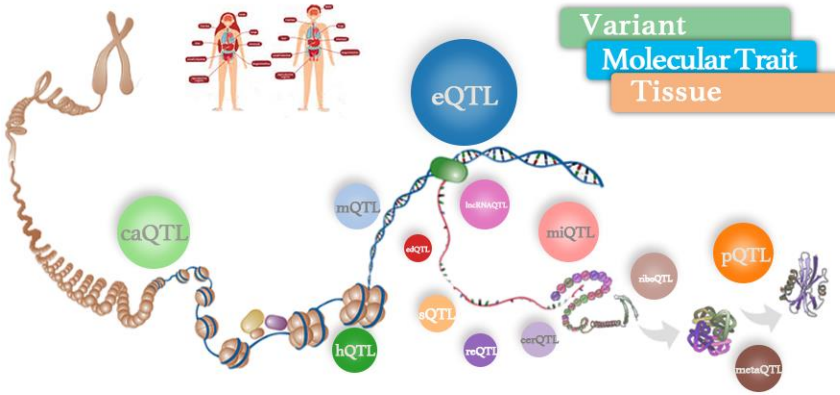


Molecular QTL mapping in humans

Volos Summer School

Thursday June 15 2023

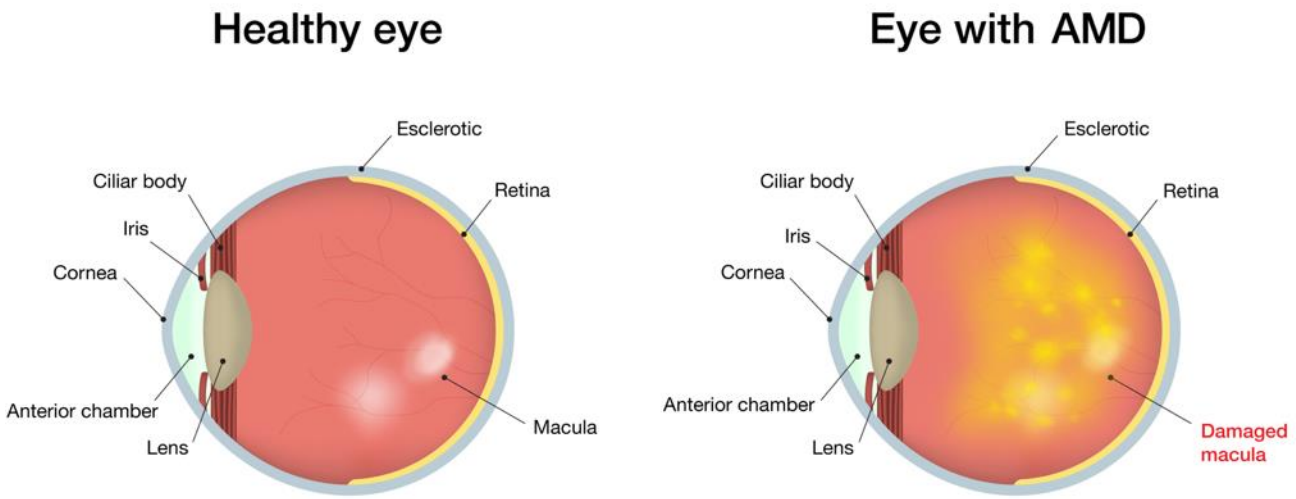
Antigone Dimas & Ozvan Bocher

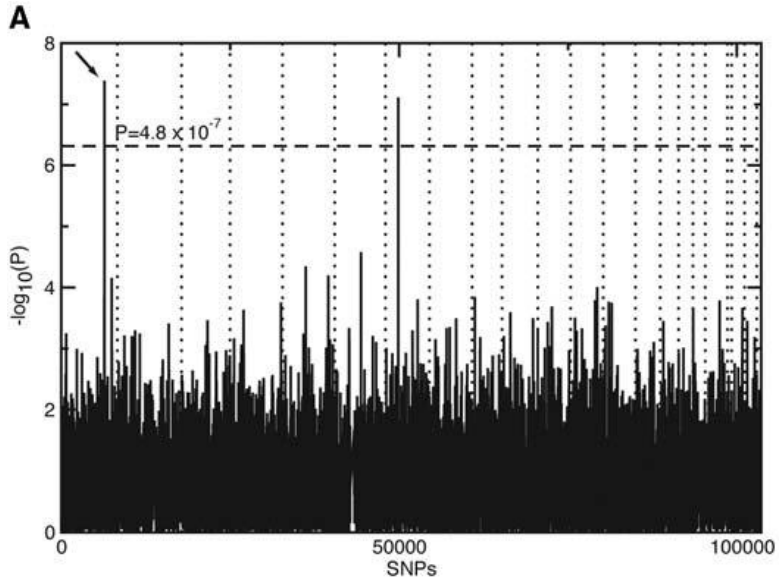
"ALEXANDER FLEMING"
Biomedical Sciences Research Center

HELMHOLTZ
MUNICH

How do genetic variants affect cellular processes?

- Few GWAS variants shape phenotypes in a straightforward way (e.g. macular degeneration).



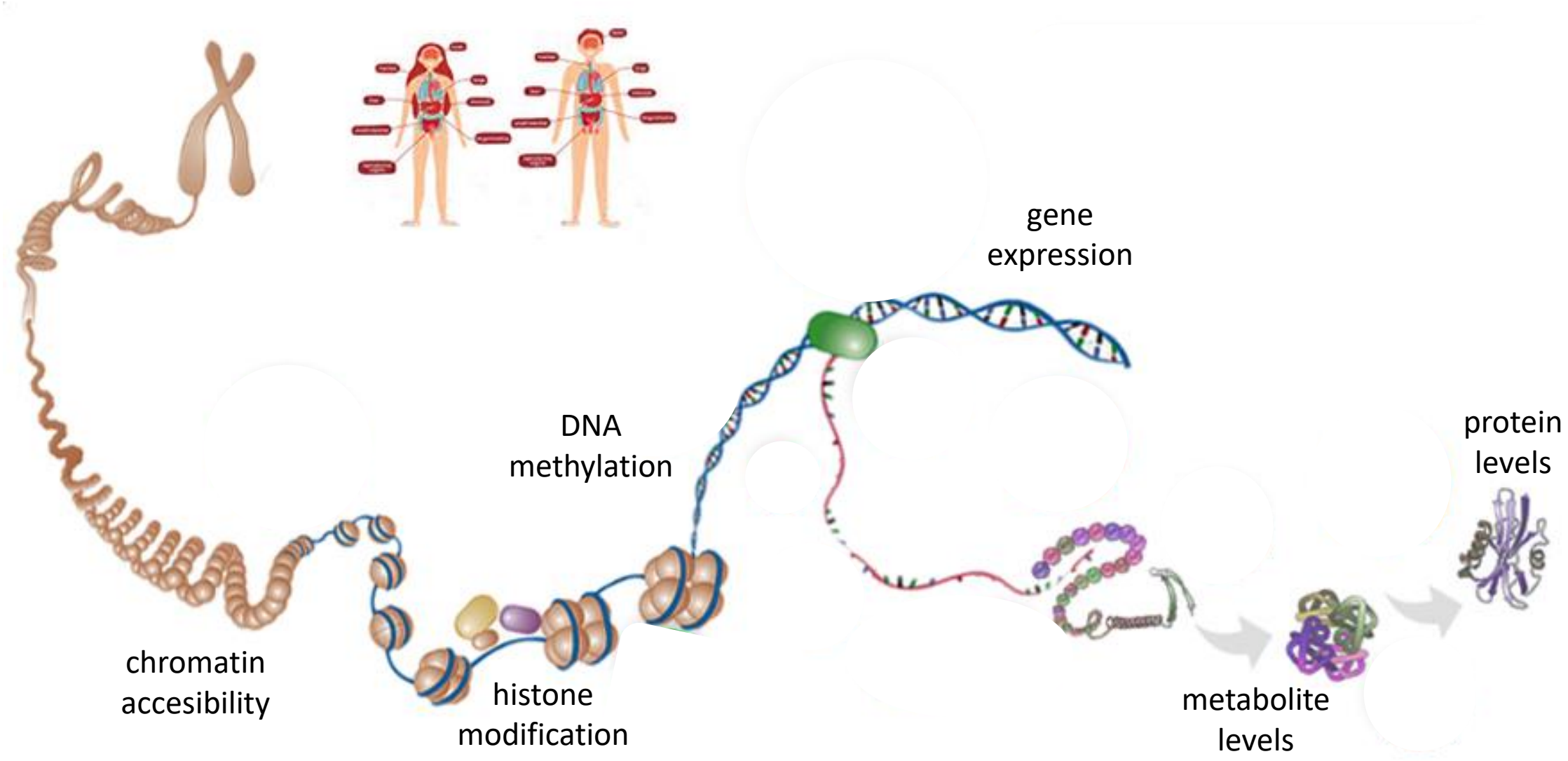


Major cause of blindness in elderly. destruction of retina's macula with central field visual loss
One of first GWAS: 96 cases, 50 controls, 116,204 SNPs
Detected tyrosine → histidine in complement factor H gene, in region binding heparin and CRP
New isoform causes aberrant inflammatory processes including inappropriate complement activation → macular degeneration

- Much of disease-associated (GWAS-detected) variation is in non-coding regions
- To understand functional effects of such genetic variants → molecular quantitative trait locus (molQTL) mapping

Klein *Science* 2005

Multiple molecular traits in cells can be quantified

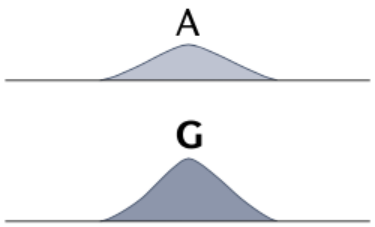


QTLbase

Quantifying molecular traits – can be measured at scale

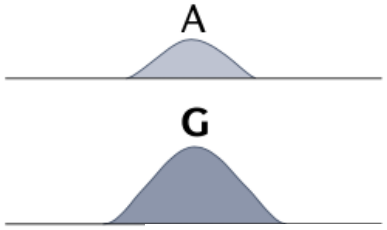
Chromatin accessibility QTL (caQTL or chQTL)

Chromatin accessibility measured by ATAC-seq, DNase I sensitivity, etc.



