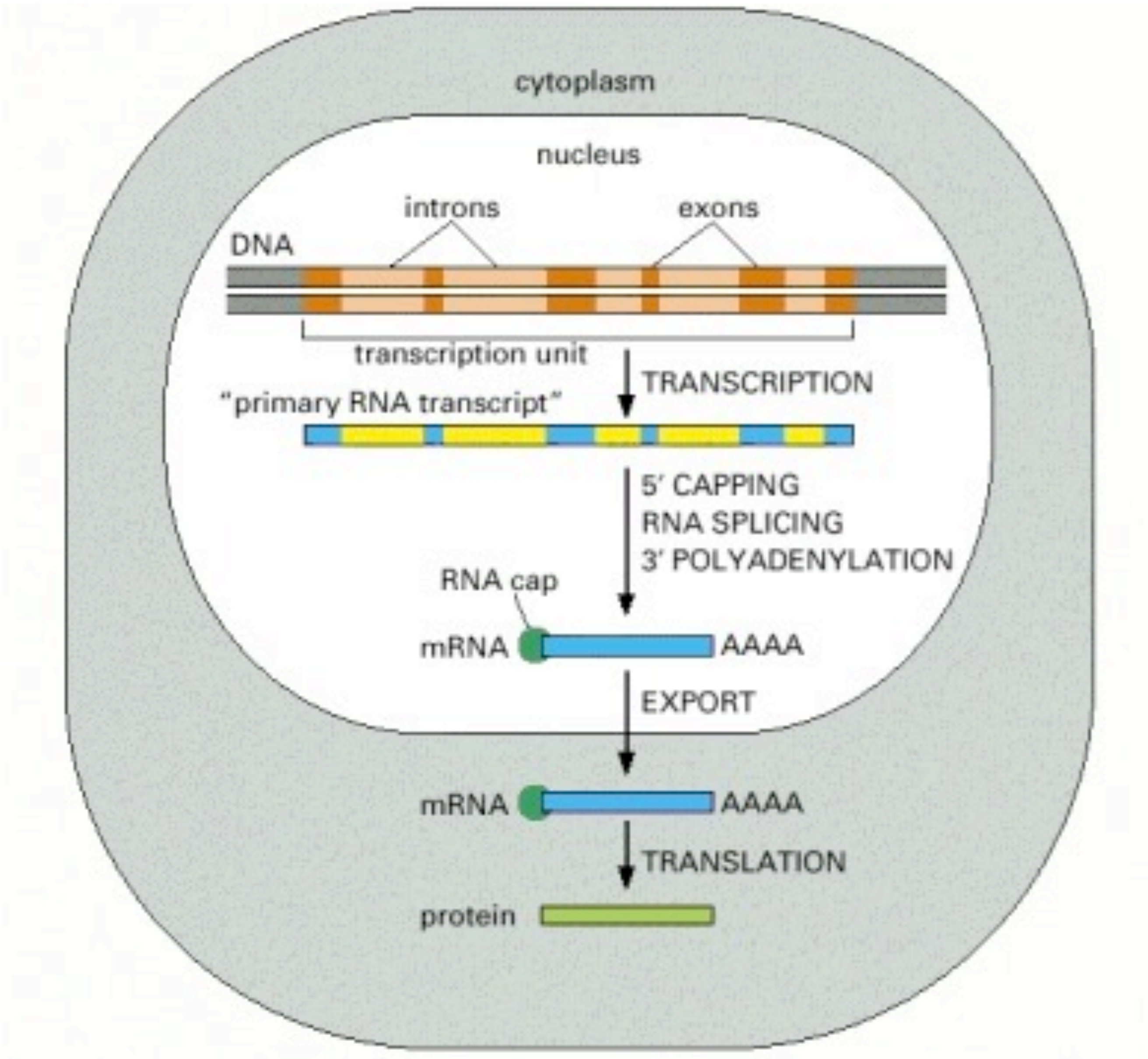
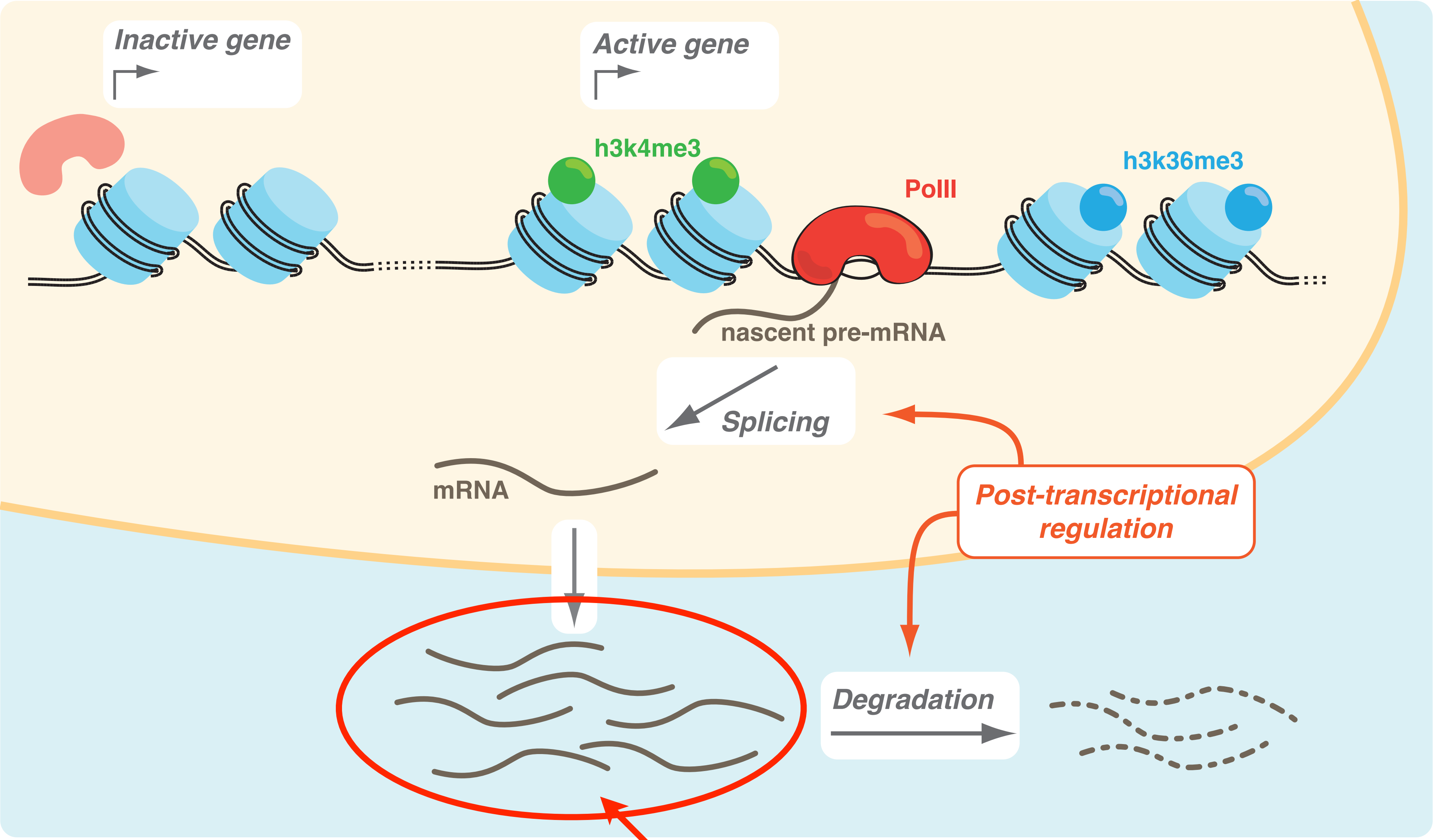
Transcription



Figures from Alberts and co, Mol Biol Cell.

Transcription



Equilibrium pool of mRNA for each gene

Simplest model

Transcription rate: $P(t)$ [#mRNA/time]
mRNA pool: $m(t)$ [#mRNA]
Degradation rate: γ [1/time]

