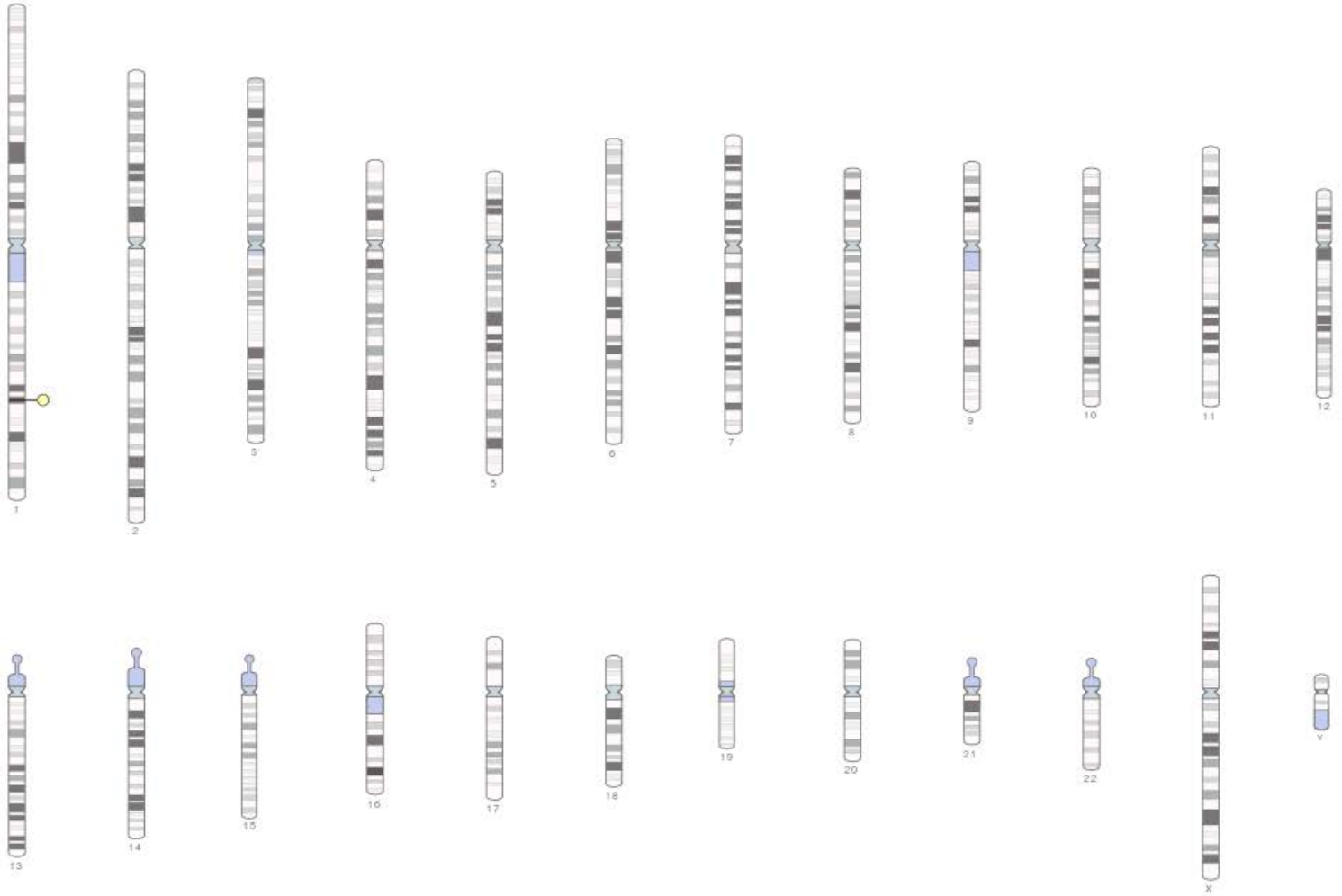
Effect of genotypes on gene expression

Patrick Deelen, 25-09-2018

umcg

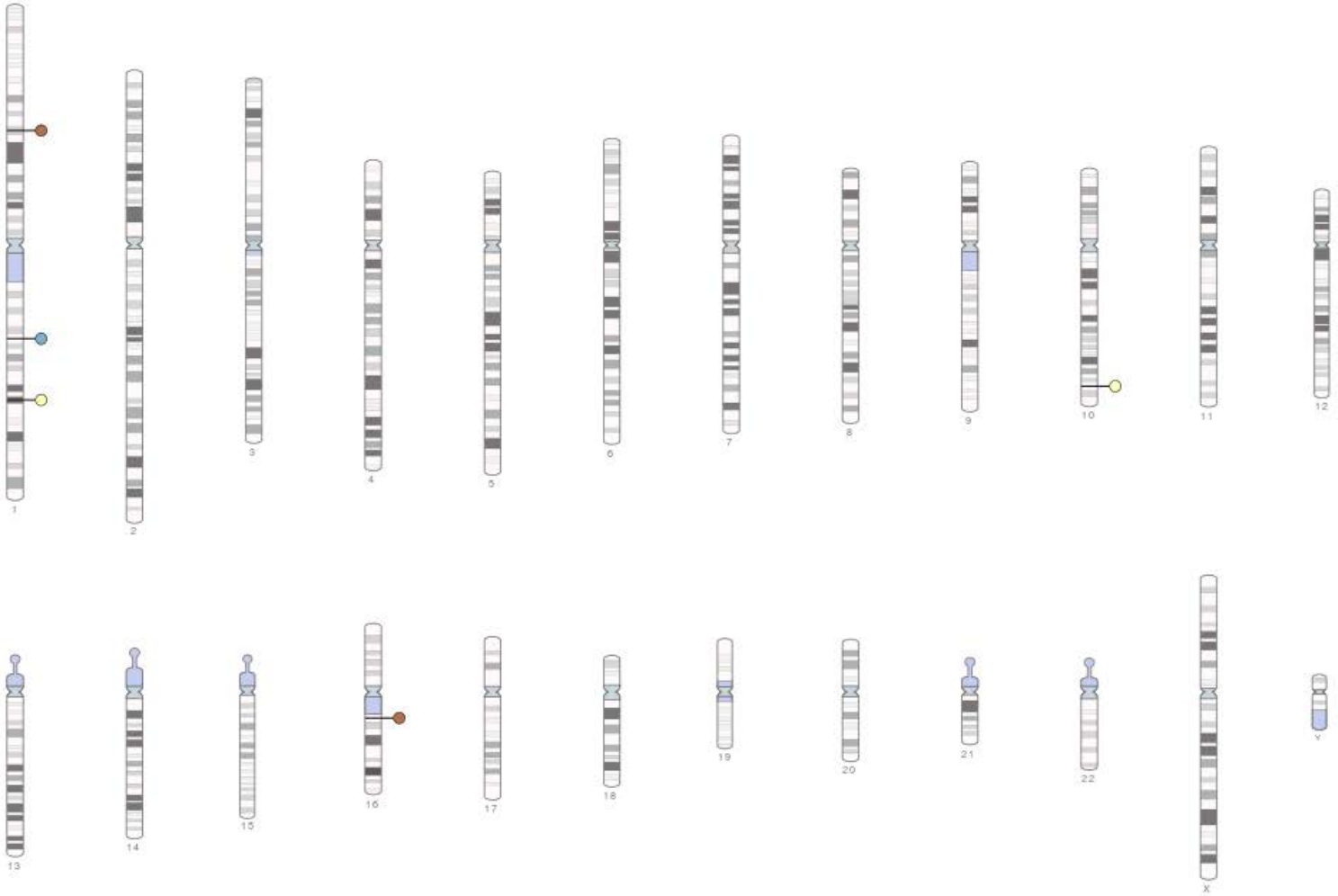
BBMRI • NL

2005



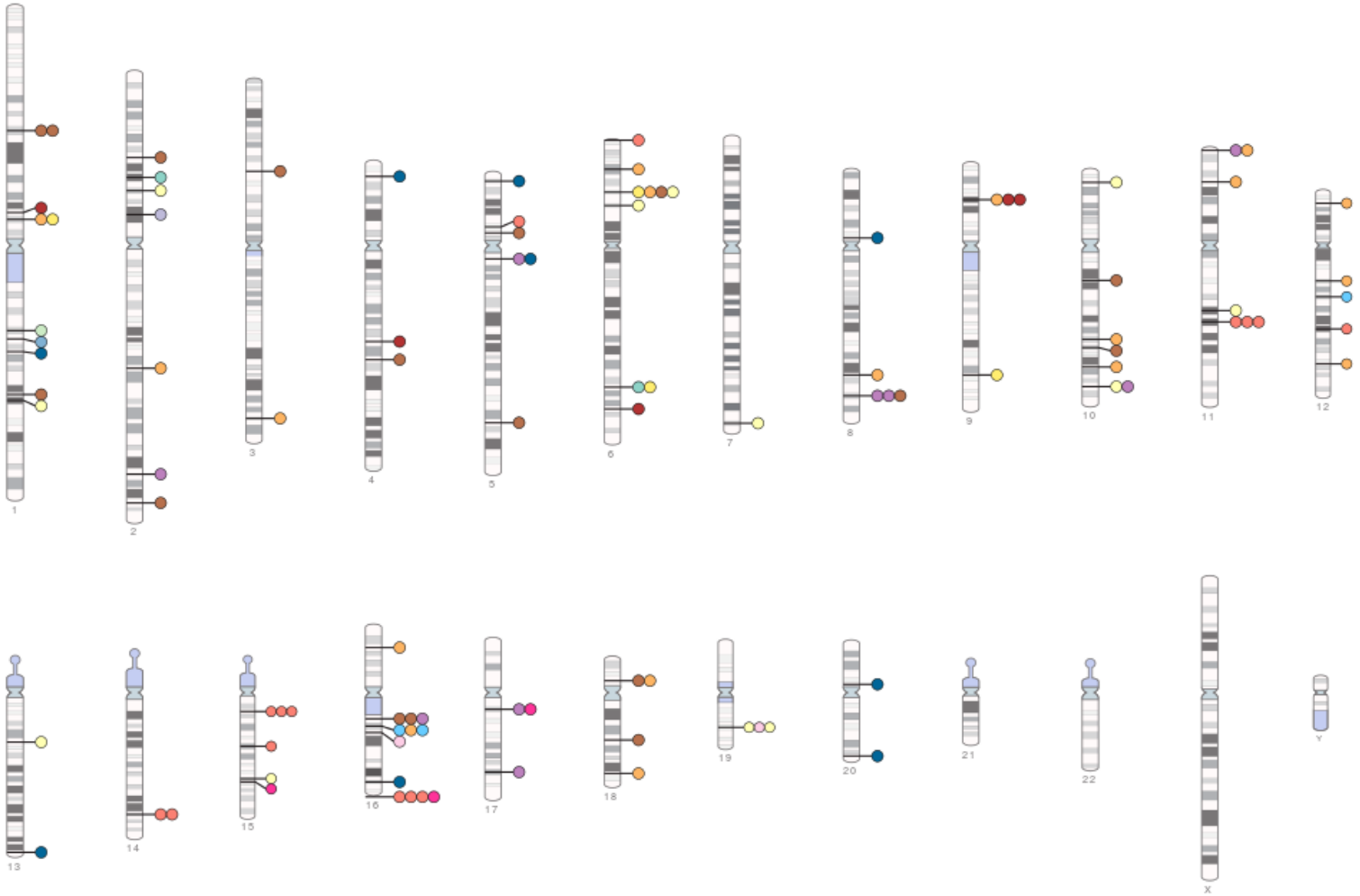
www.ebi.ac.uk/fgpt/gwas

2006



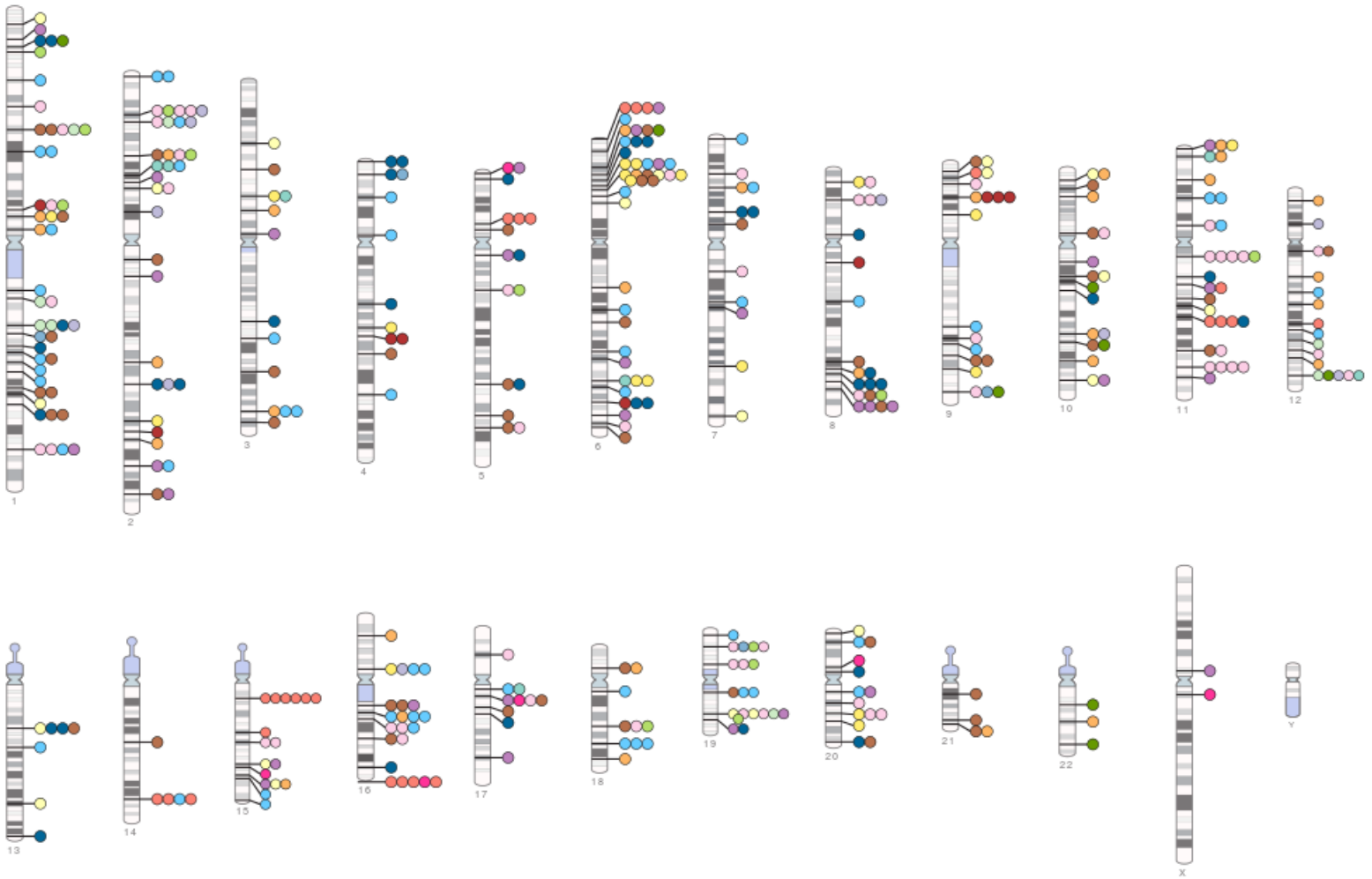
www.ebi.ac.uk/fgpt/gwas

2007



www.ebi.ac.uk/fgpt/gwas

2008



www.ebi.ac.uk/fgpt/gwas

2009

www.ebi.ac.uk/fgpt/gwas

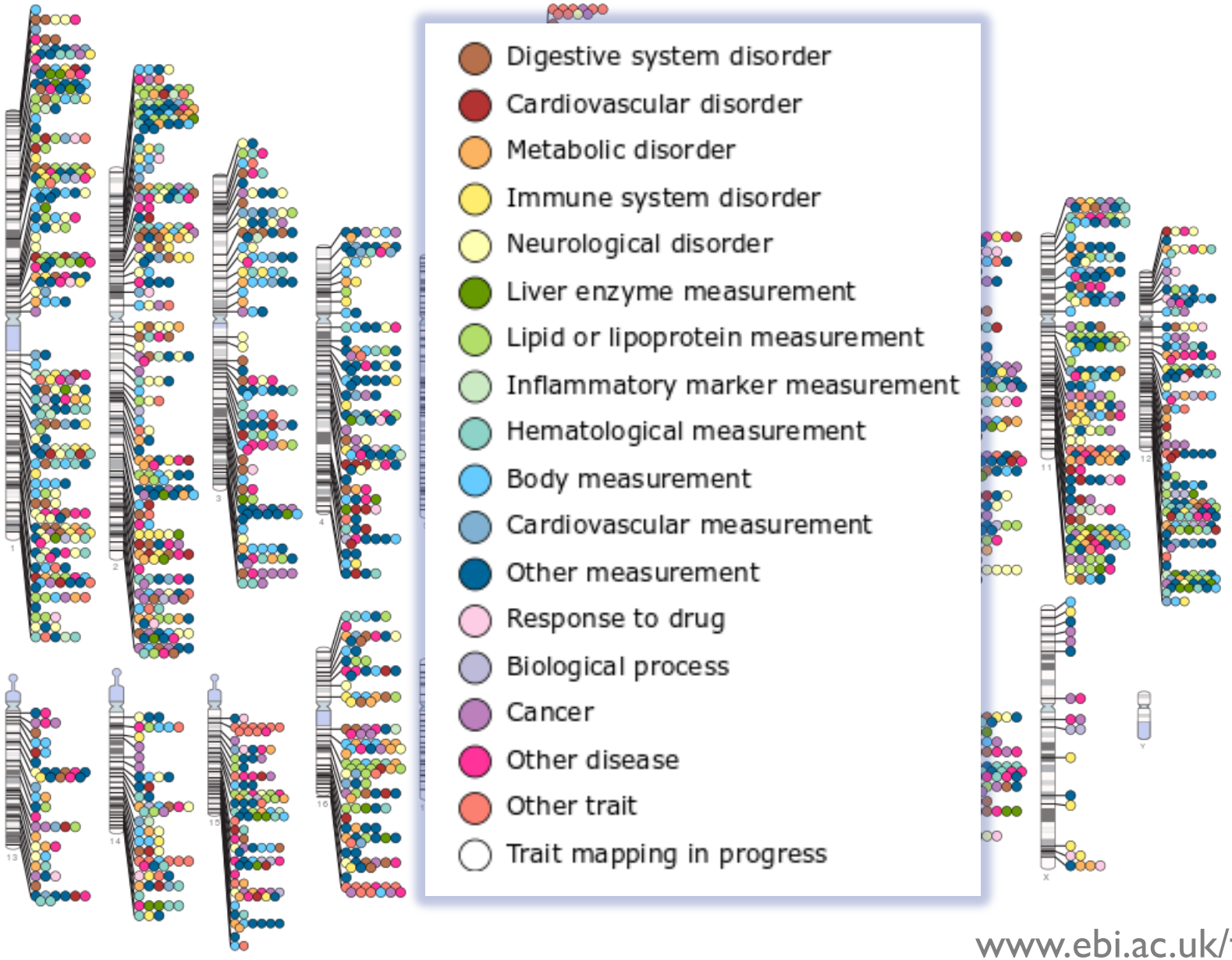
2010

www.ebi.ac.uk/fgpt/gwas

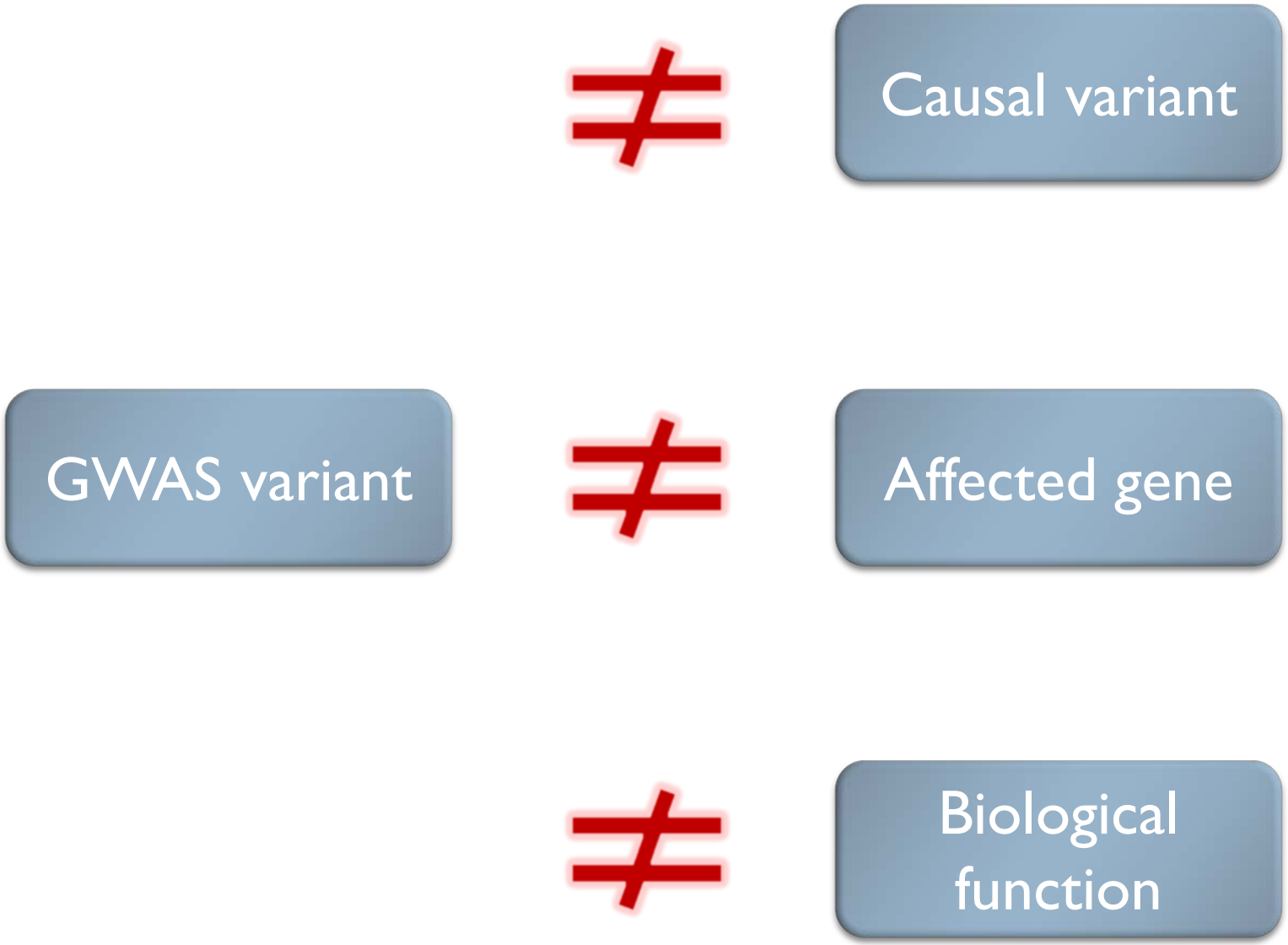
2011

www.ebi.ac.uk/fgpt/gwas

2012

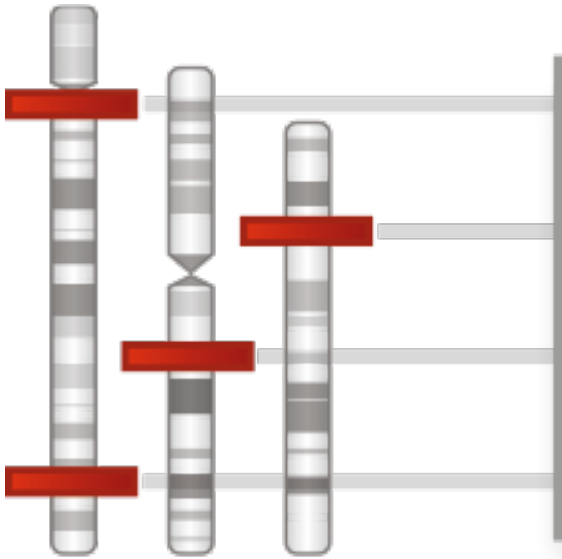


www.ebi.ac.uk/fgpt/gwas


Causal variant
GWAS variant
Affected gene
Biological function

Problem of life science community

Genetic risk factors



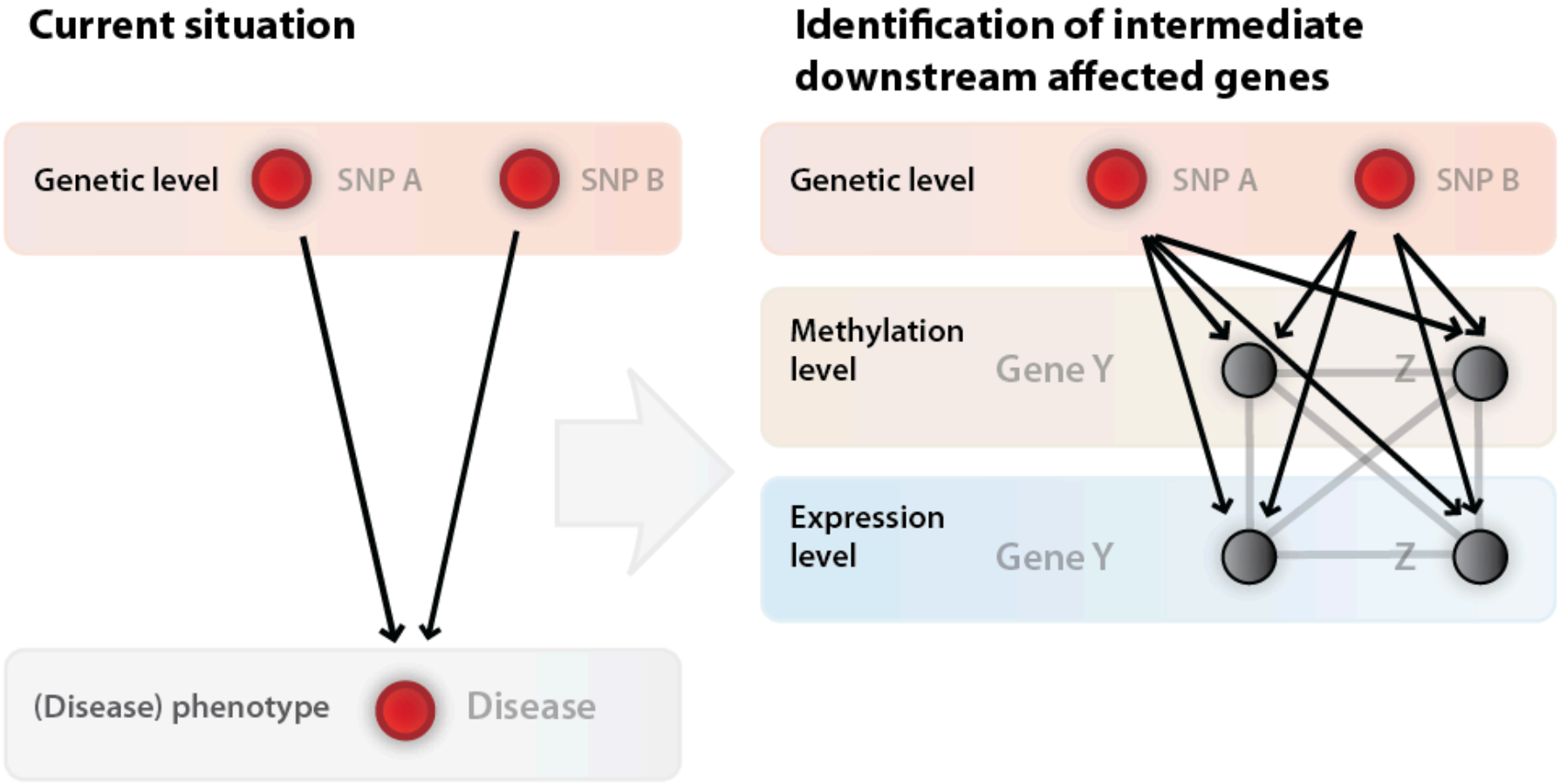
>10,000 known

Genes unknown
Pathways unknown
Cell-types unknown

Disease

>200 diseases

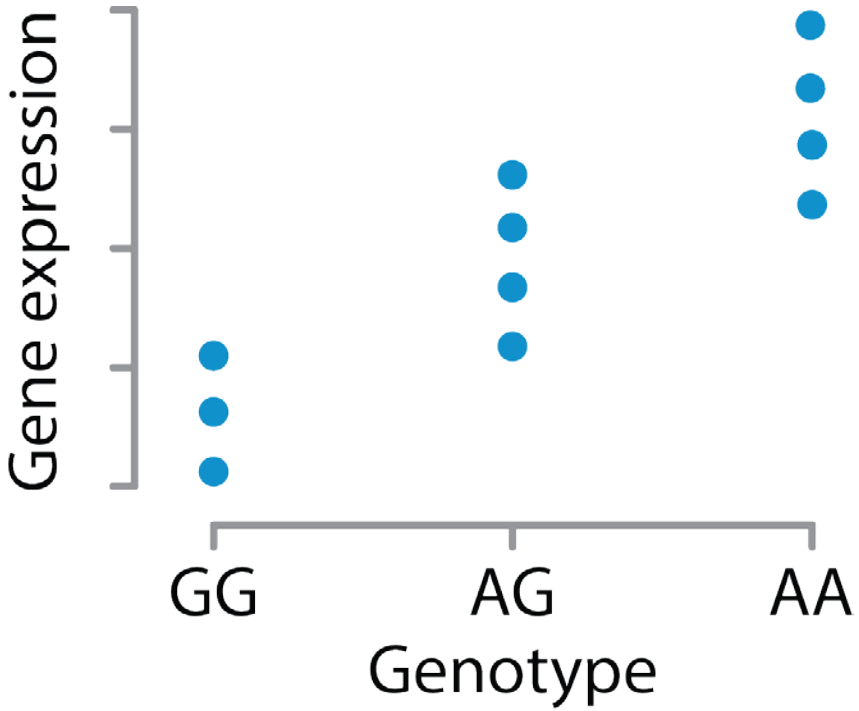
Why look at eQTL



Slide: Lude Franke

Moving from variants to genes using eQTLs

- Most GWAS hits do **not** change protein structure
