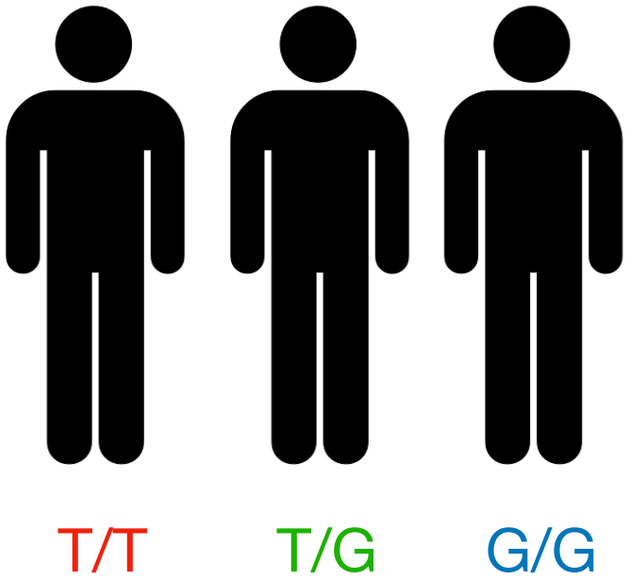
expression quantitative trait loci (eQTL)

mapping & interpretation

Ankeeta Shah
Li Lab, GGSB

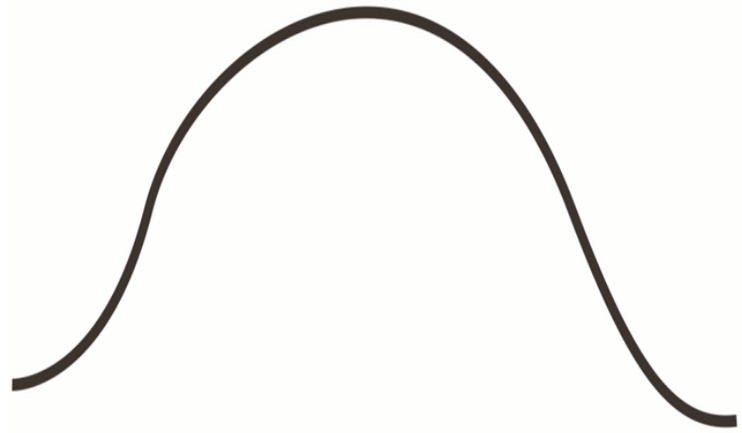
HGEN 47000 Human Genetics I
Computational Workshop II
November 20, 2019

Understanding how genetic variation contributes to phenotypic variation in humans



Genotype

Phenotype



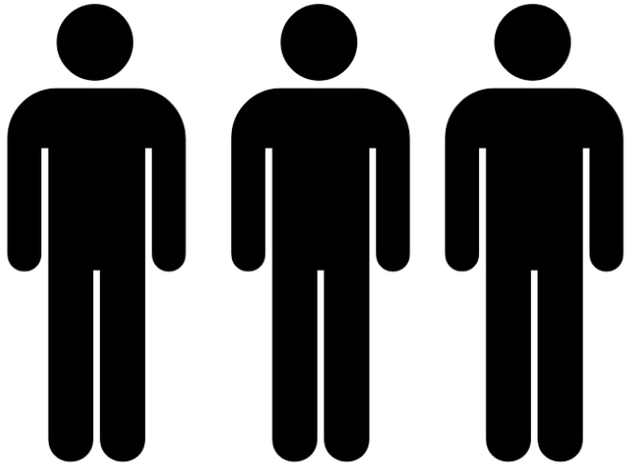
Why should we map quantitative trait loci (QTL)?

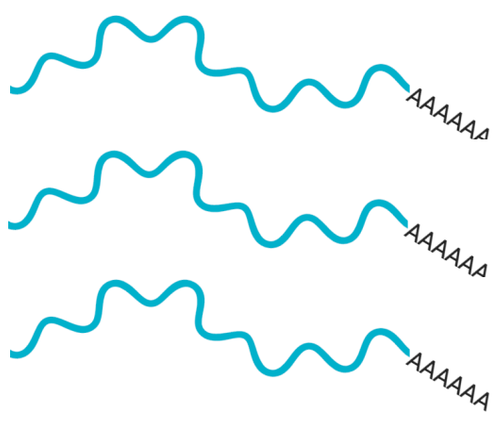
Why should we map QTL?

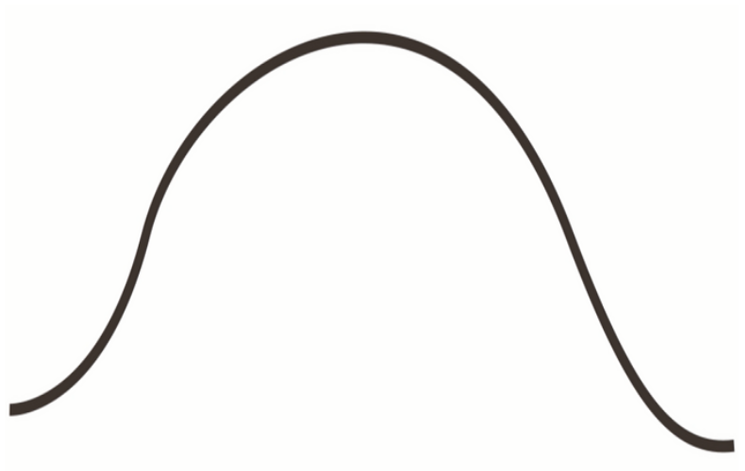
quantitative trait loci (QTL): regions of the genome that affect the levels of a heritable quantitative trait

Genotype → Molecular phenotype → Phenotype

e.g. height, CAD, etc

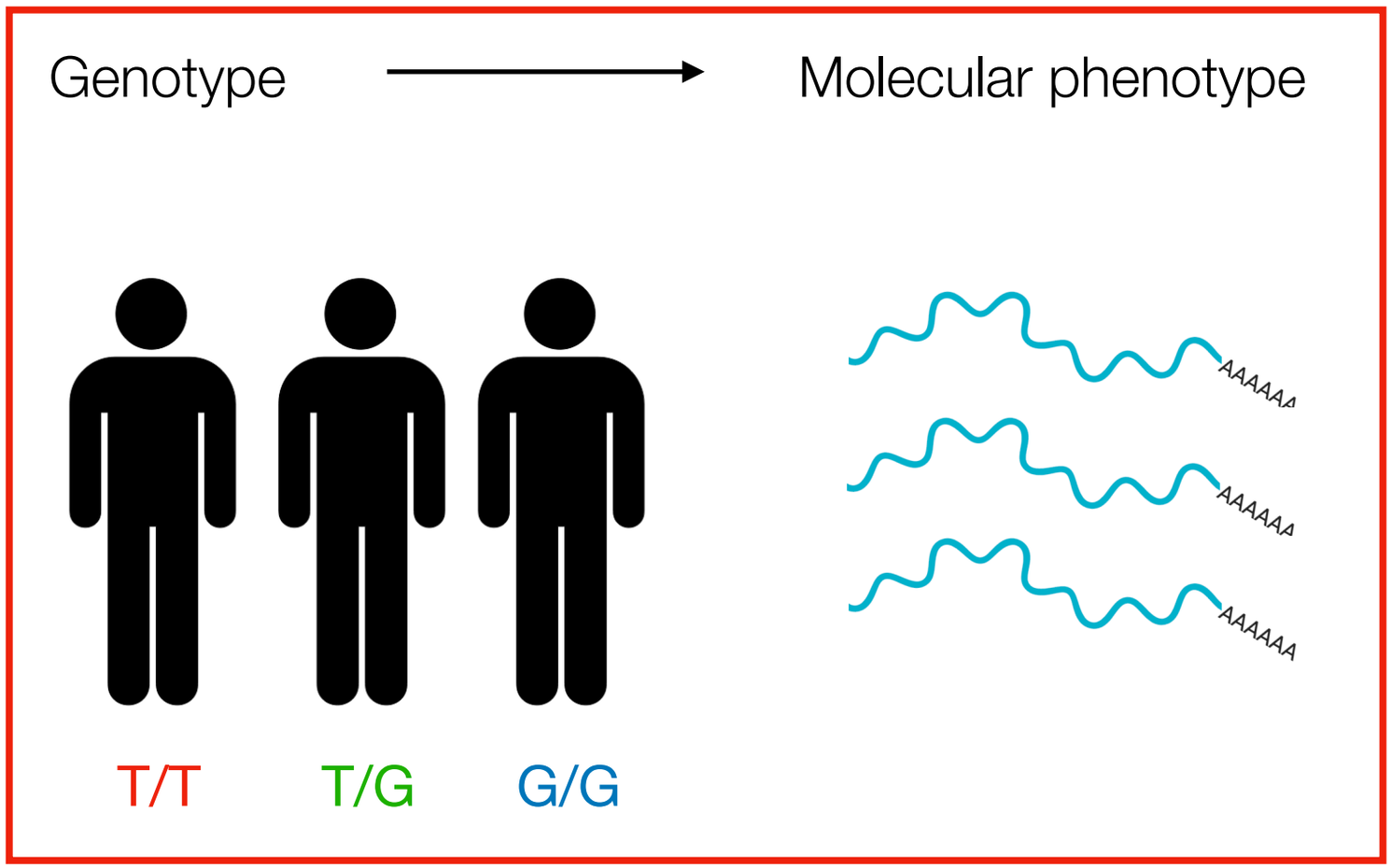






Why should we map QTL?