$$\dot{m}(t) \; = \; P(t) - \gamma m(t) \; .$$

Case 1: $P(t) = P_0$

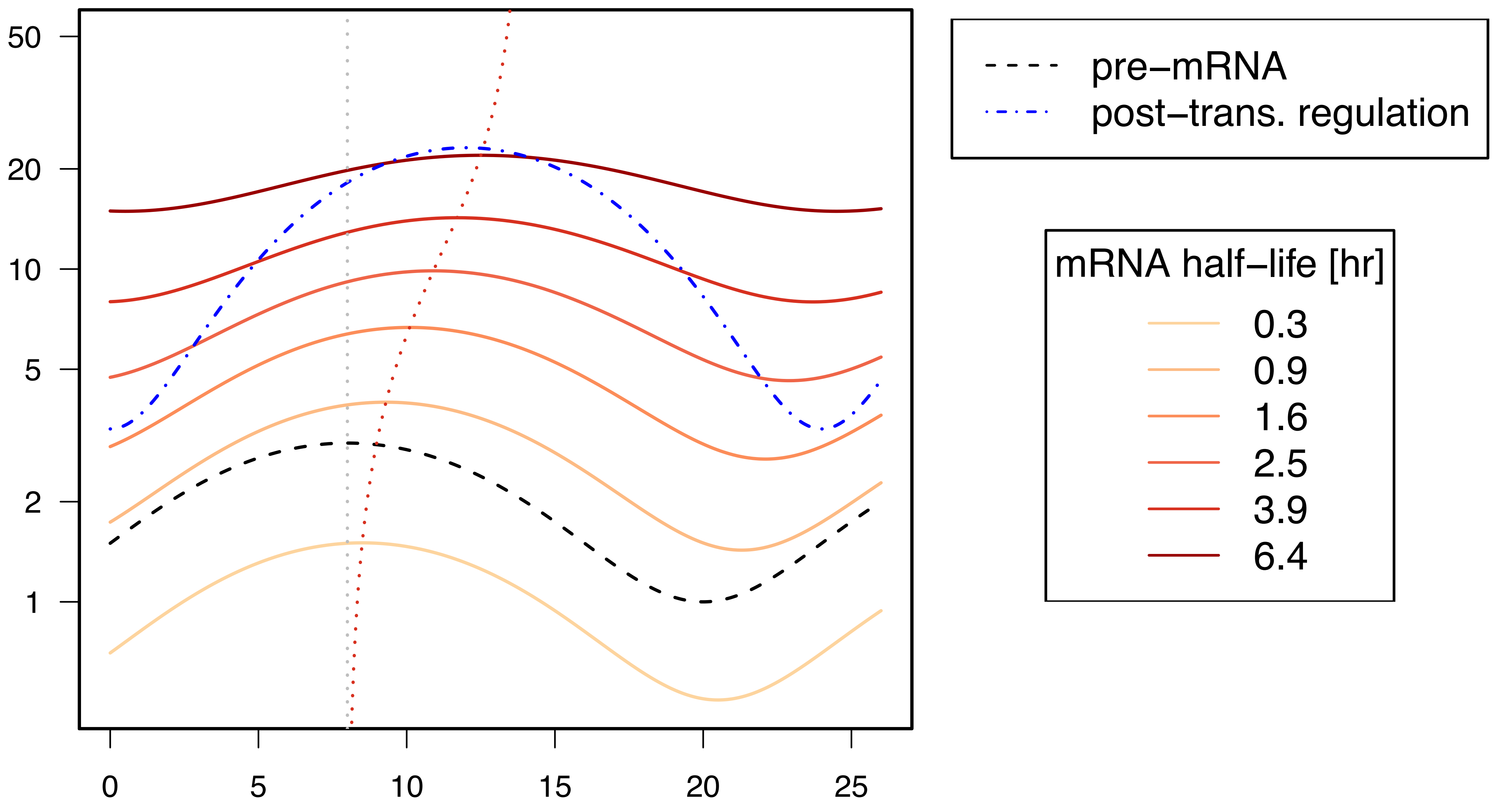
$$m(t) \; = \; \frac{P_0}{\gamma}(1 - e^{-\gamma t}) + e^{-\gamma t} m(0) \; \rightarrow \; \frac{P_0}{\gamma} \; .$$

Case 2: $P(t) = P_0 + \cos(\omega t)$

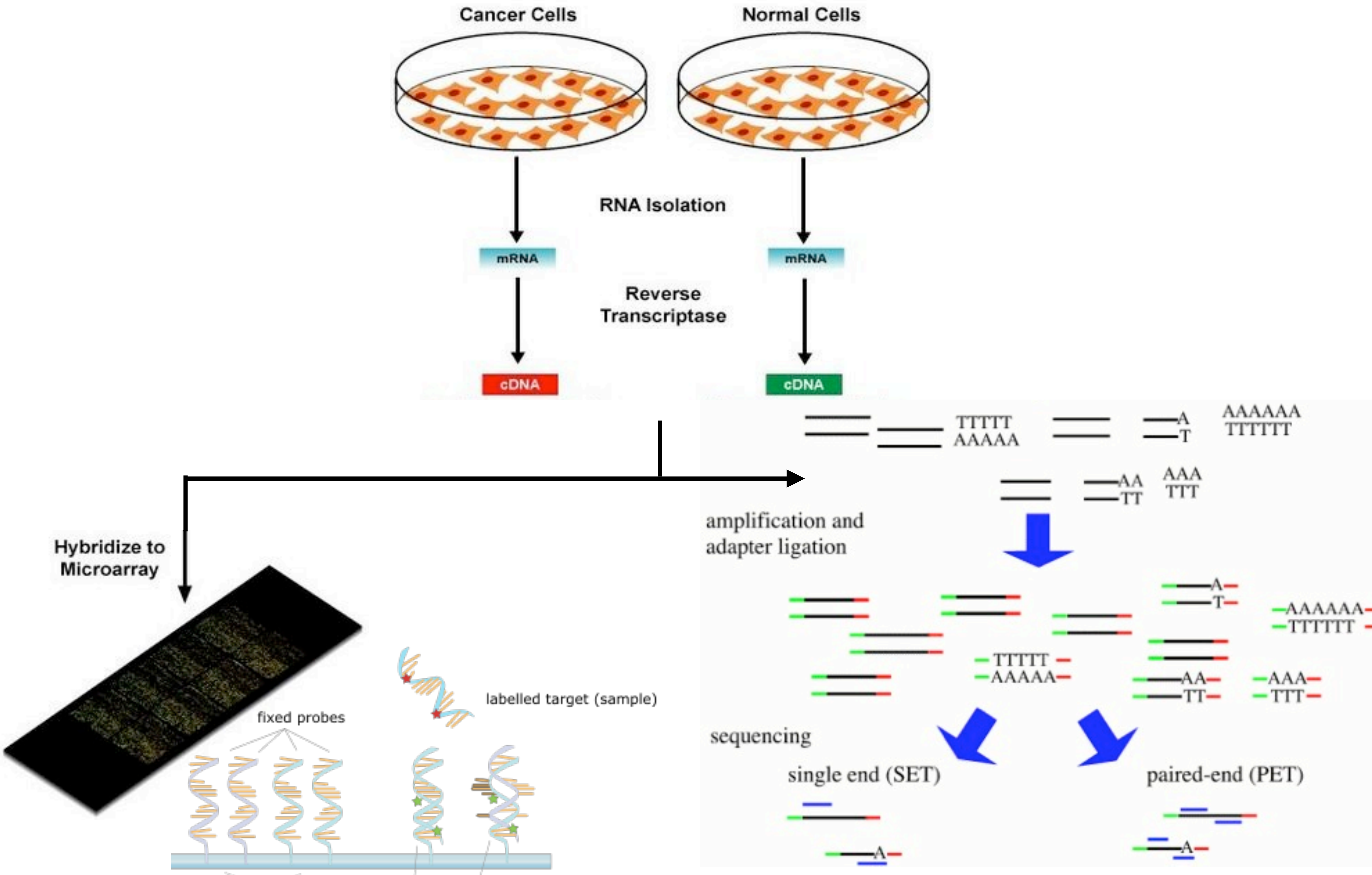
$$m(t) \; \rightarrow \; \frac{P_0}{\gamma} + \frac{1}{\sqrt{\gamma^2 + \omega^2}} \cos(\omega(t - \tau)) \; .$$

with $\sin \omega \tau = \frac{\omega}{\sqrt{\gamma^2 + \omega^2}}$

Simplest model



Measuring expression



Cancer Cells

Normal Cells

RNA Isolation

mRNA

mRNA

Reverse Transcriptase

cDNA

cDNA

Hybridize to Microarray

fixed probes

labelled target (sample)

amplification and adapter ligation

sequencing

single end (SET)

paired-end (PET)

