umcg

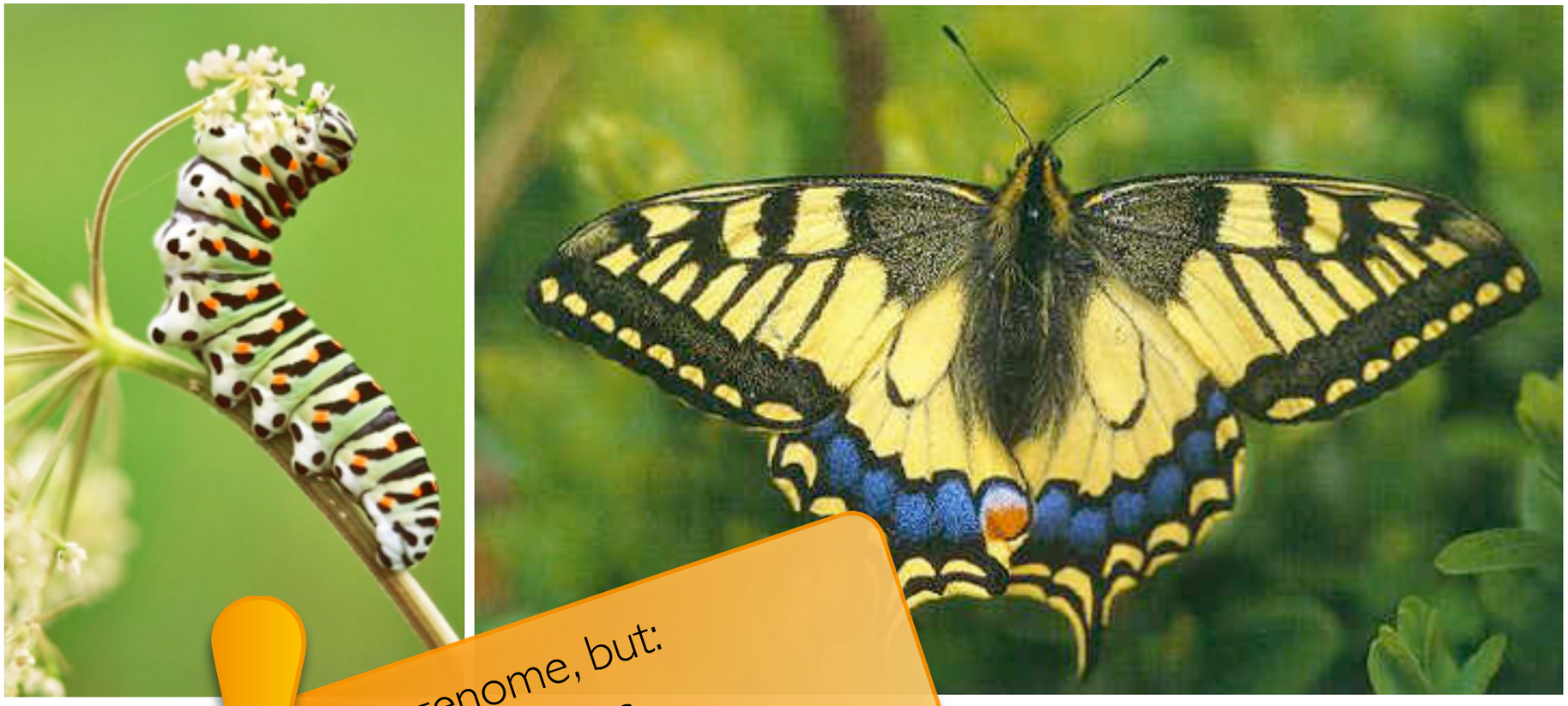
Department of genetics

Lude Franke > Gene expression analysis using microarrays and RNA-seq

Department of Genetics, UMC Groningen

From caterpillar to butterfly

The dramatic consequences of gene regulation in biology:

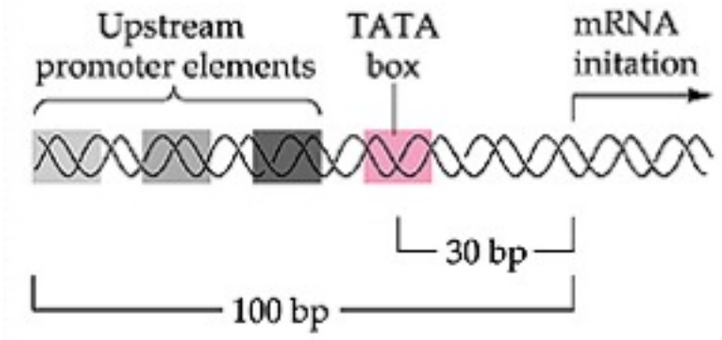


Koninginnenpage (Papilio machaon)

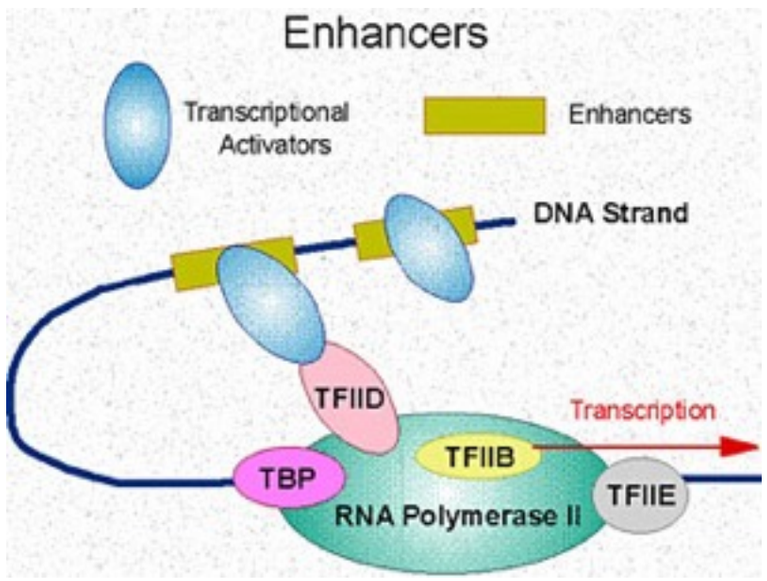
Same genome, but:
Different tissues
Different physiology
Different expression pattern

Eukaryotic gene expression regulation

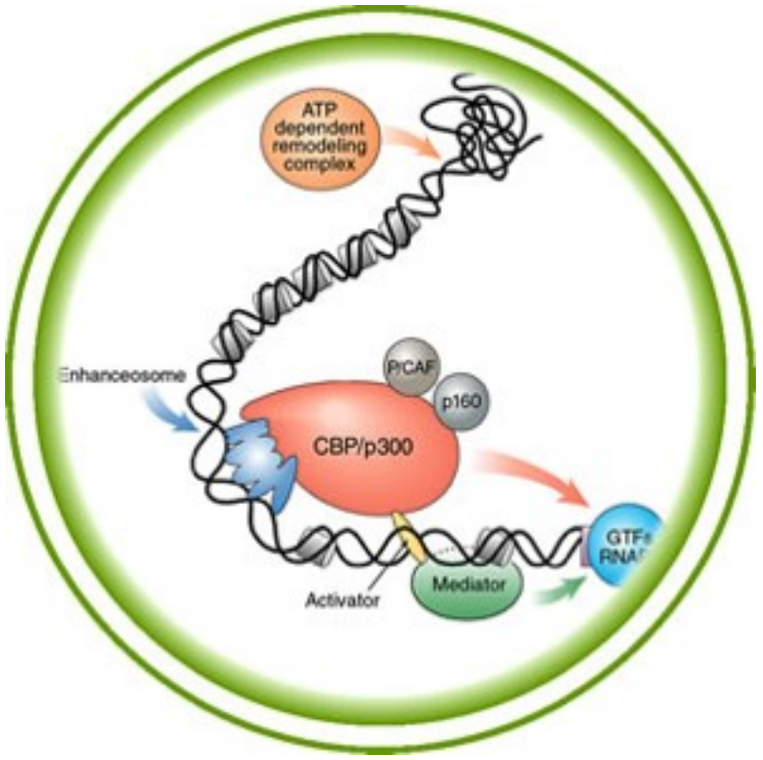
1



2



3

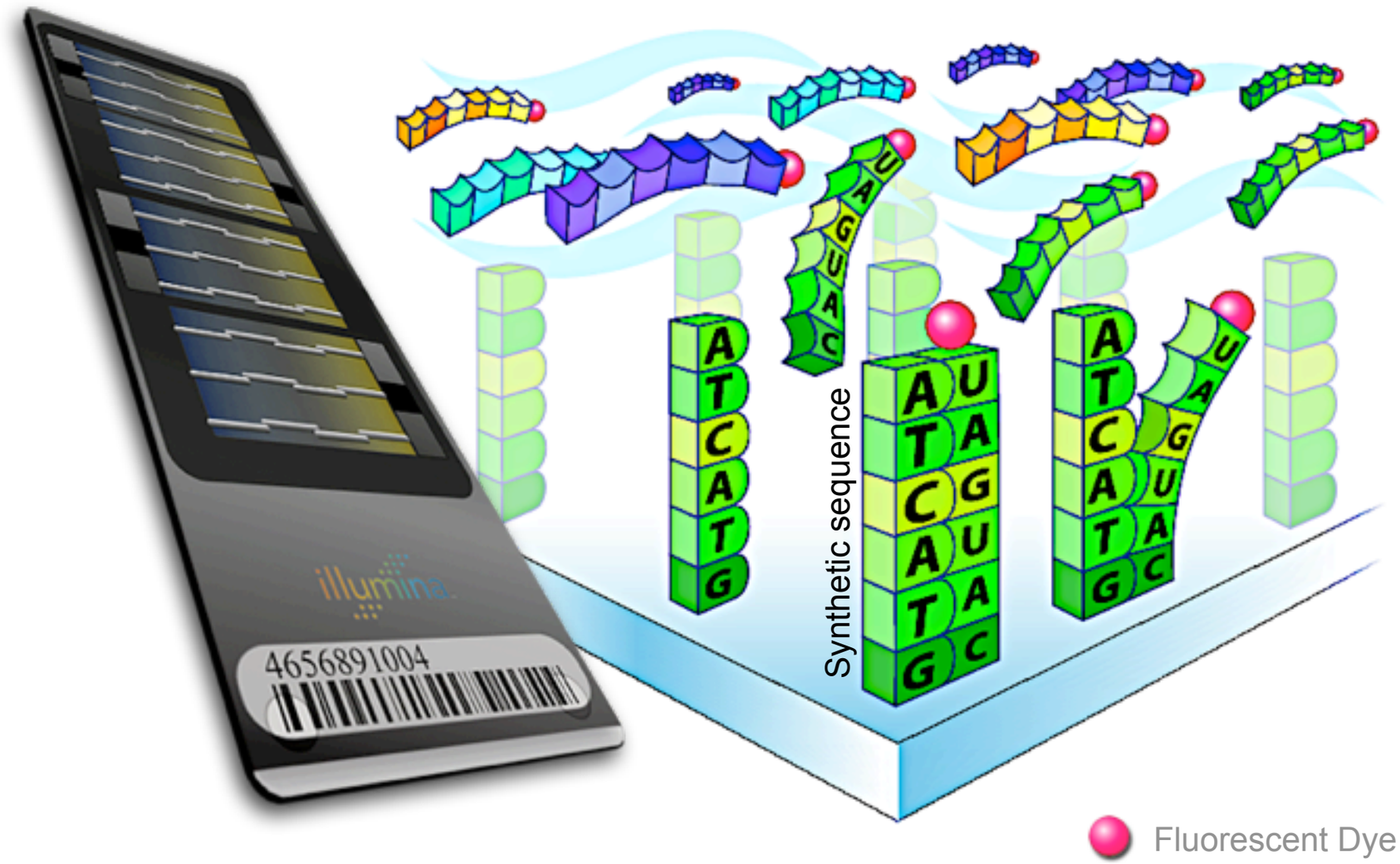


Gene expression determines a lot

Gene expression determines

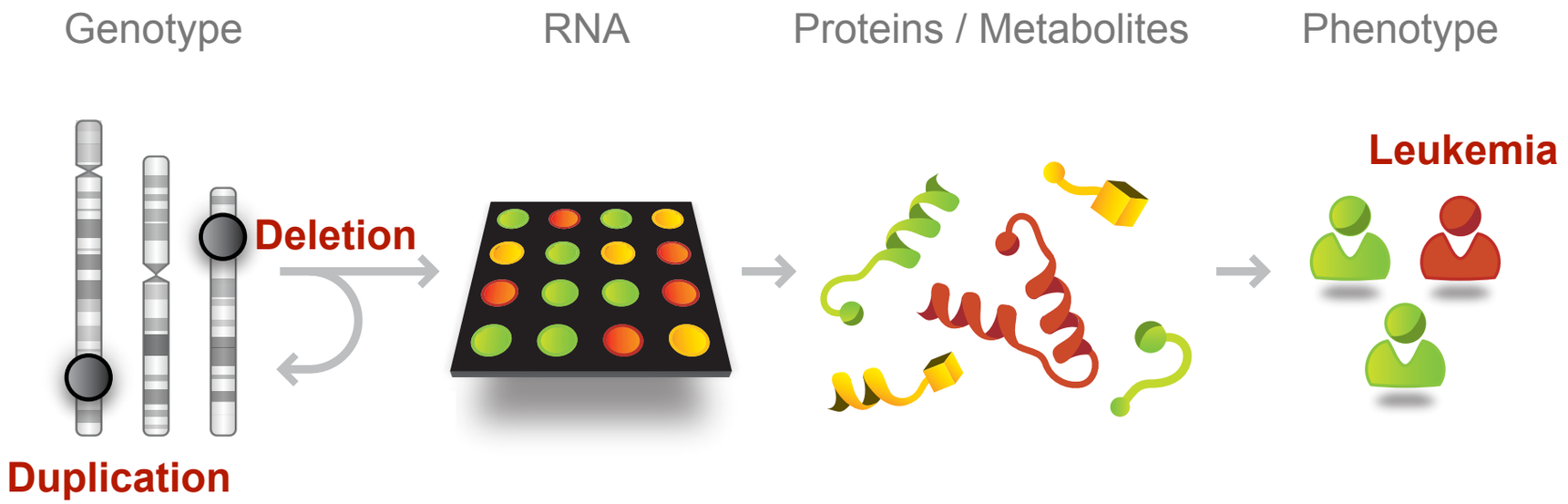
- Physiological status (nutrition, environment)
- Sex and age
- Various tissues and cell types
- Response to stimuli (drugs, signals, toxins)
- Health and disease
  - underlying pathogenic diversity
  - progression and response to treatment
  - patient classes of varying prospects

Oligonucleotide array (Microarray or GeneChip)



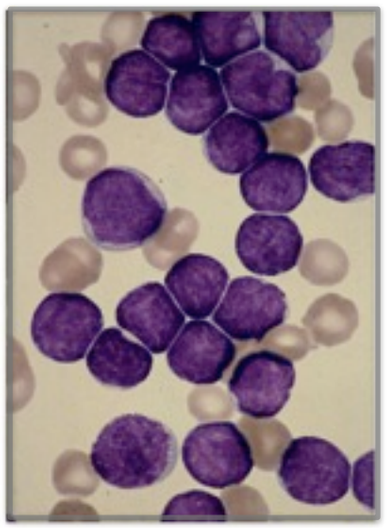
Leukemia

Cancer of the blood or bone marrow
Characterized by an abnormal proliferation (production by multiplication) of blood cells, usually white blood cells (leukocytes).



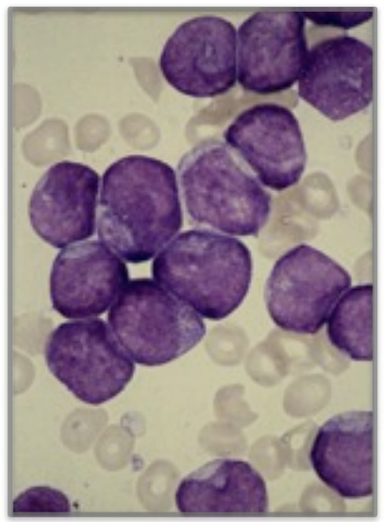
Leukemia

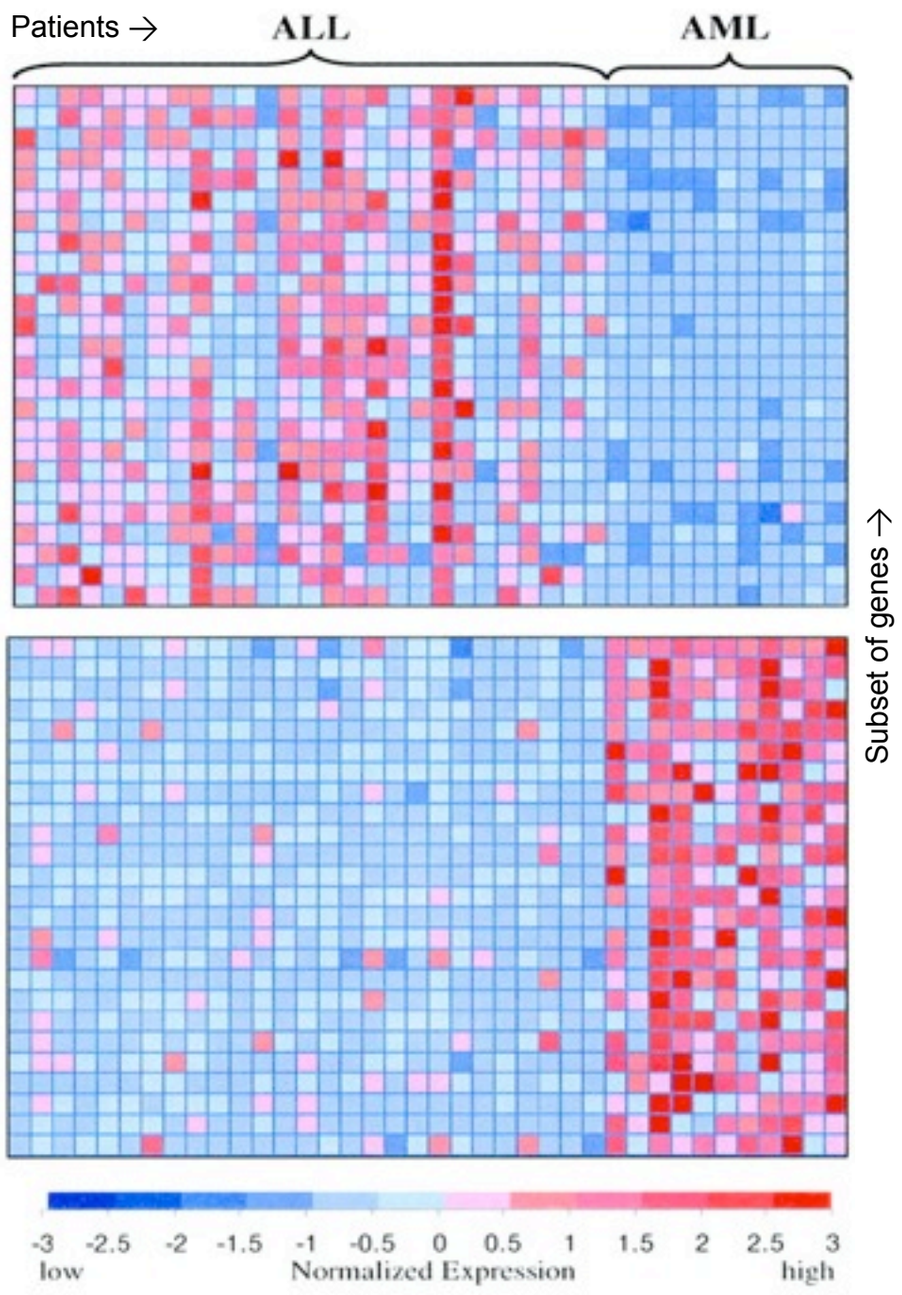
Acute lymphoblastic leukemia (ALL)



versus

Acute myeloid leukemia (AML)