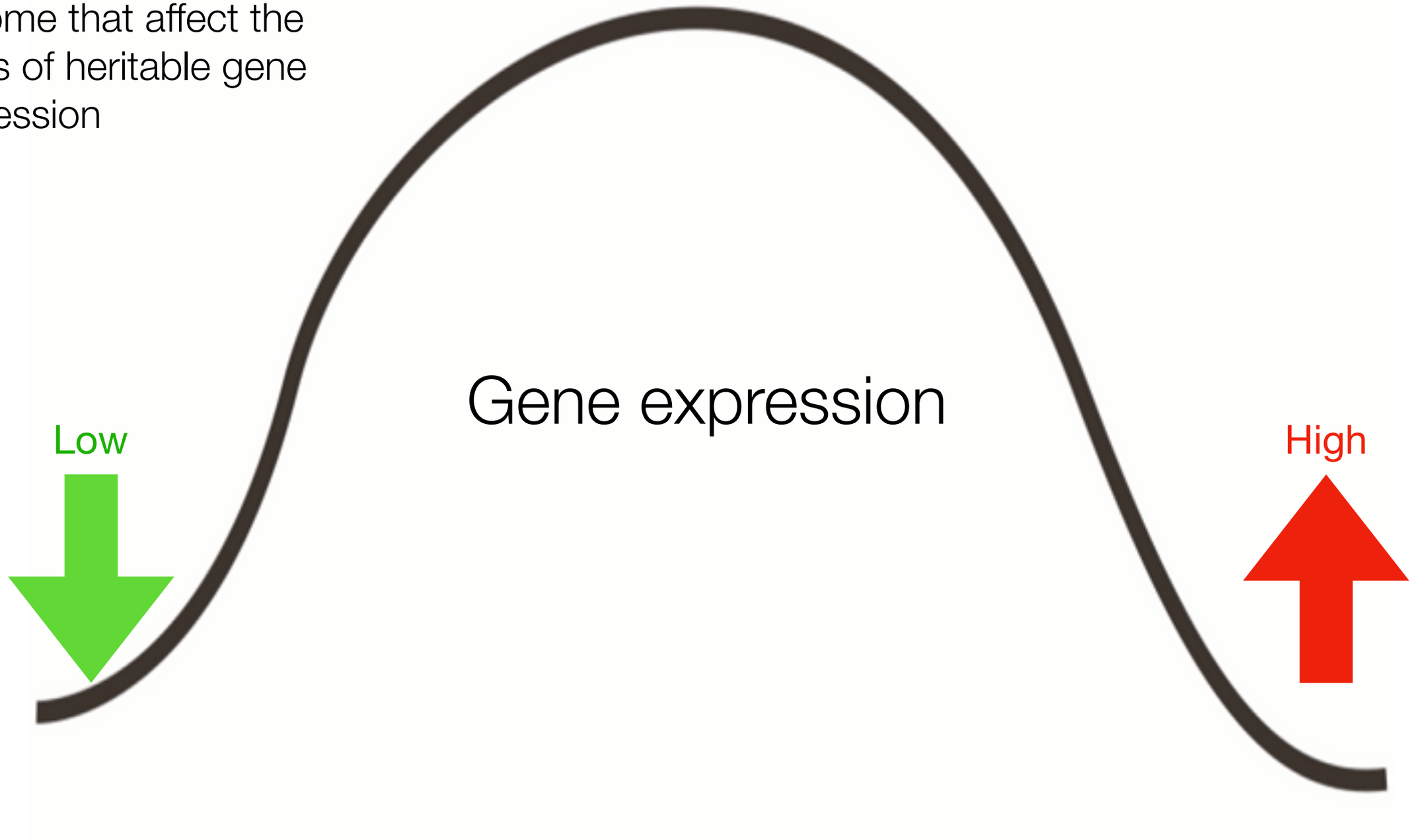
quantitative trait loci (QTL): regions of the genome that affect the levels of a heritable quantitative trait



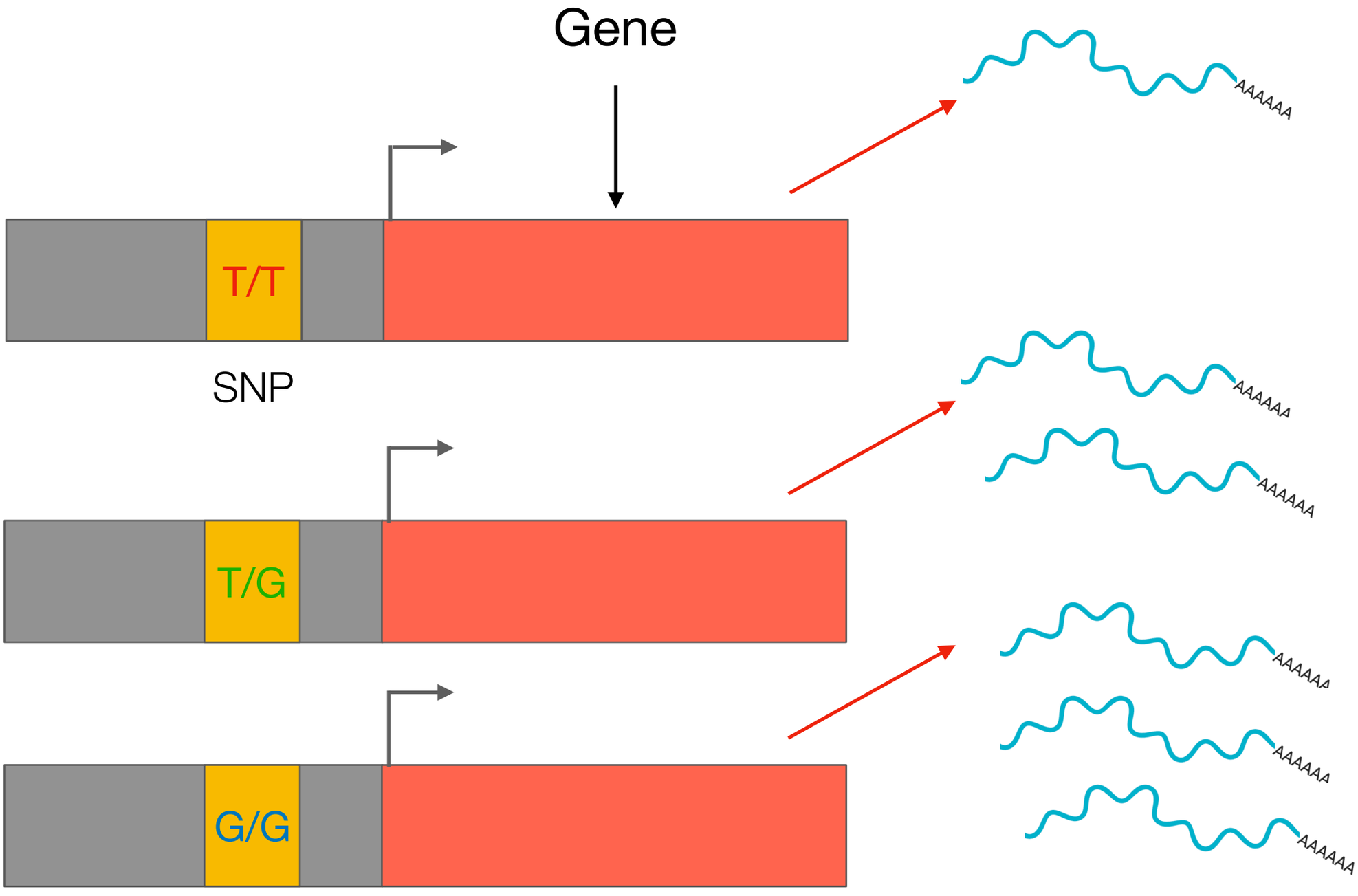
Why should we map expression QTL (eQTL)?

Gene expression is continuous (i.e. “quantitative trait”)

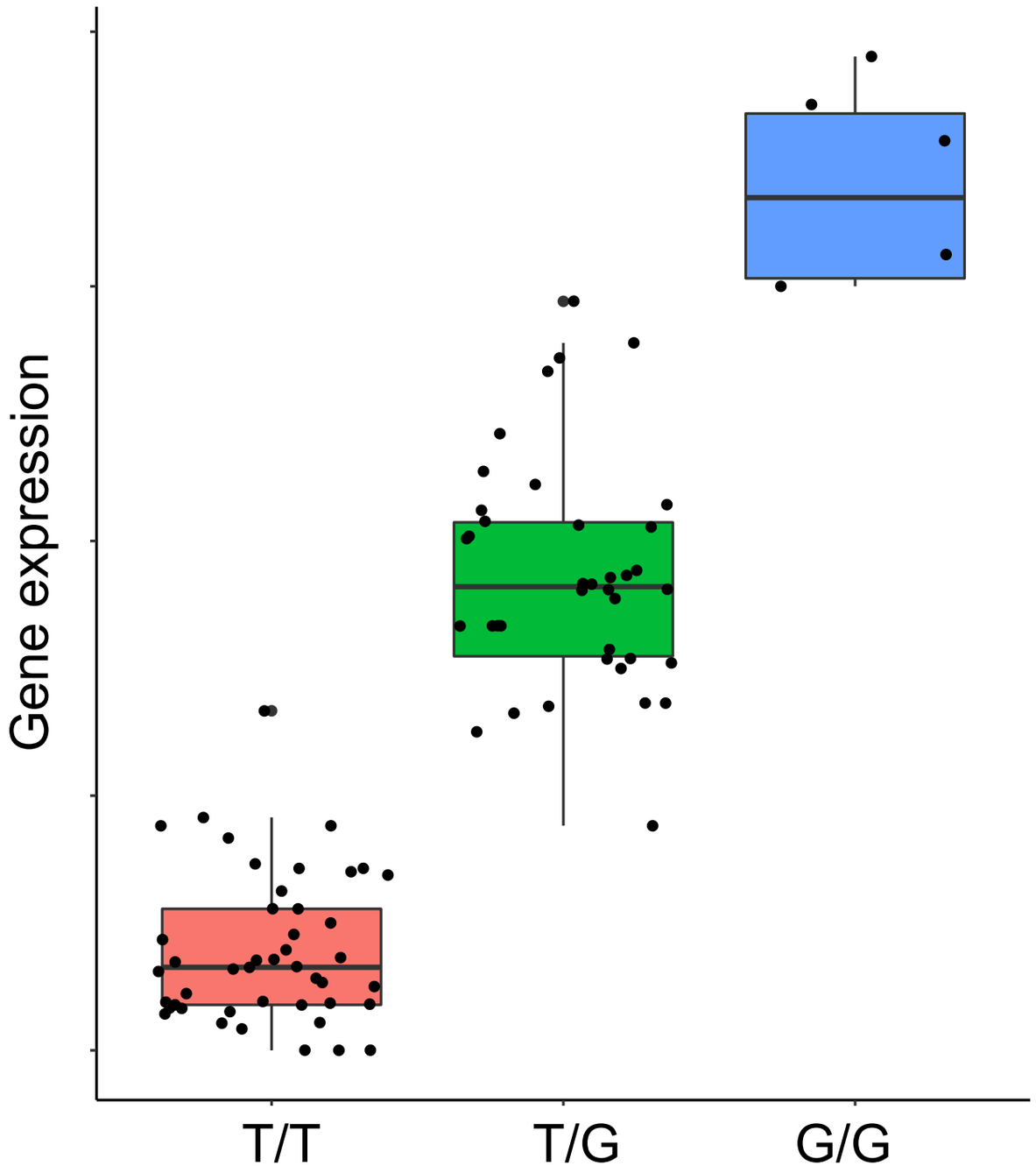
eQTLs: regions of the genome that affect the levels of heritable gene expression



