Microarrays, differential expression analysis, pathway analysis

Mikhail Dozmorov

Fall 2018

Overview

Introduction to the theory and practice of gene expression analysis

- Microarray technology
- Exploratory data analysis, QC, normalization
- Differential expression analysis
- Pathway enrichment analysis

Microarray technology

What is a Microarray?

- "A DNA microarray is a multiplex technology consisting of thousands of oligonucleotide spots, each containing picomoles of a specific DNA sequence."
- An *oligonucleotide* (from Greek prefix *oligo-*, "having few, having little") is a short nucleic acid polymer.

What Are Microarrays Used For?

- Various molecular assays

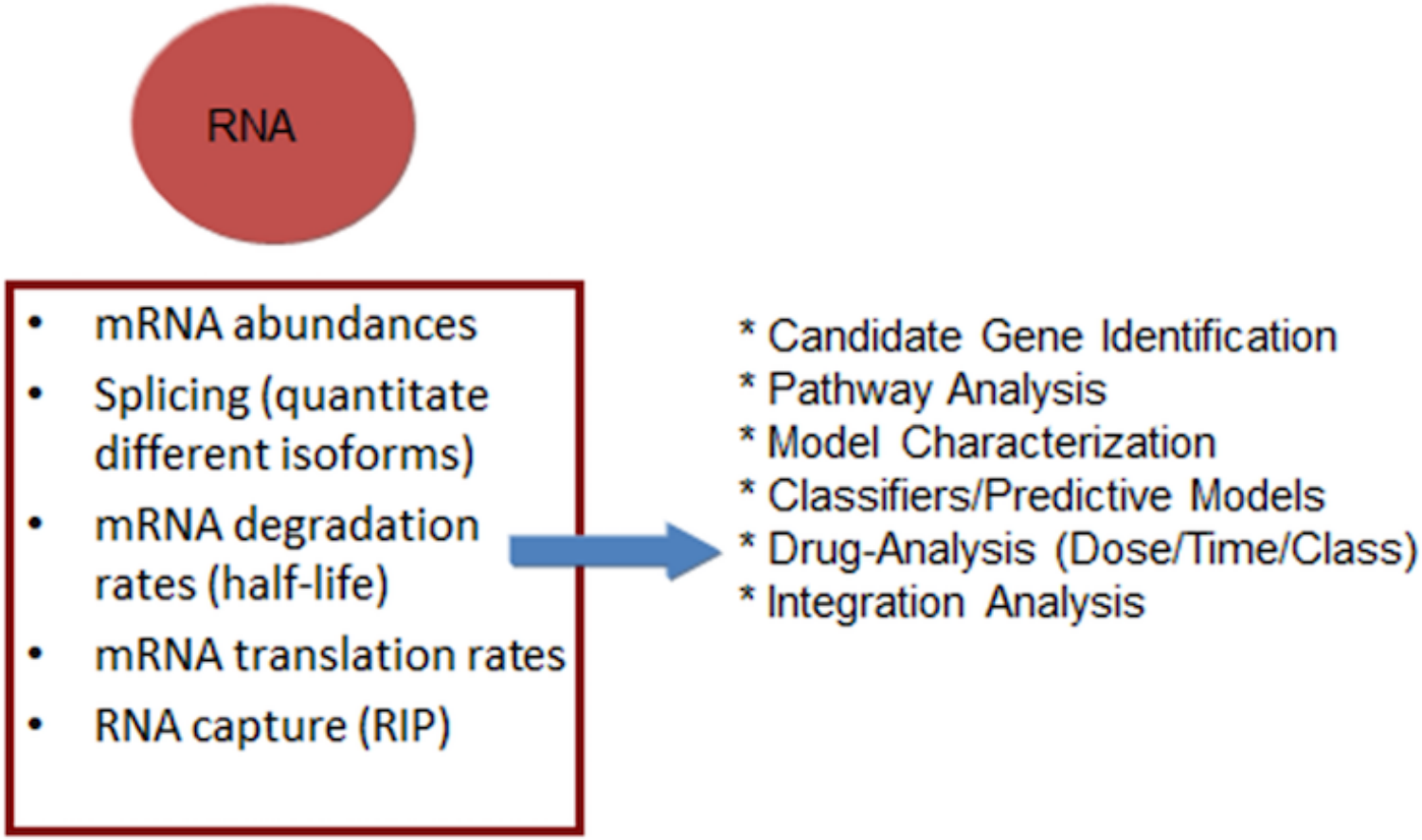
- DNA sequence (SNPs)
- DNA copy-number
- DNA capture (exome, ChIP)
- Tag quantitation (genetic screening)

- mRNA abundances
- Splicing (quantitate different isoforms)
- mRNA degradation rates (half-life)
- mRNA translation rates
- RNA capture (RIP)

- Protein arrays
- Cell based arrays
- Lipid arrays

What Are Microarrays Used For?

- Biological insights



Microarrays measure expression of all genes

- Traditional molecular biology research followed a "*one gene per experiment*" paradigm.
- With the advent of microarrays, research practice has moved from a "one gene at a time" mode to "thousands of genes per experiment".
- Allows for the study of how genes function *en masse*.

Microarrays are used in all areas of life sciences

- **Cancer research**: Molecular characterization of tumors on a genomic scale; more reliable diagnosis and effective treatment of cancer.
- **Immunology**: Study of host genomic responses to bacterial infections.
- **Model organisms**: Multifactorial experiments monitoring expression response to different treatments and doses, over time or in different cell types.

Typical comparisons

- Compare mRNA transcript levels
  - Different type of cells, tissues (e.g., liver vs. brain).
  - Treatment (Drugs A, B, and C).
  - Disease state (tumor vs. normal).
  - Different organism (yeast, different strains) different timepoints.

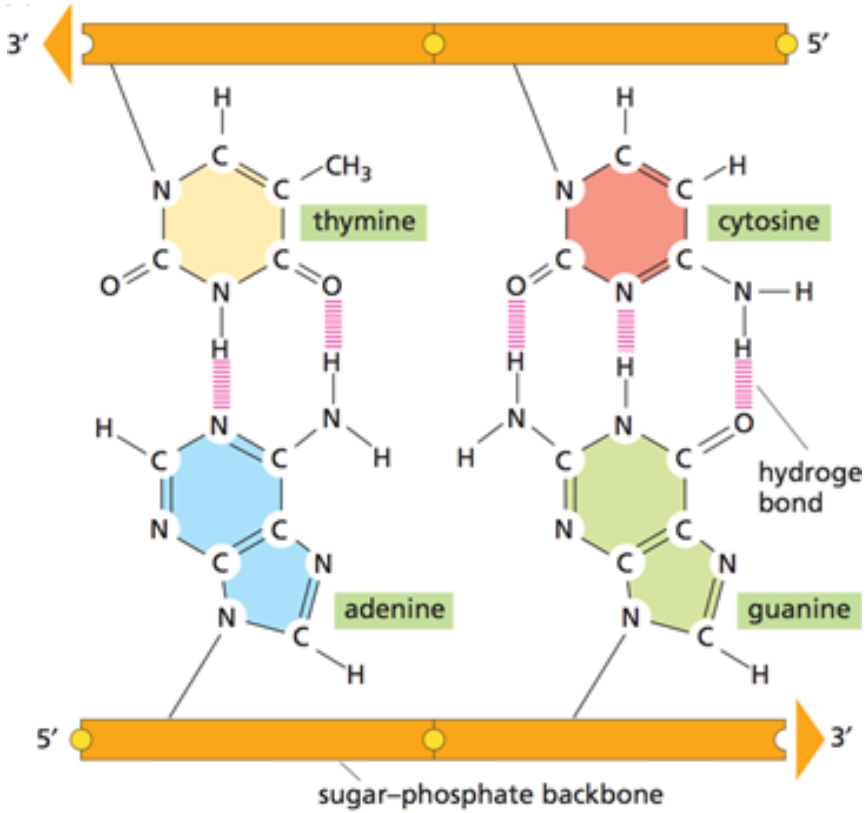
Microarrays in cancer detection

- Characterize molecular variations among tumors by monitoring gene expression.
- Divide morphologically similar tumors into different groups based on gene expression.
- Goal: microarrays will lead to more reliable tumor classification and sub-classification (therefore, more appropriate treatments will be administered resulting in improved outcomes).

Basic Design of Expression Arrays

- For **each gene** that is a target for the array, we have a **known DNA sequence**.
- Microarrays are composed of short DNA sequences complementary to the target genes.
- These sequences are attached to a slide at high density.

Complementary hybridization



Hybridization - Two single-stranded DNA molecules whose sequences are complementary to each other will exhibit a tendency to bind together to form a single double-stranded DNA molecule.

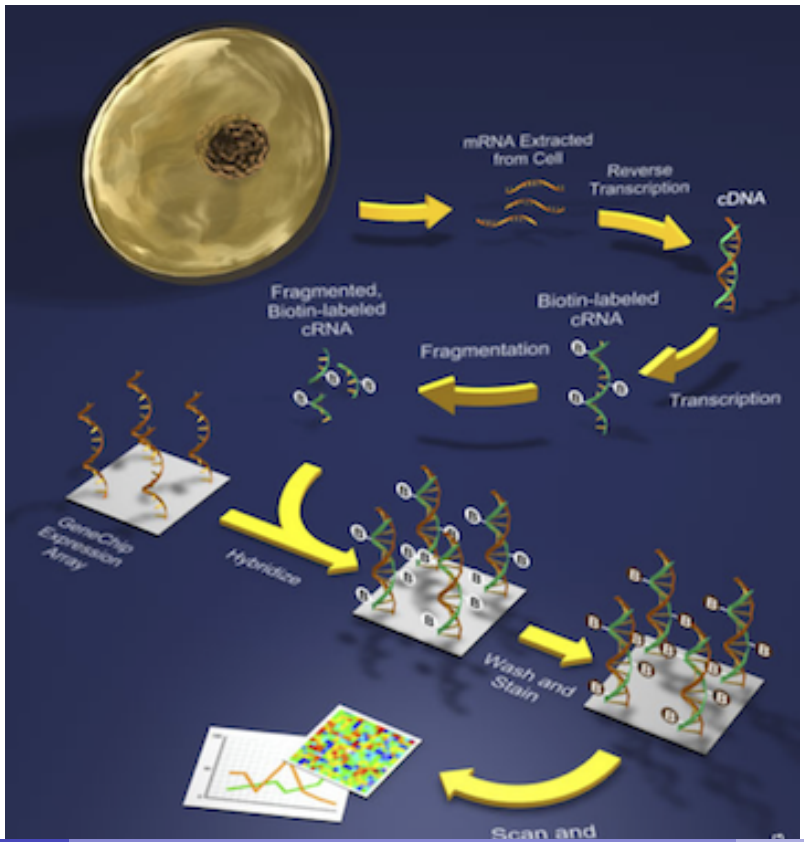
Basic Design of Expression Arrays

- mRNA is reverse transcribed to cRNA, and if a complementary sequence is on the on a chip, the cRNA will be more likely to hybridize to it.
- The cRNA is labeled with a dye that will fluoresce and generate a signal that is monotonic with the amount of the mRNA sample.
- The amount of hybridization can be **quantitatively** measured by the amount of fluorescence.

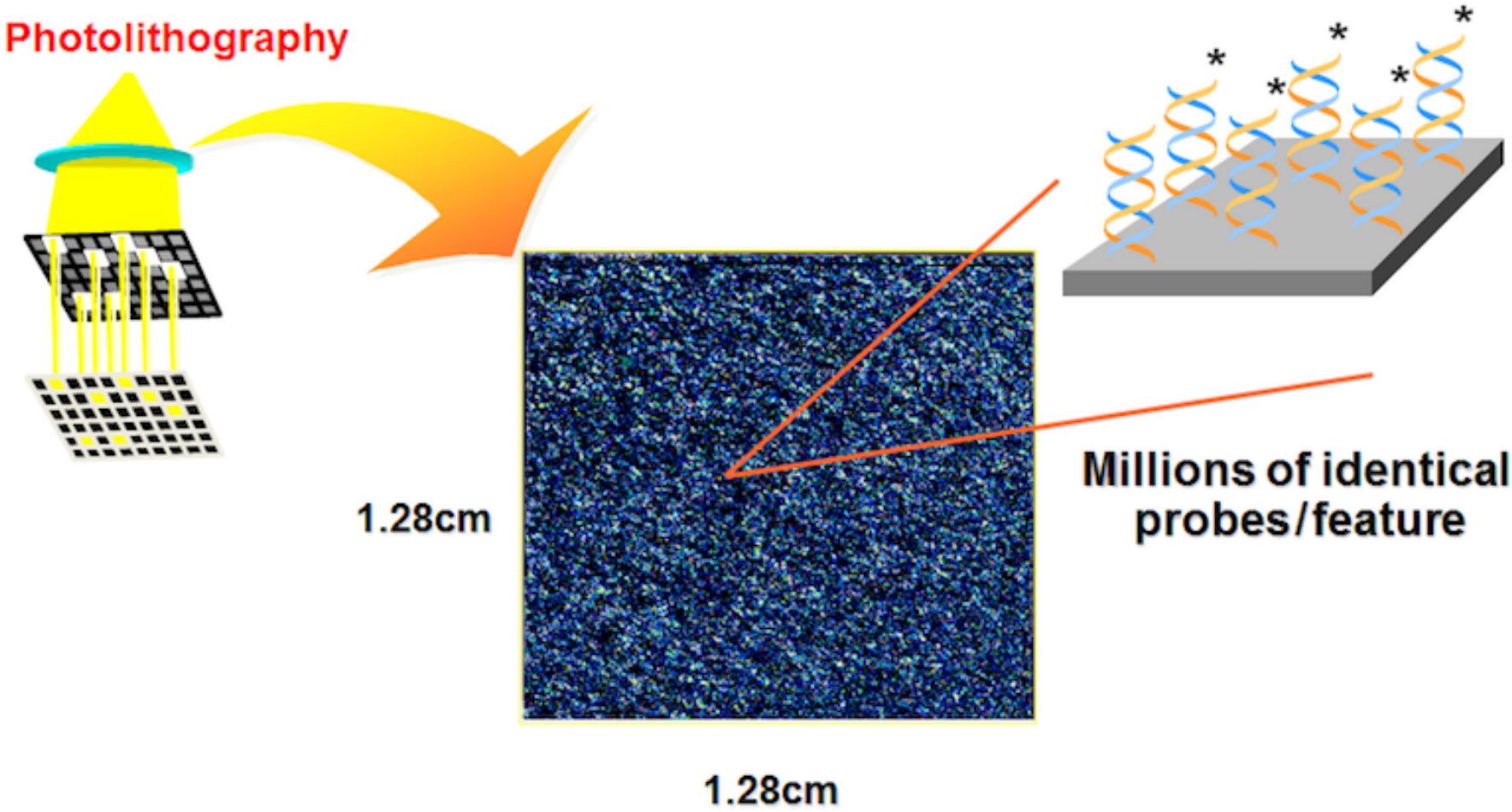
Single-channel arrays

Single-channel arrays

- mRNA extraction from one sample
- cRNA synthesis and fluorescent dye-labeling
- cRNA hybridization onto array
- Scanning and quantification of fluorescence of each spot

