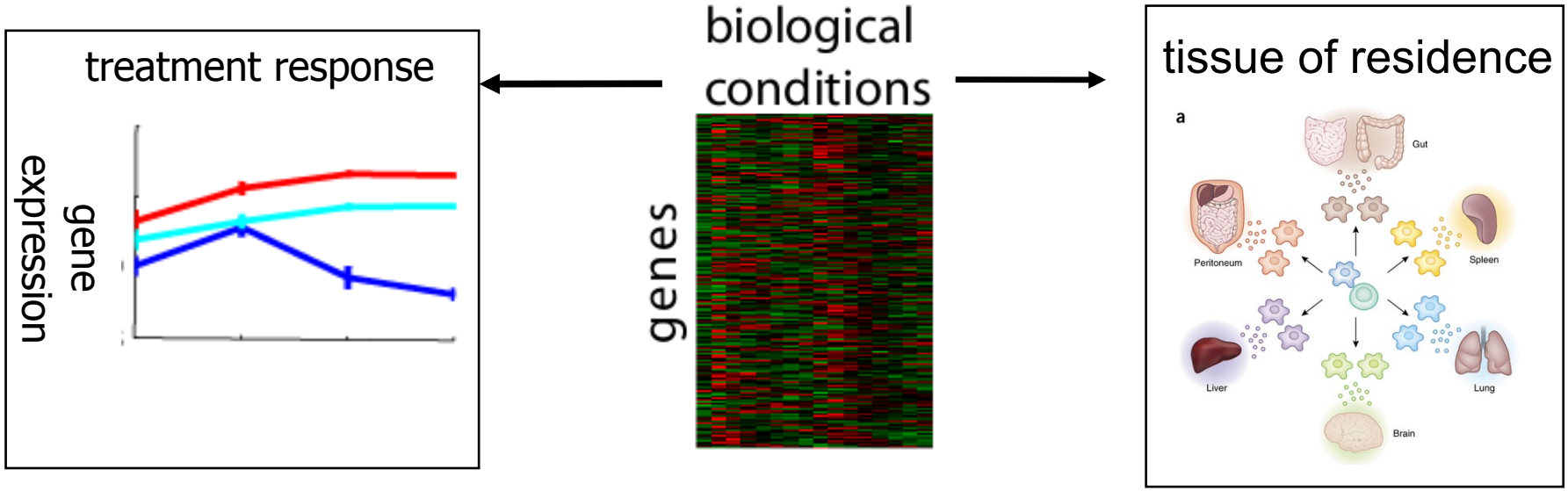
EASL Course
Bioinformatics of microarray analysis

Ivan G. Costa, Joseph Kuo, Oliver Krenkel
IZFK Research Group Bioinformatics
RWTH Aachen
www.costalab.org

AICES

IZKF Interdisziplinäres Zentrum für Klinische Forschung | RWTH AACHEN UNIVERSITY

Analysis of Gene Expression



adapted from: Amit et al. 2016

1- Which genes are up/down regulated after treatment?

- differential analysis / clustering genes

2 - Which cells are more similar?

- clustering samples / PCA

3 - How to interpret large lists of genes?

- gene ontology enrichment /gene set enrichment analysis (GSEA)

Objective of the course

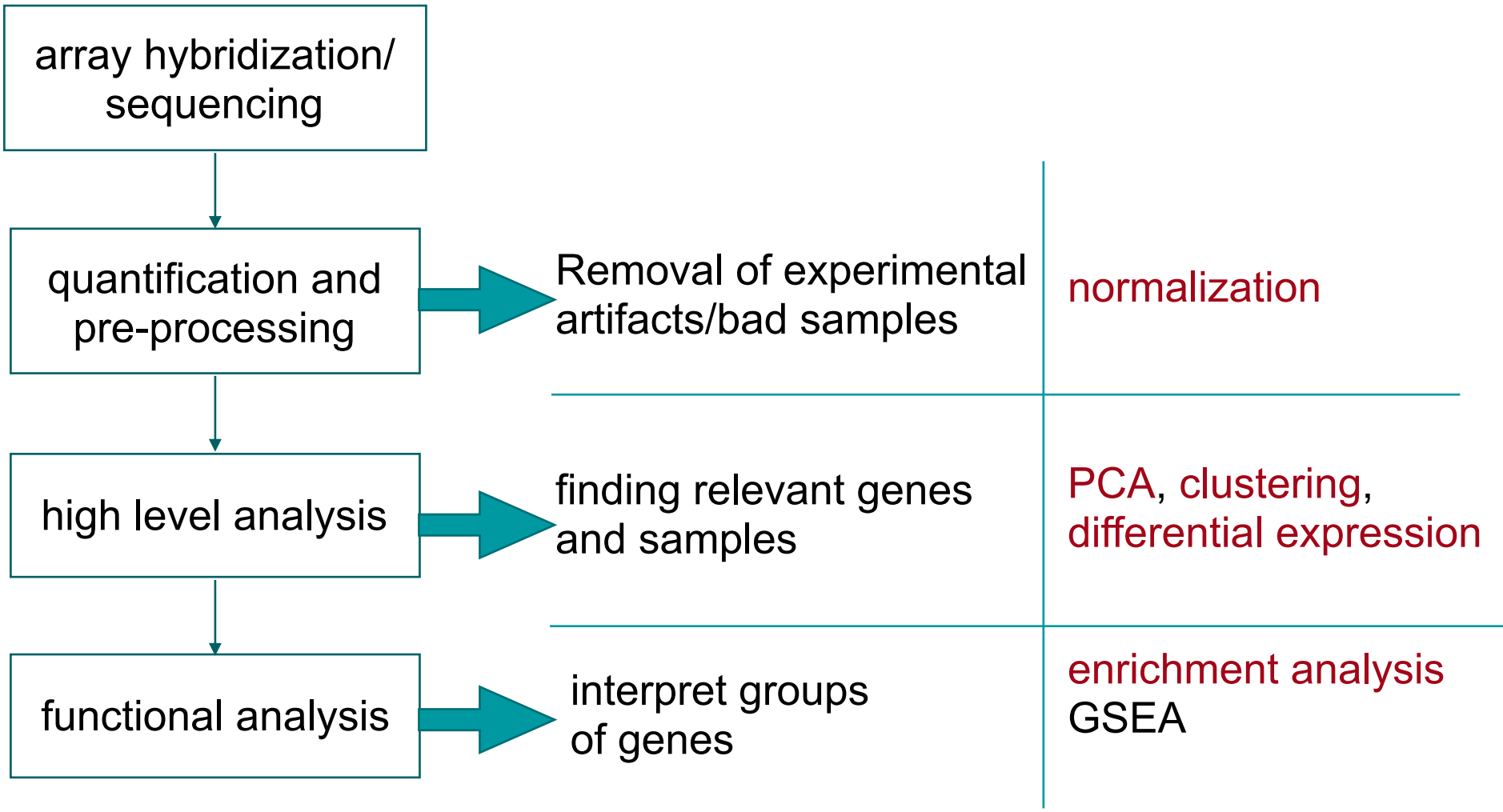
1 - Give you a (very) brief overview on the use of R/ bioconductor tools