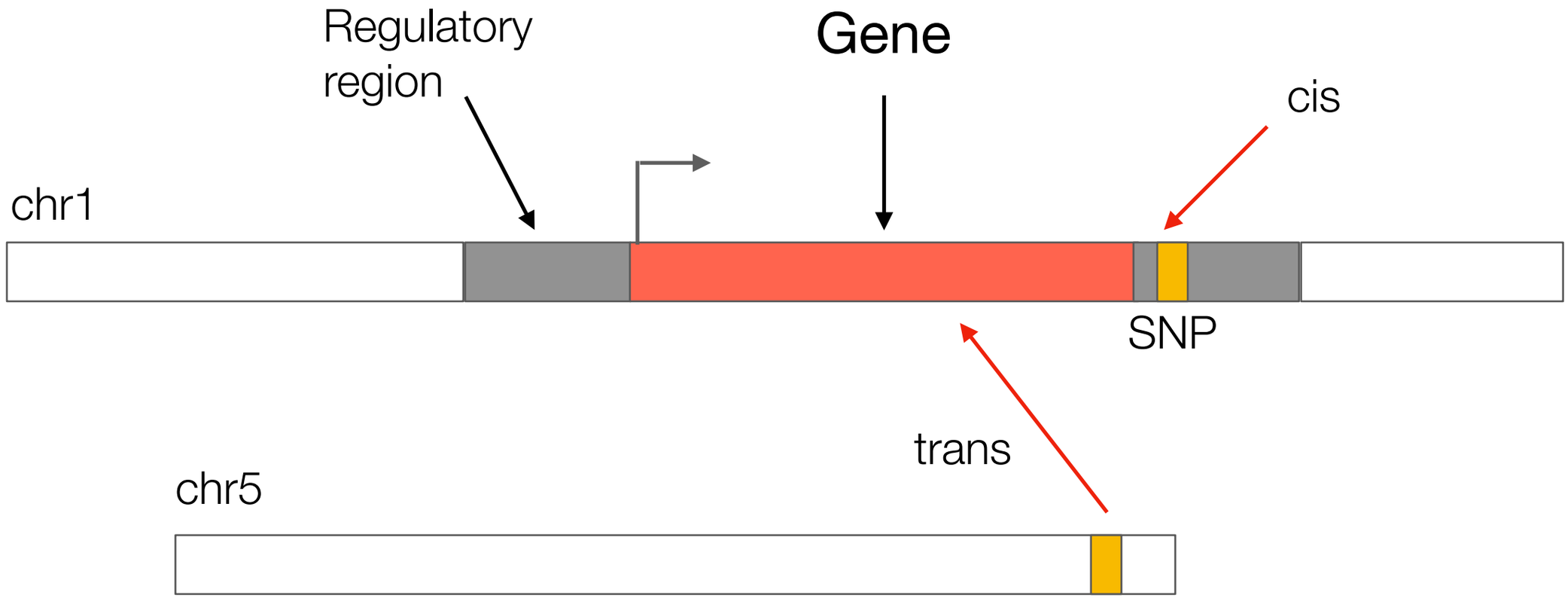
eQTL mapping



eQTL mapping



cis- vs trans-eQTL mapping



cis: test all SNPS within 1MB of the gene

trans: test all other SNPs in the genome

GEUVADIS* Consortium data

Entire dataset: Lymphoblastoid cell lines (LCLs)

Individuals from the 1000 Genomes Project
Across 5 populations:
CEPH (CEU),
Finns (FIN),
British (GBR),
Toscani (TSI),
and Yoruba (YRI))

Total of ~500 individuals with genotypes and RNA-seq

We are going to be working with the Yoruba (YRI) individuals (N = 89)

*Lappalainen et al., *Nature,* 2013

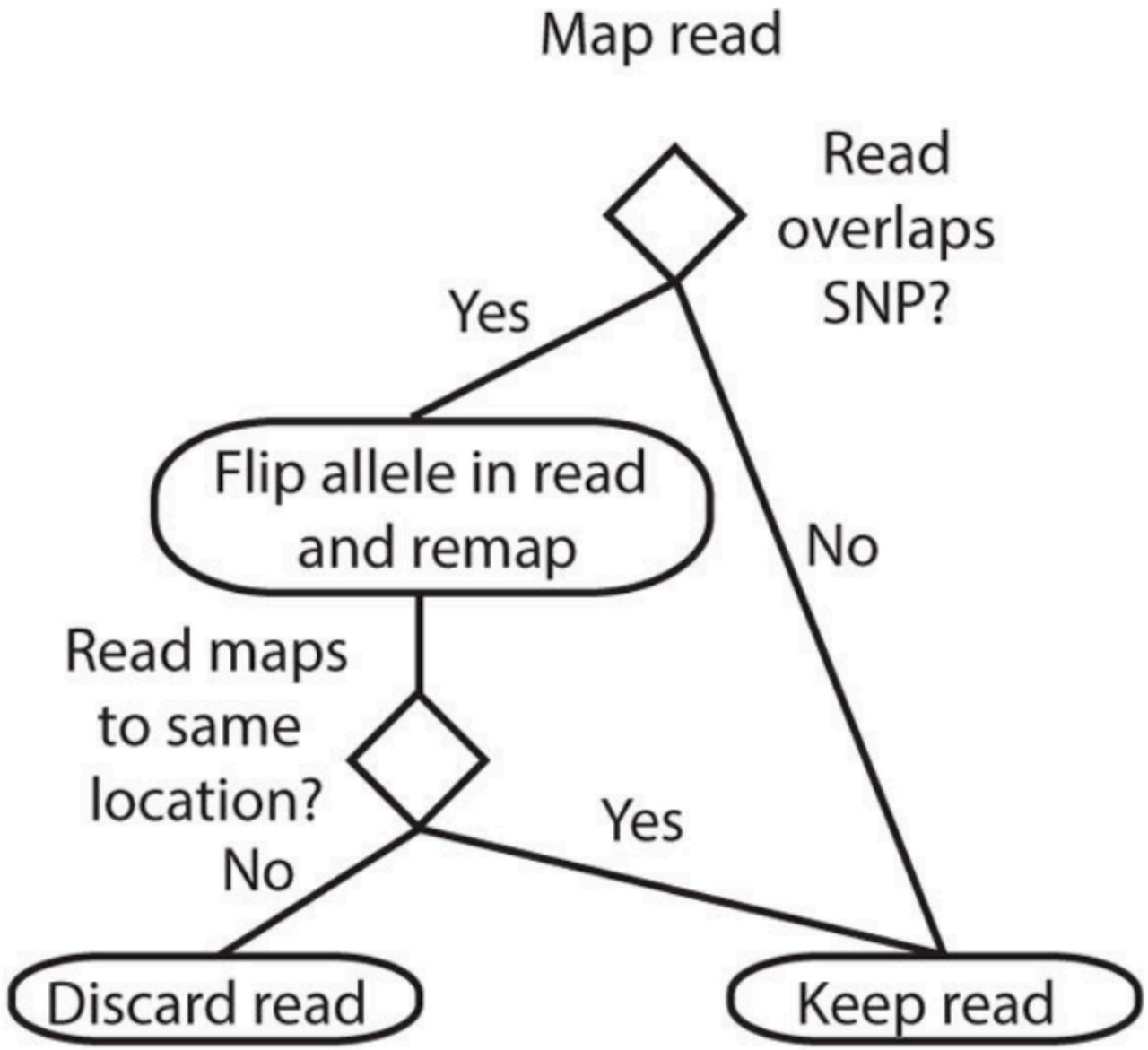
You can find the raw data here: https://www.internationalgenome.org/data-portal/data-collection/geuvadis

Data processing steps

1. **Pre-processing**:
    1. Filtering allele-specific biases
    2. Normalizing gene counts (transcripts per million, TPM)
2. **Normalization**:
    1. Quantile normalize the gene expression data (i.e. make an adjustment such that the expression data is normally distributed).
3. **Association**:
    1. Test for association between SNPs and gene expression, while adjusting for covariates
4. **Significance testing**:
    1. Permutations
    2. Multiple testing correction

Filtering allele-specific biases

Read alignment with STAR (Dobin et al., *Bioinformatics,* 2013). Note: we have to be careful to remove allele-specific mapping biases (to avoid false positives).



van de Geijn et al., *Nature Methods*, 2016

Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

2. Normalize counts for gene length (divide rows by length in kb)

Genes	Individuals		
5 kb	2	20	1
7 kb	4	35	3
6 kb	11	32	2

Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

2. Normalize counts for gene length (divide rows by length in kb)

Genes	Individuals		
5 kb	2	20	1
7 kb	4	35	3
6 kb	11	32	2

→

Genes	Individuals		
	2/5	20/5	1/5
	4/7	35/7	3/7
	11/6	32/6	2/6

Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

2. Normalize counts for gene length (divide rows by length in kb)

Genes	Individuals		
5 kb	2	20	1
7 kb	4	35	3
6 kb	11	32	2

→

Genes	Individuals		
	2/5	20/5	1/5
	4/7	35/7	3/7
	11/6	32/6	2/6

→

Genes	Individuals		
	0.4	4	0.2
	0.57	5	0.43
	1.8	5.3	0.33

Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

2. Normalize counts for gene length (divide rows by length in kb)

Genes	Individuals		
5 kb	2	20	1
7 kb	4	35	3
6 kb	11	32	2

→

Genes	Individuals		
	2/5	20/5	1/5
	4/7	35/7	3/7
	11/6	32/6	2/6

→

Genes	Individuals		
	0.4	4	0.2
	0.57	5	0.43
	1.8	5.3	0.33

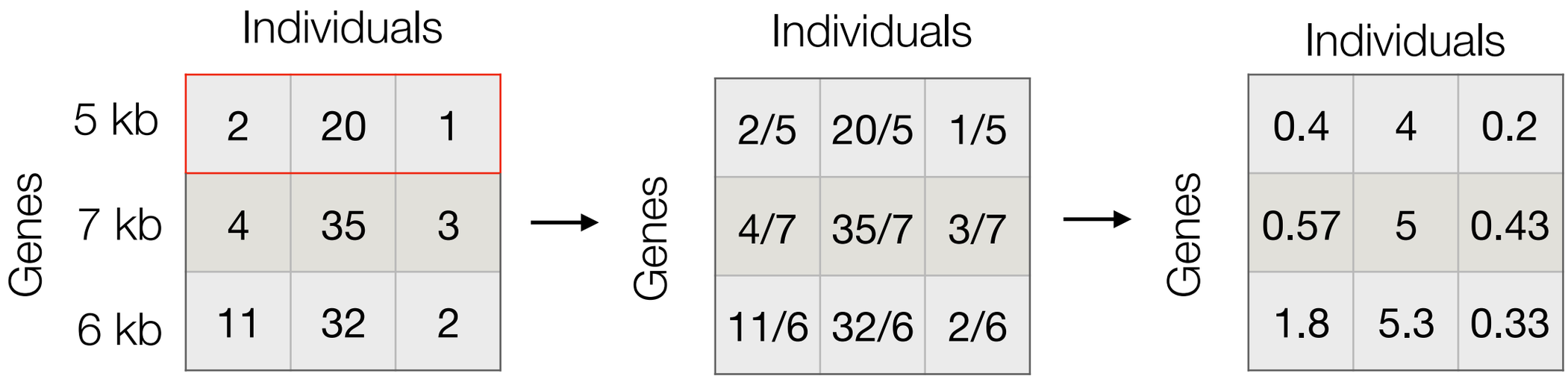
3. Normalize counts for sequencing depth (divide columns by sum columns)