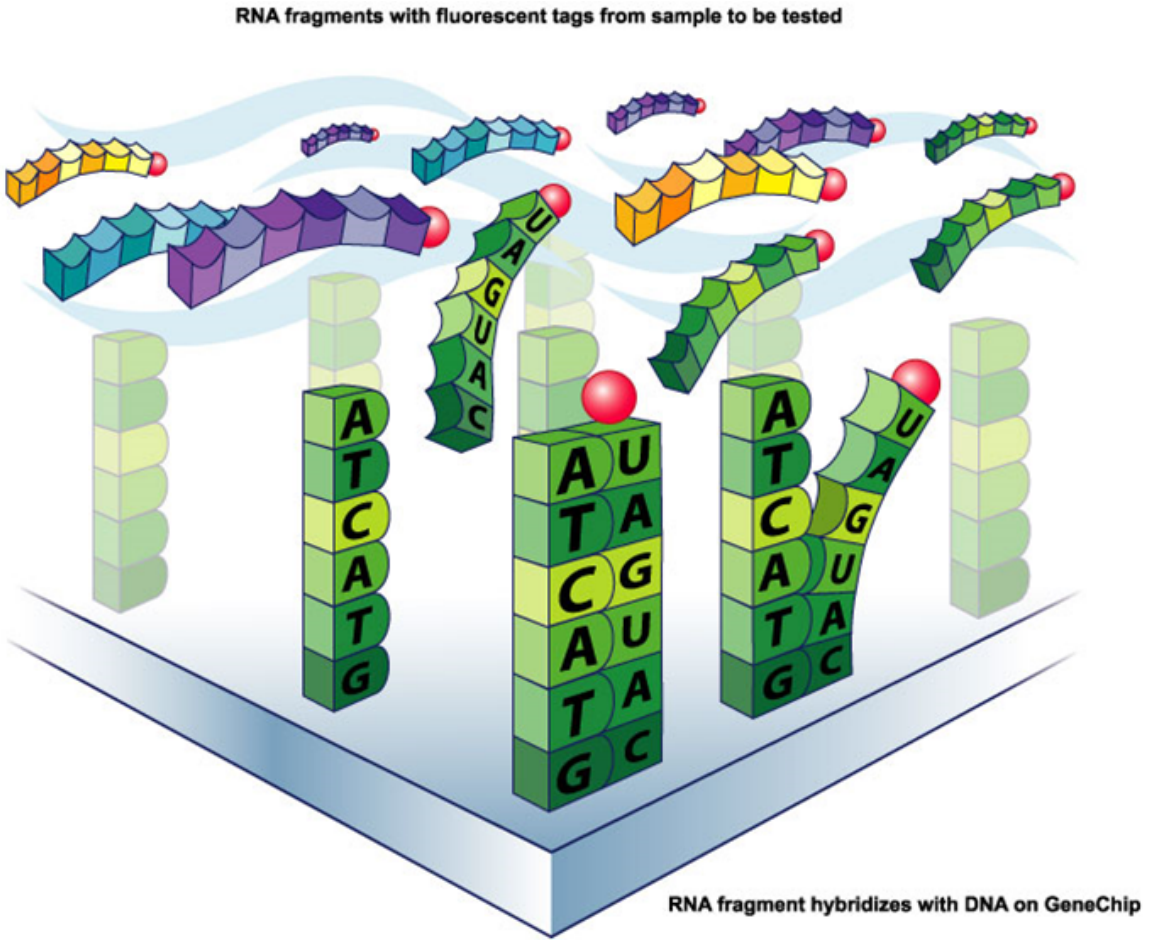


Oligonucleotide Arrays | Affymetrix arrays

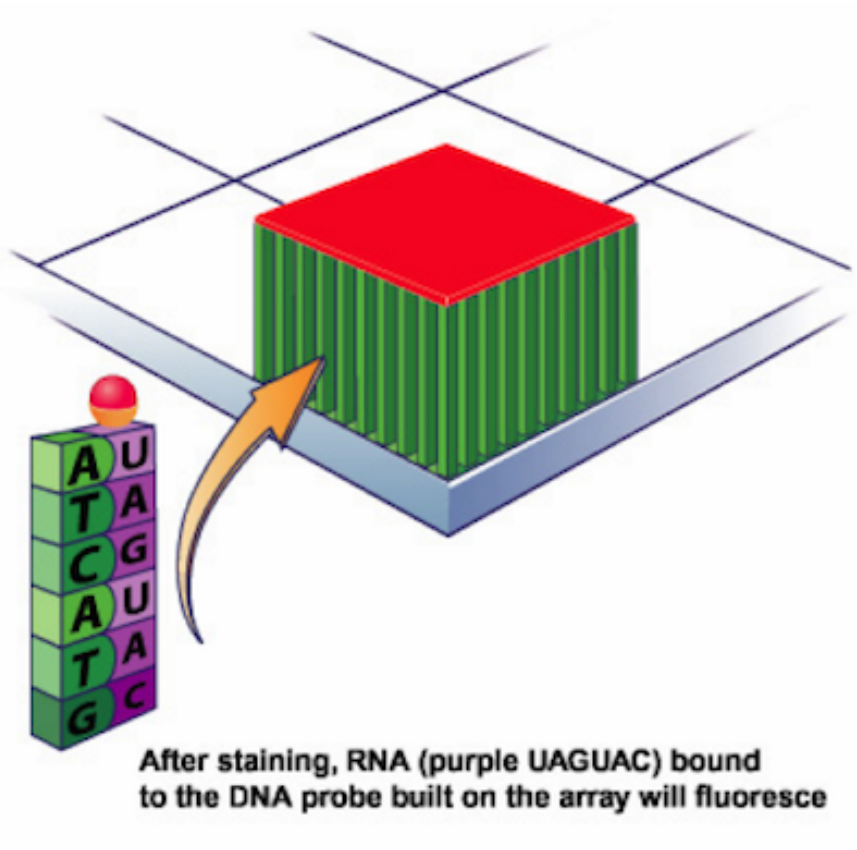


https://youtu.be/MRmpeBTwwWw

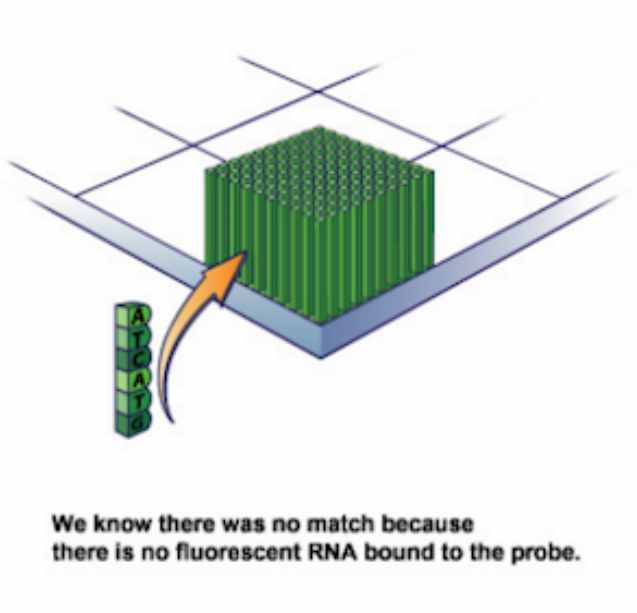
RNA Wash


RNA fragments with fluorescent tags from sample to be tested
RNA fragment hybridizes with DNA on GeneChip

RNA Wash



After staining, RNA (purple UAGUAC) bound to the DNA probe built on the array will fluoresce



We know there was no match because there is no fluorescent RNA bound to the probe.

Additional Microarray Platforms

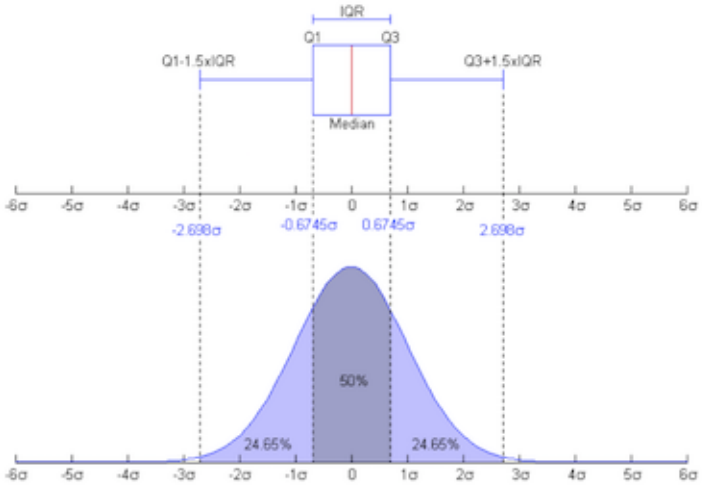
Array	Probes *on the array*	Targets *to be hybridized*	Large-scale Analysis of...
Gene Expression	DNA (cDNA, oligos: gene representatives)	mRNA/cDNA	transcriptional alterations
CGH	DNA (clones, oligos)	DNA	Genomic changes in cancers
SNP	DNA (oligos)	DNA	Genotyping; Genomic changes
Methylation	DNA (CpG island)	DNA (IP or bisulfite-treated)	Methylation-status in genes
Promoter	DNA (promoter ~1kb)	DNA (ChIP-enriched)	Transcription factor binding sites; histone modifications
Tiling	DNA	All of the above	All of the above; sequencing; gene annotation
Protein	antibody	protein	Protein expression
Tissue	tissues	proteins	Histology; protein expression (immunohistochemistry)

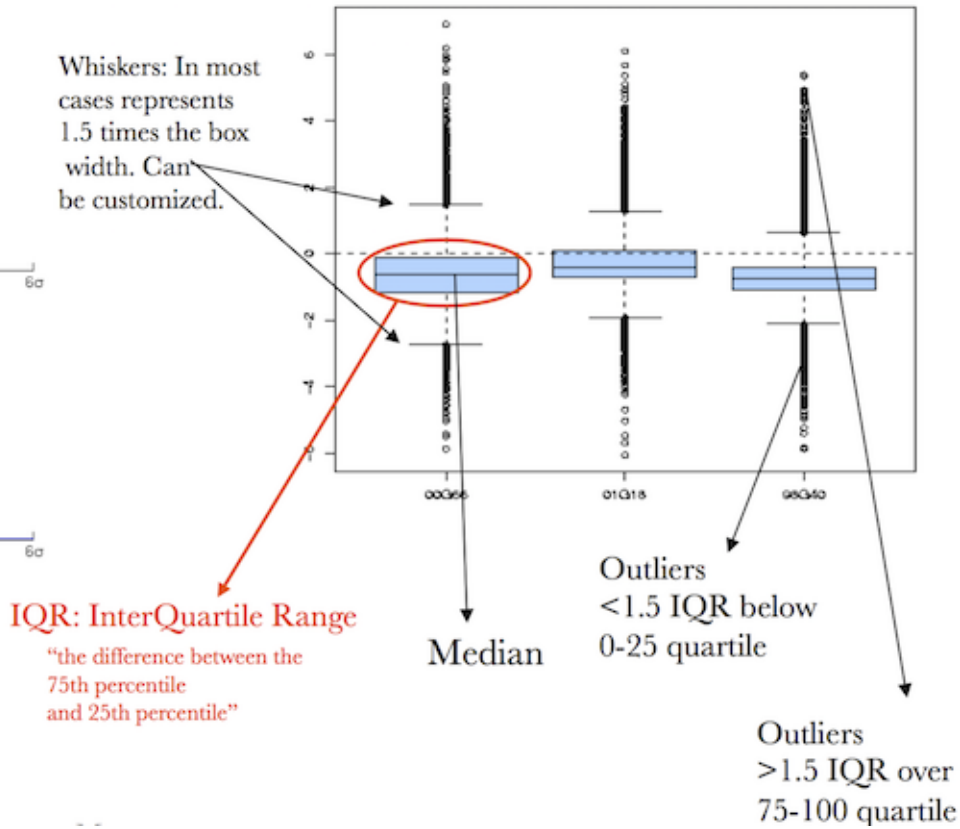
Exploratory data analysis, QC, normalization

Gene expression data

- A matrix of *n* genes (rows) x *m* samples (columns).
  - Each cell in the matrix has real number expression value for a given gene in a given sample.

