Improving the diagnostic yield in clinical genetics by recycling publicly available RNA-seq data

Expression data

Public data

Phenotypes

UMCG

Genetics Department

‘To capture something small
you need something big’

CERN

© Ruben van Leer

DNA

AC

GT

‘To capture something small you need something big’

DNA Sequencers

© Sanger Institute

‘To capture something small you needed something big’

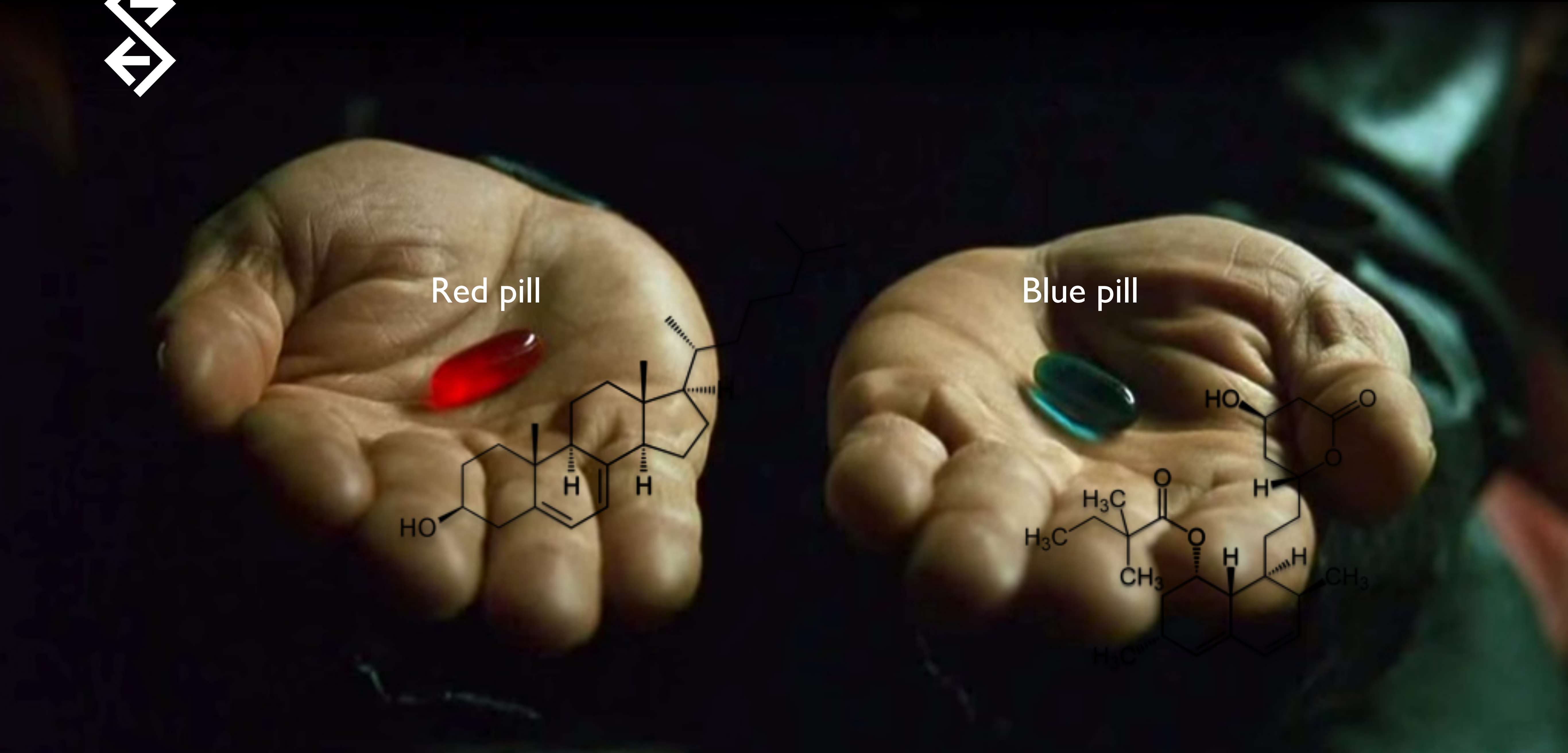
DNA Sequencer

Minion

© Oxford Nanopore

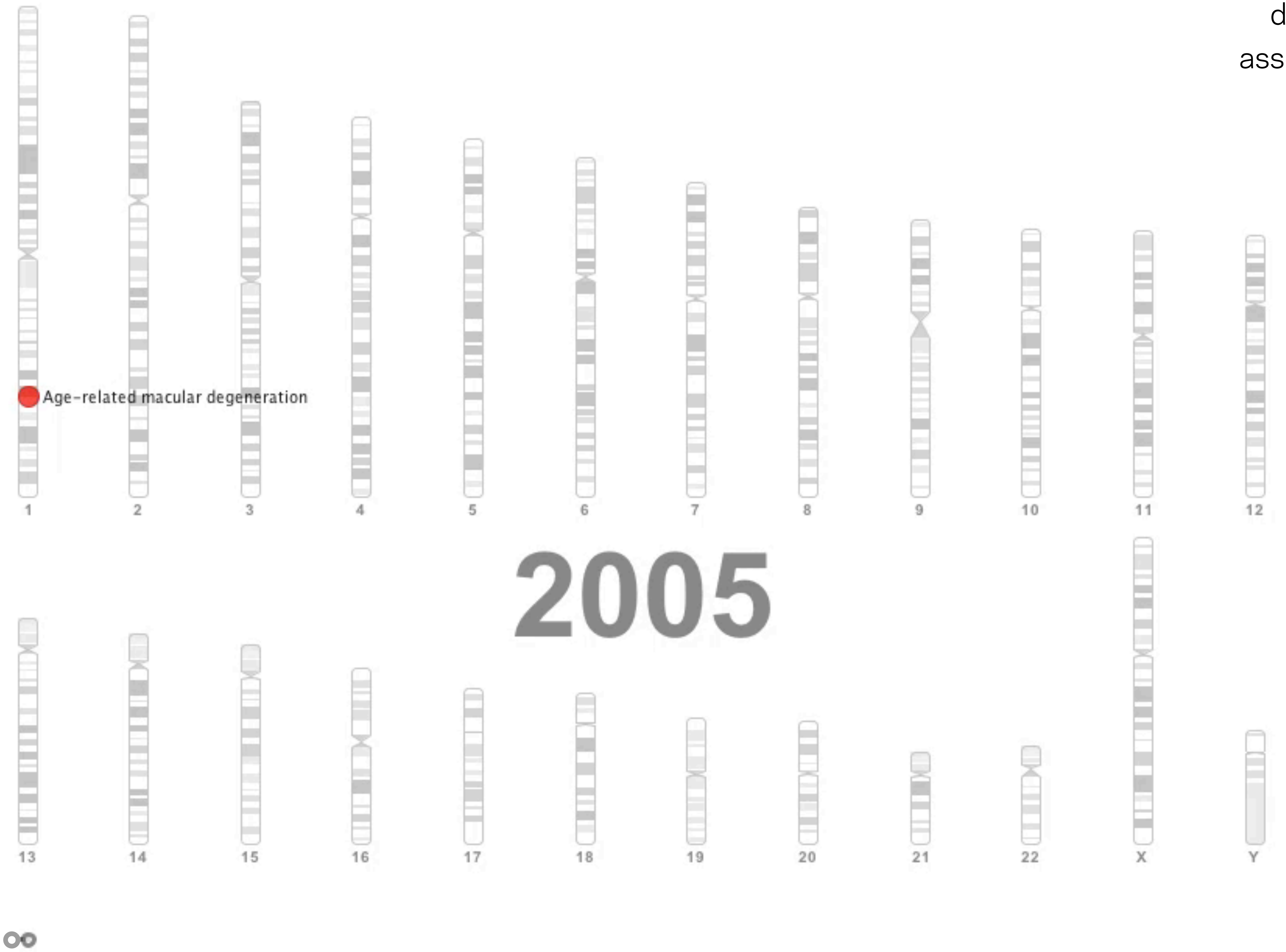
large amounts of data now available

Goal: better diagnose and treat patients



Seven years of GWAS studies

6,054 disease associations

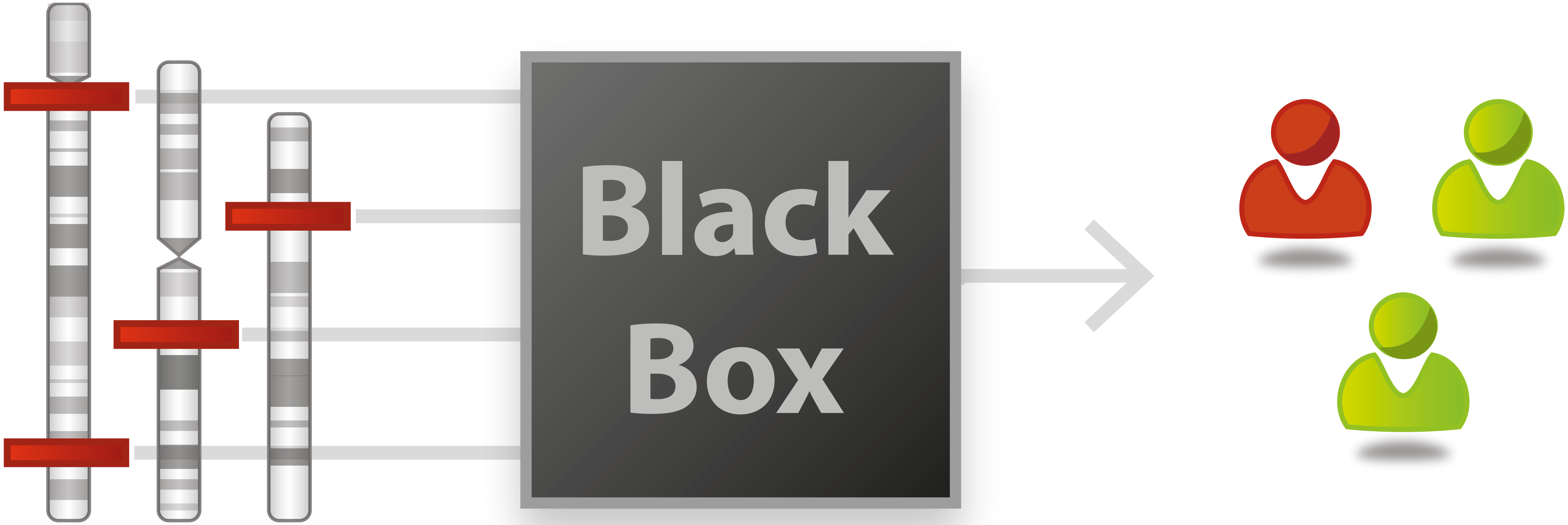


Teri Manolio *et al*: A catalog of published genome-wide association studies

Problem of life science community

Genetic risk factors

Disease



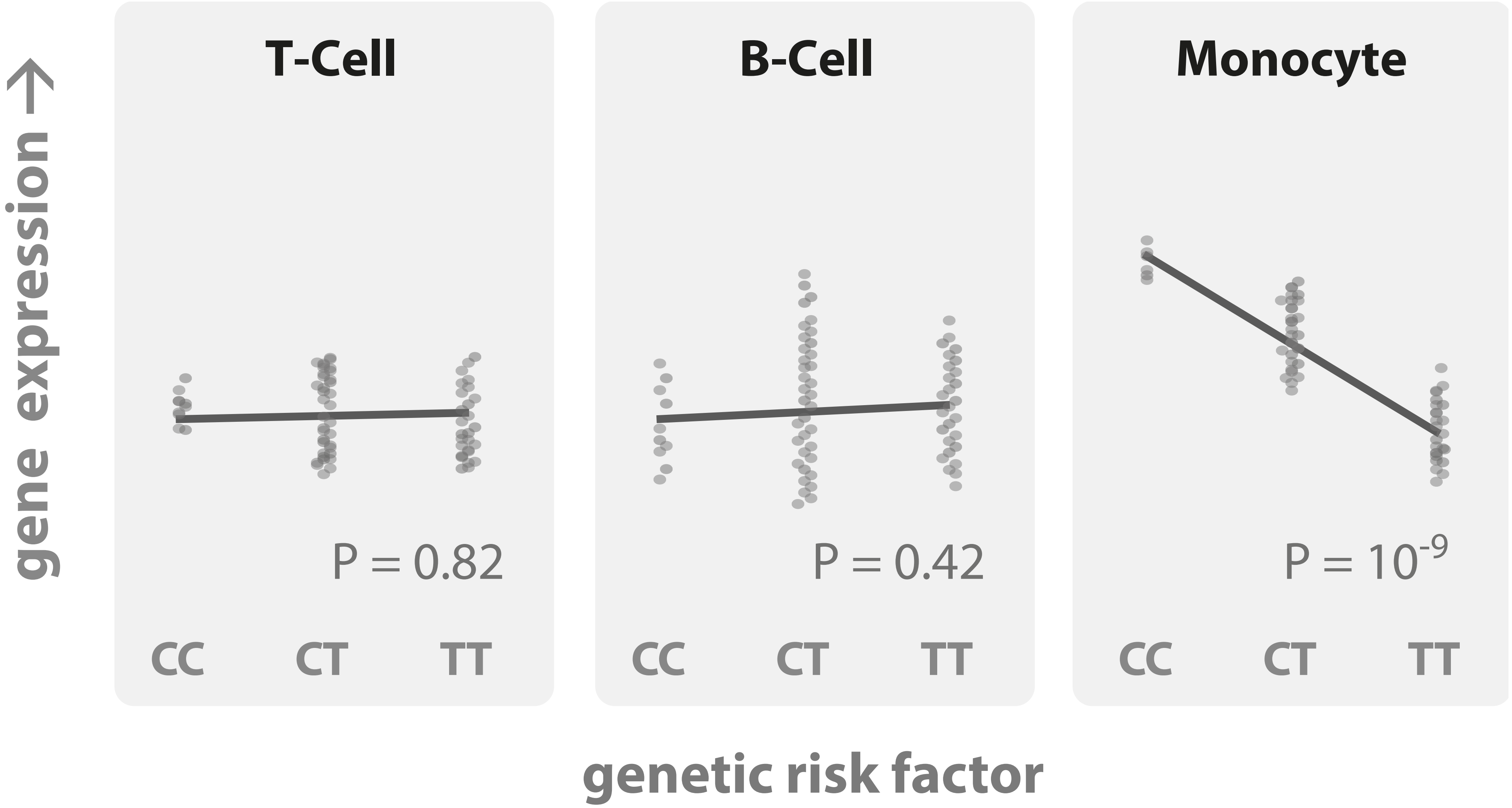
Black Box

>10,000 known

Genes unknown
Pathways unknown
Cell-types unknown

>200 diseases

Far majority of genetic risk factors affect gene expression



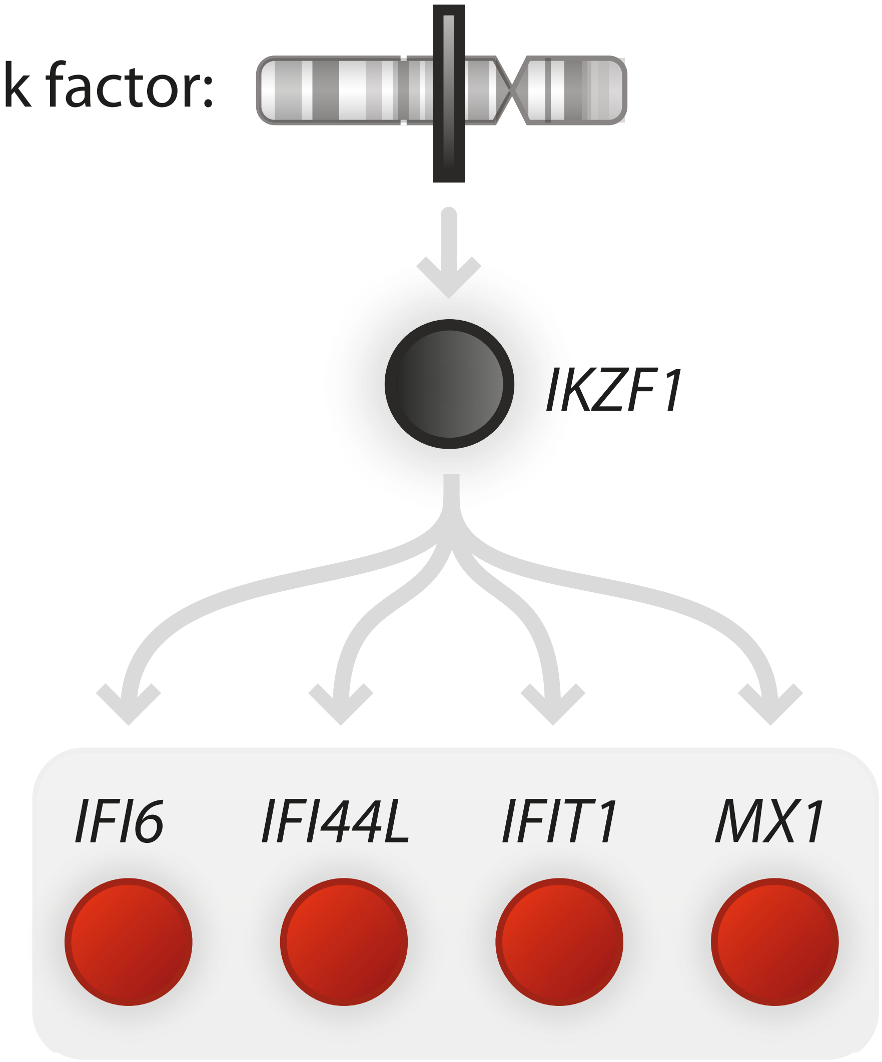
Dubois *et al*, Nature Genetics 2010

Fehrmann *et al*, PLoS Genetics 2011

Fu *et al*, PLoS Genetics 2012

Westra *et al*, Nature Genetics 2013

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



Systemic lupus erythematosis risk factor:

Chr. 7

Local expression effect:

IKZF1

Chr. 7

Type 1 interferon response:

(in Monocytes)

IFI6 *IFI44L* *IFIT1* *MX1*

Downstream *trans*-eQTL effects

Downstream effects identified for >200 genetic risk factors

New meta-analysis ongoing in 25,000 blood samples

Westra *et al*, Nature Genetics 2013

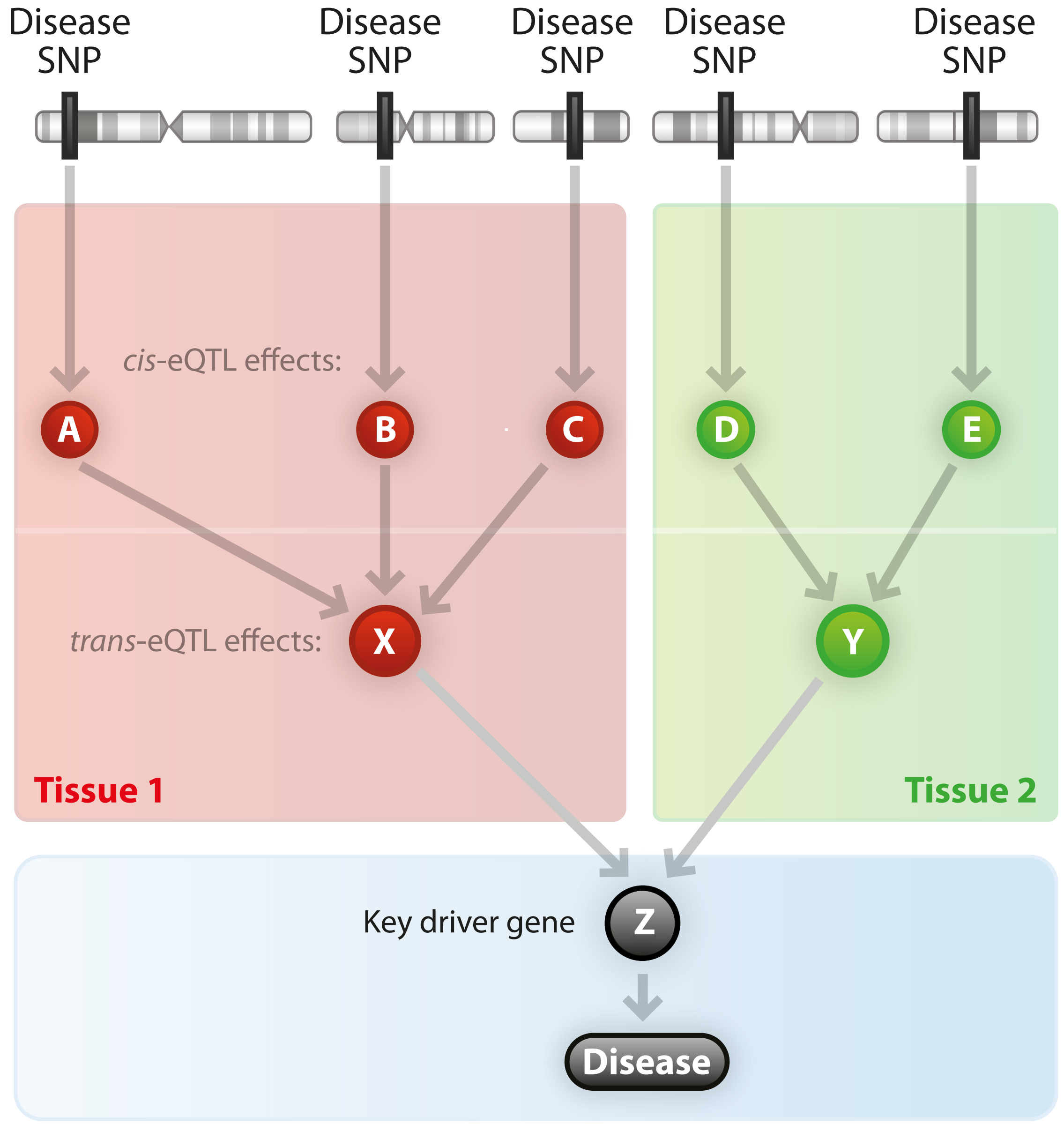
Goal

Genome-wide association studies

cis-eQTL mapping

trans-eQTL mapping

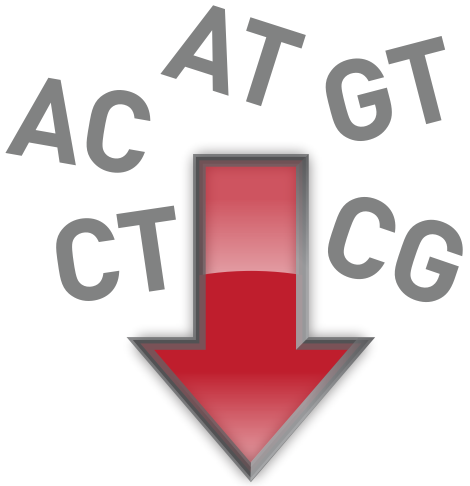
Key driver gene identification



But what about patients we see?

Patient with a severe disease.
You suspect a genetic cause.
What do you do?

- Targeted gene panel?
- Whole exome sequencing?
- Whole genome sequencing?



Problem:
Many (rare) variants
of unknown significance

Smart ways to filter?

AC
CT
AT
GT
AG

gene expression?

Transcriptome of the Netherlands

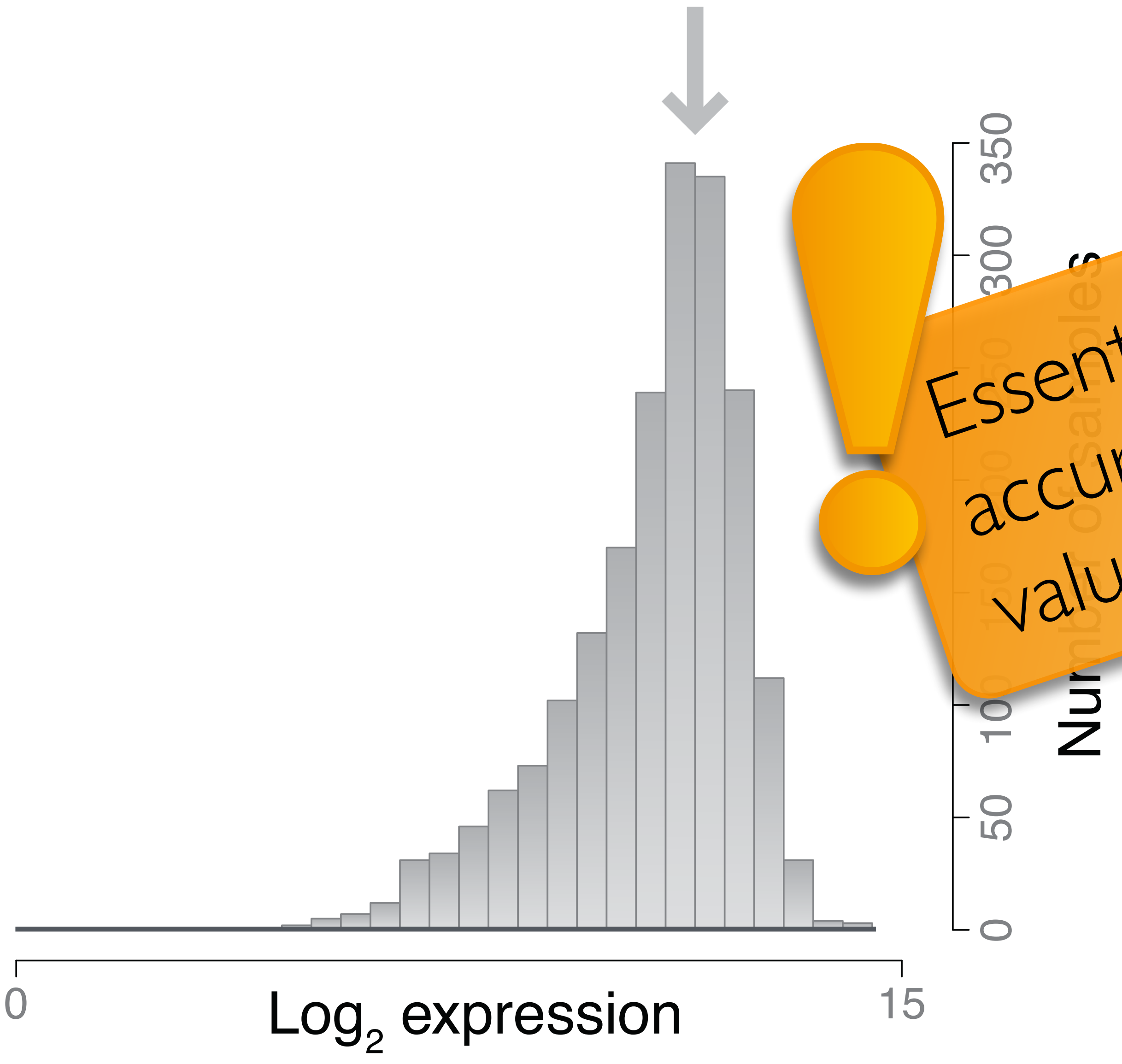
- Rare genetic variants also have effects on gene expression

- Rationale BBMRI-NL BIOS Consortium to establish ‘Transcriptome of the Netherlands’ in 5,000 population based samples

- Generate RNA-seq data on patients. Contrast these expression values to the Transcriptome of the Netherlands.

Transcriptome of the Netherlands

***TRIM51BP* gene expression distribution in the Dutch population**



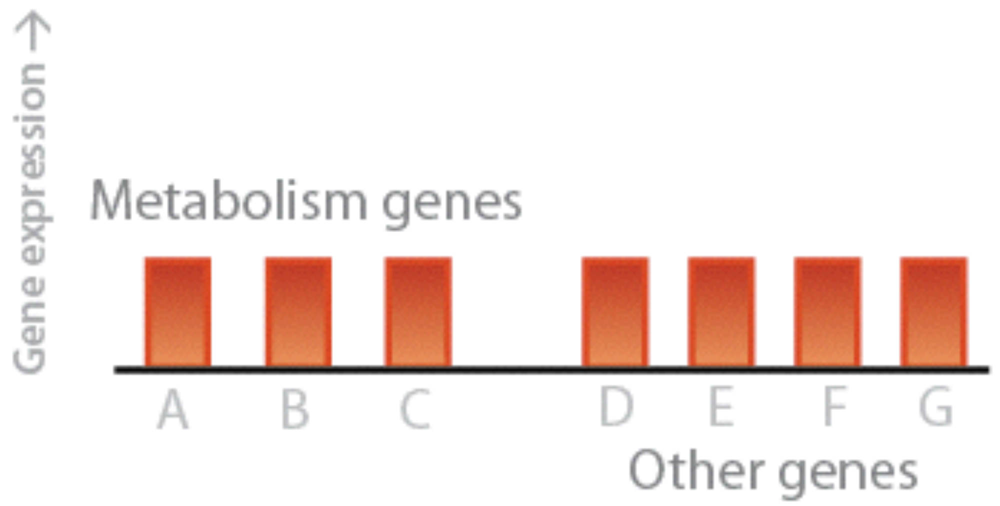
Essential to get very accurate reference values for each gene

