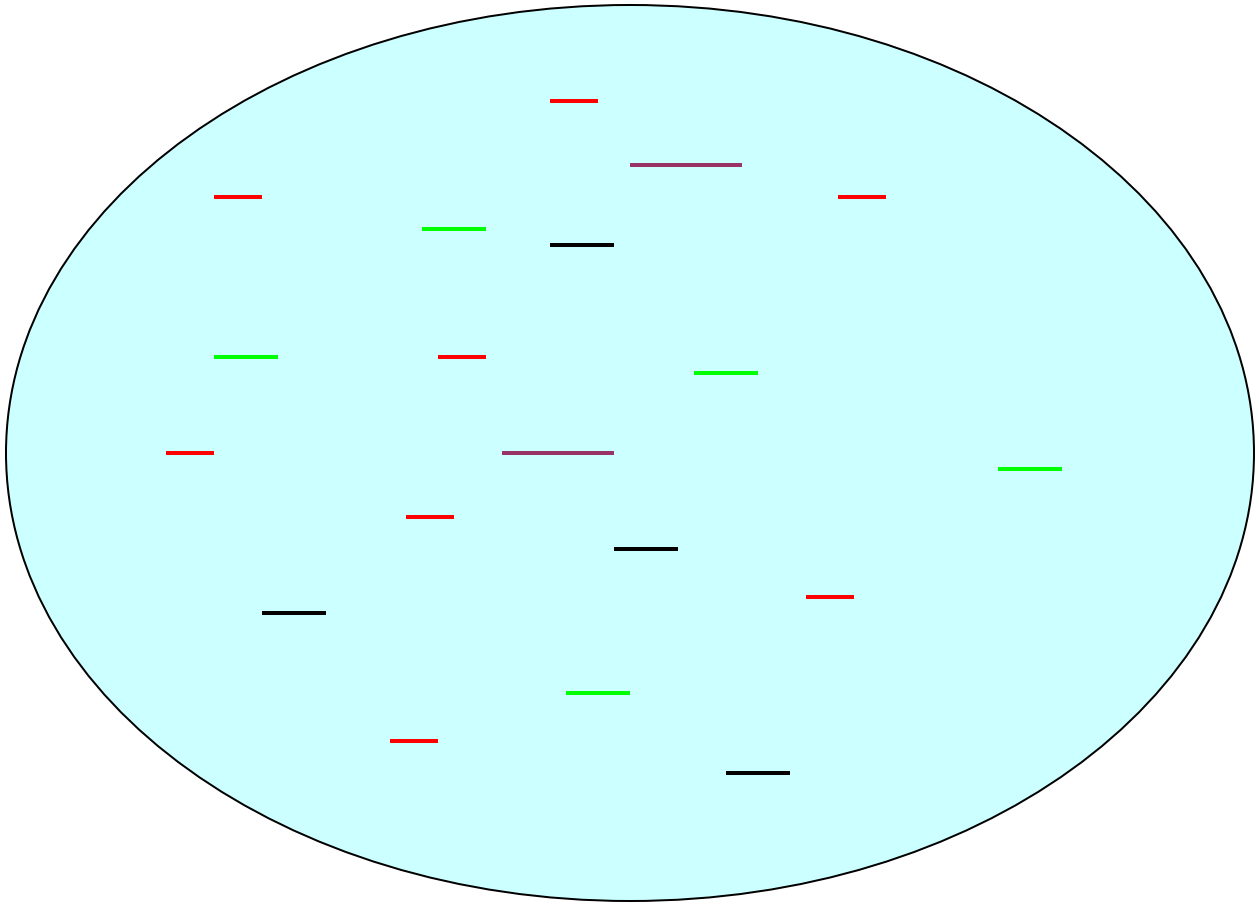
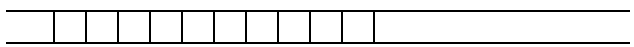
Genexpressionsanalyse

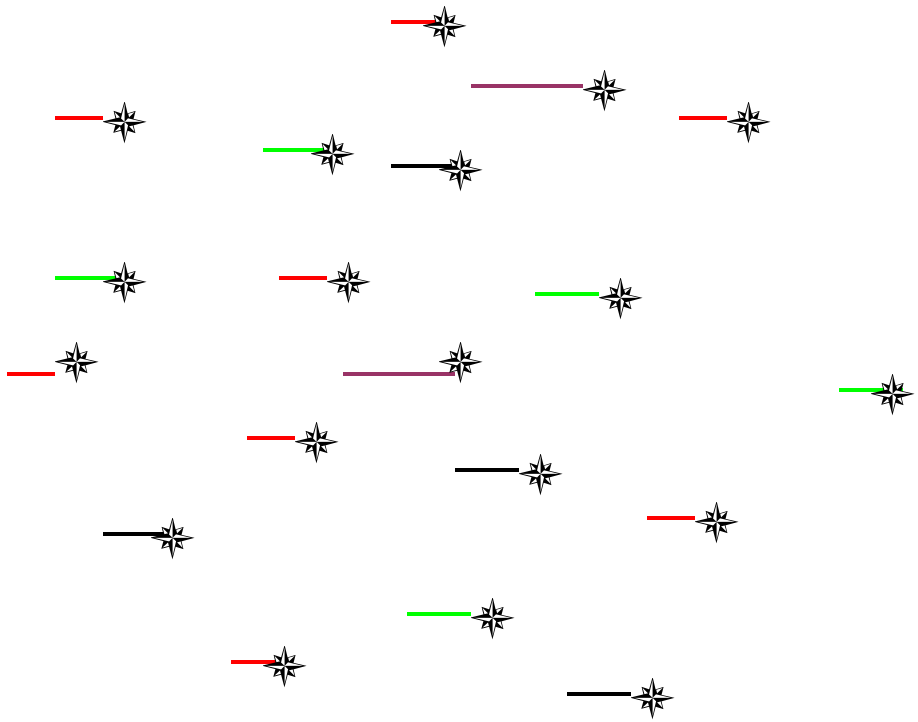
A cell and its population of genes:



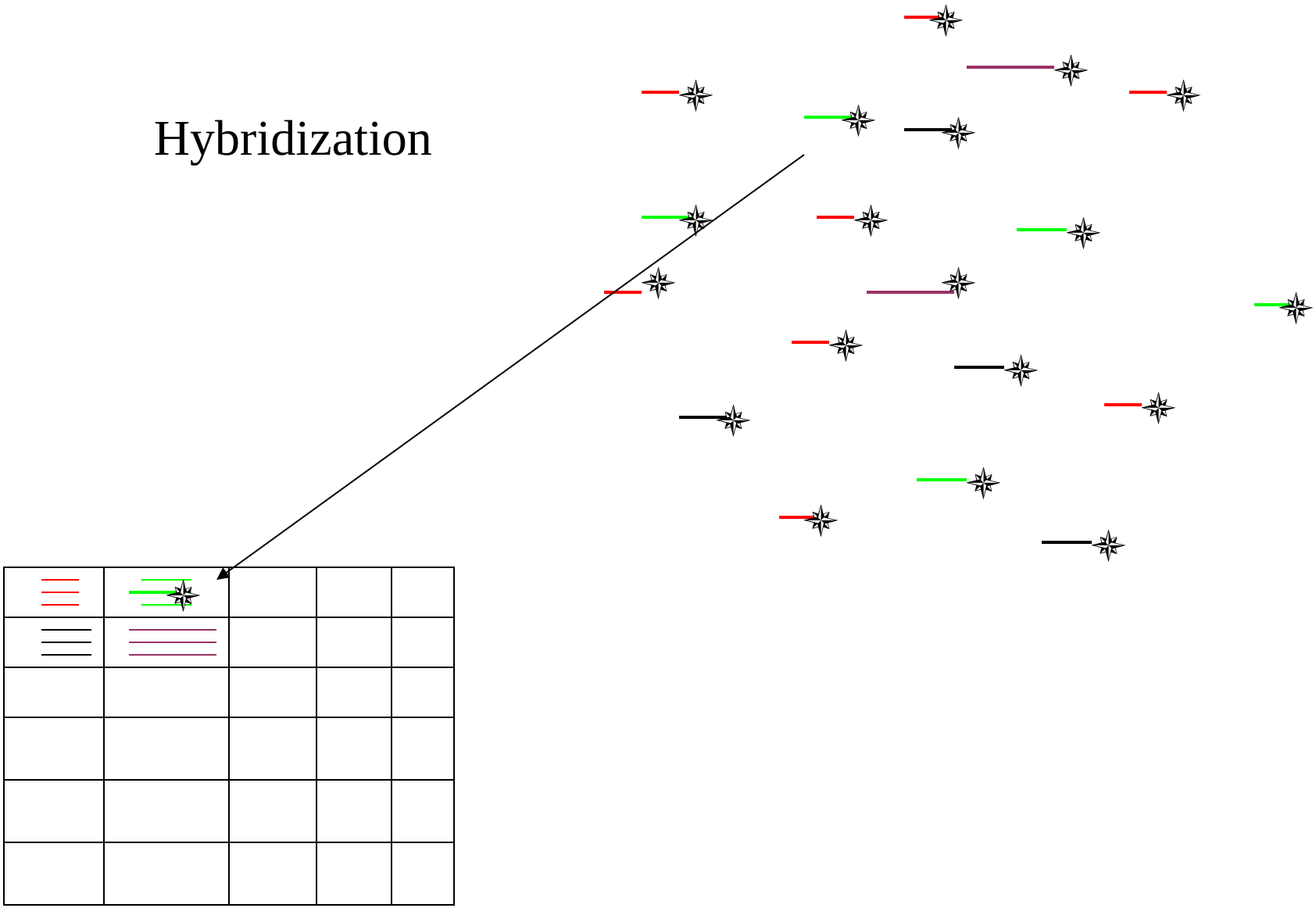
DNA forms double strands by a process called hybridization:

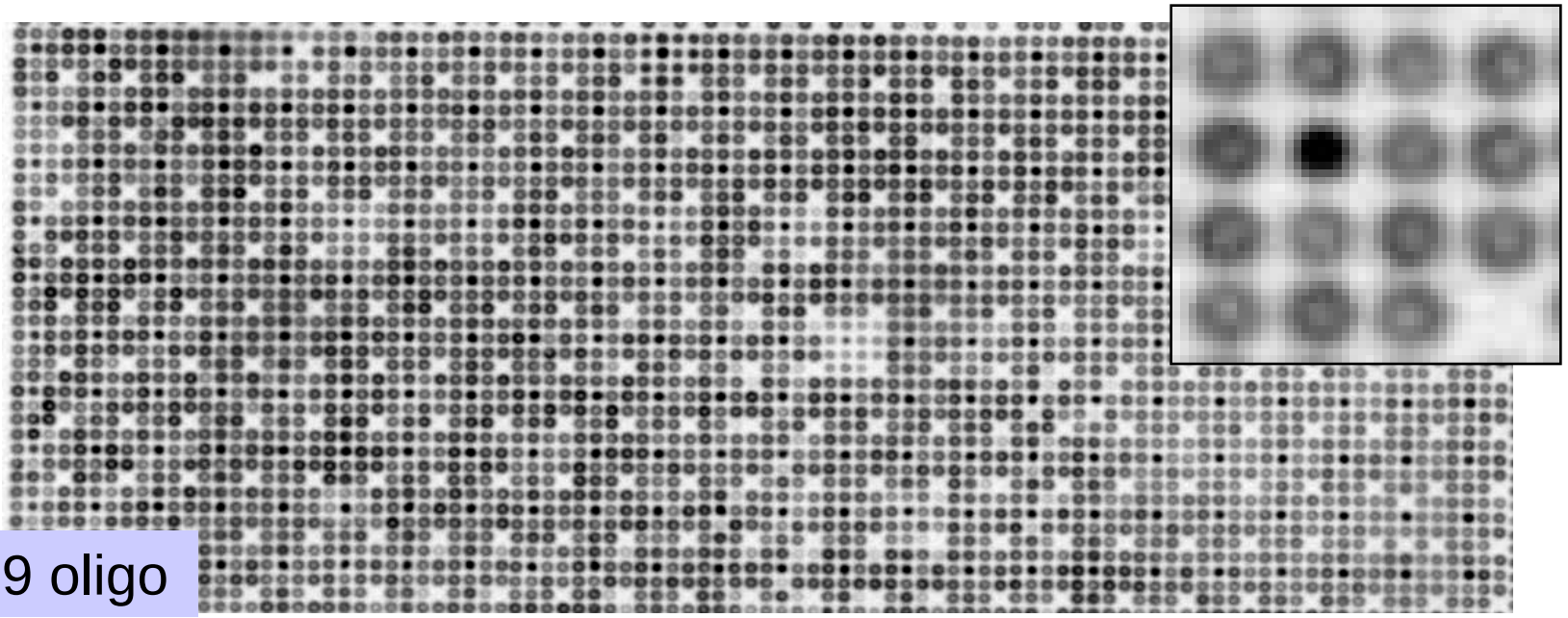


Labeling

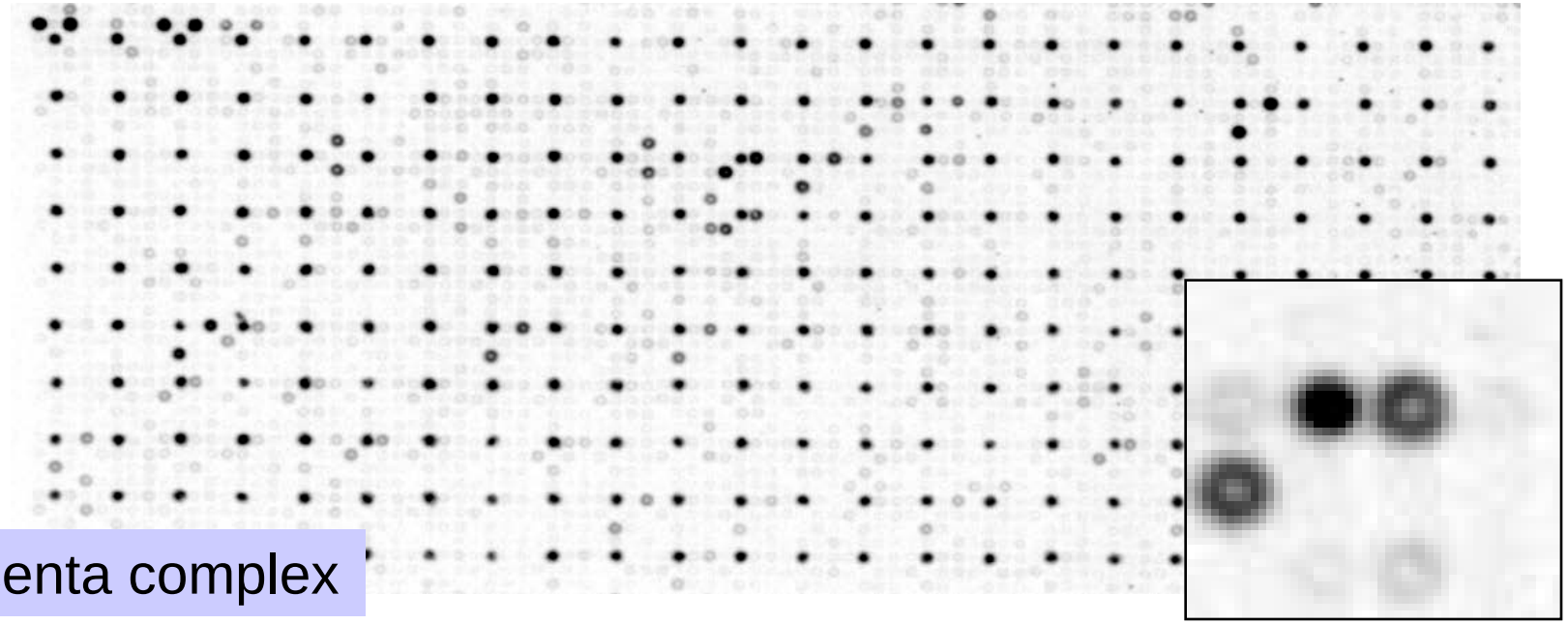


Hybridization



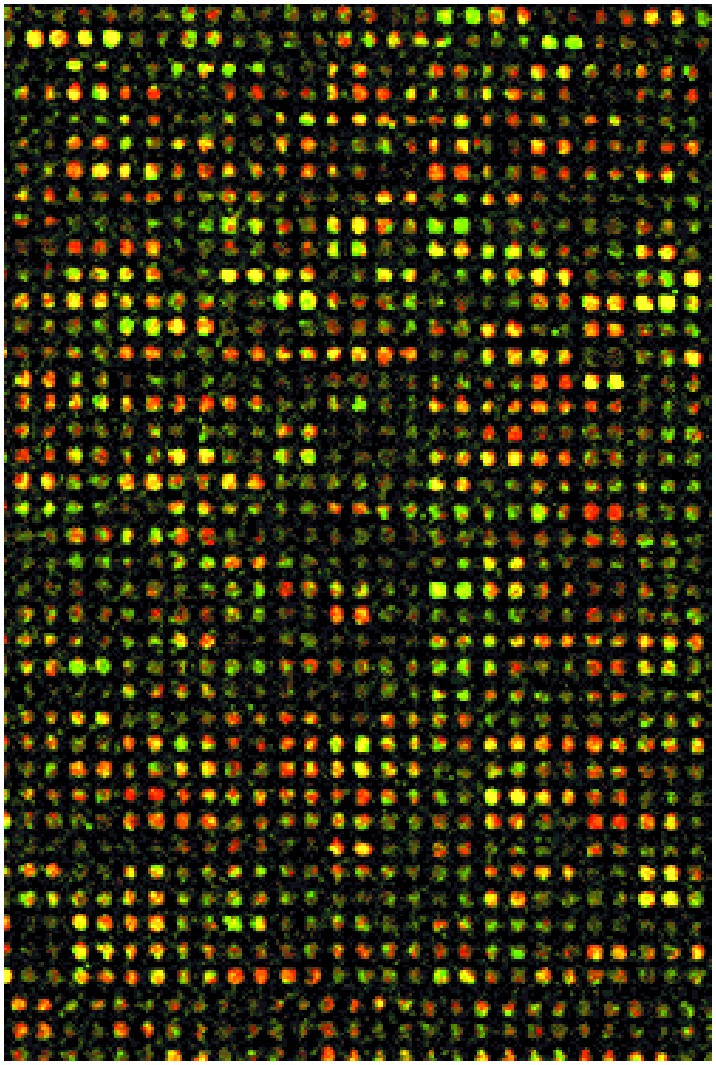


m109 oligo

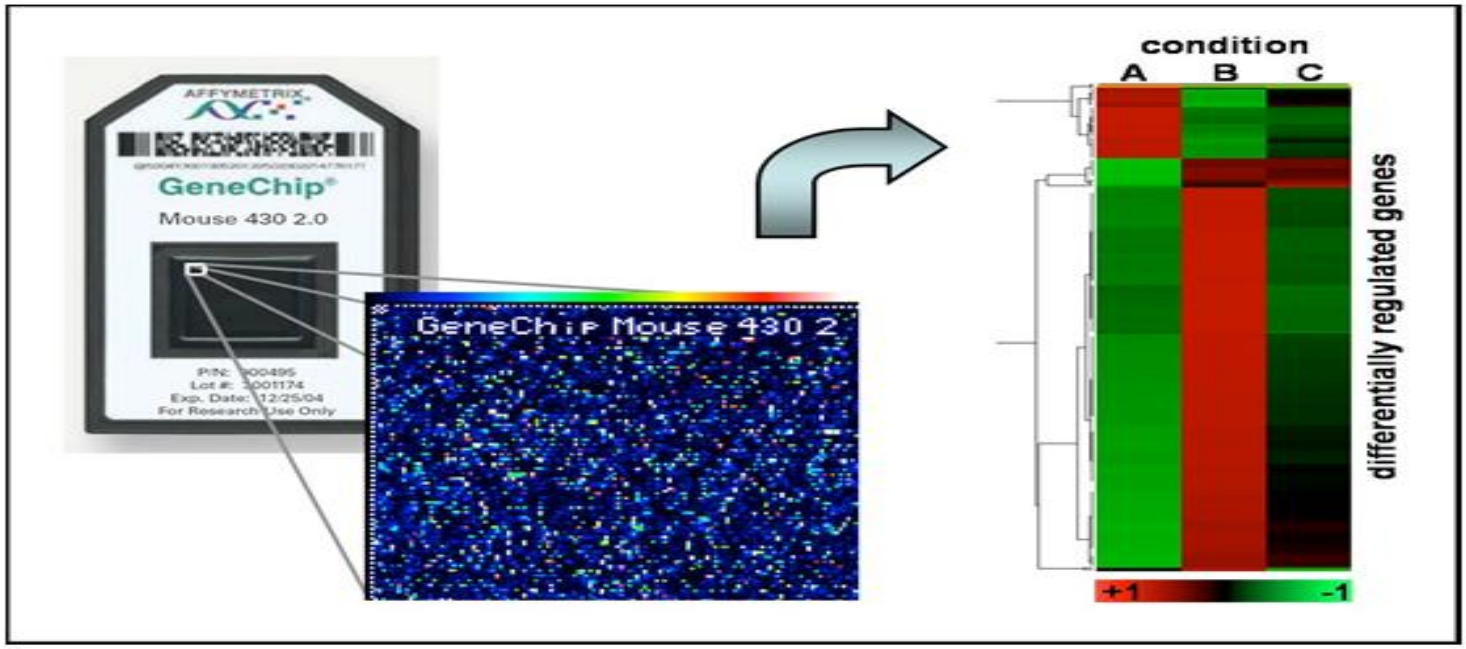


placenta complex

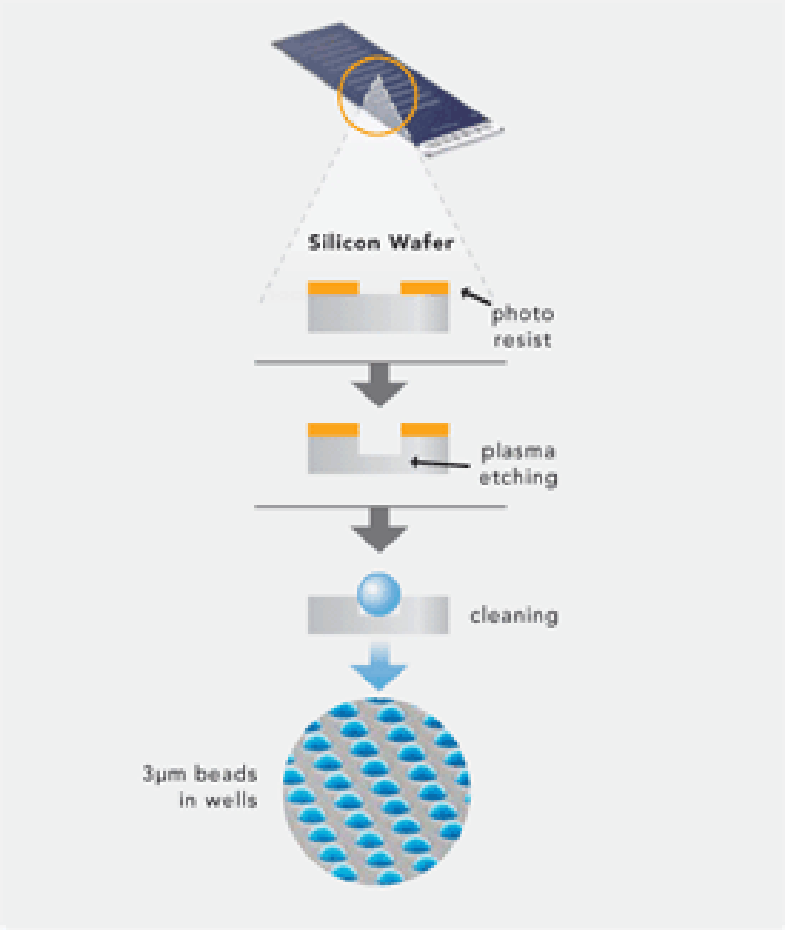
2 color fluoresent dye labeled mRNA



Affymetrix oligonucleotide arrays

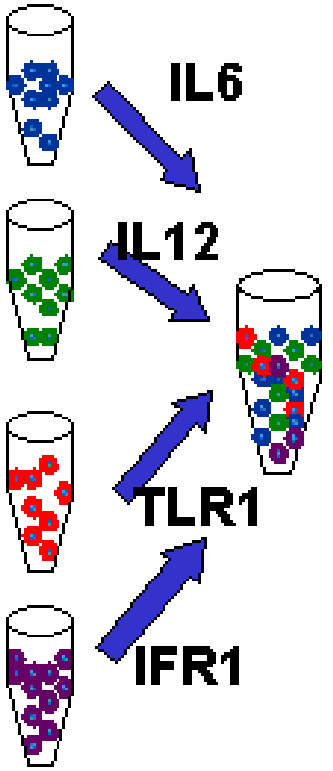

AFFYMETRIX
GeneChip®
Mouse 430 2.0
GeneChip Mouse 430 2
condition
A
B
C
differentially regulated genes
+1
-1