Boxplots



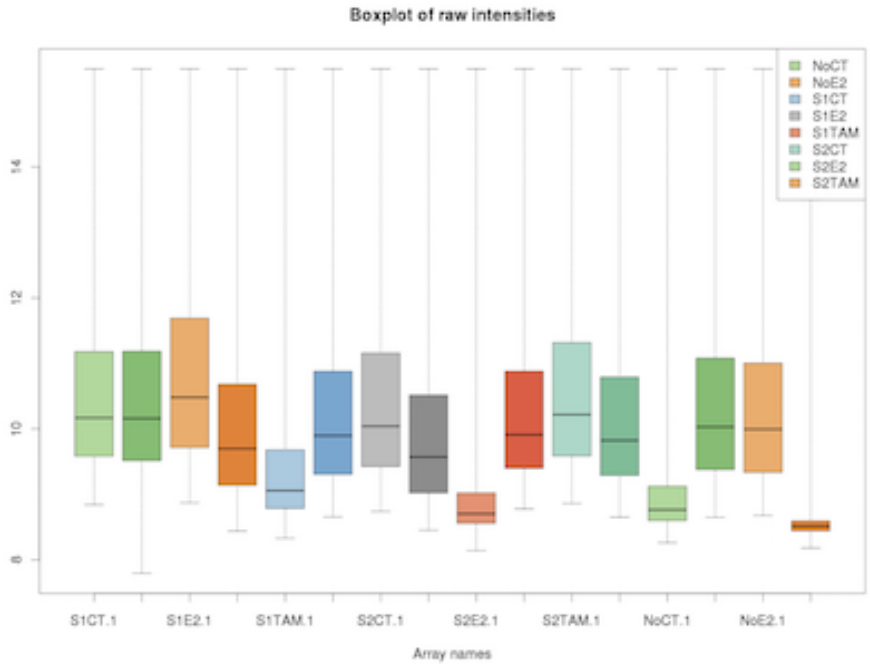


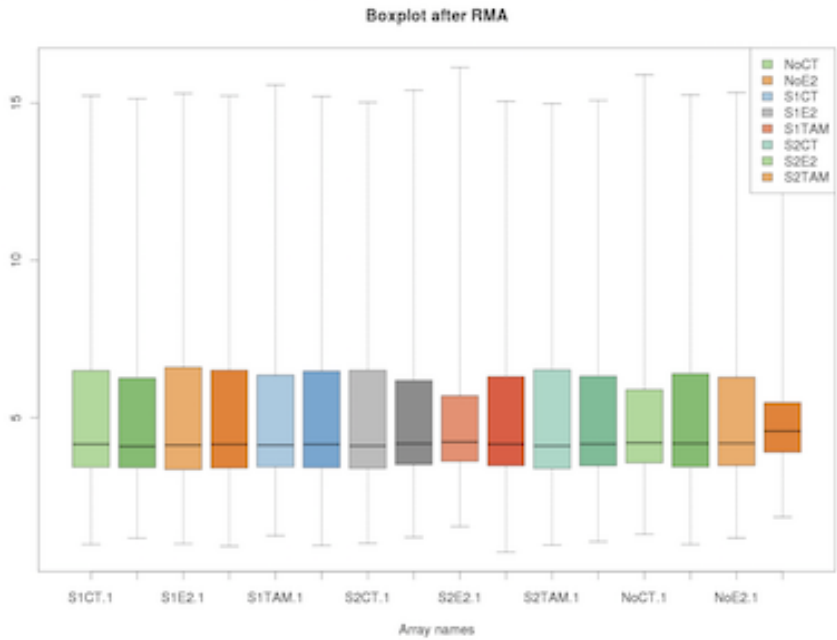
The box plot can answer the following questions:

- Does signal distribution/variation differ between subgroups?
- Are there any outliers?

Boxplots of log-intensities

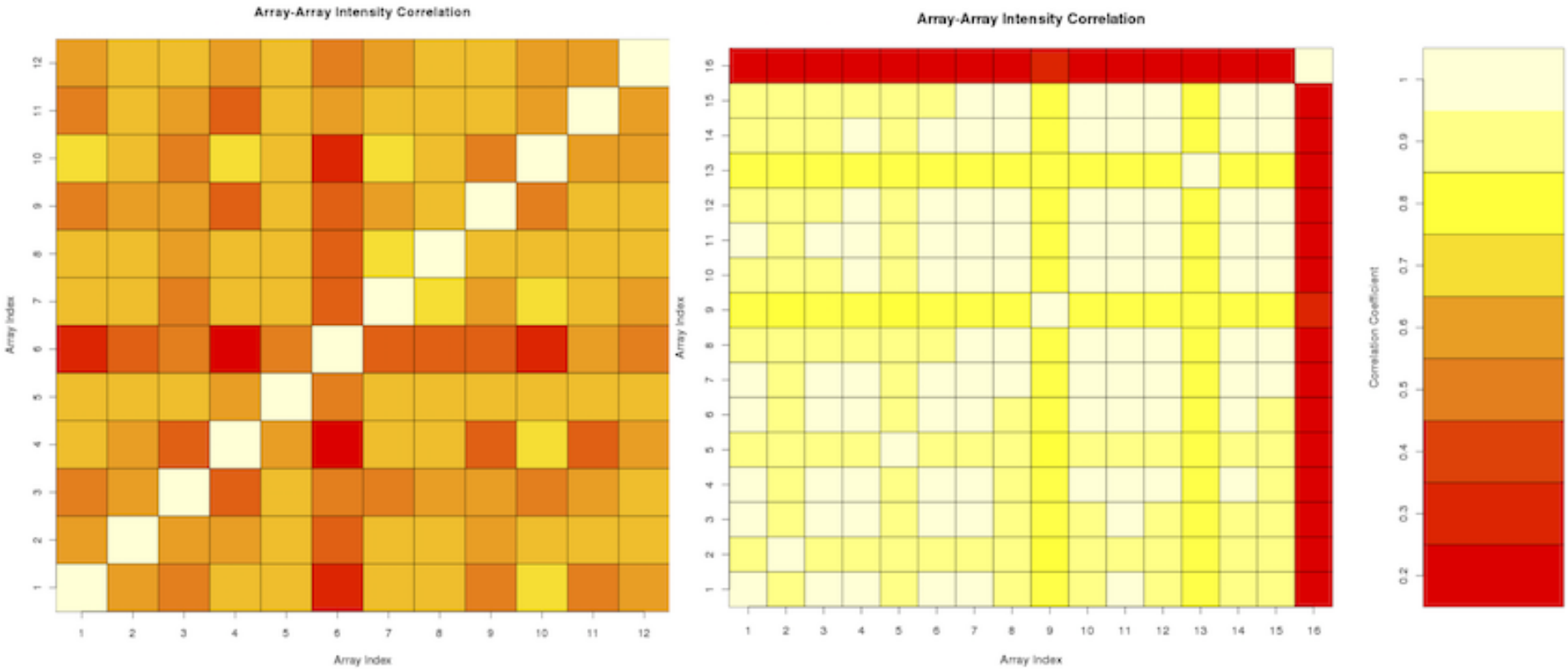
The distributions of raw PM log-intensities are not expected to be identical but still not totally different while the distributions of normalized (and summarized) probe-set log-intensities are expected to be more comparable.





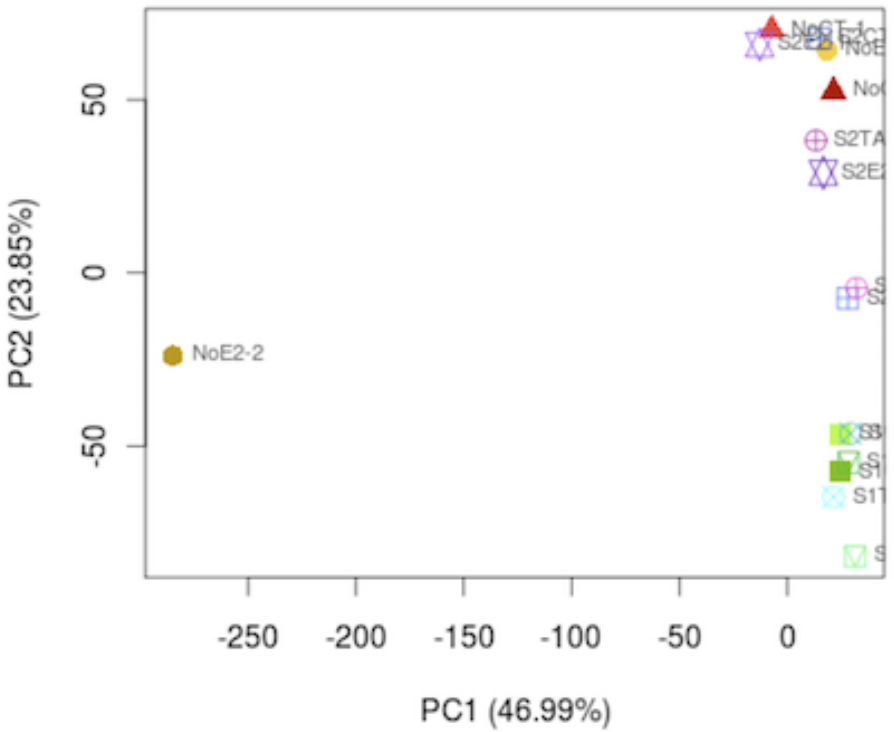
Correlation between arrays

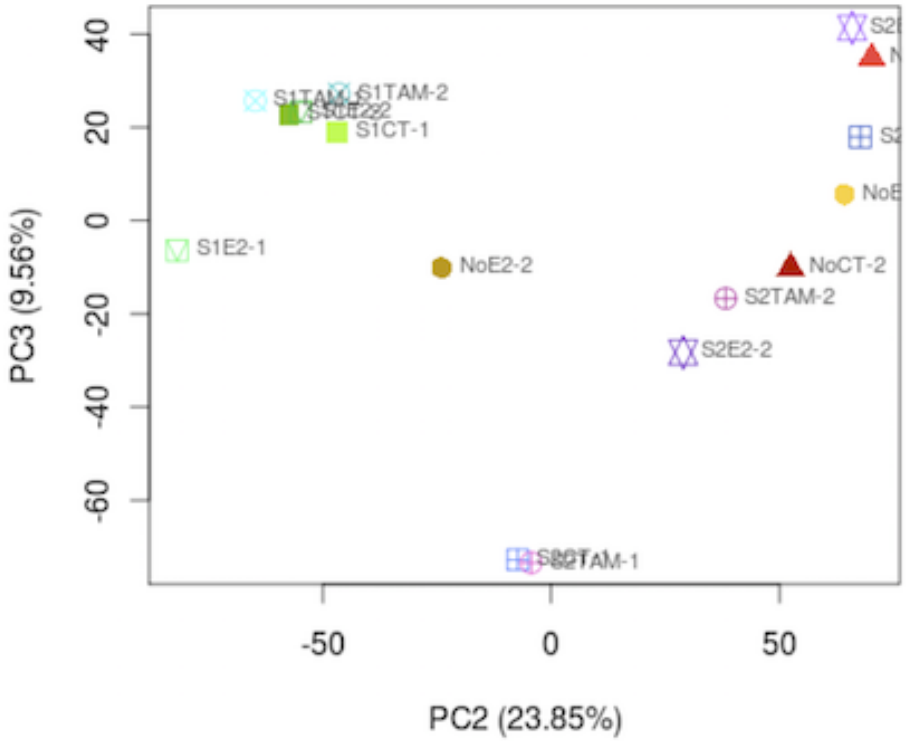
- A correlation coefficient is computed for each pair of arrays in the dataset and is visualized as a heatmap.
- Best to do at each pre-processing step, e.g., before/after normalization



Principal Components Analysis

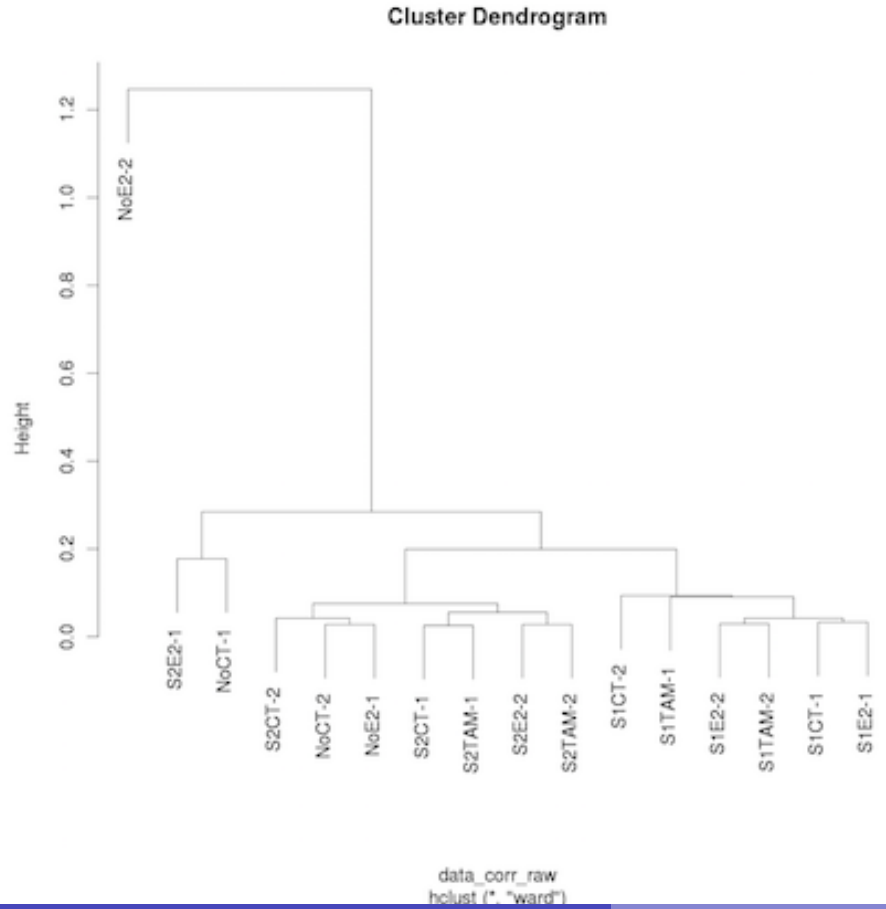
- Projects arrays onto a coordinate system that emphasizes variability among data