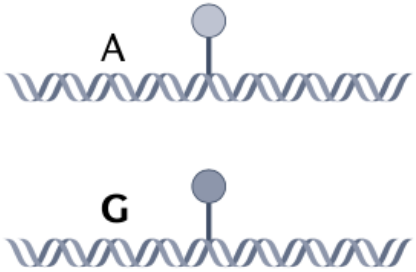
Histone modification QTL (hQTL or cQTL)

Histone mark ChIP–seq peak height



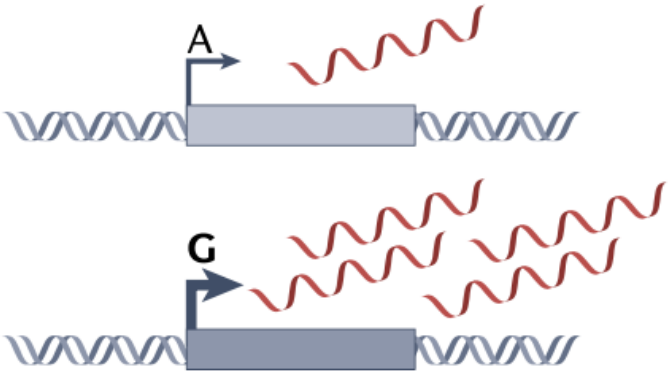
Methylation QTL (meQTL)

Methylation ratio of a CpG site



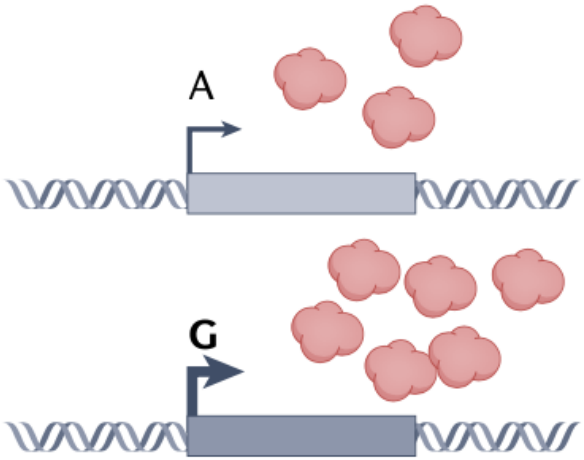
Expression QTL (eQTL)

RNA expression level of a gene or a transcript



Protein QTL (pQTL)/metabolite QTL (mQTL)

Protein expression level of a gene



Aguet et al Nat Rev Methods Primers 2023

Genetic associations for molecular traits can be mapped in a similar way to GWAS

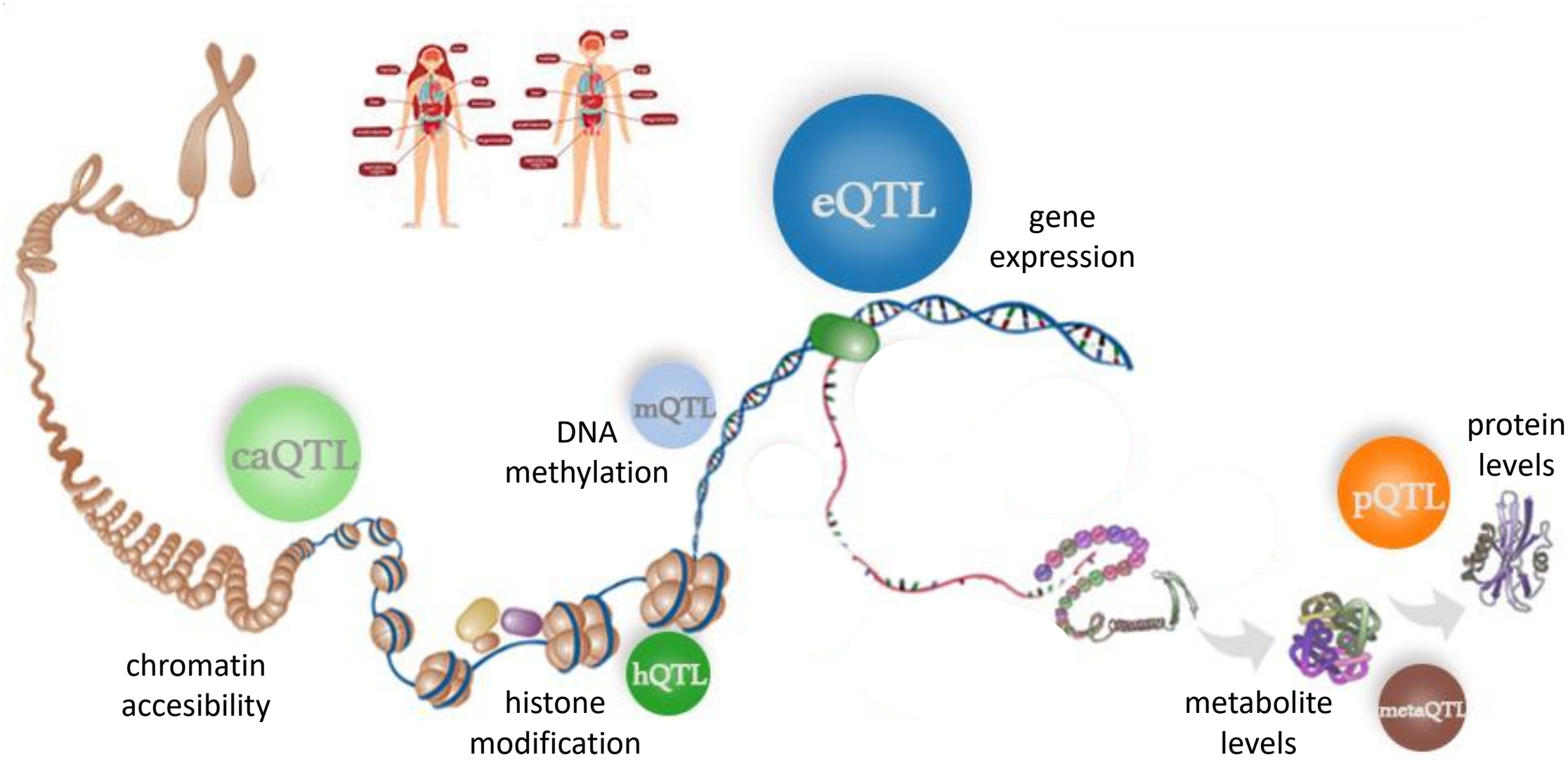
- Molecular traits have substantial genetic component → map genetic associations in similar way to GWAS

- Statistically, GWAS for a quantitative trait (e.g. height) and QTL mapping are nearly identical approaches
  → Linear regression to associate genetic variation with a quantitative phenotype in a population sample

- Typically:
  - GWAS used for *non-molecular* traits (e.g. height)
  - QTL refers to *molecular* quantitative trait locus

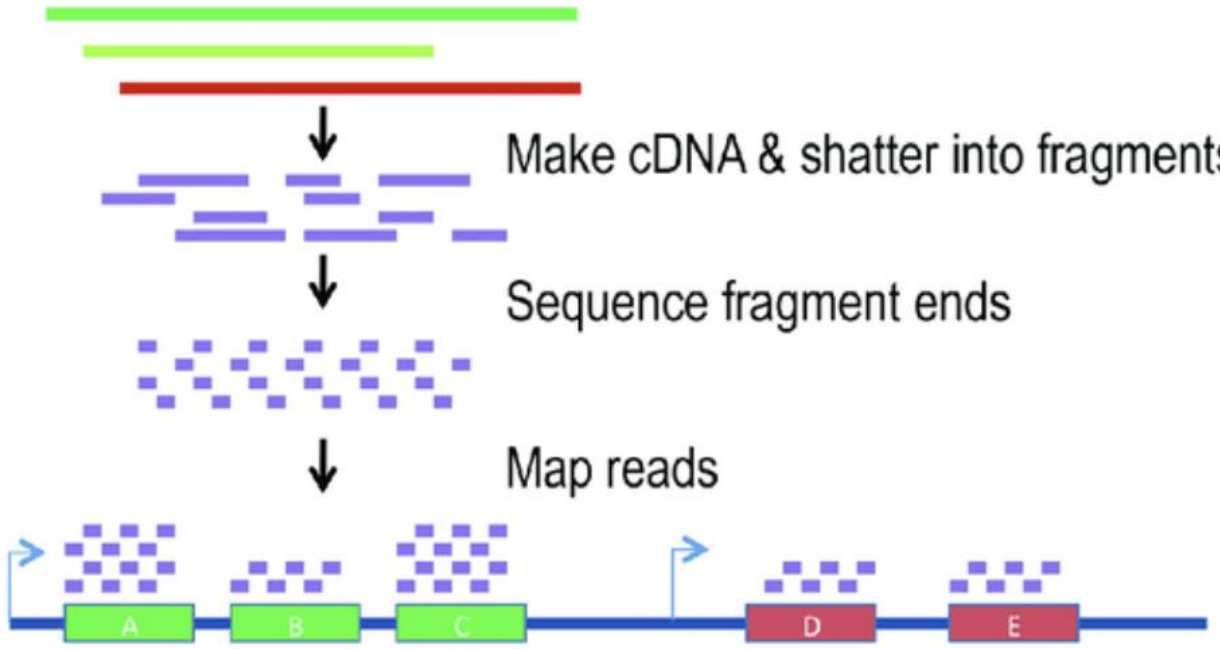
- molQTL umbrella term for loci with a genetic association for a quantitative level of a molecular trait including:
  - eQTLs – gene expression
  - mQTLs – DNA methylation
  - hQTL – histone modification
  - caQTLs – chromatin accessibility
  - pQTLs – protein levels
  - metaQTLs – metabolite levels

molQTL umbrella term for loci with a genetic association for a quantitative level of a molecular trait



QTLbase

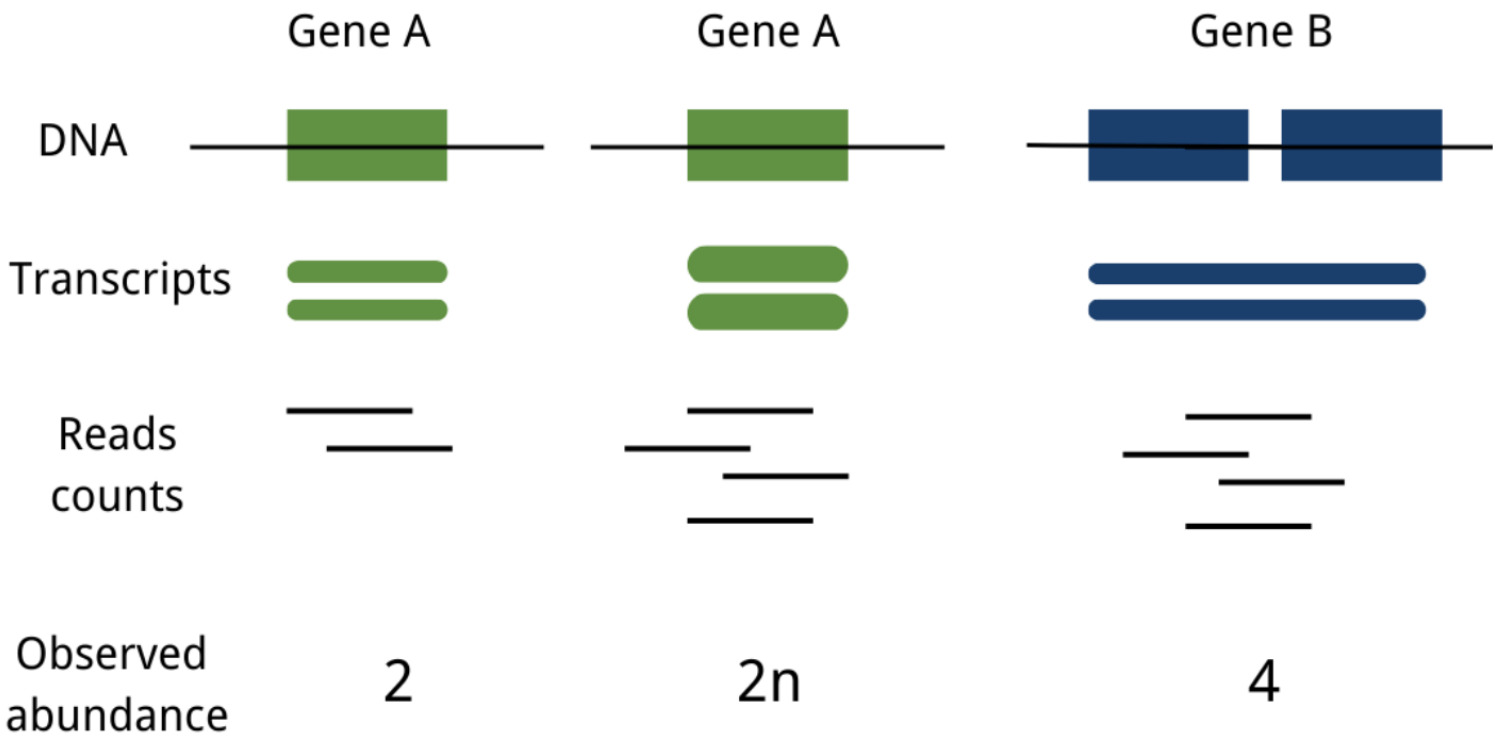
Measure gene expression through RNA-Seq in specific tissue or cell type