- They often **do** affect expression levels
- Expression quantitative trait locus - eQTL
- Genetic variation correlated to expression
- Large dataset with for same individual
  - Genotypes
  - Expression levels



eQTL dataset

Genotypes

Variant	Sample1	Sample2	Sample3
SNP 1	A/A	A/T	T/T
SNP 2	T/T	T/T	T/C
SNP 3	G/C	G/C	G/G
SNP 4	G/G	G/G	G/G

Expression levels

Gene	Sample1	Sample2	Sample3
Gene 1	10	20	30
Gene 2	20	10	18
Gene 3	10	11	12
Gene 4	10	10	10

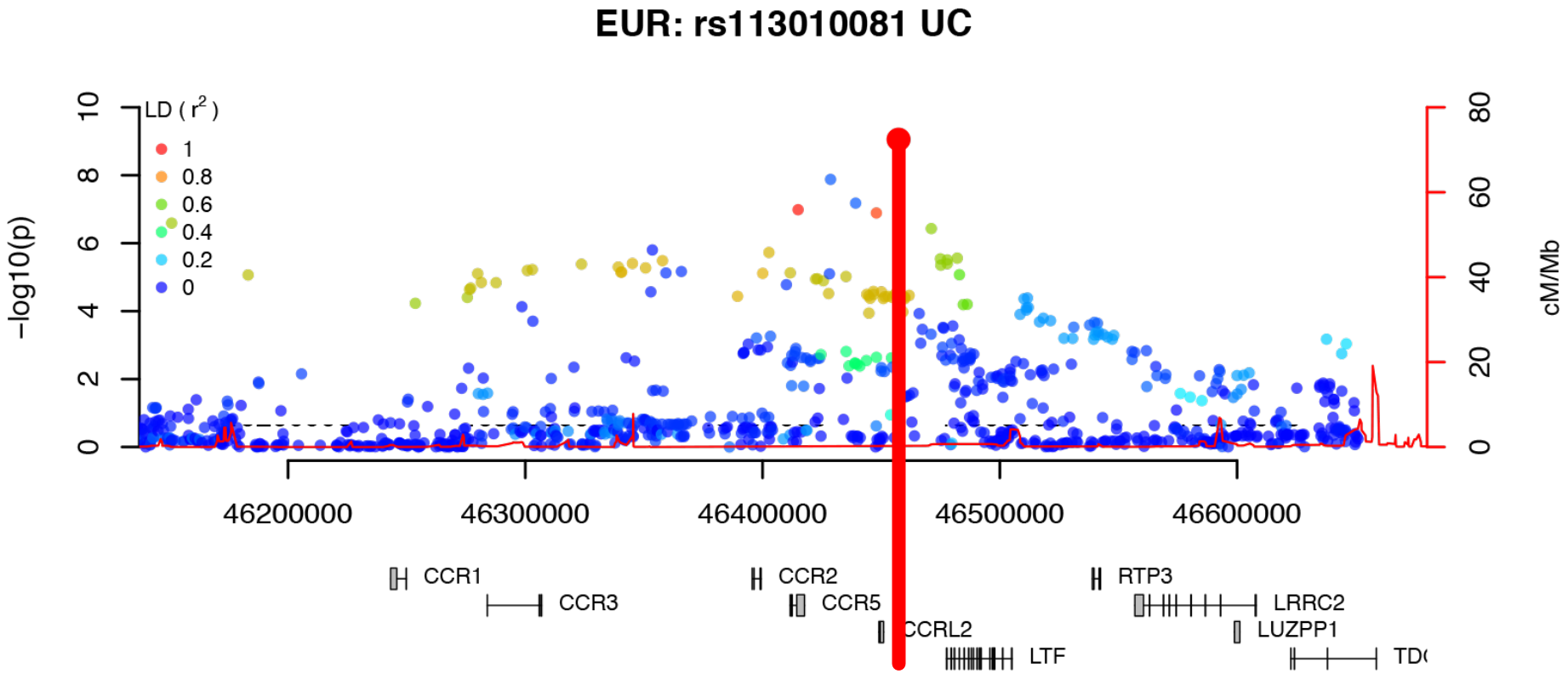
Multiple testing

- Normally we say p-value 0.05 is significant
  - 5% (1/20) chance that a result is a false positive.
- Suppose we perform 20 test
  - When using p-value cut-off of 0.05, we expect 1 false positive
- Cis eQTL mapping: 20,000 genes x 10 SNPs = 200,000 tests

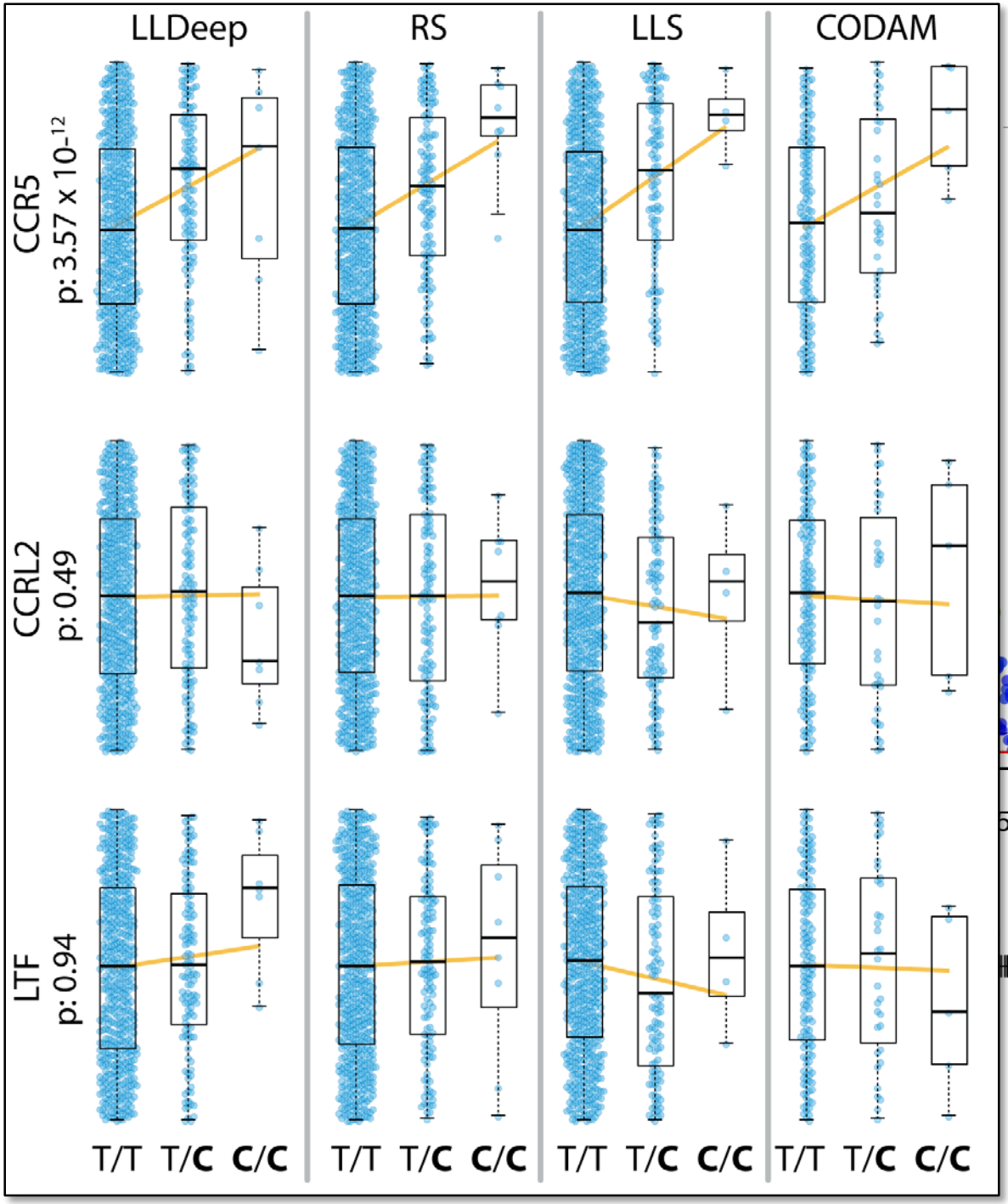
False discovery rate (FDR)

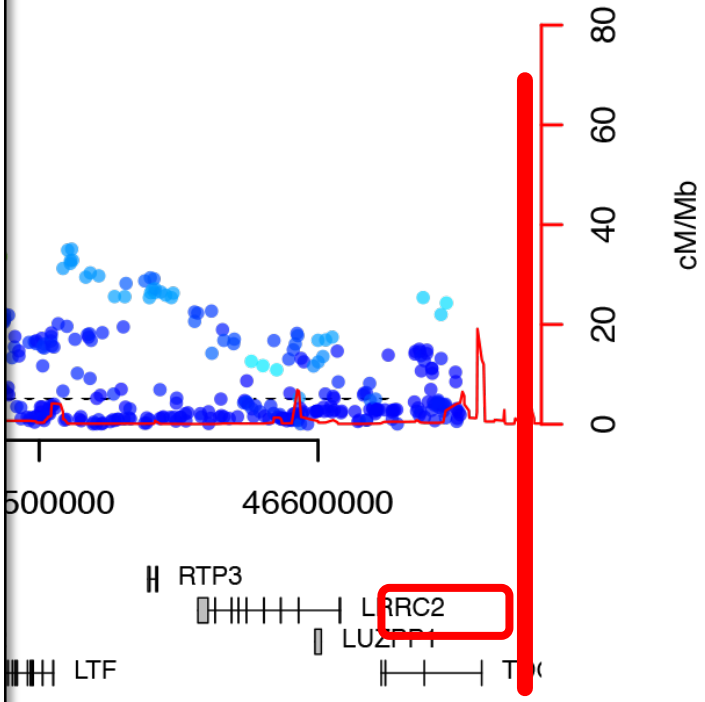
- One of the solutions when doing a lot of tests
- Expected fraction of false positives results
- Determine p-value cut-off that would result in a specific result

