Expression Quantitative Trait Locus (eQTL) mapping and use in complex trait mapping

Barbara E. Stranger, PhD

University of Colorado School of Medicine

barbara.stranger@cuanschutz.edu

University of Colorado
Anschutz Medical Campus

Quantitative Traits

Important in Agriculture, Medicine and Ecology/Evolution.

- Agriculture
  Growth, Yield, Disease Resistance, Stress Resistance
- Medicine
  Disease Susceptibility, Drug response, Diet response
- Ecology and Evolution
  Many of the above traits are essentially components of fitness.

Is there a genetic basis underlying quantitative traits?

Is there a way to robustly identify the genetic basis?

Quantitative Trait Locus (QTL)

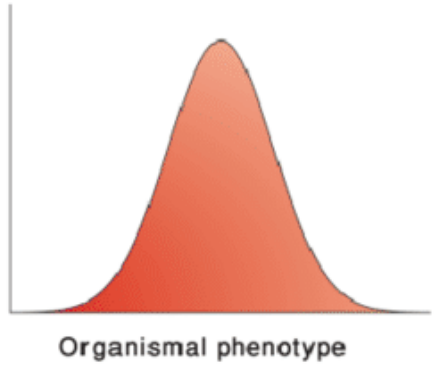
Region of the genome affecting quantitative phenotype

Phenotype is measured numerically.

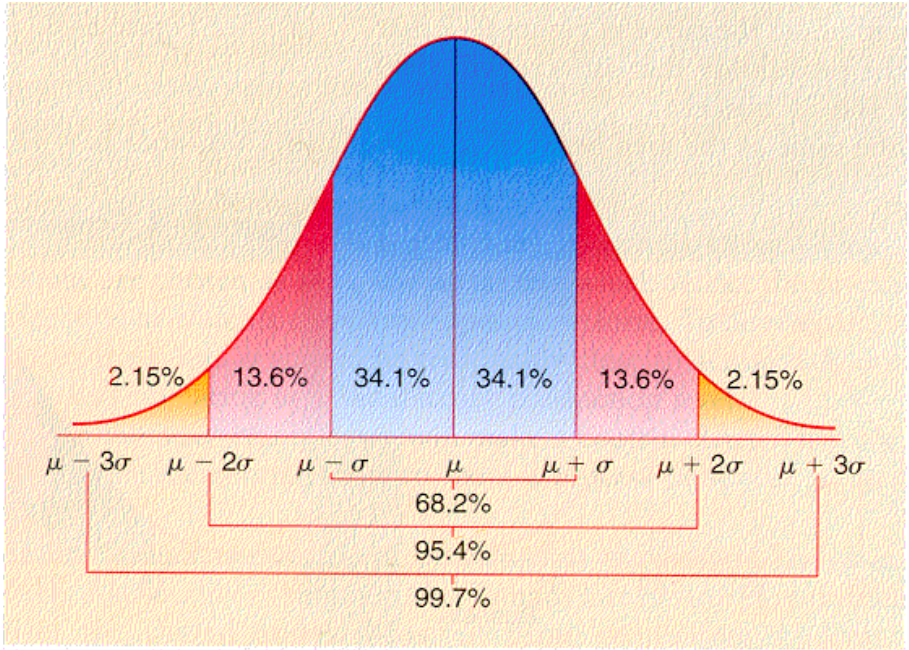
Variation in phenotype.

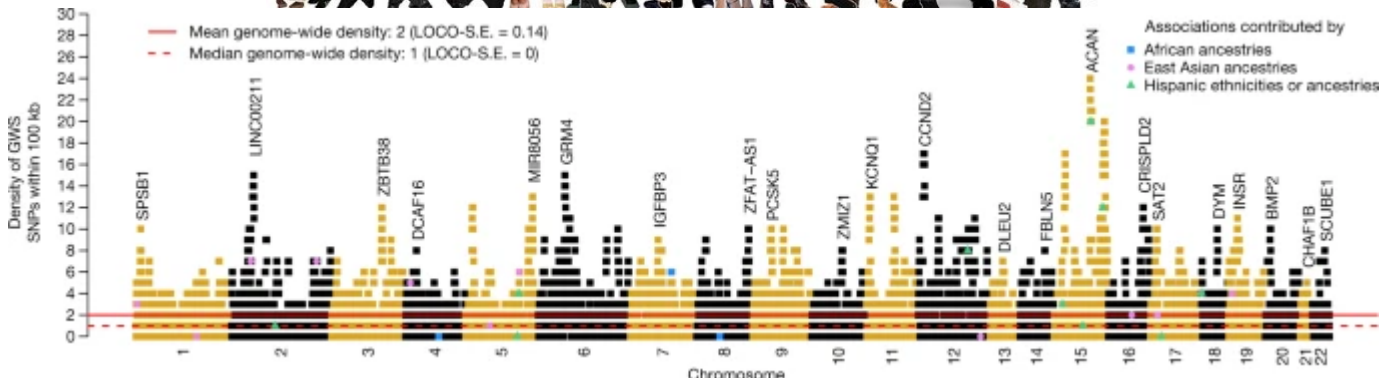
Variation in genotype.

Statistical link between genotype and phenotype



Human height





Yengo *et al*. 2022 *Nature*

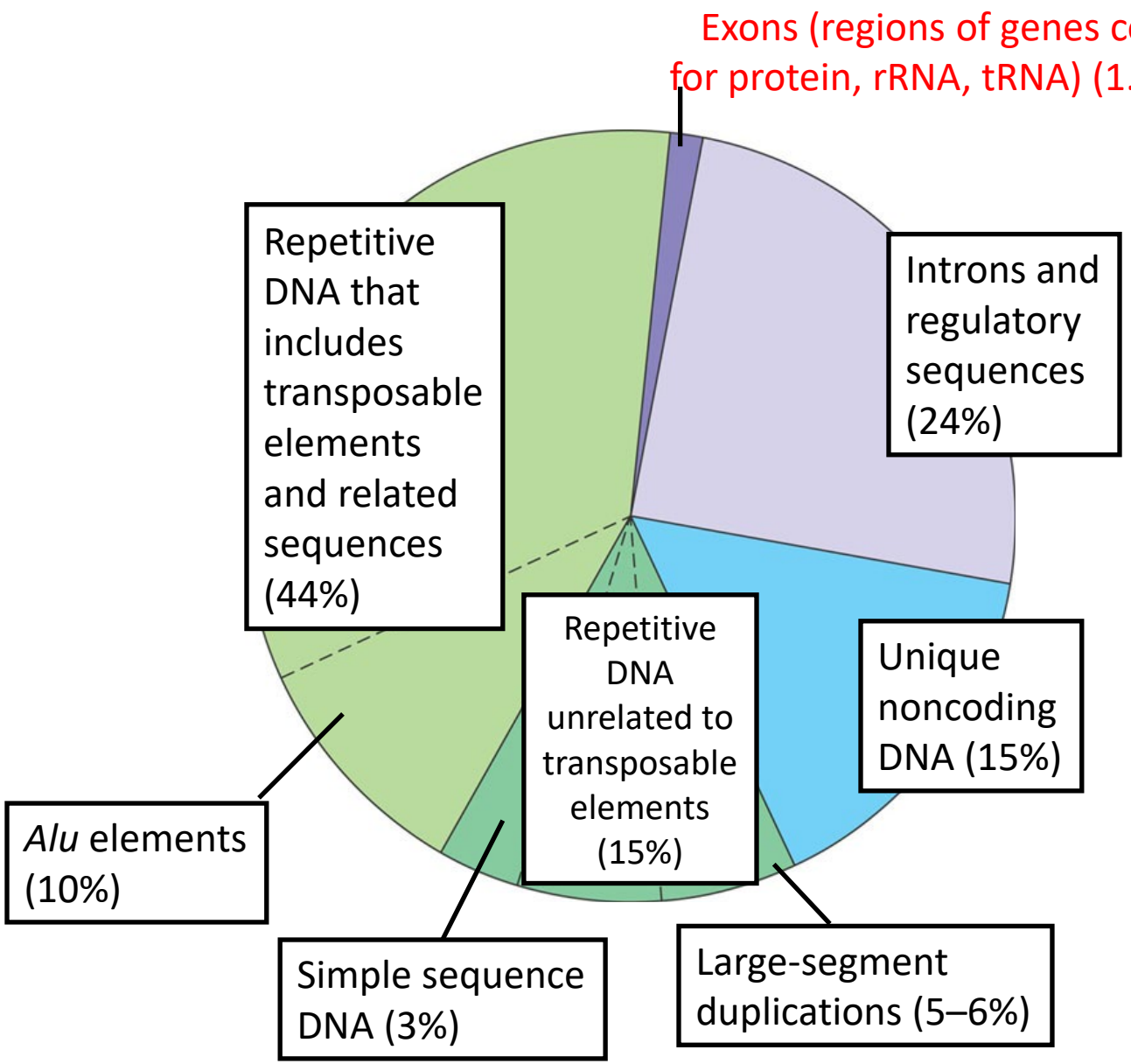
Genetic architecture of complex traits

- Where are variants found in the genome?
- What type of variants affect the trait?
- How much of trait variance is explained?

- Through which mechanisms do the variants exert their effects?
- How variations in pathway/network and molecular interactions contribute to phenotype

https://www.ebi.ac.uk/gwas/

Overview of the human genome



Majority of trait-associated variation is non-coding.

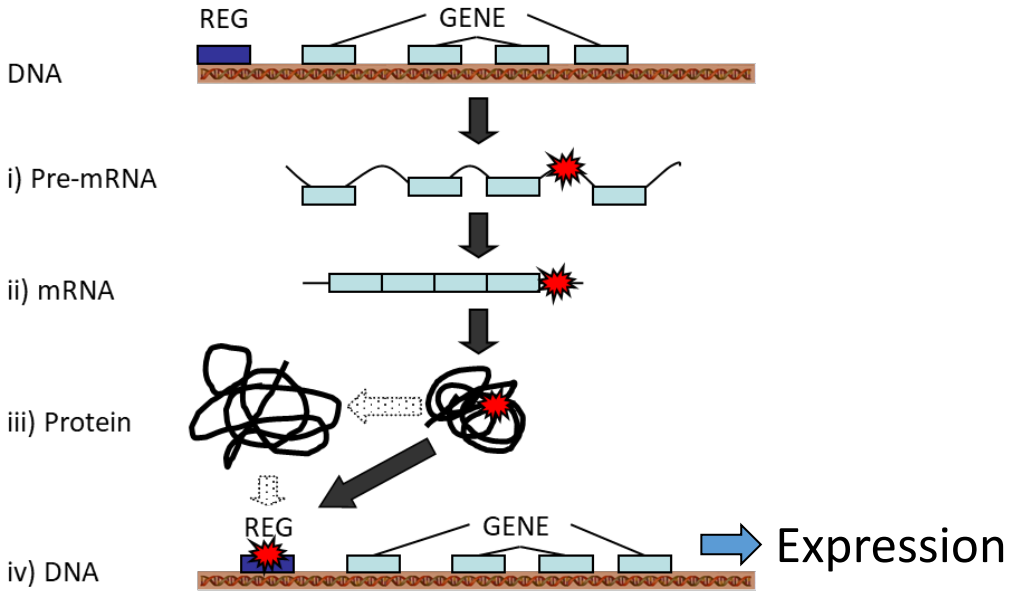
Hypothesis: most of these function by altering gene expression.

Regulatory variation

What do trait-associated variants do?

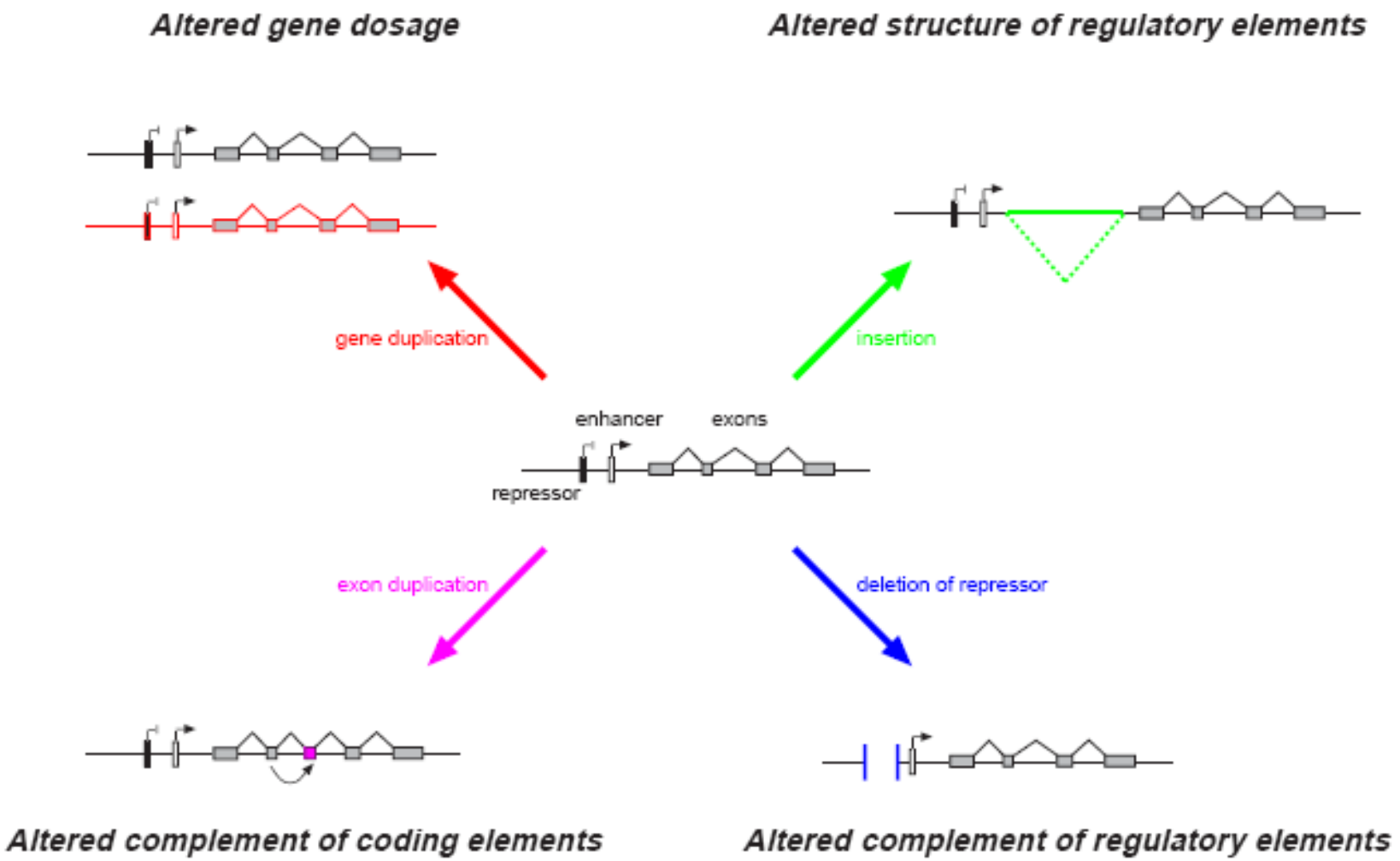
Genetic changes to:

- Coding sequence **
- Gene expression levels
- Splice isoform levels
- Methylation patterns
- Chromatin accessibility
- Transcription factor binding kinetics
- Cell signaling
- Protein-protein interactions



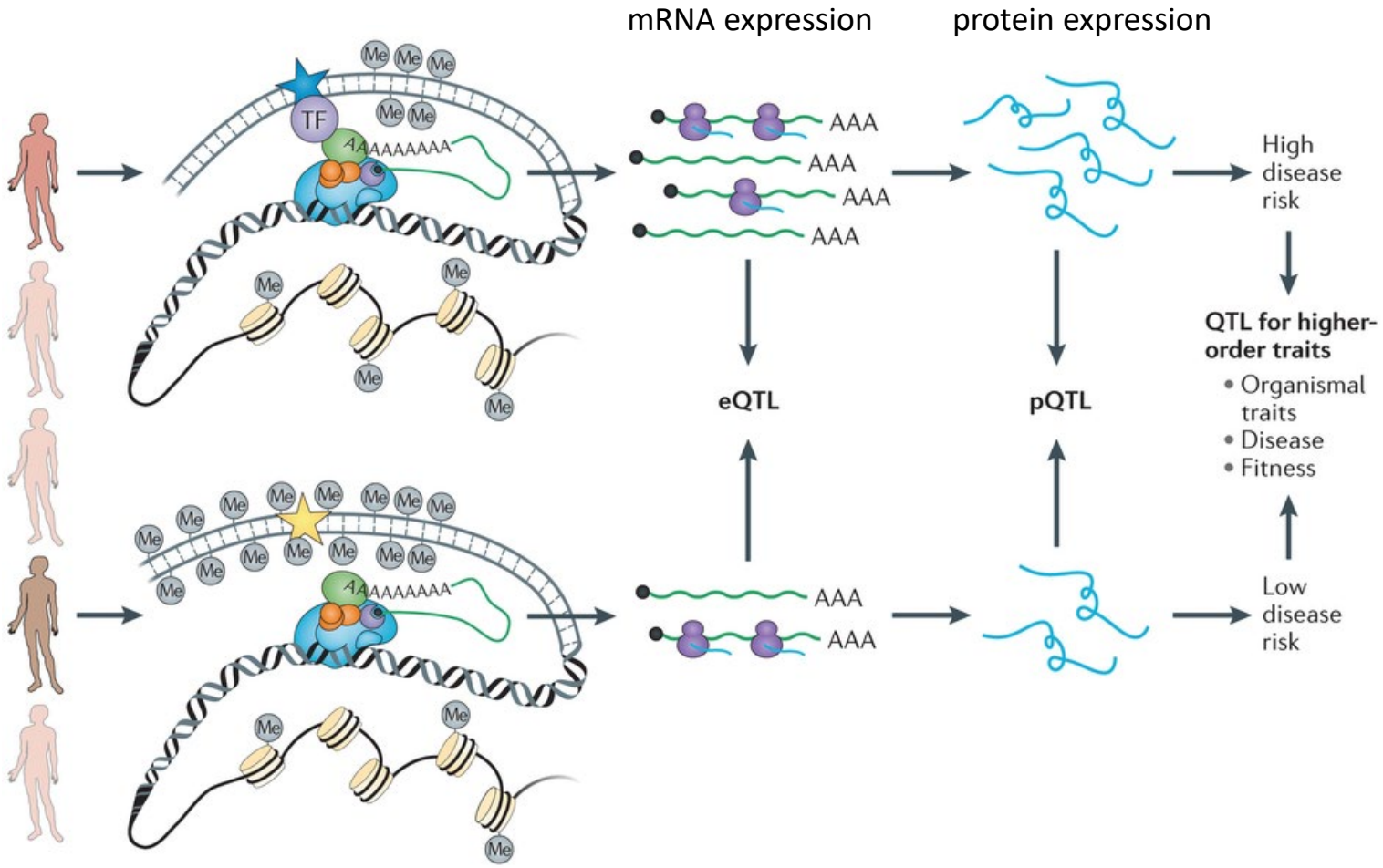
Stranger and Dermitzakis, *Human Genomics 2005*

Effects of copy number variation on gene expression



Hurles *et al.* 2008; *Trends In Genetics*

Understanding genetic-trait associations by exploring the biology that lies between the two



Albert and Kruglyak, 2015 *Nature Reviews Genetics*

Regulatory variation and gene expression

Altered patterns of gene expression → disease.

- e.g., Type 1 diabetes, Burkitt's lymphomas.

Widespread intraspecific variation.

Heritable genetic variation for transcript levels.

