

Genome variation and function 3

Marie Saitou
Assistant Professor, CIGENE, BIOVIT
17. Oct. 2022



Learning Outcomes

We will learn about:

- Methods and significance of RNA sequencing
- How to analyze genetic variants that are associated with gene expression (eQTL, expression QTL)

Discussion Review

Q: Do some literature search about the top-associated genes and assume how this gene is associated with the trait, “Short-sightedness”.

To further understand the mechanism, what experiment would you plan?

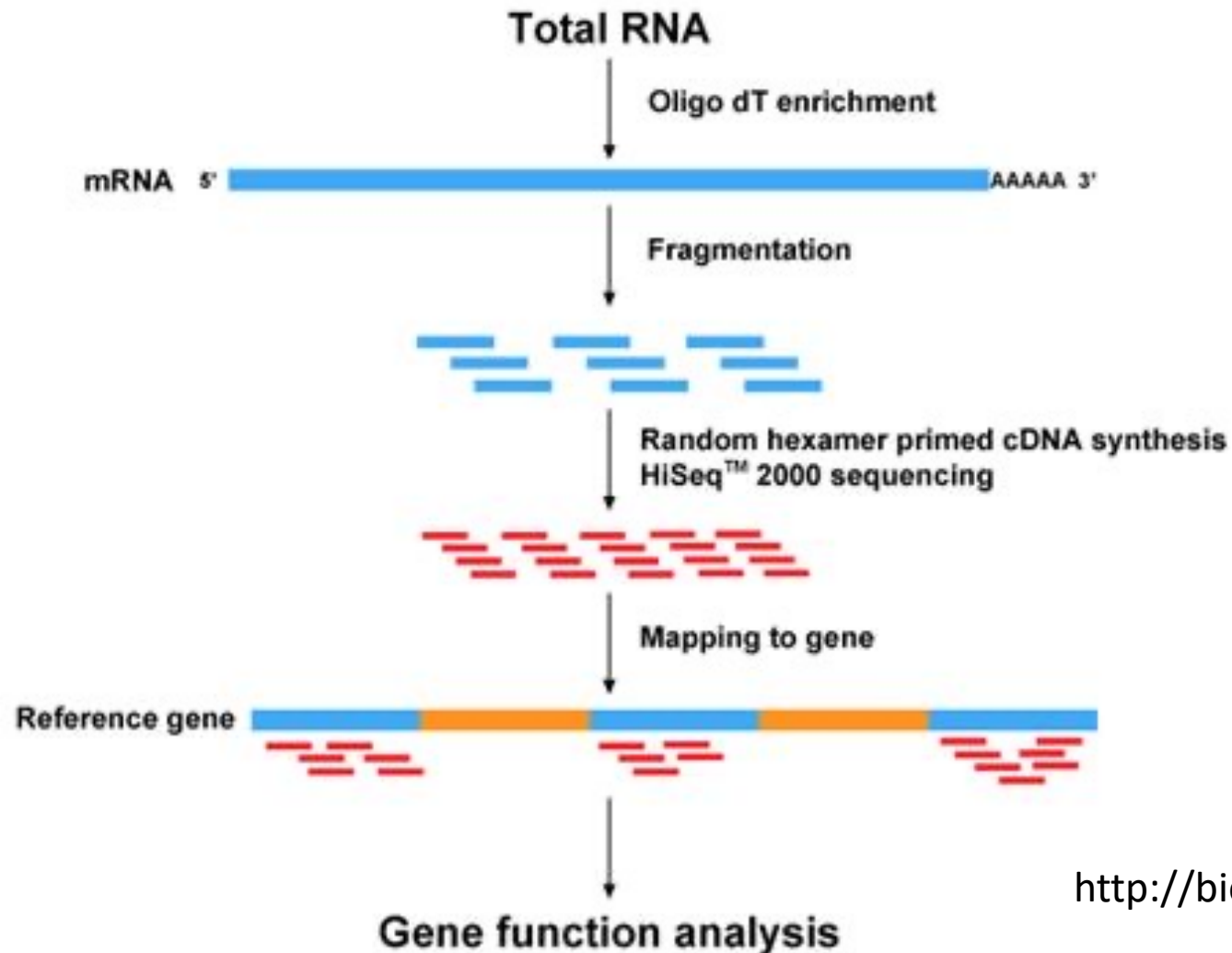
- Function: PRSS56 (serine protease: Type II transmembrane). This gene produces a protein that is involved during eye development.
- Gene-editing in cell lines, organoid (3D cell model from Induced Pluripotent Stem (iPS) Cells), or animals
- RNA-sequencing, cell development observation

Functional genomics of molecular traits

- Transcriptome analysis – gene expression in a genome-wide manner
- eQTL analysis – connect genetic variants and gene expression

RNA-sequencing

Examine the gene expression pattern in a genome-wide manner



<http://bio.lundberg.gu.se/courses/vt13/rnaseq.html>

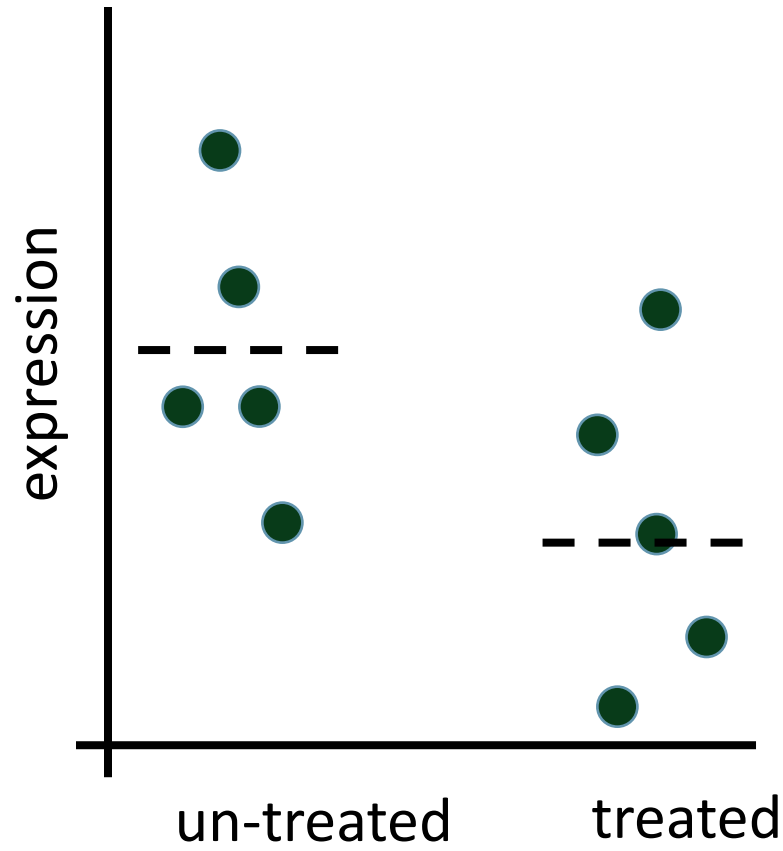
Fold change vs p-value

Fold change is
$$\frac{(\text{expression}_{\text{treated}} - \text{expression}_{\text{non-treated}})}{\text{expression}_{\text{non-treated}}}$$

p-value is the probability of obtaining results as extreme as the observed results of a hypothesis test, assuming that the null hypothesis is correct.