RNA quantification: prepare count matrix with expression levels for each gene in each individual

Magul et al BMC Genomics

RNA sequencing depth is a function of gene length



- As Gene A sequencing depth increases, more read counts are produced
- Gene B produces more gene counts at the same sequencing level as Gene A, since it has greater gene length

→ Need to normalise for gene length

Within sample normalization

$$FPKM = \frac{ExonMappedFragments * 10^9}{TotalMappedFragments * ExonLength}$$

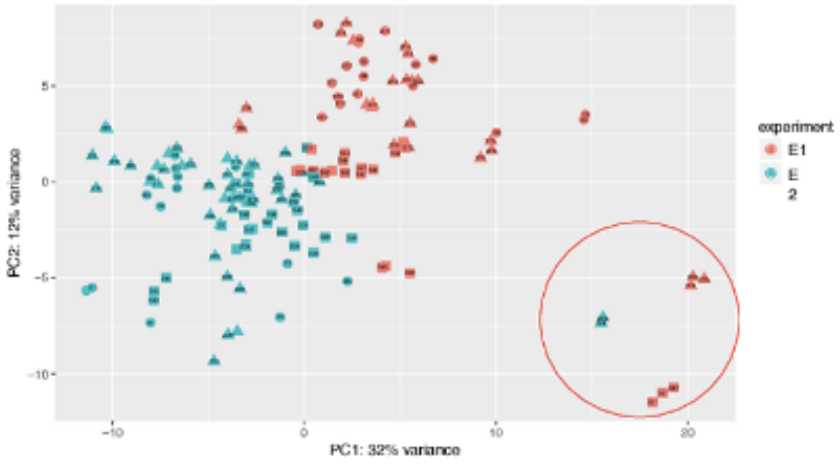
FPKM (Fragments Per Kilobase per Million mapped fragments) normalizes the abundance of transcripts from different samples to a standard that allows quantitative comparison by dividing transcript length (L) and total number of Reads (N)

$$TPM = \frac{Ni/Li * 10^6}{sum(N1/L1 + N2/L2 + \ldots + Nn/Ln)}$$

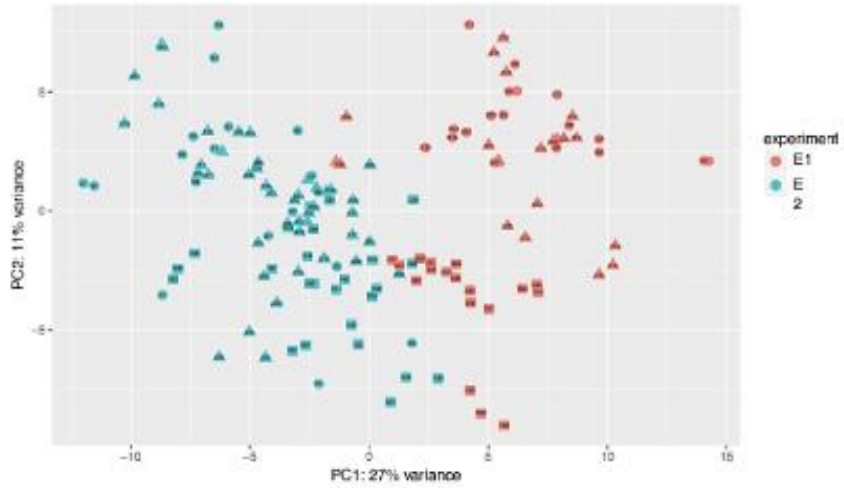
TPM (transcripts per kilobase million) is very much like FPKM, but the only difference is that at first, normalize for gene length, and later normalize for sequencing depth.

Between sample normalization

- QC to exclude problematic samples (may result in loss of power or introduction of artefacts)
- Samples mostly fail due to problems with original biospecimen (e.g. RNA degradation)
- Inspect data for outliers and batch effects using PCA



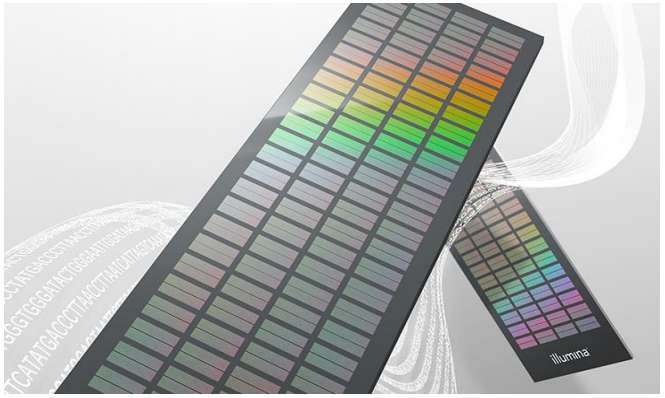
with outliers

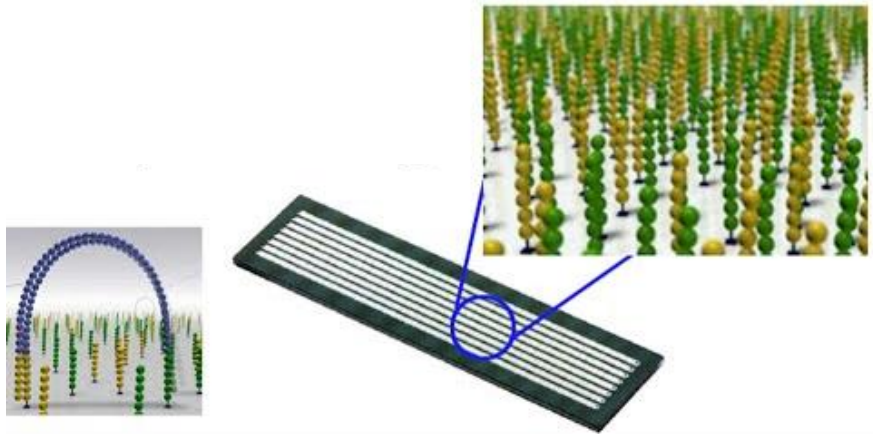


without outliers

- Inverse normal transformation of gene expression to conform to assumptions of regression model (homoscedasticity)

Identify genetic variants through SNP arrays or whole genome sequencing





SNP arrays

- Target hundreds of thousands to few million known genetic variants
- Genotype + imputation to obtain good coverage of common variants

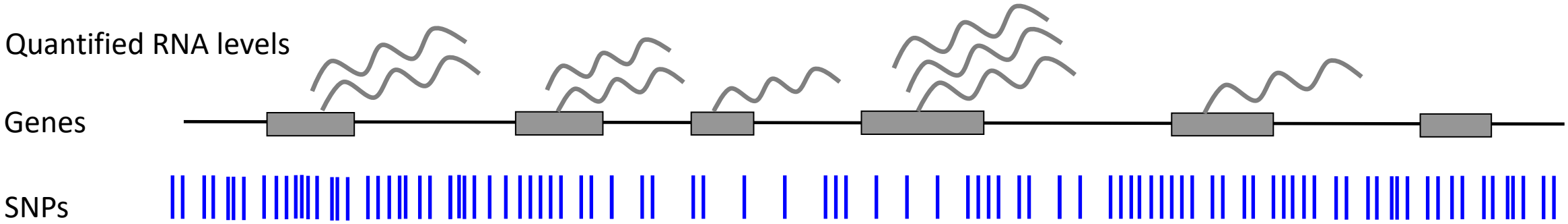
Whole genome sequencing

- Provides increased power for identifying causal variants
- Provides possibility to map QTL effects for complex genetic variants (including short tandem repeats, indels, structural variants)

- QC on genotypes similar to GWAS
- Determine minor allele frequency (MAF) threshold for study
- Studies with small sample sizes are susceptible to false positive associations resulting from variants with low MAFs
- Require higher MAF thresholds to ensure the robustness of association test
- Typically minimum sample size of ~70 samples

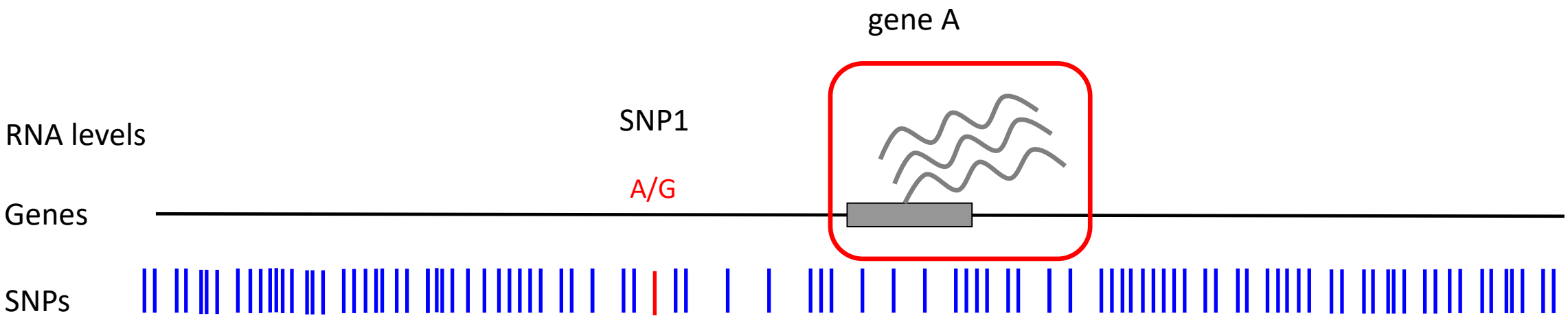
Illumina

Genome-wide association of genetic variation with levels of gene expression (molecular trait) to map eQTLs (molQTLs)



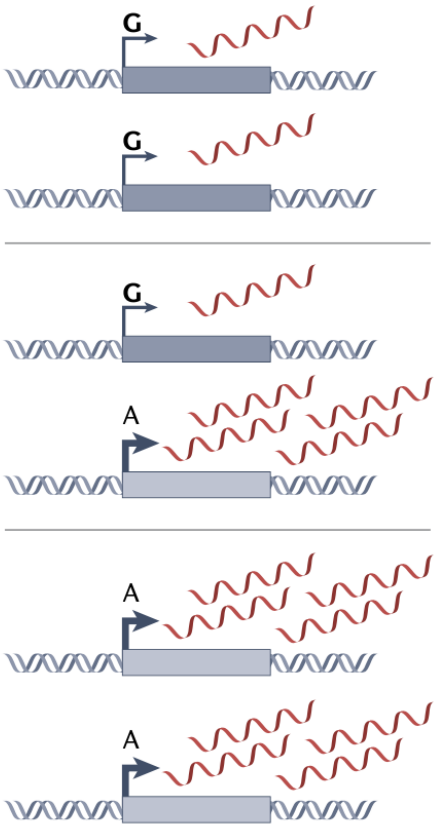
- Association of millions of SNP with tens or hundreds of thousands of molecular features across the genome (all expressed genes in a tissue) through linear regression

