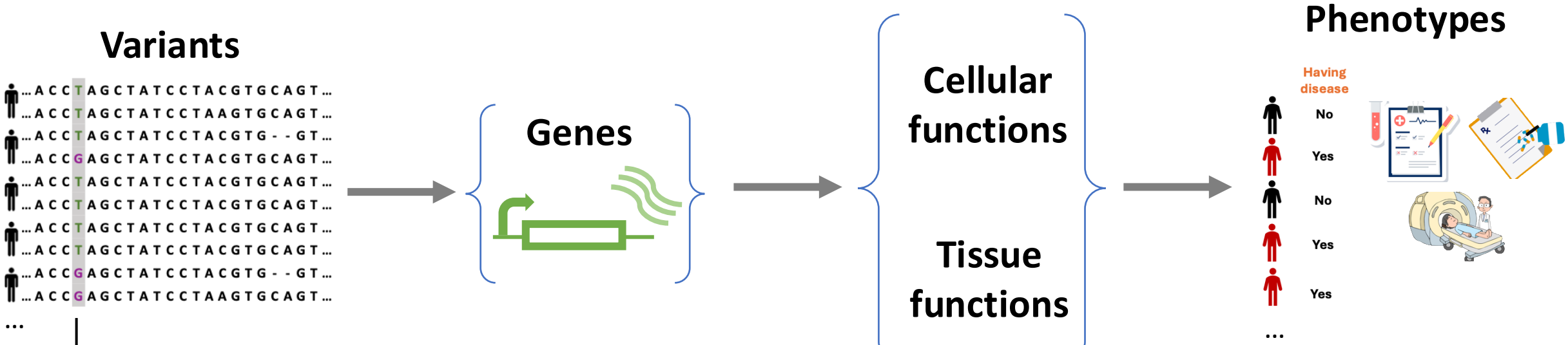
# Molecular QTL mapping: The statistical foundations

Wei Zhou, Ph.D.

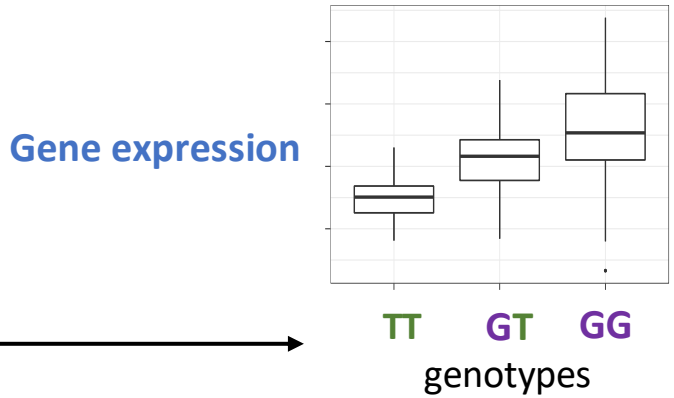
Center for Genomic Medicine, Massachusetts General Hospital
Stanley Center of Psychiatric Research, Broad Institute

The 2026 International Statistical Genetics Workshop

# Expression quantitative trait locus (eQTL) mapping detects variants that influence gene expression




**eQTL mapping** tests variant-gene expression associations



# Molecular QTLs: same core idea, different phenotype layers

| QTL | Full Name | Molecular Phenotype | Assay | Statistical Model | Key Resource |
|---|---|---|---|---|---|
| **Linear regression on normalized, continuous phenotypes** | | | | | |
| **eQTL ★** | **expression QTL** | Gene expression level | RNA-seq (bulk) | Linear regression | GTEx Consortium, Science, 2020 |
| **pQTL** | protein QTL | Protein abundance | Mass spec / SomaScan | Linear regression | Ferkingstad et al., Nat Genet 2021., Sun *et al.*, Nat Genet 2023 |
| **caQTL** | chromatin accessibility QTL | Open chromatin peaks | ATAC-seq (bulk) | Linear regression | Calderon *et al.*, Nat Genet 2019 |
| **haQTL** | histone acetylation QTL | H3K27ac / H3K4me1 marks | ChIP-seq (bulk) | Linear regression | Hou *et al.*, Nat Genet 2023 |
| **Require different models (bounded or compositional phenotypes)** | | | | | |
| **mQTL** | methylation QTL | CpG methylation beta values | Bisulfite-seq / array | Logit-transformed linear or beta regression | GoDMC Consortium, Nat Genet 2021; Oliva et al., Nat Genet 2023 |
| **sQTL** | splicing QTL | Intron excision ratios | RNA-seq (bulk) | Dirichlet-multinomial or ratio-based (PSI) | GTEx Consortium, Science, 2020 |

