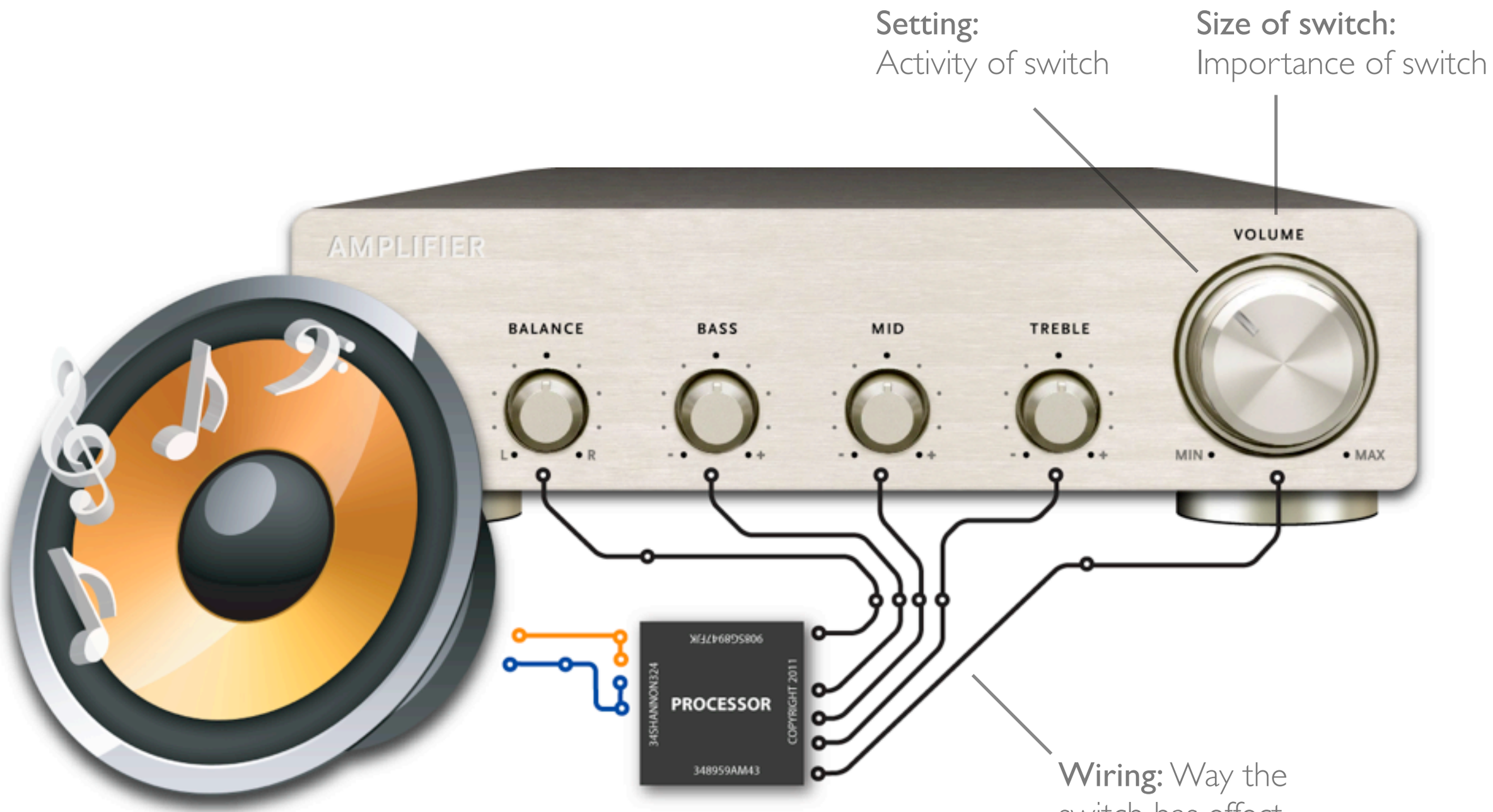
umcg

Department of genetics

Lude Franke > Integrating different omics data

Department of Genetics, UMC Groningen

Amplifier can change many aspects of music

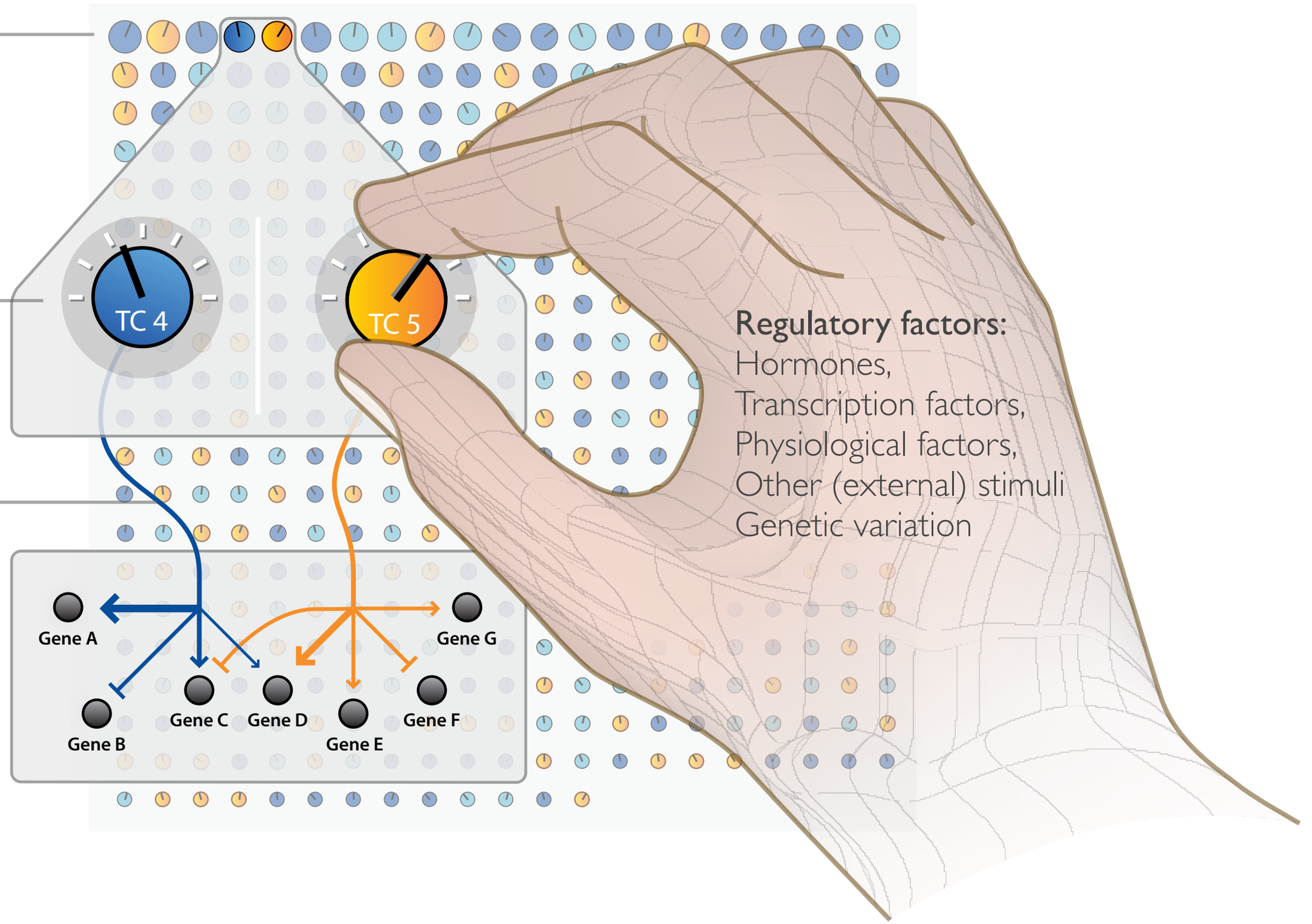


A control panel that determines gene expression?

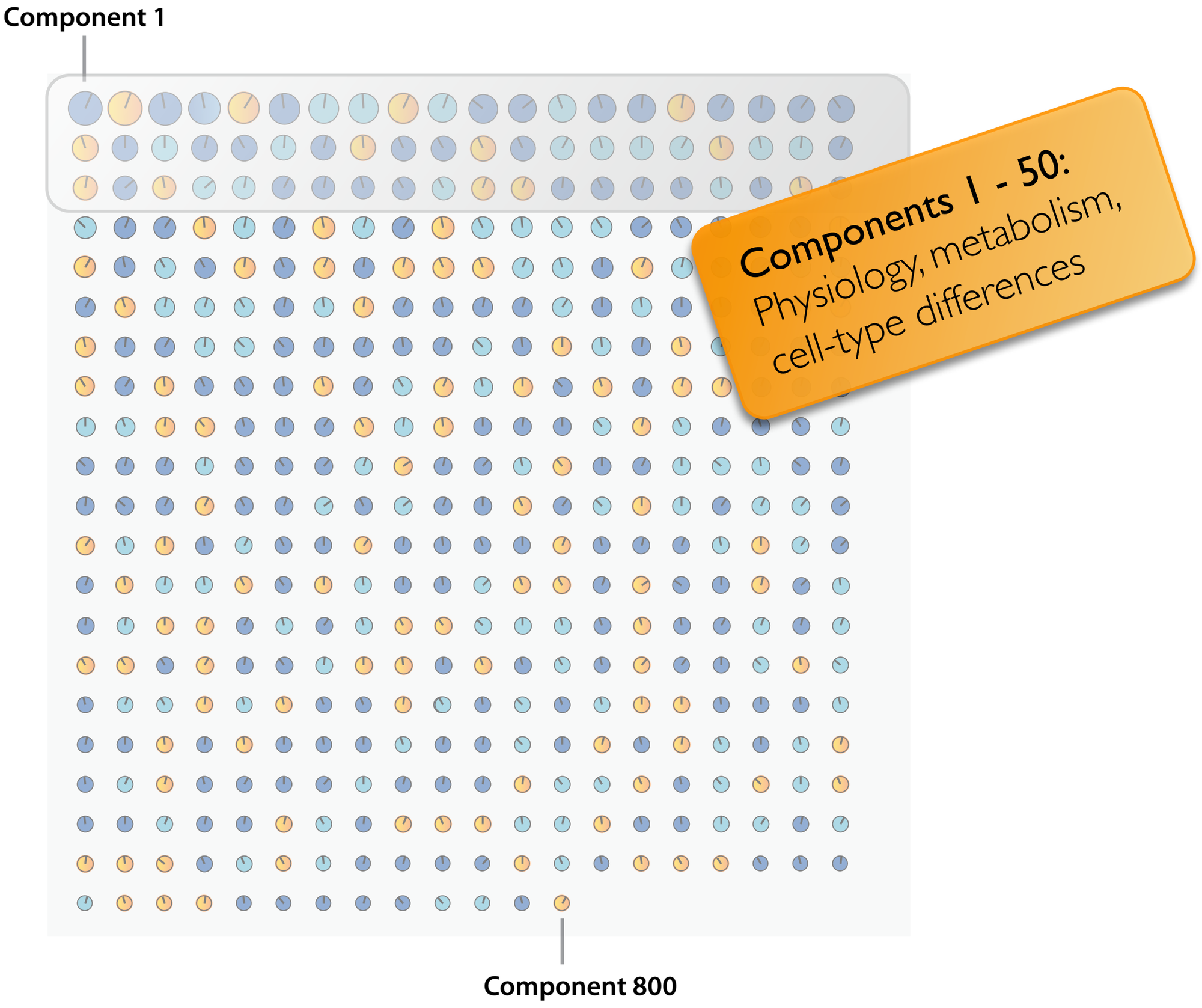
Size of switch:
Importance

Setting: State of a certain sample

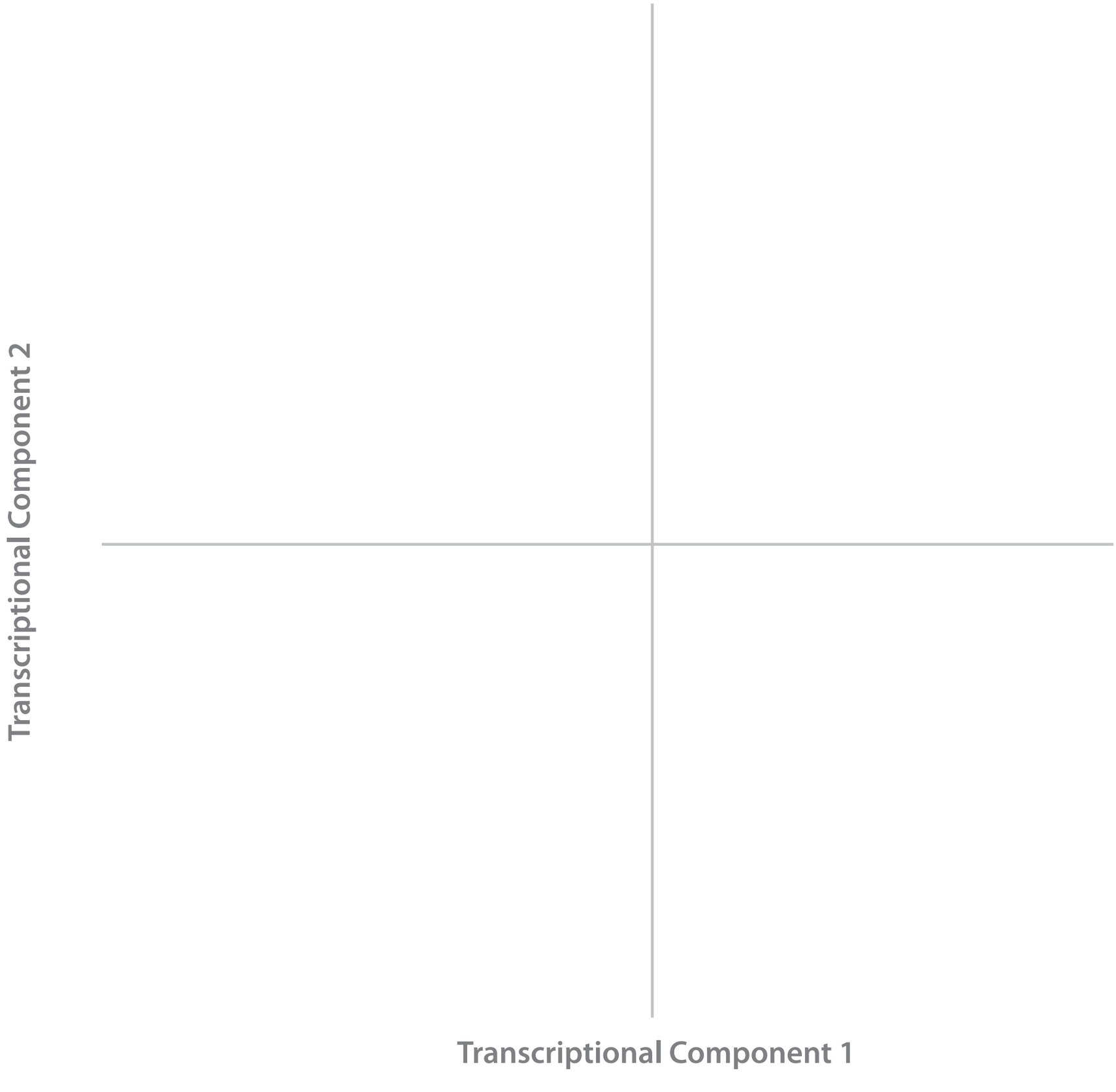
Wiring: Effect on individual genes



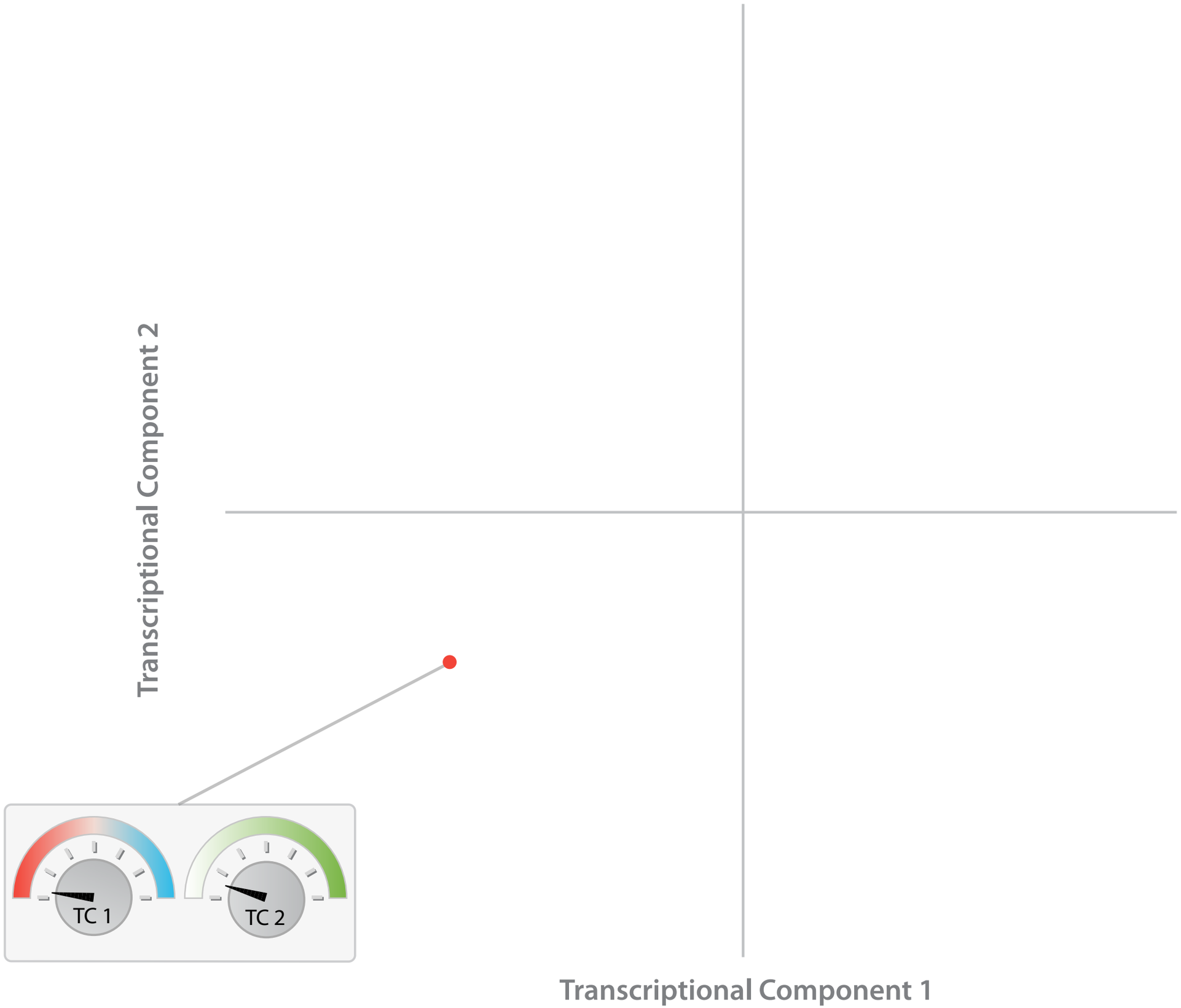
800 ‘transcriptional components’: Component 1 - 50