Genes	Individuals		
	0.4	4	0.2
	0.57	5	0.43
	1.8	5.3	0.33
	2.8	14.3	0.96

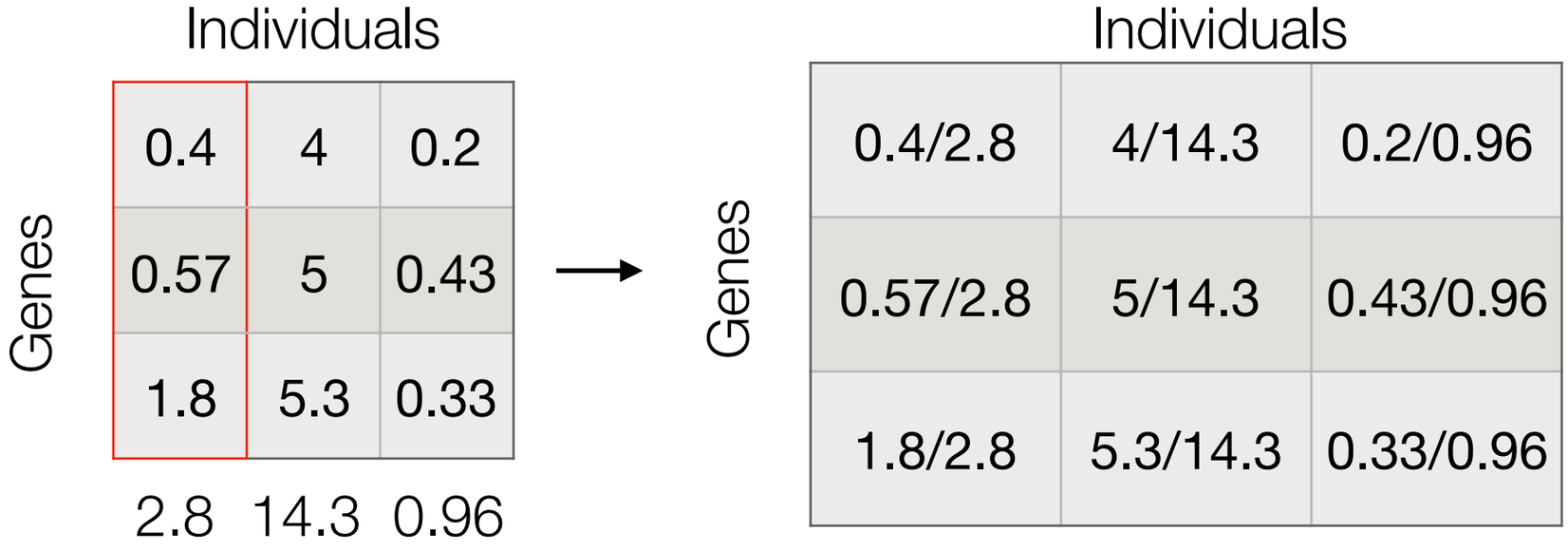
Transcripts per kilobase million (TPM)

1. Count reads overlapping genes (e.g. featureCounts, Liao et al., *Bioinformatics,* 2014)

2. Normalize counts for gene length (divide rows by length in kb)



3. Normalize counts for sequencing depth (divide columns by sum columns)



TPM CODE DEMO

Quantile normalization

There may be **systematic differences** between individuals that we can observe in gene expression data that have nothing to do with genotype.

Normalizing the data (e.g. TPM) will allow us to obtain identical distributions of gene expression across individuals.

	Individual 1	Individual 2	Individual 3	Individual 4
Gene 1	2	20	1	7
Gene 2	4	35	3	10
Gene 3	4	32	3	9
Gene 4	3	30	3	8
Gene 5	3	30	3	8

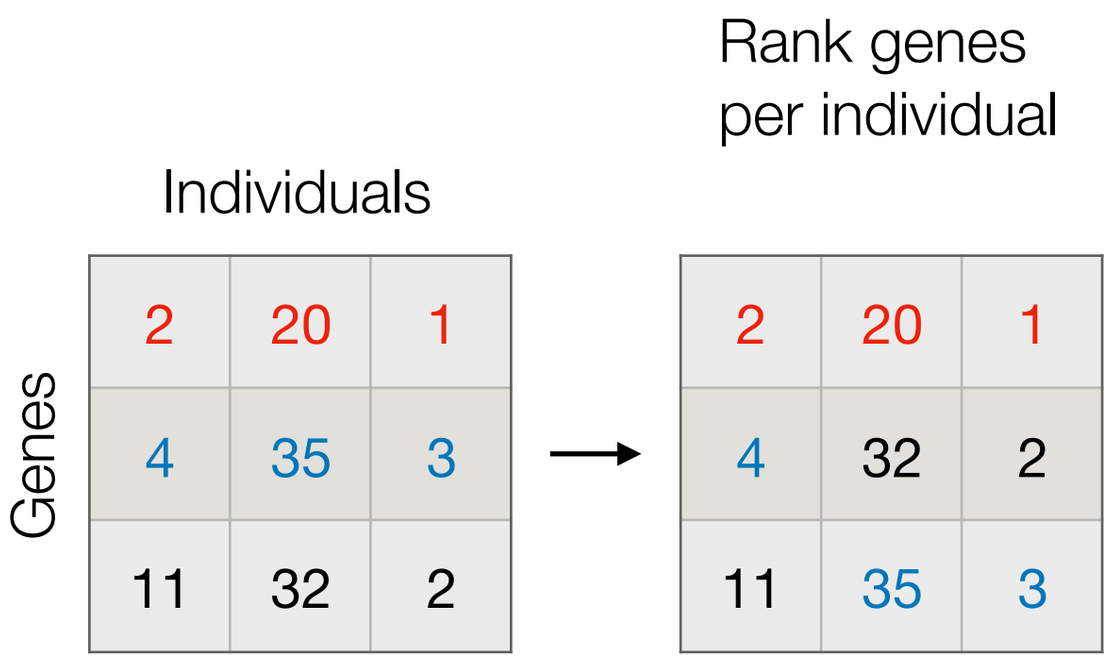
Quantile normalization

Individuals

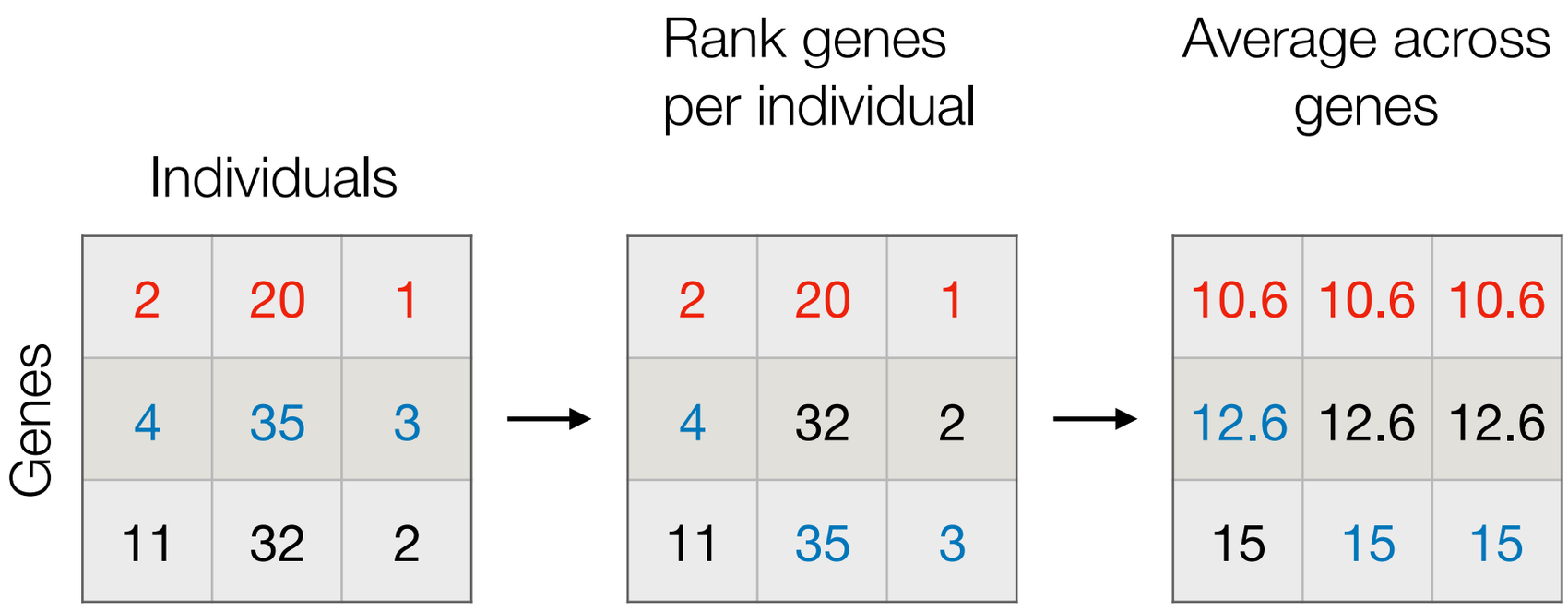
Genes

2	20	1
4	35	3
11	32	2

Quantile normalization



Quantile normalization



Quantile normalization

Individuals

Rank genes per individual

Average across genes

Re-order back to original

Genes

2	20	1
4	35	3
11	32	2

→

2	20	1
4	32	2
11	35	3

→

10.6	10.6	10.6
12.6	12.6	12.6
15	15	15

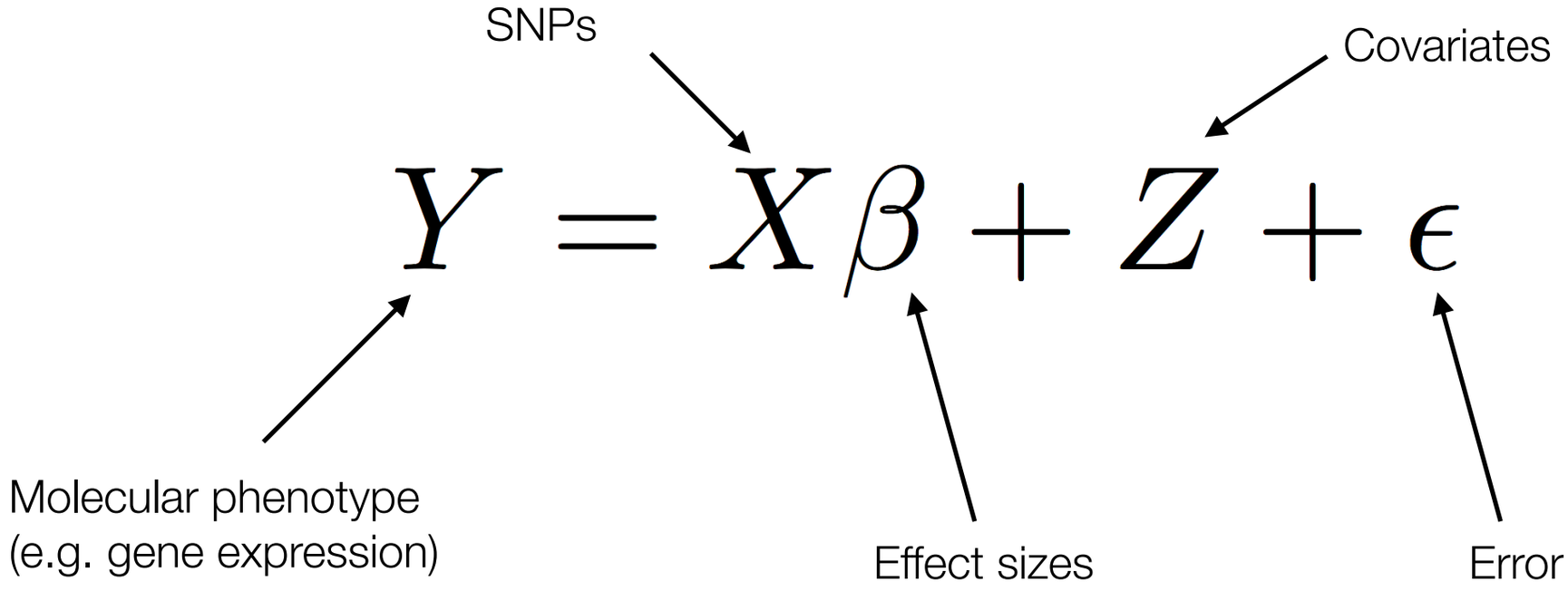
→

10.6	10.6	10.6
12.6	13.6	13.6
13.6	12.6	12.6

The data now follow a normal distribution.