# The linear model: testing variant-expression associations

$$Y = \beta_0 + \beta_1 G + \beta_2 C + \varepsilon$$

**Y**
gene expression
*(normalized read counts)*

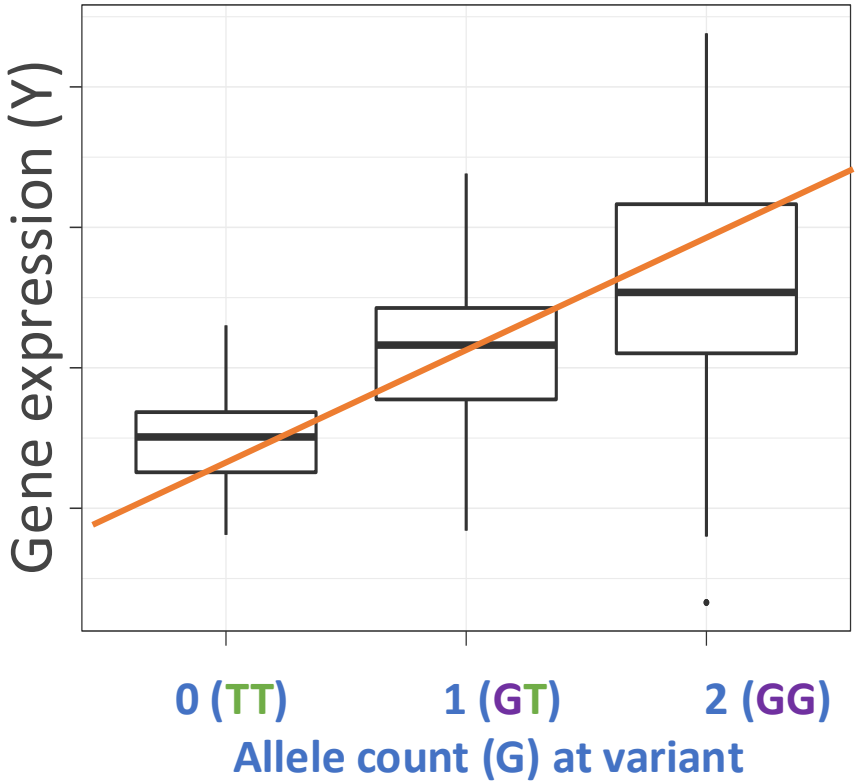
**Key term**
$\beta_1 G$
genotype effect size
*G = 0, 1, or 2 (allele count)*

**$\beta_2 C + \varepsilon$**
covariates + noise
*(sex, age, PCs, PEER factors)*

- **Test:** Is $\beta_1$ significantly different from zero?
- **Run for:** each gene × each nearby variant (cis-window: ~1 Mb)
- **Correct for multiple testing:** permutations or Benjamini-Hochberg FDR per gene

**Key assumption:**
**expression Y is approximately normally distributed and independent.**

# Visualizing eQTL detection: genotype stratifies expression




- Each box shows expression distribution for individuals with that genotype
- $\beta_1$ = slope = change in mean expression per additional testing allele
- **eQTL detected** when expression increases or decreases across genotype groups

$$Y = \beta_0 + \beta_1 G + \beta_2 C + \varepsilon$$

$$H_0: \beta_1 \neq 0 \quad \text{vs.} \quad H_1: \beta_1 = 0$$

# Covariates, multiple testing, and calling eGenes

## Covariates (C)

- Genotype PCs (population stratification)
- Age, sex, batch
- PEER factors (hidden confounders in expression)
- Cell type composition (bulk)