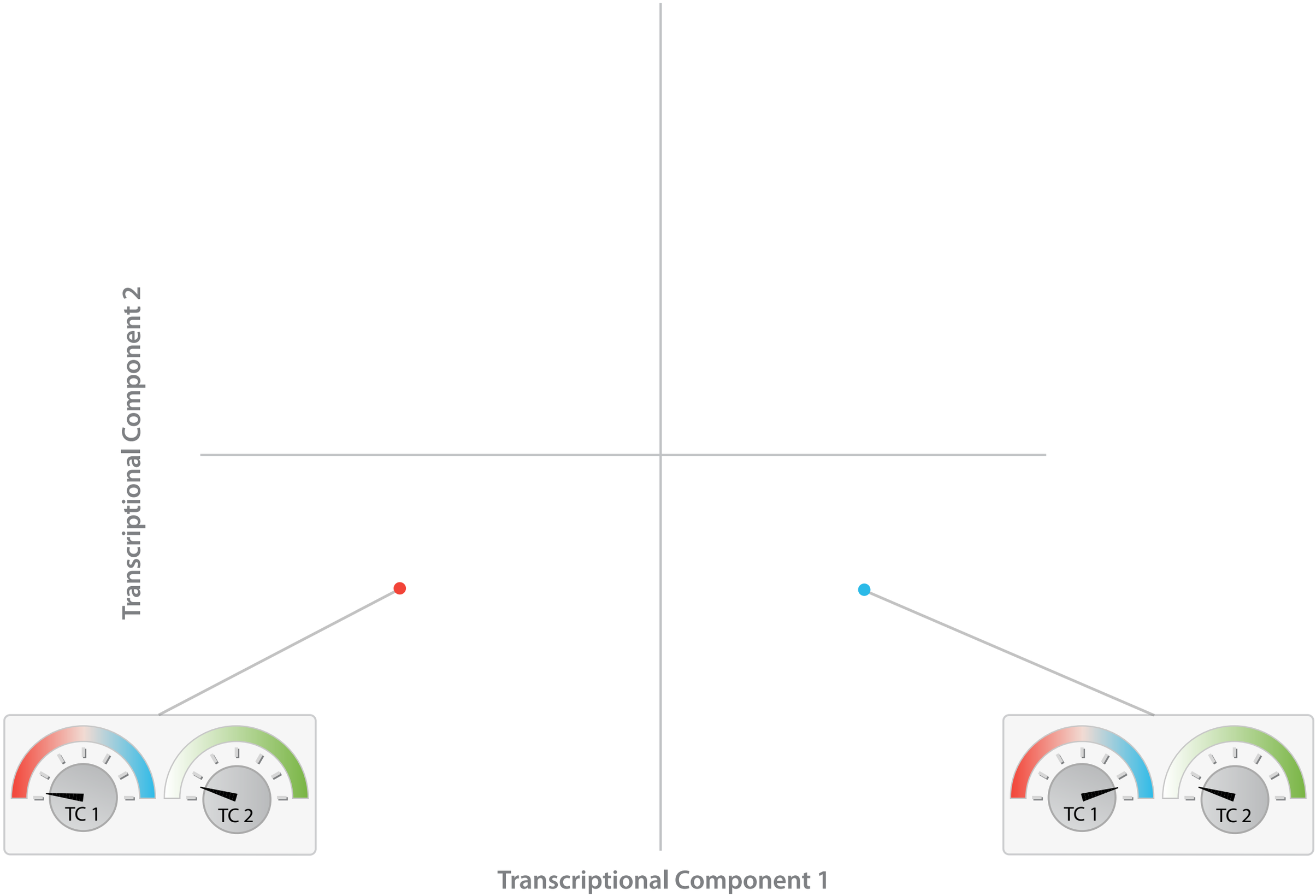
Component 1 and 2



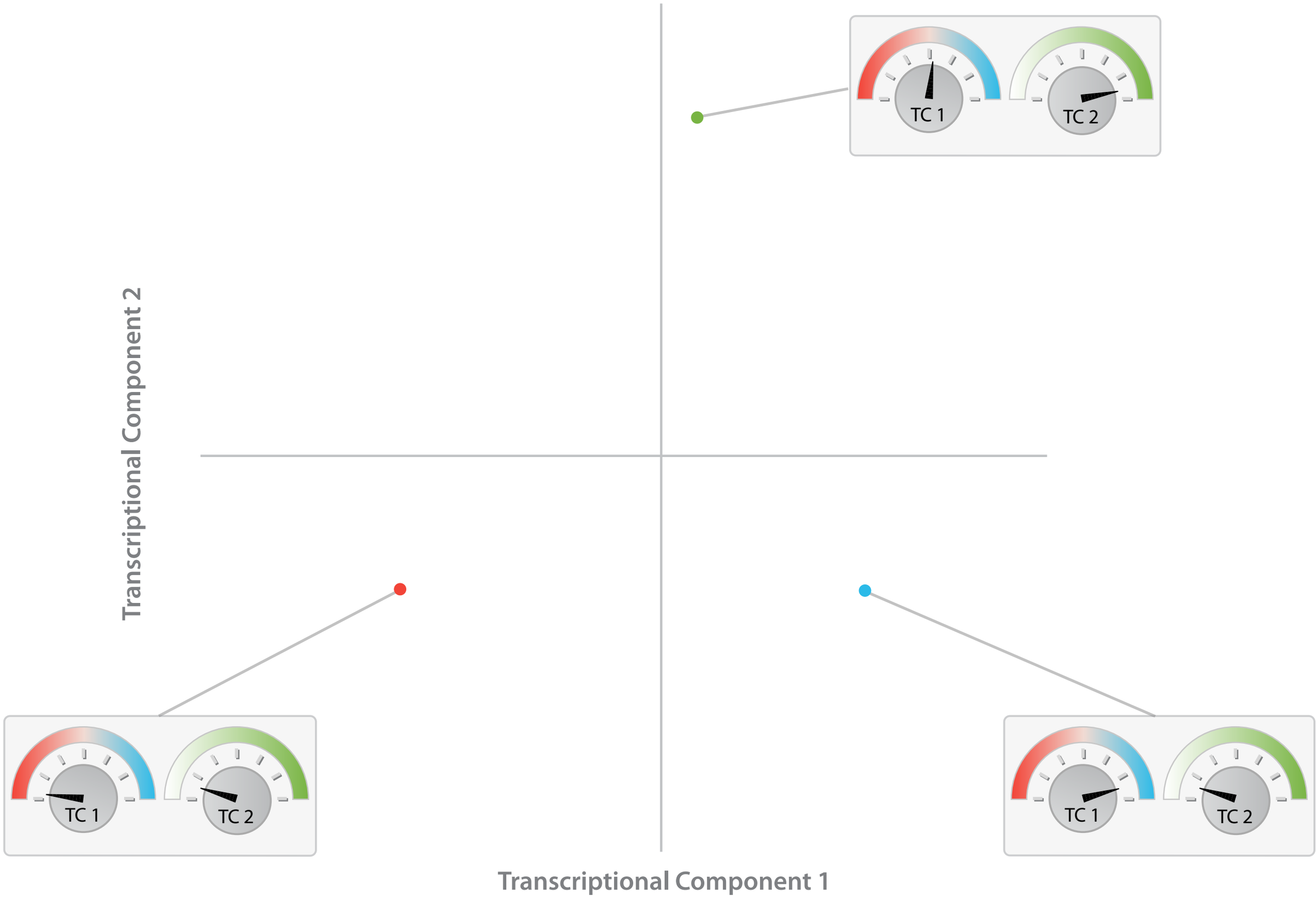
Component 1 and 2



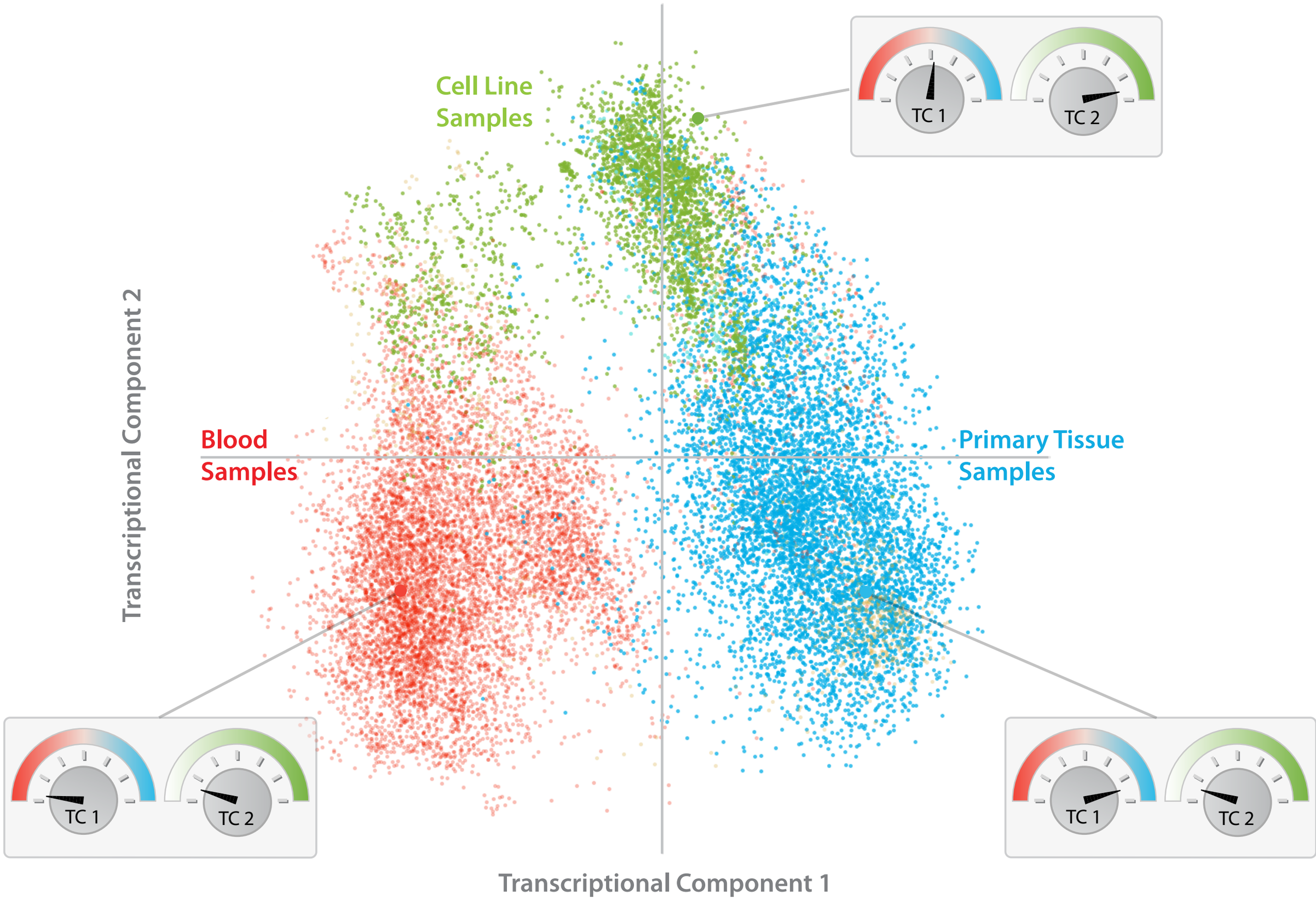
Component 1 and 2



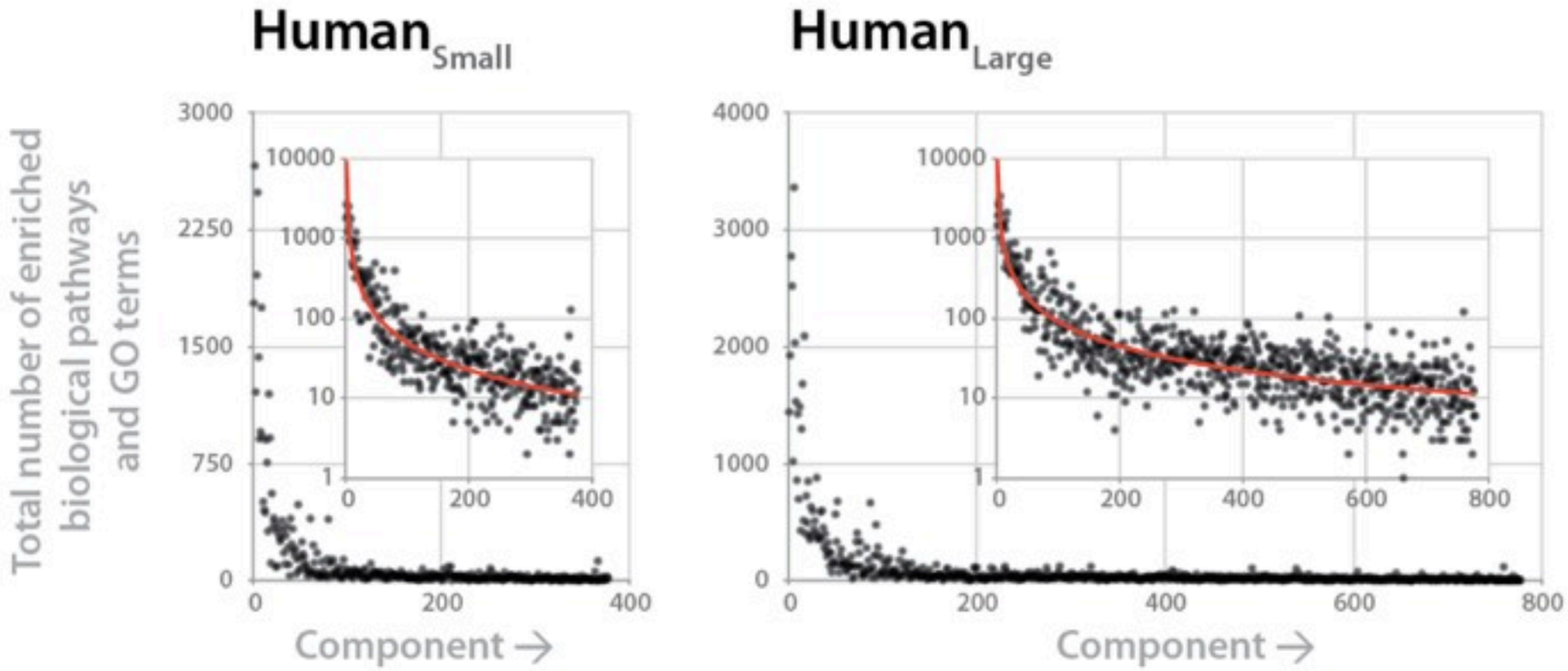
Component 1 and 2

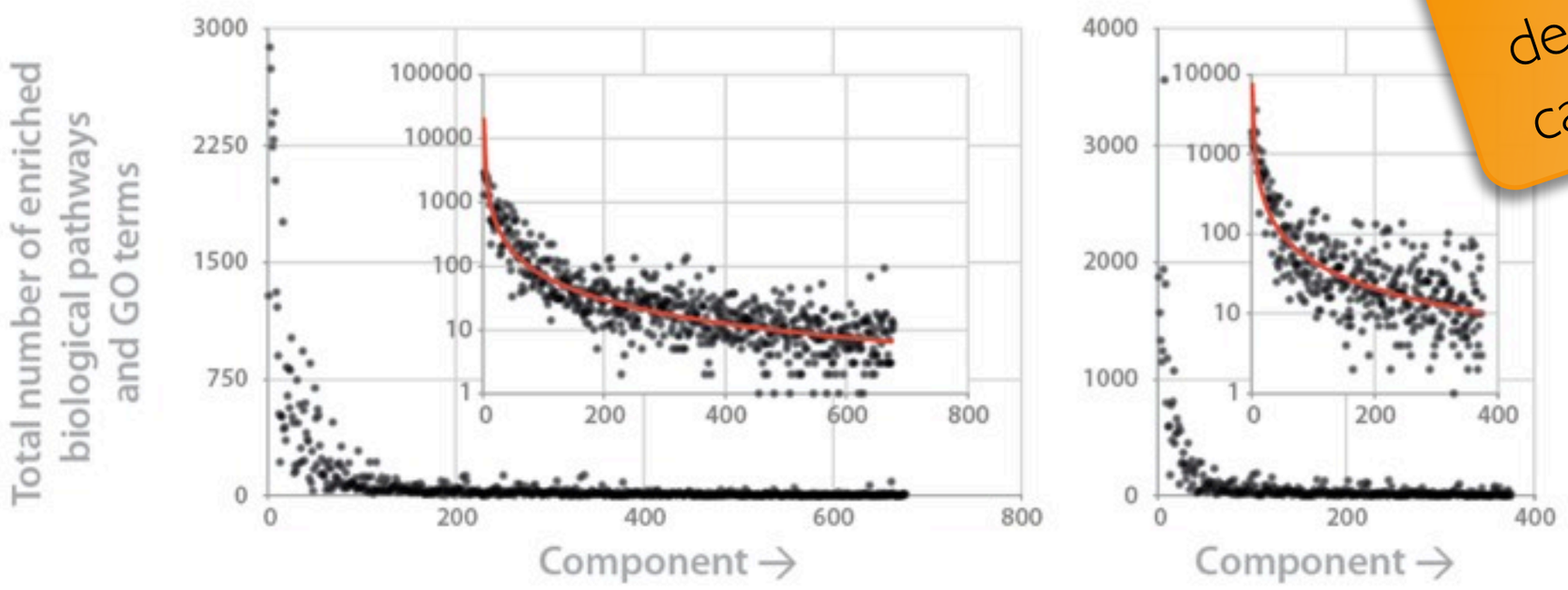


Component 1 and 2



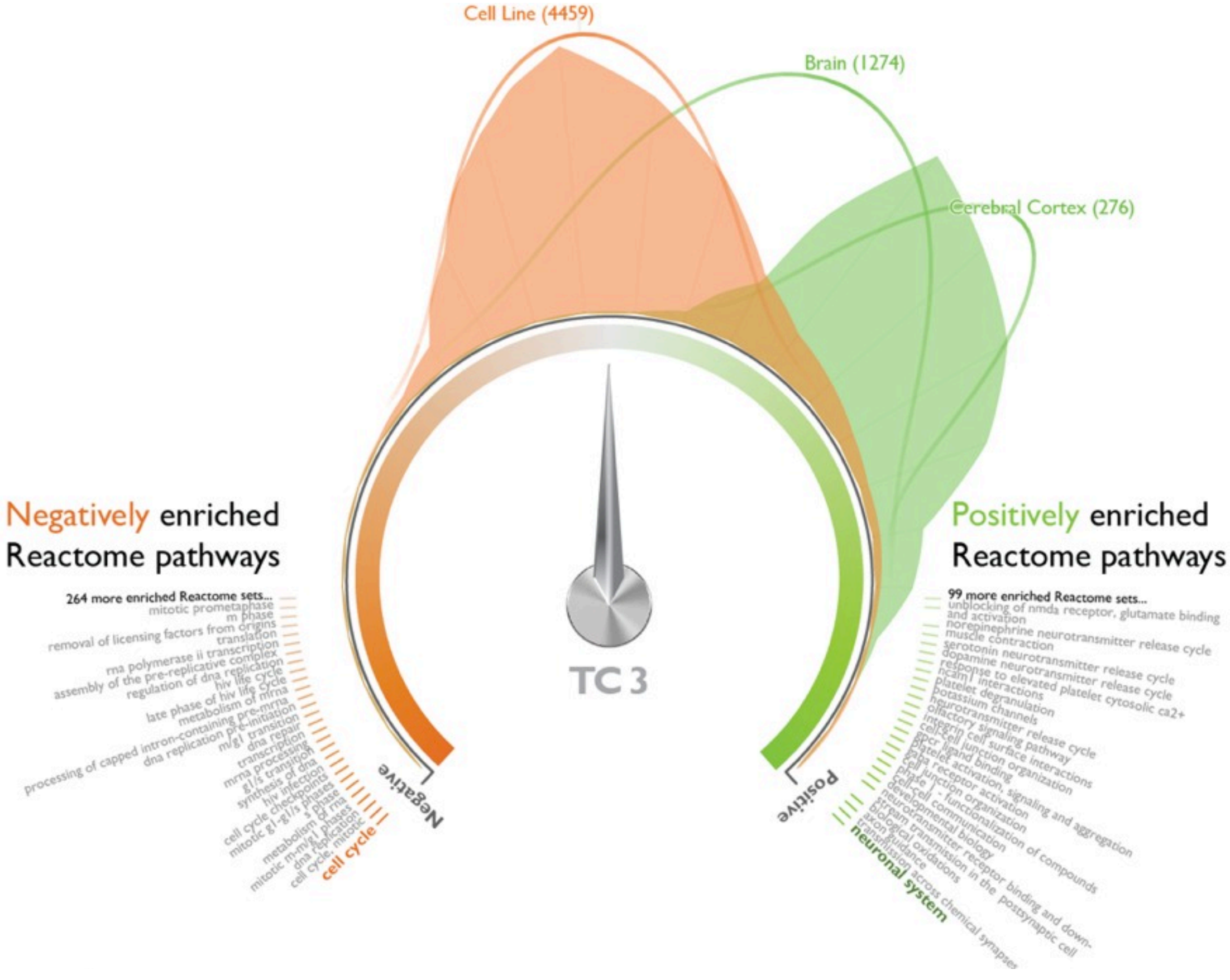
Three different species





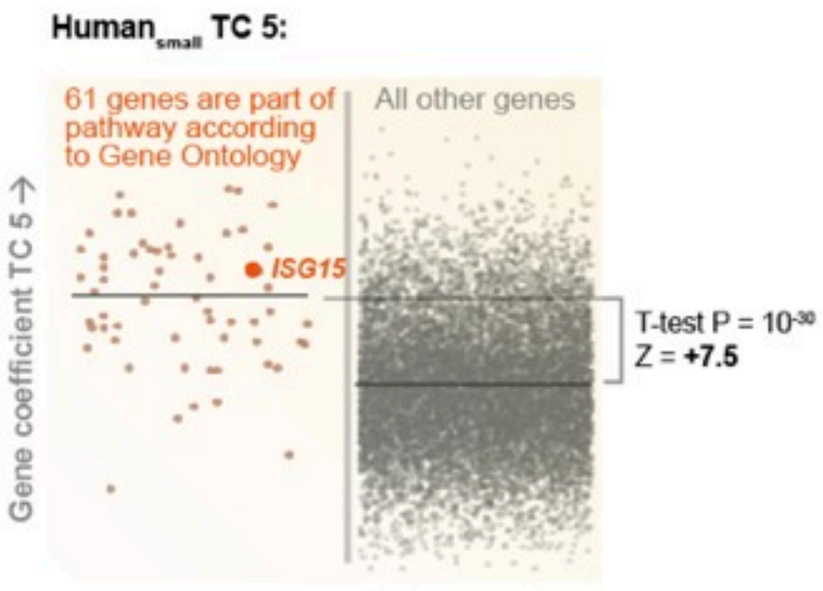
These components describe clear biology, can we use this?

