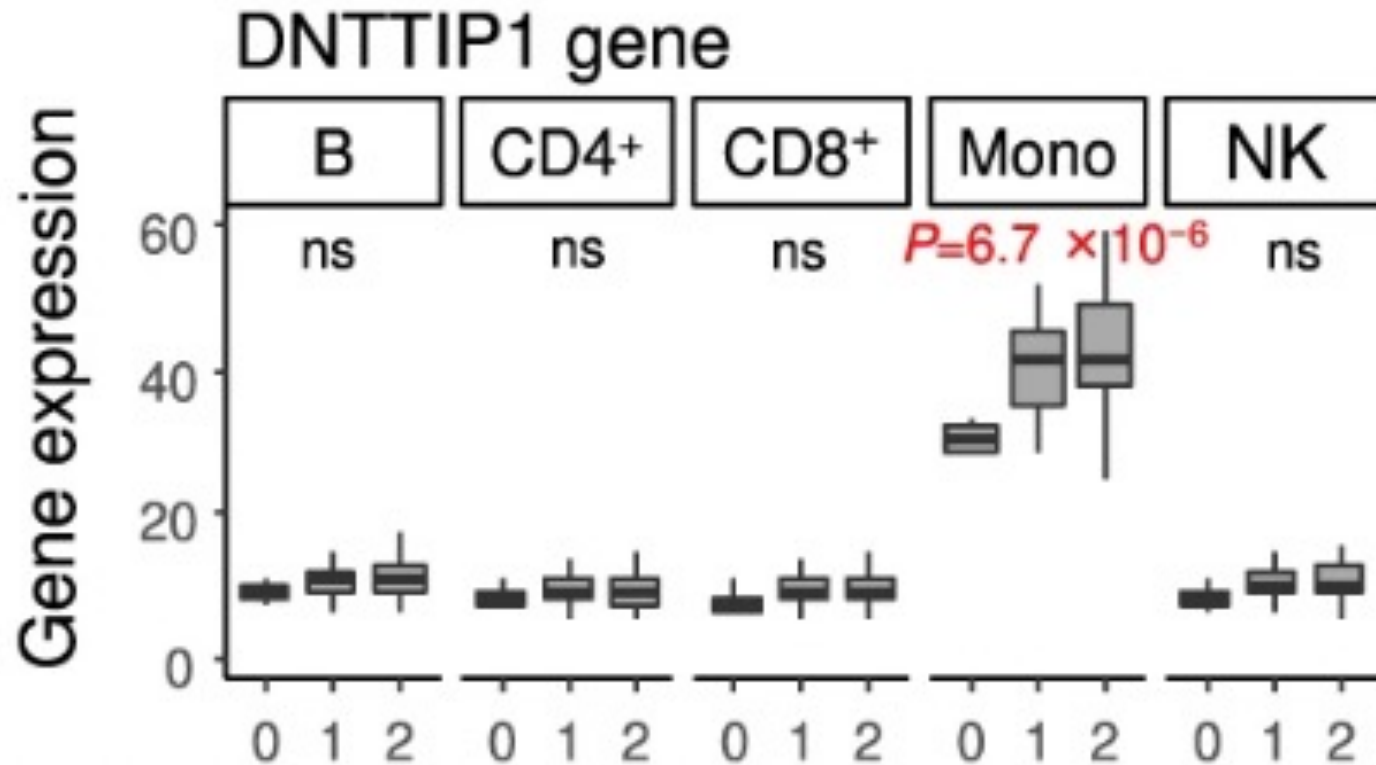


# Dynamic eQTL?

# Dynamic QTLs (cell or context-specific eQTLs)

- Most eQTL studies measure gene expression at a **single timepoint** and mostly in adult tissues
- eQTLs can be time-dependent or environment-specific
  - Cell-type-specific expression (bulk tissue studies may mask cell-specific effects)
  - Response to treatment
  - Developmental stage