2 - Show a real example with all steps necessary for gene expression analysis (based on arrays)

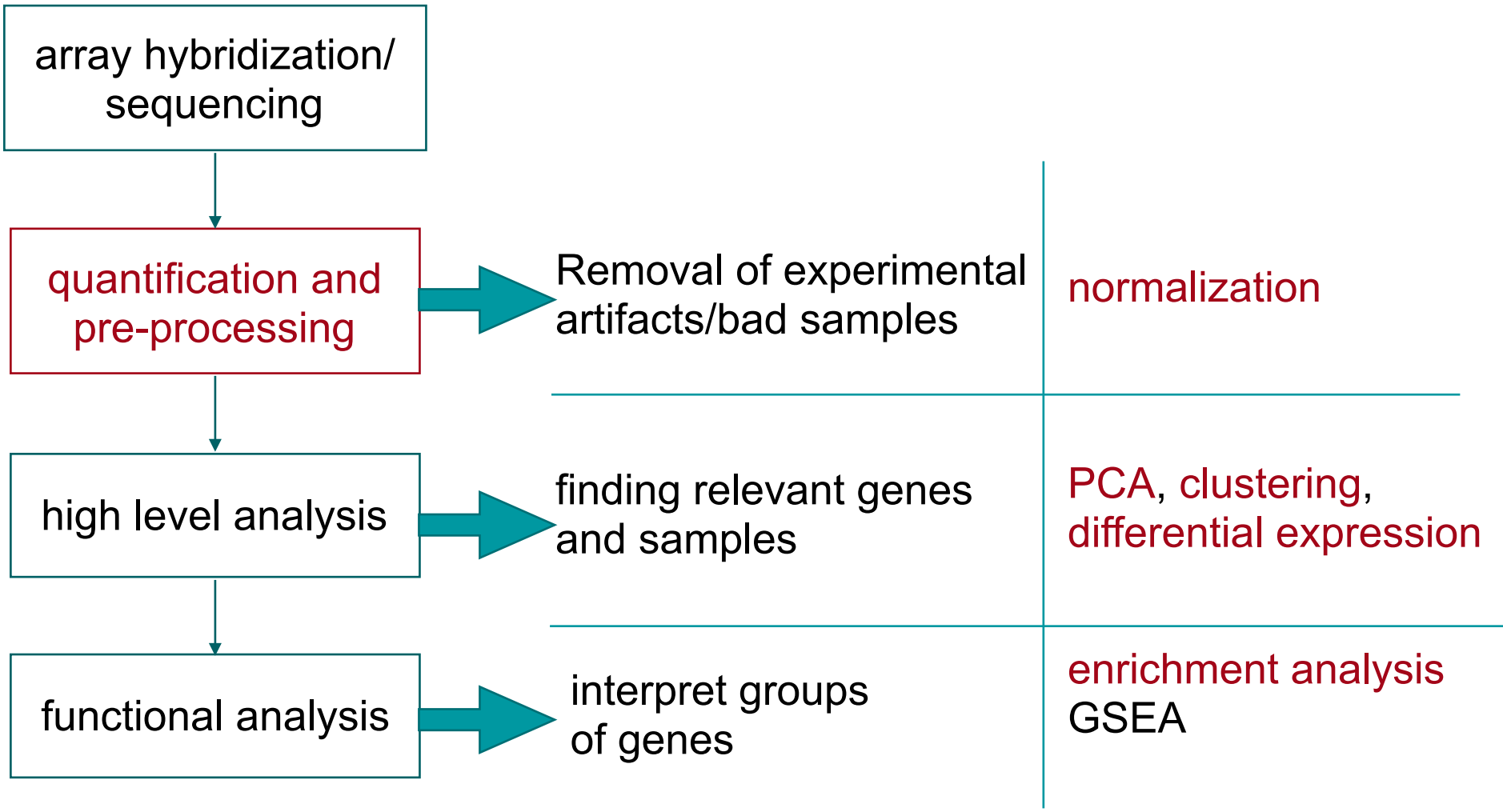
3 - Why arrays? Analysis of sequencing data is still complex / require command line “programming”.