Measuring expression

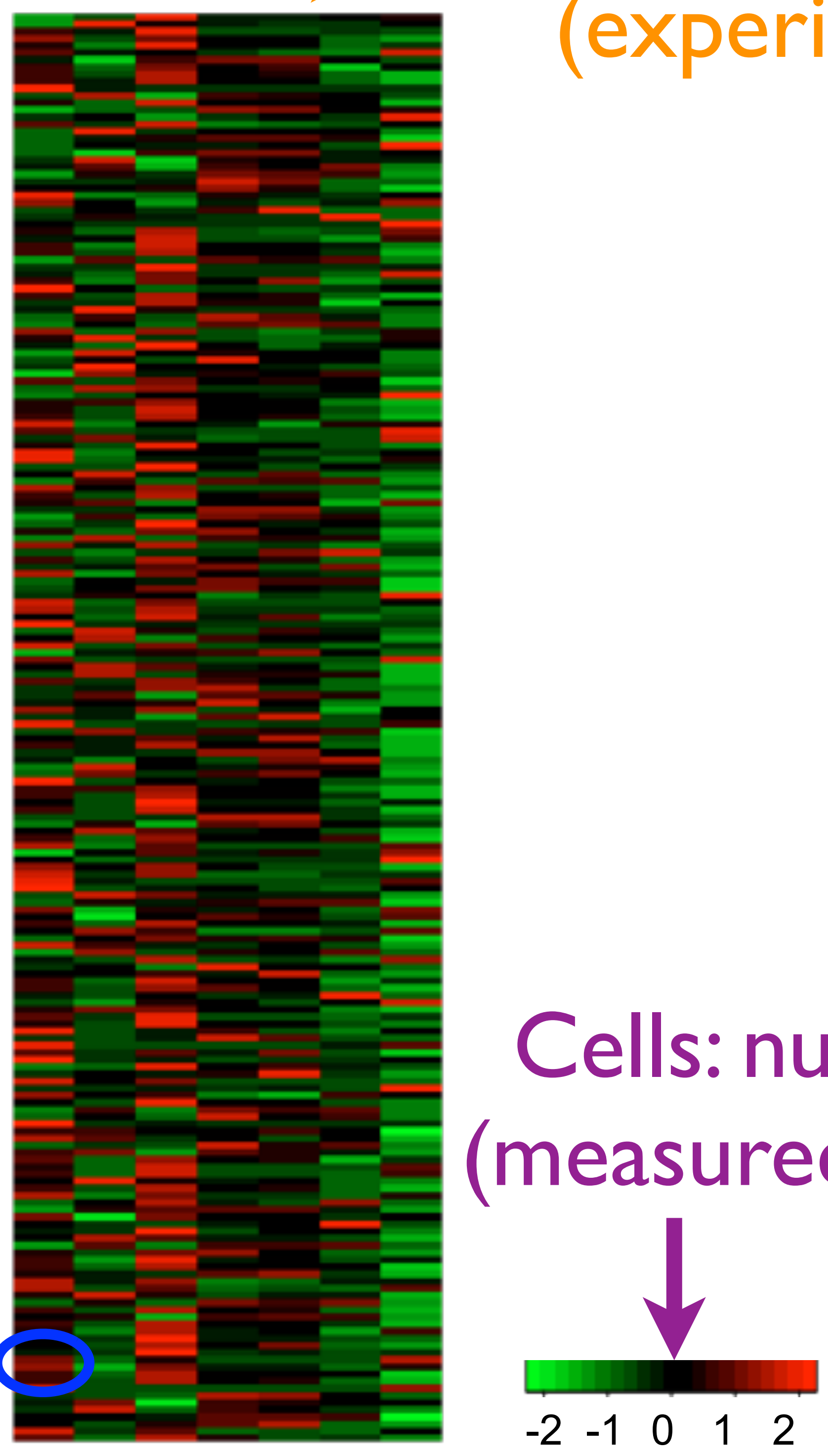
- Goal: for each possible transcript **T**, estimate number of copies of **T** per cell
- We start from a large (10^6-10^7) population of cells and take a sample of the total mRNA pool
- We then measure a signal which is proportional to the abundance of **T** in the sample
- We generally do not know the calibration but we can deduce ratios:

$$\begin{aligned}\frac{\mathrm{Expr}(g_1)}{\mathrm{Expr}(g_2)} &= \frac{N_{\mathrm{copy}}(g_1)/N_{\mathrm{cell}}}{N_{\mathrm{copy}}(g_2)/N_{\mathrm{cell}}} \\ &= \frac{\alpha \mathrm{Signal}(g_1)/N_{\mathrm{cell}}}{\alpha \mathrm{Signal}(g_2)/N_{\mathrm{cell}}} \\ &= \frac{\mathrm{Signal}(g_1)}{\mathrm{Signal}(g_2)}\end{aligned}$$

Output

Columns: conditions (experiments)

Rows: features (genes)

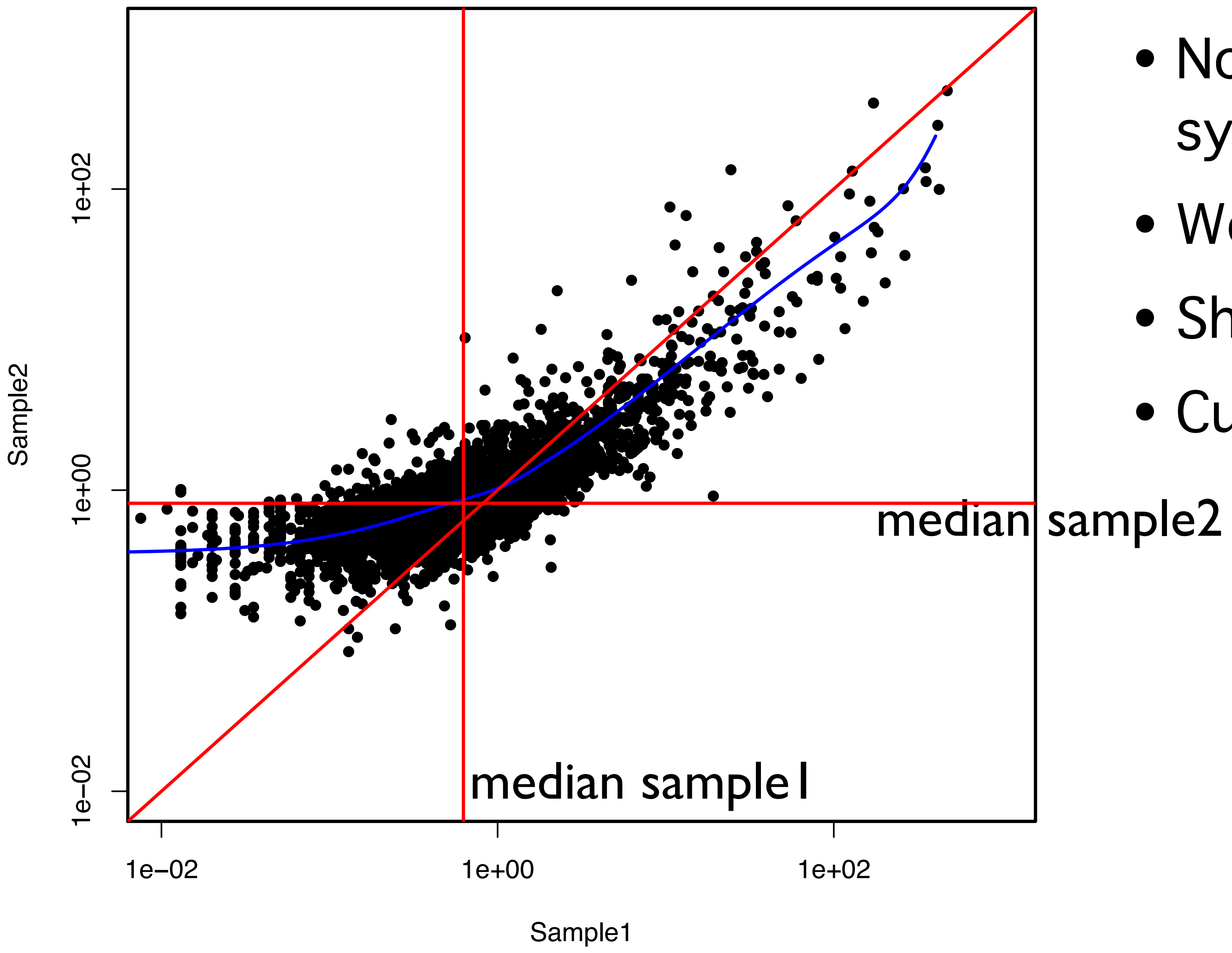


Cells: numbers (measured signal)

Signal is proportional to concentration of mRNA

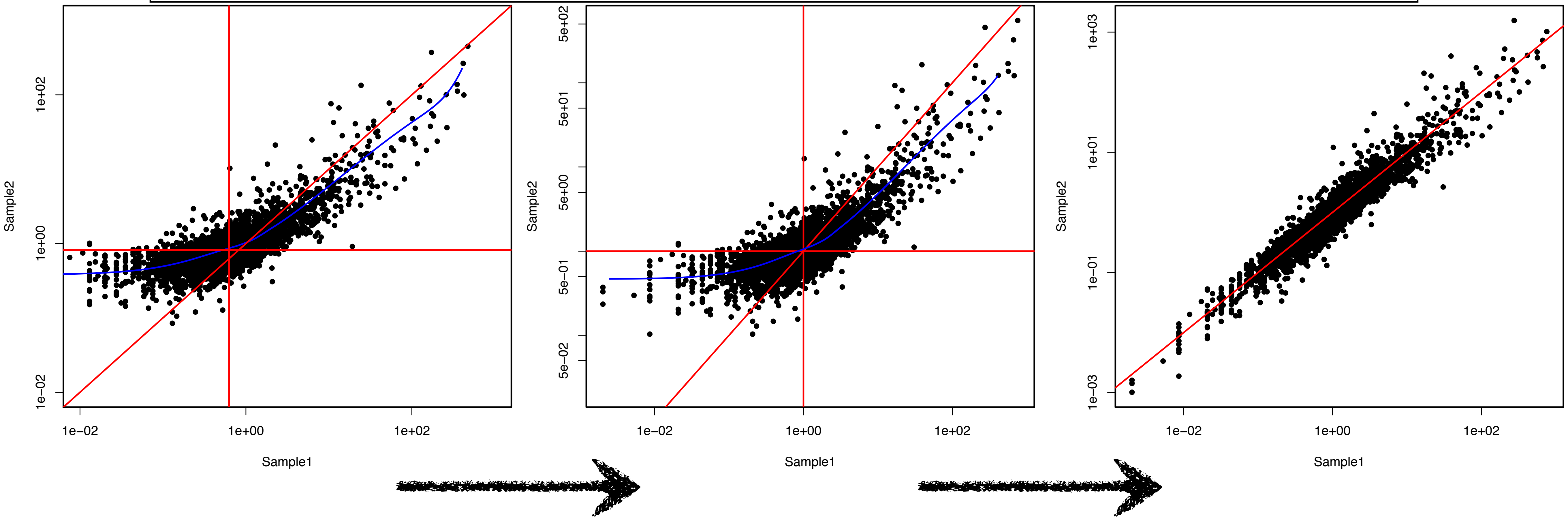
Normalization

Average variation between any 2 conditions is 0:
Systematic variation MUST BE technical artifact