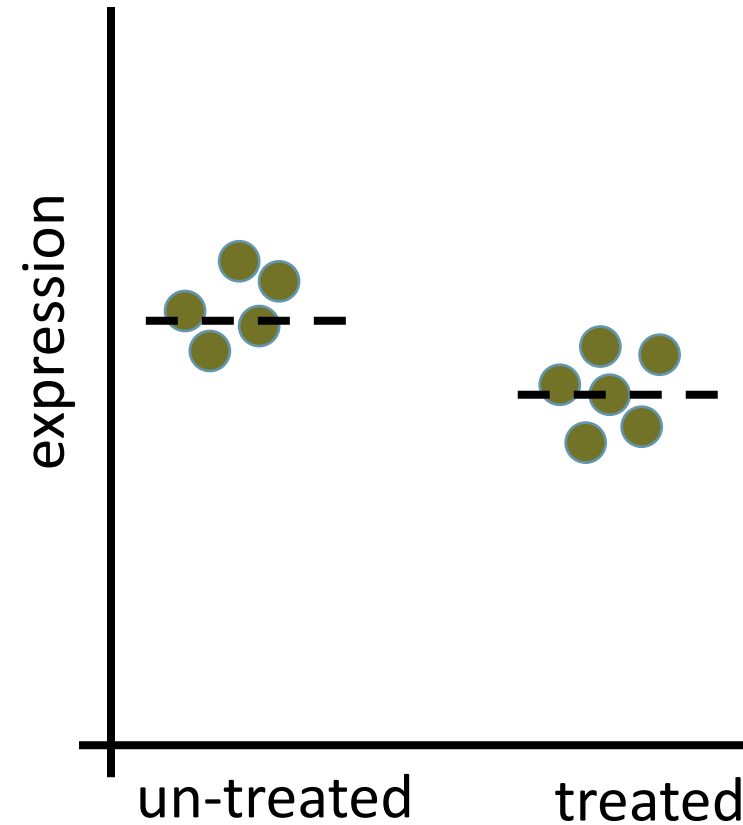
Case study:

Observed expression of gene A



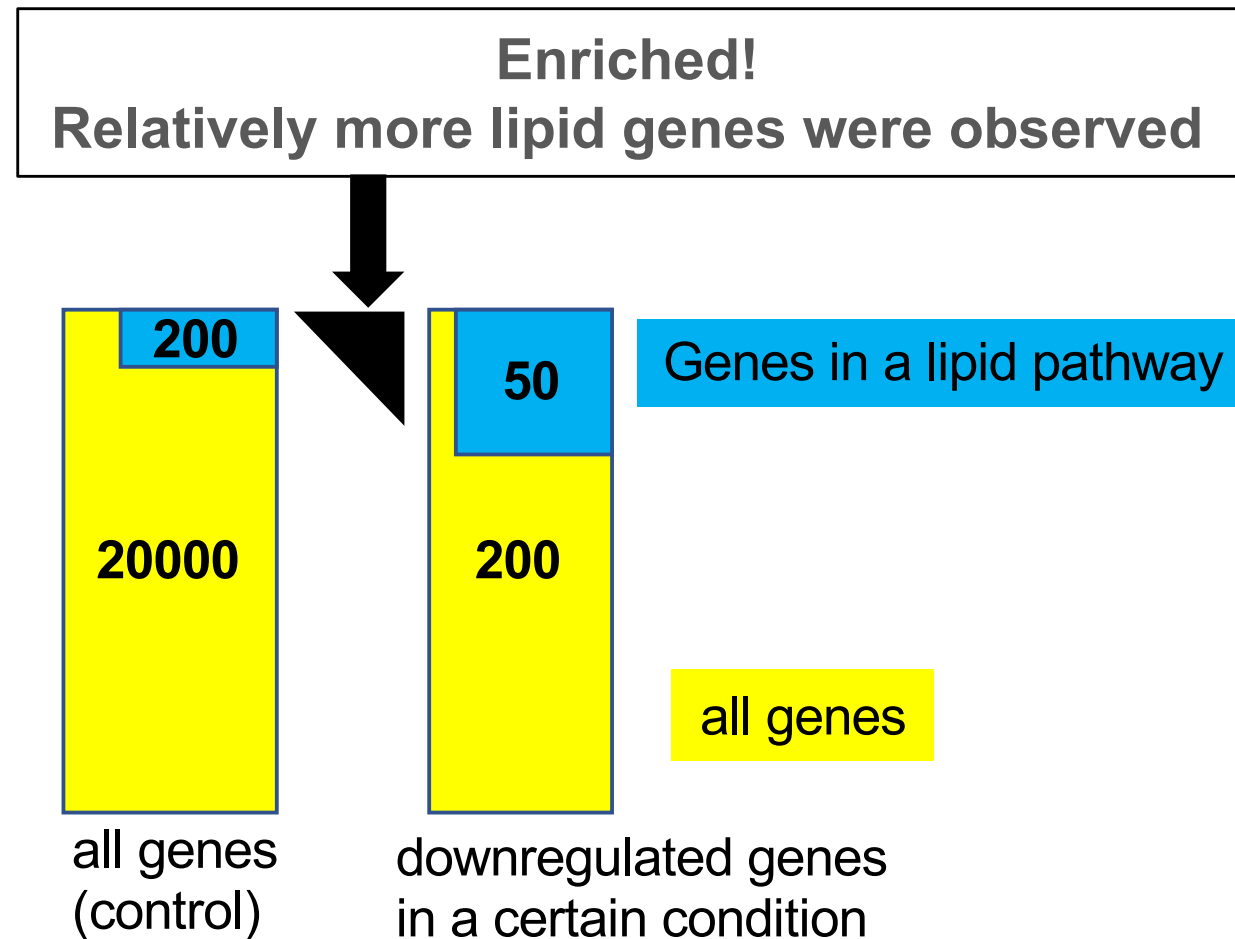
High fold change – high p-value

Observed expression of gene B



Low fold change – low p-value ⁷

Functional categorization of genes (Gene Ontology Analysis)



Transcripts Per Million (TPM)