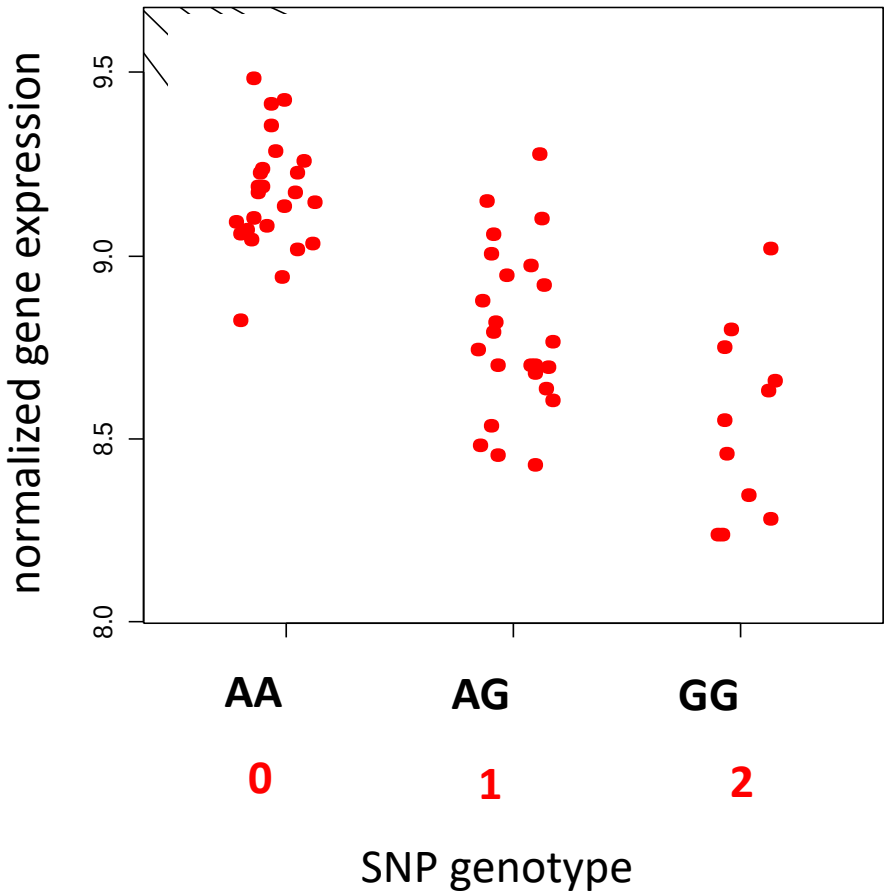
Association of SNP genotype to levels of gene expression to map eQTLs



- Test for association between SNP genotype at SNP1 with gene expression at gene A using linear regression

Linear regression to associate SNP genotype with gene expression





Linear regression (LR)

Expression ~ Genotype + Covariates*

- slope
- observed p-value
- r^2

Assumes additivity of genetic effects

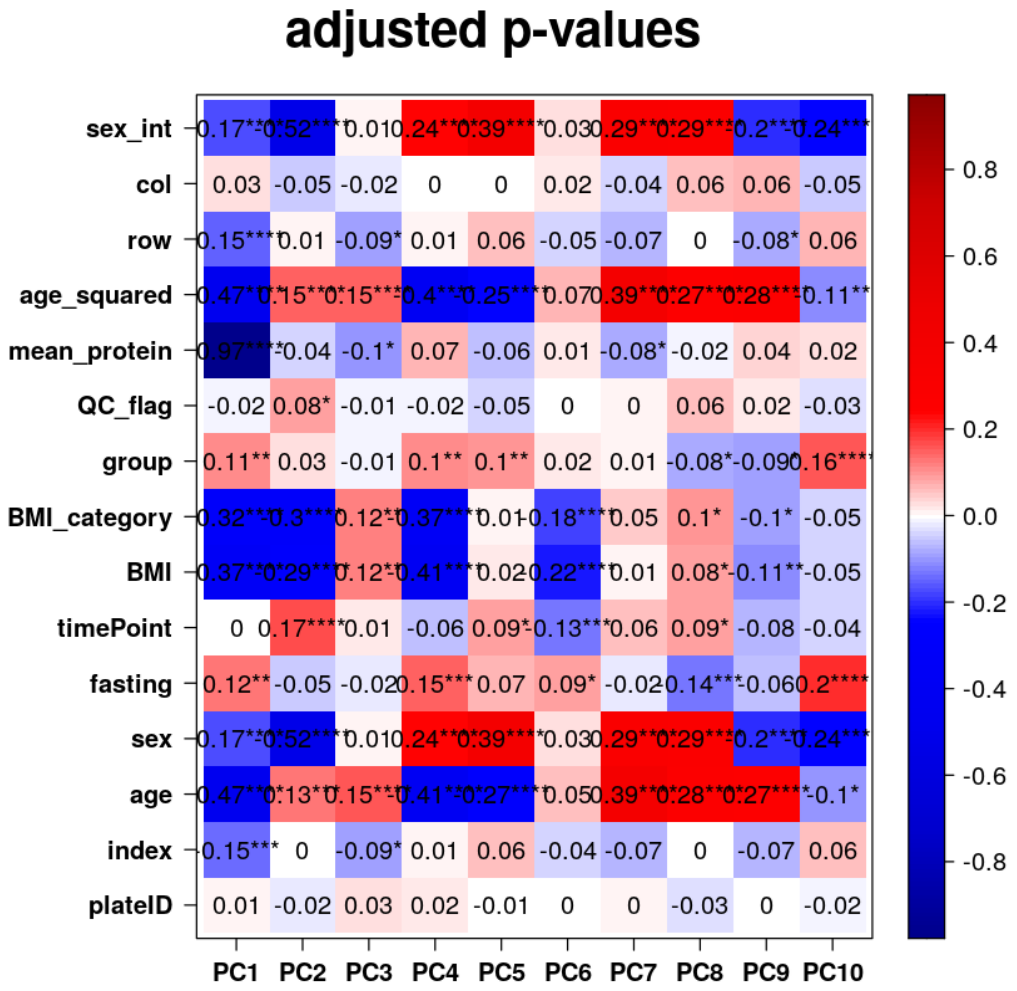
Software: FastQTL, Tensor eQTL, OSCA, Matrix eQTL

Correcting for confounding factors (covariate inclusion)

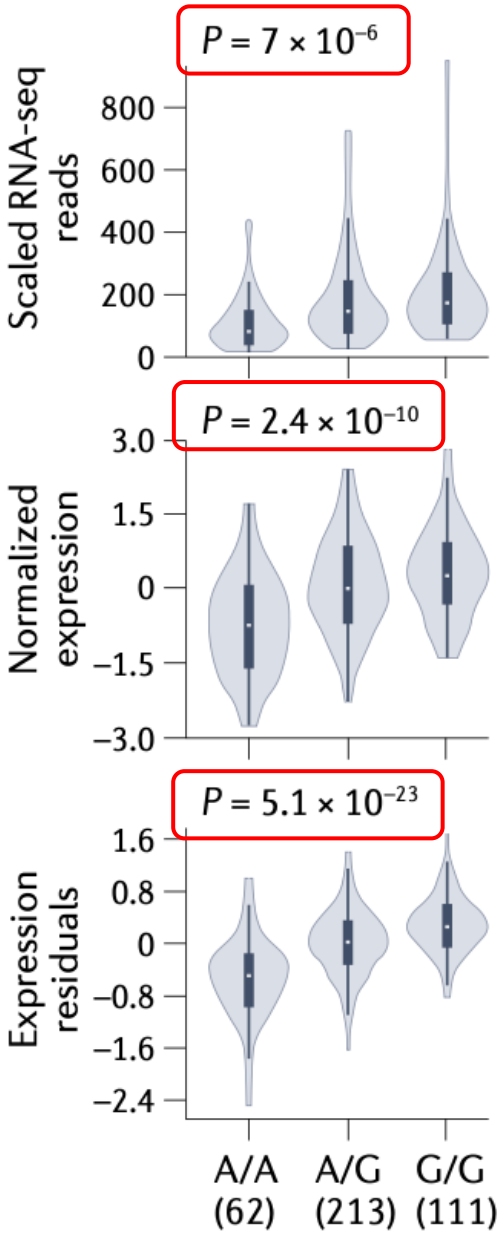
- Confounding effects can lead to loss of power, false positive associations
- Include covariates in association testing that account for effects of confounders

- Known confounding effects of e.g. age and sex on gene expression
- Better to correct using latent variables computed from the normalized expression data using:
  - principal components (PCs)
  - probabilistic estimation of expression residuals (PEER) factors

- Identify covariates to include in association testing



Gene expression as function of genotype – effect of QC, normalization and correction for covariates



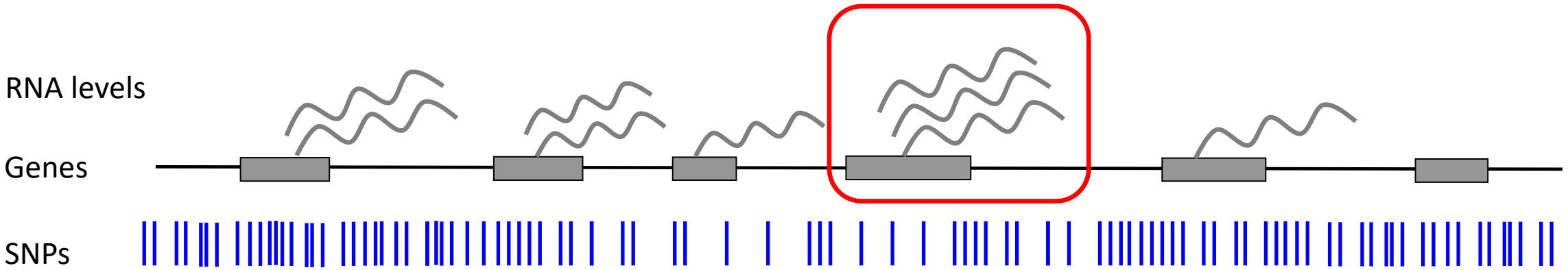
Association test on gene expression (RNA-Seq) corrected only for library size

Association test on inverse normal transformed gene expression values

Association test on gene expression residuals after covariate correction

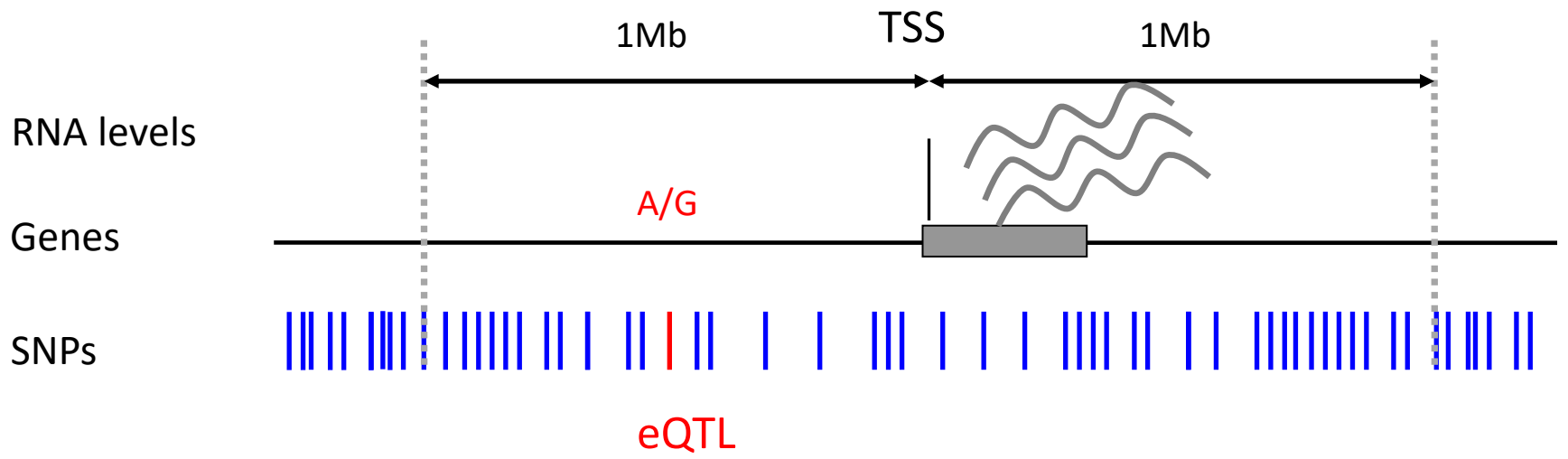
Aguet et al Nat Rev Methods Primers 2023

