Expression quantitative trait loci (*eQTLs*) (I)

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today's learning objectives

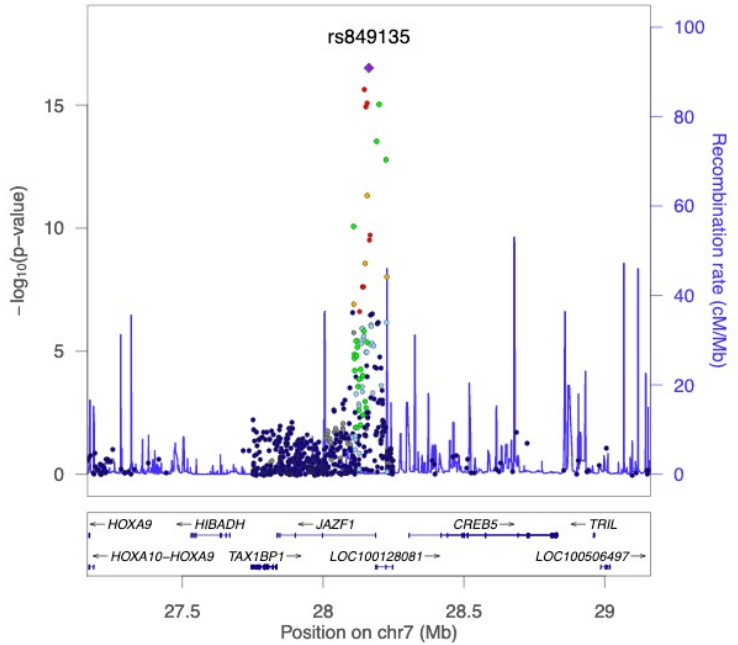
- Understand the biological machinery of eQTLs
- Describe the statistical methods for calculating and predicting eQTLs
- Understand the purpose of multiple testing correction
- Appreciate the utilization of eQTL to help identify disease susceptibility loci

Relevant papers

1) Westra & Franke. From genome to function by studying eQTLs. Biochimica et Biophysica Acta, 2014
2) Lappalainen et al., Transcriptome and genome sequencing uncovers functional variation in humans, Nature 2013

Genetic mapping study (GWAS)

Susceptibility loci for type 2 diabetes: 34,840 cases and 114,981 controls



The DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, Nat Genetics. 2012

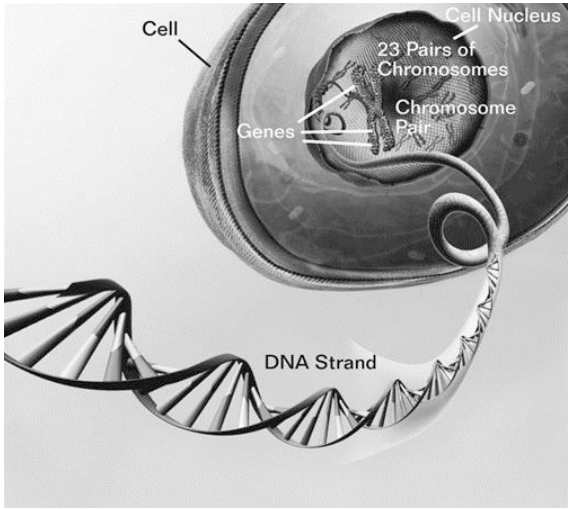
Genetic mapping study (GWAS)

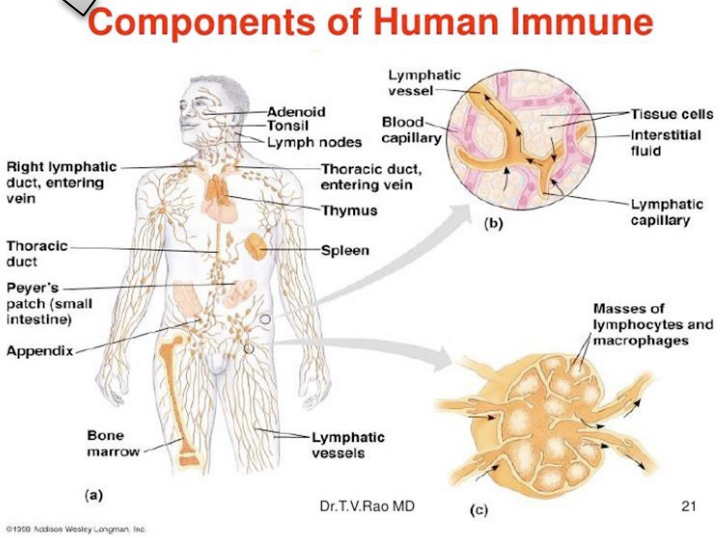
- Why are there so many SNPs with significant p-values in that region?
- How can we identify the true risk SNP(s)?

