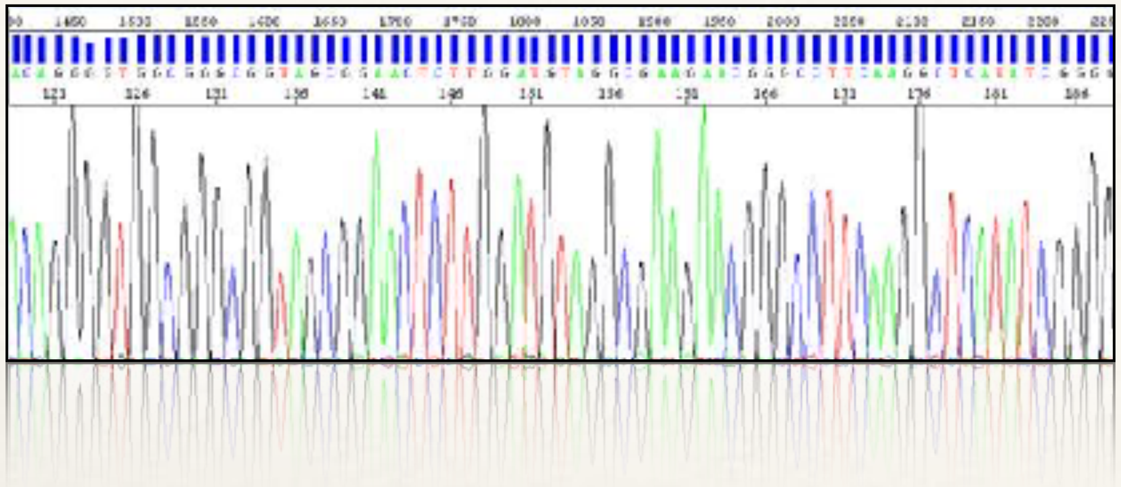
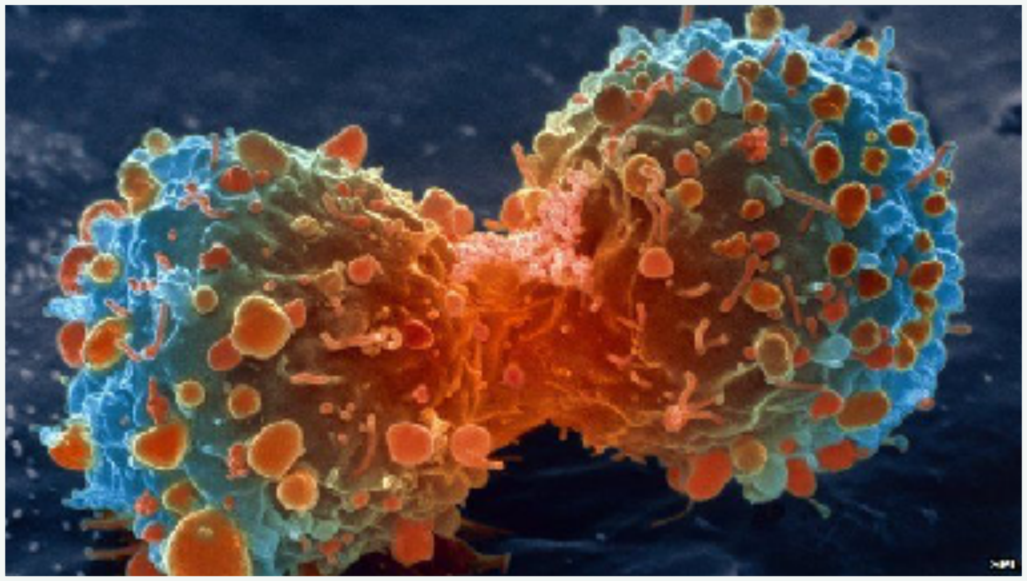


Humboldt Universität zu Berlin

Microarrays

Grundlagen der Bioinformatik
SS 2017

Lecture 6
09.06.2017



Agenda

1. mRNA: Genomic background

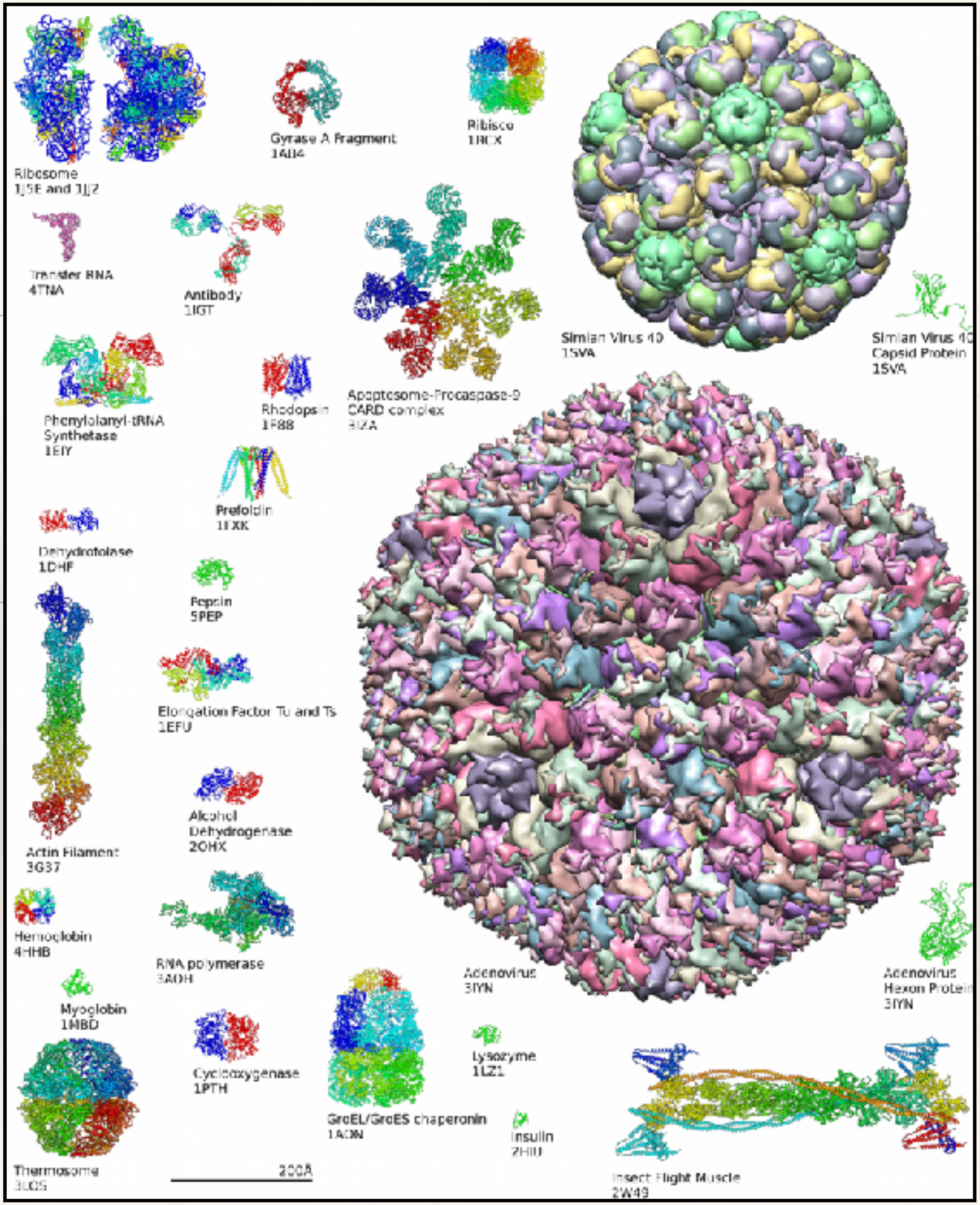
2. Overview: Microarray

3. Data-analysis: Quality control & normalization

Proteins

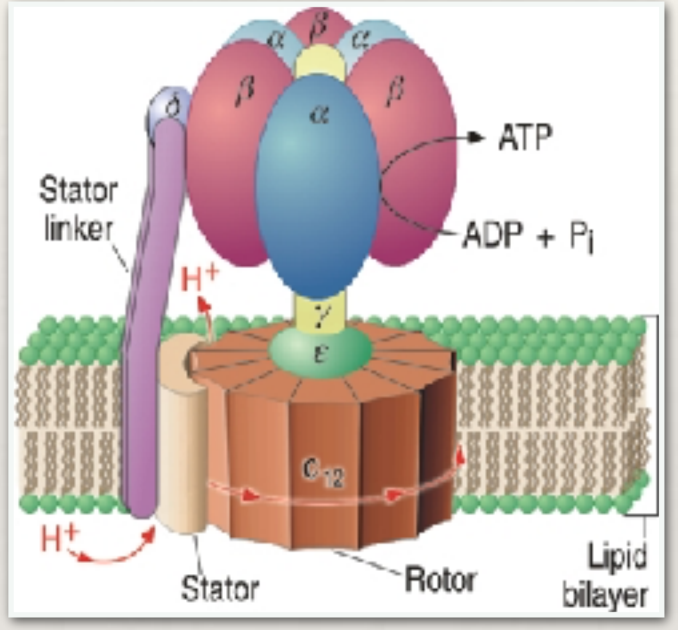
Based on mRNA

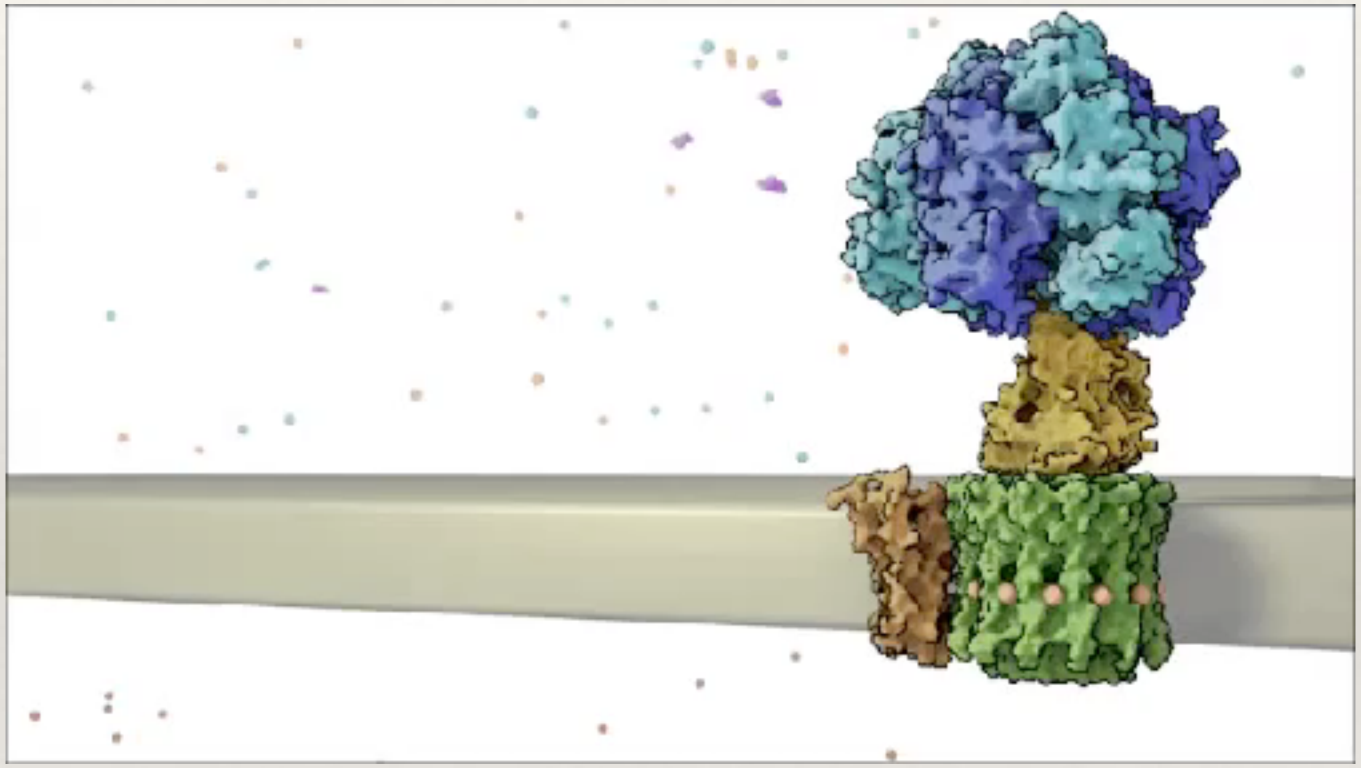
Fit particular purpose and vary with the tasks of the protein



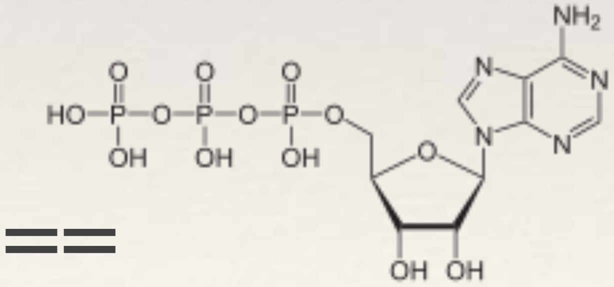
Example: Energy production

Making money (ATP)





==



ATP

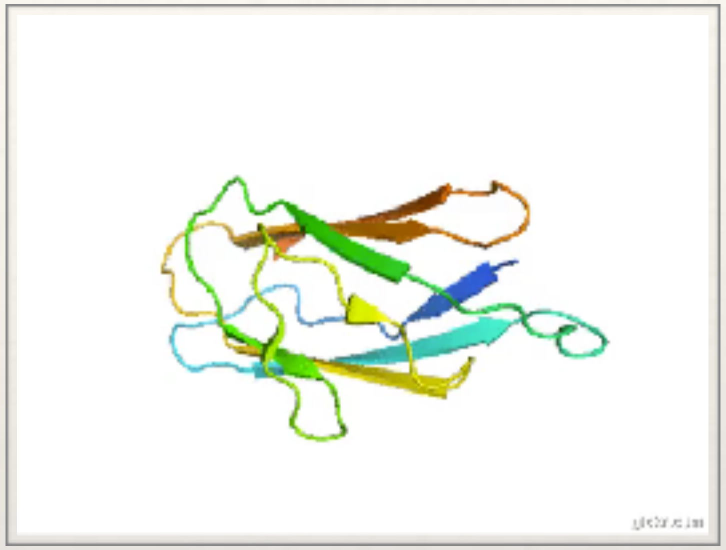
Money-press (ATP)

Source upper figure: http://vle.du.ac.in/mod/book/print.php?id=10108

Source: lehninger principles of biochemistry: international edition

Example: Transport

- Proteins transport molecules

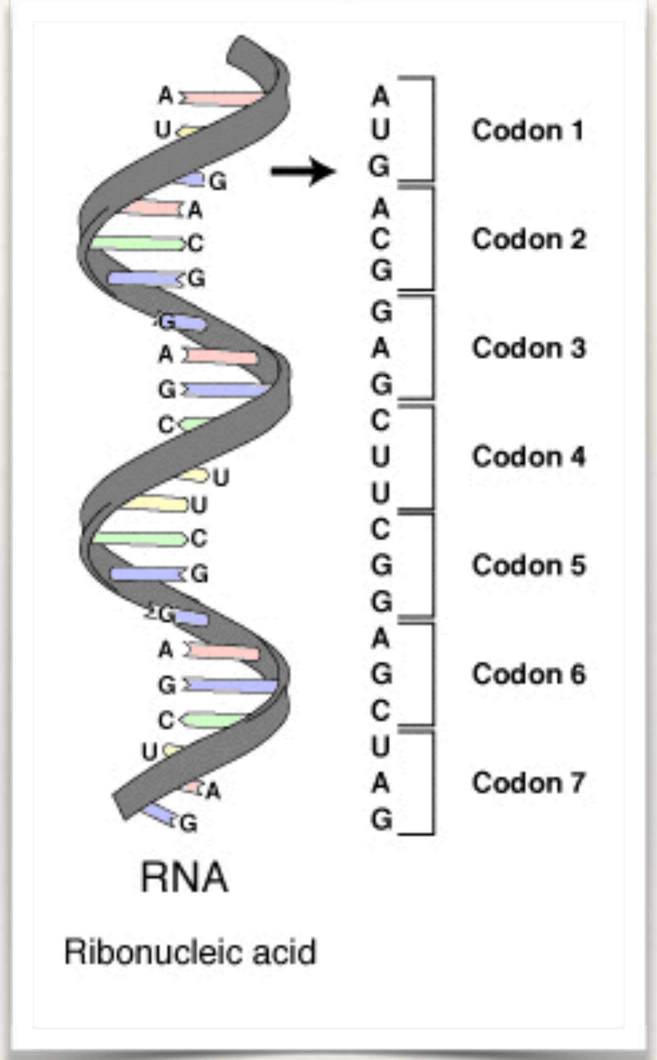


Example Proteins

Source: *Biochemistry*, 2015; Jeremy M. Berg et al.