Transcriptional component 3



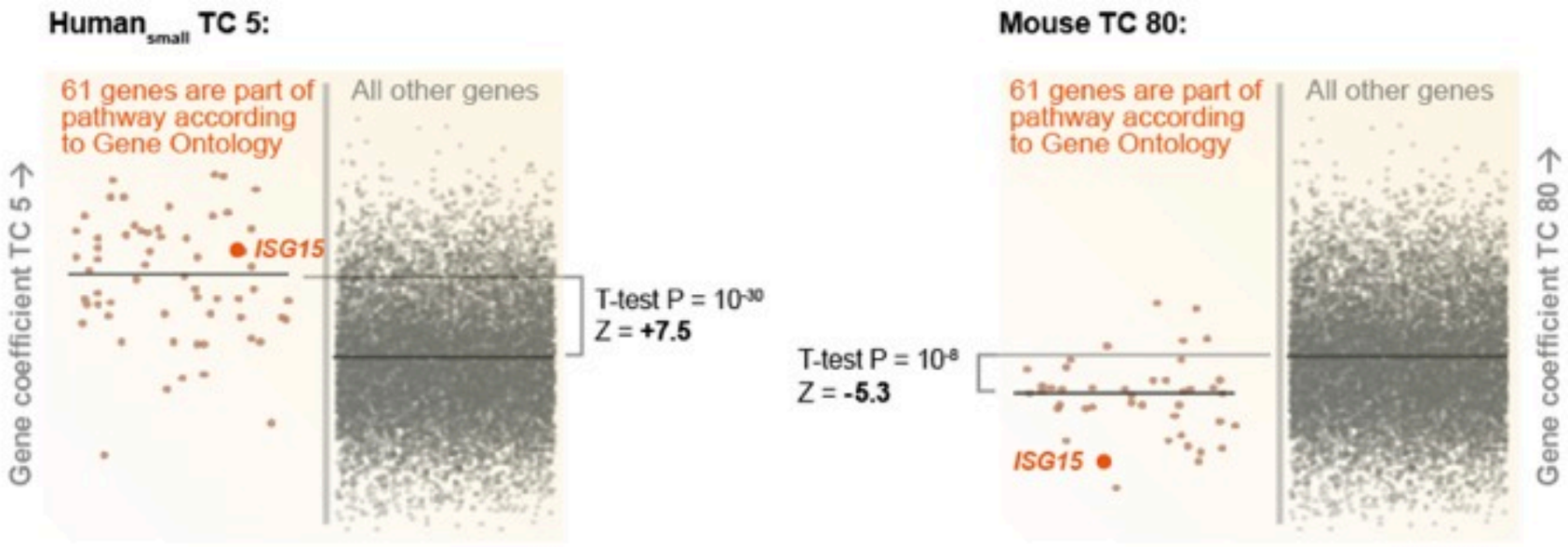
Predict function of genes

1 **Build profile for a pathway (GO: *Type I interferon-mediated signaling pathway*)**
Assess each of the 2,200 transcriptional components (TCs), perform T-Test per TC:



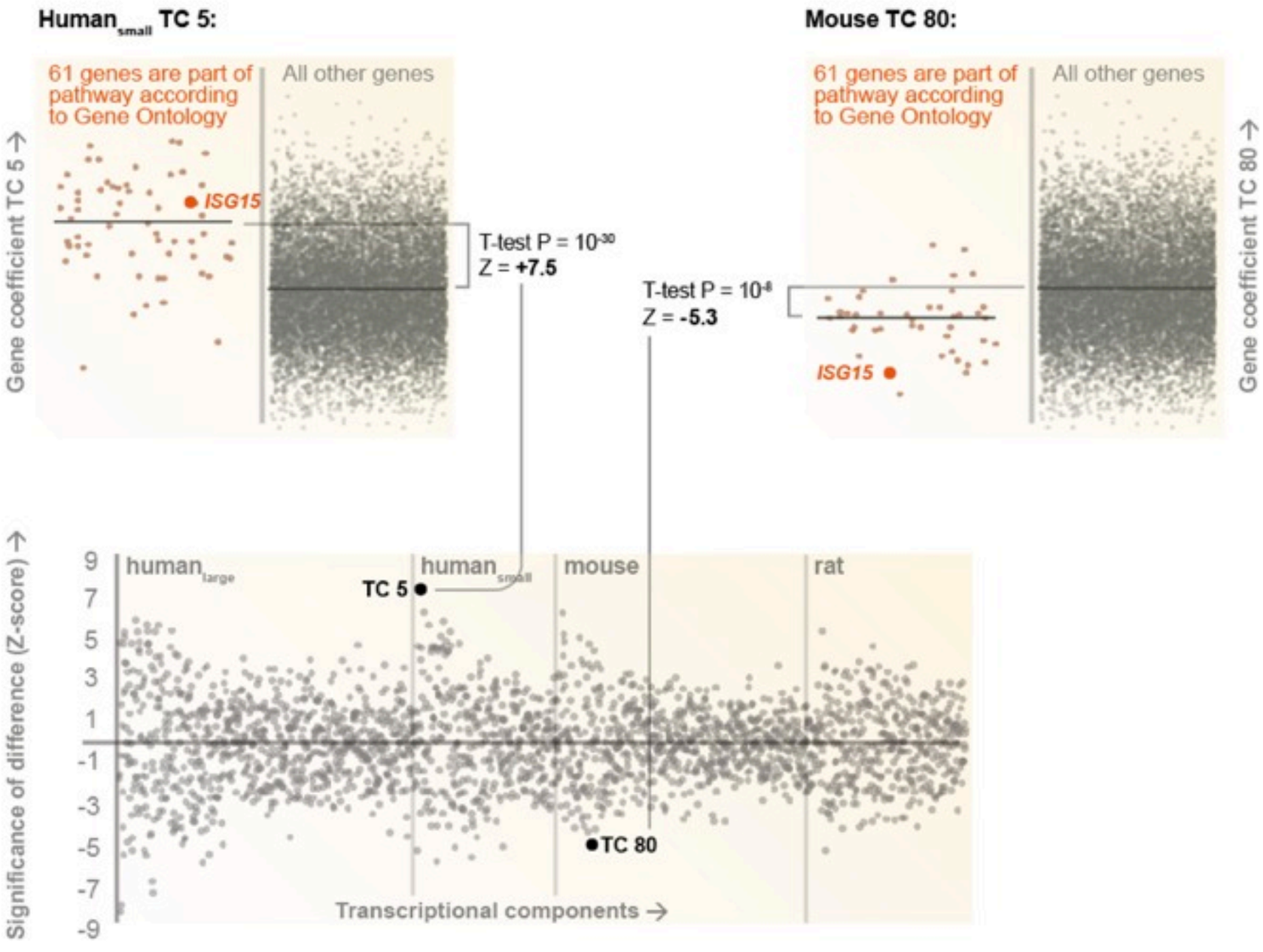
Predict function of genes

1 **Build profile for a pathway (GO: *Type I interferon-mediated signaling pathway*)**
Assess each of the 2,200 transcriptional components (TCs), perform T-Test per TC:

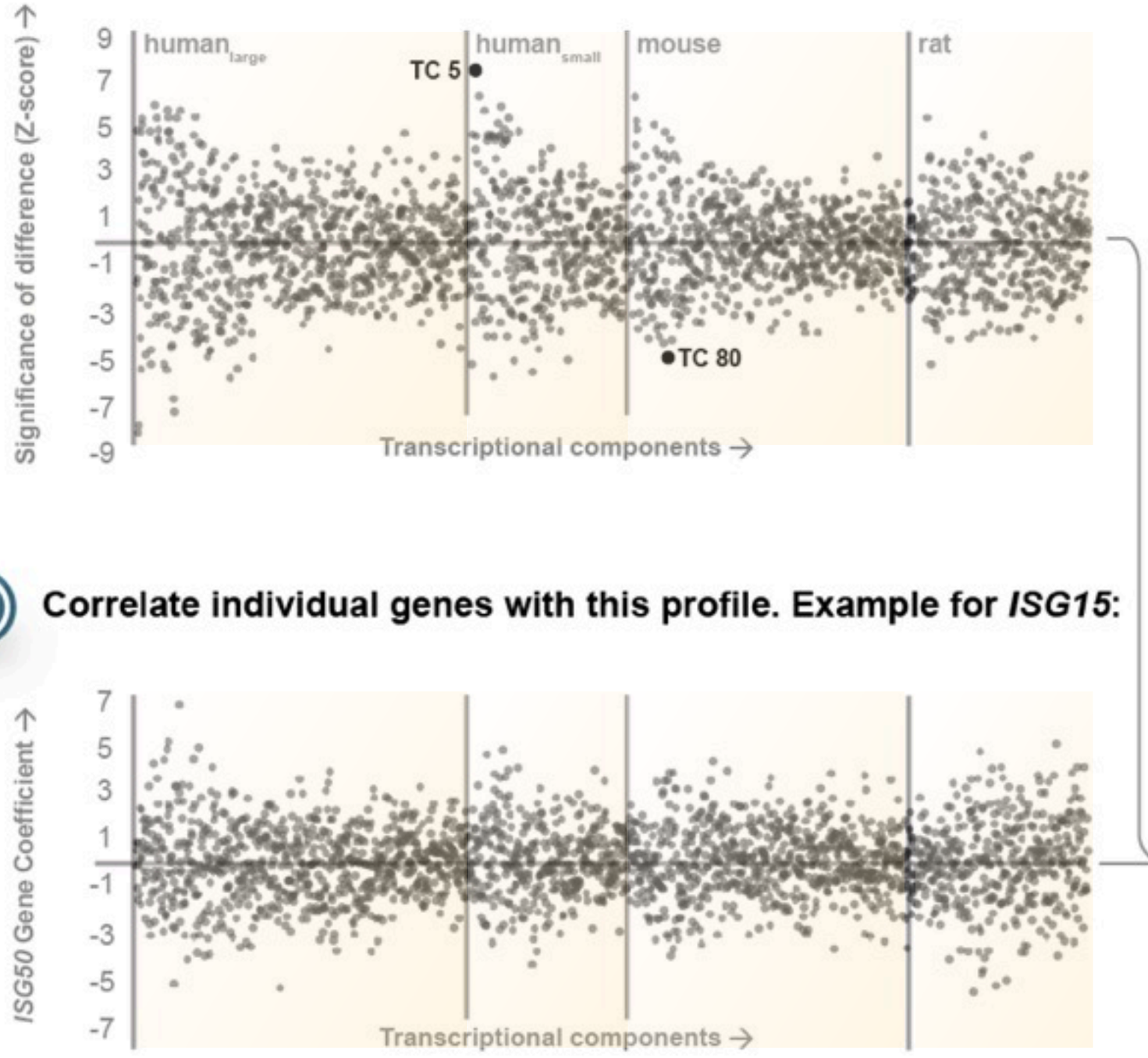


Predict function of genes

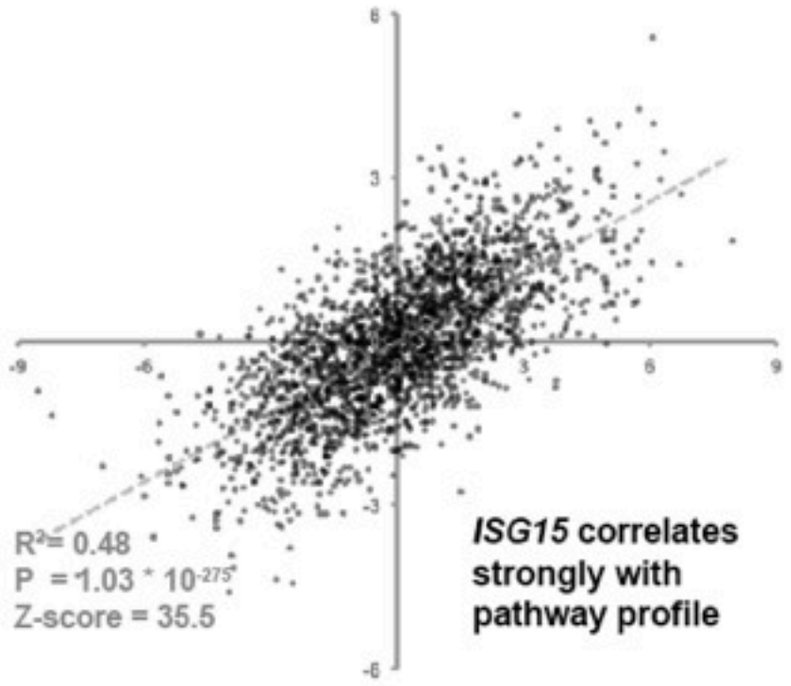
1 **Build profile for a pathway (GO: *Type I interferon-mediated signaling pathway*)**
Assess each of the 2,200 transcriptional components (TCs), perform T-Test per TC:



Predict function of genes

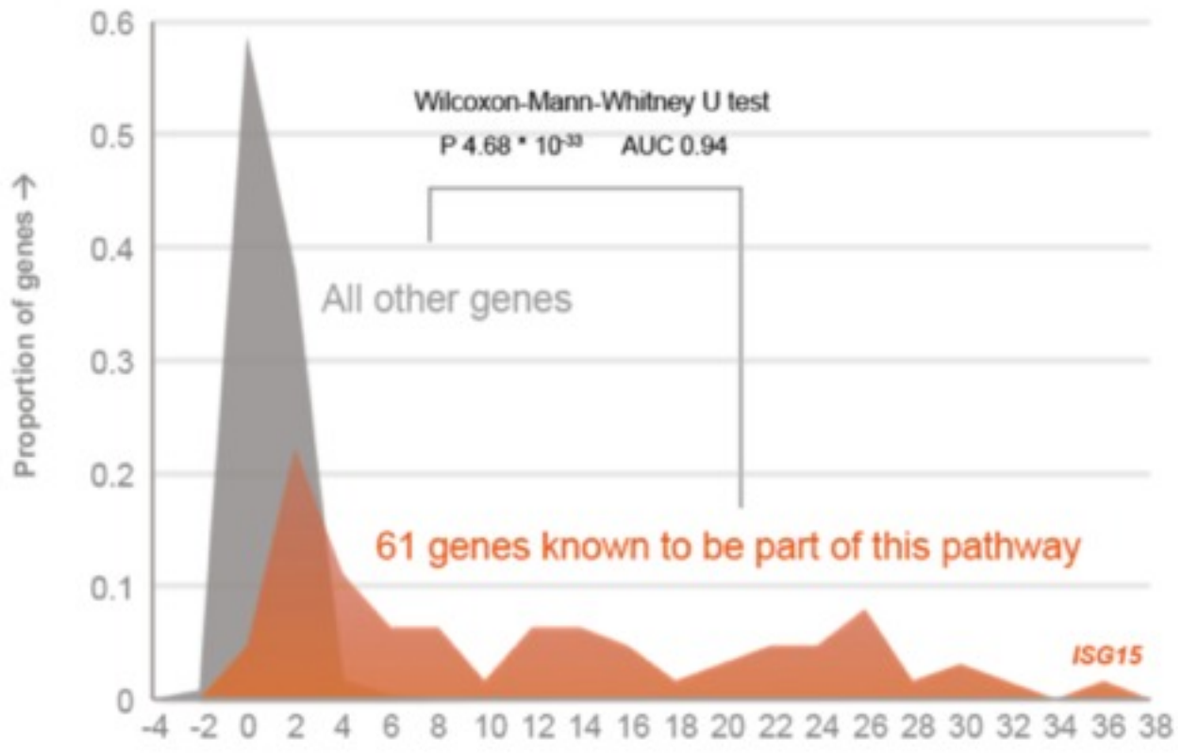


2 **Correlate individual genes with this profile. Example for *ISG15*:**



Predict function of genes

3 Assess performance

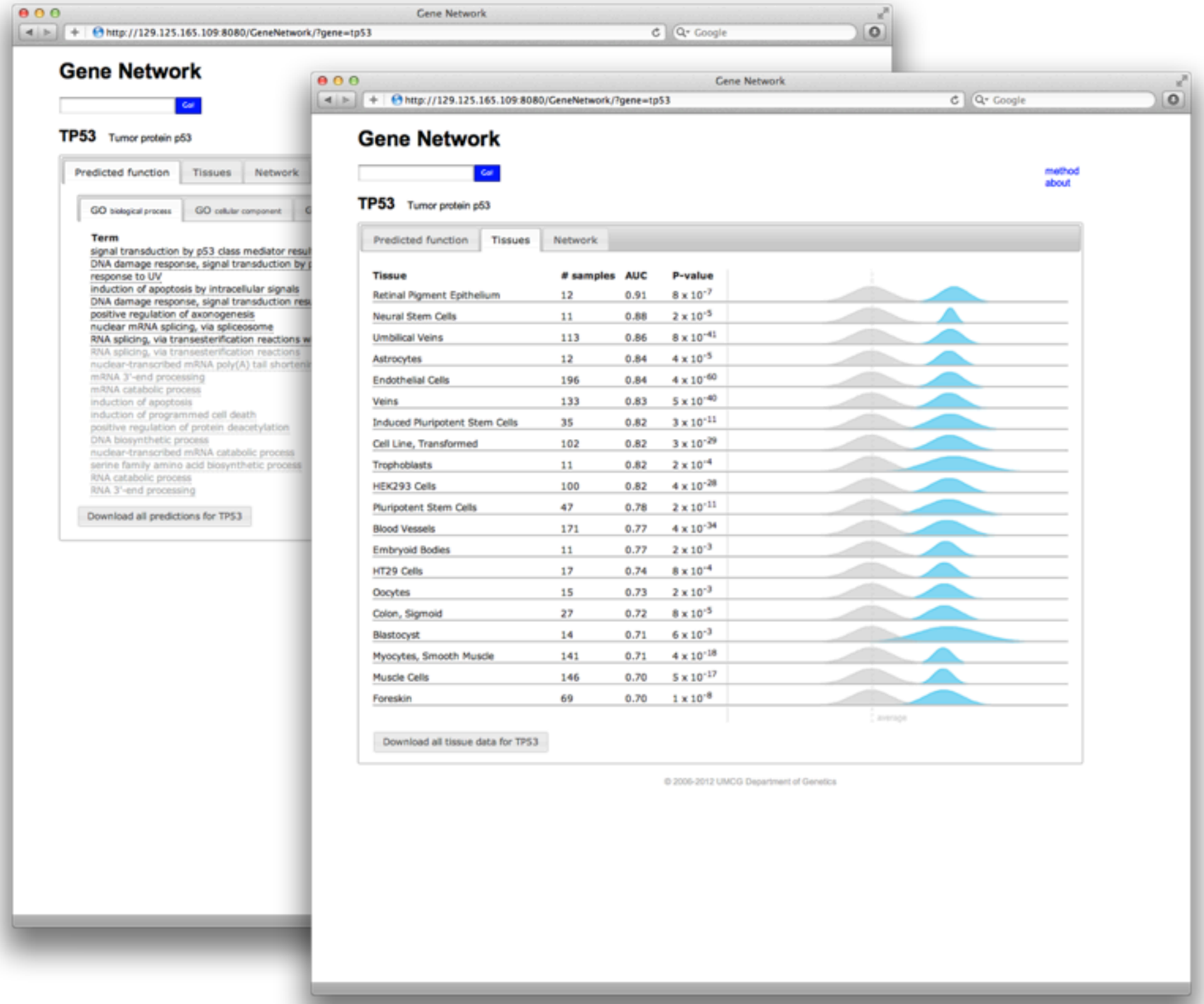


Assess predictions yourself: Look at genenetwork.nl/genenetwork

Example: TP53

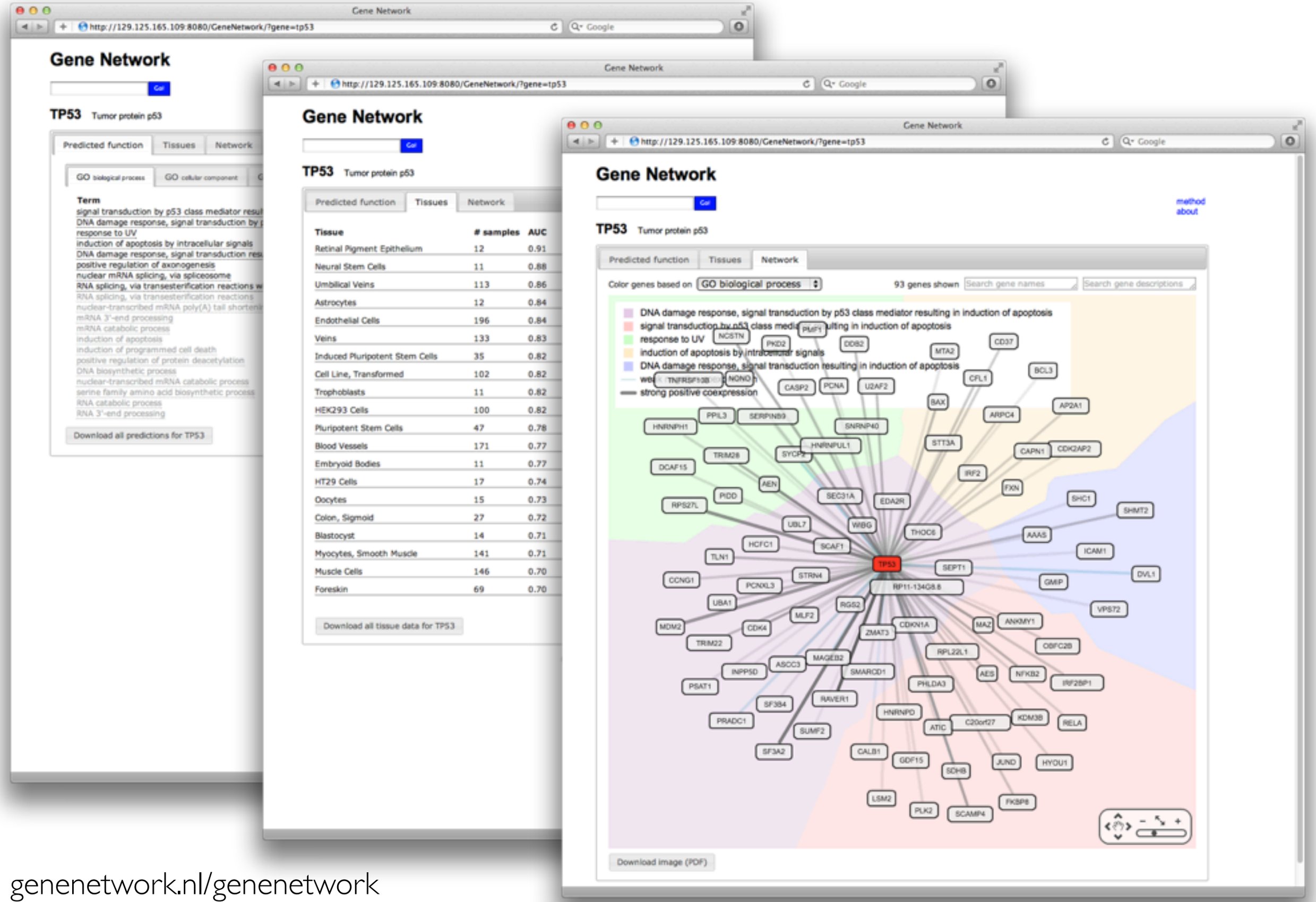
genenetwork.nl/genenetwork

Example: TP53



genenetwork.nl/genenetwork

Example: TP53



genenetwork.nl/genenetwork

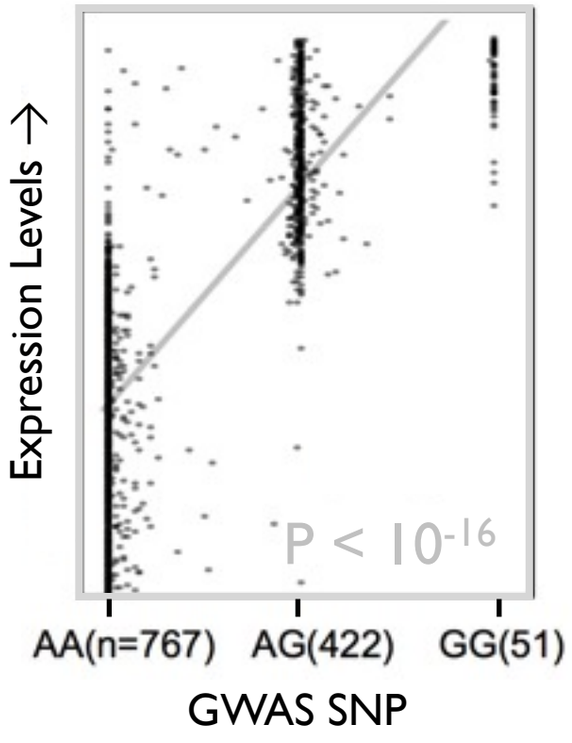
VEL blood group, gene unknown

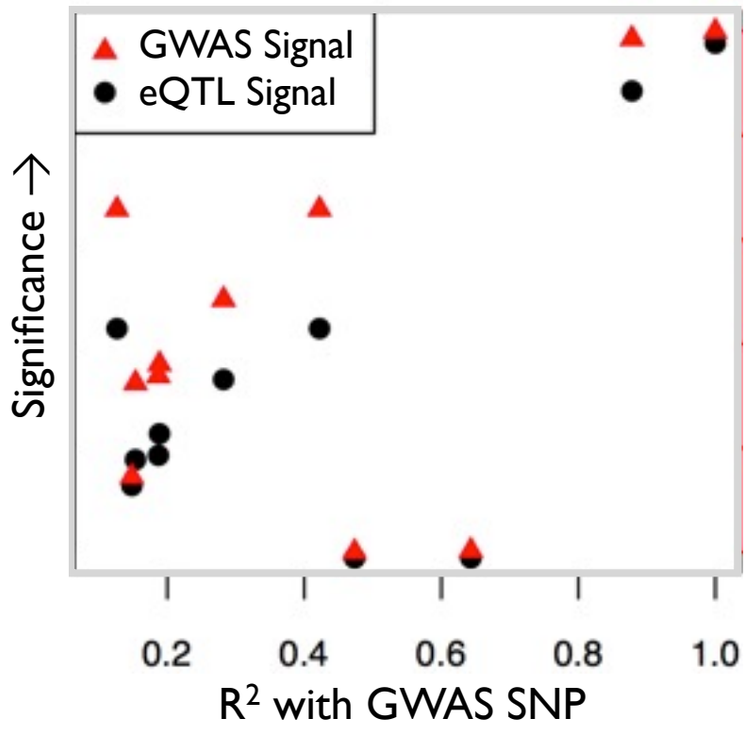
- GWAS on red blood cell traits (Van der Harst *et al*, Nature 2012)