However, high level analysis are the same!

Bioinformatics - Gene Expression Analysis

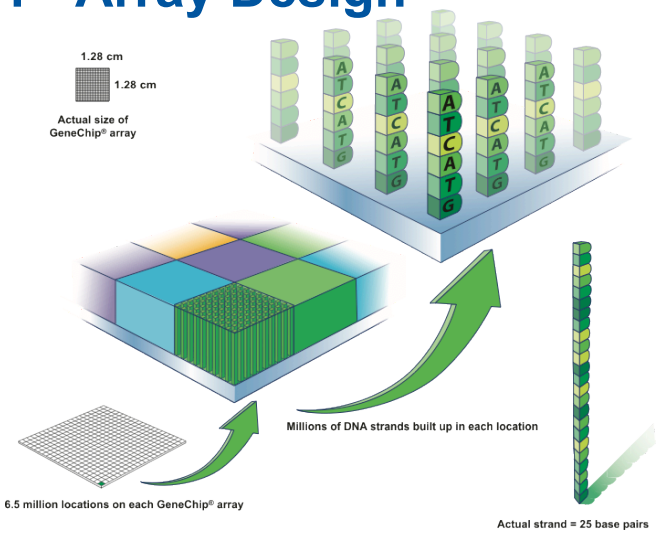


Bioinformatics - Gene Expression Analysis

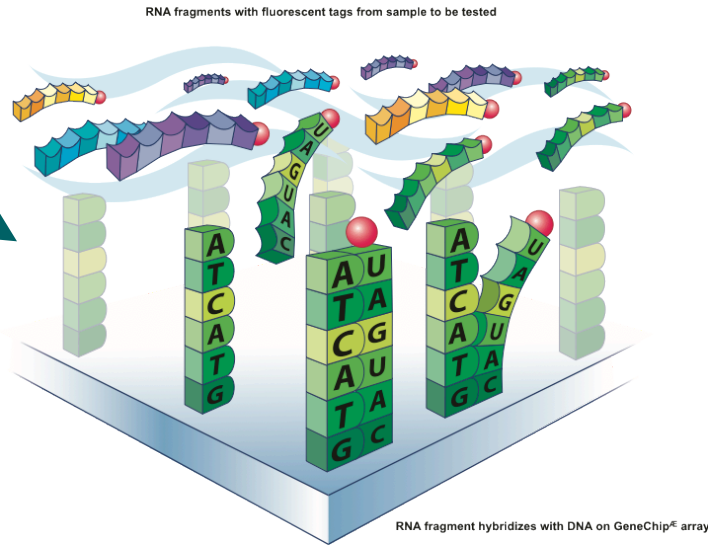


Affymetrix Arrays - Example

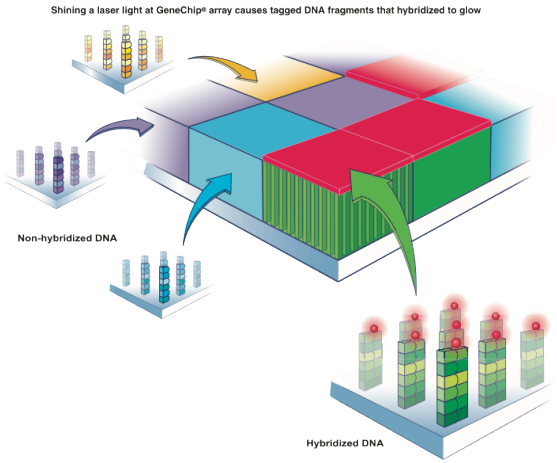
1 - Array Design



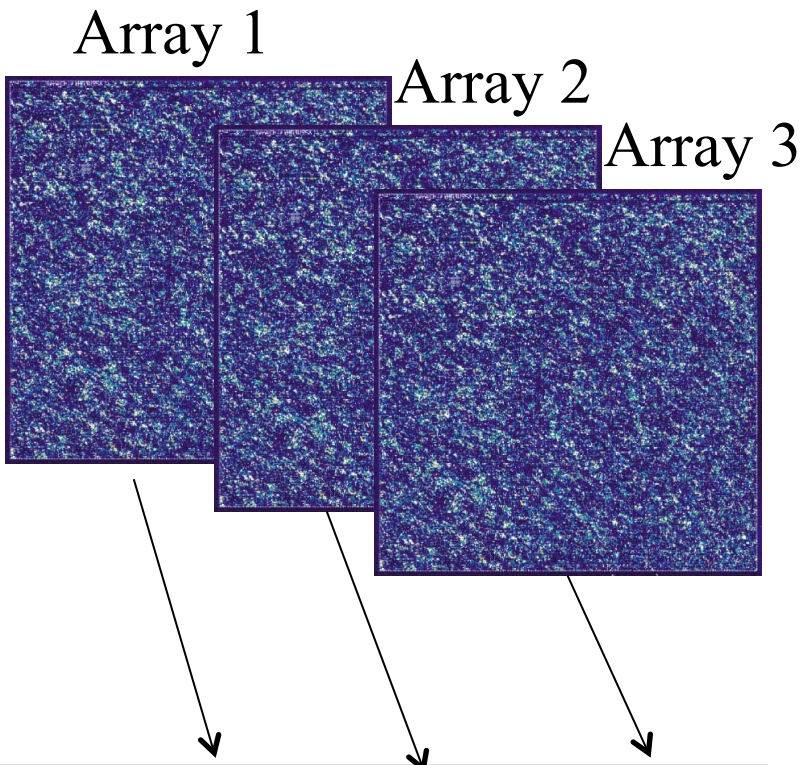
2 - cDNA Hybridization



3 -Quantification



Quantification/Pre-processing



	Array 1	Array 2	Array 3
Gene 1	100	200	500
Gene 2	3000	5000	10000
Gene 3	50	10	100
...	...	...	