Intermediate layer: functional genomics

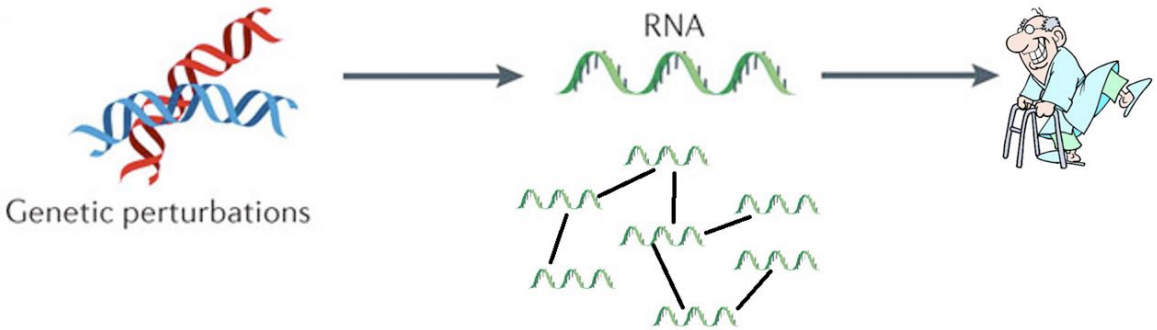
There are multiple intermediate factors between DNA to phenotypes

DNA →→→ →→→ →→→ **disease / traits**

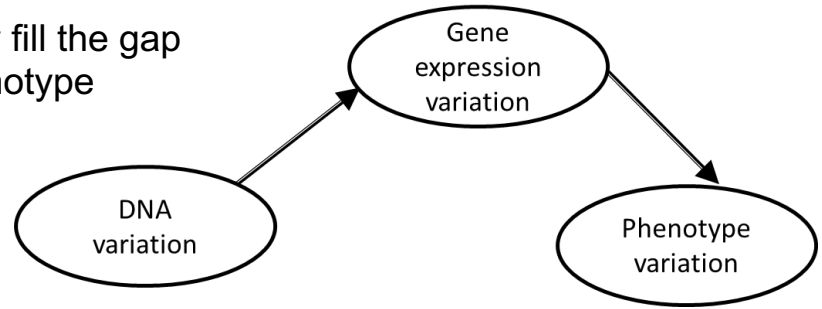




Gene expression for genetics studies



Gene expression may fill the gap between DNA to phenotype



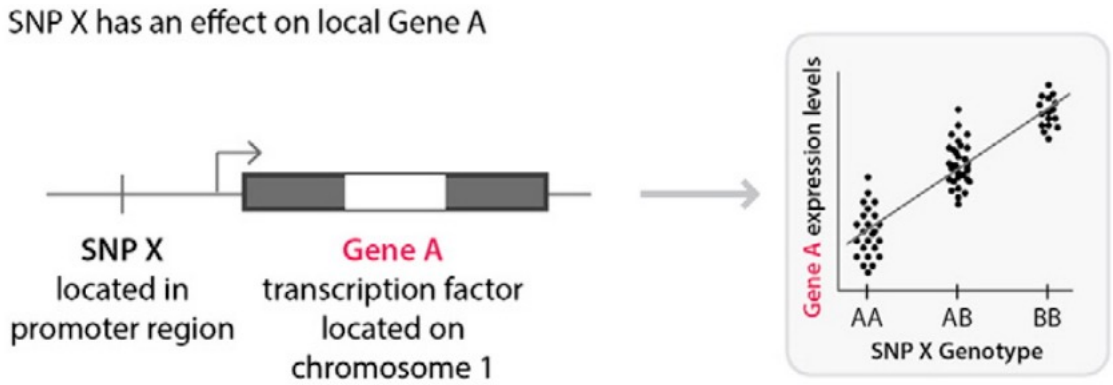
Data needed

- Genotypes
  - SNP arrays (and imputation from reference of whole genomes)
  - Common alleles in population, e.g., Minor Allele Frequency (MAF > 5%)
- Gene expression
  - RNA sequencing
  - Microarray RNA transcriptional profiling (less common now)
  - Count data: usually log(FPKM + small_value)
    - small value could be 1, 0.5, 0.1, etc.

