Introduction to Microarray Analysis

Affymetrix GeneChip technology

Katerina Taškova
Computational Biology and Data Mining Group
Faculty of Biology

11 March 2016

Goal of the talk

- Review Affymetrix GeneChip technology & terminology
- Microarray data analysis
- Test for differential expression

Method

Lecture

- Slides

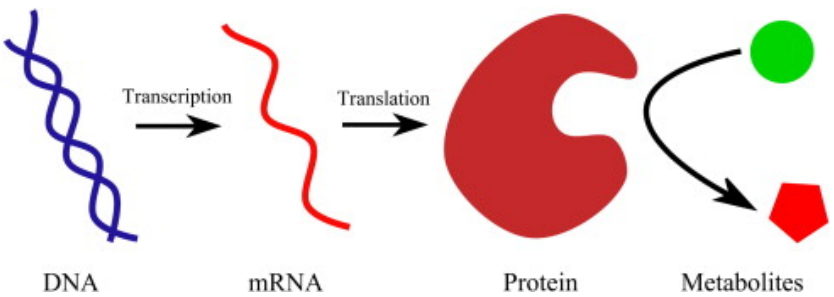
Tutorial

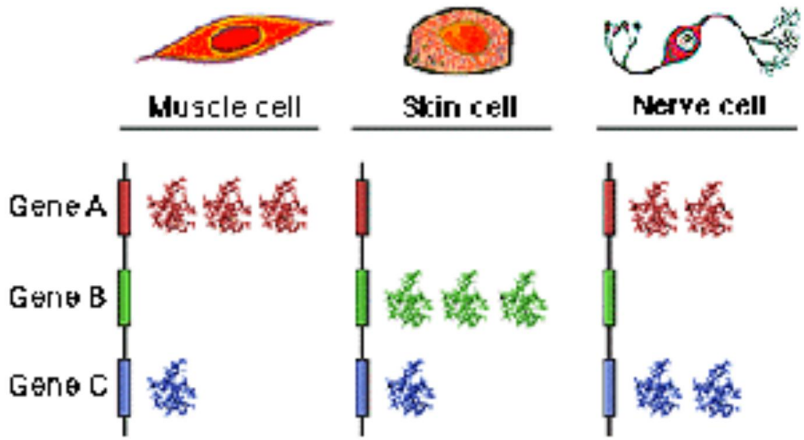
- Gene expression analysis in R/Bioconductor
- https://cbdm.uni-mainz.de/mb16/

Feel free to ask questions at any point of the lecture/tutorial

Gene expression

Genes are 'decoded' to perform different functions,
e.g. synthesis of proteins





(left) Karakach et al. (2010). Chemometrics and Intelligent Laboratory Systems
(right) http://www.ncbi.nlm.nih.gov/Class/MLACourse/Original8Hour/Genetics/

The set of expressed genes determines the phenotype of a particular cell

If we are able to find out **which** and **how much mRNA** is in the cell we should be able to find out **which genes** and **with which intensity** they are being expressed $\Rightarrow$ **microarrays**

Microarrays: multiplex lab-on-chip

2D grid on a solid substrate (plastic/glass/silicon) that profiles large amounts of biological material using high-throughput screening, multiplexed & parallel processing & detection methods. (Wikipedia)

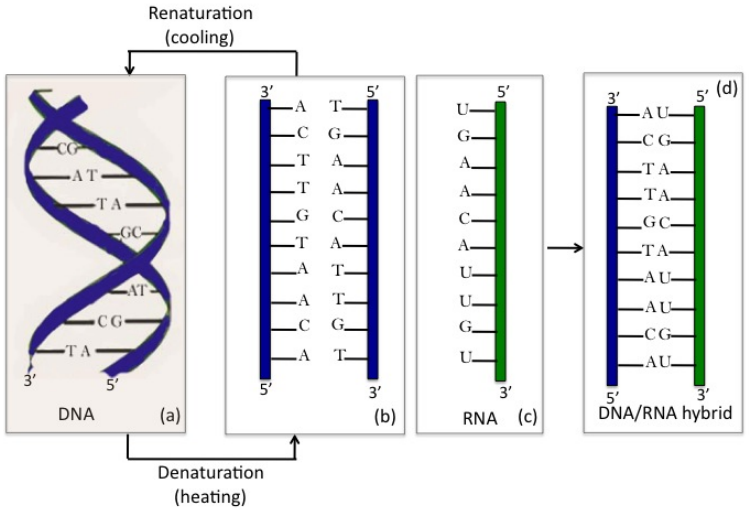
Types: **DNA**, protein, antibody, tissue, cellular

Purpose: **Gene expression analysis**, mutation analysis (SNP), comparative genomic hybridization

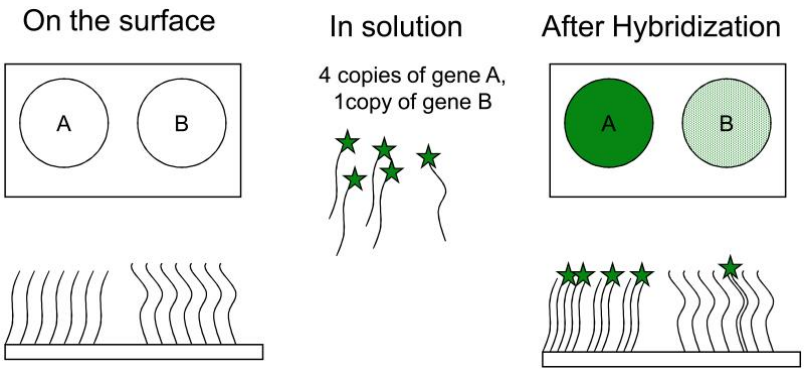
Application: Disease characterization, diagnostics development, cellular physiology, stress responses, drug discovery, toxicological research

Principle

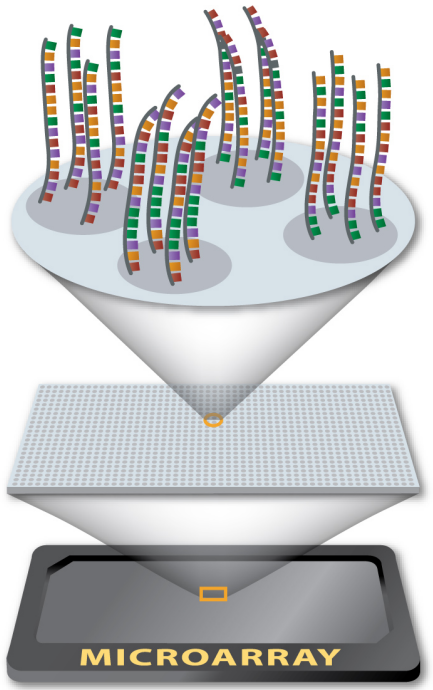
nucleic acid hybridization for a global investigation of cellualr activity