ARTICLE nature
doi:10.1038/nature11677

Seventy-five genetic loci influencing the human red blood cell

- In one locus the SNP strongly affected gene expression of SMIM1, a gene without known function:





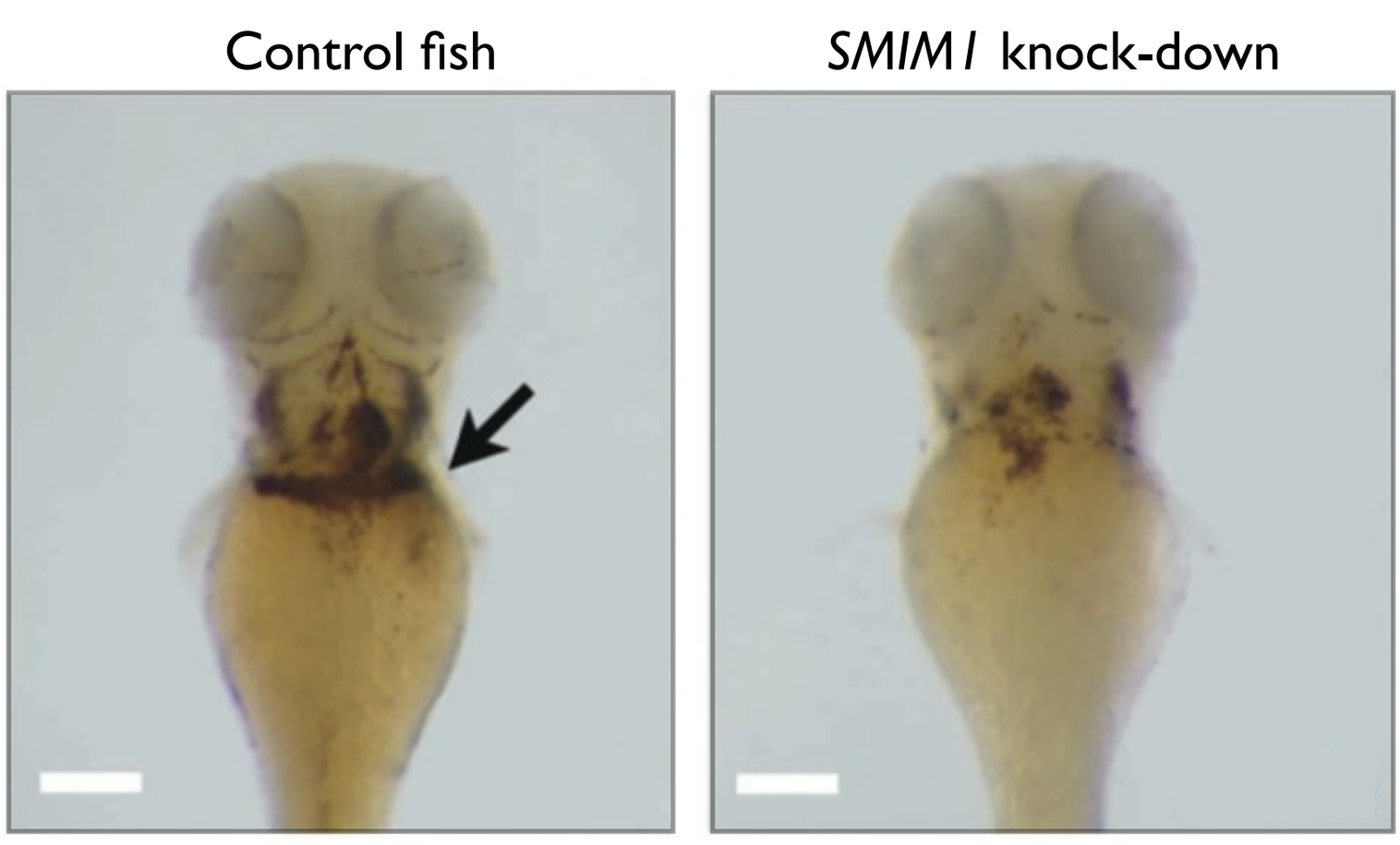
Cvejic *et al*, Nature Genetics 2013

VEL blood group, *SMIM1* gene

- Most significantly predicted GeneNetwork function:
  'Hemoglobin metabolic process' ($P = 10^{-16}$)

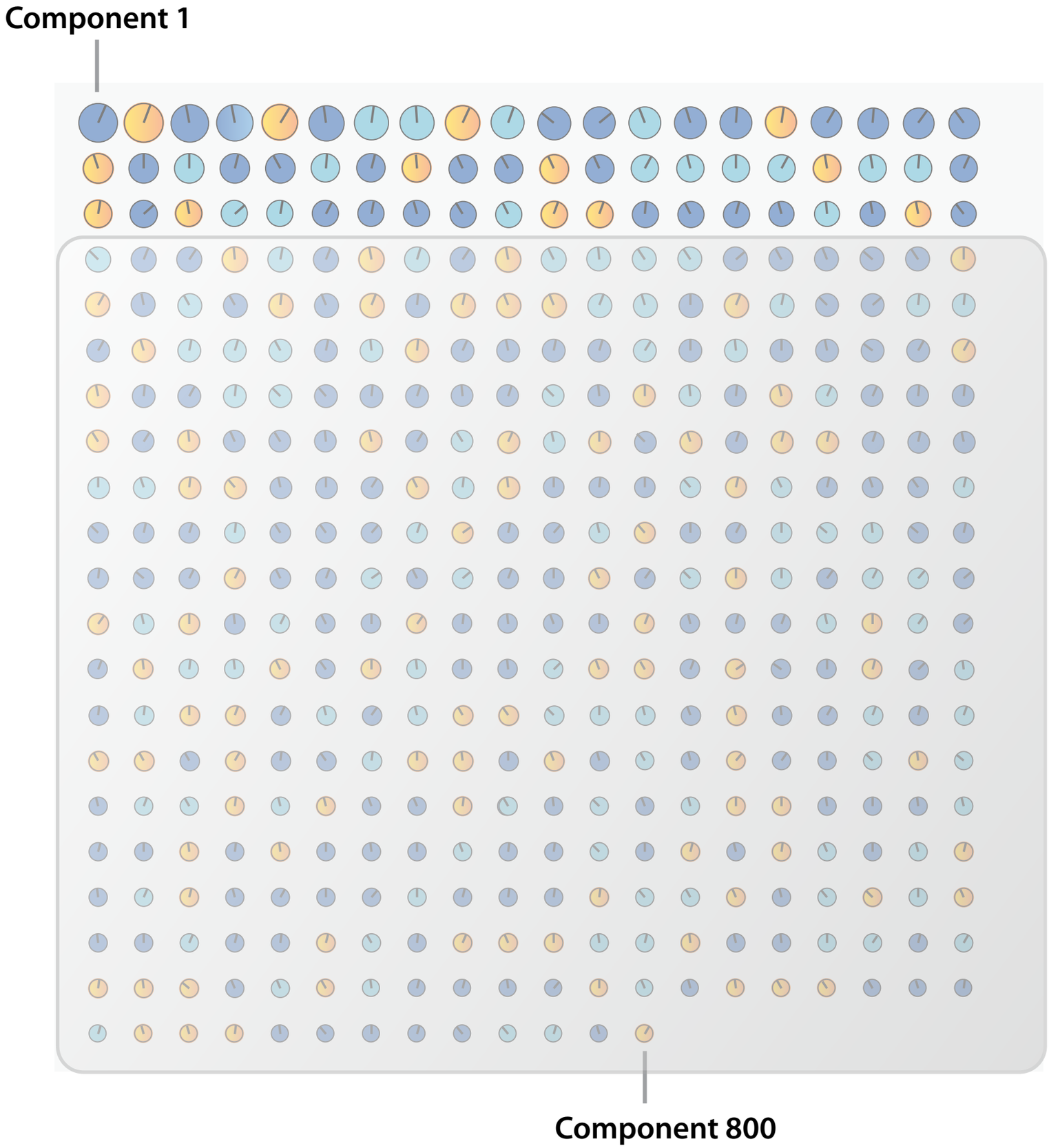
- Homozygous *SMIM1* 17bp deletion found in 63 out of 69 individual
  (Mutation frequency <1% in entire population)

- Zebrafish knock-down: reduced number of red blood cells

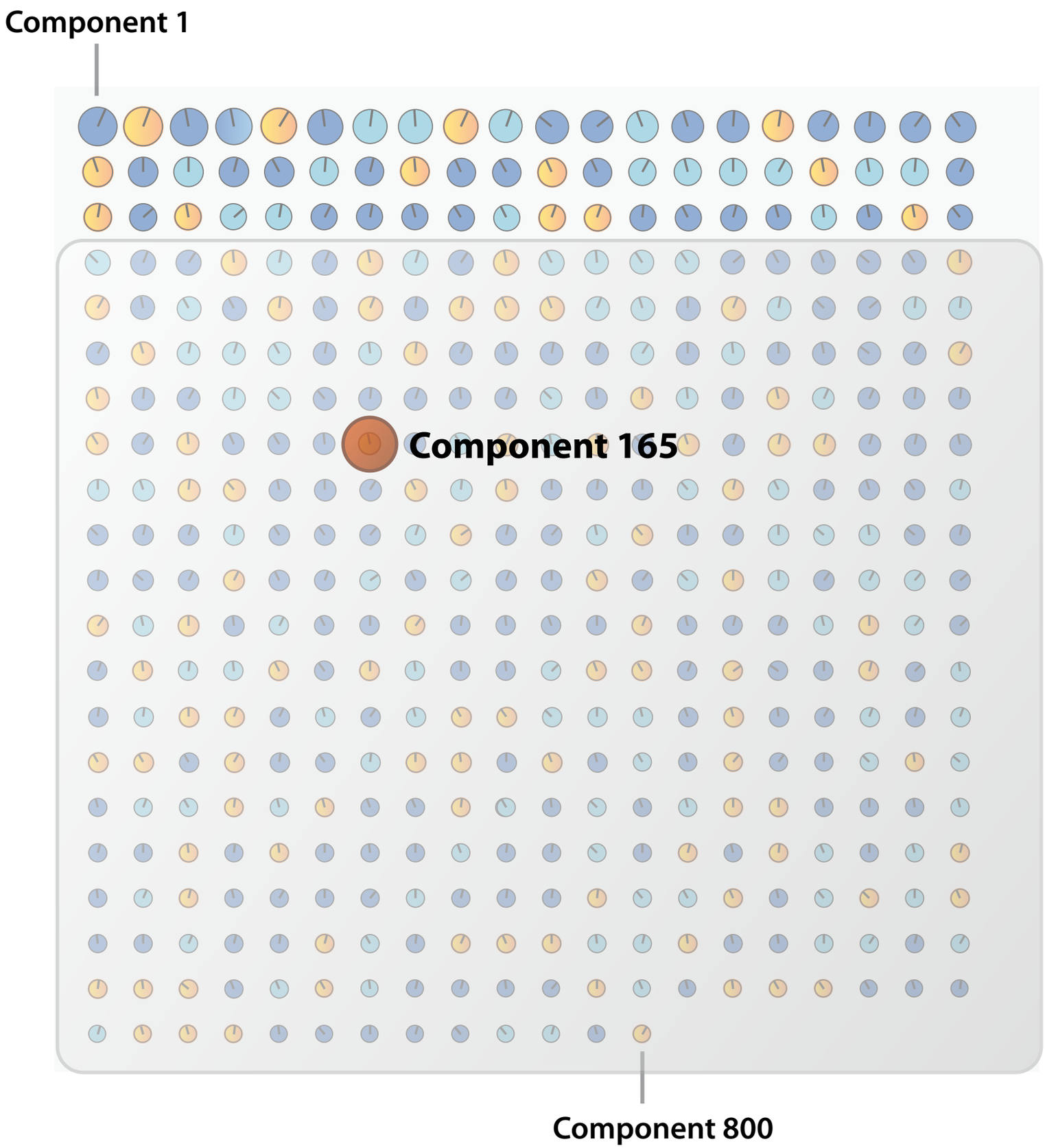


Cvejic *et al*, Nature Genetics 2013

Components 51 - 800

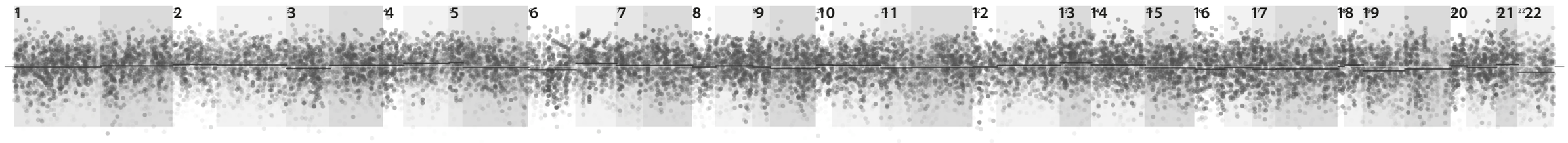


Components 51 - 800



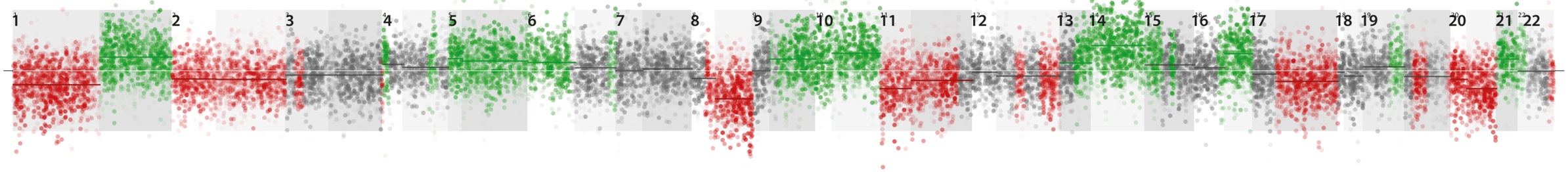
Some component show weird behaviour

TC 1: No cytogenetic effect, zero autocorrelation

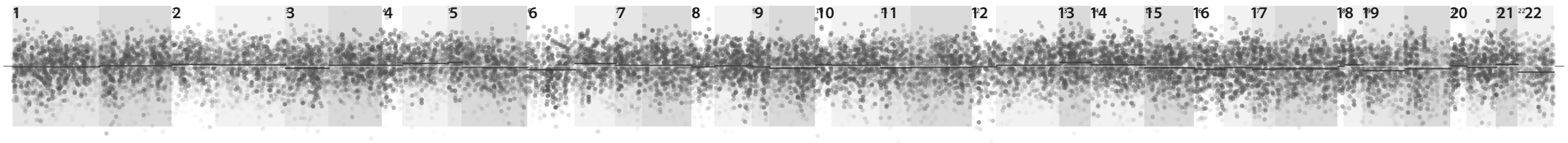


Some component show weird behaviour

TC 165: Strong cytogenetic effects, high autocorrelation

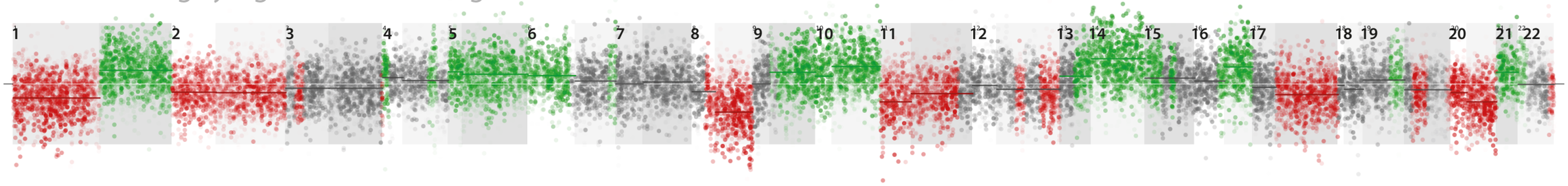


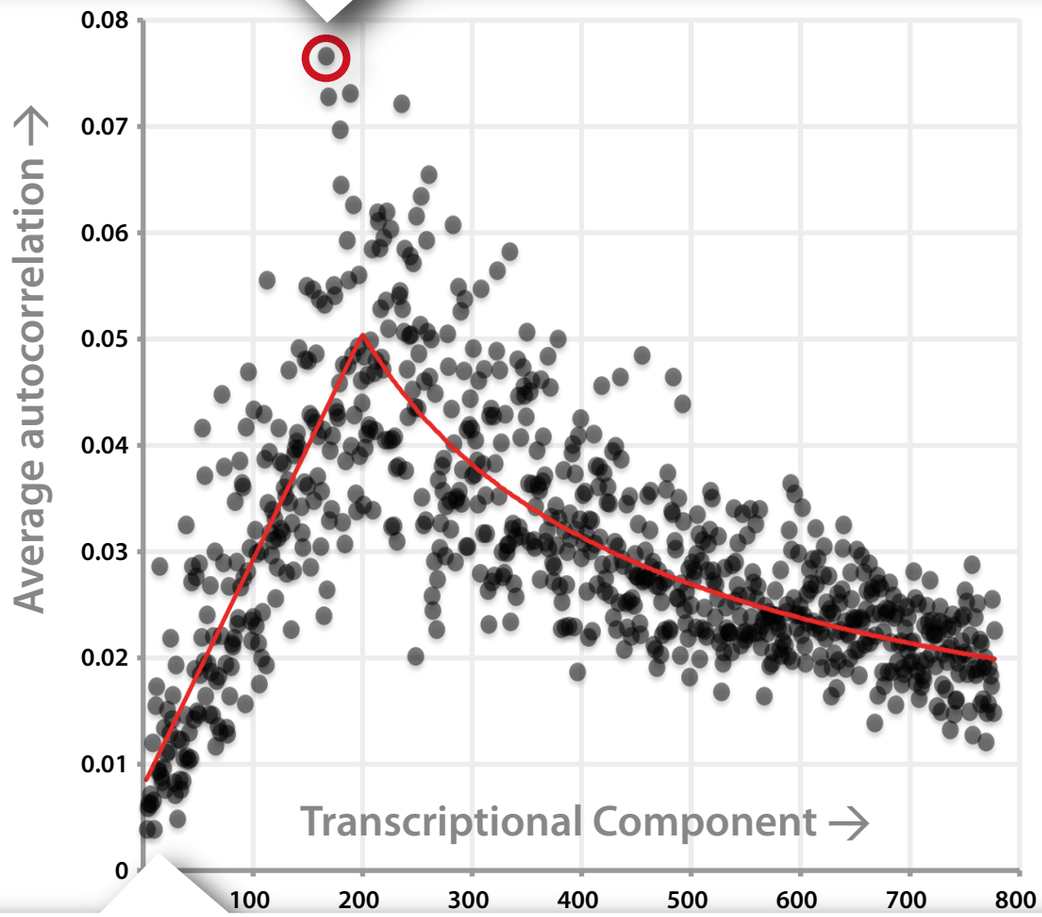
TC 1: No cytogenetic effect, zero autocorrelation



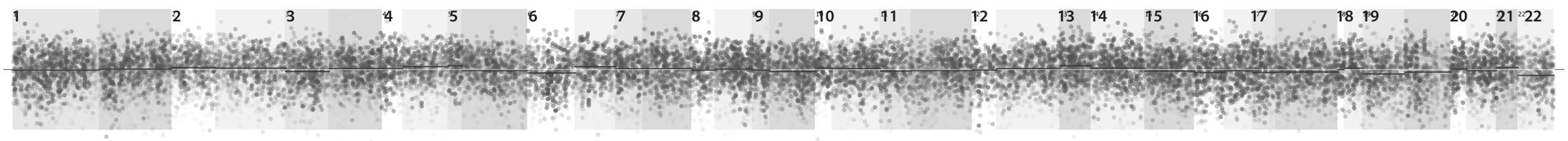
Some component show weird behaviour

TC 165: Strong cytogenetic effects, high autocorrelation





TC 1: No cytogenetic effect, zero autocorrelation



Identify non-cancer, physiological TCs

Take 50% of 37,427 samples with least genomic instability

↓

PCA on 54,000 x 54,000 probeset correlation matrix (dedicated multi-threaded optimized C++ code, ensuring numerical stability)