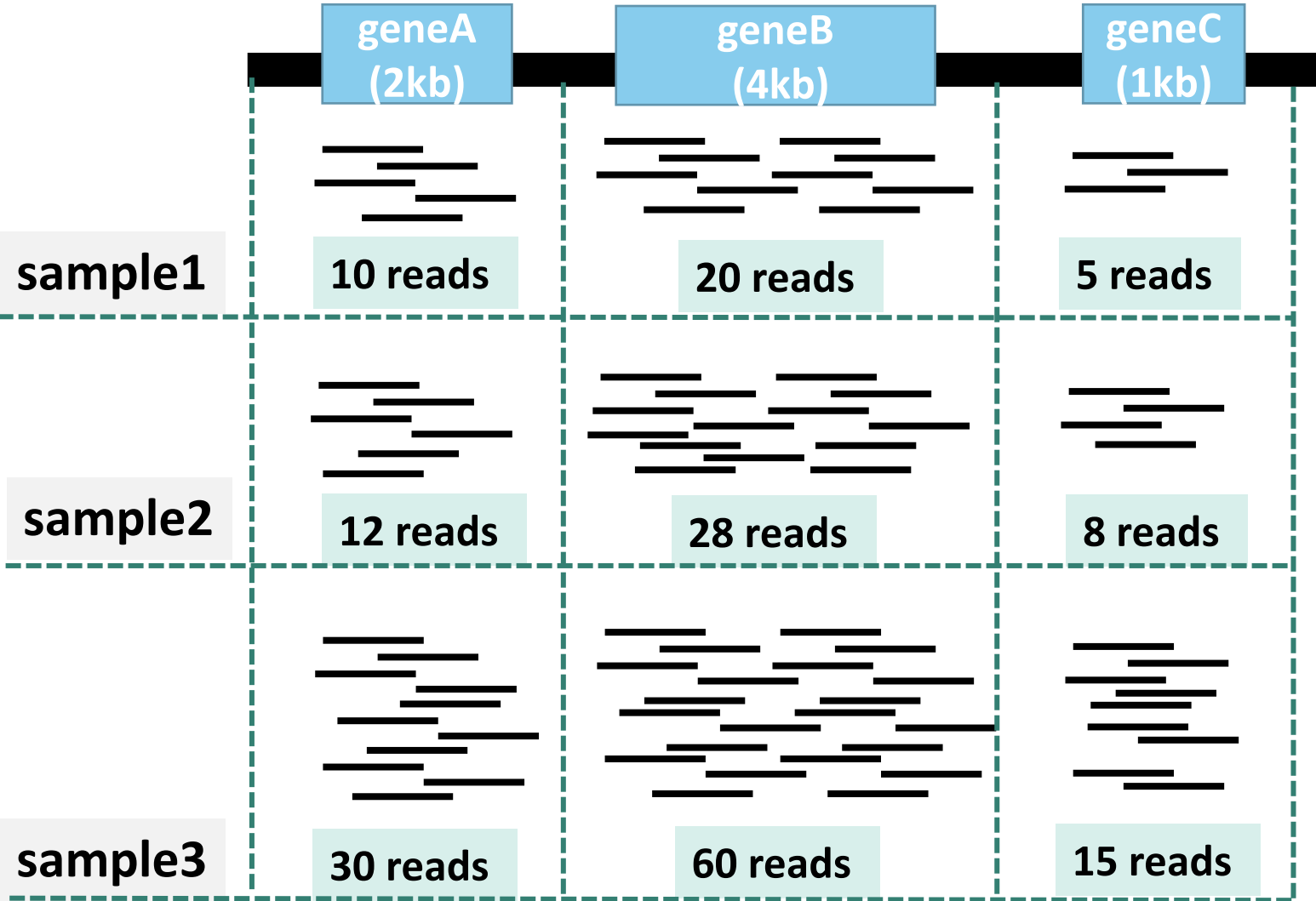
Unit of gene expression

Let's assume that we got the following read count table.

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)
sample1	10	20	5
sample2	12	25	8
sample3	30	60	15

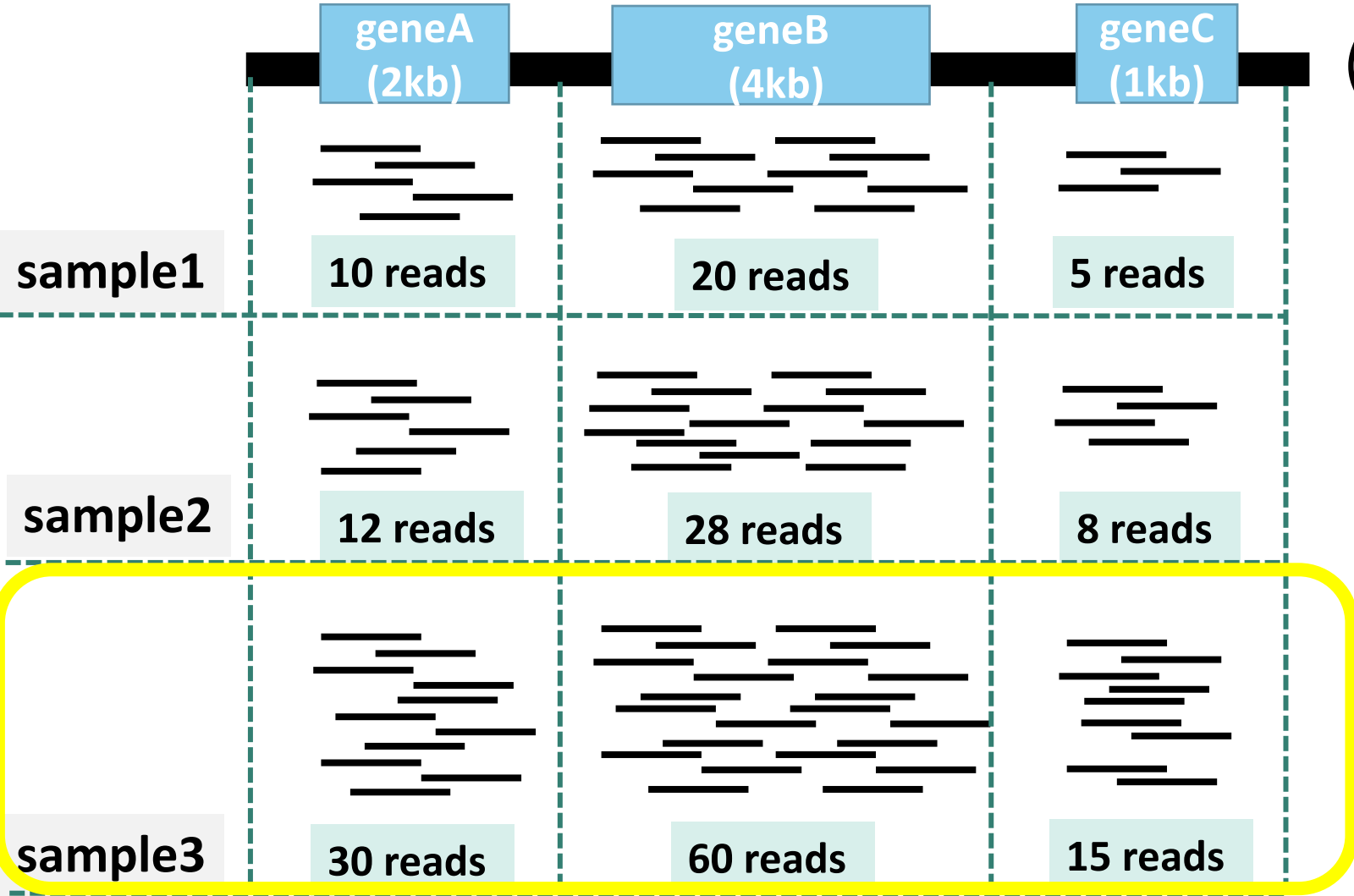
<https://www.youtube.com/watch?v=TTUrtCY2k-w&t=7s>

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)
sample1	10	20	5
sample2	12	25	8
sample3	30	60	15