Fusco and Quero (2012). Structure and Function of Food Engineering



Bumgarner (2013). Current Protocols in Molecular Biology

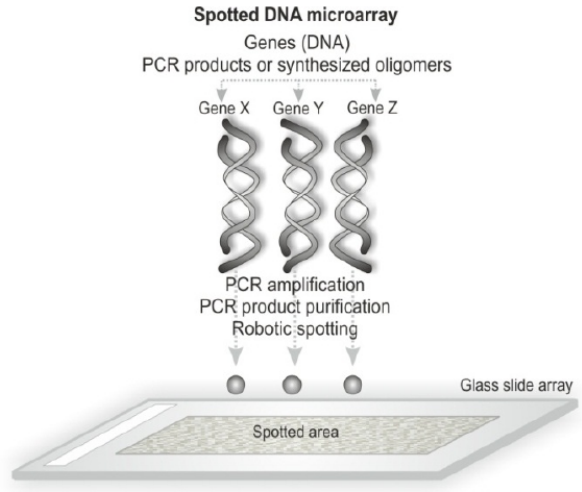


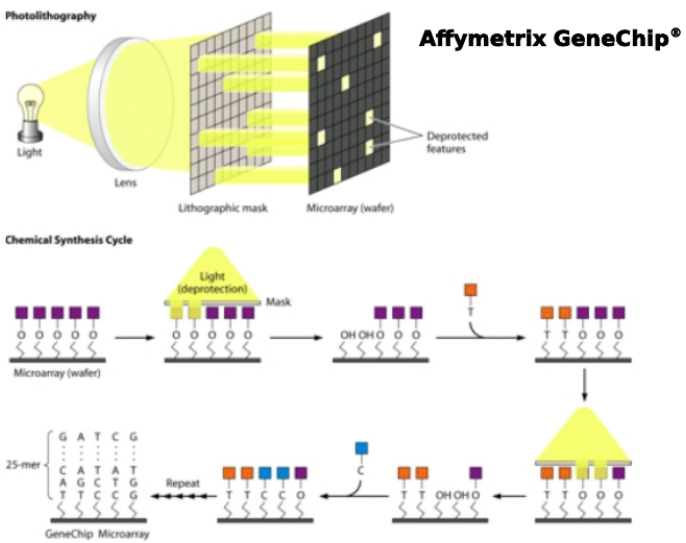
http://learn.genetics.utah.edu

Assumption: number of mRNA molecules ≈ level of gene expression

Technology

spotted vs. in situ synthesized arrays

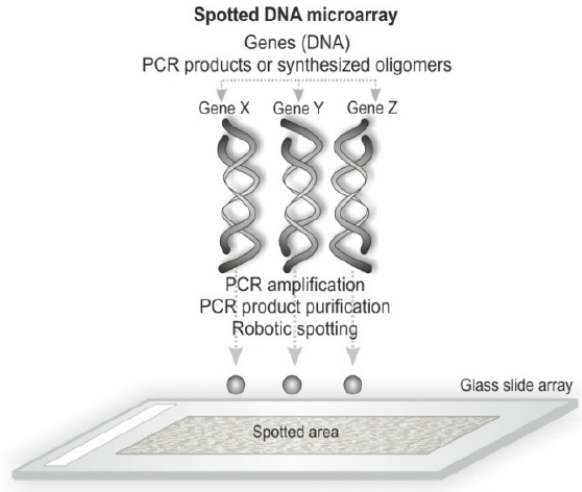


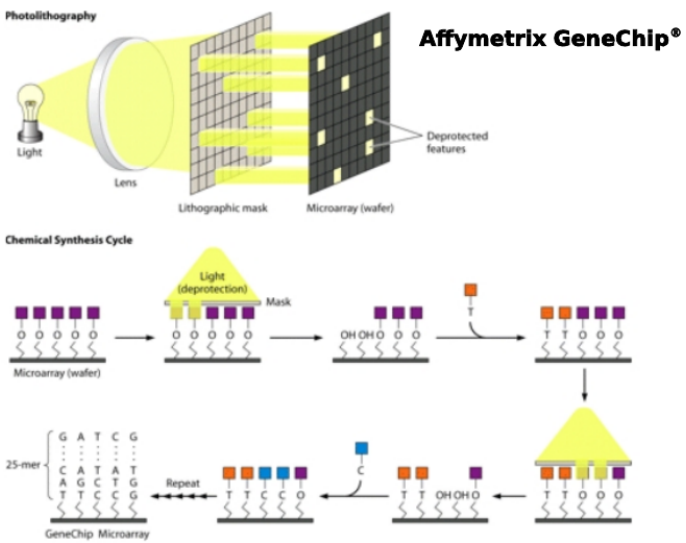


Saei and Omidi (2011). BioImpacts
Lipshutz et al. (1999). Nature Genetics

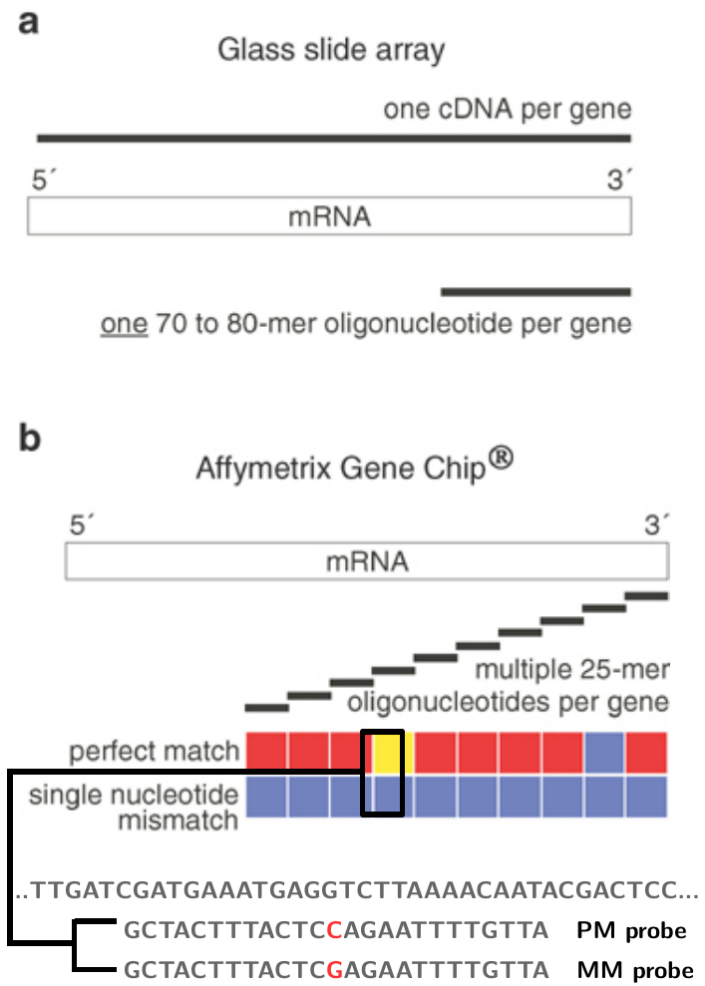
Technology

spotted vs. in situ synthesized arrays





Saei and Omidi (2011). BioImpacts
Lipshutz et al. (1999). Nature Genetics



Staal et al.(2003). Leukemia

Affymetrix GeneChip probe sets

Intended to measure expression for a specific mRNA

Complementary to a target sequence (from one or more mRNA sequences)

11 - 20 25-mer probe pairs (PM and MM) selected from the target sequence

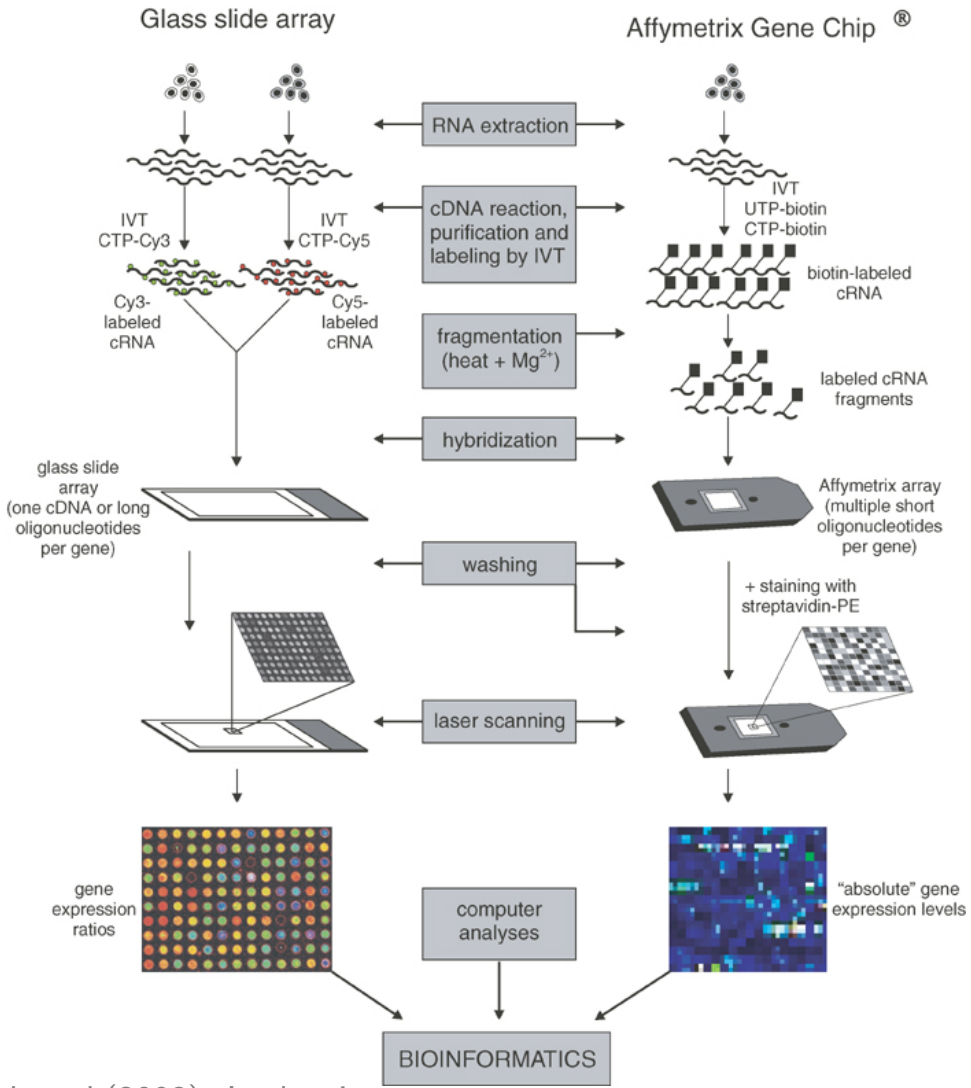
10000 - 50000 probe sets per chip, with several probe sets per gene

Probe set ID	Description
AFFX...	control probe sets, not generally used for analysis
..._at	hybridizes to unique antisense transcript for this chip
..._s_at	all probes cross hybridize to a specified set of sequences
..._a_at	all probes cross hybridize to a specified gene family
..._x_at	at least some probes cross-hybridize with other target sequences for this chip

Chip Description File (CDF) with probe locations and probe set groupings on the chip

Chip types: HG-U133 Plus 2.0, HG-95Av2, MOE 430 2.0, RAE 230A ...

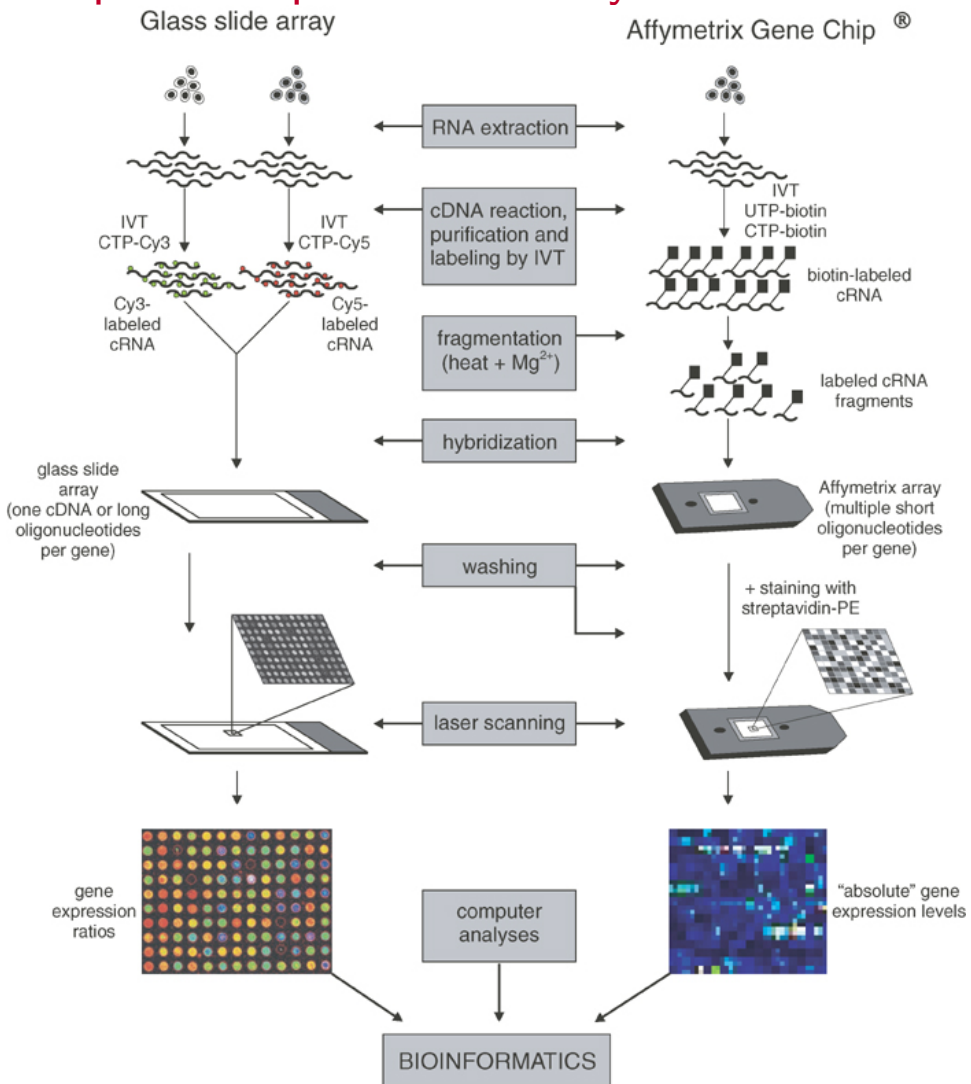
Microarray experiment



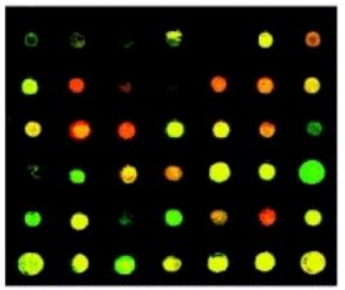
Staal et al.(2003). Leukemia

Microarray experiment

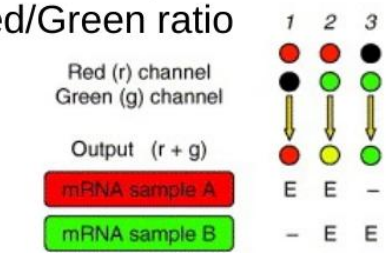
gene expression quantification by means of fluorescence intensity