Connection DNA-protein

- DNA codes for proteins
- One gene, one protein (but different iso-forms)



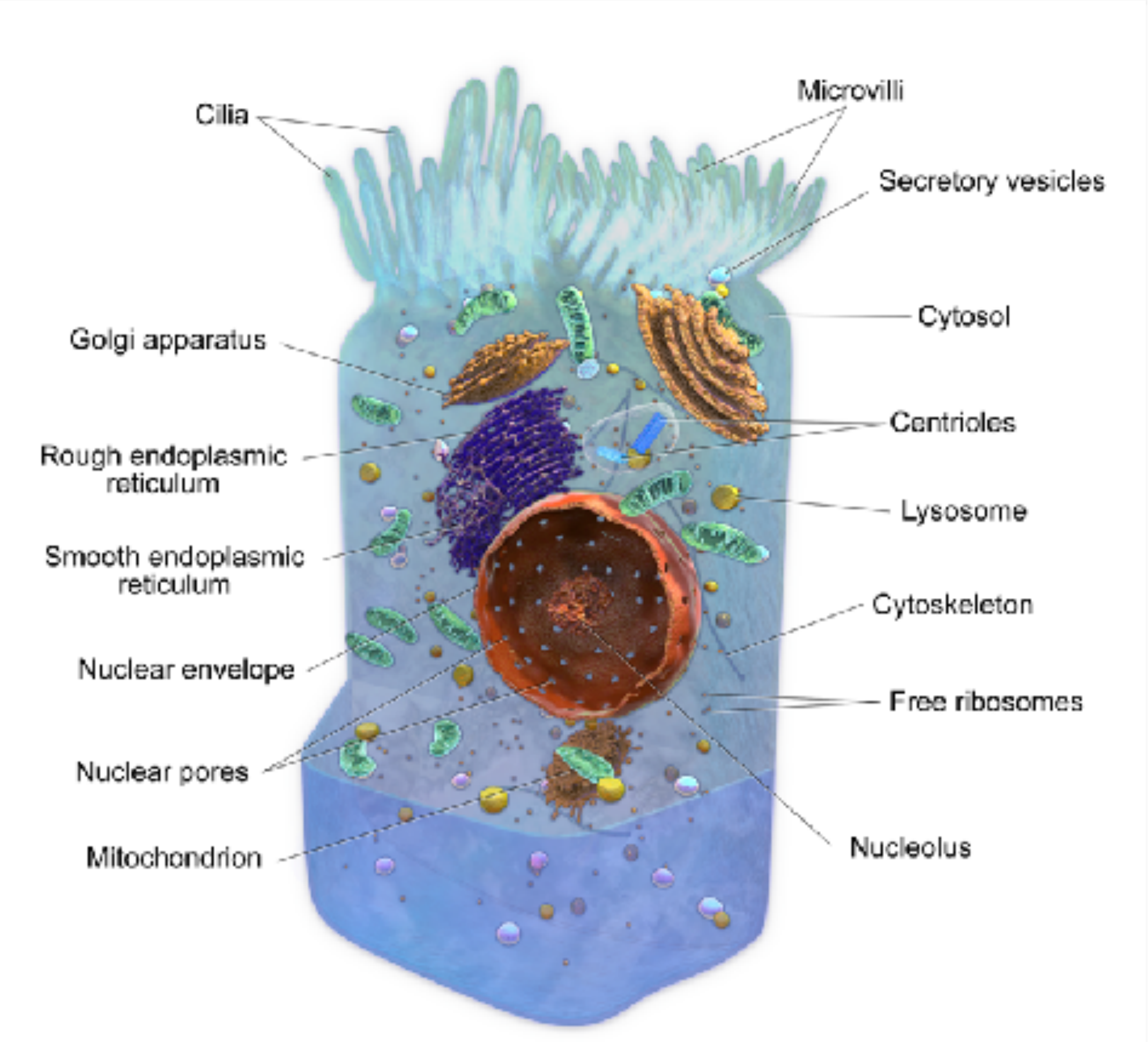
DNA -> mRNA -> Amino acids -> Protein

"RNA-codons" by TransControl - http://en.wikipedia.org/skins-1.5/common/images/magnify-clip.png. Licensed under CC BY-SA 3.0 via Commons - https://commons.wikimedia.org/wiki/File:RNA-codons.png#/media/File:RNA-codons.png

Connection DNA-protein

- **Arrays quantify mRNA**

  located in cytosol & nucleus



Anatomy of a Cell

(Only Eukaryotes)

Polymerases

- Polymerases **read DNA** and **write mRNA**

- Gene activity ~ mRNA production

A polymerase creates a new protein (primary sequence/ mRNA)

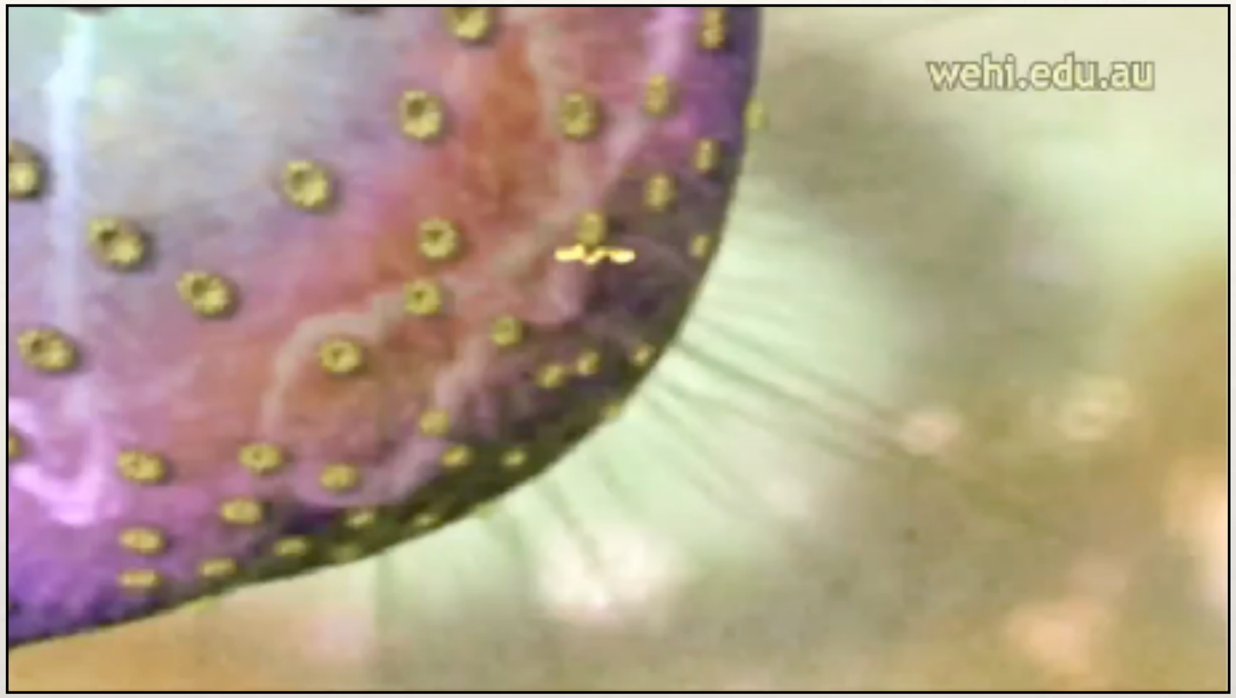
Source: *Biochemistry*, 2015; Jeremy M. Berg et al.

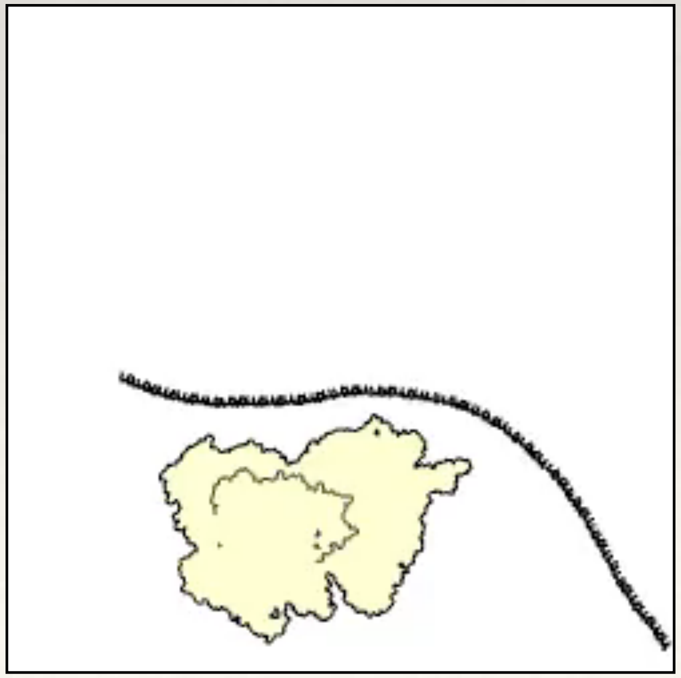
Ribosome - a protein factory

1. Export mRNA from nucleus into cytosol

2. Attach ribosome

3. Make compose amino acids into protein





Source upper animation: The Walter and Eliza Hall Institute is Australia's oldest medical research institute.

Summary Biology

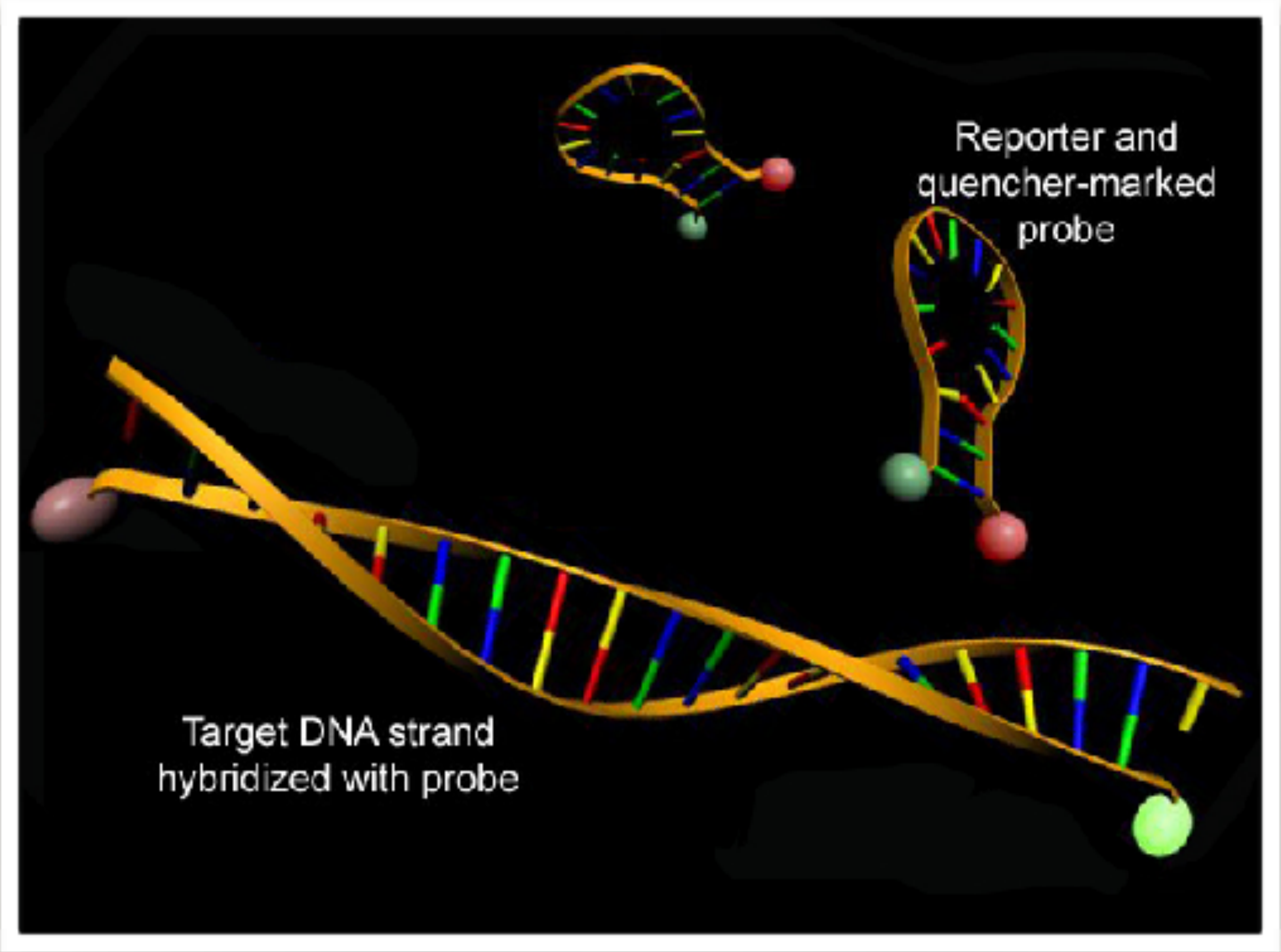
- Gene activity ~ mRNA expression

- Measure mRNA to assess gene-activity

Ideas how to meassure mRNA?