Remove non-genetic expression variation

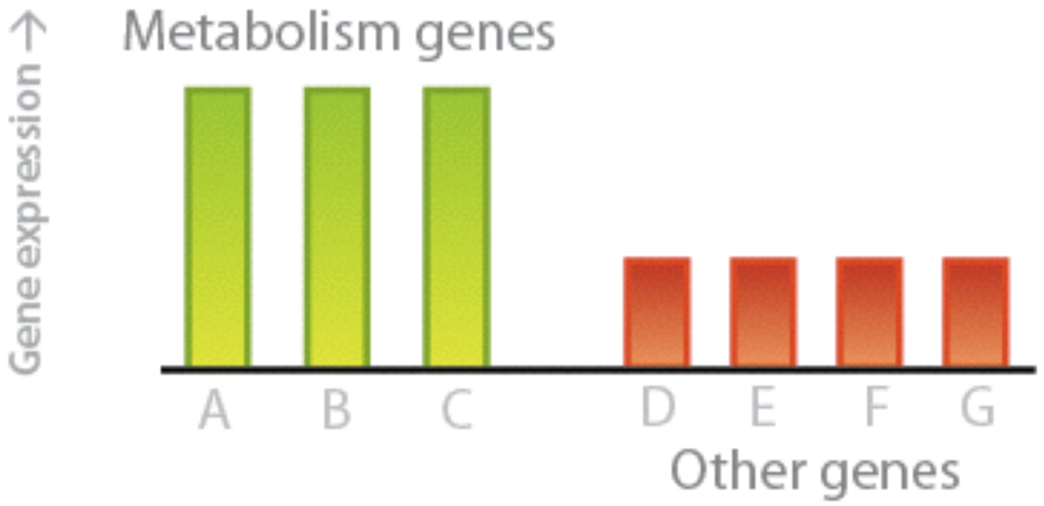
Most expression variation due to:

- Physiological state
- Metabolic state
- Environmental state

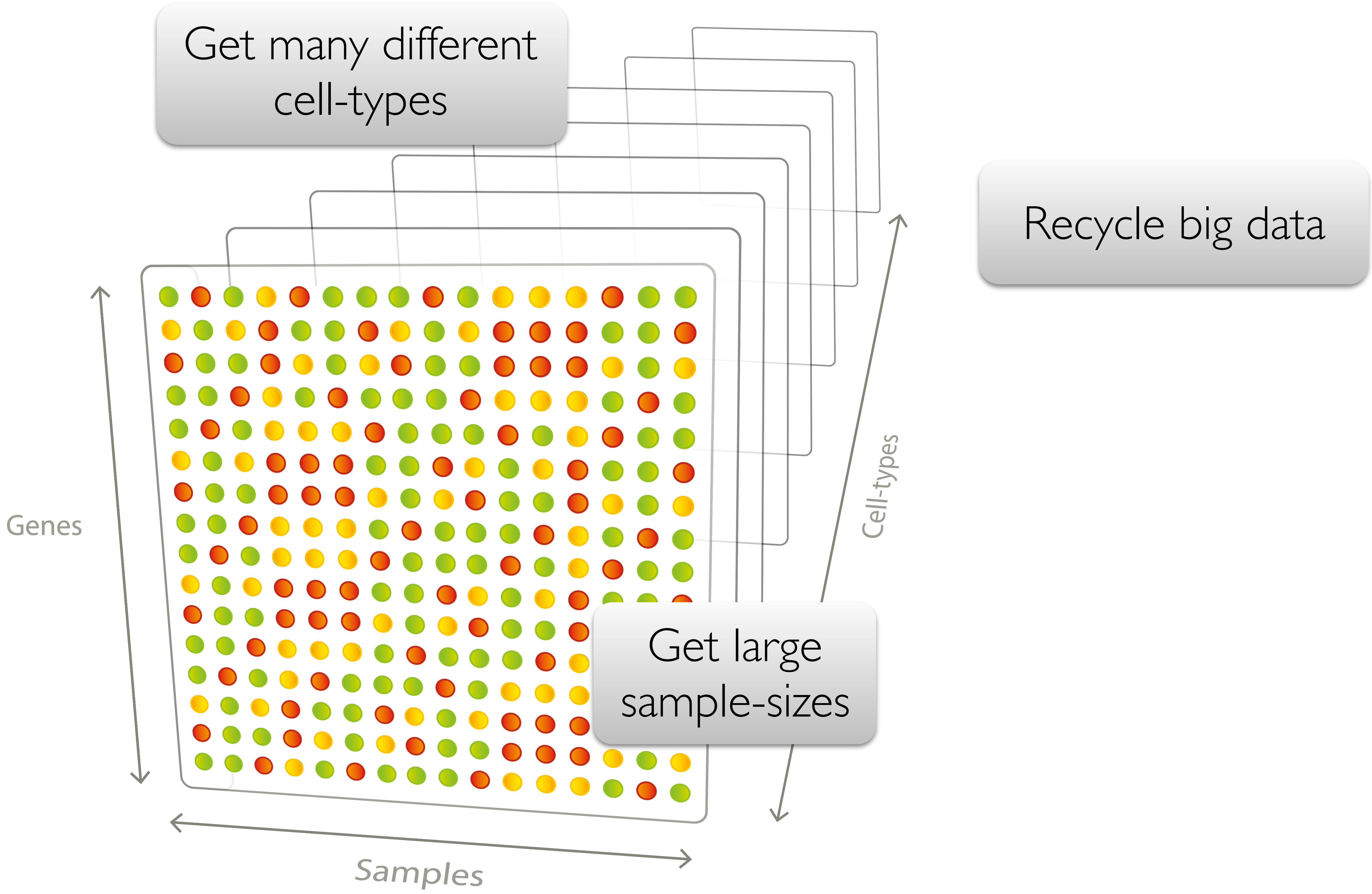
RNA blood expression
when you wake up



RNA blood expression
after nice diner



Strategies



Get many different cell-types

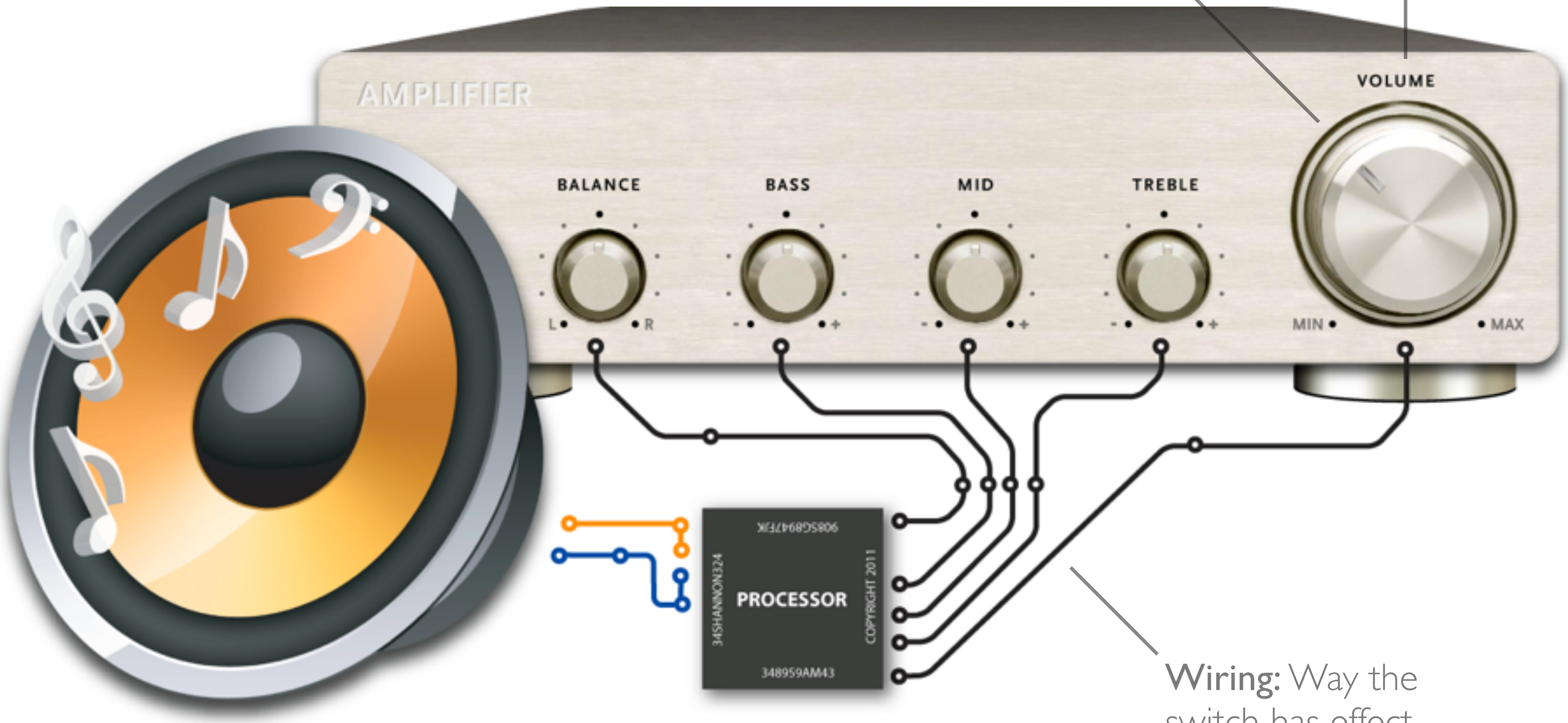
Recycle big data

Get large sample-sizes

Amplifier can change many aspects of music

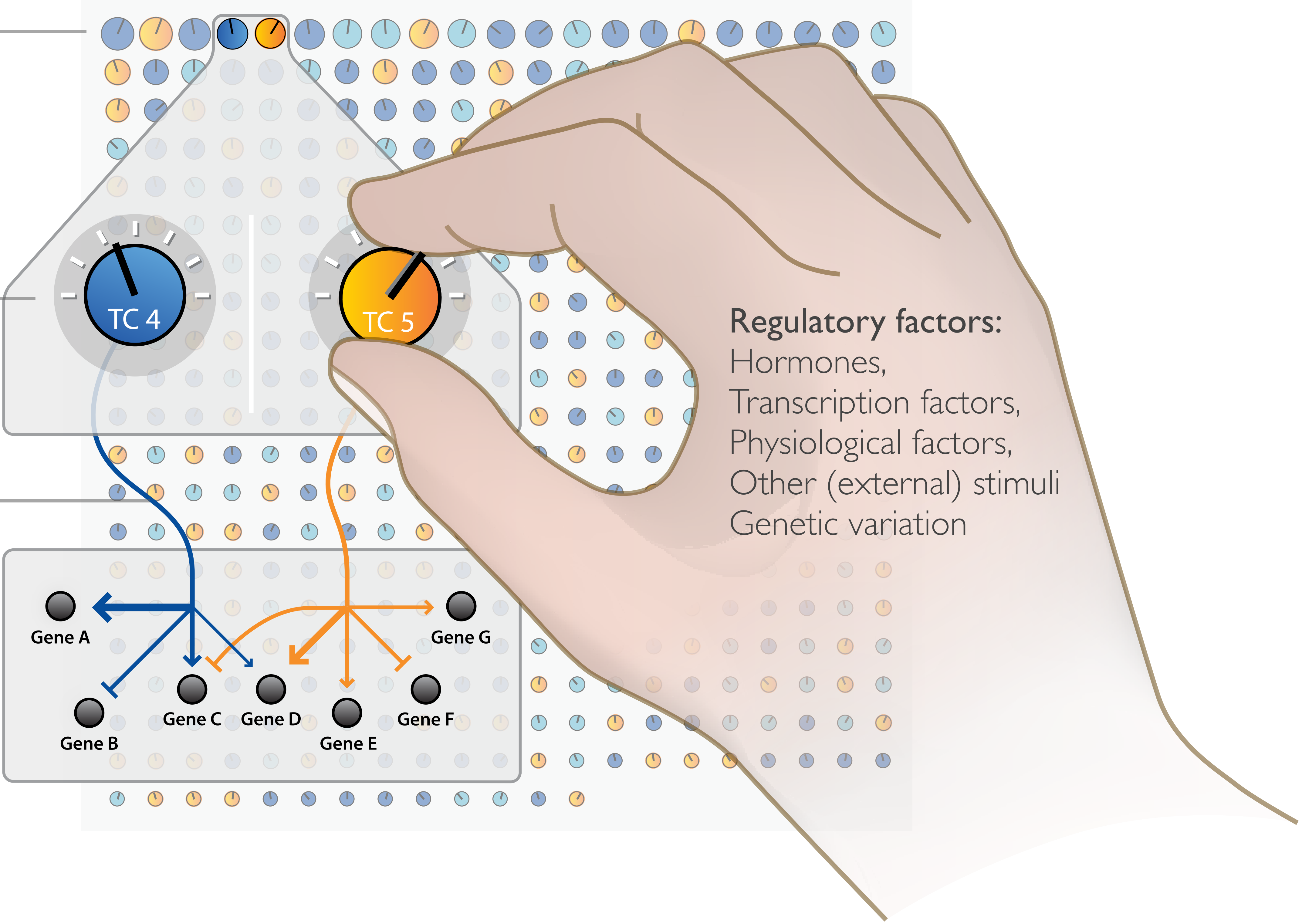
Setting:
Activity of switch

Size of switch:
Importance of switch



Wiring: Way the
switch has effect

A control panel that determines gene expression?