- Familial aggregation of expression profiles
- Median heritability 0.25 (Ouwens et al. 2019; *EJHG*)
- In humans, ~95% of protein-coding loci exhibited a genetic component for expression differences (GTEx Consortium 2020; *Science*)

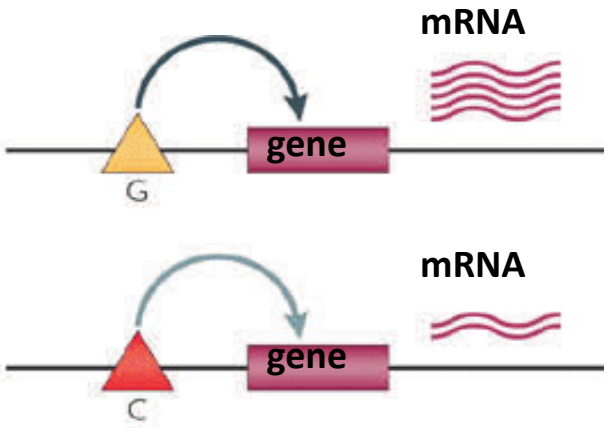
Much of the detected influential variation is located *cis*- to the coding locus.

- In humans, mouse, and maize, 25-60% of the genetic basis for intraspecific differences in transcription level are *cis*- to the coding locus
- More recent work in humans estimated ~35% heritable expression variation due to variants acting in cis-

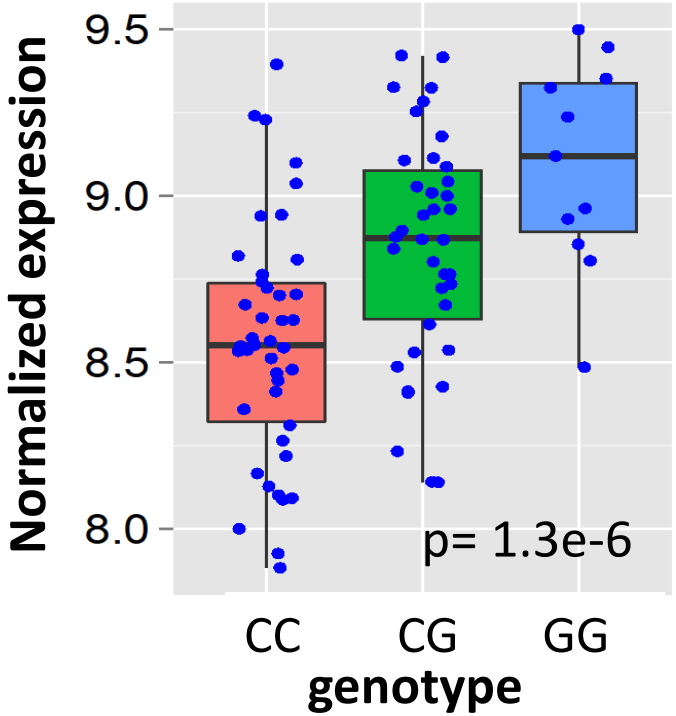
Some variants associated with disease also associate with gene expression variation

- Variants associated with Asthma, Rheumatoid arthritis, Crohn's disease, Bipolar Disorder , T1D, polygenic dyslipidaemia, lupus, blood lipid traits, etc. shown to affect gene expression.

Expression quantitative trait locus (eQTL) mapping: identify genomic regions affecting expression levels

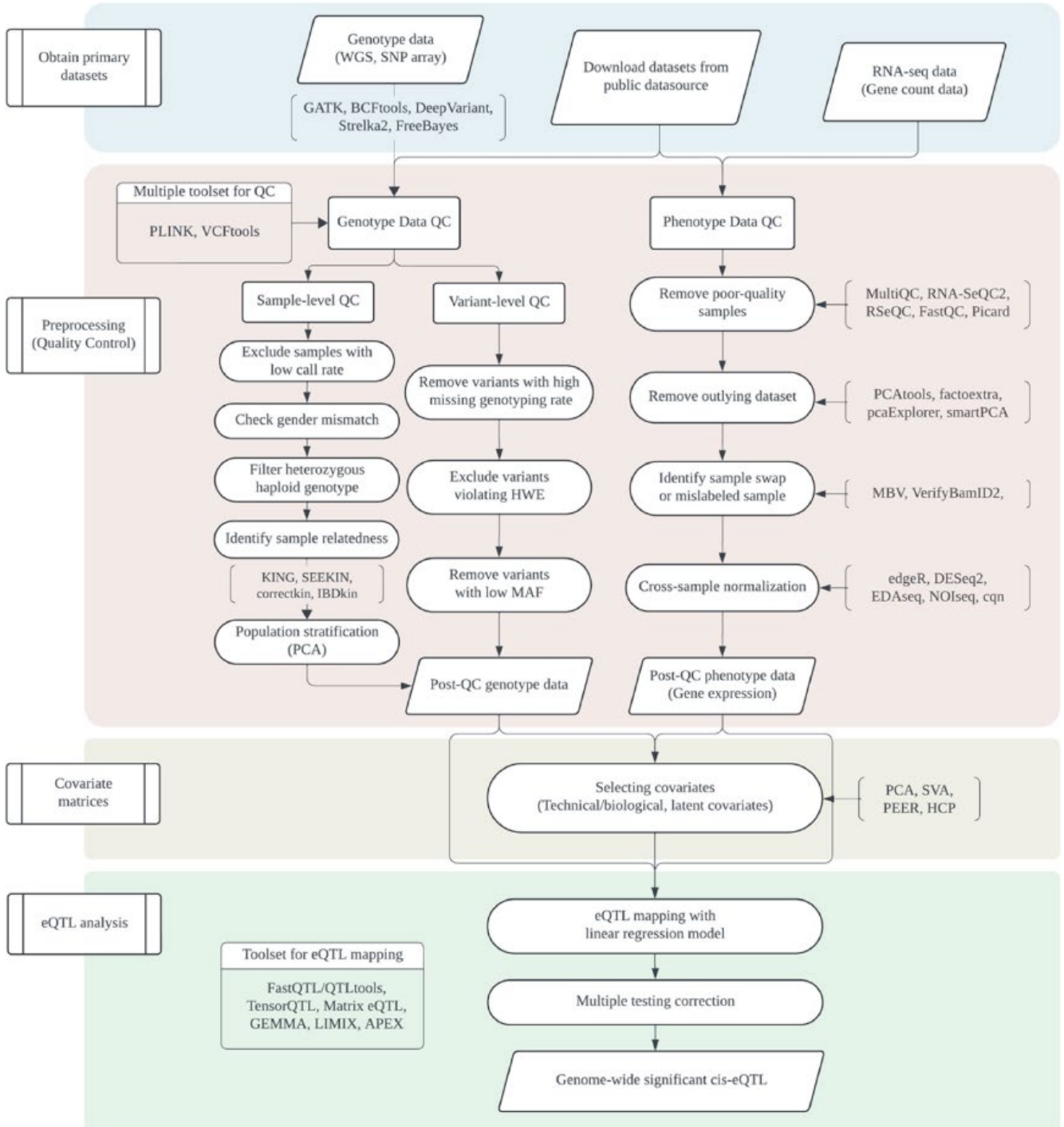


Population of individuals
Characterize transcriptomes
Characterize genomes
Perform statistical analysis



- p-value
- r^2 or rho
- Effect size: fold-change

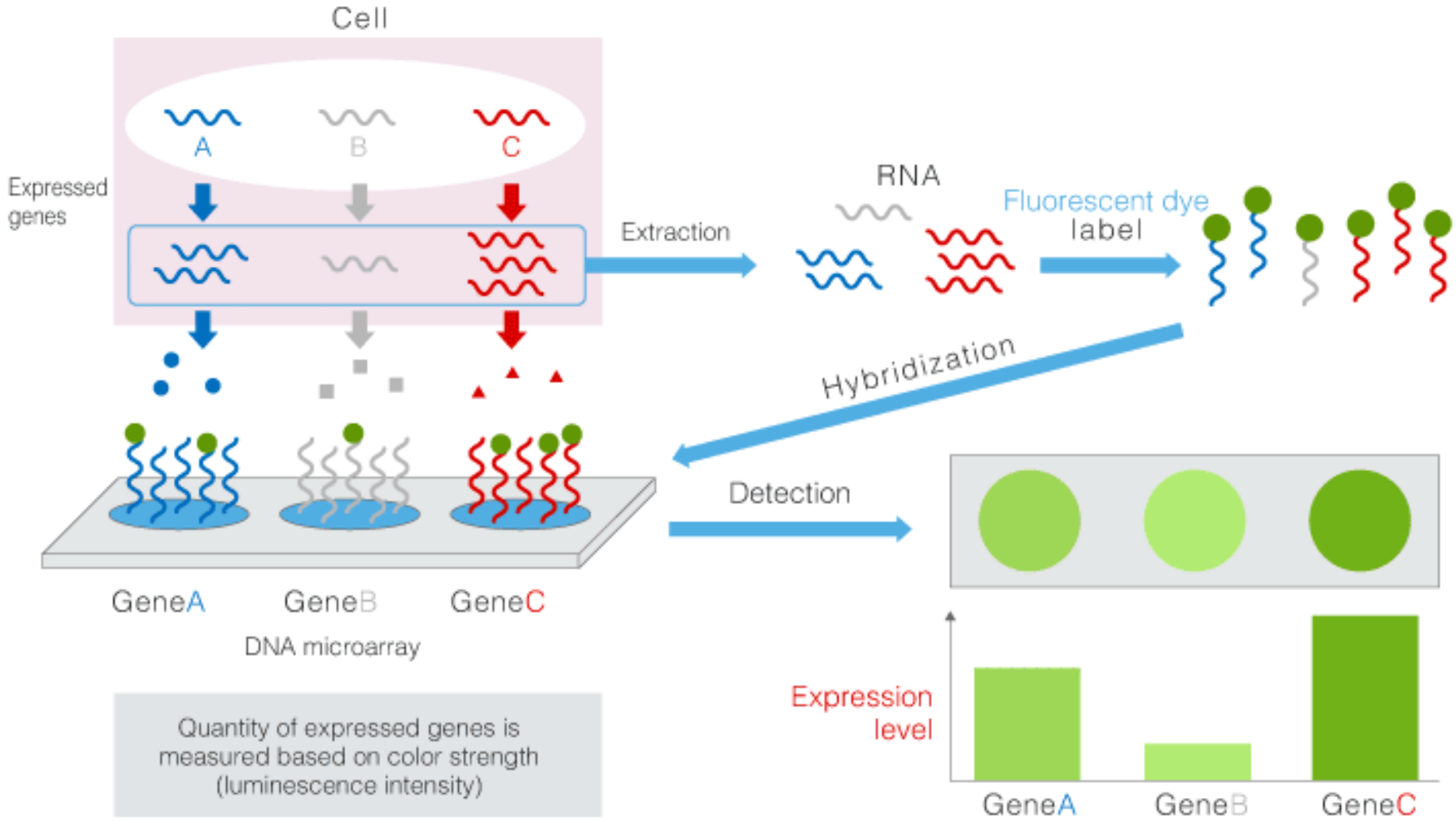
Multiple testing correction.



A brief guide to analyzing expression quantitative trait loci

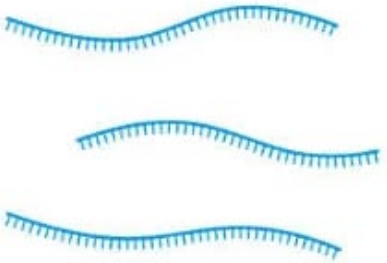
Ko *et al.* 2024, *Molecules and Cells*

mRNA quantification by array

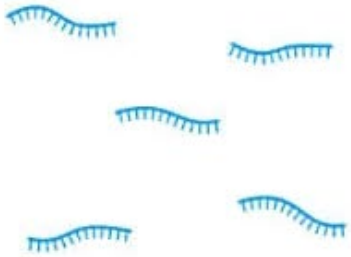


RNA Sequencing

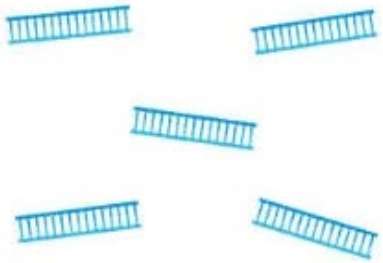
1 Isolate RNA from samples



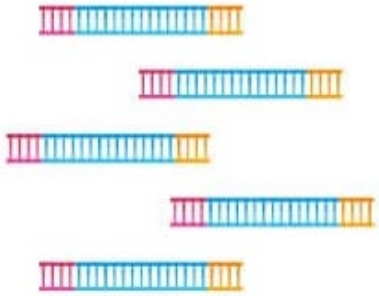
2 Fragment RNA into short segments



3 Convert RNA fragments into cDNA

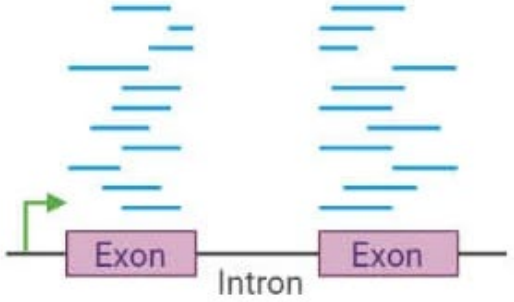


4 Ligate sequencing adapters and amplify



5 Perform NGS sequencing

6 Map sequencing reads to the transcriptome/genome



Very(!) general bulk RNA-seq workflow steps

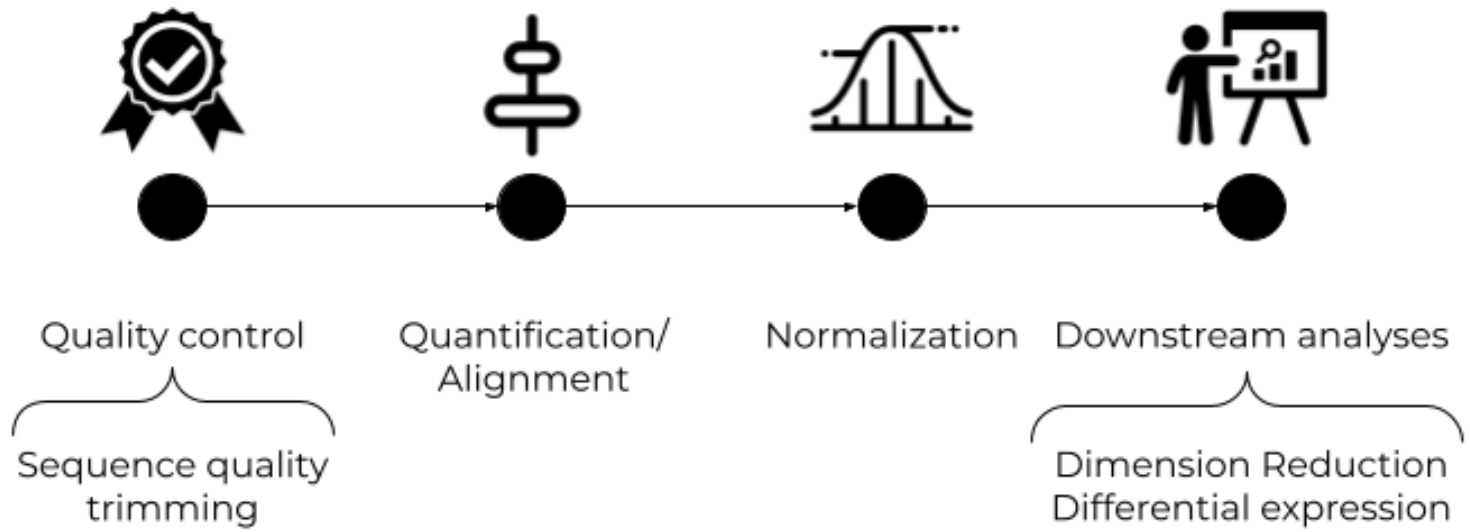


Image by Candace Savonen

RNA-seq Quality Control

countData

gene	ctrl_1	ctrl_2	exp_1	exp_1
geneA	10	11	56	45
geneB	0	0	128	54
geneC	42	41	59	41
geneD	103	122	1	23
geneE	10	23	14	56
geneF	0	1	2	0
…	…	…	…	…
…	…	…	…	…
…	…	…	…	…

Count matrix

Sample-Level QC

QC Metric	Recommended Cutoff	Reason
Read depth per sample	**>10 million mapped reads (preferably >20M)**	Ensures sufficient power for expression quantification
Uniquely mapped reads	**>70%**	Avoids samples with excessive multi-mapped reads
RIN score	**>6** (Remove degraded RNA samples)	Ensures RNA integrity
Gene body coverage	**Even distribution (avoid extreme 3′ bias)**	Ensures non-degraded transcripts
PCA/MDS outlier detection	**Remove samples >6 SD from the mean**	Identifies batch effects or outlier samples
Sex check	**Match XIST (high in females) and Y-linked gene expression with reported sex**	Detects mislabeled samples

Trimmed Mean of M-values (TMM) normalization

Goal: normalize RNA-seq count data across samples.

Adjusts for sequencing depth and RNA composition bias.

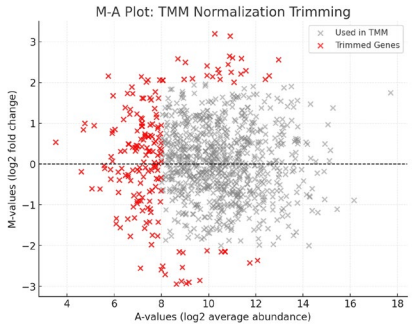
1. **Compute M-values and A-values**
   - **M-value (log-fold change):**

$$M_i = \log_2 \left(\frac{X_i}{X_i'} \right)$$

   where X_i is the count for gene i in the sample of interest and X_i' is the count for the same gene in a reference sample.

   - **A-value (average abundance):**

$$A_i = \frac{1}{2} \log_2 (X_i \times X_i')$$



2. **Trim Extreme M-values and Low-Expression Genes**
   - Removes highly differentially expressed genes that could skew normalization.
3. **Calculate a Weighted Mean**
   - Uses the remaining M-values to compute a scaling factor.
4. **Apply the Scaling Factor**
   - Adjusts raw counts for more accurate cross-sample comparisons.