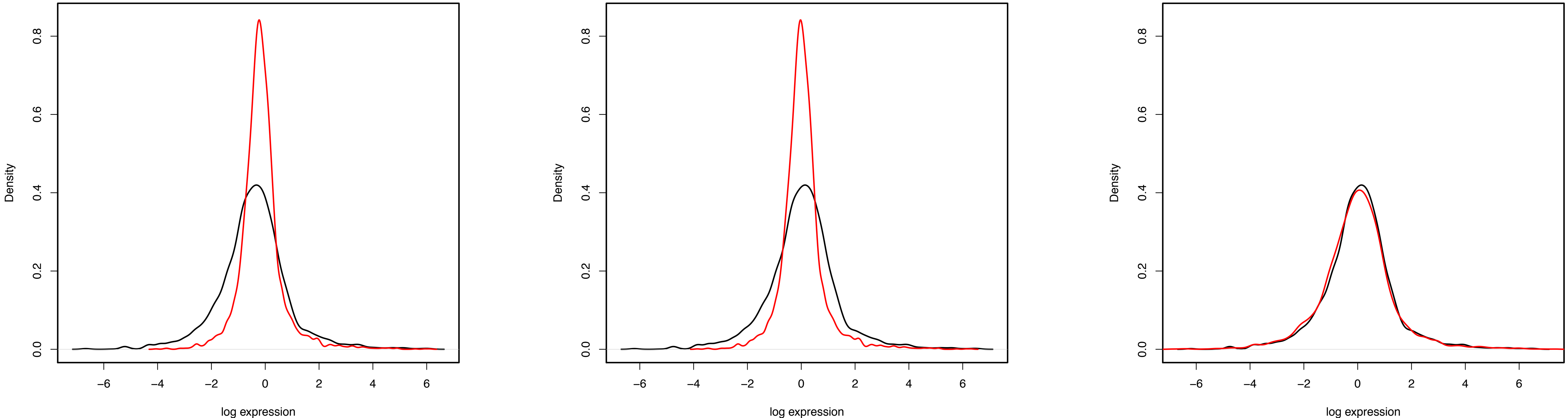


- Normalization consists in removing systematic variations
- Work in log-log coordinates
- Shift in medians
- Curved shaped

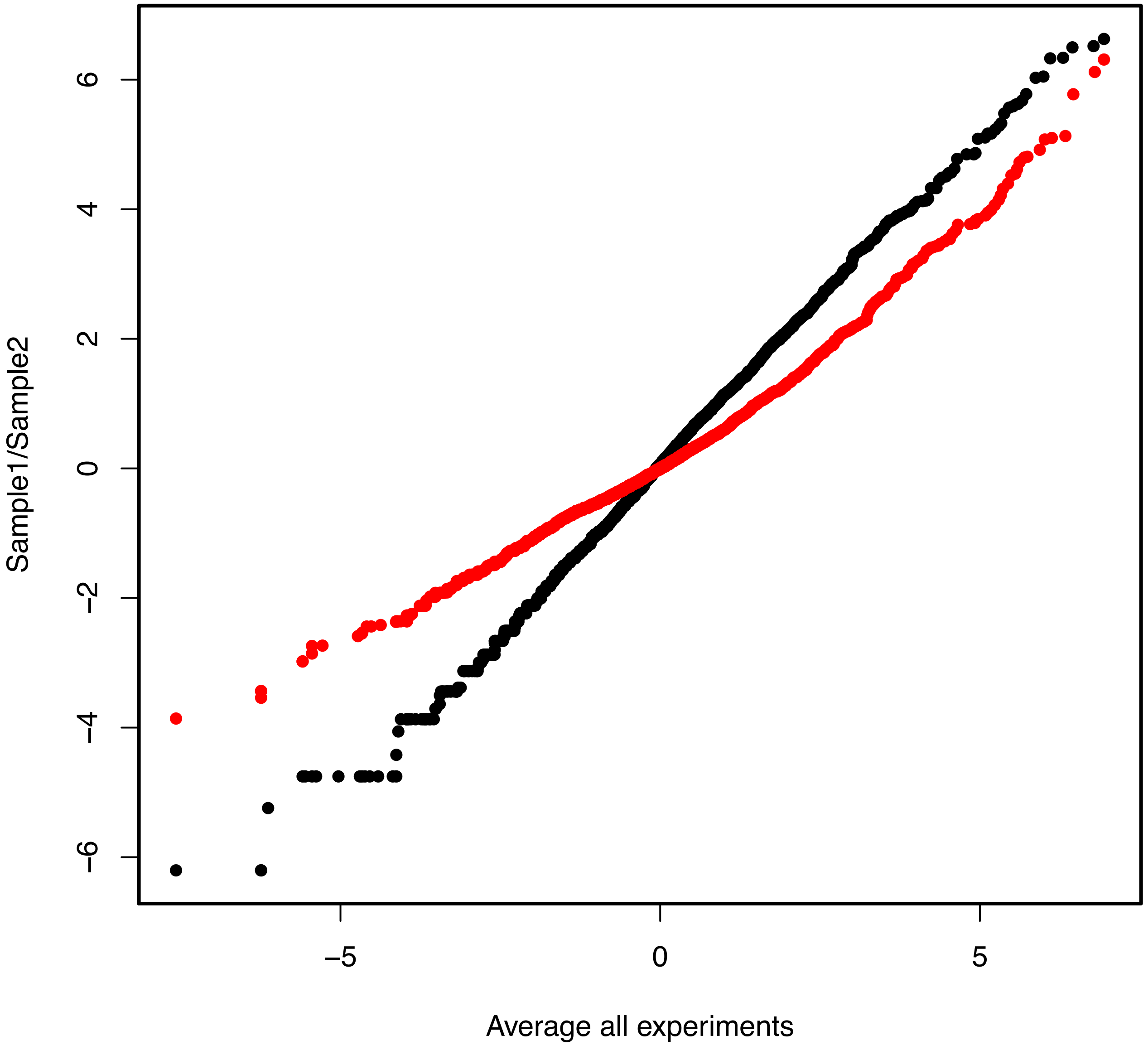
Normalization



sample=sample/median(sample) sample2=sample2/fit*sample1



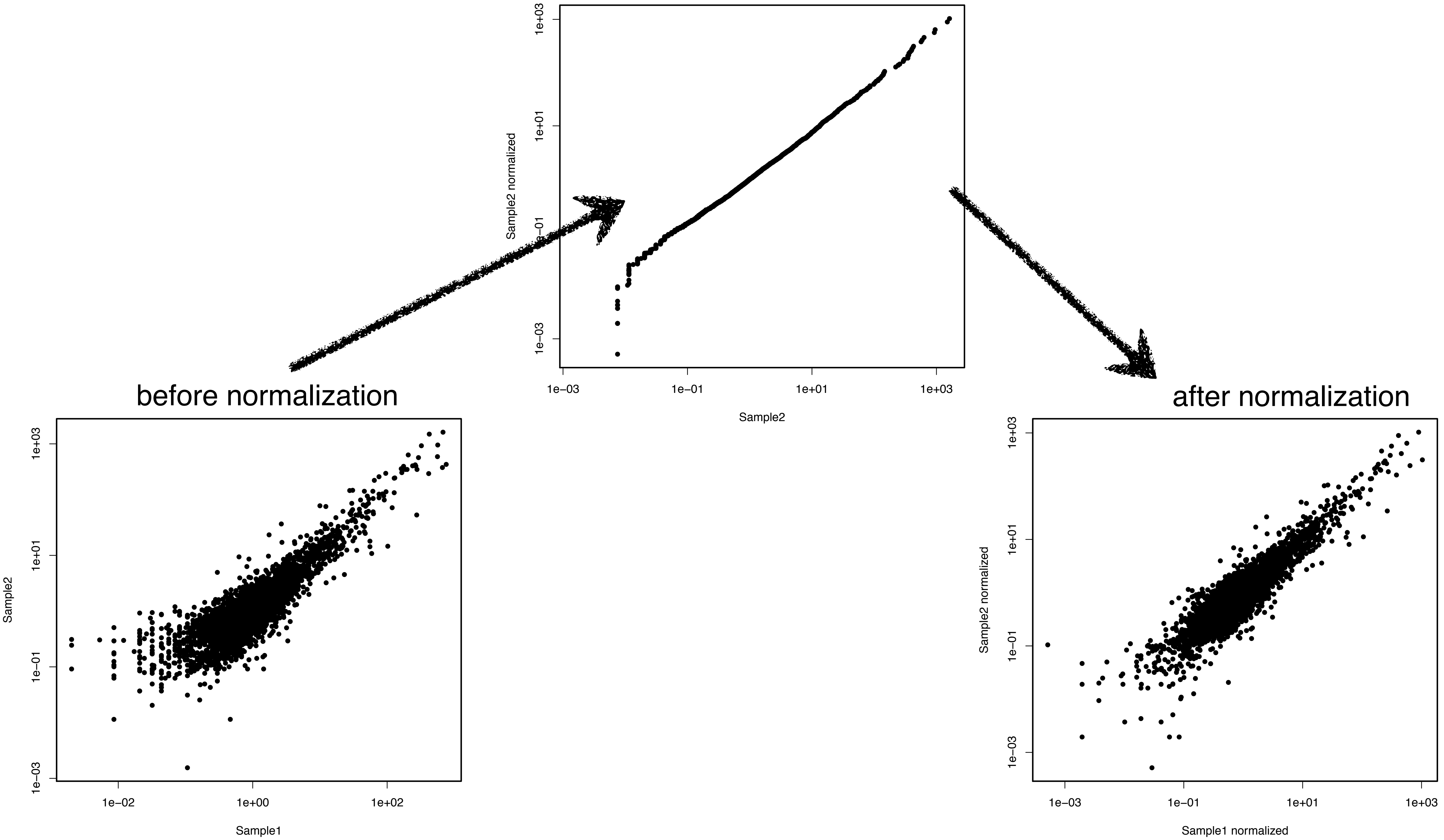
Quantile normalization



order all data

- Substitute ordered values from every sample with ordered values from average (or from specific distribution, e.g. gaussian)