Size of switch: Importance

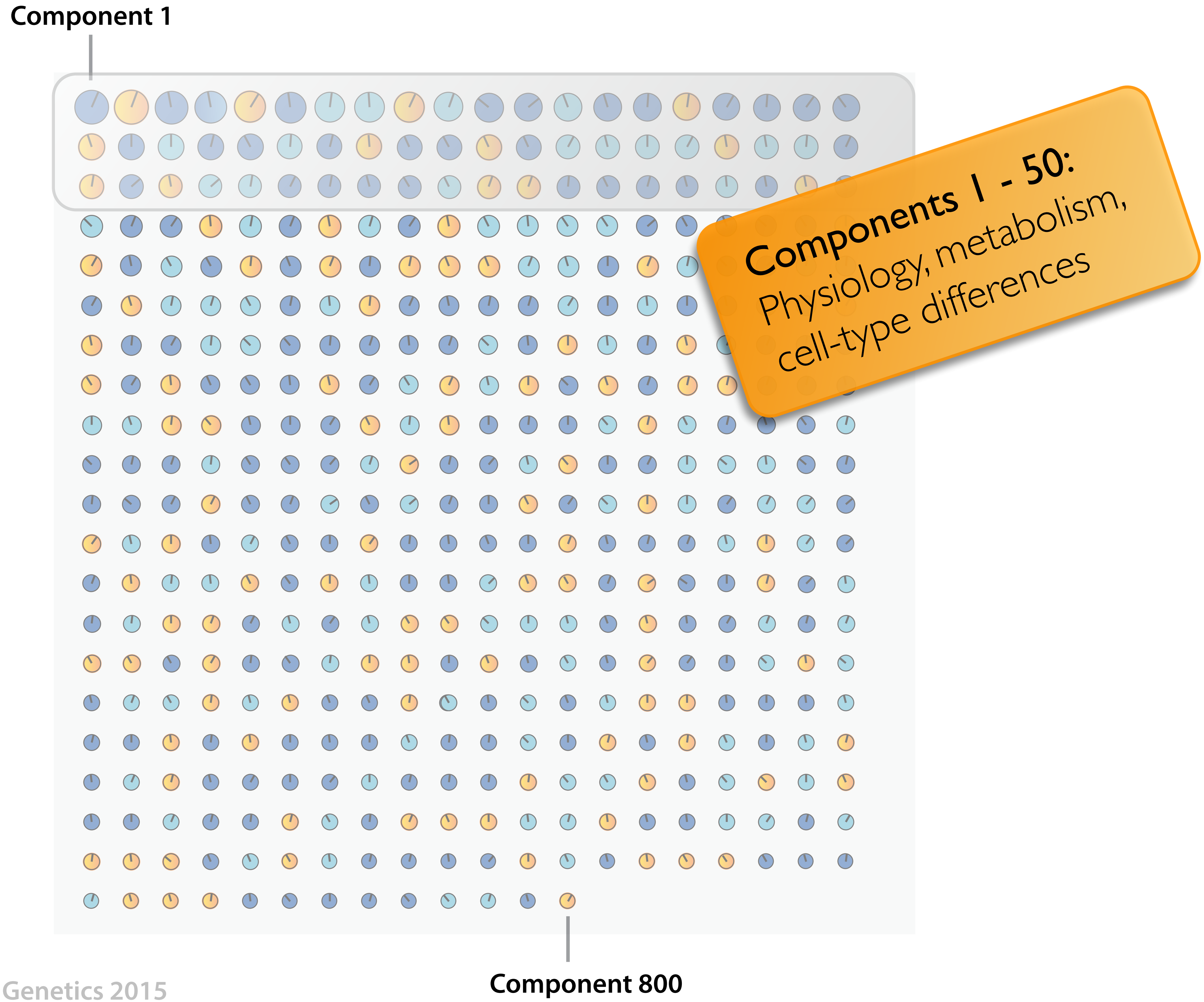
Setting: State of a certain sample

Wiring: Effect on individual genes

Regulatory factors:
Hormones,
Transcription factors,
Physiological factors,
Other (external) stimuli
Genetic variation

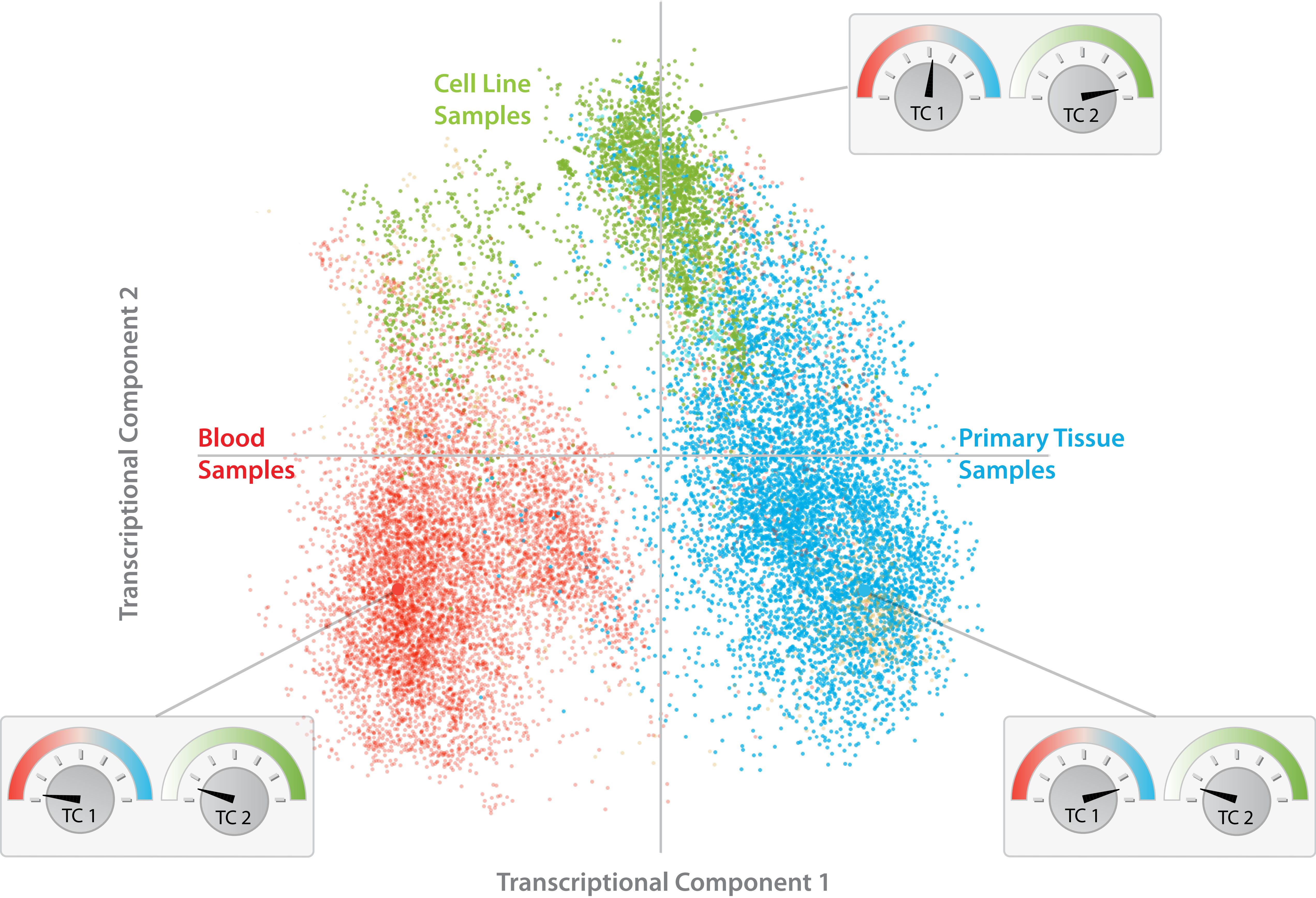
Fehrmann *et al*, Nature Genetics 2015

800 ‘transcriptional components’: Component 1 - 50

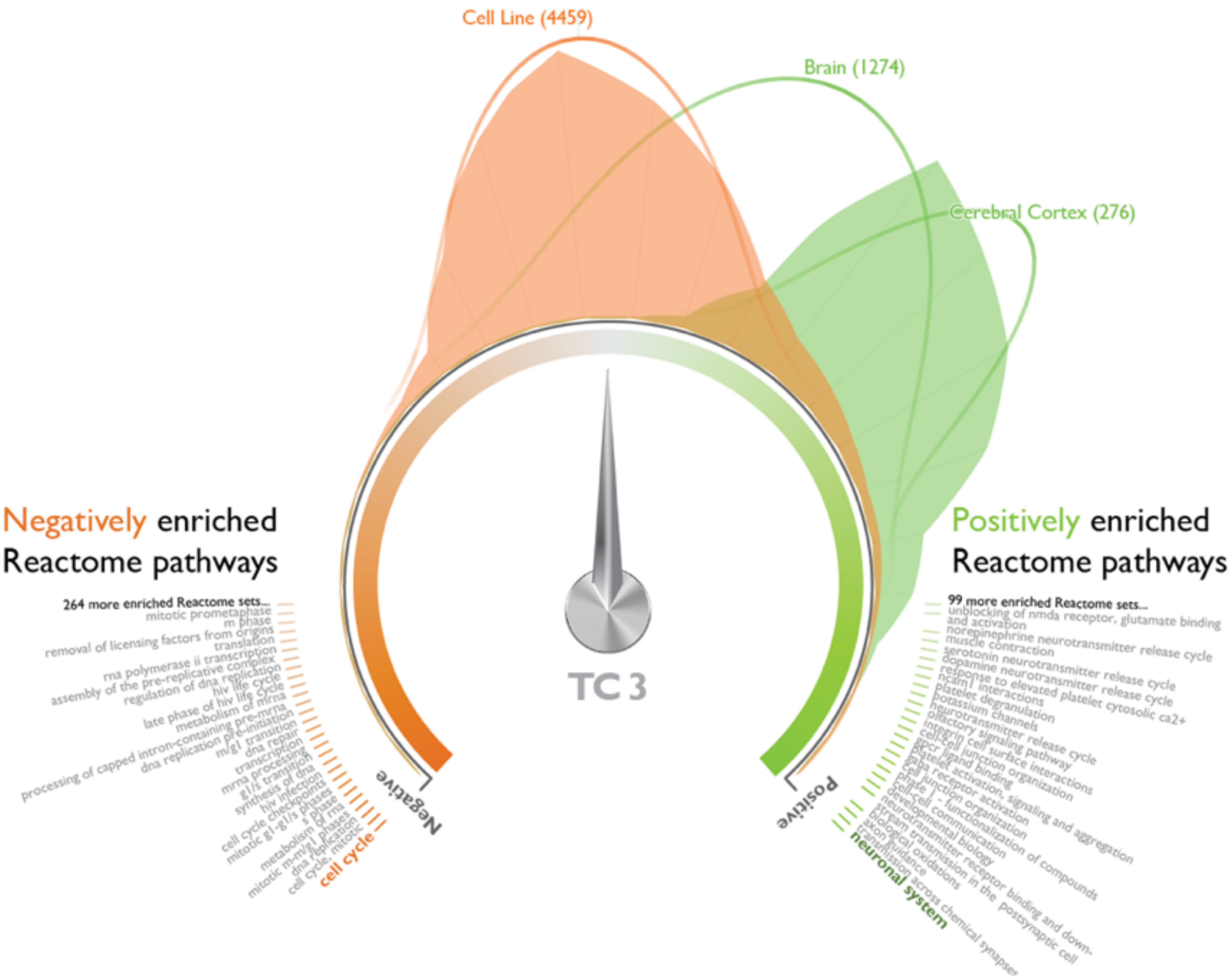


Fehrmann *et al*, Nature Genetics 2015

Component 1 and 2



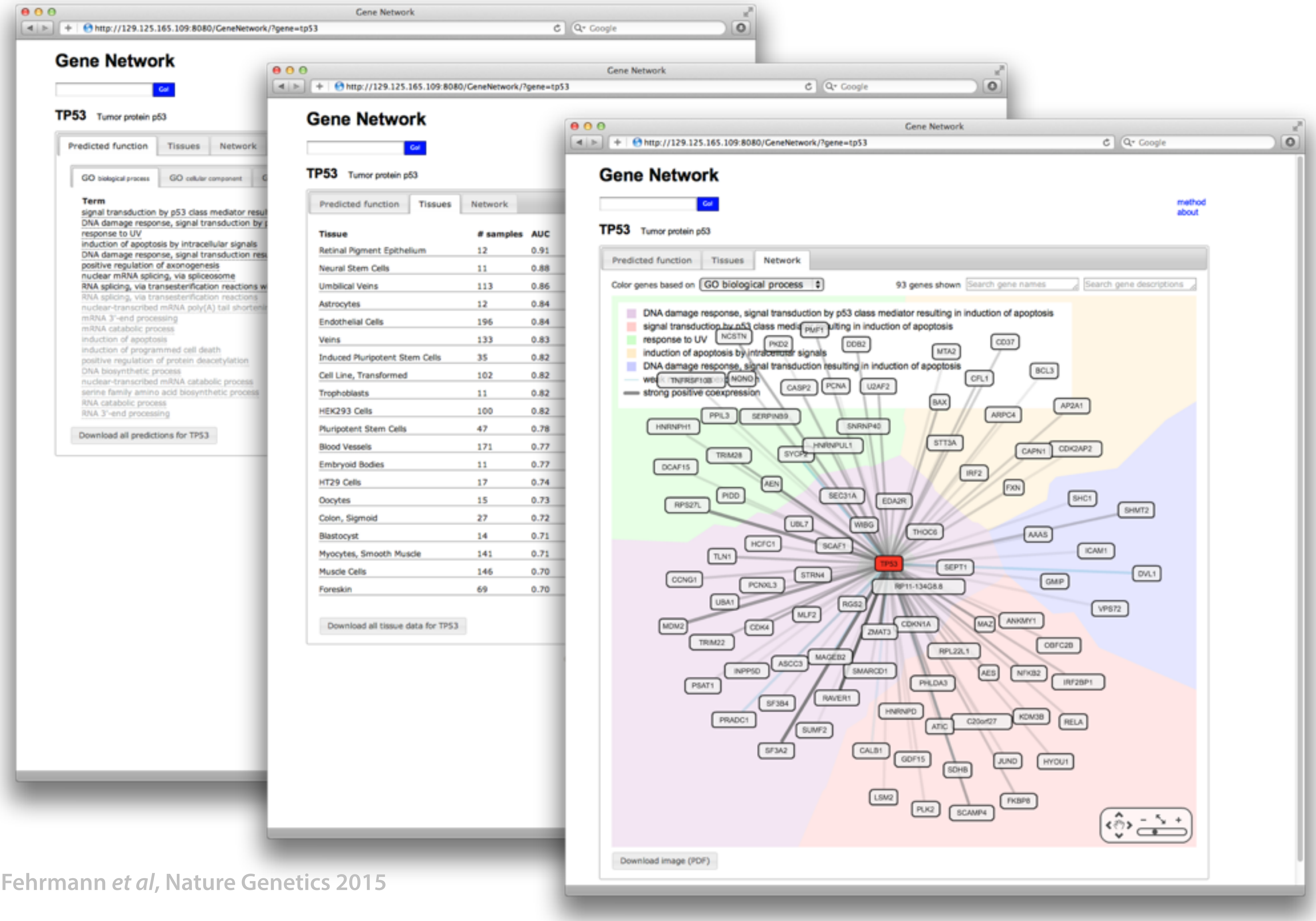
Transcriptional component 3



Negatively enriched Reactome pathways

Positively enriched Reactome pathways

Predicted gene functions: www.genenetwork.nl



Gene Network

http://129.125.165.109:8080/GeneNetwork/?gene=tp53

Gene Network

TP53 Tumor protein p53

Predicted function | Tissues | Network

GO biological process | GO cellular component

Term

signal transduction by p53 class mediator resul
DNA damage response, signal transduction by p
response to UV
induction of apoptosis by intracellular signals
DNA damage response, signal transduction res
positive regulation of axonogenesis
nuclear mRNA splicing, via spliceosome
RNA splicing, via transesterification reactions w
RNA splicing, via transesterification reactions
nuclear-transcribed mRNA poly(A) tail shorteni
mRNA 3'-end processing
mRNA catabolic process
induction of apoptosis
induction of programmed cell death
positive regulation of protein deacetylation
DNA biosynthetic process
nuclear-transcribed mRNA catabolic process
serine family amino acid biosynthetic process
RNA catabolic process
RNA 3'-end processing