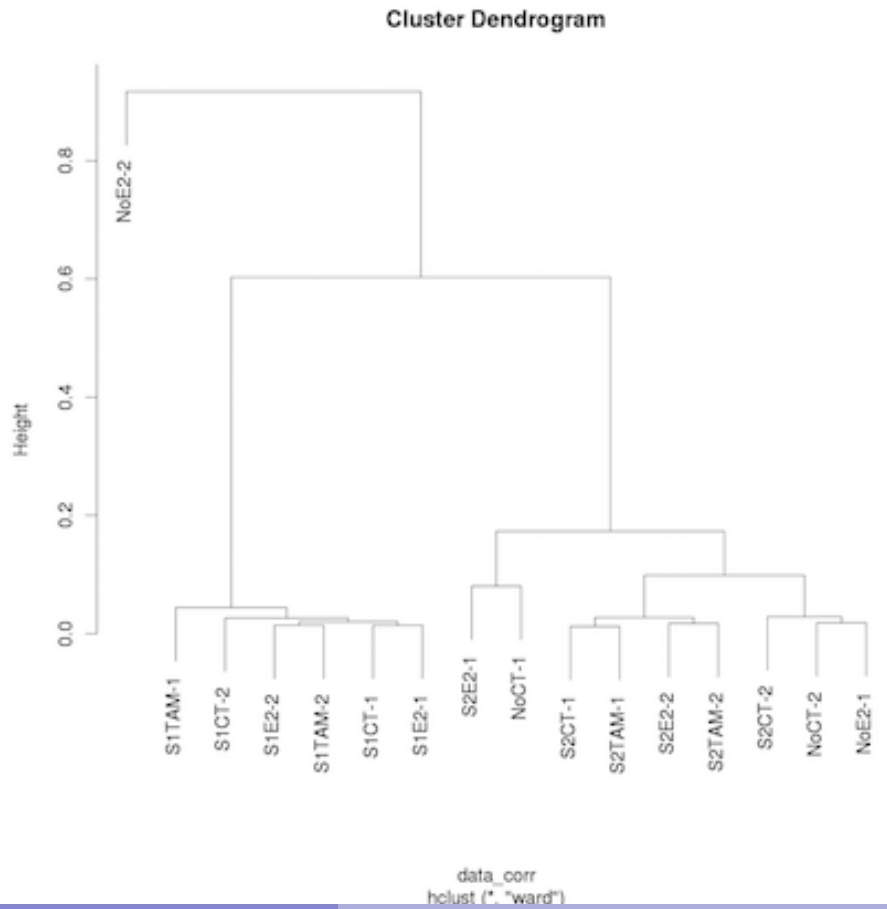




Hierarchical clustering

The Hierarchical Clustering plot is computed in two steps: first it computes an expression measure distance between all pairs of arrays and then it creates the tree from these distances.





Biological vs. technical variability in gene expression

- What is ultimately of interest in the use of gene expression microarrays is the measurement of **differences between experimental and reference states** or **between different groups** of experimental units.
- Observed differences in microarray gene expression studies, however, are recognized as arising from two sources:
  - **Biological variability** – changes in signal intensity driven by changes between biological states (healthy – disease)
  - **Technical variability** – non-biological sources of variability

Sources of technical variability

Systematic

- Amount of extracted RNA, efficiencies of RNA extraction, reverse transcription, labeling, photodetection, GC content of probes
- Similar non-bilological effect on many measurements

- Corrections can be estimated from data and accounted for by normalization

Sources of technical variability

Stochastic

- PCR yield, DNA quality, spotting efficiency, spot size, non-specific hybridization, stray signal
- Noise components & "Schmutz" (dirt)

- Too random to be explicitly accounted for – need to use error modeling

Why normalization

Main idea

- Remove the systematic bias in the data as completely as possible while preserving the variation in the gene expression that occurs because of biologically relevant changes in transcription.

- The purpose of normalization is to adjust the gene expression values so that all genes on the array *that are not differentially expressed* have similar values across all arrays.

Goal of normalization

Assumption

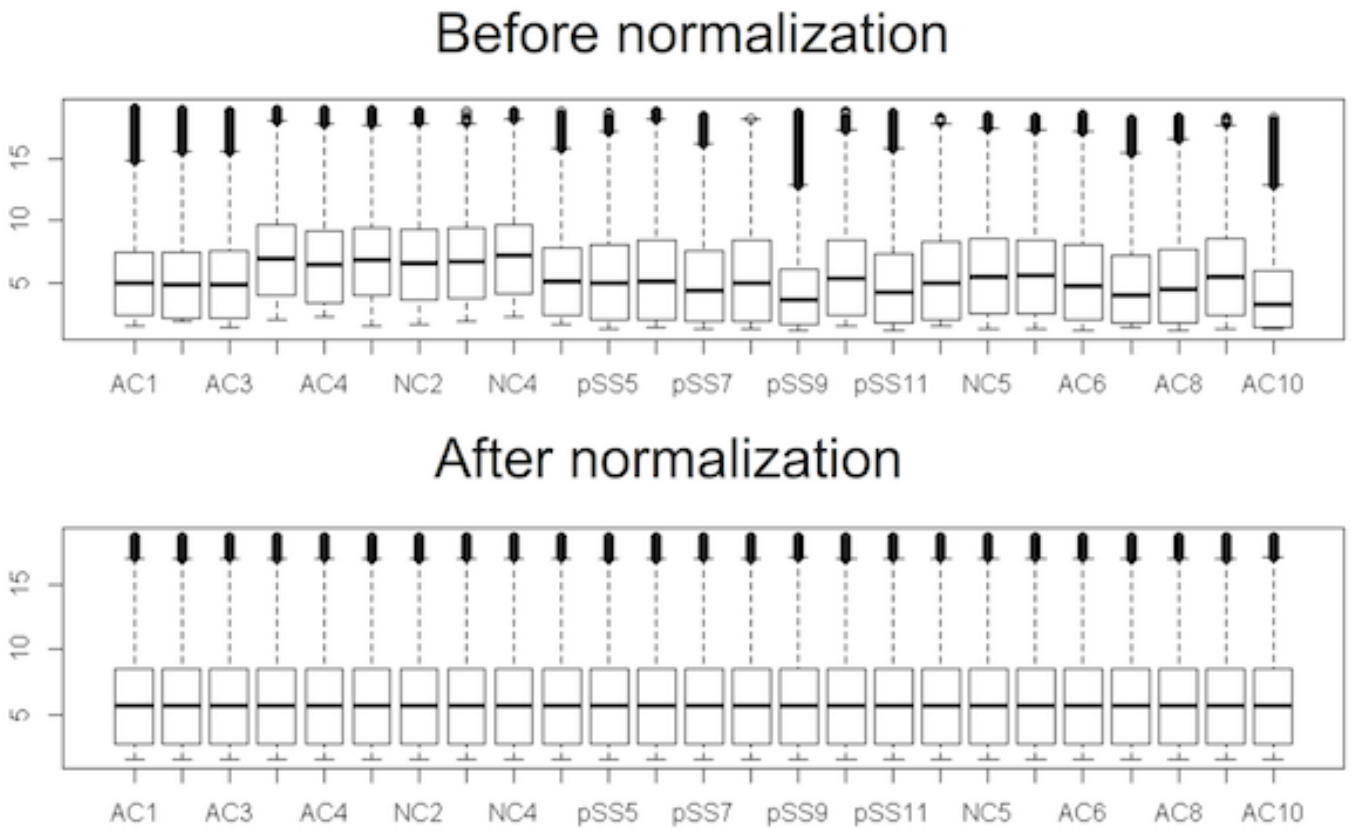
- The average gene does not change in its expression level in the biological sample being tested.
- Most genes are not differentially expressed.
- Up- and down-regulated genes roughly cancel out the expression effect.

Quantile normalization

- Motivation from quantile-quantile plot.
- Normal quantile-quantile plot consists of a plot of the ordered values in your data versus the corresponding quantiles of a standard normal distribution.
- If the normal qqplot is fairly linear, your data are reasonably Gaussian; otherwise, they are not.

Between-array normalization methods

- **Quantile normalization**: Make distribution of data equal across all samples. Final distribution is the average of each quantile across chips (Bolstad et.al., Bioinformatics, 2003).



Quantile normalization

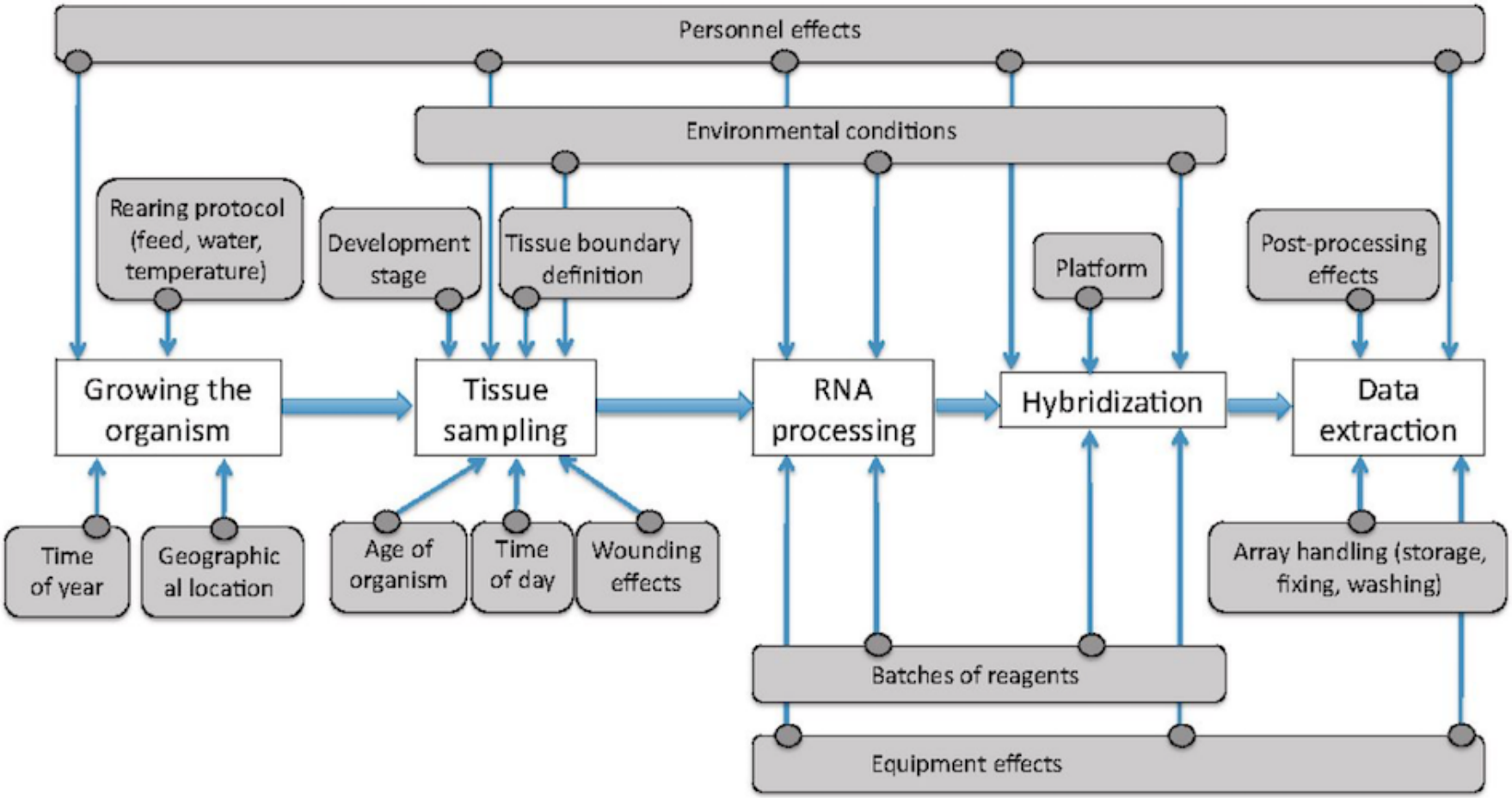
1. Given n arrays of length p, form matrix X of dimension $p??n$ where each array is a column.
2. Sort each column of X to get X_{sort}. Remember to original order
3. Take the means across rows of X_{sort} and replace the values of X by those means. The resulting matrix is X'_{sort}.
4. Get $X_{normalized}$ by rearranging each column of X'_{sort} to have the same ordering as original X.

Quantile normalization changes expression over many slides i.e. changes the correlation structure of the data, may affect subsequent analysis.