Example UC locus



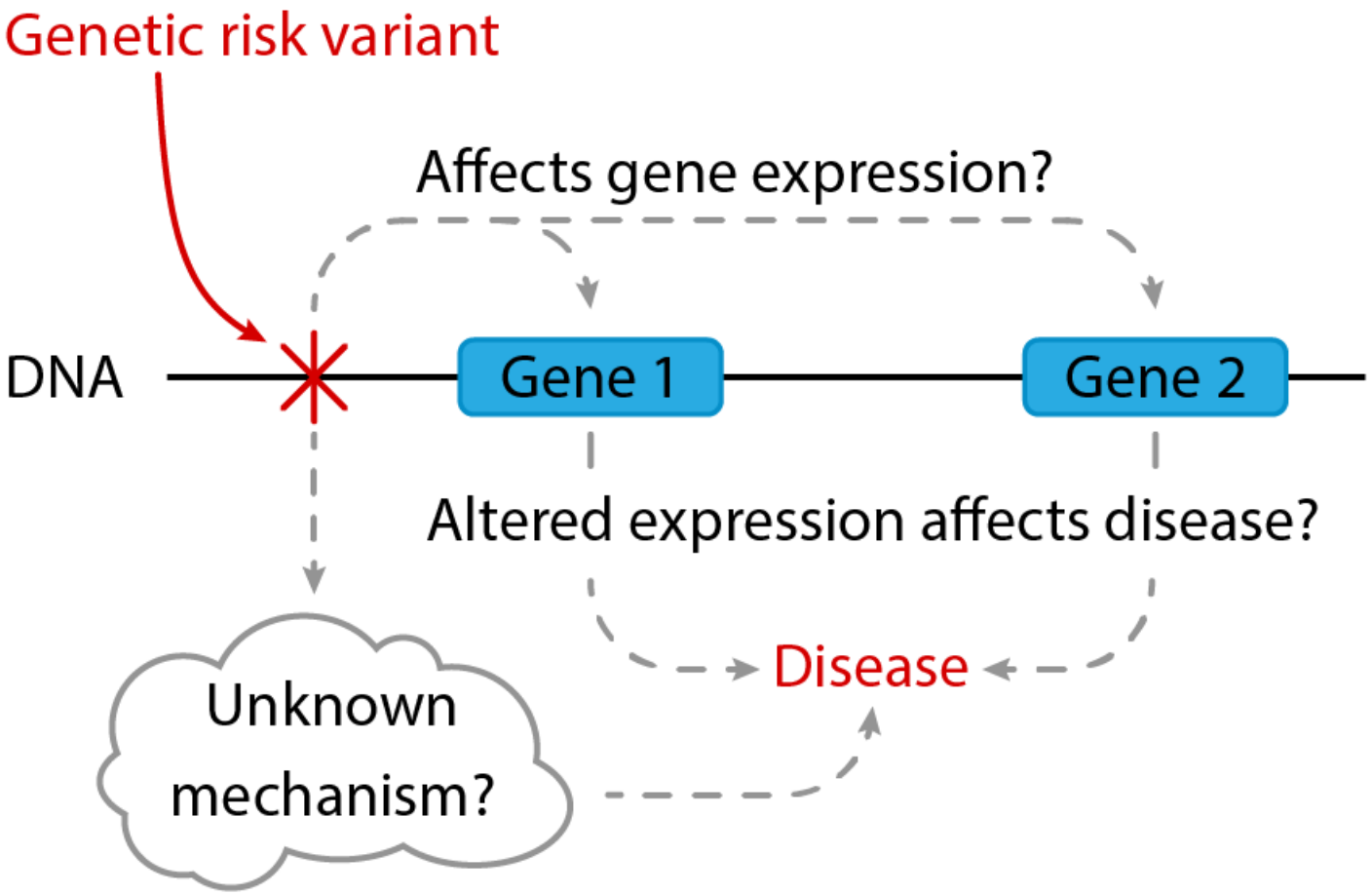
Liu et al. NG 2016



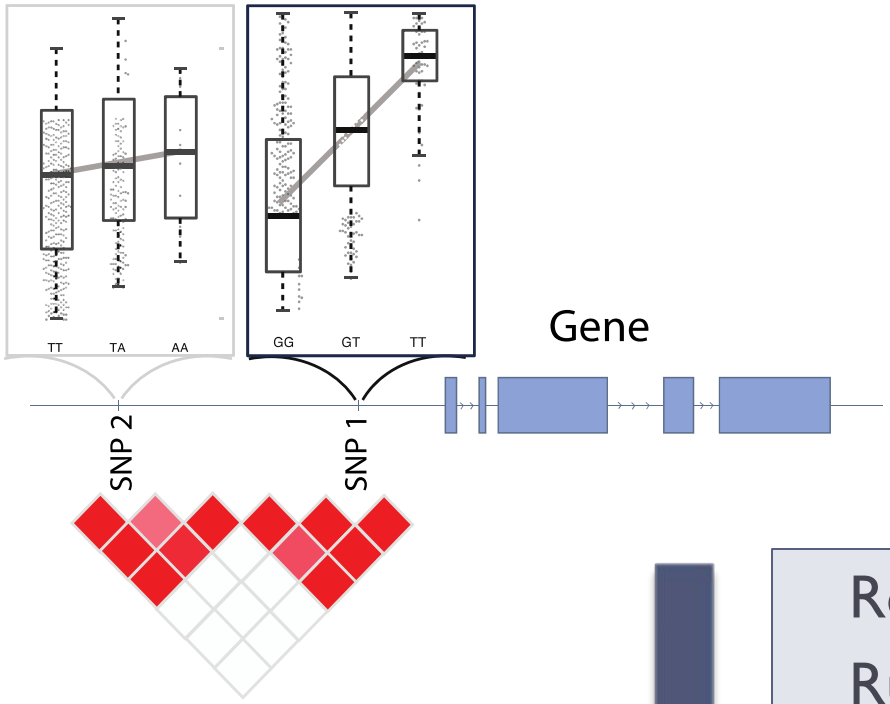


Liu et al. NG 2016

Pleiotropy

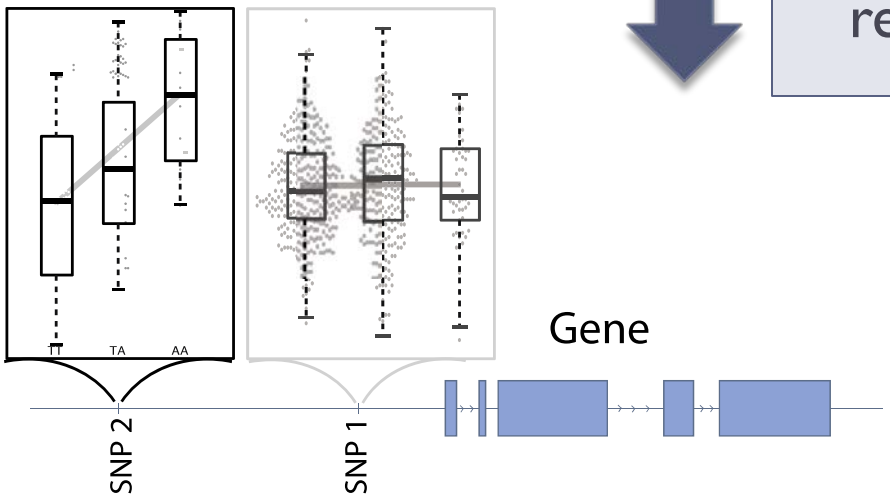

Genetic risk variant
Affects gene expression?
DNA
Gene 1
Gene 2
Altered expression affects disease?
Disease
Unknown
mechanism?