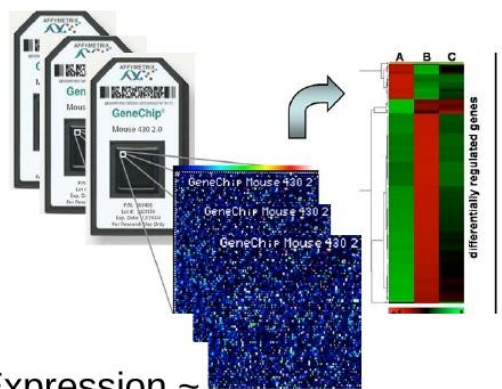


Staal et al.(2003). Leukemia



Relative expression ~
Red/Green ratio



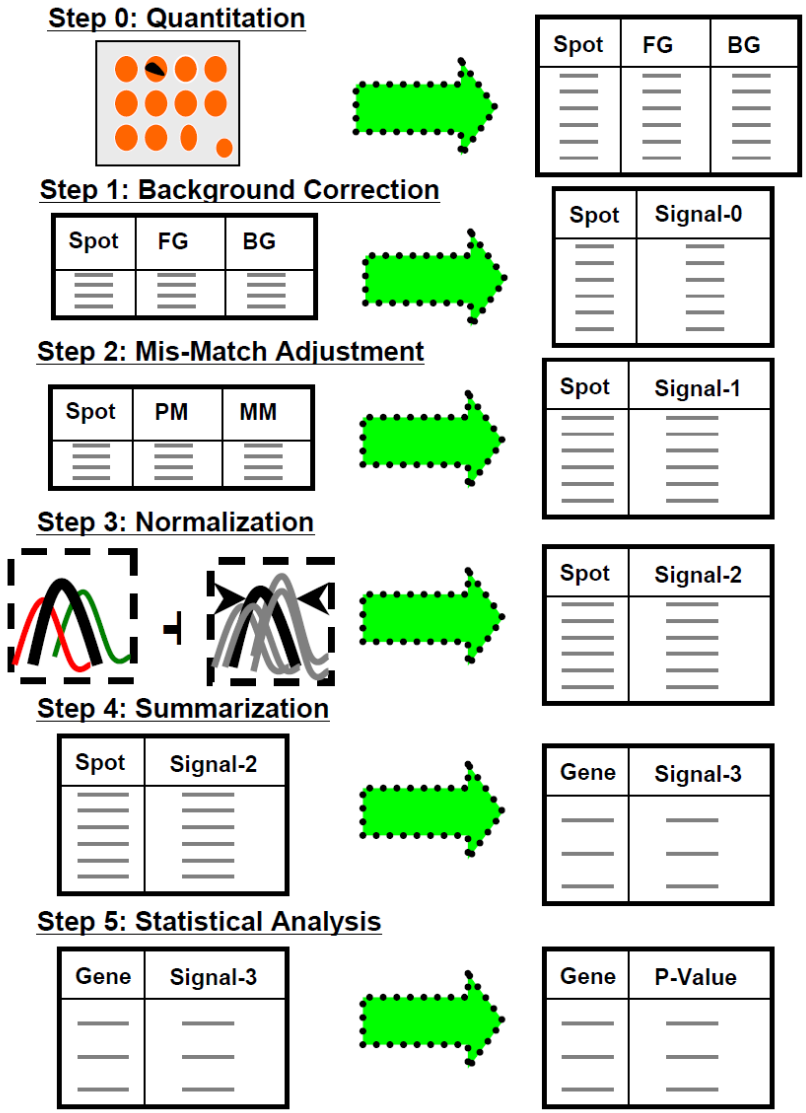


Expression ~
f(probe intensities)

Ranz et al (2006). Trends in Ecology & Evolution
http://www.utoledo.edu

Microarray data analysis

The pipleine



http://www.hypothesisjournal.com/?p=789

TIFF image $\longrightarrow$ signal estimates
Detect spots (foreground signal FG) from surrounding (background signal BG)

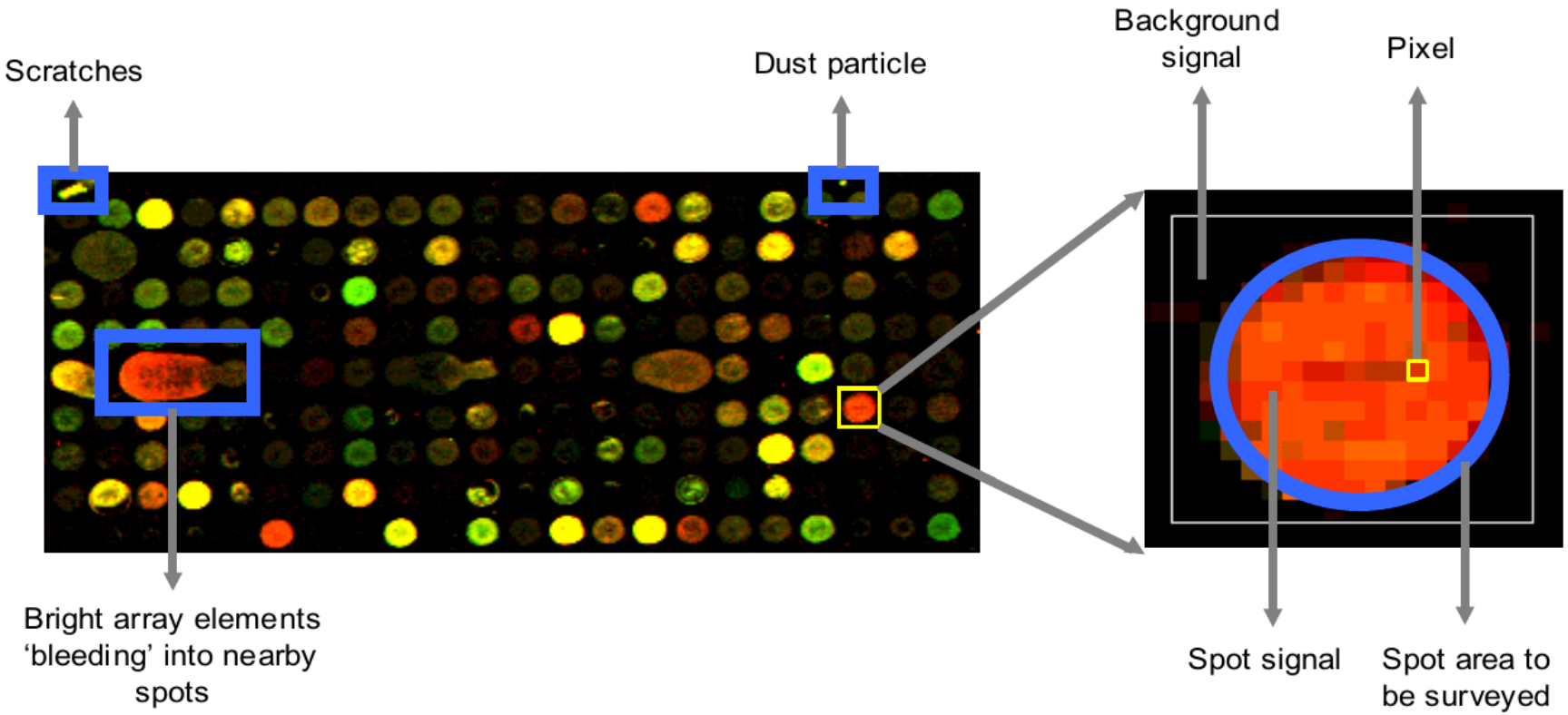
Correct for non-specific hybridization
e.g. FG-BG
e.g. PM-MM

Correct for systematic bias to make the different arrays comparable

Per gene expression estimates
Collapse the signal form the replicated spots

Select significant genes

Step 0-1 with spotted arrays



Identify **spots** (foreground signal), distinguish **spurious** (scratches, dust ...) and **background** signal

Estimate **spot intensity** as median or total intensity across all pixels

http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/

Step 1-4 with Affymetrix GeneChip

Robust Multi-array Average (RMA) method

- convolution background correction
- quantile normalization
- median-polish-based multi-array summarization
- log2-transformation of expression values

$$\begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots & y_{1m} \\ y_{21} & y_{22} & y_{23} & \cdots & y_{2m} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ y_{n1} & y_{n2} & y_{n3} & \cdots & y_{nm} \end{bmatrix} \xrightarrow{RMA} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_N \end{bmatrix}$$

From $n * m$ probe intensities to expression values for G probe sets

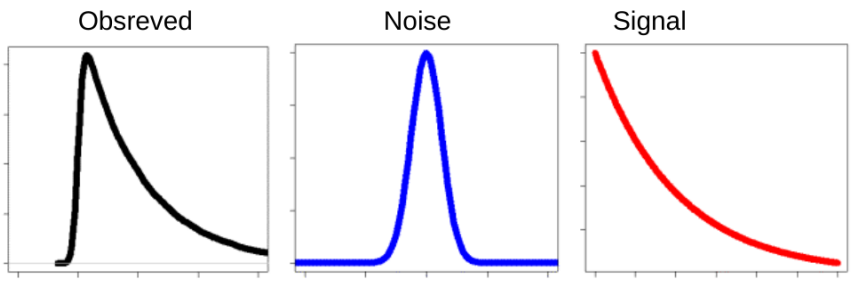
Irizarry et al. (2003) Biostatistics

Convolution background correction

Why To correct for cross-hybridization and optical detection noise

How **Perfect match** (PM) probe-level correction model
Omits Mismatch (MM) probe intensities (PM-MM < 0)

$$\underbrace{PM}_{\text{observed probe intensity}} = \underbrace{bg}_{\text{Gaussian noise component}} + \underbrace{s_{true}}_{\text{exponential signal component}}$$