## Multiple Testing

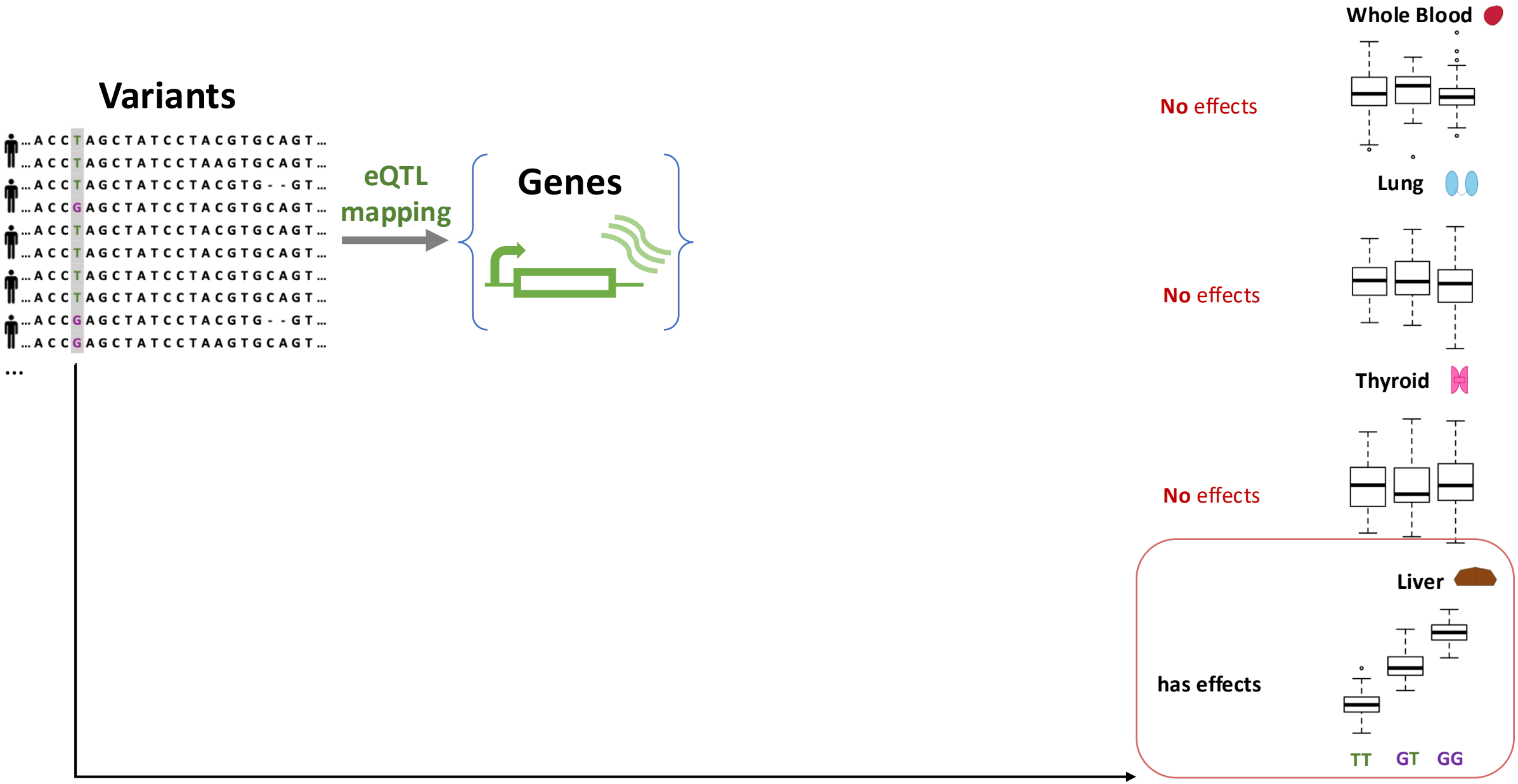
- Millions of tests: ~20K genes x ~10M variants
- **cis-window:** test variants within +/- 1 Mb of gene
- **Permutations:** get null p-value per gene, then BH-FDR across genes
- **Tools:** FastQTL, TensorQTL, Matrix eQTL