Linear mixed models (LMMs)

Goal: Use a linear model to test for association between gene expression (Y) and SNPs (X), while adjusting for covariates (Z)



Covariates:
Known factors: sex, age, batch effects, etc
Unmeasured factors: cell type composition, cell growth rate, etc

Matrix eQTL, **FastQTL**, etc

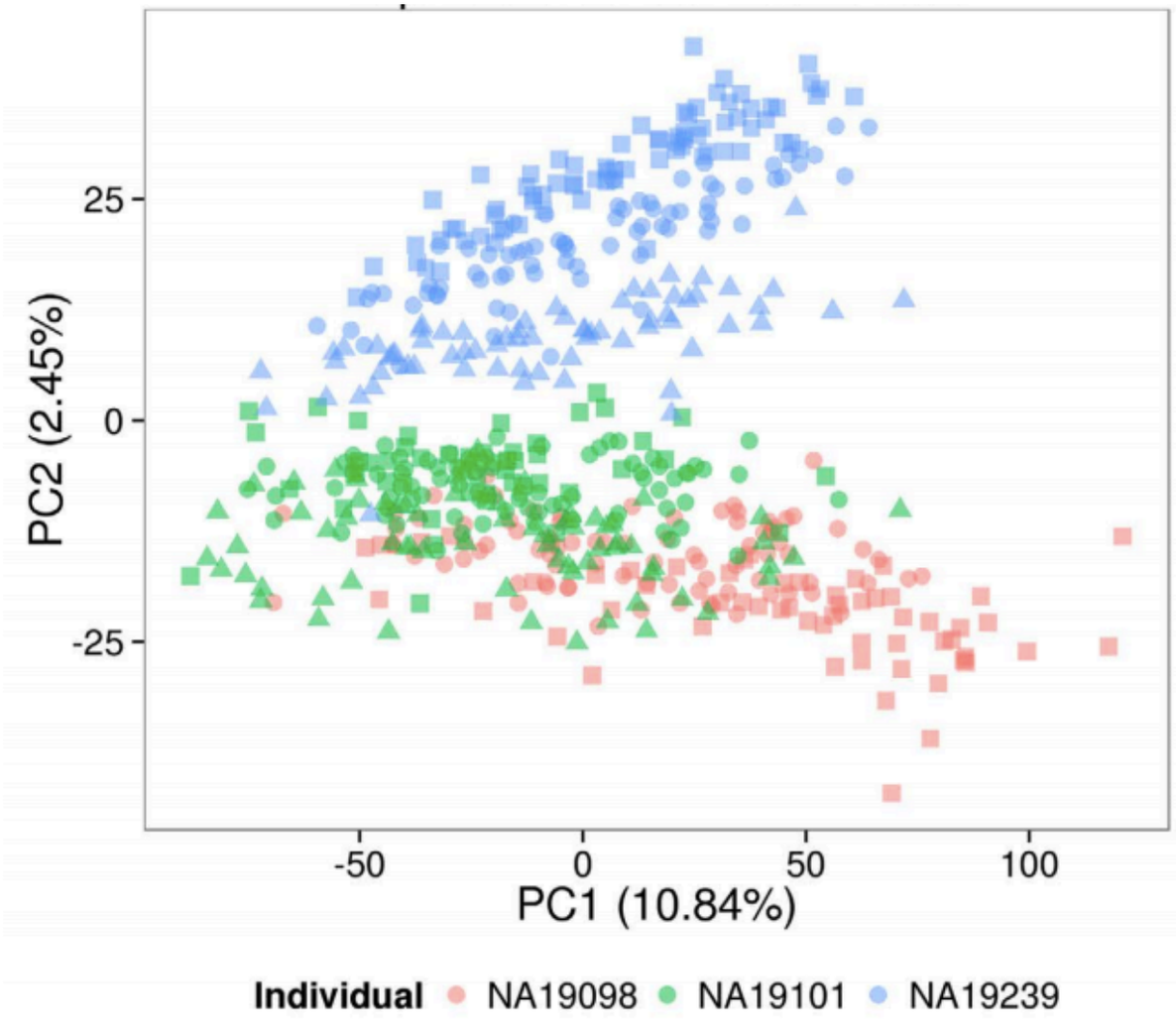
Shabalin, *Bioinformatics*, 2012
Ongen et al., *Bioinformatics*, 2016

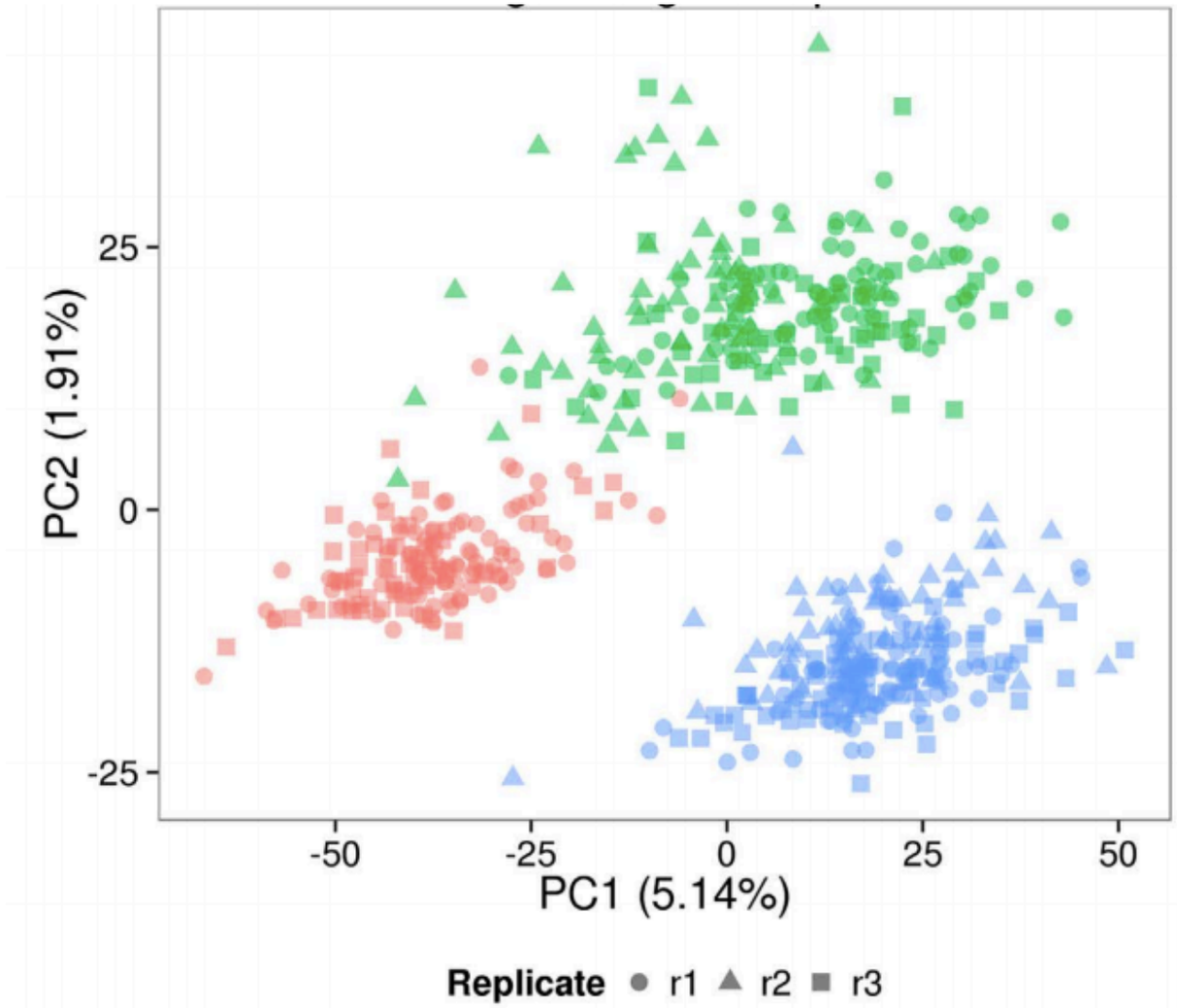
Regressing out covariates

Covariates:
Known factors: sex, age, batch effects, etc
Unmeasured factors: cell type composition, cell growth rate, etc

Top PCs should be regressed out because they do not capture cis-genotype effects (the signal / association we want to measure)

We calculate principal components to be used as covariates in linear regression.





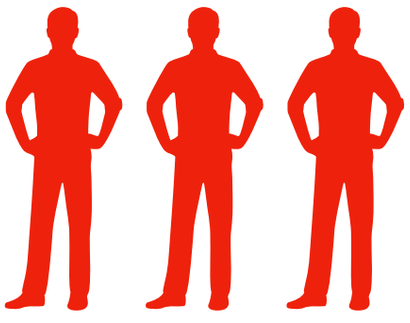
Tung et al., *Scientific Reports,* 2017

QUANTILE NORMALIZATION
COVARIATES
DEMO

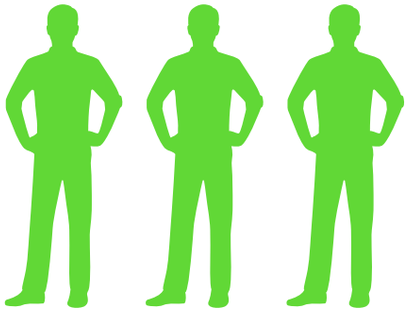
Permutation tests

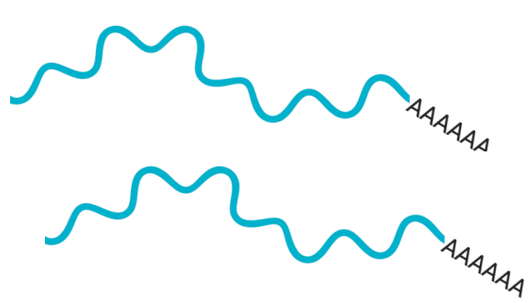
Nominal *p*-values: assigned to every SNP-gene pair that is tested (independently) } α

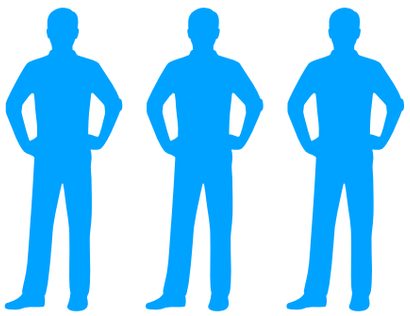
Genotypes

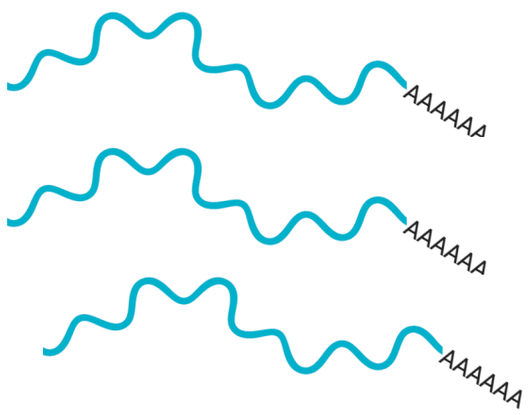


Gene Expression









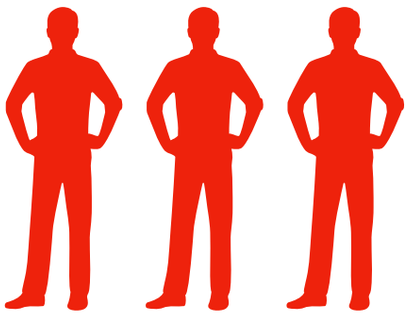
Ongen et al., *Bioinformatics*, 2016

Delaneau et al., *Nature Communications*, 2017

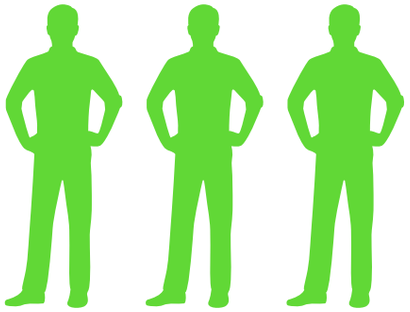
Permutation tests

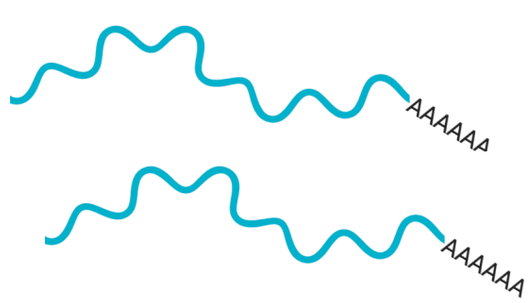
Permutation tests allow us to assess statistical significance in our dataset.