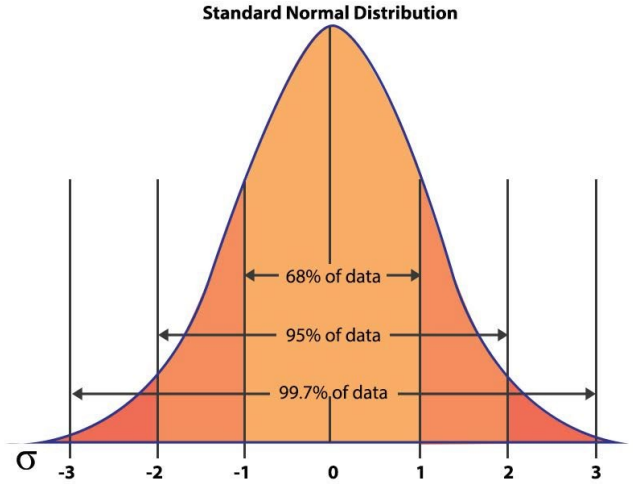
Rank-based inverse normal transformation (INT)

eQTL methods assume normally distributed expression values

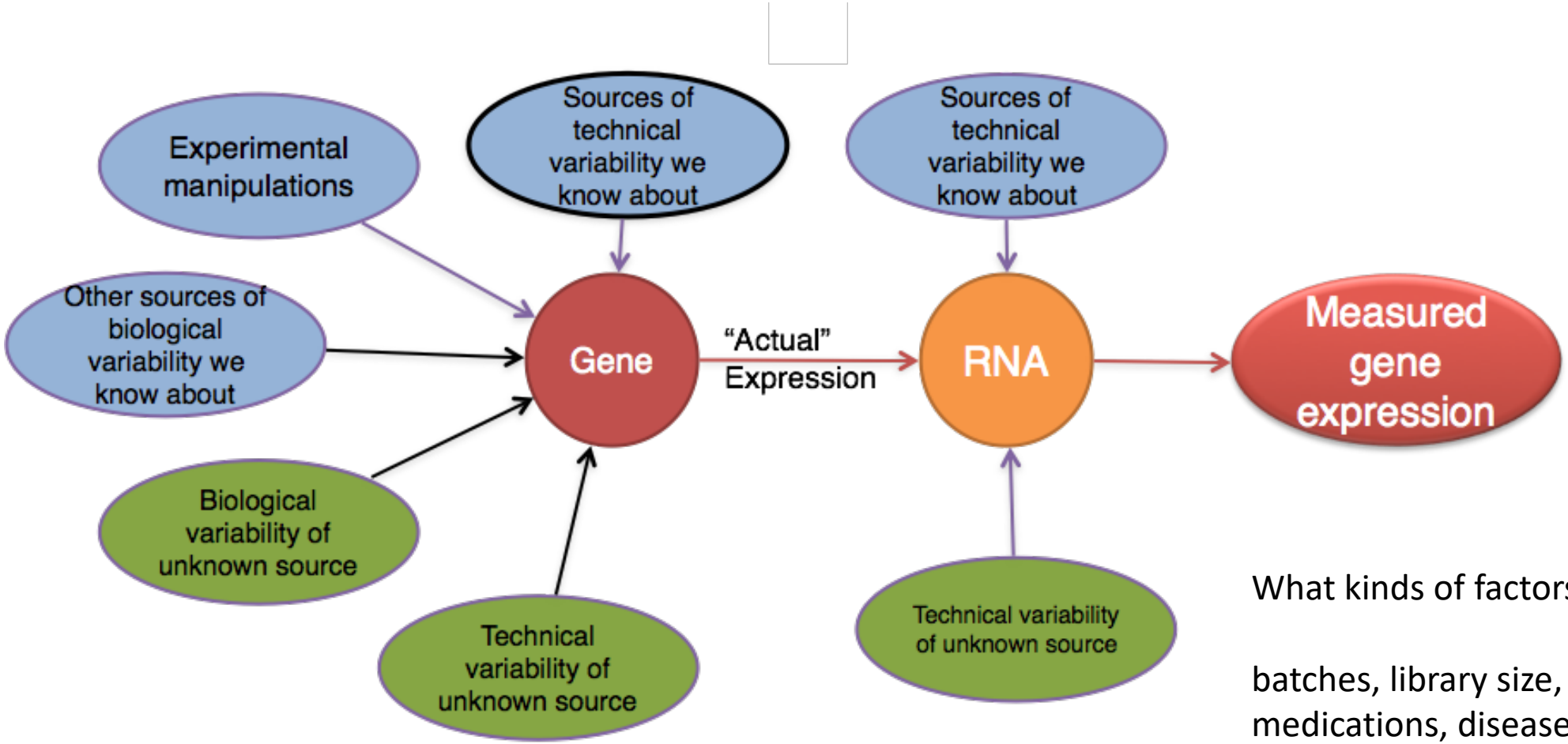
- RNA-seq counts = negative binomial or Poisson distribution →violates assumptions.
- Transformation of count data = suitable for parametric statistical models.

Typical Transformation (per gene):

Rank-based inverse normal transformation (INT), common in eQTL analysis, biobank quantitative traits



Factors influencing gene expression: known & unknown

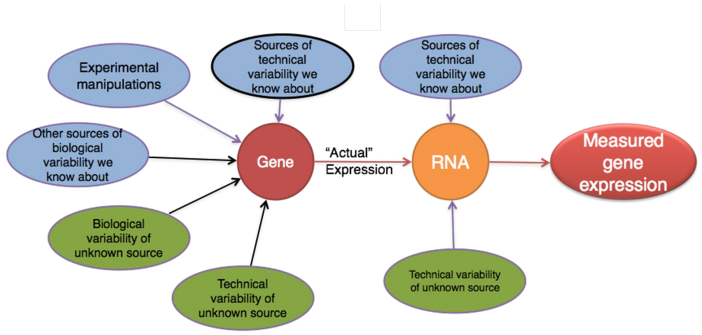


What kinds of factors?

batches, library size, RNA quality, medications, disease, infection, age, sex, etc.

Courtesy of Paul Pavlidis, UBC

Controlling for sources of variation



Why?

- trying to link genetic variation to gene expression
- known and hidden factors can obscure real genetic effects
- reduce false positives and improve statistical power

How?

Include covariates in linear regression model (known and unknown)

- typical known: age + sex + genotype principal components
- Hidden: inferred factors, learned from gene expression
  - Principal components
  - PEER (Probabilistic Estimation of Expression Residuals)

PEER (Probabilistic Estimation of Expression Residuals)

- **Bayesian method** that identifies hidden confounders (PEER factors) by modeling them as **latent variables**
- PEER factors can help remove **unwanted variation** from technical artifacts (e.g., batch effects, RNA degradation) or biological differences (e.g., cell composition) that aren't directly measured.

- Pros: Effective, both technical & biological factors
- Cons: Slow, Computationally expensive

Stegle et al. 2012, *Nature Protocols*

PEER

Core idea of PEER is to represent gene expression as:

$$y_i = X_i\beta + Z_i f + \epsilon_i$$

Where:

- y_i is the vector of gene expression for sample i,
- X_i is the design matrix of observed covariates (e.g., treatment or clinical variables),
- β is the vector of coefficients for observed variables,
- Z_i links latent factors to gene expression,
- f represents the latent factors (hidden variables),
- ϵ_i is the residual error term (unexplained variation).

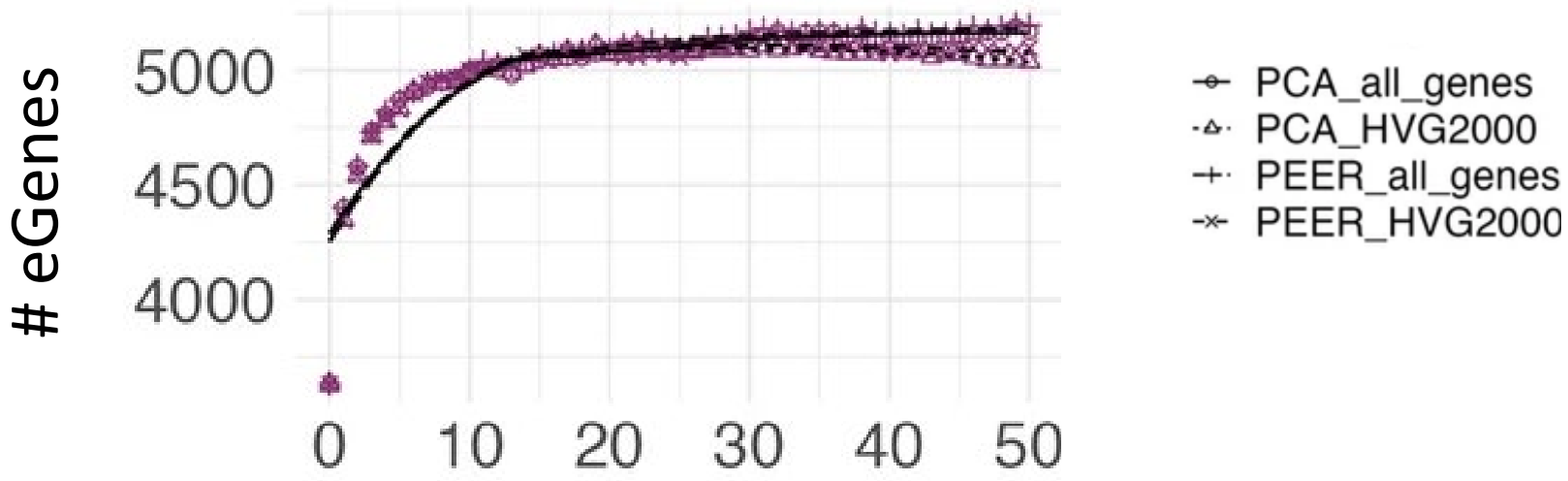
PEER infers the latent factors *f* that explain the unexplained variation in gene expression. These factors are estimated through Bayesian methods

PEER factors are then used as covariates in the eQTL model.

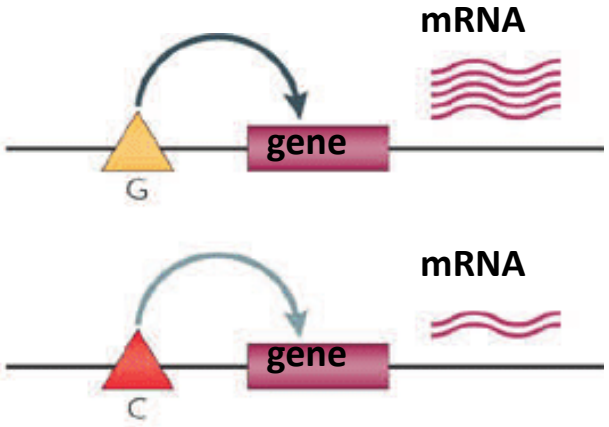
Stegle et al. 2012, *Nature Protocols*

Considerations for PEER

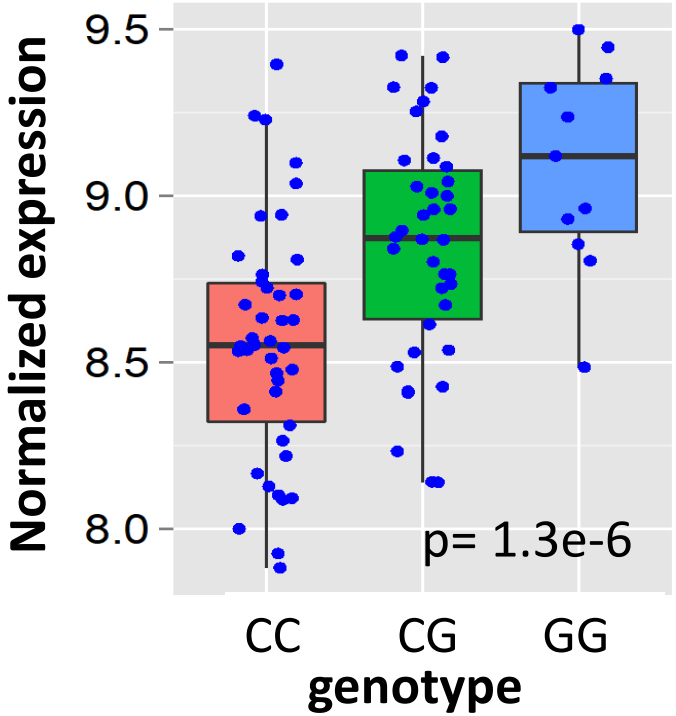
- When estimating, include known covariates (e.g., age, sex, batch)
- How many to include as covariates in eQTL model? Empirical estimation: minimize covariates, maximize discovery



Expression quantitative trait locus (eQTL) mapping: to identify genomic regions affecting expression levels



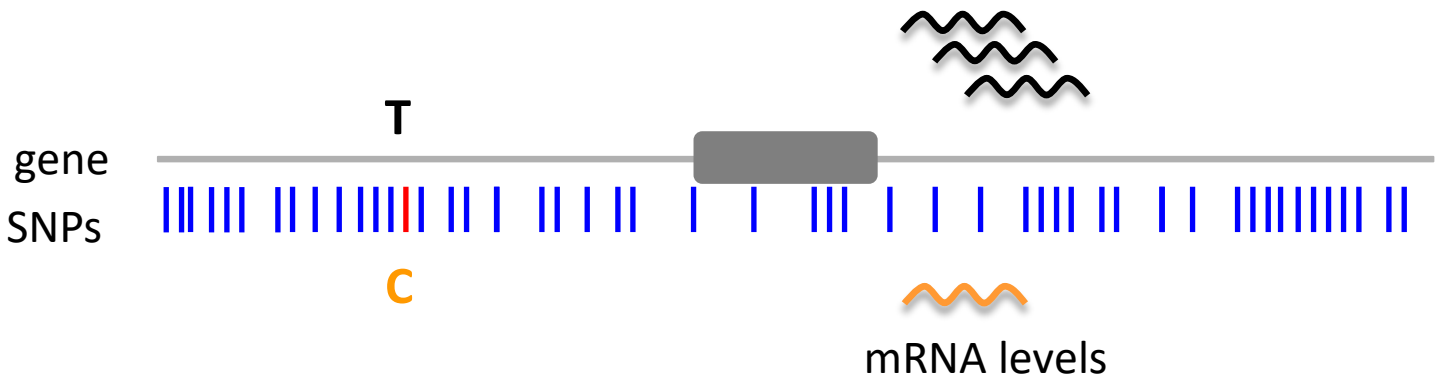
Population of individuals
Characterize transcriptomes
Characterize genomes
Perform statistical analysis



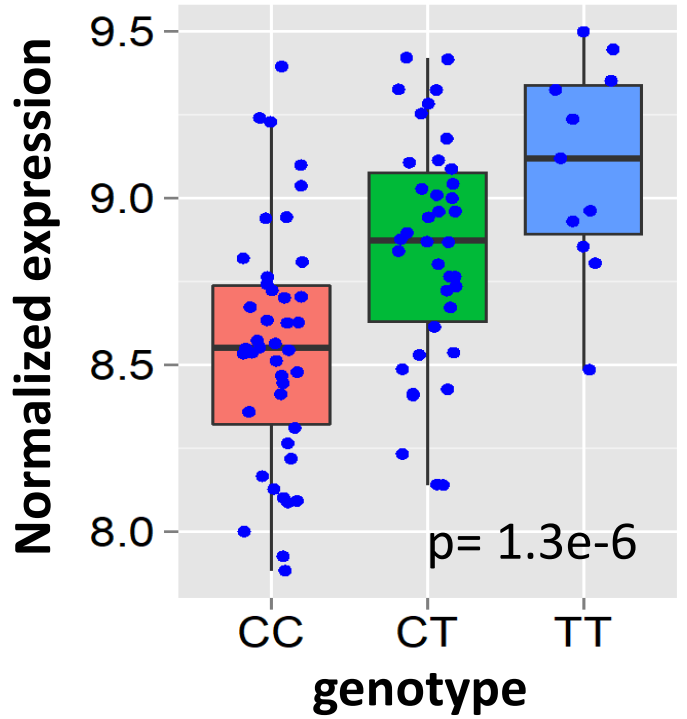
- p-value
- r^2 or rho
- Effect size: fold-change

Multiple testing correction.

SNP - expression association analysis

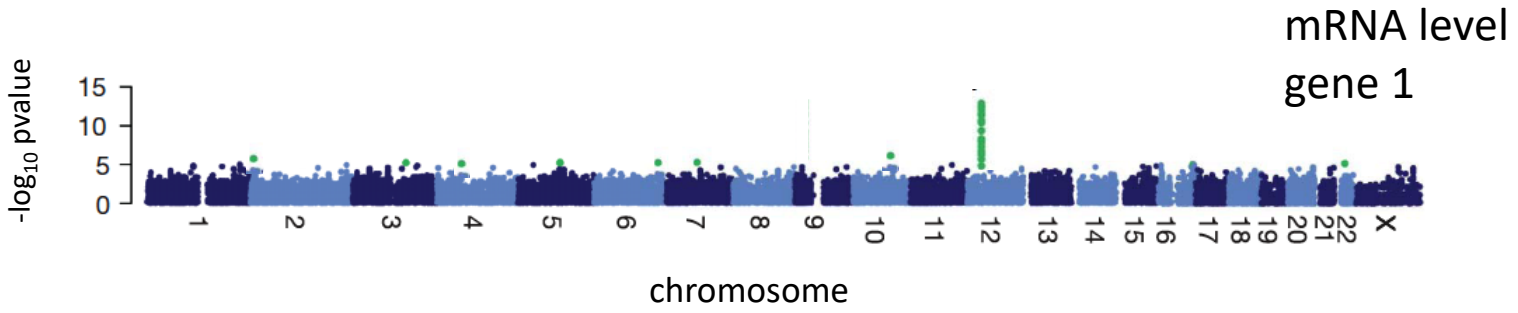


Model: $Y_i \sim \beta_0 + \beta_2 \text{Genotype} + \beta_{(1\text{-}n)} \text{Covs} + \beta_{(1\text{-}m)} \text{PEERS} + \varepsilon$



- p-value
- r^2 or rho
- Effect size: fold-change

Whole-genome eQTL analysis is a GWAS for expression of a gene

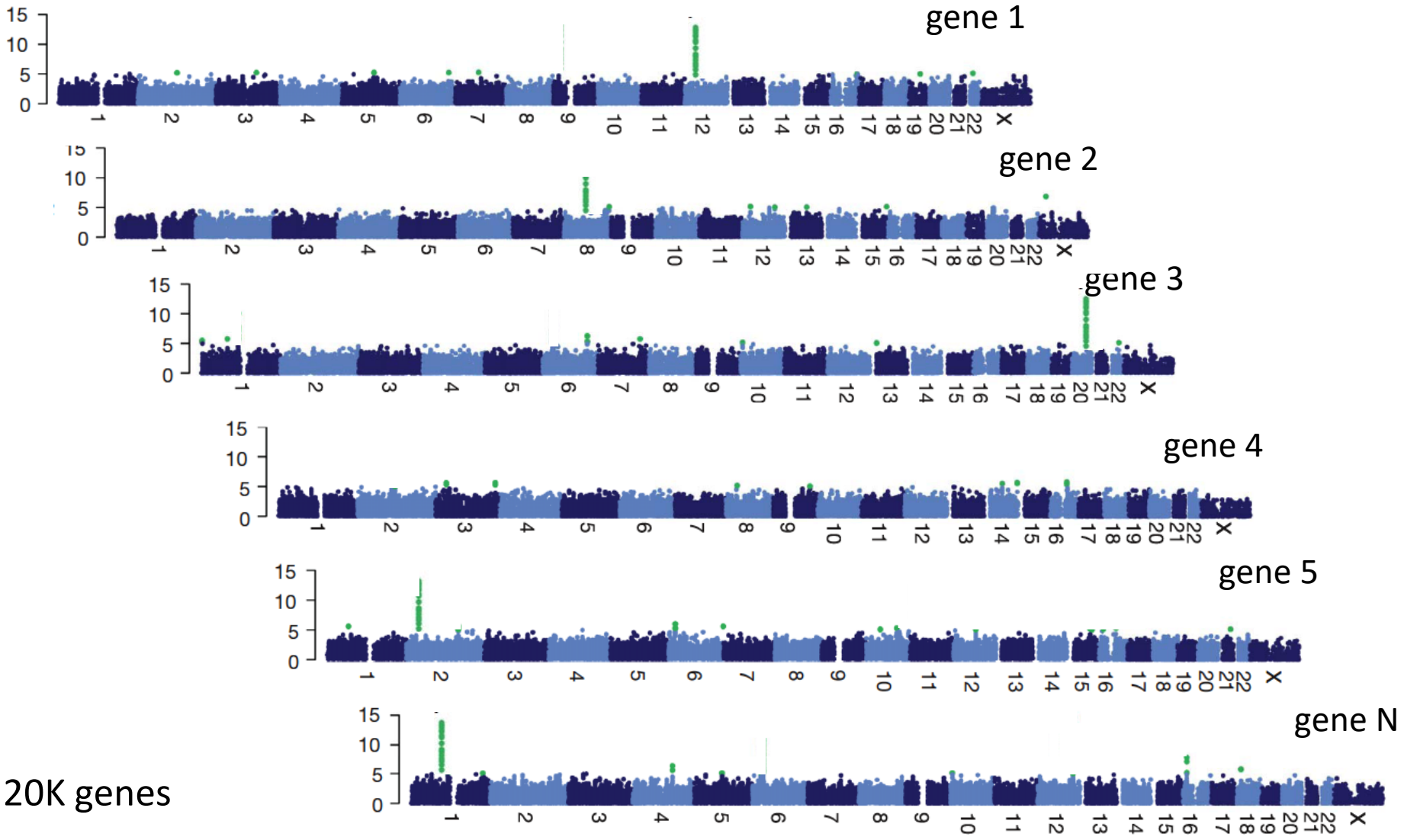


Considerations: Same as for GWAS for complex traits / disease

SNP Quality Control (e.g., missingness, HW equilibrium)
Well-quantified phenotype (robust gene expression measurements)
Population stratification
Statistical considerations (MAF, power, multiple testing, covariates)