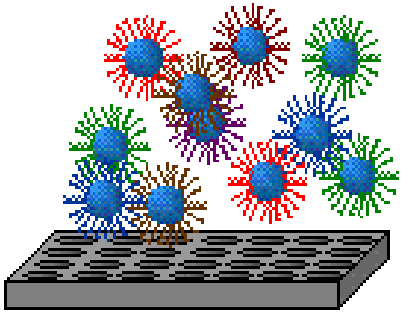
Illumina bead arrays

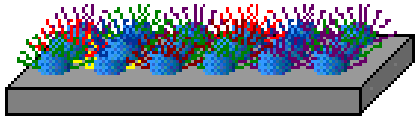


Illumina multi-sample array formats

3 micron diameter beads are coated with Oligonucleotides

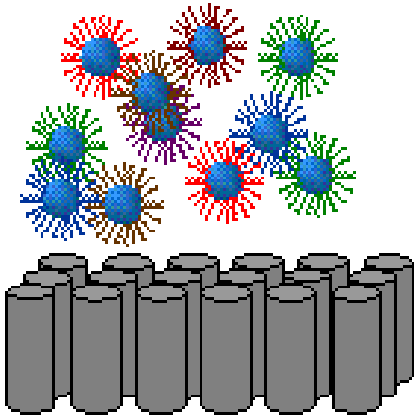


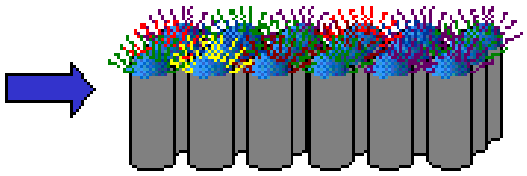


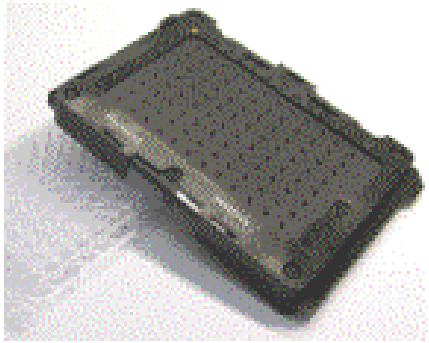


Beads embedded in a slide with multiple arrays located on each slide

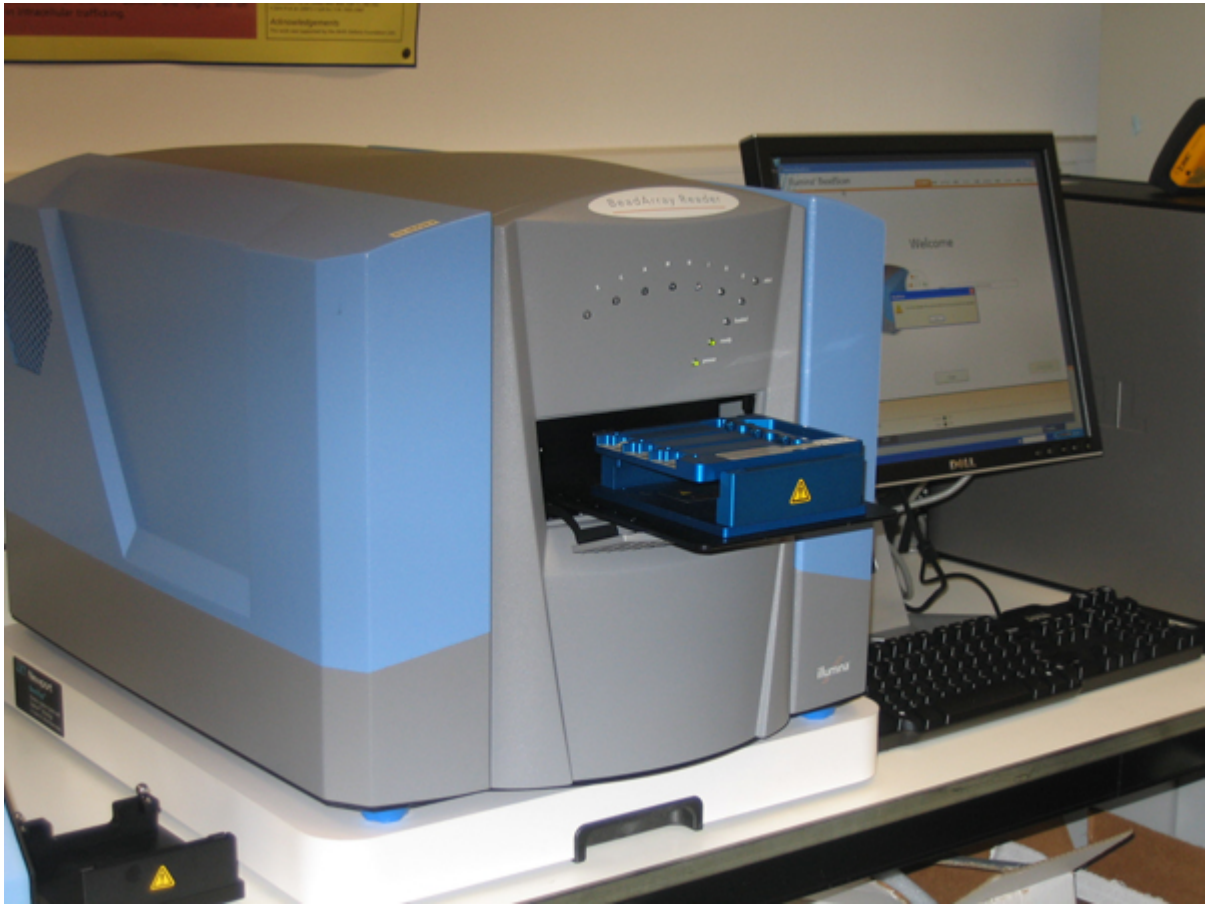
Pooled Beads are randomly located and assayed to identify the location of each bead within the array using a 29 base tag sequence







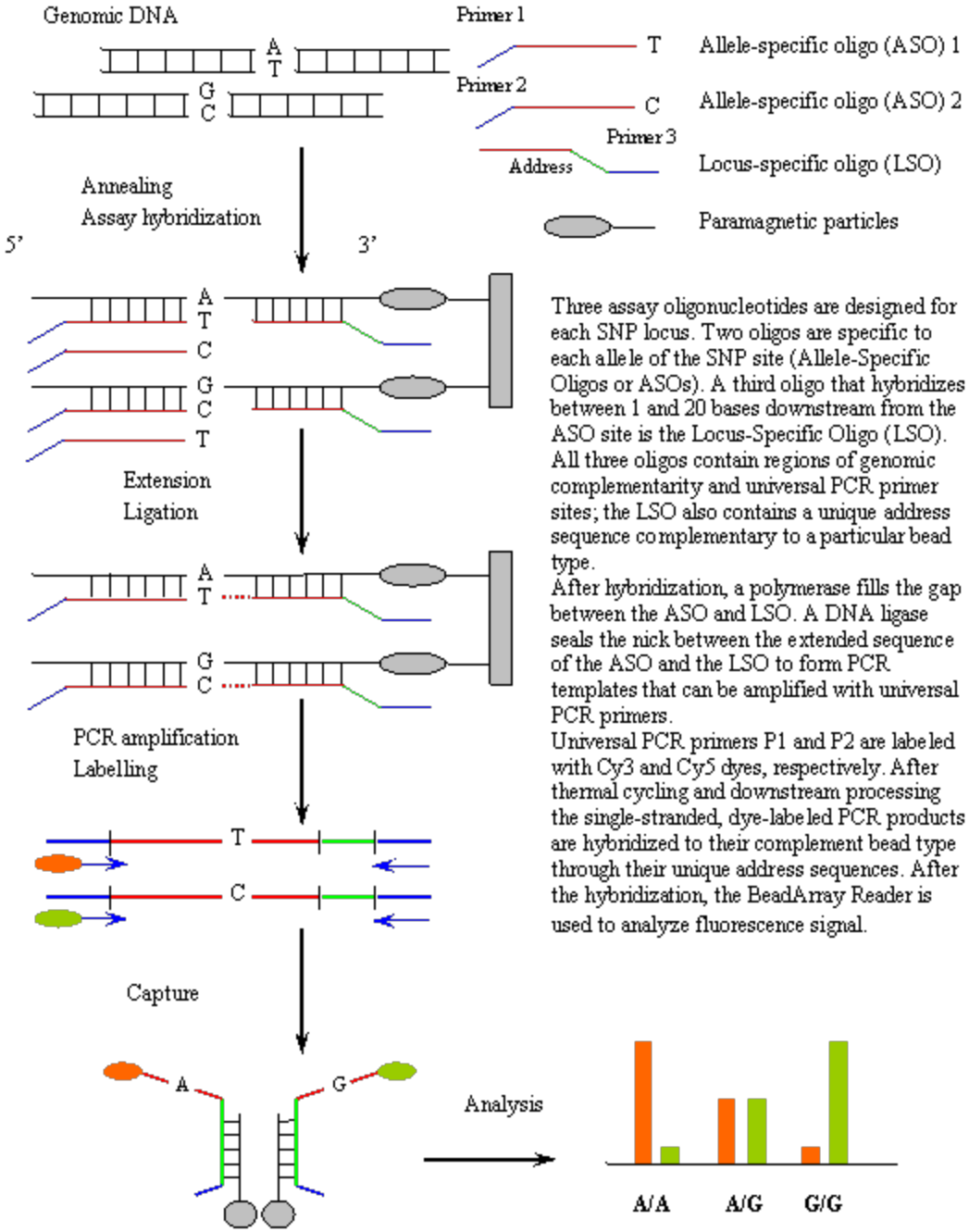
Beads located in the end of a microfibre and collected into a bundle each bundle in a 96 sample/array matrix format


BeadArray Reader
Welcome
DELL
Illumina

- Illumina flyer

SNP detection

(slide from NCBI,
http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/TechBeadArray.shtml