qPCR

- Color (quench) the mRNA
- Old-school
- More light == more mRNA

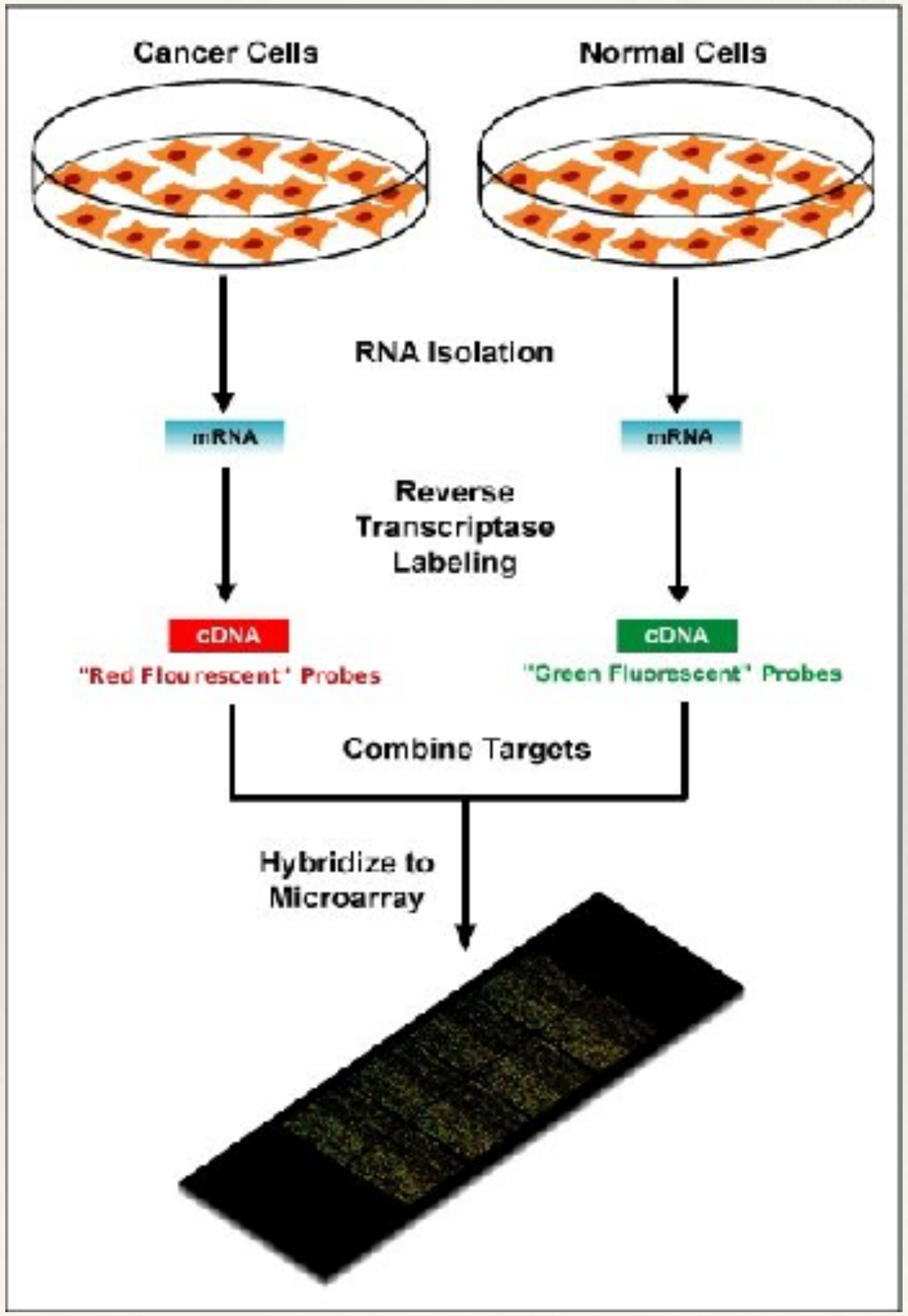

Reporter and quencher-marked probe
Target DNA strand hybridized with probe

Add shiny molecules to quantify

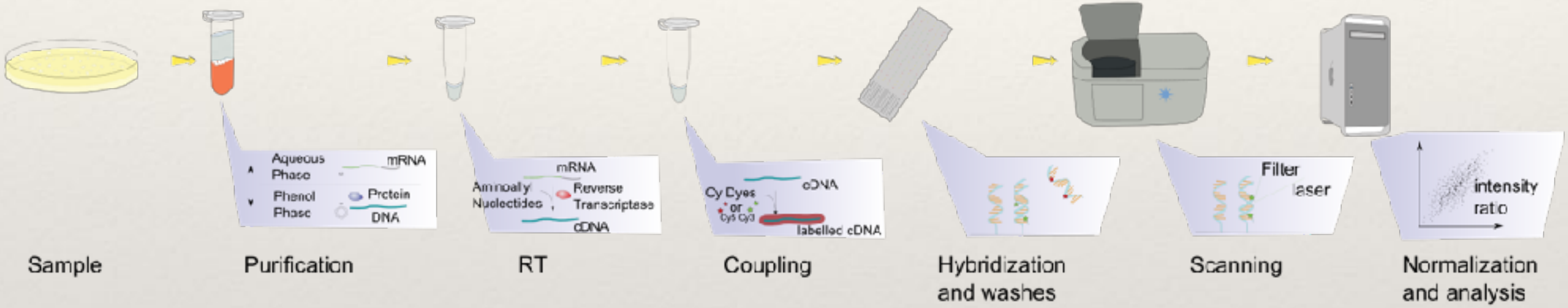
Connection experiment and data

- Identify expression differences between cohorts
  - Cancer vs healthy

- Two types:
  - Relative and absolute measurement



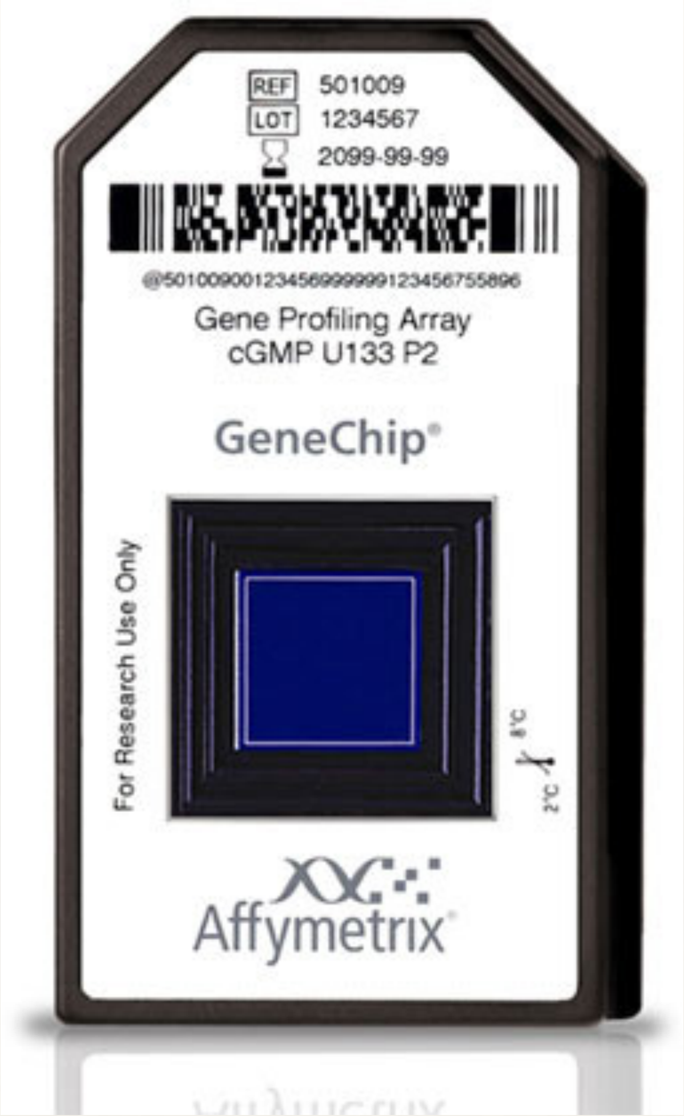
From mRNA to data



Today's topic

mRNA arrays

- Same light-principle as qPCR
- Fixed RNA sequences on a chip
- mRNA binds to chip
- More binding => more signal



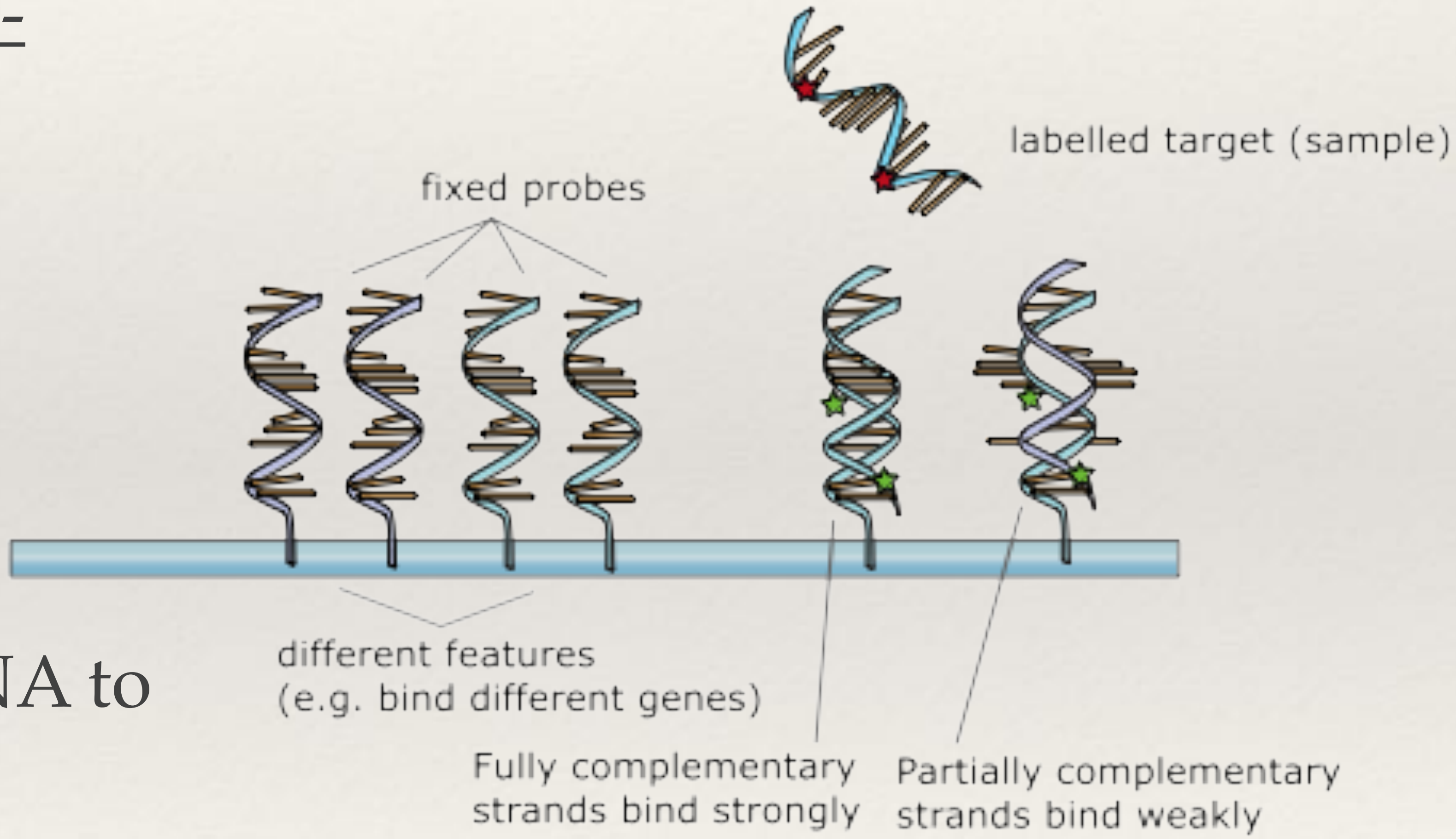
Affymetrix mRNA Array

Hybridization

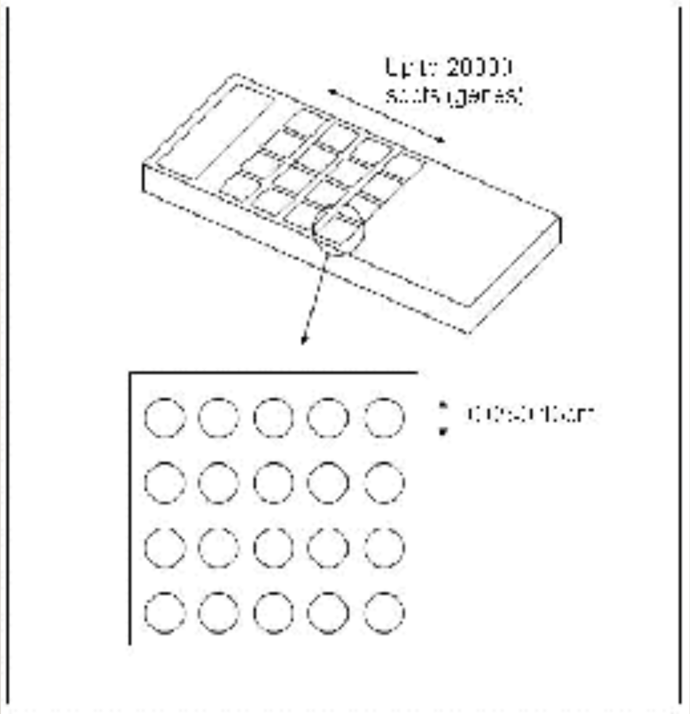
- Chip is collection of single-strand DNA-sequences

- mRNA labeled

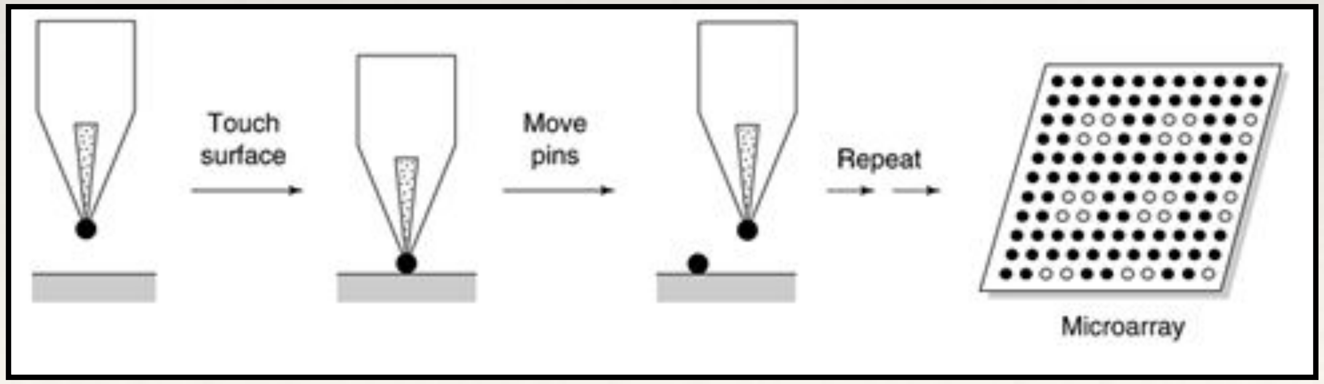
- **Hybridization** binds mRNA to probes



cDNA (spotted) array



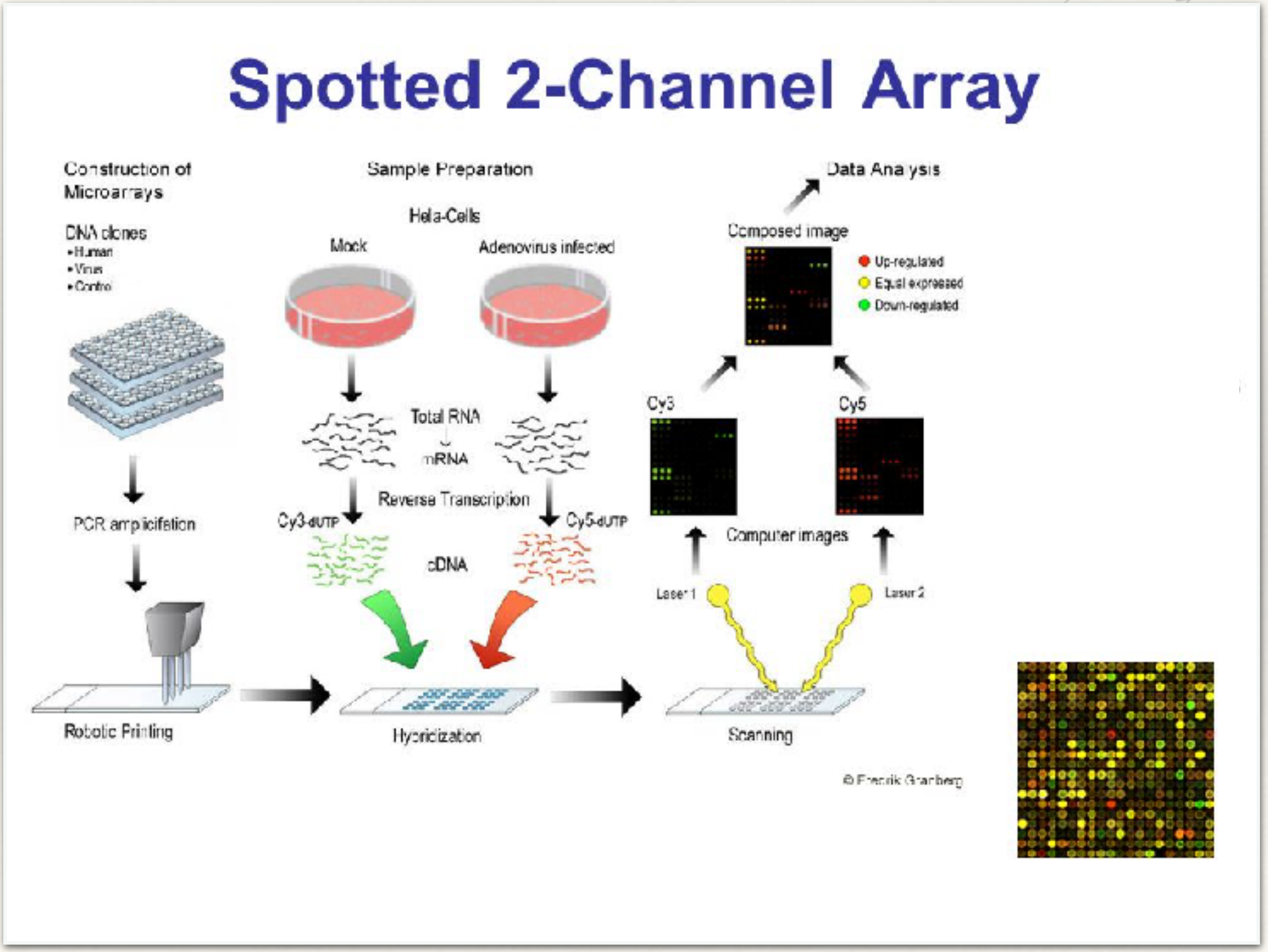
- One (long) probe = one gene
- Perfect match probe to gene
- Generic background correction



Construction: Print it!