- Sample size?
  - 462 samples in Geuvadis project, Nature 2013 (doi: 10.1038/nature12531)
  - 15,201 samples in GTEx v8, Science 2020 (doi: 10.1126/science.aaz1776)

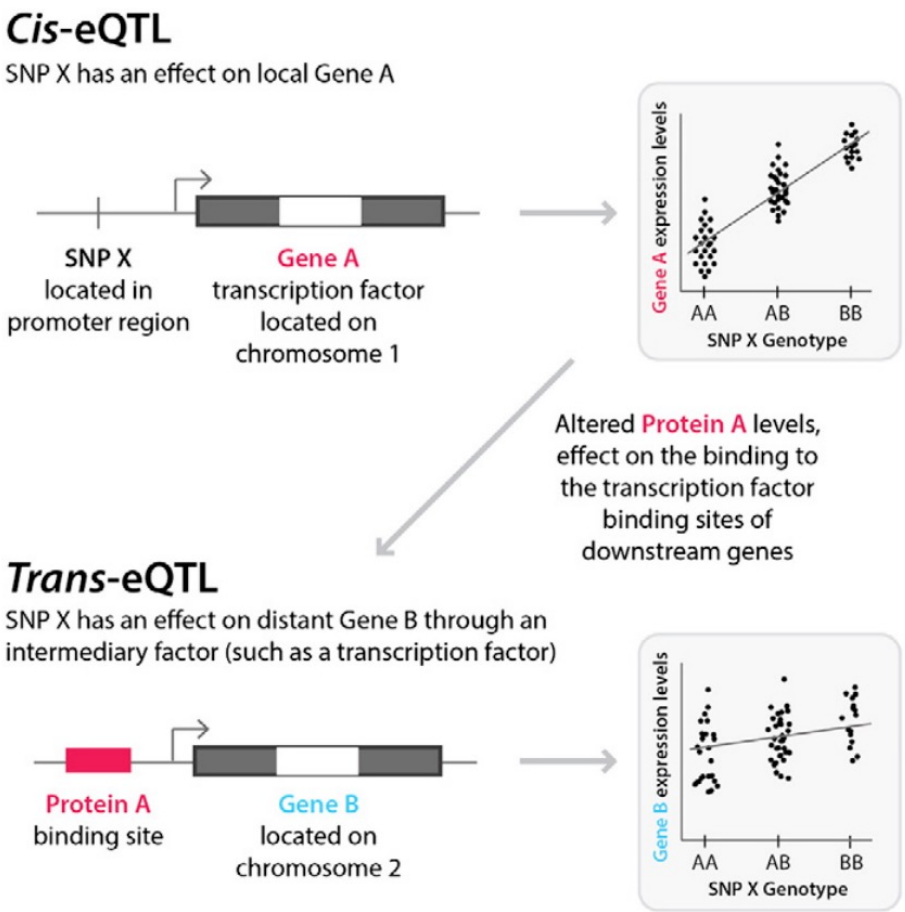
DNA variation and gene expression

Expression quantitative trait loci (eQTLs): genomic loci that contribute to variation in expression levels of mRNAs



Westra & Franke. Biochimica et Biophysica Acta, 2014

Cis-eQTL and trans-eQTL



Part 2: Statistical tests

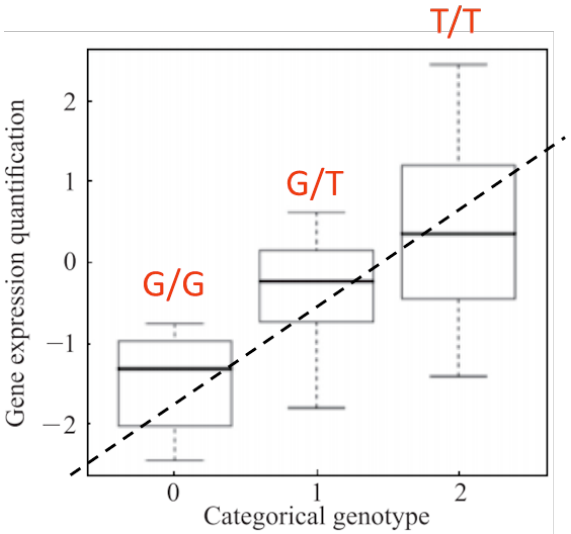
- Independent two-sample test: t test or Wilcoxon test
- Generalised linear model with likelihood ratio test
- Random effects for structured samples (next session)

Independent two-sample tests

- Independent two-sample t-test
- T/T genotype vs. non-T/T
- Comparing the gene expression level

Linear regression (generalized linear model)

- Likelihood ratio test (G for genotype: 0, 1, 2)
  - Null model likelihood L_0: $y = \beta_0 + \beta_1 \times \text{Sex}$
  - Alternative model likelihood L_1: $y = \beta_0 + \beta_1 \times \text{Sex} + \beta_G \times G$



Log Likelihood ratio: $r = 2\log(L_1/L_0)$ follows χ^2 distribution --> p value calculation

The weights (or coefficients, or effect size) $\beta_0, \beta_1, \beta_G$ will be fitted to archive maximum likelihood