↓

718 highly robust Prinicipal Components (Cronbach's Alpha > 0.7)

↓

Treat these as orthogonal covariates, correct the expression data of all 37,427 samples, take residual expression data

↓

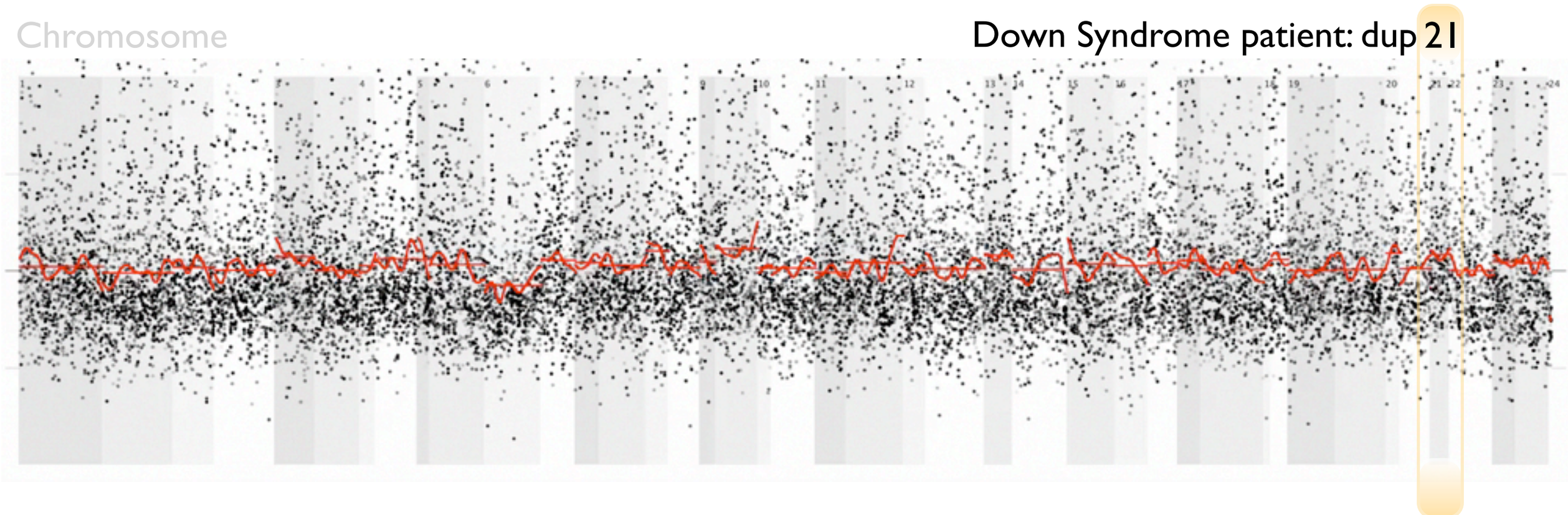
'Cytogenetic' RNA expression, only use 50% of 37,427 with most genomic instability

Detection cytogenetic aberration in expression data

Chromosome

Down Syndrome patient: dup 21

Detection cytogenetic aberration in expression data



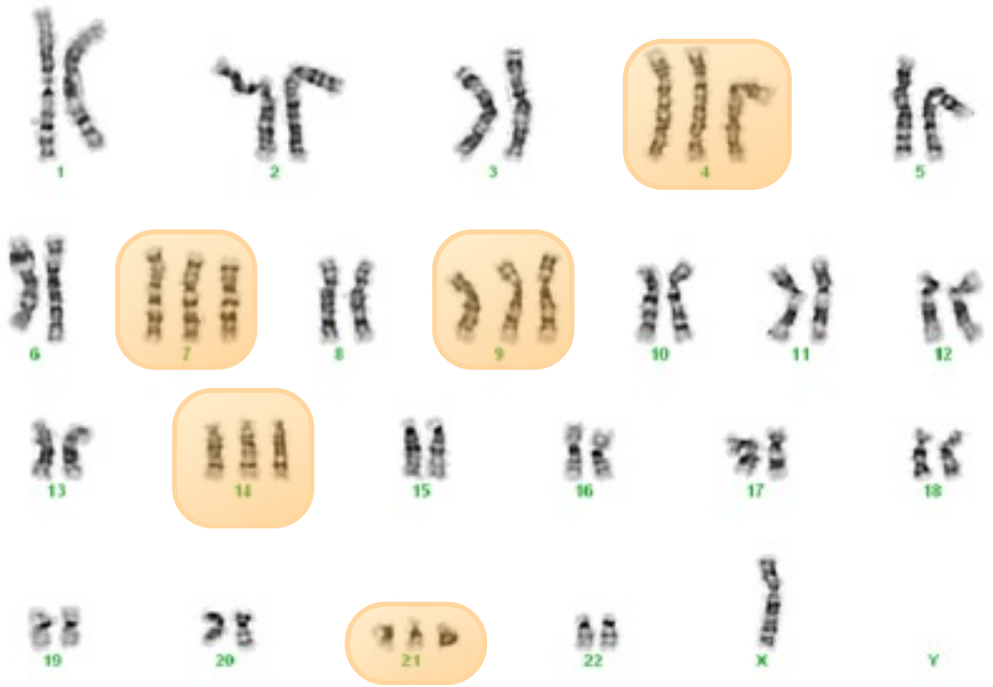
Identifying five chromosome duplications

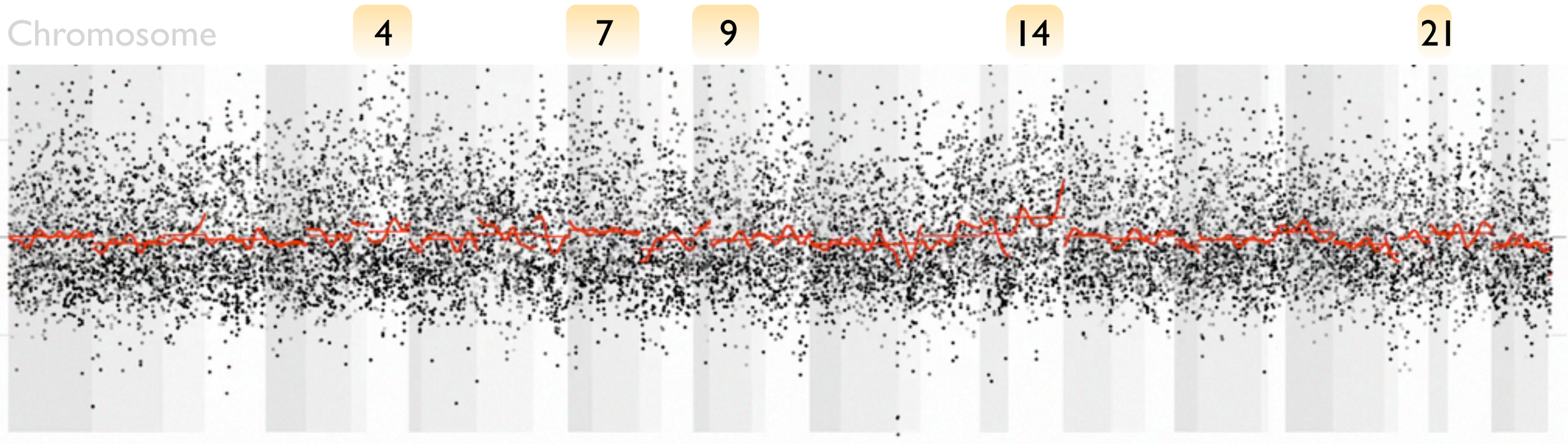
Karyogram
HapMap LCL

Chromosome	4	7	9	14	21

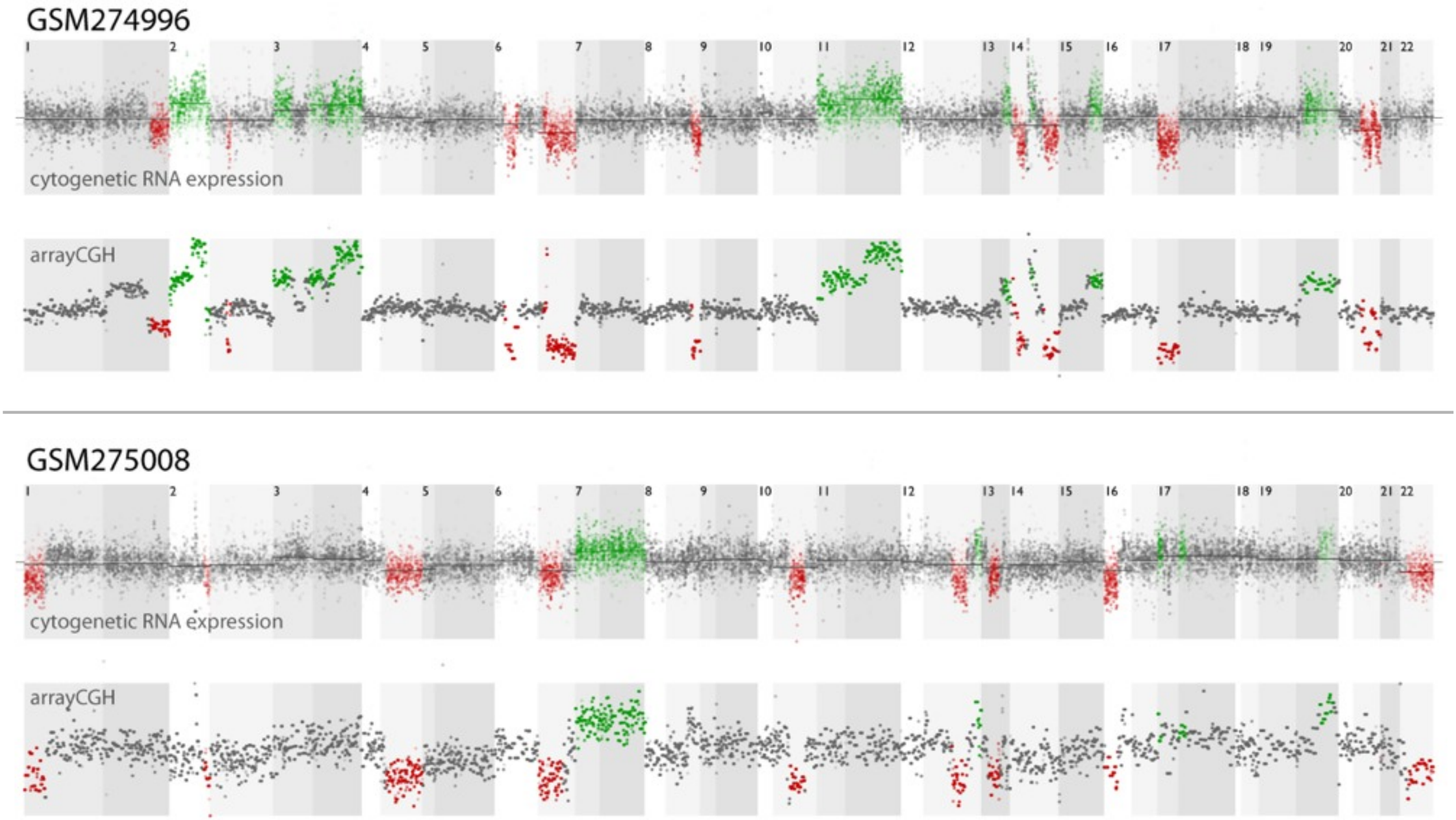
Identifying five chromosome duplications

Karyogram
HapMap LCL

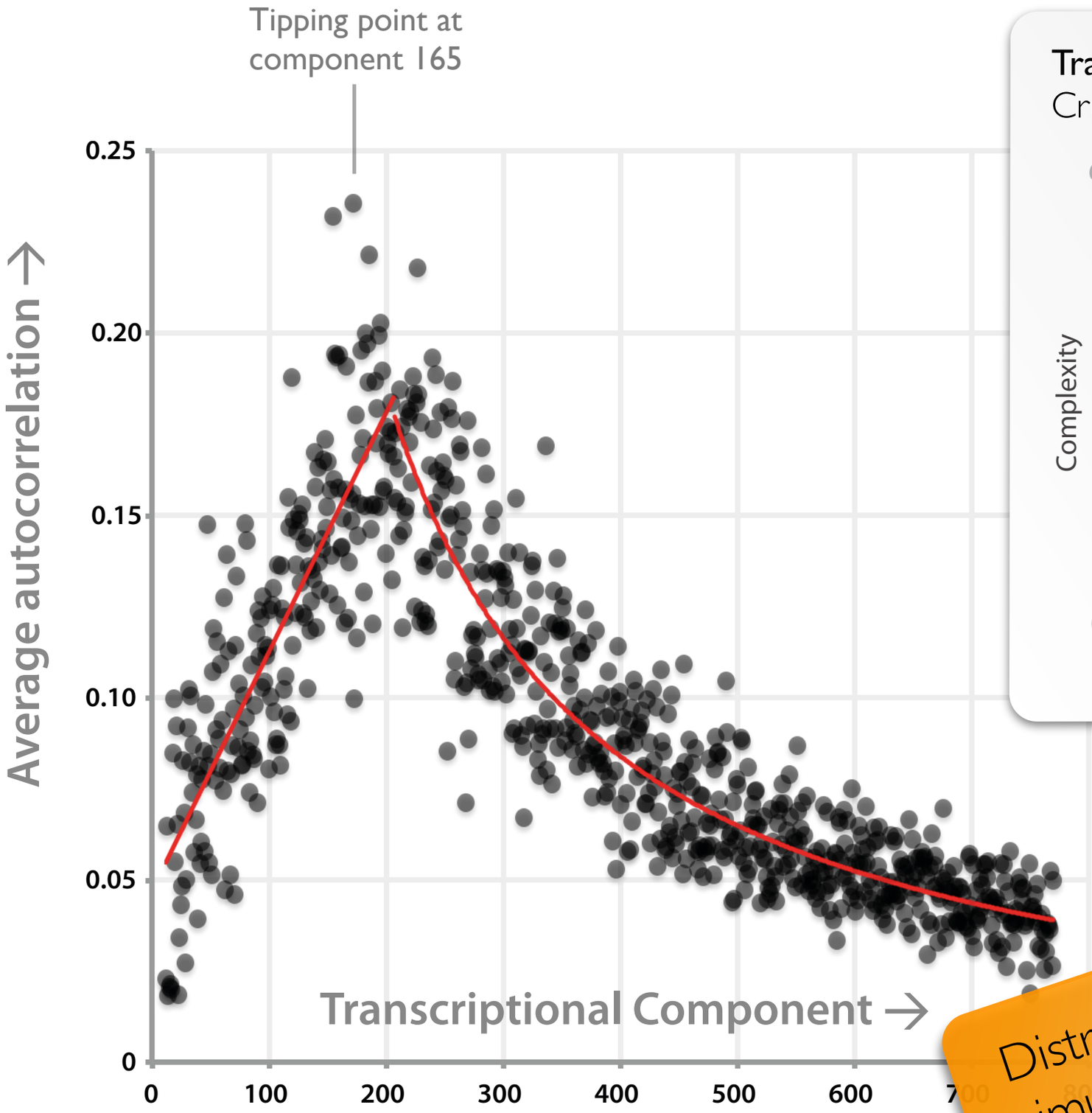




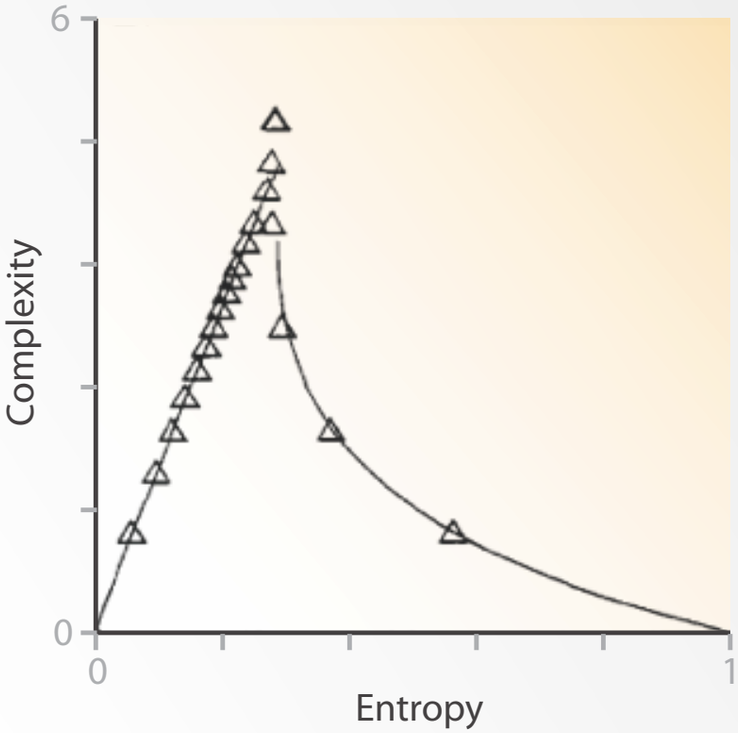
Comparison of arrayCGH and cytogenetic RNA profiles



Amount of cytogenetic aberrations



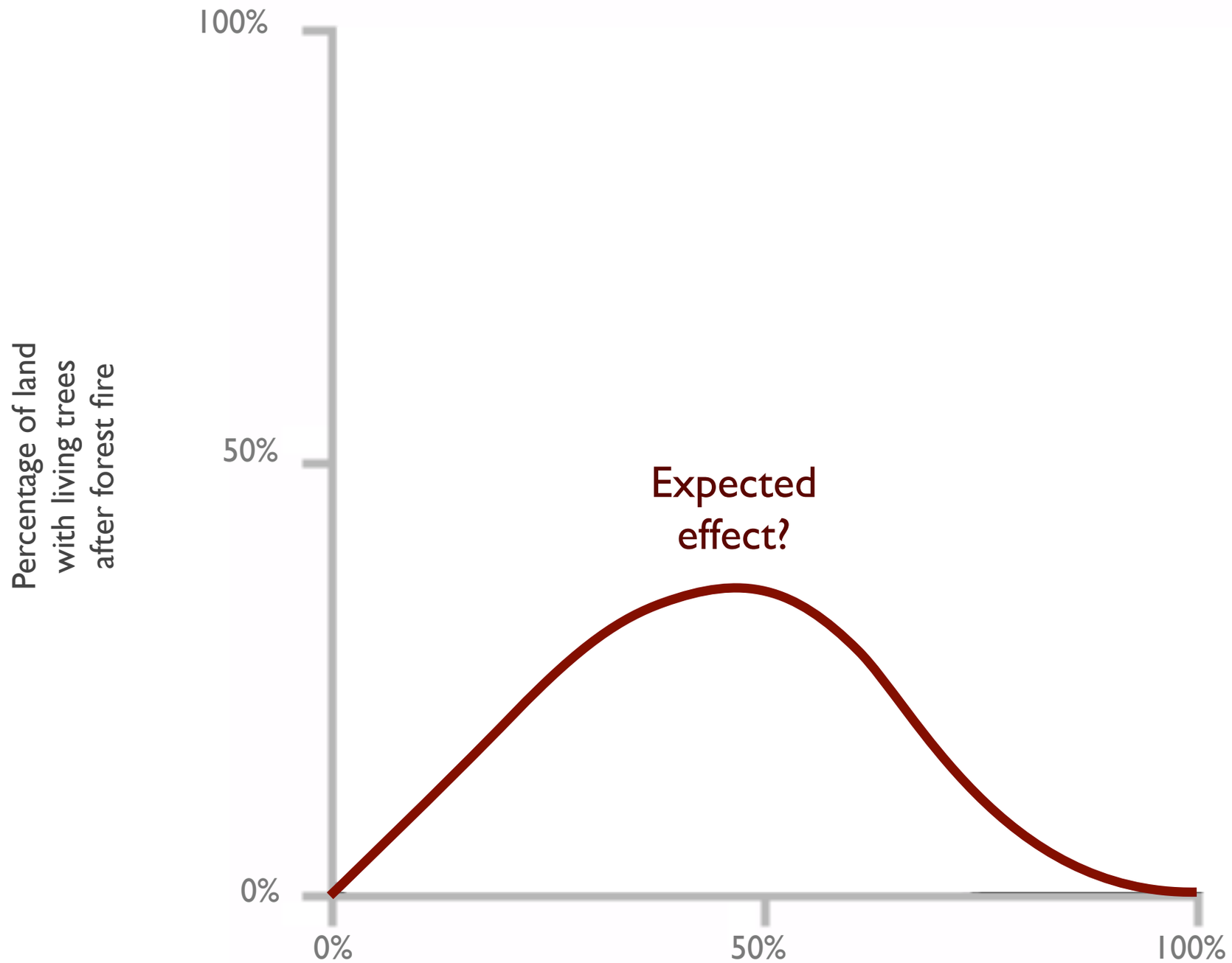
Transition to chaos in the logistic map
Crutchfield *et al*, 1990