Relative mRNA measurement

- Two samples
- Measure on same chip
- Color shows differences

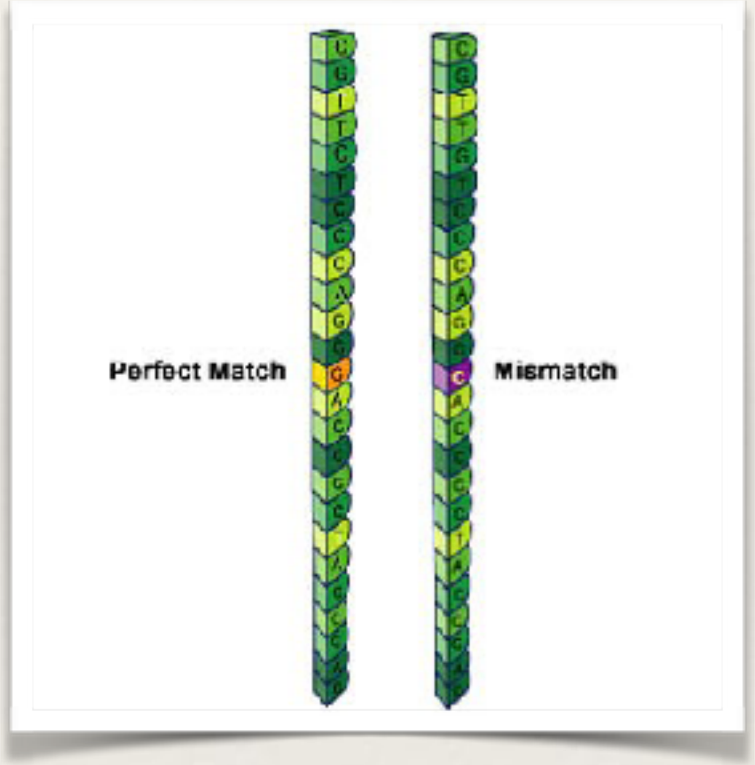


Spotted 2-Channel Array

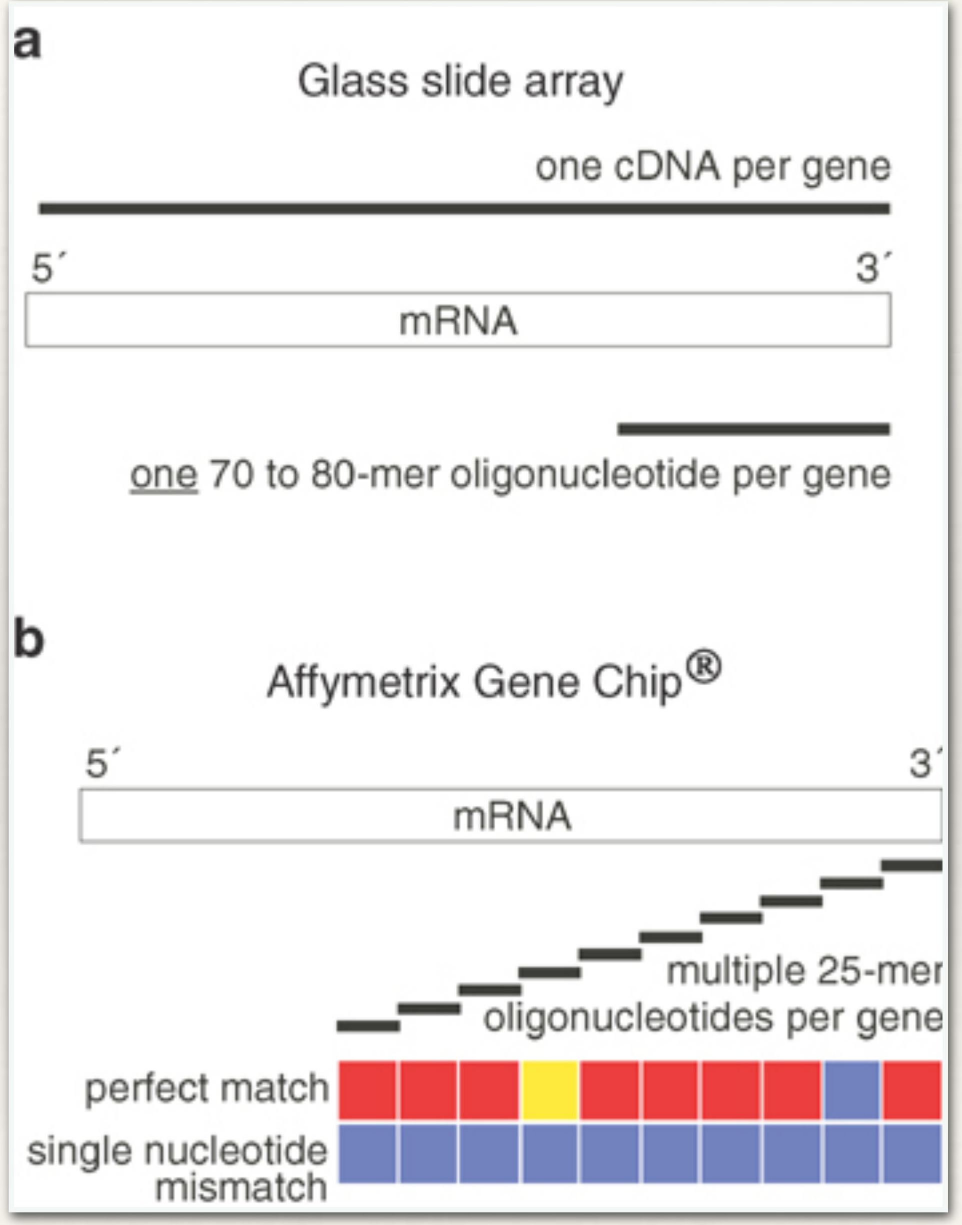
© Fredrik Granberg

Oligonucleotide array

- Probes cover only parts of genes
- Probes short: 25-60 nucleotides
- Mismatches build into probe



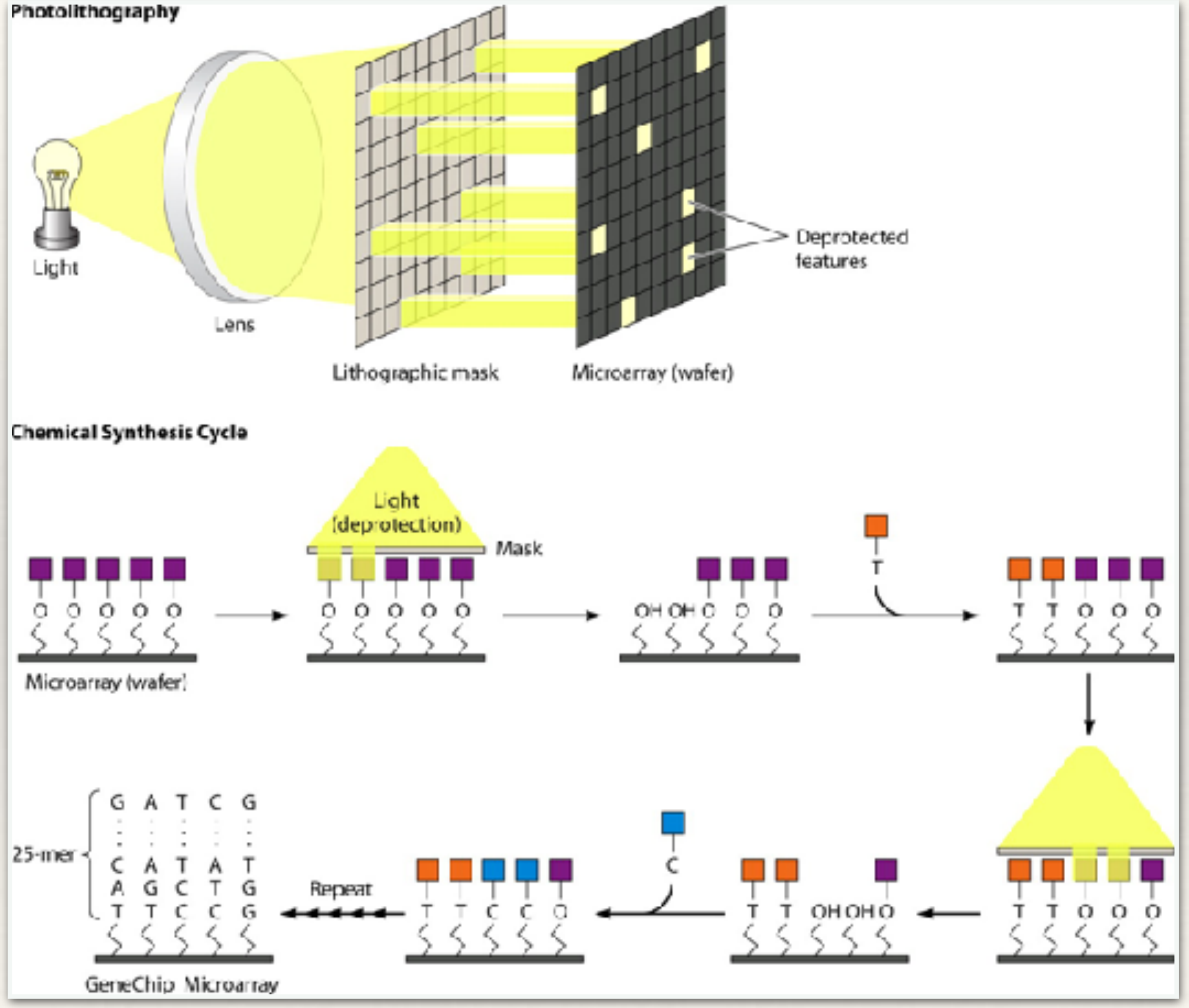
Probes contain mismatch


a
Glass slide array
one cDNA per gene
5´
3´
mRNA
one 70 to 80-mer oligonucleotide per gene
b
Affymetrix Gene Chip®
5´
3´
mRNA
multiple 25-mer oligonucleotides per gene
perfect match
single nucleotide mismatch

Difference oligo vs spotted

Constructing an oligo-Array

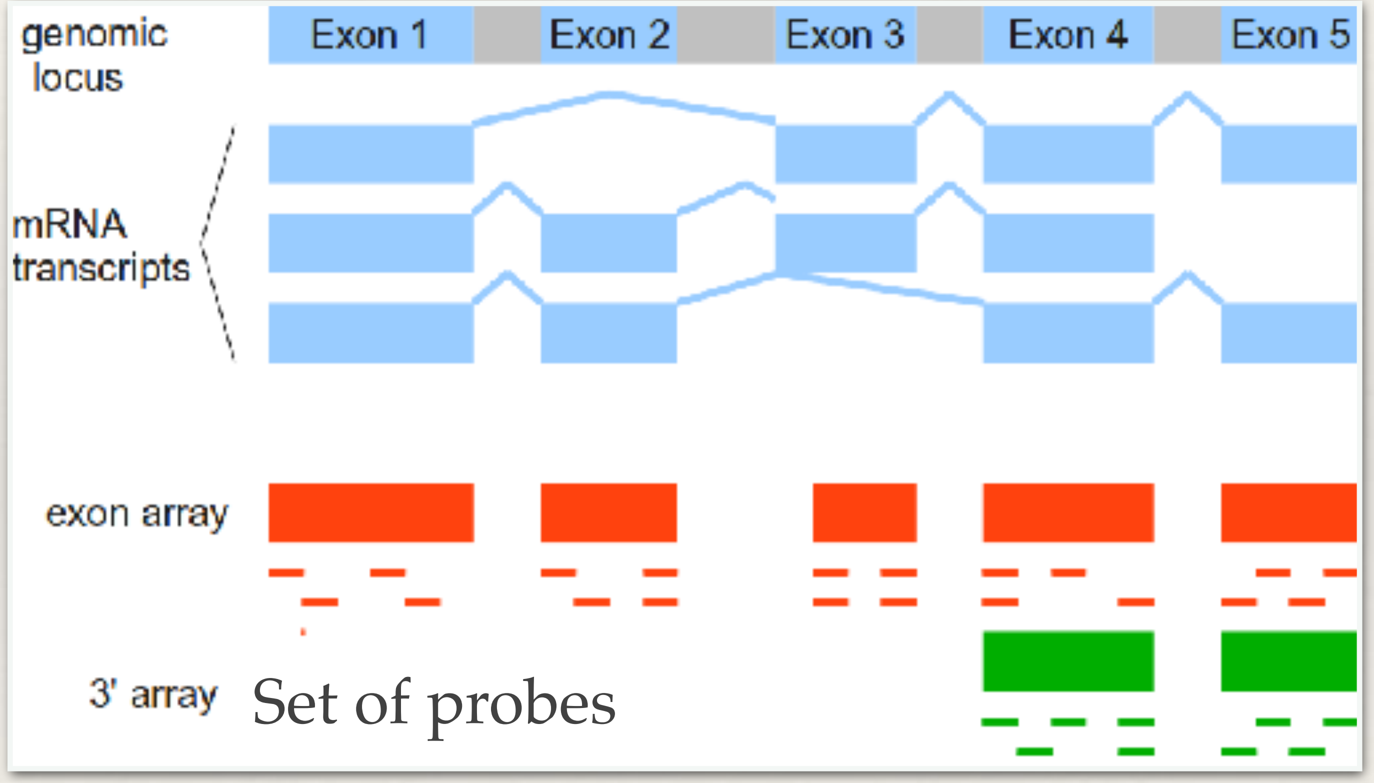
- Build probes piece by piece (nucleotide)
- Block e.g. all but As and add As
- Let As connect, wash and go to e.g. Cs



Miller MB, Tang Y-W. Basic Concepts of Microarrays and Potential Applications in Clinical Microbiology. Clinical Microbiology Reviews. 2009;22(4):611-633. doi:10.1128/CMR.00019-09.

Exon Arrays

- Exon arrays
- Measure mRNA-exons
- Allows detection of gene-isoforms



Set of probes

Focus measurement on exons

Quality control

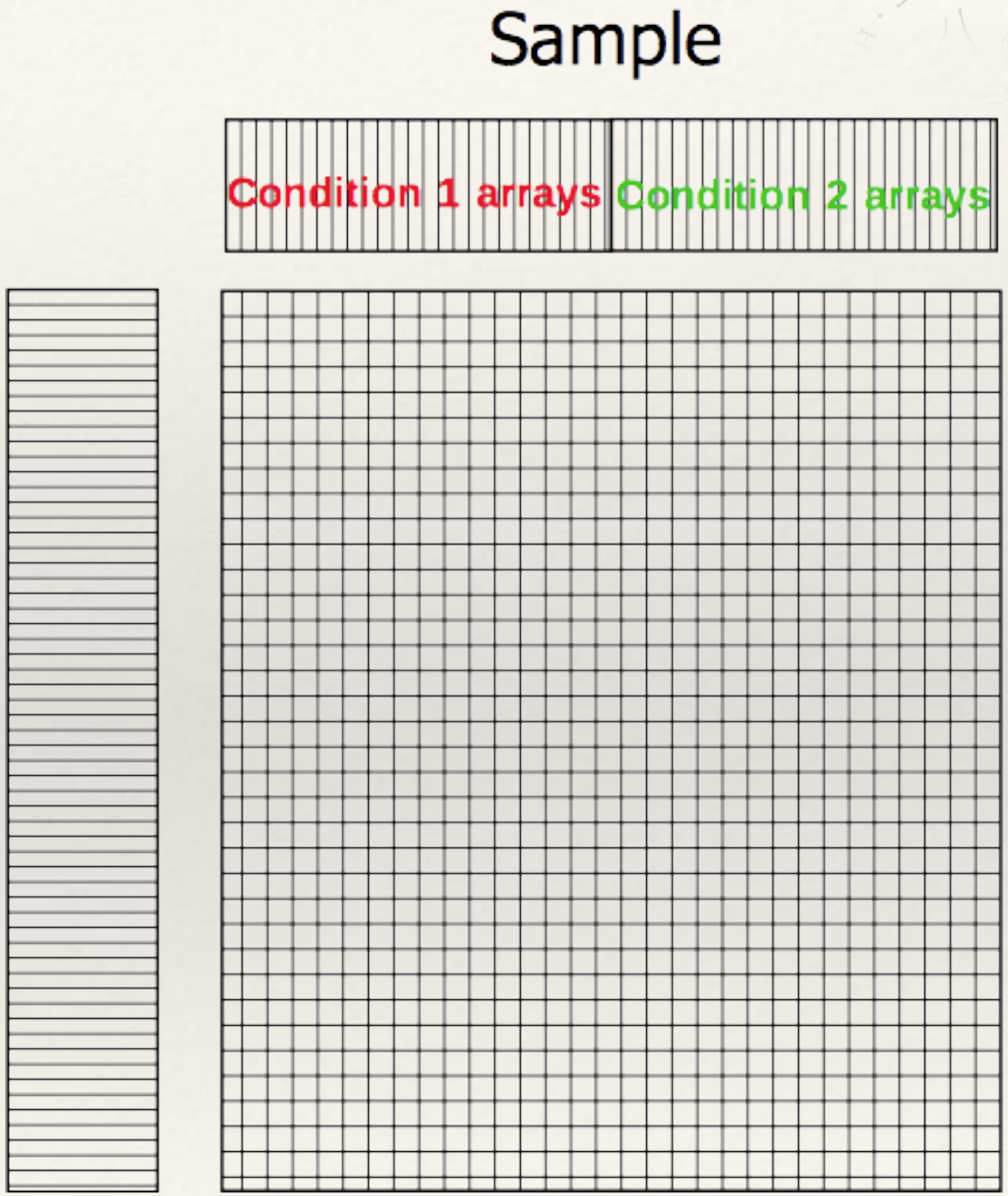
Controling for noise, biases and errors is critical

Biological source

Technical source

Analytical challenges

- Separate signal from noise
- High variance within cohorts and samples
- Technological differences
- Curse of high numbers



Cohort-concept

Benchmark

- How to measure Array performance
- Various criteria:
  - Sensitivity
  - Specificity
  - Biases etc.