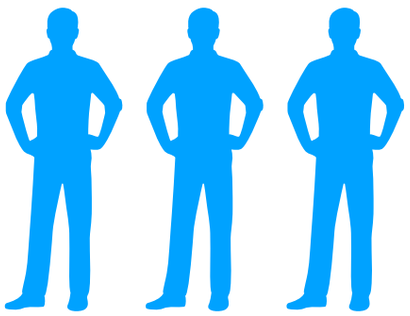
Genotypes

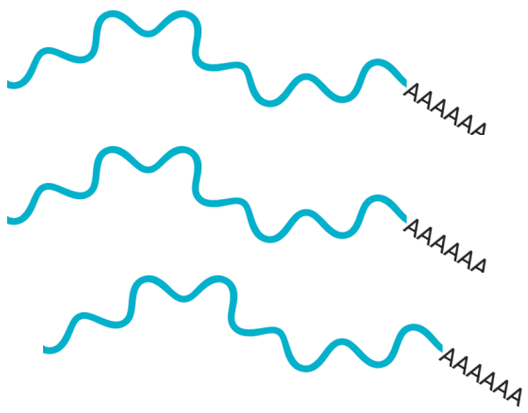


Gene Expression









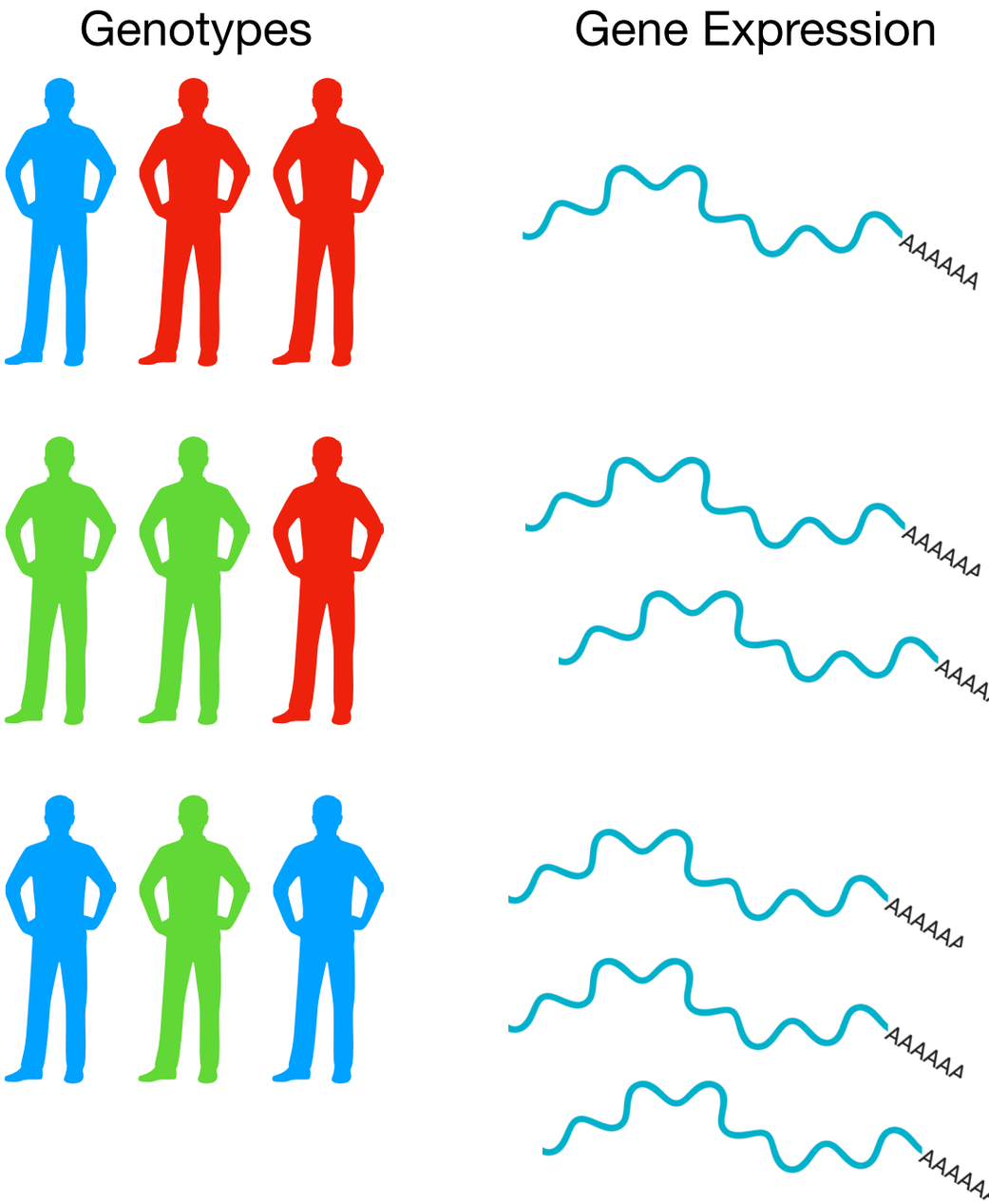
Scramble the genotypes (1,000 times)
+ re-run association

Ongen et al., *Bioinformatics*, 2016

Delaneau et al., *Nature Communications*, 2017

Permutation tests

Permutation tests allow us to assess statistical significance in our dataset.



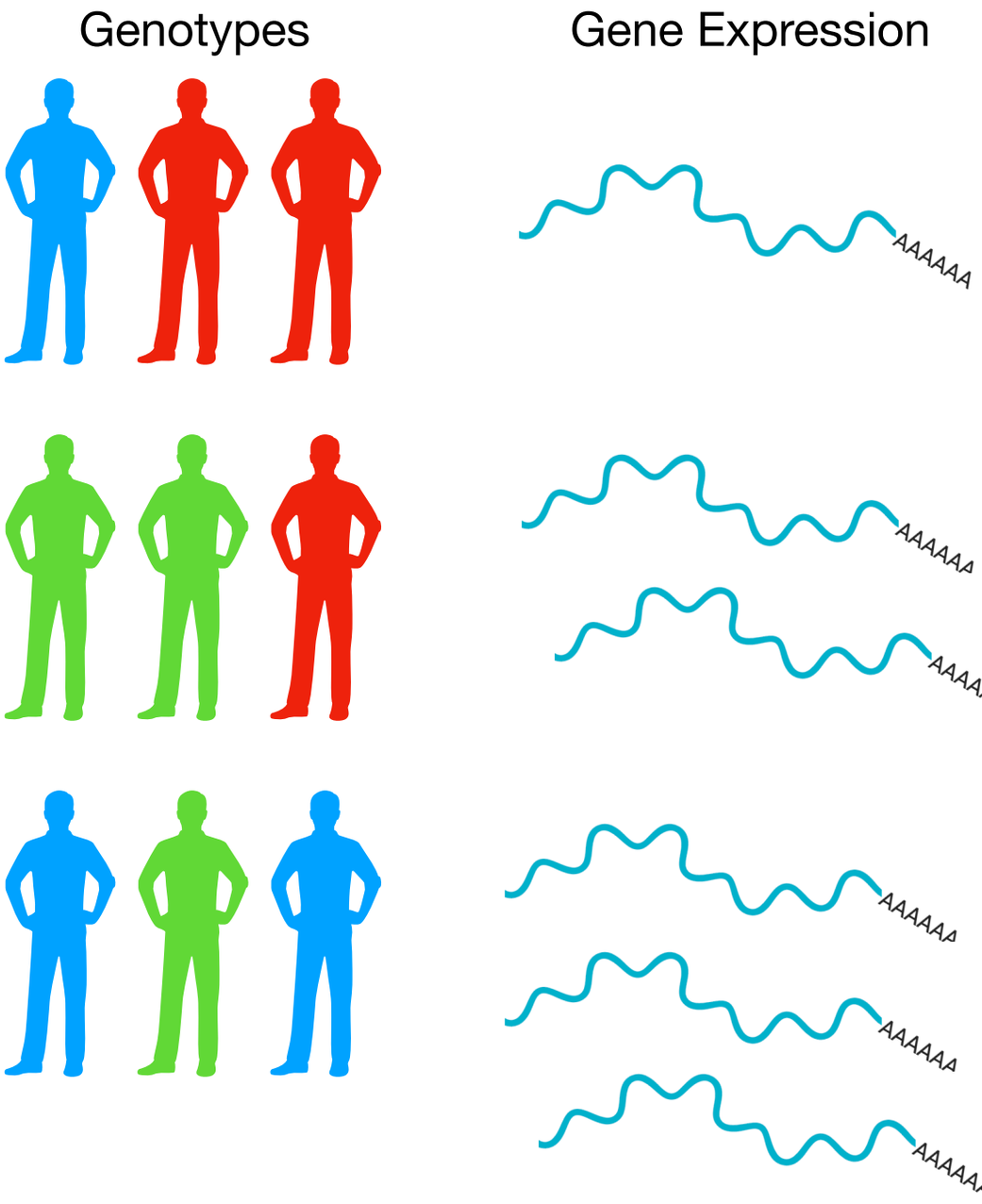
The permutation *p*-value is the proportion of permutations in which at least one SNP-gene pair *p*-value is < alpha (nominal association).

Ongen et al., *Bioinformatics*, 2016

Delaneau et al., *Nature Communications*, 2017

Permutation tests

Permutation tests allow us to assess statistical significance in our dataset.



The permutation *p*-value is the proportion of permutations in which at least one SNP-gene pair *p*-value is < alpha (nominal association).

NOTE: In a real dataset, we do an *approximation* as we cannot permute all the data.

Ongen et al., *Bioinformatics*, 2016

Delaneau et al., *Nature Communications*, 2017

ASSOCIATION DEMO

Multiple testing correction

As the number of SNP-gene pairs we test increases, the probability of getting a significant association (by random chance) increases.