Genes (molecular features) have predefined locations in the genome → association testing can be done in *cis* and *trans*



- Typical eQTL study includes:
  - 1 x 10^7 common variants (MAF>= 5%)
  - 20,000 expressed genes (molecular phenotypes)

Cis association testing looks for proximal effects to feature studied, *trans* looks for distal effects



- *Cis* testing/mapping involves testing for association between gene expression (feature) in a 2 Mb window (max) centered on the gene's TSS (i.e. the feature itself)

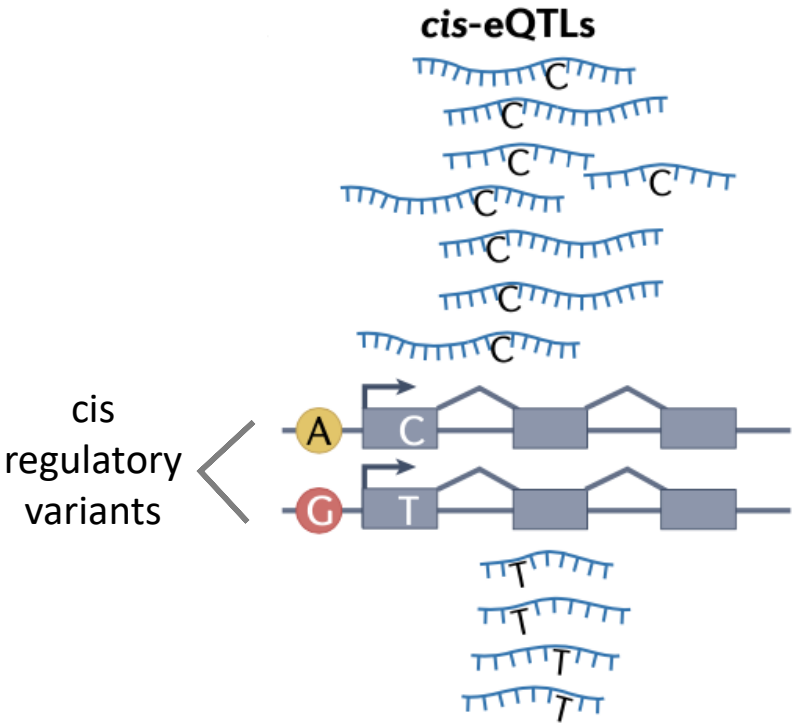
  → ~2 x 10^8 tests for all variant-phenotype pairs in *cis* (assuming 1x10^4 variants in each *cis* window)

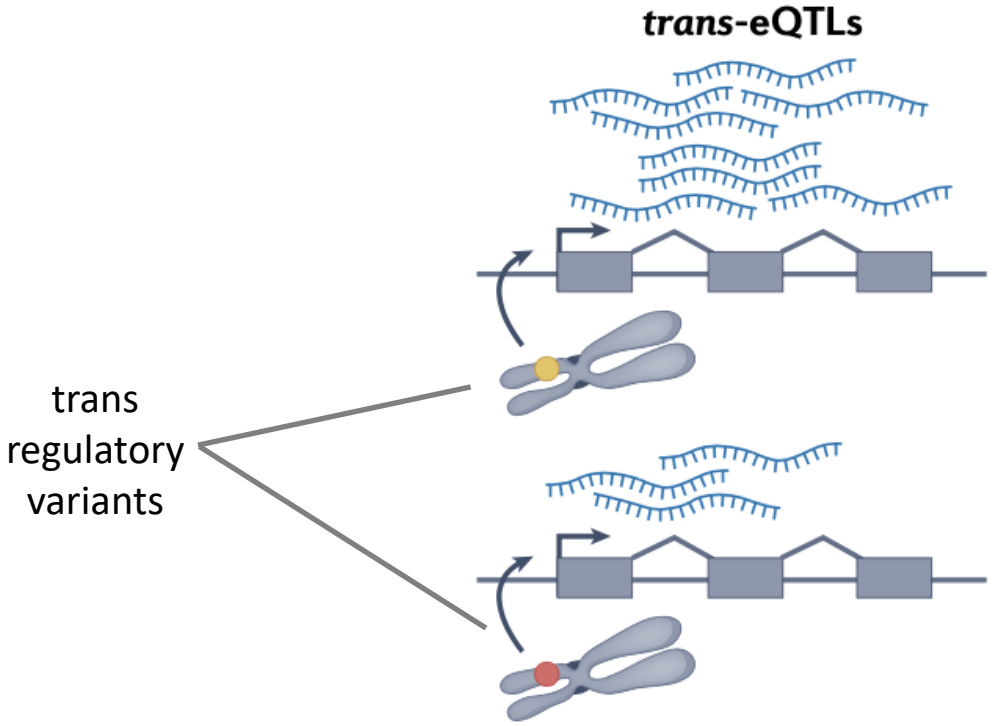
- Trans testing/mapping tests for associations between a feature and SNPs located at least 5 Mb away, or on another chromosome

  → 2 x 10^{11} tests for all variant-phenotype pairs in *trans*

Cis variants are typically close to their molecular targets, *trans* variants are usually on a different chromosome

But exact biological definition is not dependent on distance





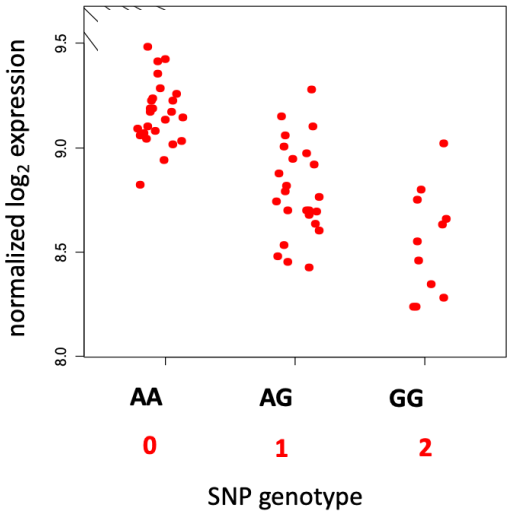
Alleles A/G of a cis-regulatory variant (eQTL) affect a molecular feature (gene expression) via cis-regulatory activity of the physically connected chromatid

Trans-acting regulatory variants usually act via intermediary molecules

Aguet et al Nat Rev Methods Primers 2023

Assigning significance in *cis* eQTL mapping

- In GWAS a fixed p-value (5 x 10^{-8}) significance threshold is typically applied
- In QTL mapping significance can be assigned through genome-wide threshold or through permutations



	SNP genotype	Normalised gene expression
Individual 1	AA	9.1
Individual 2	AA	9.3
Individual 3	AG	8.4
Individual 4	AG	8.5
…	…	…
Individual X	GG	8.2
Individual Y	GG	8.1

Linear regression

1 observed p-value

x1,000-10,000

permute

SNP genotype	Normalised gene expression
AA	9.1
AG	9.3
AG	8.4
…	8.5
GG	…
GG	8.2
AA	8.1

Linear regression

1,000-10,000 permuted p-values

- For a given gene define permutation threshold as lowest permuted p-value
- Compare observed p-value to permutation threshold
- Use p-values to compute FDR

→ To reduce the number of permutations required to accurately compute FDR, software tools (eg FastQTL) leverage property that empirical *P* values can be approximated by a beta distribution fitted to a limited set of permutations

eQTL mapping workflow

- What tissue to study?
- How many individuals?

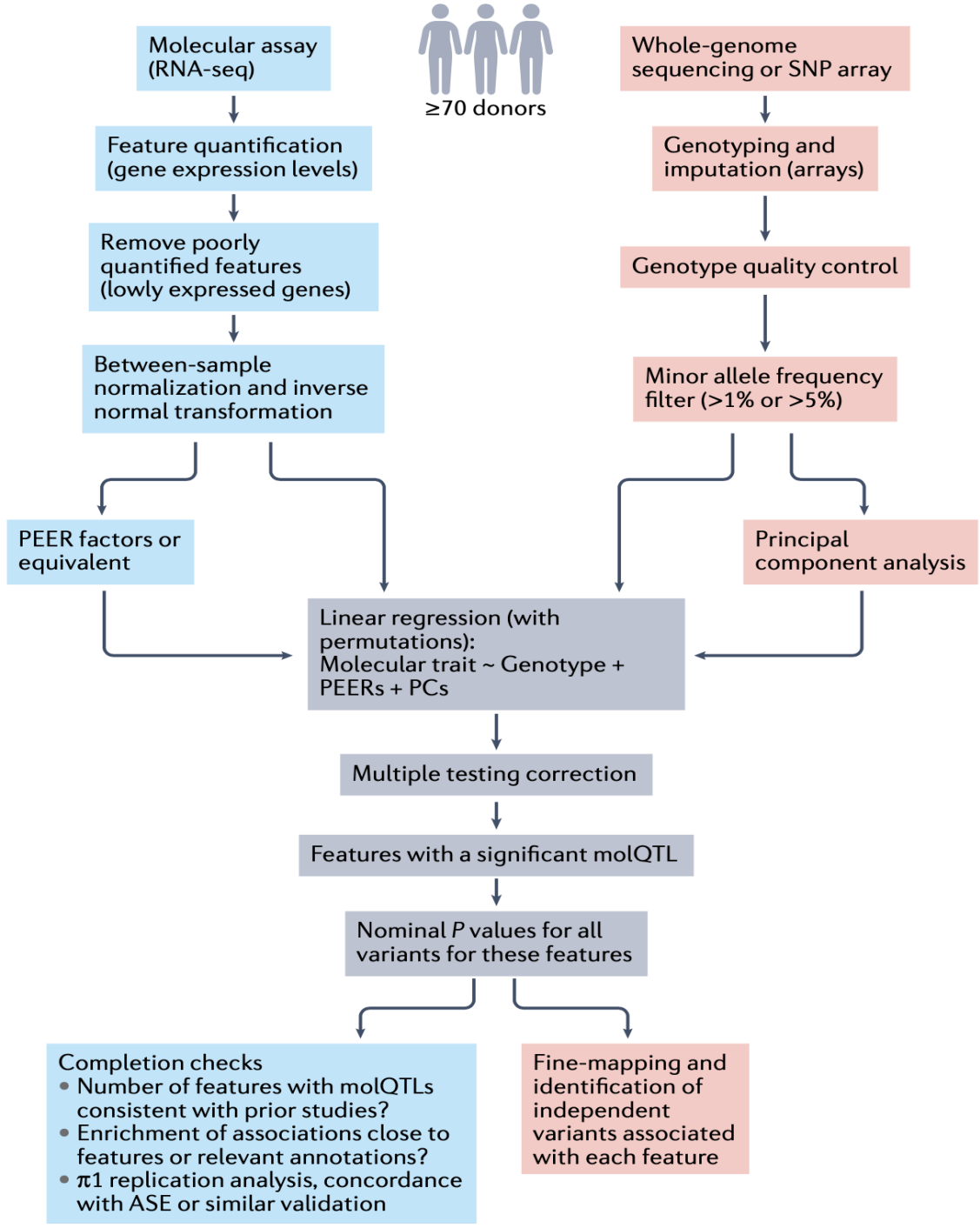
- Collect molecular trait data
- QC on molecular trait data