Interpretation of significant hits:
Significantly associated variant tags true causal variant

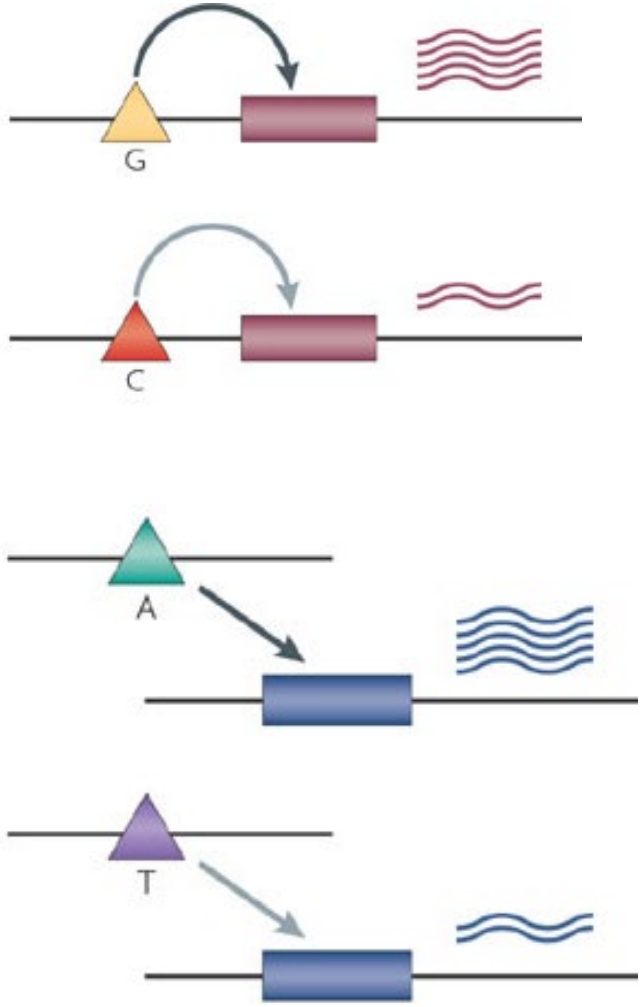
Whole-genome eQTL analysis is an independent GWAS for expression of each gene



e.g., 3M SNPs x 20K genes

Terminology

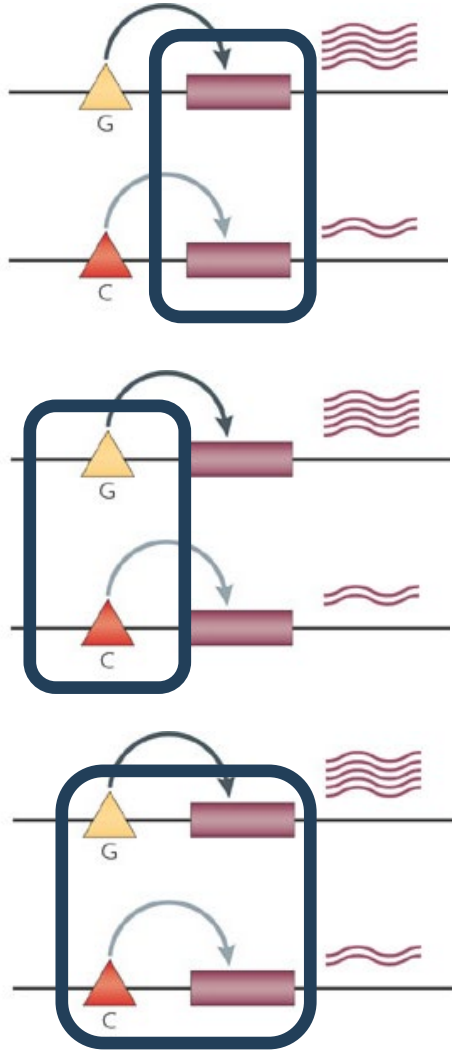
- *cis*-eQTL
  - The position of the eQTL variant maps near the physical position of the gene.
  - Promoter polymorphism?
  - Insertion/Deletion?
  - Methylation, chromatin conformation?
- *trans*-eQTL
  - The position of the eQTL variant does not map near the physical position of the gene.
  - Regulator?
  - Direct or indirect?



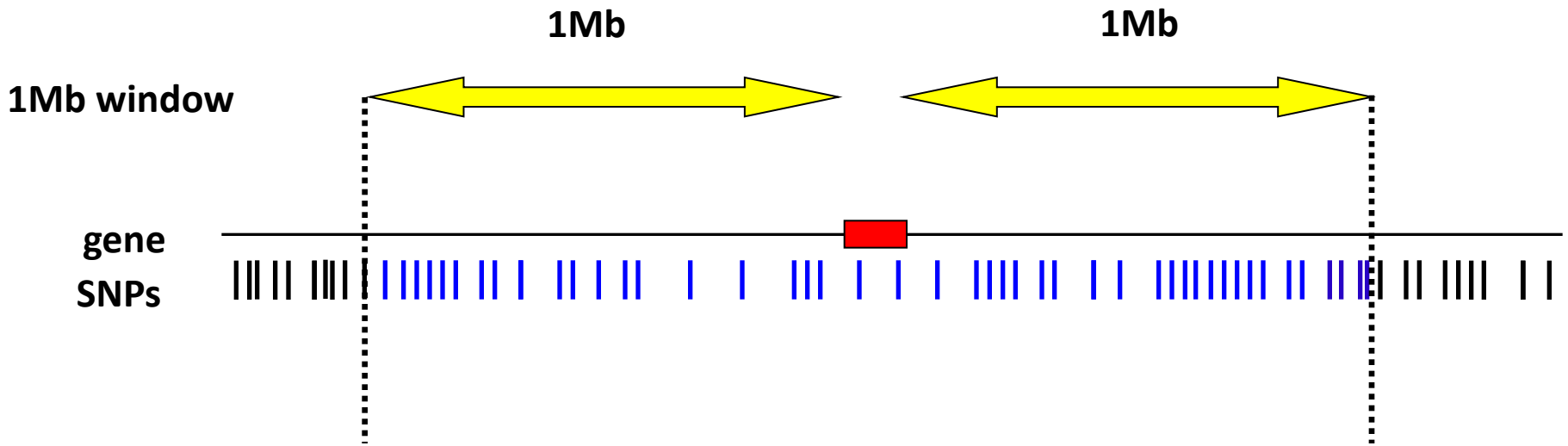
Cheung and Spielman 2009 *Nature Genetics*

More terminology

- eGene: A gene with ≥1 significantly associated SNP
- eSNP/eVariant: A SNP/variant associated with ≥1 eGene
- eQTL: A SNP-gene pair where genetic variation is associated with gene expression association, sometimes used synonymously with eSNP/eVariant



Cis- eQTL analysis:
test SNPs within a pre-defined distance of gene



Model: $Y_i \sim \boldsymbol{\beta}_0 + \boldsymbol{\beta}_2 \text{Genotype} + \boldsymbol{\beta}_{(1\text{-}n)} \text{Covs} + \boldsymbol{\beta}_{(1\text{-}m)} \text{PEERS} + \varepsilon$

probabilistic estimation of expression residuals (PEER): Stegle *et al*. 2012 *Nature Protocols*
FastQTL: Ongen *et al.* 2016 *Bioinformatics*
MatrixQTL: *Shabalin et al.* 2012 *Bioinformatics*

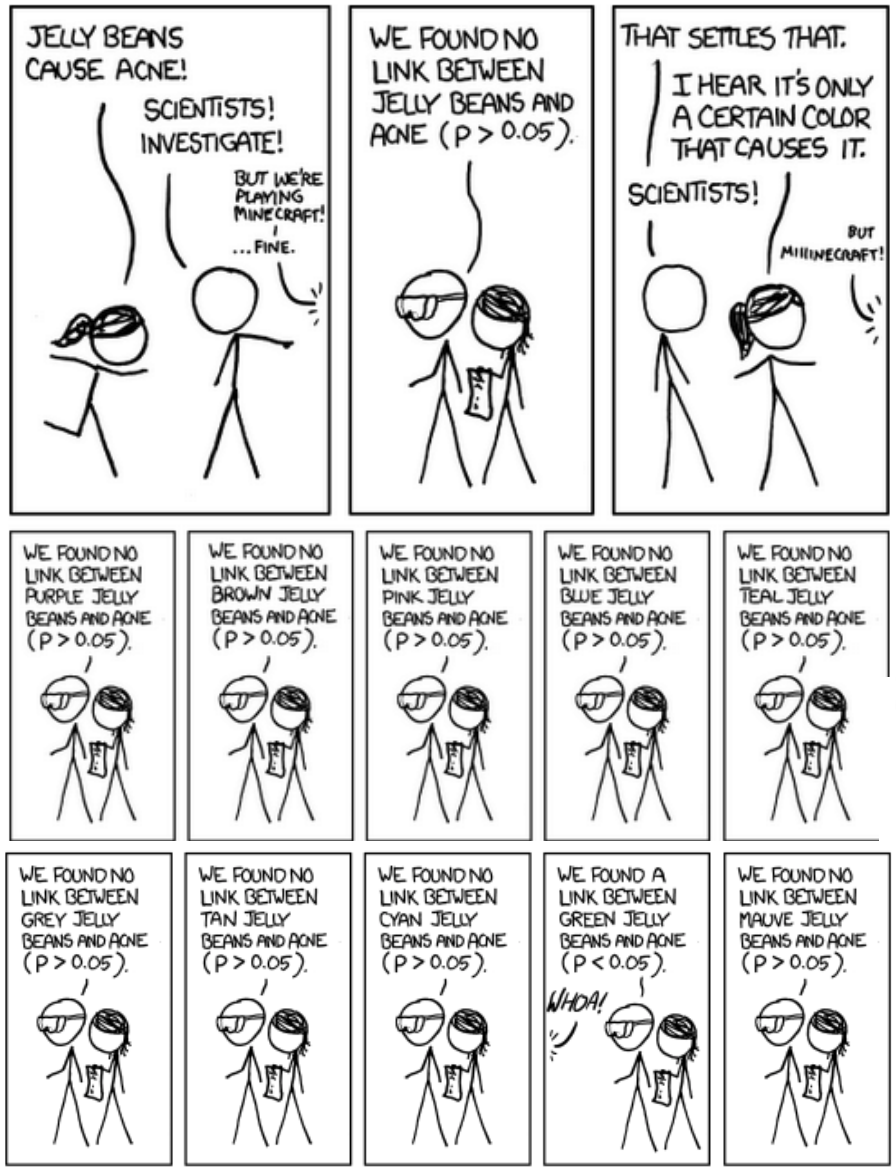
Multiple testing correction

1. Testing many SNPs for association with each gene.
2. Testing many genes for association with each SNP.

GOAL: control the false discovery rate (FDR)

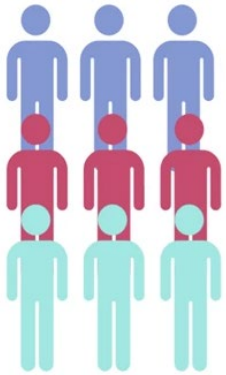
Control FDR: Benjamini-Hochberg (BH) FDR or Storey's q-value method.

permutations: (shuffling expression phenotypes, within gene)

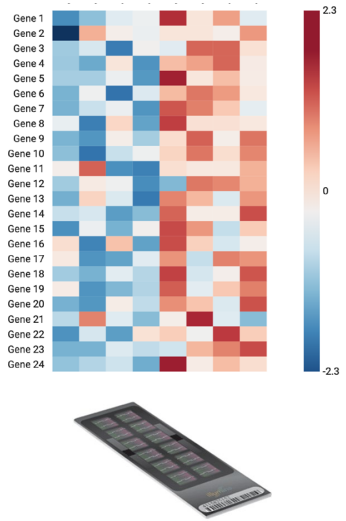


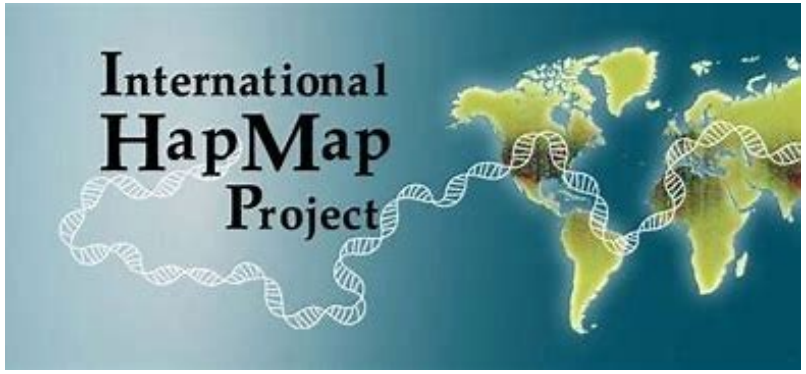
http://xkcd.com/882/

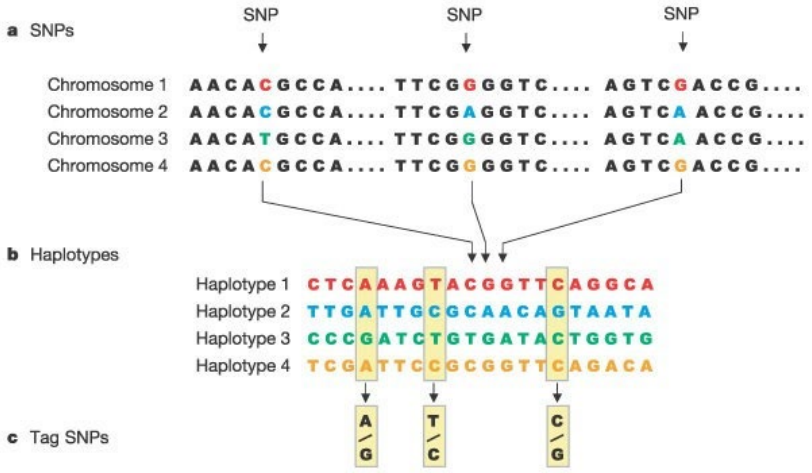
Early eQTL mapping



Lymphoblastoid cell lines (LCLs)







CEU: 109 Caucasians living in Utah USA, of northern and western European ancestry
CHB: 80 Han Chinese from Beijing, China
GIH: 82 Gujarati Indians in Houston, TX, USA
JPT: 82 Japanese in Tokyo, Japan
LWK: 82 Luhya in Webuye, Kenya
MEX: 45 Mexican ancestry in Los Angeles, CA, USA,
MKK: 138 Maasai in Kinyawa, Kenya
YRI: 108 Yoruba in Ibadan, Nigeria

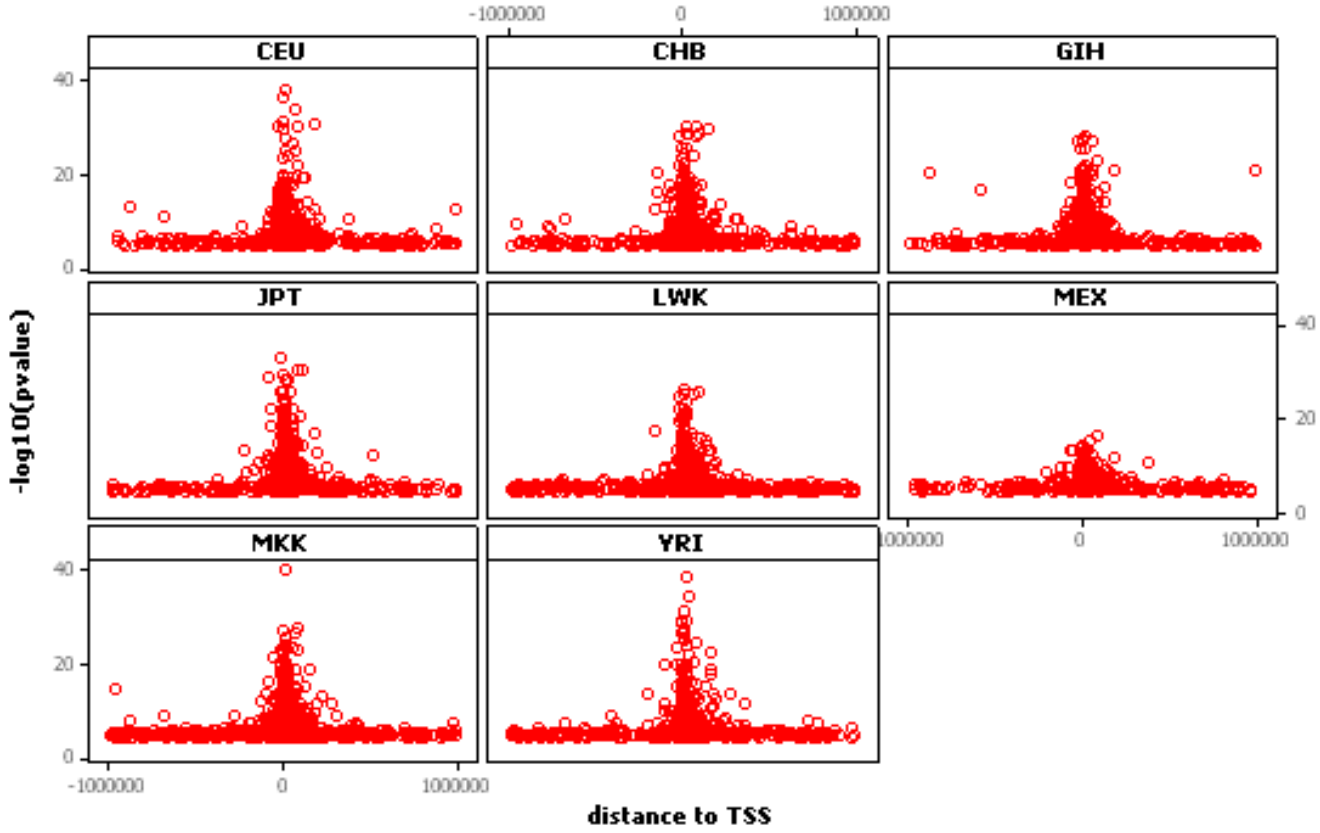
Early cis-eQTL findings

significant associations are symmetrically distributed around TSS, strongest at TSS