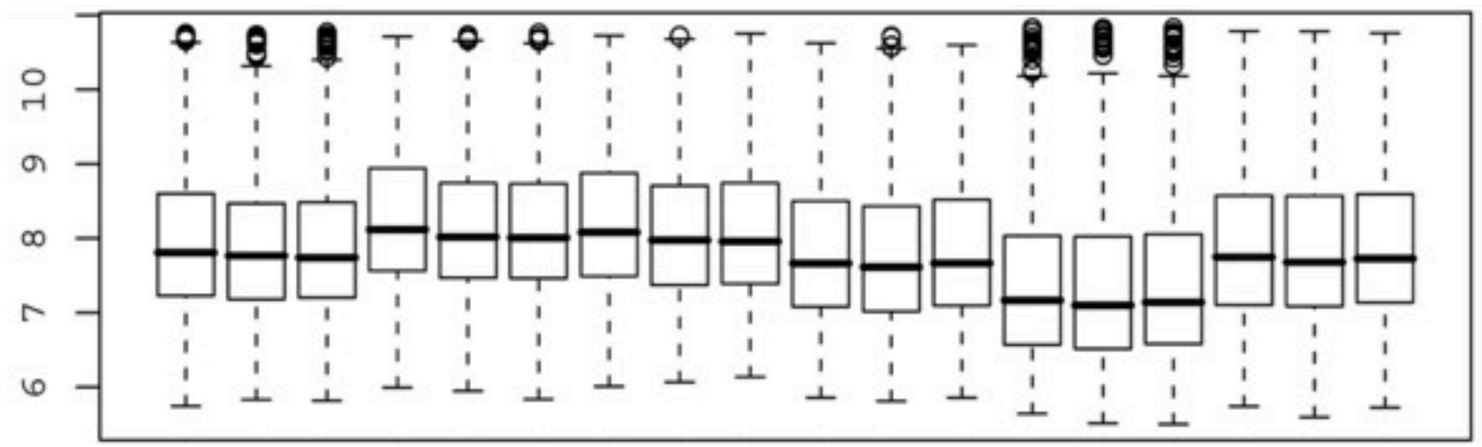
Why important:

- Crucial to distinguish between ALL & AML: Require different treatments
- Identification of diagnostic genes
- Insight in affected genes & pathways

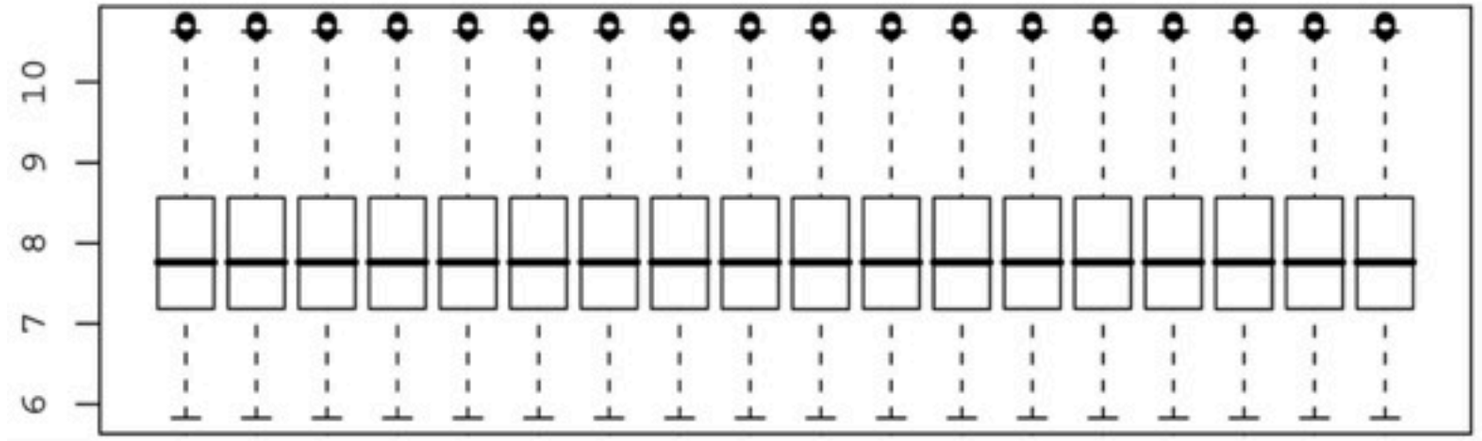
Quantile Normalization

Before normalization

Different samples



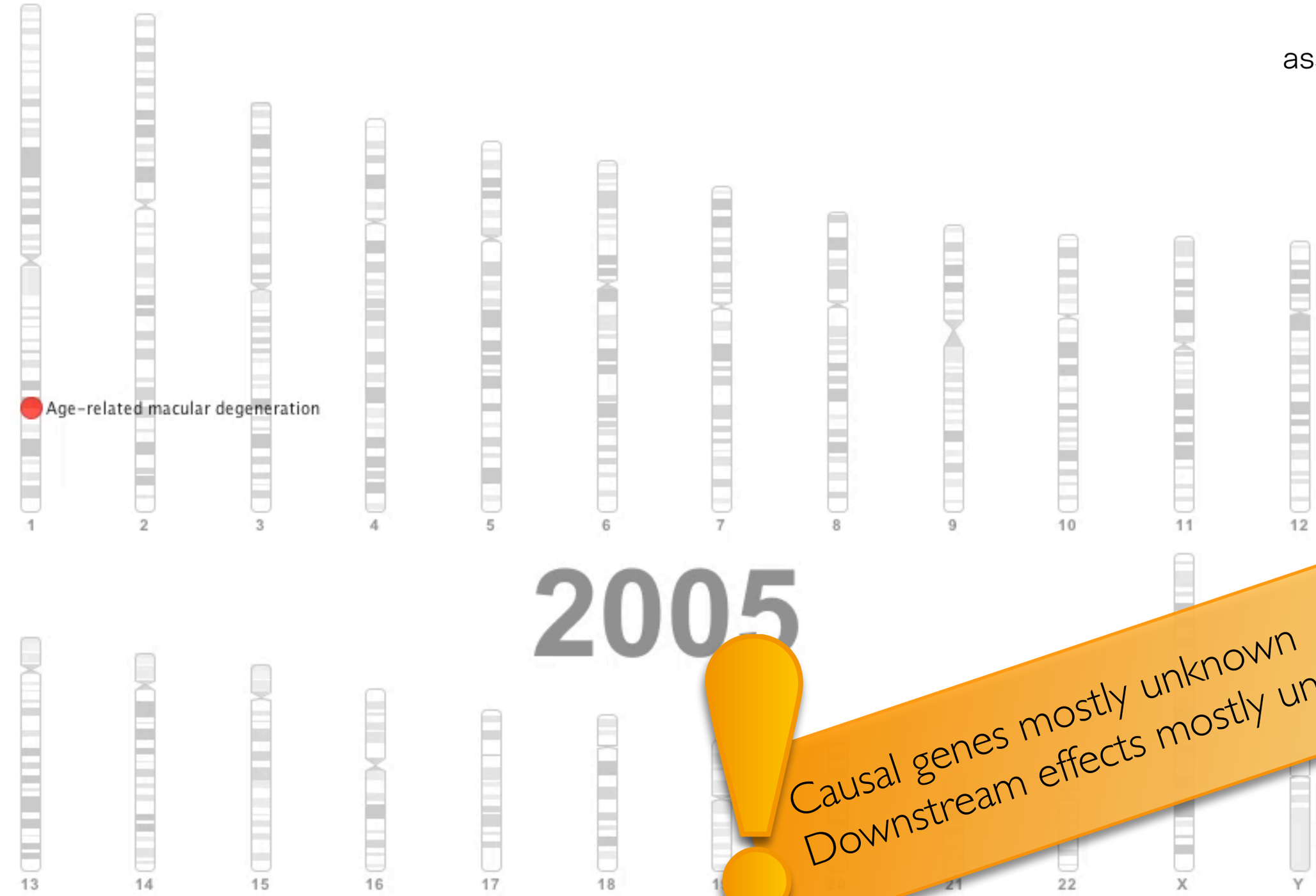
After normalization



Seven years of GWAS studies

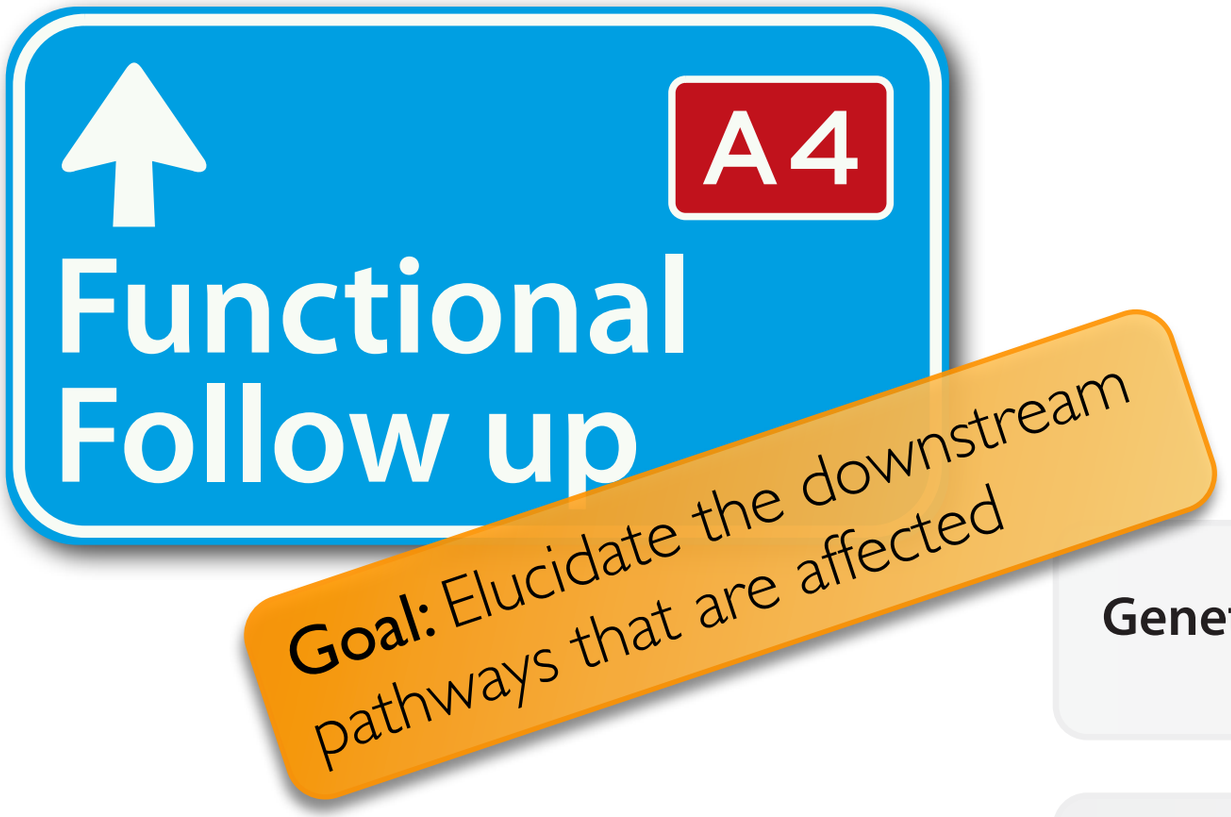
Gene atlas

6,054
disease
associations

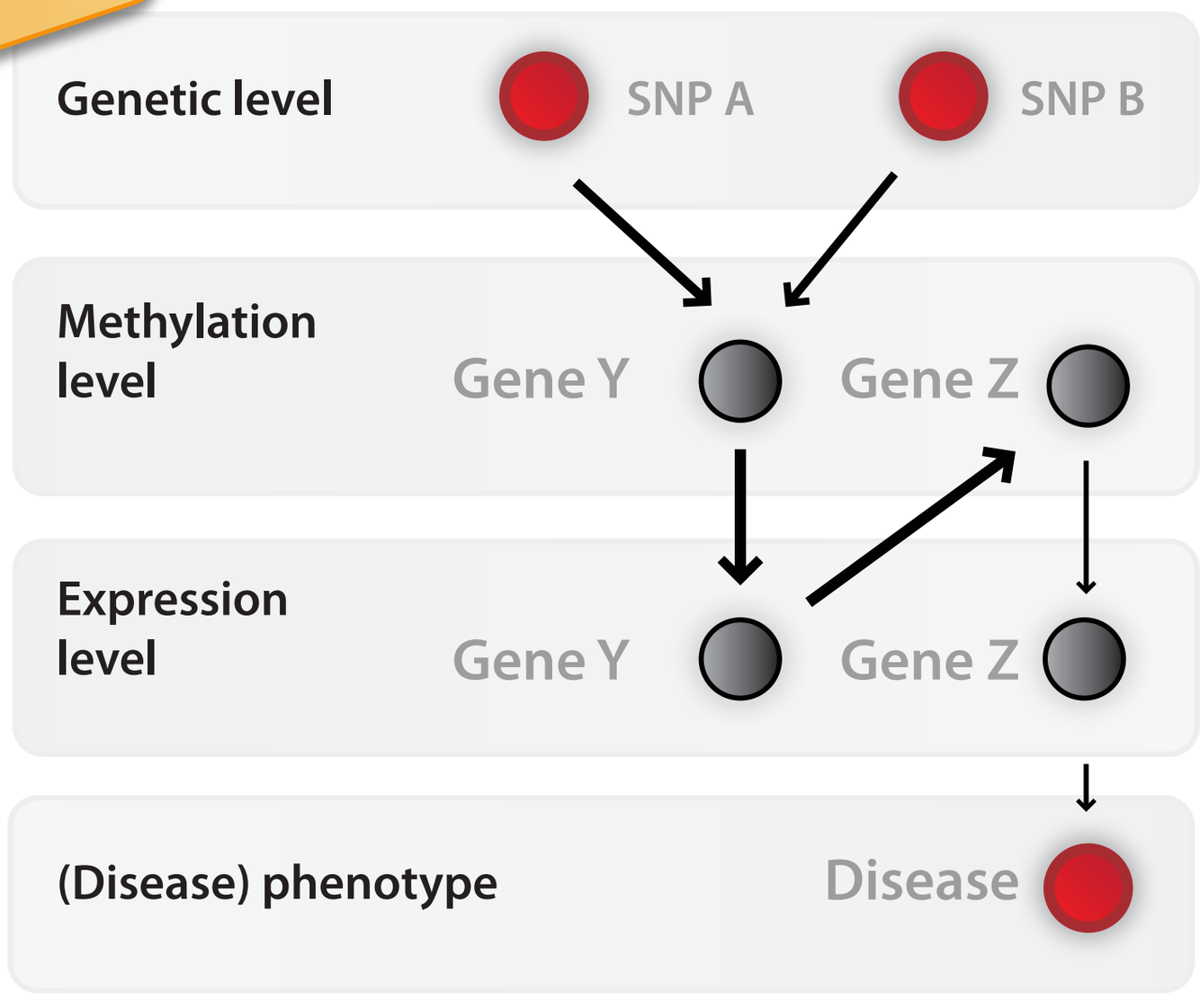


Teri Manolio *et al*: A catalog of published genome-wide association studies

Functional follow-up



Goal: Elucidate the downstream pathways that are affected



Functional follow-up

Common strategy:

Knock-down, knock-out, overexpression assay in:
- Cell line
- Model organism

Advantages:
- One single perturbation
- Different perturbations possible
- Many potential read-outs
- Cost-effective

Disadvantages:
- Effects *in vitro* the same as *in vivo*?
- Effect of knock-out the same of SNP variant?
- Effects in model organism comparable to humans?
- Possible to observe measurable effect for SNPs with a very small effect size (e.g. 1.1)?

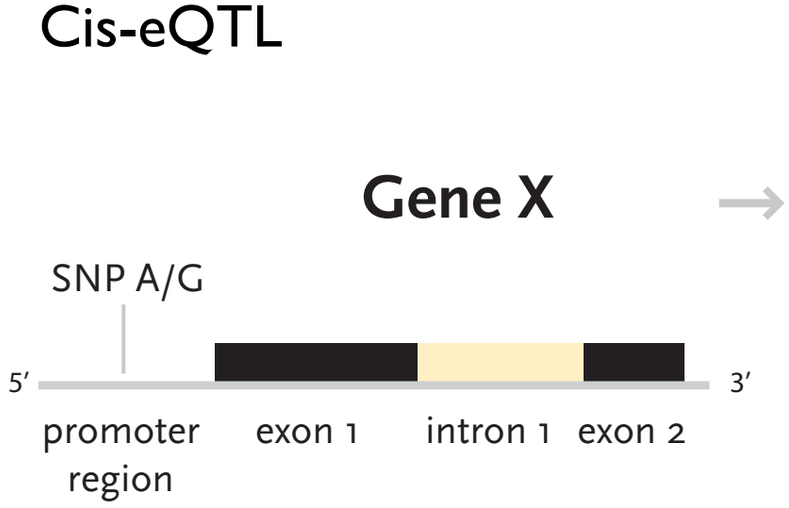
Don't we need many samples?

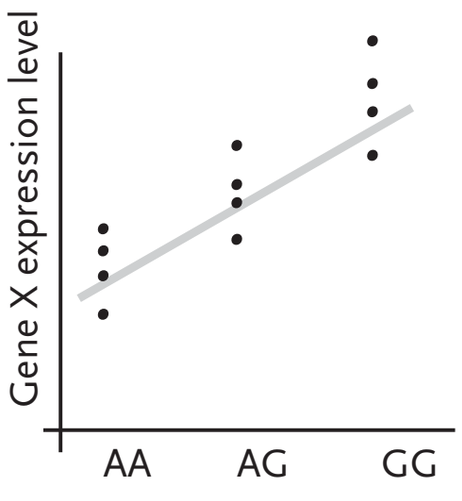
Functional follow-up: genetical genomics

Genetical genomics: High-throughput systematic approach to gain insight in the effects of genetic variants on gene, protein and metabolite expression.