1 - Quantify gene expression values

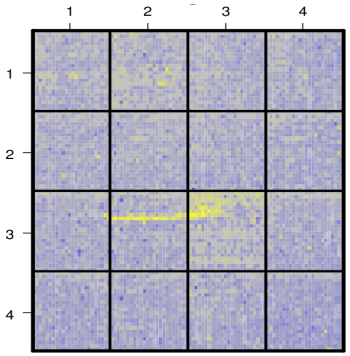
2 - Quality Control
- remove bad samples

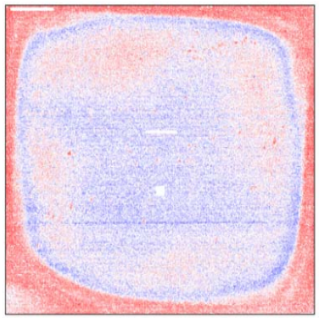
3 - Correct for Experimental artifacts
- normalization

Why is QC / Normalization important?

- Systematic errors (array wise)
  - labeling efficiency, scanning parameters, reverse transcriptase, batch effects
- Stochastic errors
  - cross-hybridization, image processing failure, error on probe sequence (manufacturer defect) (gene wise)
  - dust in array, hybridization problems (array wise)

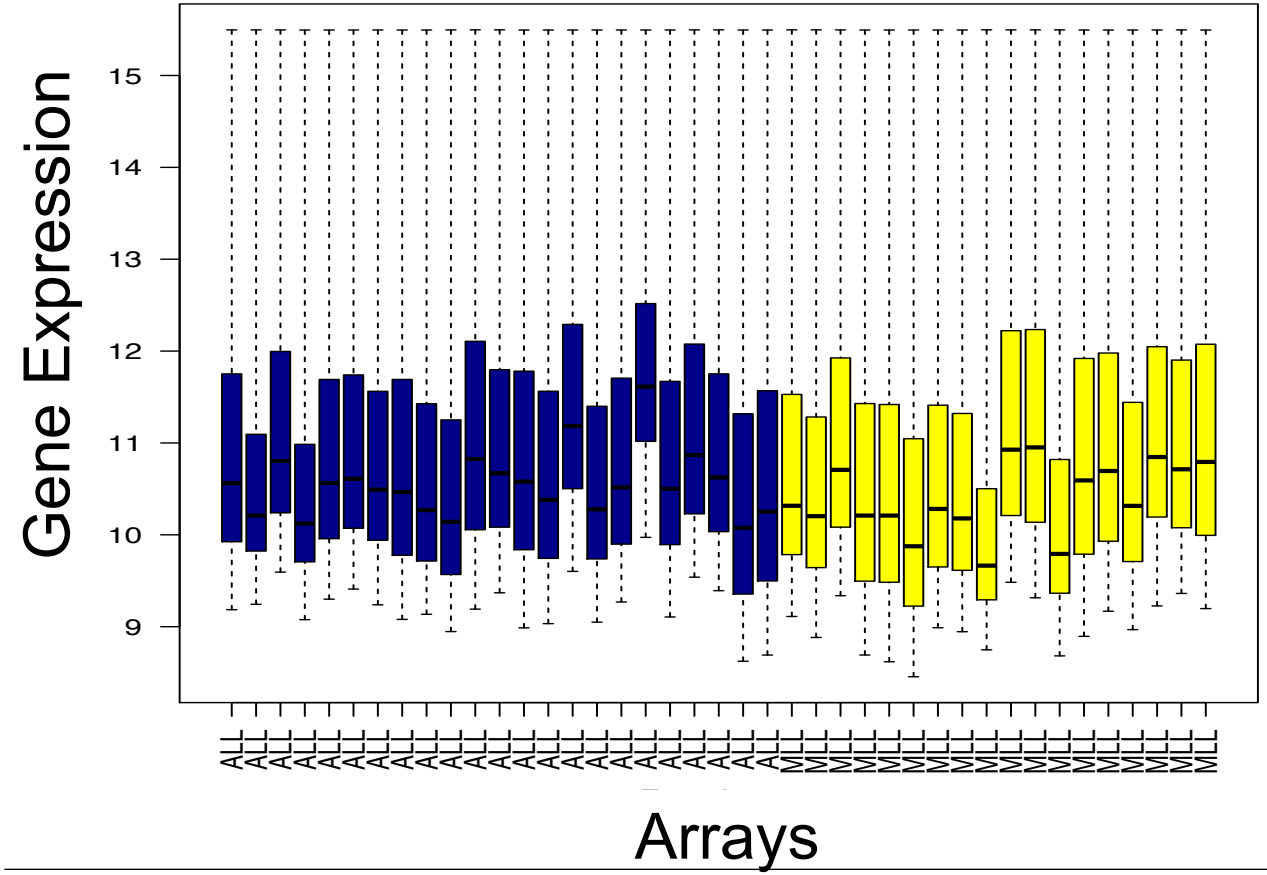
Example of Hybridization Problems

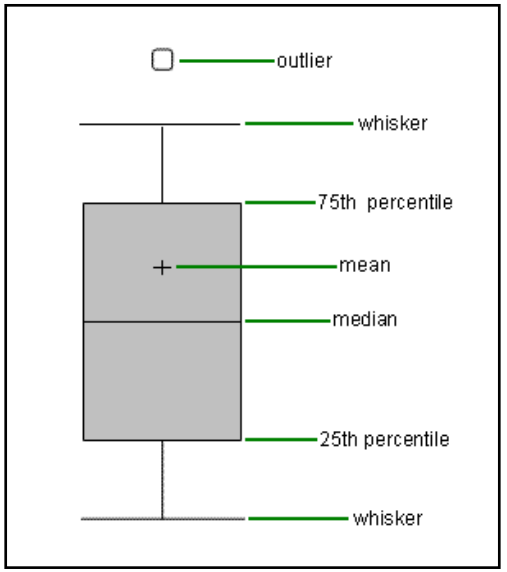




Normalization Principles

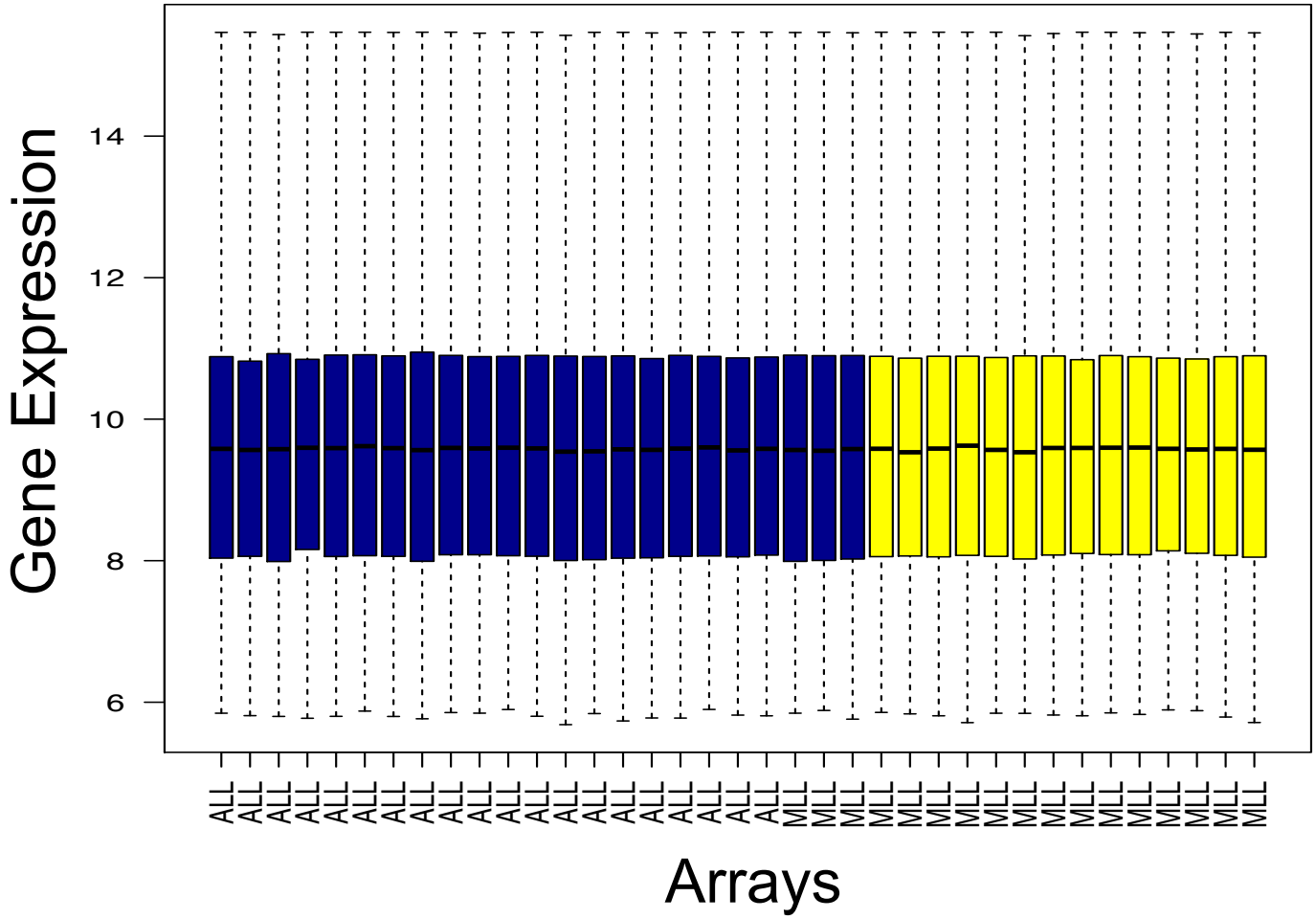
1 - Most genes don't change expression -> small/same variance
2 - Arrays are hybridized with the same amount of DNA -> same mean