## Output: eGenes

- **eGene:** a gene whose expression is influenced by at least one genetic variant
- **Lead variant:** top associated variant per gene
- GTEx (v9) detected ~16K eGenes across 49 tissues

This framework powers GTEx and all major bulk eQTL studies and assumes continuous, normally distributed expression

# Bulk RNA-seq measures gene expression in different tissues

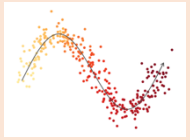


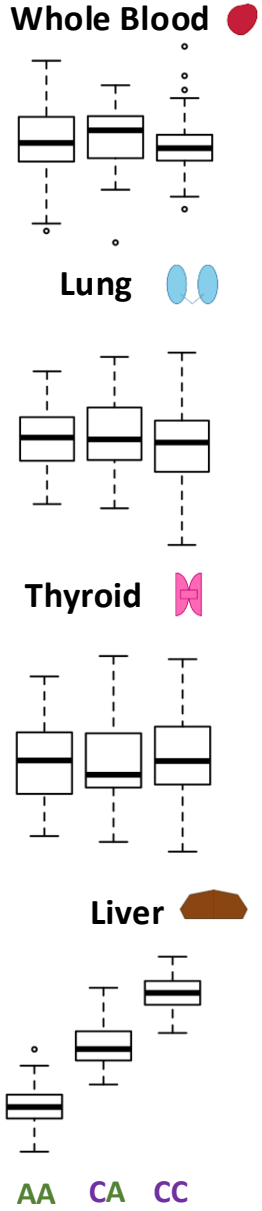
# Bulk RNA-seq measures gene expression in different tissues

## But **missing cellular variations**

**Cells vary in different aspects (cell context):**

- Cell type
- Cell state
  - ➢ under stimulations/treatments
- Cell cycle phase
  - ➢ Position in division cycle (G1, S, G2, M)
- Developmental stage
- Metabolic state





# References


- GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science. 2020 Sep 11;369(6509):1318-1330. doi: 10.1126/science.aaz1776. PMID: 32913098; PMCID: PMC7737656.
- Ferkingstad E, et al., . Large-scale integration of the plasma proteome with genetics and disease. Nat Genet. 2021 Dec;53(12):1712-1721. doi: 10.1038/s41588-021-00978-w. Epub 2021 Dec 2. PMID: 34857953.
- Sun *et al*., Plasma proteomic associations with genetics and health in the UK Biobank. Nature. 2023 Oct;622(7982):329-338. doi: 10.1038/s41586-023-06592-6. Epub 2023 Oct 4. PMID: 37794186; PMCID: PMC10567551.
- Calderon *et al*., Landscape of stimulation-responsive chromatin across diverse human immune cells. Nat Genet. 2019 Oct;51(10):1494-1505. doi: 10.1038/s41588-019-0505-9. Epub 2019 Sep 30. PMID: 31570894; PMCID: PMC6858557.
- Hou *et al*., Multitissue H3K27ac profiling of GTEx samples links epigenomic variation to disease. Nat Genet. 2023 Oct;55(10):1665-1676. doi: 10.1038/s41588-023-01509-5. Epub 2023 Sep 28. PMID: 37770633; PMCID: PMC10562256.
- GoDMC Consortium, Men et al., Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. Nat Genet. 2021 Sep;53(9):1311-1321. doi: 10.1038/s41588-021-00923-x. Epub 2021 Sep 6. PMID: 34493871; PMCID: PMC7612069.
- Oliva et al., DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. Nat Genet. 2023 Jan;55(1):112-122. doi: 10.1038/s41588-022-01248-z. Epub 2022 Dec 12. PMID: 36510025; PMCID: PMC10249665.