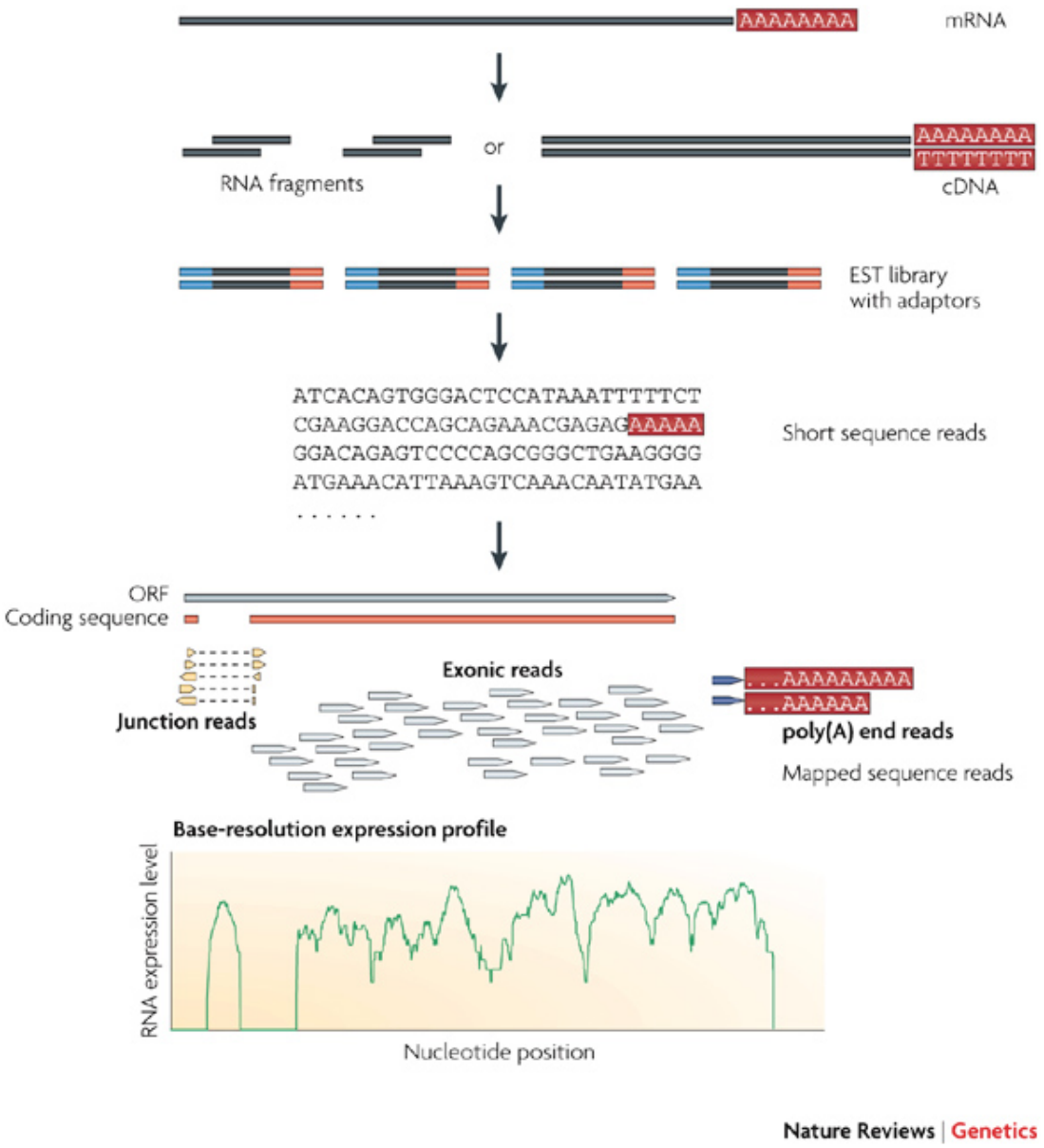
RNA-seq

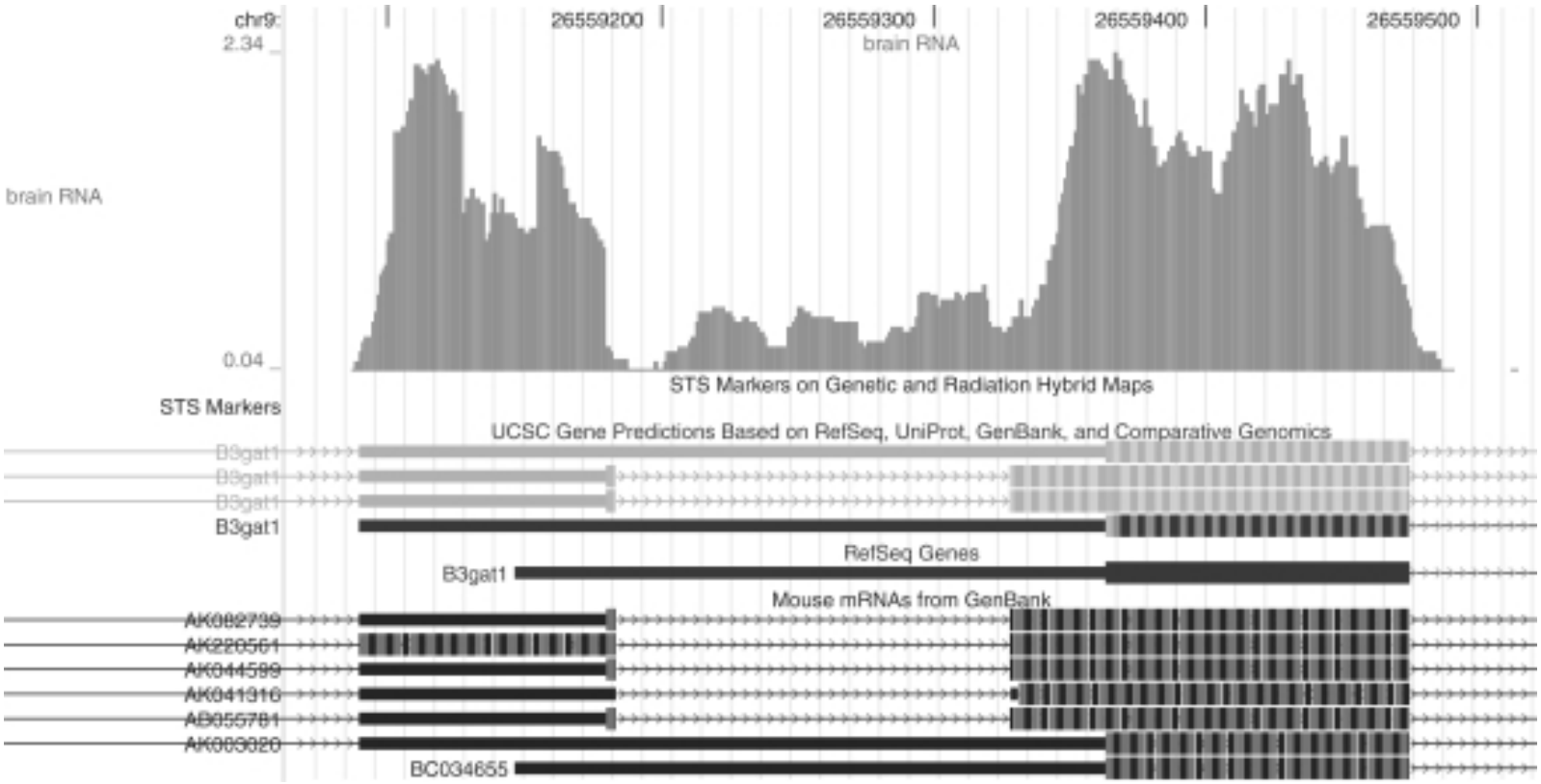
From:

RNA-Seq: a revolutionary tool for transcriptomics

Zhong Wang, Mark Gerstein & Michael Snyder

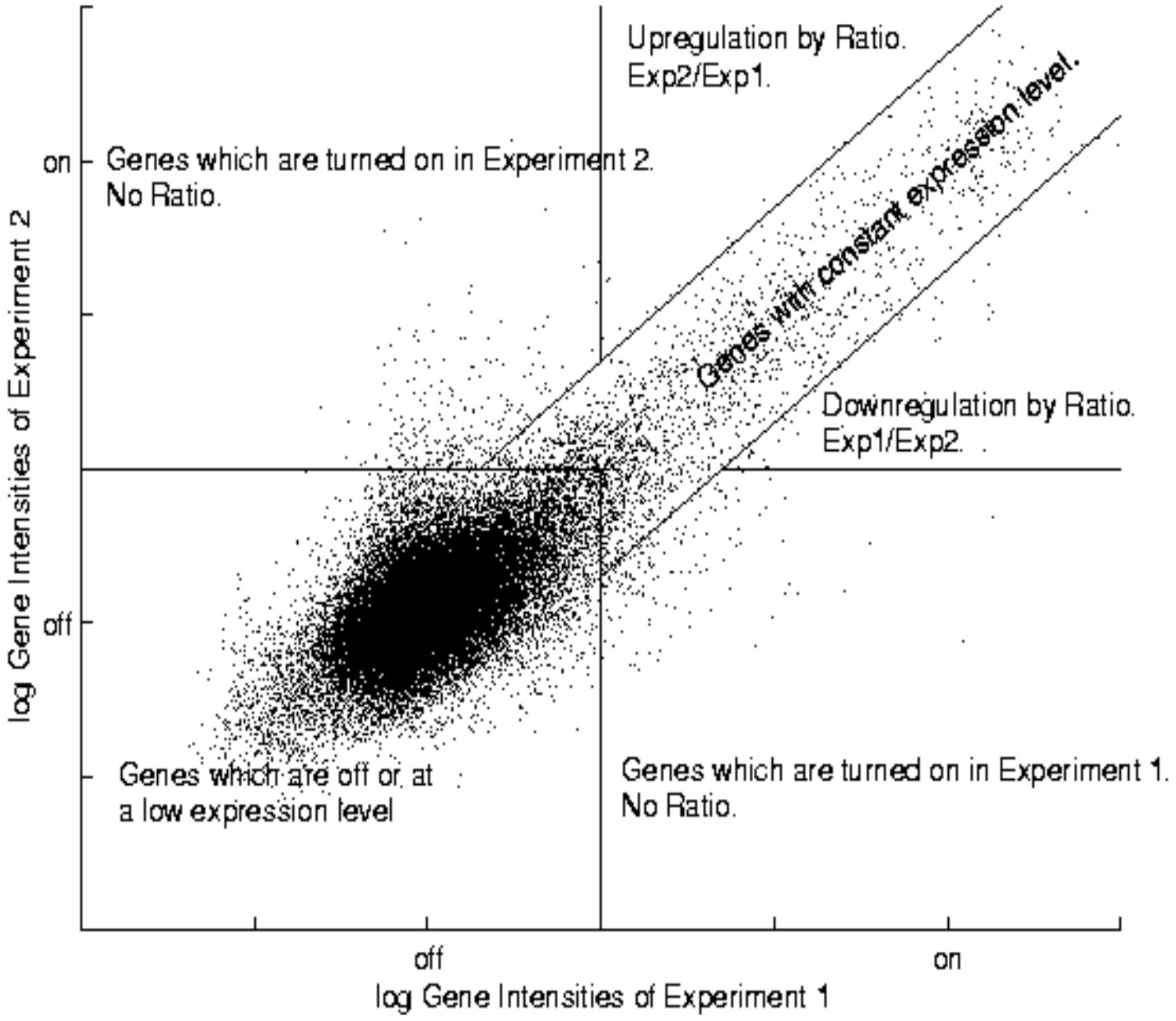
Nature Reviews Genetics 10, 57-63



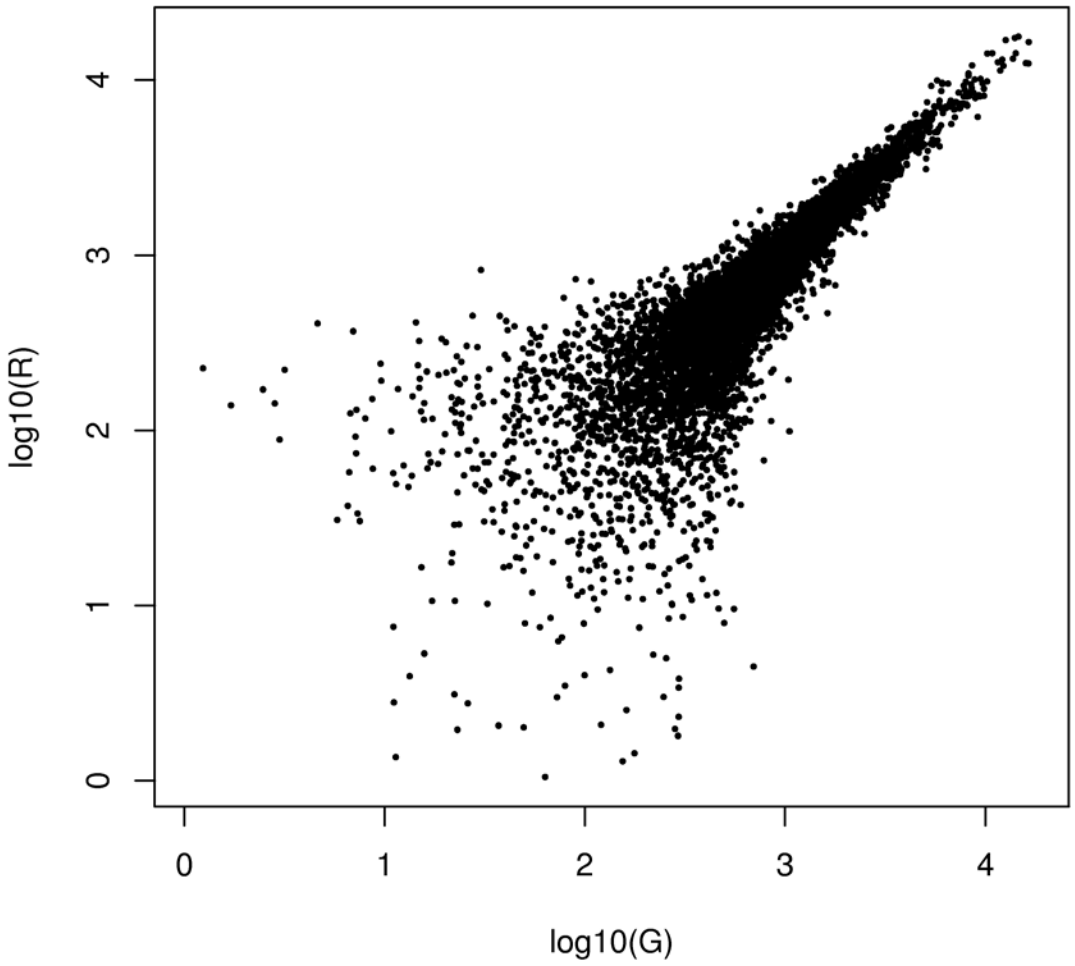

chr9:
26559200
26559300
26559400
26559500
2.34
brain RNA
0.04
brain RNA
STS Markers on Genetic and Radiation Hybrid Maps
STS Markers
UCSC Gene Predictions Based on RefSeq, UniProt, GenBank, and Comparative Genomics
B3gat1
B3gat1
B3gat1
B3gat1
RefSeq Genes
B3gat1
Mouse mRNAs from GenBank
AK082739
AK220561
AK044589
AK041916
AB055781
AK083020
BC034655