Download all predictions for TP53

Gene Network

http://129.125.165.109:8080/GeneNetwork/?gene=tp53

Gene Network

TP53 Tumor protein p53

Predicted function | Tissues | Network

Tissue	# samples	AUC
Retinal Pigment Epithelium	12	0.91
Neural Stem Cells	11	0.88
Umbilical Veins	113	0.86
Astrocytes	12	0.84
Endothelial Cells	196	0.84
Veins	133	0.83
Induced Pluripotent Stem Cells	35	0.82
Cell Line, Transformed	102	0.82
Trophoblasts	11	0.82
HEK293 Cells	100	0.82
Pluripotent Stem Cells	47	0.78
Blood Vessels	171	0.77
Embryoid Bodies	11	0.77
HT29 Cells	17	0.74
Oocytes	15	0.73
Colon, Sigmoid	27	0.72
Blastocyst	14	0.71
Myocytes, Smooth Muscle	141	0.71
Muscle Cells	146	0.70
Foreskin	69	0.70

Download all tissue data for TP53

Gene Network

http://129.125.165.109:8080/GeneNetwork/?gene=tp53

Gene Network

method
about

TP53 Tumor protein p53

Predicted function | Tissues | Network

Color genes based on: GO biological process

93 genes shown

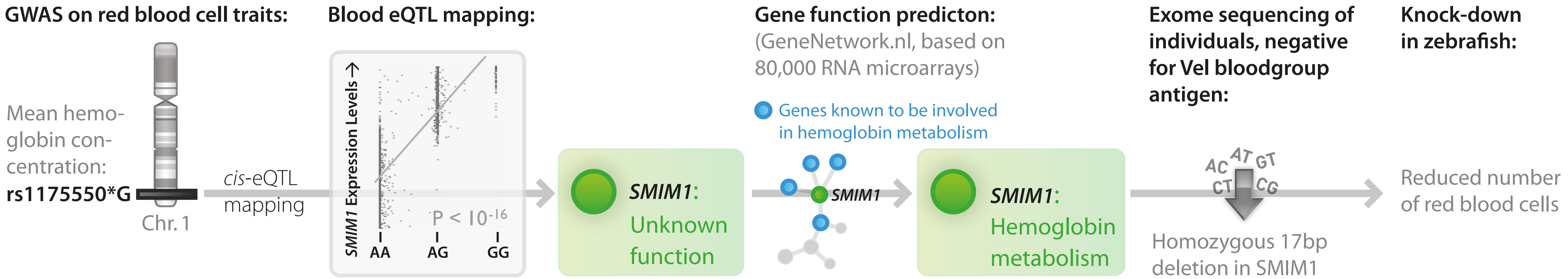
Search gene names | Search gene descriptions

- DNA damage response, signal transduction by p53 class mediator resulting in induction of apoptosis
- signal transduction by p53 class mediator resulting in induction of apoptosis
- response to UV
- induction of apoptosis by intracellular signals
- DNA damage response, signal transduction resulting in induction of apoptosis
- weak positive coexpression
- strong positive coexpression

Download image (PDF)

Fehrmann *et al*, Nature Genetics 2015

GeneNetwork gene function predictions



GWAS on red blood cell traits:

Mean hemoglobin concentration: **rs1175550*G**

Chr. 1

cis-eQTL mapping

Blood eQTL mapping:

SMIM1: Unknown function

Gene function predicton: (GeneNetwork.nl, based on 80,000 RNA microarrays)

● Genes known to be involved in hemoglobin metabolism

SMIM1: Hemoglobin metabolism

Exome sequencing of individuals, negative for Vel bloodgroup antigen:

Homozygous 17bp deletion in SMIM1

Knock-down in zebrafish:

Reduced number of red blood cells

Van der Harst *et al*, Nature 2012

Cvejic *et al*, Nature Genetics 2013

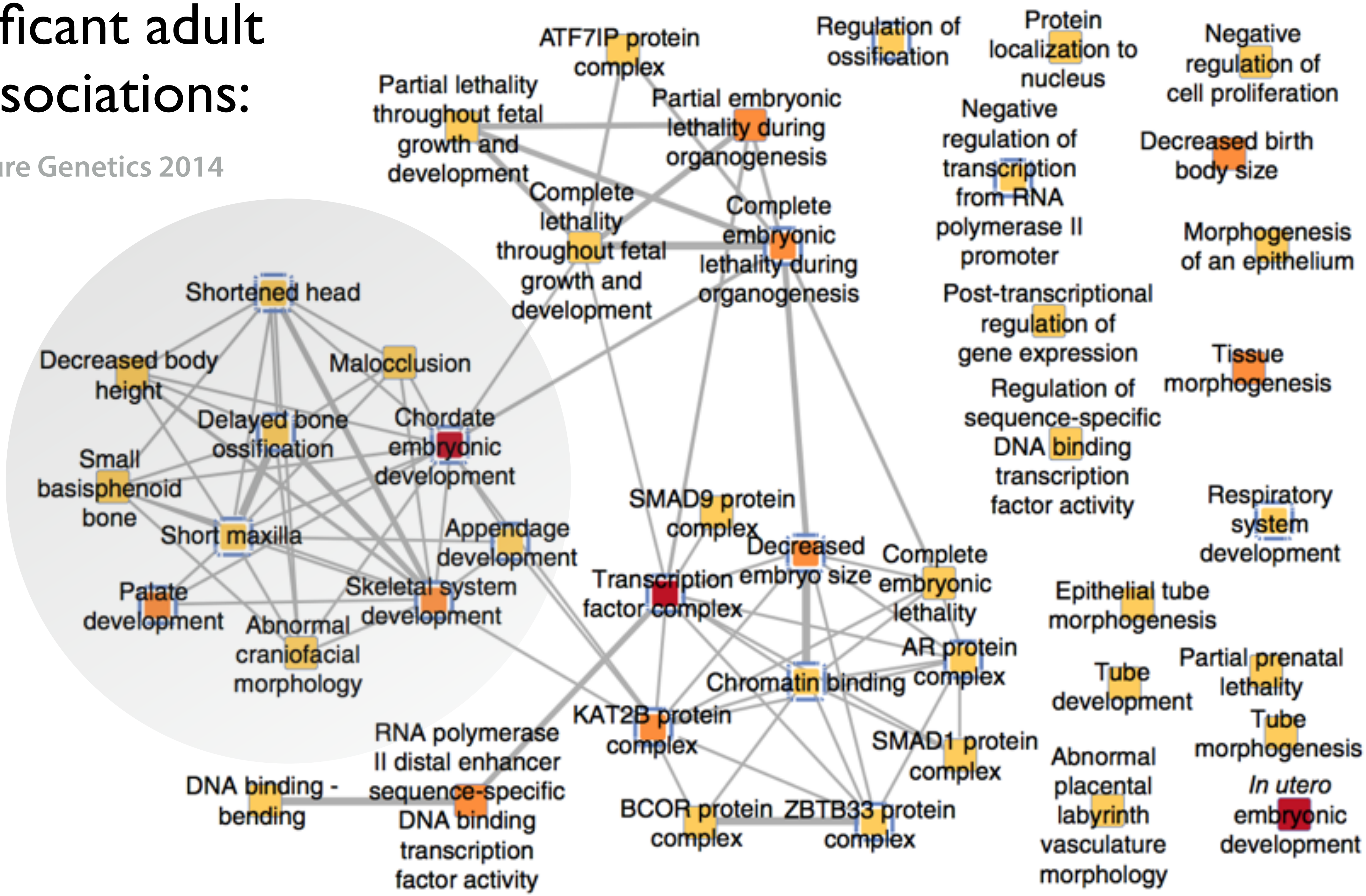
Amounts of data integrated:

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

DEPICT: New prioritisation algorithm for GWAS

697 significant adult height associations:

Wood *et al*, Nature Genetics 2014



ATF7IP protein complex
Regulation of ossification
Protein localization to nucleus
Negative regulation of cell proliferation
Partial lethality throughout fetal growth and development
Partial embryonic lethality during organogenesis
Negative regulation of transcription from RNA polymerase II promoter
Decreased birth body size
Complete lethality throughout fetal growth and development
Complete embryonic lethality during organogenesis
Morphogenesis of an epithelium
Post-transcriptional regulation of gene expression
Shortened head
Decreased body height
Malocclusion
Tissue morphogenesis
Regulation of sequence-specific DNA binding transcription factor activity
Delayed bone ossification
Chordate embryonic development
Small basisphenoid bone
Respiratory system development
SMAD9 protein complex
Short maxilla
Appendage development
Decreased embryo size
Complete embryonic lethality
Transcription factor complex
Epithelial tube morphogenesis
Palate development
Skeletal system development
Abnormal craniofacial morphology
AR protein complex
Partial prenatal lethality
Tube development
Chromatin binding complex
Tube morphogenesis
KAT2B protein complex
RNA polymerase II distal enhancer sequence-specific DNA binding transcription factor activity
SMAD1 protein complex
Abnormal placental labyrinth vasculature morphology
In utero embryonic development
DNA binding - bending
BCOR protein complex
ZBTB33 protein complex

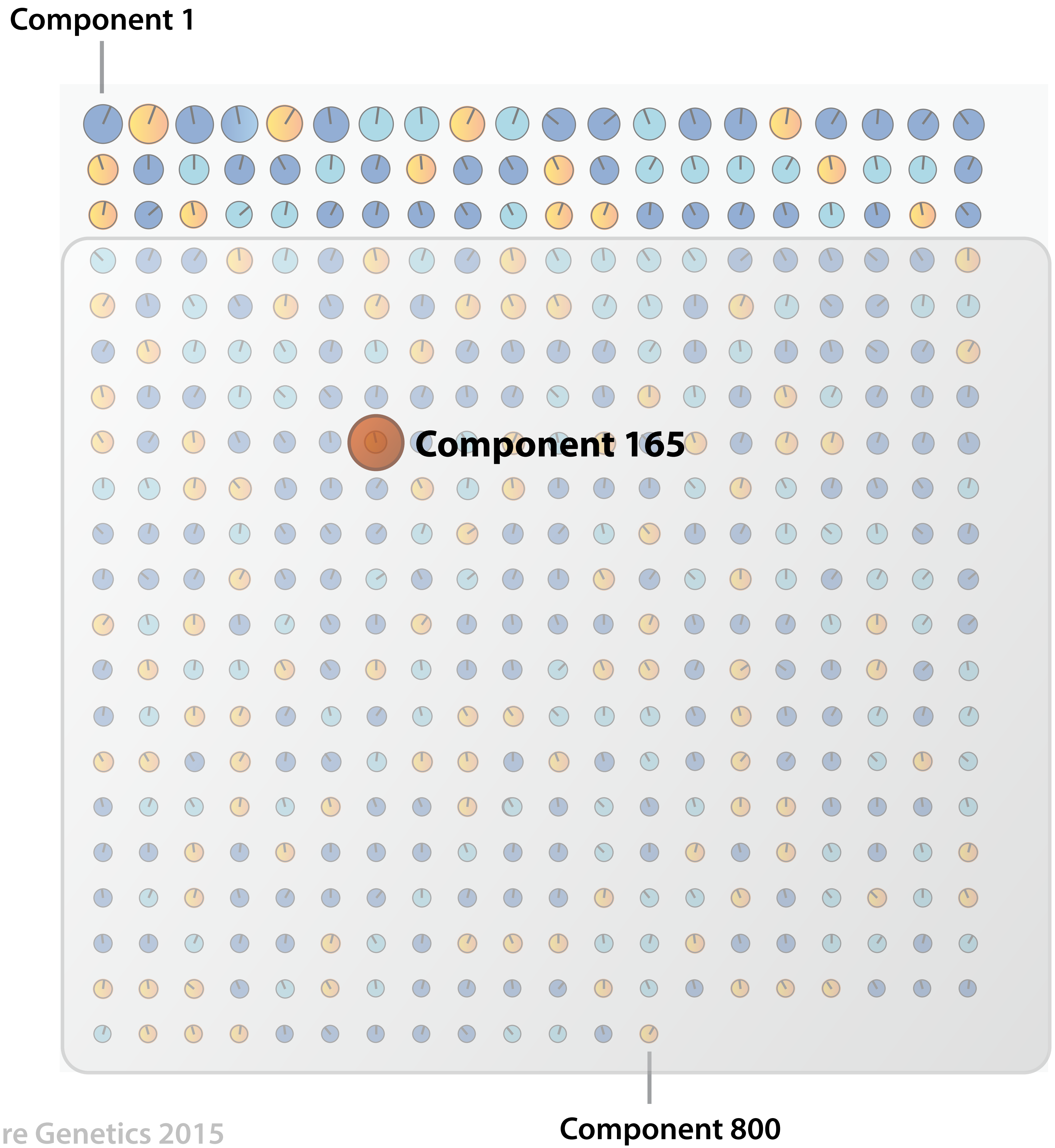
DEPICT Method:

Pers *et al*, Nature Communications 2015

DEPICT used for:

Body mass index (Locke *et al*, Nature 2015)
Waist hip ratio (Shungin *et al*, Nature 2015)
Hypospadias (Geller *et al*, Nature Genetics 2014)
Lipid Levels (Surakka, Nature Genetics 2015)