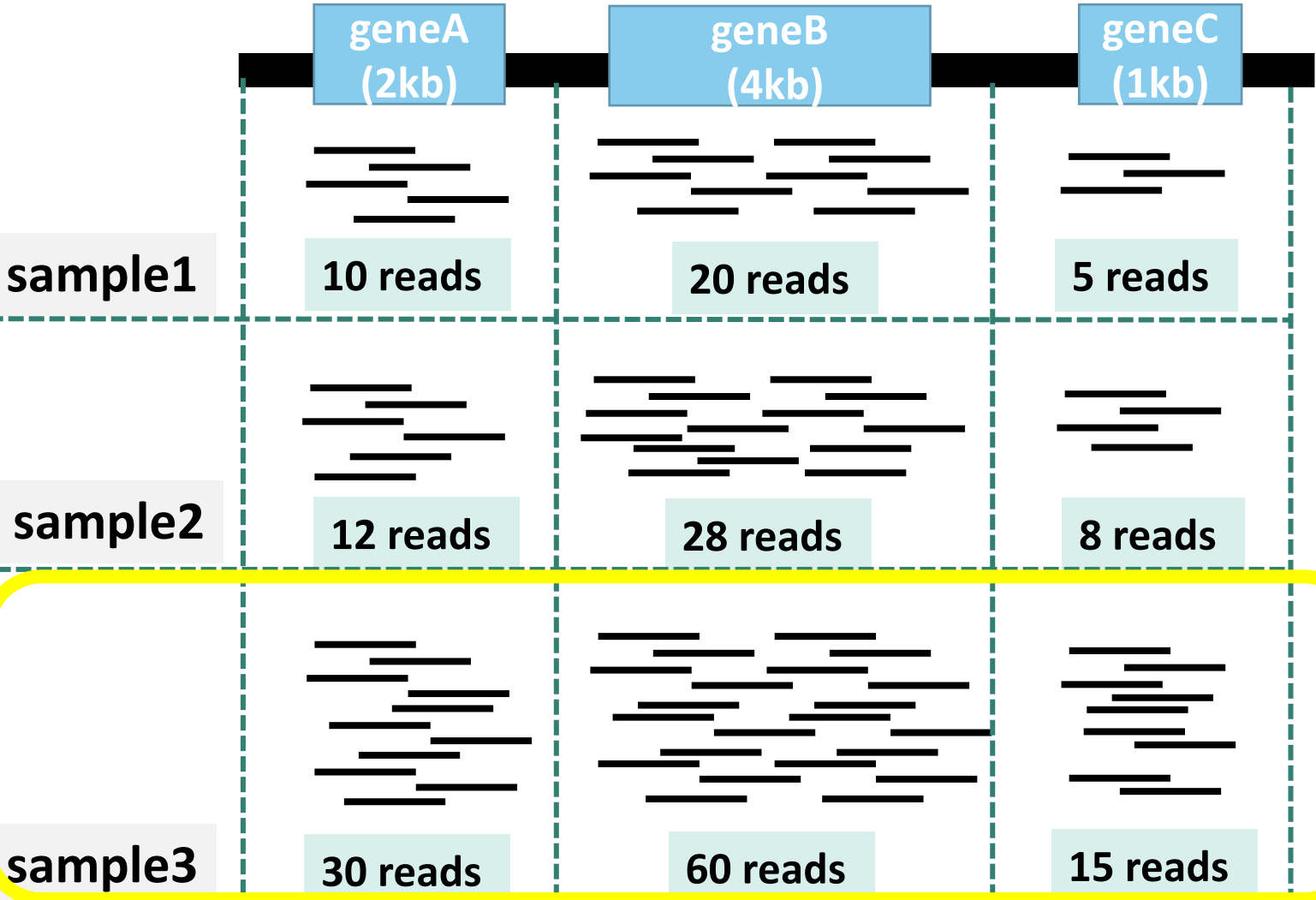
(1) Normalize the read counts by gene length

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)	sum
sample1	5	5	5	15
sample2	6	6.25	8	20.25
sample3	15	15	15	45



(2) Normalize the read counts by total read counts

gene	geneA (2kb)	geneB (4kb)	geneC (1kb)	sum
sample1	0.33	0.33	0.33	1
sample2	0.30	0.31	0.40	1
sample3	0.33	0.33	0.33	1



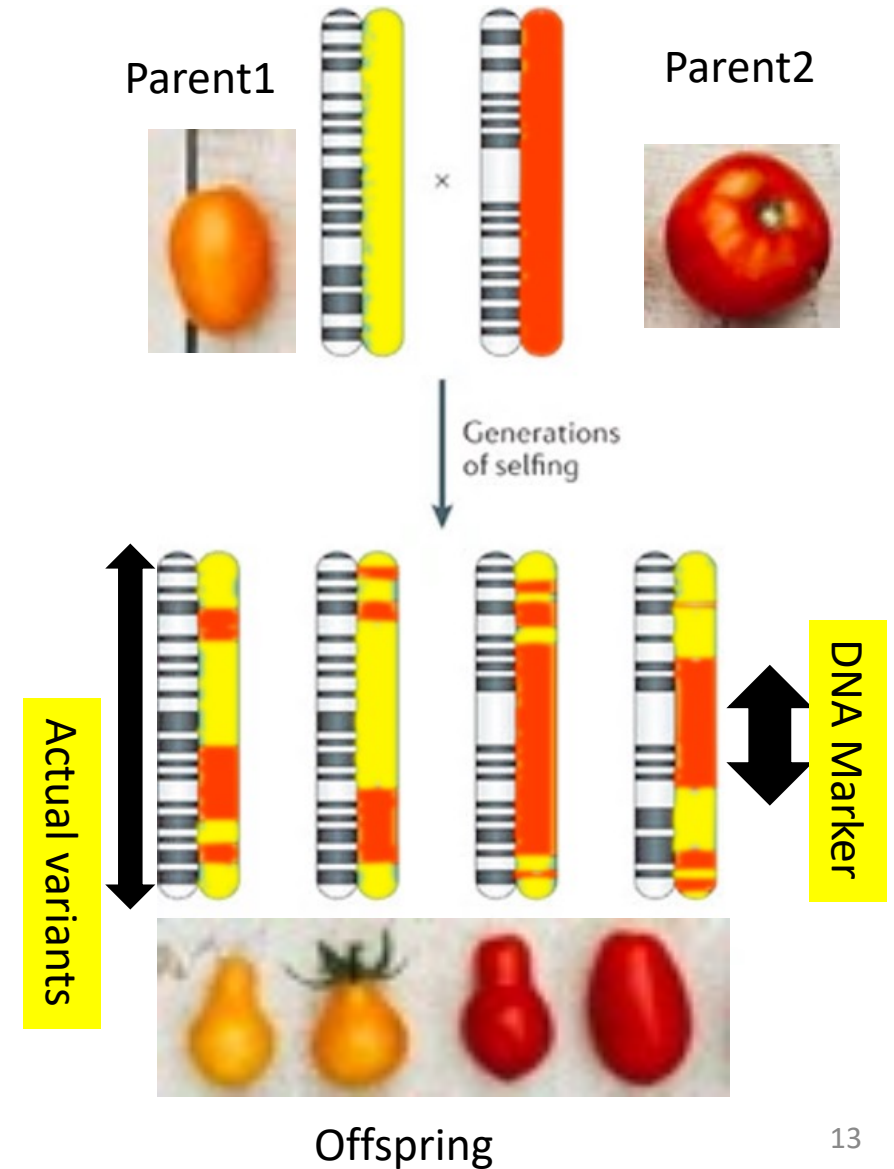
(3) total read counts will be one
(In reality, 1 Million)
= transcripts per million