# Cell-type-specific QTLs



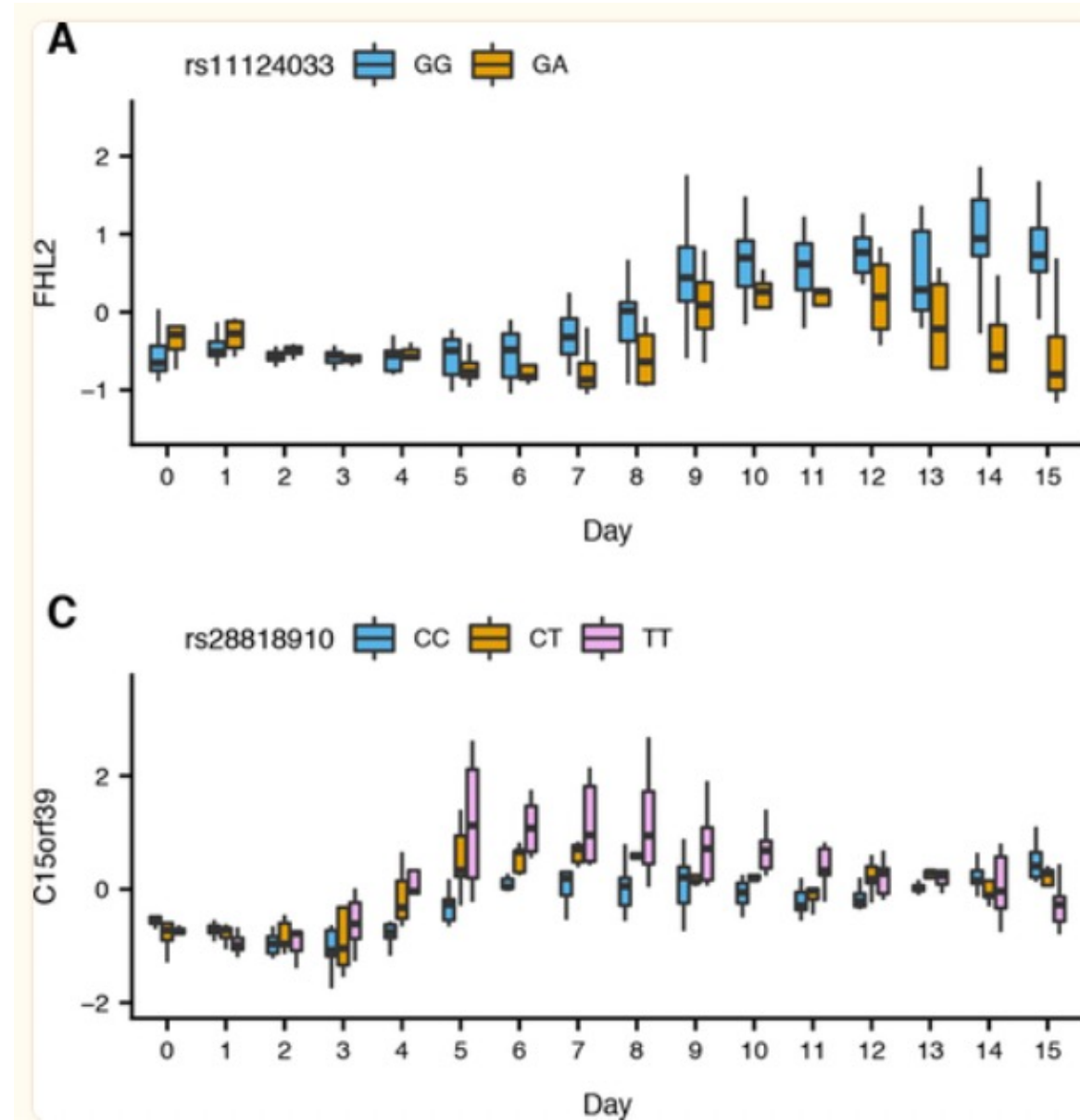
No significant  
eQTL for  
DNTTIP1 in  
GTEx v8 data