Distribution identical to simulations in complexity theory

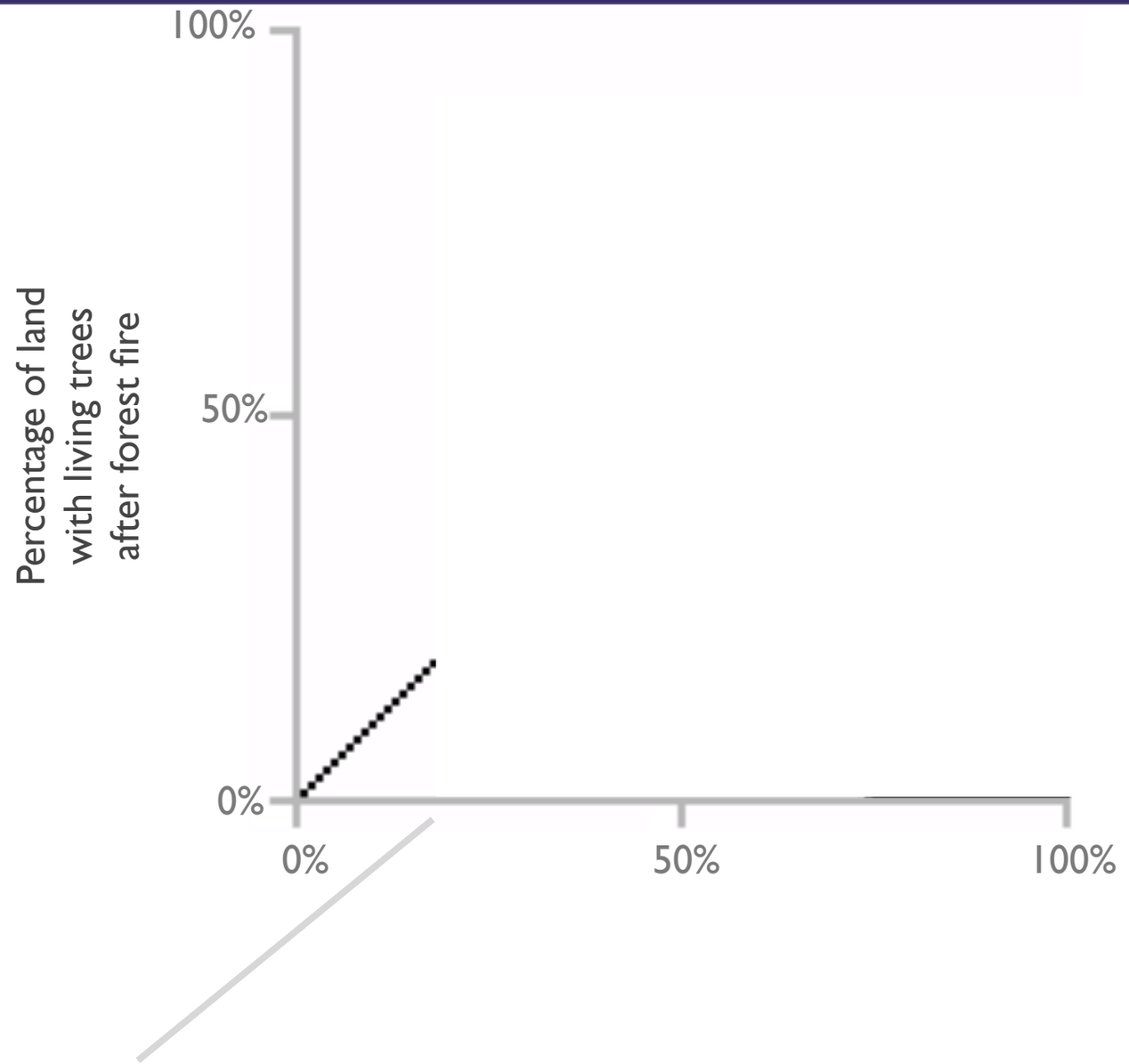
Complexity: Forest fire

cdc



Complexity: Forest fire

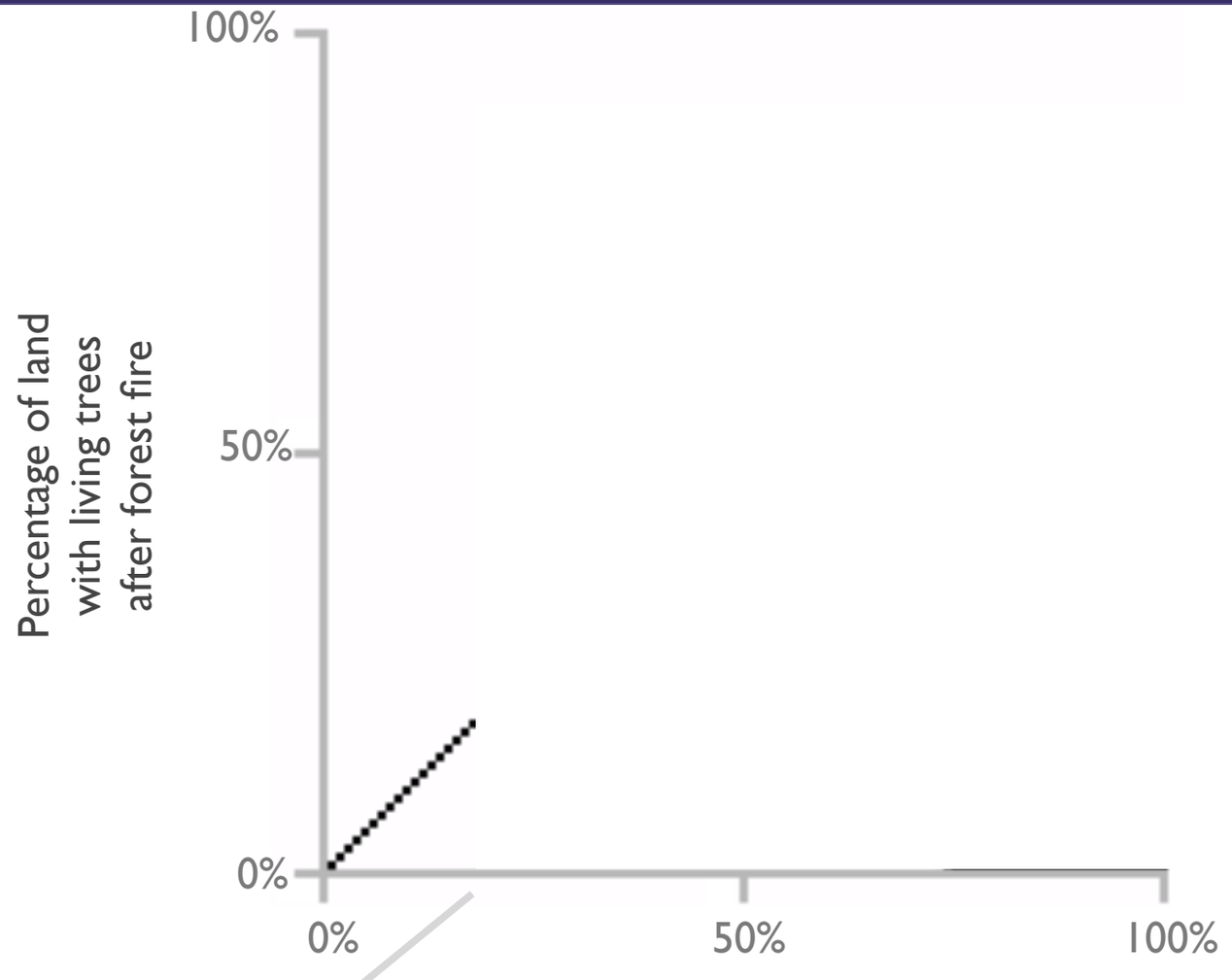
cdc

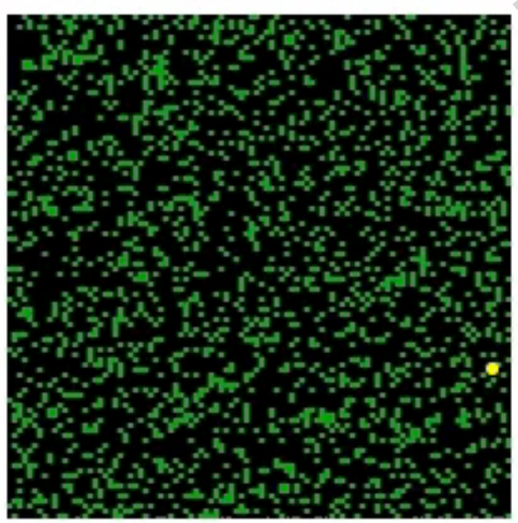


20%

Complexity: Forest fire

cdc

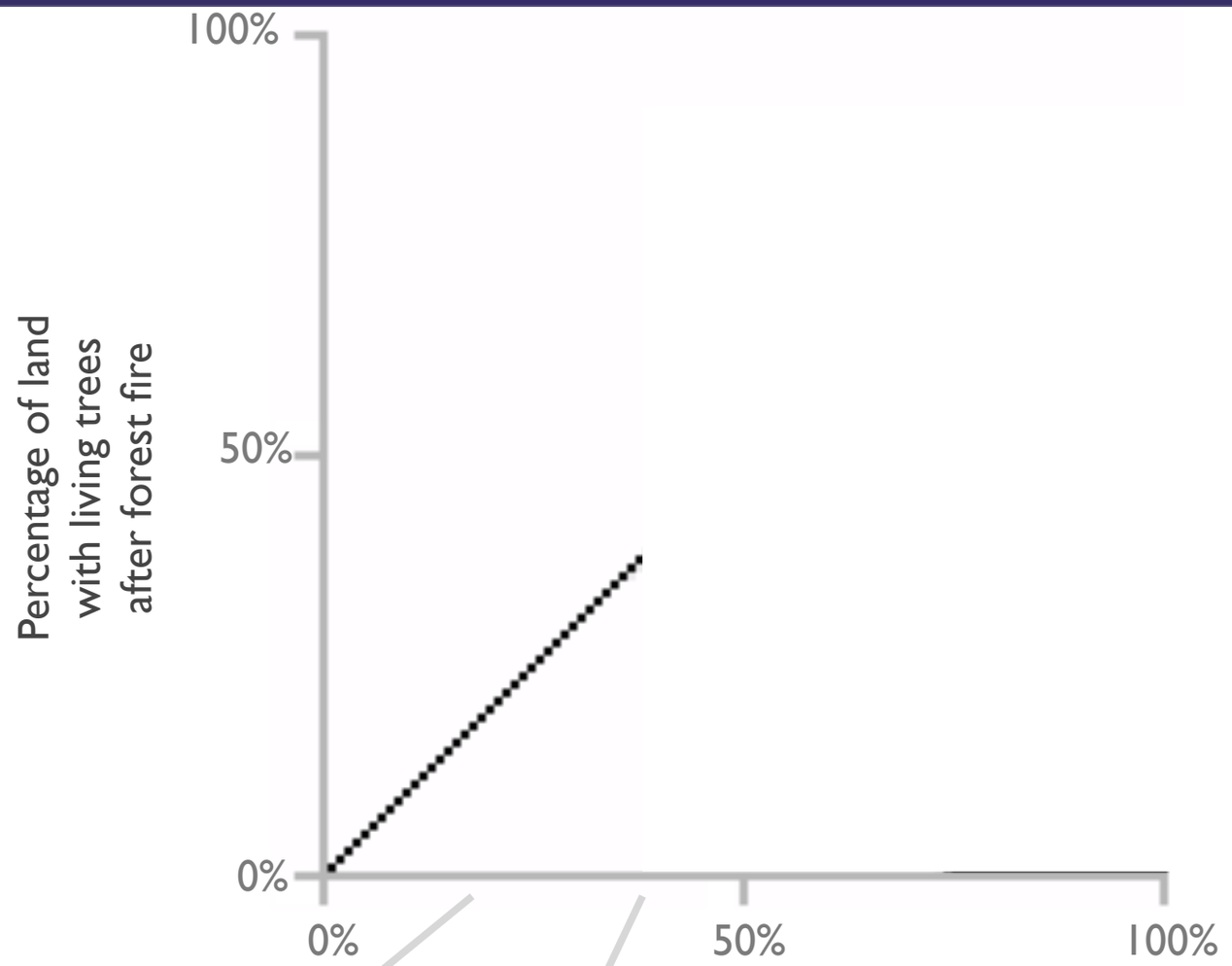


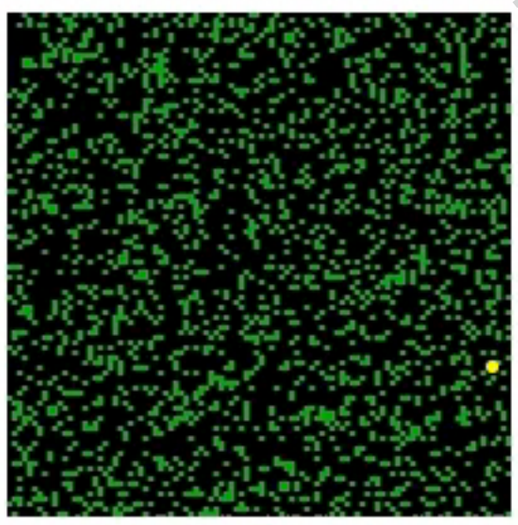


20%

Complexity: Forest fire

cdc



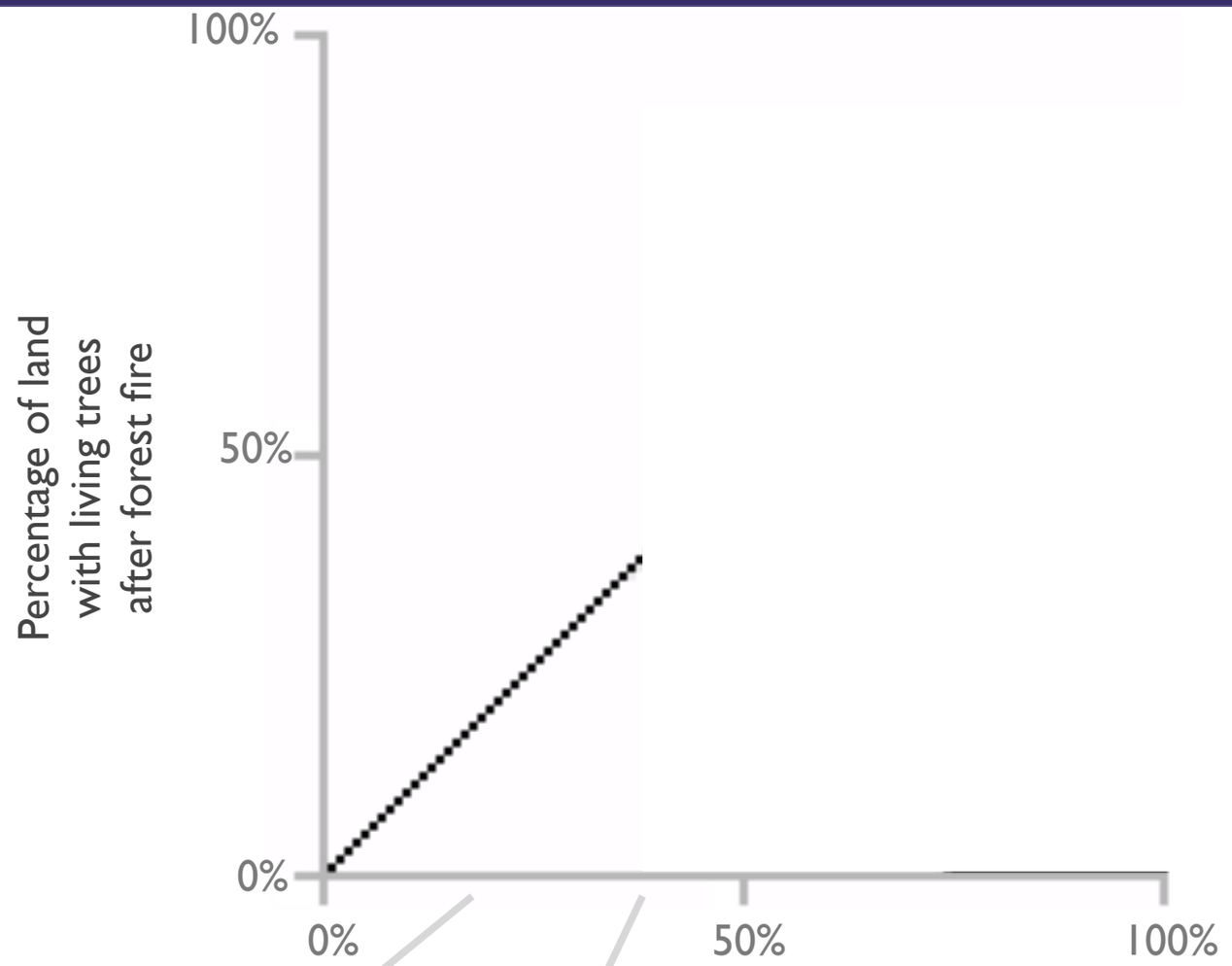


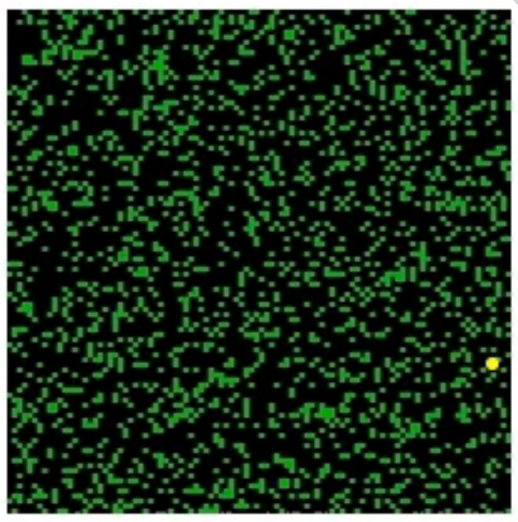
20%

40%

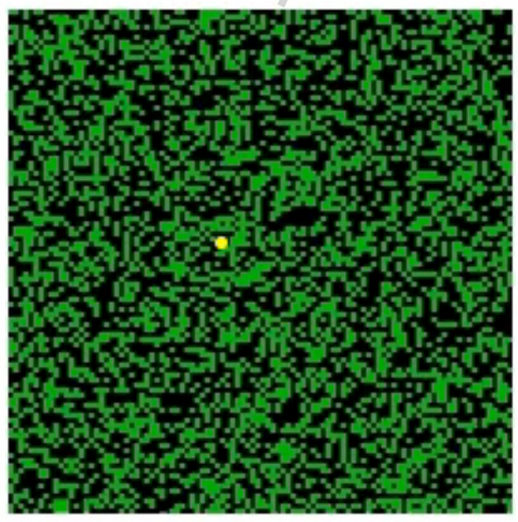
Complexity: Forest fire

cdc





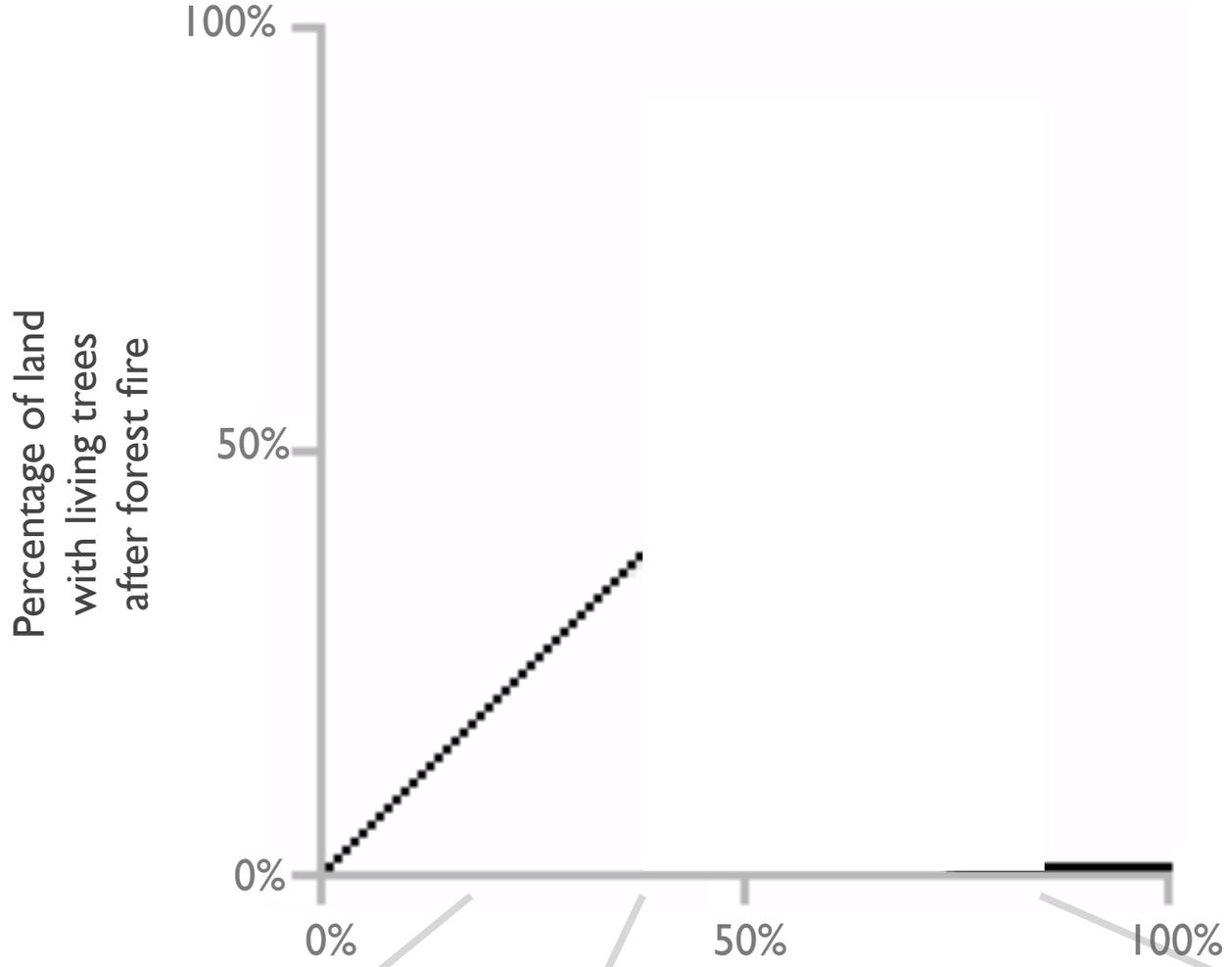
20%

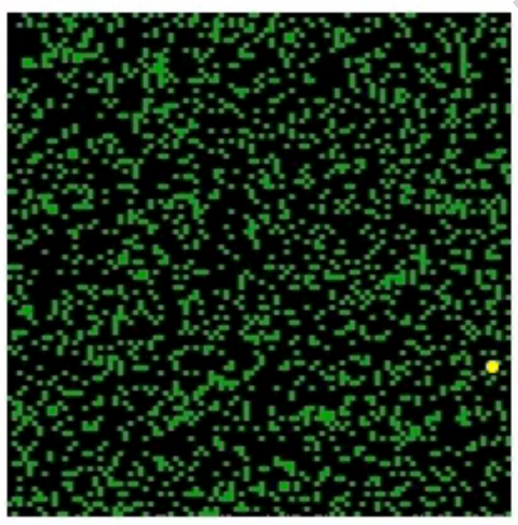


40%

Complexity: Forest fire

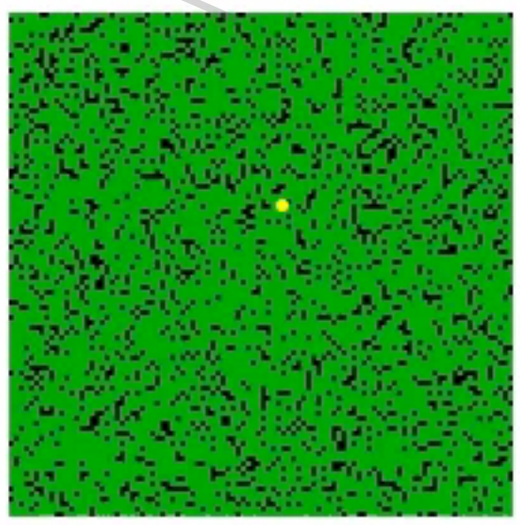
cdc





20%

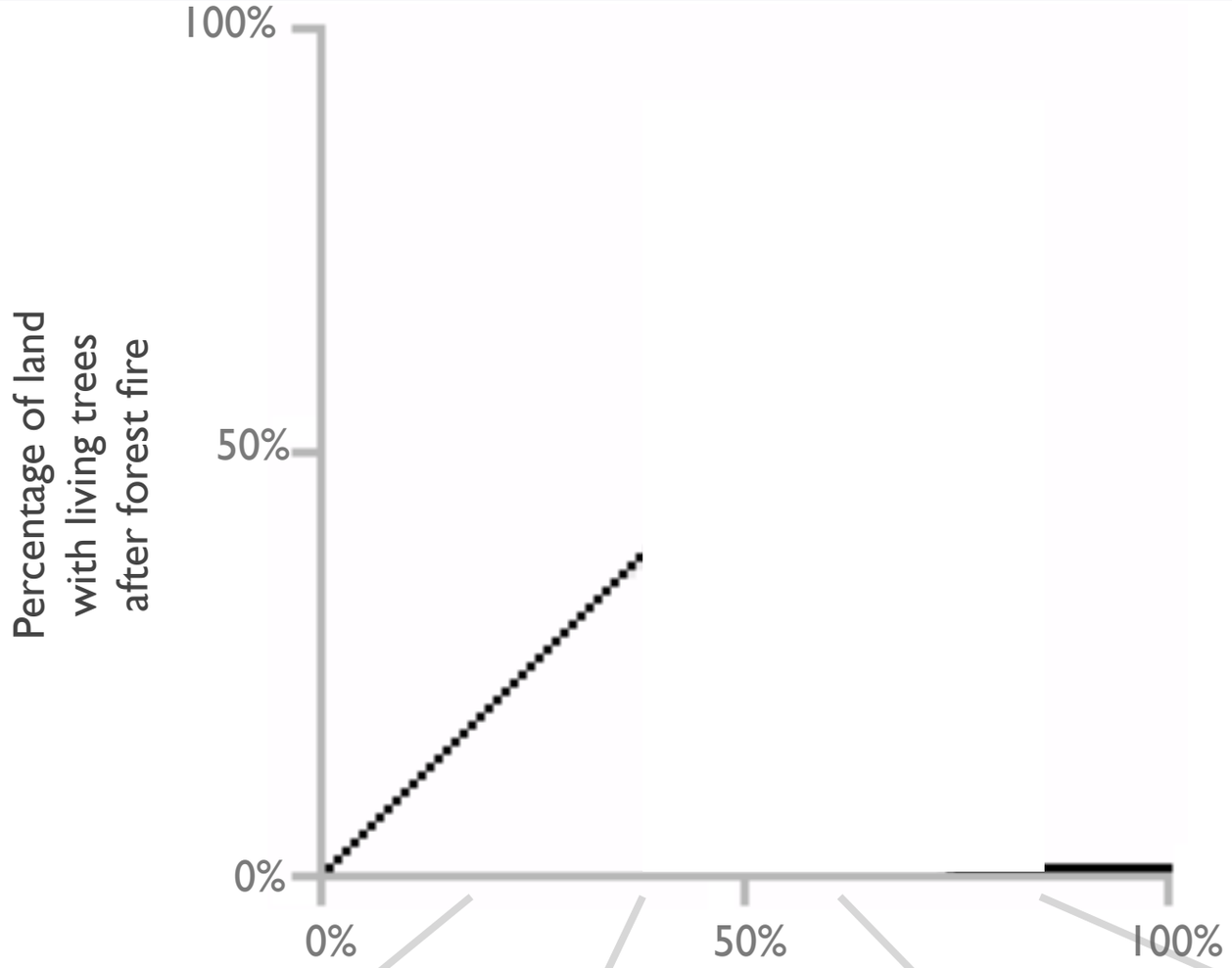
40%

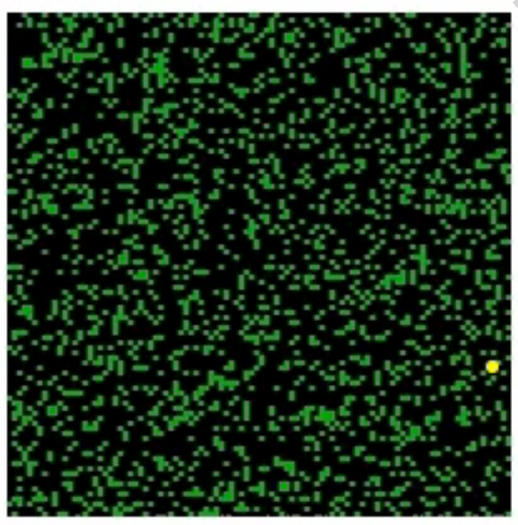


80%

Complexity: Forest fire

cdc