Estimation of gene expression from RNA-seq: **RPKM values**

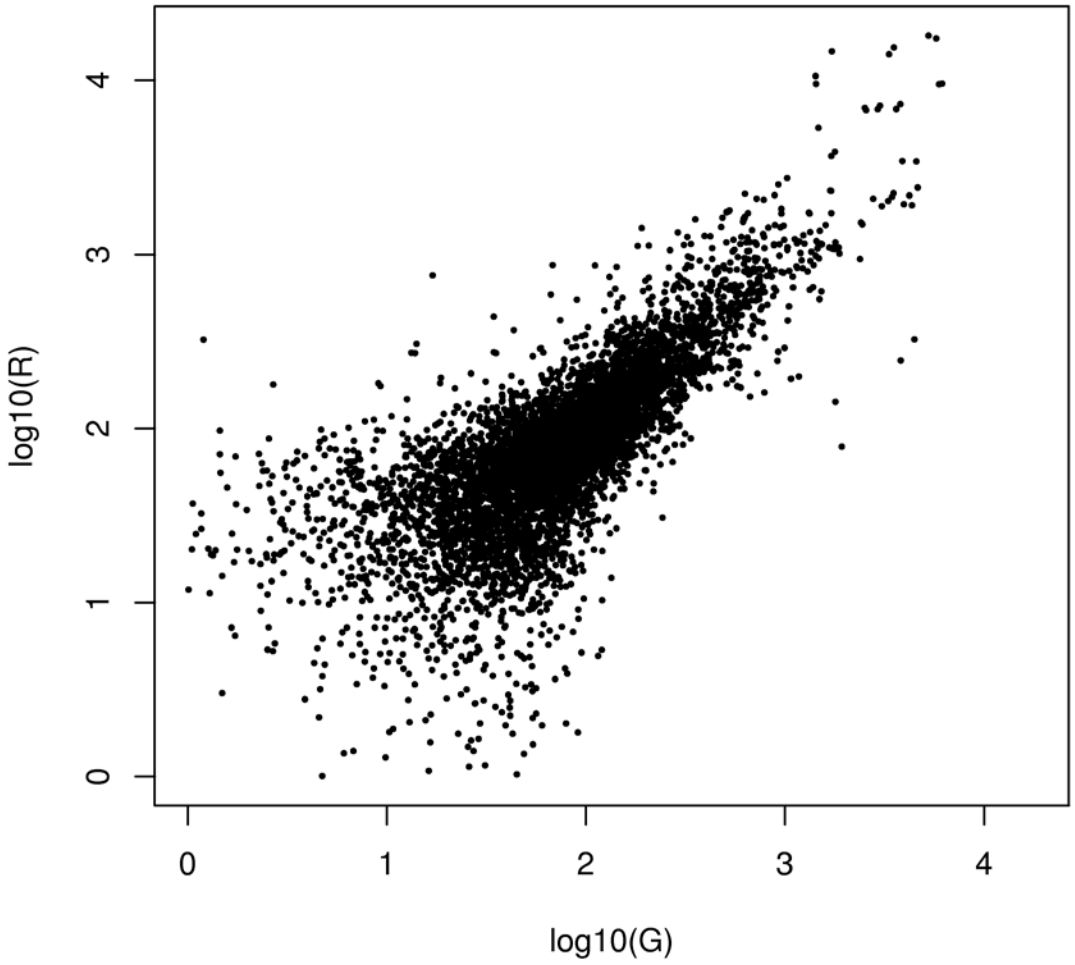
Number of reads which map per kilobase of exon per million mapped reads for each gene



Two samples from the same kidney carcinoma



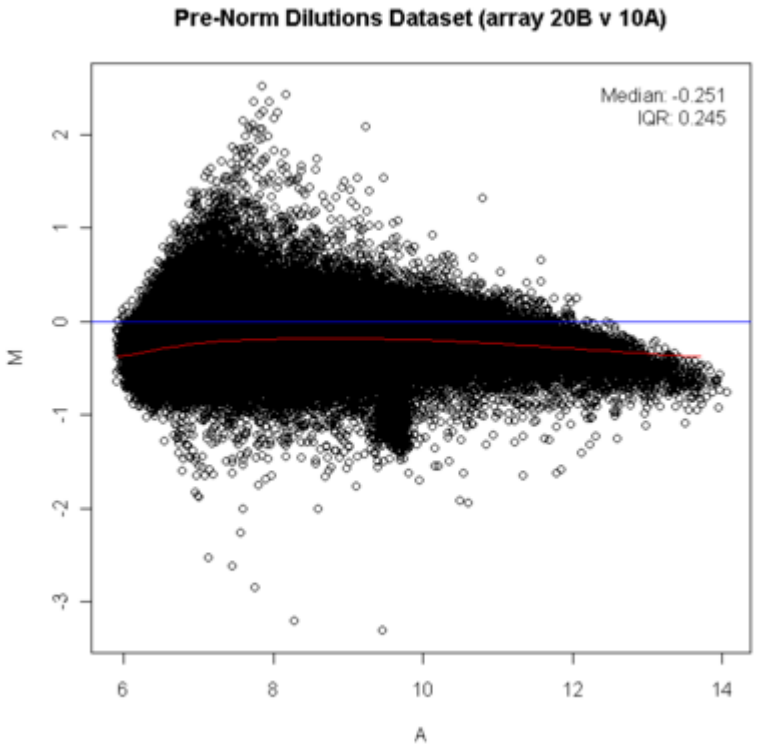
Cancer – normal comparison



MA-plot: Minus vs. Average

$$M = log_2(R/G) = log_2(R) - log_2(G)$$

$$A = \frac{1}{2} log_2(RG) = \frac{1}{2}(log_2(R) + log_2(G))$$



(Wikipedia)

Normalization of microarray data

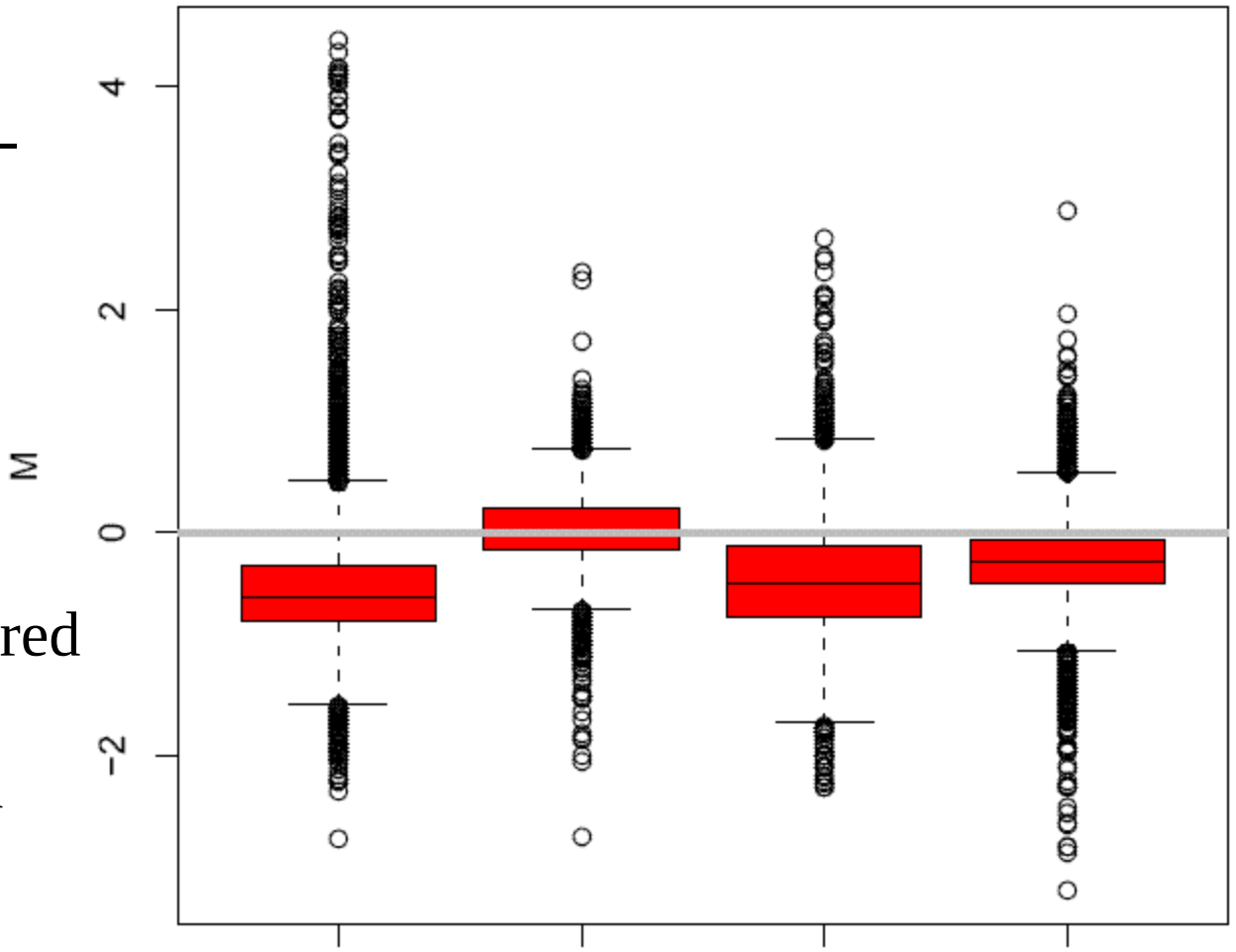
Anja von Heydebreck

Dept. Computational Molecular Biology, MPI for Molecular Genetics, Berlin

Systematic differences between arrays

The boxplots show distributions of log-ratios from 4 red-green 8448-clone cDNA arrays hybridised with zebrafish samples.

Some are not centered at 0, and they are different from each other.



Experimental variation

amount of RNA in the sample
efficiencies of
-RNA extraction
-reverse transcription
-labeling
-photodetection

Systematic

- **similar effect on many measurements**
- **corrections can be estimated from data**

Normalization

Normalization:

Correction of
systematic effects
arising from variations in the
experimental
process

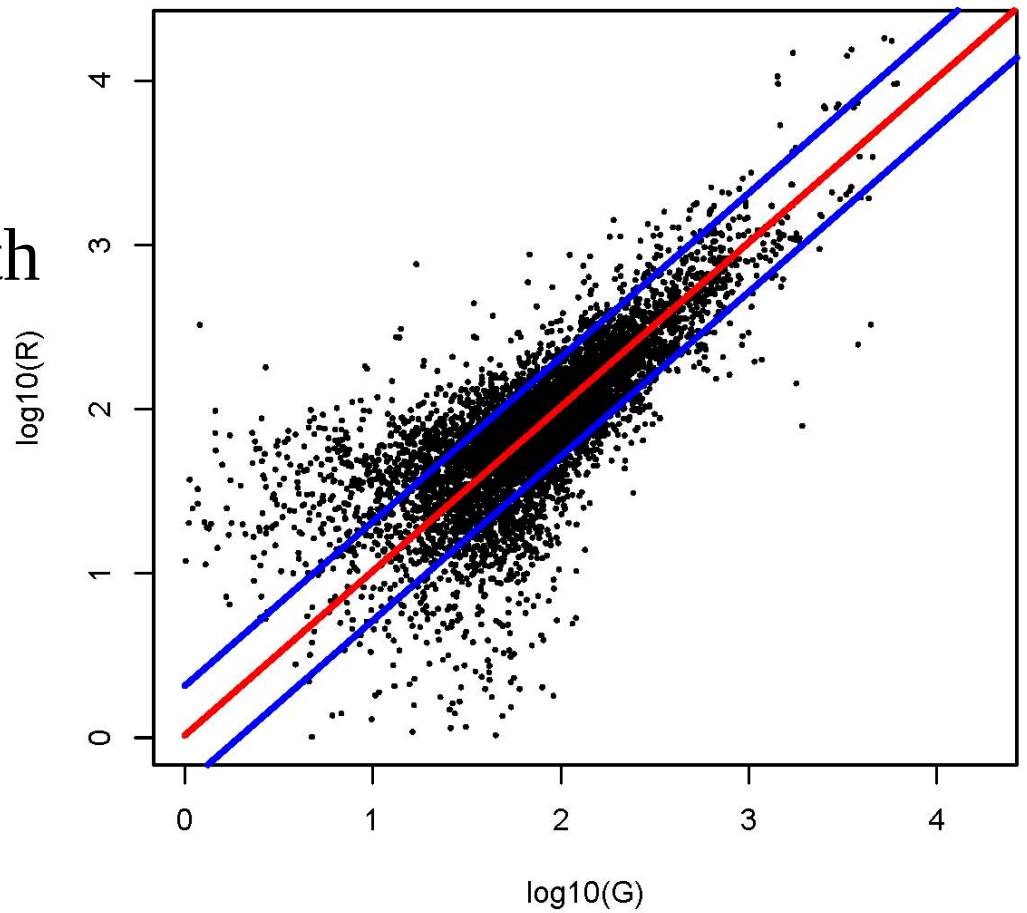
Ad-hoc normalization procedures

- 2-color cDNA-arrays: multiply all intensities of one channel with a constant such that the median of log-ratios is 0 (equivalent: shift log-ratios). Underlying assumption: equally many up- and downregulated genes.
- One-color arrays (Affy, radioactive): multiply intensities from each array k with a constant c_k, such that some measure of location of the intensity distributions is the same for all arrays (e.g. the trimmed mean (Affy *global scaling*)).