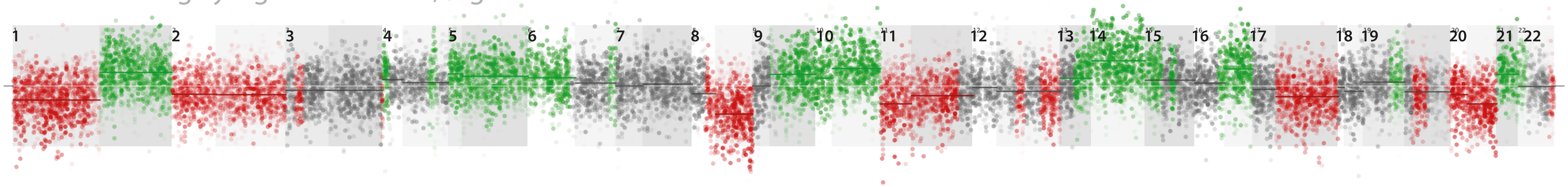
Components 51 - 800

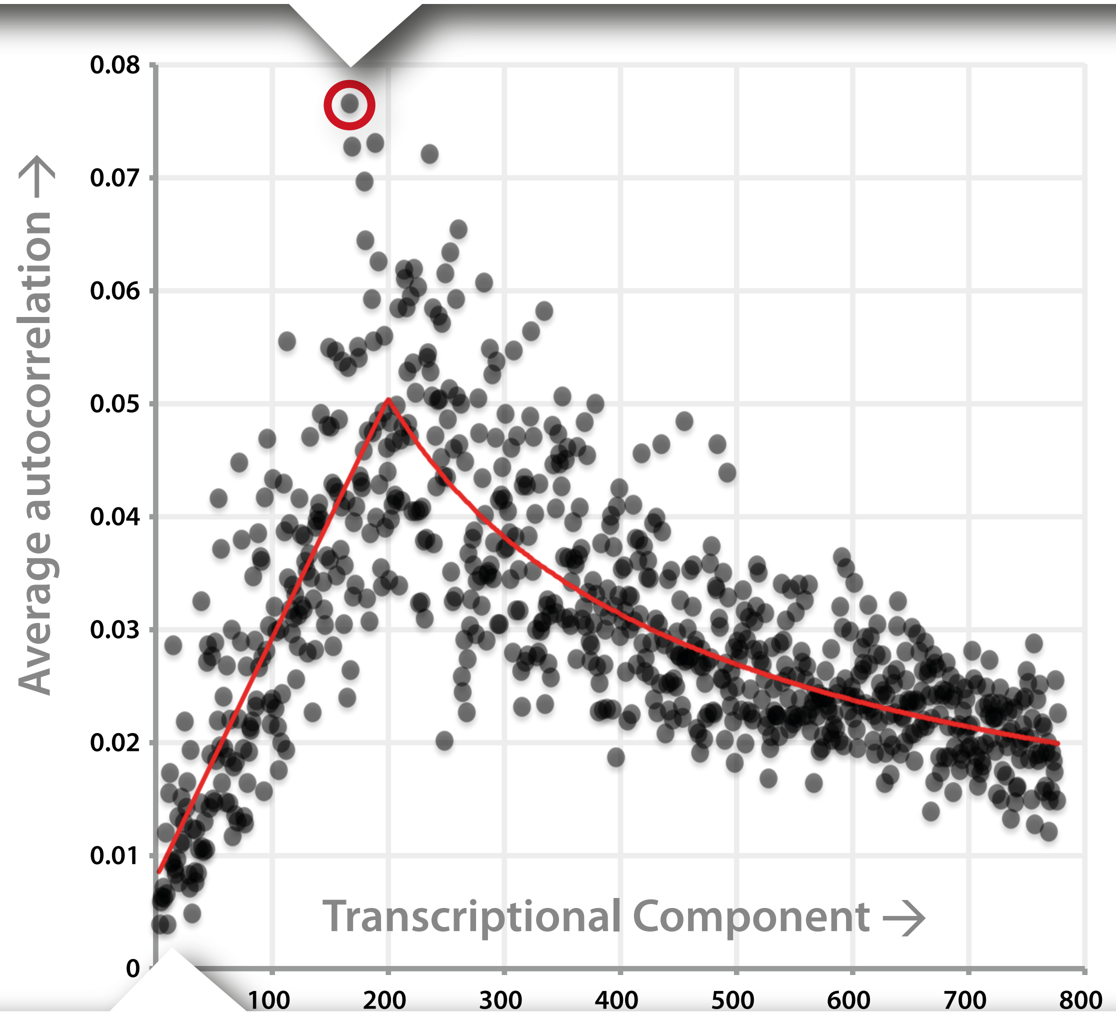


Fehrmann *et al*, Nature Genetics 2015

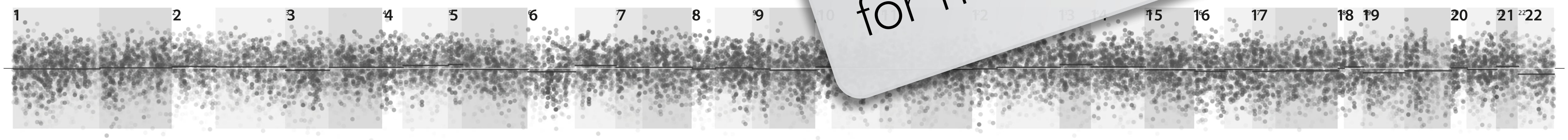
Some component show weird behaviour

TC 165: Strong cytogenetic effects, high autocorrelation



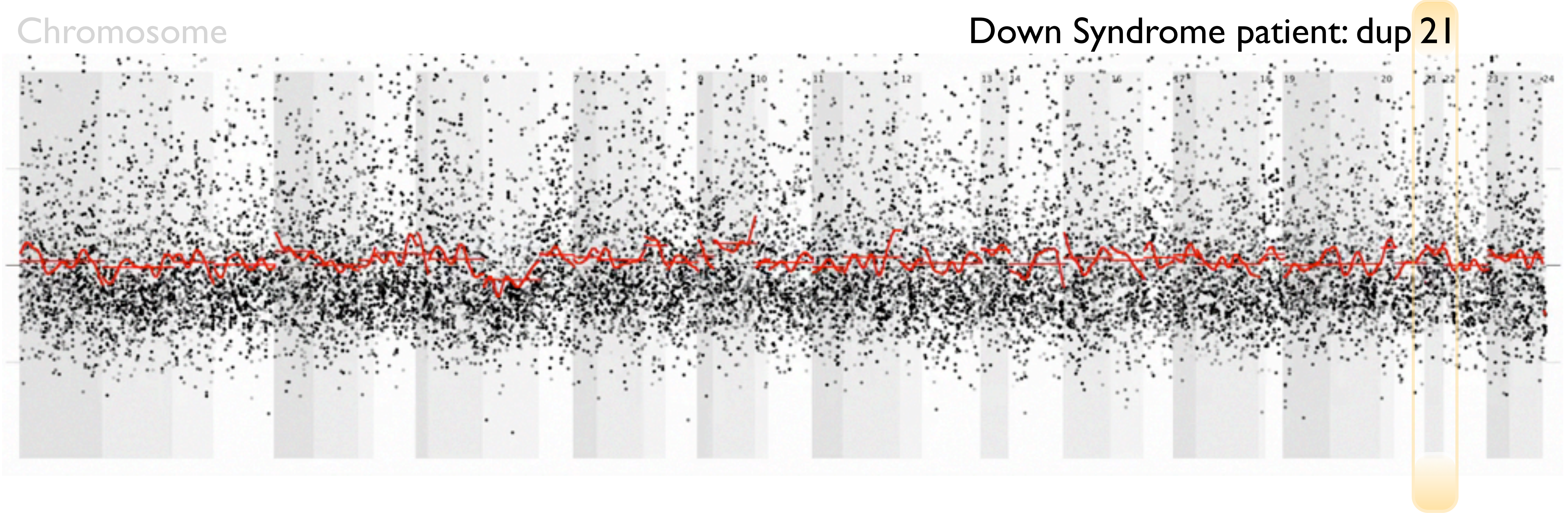


TC 1: No cytogenetic effect, zero autocorrelation



Redo analysis in healthy samples, correct cancer data for healthy components

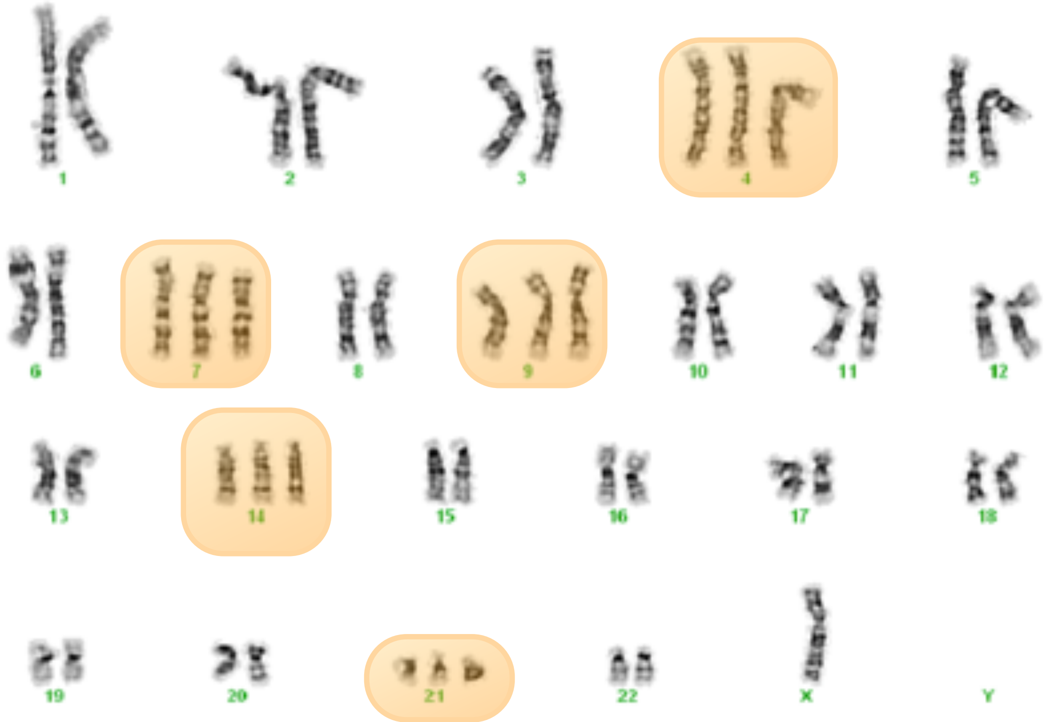
Detection cytogenetic aberration in expression data

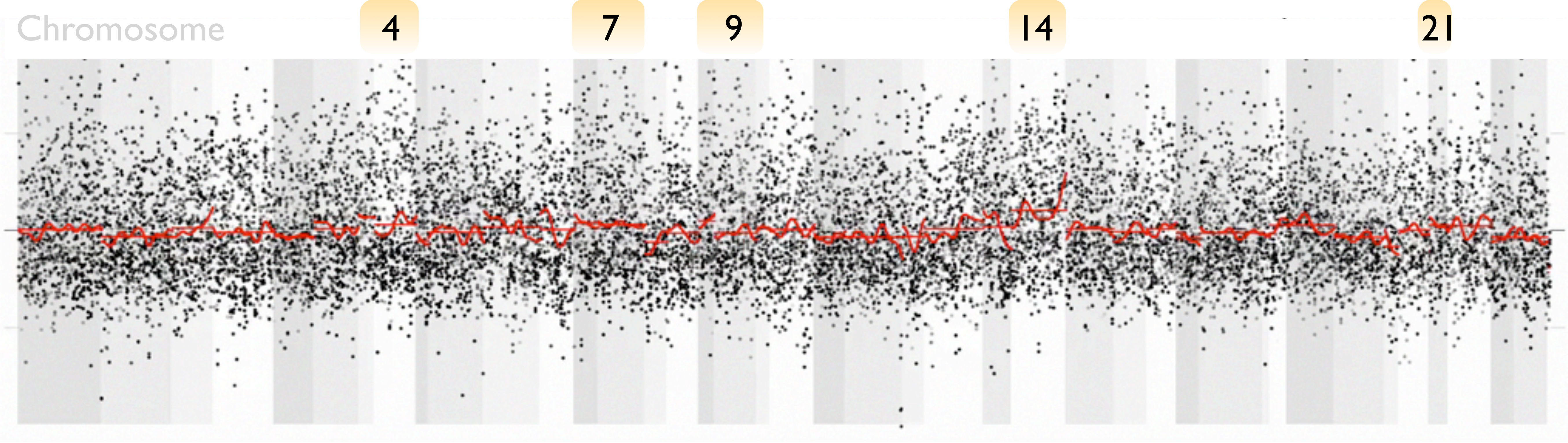


Down Syndrome patient: dup 21

Identifying five chromosome duplications

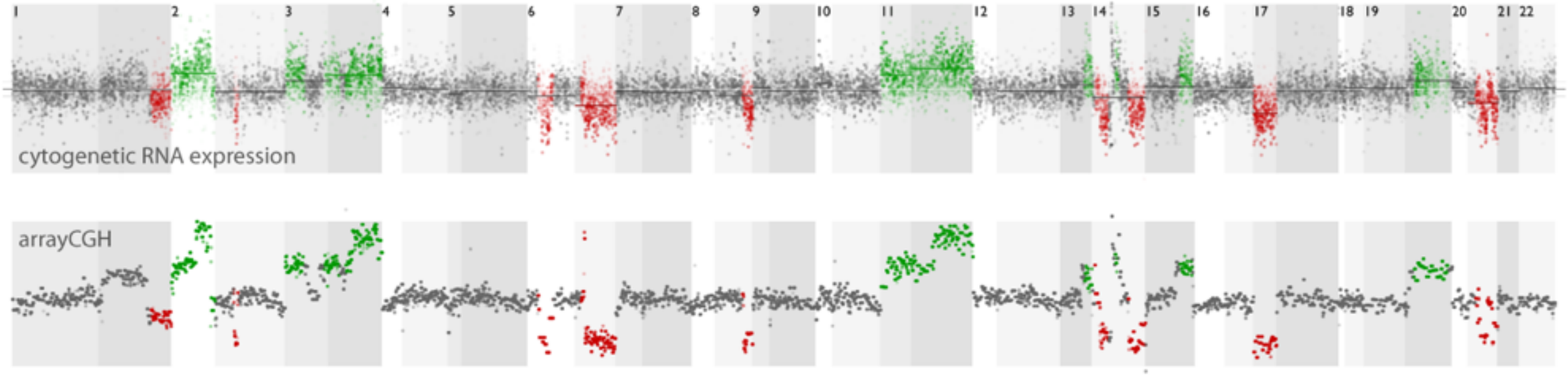
Karyogram
HapMap LCL



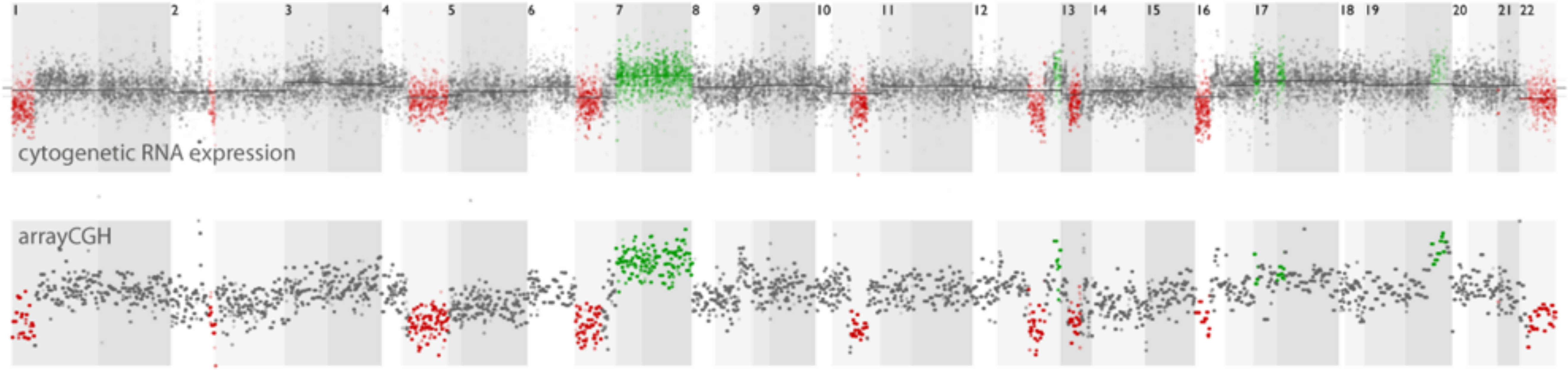


Comparison of arrayCGH and cytogenetic RNA profiles

GSM274996



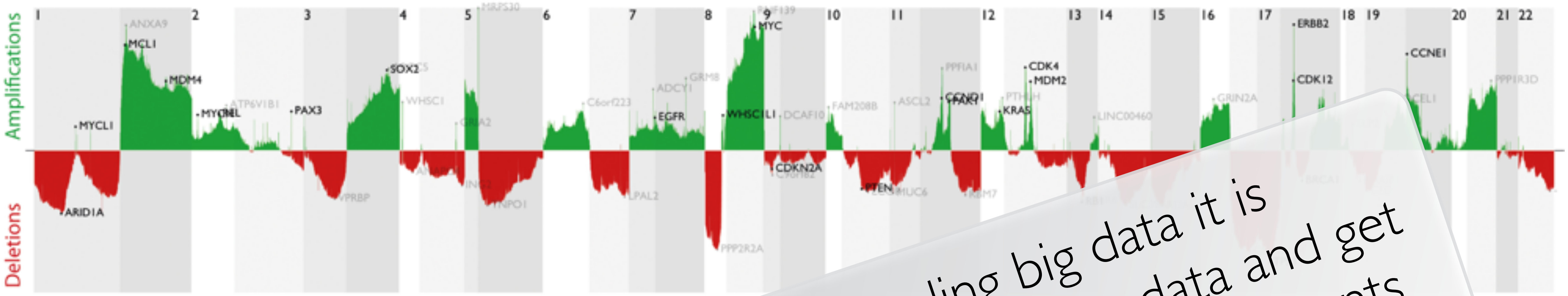
GSM275008



Fehrmann *et al*, Nature Genetics 2015

Known driver genes in amplification and deletion peaks

Average somatic copy number aberration profile of 16,172 primary tumor samples (GPL570 + GPL96 platforms)