Multiple testing correction

As the number of SNP-gene pairs we test increases, the probability of getting a significant association (by random chance) increases.

Approach 1: we need to correct the p-values by the number of variants tested (Bonferroni correction). This tends to be overly stringent and can result in many false negatives in QTL mapping.

Approach 2: false discovery rate (FDR) cutoffs can be applied. FDR is defined as the proportion of false positives among all significant results. In QTL mapping, we tend to assume an FDR < 5% or FDR < 10%.

Multiple testing correction

As the number of SNP-gene pairs we test increases, the probability of getting a significant association (by random chance) increases.

Approach 1: we need to correct the p-values by the number of variants tested (Bonferroni correction). This tends to be overly stringent and can result in many false negatives in QTL mapping.

Approach 2: false discovery rate (FDR) cutoffs can be applied. FDR is defined as the proportion of false positives among all significant results. In QTL mapping, we tend to assume an FDR < 5% or FDR < 10%.

Benjamini-Hochberg (BH) procedure: controls the FDR at the value you specify (e.g. 10%)

	P-value	Rank	(Rank / Tests)* FDR cutoff	
	0.001	1	0.004	
	0.005	2	0.008	
	0.007	3	0.012	
	…	…	…	
*	0.09	24	0.096	Highest p-val smaller than critical value
	0.12	25	0.1	

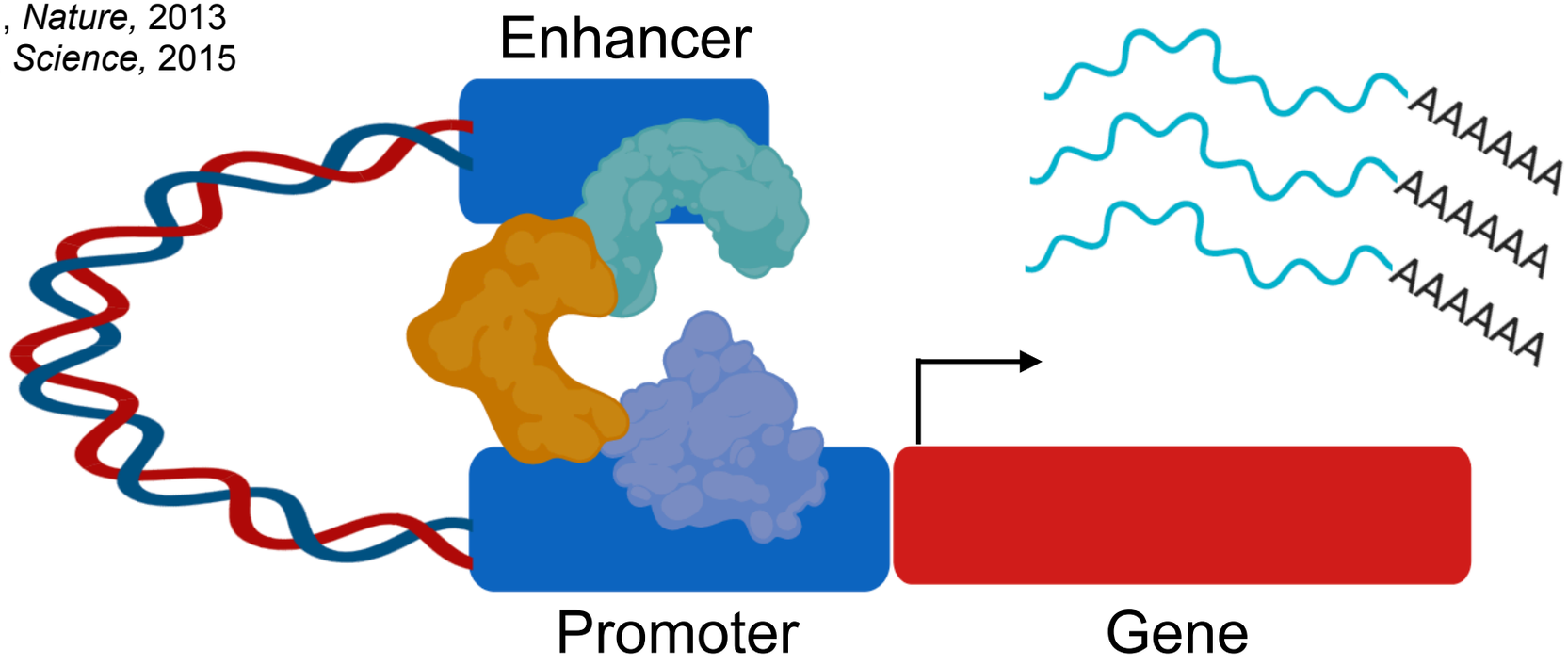
*All above (including this red row) are considered significant

MTC DEMO

What we have learned through extensive eQTL analysis…

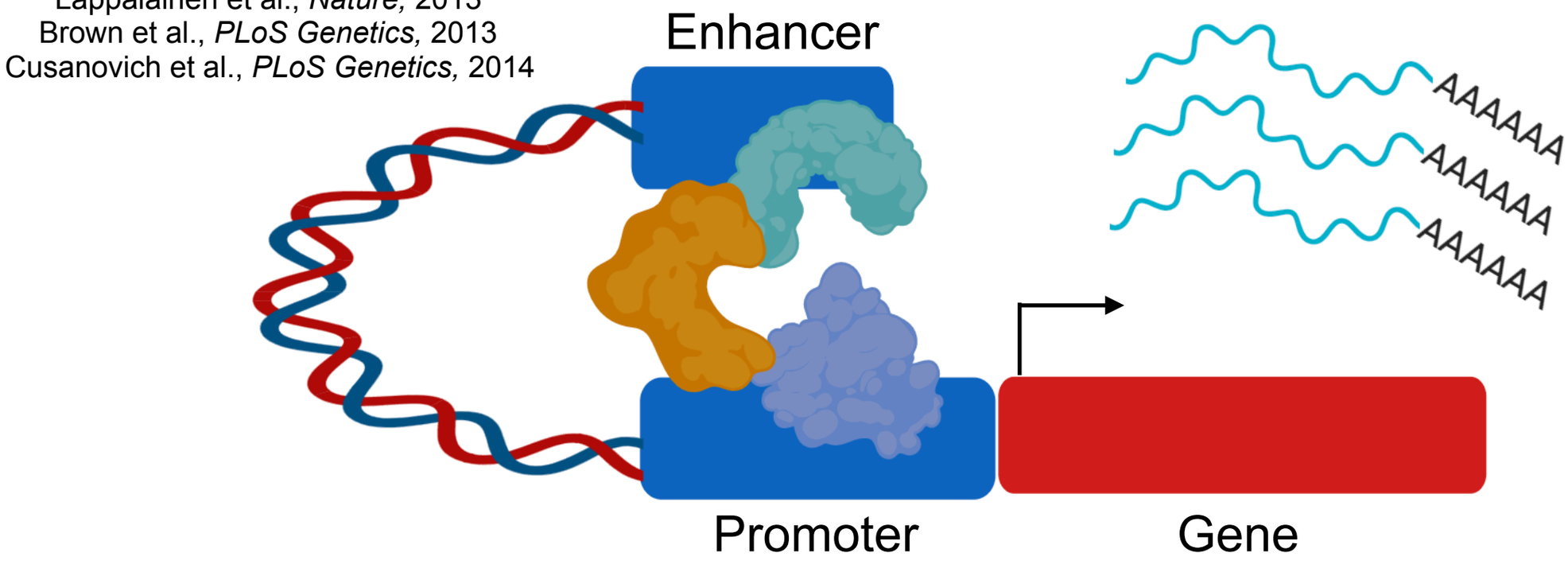
eQTL are enriched in enhancers and promoters

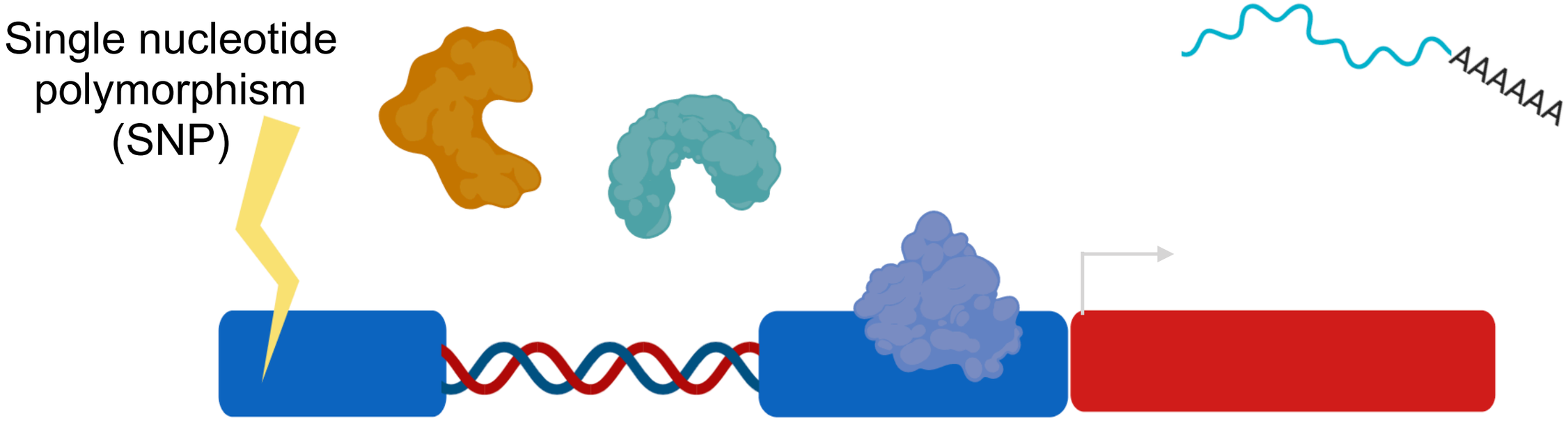
Lappalainen et al., *Nature,* 2013
GTEx Consortium, *Science,* 2015