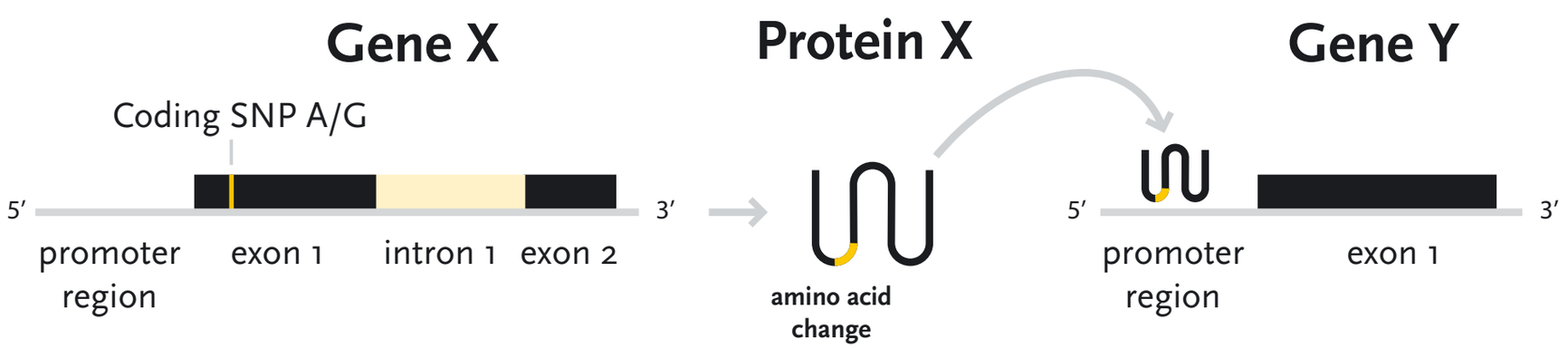
Genetical genomics: What is an eQTL?

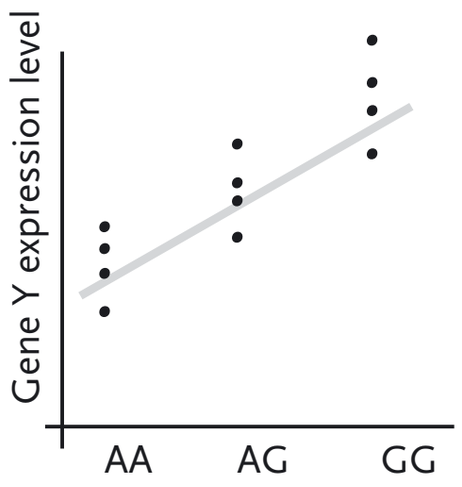
Cis-eQTL





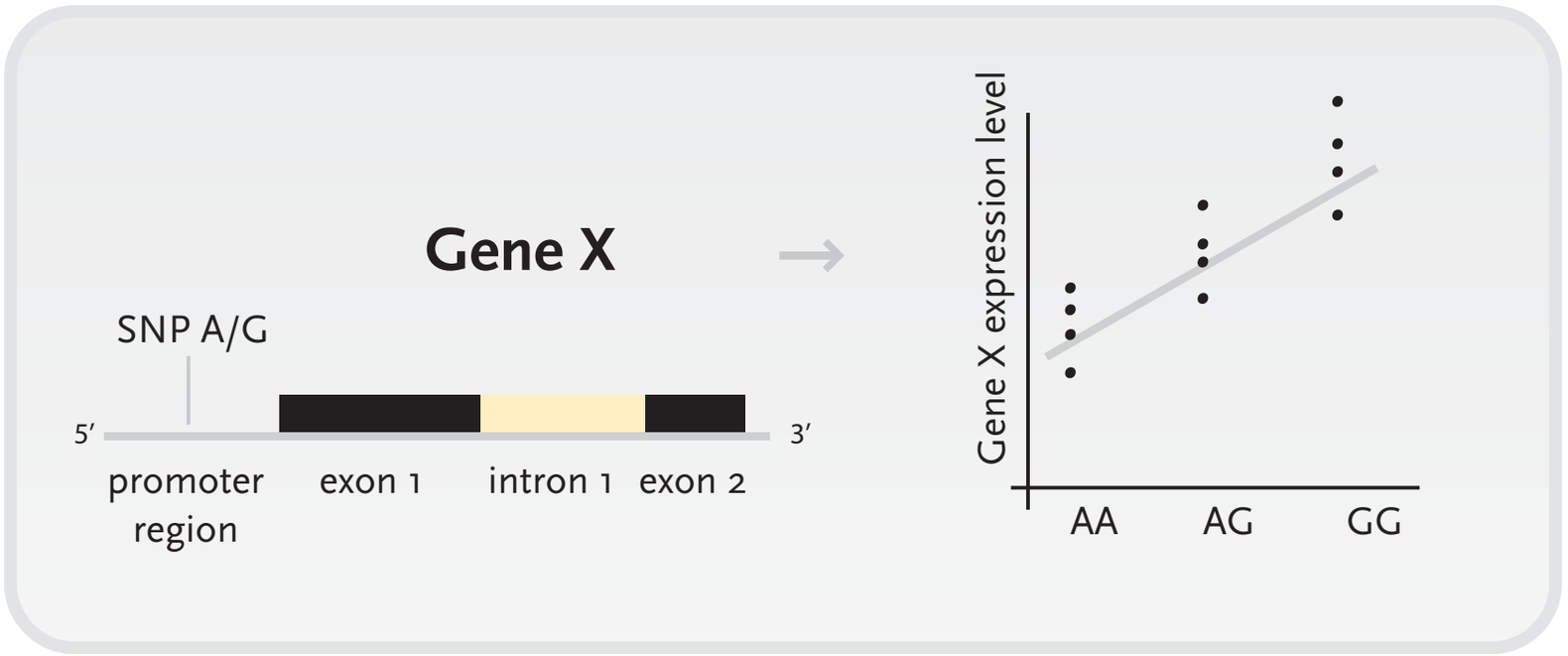
Trans-eQTL





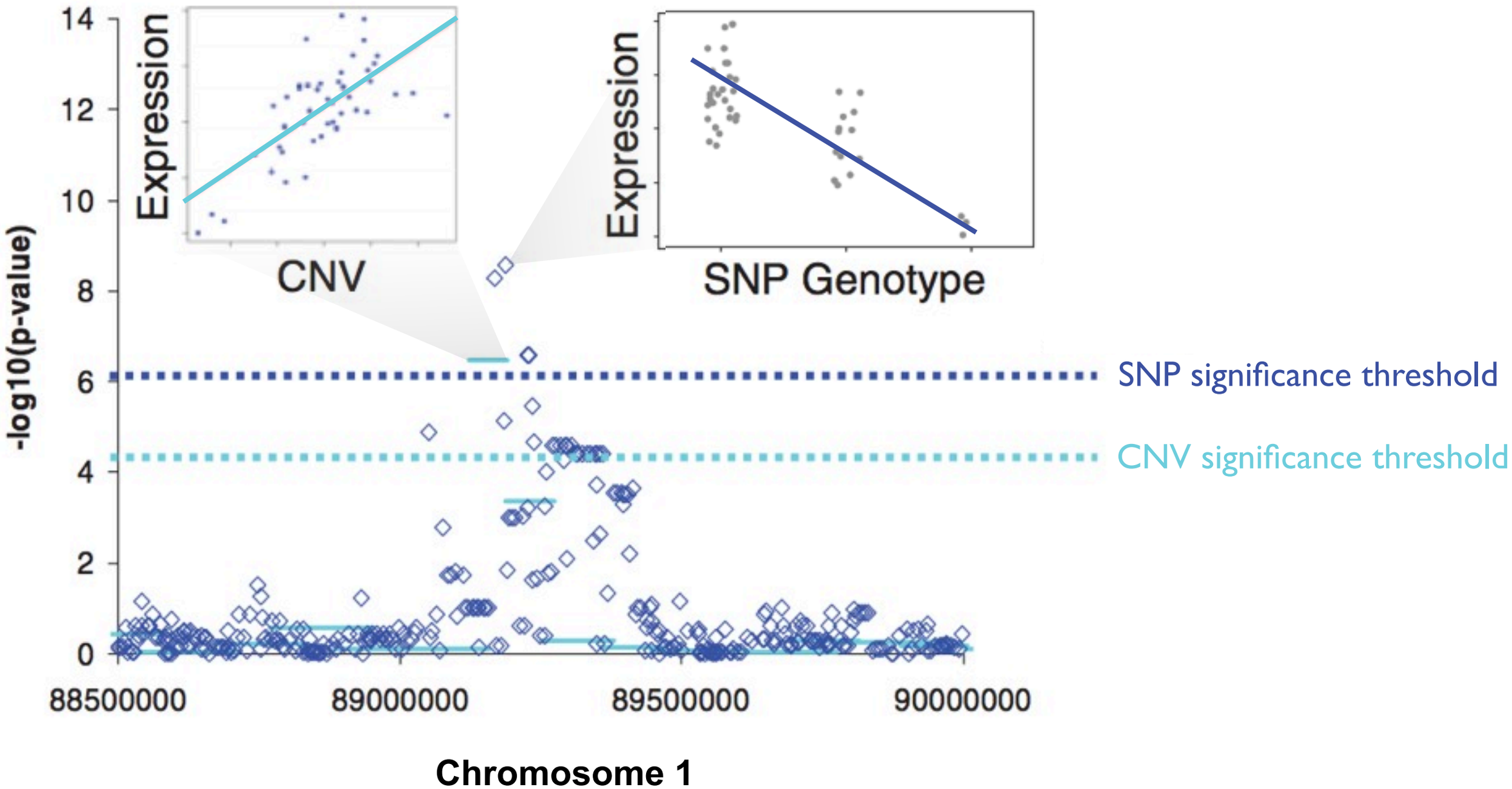
cis-eQTLs

A few *cis*-eQTL examples:



cis-eQTLs: CNVs and SNPs

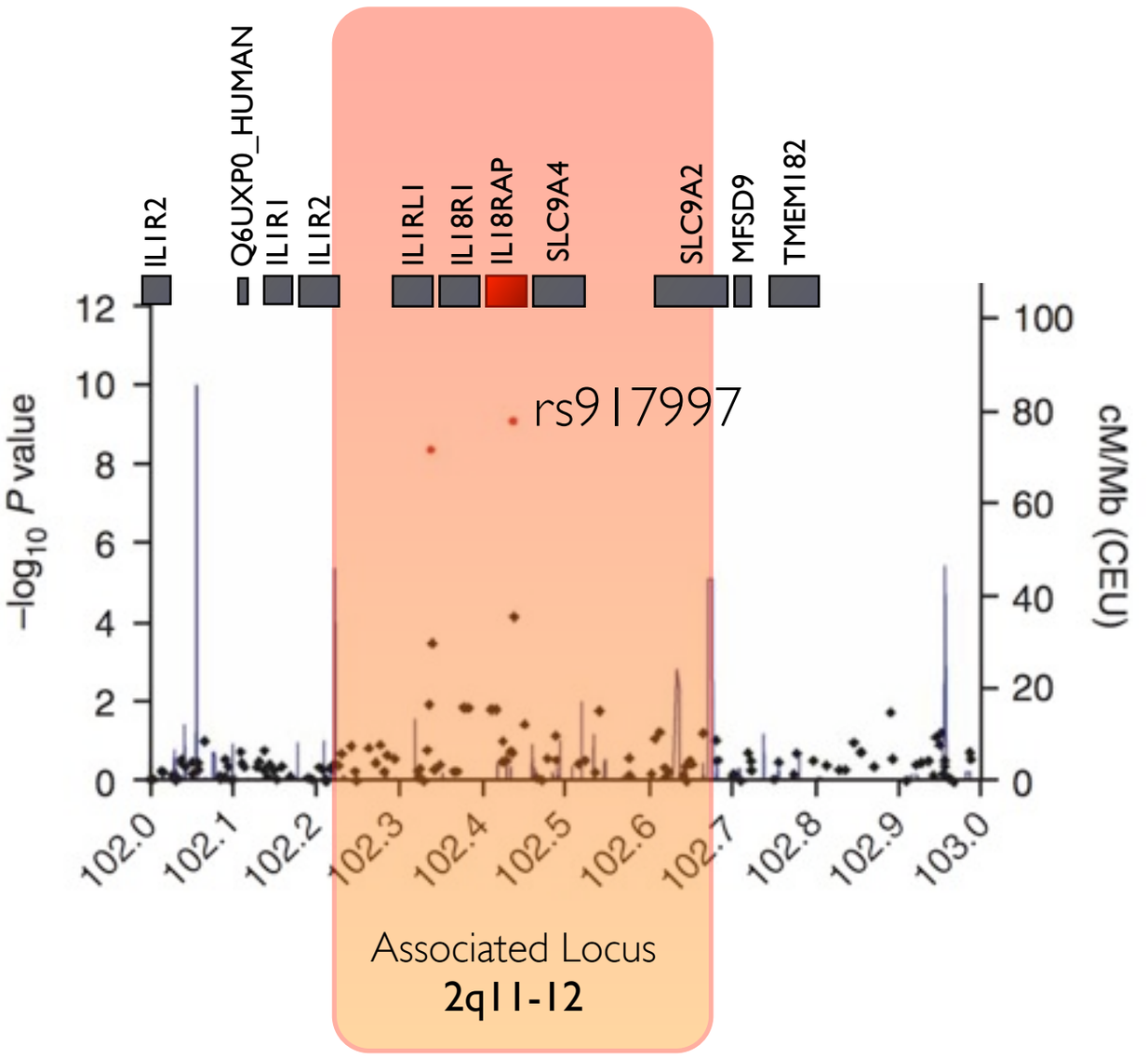
Effects of CNVs on *cis*-gene expression (GBP3, 45 JPT samples)

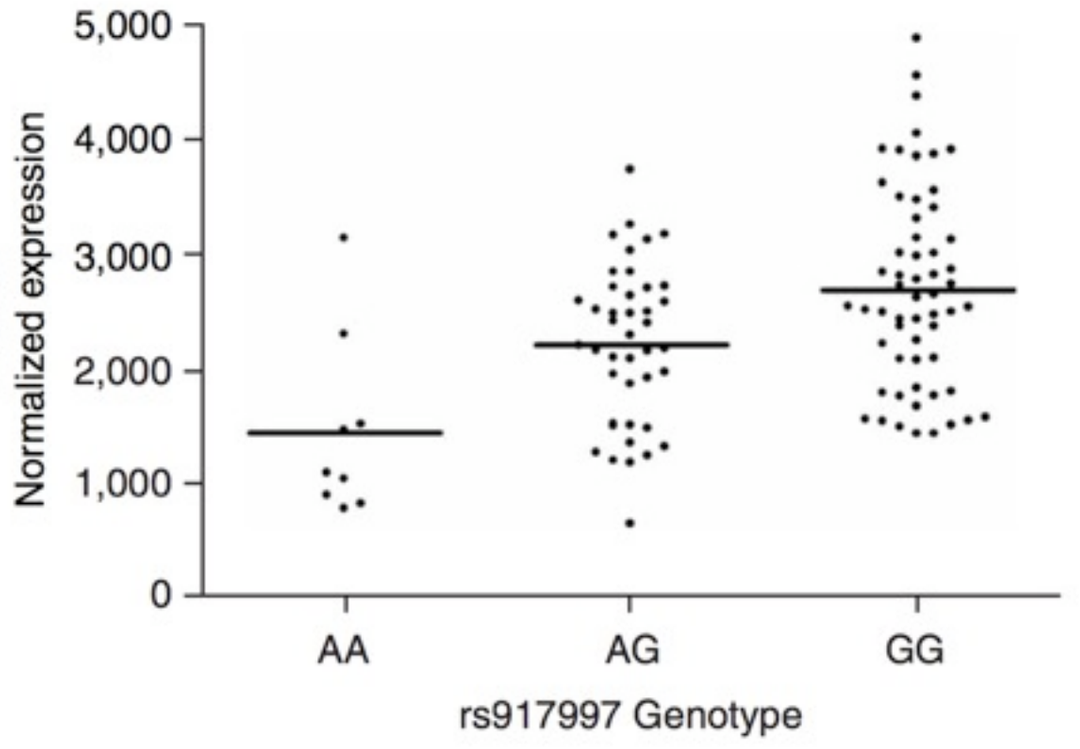


Stranger *et al*, Science 2007

cis-eQTLs: celiac disease

Hunt *et al*, 2008, Nature Genetics: 9 associated loci (*cis*-eQTL analysis in 119 samples)



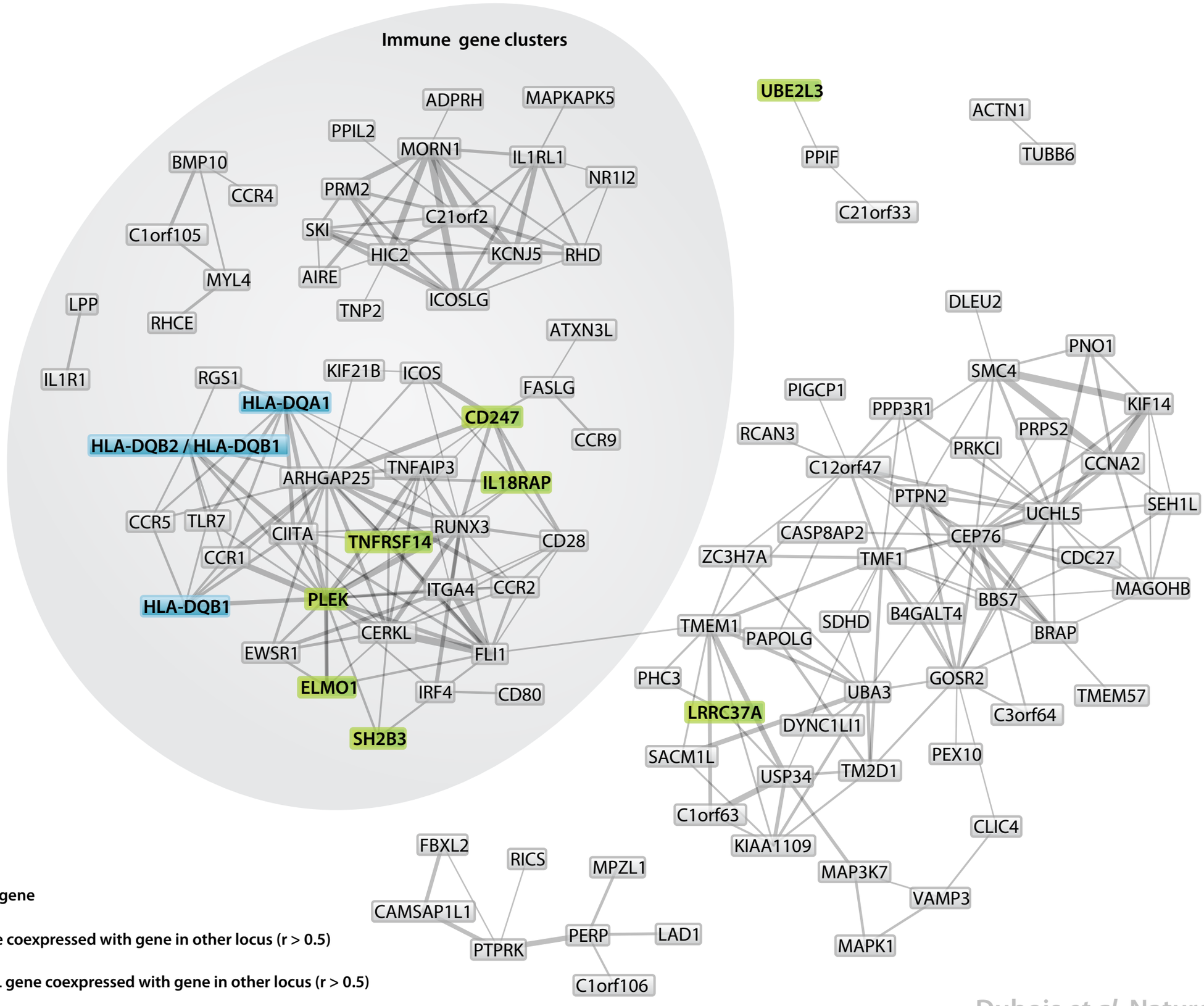


Dubois *et al*, 2010, Nature Genetics: 40 associated loci

cis-eQTL study conducted in peripheral blood (1,469 unrelated individuals):

50% of loci affect gene expression in *cis*.