Denser SNP maps → better resolution

cross-population meta-analysis increased discovery

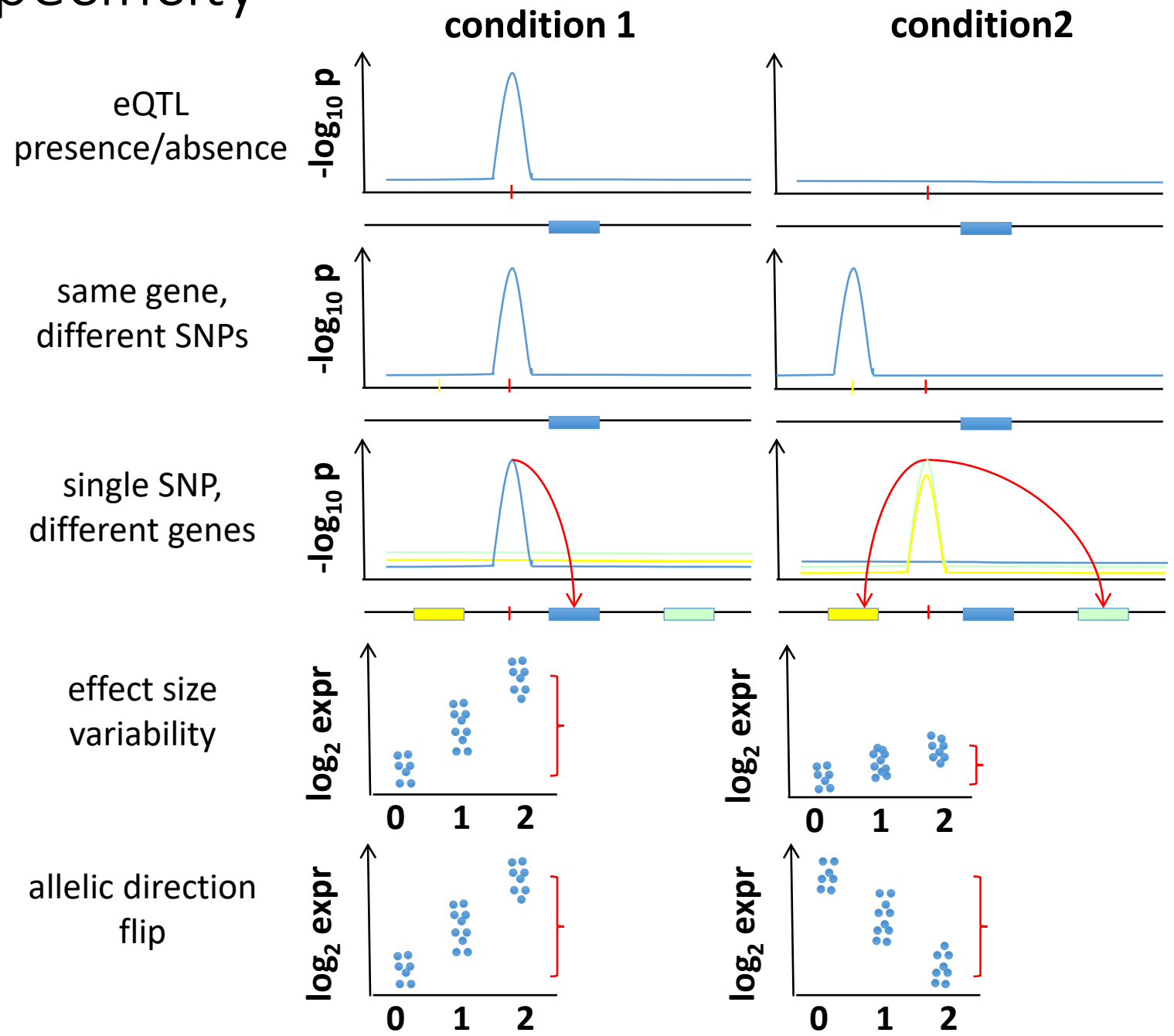
For population-shared eQTLs, effect sizes and direction largely similar across populations



HapMap LCLs cis-eQTLs
Most significant SNP per gene

Stranger *et al.*, 2012 *PLoS Genetics*

Context specificity



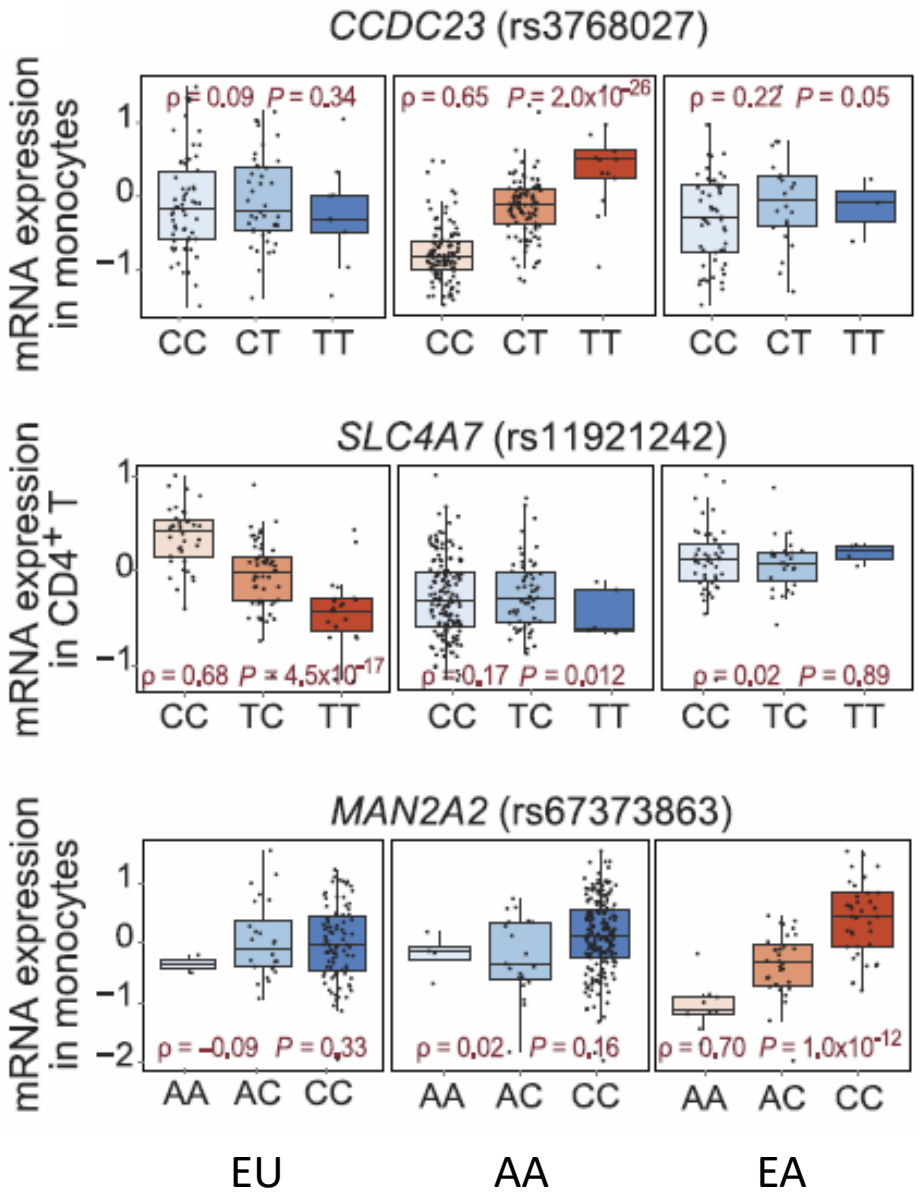
Contexts of interest

- Populations of different ancestry (HapMap, Geuvadis)
- Tissue or cell type (primary blood cells, fat, skin)
- Baseline vs stimulated
- Males vs females
- Differential with respect to age
- Impacted by environment (exogenous, endogenous, pertubations)

Population-specific cis-eQTLs (4-6%)

Little population-specificity of presence/absence or fold-change modulation

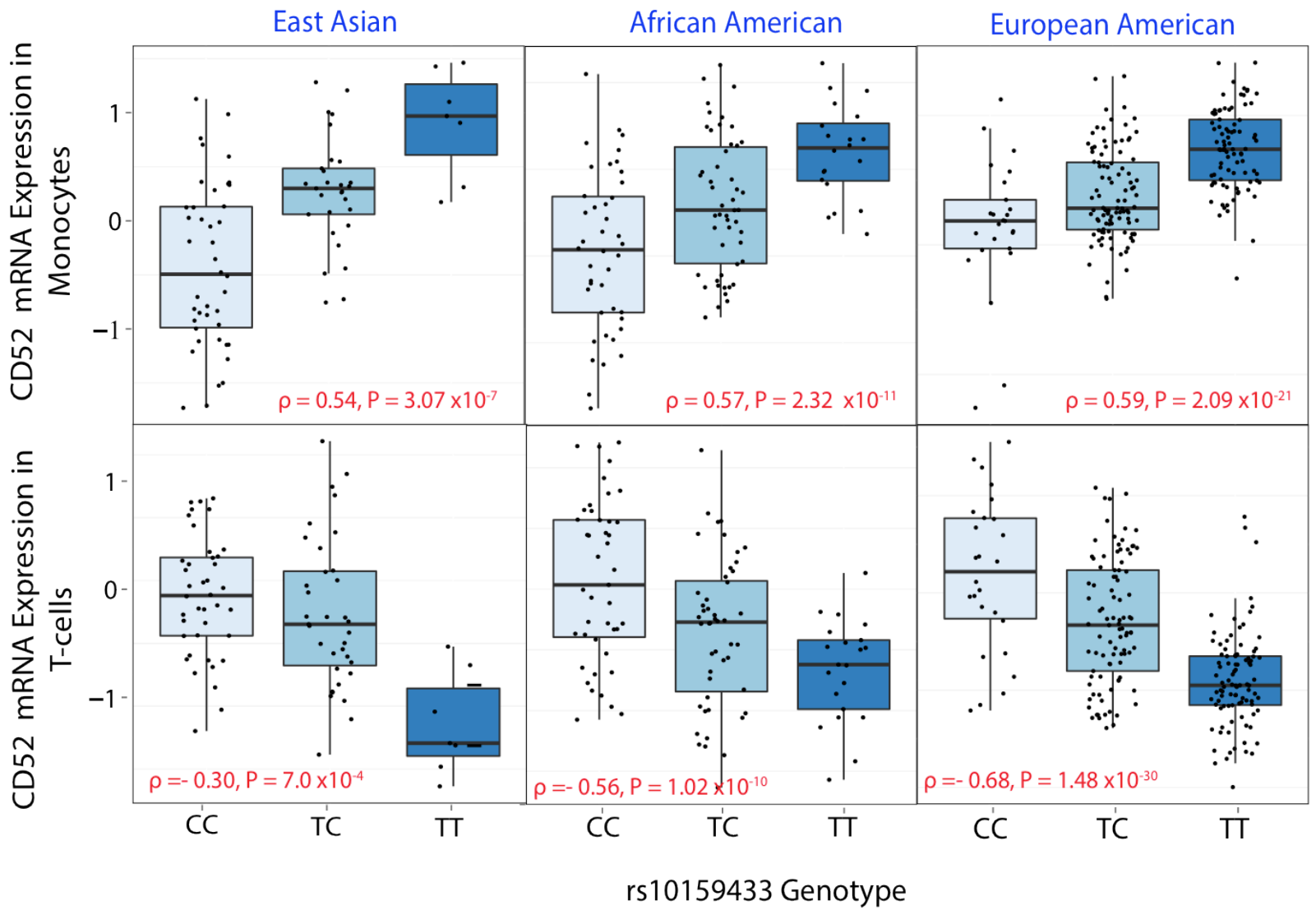
BUT those differences might be **highly relevant** for population-diverged phenotypes



Raj *et al.*, 2014 *Science*

CD52 *cis*-eQTL shows directional regulatory effects across cell types

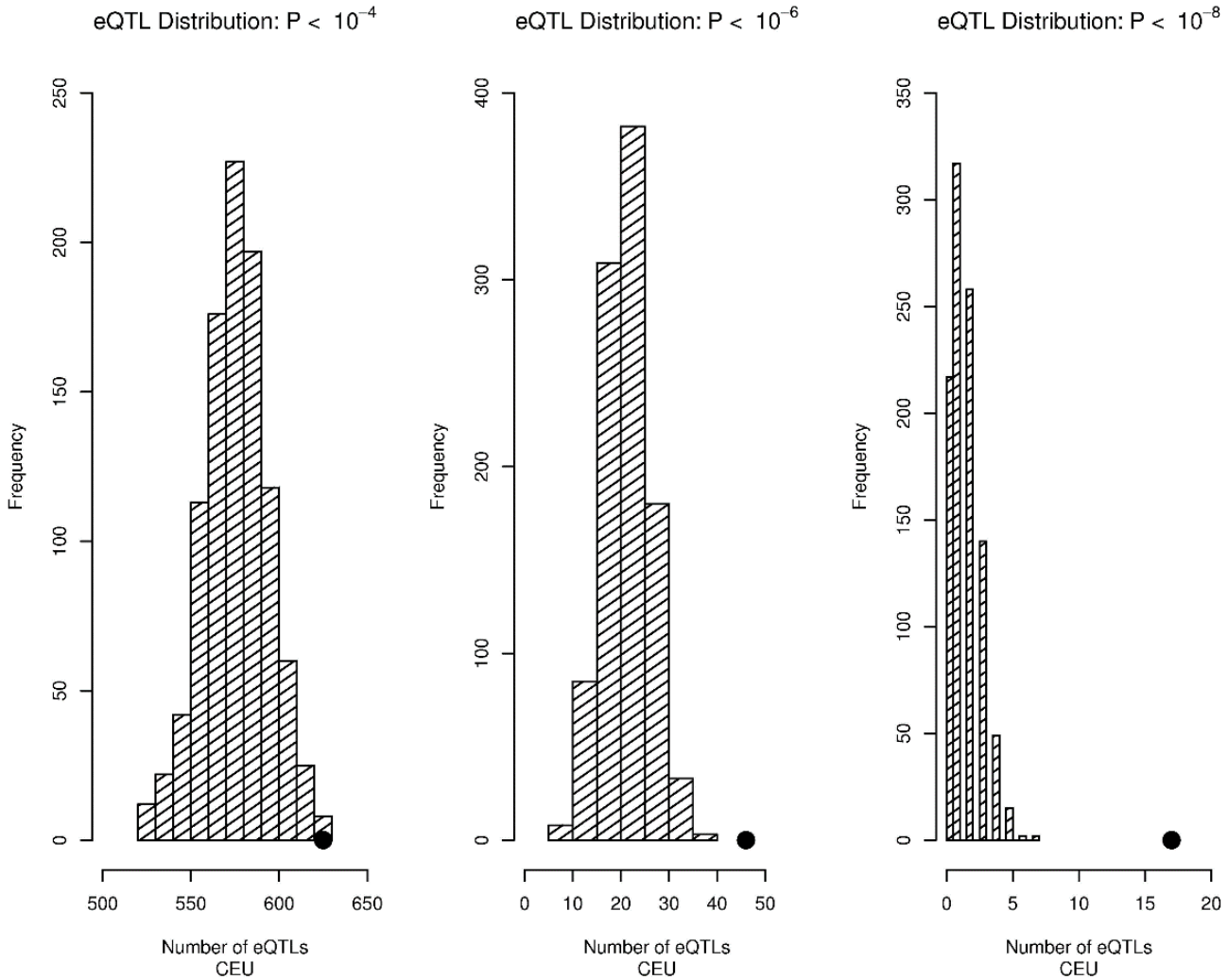
40% cis-eQTLs were cell-type specific



ImmVar

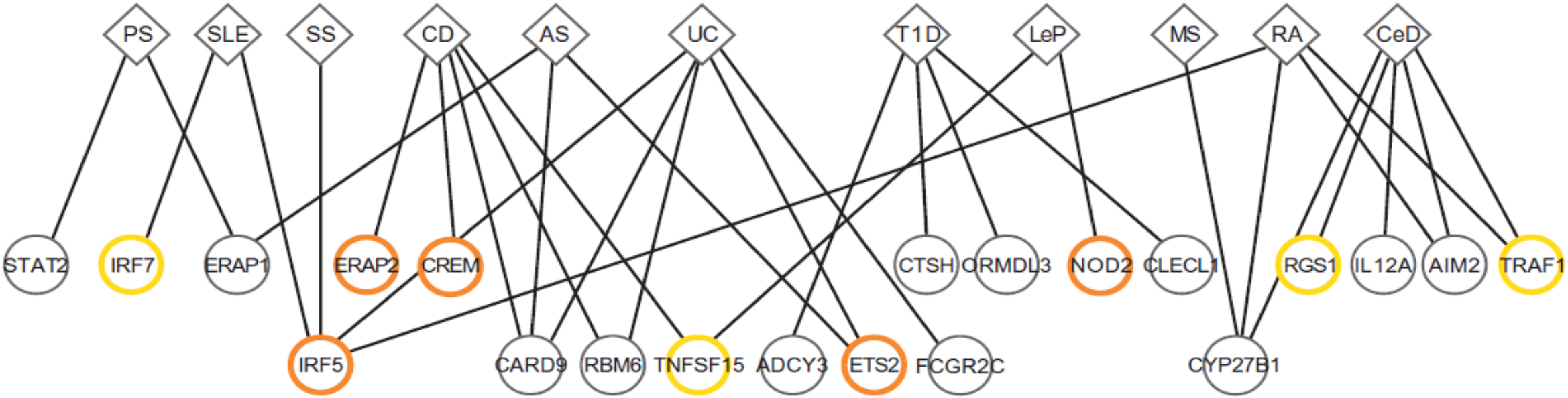
CD52 lymphocyte cell-surface glycoprotein, function in anti-adhesion, role in lymphoma. It is the protein targeted by alemtuzumab, a monoclonal antibody used for the treatment of chronic lymphocytic leukemia

Trait-associated SNPs are more likely to be eQTLs



Nicolae *et al.,* 2010, *PLoS Genetics*

Autoimmune and Infectious disease SNPs from GWAS are eQTLs



PS = Psoriasis
SLE = Systemic lupus erythematosus
SS = Systemic sclerosis
CD = Crohn's disease
AS = Ankylosing spondylitis
UC = Ulcerative colitis

T1D = Type 1 diabetes
LeP = Leprosy
MS = Multiple sclerosis
RA = Rheumatoid arthritis
CeD = Celiac's disease

orange: cis-reQTLs, yellow: stimulus-specific cis-eQTLs

Lee *et al.* 2014 *Science*

Geuvadis

- First large-scale transcriptome sequencing eQTL study
- First large-scale eQTL study with DNA sequencing data (1000genomes)
- LCLs from 462 individuals from multiple human populations
- microRNAs, mRNAs

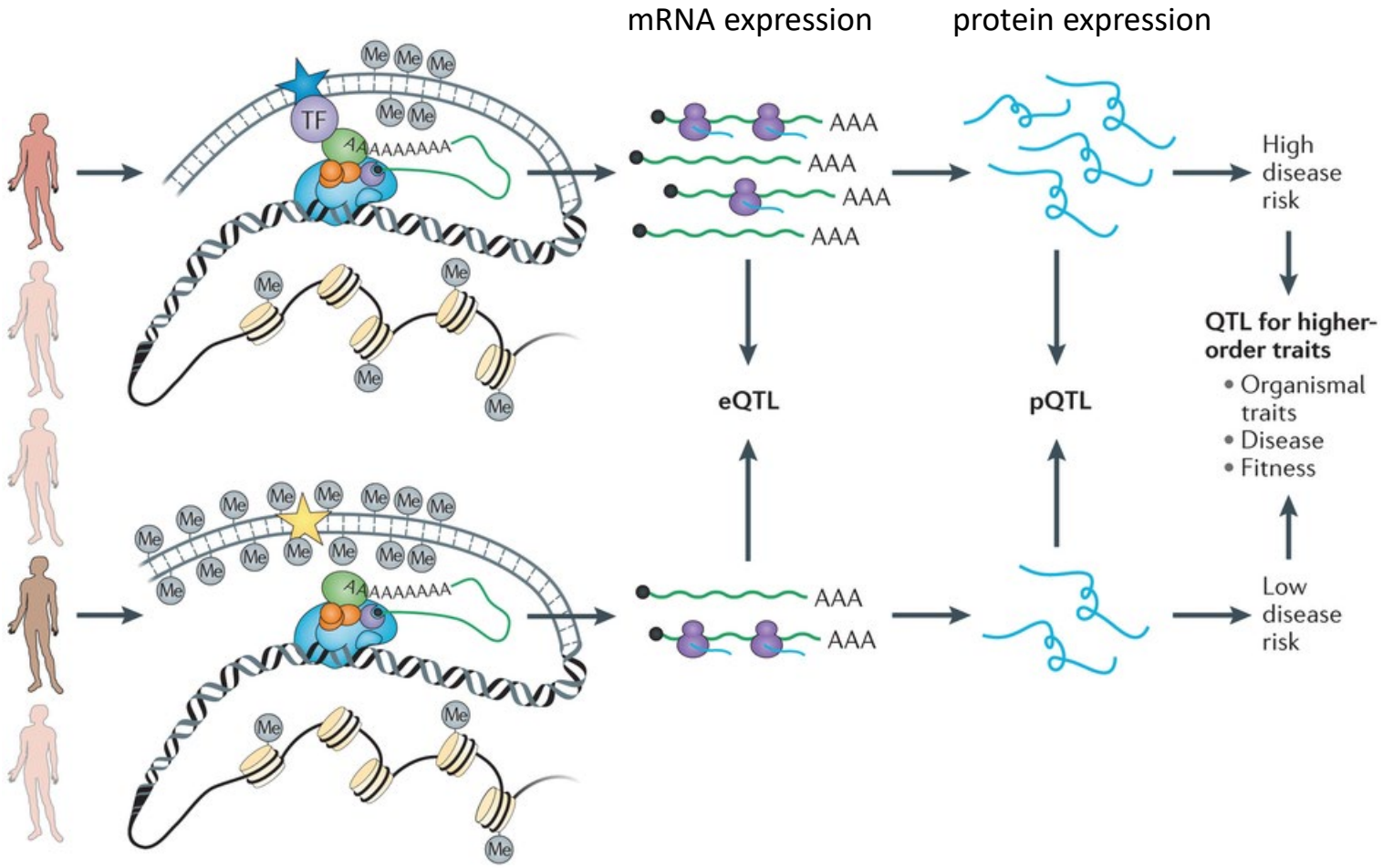
- widespread genetic variation affecting the regulation of most genes
- eQTLs for transcript structure and expression level are equally common, genetically largely independent

Lappalainen *et al.* 2013 *Nature*

Summary: early eQTL Discoveries

- Ubiquitous cis-eQTLs: Common variation impacts gene expression levels in every study exploring it, #eGenes increases with sample size, resolution increases with DNA sequencing
- Context-specificity: Ancestry, tissue type, cell type, activation status
- Disease interpretation: Genetic basis of complex traits may influence gene expression and suggest causal genes
- Resource: Data sharing facilitates discovery

Understanding genetic-trait associations by exploring the biology that lies between the two



Context:

Ancestry

Cell/Tissue type

Activation State

Sex

Disease

CONTEXT IS IMPORTANT!