log-log plot of intensities from the two channels of a microarray

comparison of kidney cancer with normal kidney tissue, cDNA microarray with 8704 spots

- red line: median normalization
- blue lines: two-fold change

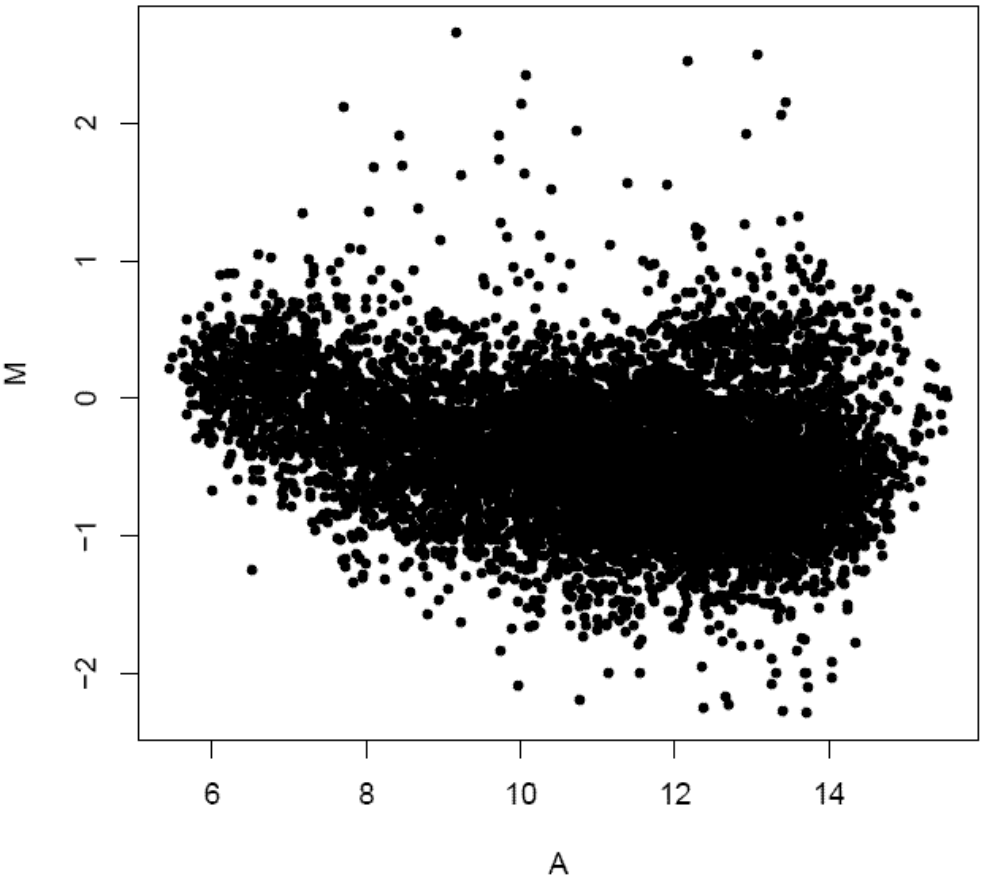


Assumptions for normalization

- When we normalize based on the observed data, we assume that the majority of genes are unchanged, or that there is symmetry between up- and downregulation.
- In some cases, this may not be true. Alternative: use housekeeping genes, which supposedely don’t change, or (spiked) controls, and base normalization on them.

1. Loess normalization

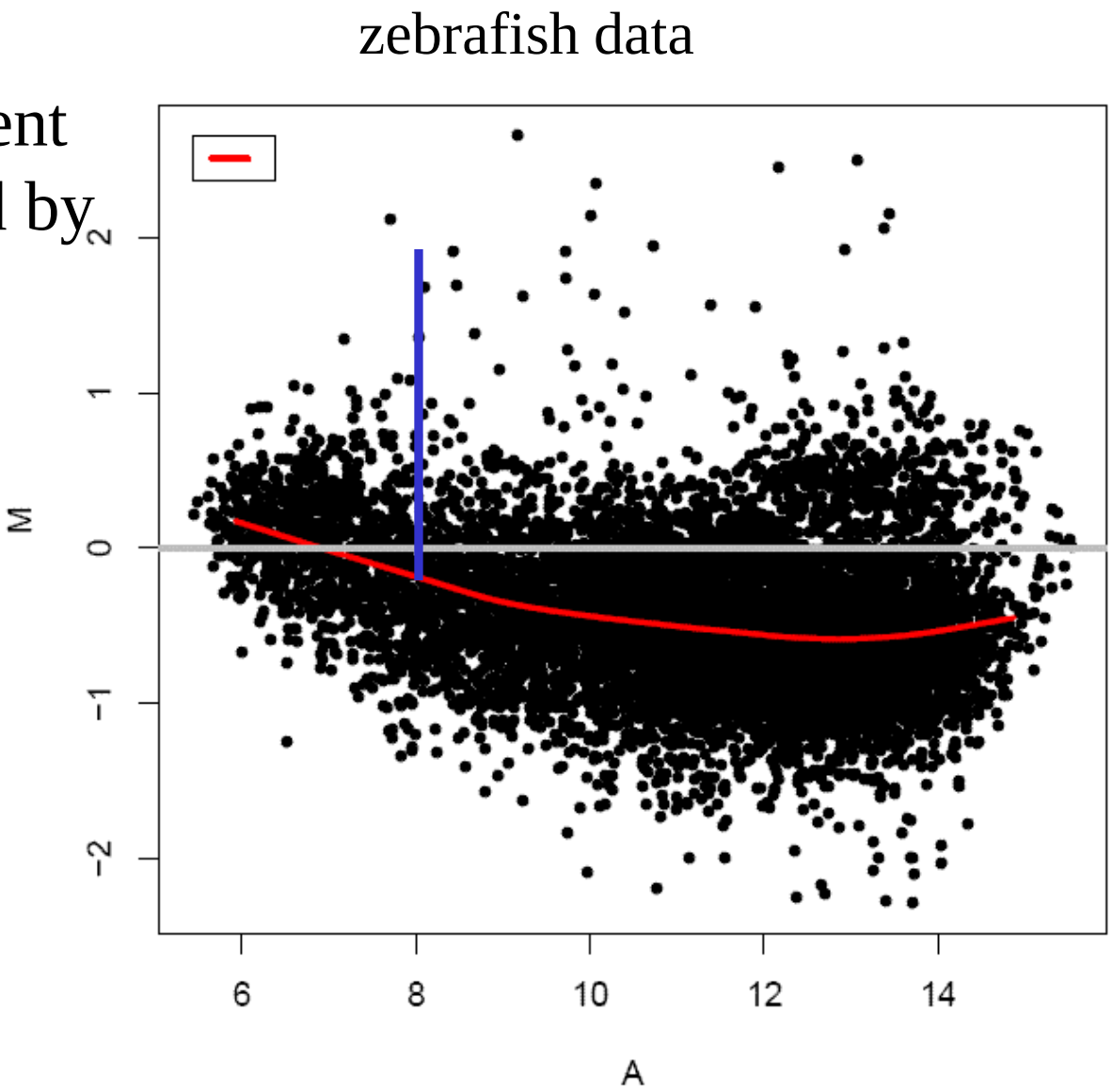
- **M-A plot**
  (**minus** vs. **add**):
  $\log(R) - \log(G) = \log(R/G)$
  vs.
  $\log(R) + \log(G) = \log(RG)$
- With 2-color-cDNA arrays, often “banana-shaped” scatterplots on the log-scale are observed.



Loess normalization

• Intensity-dependent trends are modeled by a regression curve, $M = f(A) + \varepsilon$.

• The normalized log-ratios are computed as the residuals ε of the loess regression.

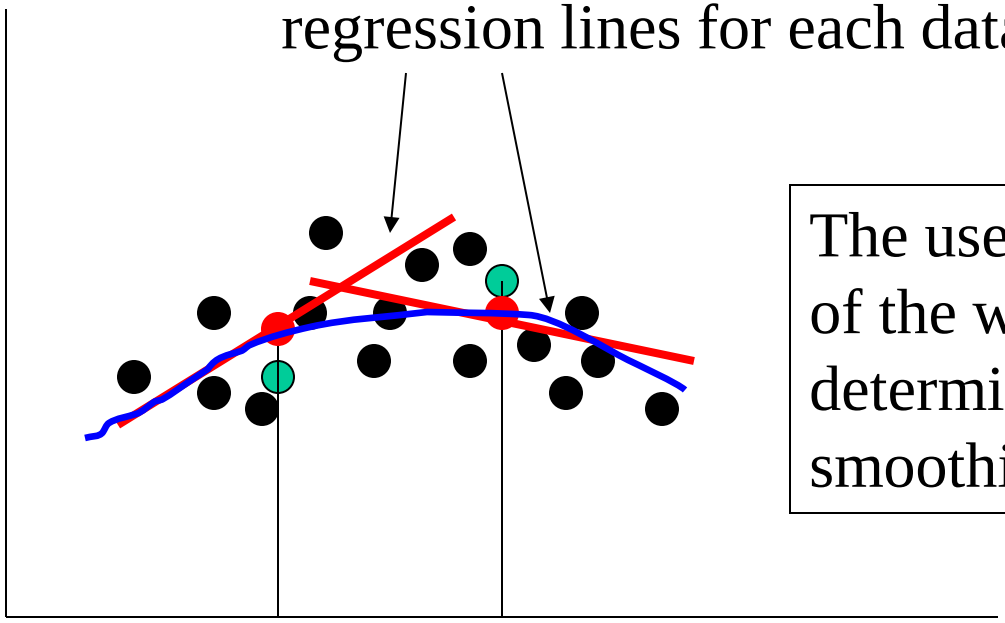


Loess regression

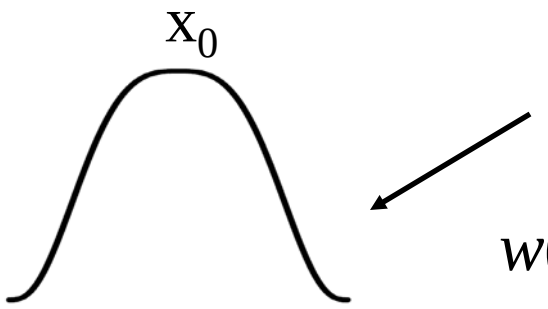
- Locally weighted regression.
- For each value x_i of X, a linear or polynomial regression function f_i for Y is fitted based on the data points close to x_i. They are weighted according to their distance to x_i.
- Local model: $Y = f_i(X) + \varepsilon$.
- Fit: Minimize the weighted sum of squares $\Sigma w_j (x_j)(y_j - f_i(x_j))^2$
- Then, compute the overall regression as:

$$Y = f(X) + \varepsilon, \text{ where } f(x_i) = f_i(x_i).$$

Loess regression



The user-defined width c of the weight function determines the degree of smoothing.