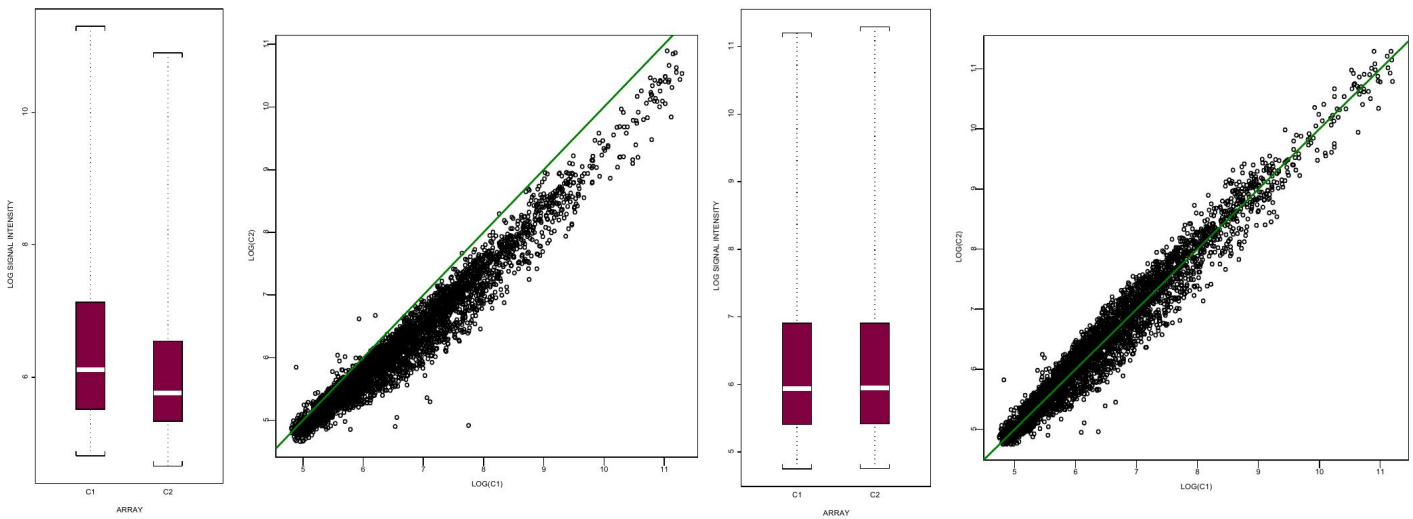
Assumes one **global background** for each array
(we avoid noise from the MM signal but we lose accuracy)

Signal estimate: **conditional expectation** $\mathrm{E}(s_{true}|PM, bg)$

Quantile normalization

Why Correct for biases form non-biological sources (RNA quantity, labeling efficiency, scanner setup)

Principle Most of the genes are either not or equally expressed in any condition, **while only a small number of genes show expression changes between conditions**

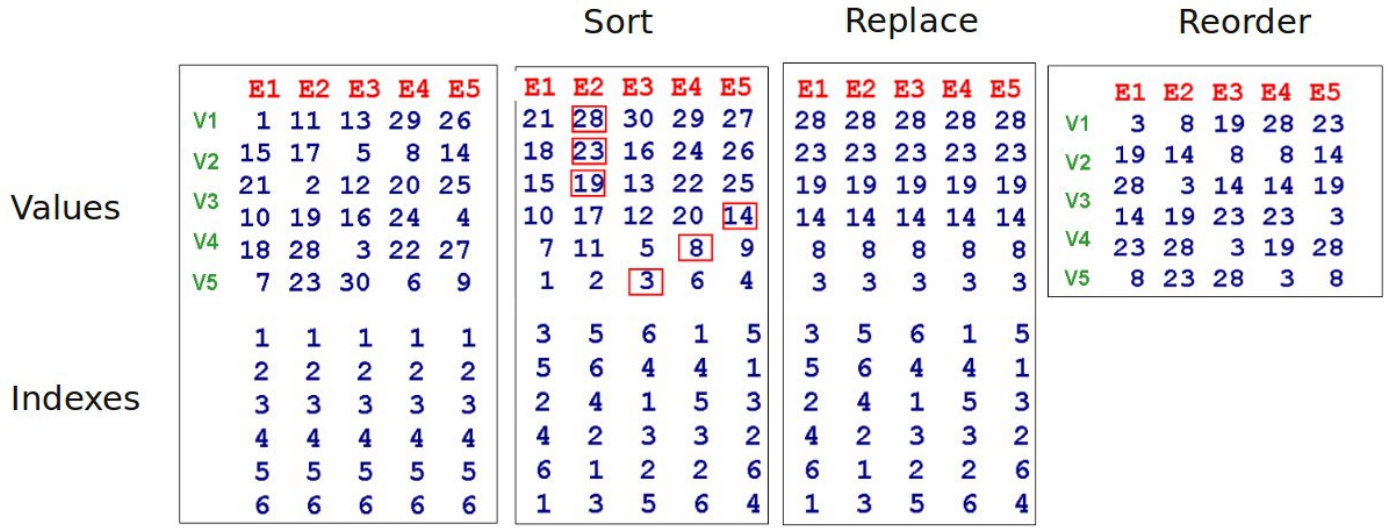


http://www.rci.rutgers.edu/ cabrera/DNAMR/

Quantile normalization

Why Correct for biases form non-biological sources (RNA quantity, labeling efficiency, scanner setup)

How Apply nonparametric nonlinear transformation of the background-corrected signal to enforce same empirical distribution of the intensities across arrays



Note: Data are first sorted by columns, then the row-wise medians of are calculated (red squares) and used to replace the row values, finally the elements of each column is reordered to theirs original (before sorting) position. Image source: http://pedagogix-tagc.univ-mrs.fr

Median-polish multi-array summarization

Why Estimate single expression values per probe set

Principle **Gene-wise linear additive probe model**

$$\underbrace{log_2(Y_{gij})}_{\text{background corrected \& normalized PM intensity}} = \underbrace{\theta_{gi}}_{\text{log-scale gene expression}} + \underbrace{\alpha_{gj}}_{\text{probe affinity effect}} + \underbrace{\varepsilon_{gij}}_{\text{measurement error}}$$

probe set/gene $g \in [1, N]$, e.g. $G = 12000$
probe pair $i \in [1, I]$ e.g. $I = 16$
array $j \in [1, J]$, e.g. $J = 8$

How Median polish-based robust estimation of model parameters

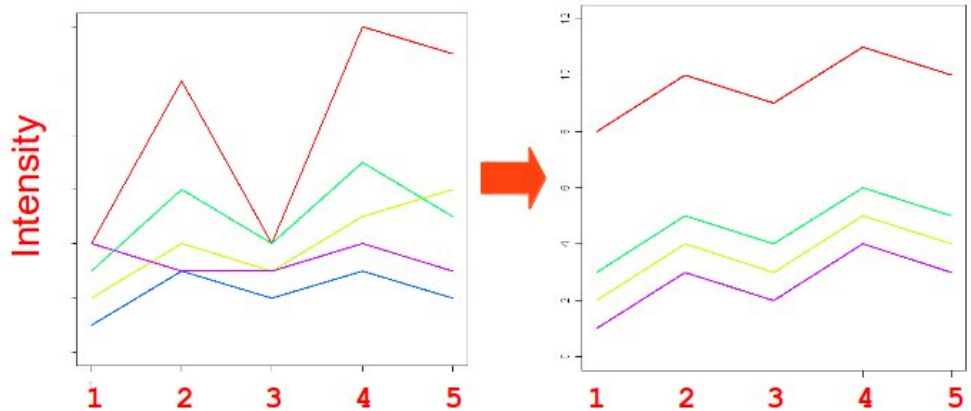
2-way data $\rightarrow$ grand effect + row effect + column effect

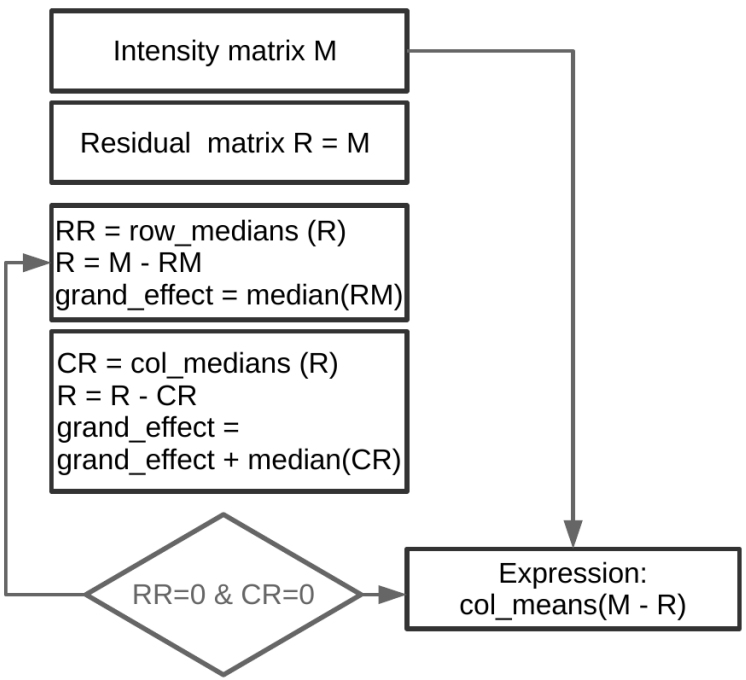
extract the effects by **medians** (robust to outliers)

Median-polish multi-array summarization

GeneChips

Probes	1	2	3	4	5	
1	4	10	4	12	11	10
2	2	4	3	5	6	4
3	3	6	4	7	5	5
4	1	3	2	3	2	2
5	4	3	3	4	3	3





Output

GeneChips

Probes	1	2	3	4	5
1	8	10	9	11	10
2	2	4	3	5	4
3	3	5	4	6	5
4	1	3	2	4	3
5	1	3	2	4	3

http://pedagogix-tagc.univ-mrs.fr

Logarithmic transformation of expression values

Advantage

Convenient for interpretation of expression ratios

1. up-regulation and down-regulation are comparable

$$fold\_change = \frac{gene\_A}{gene\_B}$$

$$fold\_change > 1, up - regulation$$

$$fold\_change < 1, down - regulation$$

$$\frac{16}{8} = 2 \xrightarrow{log_2} log_2 16 - log_2 8 = 1$$

$$\frac{8}{16} = 0.5 \xrightarrow{log_2} log_2 8 - log_2 16 = -1$$

2. mapping space is continuous

fold change: [0,1] $\xrightarrow{log_2}$ $[-\infty, +\infty]$

Disadvantage

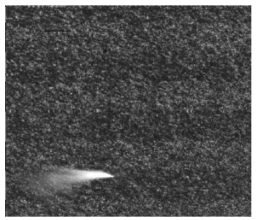
Removes absolute expression levels

$$fold\_change = 2 = \frac{160}{80} = \frac{16}{8}$$

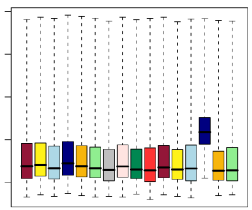
Quality assessement

Artifacts with image & data analysis, problems with experimental design ...