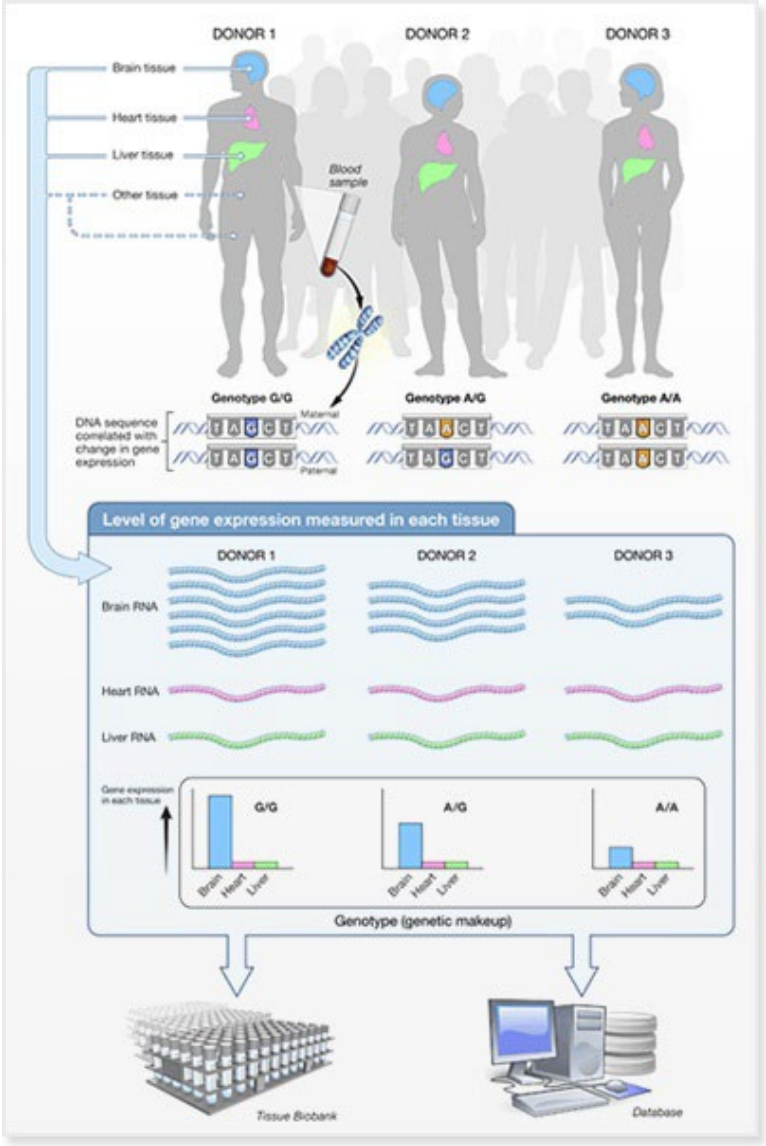
Albert and Kruglyak, 2015 *Nature Reviews Genetics*

NIH Genotype-Tissue Expression (GTEx) Project

Launched in 2010

Goal: Standardized collection and profiling of 50 tissues from 1000 deceased donor patients

RNA-Seq and genotyping



Primary scientific goals:

Determine tissue specificity of eQTLs and splicingQTLs (sQTLs)

Characterize trans-eQTLs

Create resource: Publicly-available database of eQTLs

Consortium and expanded data (whole genome sequencing +) have enabled **MUCH** more than this

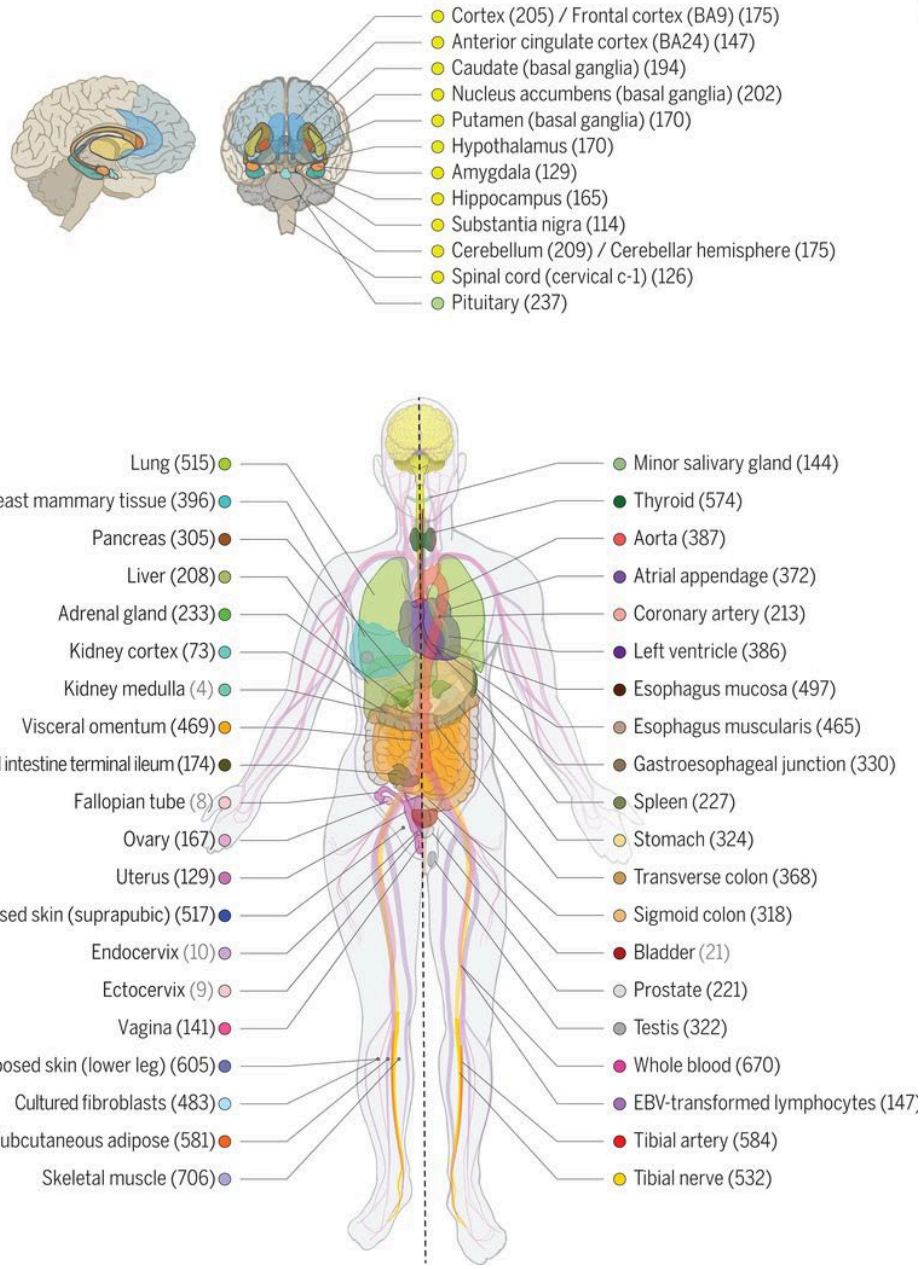
GTEx

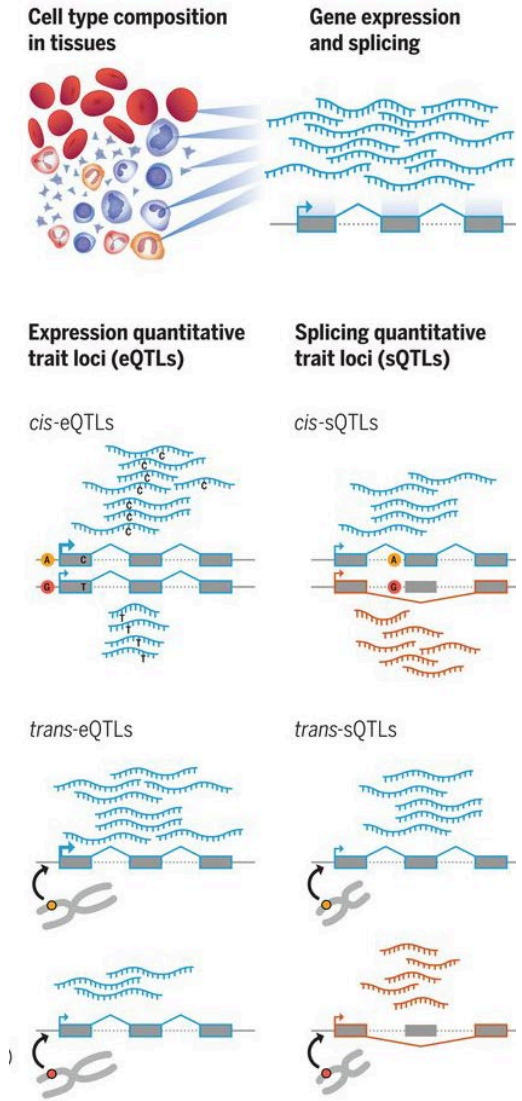
838 donors and 17,382 tissue samples

54 tissues (including 11 brain regions and two cell lines)

85.3% EUR-American
12.3% Af-American
1.4% As-American
1.9% Latino/Hispanic

RNA-seq (82.6M reads)
WGS (32x)



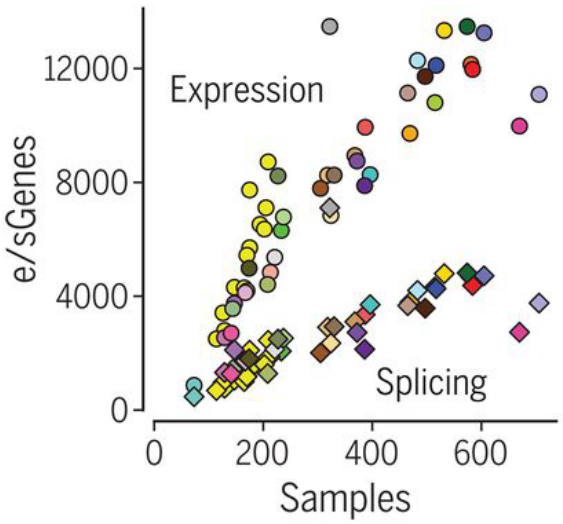


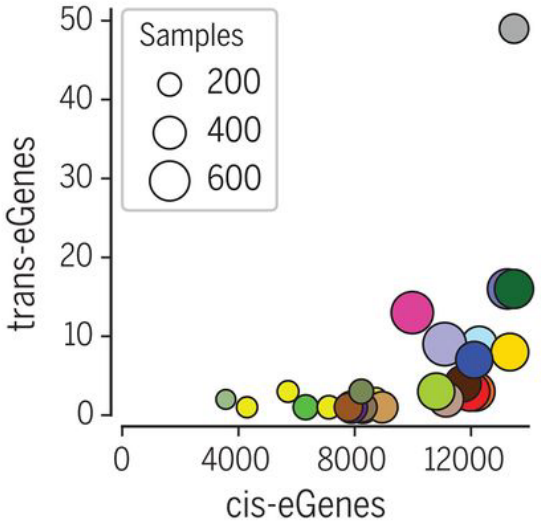
eQTL analysis: 49 tissues or cell lines that had at least 70 individuals

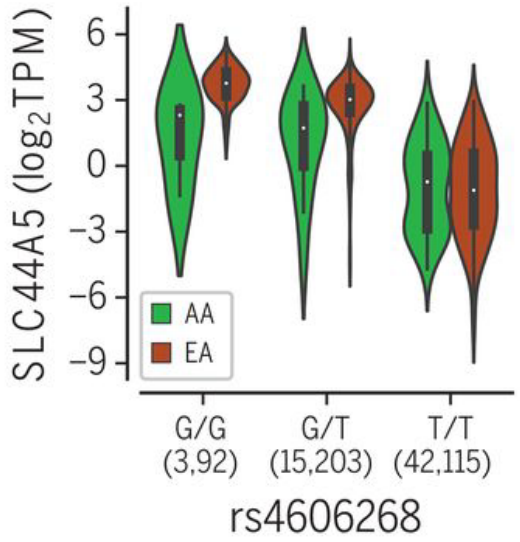
GTEx

The GTEx Consortium *Science* 2020;369:1318-1330

GTEx eQTLs







cis-eQTLs: 18,262 (94.7%) protein-coding and 5006 (67.3%) lincRNA genes

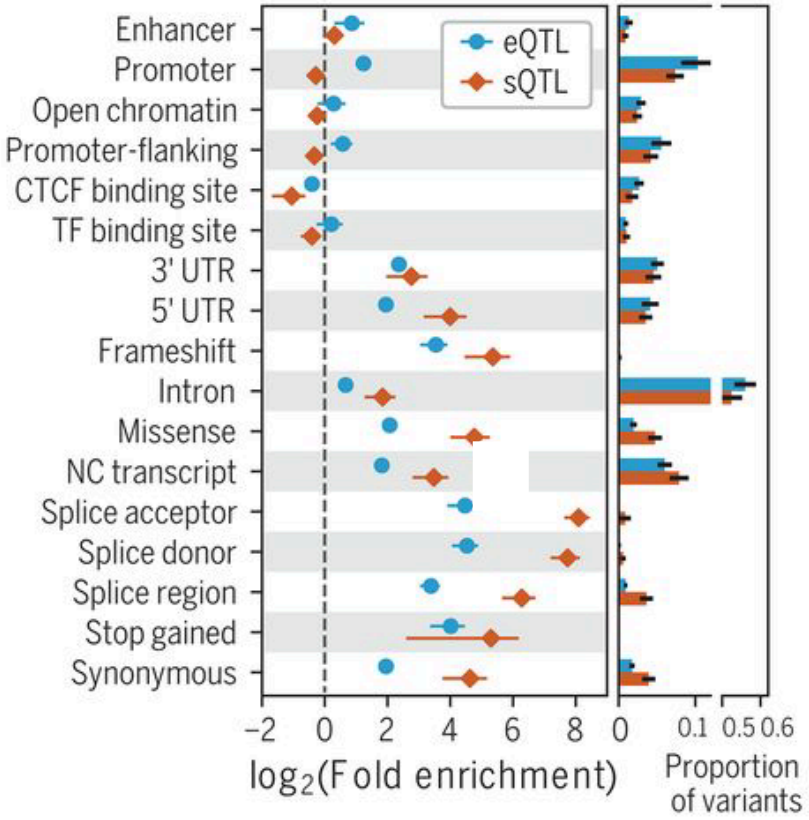
cis-sQTLs: 12,828 (66.7%) protein-coding and 1600 (21.5%) lincRNA genes

trans-eQTLs: 143 eGenes

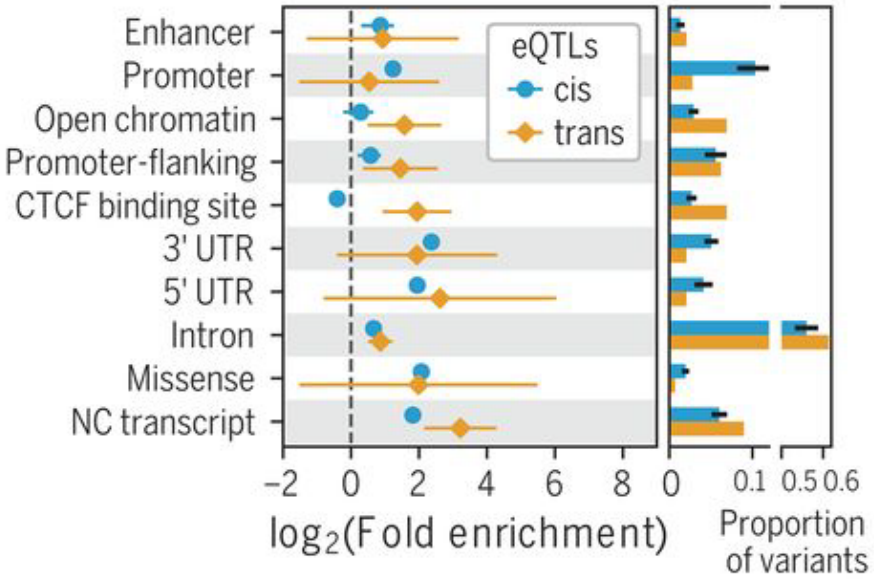
Population-biased eQTLs in 31 tissues with ss>20 EA and AA: 178 pb-eQTLs for 141 eGenes

Aguet, F. *et al*. 2020 *Science*.