Quantile normalization

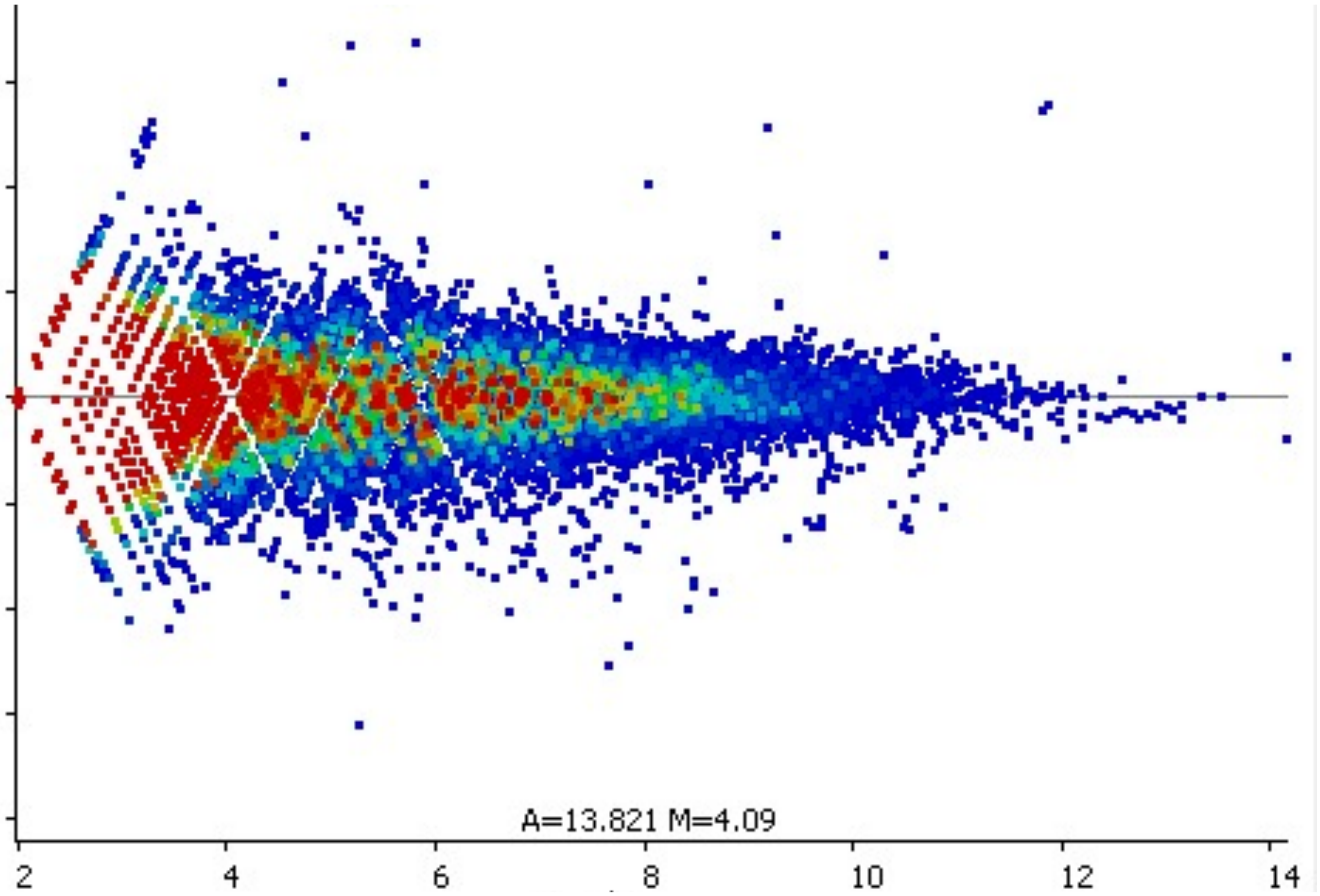


before normalization

after normalization

MA plot: differential expression

Log2(Sample1/Sample2)

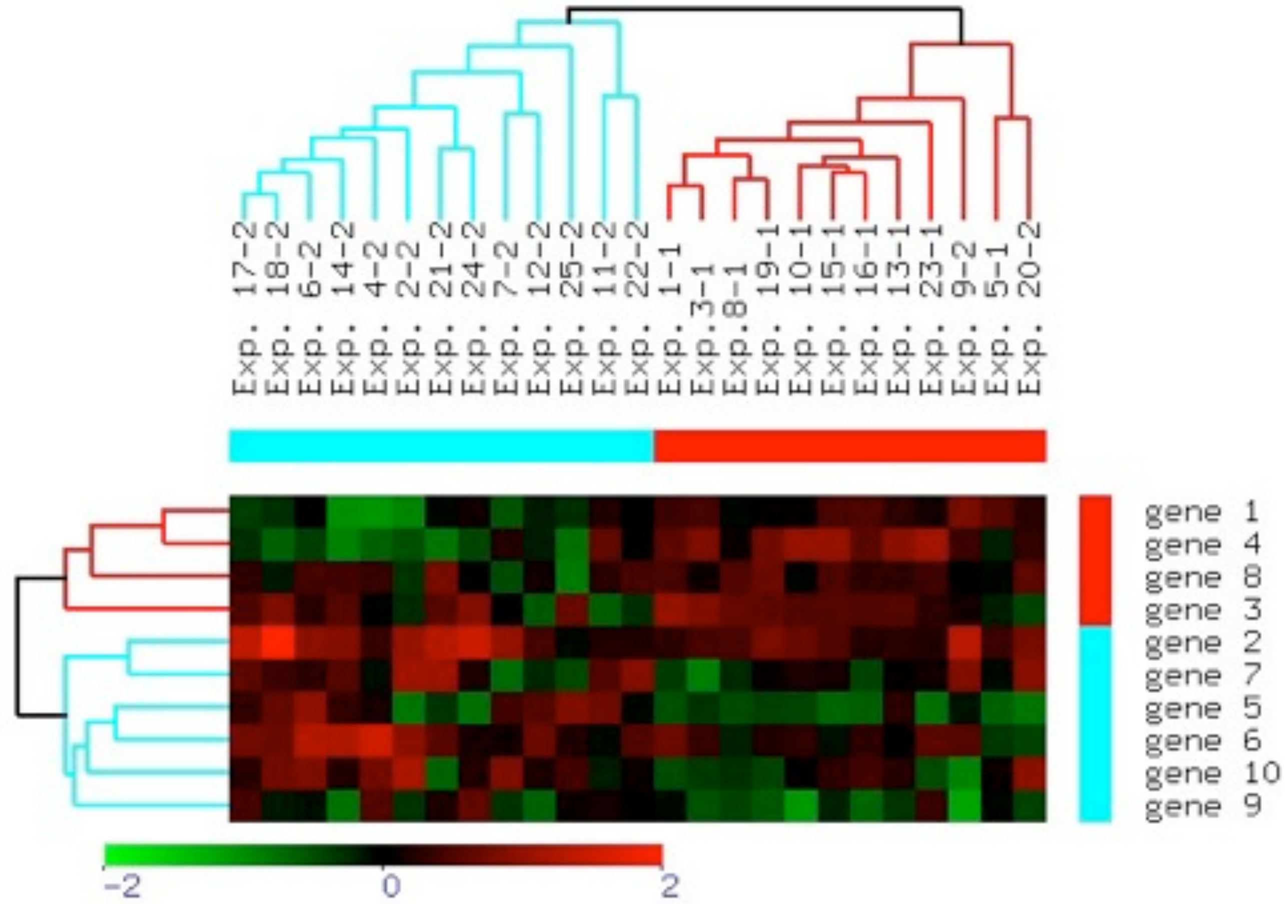


Log10(Sample1*Sample2)/2

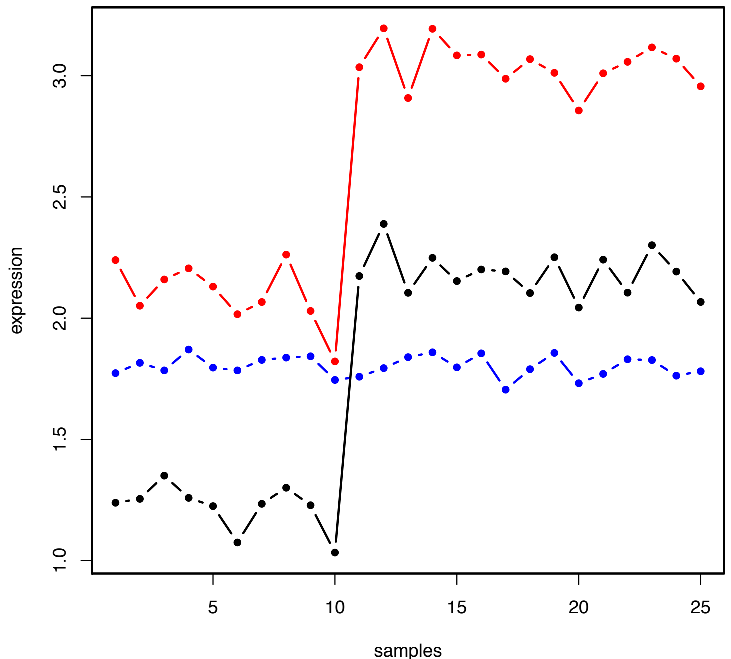
differential expression

absolute expression

Clustering



- Same algorithm as UPGMA
- Distance matrix is 1-cor(gene$_i$,gene$_j$)
- Update matrix with distance to average of two groups weighted by size
- Do the same for columns (rotate matrix)