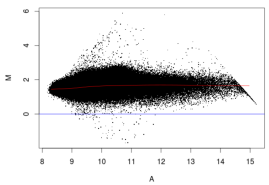
Array surface images



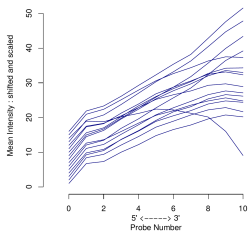
Intensity/expression boxplots



MA plots



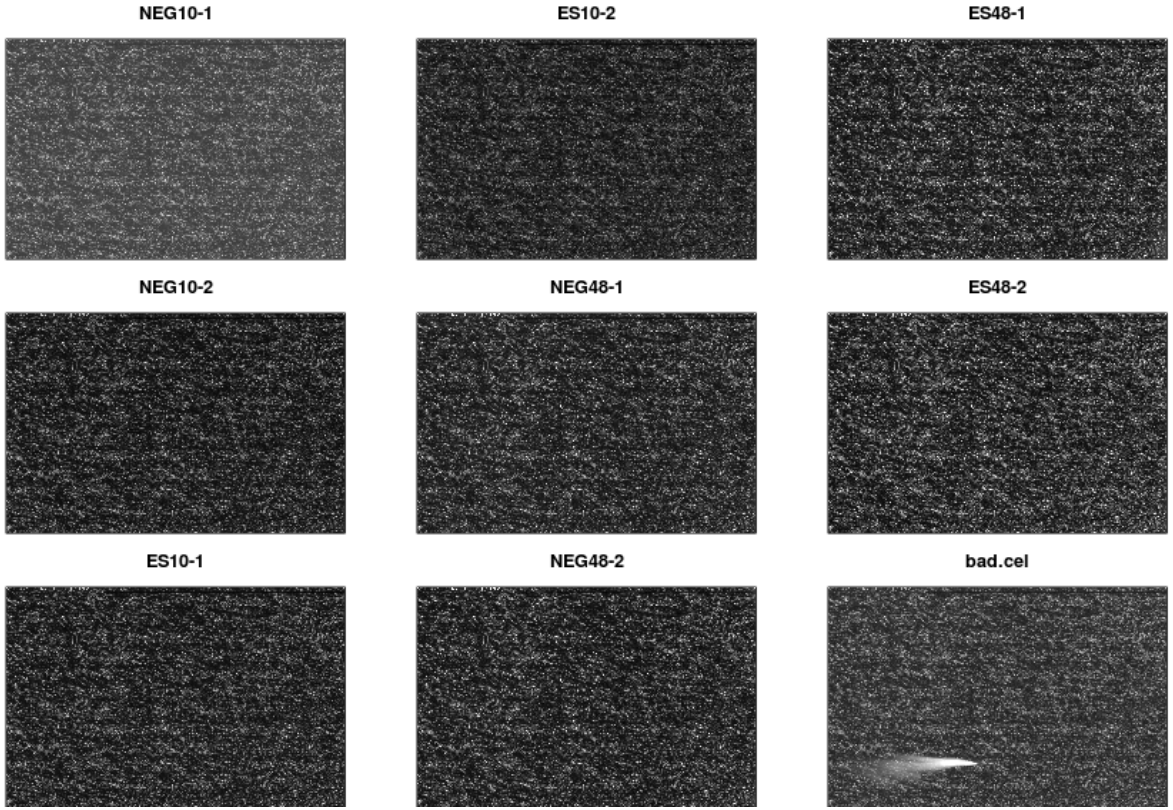
RNA degradation plots



Gentleman et al.(2005) Bioinformatics and Computational Biology Solutions using R and Bioconductor. Springer NY
Heber and Sick (2006) Journal of Integrative Biology

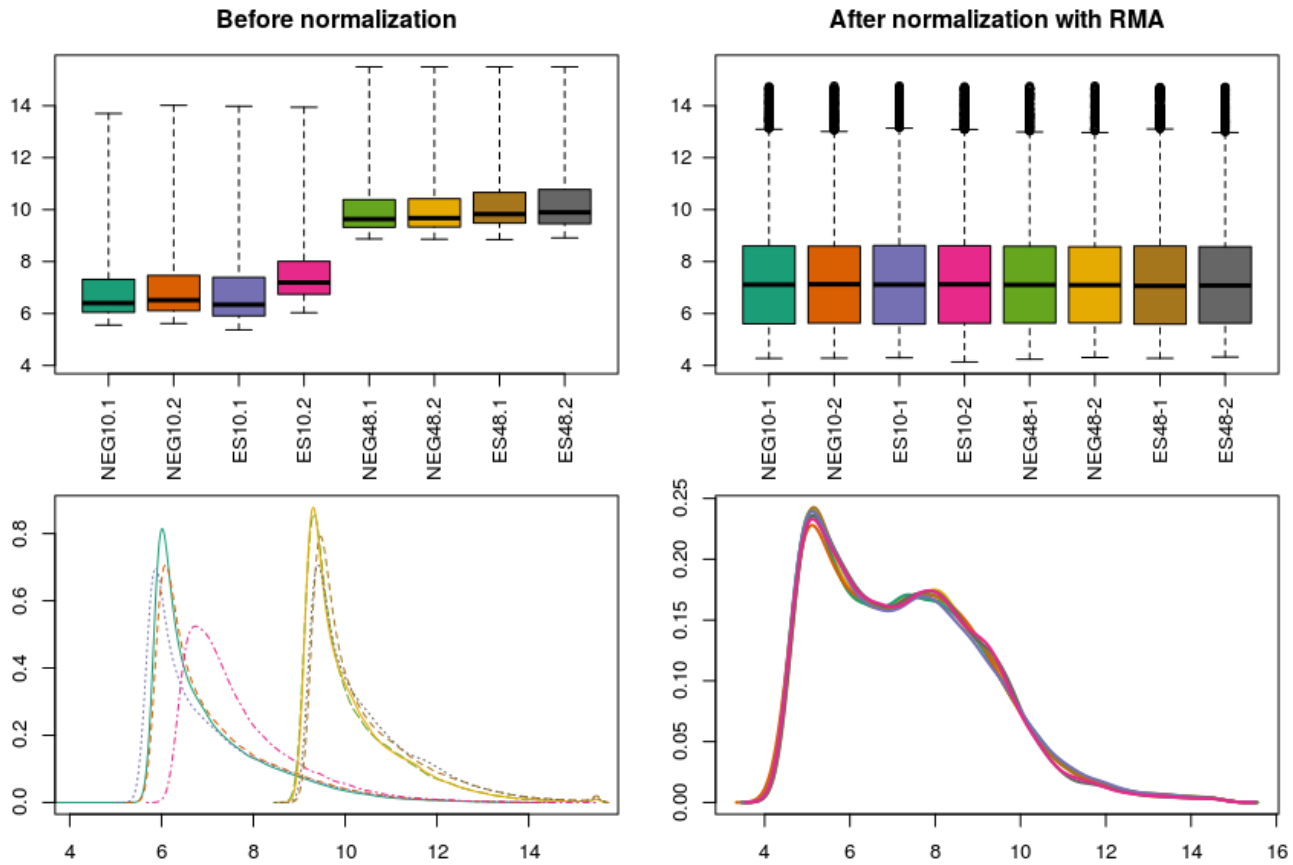
Array surface images

To inspect the spatial distribution of the raw intensities on a chip for spatial artifacts



Intensity/expression boxplots

To summarize probe intensity and gene expression distributions
Pinpoint arrays that show different spread and location

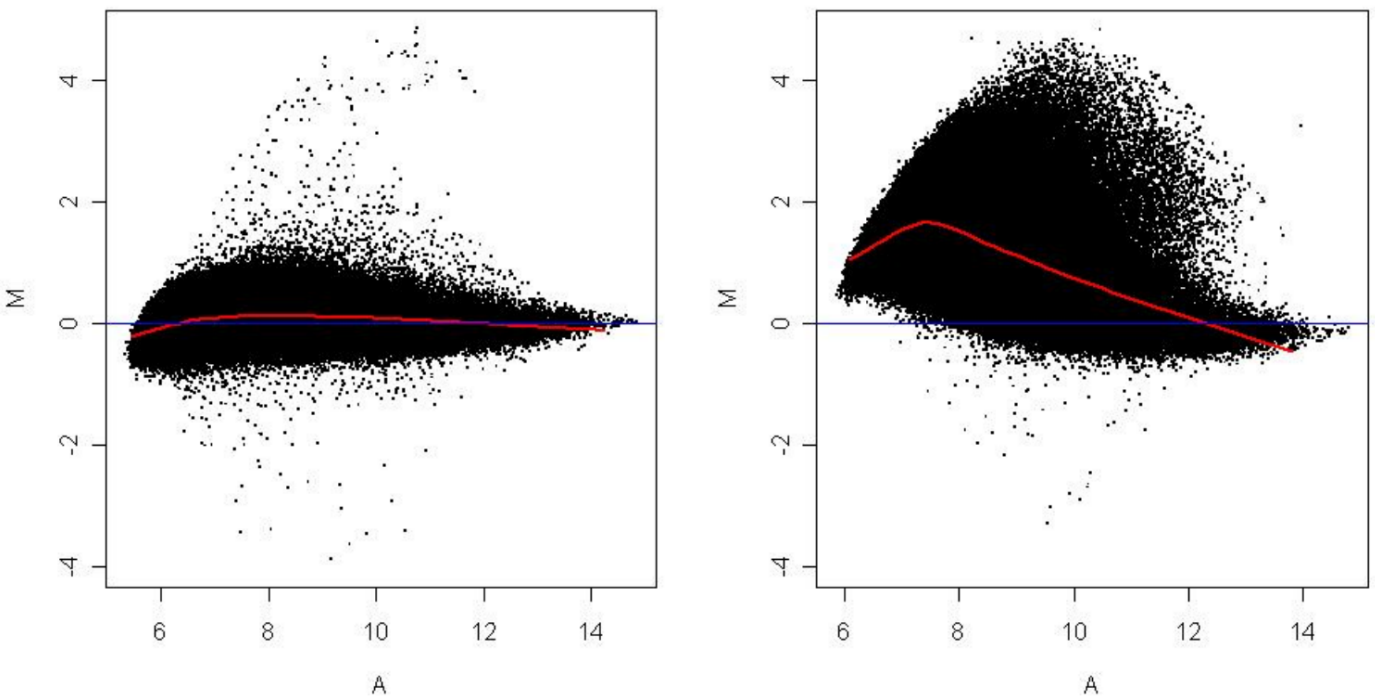


MA (scatter) plots

$M_{gij} = log_2(PM_{gij}) - log_2(PM_{gi*})$

Log fold intensity change between array i and a reference array *, with intensities equal to probe-wise medians over all arrays