Detecting independent eQTLs

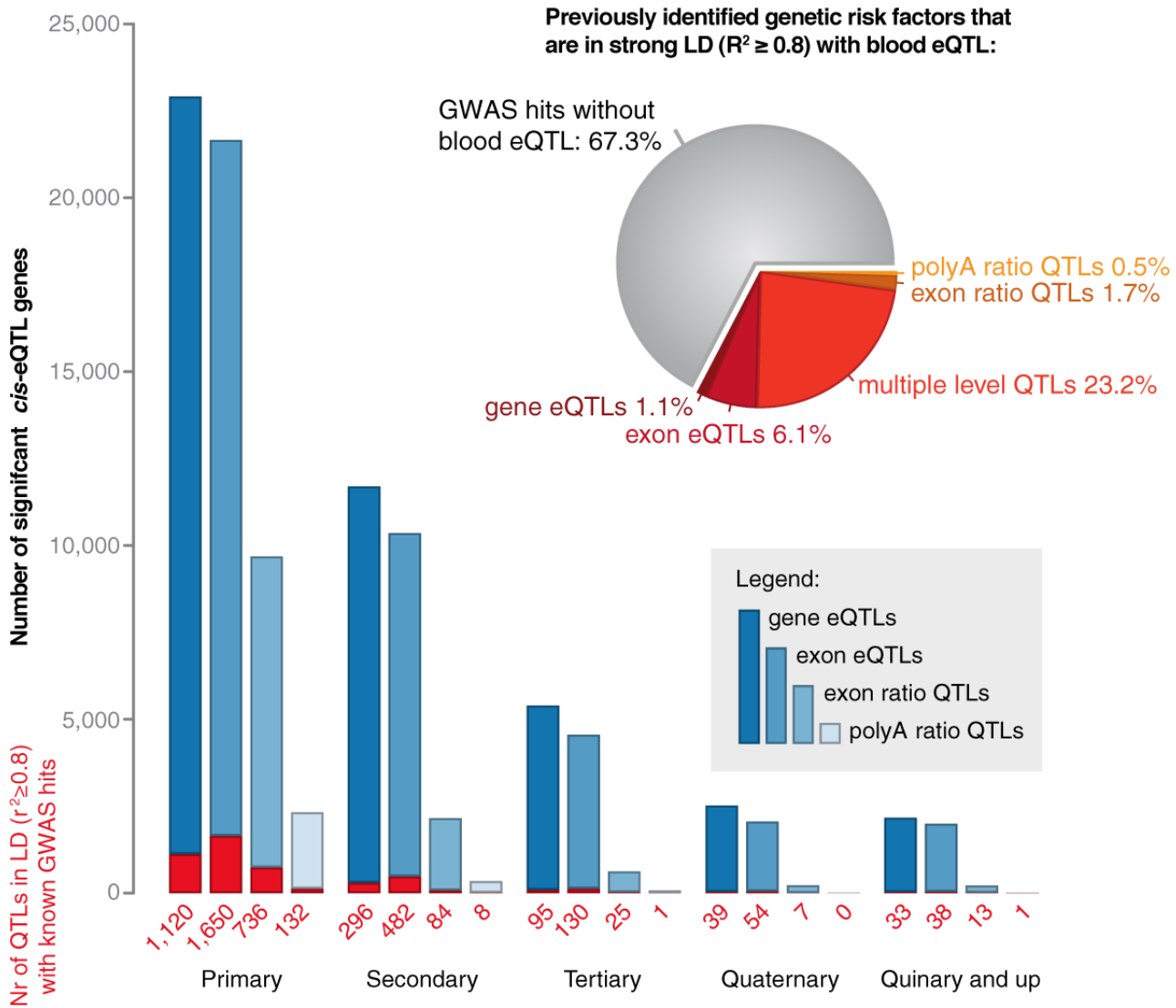


Regress out the effect of the top SNP

Run secondary eQTL mapping on residual expression

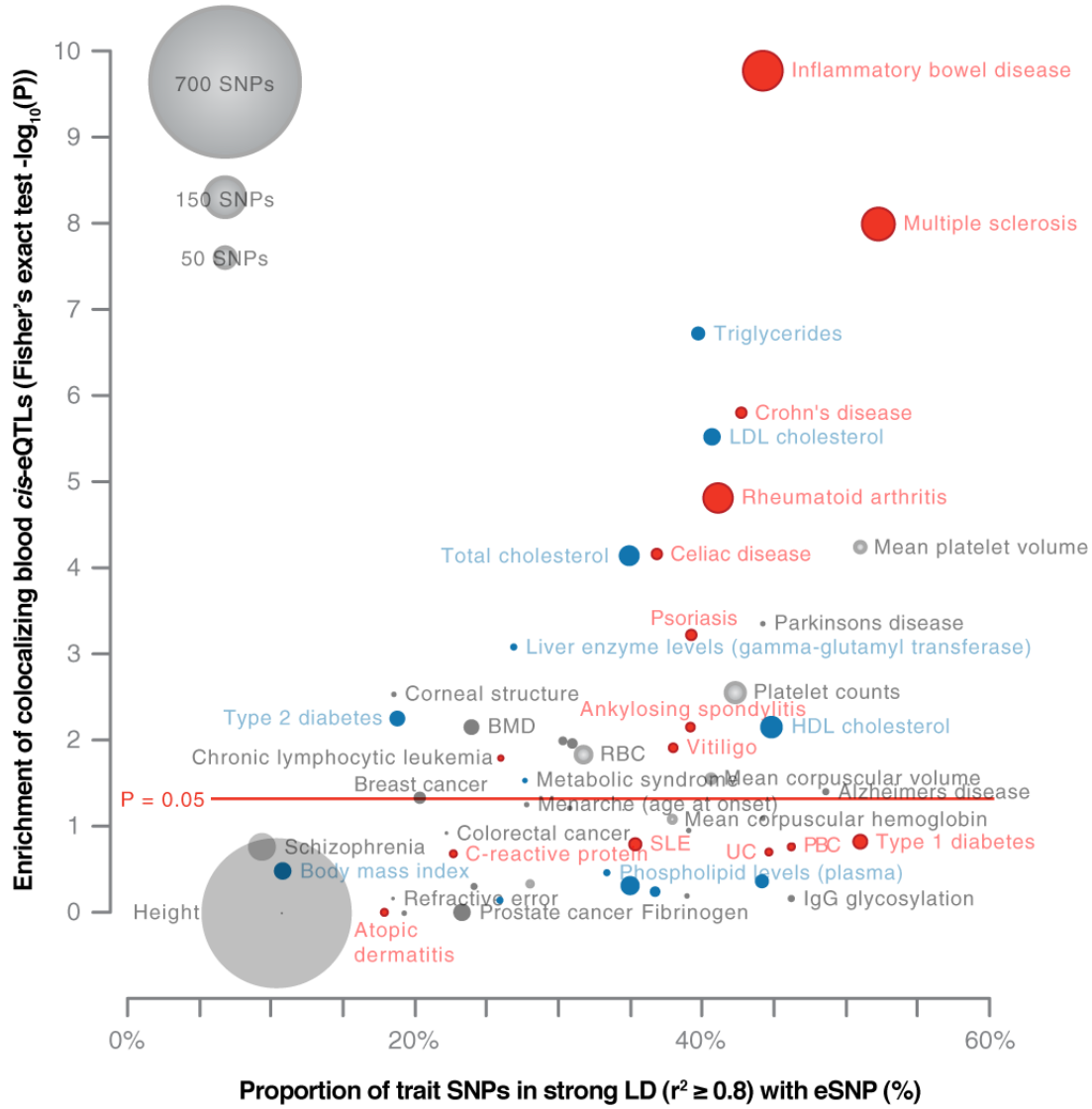


Many genetic risk factors affect gene expression



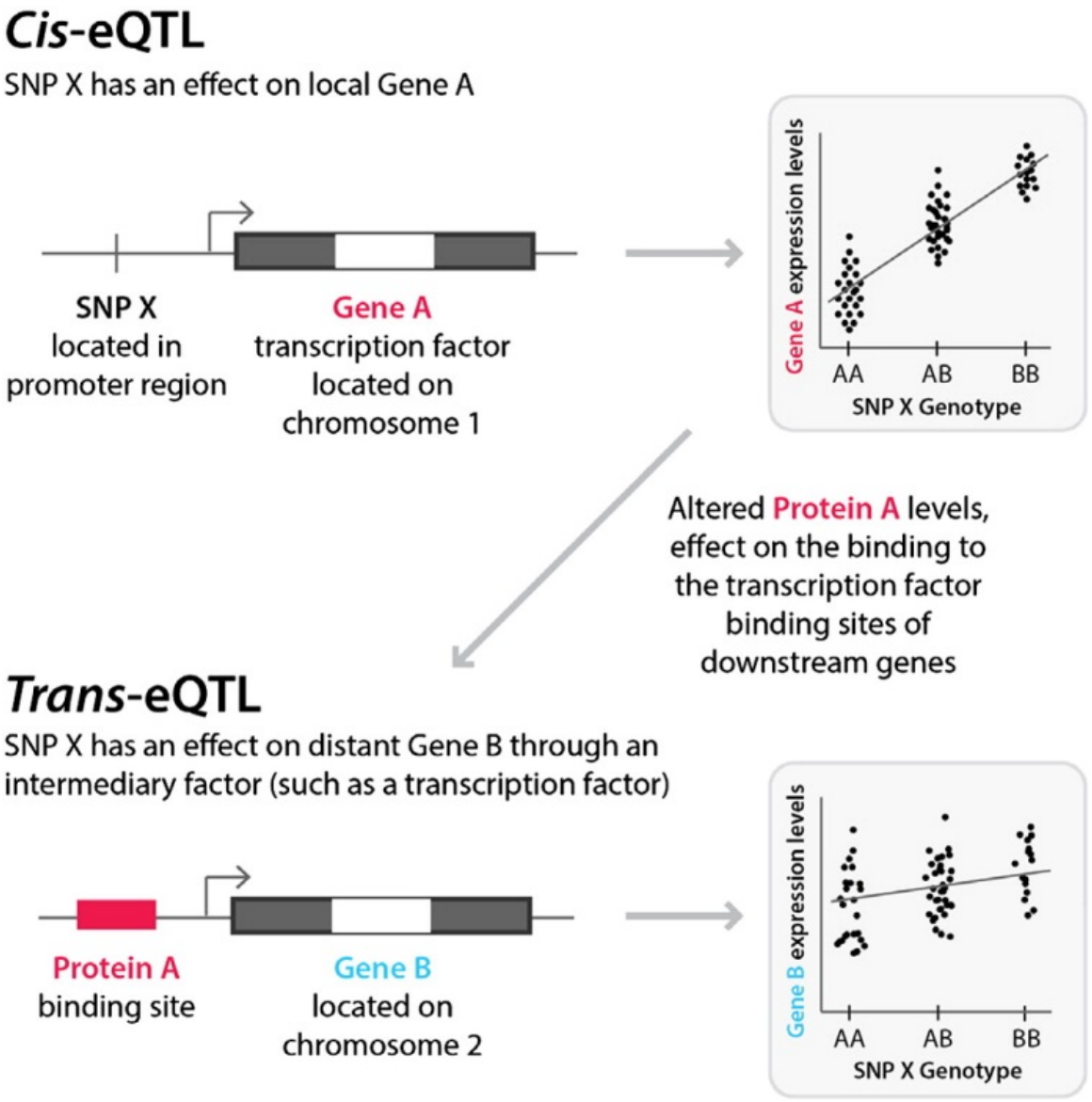
Zhernakova et al. NG 2016

Enrichments of eQTL



Zhernakova et al. NG 2016

Local vs distal effects



Westra *et al.* 2014 Biochimica et Biophysica Acta

Goal

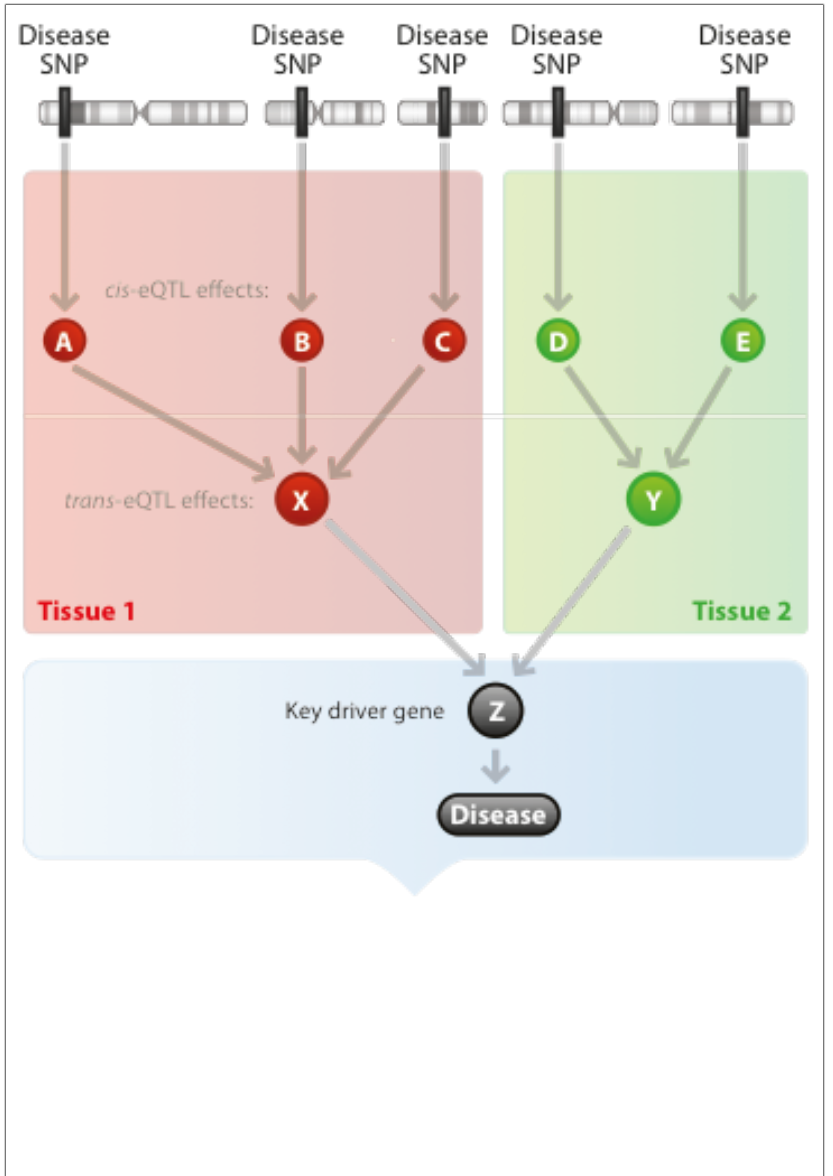
Big-data techniques

Genome-wide association studies

cis-eQTL mapping

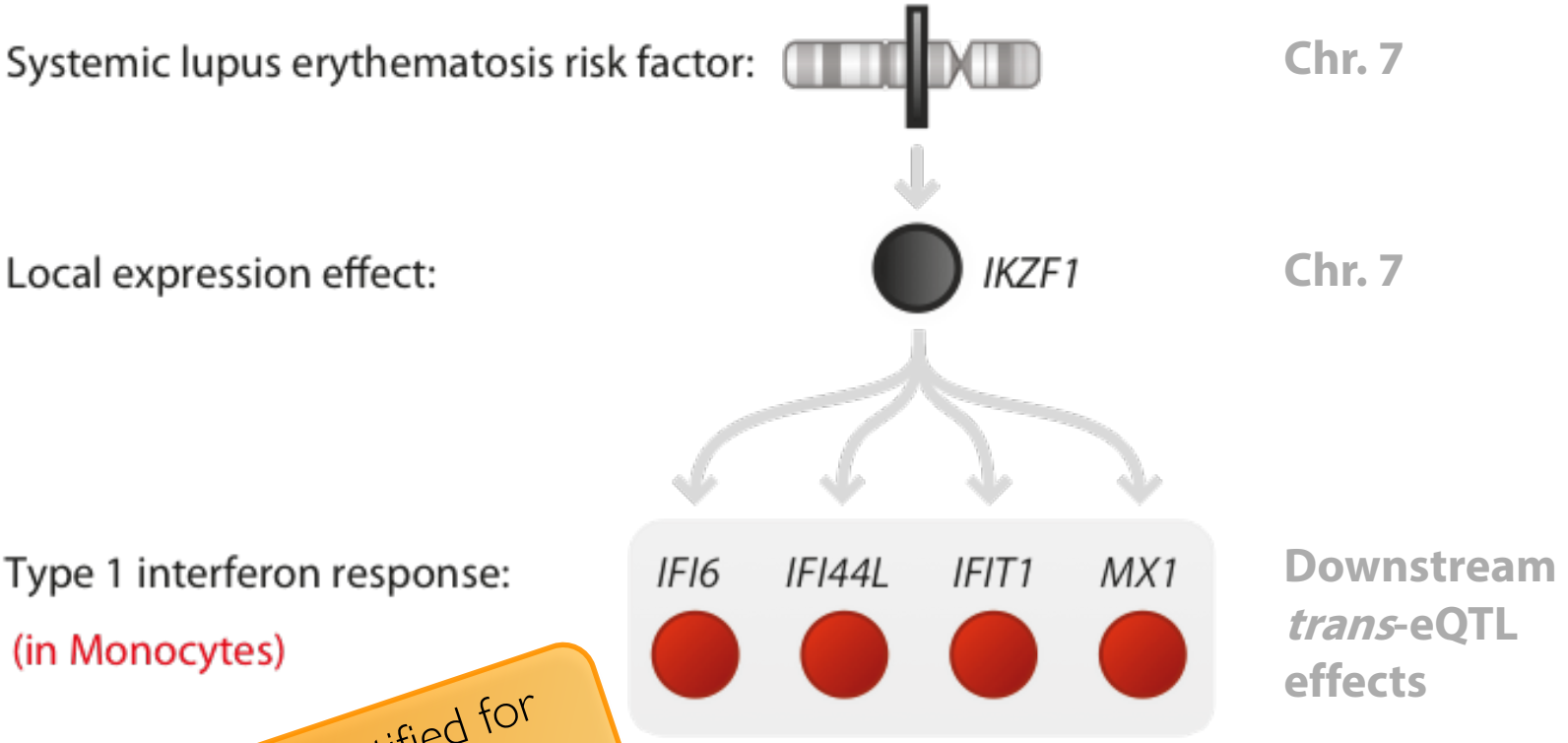
trans-eQTL mapping

Key driver gene identification



Get larger sample-sizes: meta-analysis in 5,311 samples

Lude Franke



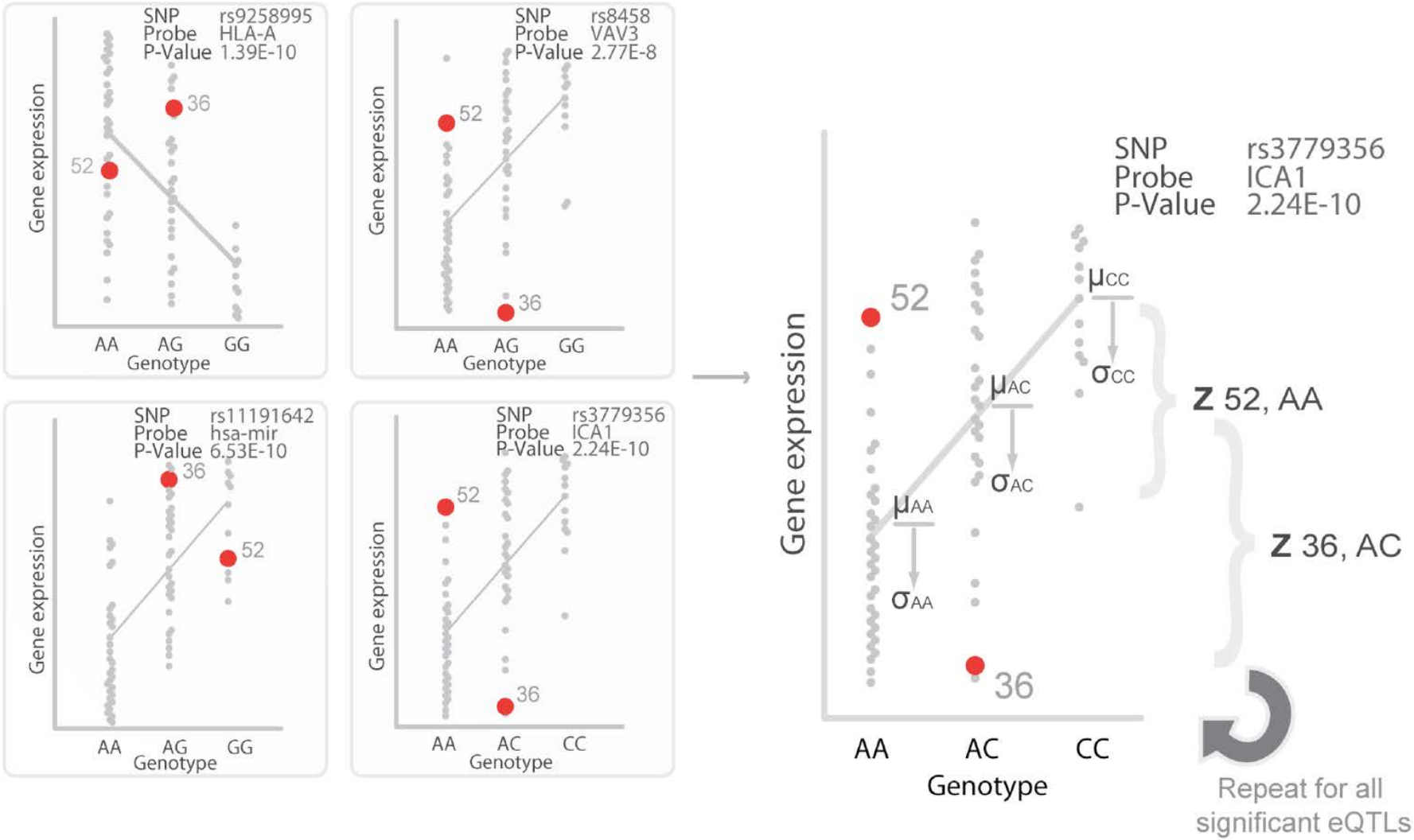
Downstream effects identified for >200 genetic risk factors

New meta-analysis ongoing in 25,000 blood samples ongoing

Westra *et al*, Nature Genetics 2013

Sample mix-ups: how to identify them

Lude Franke

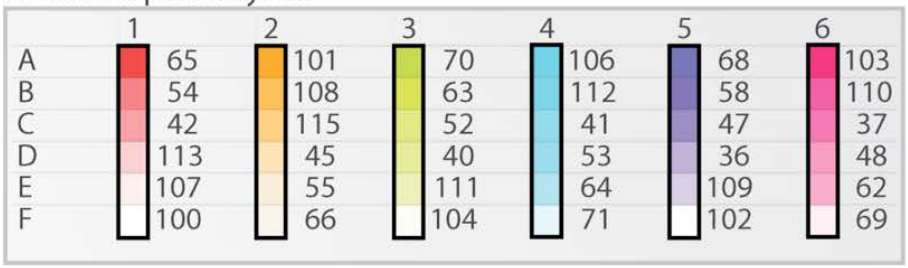


Westra et al, Bioinformatics, 2011

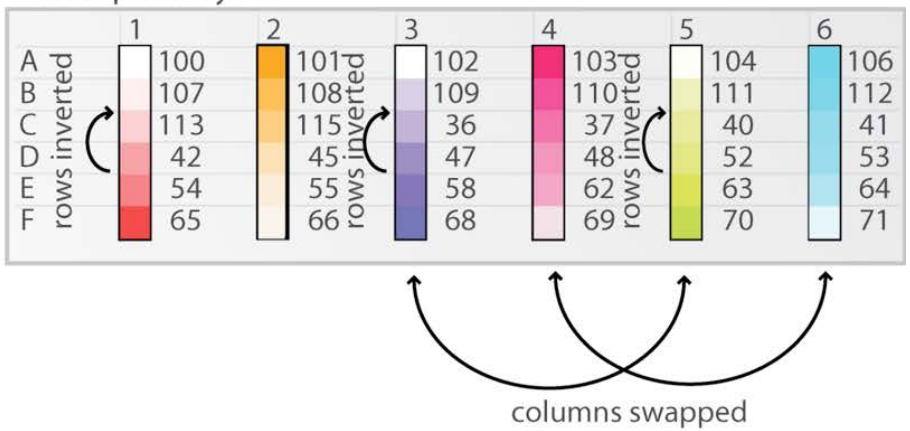
What happened to our data

Lude°Franke

Assumed plate layout



Actual plate layout

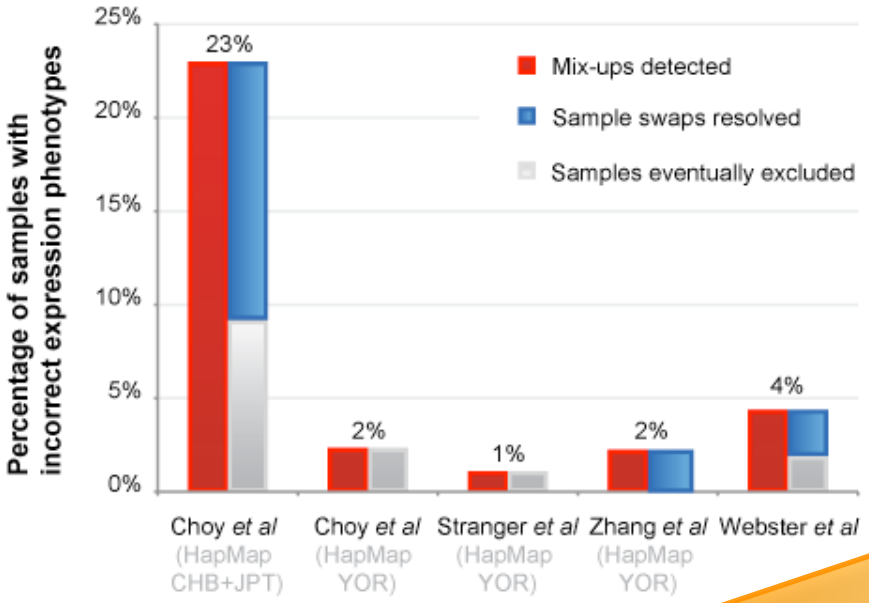


Westra et al, Bioinformatics, 2011

Sample mix-ups: do they happen?

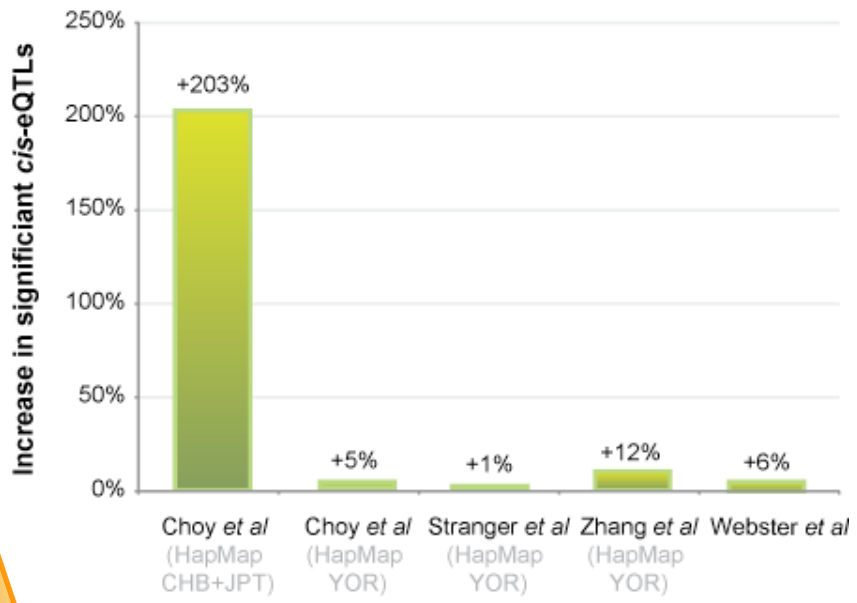
Lude Franke

eQTL datasets with mix-ups



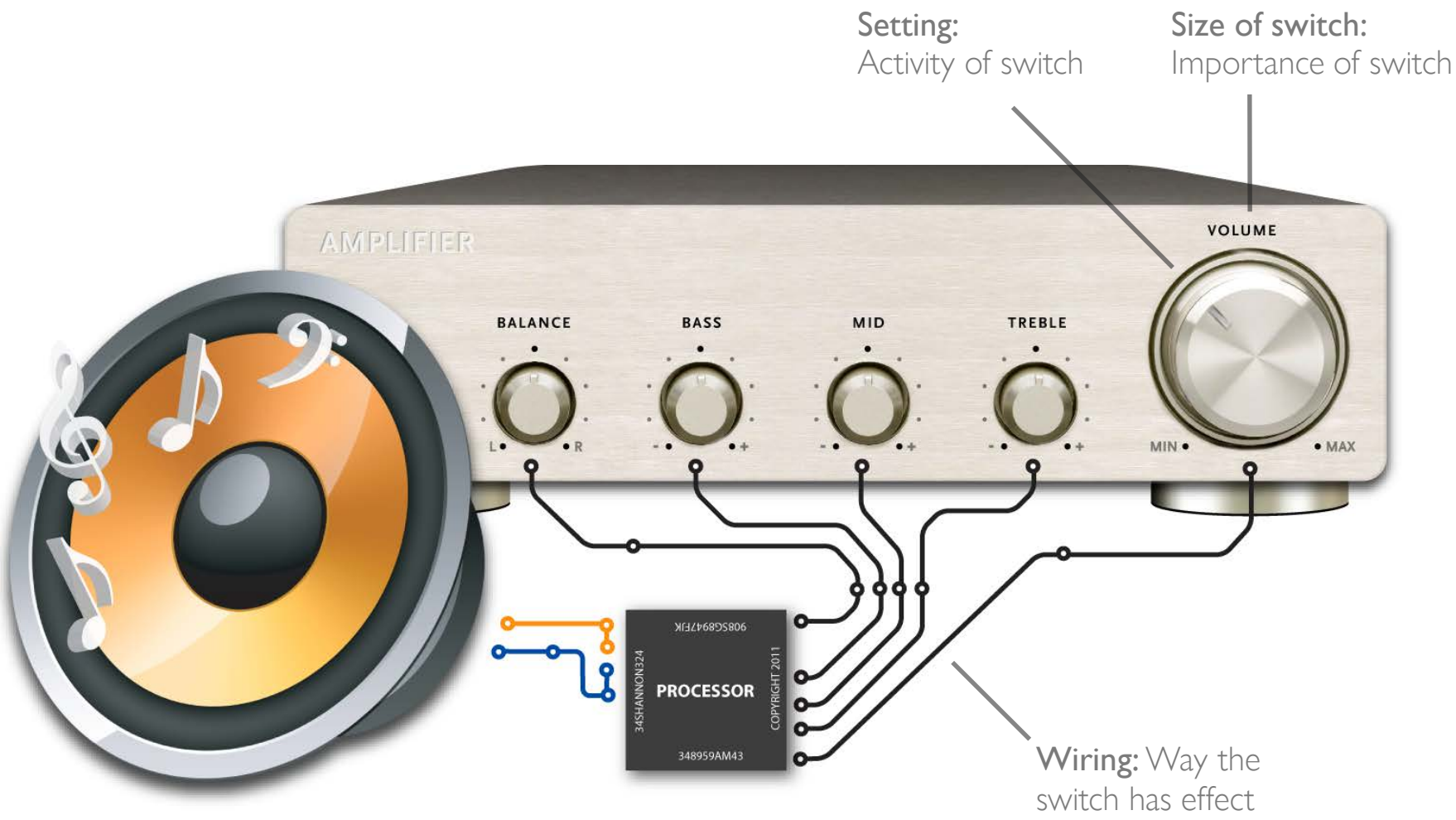
On average 3% of eQTL samples are mixed-up

Effect of correcting for these mix-ups

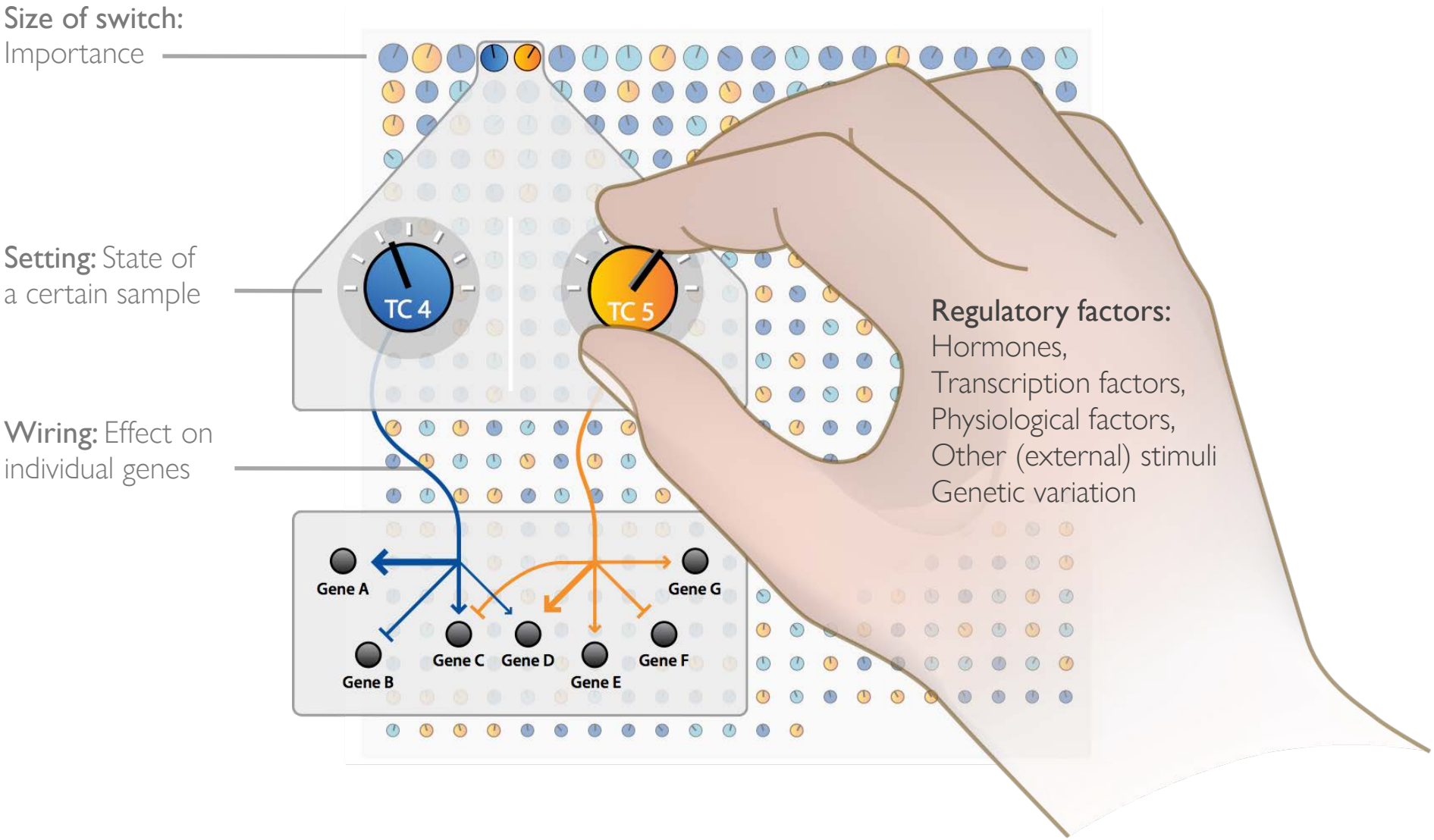


Westra et al, Bioinformatics, 2011

Amplifier can change many aspects of music

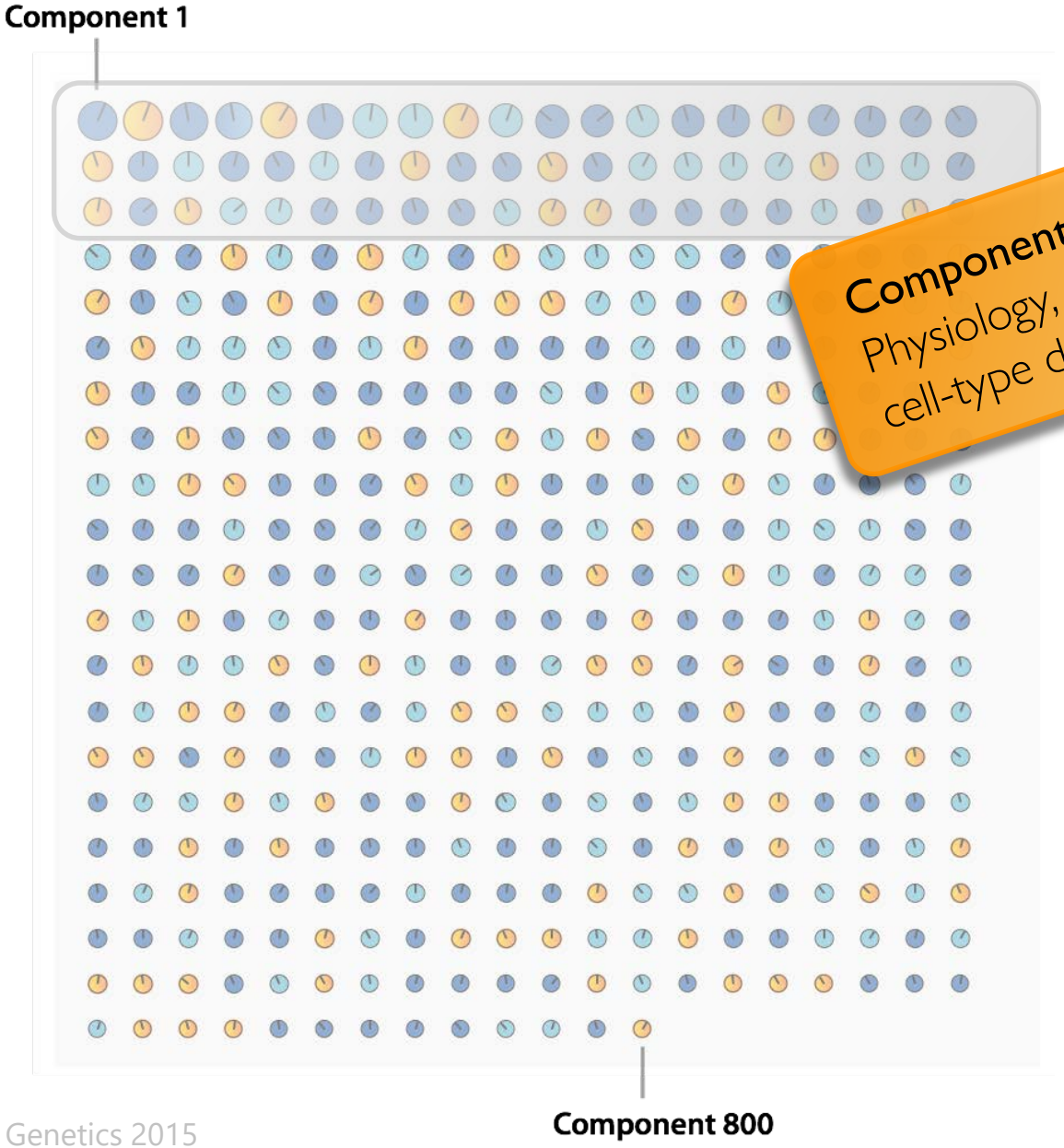


A control panel that determines gene expression?