Linear regression – more factors

- Likelihood ratio test (additional covariate x_2, e.g., BMI)
  - Null model likelihood L_0: $y = \beta_0 + \beta_1 \times \text{Sex} + \beta_2 \times x_2$
  - Alternative model likelihood L_1: $y = \beta_0 + \beta_1 \times \text{Sex} + \beta_2 \times x_2 + \beta_G \times G$

- Likelihood ratio test (multiple additional covariates $x_{1\ldots K}$)
  - Null model likelihood L_0: $y = \beta_0 + \sum_{k=1}^{K} \beta_k \times x_k$
  - Alternative model likelihood L_1: $y = \beta_0 + \sum_{k=1}^{K} \beta_k \times x_k + \beta_G \times G$

- Other factors may also contribute to variability in gene expression

Log Likelihood ratio: $r = 2\log(L_1/L_0)$ follows χ^2 distribution --> p value calculation

Linear regression – interaction

- Likelihood ratio test
  - Null model likelihood L_0: $y = \beta_0 + \beta_1 \times \text{Sex}$
  - Alternative model 1 likelihood L_1: $y = \beta_0 + \beta_1 \times \text{Sex} + \beta_G \times G$

- Adding and interaction variable: $z = \text{Sex} \times G$
  - Alternative model 2 likelihood L_2: $y = \beta_0 + \beta_1 \times \text{Sex} + \beta_G \times G + \beta_2 \times [\text{Sex} \times G]$
  - $y = \beta_0 + \beta_1 \times \text{Sex} + [\beta_G + \beta_2 \times \text{Sex}] \times G$
  - $= \beta_0 + \beta_1 \times \text{Sex} + [\beta_G \mathbb{I}(\text{Sex} = 0) + (\beta_G + \beta_2)\mathbb{I}(\text{Sex} = 1)] \times G$

- This is not physical interaction, but merely gender has different effect sizes
  - If sex =0: effect size is β_G; if sex = 1: effect size is $\beta_G + \beta_2$.

Log Likelihood ratio: $r_1 = 2\log(L_1/L_0)$ follows χ^2 distribution --> p value calculation
Log Likelihood ratio: $r_2 = 2\log(L_2/L_1)$ follows χ^2 distribution --> p value calculation

Multiple testing correction

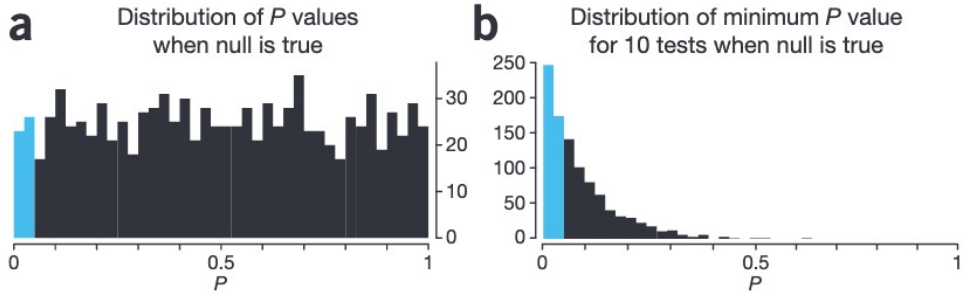
- Assume 100,000 SNPs and their cis- genes to test
  - For each SNP-gene pair, one likelihood ratio test each pair
  - We will perform 100,000 SNPs, by chance what is the lowest *p* value we will have even there is no eQTL? 1, 0.5, 0.1, or 0.00001

- What is the distribution of *p* value if the null model is true?
  - Not peak at 1, nor between 0.5 to 1
  - Under the null, the p value actually follows a uniform distribution in [0, 1].

Multiple testing correction

- What is the distribution of *p* value if the null model is true?
  - Under the null, the chance we see p value < 0.05 is 5%
  - By performing 10 times, the chance to have the lowest p value < 0.05 is 40%

- Multiple testing correction
  - None perfect methods, but some are practically useful
  - Benjamini-Hochberg correction, namely, False Discovery Rate (FDR) is commonly used