Normalization Results

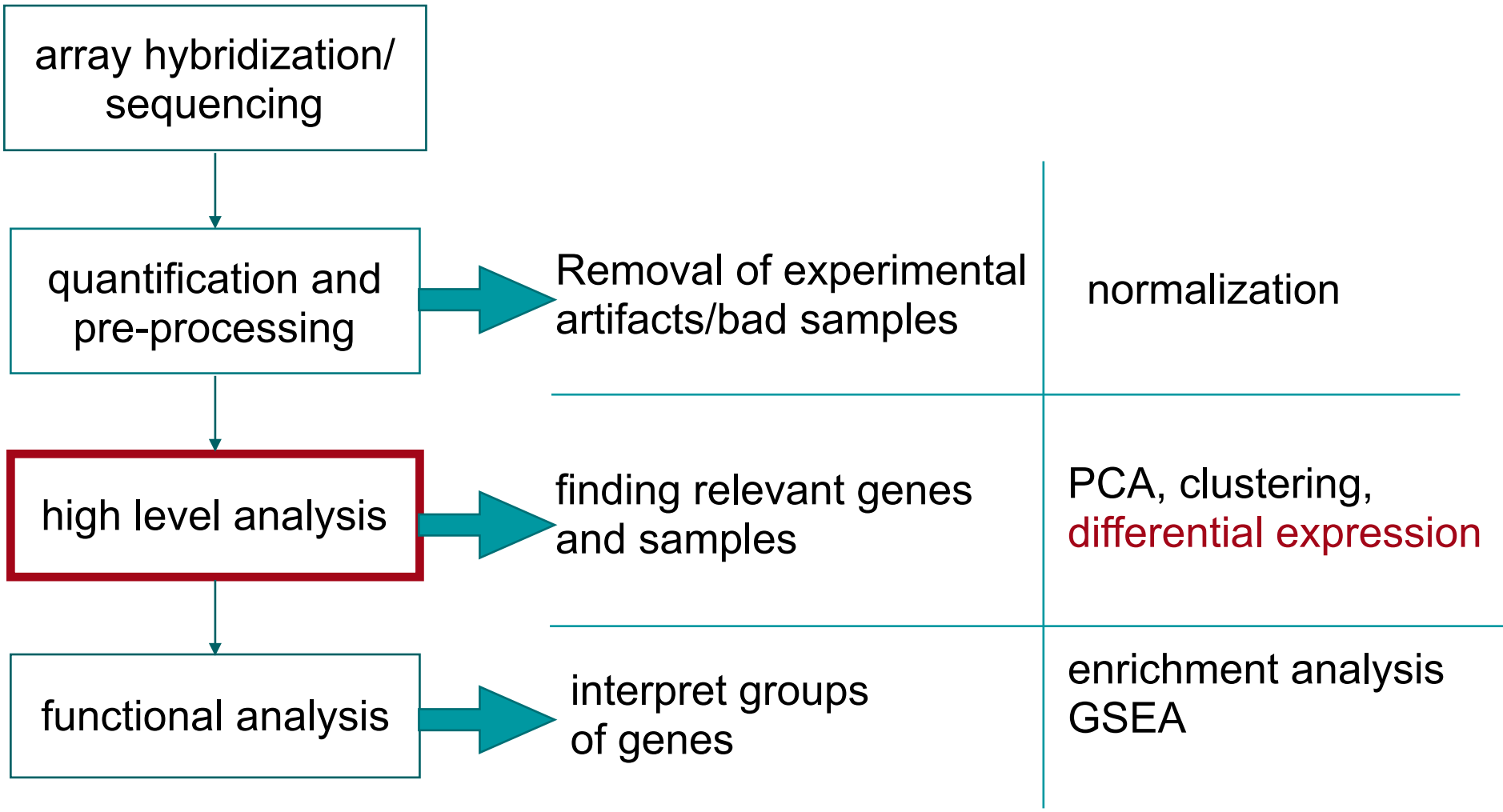
Application of BetweenArray normalization from limma package



Quantification/Pre-processing - Resume

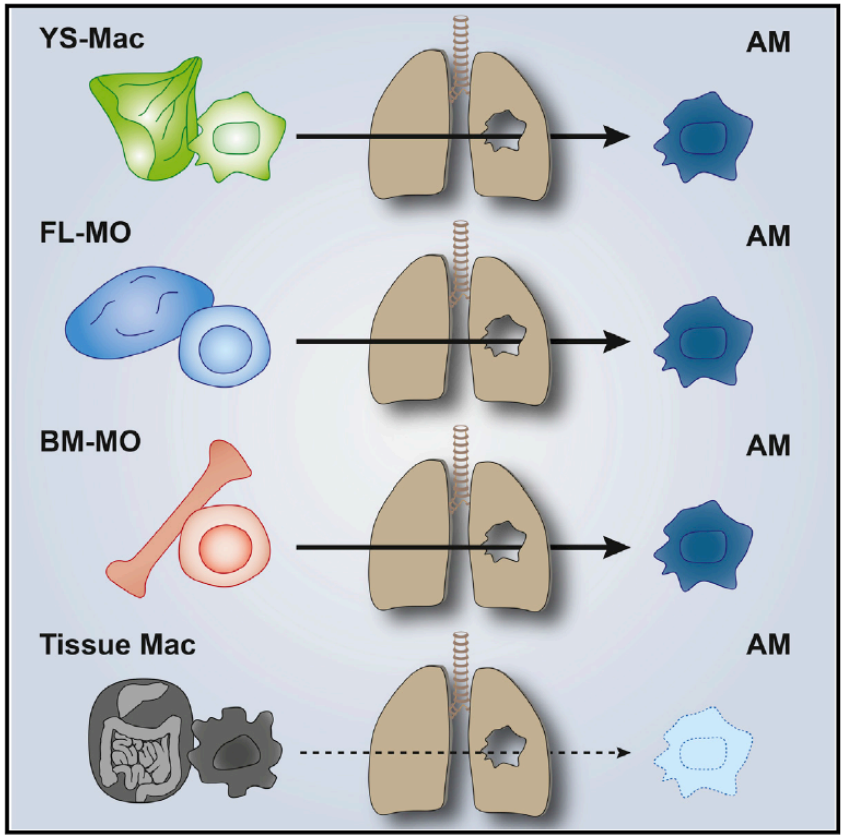
- Normalization is important to confirm the quality and consistency of data
- Boxplots hould also be performed after all steps to assure data standards
- Exclusion of “bad samples” has positive effect on downstream analysis
- **In doubt, consult a bioinformatician!**