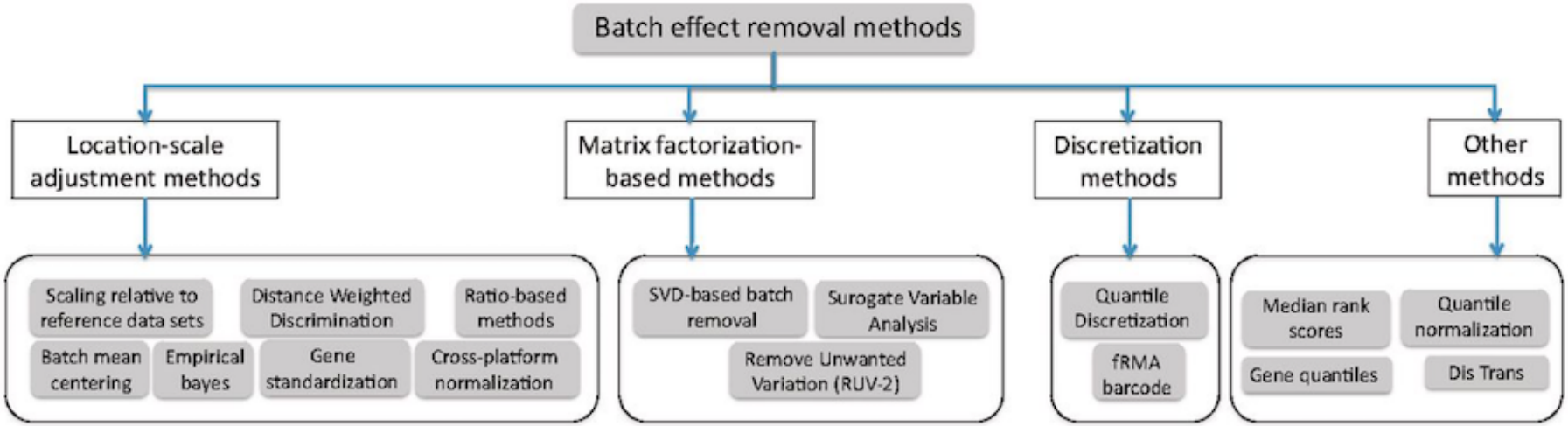
Batch effects

- Batch effects are widespread in high-throughput biology. They are artifacts not related to the biological variation of scientific interests.
- For instance, two microarray experiments on the same technical replicates processed on two different days might present different results due to factors such as room temperature or the two technicians who did the two experiments.
- Batch effects can substantially confound the downstream analysis, especially meta-analysis across studies.

Batch sources



Batch removal methods



Lazar et.al., "**Batch effect removal methods for microarray gene expression data integration: a survey**" Brief Bioinform 2013
http://bib.oxfordjournals.org/content/14/4/469.long

The effect of batch removal

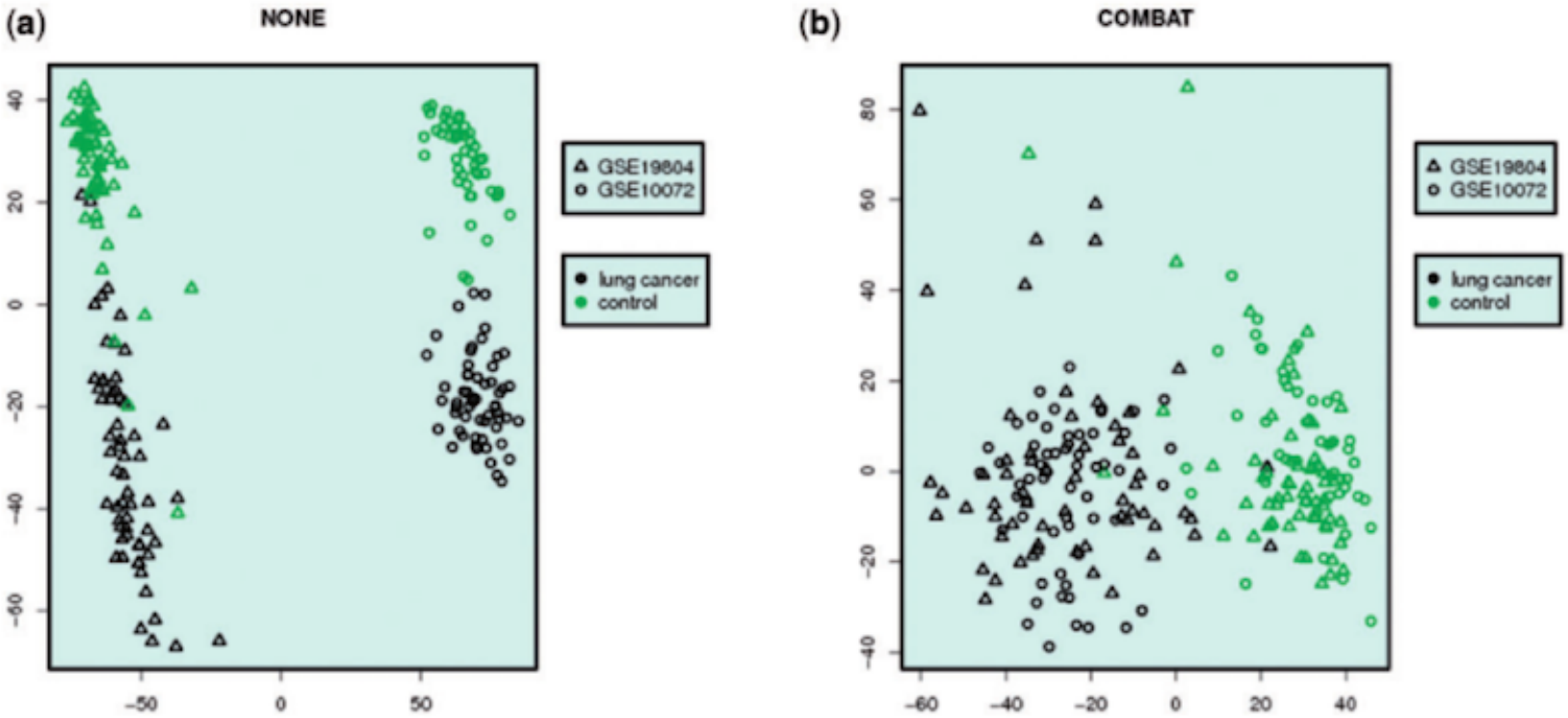


Figure 7: Illustration of PCA plots as validation tools for batch effect removal. Plot of first two principal components: (a) before batch effect removal and (b) after batch effect removal (using EB method).

Lazar et.al., "**Batch effect removal methods for microarray gene expression data integration: a survey**" Brief Bioinform 2013 http://bib.oxfordjournals.org/content/14/4/469.long

What to use

"ComBat, an Empirical Bayes method, outperformed the other five programs by most metrics"

- Chen C et.al., "**Removing Batch Effects in Analysis of Expression Microarray Data: An Evaluation of Six Batch Adjustment Methods**" PLoS ONE 2011 http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0017238

Differential expression analysis

Supervised Learning

- **Class comparison/ Feature selection**
  - T-test / Wilcoxon rank sum test
  - F-test / Kruskal-Wallis test
  - Adjustment for multiple comparisons
- **Class Prediction**
  - K nearest neighbors
  - Compound Covariate Predictors
  - Classification trees
  - Support vector machines
  - etc.

Hypothesis testing

- The hypothesis that two means μ_1 and μ_2 are equal is called a null hypothesis, commonly abbreviated H_0.
- This is typically written as $H_0 : \mu_1 = \mu_2$
- Its antithesis is the alternative hypothesis, $H_A : \mu_1 \neq \mu_2$

Hypothesis testing

- If the sample means calculated are identical, we would suspect the null hypothesis is true.
- Even if the null hypothesis is true, we do not really expect the sample means to be identically equal because of sampling variability.
- We would feel comfortable concluding H_0 is true if the chance difference in the sample means should not exceed a couple of standard errors.

P-value

- The p-value for a hypothesis test is the probability, computed under the condition that the null hypothesis is true, of the test statistic being at least as extreme as the value of the test statistic that was actually obtained.
  - A large p-value (close to 1) indicates a value of t near the center of the t-distribution.
  - A small p-value indicates a value of t in the far tails of the t-distribution.

Hypothesis testing

- The **mean** μ_X of a random variable X is a measure of central location of the density of X.
- The **variance** of a random variable is a measure of spread or dispersion of the density of X.
- $Var(X) = E[(X - \mu)^2] = \sum \frac{(X-\mu)^2}{(n-1)} = \sigma^2$
- Standard deviation $= \sqrt{Var(X)} = \sigma$

Two-sample comparison

Let us consider the simplest case: two-sample comparison. Our goal is to find the list of genes that are differentially expressed. Suppose we have:

- n_1 samples in group 1
- n_2 samples in group 2
- For each gene, $n_1 + n_2$ expression levels are recorded for all the samples

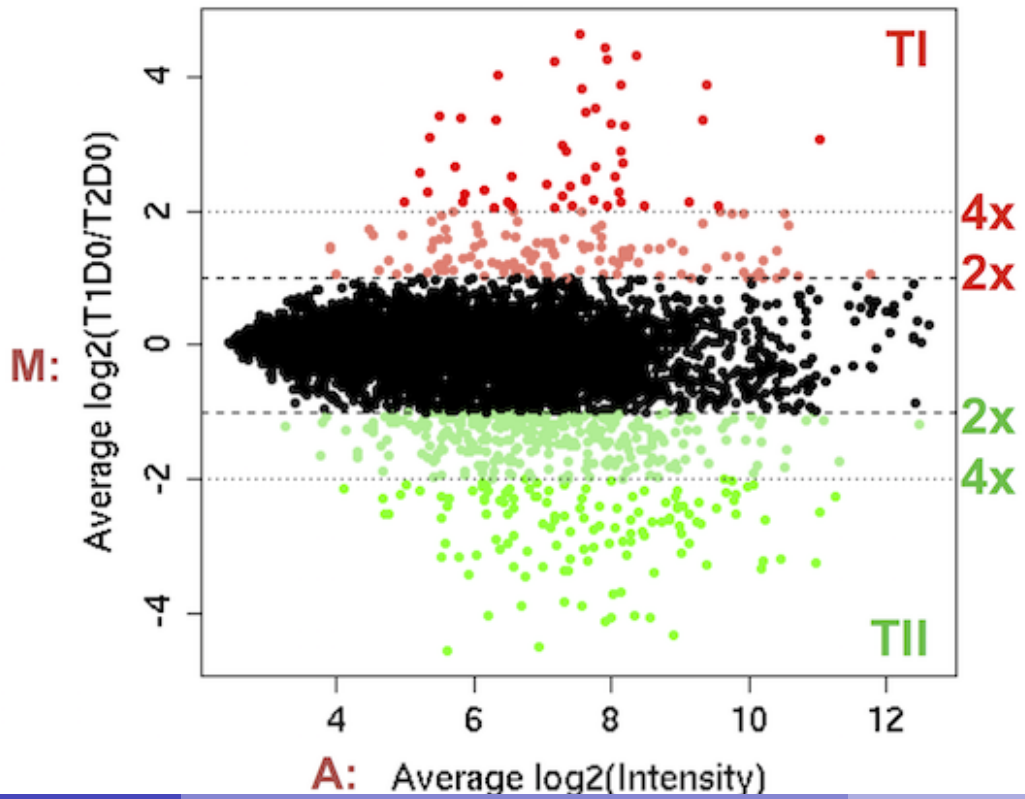
Determine which genes have differential expression between the two groups of samples.

Differential expression

- Many microarray experiments are carried out to find genes which are **differentially expressed between two (or more) samples**.
- Initially, comparative microarray experiments were done with few, if any, replicates, and statistical criteria were not used for identifying differentially expressed genes. Instead, simple criteria were used such as fold-change, with 2-fold being a popular cut-off.
- The simplest experiment involves comparing two samples on one array with two-color technology or two arrays if using one-color technology

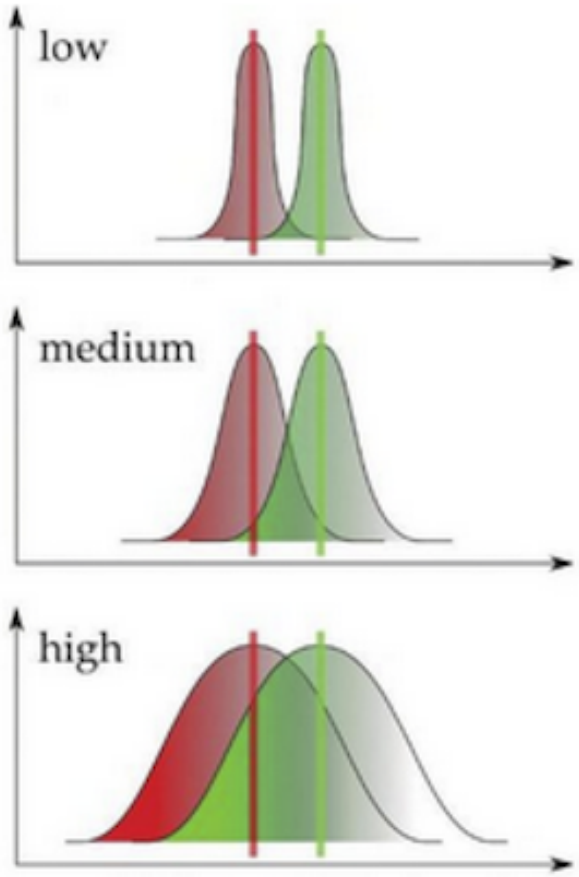
DE by Average Fold-Change (M)

- Simple fold-change rules give no assessment of statistical significance.
- Need to construct test statistics incorporating variability estimates (from replicates).



Variability and gene expression

- Simplest method, fold change, does not take gene variability into account.



Two-sample comparison, T-test

Let the mean and standard deviation expression levels for samples in two groups be

$\bar{x_i} = \frac{1}{n_i} \sum_{j=1}^{n_i} x_{ij}$, and $s_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \bar{x_i})^2$

The two-sample pooled t-statistics is given by

$$t = \frac{\bar{x_2} - \bar{x_1}}{s_p \sqrt{1/n_1 + 1/n_2}}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

is the pooled estimate of the standard deviation.