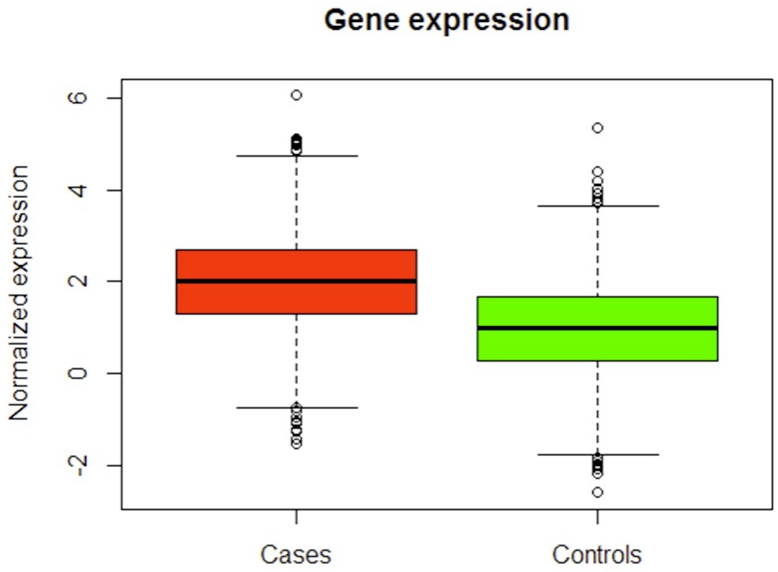
FDR: For a given FDR α, find the largest k that the kth $P_k < \frac{k}{n\_test} \alpha$

Require large sample size or do smaller number of tests

Altman & Krzywinski. P values and the search for significance. Nature methods 2017