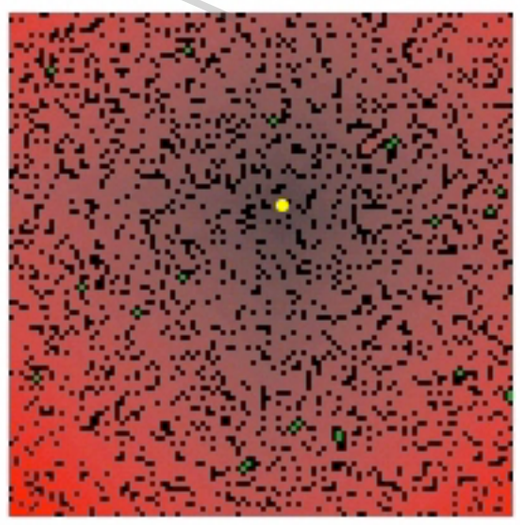




20%

40%

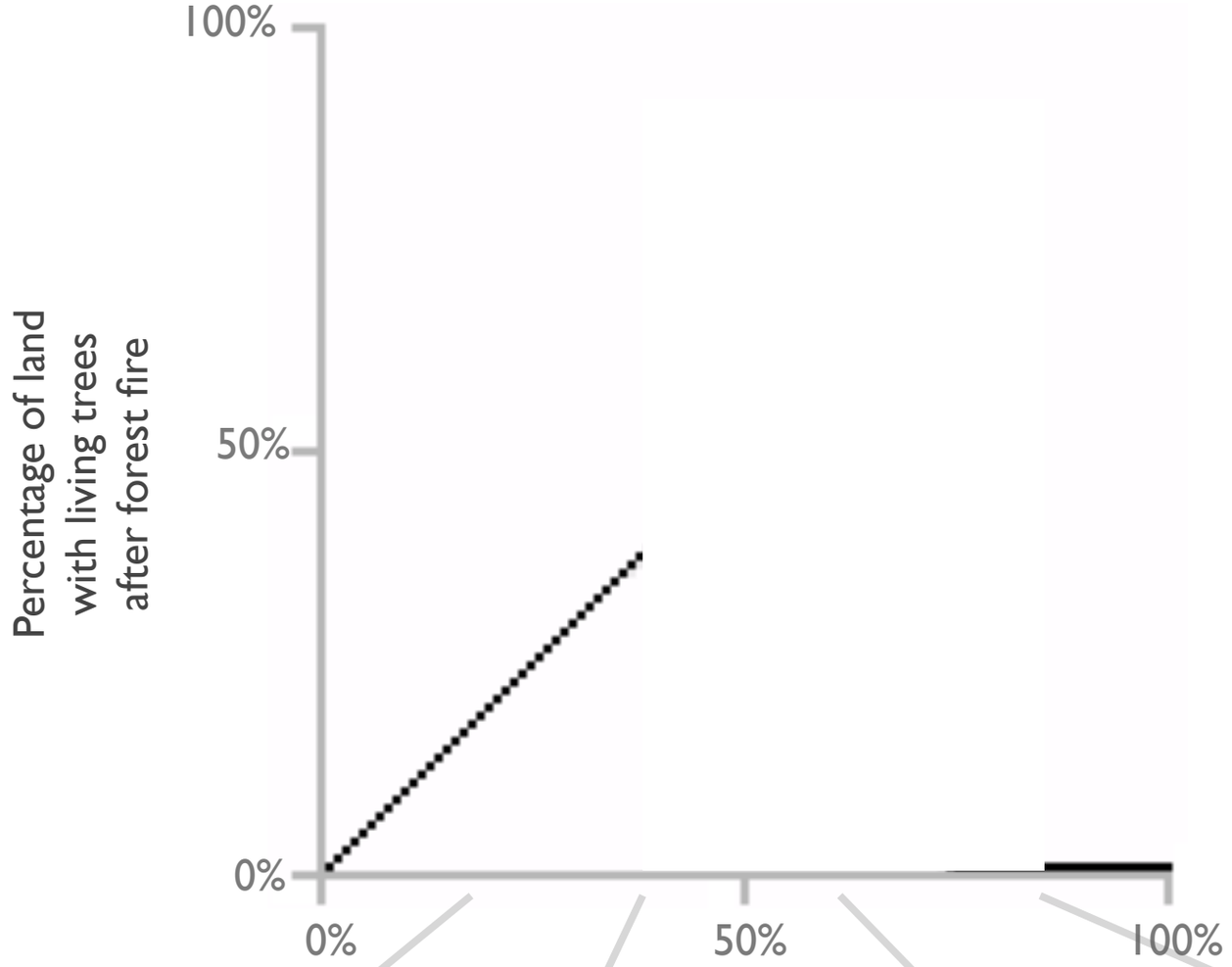
60%

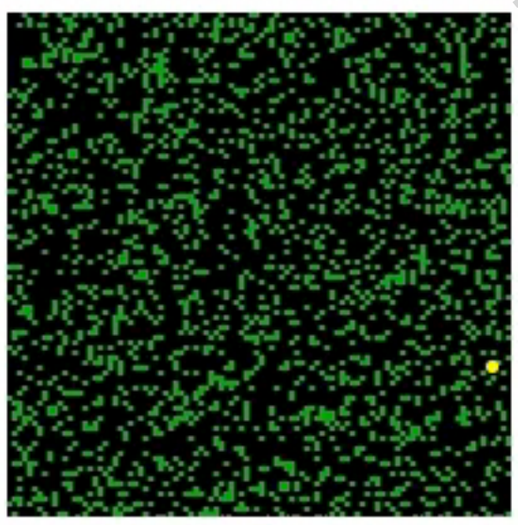


80%

Complexity: Forest fire

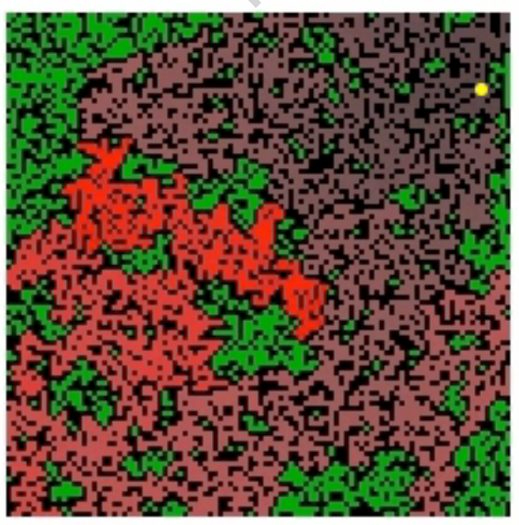
cdc



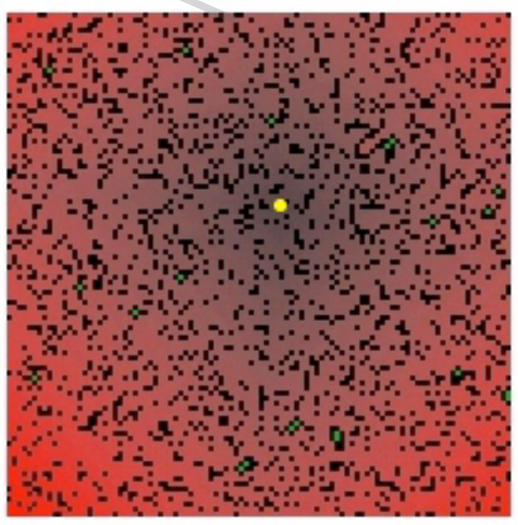


20%

40%



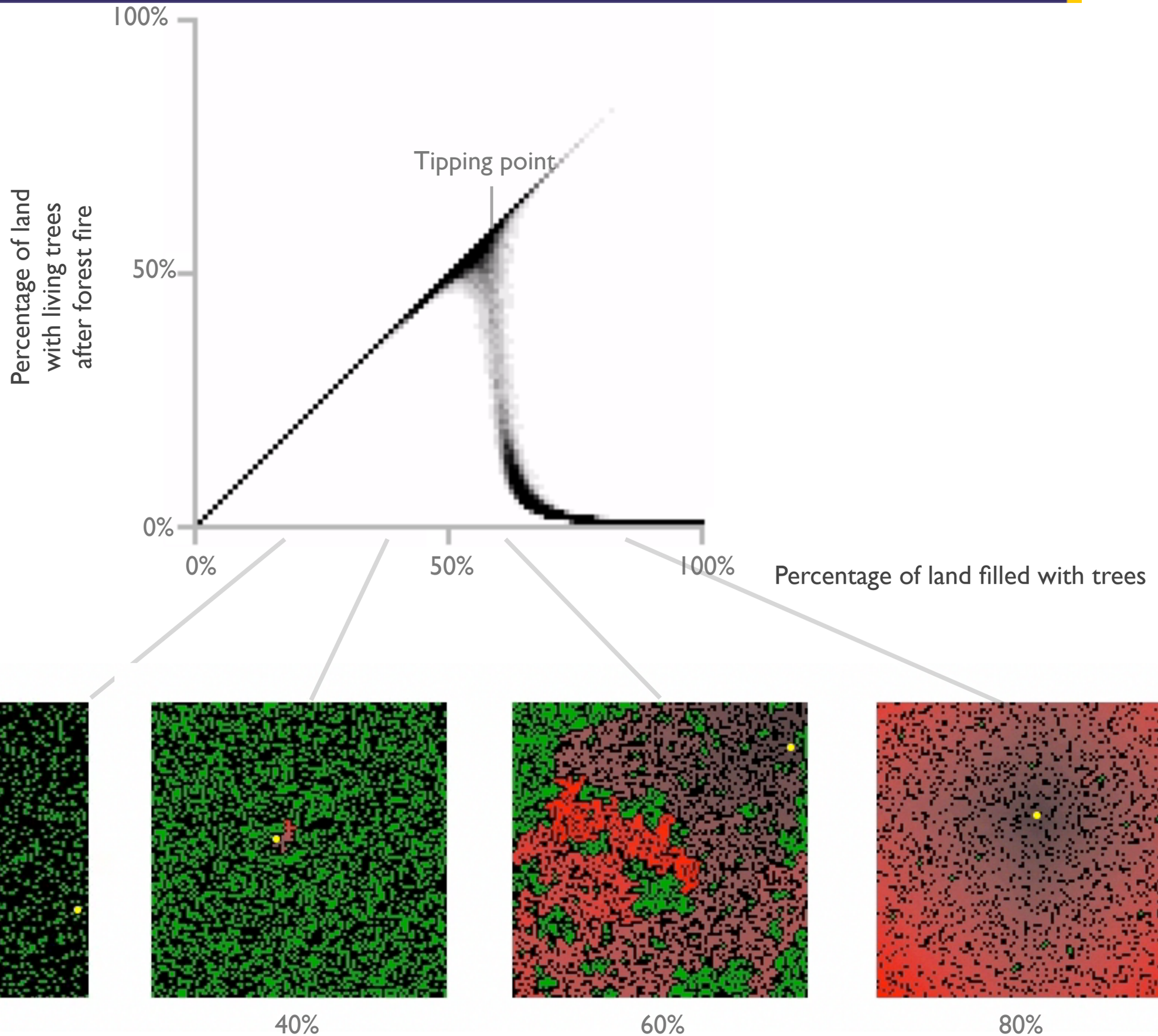
60%



80%

Complexity: Forest fire

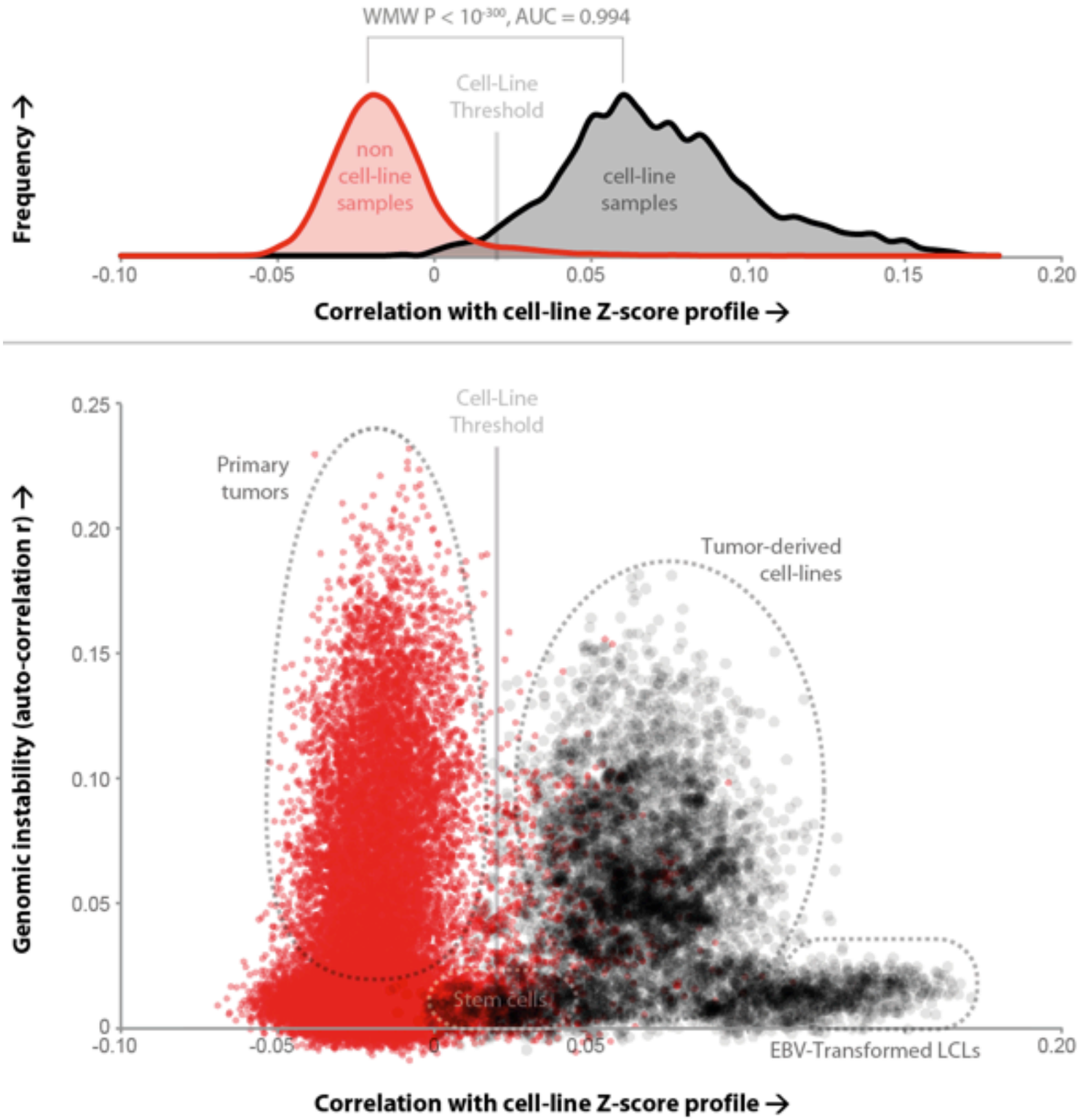
cdc



Find out which are cell line samples

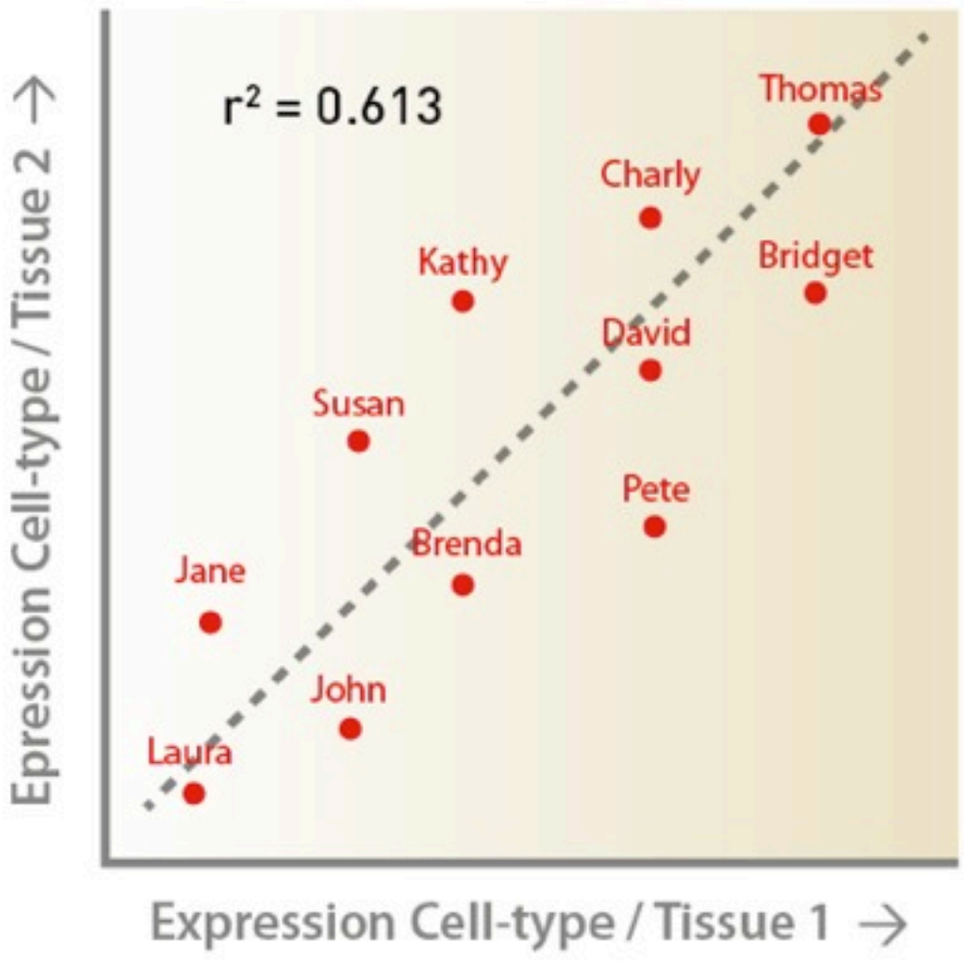
- Text mining analysis on Affymetrix U133 Plus 2.0 platform (54,000 probesets): 7,319 cell line samples
- Find additional samples that have been missed by text mining but which look similar to cell lines. Can we do this?
- Strategy: Perform for every component a T-Test, and denote T value per component. We have 777 components, and thus have a vector of 777 T values. Subsequently we correlate all 37,427 samples with this profile.