# Dynamic QTLs

Dynamic genetic regulation of gene expression during cellular differentiation

- iPSC differentiation into cardiomyocytes.
- eQTL analysis at 16 time points

Strober et al Science 2019

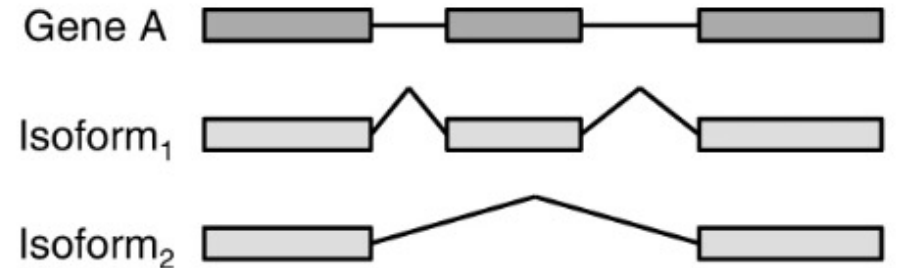


# Splice QTLs



# 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



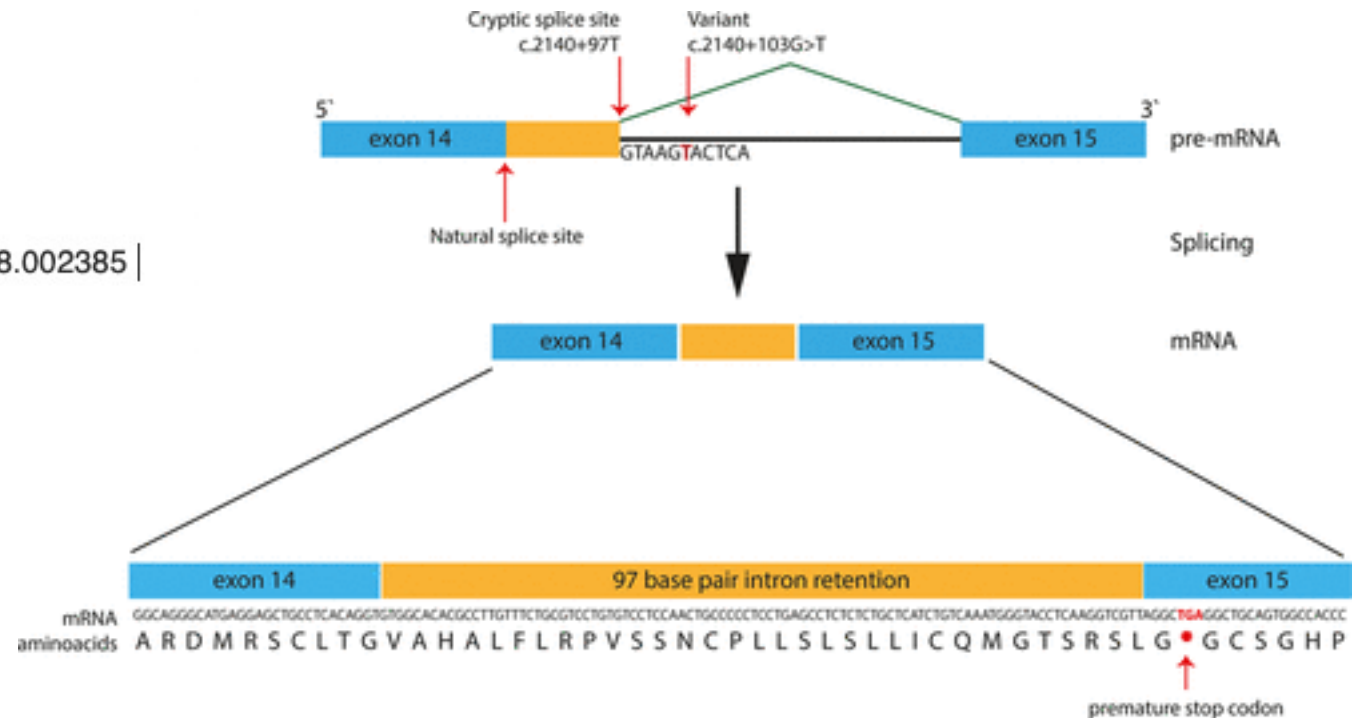
# Splice variants

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

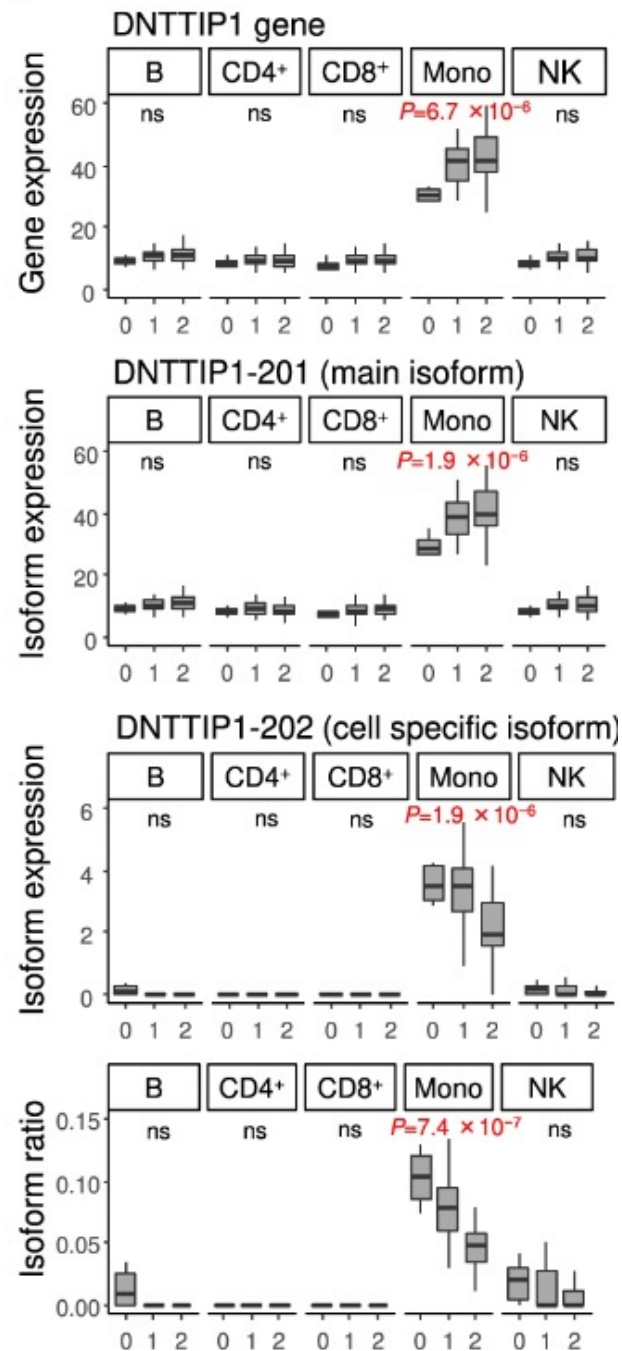
### Time to Widen the Scope?