Mean log intensity $A_{gij} = 0.5 \cdot (log_2(PM_{gij}) + log_2(PM_{gi*}))$



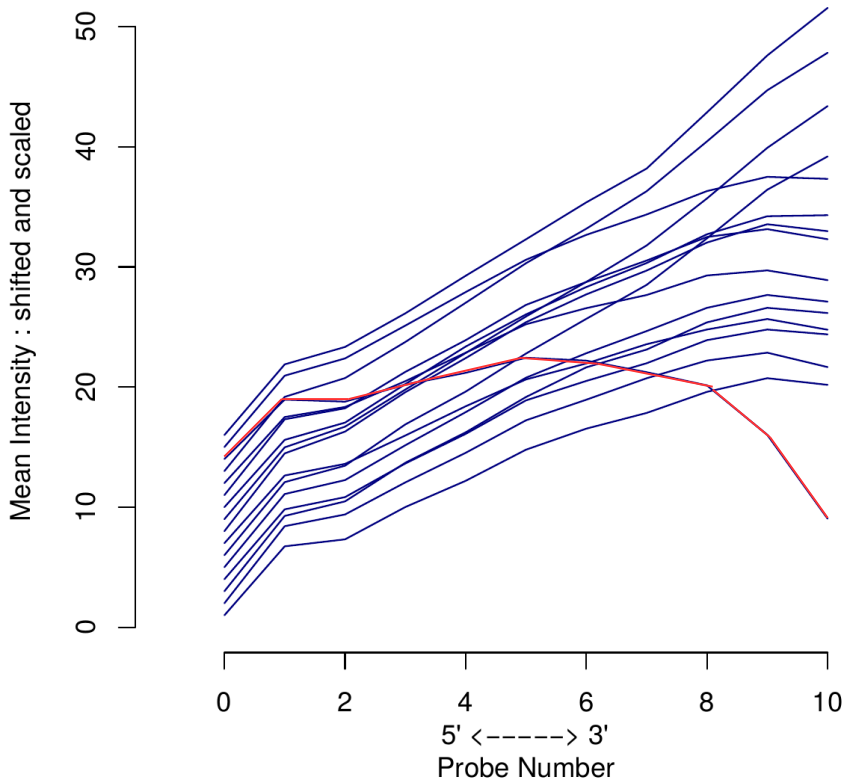
Note: One can use MA plot to also plot the expression estimates after RMA processing and check the effect of the normalization.

RNA degradation plots

RNA degradation starts at the 5‘ sequence end, therefore intensities of probes at the 3‘ probe set end are higher than those of the 5‘ end probes

Plot the mean intensity for each probe position within probe sets

Look for **high slope** and/or **disagreement between arrays**



Statistical analysis

Objective	Description
Class comparison	Which transcripts (genes) are **differentially expressed** between two conditions?
Class discovery	Are there meaningful patterns in the data (e.g. groups)?
Class prediction	Do RNA transcripts predict predefined groups, such as disease subtypes?
Pathway analysis	Find genes whose co-regulation reflects their participation in a common biochemical process?

Diffrential expression

Identify those genes that show **significantly up-regulated** or **down-regulated expression levels** across two or more **predefined classes**

- diseased vs. normal cells
- between different cell types
- between different tissues
- before and after drug treatment
- between patients with different diets
- ...

What are the criteria for statistical significance?

Gene selection by mean log fold change

	Treatment			Control			$\bar{X}_T$	$\bar{X}_C$	M
	T1	T2	T3	C1	C2	C3			
Gene1	17	16	15	8	10	12	16	10	6
Gene2	17	18	16	16	16	16	17	16	1
Gene3	8	20	8	4	3	5	12	4	8
...	...	...	...	...	...	...	...	...	...
Gene10000	18	17	19	15	17	16	18	16	2

Note: The values in the table are expression estimates on a log2 scale (as obtained by RMA). Otherwise, you will have to log2-transform the data before you calculate $\bar{X}_T$, $\bar{X}_C$, M and A

$$\bar{X}_T = \frac{T1 + T2 + T3}{3} \quad M = \bar{X}_T - \bar{X}_C$$

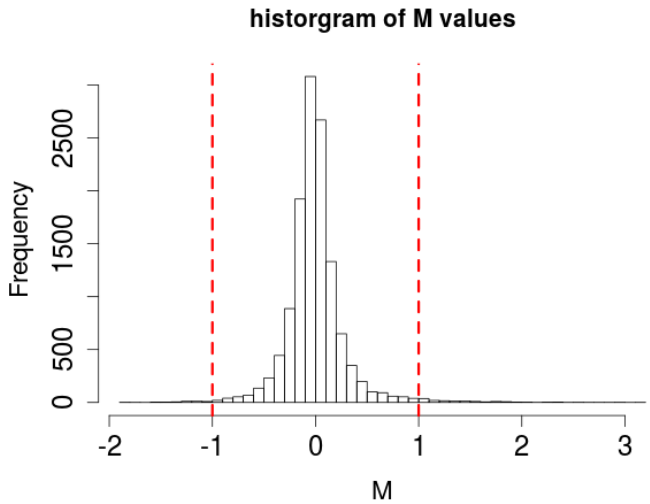
$$\bar{X}_C = \frac{C1 + C2 + C3}{3} \quad A = \frac{\bar{X}_T + \bar{X}_C}{2}$$

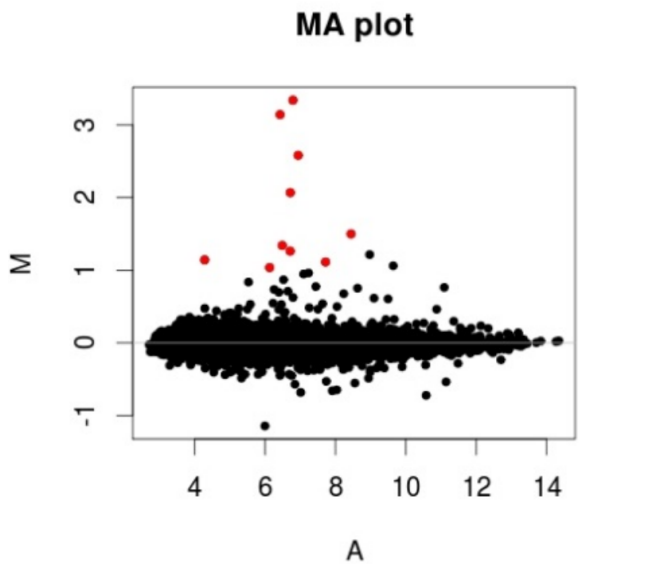
Issues

Arbitrary cutoff for M (e.g. $|M| > 1$)

Genes have different level of variation

M depends on over-all gene expression A





Gene selection by t-test

Assess the statistical significance of the observed difference in mean values between two groups

$$t = \frac{\textbf{difference of means}}{\textbf{variability}} = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$

Assumes independent experimental replicates
Assumes identically normally distributed data
Allows different group sizes $n_T \neq n_C$
Obtain P value from t using a table
As $n_T + n_C \nearrow$ P gets smaller
var_T is the variance of the data in the treatment group
var_C is the variance of the data in the control group

Gene selection by t-test

	Treatment			Control						
	T1	T2	T3	C1	C2	C3	$\bar{X}_T$	$\bar{X}_C$	M	t-test P
Gene1	17	16	15	8	10	12	16	10	6	0.0002
Gene2	17	18	16	16	16	16	17	16	1	0.9234
Gene3	8	20	8	4	3	5	12	4	8	0.5
...	...	...	...	...	...	...	...	...	...	...
Gene10000	18	17	19	15	17	16	18	16	2	0.0001

Small P-value (< 0.05) **big** mean log fold change
Small P-value (< 0.05) **trivial** mean log fold change
Large P-value (> 0.05) **big** mean log fold change
Large P-value (> 0.05) **trivial** mean log fold change

Hypothesis to test at significance level 0.05

$H_{alternative} : |\bar{X}_T - \bar{X}_C| > 0$

Gene g is regulated in the treatment group relative to the control group

$H_{null} : \bar{X}_T - \bar{X}_C = 0$

There is no difference in expression of gene g between the two groups

Gene g is differently expressed if t-test P $\leq$ 0.05 (H_{null} is rejected)

Gene selection by limma moderated t-test

In order to estimate the **gene-specific within-group variance** (var_{gene}) t-test needs **many replicates**, otherwise genes can have **small P-values by chance**

Rather than estimating within-group variability for each gene, **pool the global information from all other genes** when you have few replicates