eQTL (expression quantitative trait locus) analysis

historically... **QTL analysis:**

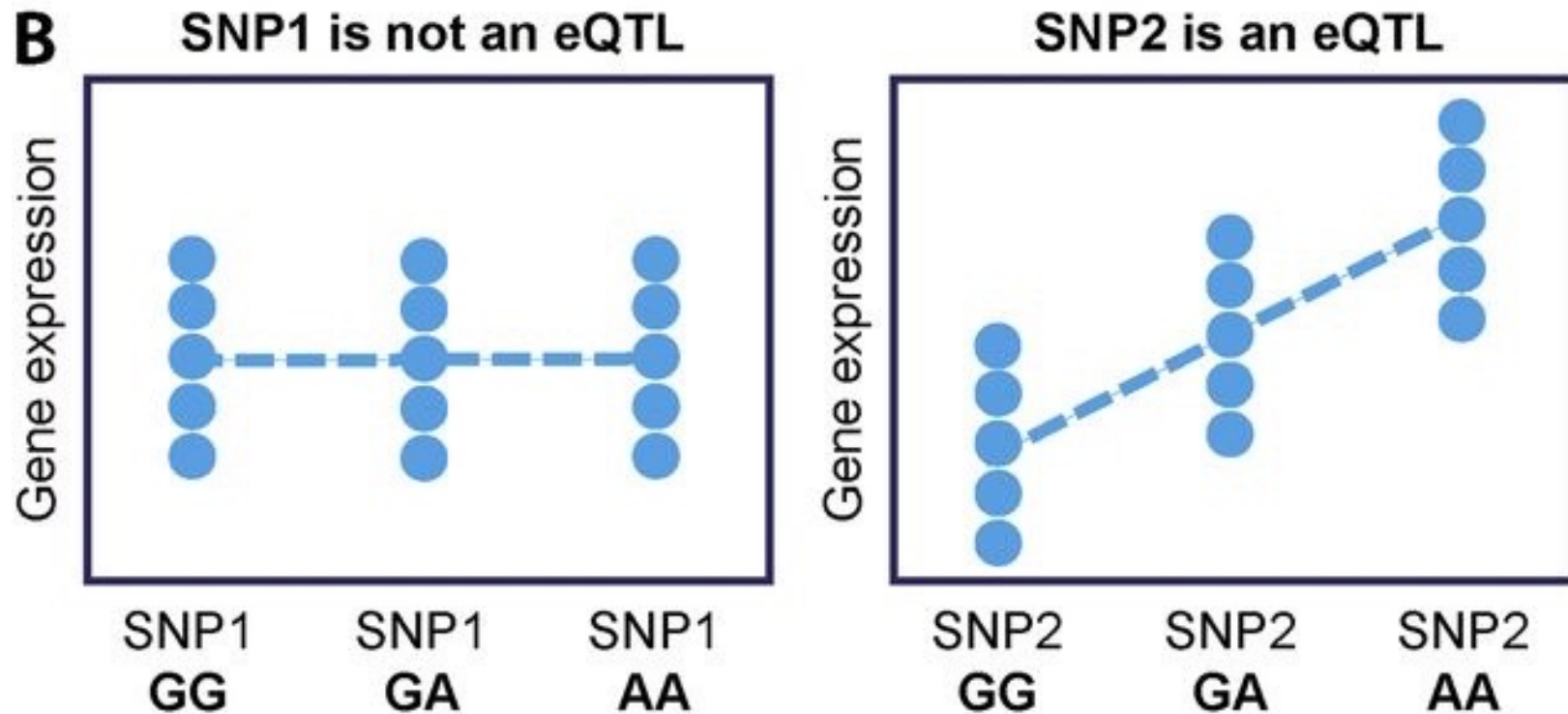
Conventional analysis based on limited number of genetic markers

- (1) two parental breeds are crossed
- (2) the resulting F_1 generations are self-fertilized for several generations, resulting in inbred lines
- (3) use genetic markers (100-) to distinguish between parental lines
- (4) the phenotypes and genotypes of the offspring are scored



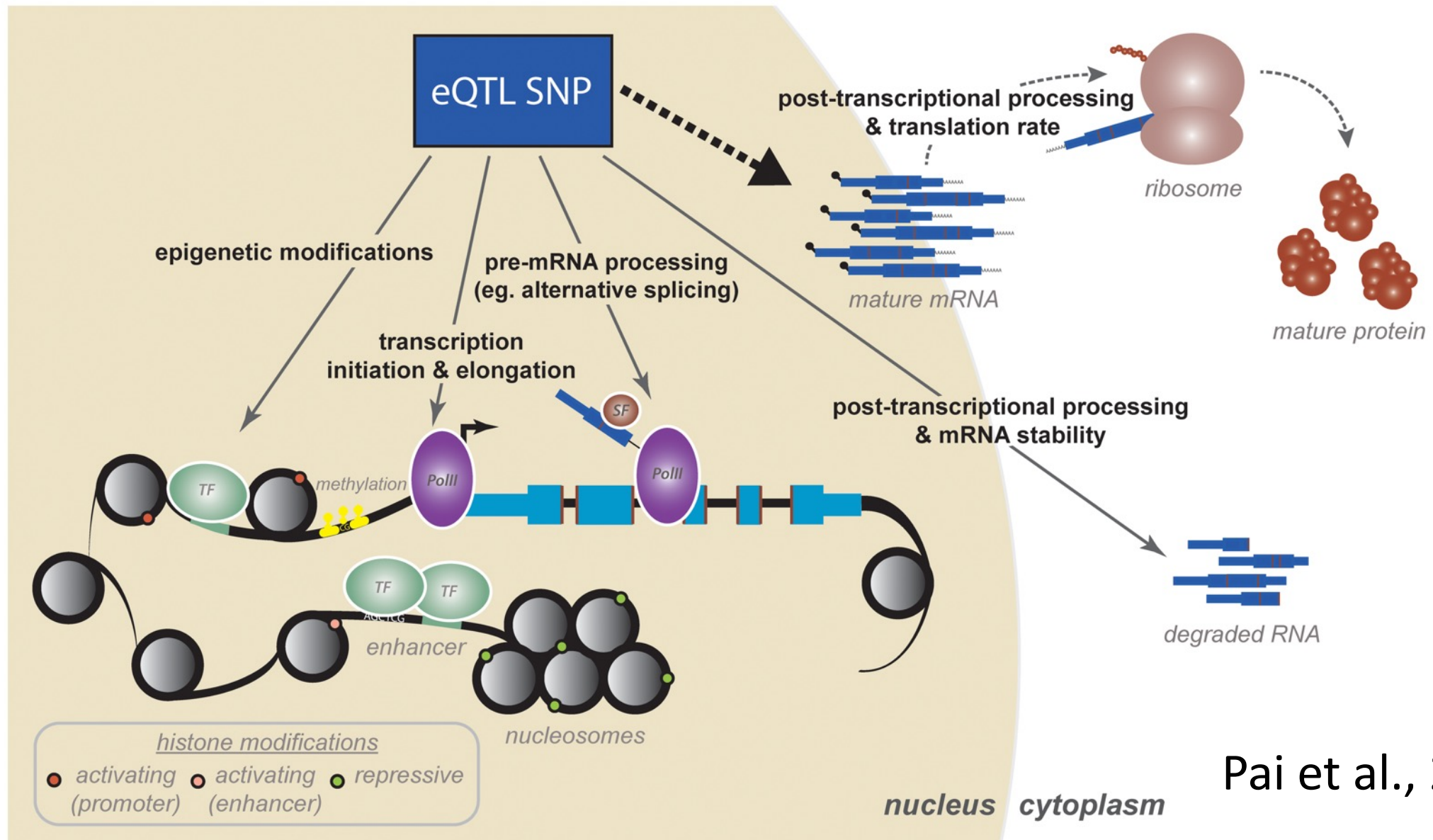
Expression Quantitative Trait Loci analysis

Variant -> **Gene expression** -> Phenotype



J Exp Med (2017)