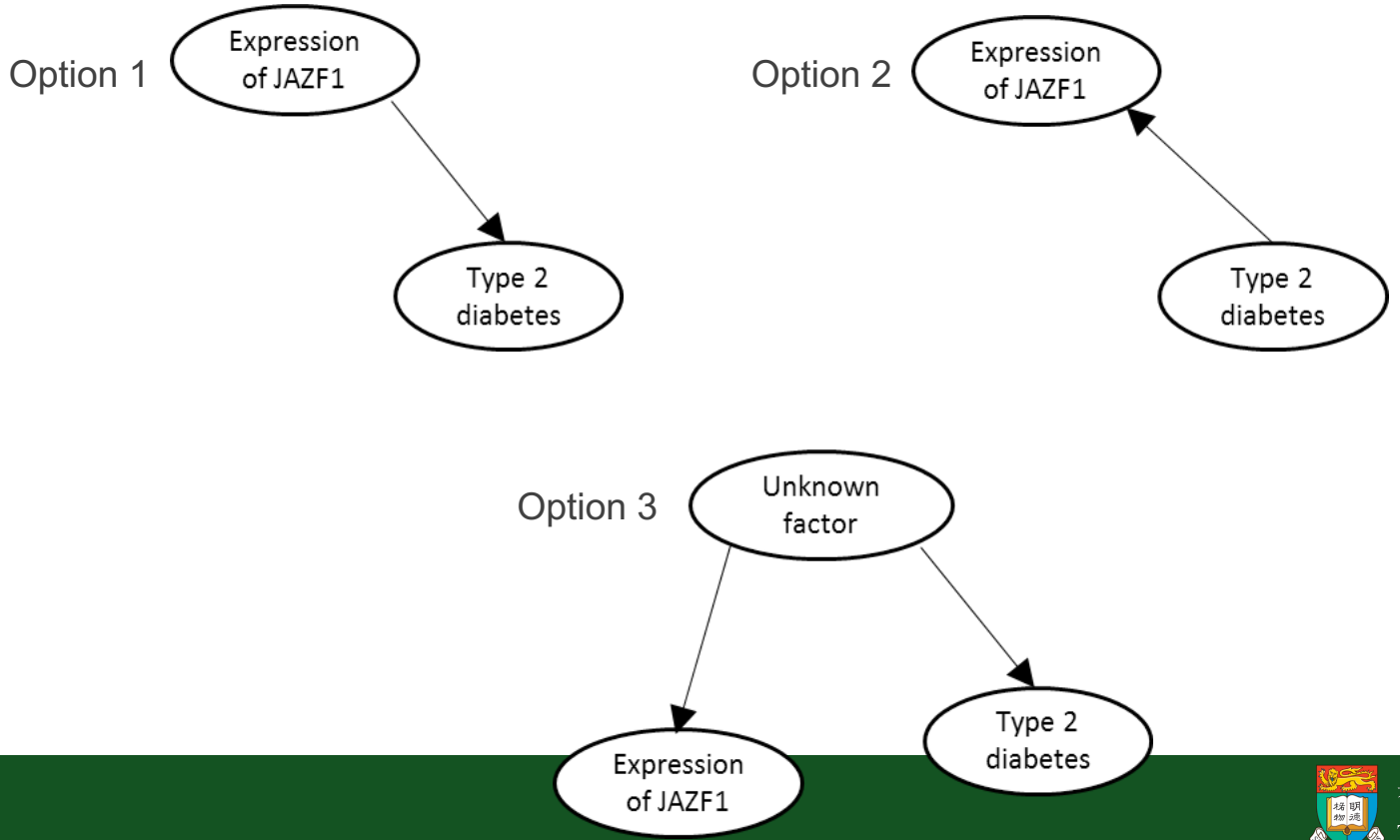
Part 3: Interpretation of eQTLs

Expression of JAZF1 in patients and healthy controls



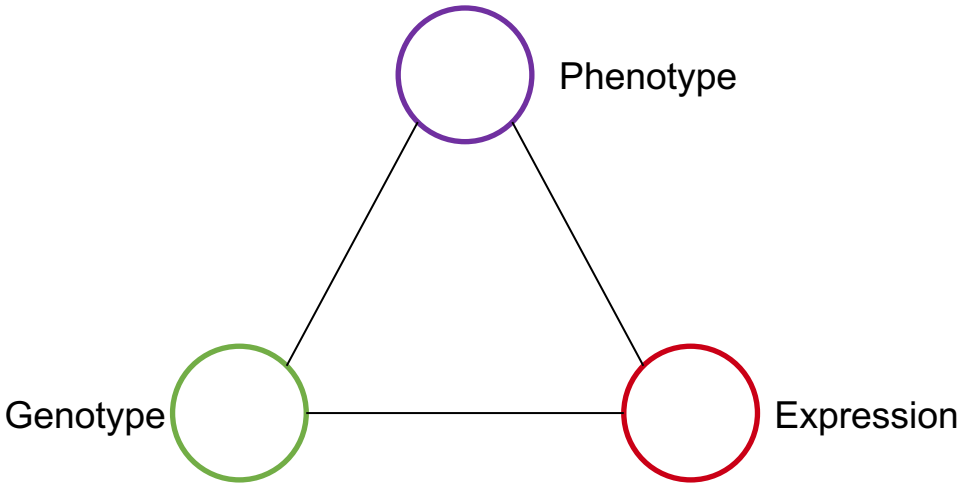
Interpretation of eQTLs: example on JAZF1

- Is the expression change a cause or an outcome?

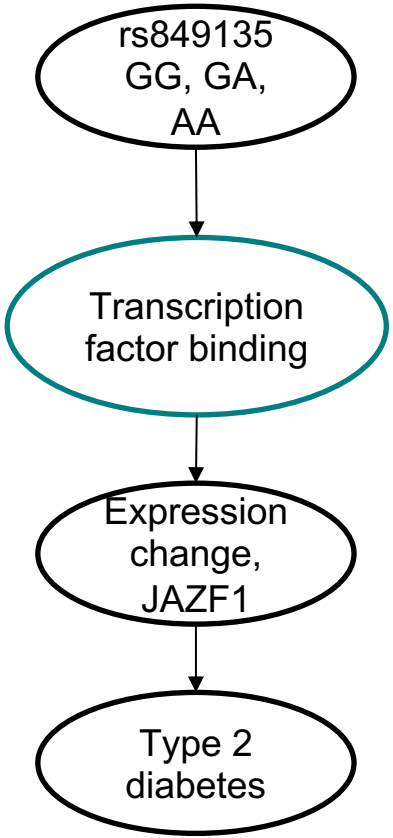


Interpretation of eQTLs: example on JAZF1

- eQTL and causality?