Functional mechanisms of genetic regulatory effects



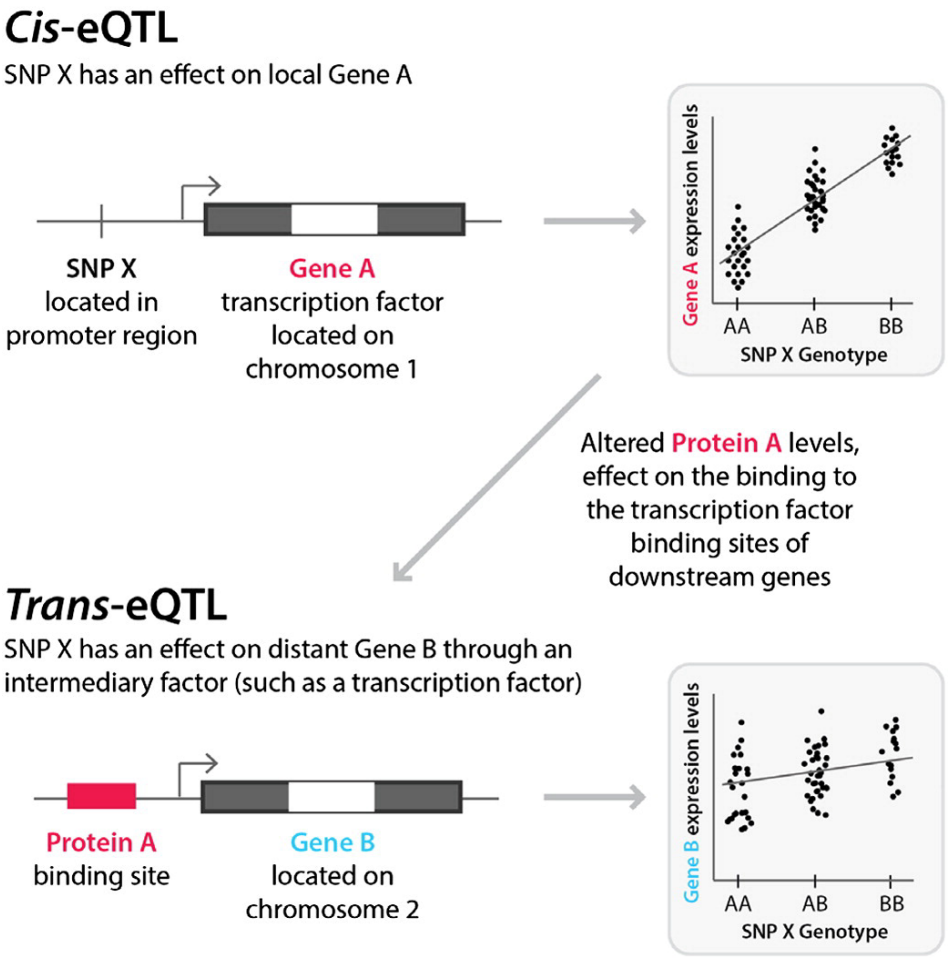
cis-eQTLs and sQTLs enriched in functional elements



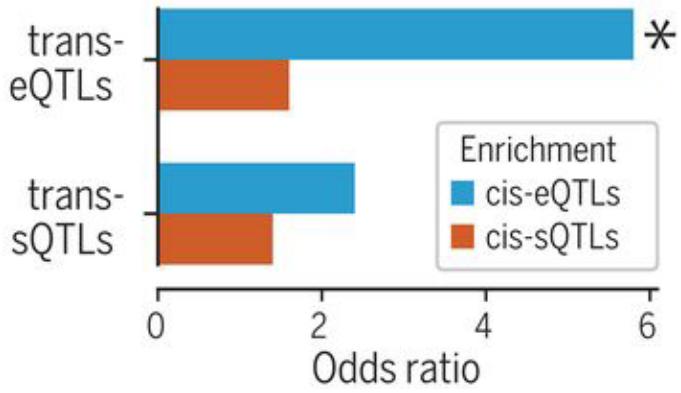
trans-eQTLs: disruption of CTCF binding may underlie distal genetic effects, potentially via its effect on interchromosomal chromatin interactions

GTEx

Trans-eQTLs caused by cis-eQTLs?



trans-eVariants enriched for cis-eVariants in the same tissue. Much less so by sQTLs



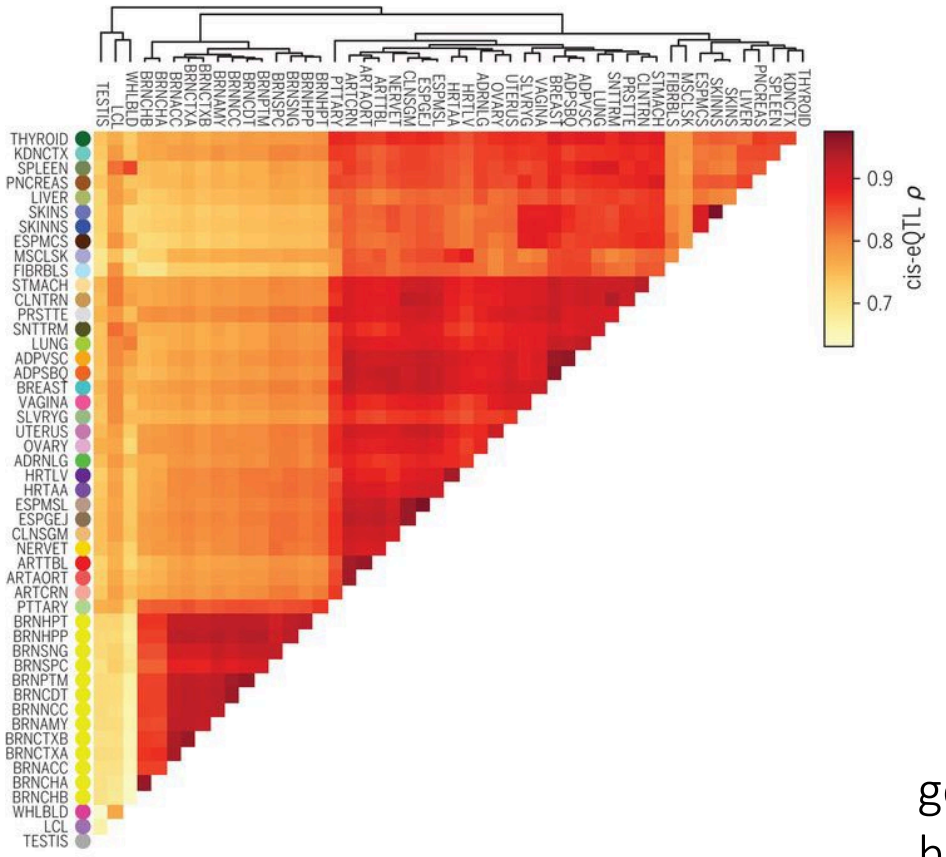
Mediation analysis: 77% of trans-eQTLs are mediated by cis-eQTLs.

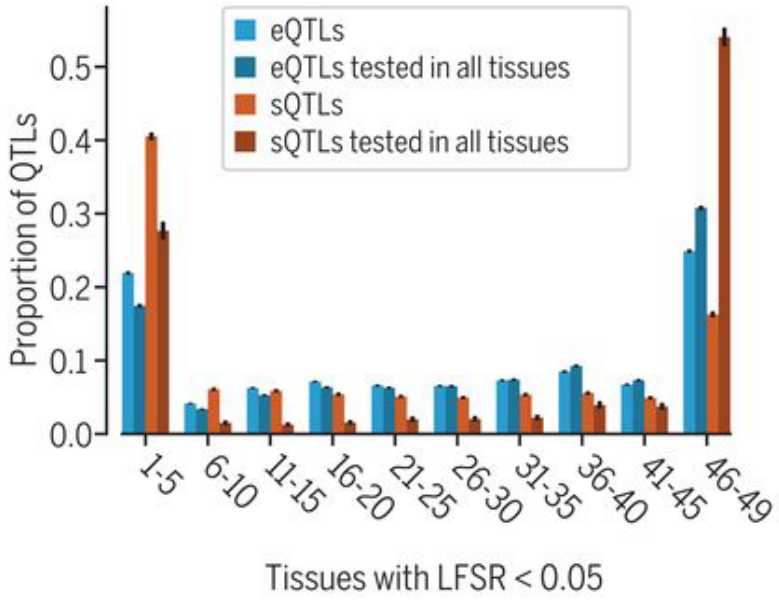
Tissue specificity of cis-QTLs

Tissue clustering cis-eQTL effect sizes:

brain regions form a separate cluster, and testis, lymphoblastoid cell lines, whole blood tend to be outliers.

Blood is not an ideal proxy for most tissues. Skin may better capture effects in other tissues.





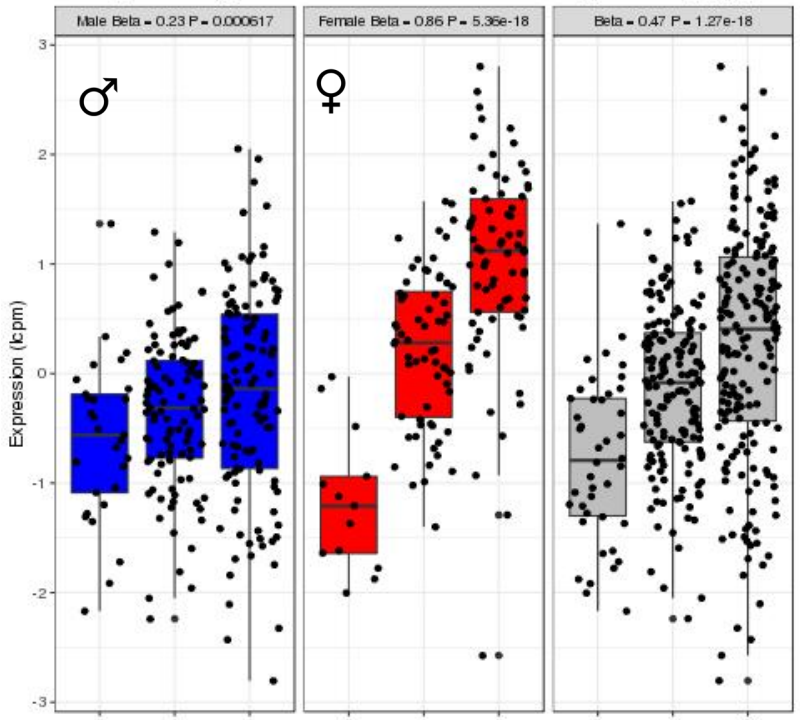
genetic regulatory effects tend to be either highly tissue specific or highly shared

Aguet, F. *et al*. 2020 *Science*.

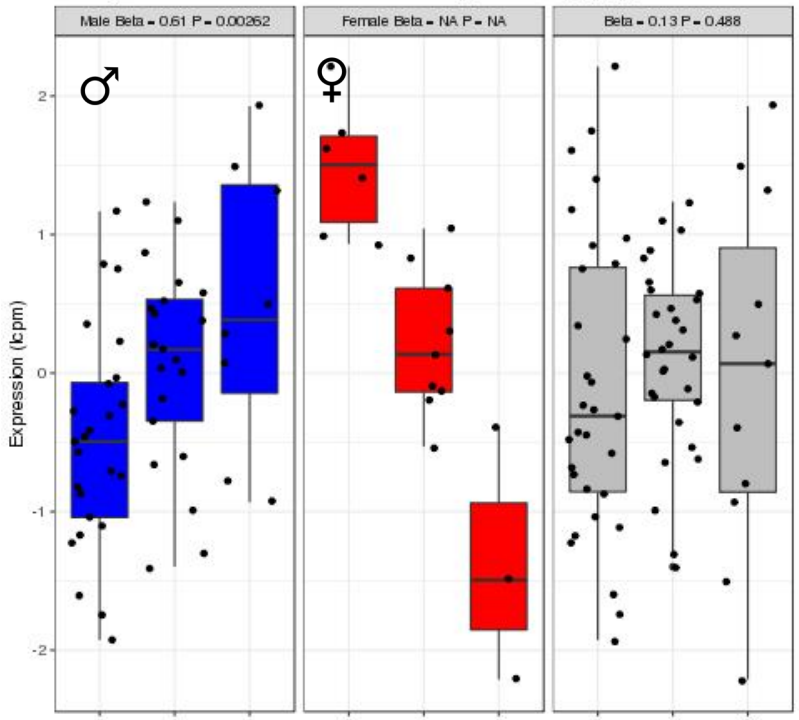
GTEx

Sex effects in the GTEx transcriptome sex-biased eQTLs

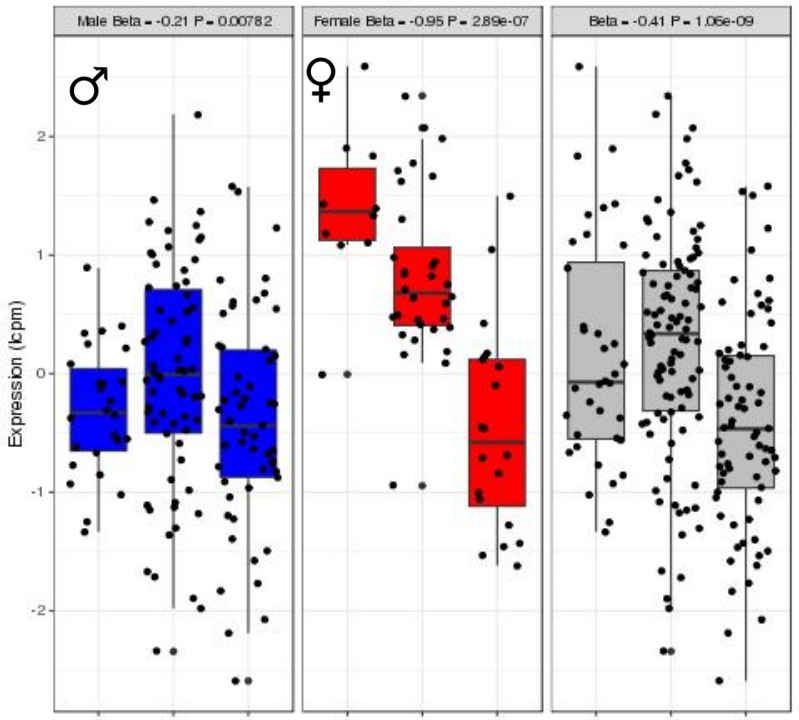
Breast: LINC00920



Kidney Cortex: CSPG5



Liver: HKDC1

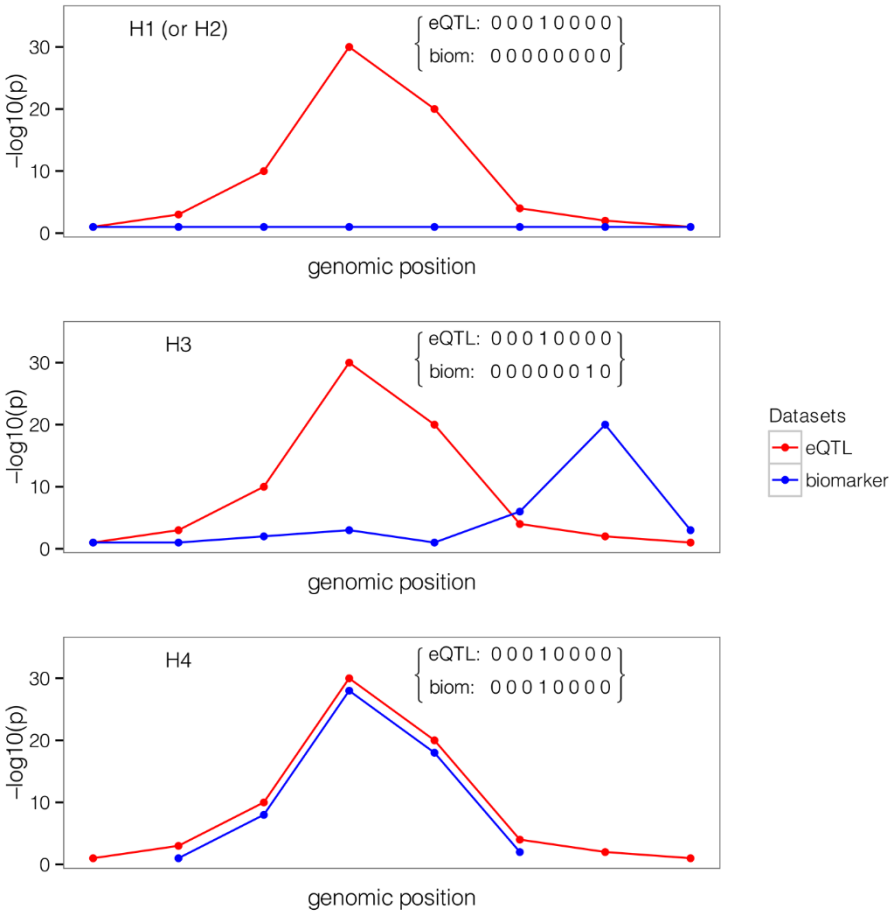


Model: $Y_i \sim \beta_0 + \beta_1 Sex + \beta_2 Genotype + \beta_{(1\text{-}n)} BasalCovs + \beta_{(1\text{-}m)} PEERS + \beta_3 Genotype \times sex + \varepsilon$

Oliva *et al*. 2020 *Science*

GTEx

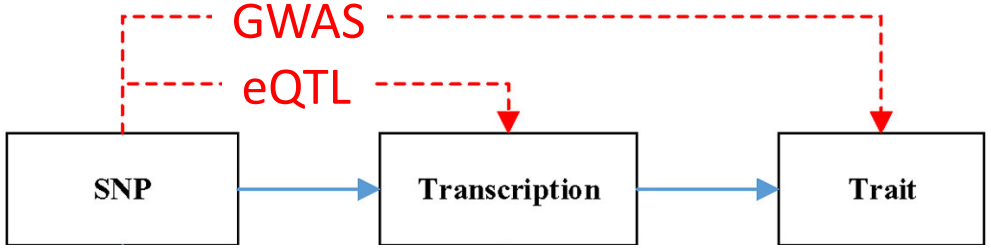
Colocalization GWAS + eQTLs: *coloc*



Giambartolomei *et al.* 2014 *PLoS Genetics*

Statistical approach to test likelihood that a GWAS association and eQTL have sa causal variant

Creates hypothesis:
SNP → gene expression → trait



Formal test: Mediation analysis

GTEx

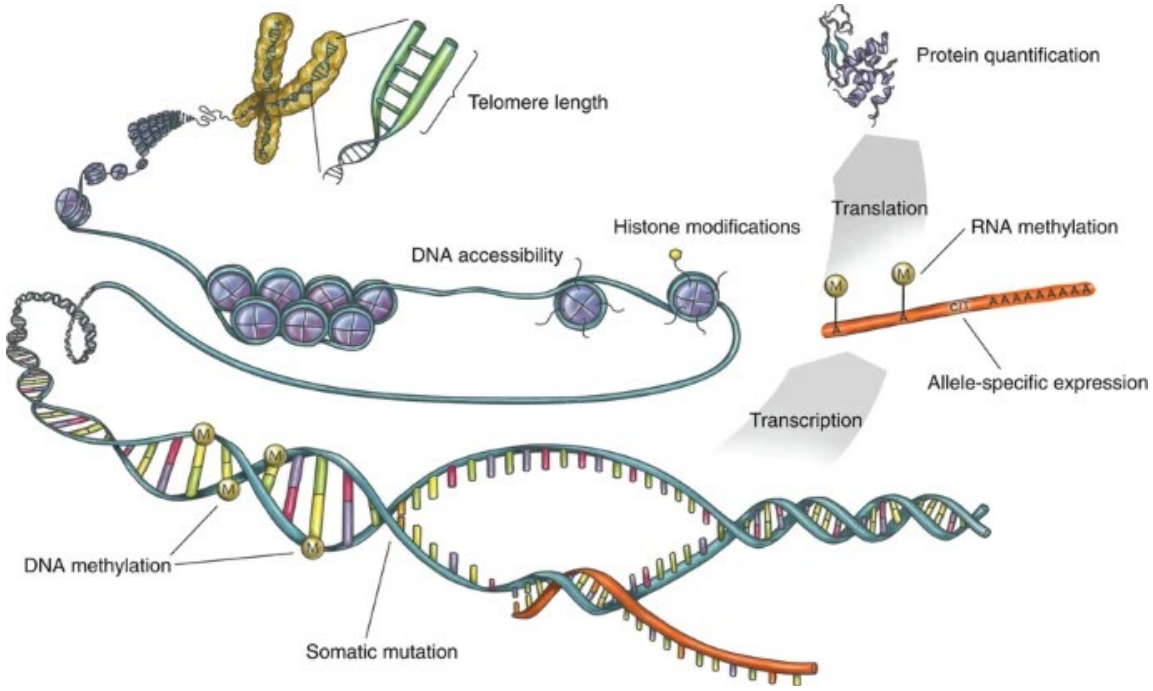
GTEx Summary

- Large resource of gene expression and eQTLs and sQTLs
- Allelic heterogeneity: multiple independent variants impacting expression levels
- trans-eQTLs mediated by cis-eQTLs
- Cell-type specific genetic regulation
- Relationship to complex traits (1000’s of hypotheses)
- Interaction QTLs (population/sex) present but will require larger sample sizes