Predict cell line samples

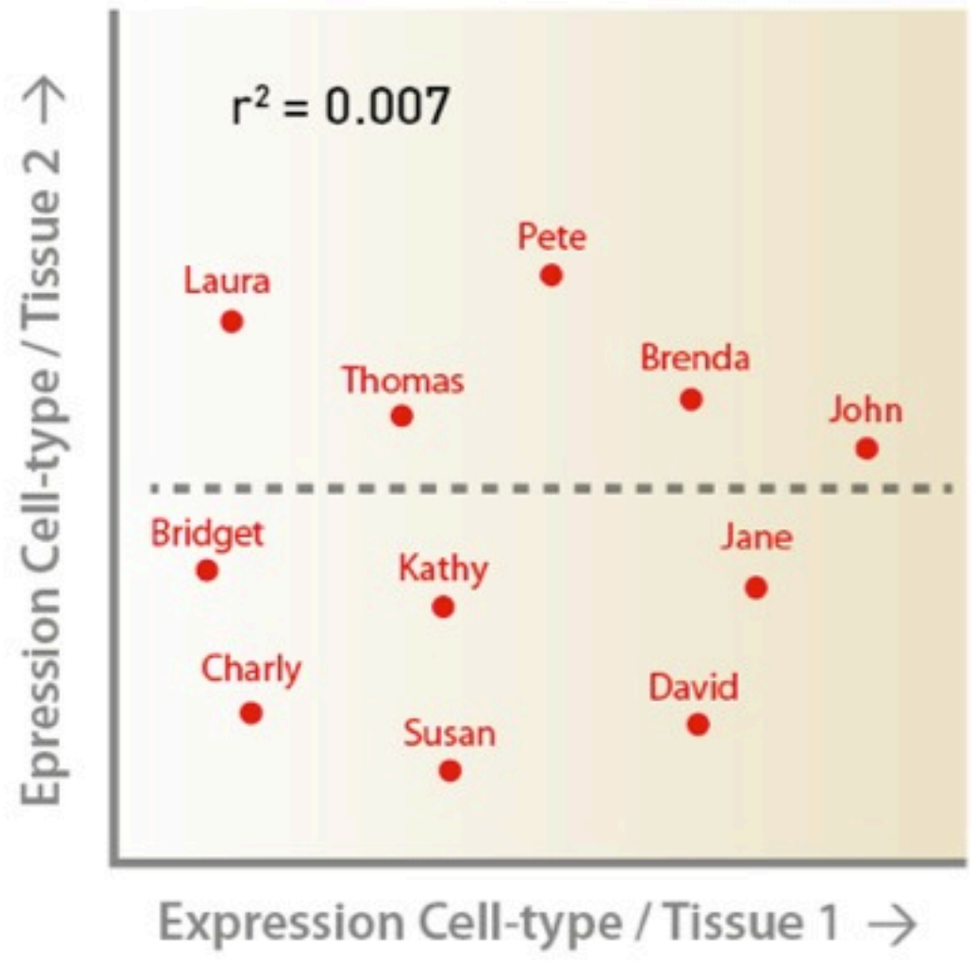


Identify related individuals

Probeset, informative for relating genetically identical samples:

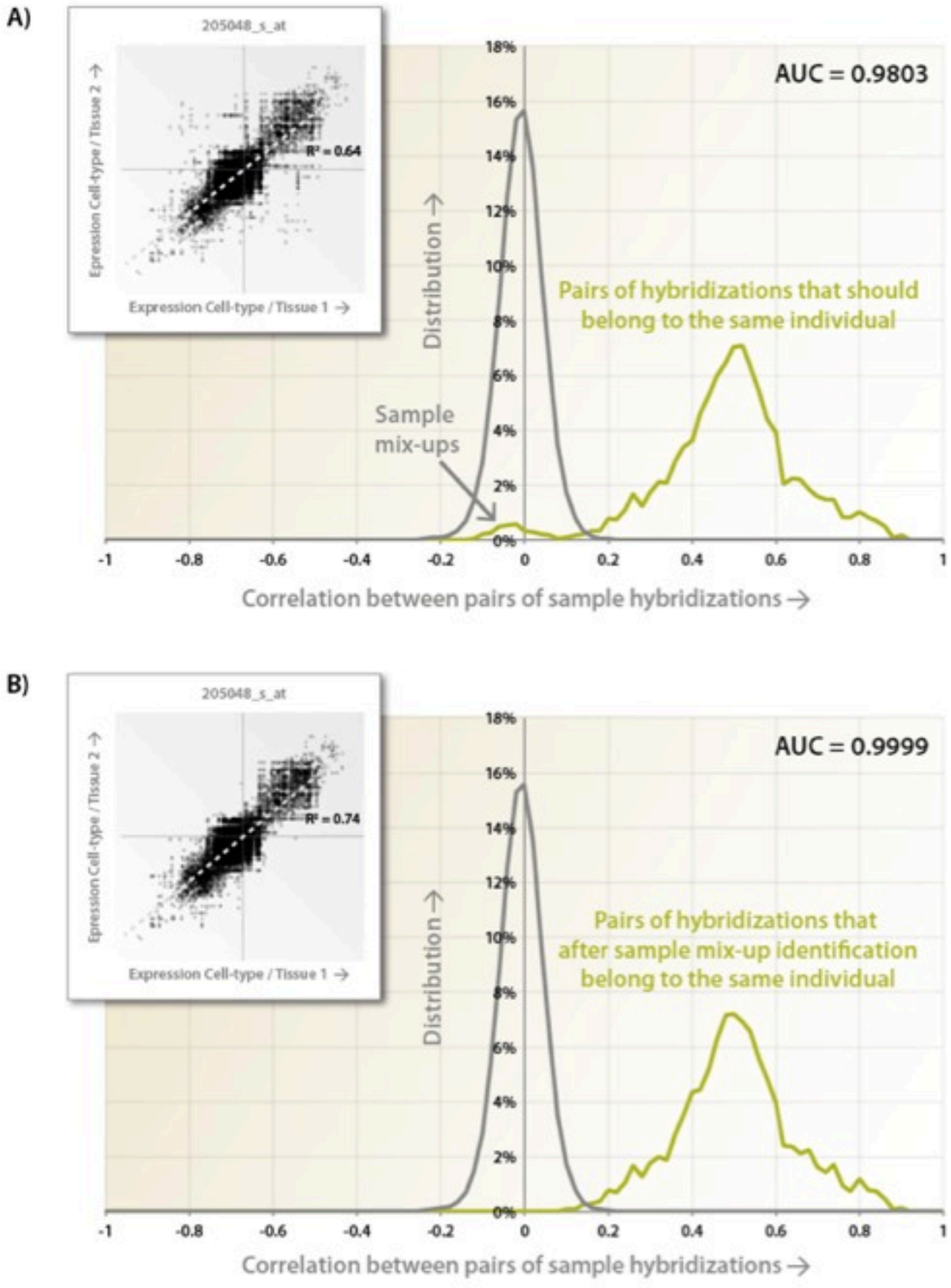


Probeset, not informative for relating genetically identical samples:



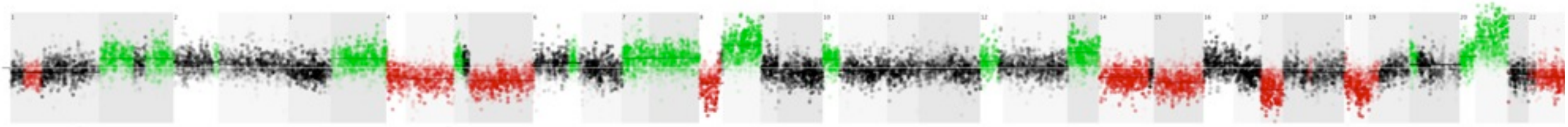
Extending upon sample mix-up identification method: Westra *et al*, Bioinformatics 2011

Identify related individuals

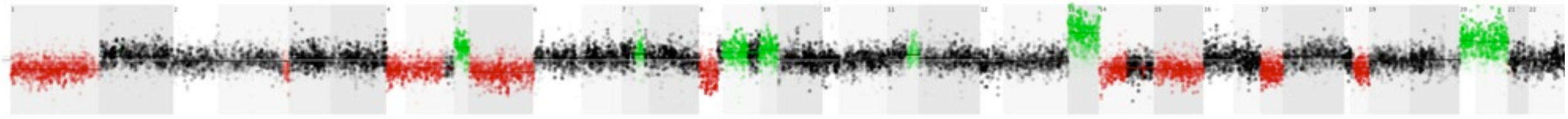


Combinations of dels/dups in many tumors

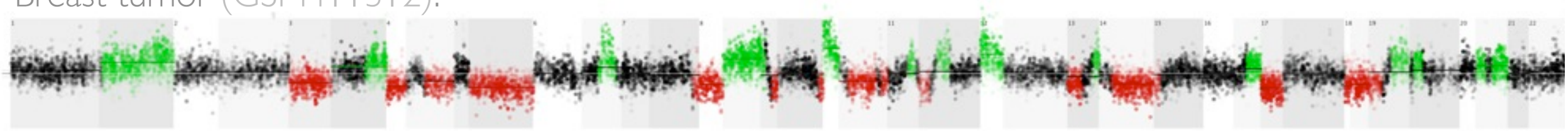
Profile most often identified in 16,172 cancer samples:



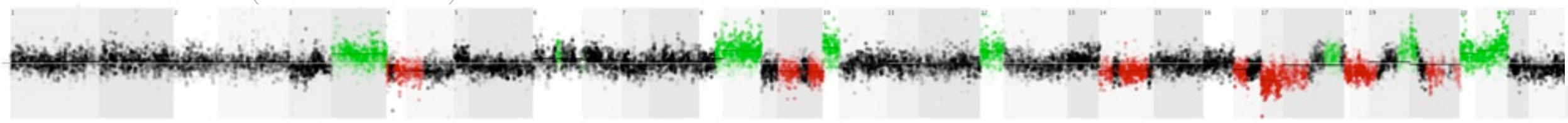
Colorectal tumor (GSM358505):



Breast tumor (GSM411312):

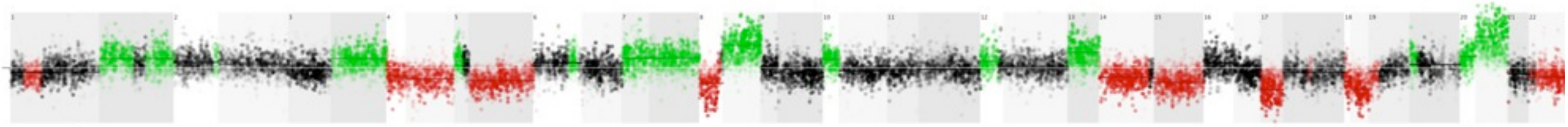


Ovarian tumor (GSM249825):

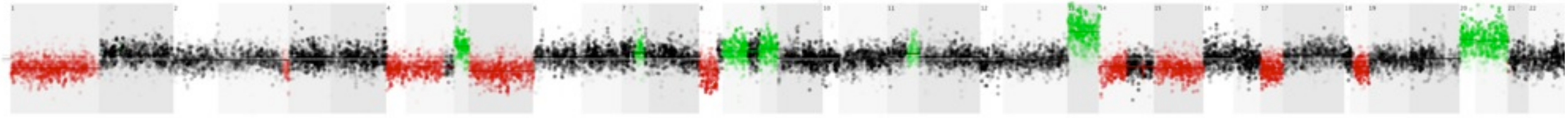


Combinations of dels/dups in many tumors

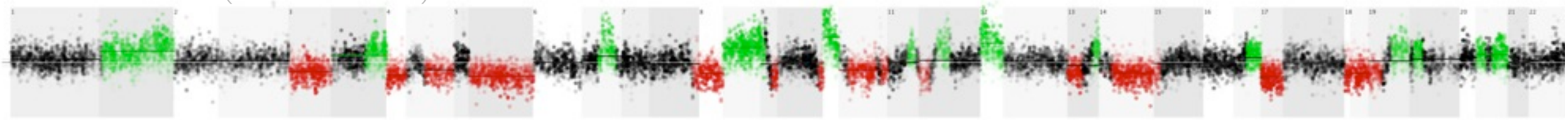
Profile most often identified in 16,172 cancer samples:



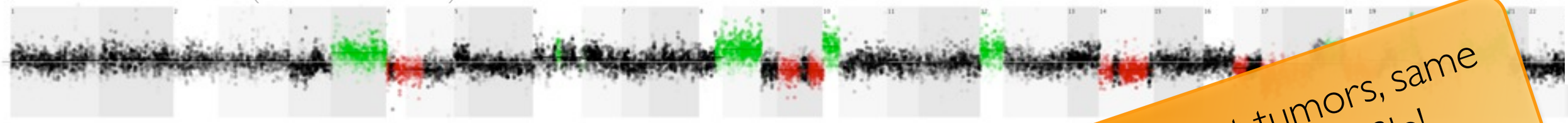
Colorectal tumor (GSM358505):



Breast tumor (GSM411312):

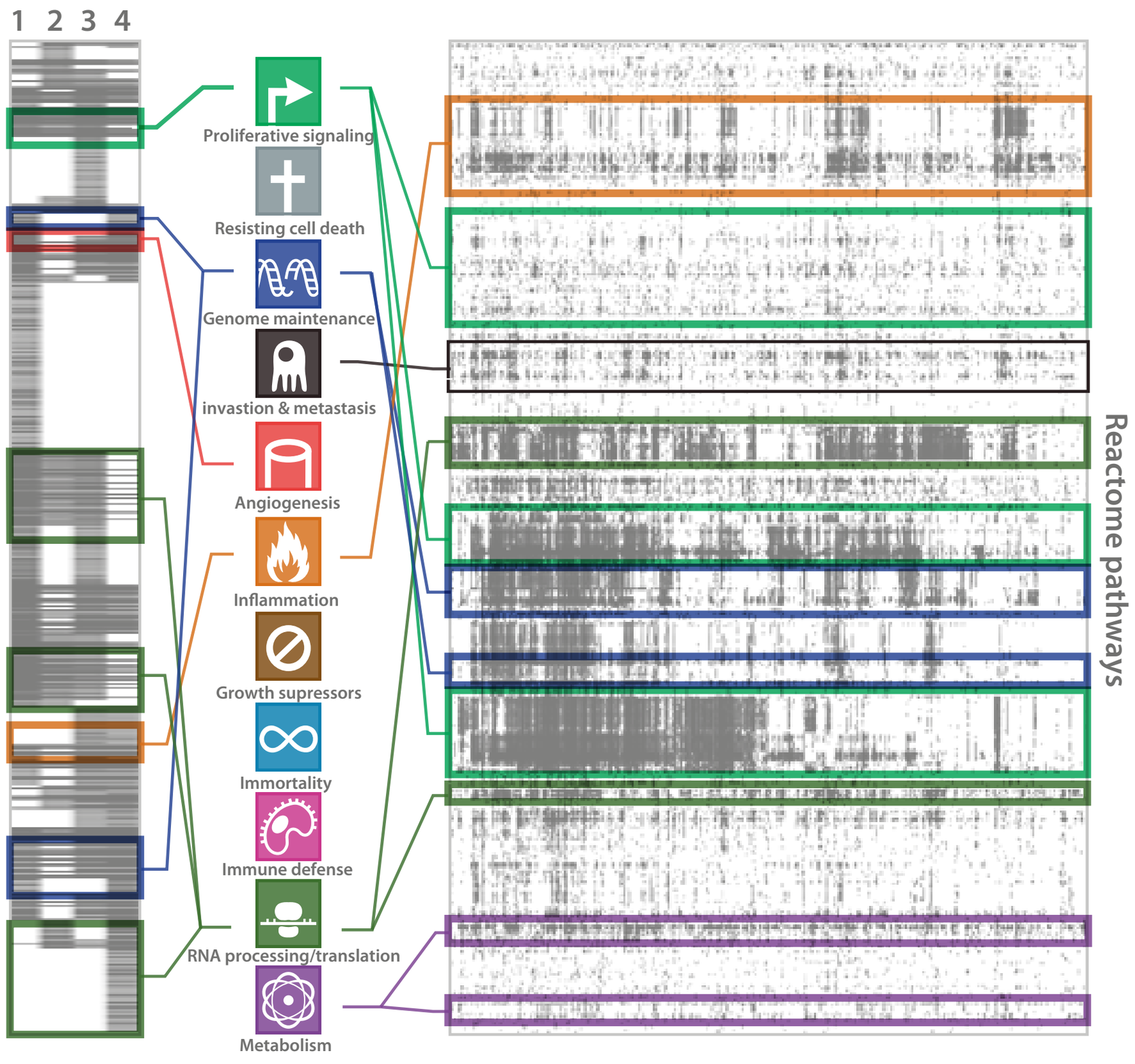


Ovarian tumor (GSM249825):

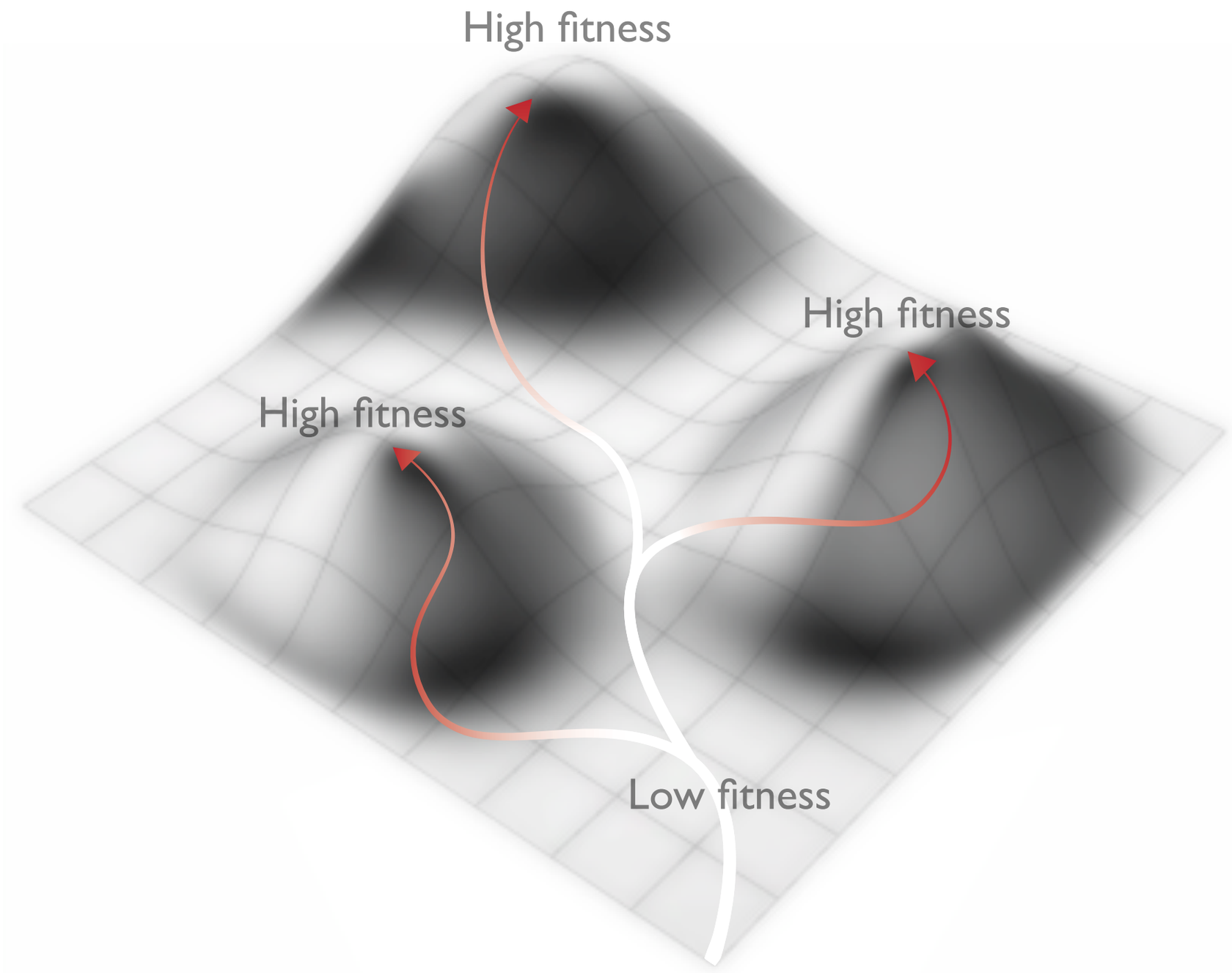


Different tumors, same cytogenetic profile!

Trans-eQTL mapping in 16,172 samples



Strong dependencies exist within cancers



Conclusions

- Reanalysis of publicly available data can reveal new insight into transcriptional regulation and genomic instability in cancer
- Many avenues for bioinformaticians and statistical geneticists!