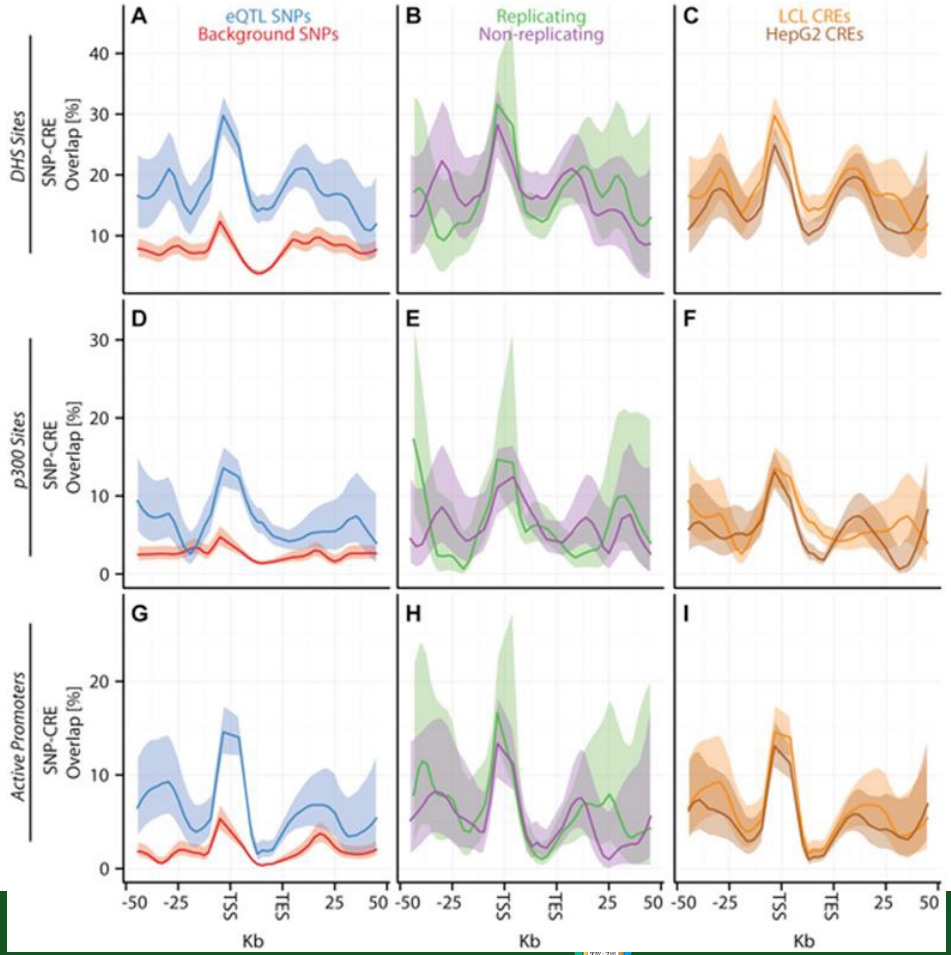


Possible pathogenic mechanism



Link between eQTLs and epigenetics

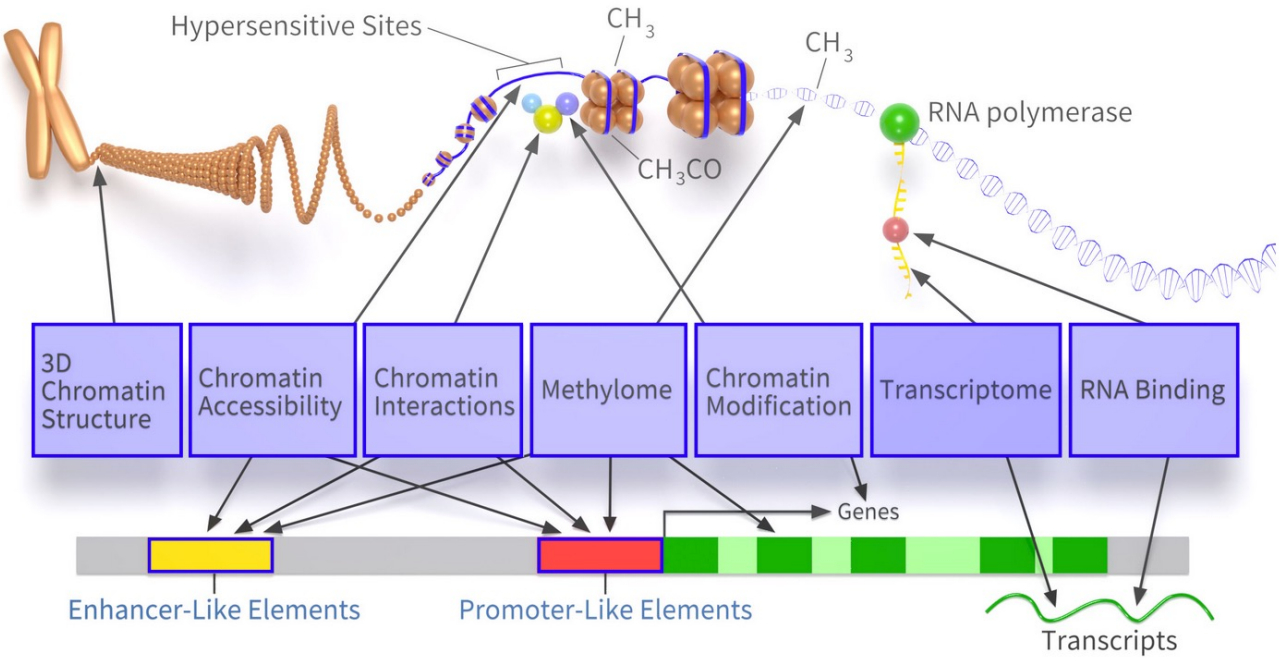
eQTL SNPs are enriched within activating cis-regulatory elements



TSS: transcription start site
TES: transcription end site

Epigenetics

study of mechanisms that involve mitotically and/or meiotically heritable changes in DNA other than changes in nucleotide sequence