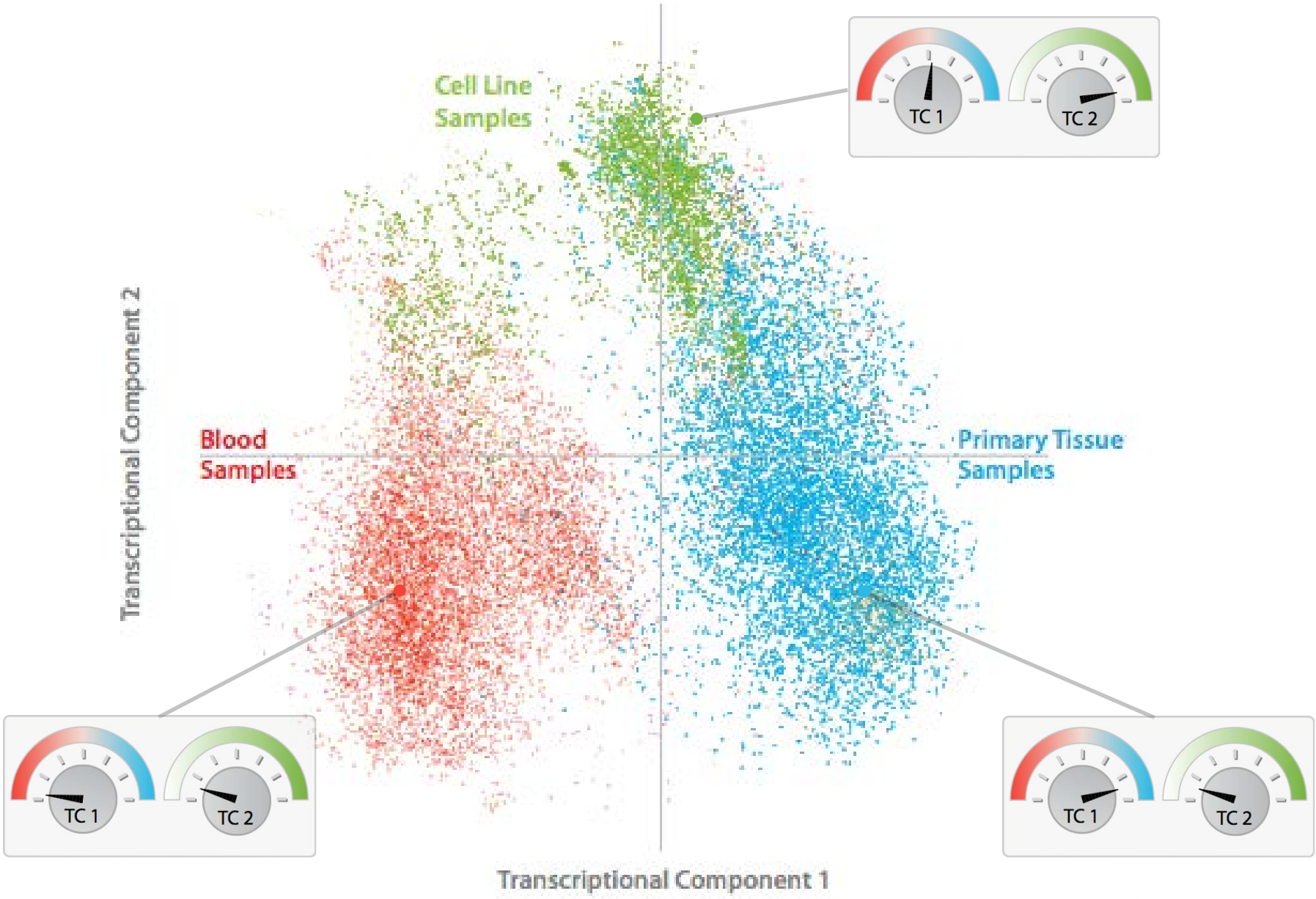
Fehrmann *et al*, Nature Genetics 2015

800 'transcriptional components': Component 1 - 50

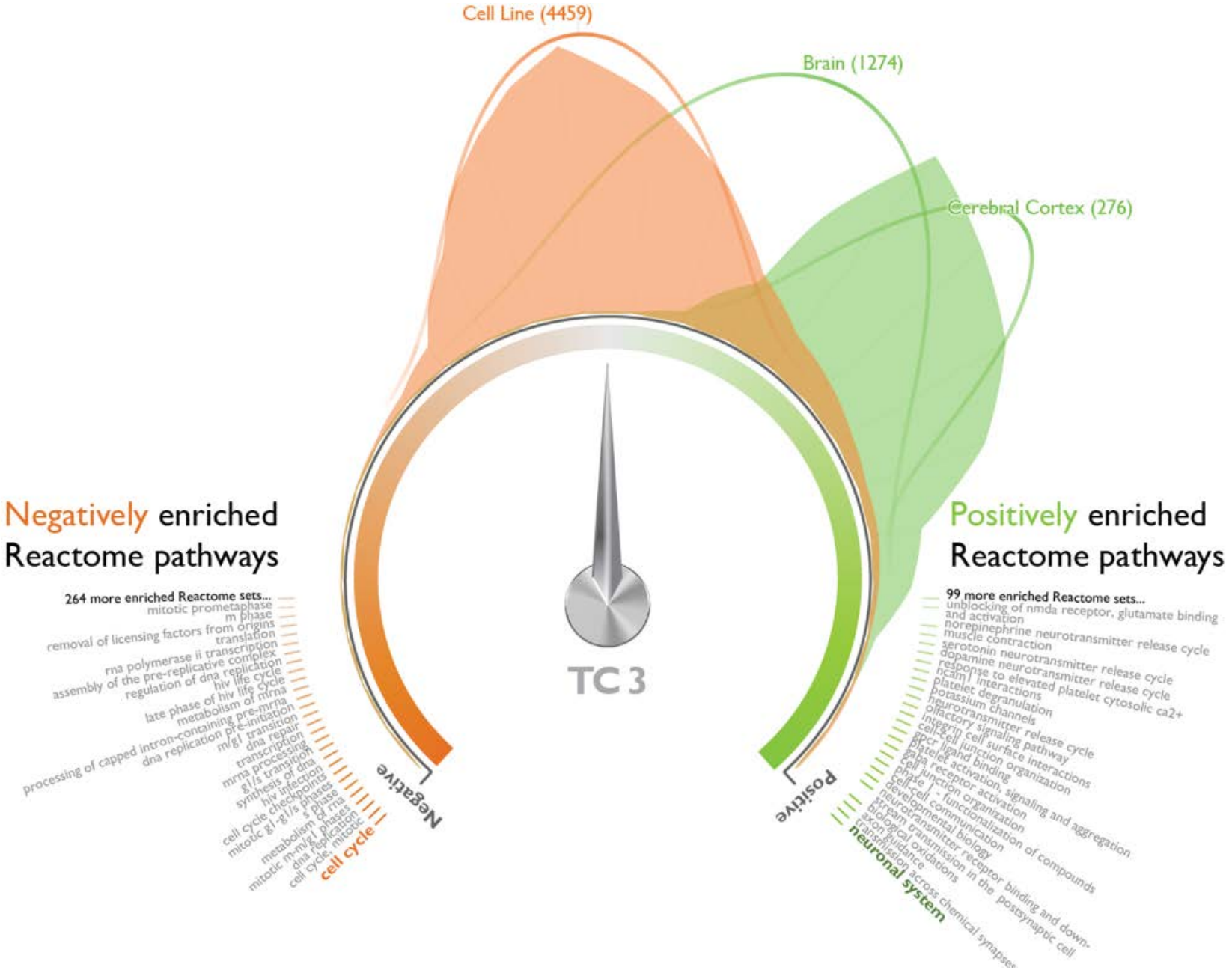


Fehrmann *et al*, Nature Genetics 2015

Component 1 and 2



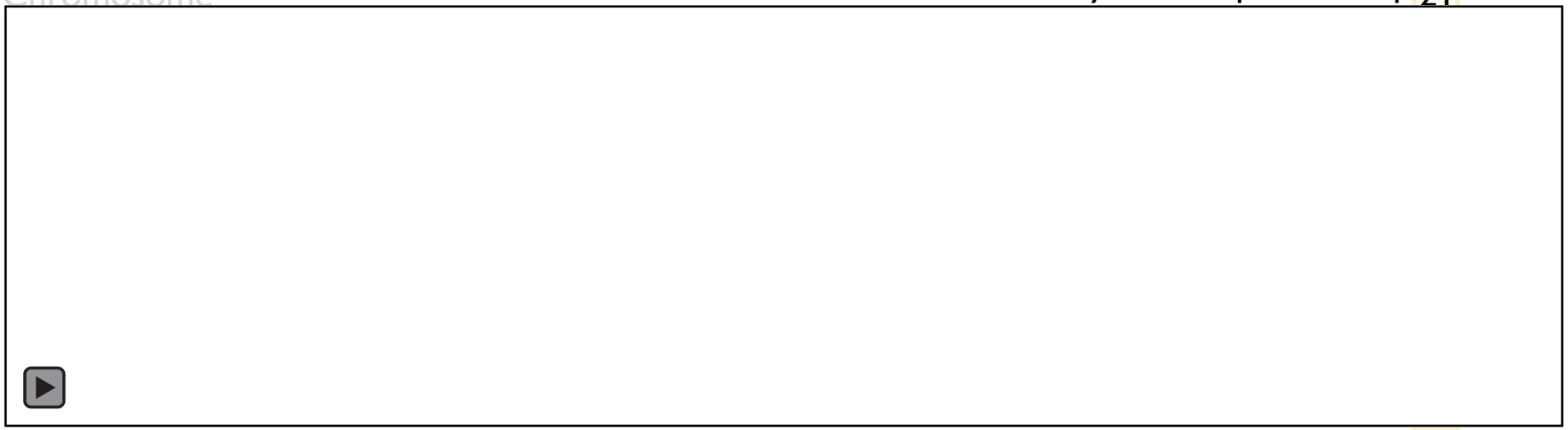
Transcriptional component 3



Detection cytogenetic aberration in expression data

Chromosome

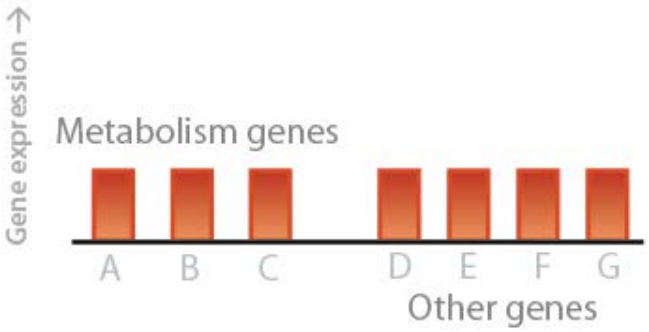
Down Syndrome patient: dup 21



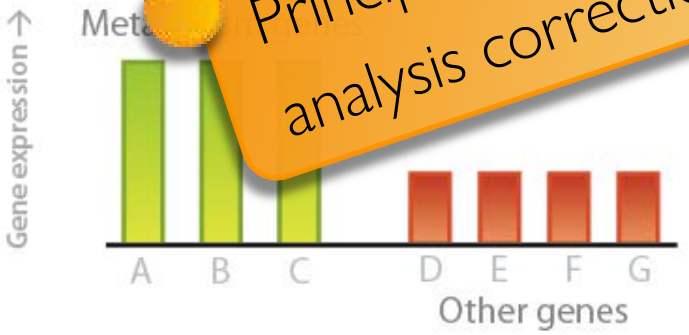
Large proportion of expression variation is determined by genetic variation but due to e.g.:

- Physiological state of samples
- Environmental state of samples (e.g. fasting vs. non-fasting)

RNA blood expression
when you wake up



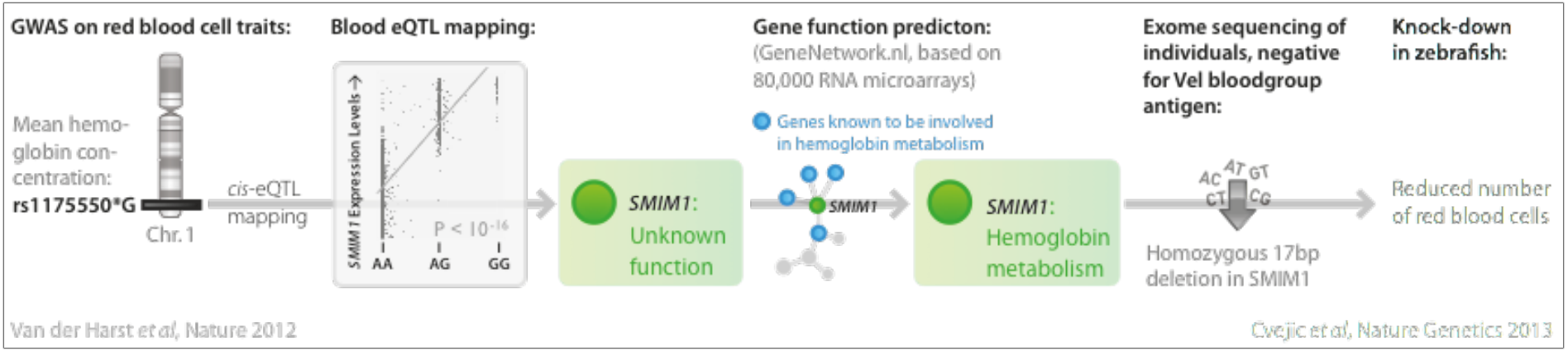
RNA b
expressi
after din



Get rid of this 'noise'?
Principal component analysis correction

Dubois *et al*, Nature Genetics 2010

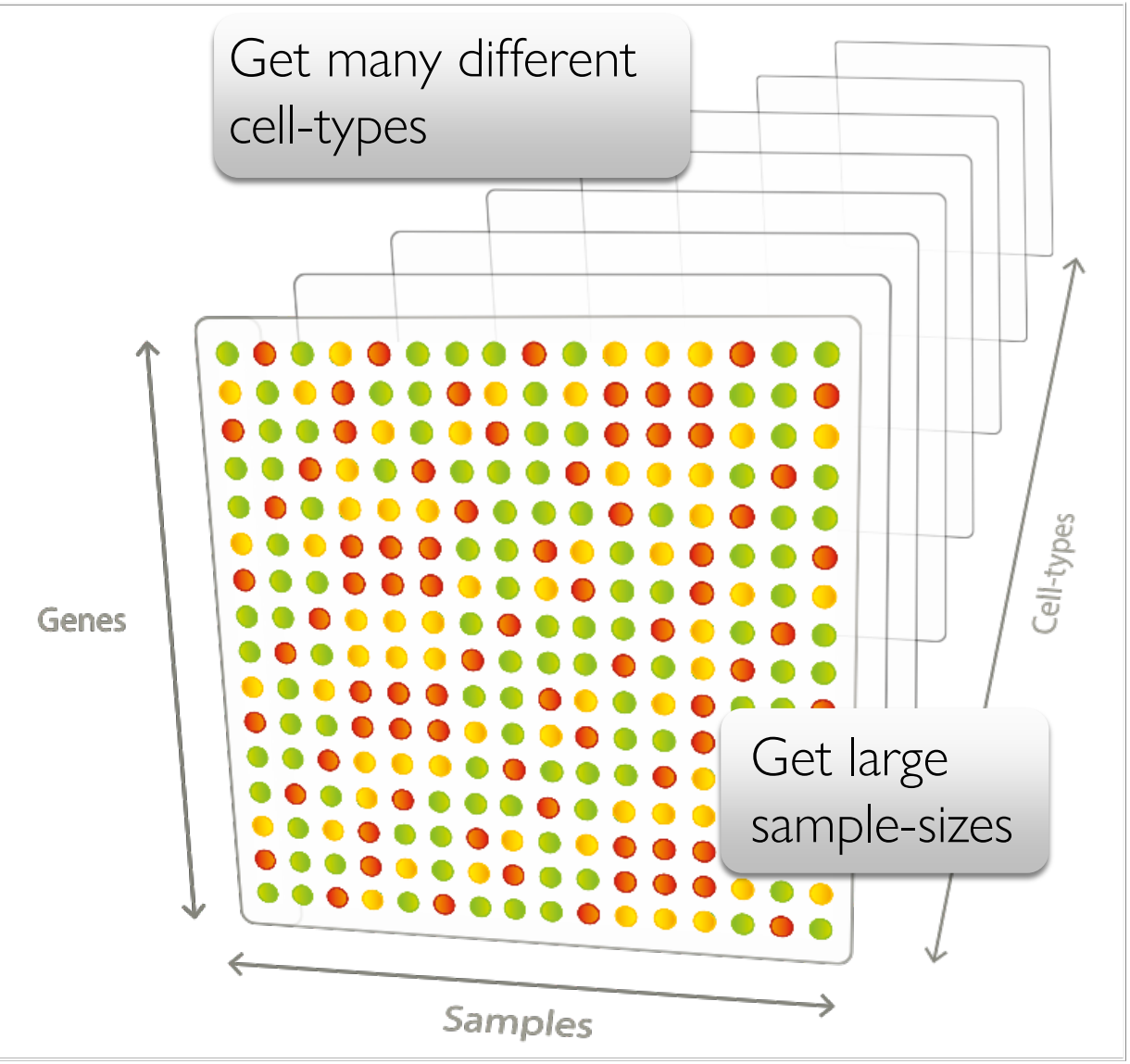
GeneNetwork gene function predictions



Amounts of data integrated:

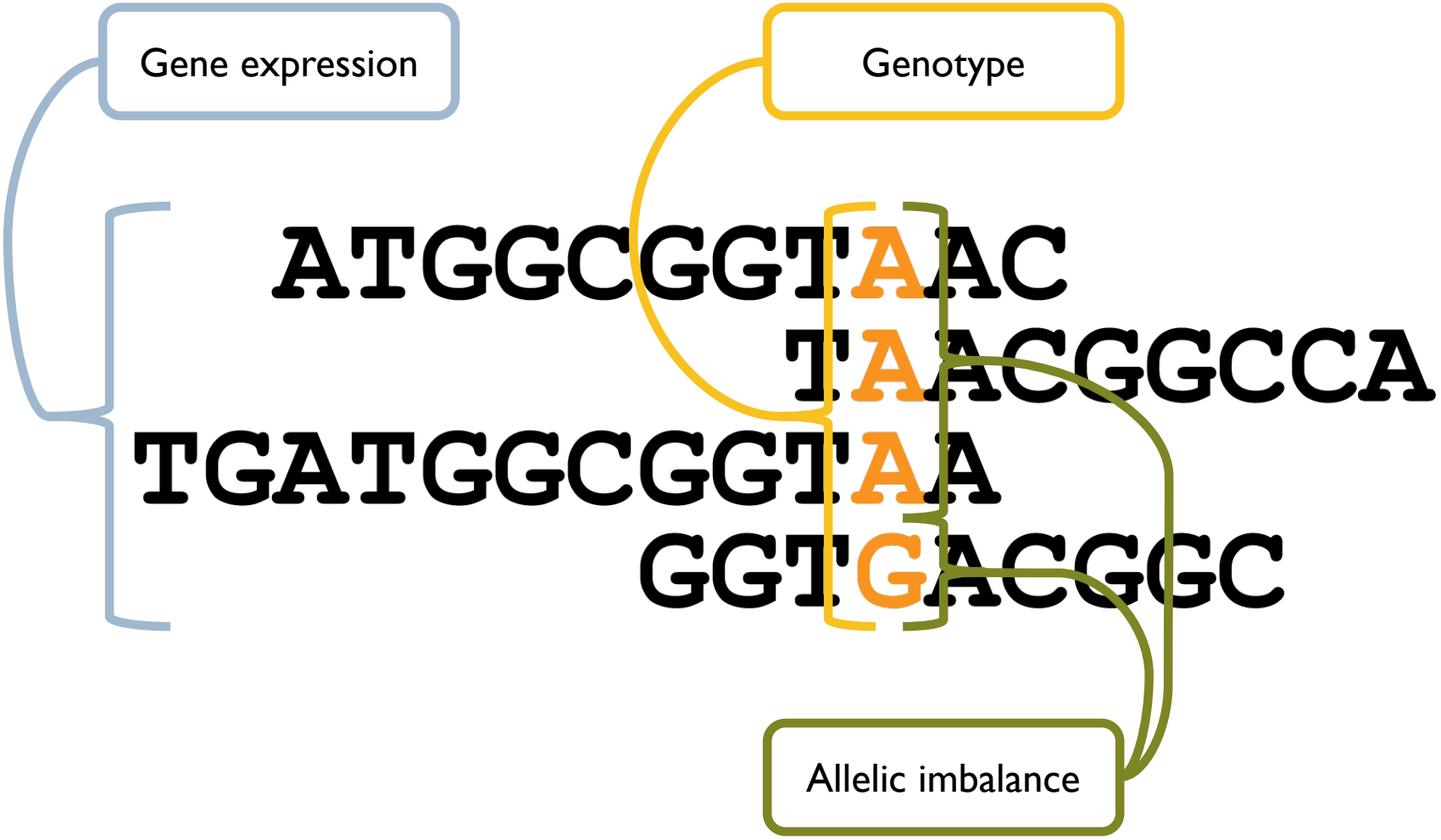
GWAS in 135,000 samples	eQTL mapping in 1,500 samples	Transcriptomics in 80,000 samples	Exome sequencing	Wet lab proof

Using public data

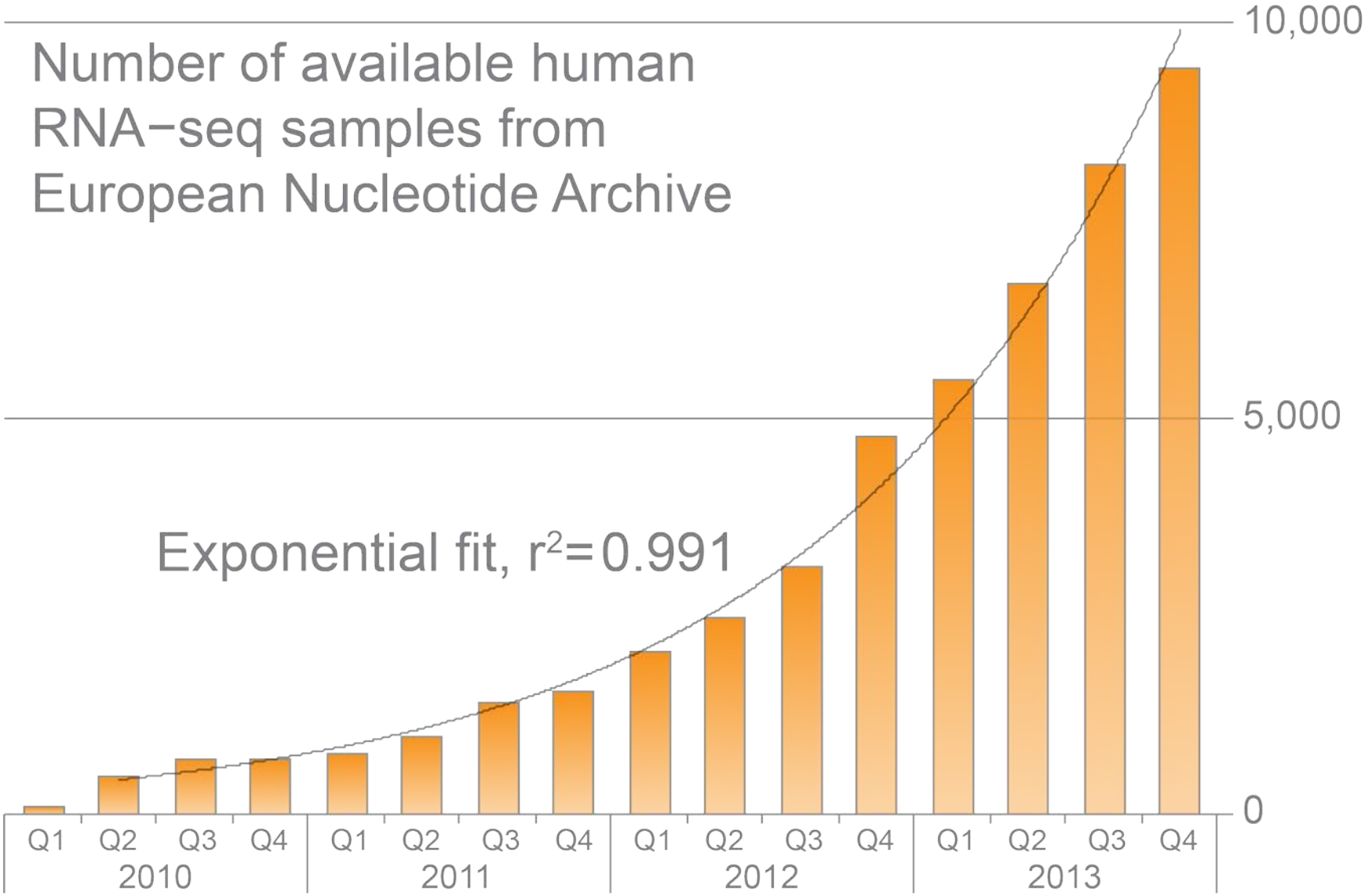


Recycle publicly available gene expression data?

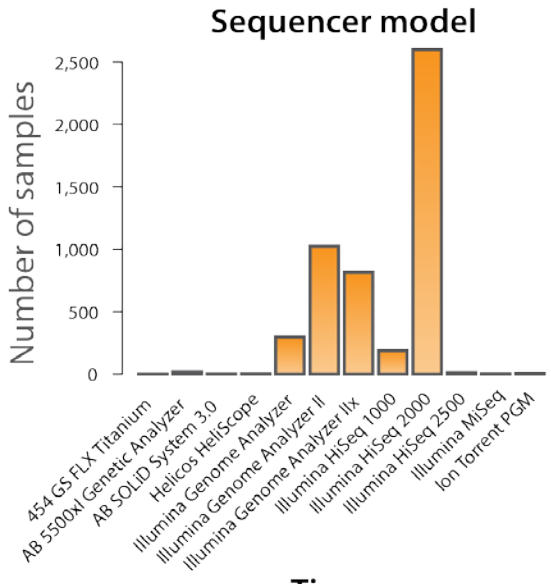
RNA sequencing

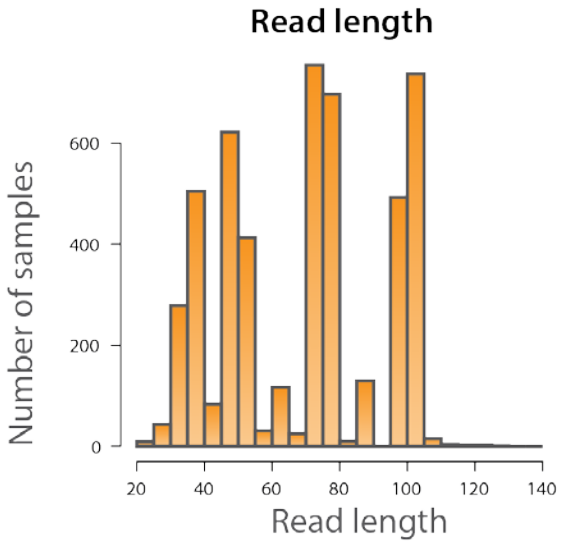


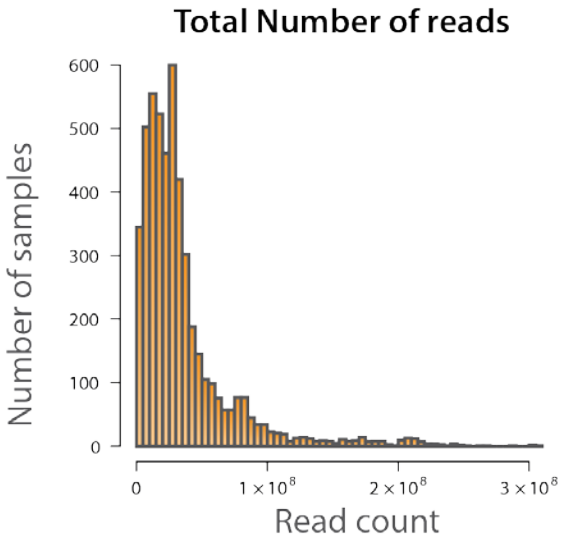
Available data is growing rapidly

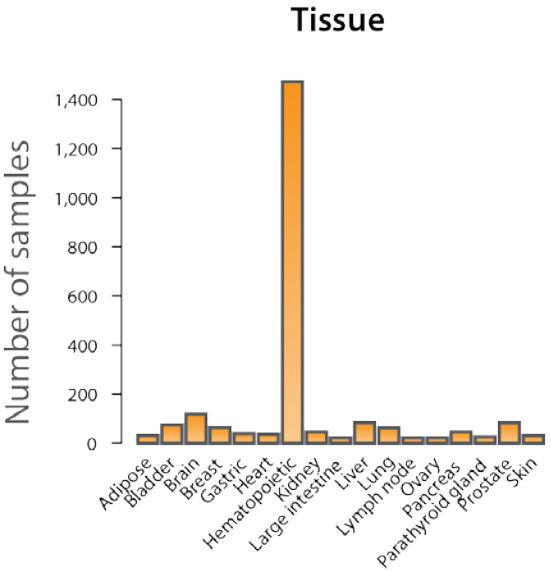


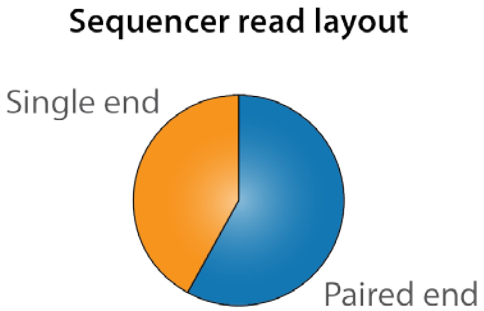
Heterogeneous dataset

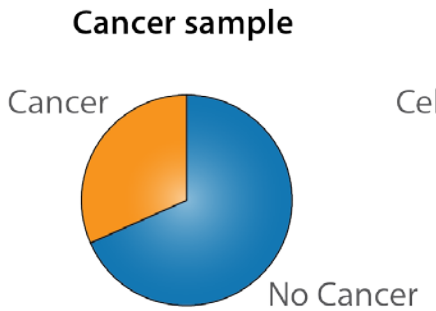


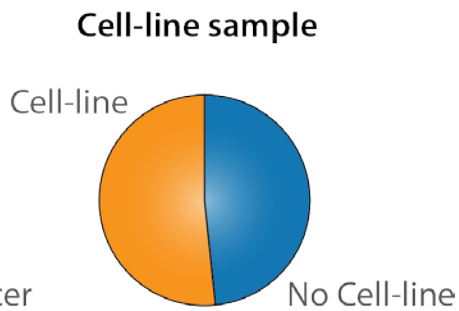




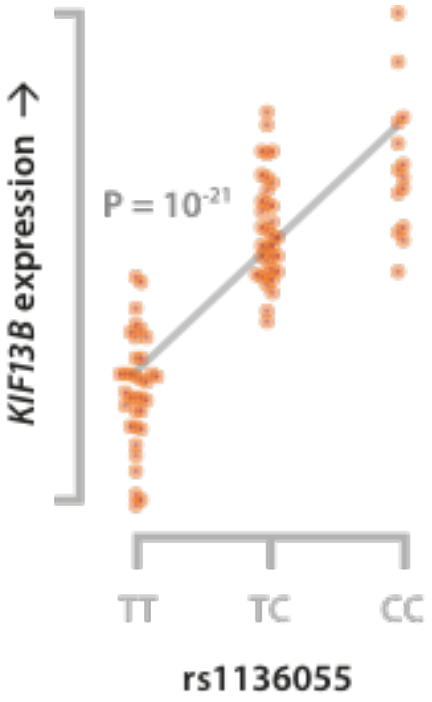






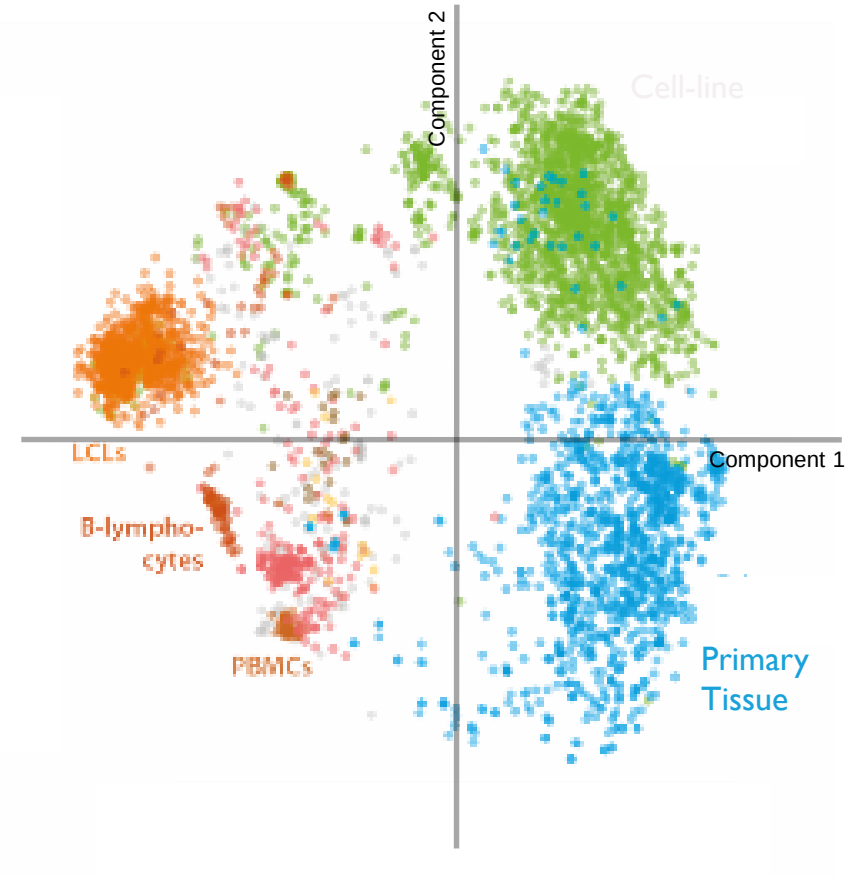


Derive SNP genotypes from RNA-seq data



Infer genotypes

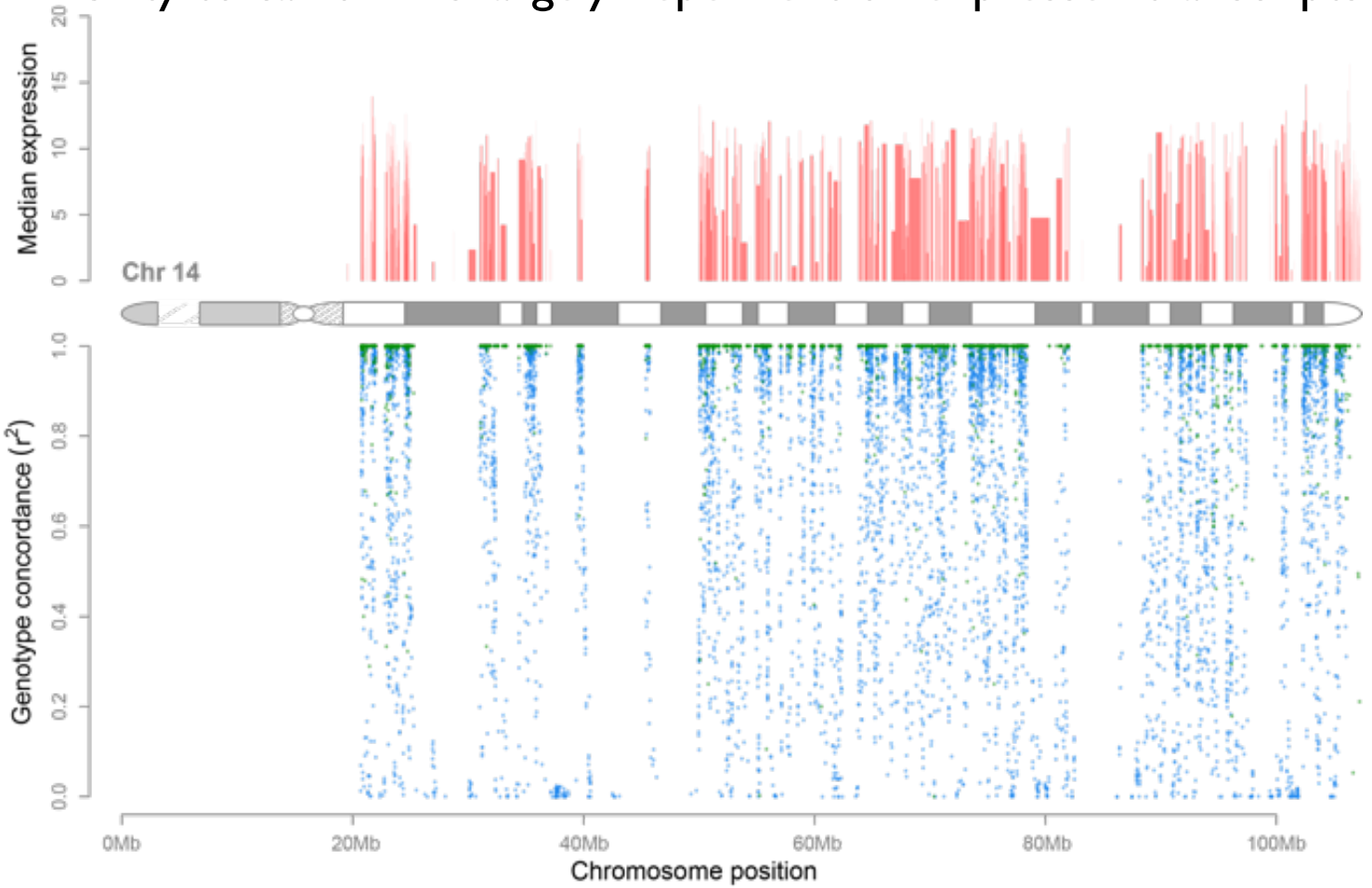
Public RNA-seq data (5,000 samples)



Deelen *et al*, Genome Medicine 2015

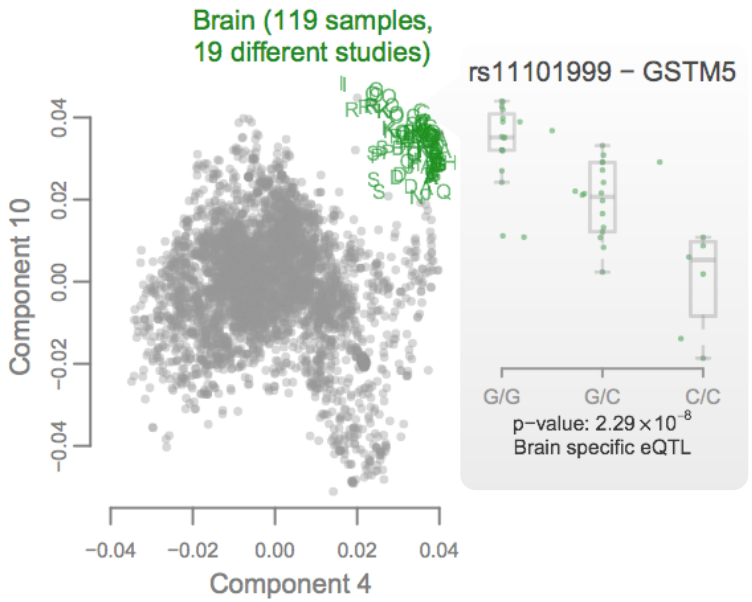
Calling genotypes in RNA-seq data

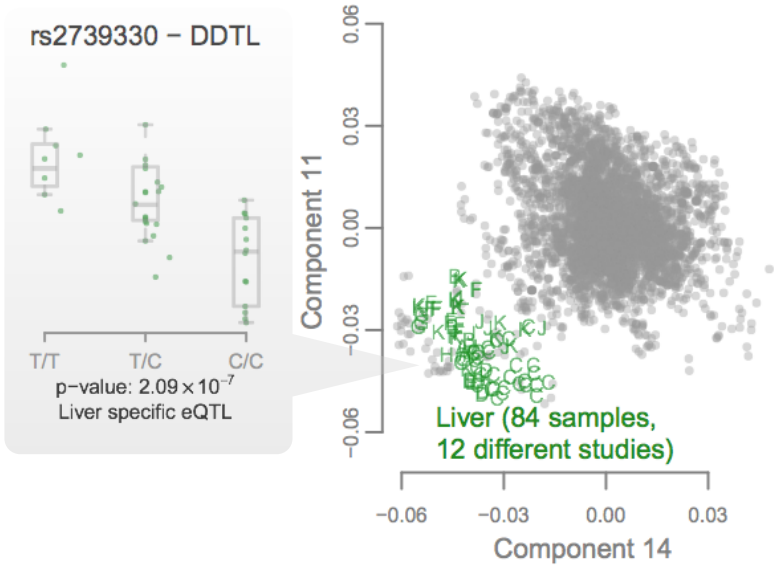
Ability to call SNP is largely dependent on expressed transcripts