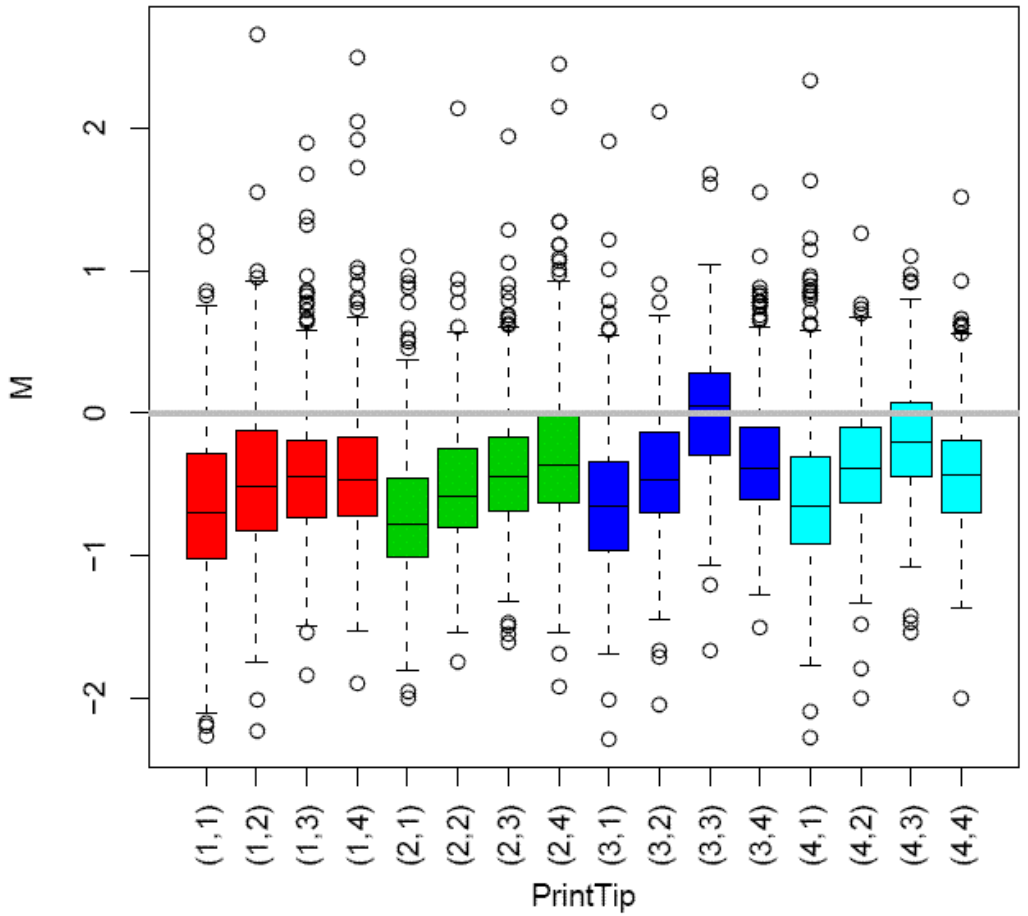


tricubic weight function

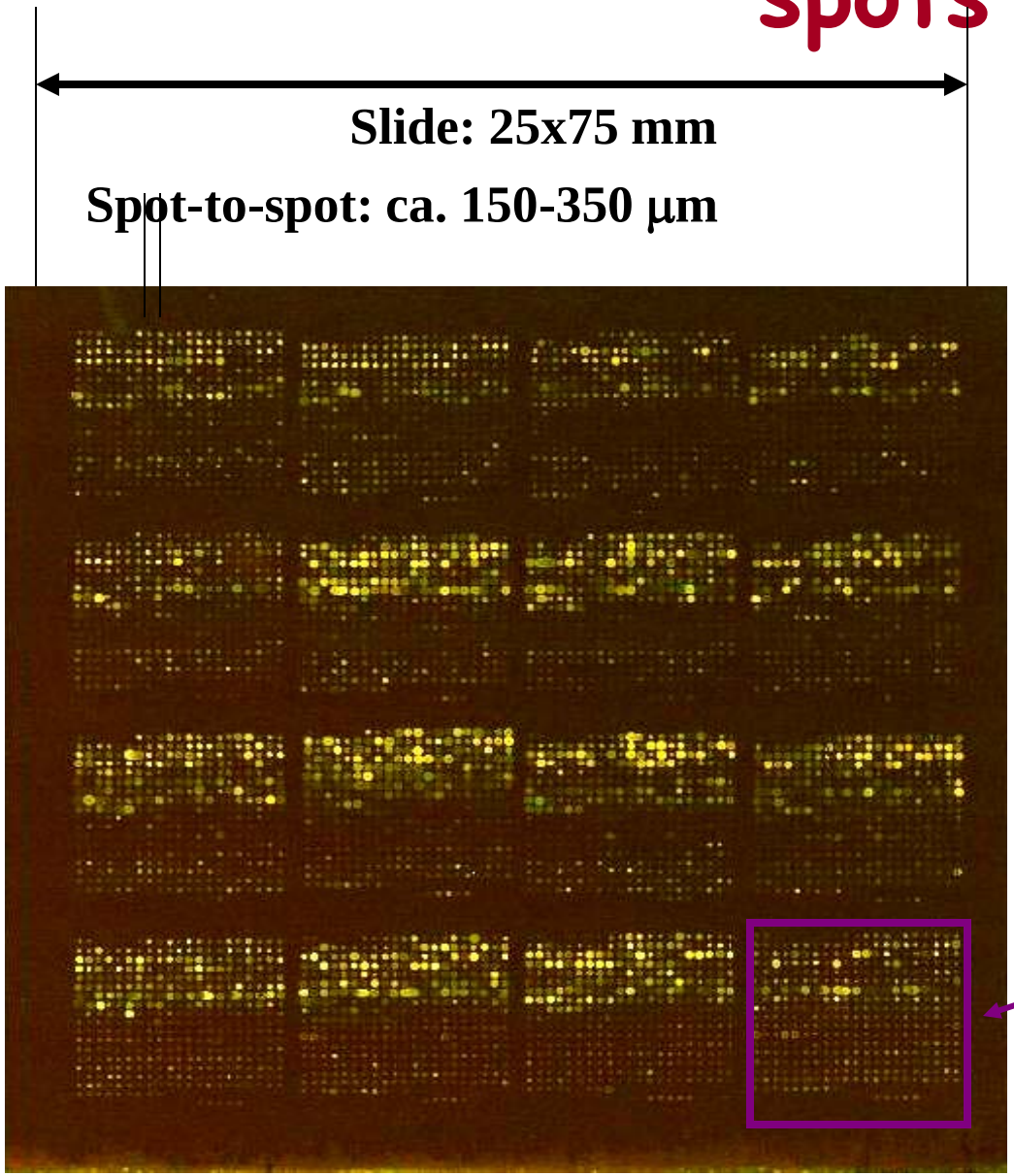
$$w(x) = [1 - (\frac{|x - x_0|}{c})^3]^3, |x - x_0| \leq c$$

Print-tip normalization

- With spotted arrays, distributions of intensities or log-ratios may be different for spots spotted with different pins, or from different PCR plates.
- Normalize the data from each (e.g. print-tip) group separately.



Print-tips correspond to localization of spots



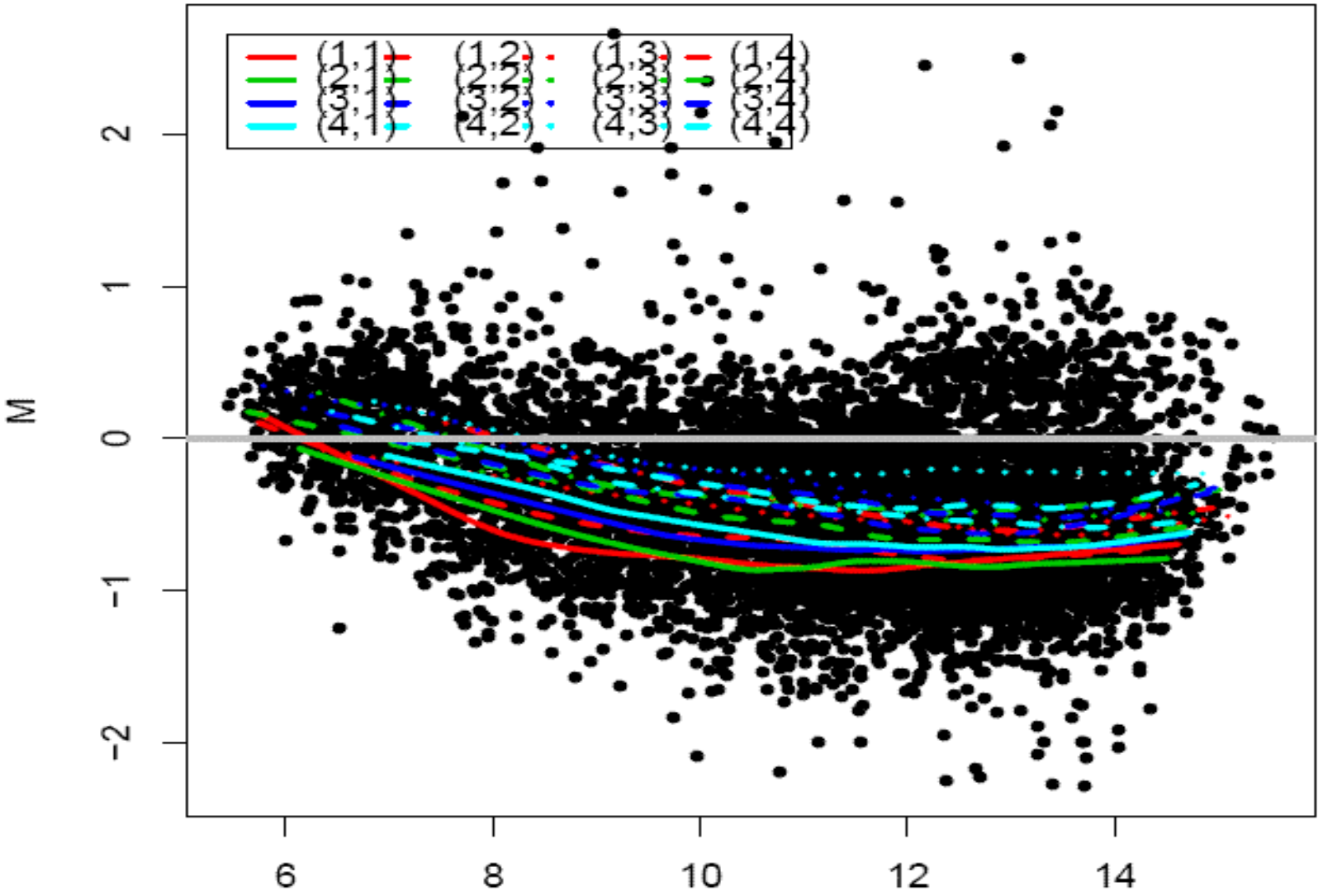
4x4 or 8x4 sectors

17...38 rows and columns per sector

ca. 4600…46000 probes/array

sector: corresponds to one print-tip

Print-tip loess normalization

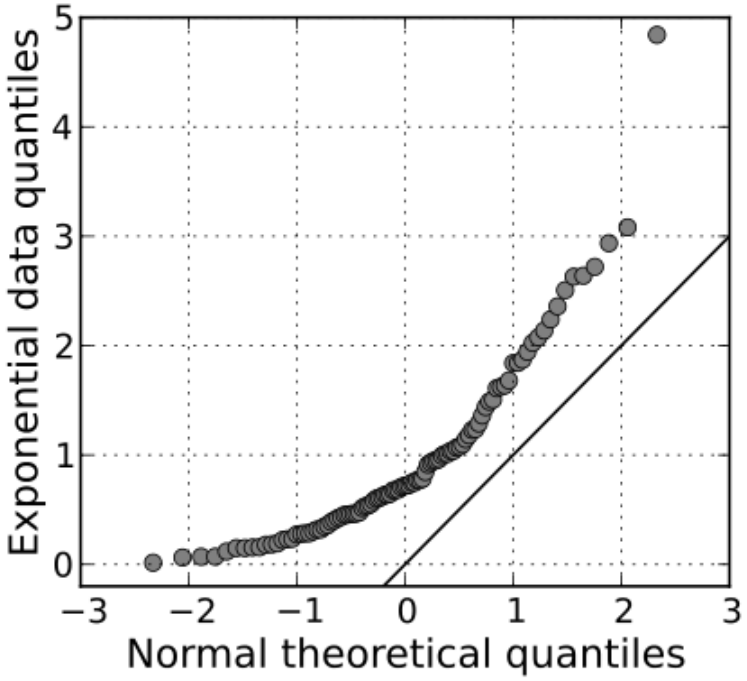


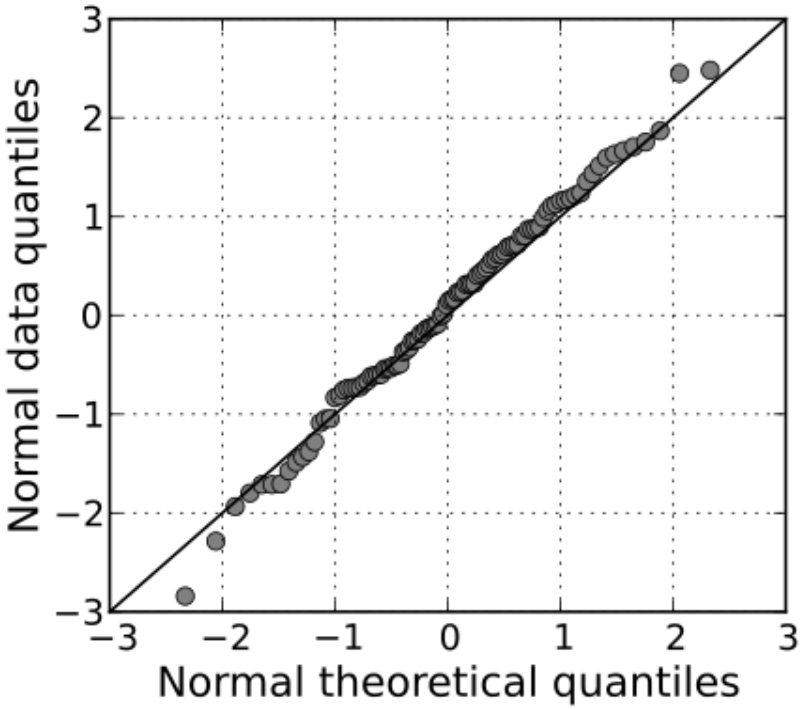
Q-Q Plots

- [Wikipedia] A **Q–Q plot** is a plot, which is a graphical method for comparing two probability distributions by plotting their quantiles against each other.
- A point (x,y) on the plot corresponds to one of the quantiles of the second distribution (y-coordinate) plotted against the same quantile of the first distribution (x-coordinate)

Q-Q Plots, continued

- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$.
- Q–Q plots can be used to compare collections of data, or theoretical distributions. The use of Q–Q plots to compare two samples of data can be viewed as a non-parametric approach to comparing their underlying distributions.