- Collect genotype data
- QC on genotype data

- Explore data to determine factors affecting molecular trait distribution (covariates, PEERs) and correct for genetic structure (PCs)

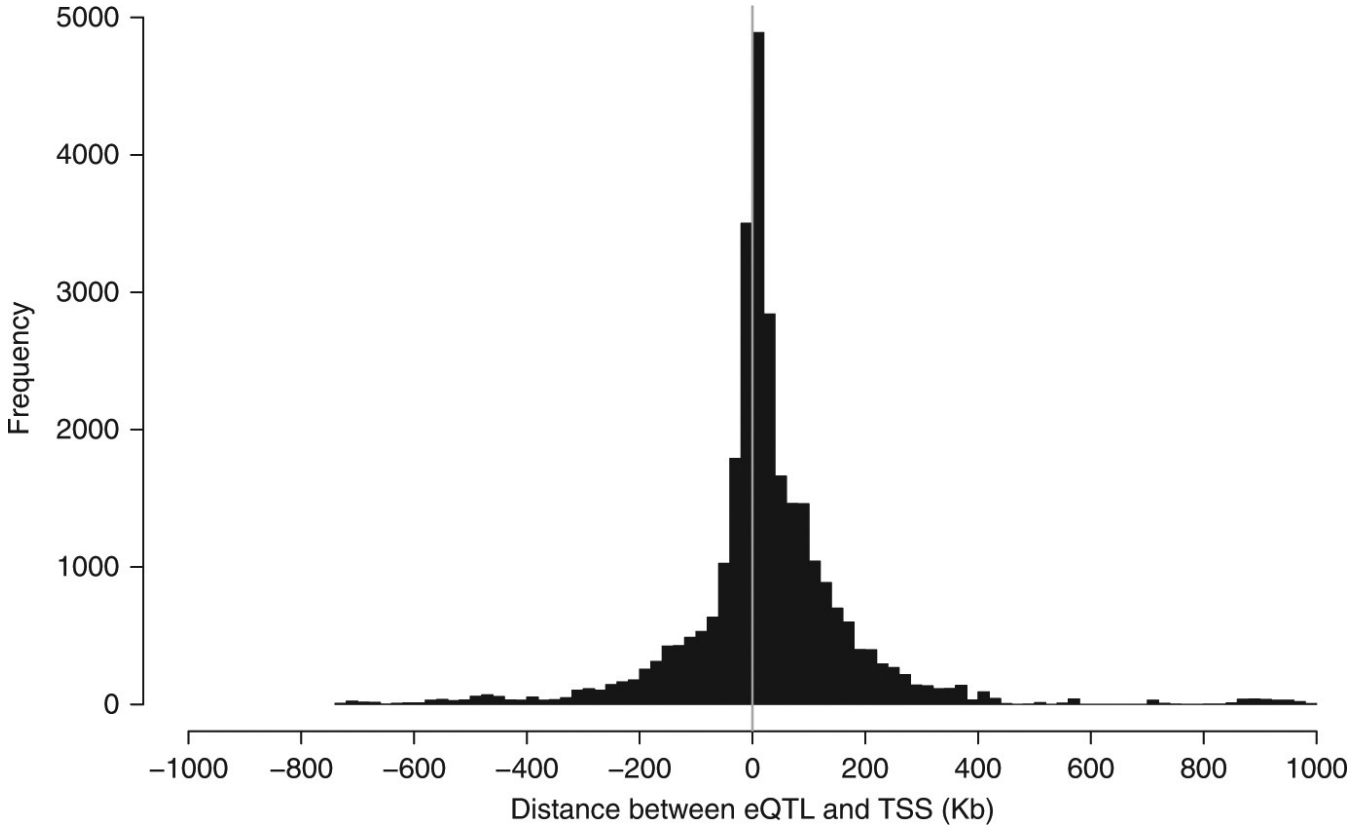
- Run association through linear regression with covariates

- Process output of regression model to identify significant associations, correcting for large number of tests performed



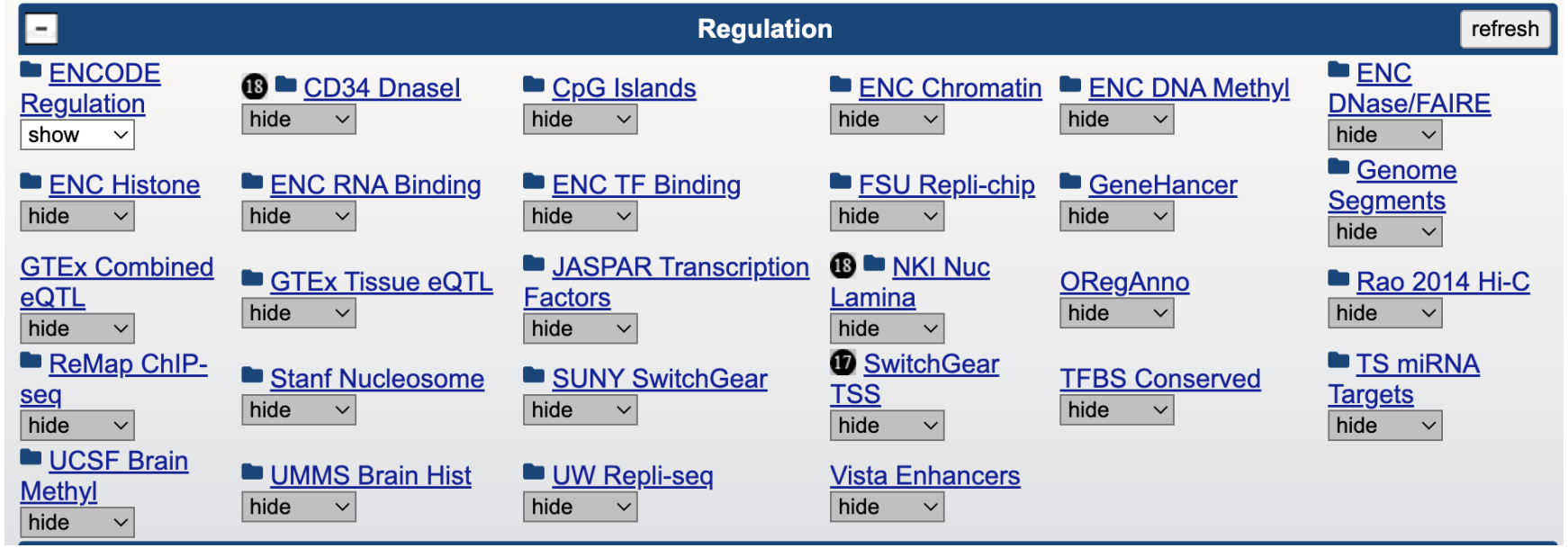
Distribution of eQTLs across the genome

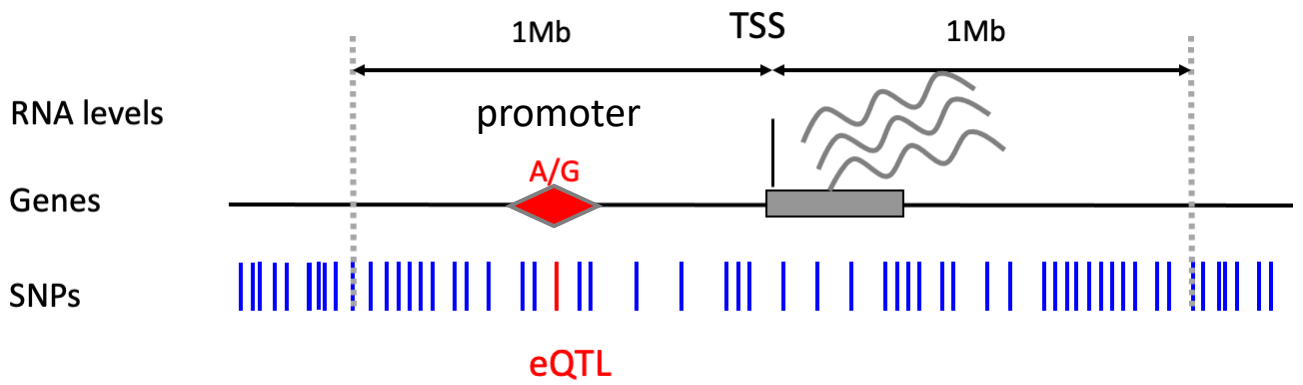
- Following association testing perform series of sanity checks on results
  - E.g. map distance of *cis* eQTLs to the TSS expecting clustering of most eQTLs around TSS since region contains a great proportion of regulatory elements



Sanity checks - mapping *cis* eQTLs on genomic features

- Wealth of publicly available information on genomic features in different tissues and cell types provides opportunity to ask: do eQTLs map in known functional annotations?
- Select genomic features from appropriate tissue/cell type and perform enrichment analysis