T-test assumptions

- Data must be independent random samples from their respective populations
- Sample size should either be large or, in the case of small sample sizes, the population distributions must be approximately normally distributed.
- When assumptions are not met, non-parametric alternatives are available (Wilcoxon Rank Sum/Mann-Whitney Test)

Welch's t-test

Does not assume equal variances for each group

$$t_g^{Welch} = \frac{\bar{y_{g}}1. - \bar{y_{g}}2.}{\sqrt{\frac{S_{g1}^2}{n_1} + \frac{S_{g2}^2}{n_2}}}$$

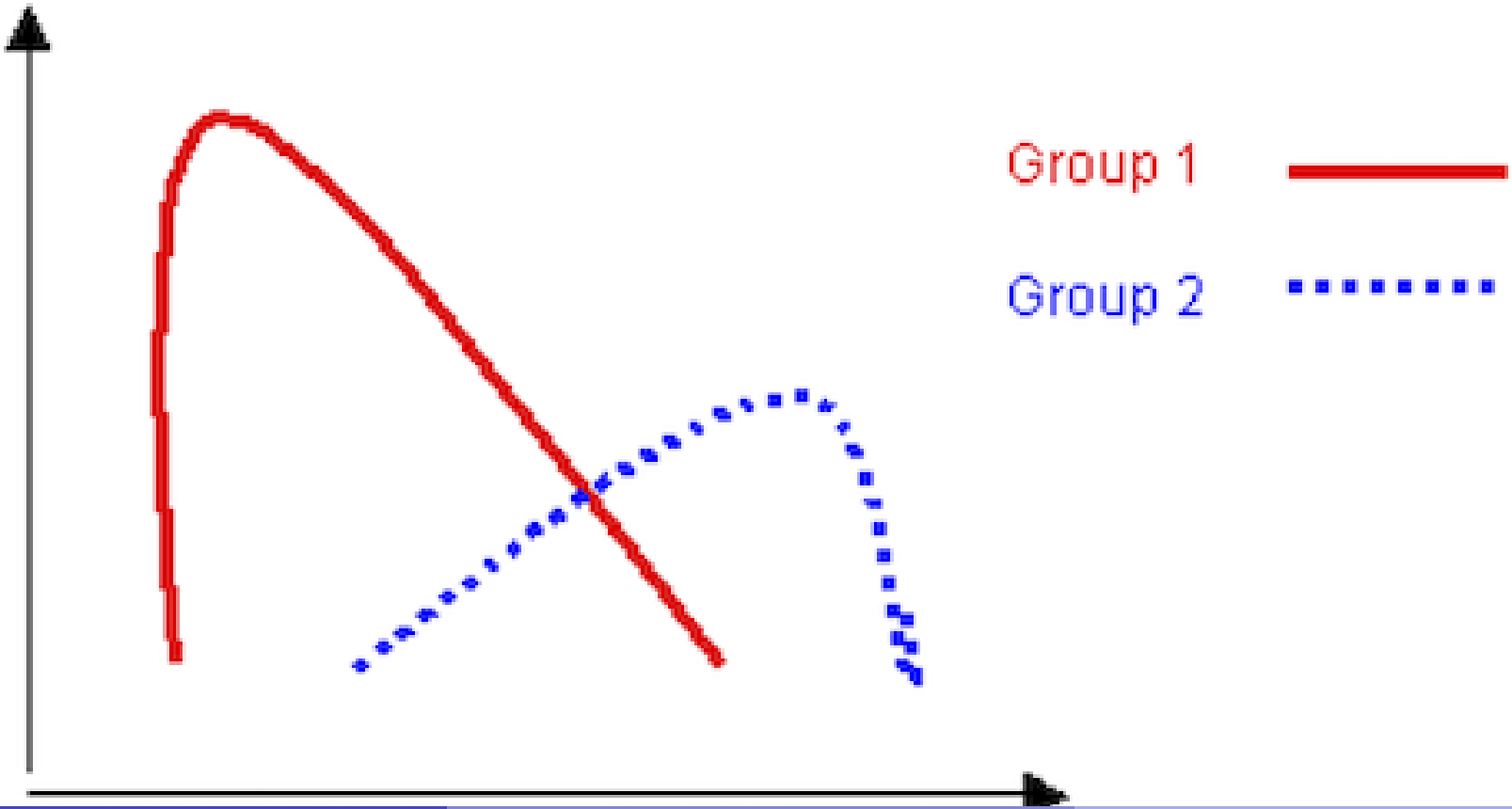
The variances S_{g1}^2 and S_{g2}^2 are then estimated independently in both groups for each gene

Non-parametric tests

- Non-normally distributed data
- More robust to outliers
- Less power
- Used when t-test assumptions cannot be met

Non-parametric tests

- **Mann-Whitney test (or Wilcoxon rank-sum test)**
  - differences in the sums of ranks between 2 populations
  - even if the medians are the same, there can be a statistically significant difference from the distribution of ranks



Anova: Analysis of Variance

Doing multiple two-sample t-tests would result in an increased chance of committing a Type I error.

For this reason, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.

Classes of ANOVA models

1. **Fixed-effects model**: a statistical model that represents observed quantities as non-random
2. **Random-effects model**: used when the treatments are not fixed
3. **Mixed model**: contains both fixed and random effects

Anova: Analysis of Variance

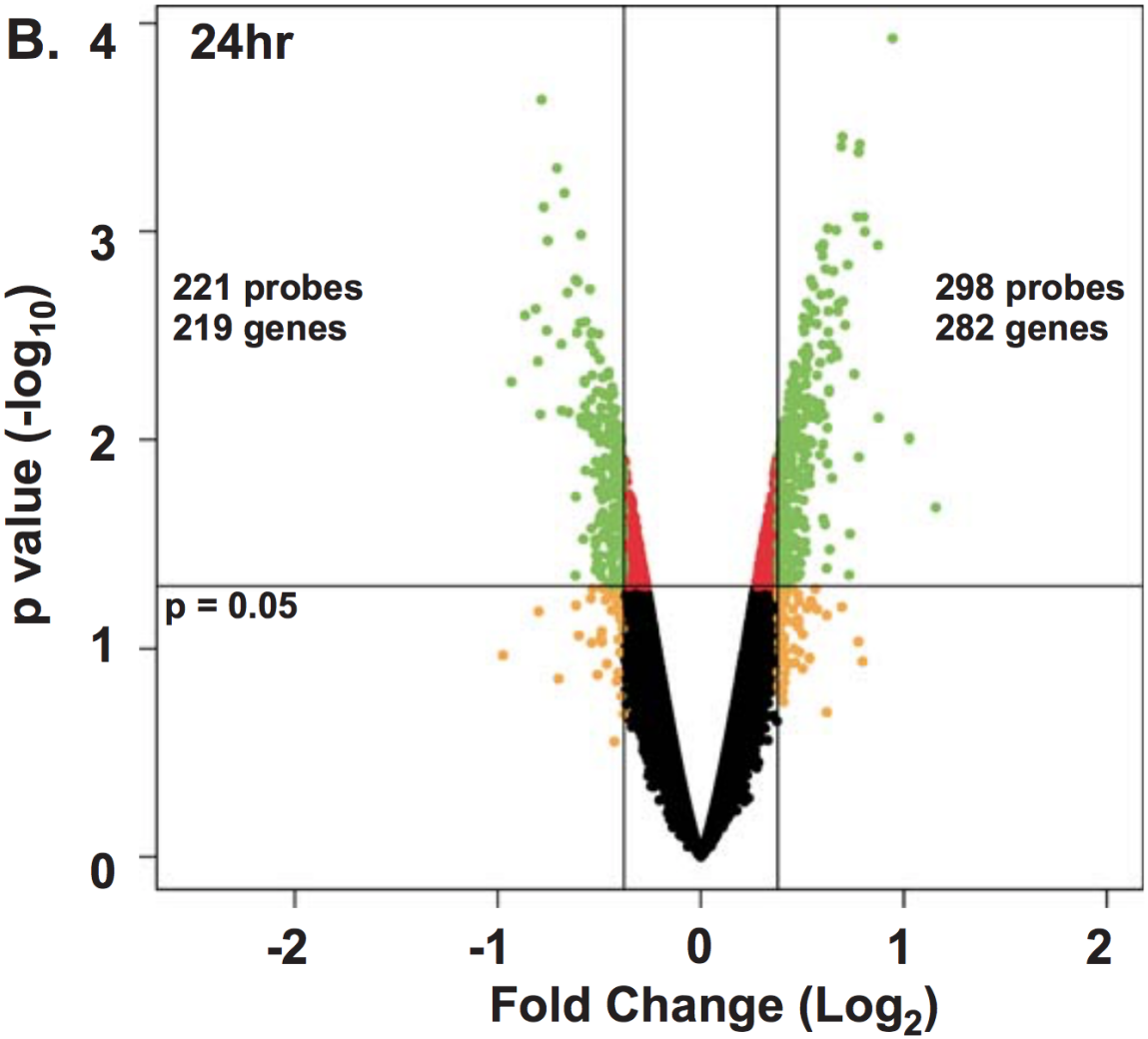
Common Designs and Tests

- **One-way ANOVA** is used to test for differences among two or more independent groups (means). When there are only two means to compare, the t-test and the ANOVA F-test are equivalent; the relation between ANOVA and t is given by $F = t^2$.
- **Factorial ANOVA** is used when the experimenter wants to study the interaction effects among the treatments.
- **Repeated measures ANOVA** is used when the same subjects are used for each treatment (e.g., in a longitudinal study).

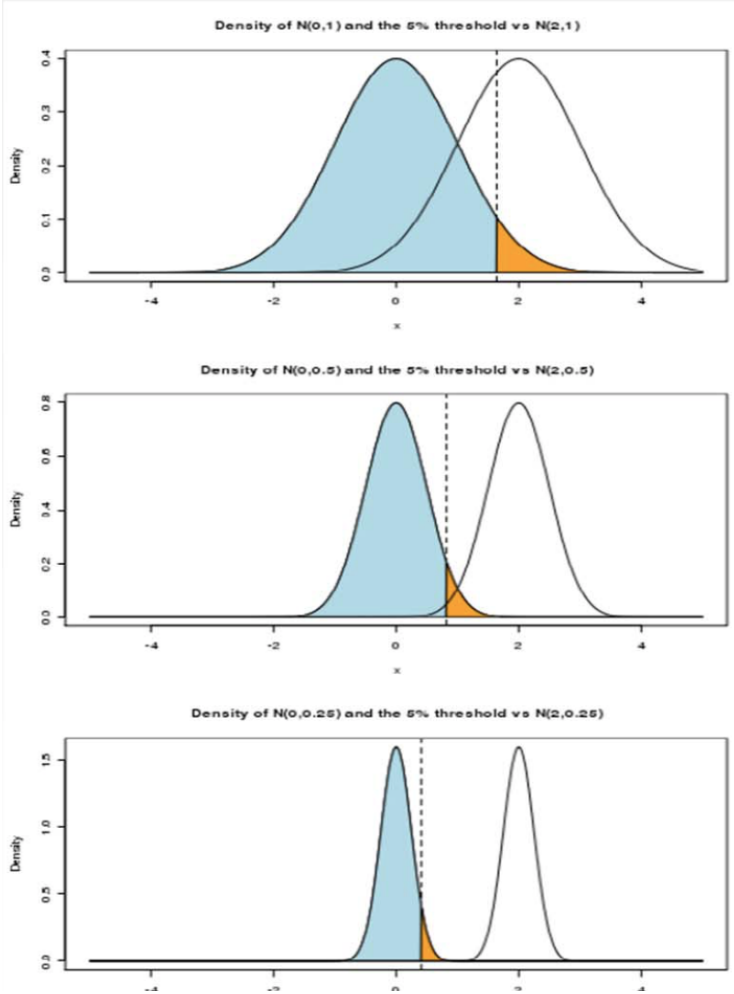
Volcano plot

- A diagnostic plot to visualize the test results.
- Scatter plot of the statistical significance (log p-values) vs. biological significance (log fold change).
- Ideally the two should agree with each other.

Volcano plot



Sample size and power calculations



- Effects of changing: variance, and mean difference (or effect size).
  Given a fixed sample size, n.
  - Larger variance $\rightarrow$ Lower Power
  - Smaller effect size $\rightarrow$ Lower Power
- *Given variance, effect size.*
  Increase sample size $\rightarrow$ Increase Power
- Can also change power by manipulating the "trade-off" between Type I and II error
  - Larger $\alpha \rightarrow$ Larger Power
  - Larger $\beta \rightarrow$ Lower Power

General framework for differential expression

- Linear models
- Model the expression of each gene as a linear function of explanatory variables (Groups, Treatments, Combinations of groups and treatments, Etc. . .)

$$y = X\beta + \epsilon$$

- y - vector of observed data
- X - design matrix
- β - vector of parameters to estimate

Example of a design matrix

Normal sample x 2

Cancer Sample x 2

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

β_1 = normal log-expression

β_2 = cancer – wt

$E[y_1]=E[y_2]=\beta_1$

$E[y_3]=E[y_4]=\beta_1+\beta_2$

Example of a design matrix

More examples

- 6 samples
- 2 groups + drug treatment
- Group and treatment effect are additive

$$y = X\beta + \epsilon$$

Group1	Group 2- Group 1	Drug dose
1	0	0.25
1	0	1
1	0	4
1	1	0.25
1	1	1
1	1	4

3 coefficients to estimate

Linear model parameter estimation

Model is specified – how do we find the coefficients? $y = X\beta + \epsilon$

- Minimize squared error $\epsilon'\epsilon = (Y - X\beta)'(Y - X\beta)$
- Take derivative $\frac{d}{d\beta}((Y - X\beta)'(Y - X\beta)) = -2X'(Y - X\beta)$
- Set to 0, $-2X'(Y - X\beta) = 0$
- Solve $X'Y = (X'X)\beta$ $\beta = (X'X)^{-1}X'Y$

Limma method

- Generalized the hierarchical model of Lonnstedt and Speed (2002) into a practical approach for general microarray experiments.
- The model borrows information across genes to smooth out variances and uses posterior variances in a classical t???test setting.
- Can be used to compare two or more groups.
- Can be used for multifactorial designs.
  - e.g. genotype and treatment.
- Uses empirical Bayes analysis to improve power in small sample sizes.

Limma method

Smyth et al. (2004) Statistical Applications in Genetics and Molecular Biology.

- Uses a Bayesian hierarchical model in multiple regression setting.
- Borrows in formation from all genes to estimate gene specific variances.
- As a result, variance estimates will be "shrunk" toward the mean of all variances. So very small variance scenarios will be alleviated.
- Implemented in Bioconductor package "limma".

http://bioinf.wehi.edu.au/limma/

https://bioconductor.org/packages/release/bioc/html/limma.html

Significance analysis of microarrays (SAM)

- A clever adaptation of the t-ratio to borrow information across genes.
- SAM seeks to control the proportion of false rejections among the set of rejected hypotheses (FDR).
- Permutation method is used to calculate the null distribution of the modified t-statistics.