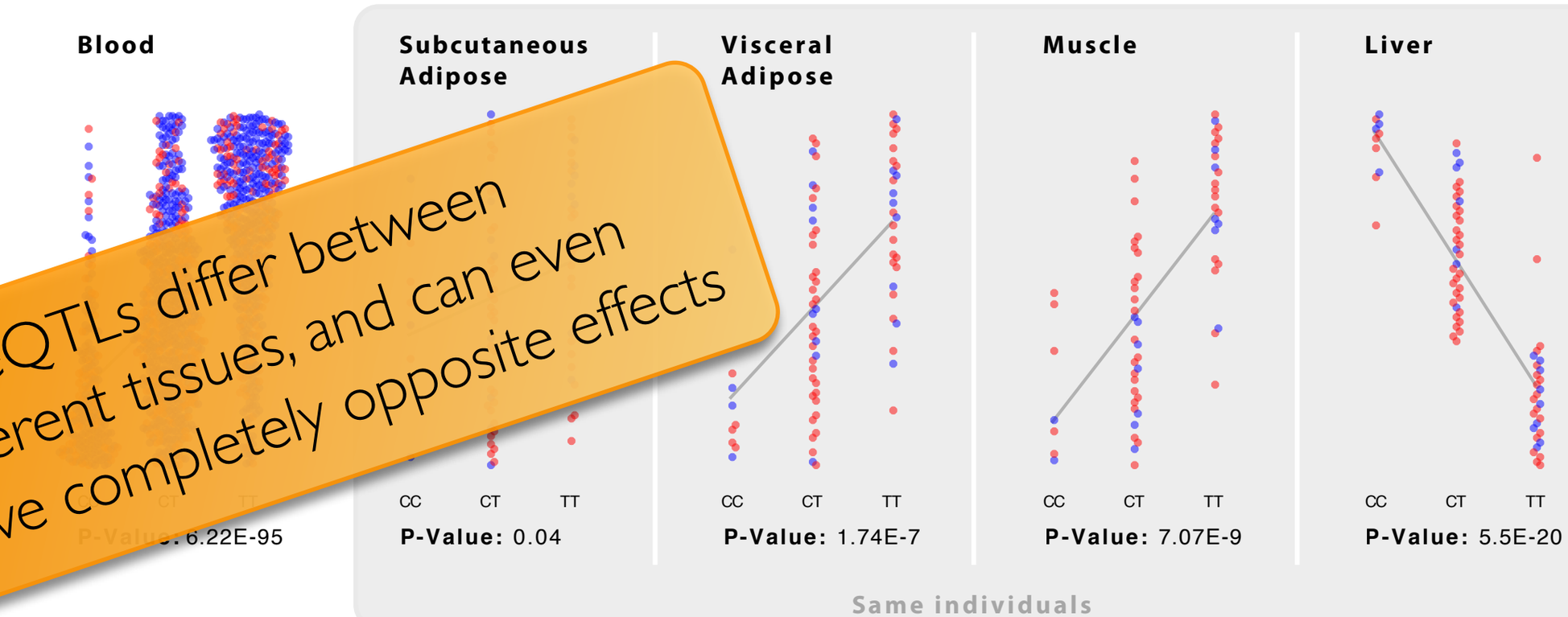
cis-eQTLs: celiac disease



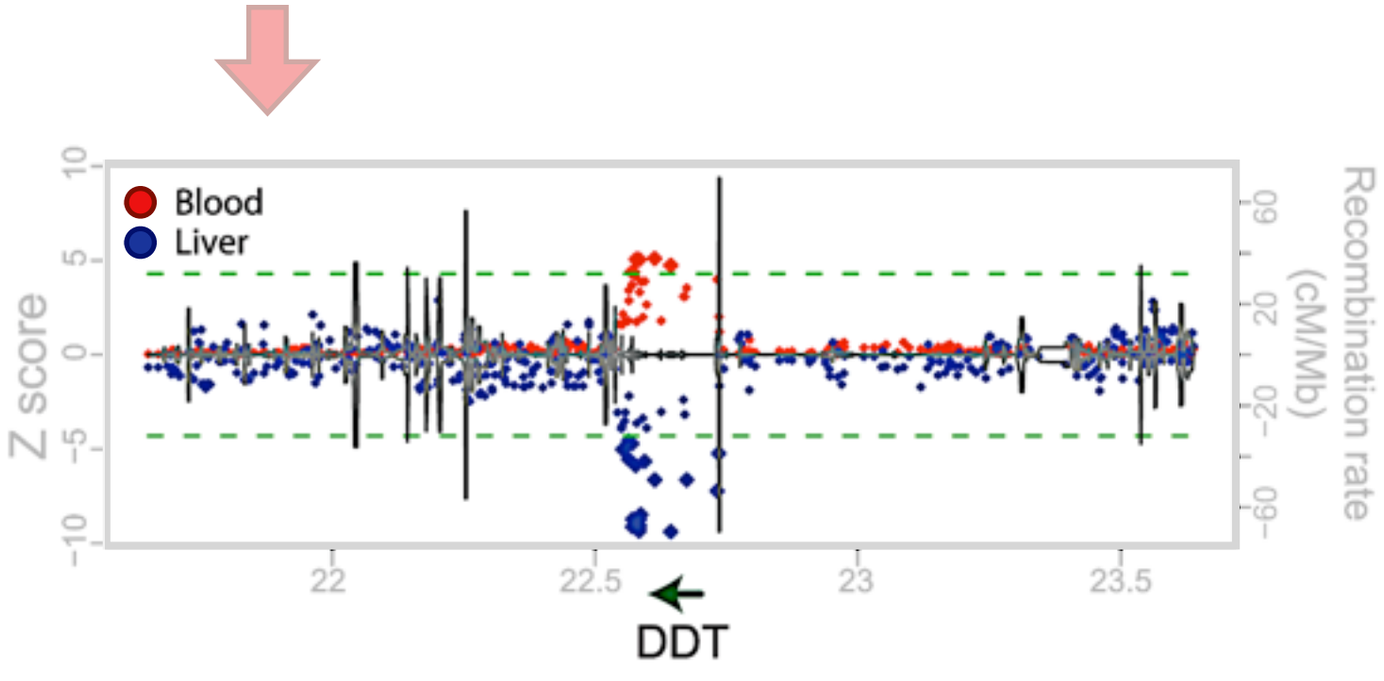
Dubois *et al*, Nature Genetics 2010

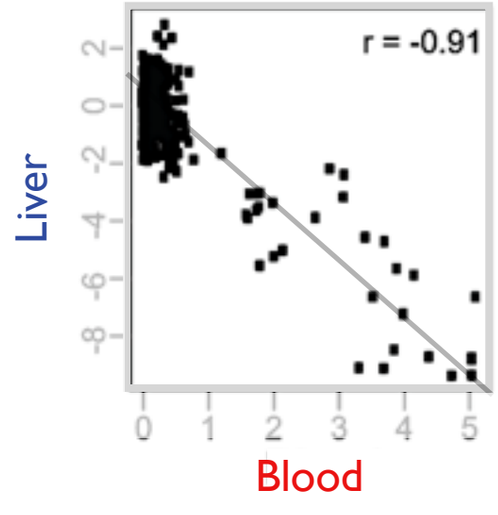
cis-eQTLs: Effects can differ between tissues

SNP rs2186366 (chr. 22, 22584113 bp) affects DDT



Cis-eQTLs differ between different tissues, and can even have completely opposite effects





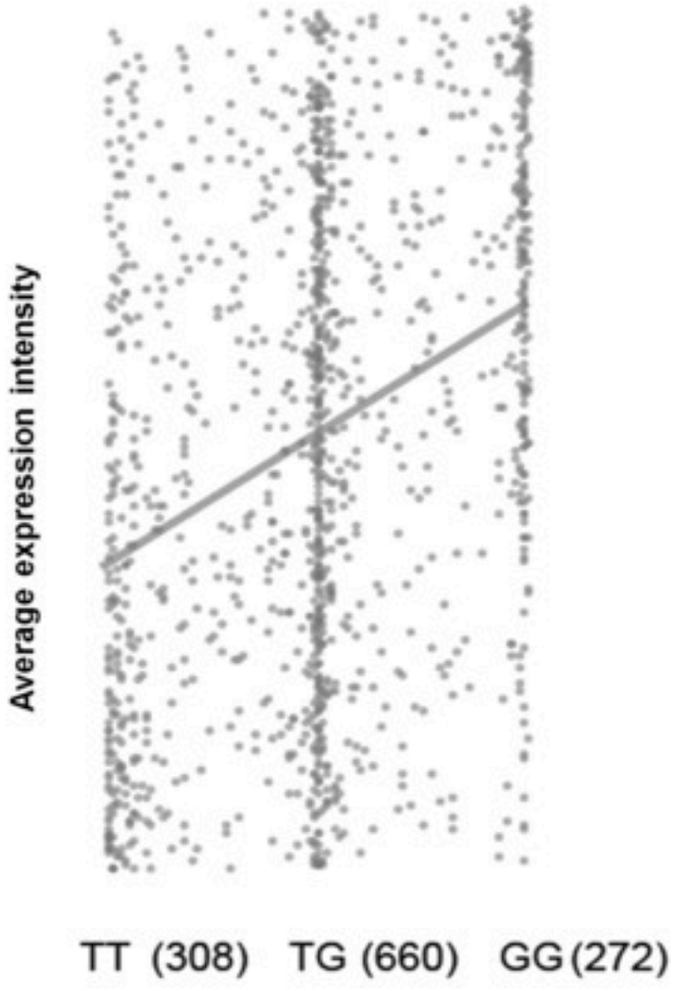
Fehrmann et al, PLoS Genetics 2011
Fu *et al*, PLoS Genetics 2012

Genetic variants can affect non-coding genes (lincRNAs)

Effect of age related macula degeneration SNP rs13278062

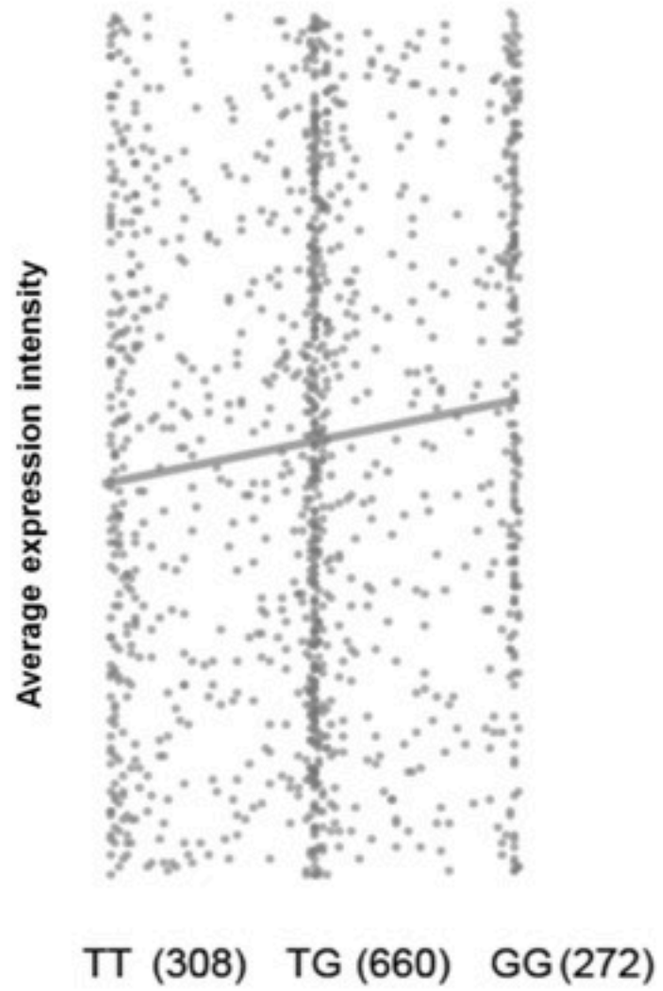
LincRNA LOC389641

$P = 4.31 \times 10^{-32}$



Protein coding TNFRSF10A

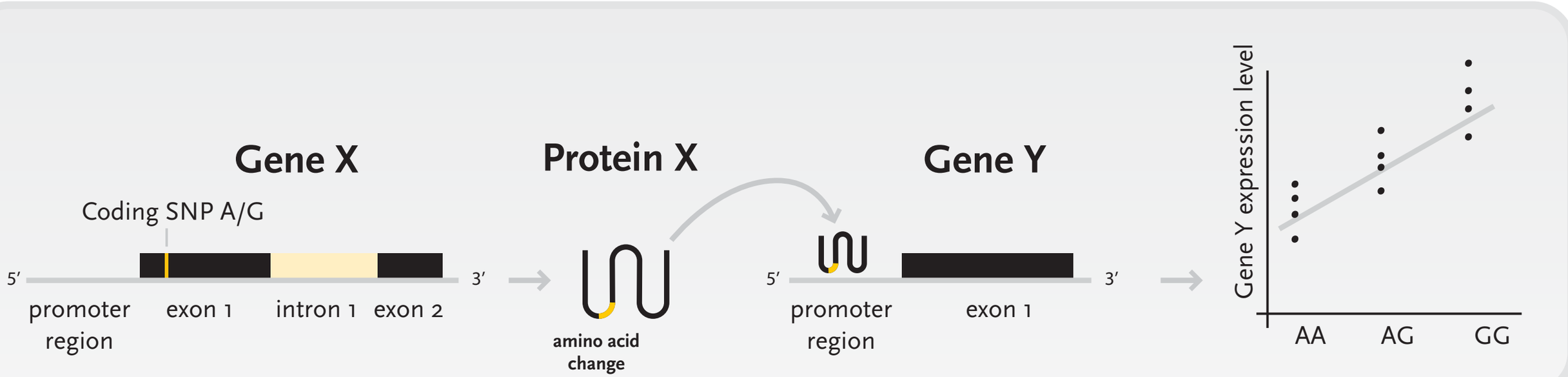
$P = 4.21 \times 10^{-4}$



Kumar *et al*, PLoS Genetics 2013

trans-eQTLs

A few *trans*-eQTL examples



trans-eQTLs: T1D SNP affects anti-viral gene network

nature | September 2010

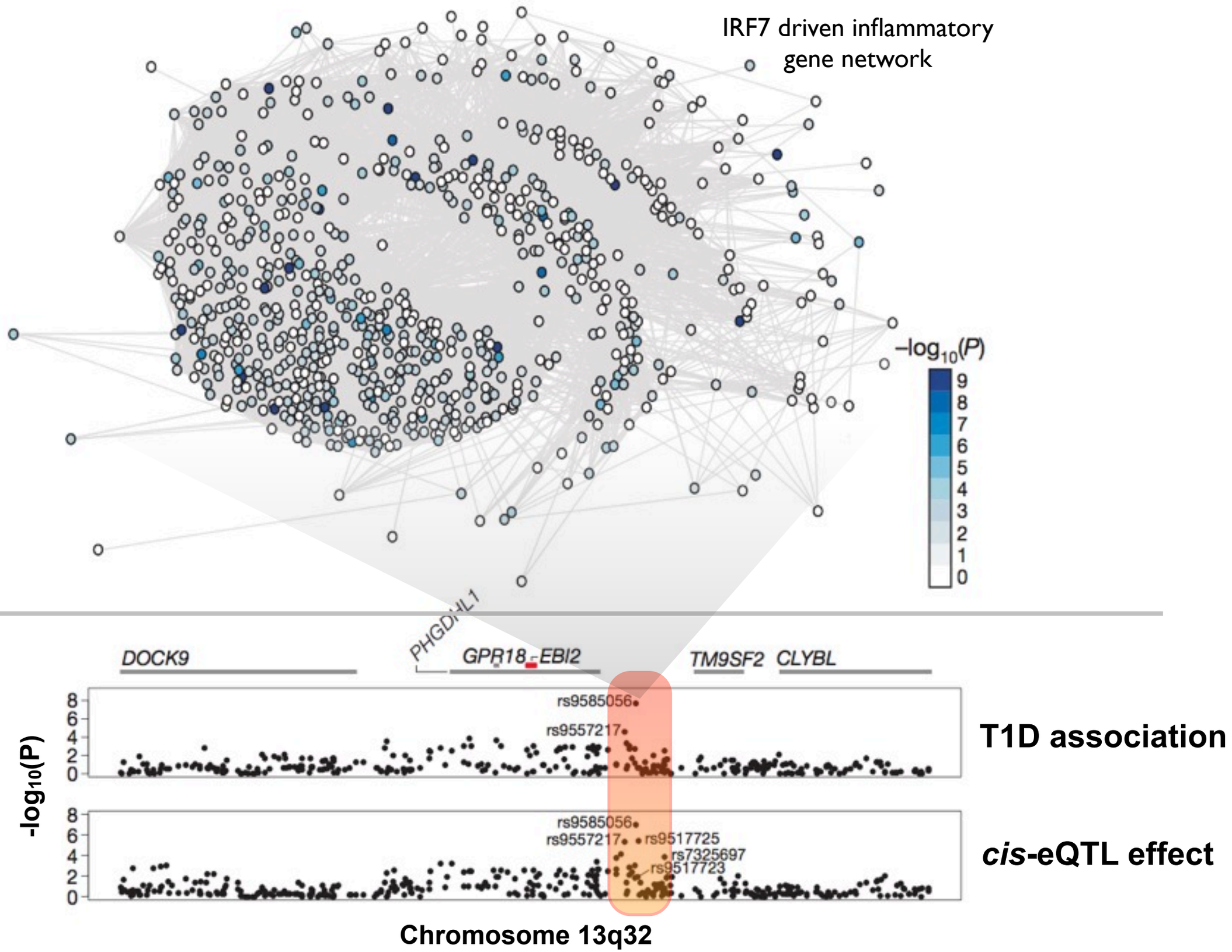
A *trans*-acting locus regulates an anti-viral expression network and type 1 diabetes risk

Matthias Heinig[1,2]*, Enrico Petretto[3,4]*, Chris Wallace[5], Leonardo Bottolo[3,4], Maxime Rotival[6], Han Lu[3], Yoyo Li[3], Rizwan Sarwar[3], Sarah R. Langley[3], Anja Bauerfeind[1], Oliver Hummel[1], Young-Ae Lee[1,7], Svetlana Paskas[1], Carola Rintisch[1], Kathrin Saar[1], Jason Cooper[5], Rachel Buchan[3], Elizabeth E. Gray[8], Jason G. Cyster[8], Cardiogenics Consortium†, Jeanette Erdmann[9], Christian Hengstenberg[10], Seraya Maouche[6], Willem H. Ouwehand[11,12], Catherine M. Rice[12], Nilesh J. Samani[13], Heribert Schunkert[9], Alison H. Goodall[13], Herbert Schulz[1], Helge G. Roider[2], Martin Vingron[2], Stefan Blankenberg[14], Thomas Münzel[14], Tanja Zeller[14], Silke Szymczak[15], Andreas Ziegler[15], Laurence Tiret[6], Deborah J. Smyth[5], Michal Pravenec[16], Timothy J. Aitman[3], Francois Cambien[6], David Clayton[5], John A. Todd[5], Norbert Hubner[1,17] & Stuart A. Cook[3,18]

rs9585056 near Epstein–Barr virus induced gene 2 (*EBI2*):