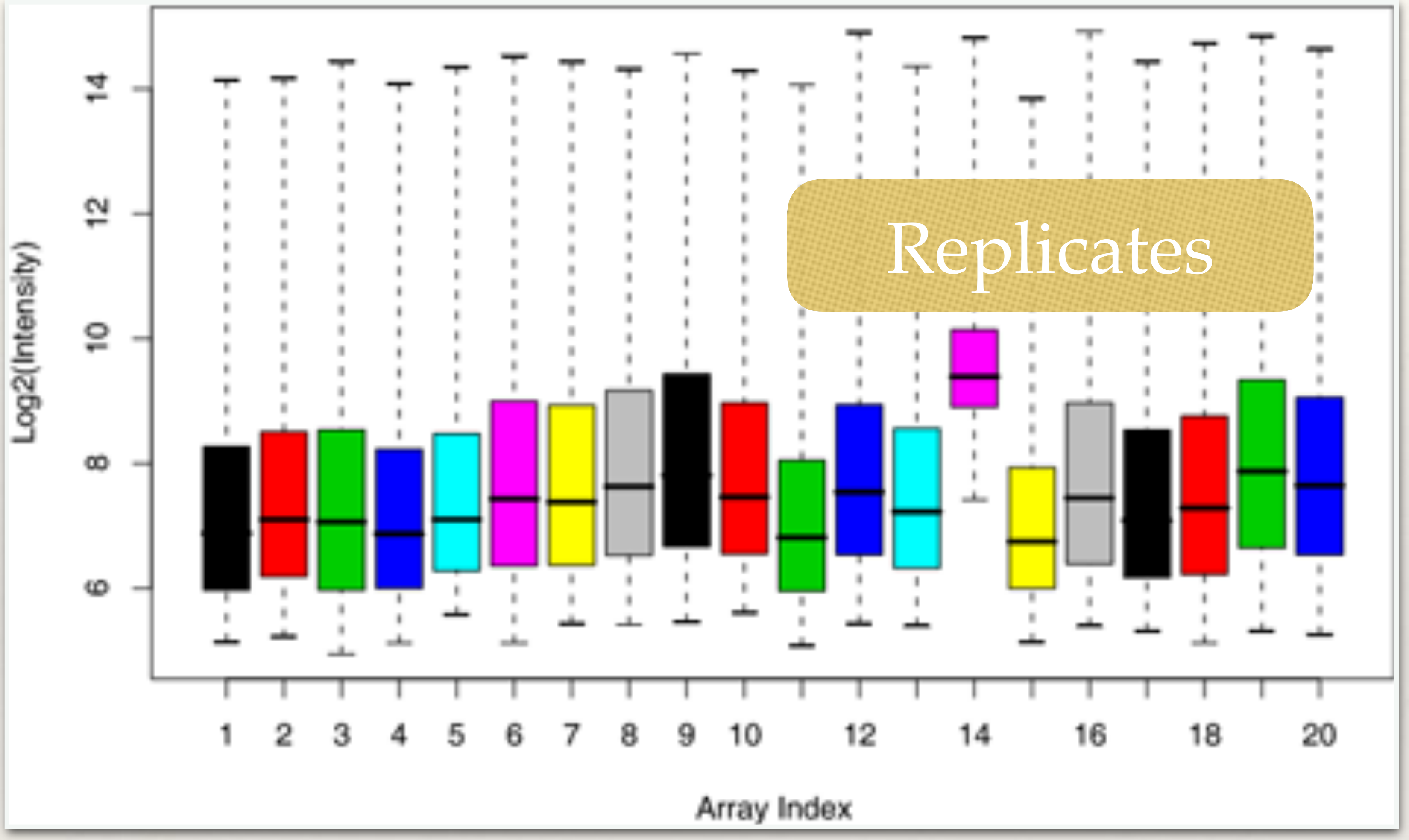
$$SN = \frac{TP}{TP + FN}$$

Sensitivity:
Correctly count number of mRNA-molecules

$$SP = \frac{TN}{TN + FP}$$

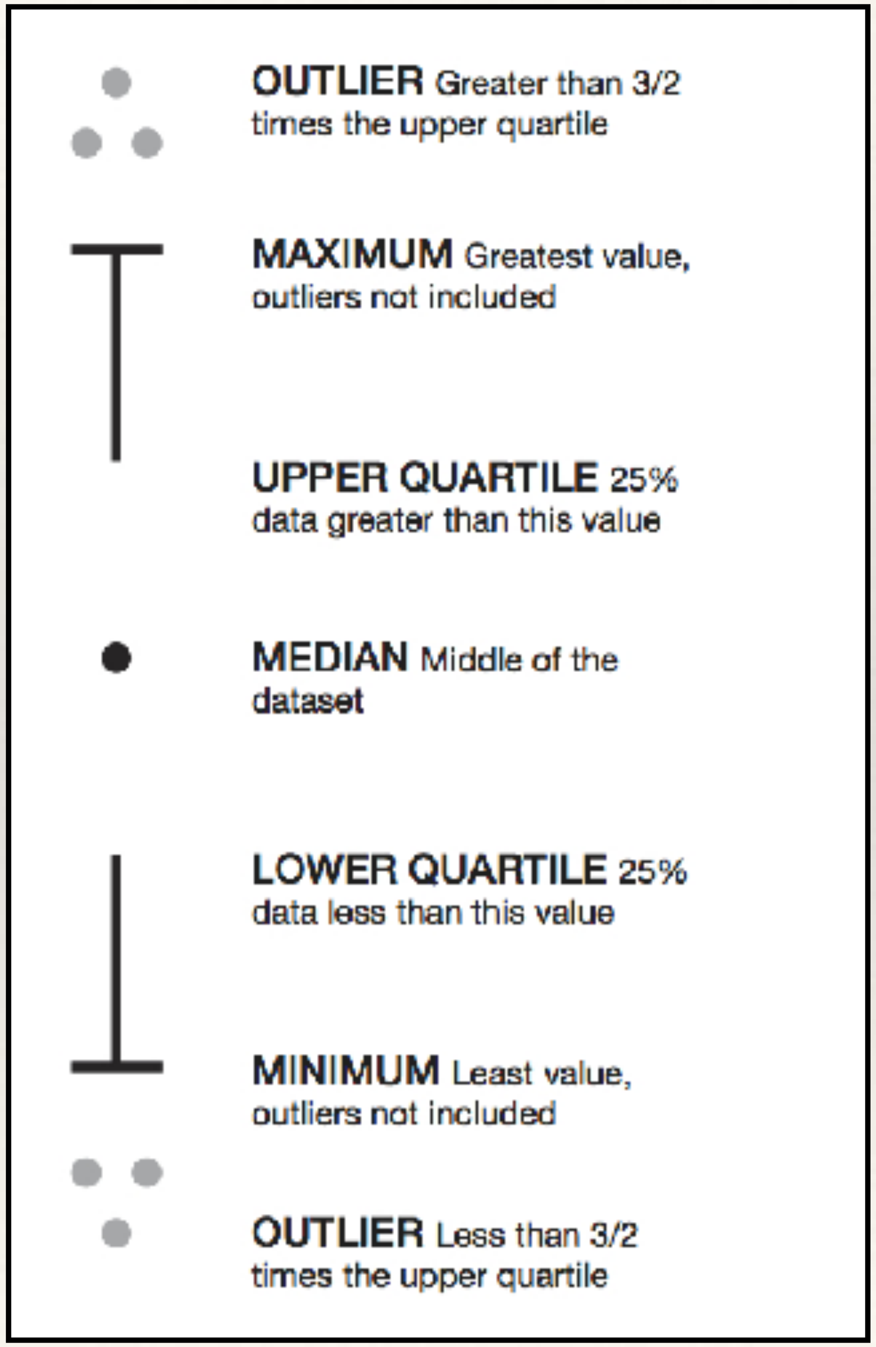
Specificity:
Do not count other mRNA-molecules

Quality control



Replicates

Compare raw signal within cohort
Identify e.g. outlier



OUTLIER Greater than 3/2 times the upper quartile

MAXIMUM Greatest value, outliers not included

UPPER QUARTILE 25% data greater than this value

MEDIAN Middle of the dataset

LOWER QUARTILE 25% data less than this value

MINIMUM Least value, outliers not included

OUTLIER Less than 3/2 times the upper quartile

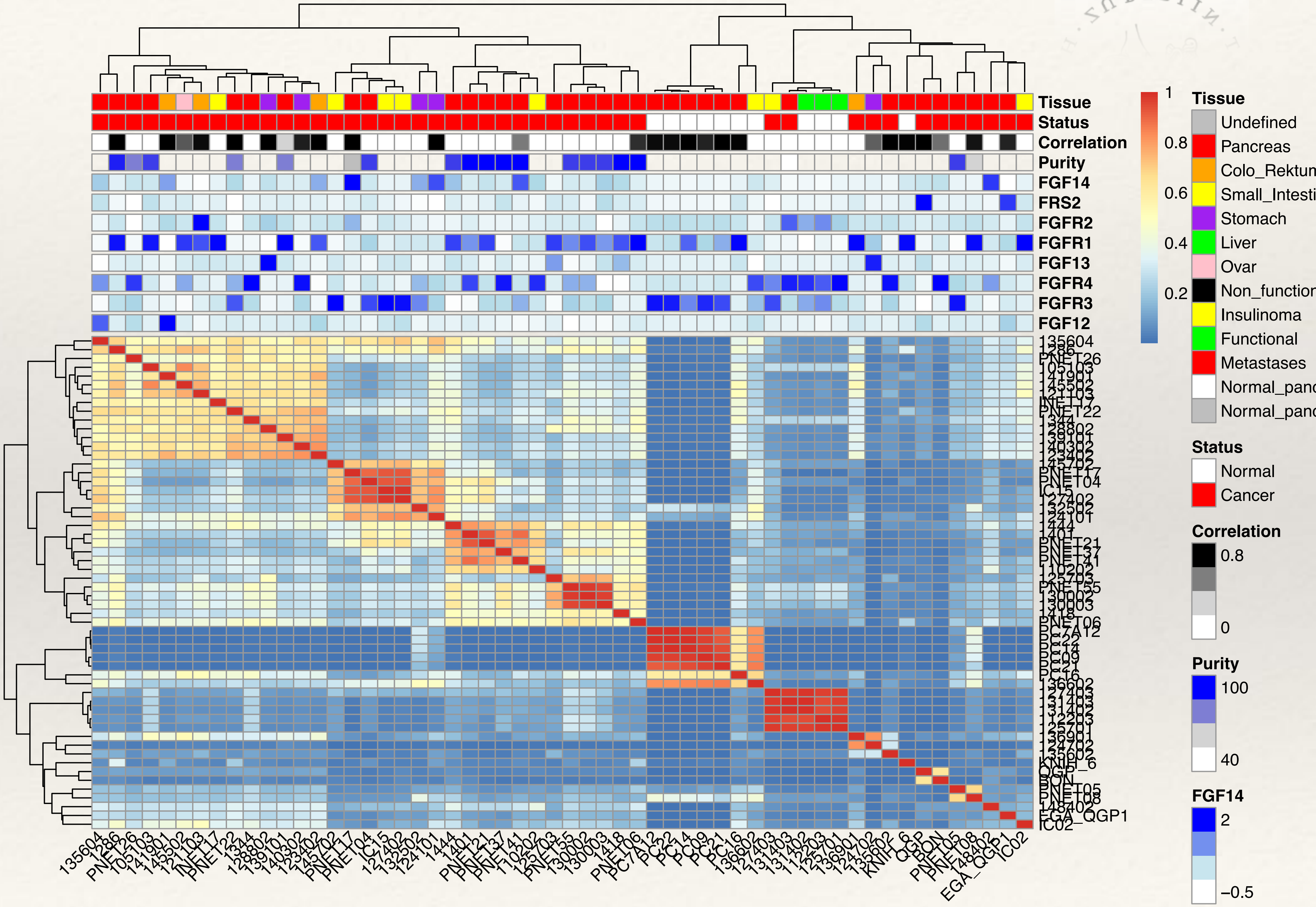
Replication

- Compensate noise
- Understand biology
- Variance-estimation

- Technical replicate
  - Estimate technical variance
- Biological replicate
  - Estimate biological variance

Correlation plot

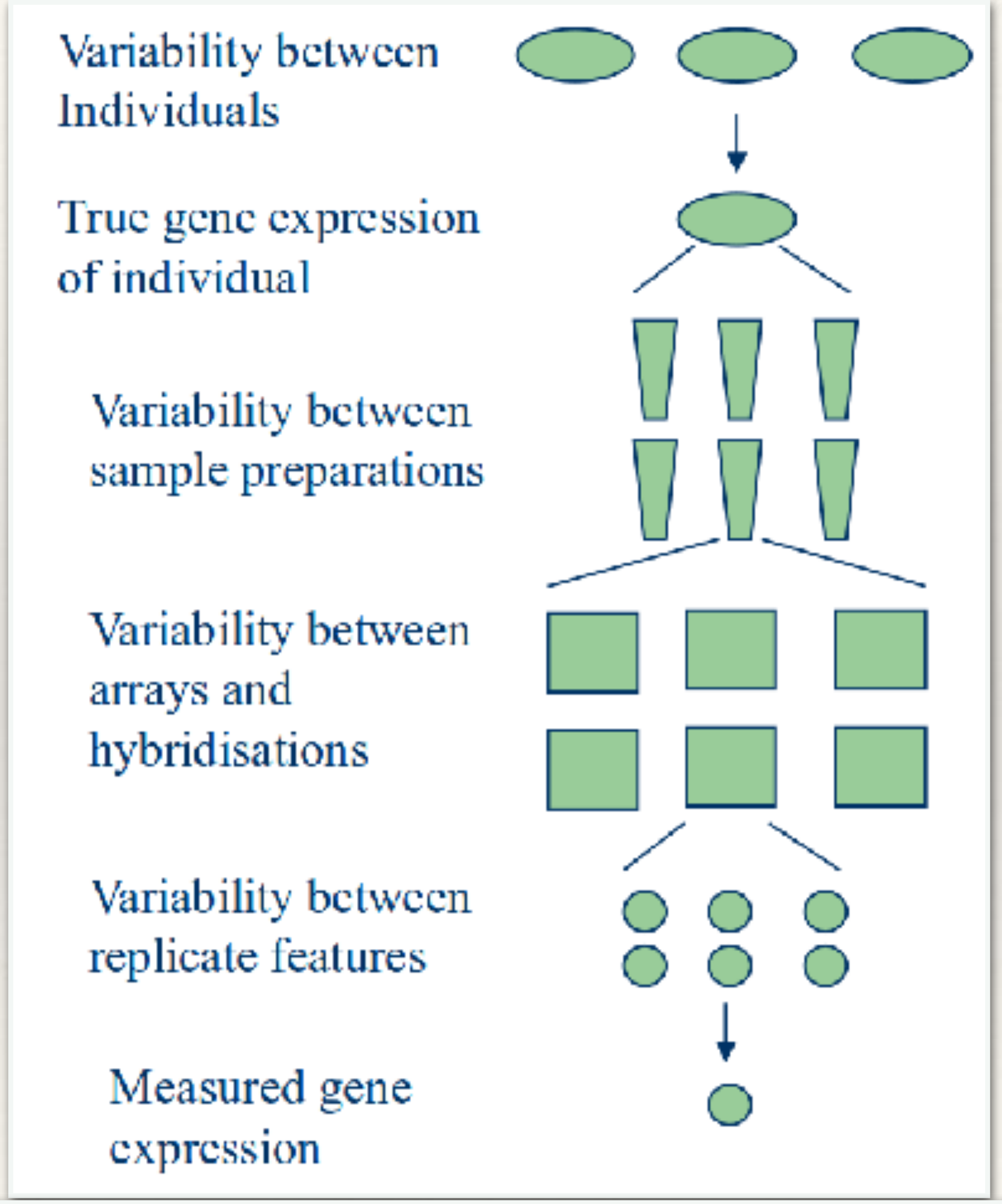
- Group similar samples
- Correlation allows interpretation



Messy real world data

Data-normalization

- Make data comparable
- Data is <u>not directly comparable</u>
- Identify true values
- Identify true variance


Variability between Individuals
True gene expression of individual
Variability between sample preparations
Variability between arrays and hybridisations
Variability between replicate features
Measured gene expression