How eQTL SNPs work

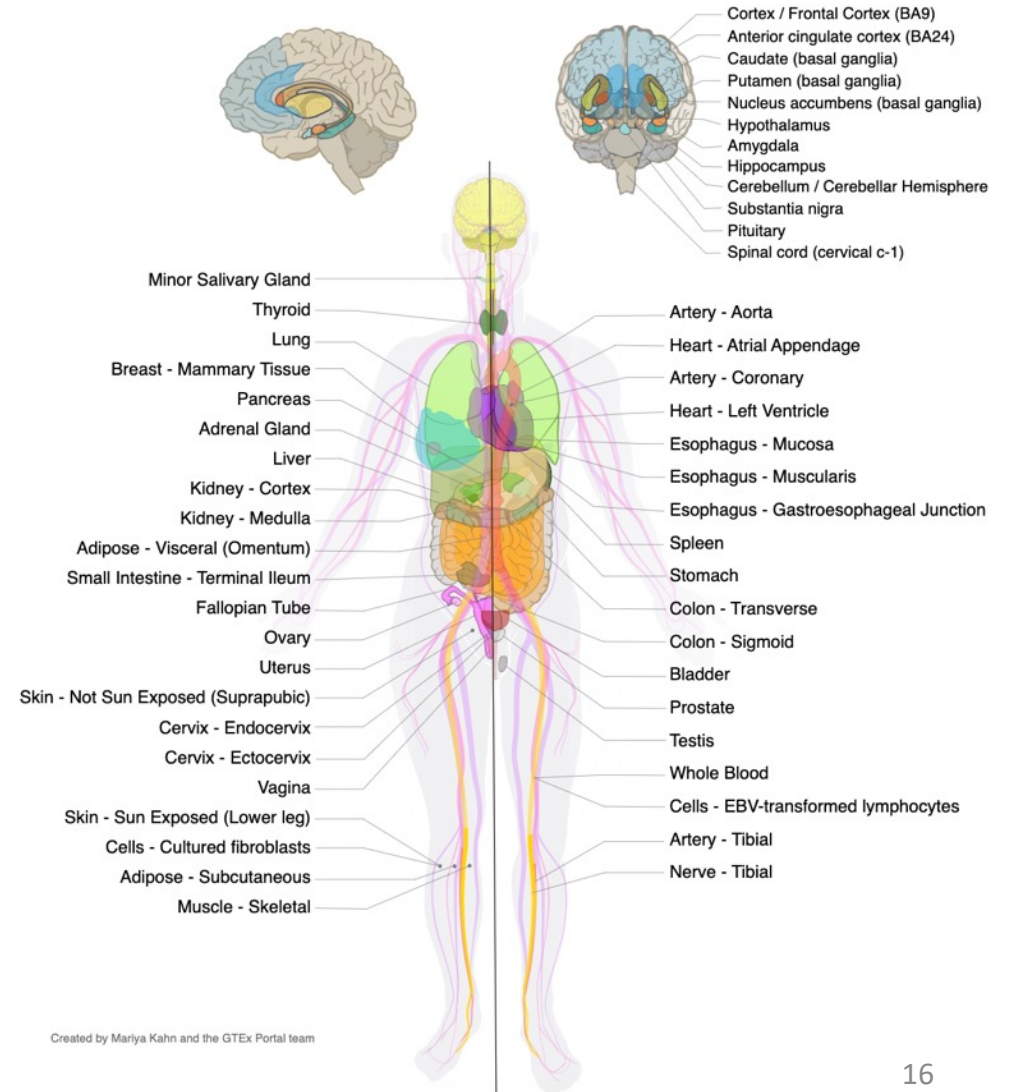


Pai et al., 2015

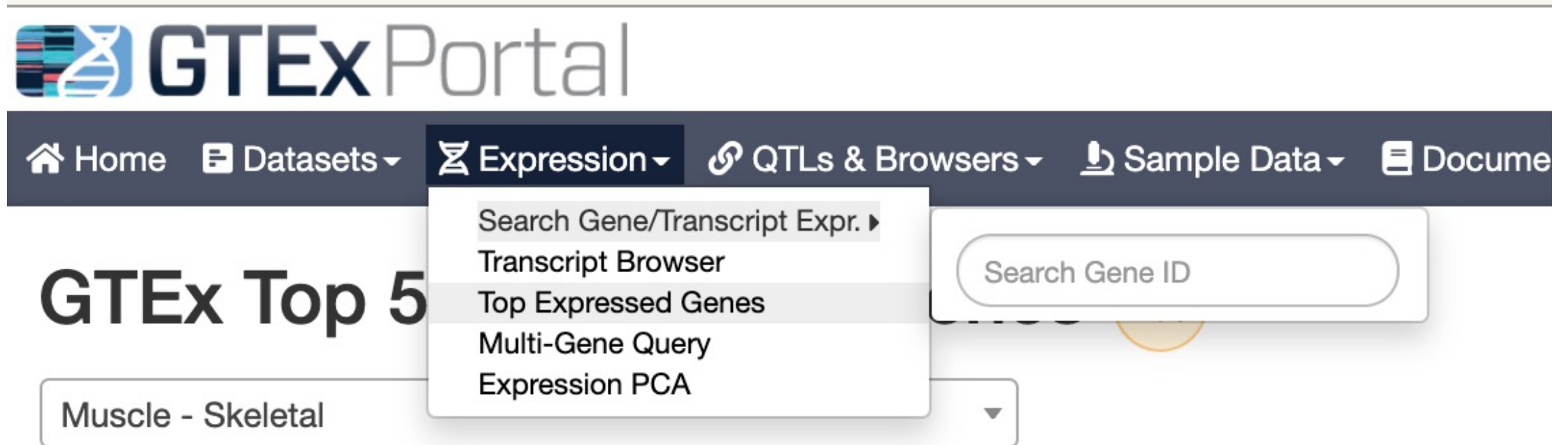
Let's explore the GTEx Portal...

Human Gene expression/splicing/eQTL database

- 54 tissues, 948 Donors, 17382 samples



Let's explore the GTEx Portal...



The screenshot shows the GTEx Portal interface. At the top is the logo with a DNA helix icon and the text "GTEx Portal". Below it is a dark navigation bar with links: Home, Datasets, Expression (selected), QTLs & Browsers, Sample Data, and Documents. A dropdown menu for "Expression" is open, showing options: Search Gene/Transcript Expr. (with a right arrow), Transcript Browser, Top Expressed Genes (highlighted), Multi-Gene Query, and Expression PCA. To the right of the menu is a search box labeled "Search Gene ID". Below the navigation bar, the text "GTEx Top 5" is visible, followed by a dropdown menu currently showing "Muscle - Skeletal".