- Affects interferon regulatory factor 7 (IRF) driven inflammatory network
- Associated with type 1 diabetes ($P = 7 \times 10^{-10}$)

trans-eQTLs: T1D SNP affects anti-viral gene network



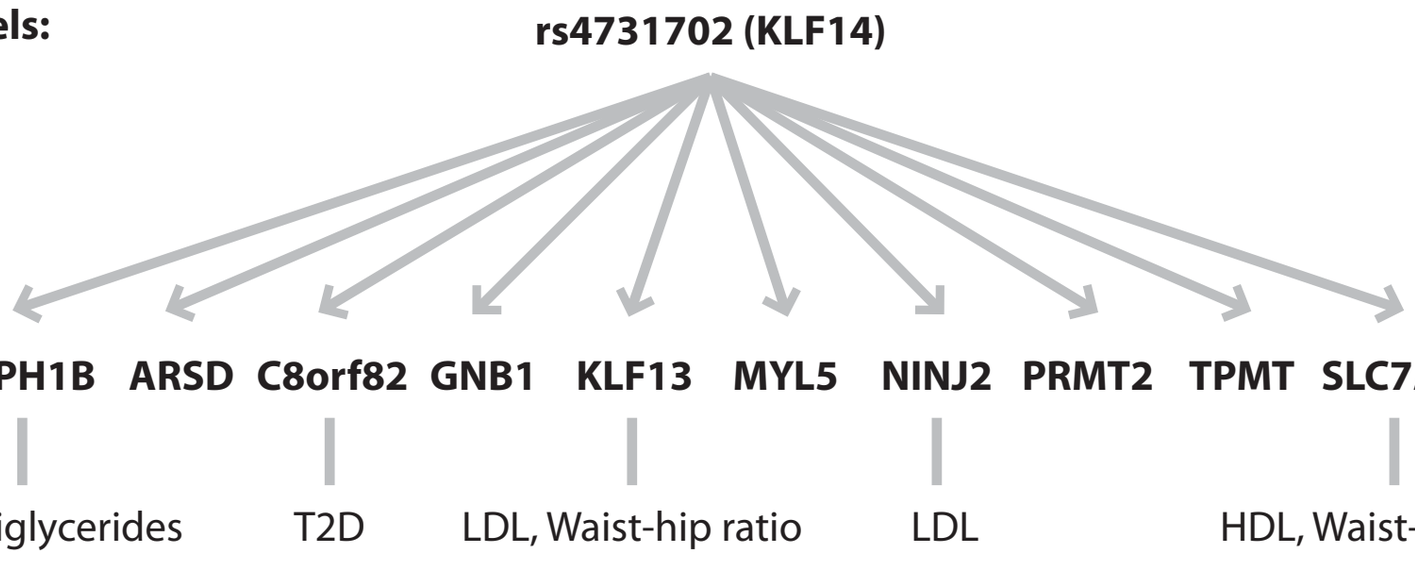
trans-eQTLs: HDL and T2D SNP: master regulator

nature genetics | Small *et al*, 2011

Identification of an imprinted master *trans* regulator at the *KLF14* locus related to multiple metabolic phenotypes

Kerrin S Small[1,2,10], Åsa K Hedman[3,10], Elin Grundberg[1,2,10], Alexandra C Nica[4], Gudmar Thorleifsson[5], Augustine Kong[5], Unnur Thorsteindottir[5,6], So-Youn Shin[2], Hannah B Richards[7], the GIANT Consortium[8], the MAGIC Investigators[8], the DIAGRAM Consortium[8], Nicole Soranzo[1,2], Kourosh R Ahmadi[1], Cecilia M Lindgren[3], Kari Stefansson[5,6,10], Emmanouil T Dermitzakis[4,10], Panos Deloukas[2,10], Timothy D Spector[1,10] & Mark I McCarthy[3,7,9,10] for the MuTHER Consortium[8]

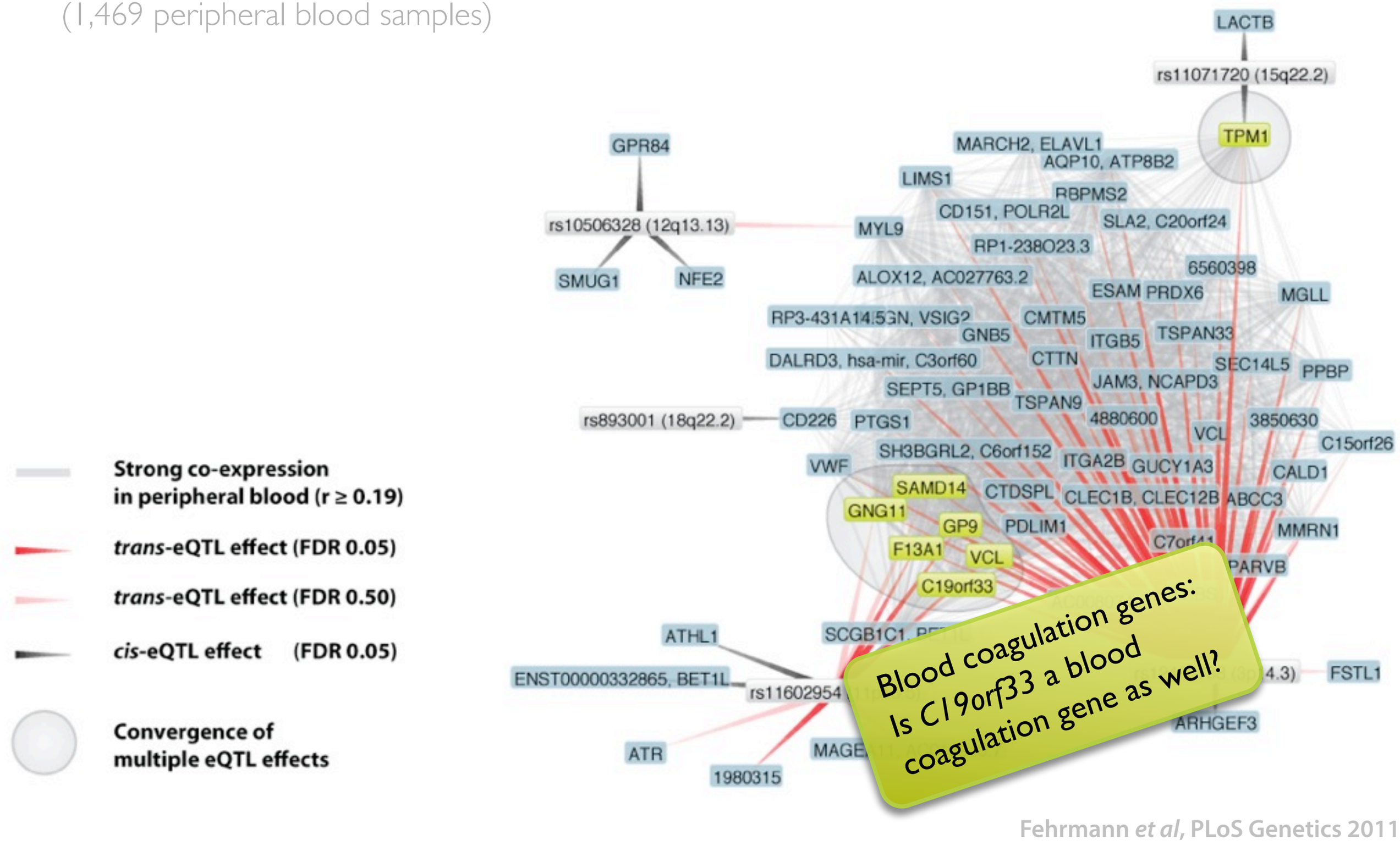
rs4731702 associated with T2D and HDL levels:



trans-eQTLs: Mean platelet volume & blood coagulation

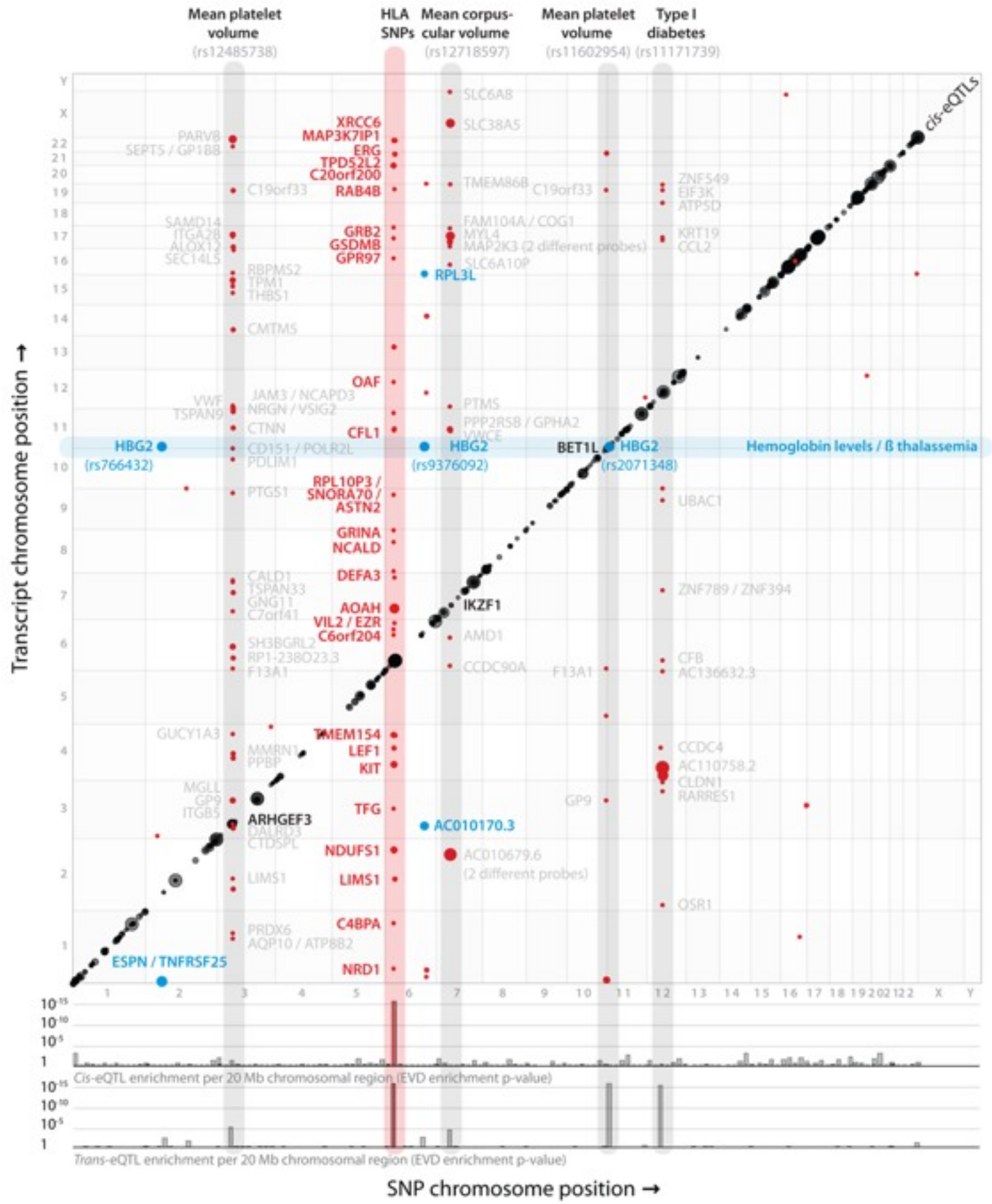
Trans effects of mean platelet volume SNPs

(1,469 peripheral blood samples)



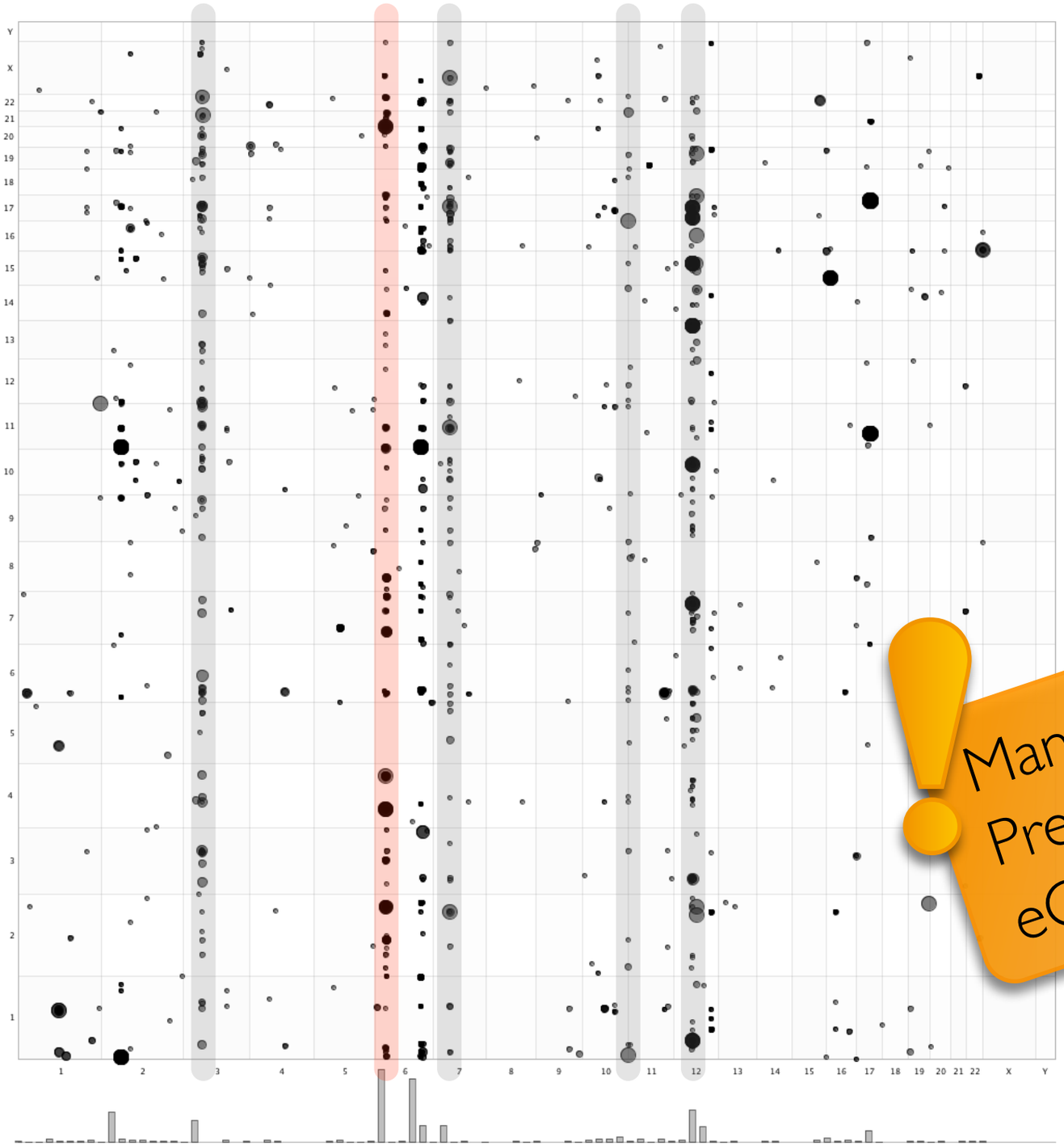
Fehrmann *et al*, PLoS Genetics 2011

Many *trans*-eQTLs found in 1,469 samples



Fehrmann *et al*, PLoS Genetics 2011

Scaling up to 5,300 samples: Work in progress



Many more *trans*-eQTLs. Previously identified *trans*-eQTLs can be replicated

Westra *et al*, In preparation

Scaling up: eQTL mapping in 7,508 primary blood samples

Discovery:

Dataset	Country	Sample Size
Groningen	The Netherlands	1,469
Rotterdam Study	The Netherlands	762
Estonian Biobank	Estonia	891
SHIP-Trend	Germany	963
DILGOM	Finland	509
InChianti	United Kingdom / Italy	611
Heart and Vascular Health Study	USA	106
	Meta-analysis	**5,311**

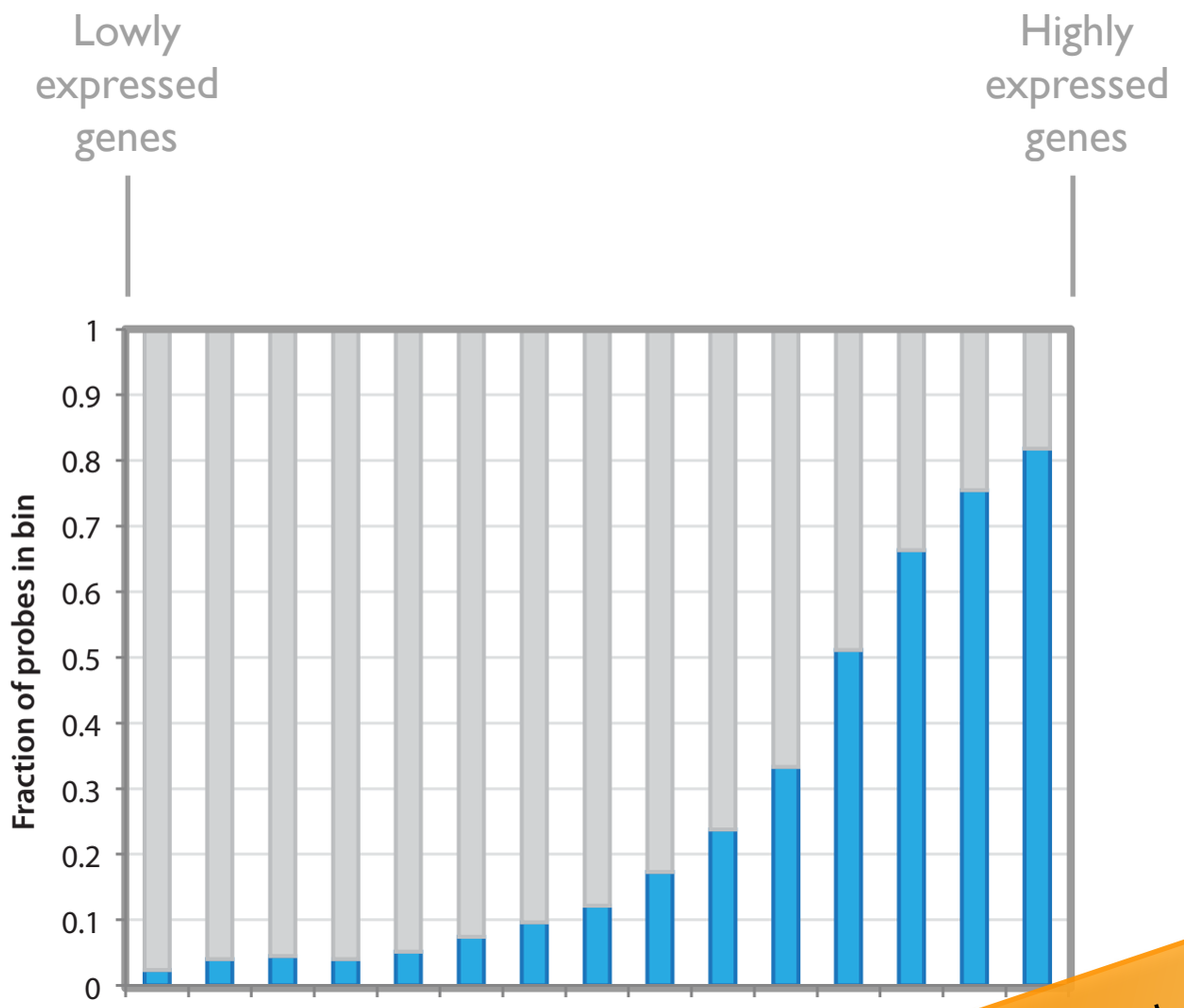
Replication:

Dataset	Country	Sample Size
KORA F4	Germany	740
BSGS	Australia	892
Monocytes (Julian Knight)	United Kingdom	283
B-Cells (Julian Knight)	United Kingdom	282
	Total	**2,197**

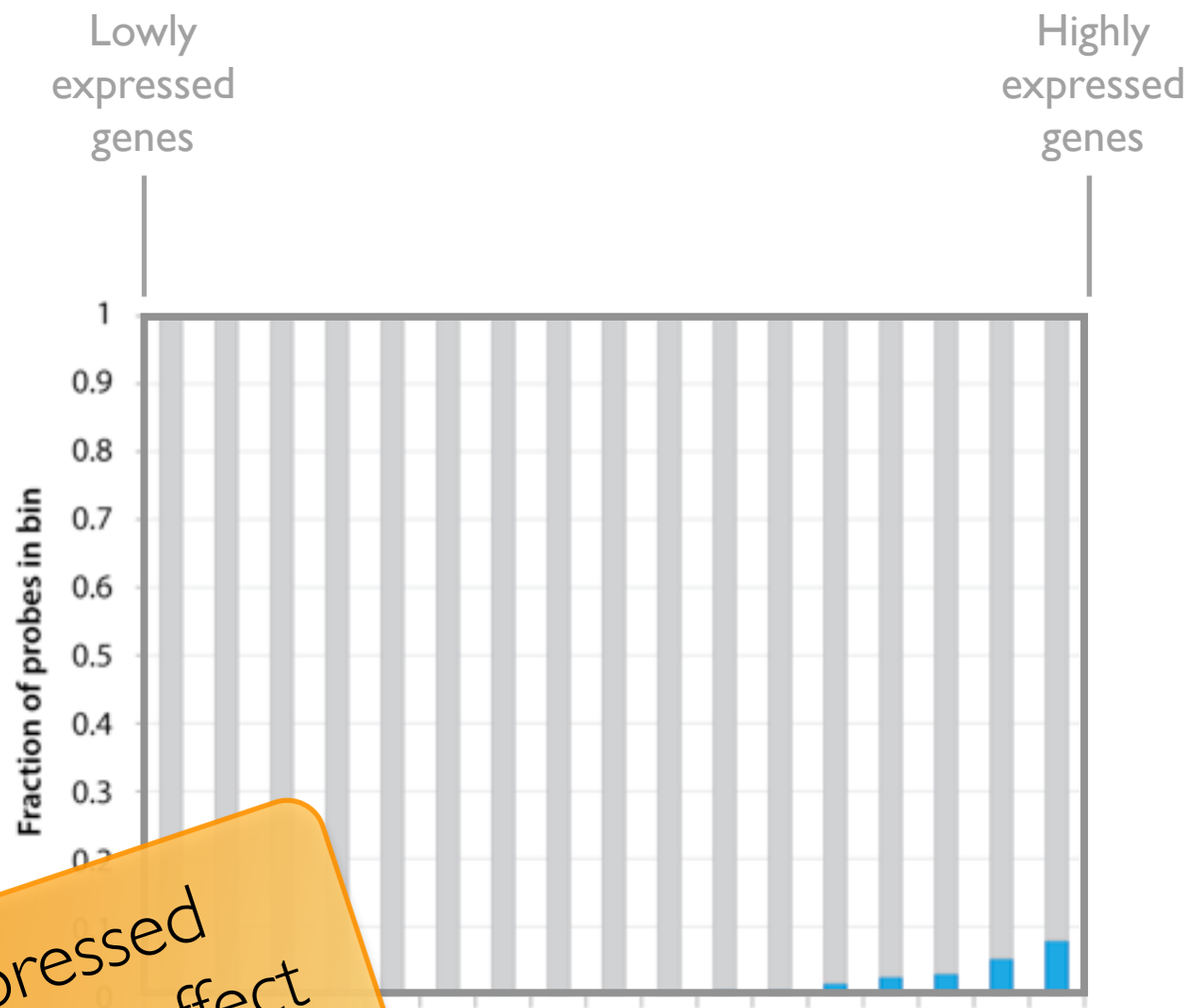
Westra *et al*, revision in preparation

eQTL properties

Detectability of *cis*-eQTLs depends on expression level



Detectability of *trans*-eQTLs depends on expression level

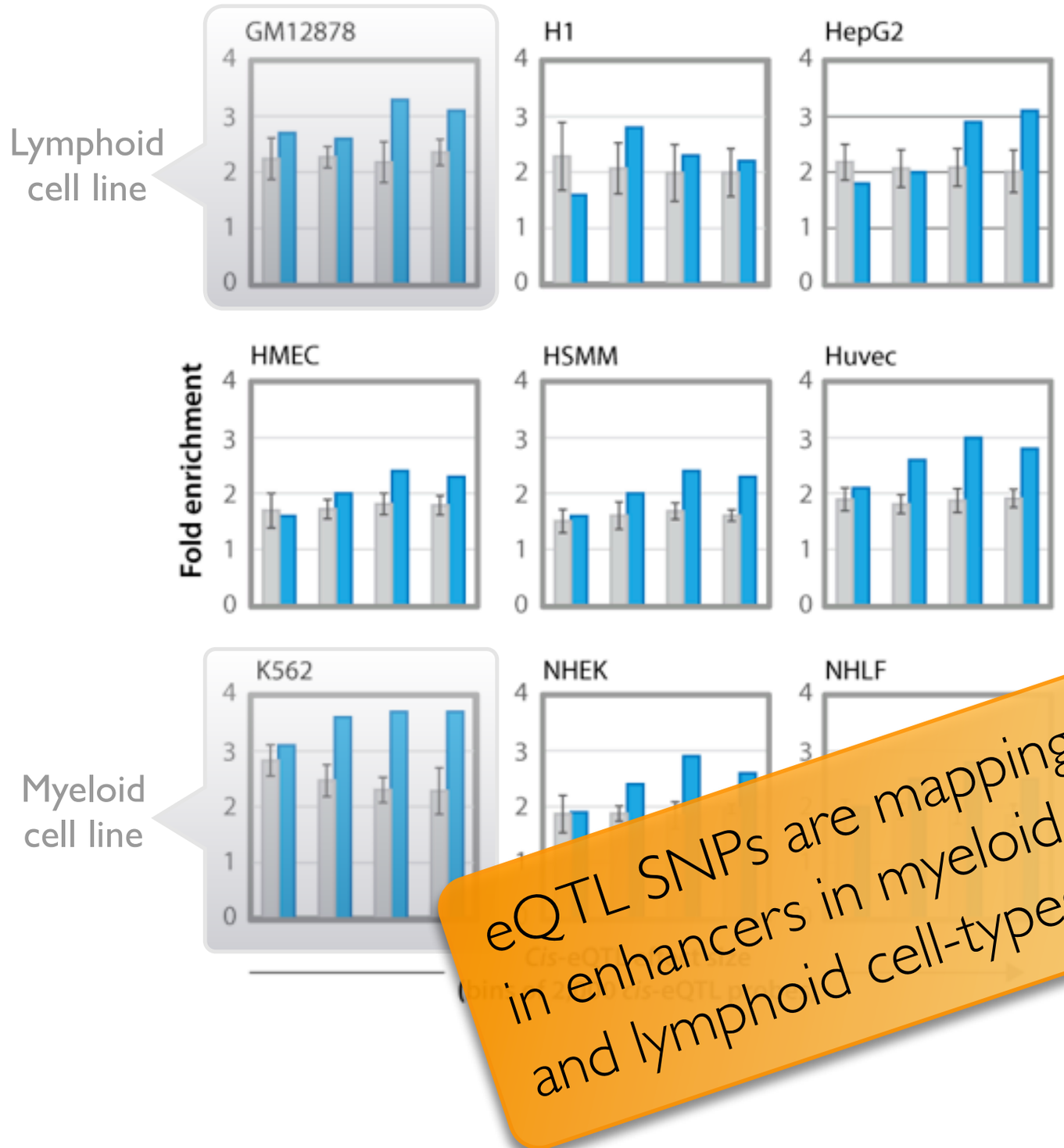


Most highly expressed genes show eQTL effect

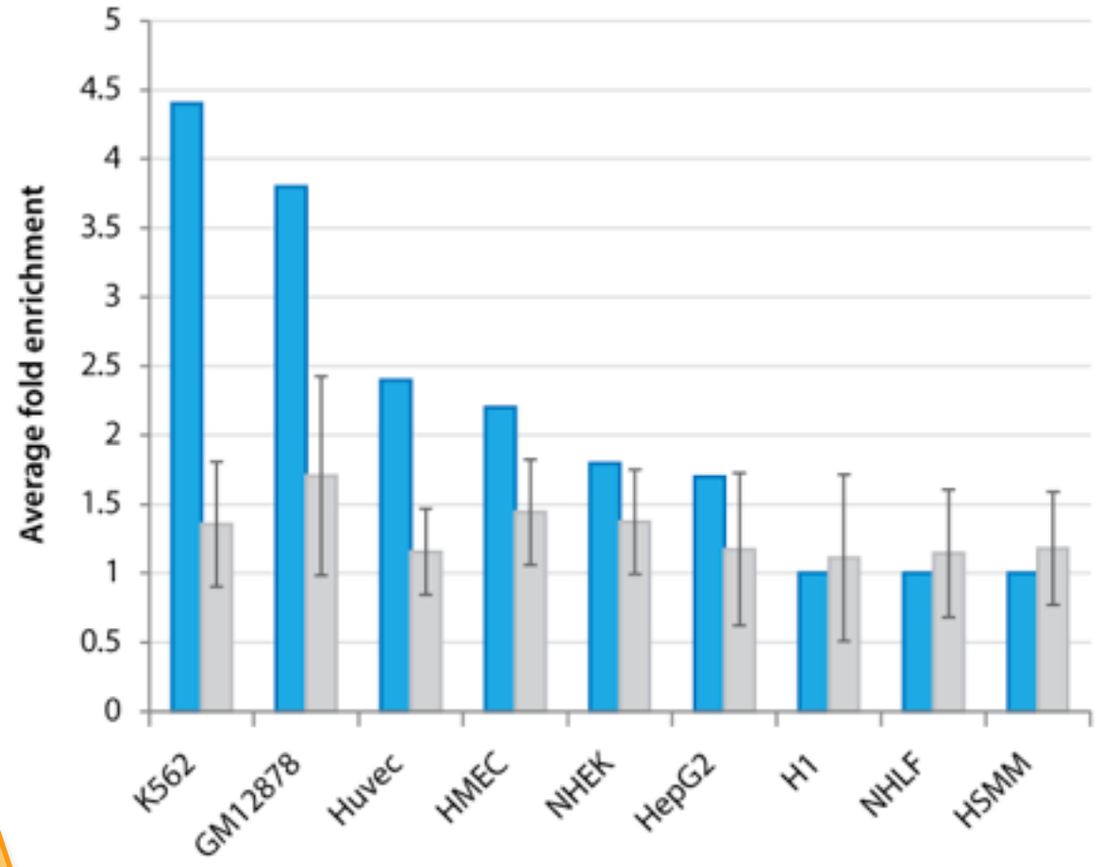
Westra *et al*, revision in preparation

eQTL properties

cis-eQTL enhancer enrichment (Haploreg)



trans-eQTL enhancer enrichment



eQTL SNPs are mapping in enhancers in myeloid and lymphoid cell-types

Westra *et al*, revision in preparation

One *trans*-eQTL highlighted: SLE

Systemic Lupus Erythamatosis (~0.1% prevalance)

2003:

Interferon-inducible gene expression signature in peripheral blood cells of patients with severe lupus

PNAS

Emily C. Baechler[†], Franak M. Batliwalla[‡], George Karypis[§], Patrick M. Gaffney[†], Ward A. Ortmann[†], Karl J. Espe[†], Katherine B. Shark[†], William J. Grande[†], Karis M. Hughes[†], Vivek Kapur[¶], Peter K. Gregersen[‡], and Timothy W. Behrens[*†]||

Interferon and Granulopoiesis Signatures in Systemic Lupus Erythematosus Blood

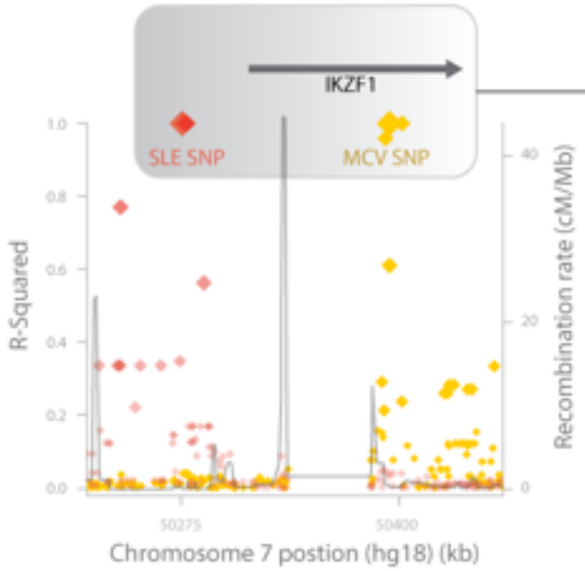
JEM

Lynda Bennett,[1,2] A. Karolina Palucka,[1] Edsel Arce,[1,2] Victoria Cantrell,[1,2] Josef Borvak,[1] Jacques Banchereau,[1] and Virginia Pascual[1,2]

Strongly differentially expressed genes: *MX1*, *IFI44L*, *C1QB* & More

Cause or consequence of SLE?

Westra *et al*, revision in preparation



Systemic lupus erythematosus allele rs4917014 (T) affects IKZF1 in cis and many genes in *trans*:

cis-eQTLs

Mean corpuscular volume
SNP rs12718597

$P = 8.14*10^{-13}$

$P = 6.29*10^{-06}$

MCV *trans*-eQTL effects

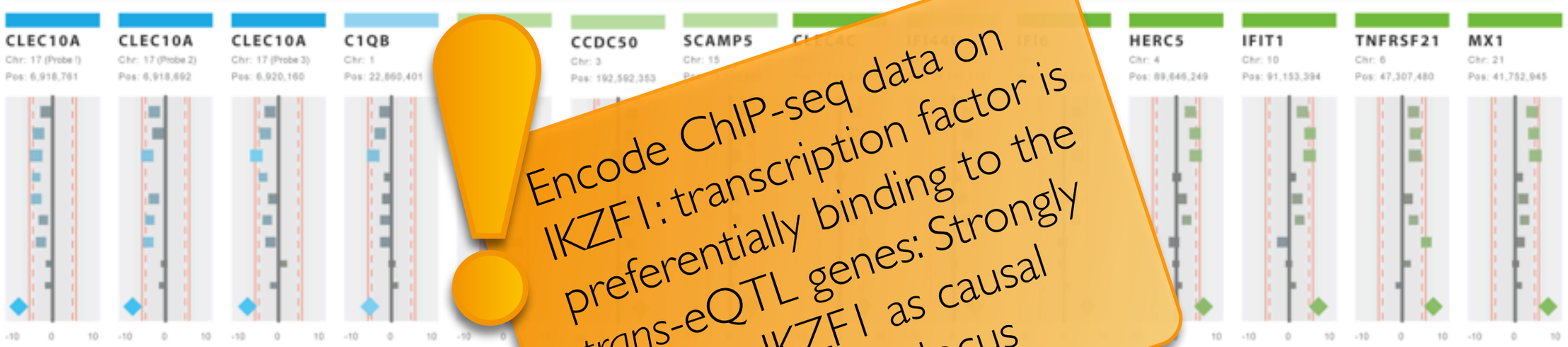
Genes involved in hemoglobin and erythrocyte metabolic processes:

AC010679.1, ALDH5A1, AP2S1, B4GALT3, C19orf62, C1orf128, C22orf13, C5orf4, CCBP2, CSDA, E2F2, EIF2AK1, EIF3S9, FAM104A, FBXO7, GCAT, GPR146, HAGH, HEMGN, HK1, HPS1, KCNH2, KLC3, KRT1, LGALS3, MAP2K3, MARCH8, MCOLN1, MSI2, OSBP2, PDLIM7, PFDN5, PLEK2, PPP2R5B, PTMS, RAP1GAP, RIOK3, RP11-529I10.4, RPIA, SESN3, SIAH2, SLC38A5, SLC6A8, SLC7A5, STOML2, TFDP1, TGM2, TMEM86B, TSTA3, VWCE

***Trans*-eQTL Legend:**

- Increases expression (FDR < 0.05)
- Increases expression (FDR < 0.5)
- Decreases expression (FDR < 0.5)
- Decreases expression (FDR < 0.05)

SLE *trans*-eQTL effects



Encode ChIP-seq data on IKZF1: transcription factor is preferentially binding to the *trans*-eQTL genes: Strongly suggests IKZF1 as causal gene in this SLE locus

Replication cohorts:												
B-Cells		✓					✓	✓	✓	✓	✓	
Monocytes	✓	✓	✓				✓	✓	✓	✓	✓	
Peripheral blood (KORA F4)	✓	✓	✓				✓	✓	✓	✓	✓	
Peripheral blood (BSGS)			✓	✓	✓		✓	✓	✓	✓	✓	

Genes involved in complement

Type I Interferon response genes
Enrichment of IKZF1 binding (Wilcoxon P = 0.05)

Conducting an eQTL study

Do it yourself:

How to conduct
an eQTL study

Conducting an eQTL study

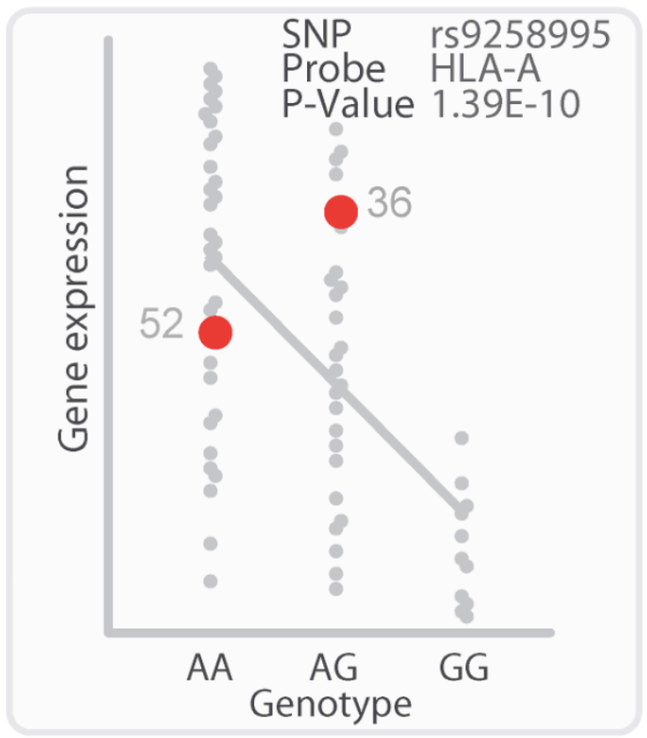
Conducting an eQTL study is not very difficult:

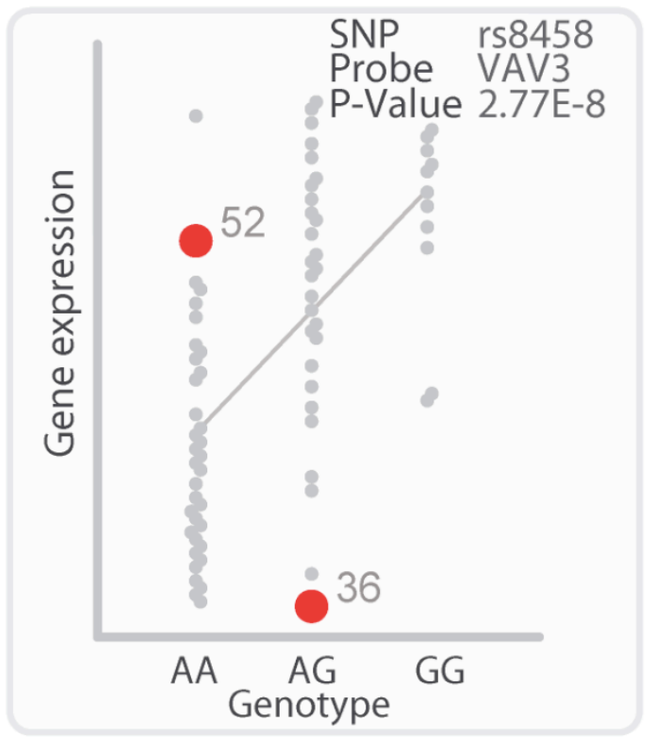
- Collect genotype data
- Collect gene expression data
- Correlate SNP genotypes with expression levels.