eQTL often affect transcription factor binding sites

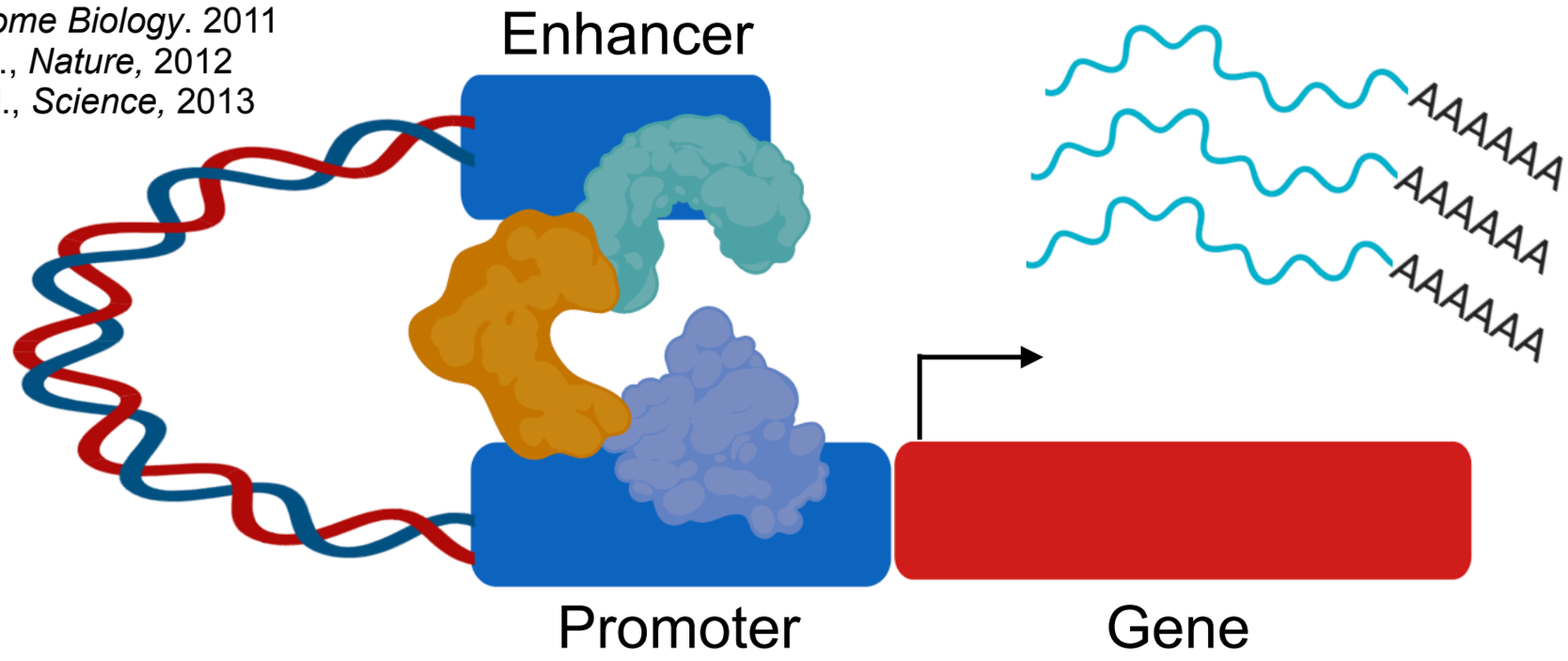
Lappalainen et al., *Nature,* 2013
Brown et al., *PLoS Genetics,* 2013
Cusanovich et al., *PLoS Genetics,* 2014

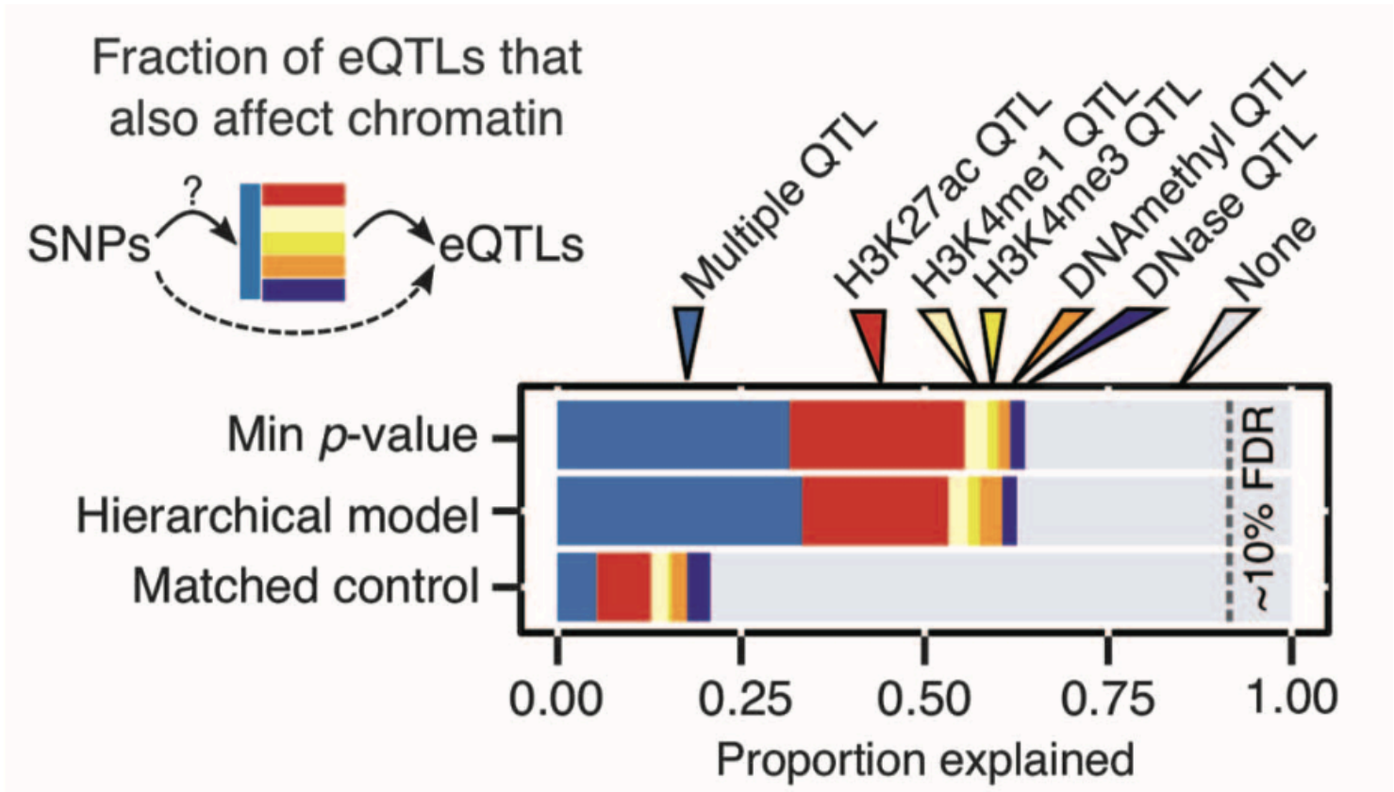




Many eQTL can be explained by chromatin-level phenotypes

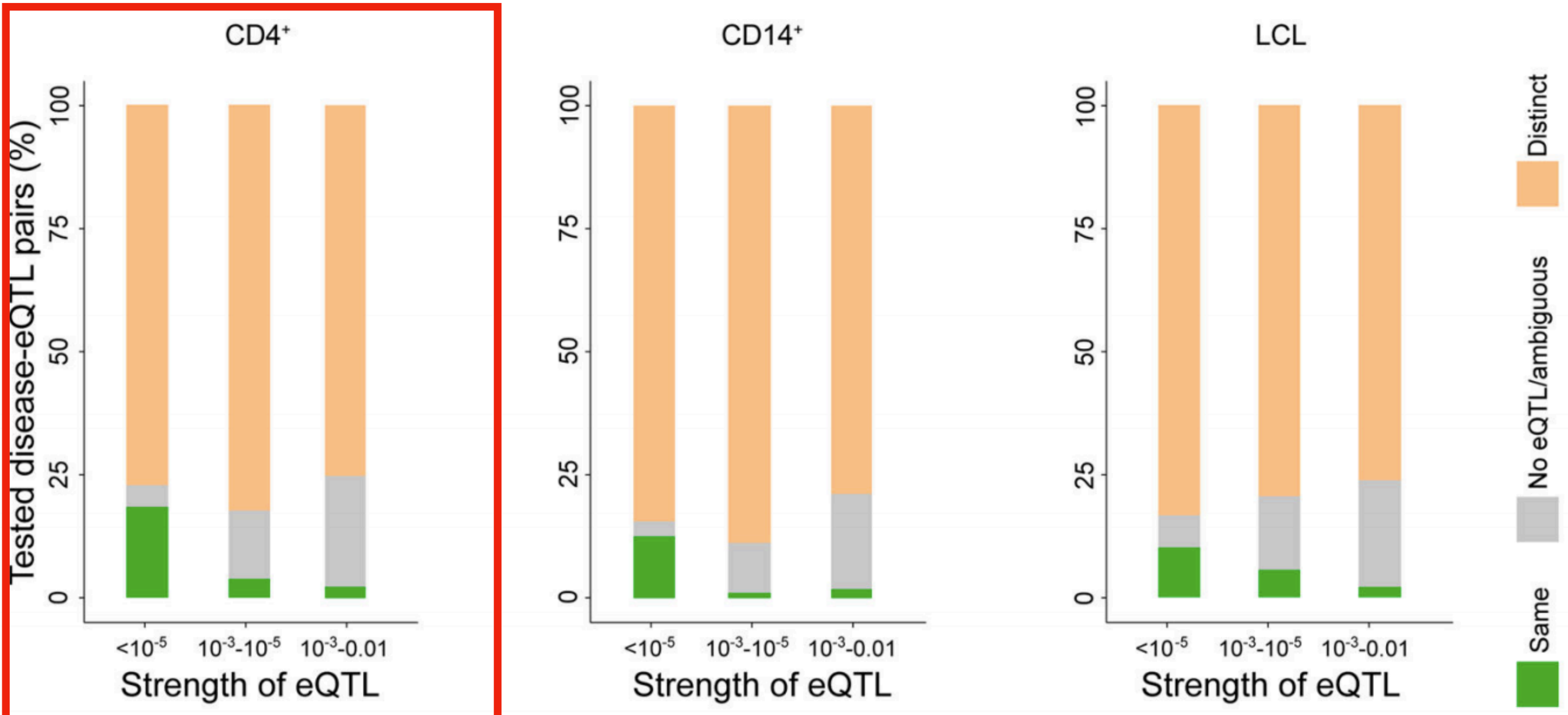
Bell et al., *Genome Biology*. 2011
Degner et al., *Nature,* 2012
McVicker et al., *Science,* 2013





Li et al., *Science,* 2016

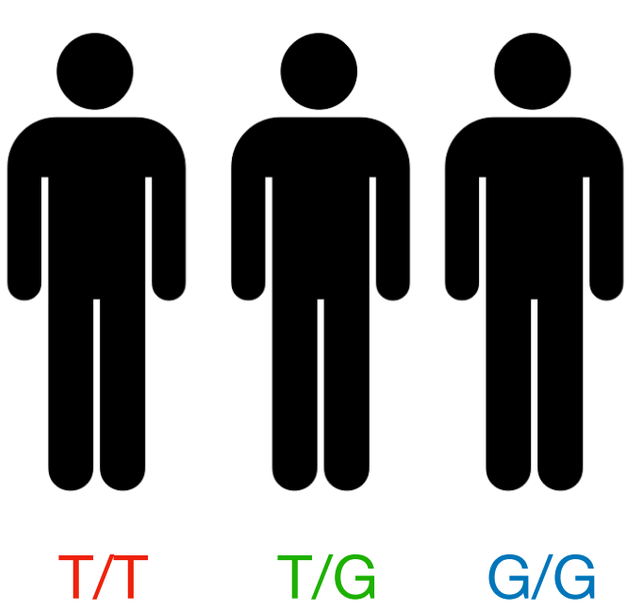
75% of *disease*-associated variants disrupt gene regulation in a manner ***independent*** of total mRNA levels / gene expression



Chun et al., *Nature Genetics*, 2017

Expanding the repertoire of molecular phenotypes

Genotype → Molecular phenotype → Phenotype



Molecular phenotype

Gene expression

Splicing

Polyadenylation

Protein

…

Phenotype

e.g. height, CAD, etc