<https://gtexportal.org/home/>

Example: highly expressed genes in muscle

TPM 0.0 7.1 64 5.2e+2 4.2e+3 3.4e+4 TPM – blue=high expression

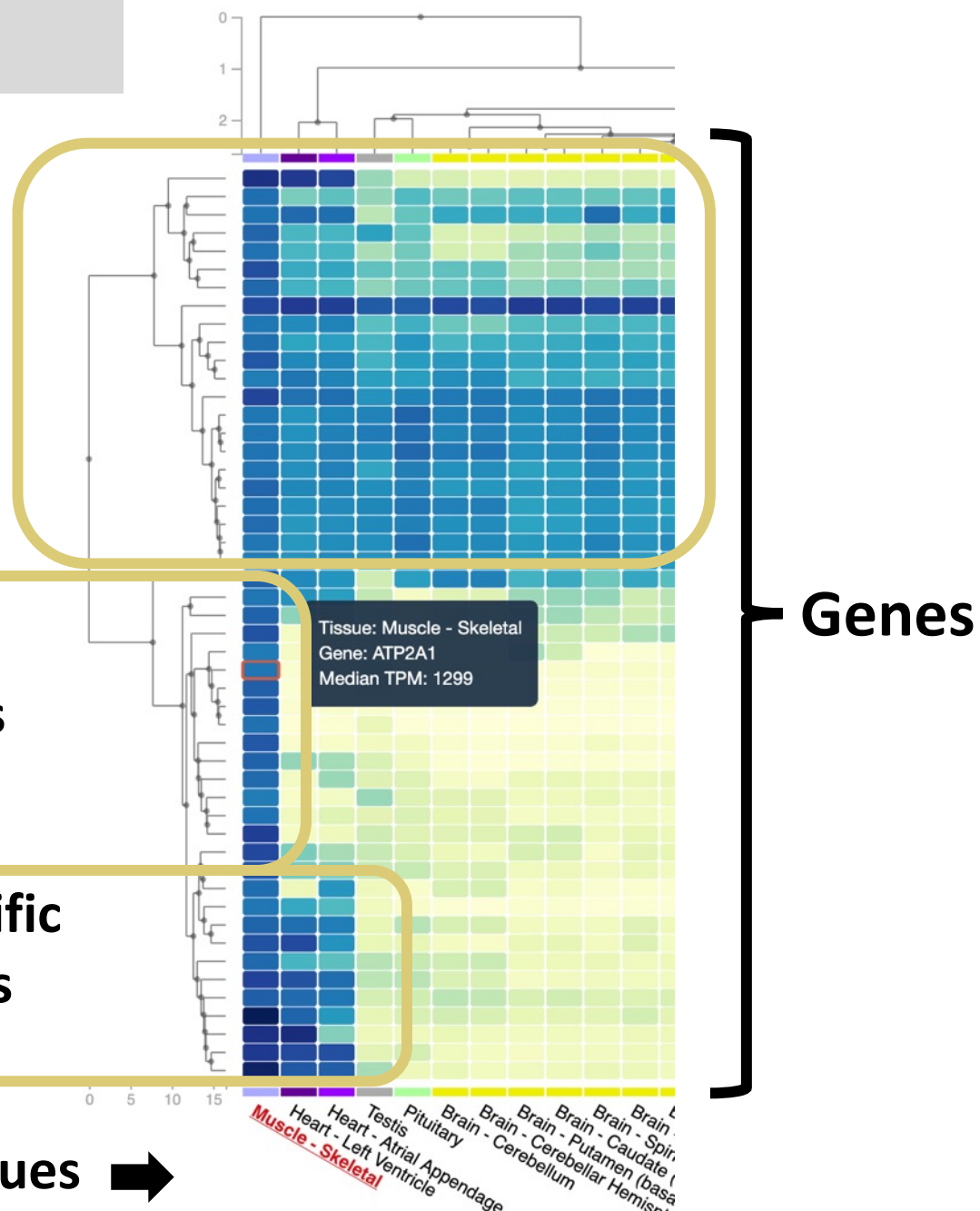
Housekeeping genes
(expressed everywhere)

Muscle-specific
highly expressed genes

Muscle and heart-specific
highly expressed genes

CKM gene

Tissues →





SUBSET

None

Sex

SCALE

Log

Linear

TISSUE SORT



MEDIAN SORT

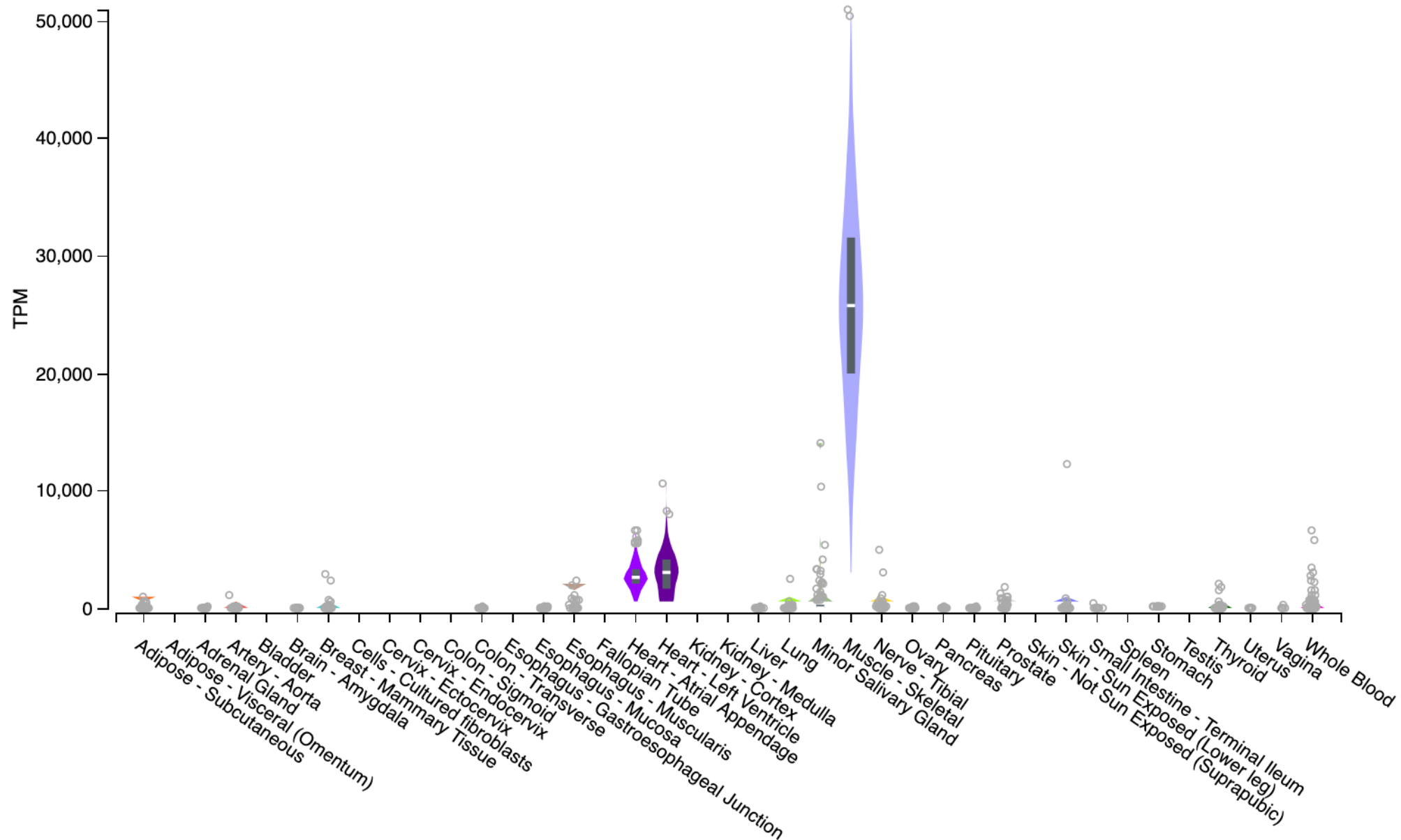


OUTLIERS

On

Off

Gene expression for CKM (ENSG00000104879.4)