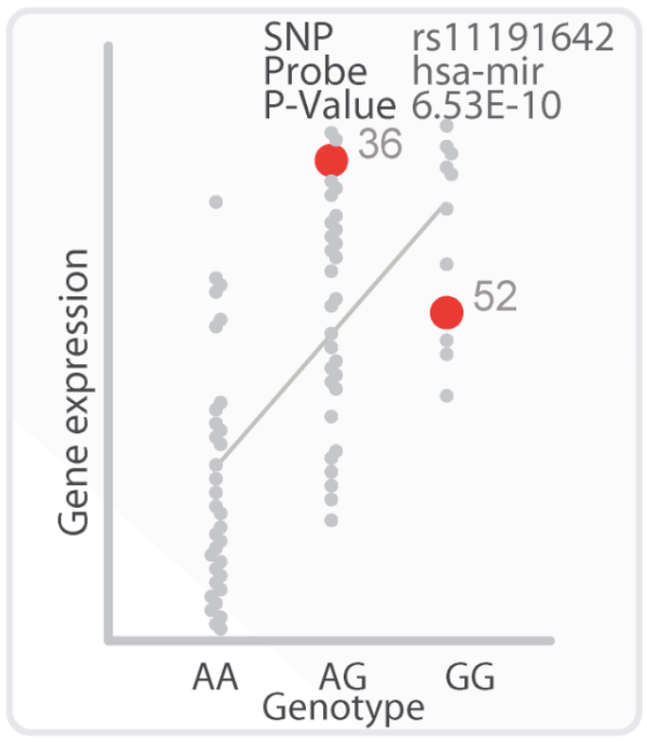
Two considerations:

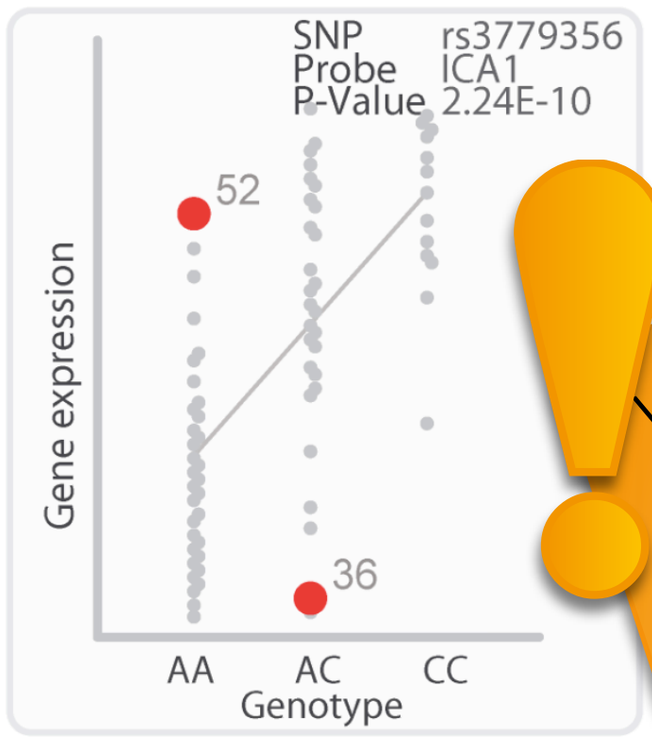
- Sample mix-ups might have actually have happened
- A considerable amount of expression variation is not genetically determined, but due to differences in physiological or metabolic state

Sample mix-ups: how to identify them





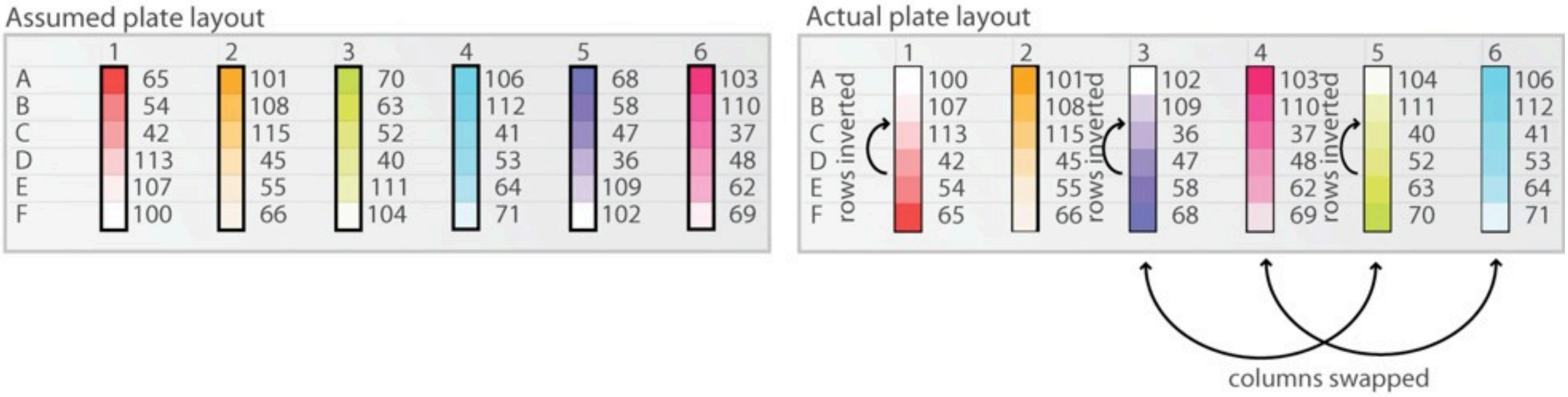




What is going on with sample 36 and 52? Sample mix-up?

Westra et al, Bioinformatics, 2011

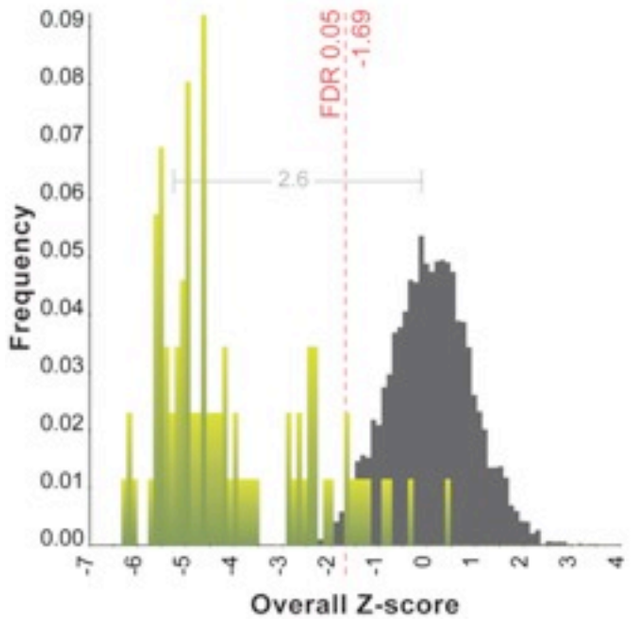
What happened to our data



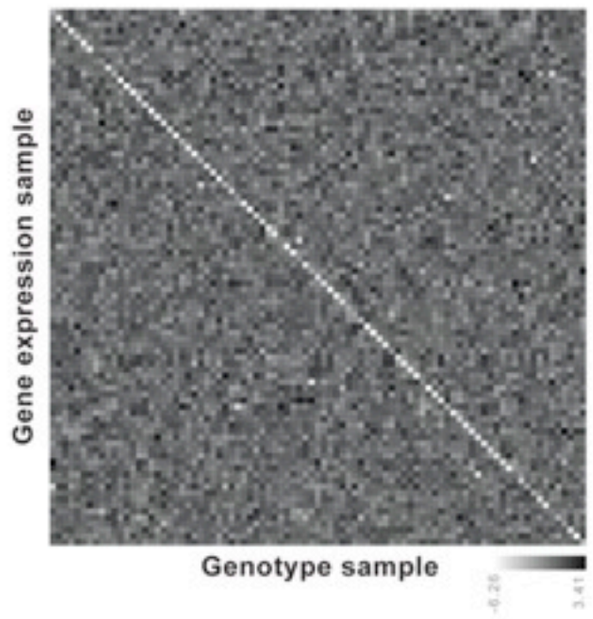
Westra et al, Bioinformatics, 2011

Sample mix-ups: how to identify them

Frequency distribution
before sample mix-up correction:

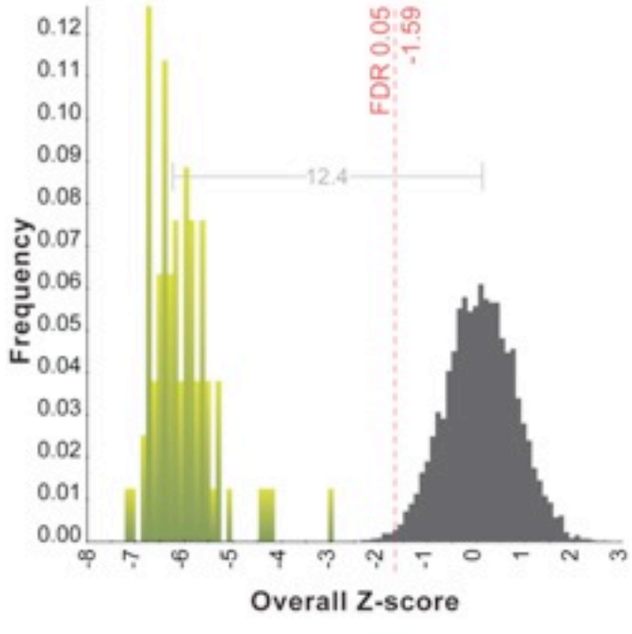


Overall Z-score heatmap
before sample mix-up correction:

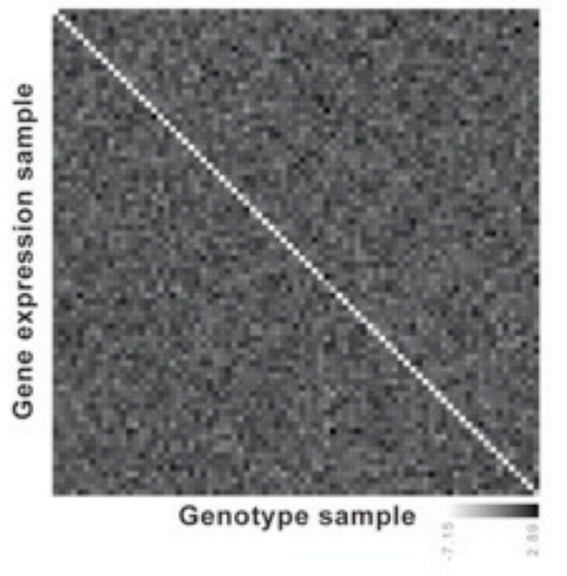


Legend
- Self - self
- Self - other (null distribution)
- Signal-to-noise ratio

Frequency distribution
after sample mix-up correction:

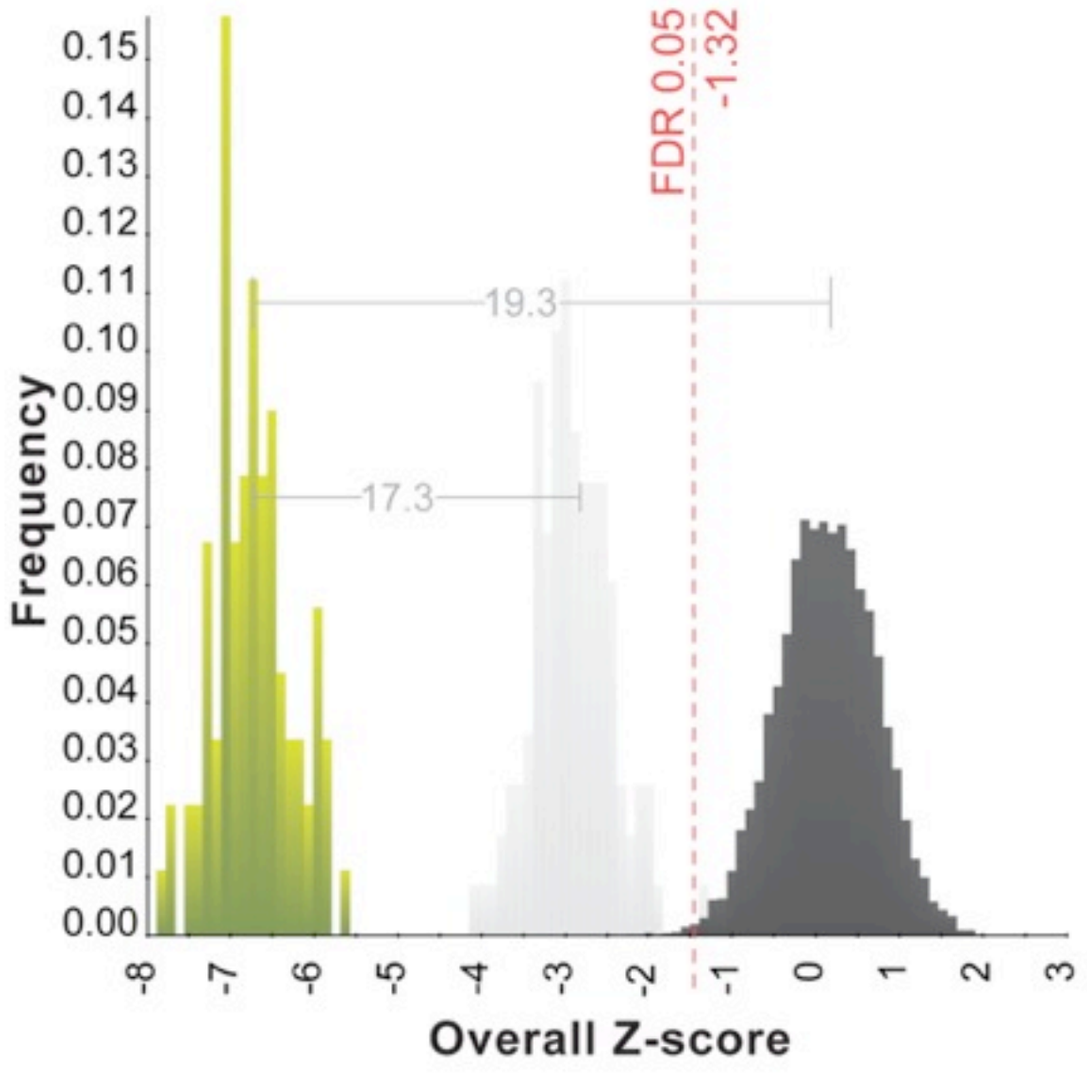


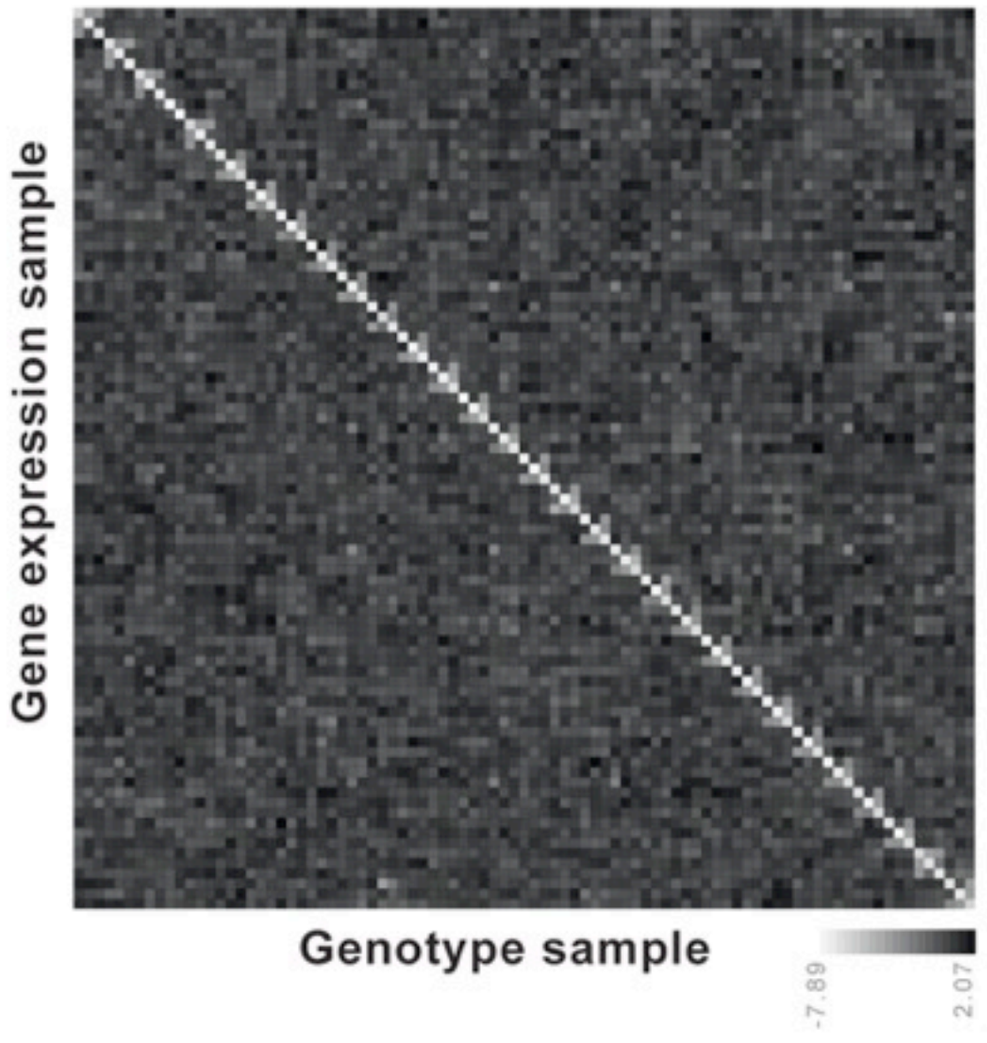
Overall Z-score heatmap
after sample mix-up correction:



Choy *et al*, PLoS Genetics 2009
Westra et al, Bioinformatics, 2011

Sample mix-ups: related samples

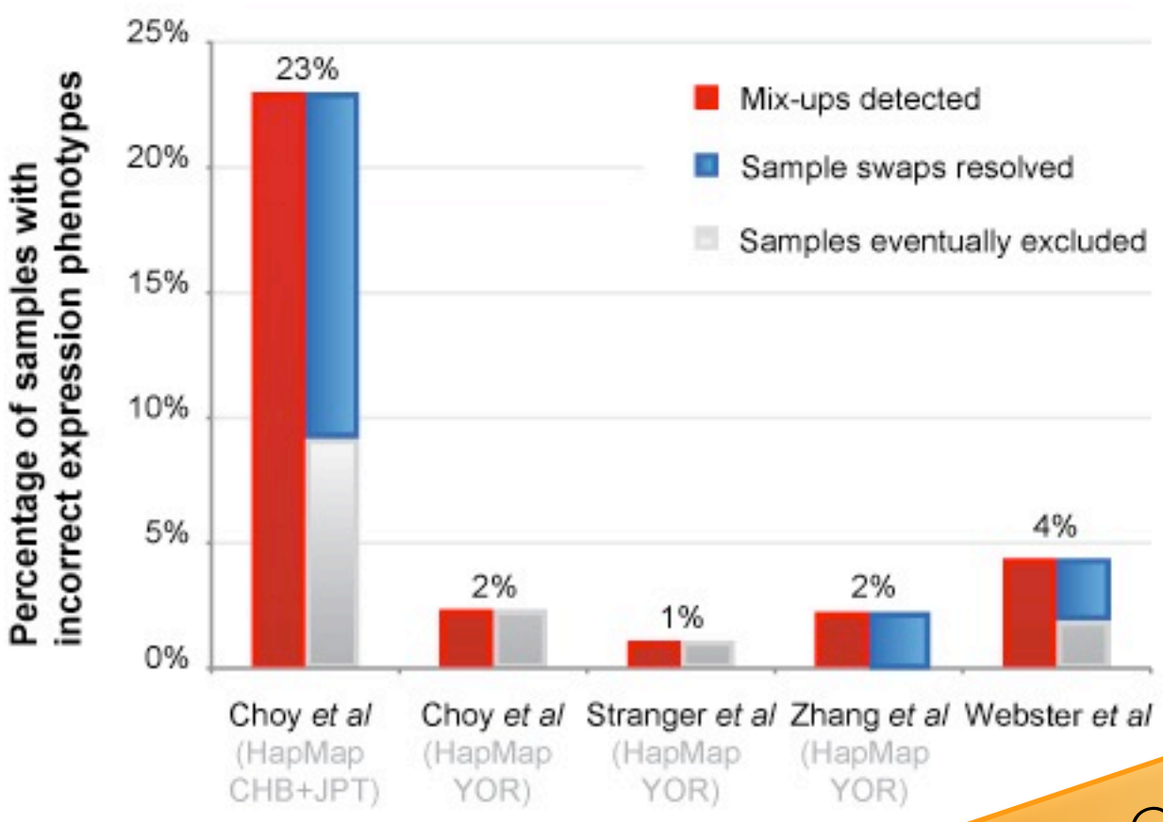




Stranger *et al*, Science 2007
Westra et al, Bioinformatics, 2011

Sample mix-ups: do they happen?

eQTL datasets with mix-ups



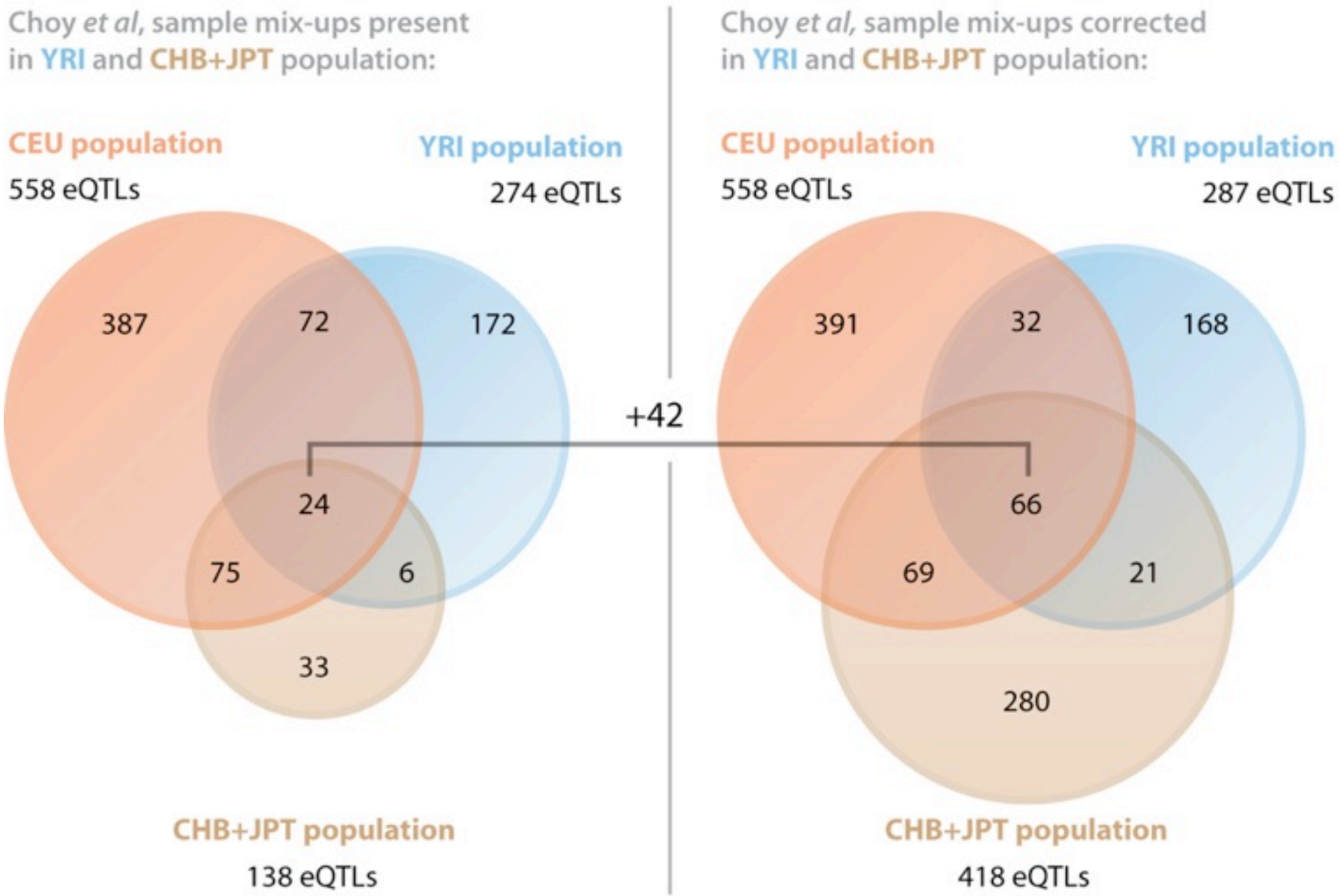
On average 3% of eQTL samples are mixed-up

Effect of correcting for these mix-ups

Westra et al, Bioinformatics, 2011

Sample mix-ups: effect of sample mix-up correction

Choy *et al* dataset, all 270 HapMap samples



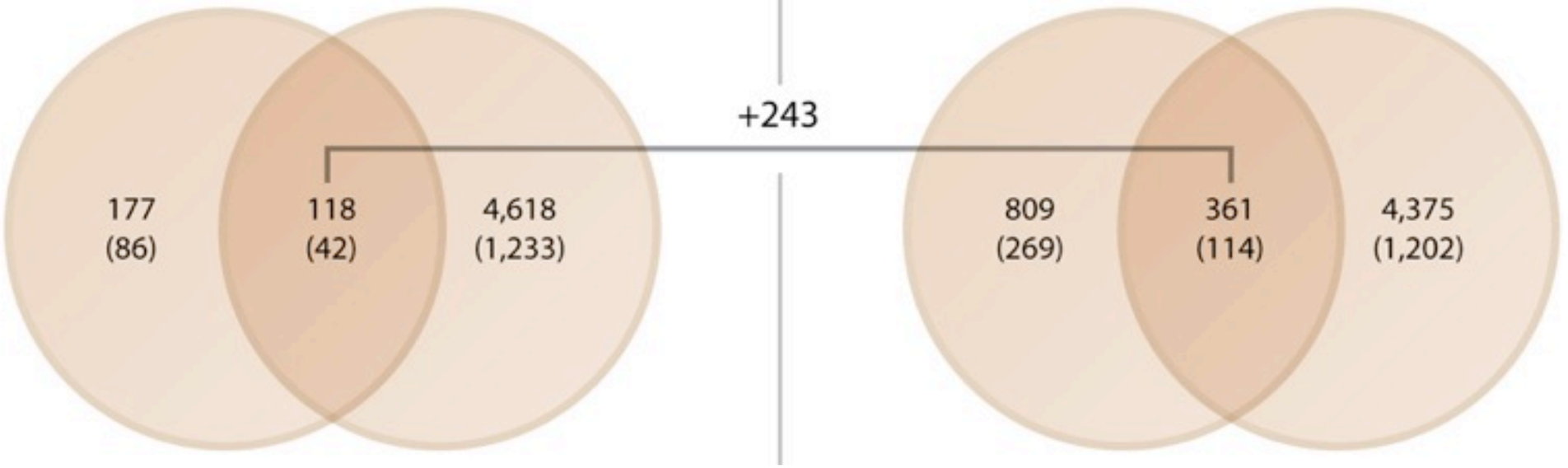
Choy *et al*, PLoS Genetics 2009
Westra et al, Bioinformatics, 2011

Comparing same samples using different platforms

Comparison between different eQTL studies
on the same HapMap CHB+JPT population.
Sample mix-ups present in Choy CHB + JPT population

Choy CHB+JPT pop.
295 unique SNP-gene combinations
(122 unique eQTL genes)

Stranger CHB`+JPT pop.
4,736 unique SNP-gene combinations
(1,244 unique eQTL genes)



Comparison between different eQTL studies
on the same HapMap CHB+JPT population.
Sample mix-ups corrected in Choy CHB+JPT population

Choy CHB+JPT pop.
1,170 unique SNP-gene combinations
(361 unique eQTL genes)

Stranger CHB`+JPT pop.
4,736 unique SNP-gene combinations
(1,244 unique eQTL genes)

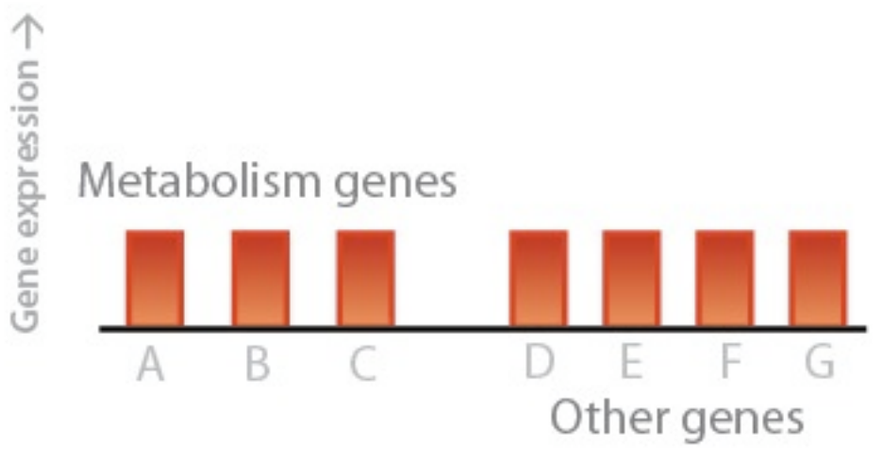
Stranger *et al*, Science 2007
Choy *et al*, PLoS Genetics 2009
Westra et al, Bioinformatics, 2011

Remove non-genetic expression variation

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



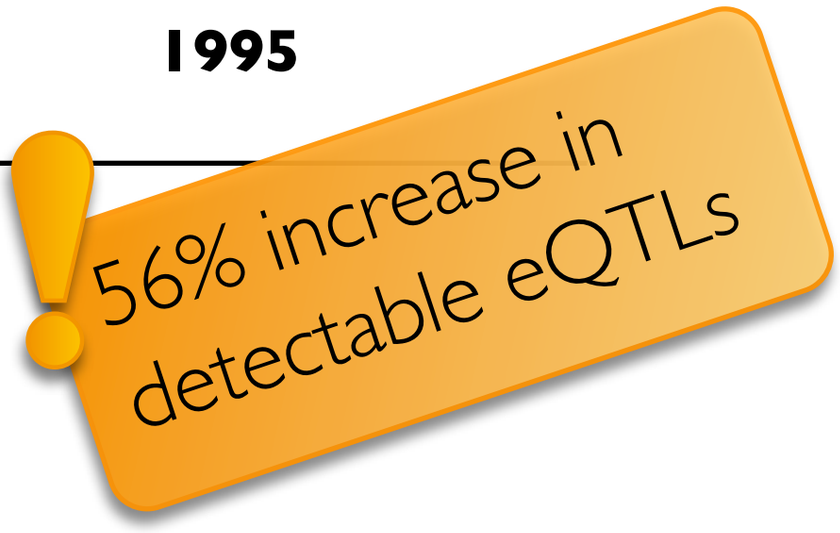
RNA b
expressi
after din

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

Dubois *et al*, Nature Genetics 2010

Correct for sample-mixups and non-genetic components

HapMap Population	Original data	After sample mix-up identification and correction	After mix-up correction and removal of physiological and metabolic components
CEU	558	558	717
YRI	274	287	383
CHB+JPT	138	418	661
Meta-analysis	1277	1508	**1995**



nature | April 2010

Transcriptome genetics using second generation sequencing in a Caucasian population

Stephen B. Montgomery[1,2], Micha Sammeth[3], Maria Gutierrez-Arcelus[1], Radoslaw P. Lach[2], Catherine Ingle[2], James Nisbett[2], Roderic Guigo[3] & Emmanouil T. Dermitzakis[1,2]

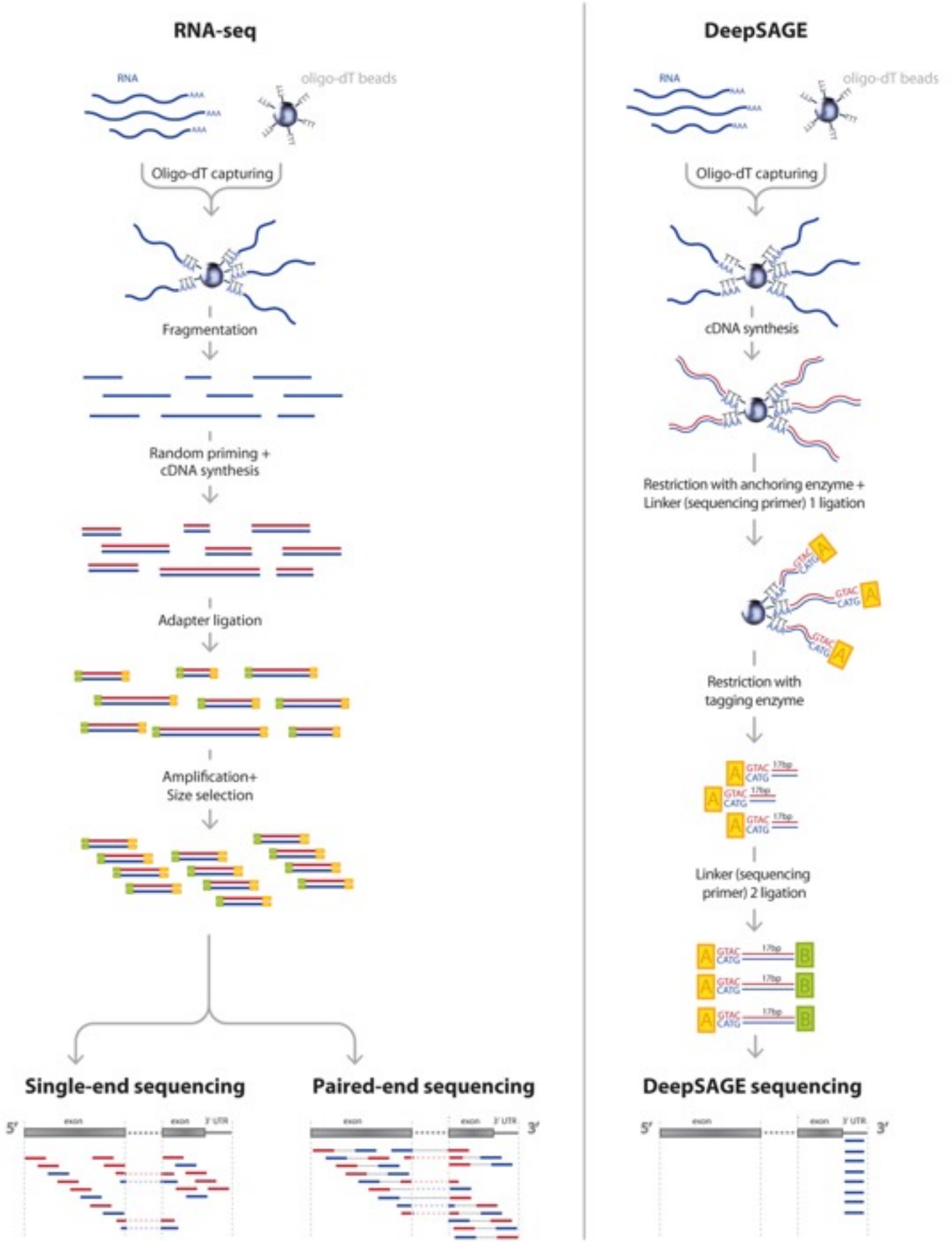
Understanding mechanisms underlying human gene expression variation with RNA sequencing

Joseph K. Pickrell[1], John C. Marioni[1], Athma A. Pai[1], Jacob F. Degner[1], Barbara E. Engelhardt[2], Everlyne Nkadori[1,3], Jean-Baptiste Veyrieras[1], Matthew Stephens[1,4], Yoav Gilad[1] & Jonathan K. Pritchard[1,3]

Observations:

- RNA-sequencing can interrogate entire transcriptome
- Possible to investigate allele-specific expression
- With limited read-depth accurate gene expression levels can be established

eQTLs in single-end, paired-end and deepSAGE data



Zhernakova *et al*, submitted