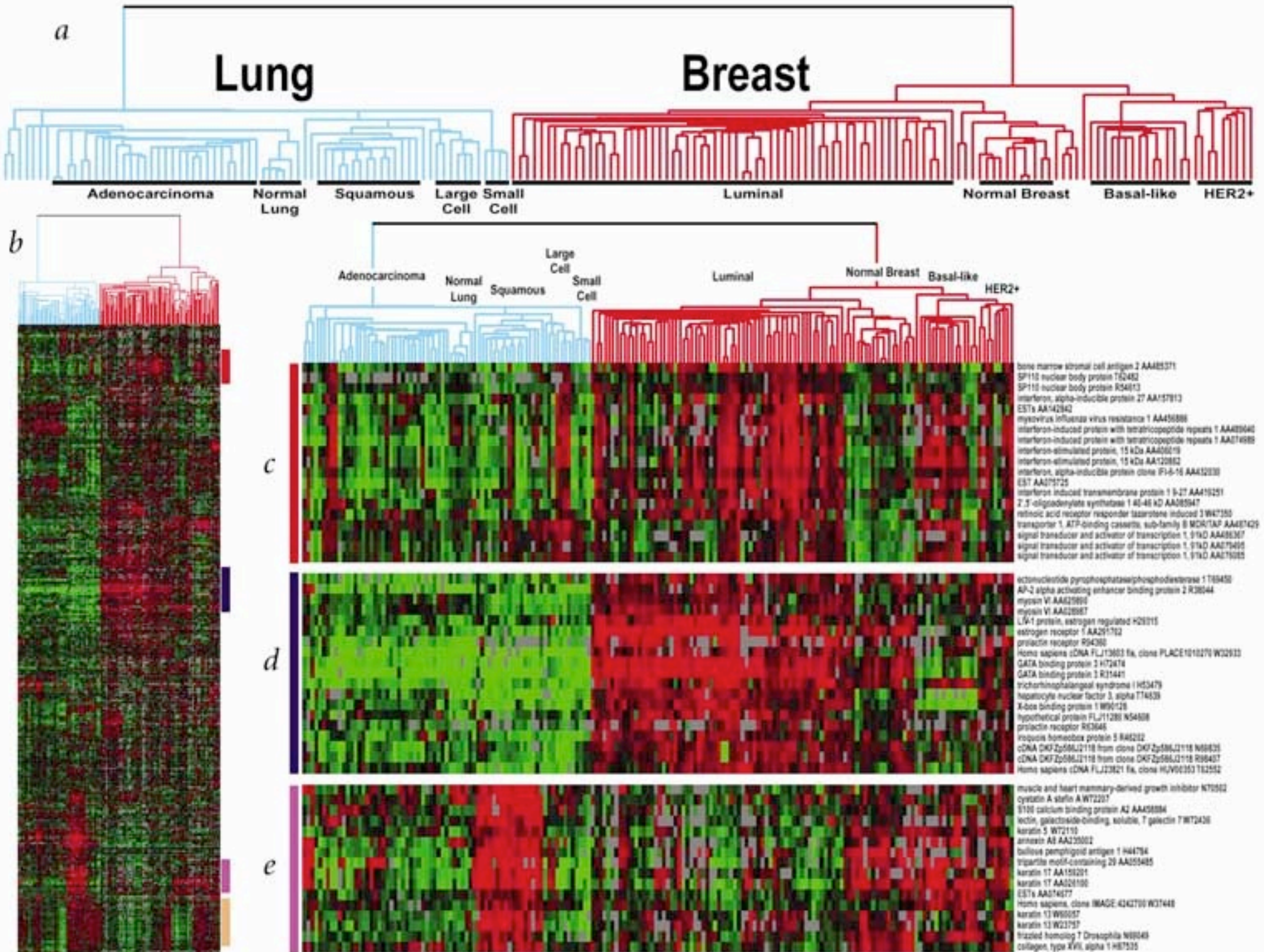
distance

	1	2
2	4.3	
3	2.4	4.9

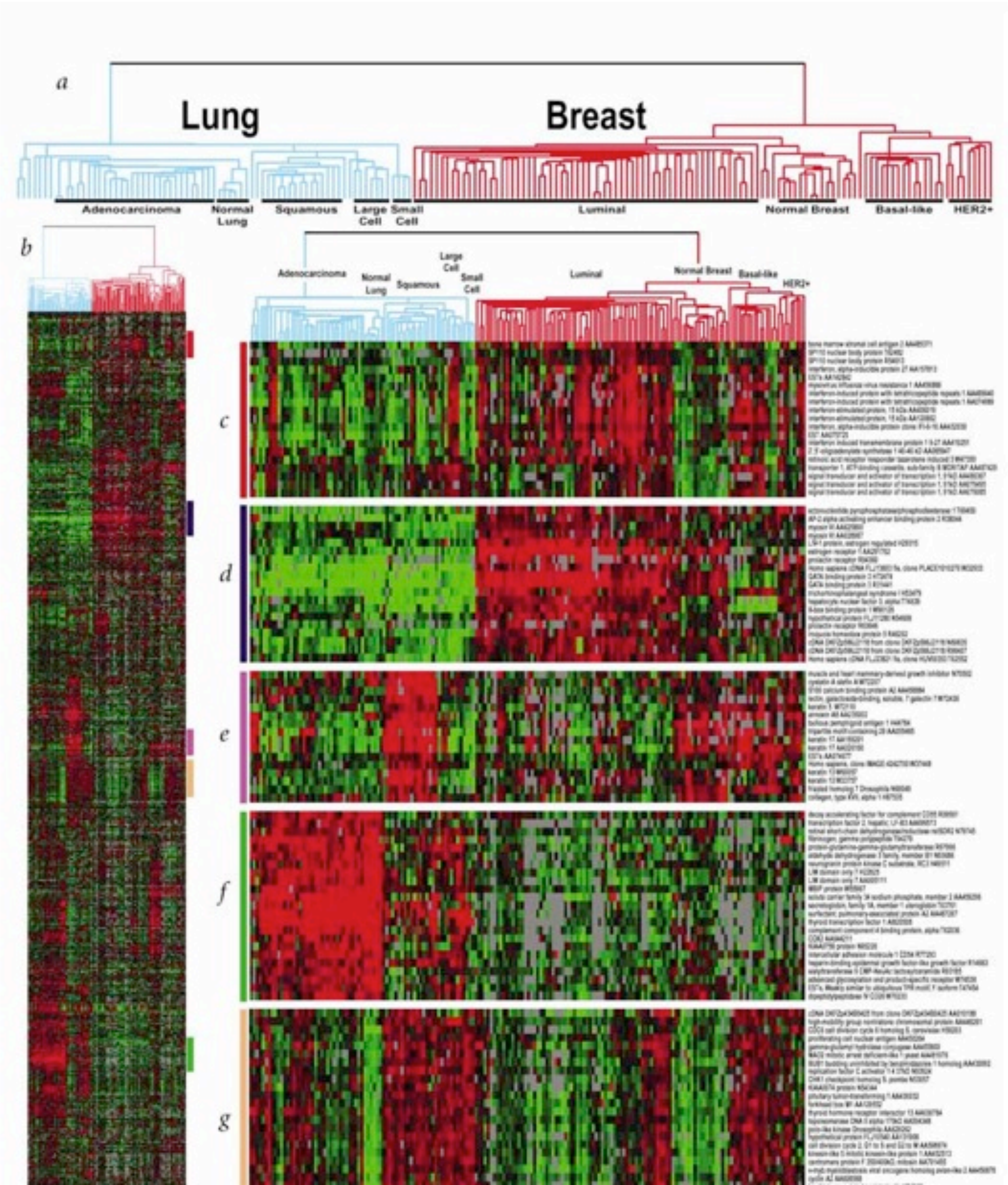
correlation

	1	2
2	0.99	
3	-0.06	-0.03

Clustering


a
Lung
Breast
Adenocarcinoma
Normal Lung
Squamous
Large Cell
Small Cell
Luminal
Normal Breast
Basal-like
HER2+
b
Adenocarcinoma
Normal Lung
Squamous
Large Cell
Small Cell
Luminal
Normal Breast
Basal-like
HER2+
c
d
e
bone marrow stromal cell antigen 2 AA485371
SP110 nuclear body protein T62482
SP110 nuclear body protein R54613
interferon, alpha-inducible protein 27 AA157813
ESTs AA142942
myxovirus influenza virus resistance 1 AA456886
interferon-induced protein with tetratricopeptide repeats 1 AA489640
interferon-induced protein with tetratricopeptide repeats 1 AA074989
interferon-stimulated protein, 15 kDa AA406019
interferon-stimulated protein, 15 kDa AA120862
interferon, alpha-inducible protein clone IFI-6-16 AA452030
EST AA075725
interferon induced transmembrane protein 1 9-27 AA419251
2',5'-oligoadenylate synthetase 1 40-46 kD AA085947
retinoic acid receptor responder tazarotene induced 3 W47350
transporter 1, ATP-binding cassette, sub-family B MDR/TAP AA487429
signal transducer and activator of transcription 1, 91kD AA486367
signal transducer and activator of transcription 1, 91kD AA079495
signal transducer and activator of transcription 1, 91kD AA076085
ectonucleotide pyrophosphatase/phosphodiesterase 1 T69450
AP-2 alpha activating enhancer binding protein 2 R38044
myosin VI AA625890
myosin VI AA028987
LIV-1 protein, estrogen regulated H29315
estrogen receptor 1 AA291702
prolactin receptor R94360
Homo sapiens cDNA FLJ13603 fis, clone PLACE1010270 W32933
GATA binding protein 3 H72474
GATA binding protein 3 R31441
trichorhinophalangeal syndrome I H53479