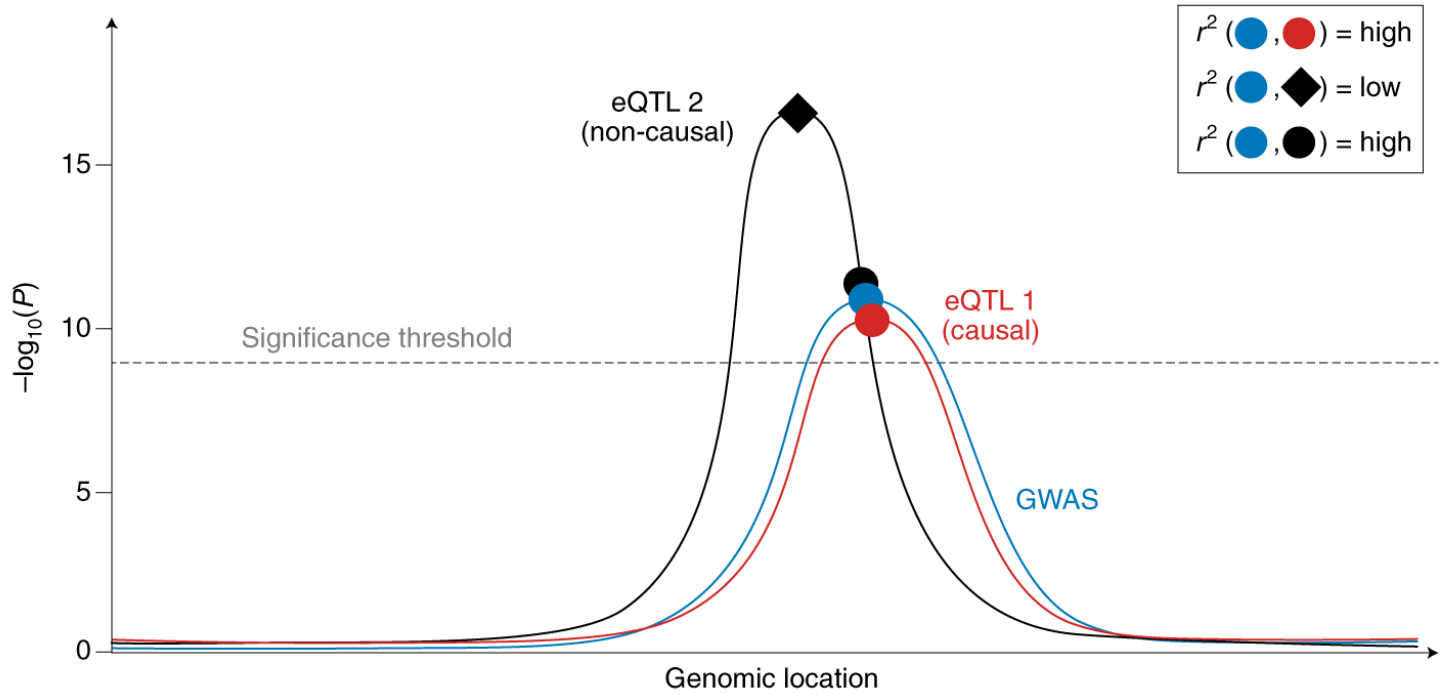




UCSC Genome Browser

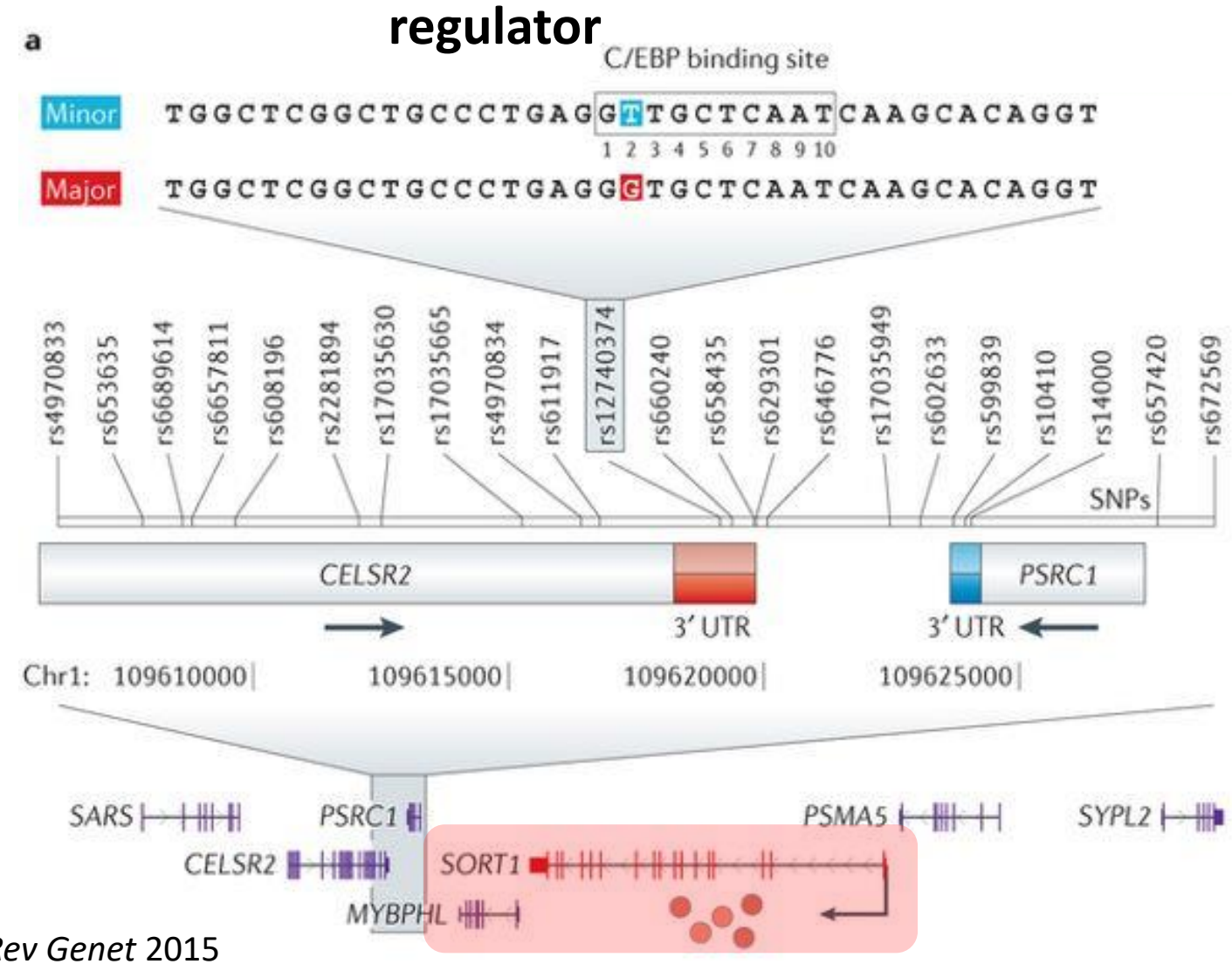
eQTL – GWAS SNP co-localization – is eQTL the causal variant?

- Molecular changes captured by eQTLs may have causal role in shaping GWAS traits
- To address whether eQTLs underlie GWAS traits, can ask the question: is a GWAS signal and an eQTL that map to a specific locus driven by shared causal variants?
- First quantify support for each variant being causal for each trait (GWAS, eQTL)
- Then aggregate this info across variants to estimate global support for co-localization



Liu et al Nat Genet 2019

Integrating eQTL and GWAS information to understand how genes are regulated: key to unlocking disease mechanisms



Albert *Nat Rev Genet* 2015

molQTL resources

Study/resource	Type	Number of donors	Population ancestries	Biospecimens	Molecular phenotypes
eQTL Catalogue[76]	Aggregated database of reanalysed data	73–948 per study, total 8,193	88.5% European ancestries	Diverse	Transcriptome phenotypes
GTEx[58]	Consortium with centralized data production and analysis	73–706	American; 85% European and 11% African ancestries	49 postmortem tissues	Gene expression and splicing, others in smaller scale
eQTLGen[5]	Consortium with federated analysis	31,684	Predominantly European	Whole blood	Gene expression
GoDMC[42]	Consortium with federated analysis	32,851	European ancestries	Whole blood	DNA methylation
Hawe et al.[43]	Research project	6,994	European and South Asian ancestries	Whole blood	DNA methylation
Ferkingstad et al.[50]	Single-cohort study	35,559	Icelandic	Plasma	Aptamer proteomics
Jerber et al.[158]	Research project	215	European ancestries	In vitro differentiated iPSCs	scRNA-seq
Yazar et al.[153]	Research project	982	European ancestries	PBMCs	scRNA-seq

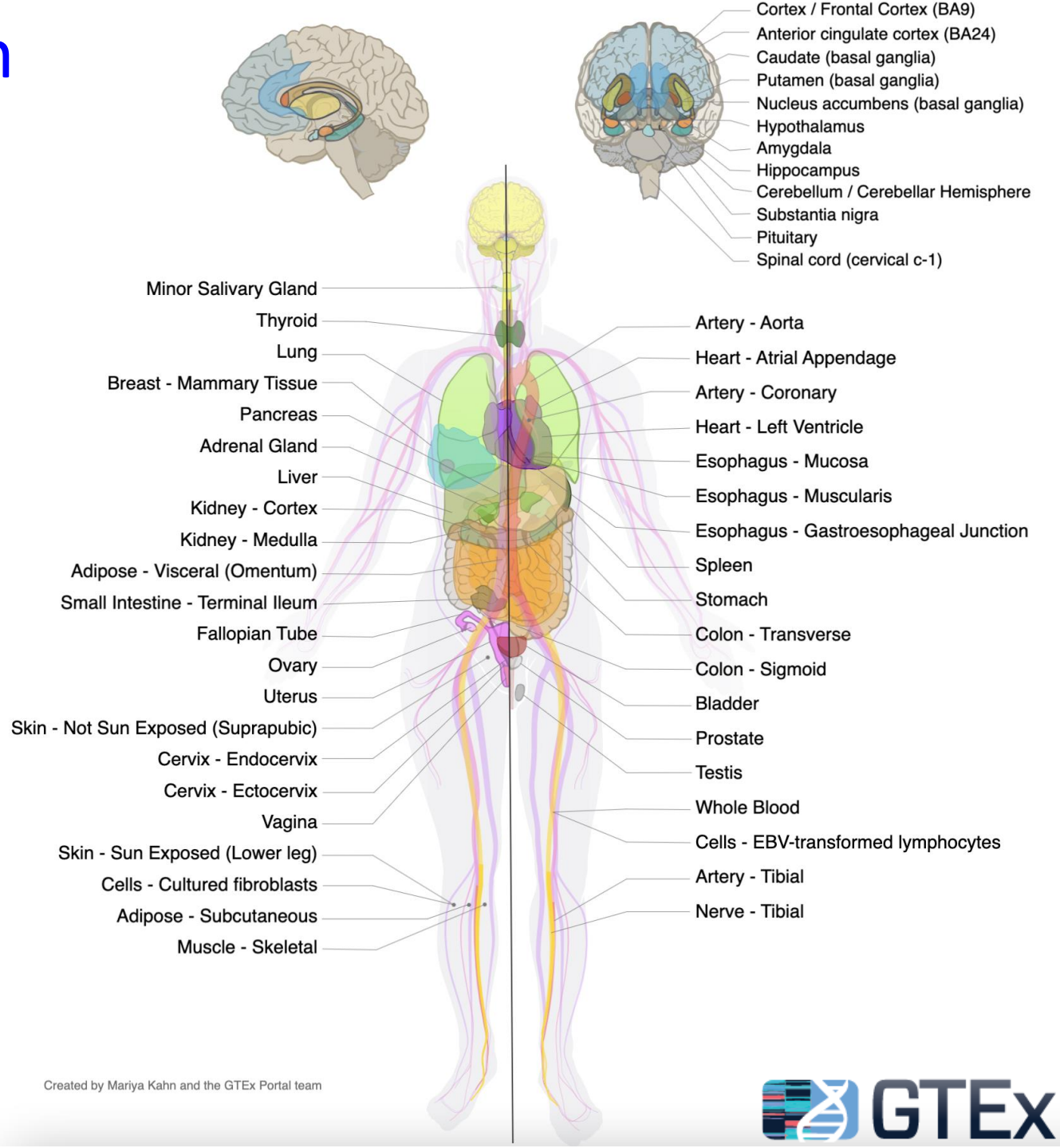
iPSC, induced pluripotent stem cell; molQTL, molecular quantitative trait locus; PBMC, peripheral blood mononuclear cell; scRNA-seq, single-cell RNA sequencing.