Make your data great again

Data-normalization

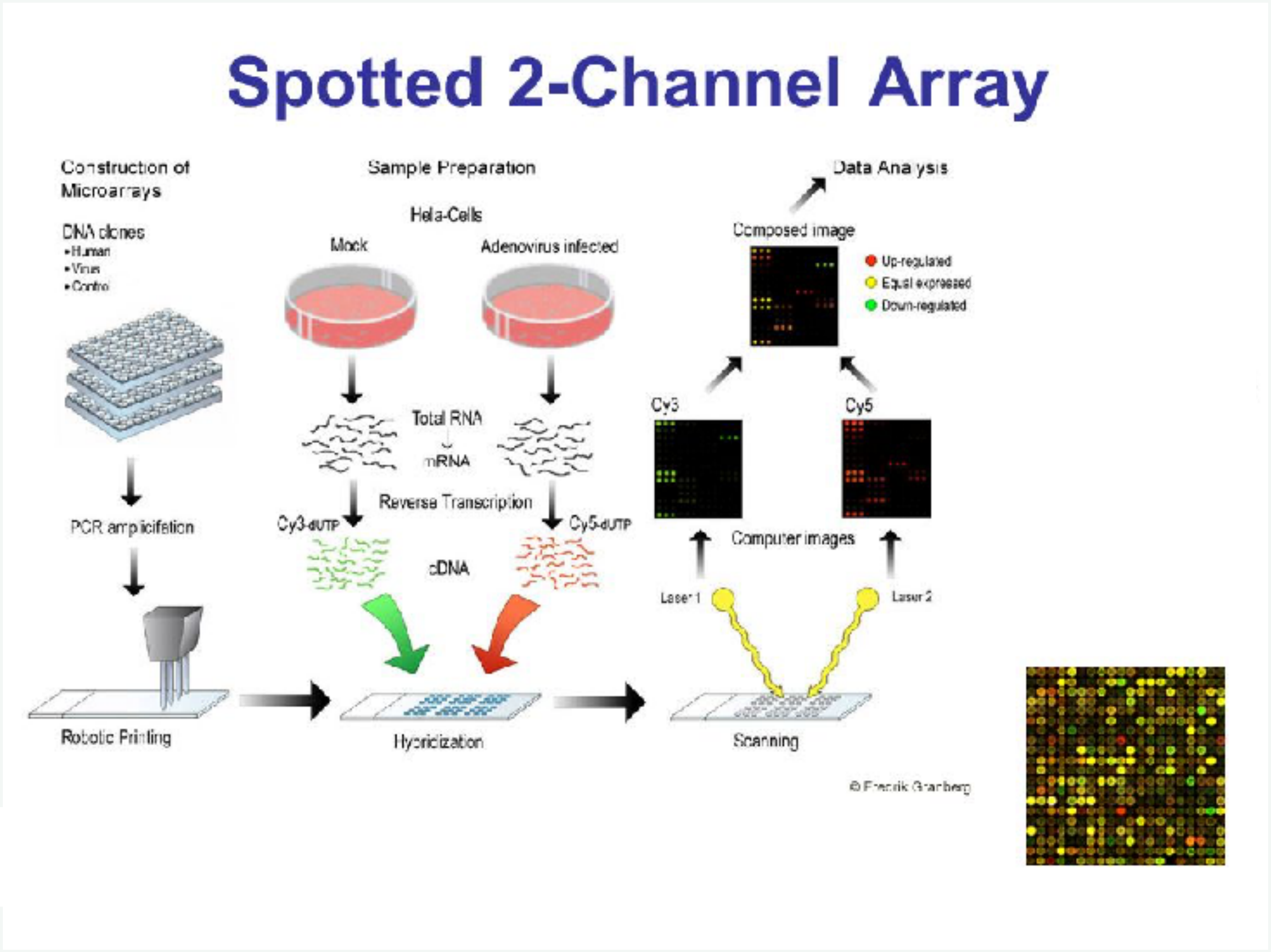
Key assumption

- E.g. 2 x mRNA amount leads to 2 x signal intensity

Noise and bias are linear effects

- Quantify the linear effects and correct them

Reminder

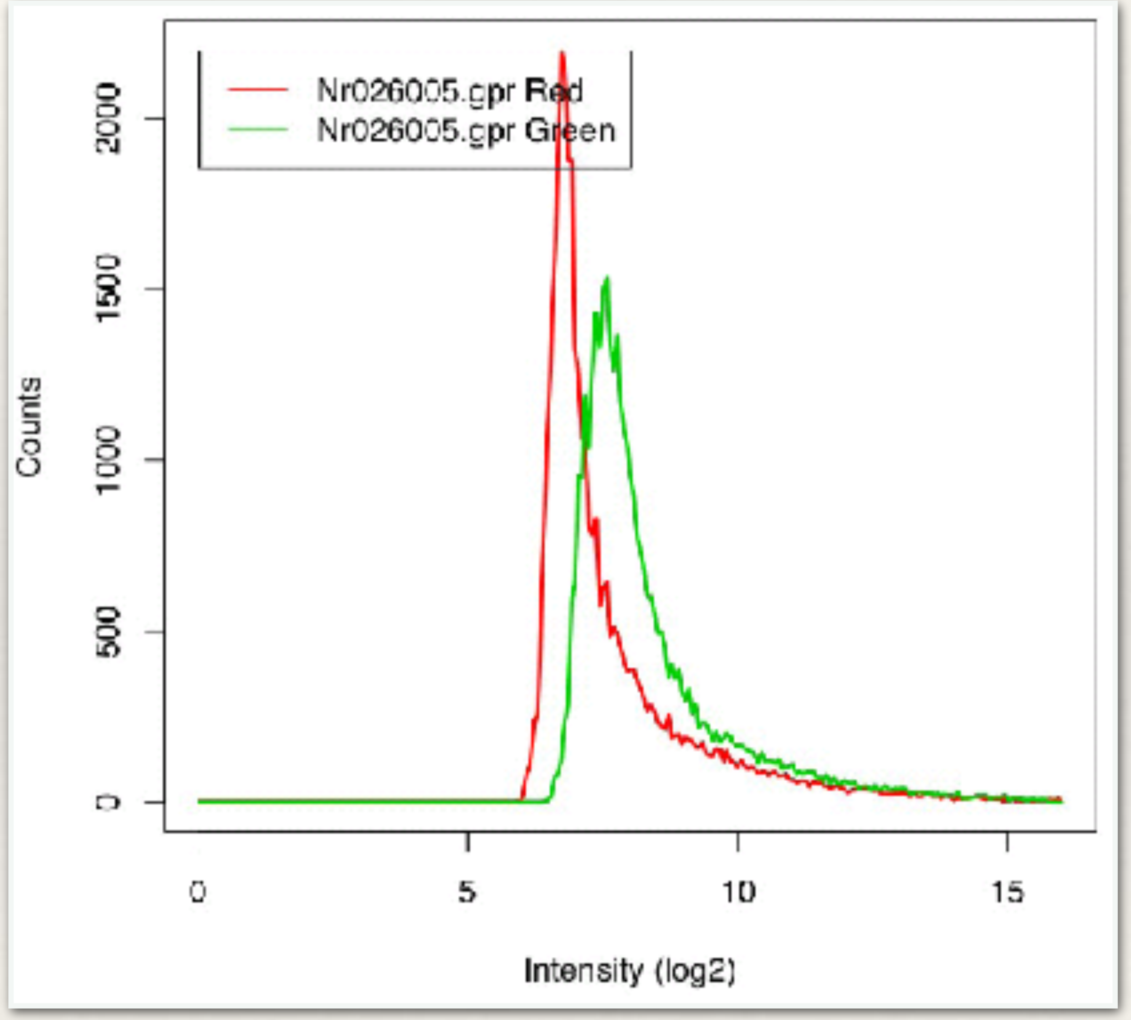


Spotted 2-Channel Array

Example Dye-correction

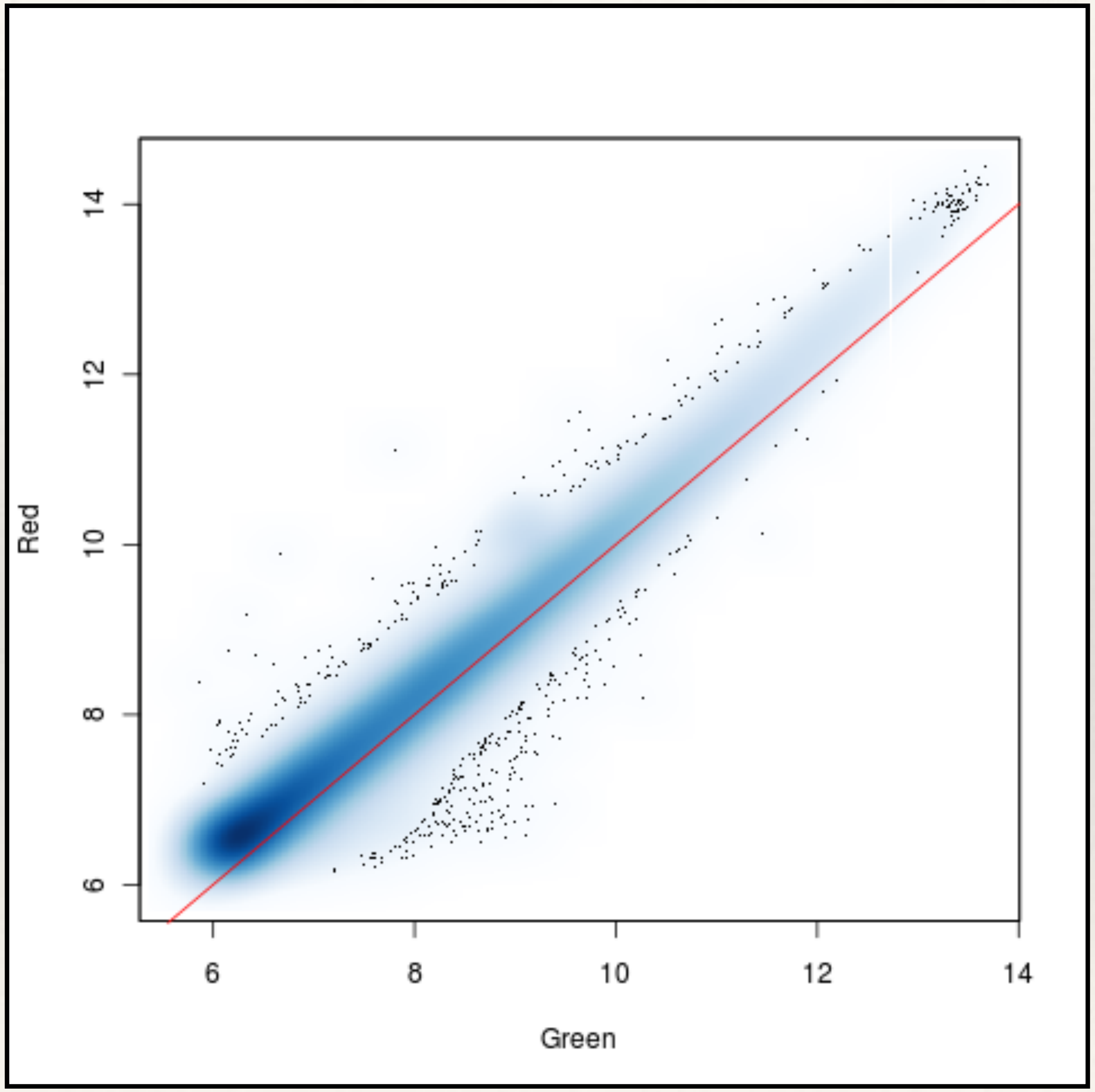
- 2-color spotted array

- Green dye brighter than red



Scatter plot

- Dot = Gene
- Describe data
- Visualize bias

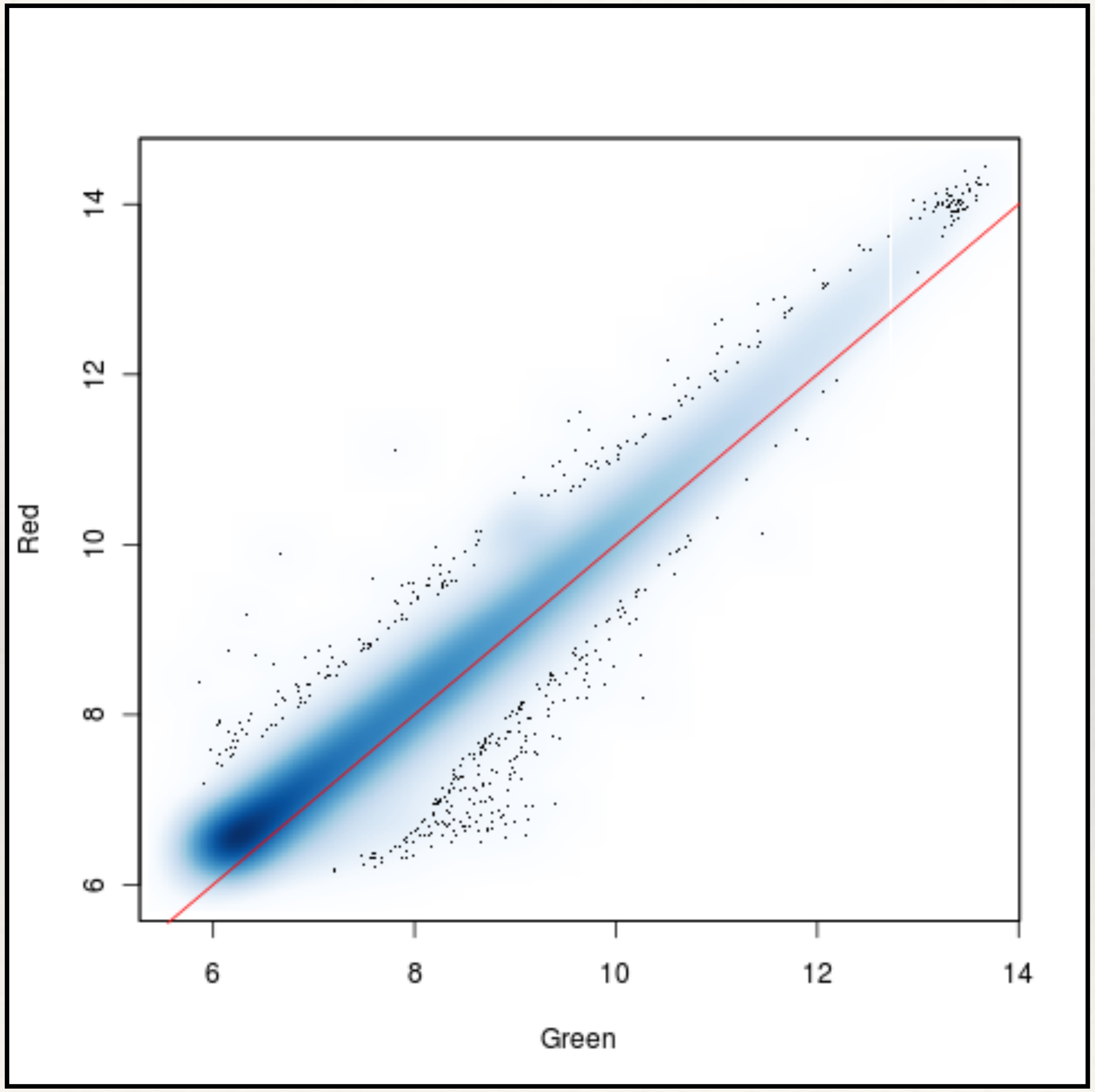


Shown: Same input, different channel

Solution compensate by calculation

1. Find formula to describe bias

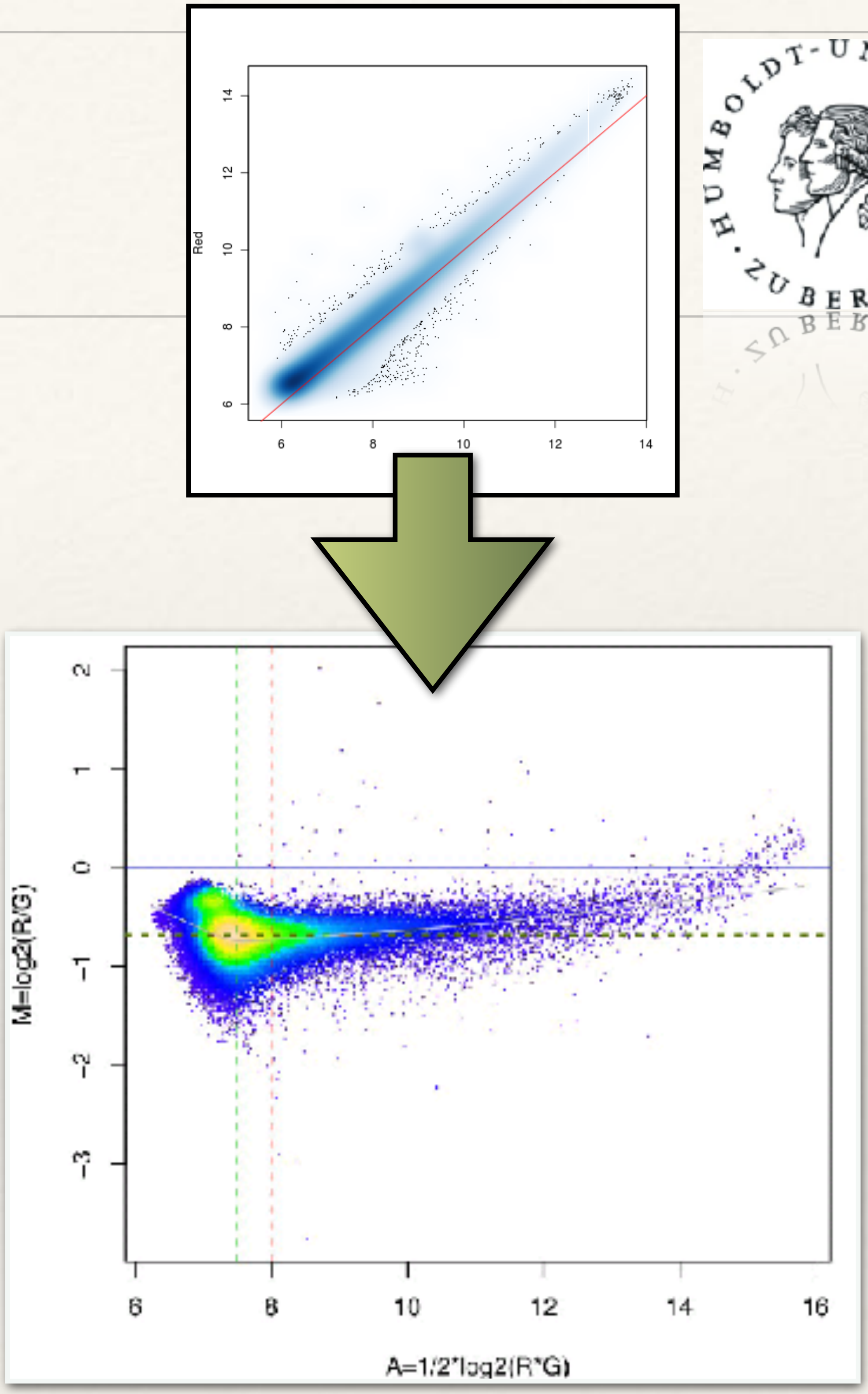
2. ‚Correct' bias (fiddle numbers)