hepatocyte nuclear factor 3, alpha T74639
X-box binding protein 1 W90128
hypothetical protein FLJ11280 N54608
prolactin receptor R63646
iroquois homeobox protein 5 R46202
cDNA DKFZp586J2118 from clone DKFZp586J2118 N66635
cDNA DKFZp586J2118 from clone DKFZp586J2118 R98407
Homo sapiens cDNA FLJ23821 fis, clone HUV00353 T62552
muscle and heart mammary-derived growth inhibitor N70502
cystatin A stefin A W72207
S100 calcium binding protein A2 AA458884
lectin, galactoside-binding, soluble, 7 galectin 7 W72426
keratin 5 W72110
annexin A8 AA235002
bullous pemphigoid antigen 1 H44784
tripartite motif-containing 29 AA055485
keratin 17 AA150201
keratin 17 AA026100
ESTs AA074677
Homo sapiens, clone IMAGE:4242700 W37448
keratin 13 W60057
keratin 13 W23757
frizzled homolog 7 Drosophila N68049
collagen, type XVII, alpha 1 H67535

Clustering



- Similar expression patterns across many conditions probably imply a common set of regulators
- Looking for a shared set of functional annotations can help find the regulators

Linear models

	treat+WT	treat+KO	no treat+WT	no treat+KO
g	M_{11}	M_{12}	M_{13}	M_{14}

For each gene, make a linear relation between effect (expression) and factors (conditions)

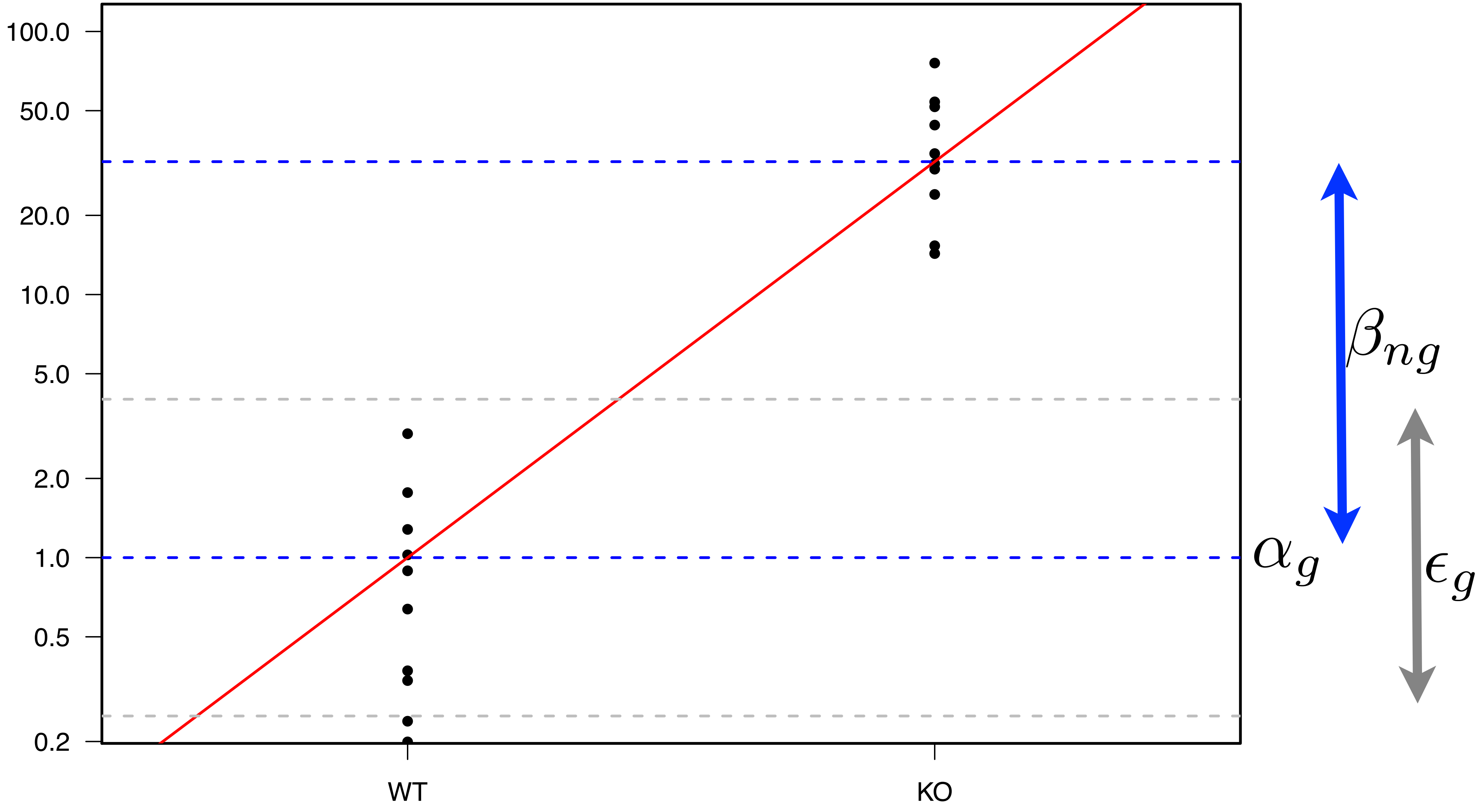
$$\log(M_{cg}) \quad = \quad \alpha_g + \sum_n I_{cn}\beta_{ng} + \epsilon_g \ ,$$

Design matrix:

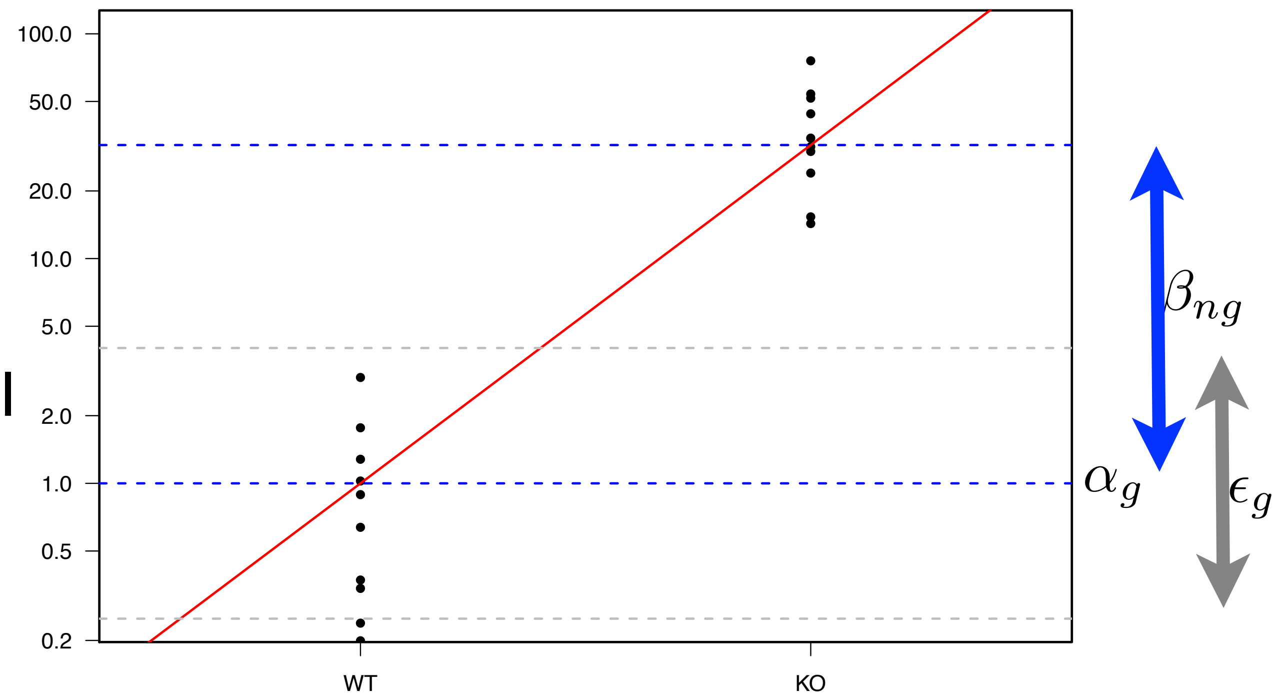
	1	2	3	4
treat	1	1	0	0
KO	0	1	0	1

Linear models

$$\log(M_{cg}) \quad = \quad \alpha_g + \sum_n I_{cn}\beta_{ng} + \epsilon_g$$



Hypothesis Testing



- "Null Hypothesis" H0:
  - average expression of *g* in WT and KO are equal
- Compute a "statistic":

$$t(g) \; = \; \frac{\beta^2}{\epsilon^2_{\mathrm{WT}} + \epsilon^2_{\mathrm{KO}}}$$