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

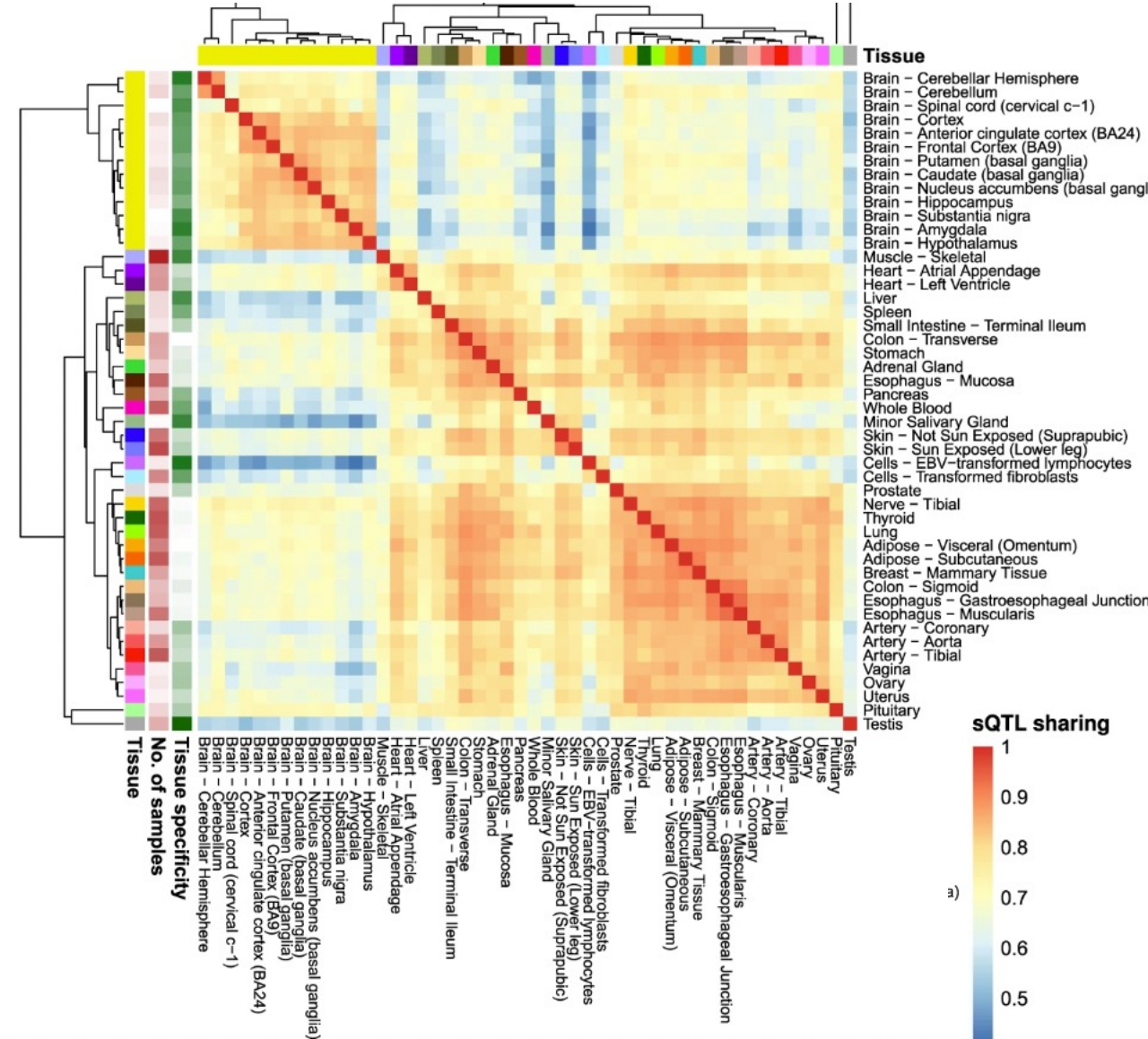
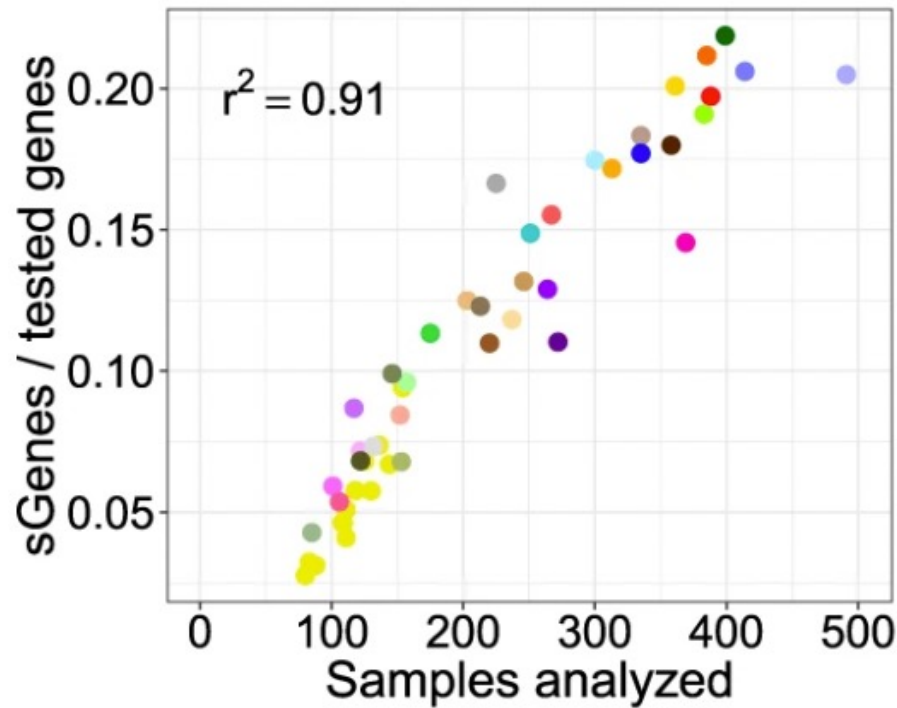
Originally published 11 Dec 2018 | <https://doi.org/10.1161/CIRCGEN.118.002385> | Circulation: Genomic and Precision Medicine. 2018;11:e002385