Bioinformatics - Gene Expression Analysis



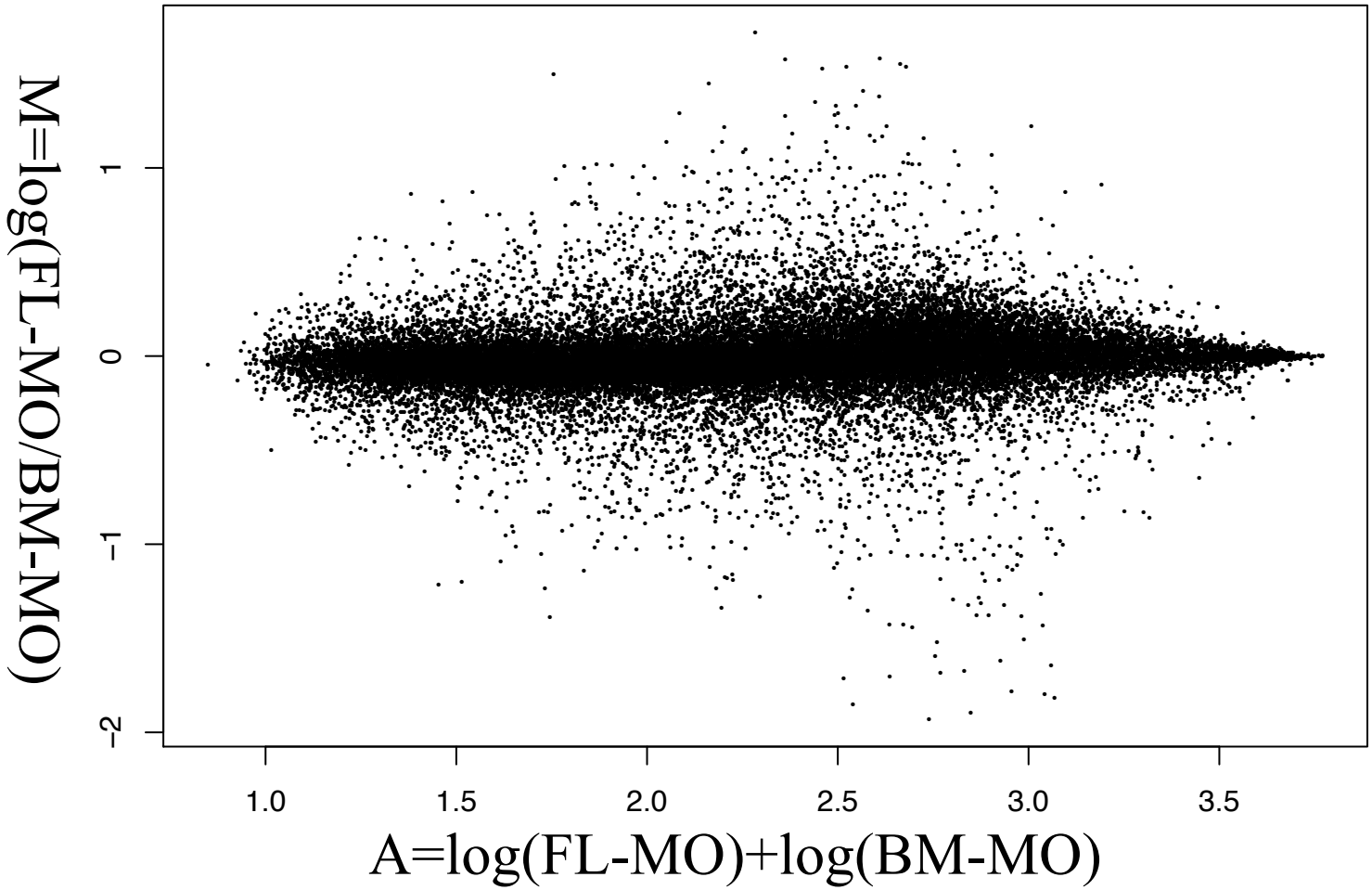
Differential Expression Analysis

- Identify genes related to a particular condition
  - example - van de Laar, et al. 2016, Immunity, 2016.
- We will consider:
  - You Sac Macrophages (YS-Mac)
  - Fetal Liver Monocytes (FL-MO)
  - Bone Marrow Monocytes (BM-MO)
    - 4 replicates per condition