Exponential data quantiles
Normal theoretical quantiles
5
4
3
2
1
0
−3
−2
−1
0
1
2
3


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Quantile Normalization

A comparison of normalization methods for high density oligonucleotide array data based on variance and bias

Bolstad, Irizarry, Astrand, Speed

Bioinformatics 2003, vol 19(2), 185-193

Quantile Normalization

	Hyb 1	Hyb 2	Hyb 3	...
Gen 1	**...**			
Gen 2		**...**		
Gen 3			**...**	
Gen 4				
...				

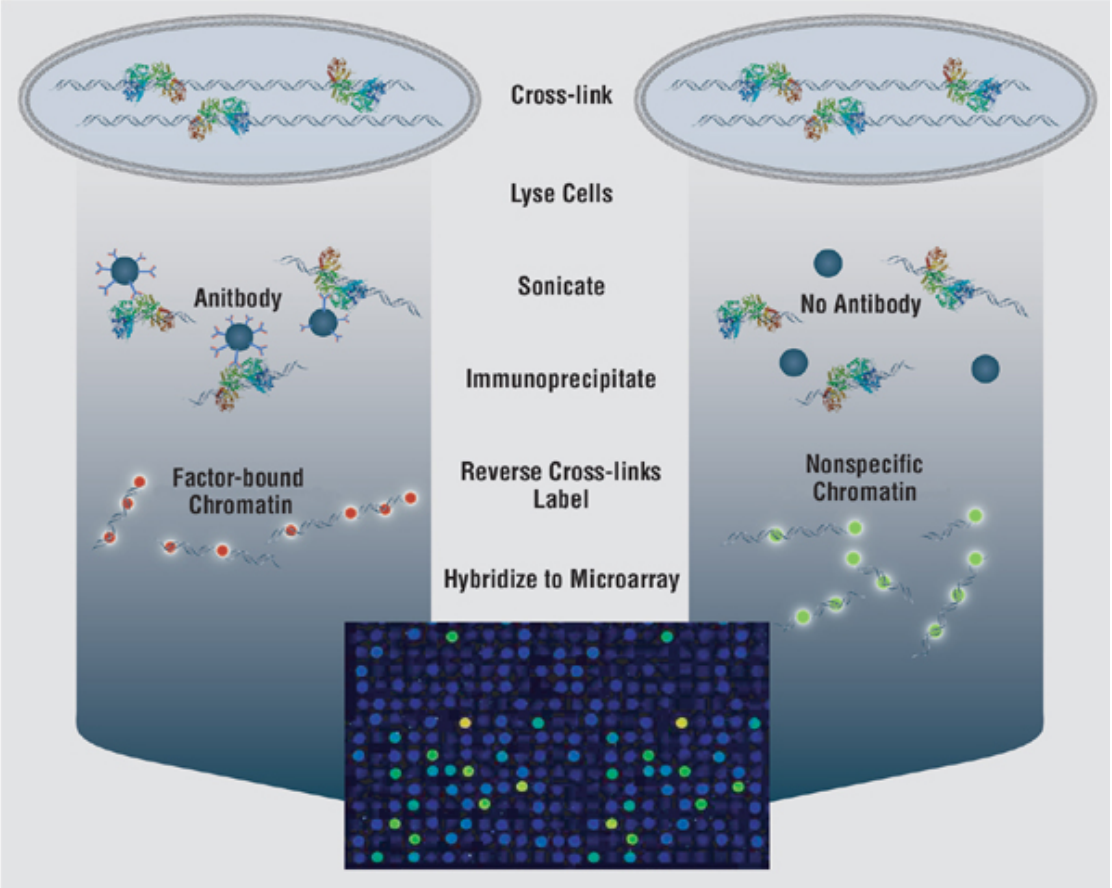
Quantile normalization: Algorithm

1. Given n arrays of length p, form X of dimension $p \times n$ where each array is a column;
2. sort each column of X to give Xsort;
3. take the means across rows of Xsort and assign this mean to each element in the row to get Xsort;
4. get Xnormalized by rearranging each column of Xsort to have the same ordering as original X

Questions

- Differential genes between two conditions (e.g., healthy/diseased)
- Coexpressed genes: pathways, coregulation
- Genes characteristic for particular conditions (e.g., tumor staging)

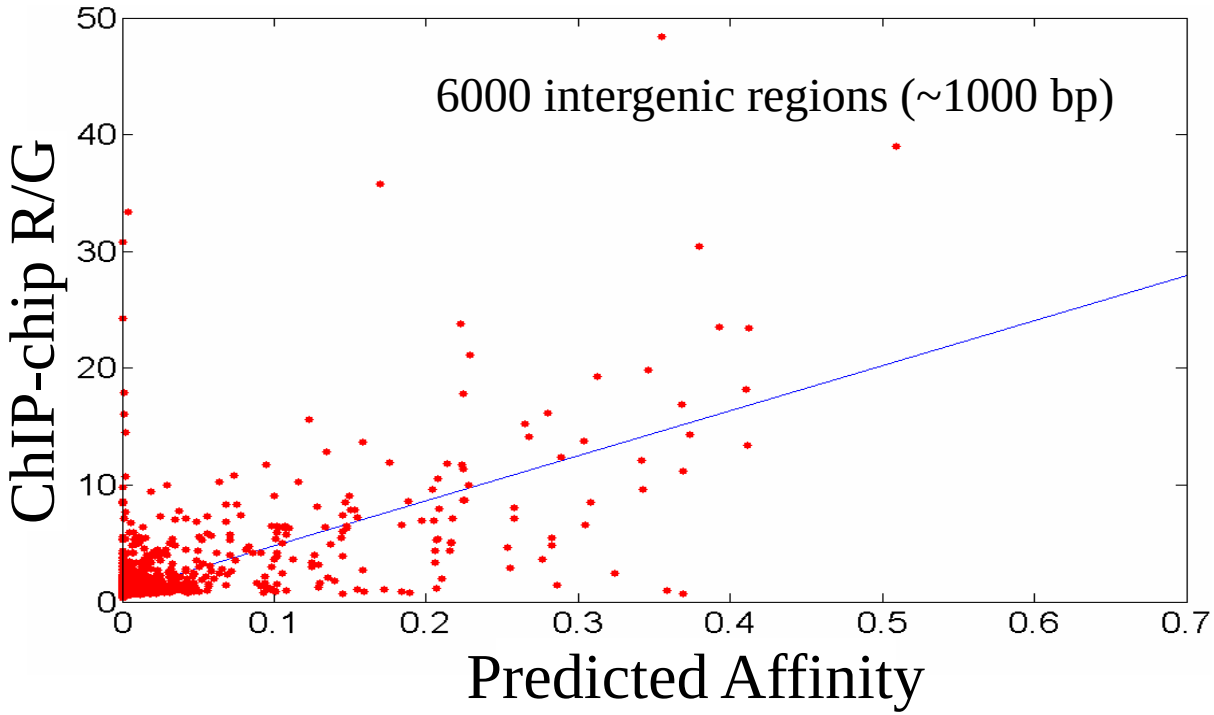
Chromatin IP (ChIP chip)



Tuning the model

ChIP-chip data for yeast: *bound/unbound binding ratios (R/G)*

- ~ 200 transcription factors
- 32 transcription factor descriptions (*TRANSFAC*)

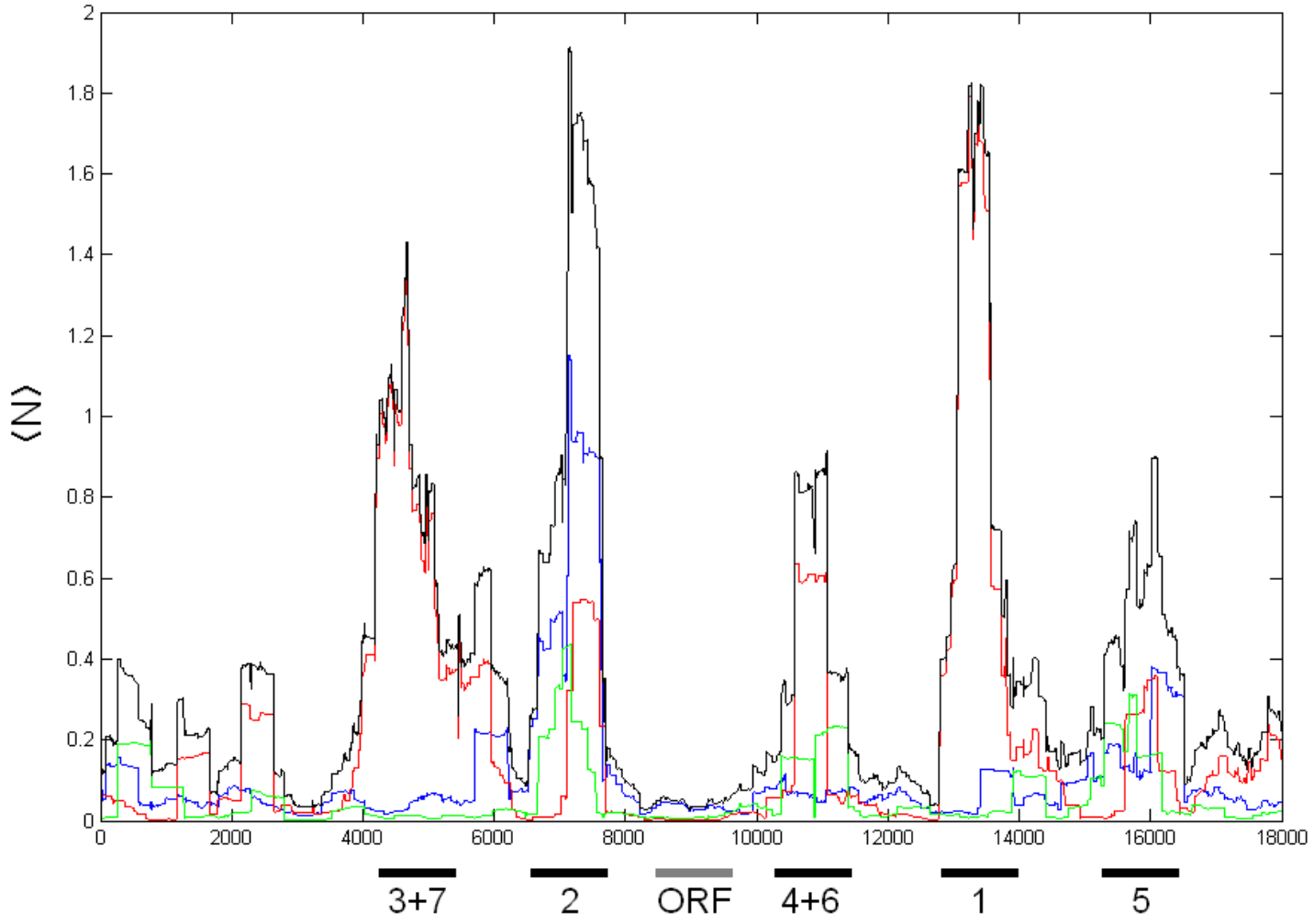


choose parameters, λ and R_0, to achieve ***optimal correlation*** *with R/G-ratios*

Cbf1

Pearson corr = 0.67

Drosophila: Eve-2 Promoter



Krueppel

Hunchback

Bicoid

Kr + Hu + Bcd

Scan with
500 bp region