# Cell-type-specific sQTLs

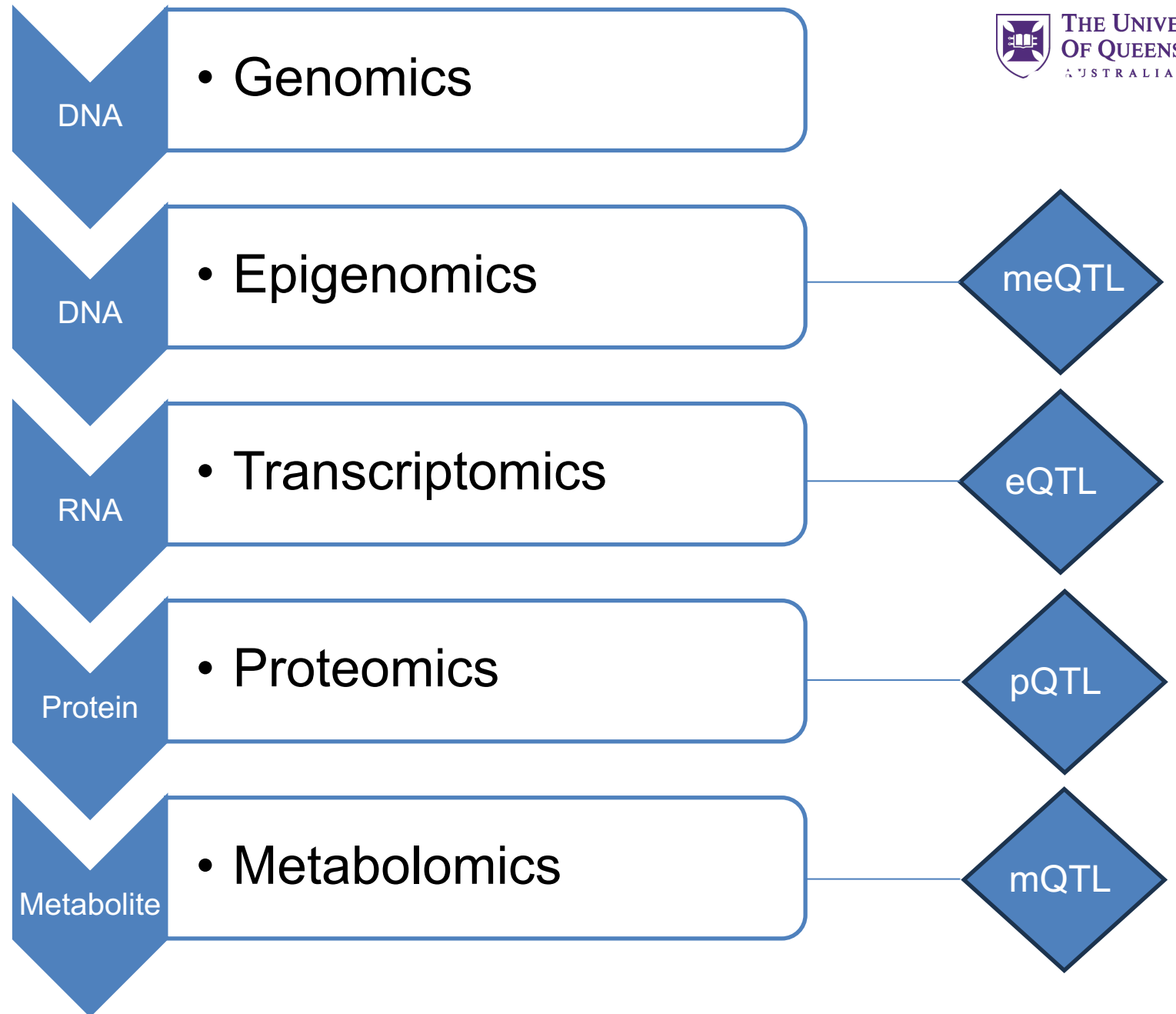


# sQTL sharing across tissues



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



# Association does not mean causality

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