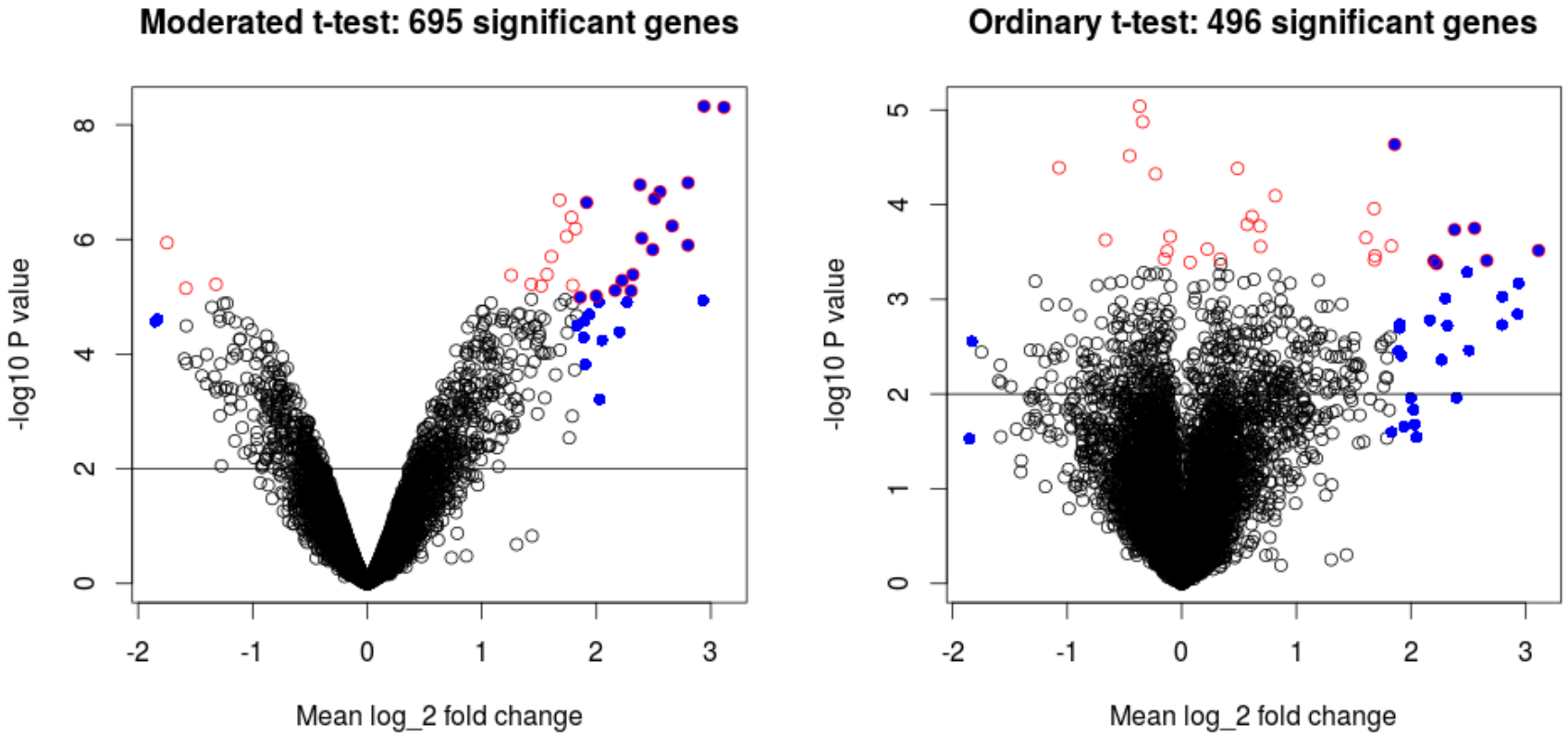
Moderated t-statistics is calculated using moderated variance $var_{t-moderated}$ estimated by empirical Bayes approach, that shrinks the gene-specific variance towards the global (across all genes) variance (var_{global})

$$var_{t-moderate} = f(var_{gene}, var_{global})$$

When many replicates are available the two statistics give similar results $var_t \sim var_{t-moderate} \sim var_{gene}$

Ritchie et al. (2015) Nucleic Acid Research

Gene selection by limma moderated t-test



Two replicates per group (estrogen dataset - see tutorial)
Both test performed at significance level of 0.01 (the black horizontal line)
Red circles represent the 30 genes with smallest P-value
Blue dots represent the 30 genes with highest absolute mean log fold change

Moderated t-test finds more differentially expressed genes than t-test

Multiple testing adjustment

What the significance level of 0.05 means?

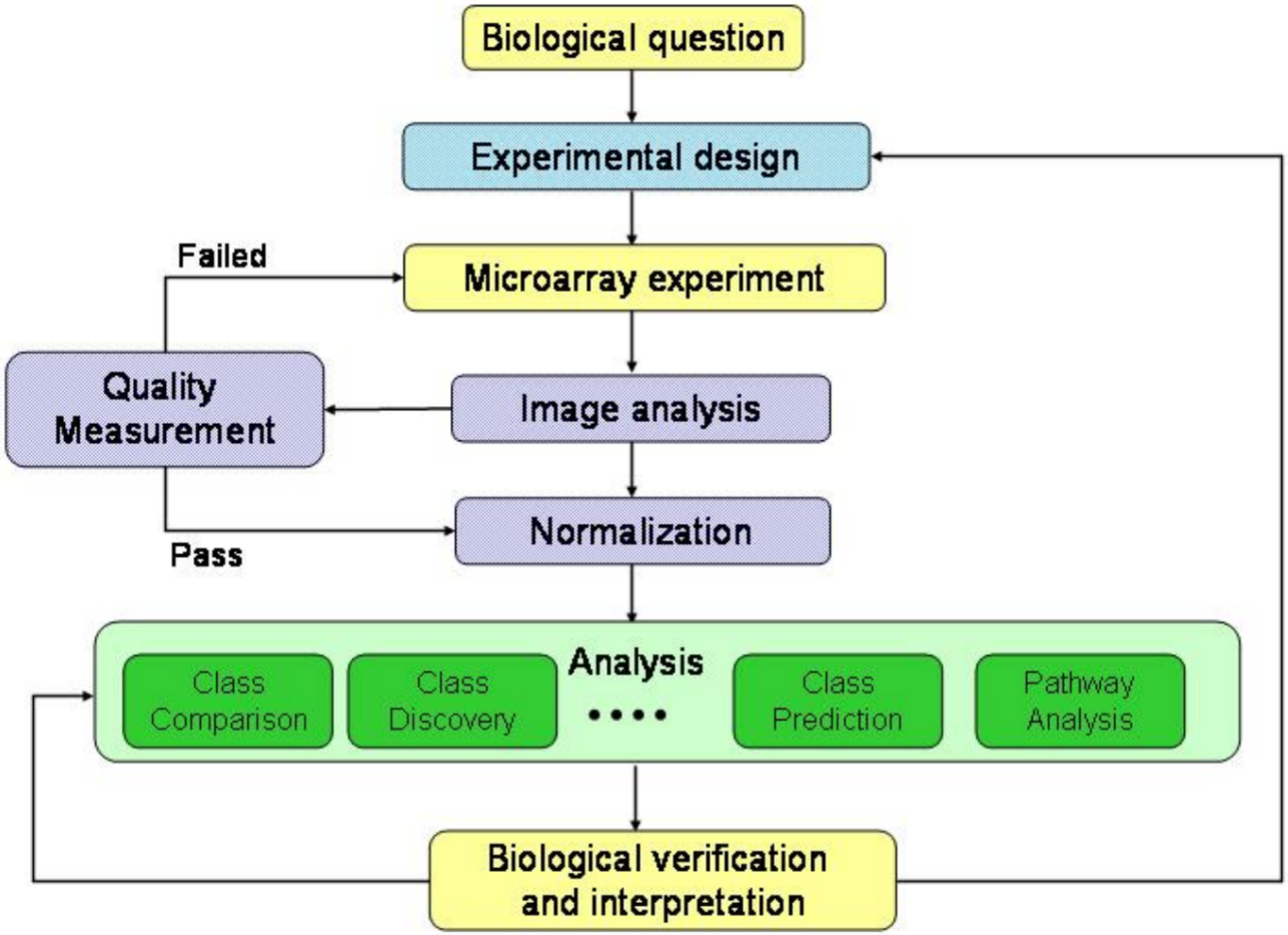
You have data for 10000 genes and even if none of the genes is truly differently expressed, you will expect to see $0.05 \cdot 10000 = 500$ genes by chance as regulated.

We can address this problem by P-value adjustment

Bonferoni	$P_{adj} = P \cdot 10000$ too conservative
Benjamini & Hochberg	Controls the false discovery rate FDR

FDR	Significant	False discoveries
0.1	100	$0.1 \cdot 100 = 10$
0.05	40	$0.05 \cdot 40 = 2$
0.01	40	$0.01 \cdot 40 = 4$

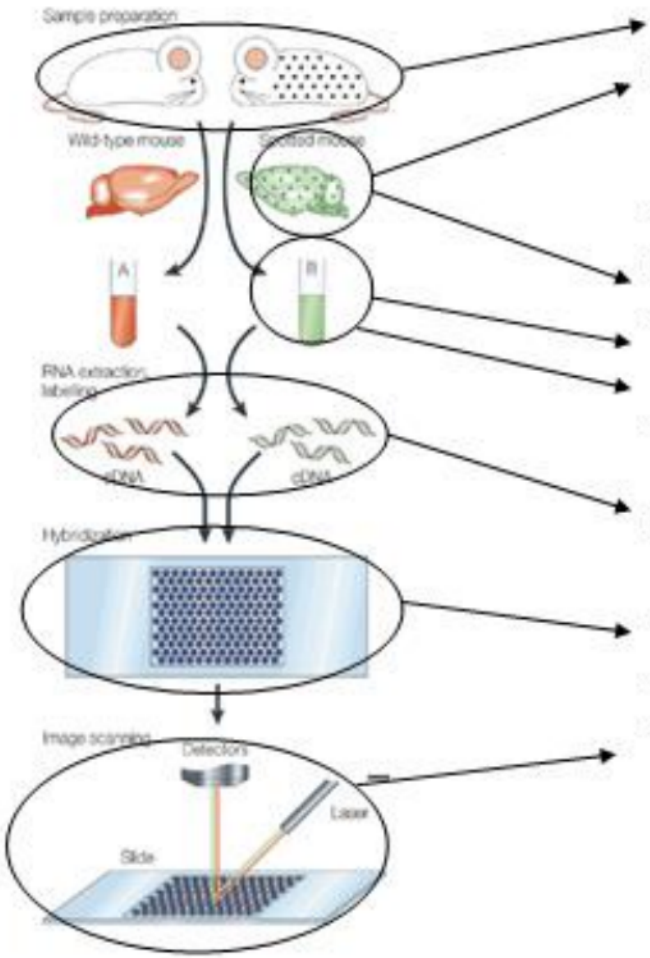
Integrated view



Sánchez and Ruíz de Villa (2008) A Tutorial Review of Microarray Data Analysis

Experimental design

Sources of data variability: systematic vs. random



- Biological Heterogeneity in Population.
- Specimen Collection/ Handling Effects.
  - Tumor: surgical bx, FNA.
  - Cell Line: culture condition, confluence level.
- Biological Heterogeneity in Specimen.
- RNA extraction.
- RNA amplification.
- Fluor labeling.
- Hybridization.
- Scanning.
- – PMT voltage.
- – laser power.