Aguet et al Nat Rev Methods Primers 2023

Genotype-Tissue Expression (GTEx) resource

- Public resource to study tissue-specific expression and regulation
  - 54 tissue sites
  - ~1000 individuals
  - WGS, WES, RNA-Seq

- GTEX Portal provides open access to data including gene expression, QTLs, and histology images



www.gtexportal.org

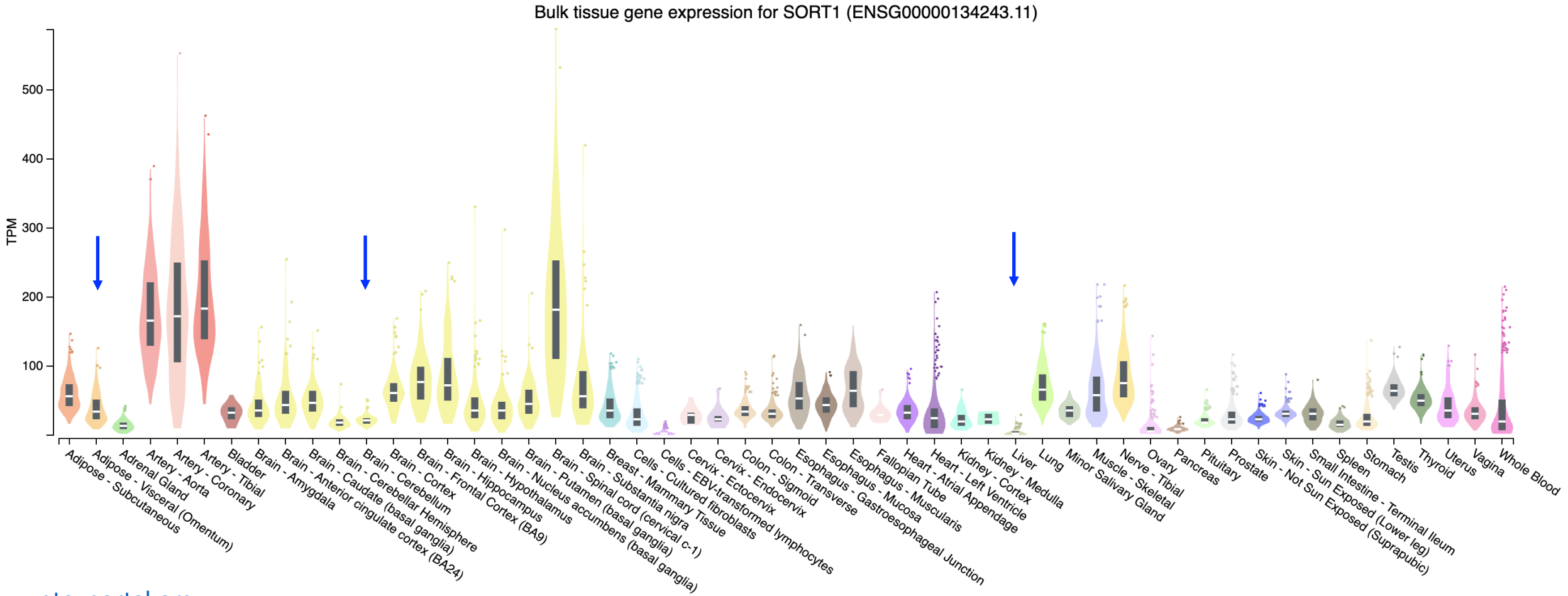
SORT1 gene expression levels across 54 tissues

Bulk tissue gene expression for SORT1 (ENSG00000134243.11)

Data Source: GTEx Analysis Release V8 (dbGaP Accession phs000424.v8.p2)

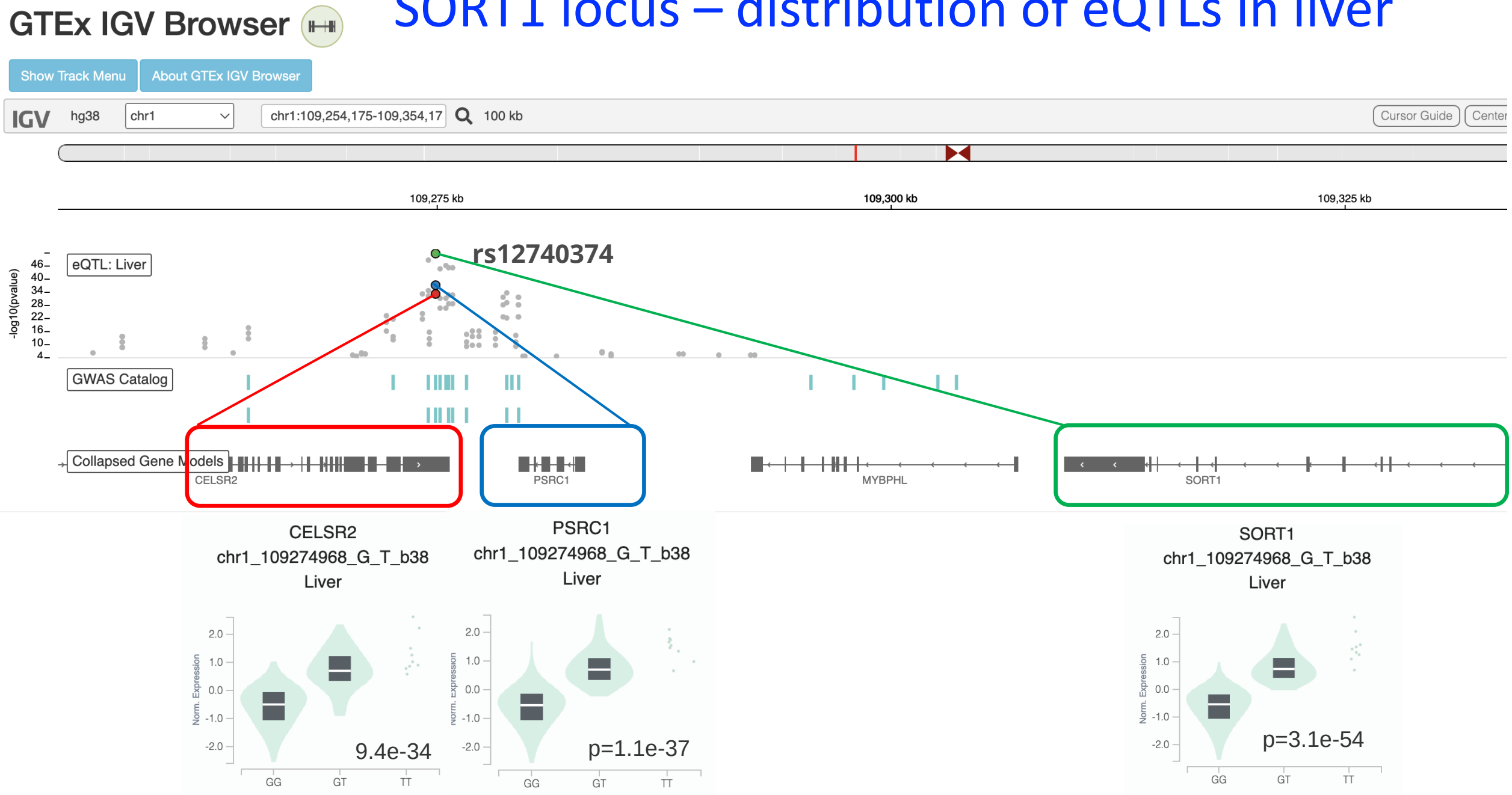
Data processing and normalization

SUBSET None Sex SCALE Log Linear TISSUE SORT MEDIAN SORT OUTLIERS On Off



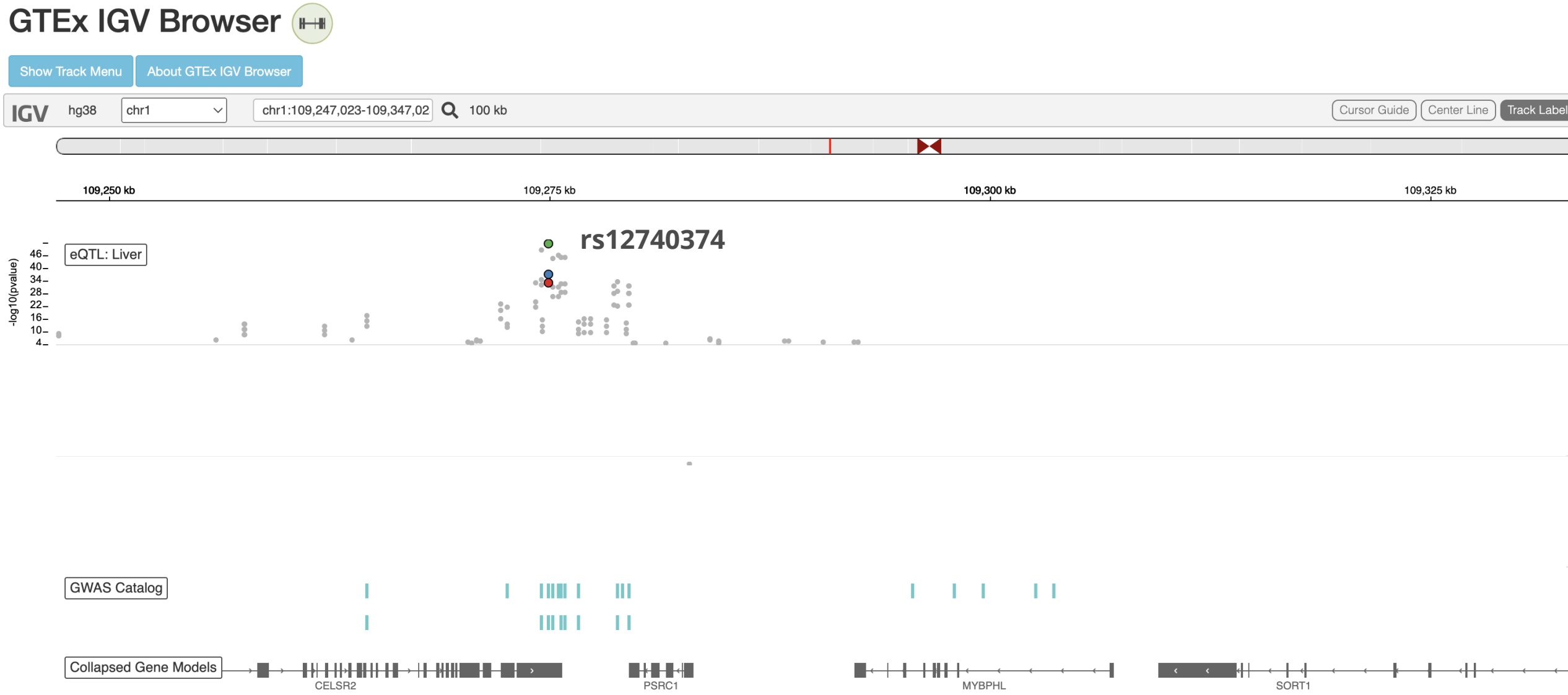
www.gtexportal.org

SORT1 locus – distribution of eQTLs in liver



www.gtexportal.org

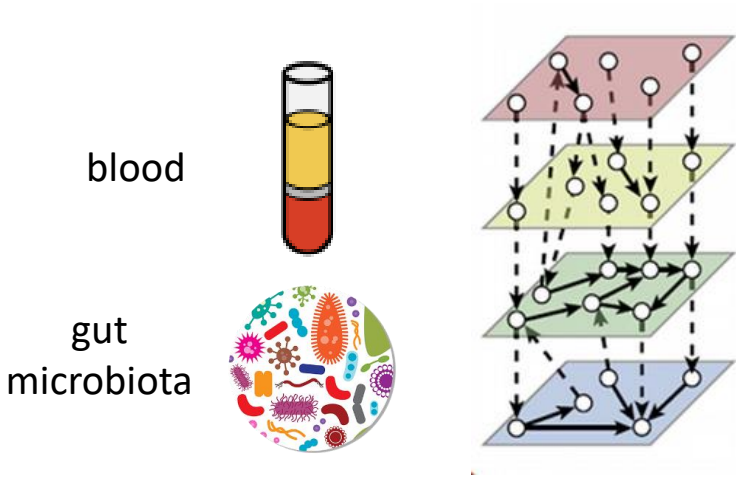
SORT1 locus – distribution of eQTLs in liver, brain, adipose tissue



www.gtexportal.org

FASTBIO: exploring the molecular impact of dietary restriction and its impact on human health

Harnessing the power of the immune system through metabolic reprogramming



Molecular phenotypes

Genotypes (DNA)
Transcriptome (RNA)
Methylome
Proteome
Metabolome
Gut microbiome

Mapping eQTLs, pQTLs, metaQTLs to understand how genetic effects influence response to dietary restriction

Multi-omics integration to address effects of dietary restriction on health, disease and response to drugs

Using diet as a complementary therapeutic means

Positions available
Contact: dimas@fleming.gr

"ALEXANDER FLEMING"
Biomedical Sciences Research Center

HELMHOLTZ MUNICH