Example – CKMT1A gene expression and a variant at chr5:43504700

CopyCSV

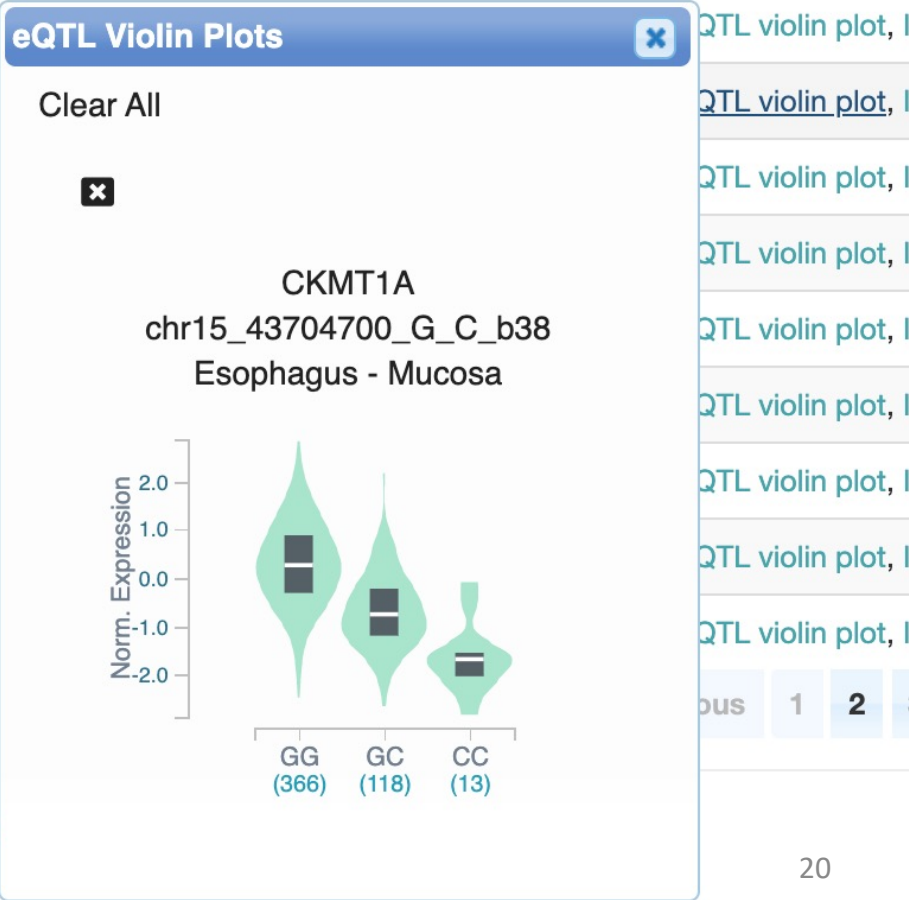
Search:

Gencode Id	Gene Symbol	Variant Id	SNP	P-Value	NES	Tissue	Actions
ENSG00000223572.9	CKMT1A	chr15_43704700_G_C_b38	rs2288768 dbSNP	5.2e-65	-0.88	Esophagus - Mucosa	eQTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43706786_A_G_b38	rs199879510 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43709310_A_G_b38	rs200795659 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43696406_G_A_b38	rs28578041 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43707916_G_C_b38	rs74831743 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43728977_T_C_b38	rs12441127 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43719545_T_C_b38	rs8023696 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43720567_C_T_b38	rs11856818 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43720983_G_A_b38	rs8029210 dbSNP				QTL violin plot, I
ENSG00000223572.9	CKMT1A	chr15_43733660_G_C_b38	rs7171750 dbSNP				QTL violin plot, I

Showing 1 to 10 of 7,303 entries

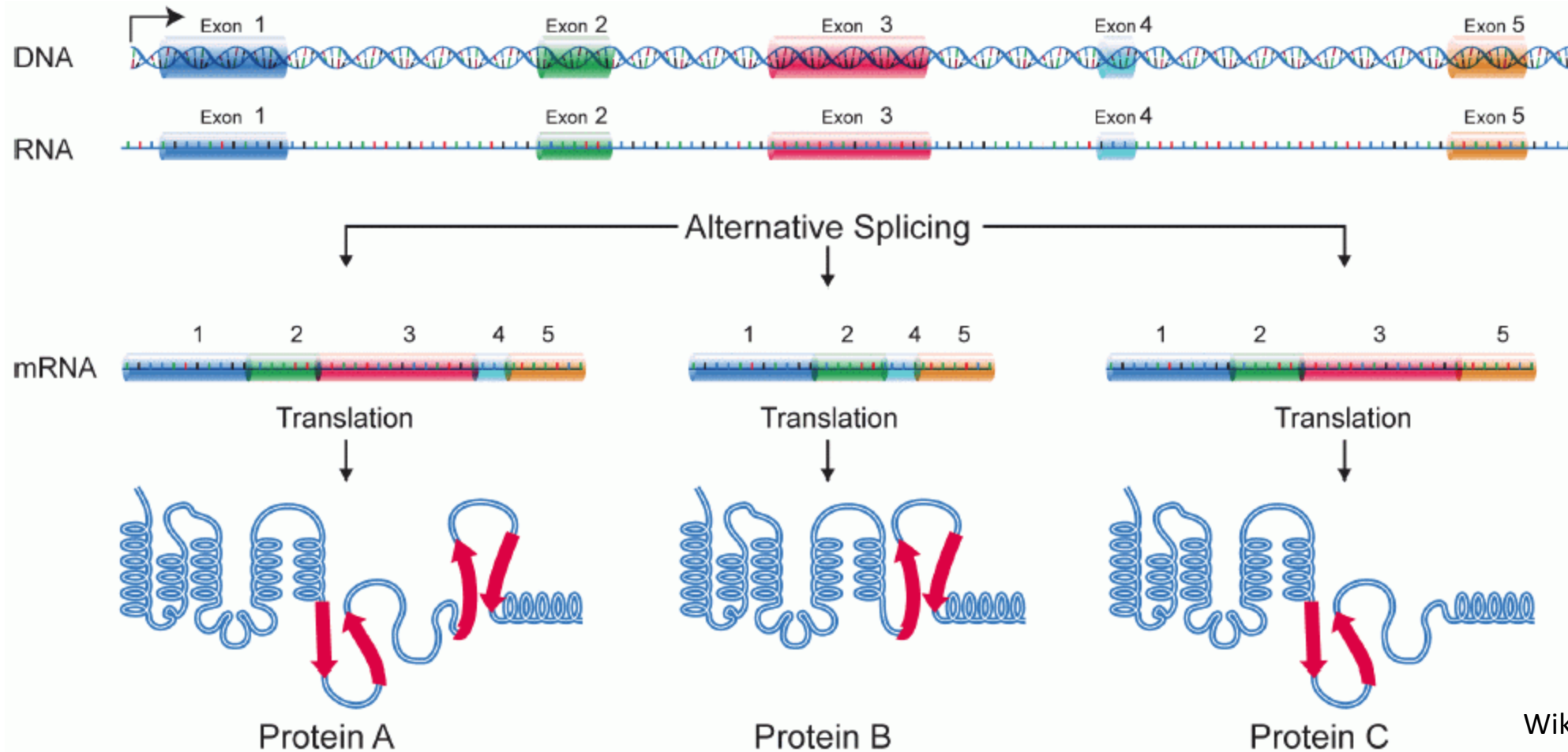
Significant Single-Tissue sQTLs for CKMT1A (ENSG00000223572.9) in all tissue

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



sQTL – splicing QTL

Alternative Splicing allows a single gene to code for multiple proteins



Wikipedia

Hands-on exercise