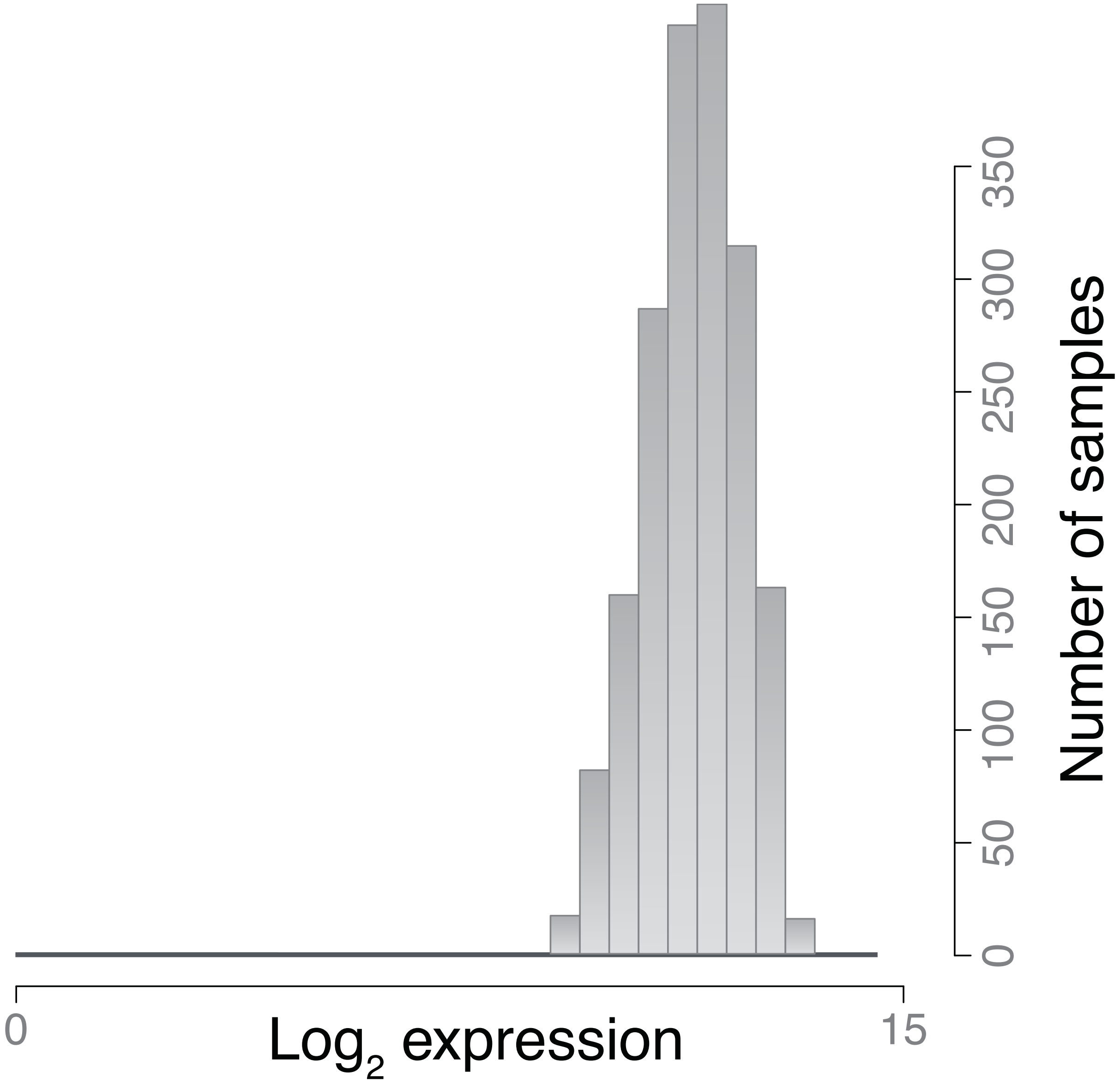


By recycling big data it is possible to clean data and get very accurate measurements

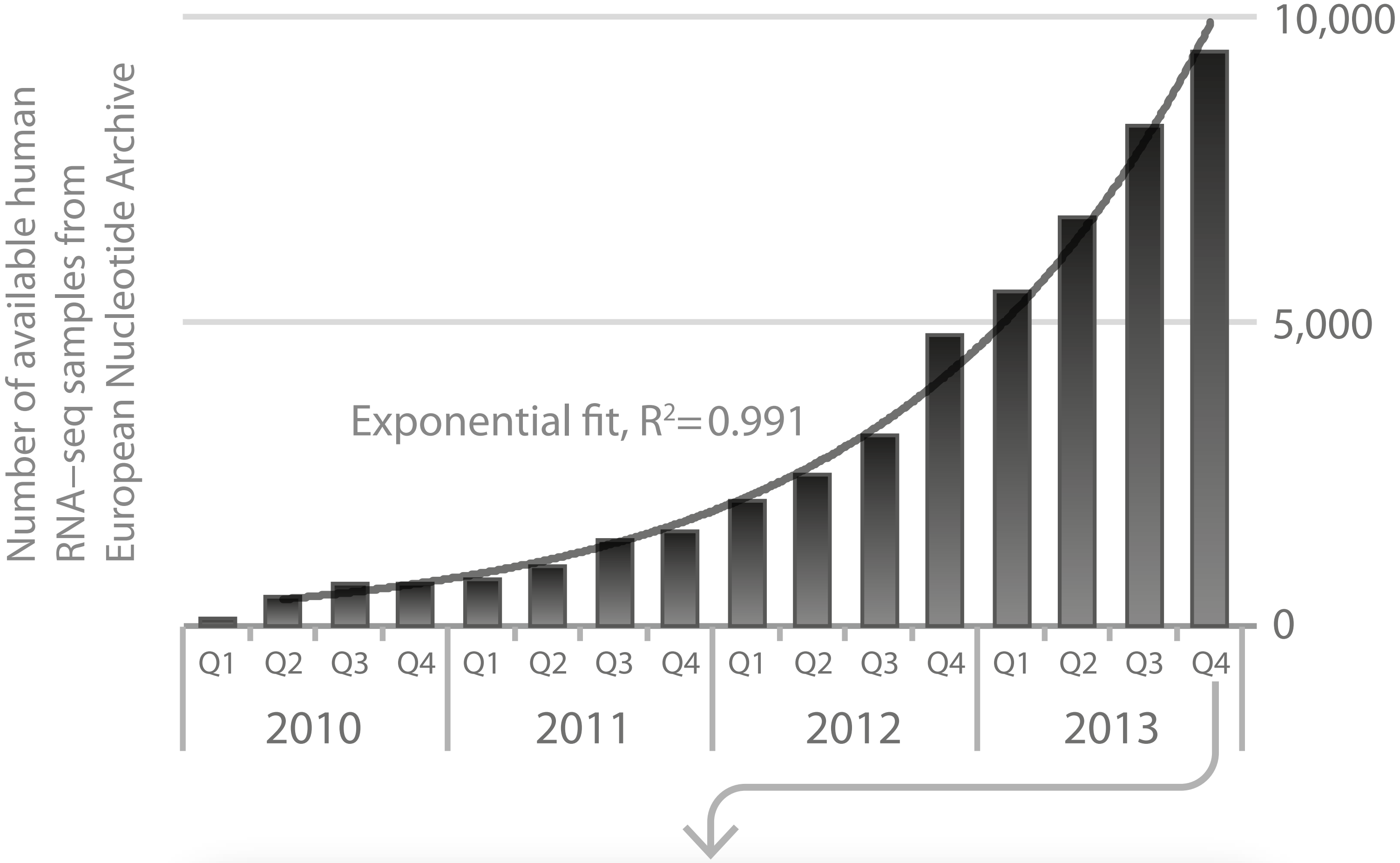
Fehrmann *et al*, Nature Genetics 2015

More accurate reference values for genes

***TRIM51BP* gene expression distribution in the Dutch population**



Explosion of publicly available RNA-seq data



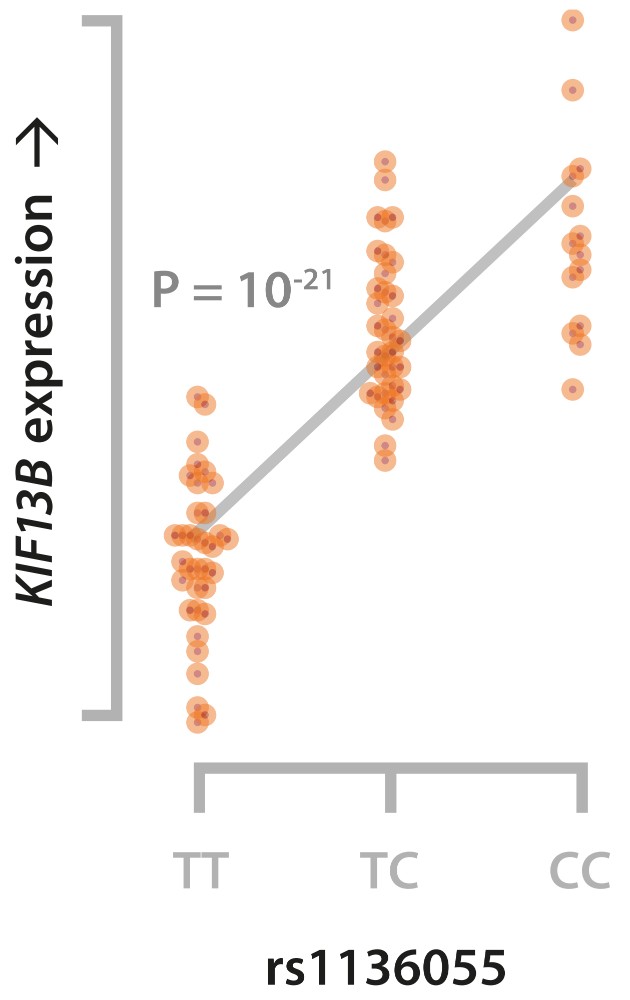
9,527 public human RNA-seq runs from ENA

Read alignment, epression quantification, normalization and PCA:

- 4,028 runs with low mapping statistics removed
- 521 expression outliers removed

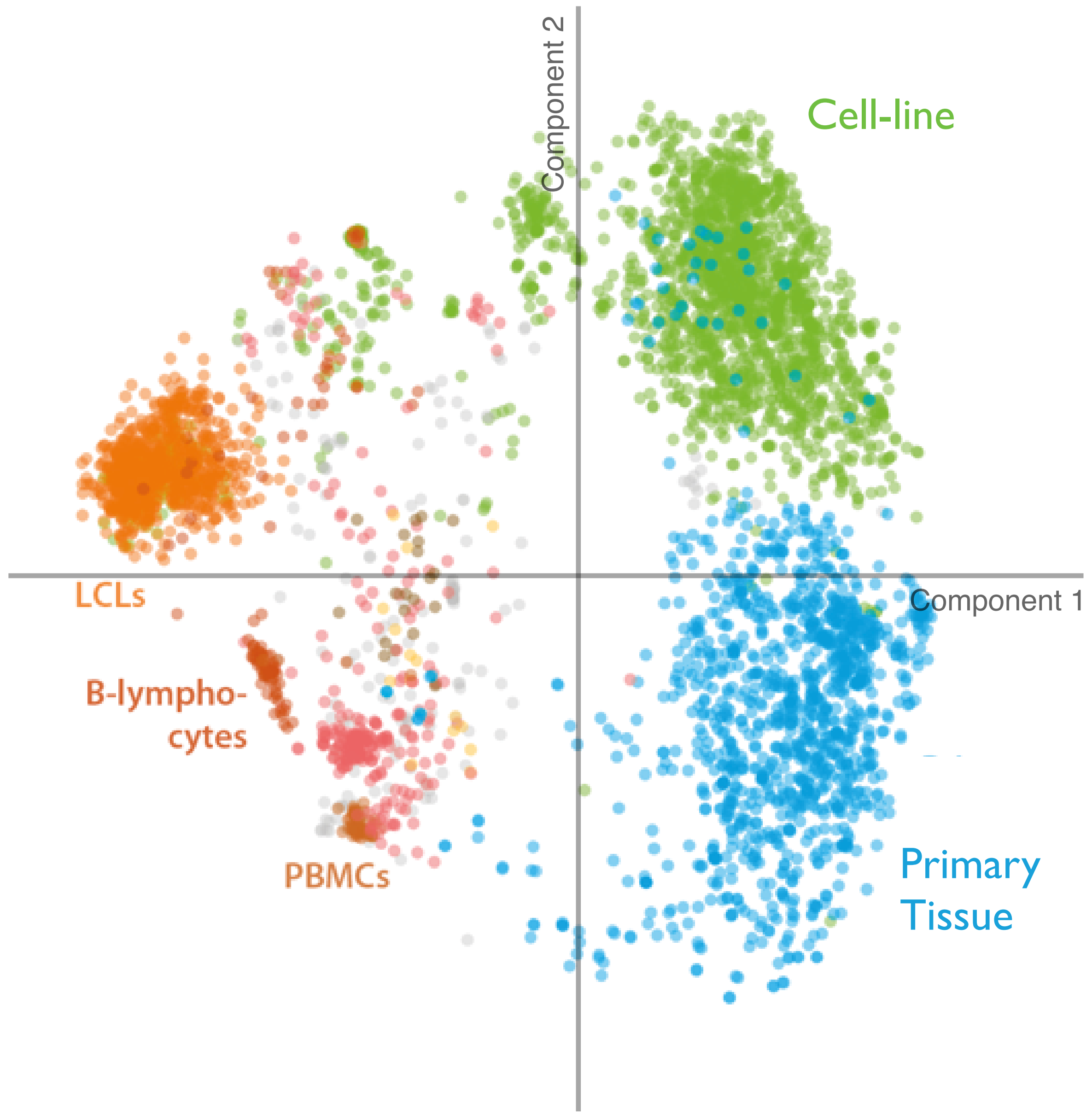
4,978 samples (used for expression clustering)