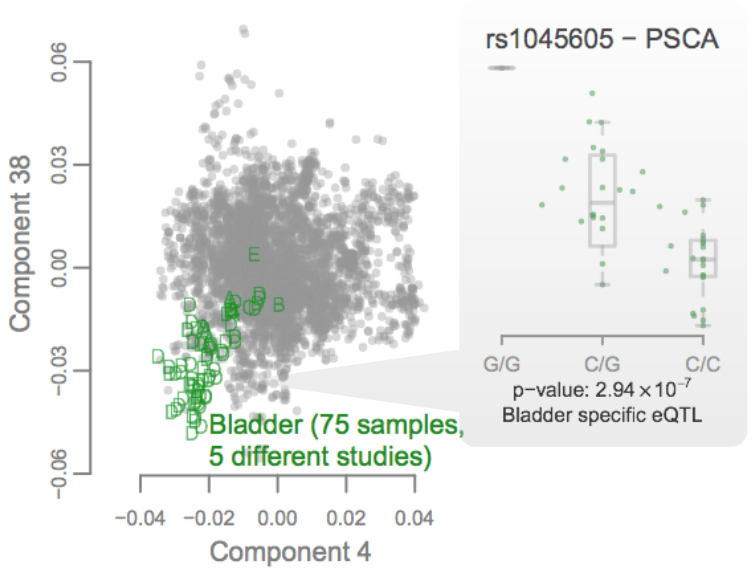


Deelen *et al*, Genome Medicine 2015

Tissue-specific eQTL mapping for free





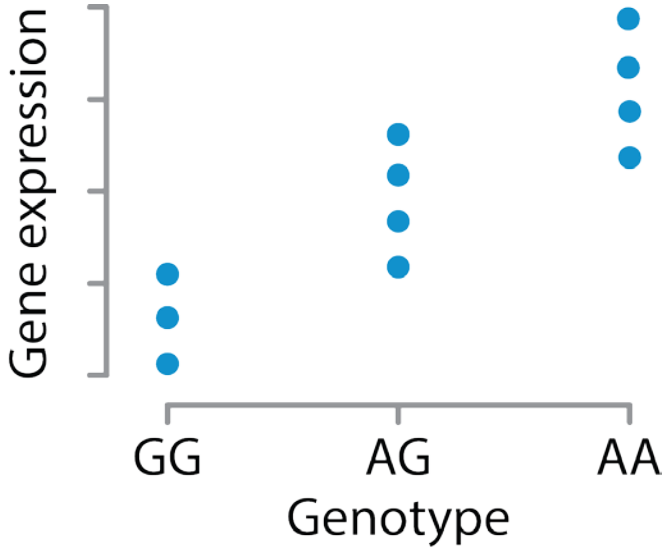


Deelen *et al*, Genome Medicine 2015

Investigation of rare variants

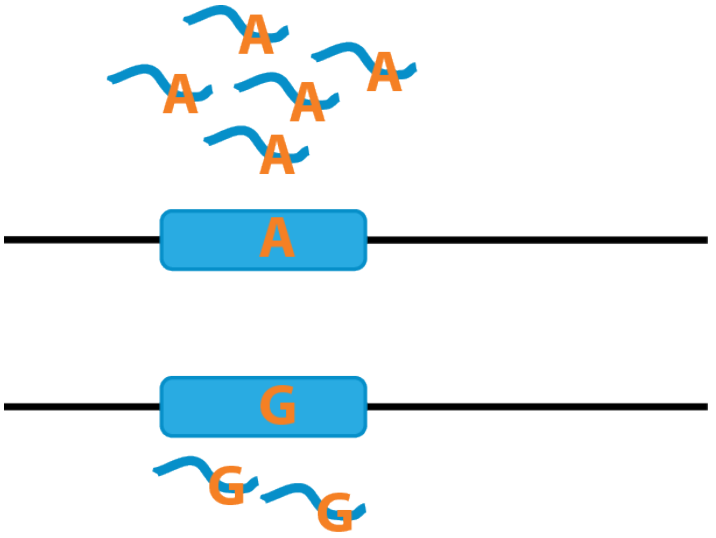
eQTL

- Correlation between genotype and expression
- Common variants



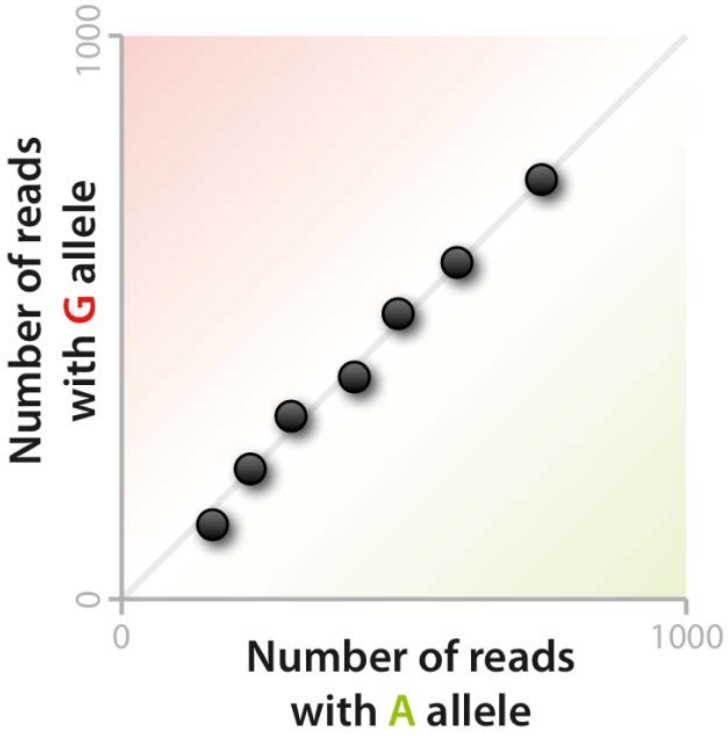
Allele Specific Expression

- Imbalance in expression
- Rare variants

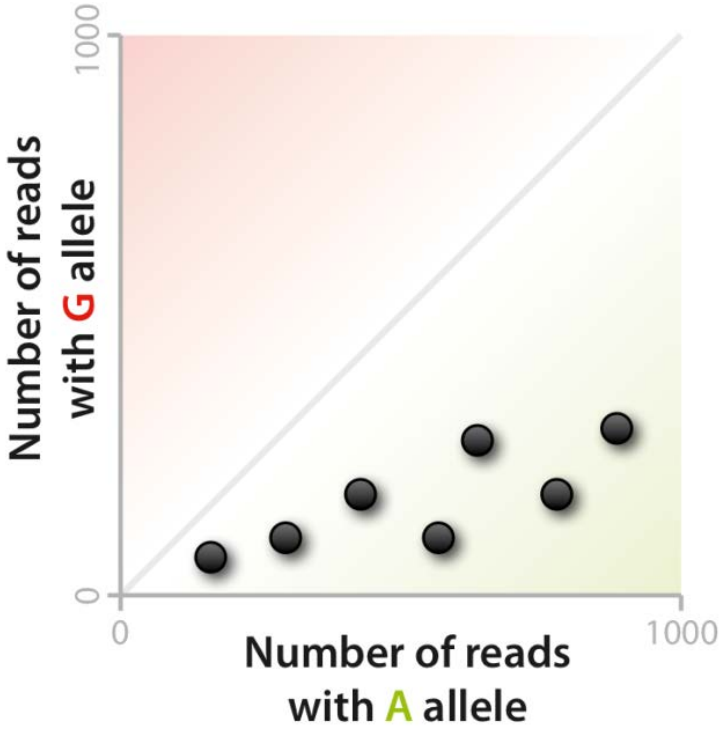


Allele Specific Expression (ASE)

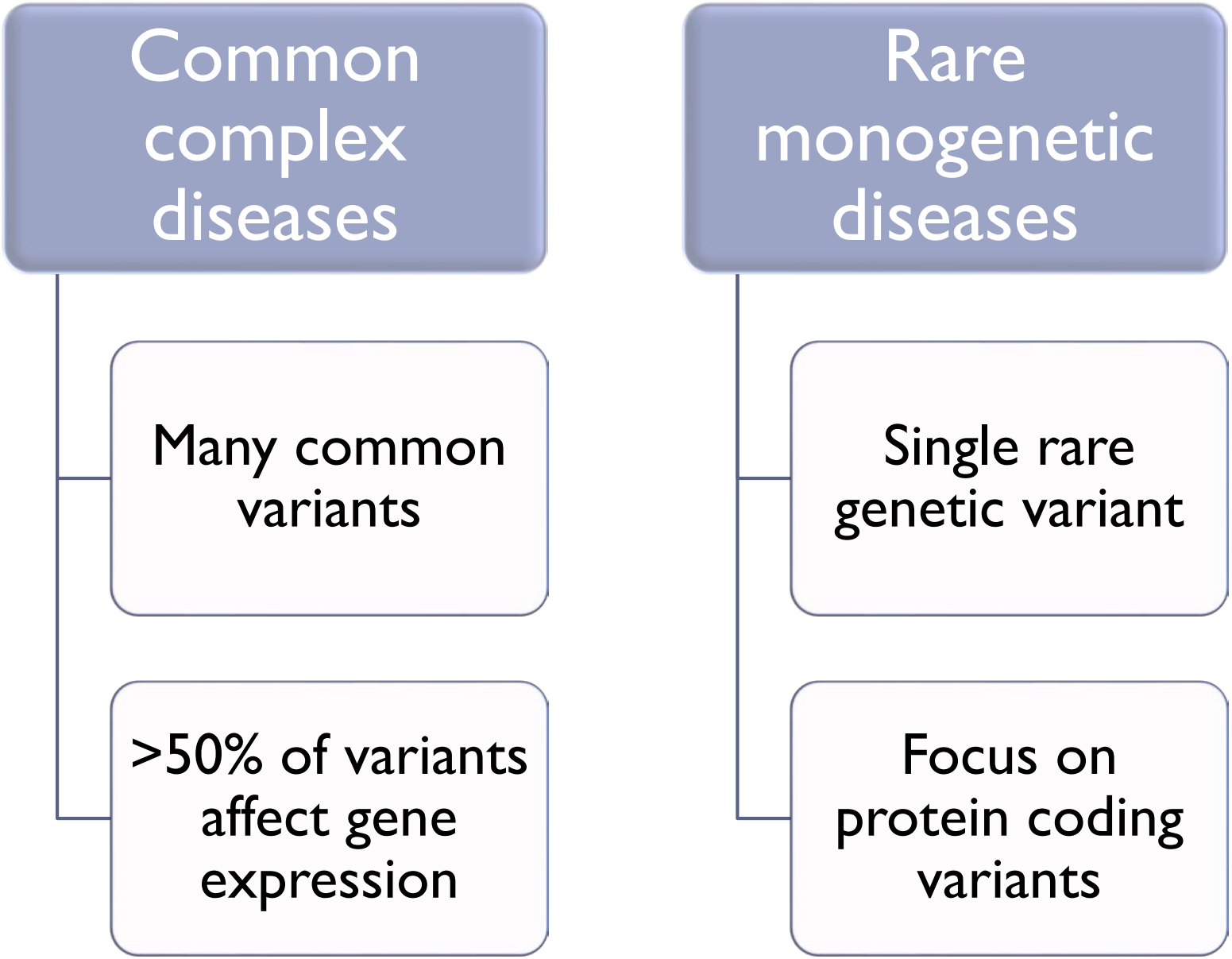
Alleles equally expressed



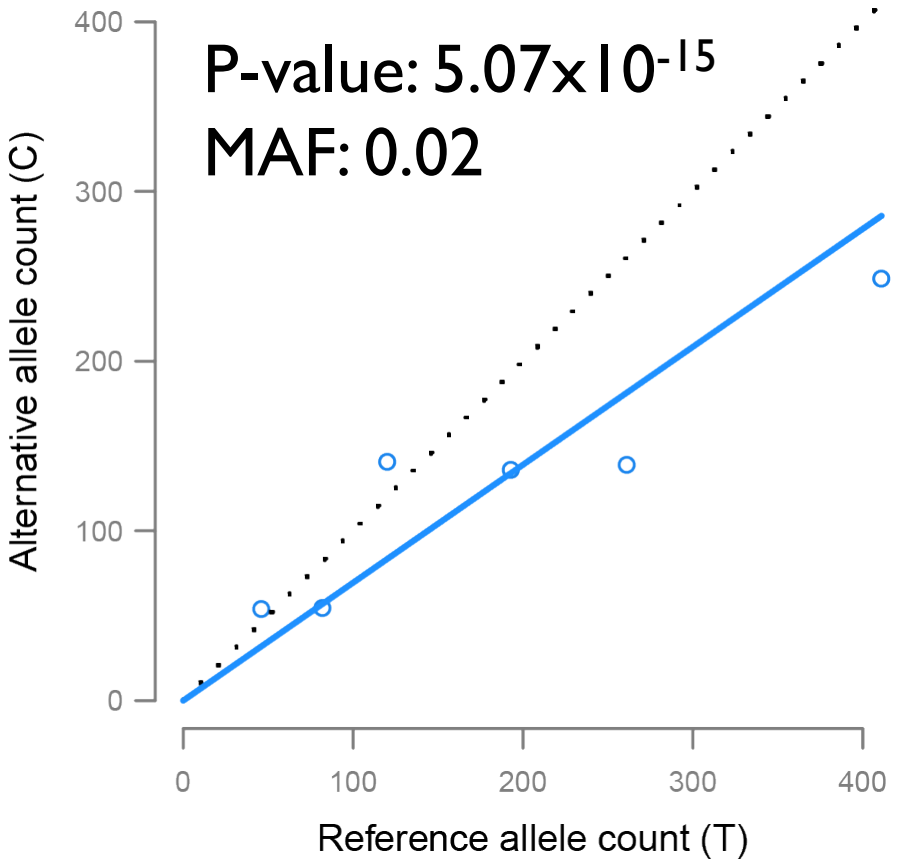
Bias towards A allele

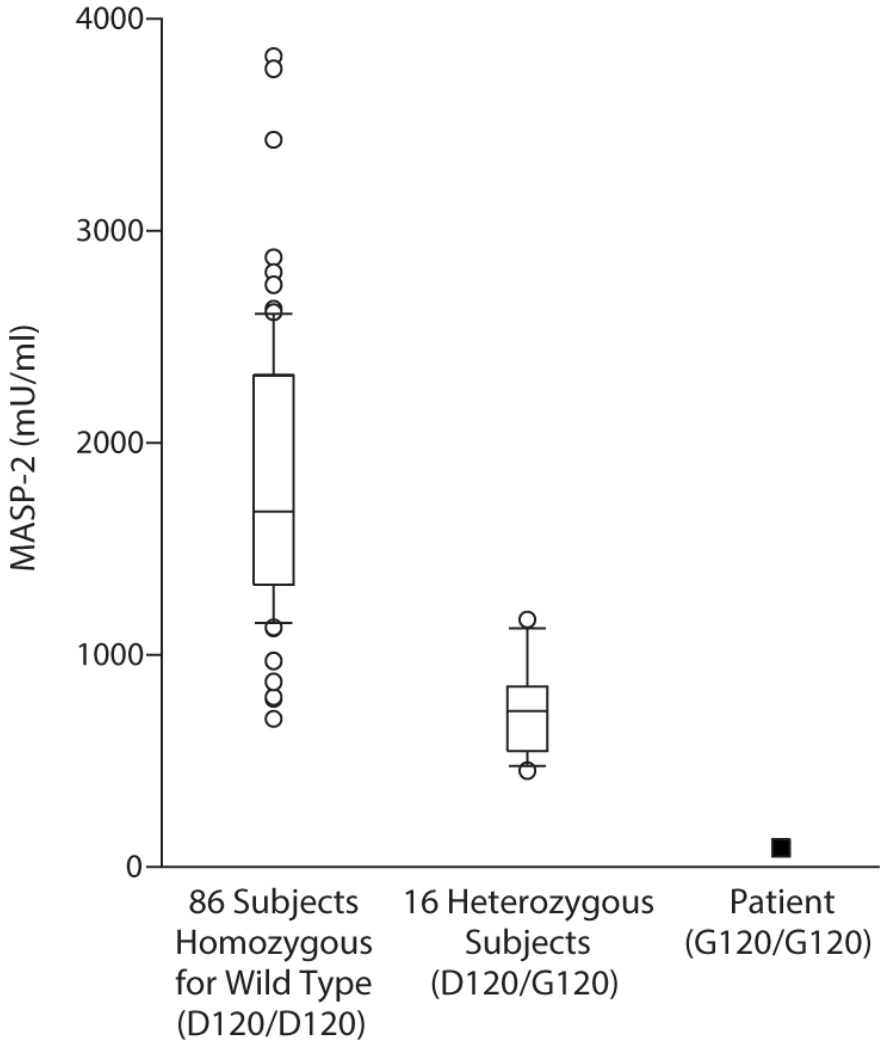


Why focus on rare variants?