Shift the signal according to intensity

MA-Plot

- Difference-to-intensity plot
- Discretizes bias
- Limited by data-quality

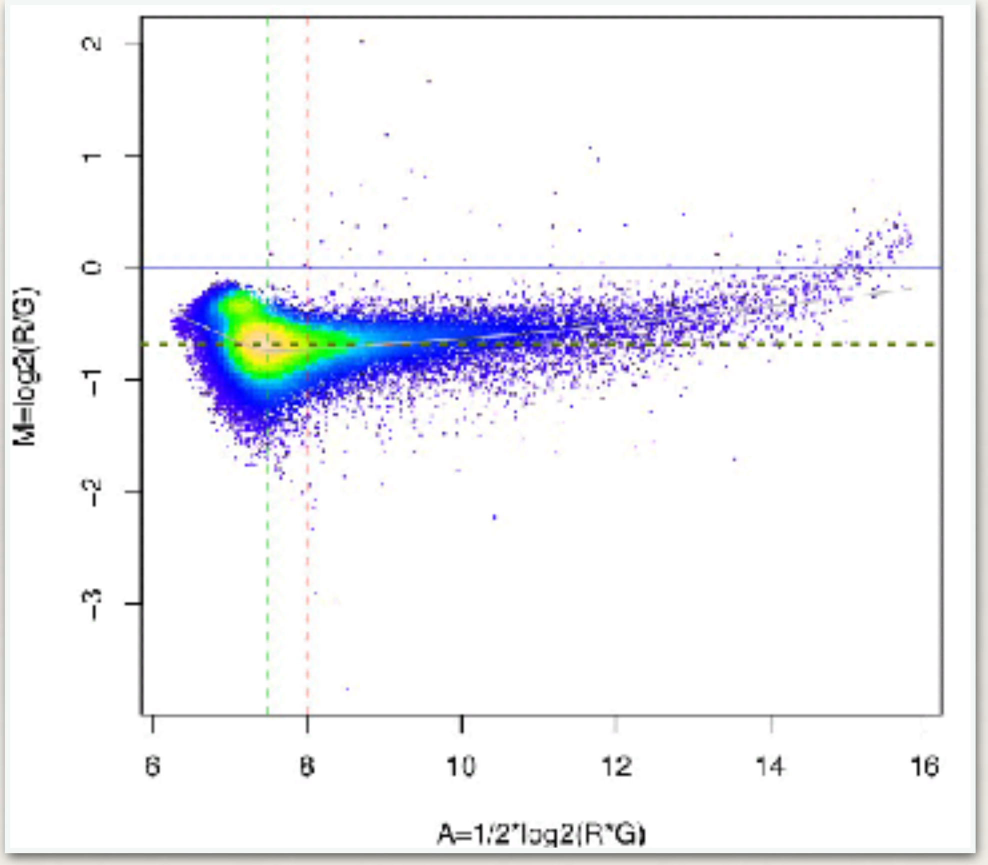


Shift the signal according to intensity

M-part

M-Part

1. Log2 of expression difference-ratio $\boxed{M = \log_2(R/G) = \log_2(R) - \log_2(G)}$



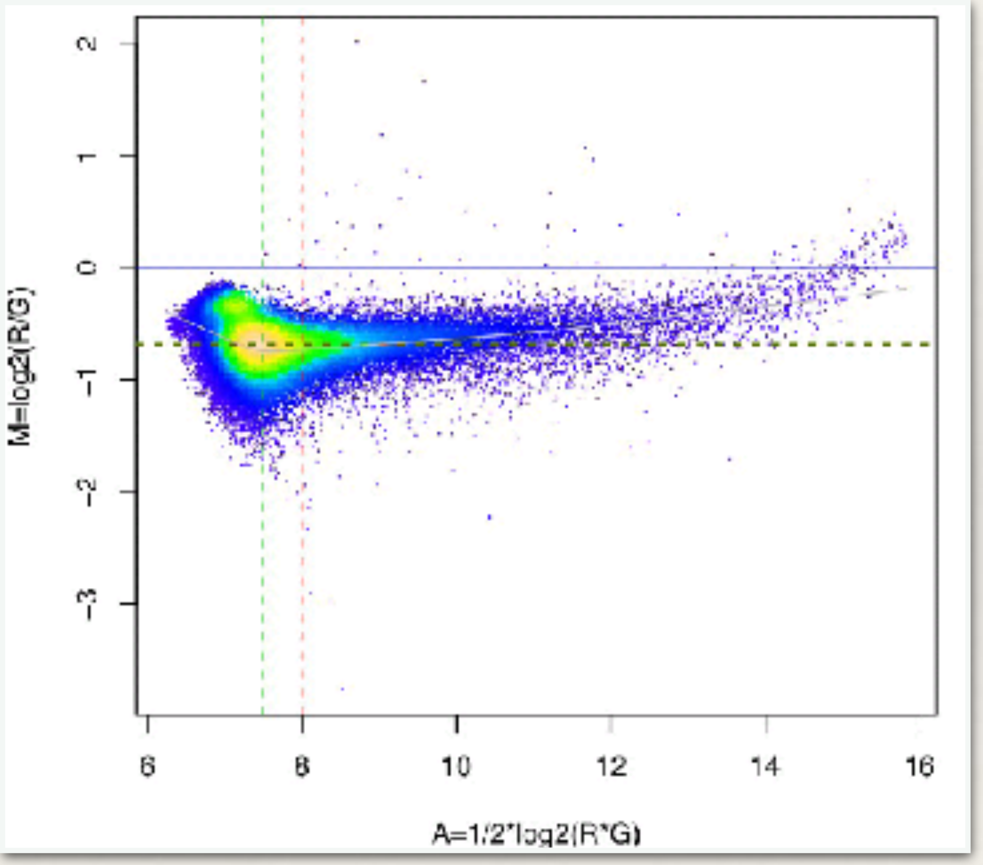
Shift the signal according to intensity

A-part

A-Part

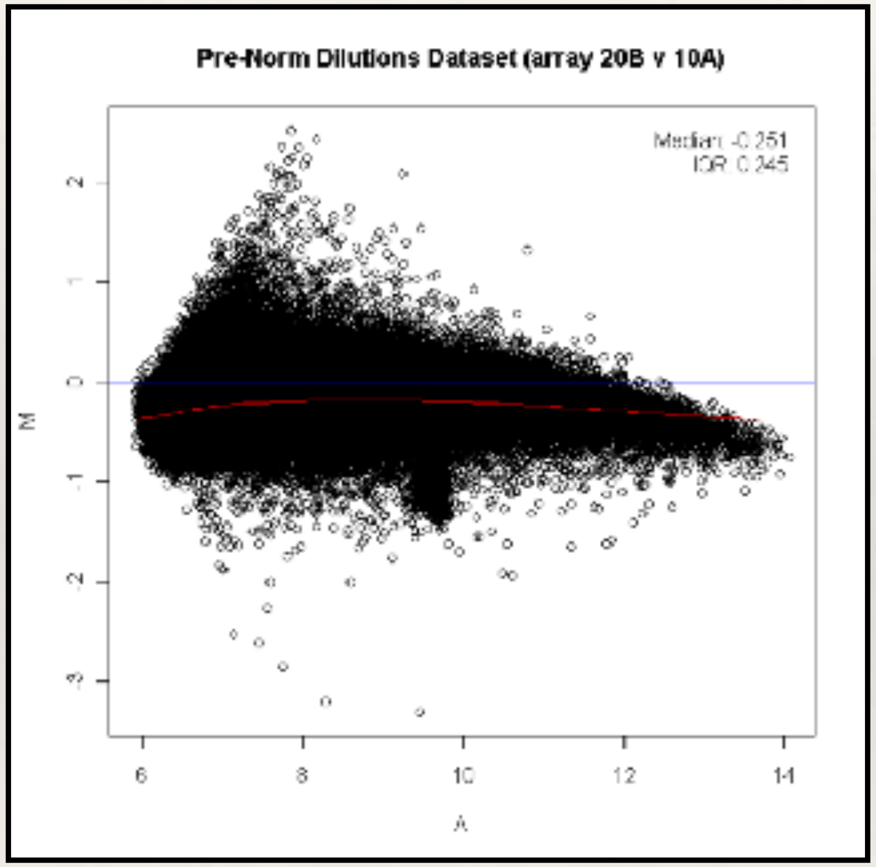
1. Log2 of expression difference-ratio $\boxed{M = \log_2(R/G) = \log_2(R) - \log_2(G)}$

2. Logarithm of intensity mean value $\boxed{A = \frac{1}{2}\log_2(RG) = \frac{1}{2}(\log_2(R) + \log_2(G))}$

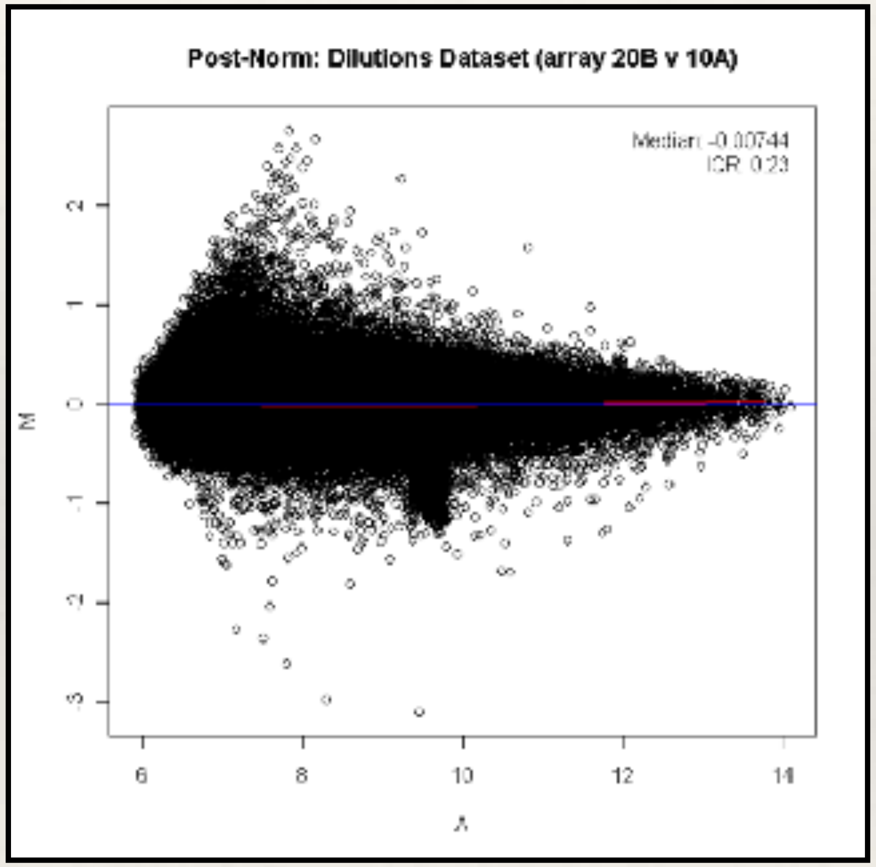


Shift the signal according to intensity

Result MA (LOESS)-correction



Before



After

Z-transformation

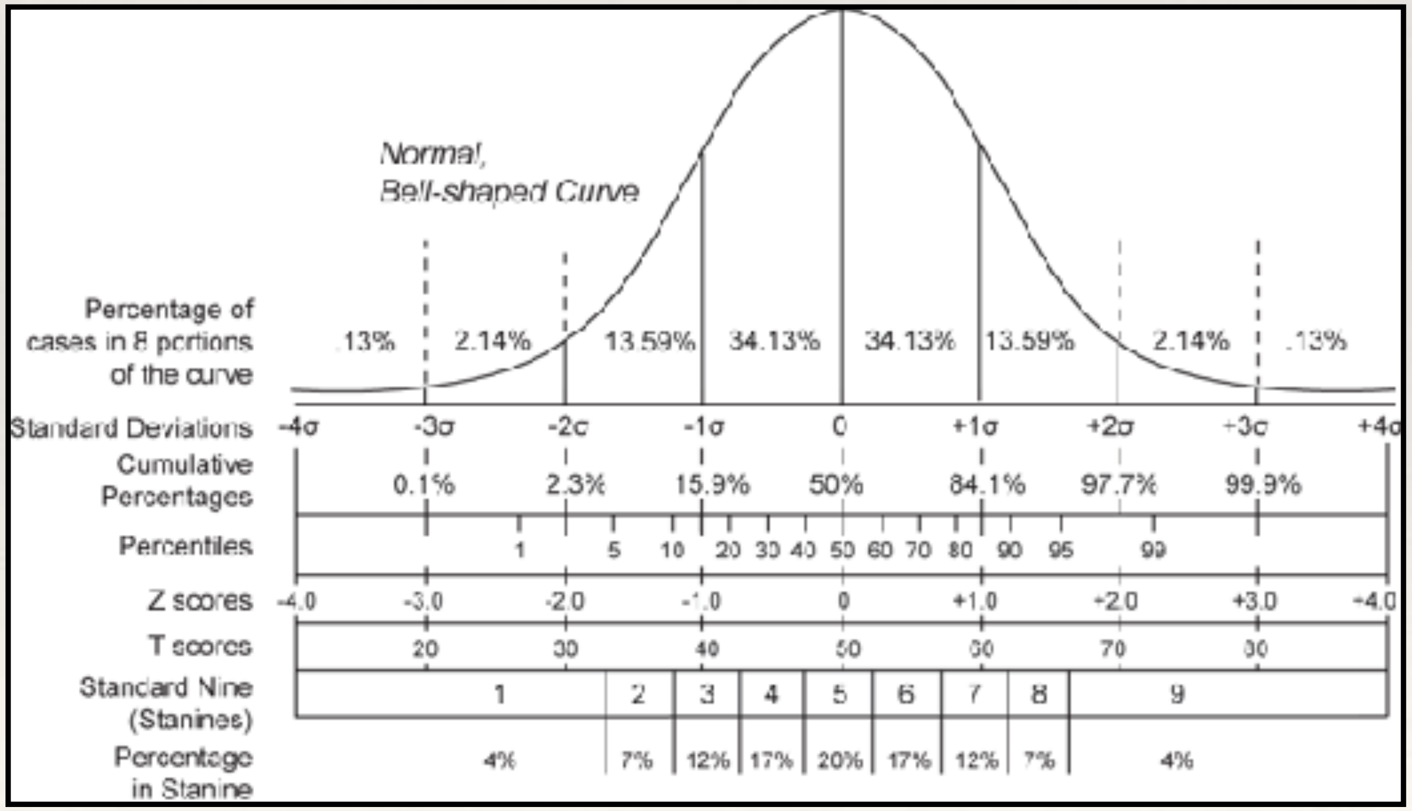
$$z = (x - mean_{est}) / sd_{est}$$

Standardized value

Sample mean

Sample deviance

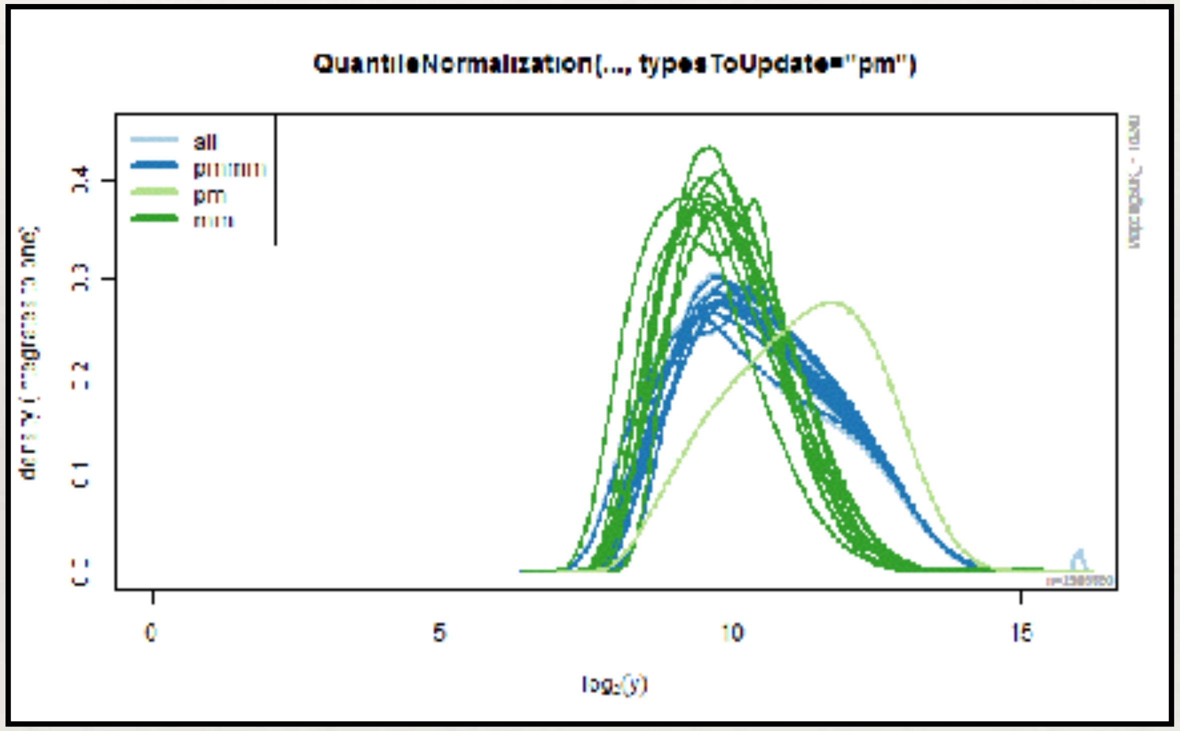
- Normalization requires z-scaling of samples
- Independent of units
- Allows identification of distribution