Derive SNP genotypes from RNA-seq data



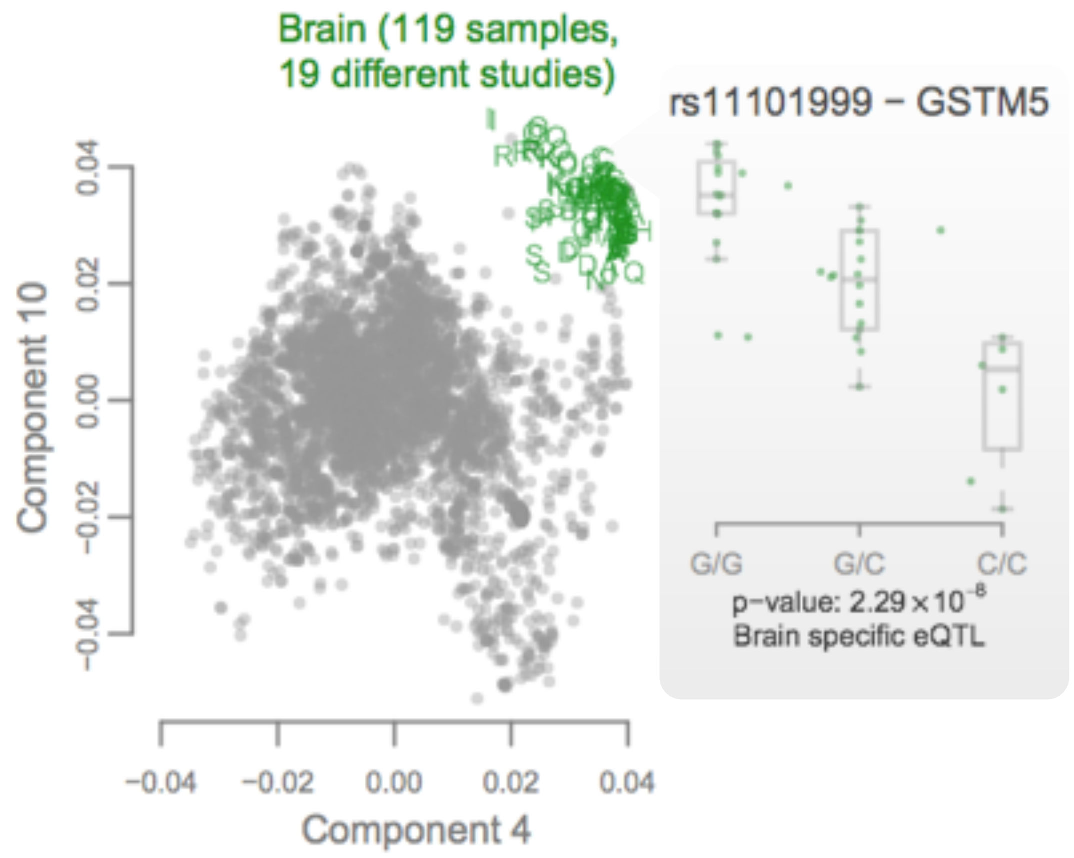
Infer genotypes

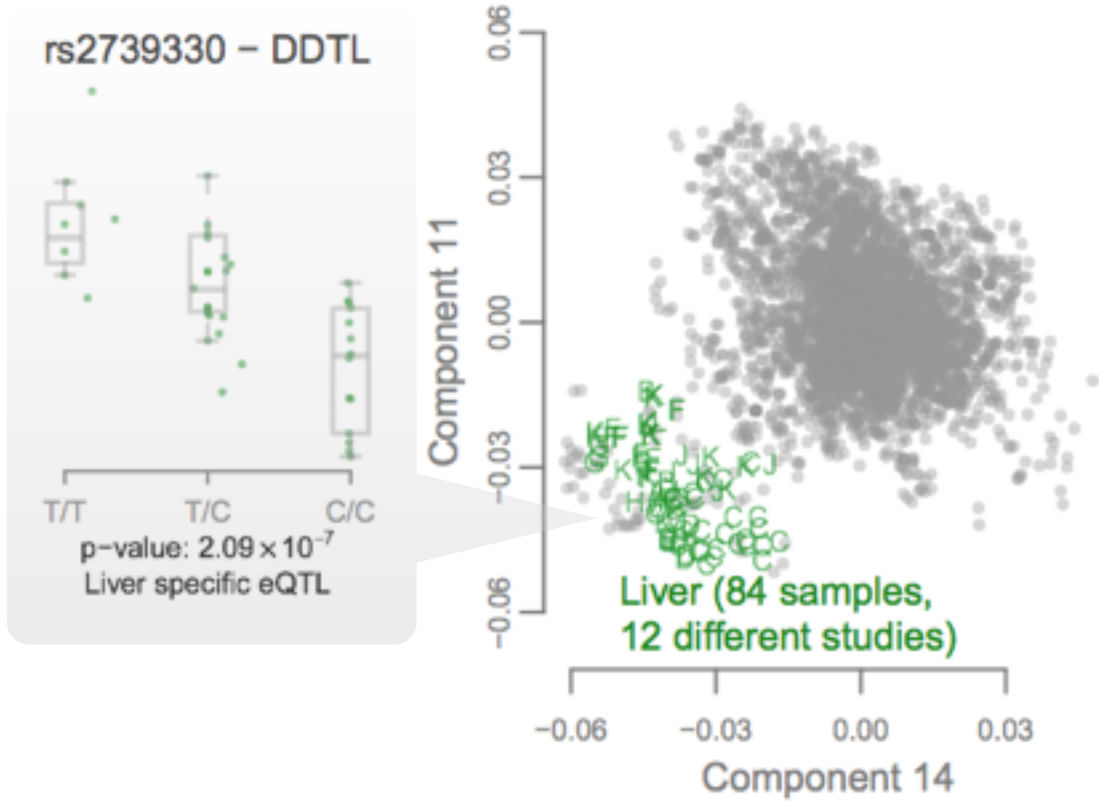
Public RNA-seq data (5,000 samples)

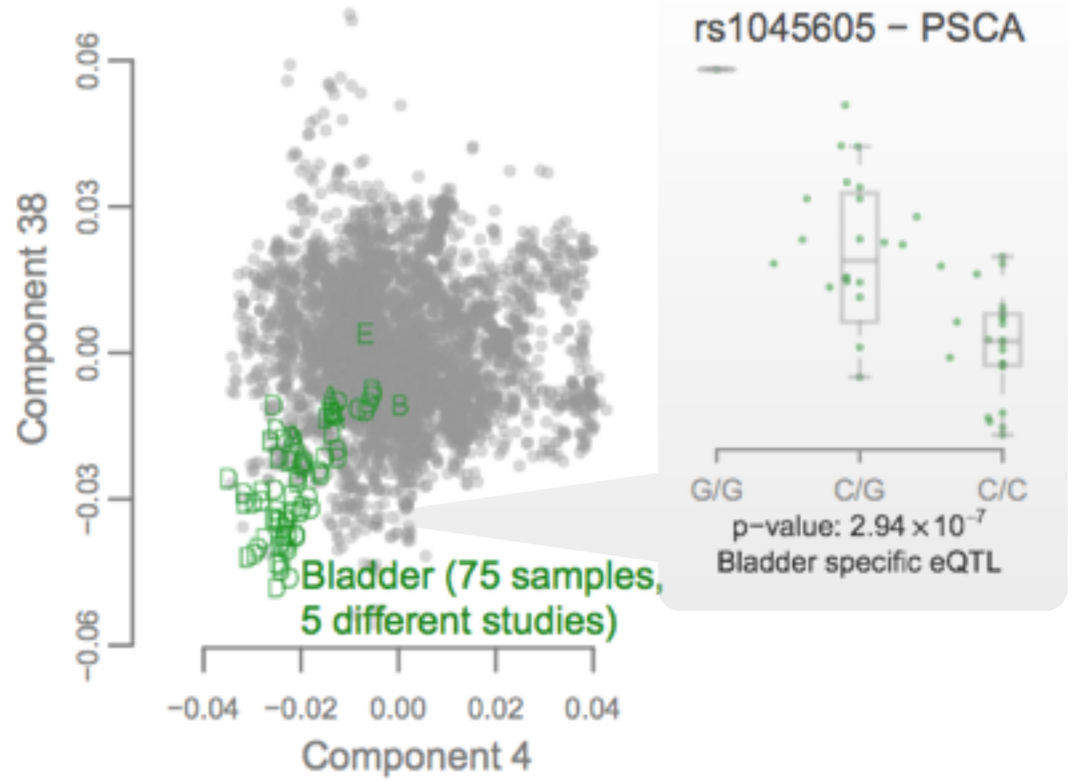


Deelen *et al*, Genome Medicine 2015

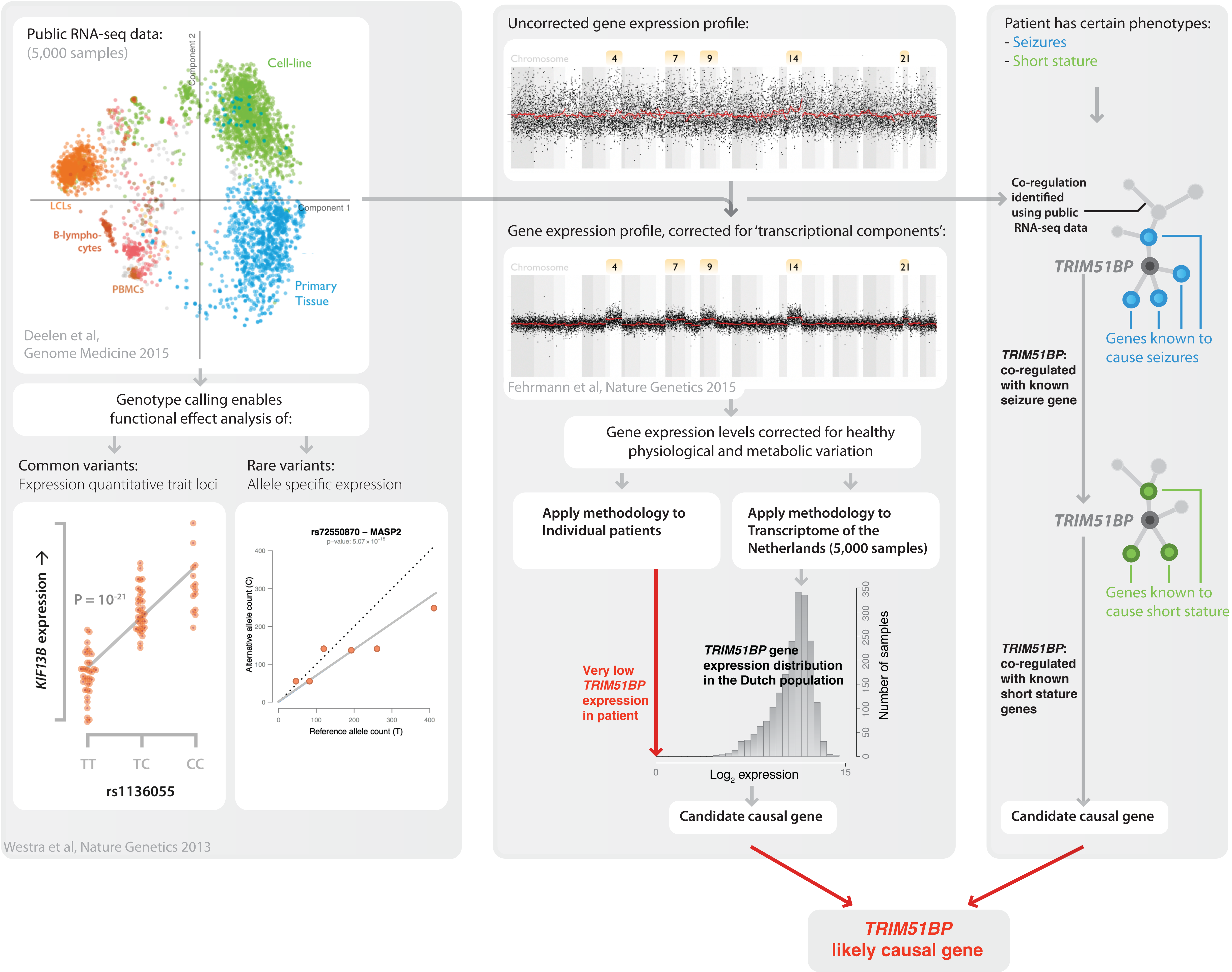
Tissue-specific eQTL mapping for free







Deelen *et al*, Genome Medicine 2015



Public RNA-seq data:
(5,000 samples)

Deelen et al, Genome Medicine 2015

Genotype calling enables functional effect analysis of:

Common variants:
Expression quantitative trait loci

Rare variants:
Allele specific expression

KIF13B expression →
P = 10^{-21}
TT
TC
CC
rs1136055
rs72550870 – MASP2
p-value: 5.07 × 10^{-15}
Alternative allele count (C)
Reference allele count (T)
0
100
200
300
400

Westra et al, Nature Genetics 2013

Uncorrected gene expression profile:

Gene expression profile, corrected for ‘transcriptional components’:

Fehrmann et al, Nature Genetics 2015

Gene expression levels corrected for healthy physiological and metabolic variation

Apply methodology to Individual patients

Apply methodology to Transcriptome of the Netherlands (5,000 samples)

Very low *TRIM51BP* expression in patient

***TRIM51BP* gene expression distribution in the Dutch population**

Candidate causal gene

Patient has certain phenotypes:
- Seizures
- Short stature

Co-regulation identified using public RNA-seq data

TRIM51BP
Genes known to cause seizures
TRIM51BP
Genes known to cause short stature

***TRIM51BP*: co-regulated with known seizure gene**

***TRIM51BP*: co-regulated with known short stature genes**

Candidate causal gene

***TRIM51BP* likely causal gene**

Acknowledgements >

UMC Groningen	Juha Karjalainen Dasha Zhernakova Patrick Deelen Marc Jan Bonder Sipko van Dam Morris Swertz	Freerk van Dijk Niek de Klein Urmo Vosa Annique Claringbould Rudolf Fehrmann Cisca Wijmenga
BBMRI-NL BIOS Consortium	Peter-Bram ’t Hoen	Bas Heijmans
eQTLGen Consortium	Tonu Esko	Tune Pers
Target Project CIT RUG	Haije Wind	

Funding >

erc

SEVENTH FRAMEWORK PROGRAMME

NWO

BBMRI•NL