Enhancing GTEx (eGTEx)

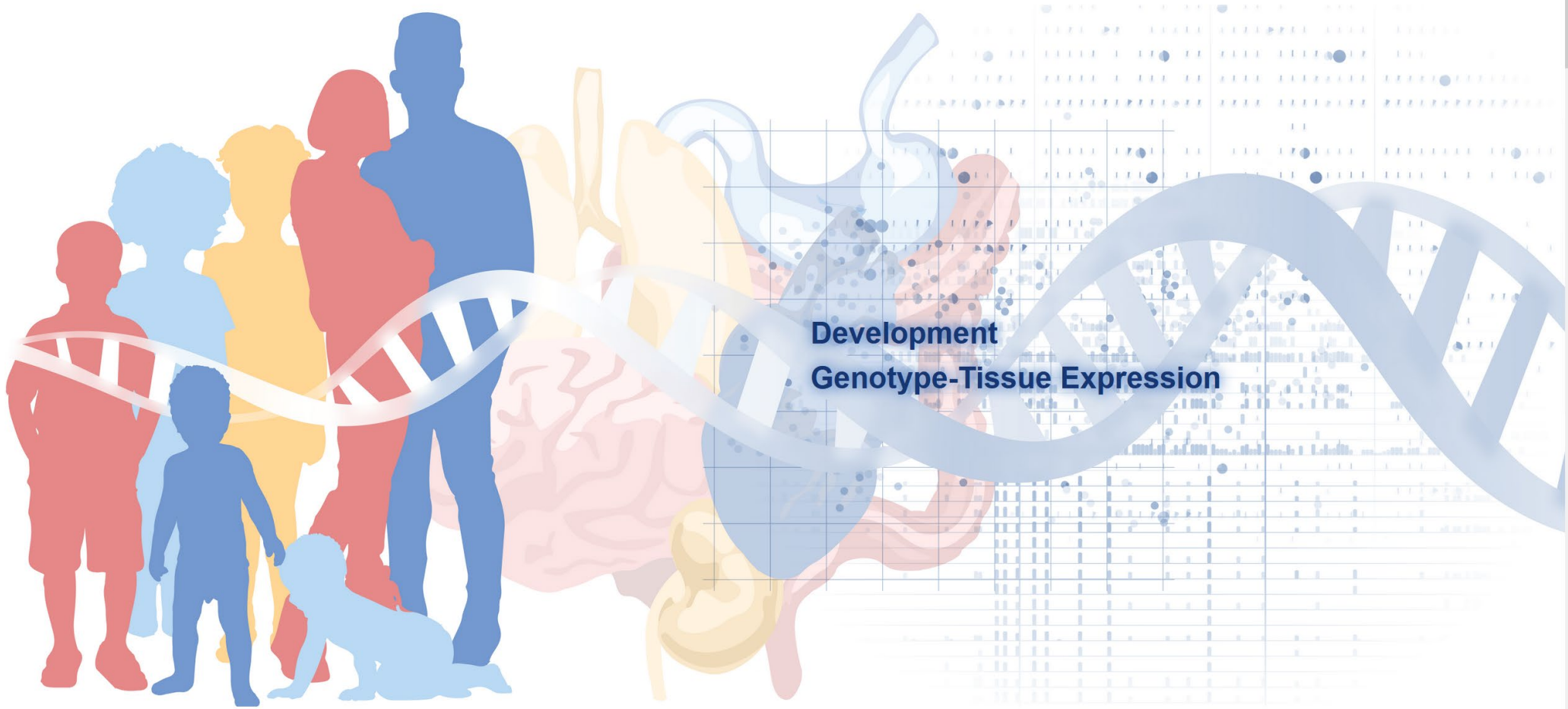
Adding additional data to GTEx samples

- DNA methylation of brain
- DNaseI hs
- bisulfite sequencing 8 tissues + H3K27ac
- somatic mutations across tissues
- targeted allele specific expression
- telomere length across tissues
- protein QTLs 3 tissues (mass spec)



Stranger *et al.* 2017 *Nature Genetics*

dGTEx Development Genotype-Tissue Expression



Postnatal
0-2 yo

Early Childhood
2-8 yo

Pre-pubertal
~8-12.5 yo

Post-pubertal
~12.5-18 yo

The Genotype-Tissue Expression (GTEx) Portal is a comprehensive public resource for researchers studying tissue and cell-specific gene expression and regulation across individuals, development, and species, with data from 3 NIH projects.

GTEx

The Adult GTEx project is a comprehensive resource of WGS, RNA-Seq, and QTL data from samples collected from 54 non-diseased tissue

dGTEx

The Developmental GTEx (dGTEx) project is a new effort to study development-specific genetic effects on gene expression and to establish a

NHPdGTEx

The Non-Human Primate Developmental GTEx (NHP-dGTEX) project is a complement to dGTEx in 2 translational non-human primate

eQTLGen Consortium

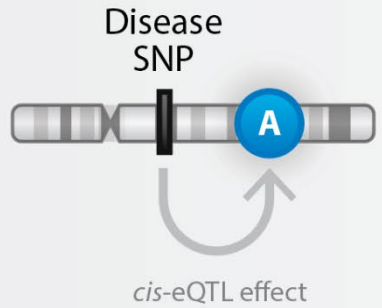
31,684 blood samples

10,317 trait-associated SNPs

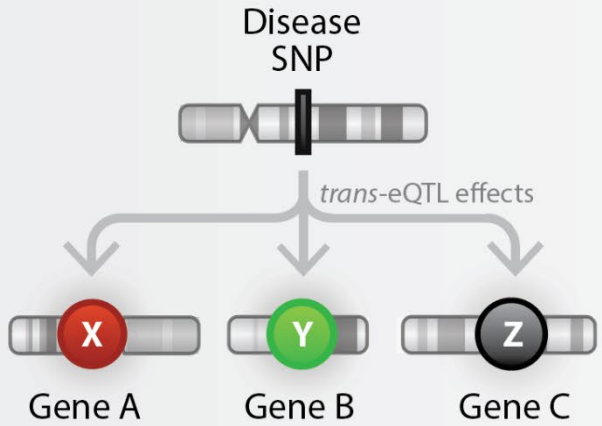
11M SNPs (MAF ≥ 1%)

19,960 genes studied

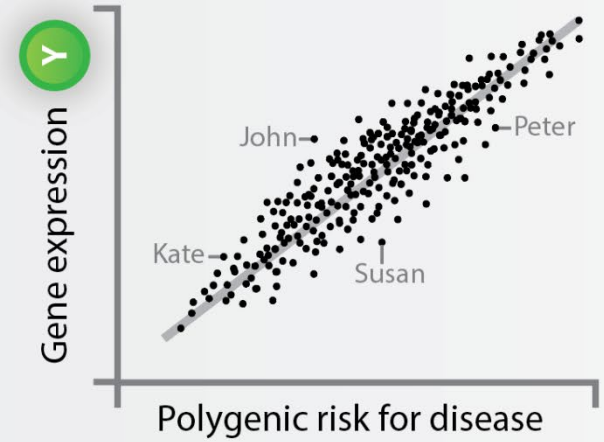
cis-eQTL analysis:
11M SNPs studied
(Window size 1Mb, MAF ≥ 1%)



trans-eQTL analysis:
10,317 trait-associated
SNPs studied



Polygenic risk score analysis:
1,263 traits studied



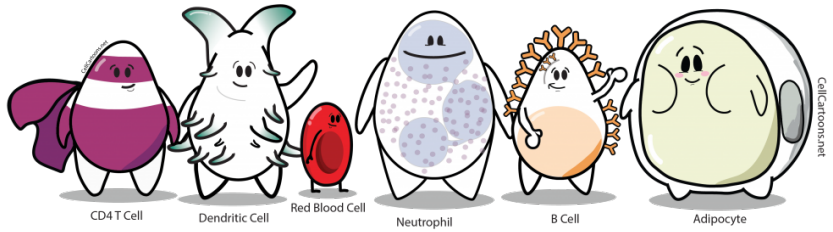
cis-eQTL analysis results:
16,989 (88.3%) cis-eQTL genes

trans-eQTL analysis results:
6,298 (31%) trans-eQTL genes
3,853 (36%) genetic risk factors

Polygenic score analysis results:
2,658 (13%) eQTS genes
689 (54%) traits affect gene expression

Moving forward

- Context-specific experimental designs
- Context-specific analysis and reporting
- Novel statistical approaches
- Context-specific annotations
- Single-cell approaches
- Larger-sample sizes (esp. for trans-eQTLs)
- Machine learning to predict function



eQTL resources

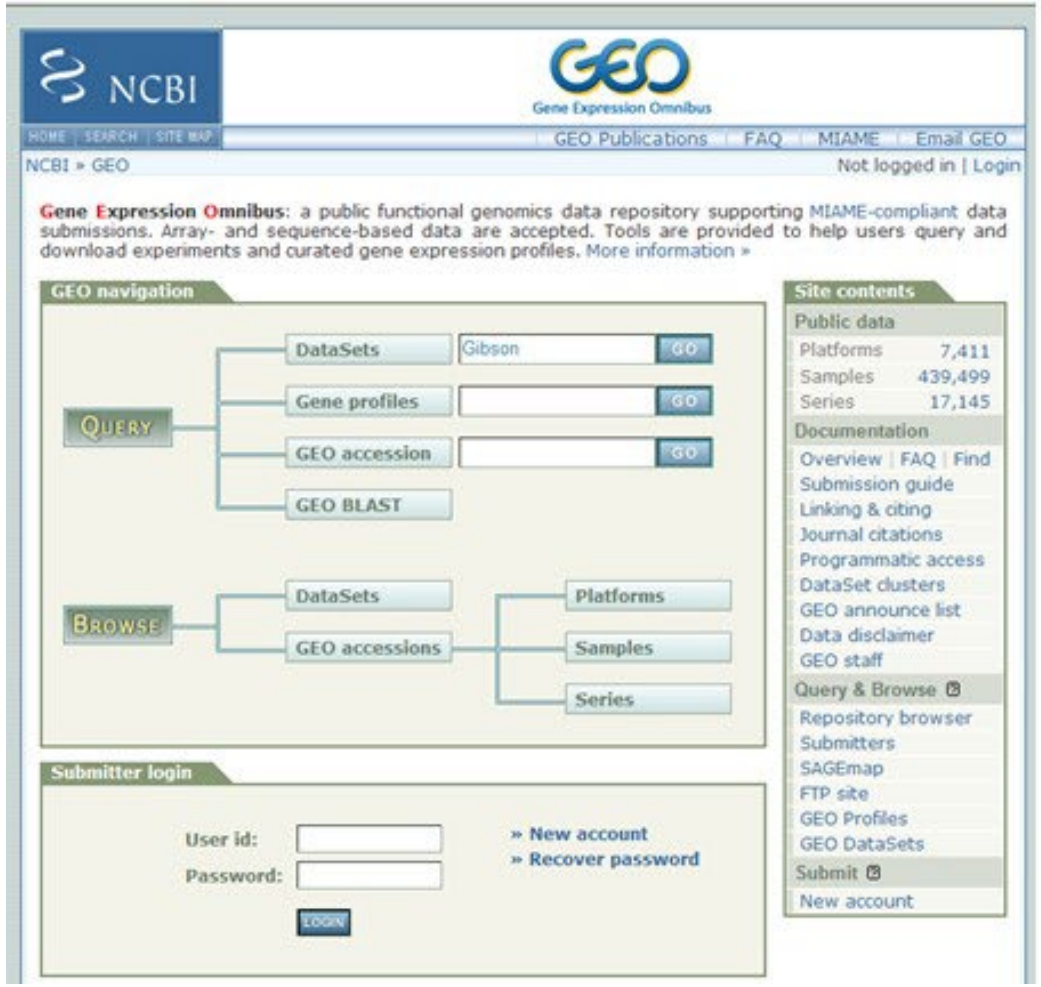
- Genotype-Tissue Expression Project: https://www.gtexportal.org/home/
- eQTLGen consortium: https://www.eqtlgen.org/
- Fivex eQTL browser: https://fivex.sph.umich.edu/about
- Open Targets: https://www.opentargets.org/
- scQTLbase: http://bioinfo.szbl.ac.cn/scQTLbase/
- singleQ: http://www.sqraolab.com/scqtl

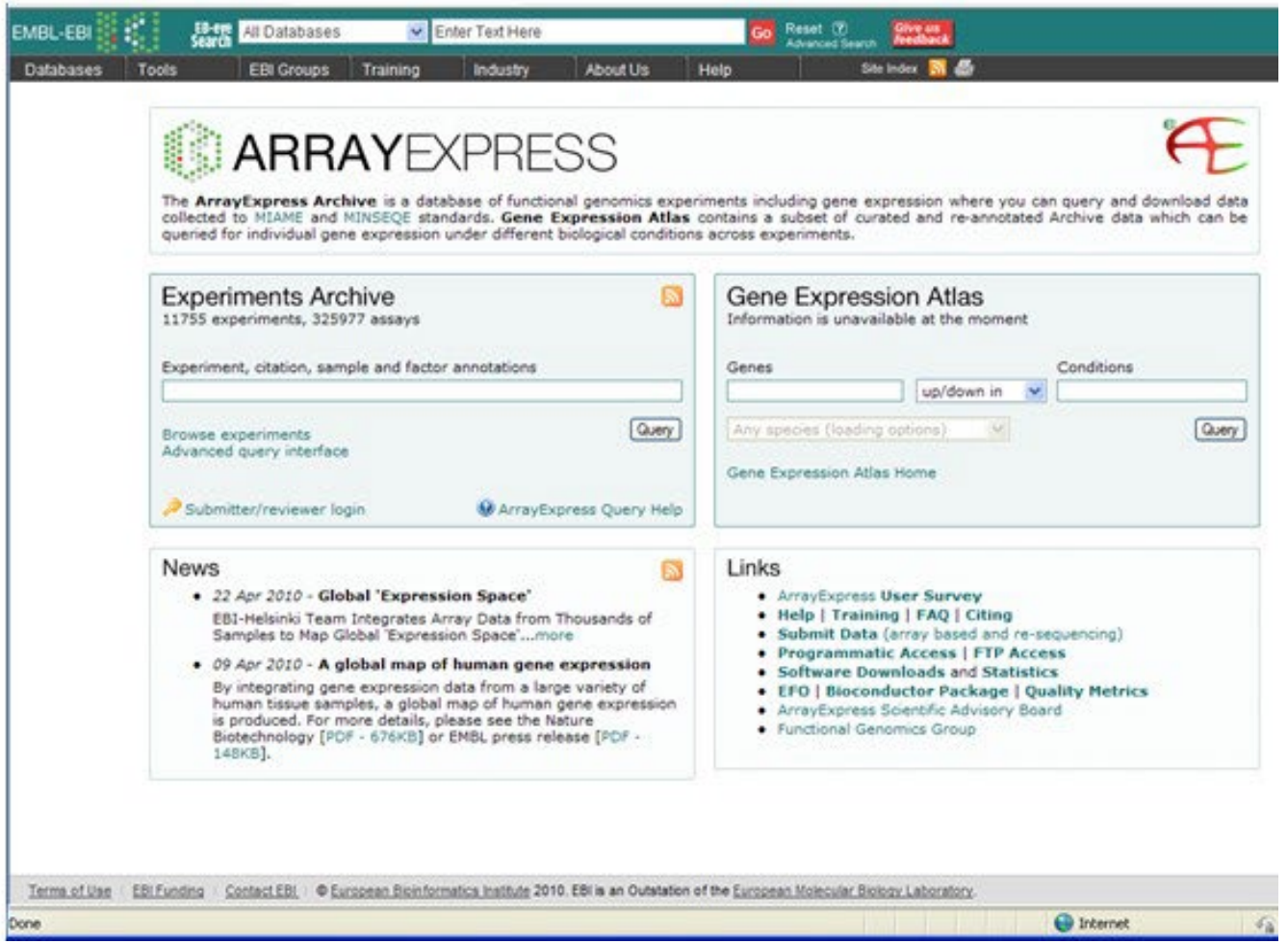
eQTL software tools

- Matrix eQTL: *Shabalin et al.* 2012 *Bioinformatics*
  - Pros: Supports cis- and trans-, Fast and scalable, Memory-efficient, highly parallelizable
  - Cons: Does not handle random effects or more complex mixed models, Limited visualization capabilities
- FastQTL: Ongen *et al.* 2016 *Bioinformatics*
  - Pros: extremely fast, memory efficient for cis-eQTLs
  - Cons: not for trans-, does not support complex models
- QTLtools: Delaneau *et al.* 2017 *Nature Comm*
  - Pros: more flexible to other data types (methylation, proteins), supports mixed models, built-in support for permutation tests, which help assess statistical significance
  - Cons: slower than Matrix eQTL (permutations), more complex model specification
- tensorQTL: Taylor-Weiner *et al.* 2019 *Genome Biology*
  - Pros: GPU-based, can capture complex interactions between multiple variables, powerful for datasets with rich, multi-dimensional data beyond just SNP-expression pairs.
  - Cons: computationally intense, more complex to implement and interpret, requiring knowledge of tensor decomposition methods, slow for basic eQTL analysis tasks.

Extra slides

Where to obtain transcriptome datasets





Example of PEER Factors in Action

- Imagine you measure **gene expression** from 100 individuals and want to find eQTLs. However, some individuals' samples were processed on different days, some had slightly degraded RNA, and some had more immune cells in their blood than others.
- Without correction, these factors create **systematic noise** in your expression data, making it harder to detect true genetic effects. PEER analyzes expression patterns to find these hidden influences and removes them before eQTL mapping.
- **Before PEER correction**: Expression levels for a gene are highly variable due to technical and biological confounders.
- **After PEER correction**: The expression levels are **adjusted**, making genetic associations easier to detect.

More details for TMM

Choose a Reference Sample

- One sample is selected as a reference (typically the sample with median total library size).
- All other samples are compared to this reference to compute M-values and A-values.

Compute M-values and A-values Per Sample

- For each gene, M-values (log-fold change) and A-values (average abundance) are computed per sample relative to the reference.
- This means that each sample gets its own M-A plot when calculating its normalization factor.

Trim Low-Expression and Extreme M-Value Genes (Per Sample)

- Genes with low A-values (low abundance) or extreme M-values (large fold changes) are excluded from the scaling factor calculation.
- This trimming is done for each sample separately.

Calculate the TMM Scaling Factor Per Sample

- The weighted mean of M-values is computed after trimming.
- This results in a TMM normalization factor for each sample.

Apply Normalization Factors Across All Samples

- Once all samples have their own TMM factor, the raw counts are adjusted accordingly across all samples.