Outlook: P-values and standardized values

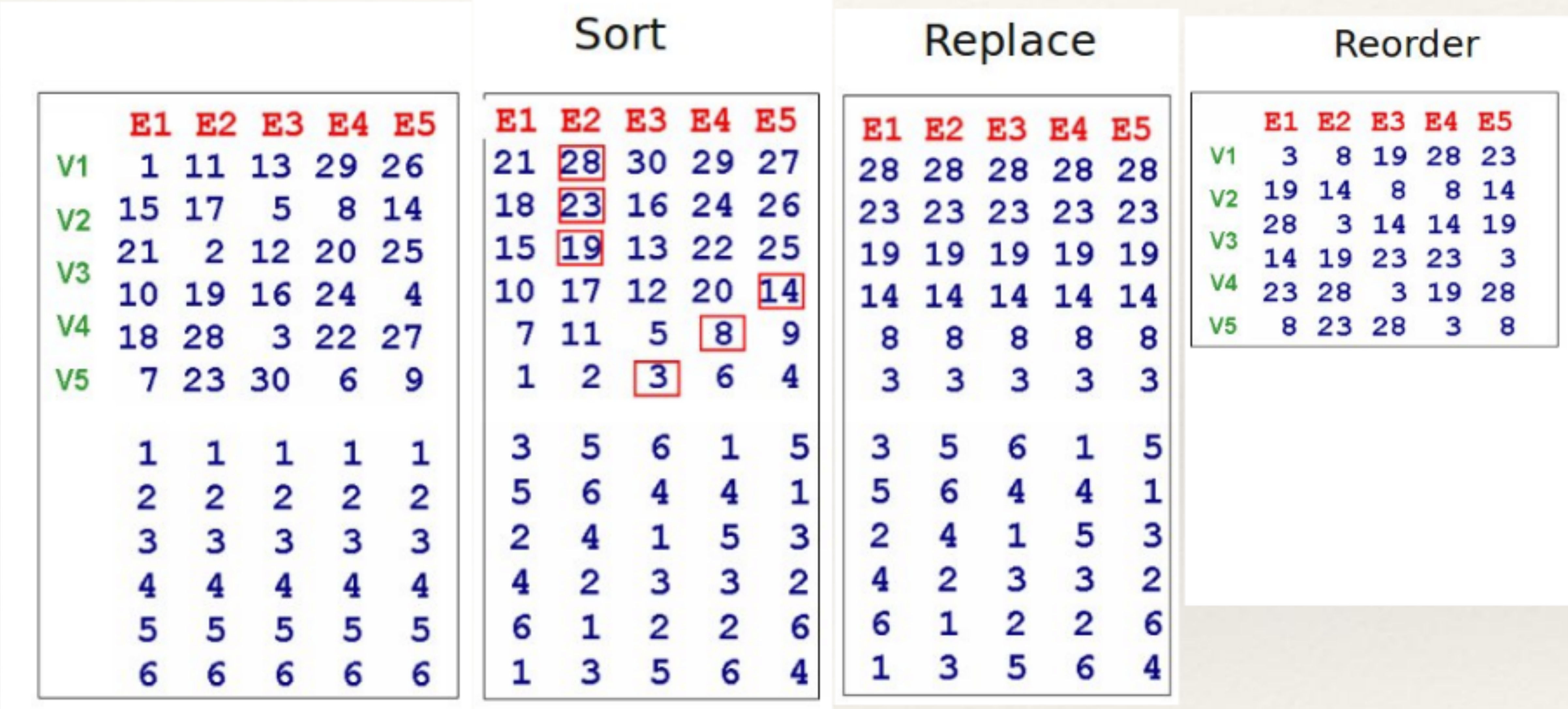
Quantile-Normalization

1. Create genes-samples matrix
2. Z-transformation
3. Sort columns
4. Replace values by row-median
5. Reorder (unsort) values



Raw array data incomparable

Quantile-Normalization



Vaules

Indices

	E1	E2	E3	E4	E5
V1	1	11	13	29	26
V2	15	17	5	8	14
V3	21	2	12	20	25
	10	19	16	24	4
V4	18	28	3	22	27
V5	7	23	30	6	9
	1	1	1	1	1
	2	2	2	2	2
	3	3	3	3	3
	4	4	4	4	4
	5	5	5	5	5
	6	6	6	6	6

Sort

E1	E2	E3	E4	E5
21	$\boxed{28}$	30	29	27
18	$\boxed{23}$	16	24	26
15	$\boxed{19}$	13	22	25
10	17	12	20	$\boxed{14}$
7	11	5	$\boxed{8}$	9
1	2	$\boxed{3}$	6	4
3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

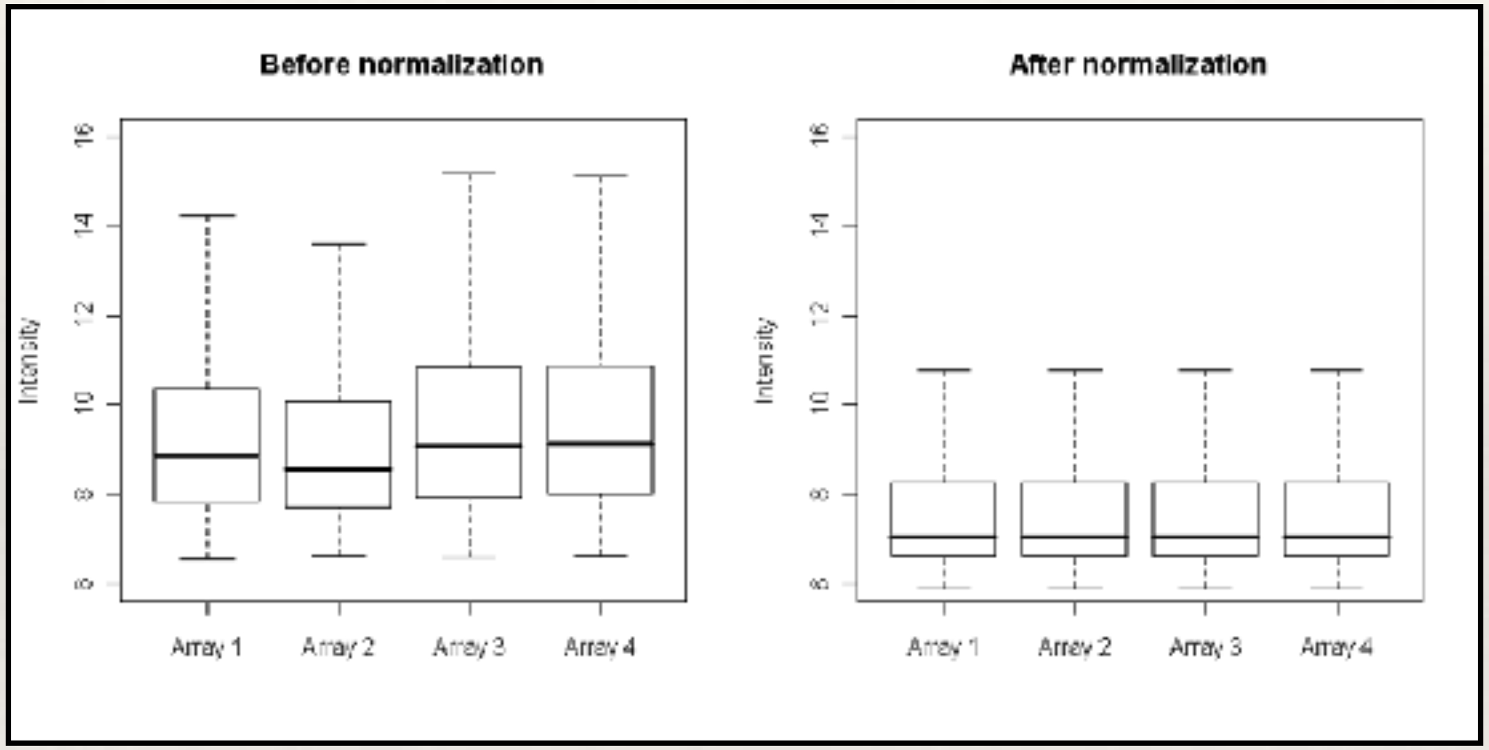
Replace

E1	E2	E3	E4	E5
28	28	28	28	28
23	23	23	23	23
19	19	19	19	19
14	14	14	14	14
8	8	8	8	8
3	3	3	3	3
3	5	6	1	5
5	6	4	4	1
2	4	1	5	3
4	2	3	3	2
6	1	2	2	6
1	3	5	6	4

Reorder

	E1	E2	E3	E4	E5
V1	3	8	19	28	23
V2	19	14	8	8	14
V3	28	3	14	14	19
	14	19	23	23	3
V4	23	28	3	19	28
V5	8	23	28	3	8

Array data-analysis results



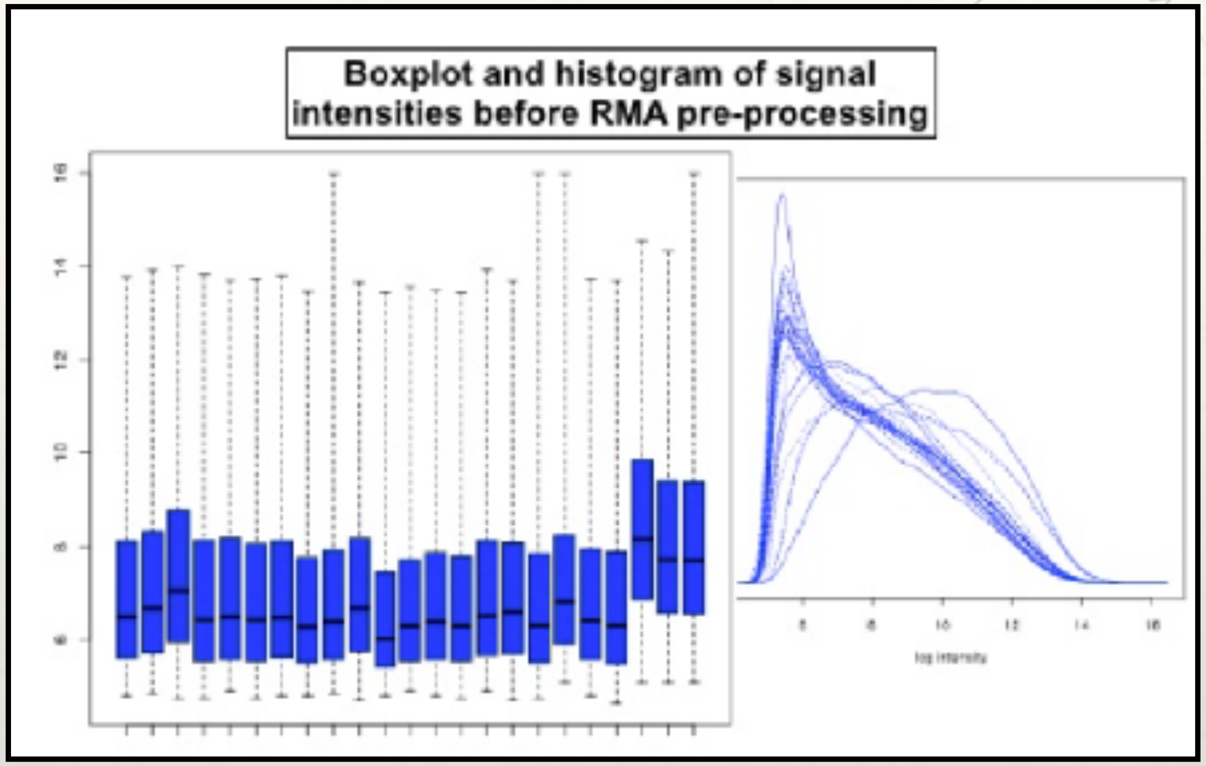
Quantile-normalization

Outlook - RMA

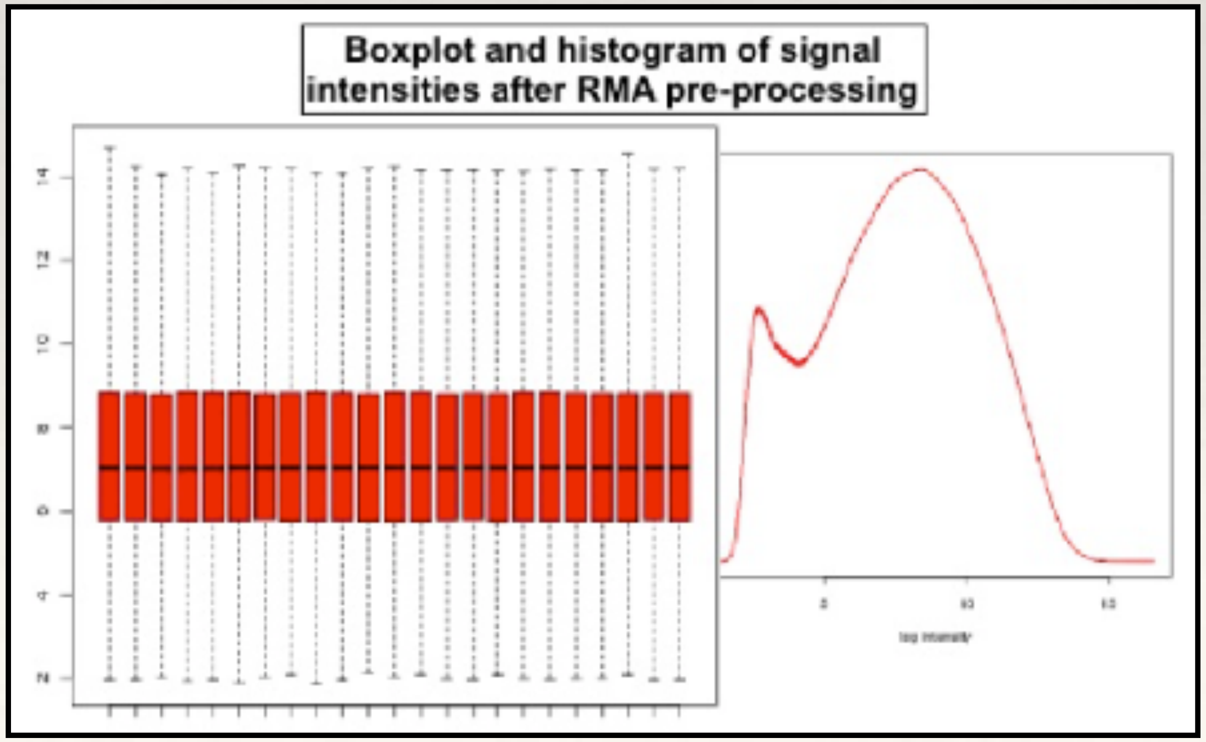
- Excercise next week:

  Robust Multichip Average (RMA) algorithm

1. Z-score transformation
2. Background-correction
3. Quantile-normalization
4. Media-polish



Boxplot and histogram of signal intensities before RMA pre-processing



Boxplot and histogram of signal intensities after RMA pre-processing

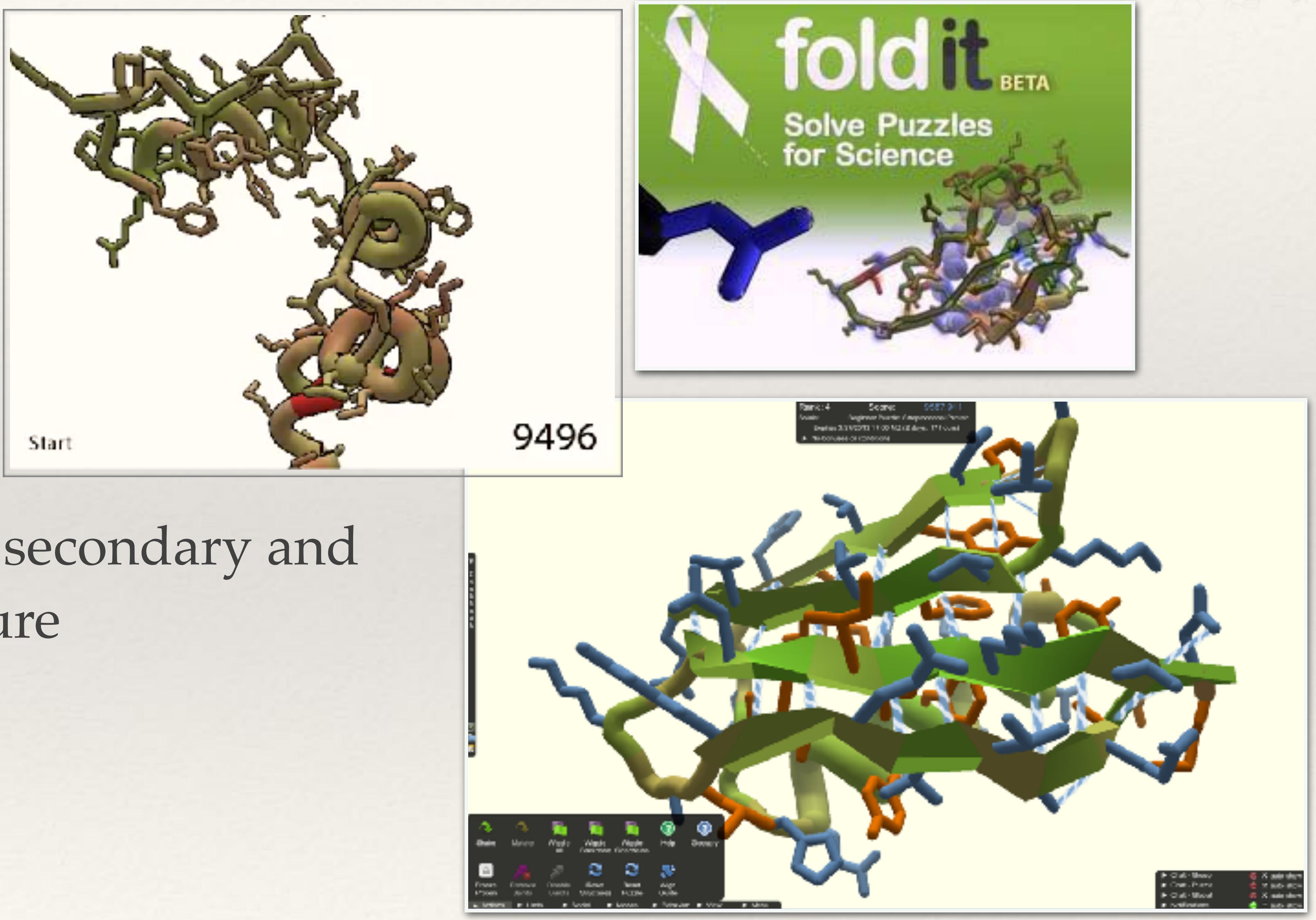
Today's summary

- Biology:
- mRNA expression = gene-activity
- Explain cause of e.g. cancer by
  - Comparing cohorts

- Technology:
- Arrays measure mRNA-expression
- Numerous challenges e.g. biases -> require correction

Try it yourself

- www.fold.it



- Fold proteins' secondary and ternary structure