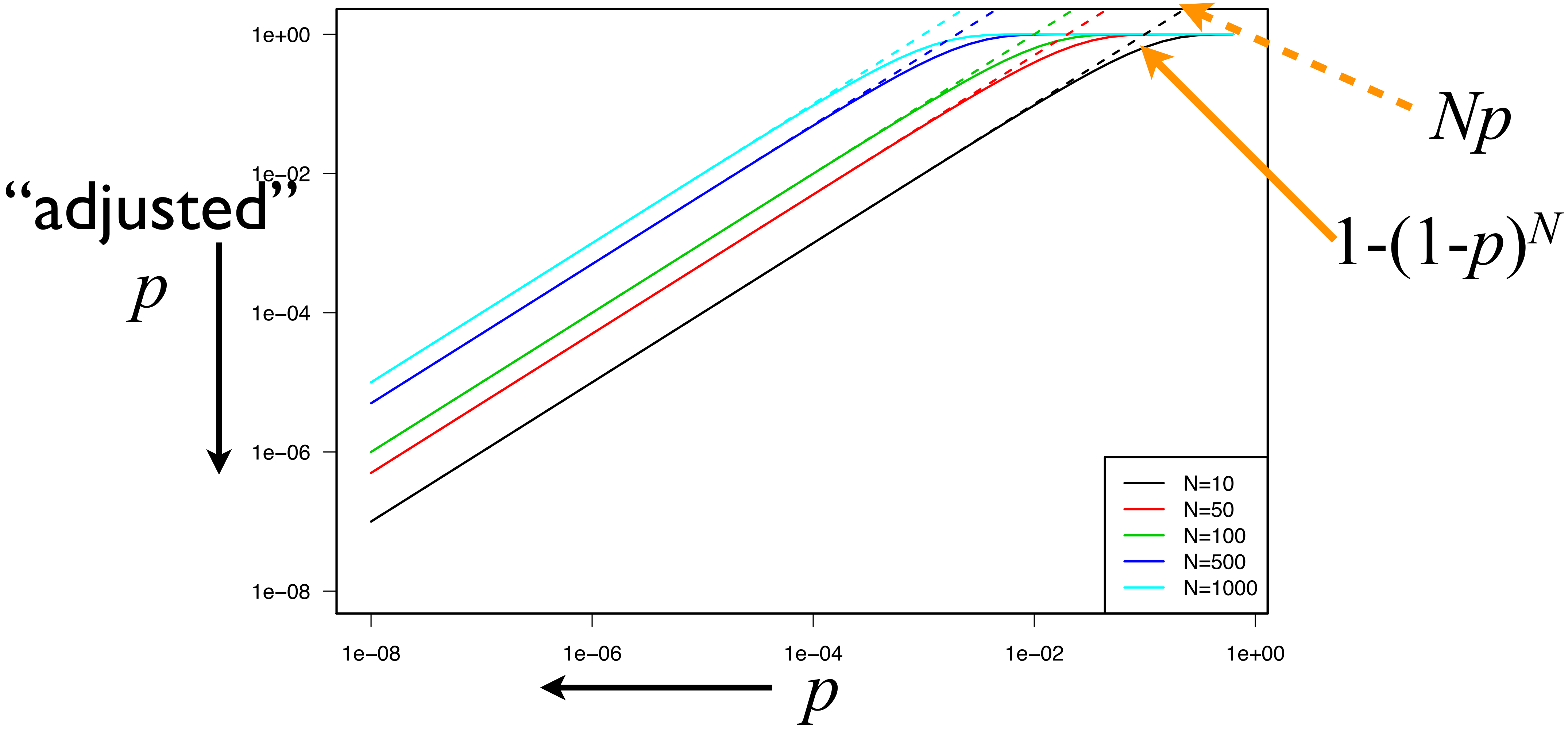
- Model the data (normal distribution), compute a "p-value":
  - probability of observing a variation larger than *t(g)* if H0 AND the model are true

$$p(g) \; = \; \mathrm{Prob}(t \geq t(g) | H0)$$

Multiple testing

- 25’000 genes ⇒ 25’000 p-values. Suppose all tests are independent:

$$\begin{aligned}\mathrm{Prob}(\exists n : t(g_n) \geq t|H0) &= 1 - \mathrm{Prob}(\forall n : t(g_n) < t|H0)\\ &= 1 - \big(1 - \mathrm{Prob}(t(g) \geq t|H0)\big)^N\end{aligned}$$



“adjusted” p

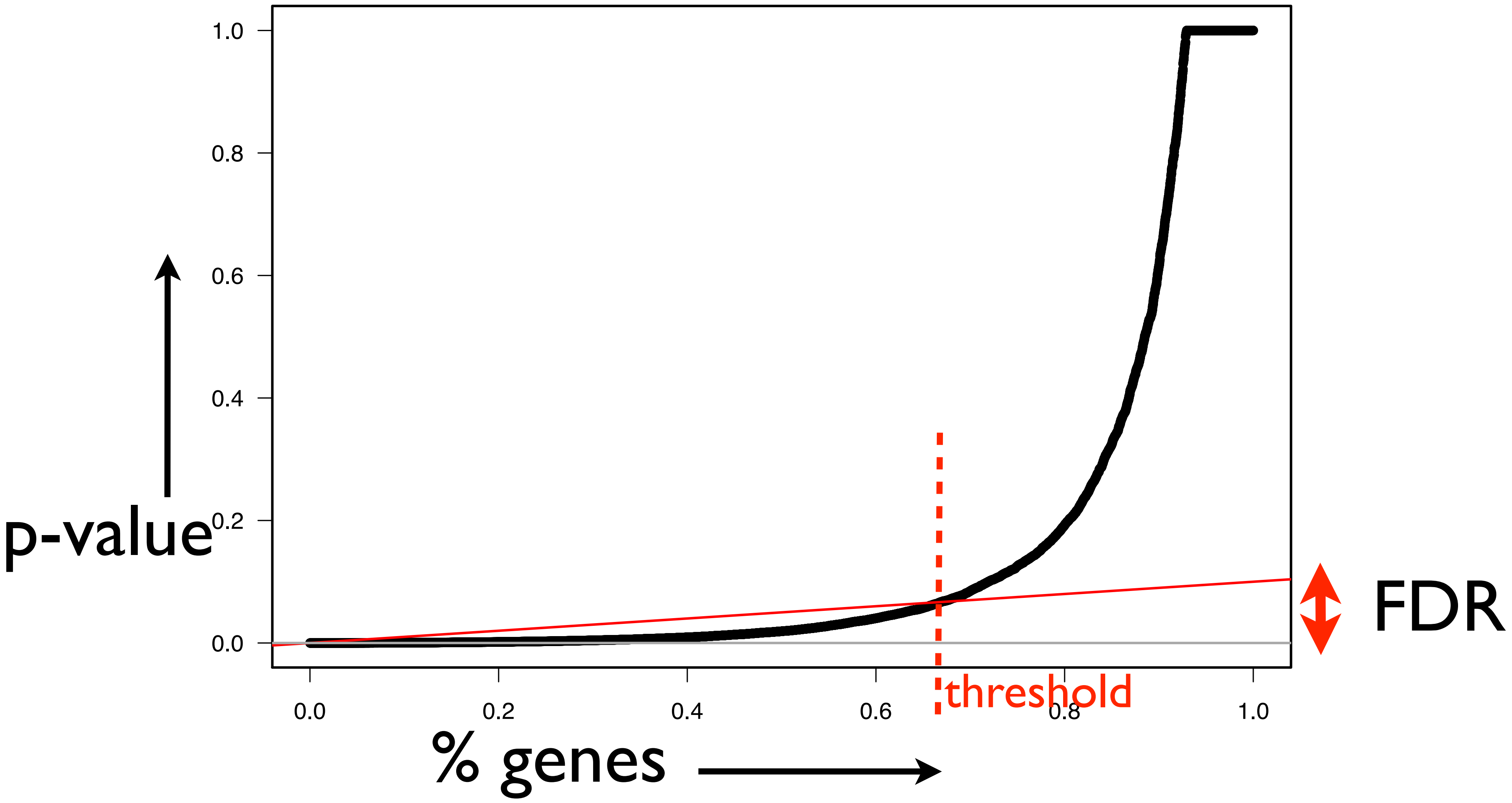
p

Np

$1\text{-}(1\text{-}p)^N$

False Discovery Rate

- If we detect 200 differential genes out of 25’000, we may accept a small proportion of false positive
- False Discovery Rate (FDR): proportion of True H0 among set of 200
- Benjamini–Hochberg procedure: find largest *k* such that $P(g_k)/k < FDR/N$


p-value
% genes
threshold
FDR