Sánchez and Ruíz de Villa (2008) A Tutorial Review of Microarray Data Analysis

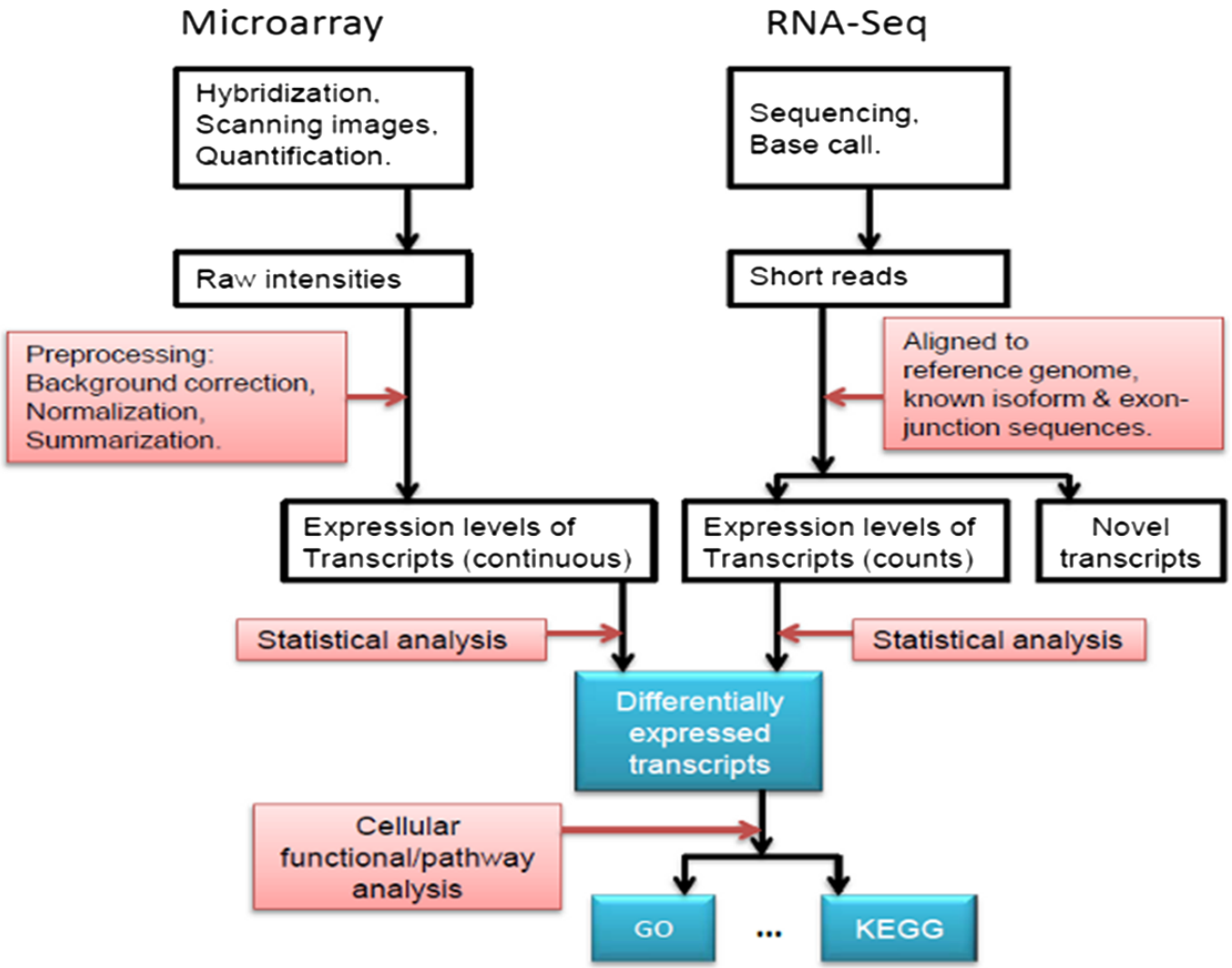
Experimental design

The number of samples determine the data analysis approach
Tradeoff between cost and reproducibility: 3^+ biological replicates per condition is a minimum!

Biological replicates	Recreate the experiment several times to get a sense of biological (population-level) variability
Technical replicates	Repeat hybridization with several chips to get a sense of microarray (measurement-level) variability

Why use DNA microarrays in the era of Next Generation Sequencing technology?

Analysis overview



Fang et al. (2012) Cell & Bioscience

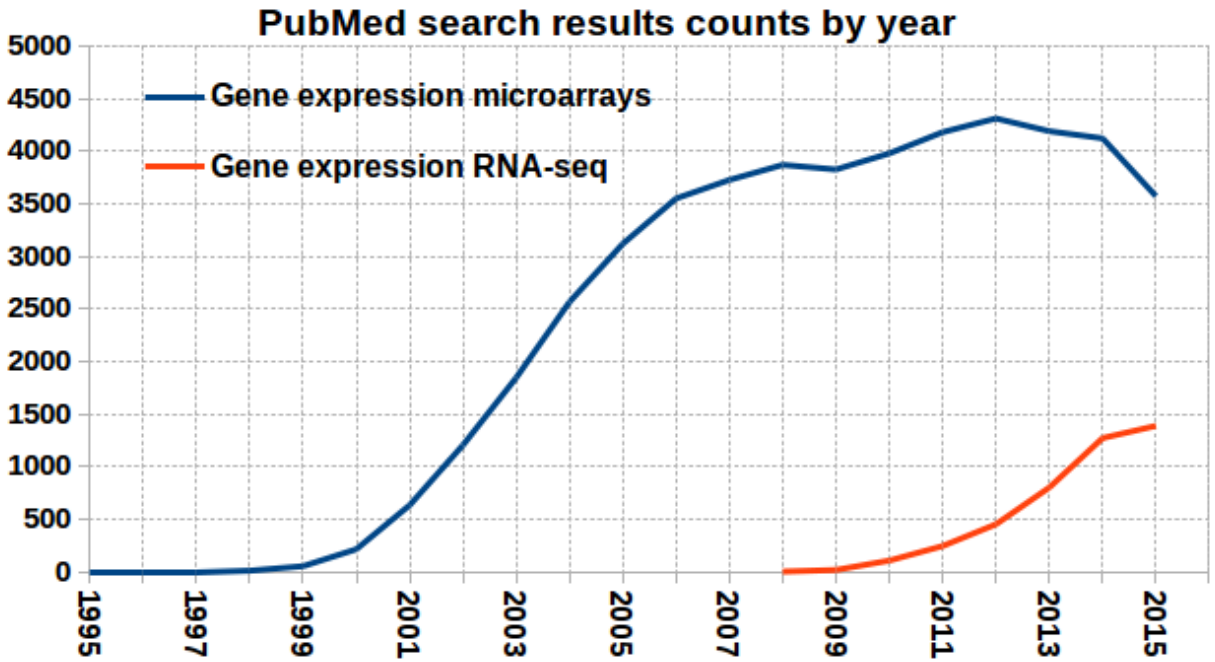
What the literature search says?

RNA-seq: direct sequencing of transcripts by high-throughput sequencing technologies

NCBI Gene Expression Omnibus: 66149 public data records

- Expression profiling by array: 44312
- Expression profiling by RNA-seq 6819

NCBI PubMed



Objective comparison

Microarrays

+ easier and mature protocols for sample preparation &data analysis
+ lower cost (100-200/sample)
+ yield higher throughput when processing a large number of samples
– cross-hybridization
– probe design bias & probe annotations
– limited ability to quantify lowly/highly expressed genes

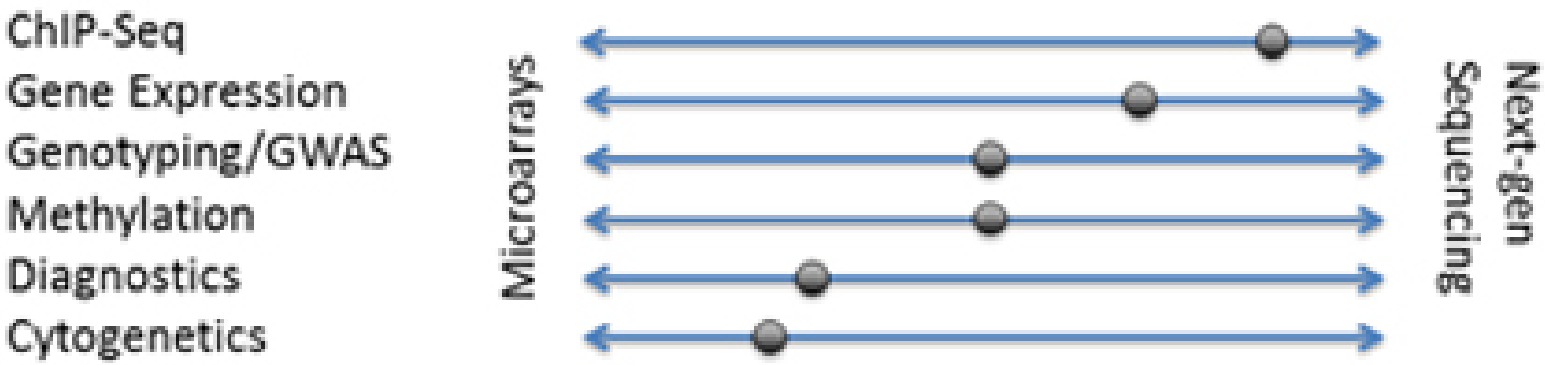
RNA-seq

+ precise and not subject to cross-hybridization
+ higher accuracy and wider dynamic range
+ discovery of novel transcripts, allele-specific expression and splice junctions
– complicated/time-consuming library preparation & data analysis
– higher cost (300-1000/sample)

Objective comparison

Trends based on application needs

research goals, access to technology, maturity of applications, cost per sample, and desired throughput



http://www.genengnews.com/gen-articles/next-generation-sequencing-vs-microarrays/4689/

Lets have some fun with R/Bioconductor!