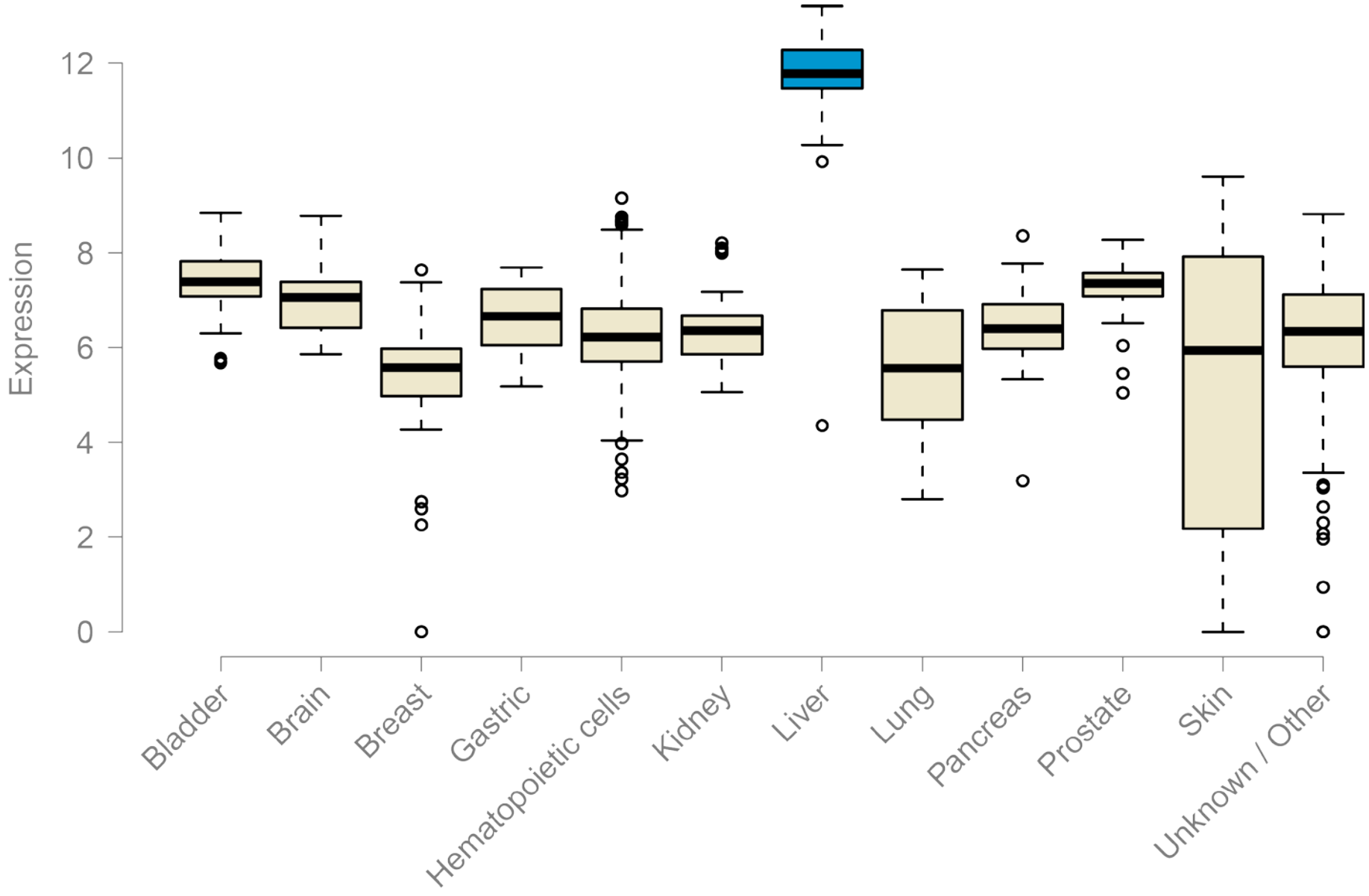
Example: MASP2 deficiency





Stengaard-Pedersen et.al NEJM 2003

MASP2 expression levels



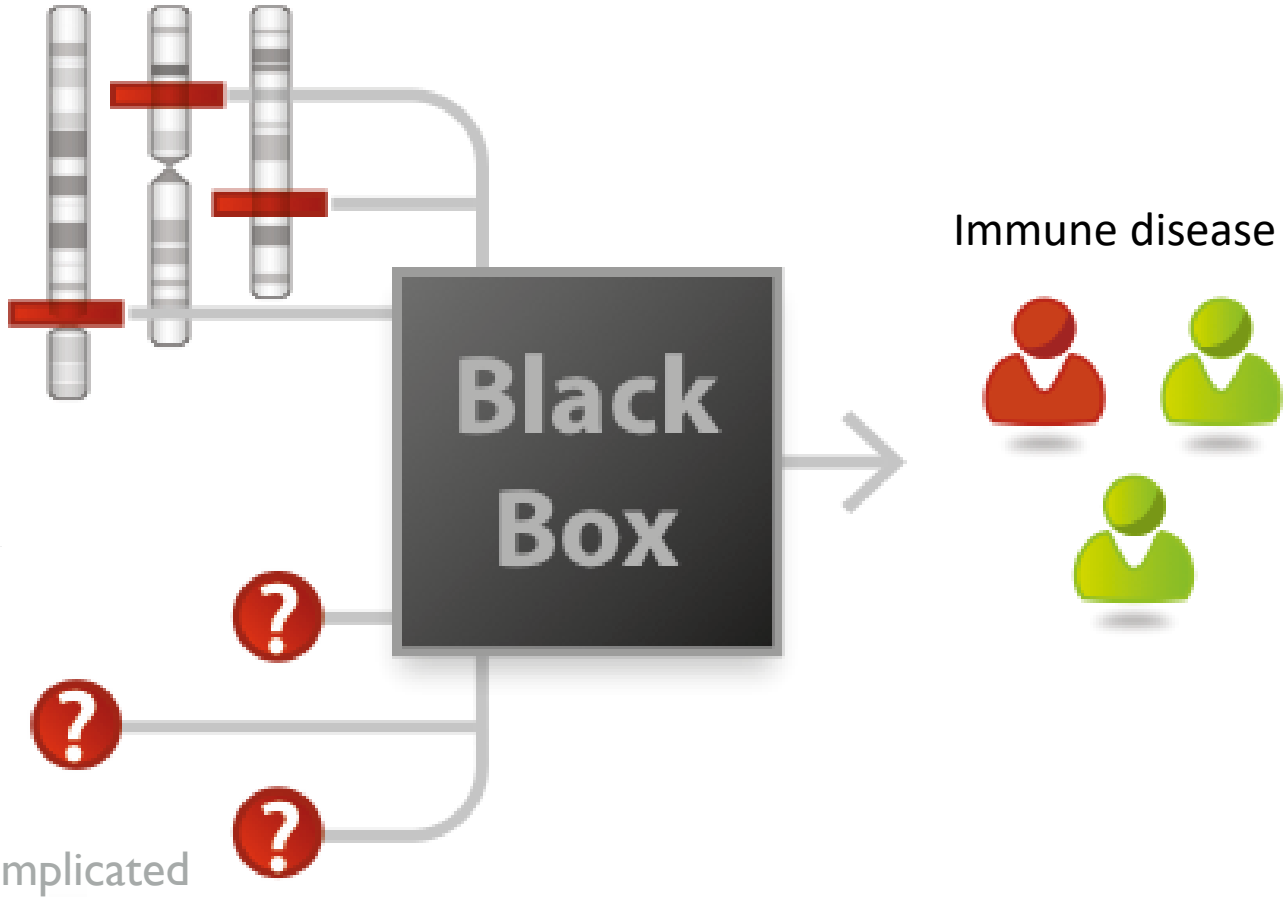
Interplay of genetic & environmental risk factors unknown

Genetic
risk factors

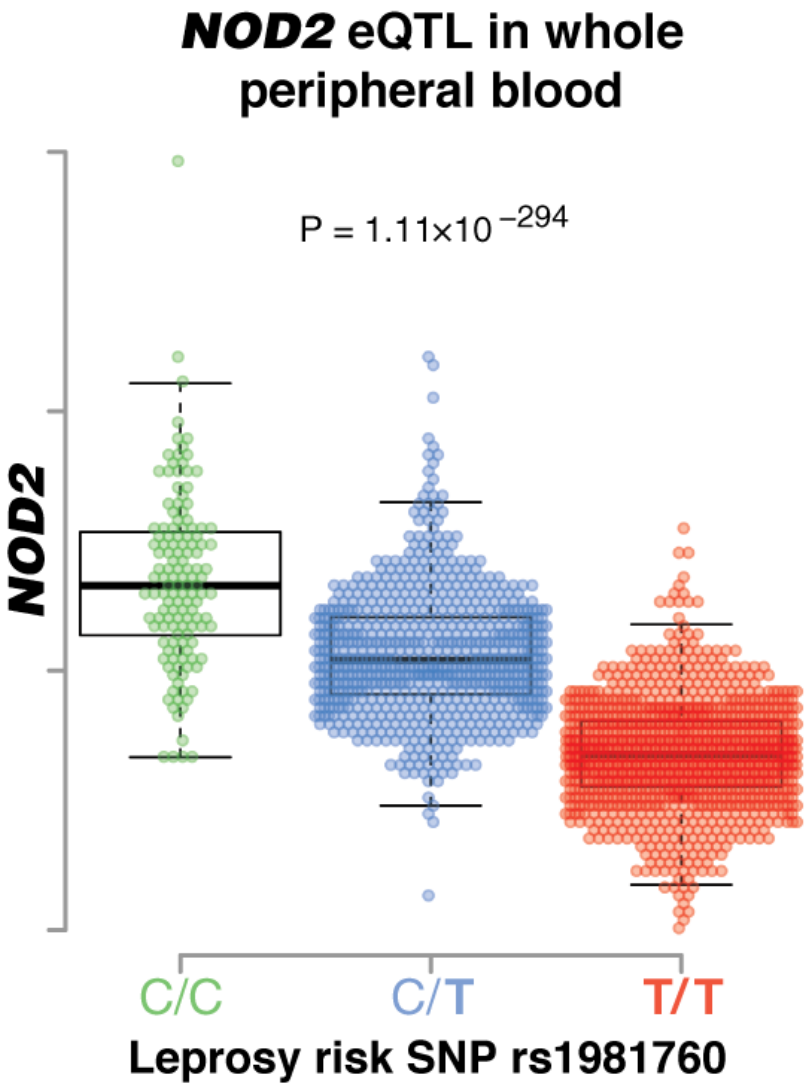
>1,000 known,
But small effects

Environmental
risk factors

Mostly unknown,
Viruses / bacteria implicated

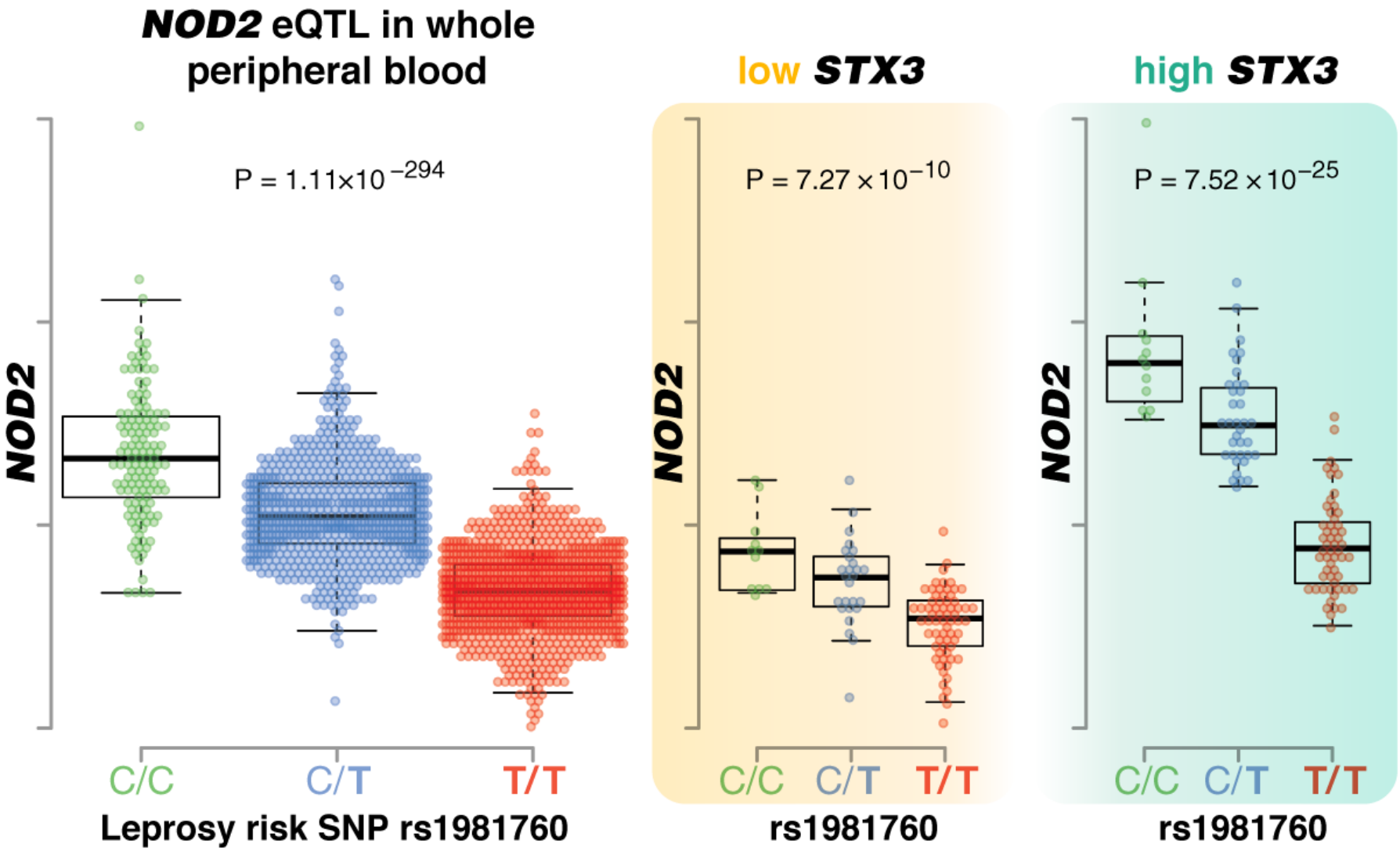


Detecting cell-type dependent eQTLs in whole blood



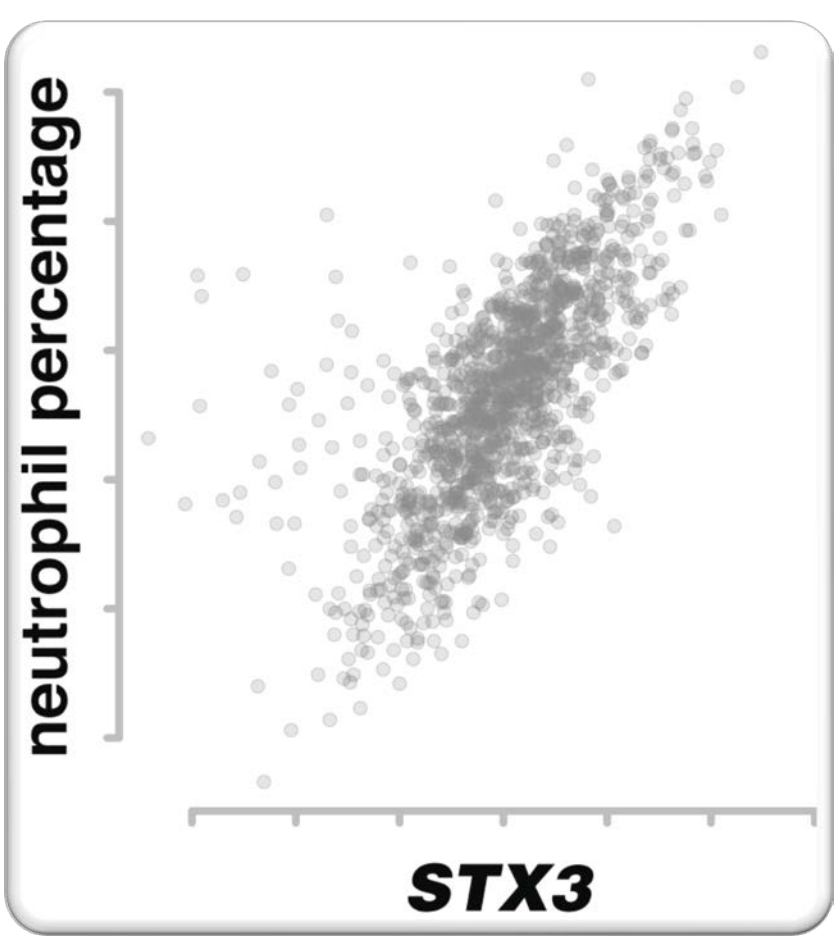
Zhernakova et al. NG 2016

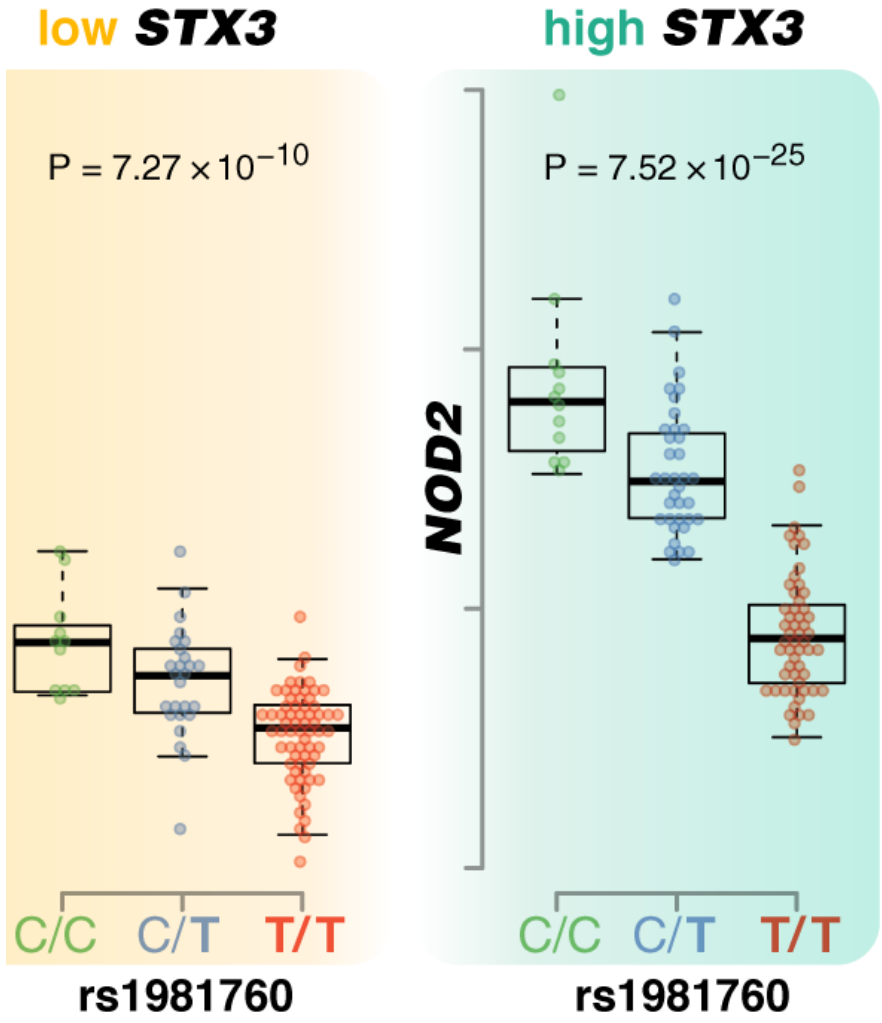
Detecting cell-type dependent eQTLs in whole blood



Zhernakova et al. NG 2016

Detecting cell-type dependent eQTLs in whole blood





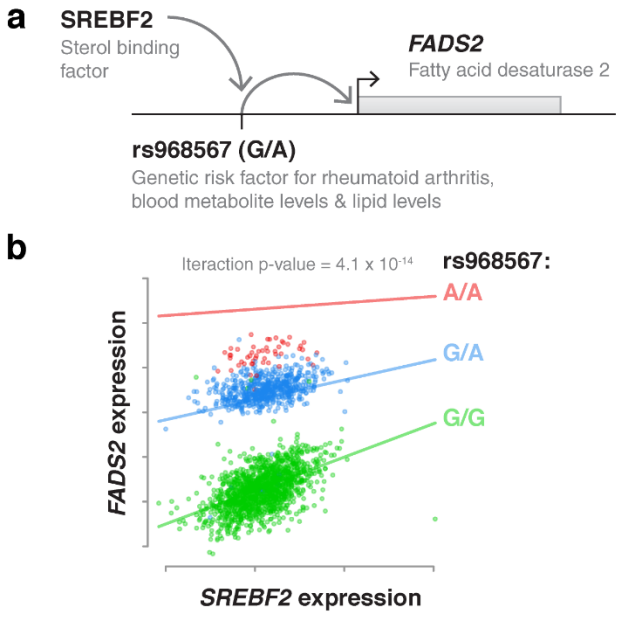
Zhernakova et al. NG 2016

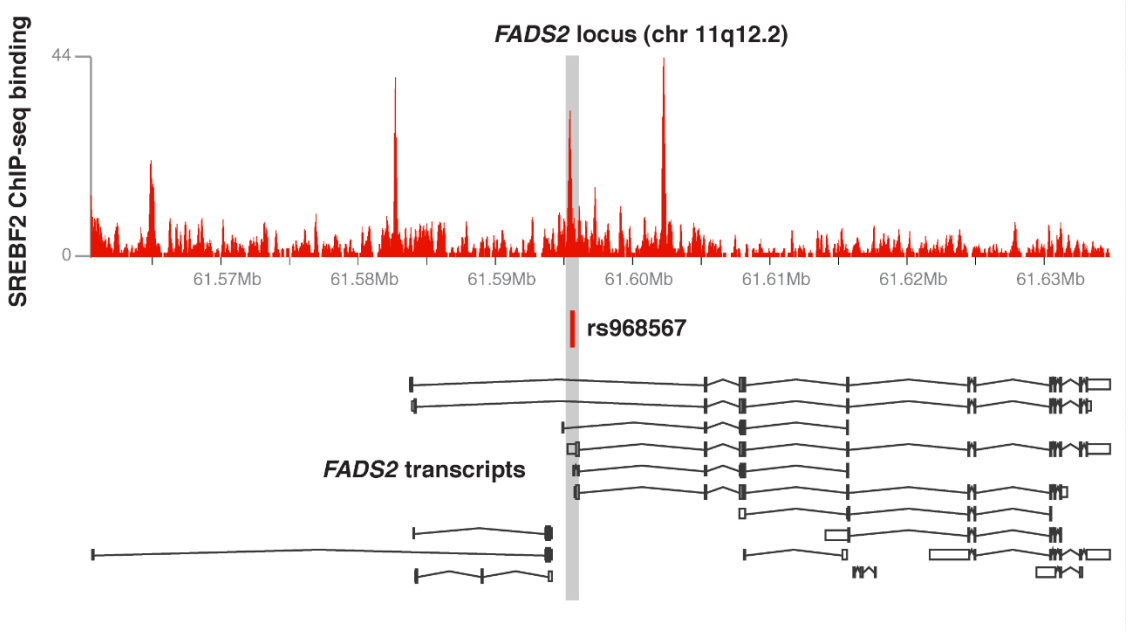
Many different cell-types show specific eQTL effects

Module number	Module description	Number of affected eQTLs	# eQTLs in strong LD with known GWAS hits	GO biological process top enriched pathway
1	Neutrophils 1	917	75	Detection of bacterium
2	CD4+ T-cells	337	25	T cell selection
3	NK cells / CD8+ T-cells	226	19	Cellular defense response
4	Erytrocytes	188	8	Hemoglobin metabolic process
5	Monocytes / Macrophages	181	11	Defense response to virus
6	Growth factor	156	10	Nerve growth factor receptor signaling pathway
7	Type 1 interferon	145	11	Regulation of defense response
8	Neutrophils 2	121	3	Detection of bacterium
9	B-cells	123	11	B cell receptor signaling pathway
10	Eosinophil	120	7	Regulation of myeloid leukocyte mediated immunity

Zhernakova et al. NG 2016

Detection of regulatory networks





Zhernakova et al. NG 2016

Conclusions

- GWAS allows identification of genomic regions that harbor a genetic risk factor for complex traits and diseases
  - They do not directly point to function
- eQTLs can be used to link genes to disease
- Combing eQTL with environment allows deeper insight into systems