V. G. Tusher et.al. "Significance Analysis of Microarrays Applied to the Ionizing Radiation Response" PNAS 2001
http://www.pnas.org/content/98/9/5116.long

http://www-stat.stanford.edu/~tibs/SAM/

Implemented as Bioconductor package `siggenes`, and Excel plugin.

SAM t-test

- With small sample sizes low and high variance can occur by chance.
- Variance depends on expression level.
- Try to remove (or minimize) the dependence of test statistics on variances (because small variance tend to lead to bigger test statistics).
- Solution: add a small constant to the denominator in calculating t statistics:

$$d_i = \frac{\bar{y}_i - \bar{x}_i}{s_i + s_0}$$

- $\bar{y}_i, \bar{x}_i$ - Means of two groups for gene i.
- s_i - Standard deviation for gene i, assuming equal variance in both groups.
- s_0 - "Exchangeability factor" estimated using all genes.

Multiple testing problem

- With thousands of genes on a microarray we're not testing one hypothesis, but many hypotheses – one for each gene.
- Analysis of 20,000 genes using commonly accepted significance level $\alpha = 0.05$ will identify 1,000 differentially expressed genes simply by chance.
- If probability of making an error in one test is 0.05, probability of making at least one error in ten tests is

$$(1 - (1 - 0.05)^{10}) = 0.40126$$

Naomi Altman & Martin Krzywinski "Points of significance: P values and the search for significance", Nat. Methods 2016, http://www.nature.com/nmeth/journal/v14/n1/full/nmeth.4120.html

False positive vs. False discovery rates

False positive rate is **the rate at which truly null genes are called significant**

$$FPR \approx \frac{false\ positives}{truly\ null} = \frac{V}{m_0}$$

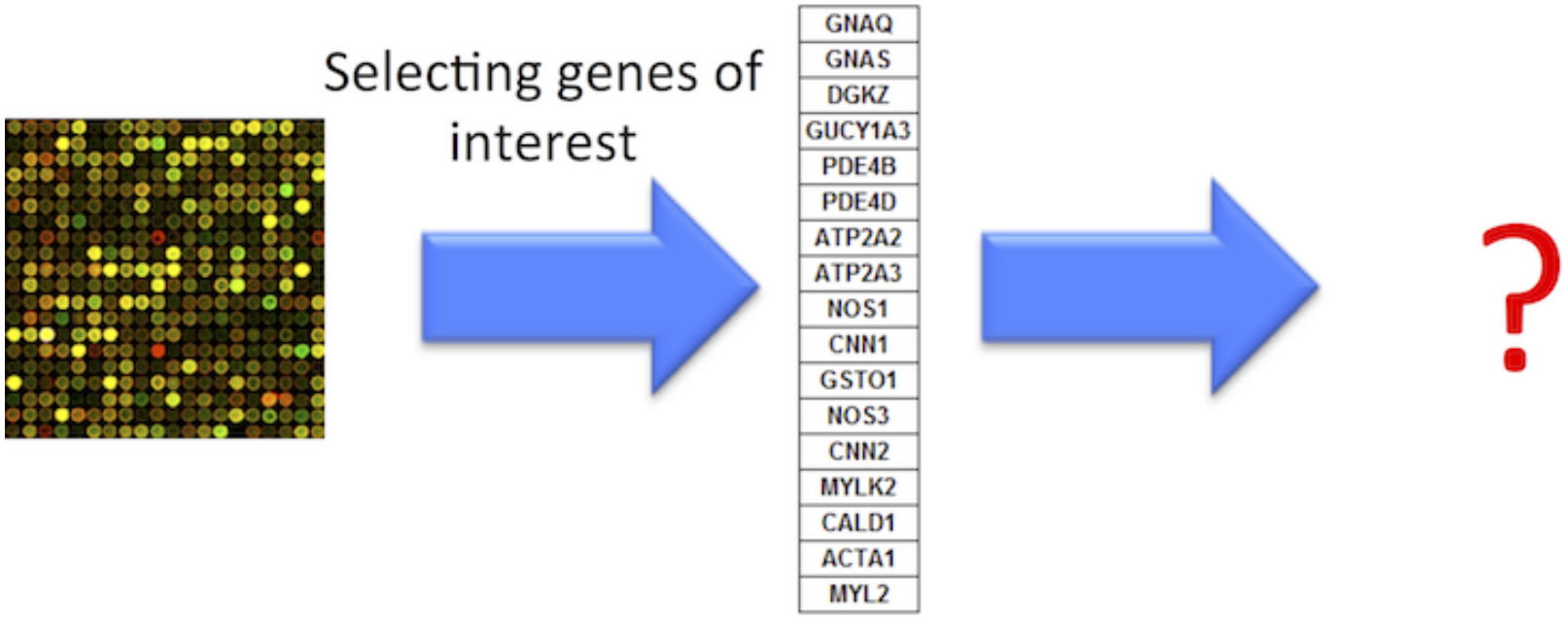
False discovery rate is **the rate at which significant genes are truly null**

$$FDR \approx \frac{false\ positives}{called\ significant} = \frac{V}{R}$$

Pathway enrichment analysis

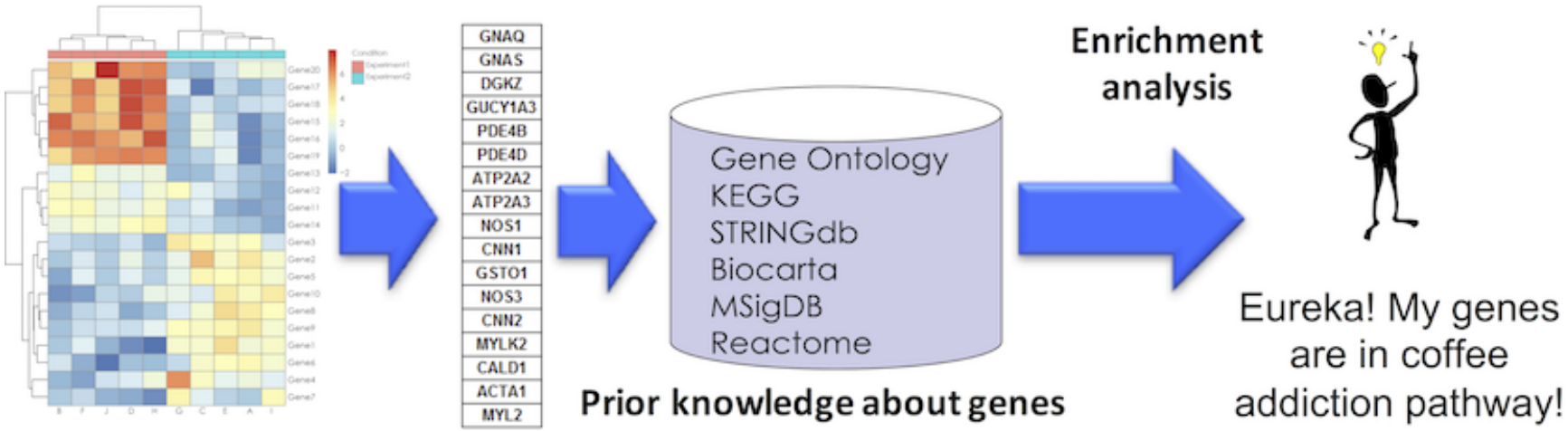
Why enrichment analysis?

- Human genome contains ~20,000-25,000 genes.
- Each gene has multiple functions.
- If 1,000 genes have changed in an experimental condition, it may be difficult to understand what they do.



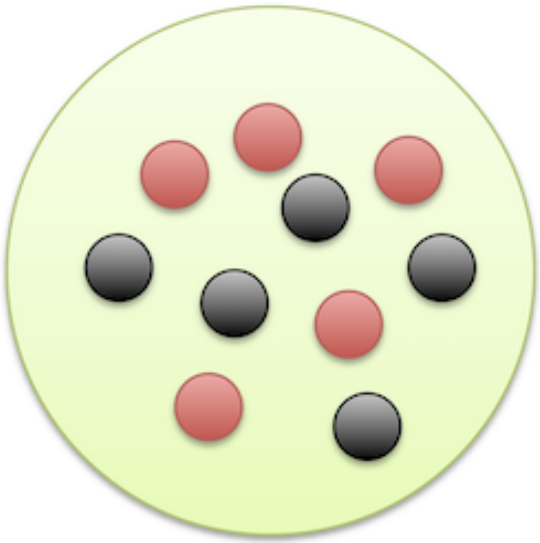
Why enrichment analysis?

- Translating changes of **hundreds/thousands of differentially expressed genes** into a few biological processes (reducing dimensionality).
- High level understanding of the biology behind gene expression – **Interpretation!**

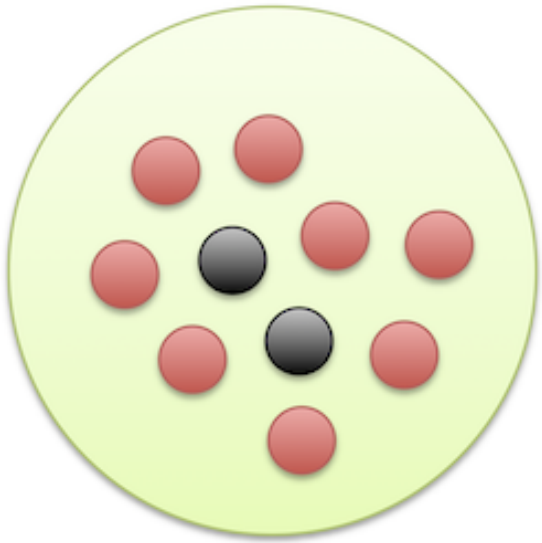


What is enrichment analysis? – statistical definition

Enrichment analysis – detection whether a group of objects has certain properties more (or less) frequent than can be expected by chance.



Jar 1



Jar 2

Classification of genes

Gene set - *a priori* classification of genes into biologically relevant groups (sets).

- Members of the same biochemical pathways.
- Genes annotated with the same molecular function.
- Transcripts expressed in the same cellular compartments.
- Co-regulated/co-expressed genes.
- Genes located on the same cytogenetic band.
- ...

Gene sets

- **Gene ontology** - provides controlled vocabularies of terms for the description of gene products.
  - **Molecular Function** - the tasks performed by individual gene products; examples are *transcription factor* and *DNA helicase*.
  - **Biological Process** - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by ordered assemblies of molecular functions.
  - **Cellular Component** - subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *origin recognition complex*.

Gene sets

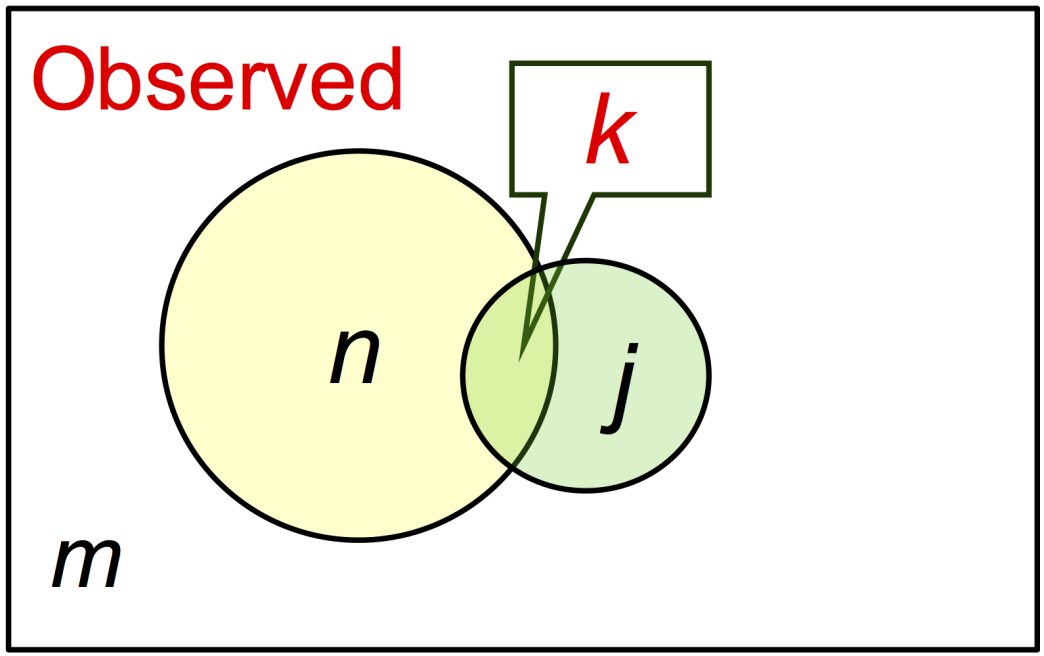
- **MSigDb** - Molecular Signatures Database
  - **H**, **hallmark gene sets** are coherently expressed signatures derived by aggregating many MSigDB gene sets to represent well-defined biological states or processes.
  - **C1**, **positional gene sets** for each human chromosome and cytogenetic band.
  - **C2**, **curated gene sets** from online pathway databases, publications in PubMed, and knowledge of domain experts.
  - **C3**, **motif gene sets** based on conserved *cis*-regulatory motifs from a comparative analysis of the human, mouse, rat, and dog genomes.
  - **C4**, **computational gene sets** defined by mining large collections of cancer-oriented microarray data.
  - **C5**, **GO gene sets** consist of genes annotated by the same GO terms.
  - **C6**, **oncogenic signatures** defined directly from microarray gene expression data from cancer gene perturbations.
  - **C7**, **immunologic signatures** defined directly from microarray gene expression data from immunologic studies.