Genotype-Tissue Expression (GTEx) Project

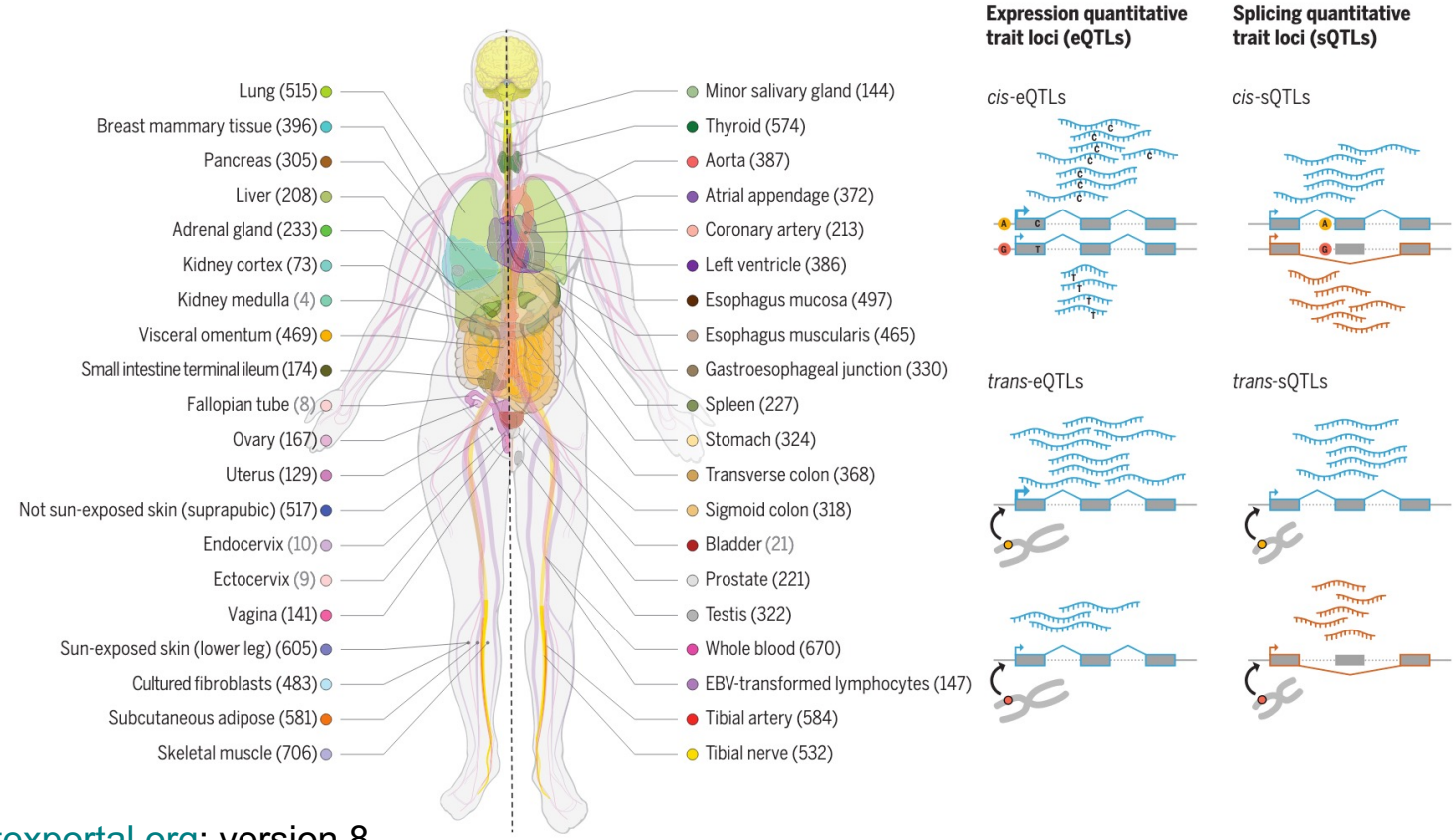
To establish a resource database and associated tissue bank for the scientific community to study the relationship between **genetic variation and gene expression** in human tissues

Goals

- Generate **public resource** with tissue-specific eQTLs and gene and isoform expression data across multiple human tissues
- **Contribute to understanding** of effects of genetic variation on gene expression and regulation
- Assist in **interpretation of disease/trait GWAS signals**

https://gtexportal.org

Genotype-Tissue Expression (GTEx) Project



https://gtexportal.org; version 8

Other resources

- EBI eQTL catalogue (re-computed 19 eQTL publications)
  - Data base: https://www.ebi.ac.uk/eqtl/Datasets/
  - Paper: https://www.biorxiv.org/content/10.1101/2020.01.29.924266v1

Questions

- Understand the biological machinery of eQTLs
- Describe the statistical methods for calculating and predicting eQTLs
- Understand the purpose of multiple testing correction
- Appreciate the utilization of eQTL to help identify disease susceptibility loci

Relevant papers

1) Westra & Franke. From genome to function by studying eQTLs. Biochimica et Biophysica Acta, 2014
2) Lappalainen et al., Transcriptome and genome sequencing uncovers functional variation in humans, Nature 2013