Gene sets

- **KEGG: Kyoto Encyclopedia of Genes and Genomes** - a collection of biological information compiled from published material = curated database. https://www.genome.jp/kegg/
- **Reactome** - curated human pathways encompassing metabolism, signaling, and other biological processes. https://reactome.org/

Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category



Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

The expected value of k would be $k_e = (n/m) * j$.

If $k > k_e$, functional category is said to be enriched, with a ratio of enrichment $r = k/k_e$

Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

	Diff. exp. genes	Not Diff. exp. genes	Total
In gene set	k	j-k	j
Not in gene set	n-k	m-n-j+k	m-j
Total	n	m-n	m

Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

What is the probability of having k or more genes from the category in the selected n genes?

$$P = \sum_{i=k}^{n} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Overrepresentation analysis, Hypergeometric test

- m is the total number of genes
- j is the number of genes are in the functional category
- n is the number of differentially expressed genes
- k is the number of differentially expressed genes in the category

$k < (n/m) * j$ - underrepresentation. Probability of k or less genes from the category in the selected n genes?

$$P = \sum_{i=0}^{k} \frac{\binom{m-j}{n-i}\binom{j}{i}}{\binom{m}{n}}$$

Overrepresentation analysis, Hypergeometric test

1. Find a set of differentially expressed genes (DEGs)
2. Are *DEGs in a set* more common than *DEGs not in a set*?

- Fisher test `stats::fisher.test()`
- Conditional hypergeometric test, to account for directed hierachy of GO `GOstats::hyperGTest()`

Example: https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

GSEA: Gene set enrichment analysis

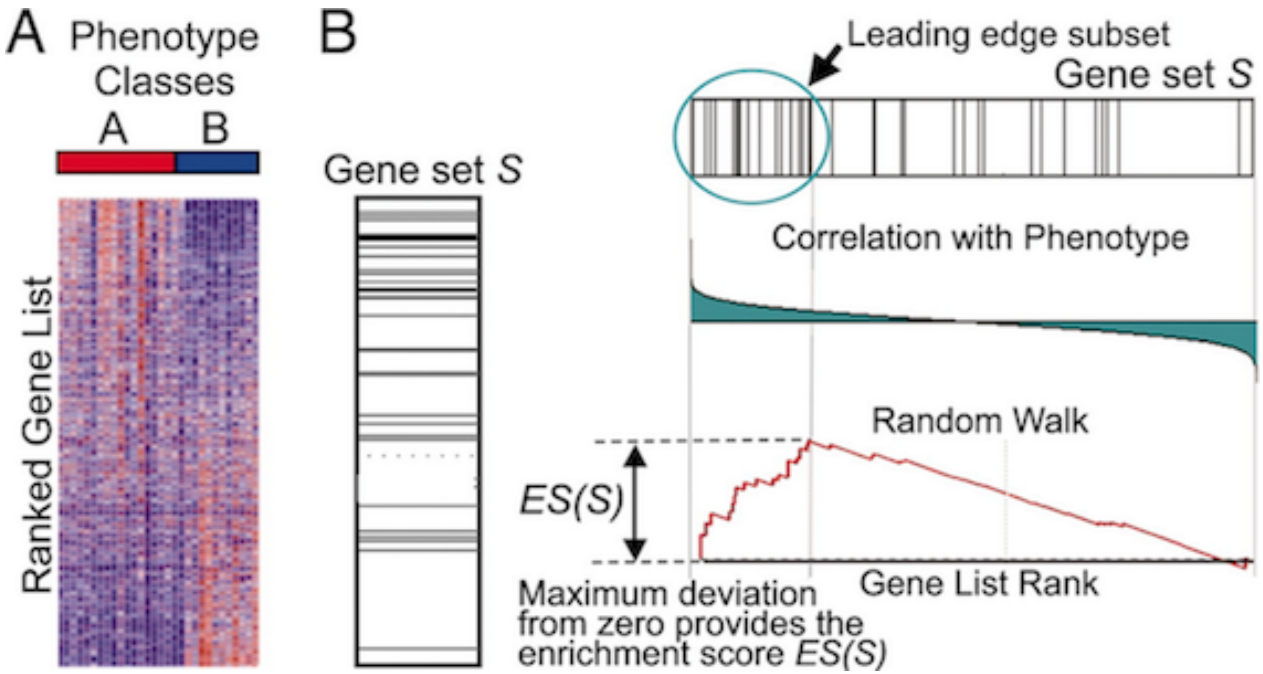
- **Gene set analysis (GSA)**. Mootha et al., 2003; modified by Subramanian, et al. "**Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles.**" PNAS 2005 http://www.pnas.org/content/102/43/15545.abstract
- Main rationale – functionally related genes often display a coordinated expression to accomplish their roles in the cells.
- Aims to identify gene sets with "subtle but coordinated" expression changes that would be missed by DEGs threshold selection.

GSEA: Gene set enrichment analysis

- The null hypothesis is that the **rank ordering** of the genes in a given comparison is **random** with regard to the case-control assignment.
- The alternative hypothesis is that the **rank ordering** of genes sharing functional/pathway membership is **associated** with the case-control assignment.

GSEA: Gene set enrichment analysis

1. Sort genes by log fold change
2. Calculate running sum - increment when gene in a set, decrement when not
3. Maximum of the runnig sum is the enrichment score - larger means genes in a set are toward top of the sorted list
4. Permute subject labels to calculate significance p-value

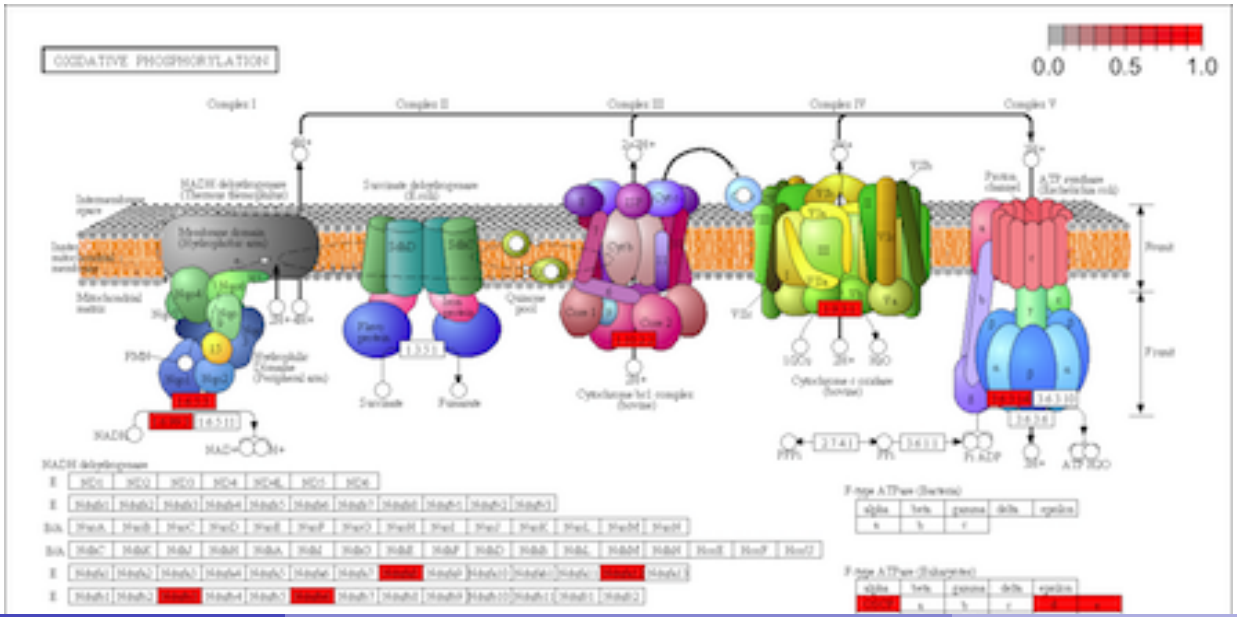


Gene set enrichment analysis | DIY

- **clusterProfiler** (https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html) - statistical analysis and visualization of functional profiles for genes and gene clusters.
- **limma** (https://bioconductor.org/packages/release/bioc/html/limma.html) - Linear Models for Microarray Data, includes functional enrichment functions `goana`, `camera`, `roast`, `romer`.
- **GOstats** (https://www.bioconductor.org/packages/2.8/bioc/html/GOstats.html) - tools for manimpuating GO and pathway enrichment analyses. https://github.com/mdozmorov/MDmisc/blob/master/R/gene_enrichment.R

Gene annotation databases

- **annotables** (https://github.com/stephenturner/annotables) - R data package for annotating/converting Gene IDs.
- **msigdf** (https://github.com/stephenturner/msigdf) - Molecular Signatures Database (MSigDB) in a data frame.
- **pathview** (https://www.bioconductor.org/packages/devel/bioc/html/pathview.html) - a tool set for pathway based data integration and visualization.



References | Microarrays

- Nguyen, Danh V., A. Bulak Arpat, Naisyin Wang, and Raymond J. Carroll. "DNA Microarray Experiments: Biological and Technological Aspects." Biometrics 58, no. 4 (December 2002): 701–17. - Full microarray technology description. Very thorough and comprehensive.
- Tilstone, Claire. "DNA Microarrays: Vital Statistics." Nature 424, no. 6949 (August 7, 2003): 610–12. doi:10.1038/424610a. - Microarrays and statisticians, the importance of analysis
- Tumor Analysis Best Practices Working Group. "Expression Profiling–Best Practices for Data Generation and Interpretation in Clinical Trials." Nature Reviews. Genetics 5, no. 3 (March 2004): 229–37. doi:10.1038/nrg1297. - Microarray overview. Best practices for technology and analysis.

References | QC

- Fan, Jianqing, and Yi Ren. "Statistical Analysis of DNA Microarray Data in Cancer Research." Clinical Cancer Research: An Official Journal of the American Association for Cancer Research 12, no. 15 (August 1, 2006): 4469–73. doi:10.1158/1078-0432.CCR-06-1033. - Steps in microarray data analysis, from preprocessing to differential expression and time course. Brief.
- Bolstad, B. M., R. A. Irizarry, M. Astrand, and T. P. Speed. "A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias." Bioinformatics (Oxford, England) 19, no. 2 (January 22, 2003): 185–93. - Normalization methods description.
- Carvalho, Benilton S., and Rafael A. Irizarry. "A Framework for Oligonucleotide Microarray Preprocessing." Bioinformatics (Oxford, England) 26, no. 19 (October 1, 2010): 2363–67. doi:10.1093/bioinformatics/btq431. - Preprocessing for different microarray types - Affy, Illumina, Nimblegen - , and platforms - SNP, Exon, Expression, Tiling. Probe affinity effect figure

References | Differential expression

- Krzywinski, Martin, and Naomi Altman. "Points of Significance: Comparing Samples—part I." Nature Methods 11, no. 3 (March 2014): 215–16.
- Cui, Xiangqin, and Gary A. Churchill. "Statistical Tests for Differential Expression in CDNA Microarray Experiments." Genome Biology 4, no. 4 (2003): 210. - Differential expression analysis of microarrays, from fold-change to t-test, its moderated versions SAM and limma, and ANOVA.
- Krzywinski, Martin, and Naomi Altman. "Points of Significance: Analysis of Variance and Blocking." Nature Methods 11, no. 7 (July 2014): 699–700.
- Krzywinski, Martin, and Naomi Altman. "Points of Significance: Power and Sample Size." Nature Methods 10, no. 12 (November 26, 2013): 1139–40. doi:10.1038/nmeth.2738.
- Tong, Tiejun, and Hongyu Zhao. "Practical Guidelines for Assessing Power and False Discovery Rate for a Fixed Sample Size in Microarray Experiments." Statistics in Medicine 27, no. 11 (May 20, 2008): 1960–72. doi:10.1002/sim.3237. Power analysis. t-statistics, FDR types and definitions, then derivation of power calculations.
- Leek, Jeffrey T., Robert B. Scharpf, H??ctor Corrada Bravo, David Simcha, Benjamin Langmead, W. Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A. Irizarry. "Tackling the Widespread and Critical Impact of Batch Effects in High-Throughput Data." Nature Reviews. Genetics 11, no. 10 (October 2010): 733–39. https://doi.org/10.1038/nrg2825. - Batch effect, types, sources, examples of wrong conclusions, SVA and ComBat methods

References | Functional enrichment

- Bard, Jonathan B. L., and Seung Y. Rhee. "Ontologies in Biology: Design, Applications and Future Challenges." Nature Reviews. Genetics 5, no. 3 (March 2004): 213–22. doi:10.1038/nrg1295. - Ontologies review
- Ackermann, Marit, and Korbinian Strimmer. "A General Modular Framework for Gene Set Enrichment Analysis." BMC Bioinformatics 10, no. 1 (2009): 47. doi:10.1186/1471-2105-10-47. - All steps for enrichment analysis, methods, statistics, GSEA.
- Efron, Bradley, and Robert Tibshirani. "On Testing the Significance of Sets of Genes." The Annals of Applied Statistics, 2007, 107–29. - maxmean statistics for enrichment analysis. Comparison with GSEA.
- Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Bioinformatics Enrichment Tools: Paths toward the Comprehensive Functional Analysis of Large Gene Lists." Nucleic Acids Research 37, no. 1 (January 2009): 1–13. doi:10.1093/nar/gkn923. - Gene enrichment analyses tools. Statistics, concept of background. 68 tools, table
- Hung, Jui-Hung, Tun-Hsiang Yang, Zhenjun Hu, Zhiping Weng, and Charles DeLisi. "Gene Set Enrichment Analysis: Performance Evaluation and Usage Guidelines." Briefings in Bioinformatics 13, no. 3 (May 2012): 281–91. doi:10.1093/bib/bbr049. - Details of GSEA. Statistics, correction for multiple testing. Lack of gold standard - concept of mutual coverage.

Thank you

This lecture, including PDF, is on GitHub:
https://github.com/mdozmorov/microarray_overview

Questions?