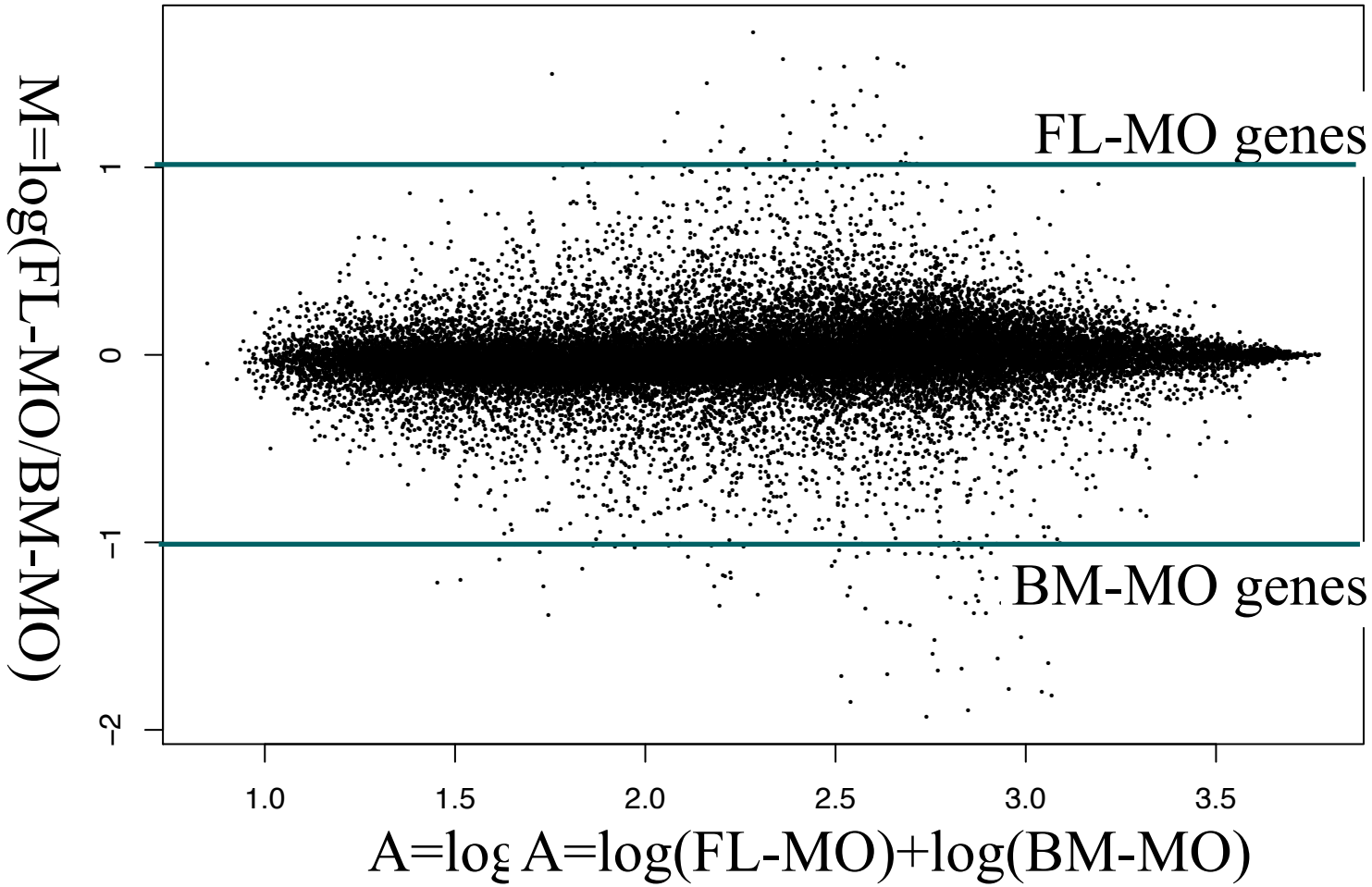
Source: van de Laar, et al. 2016, Immunity, 2016.

Differential Expression - Example



Differential Expression - Example

- Fold change analysis - change > |log2(2)|

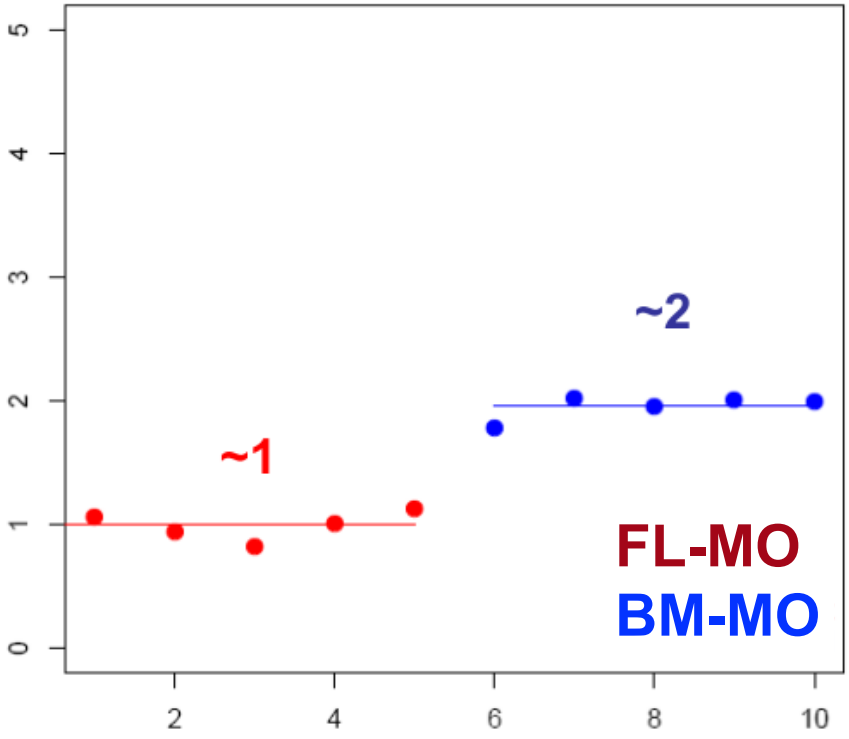


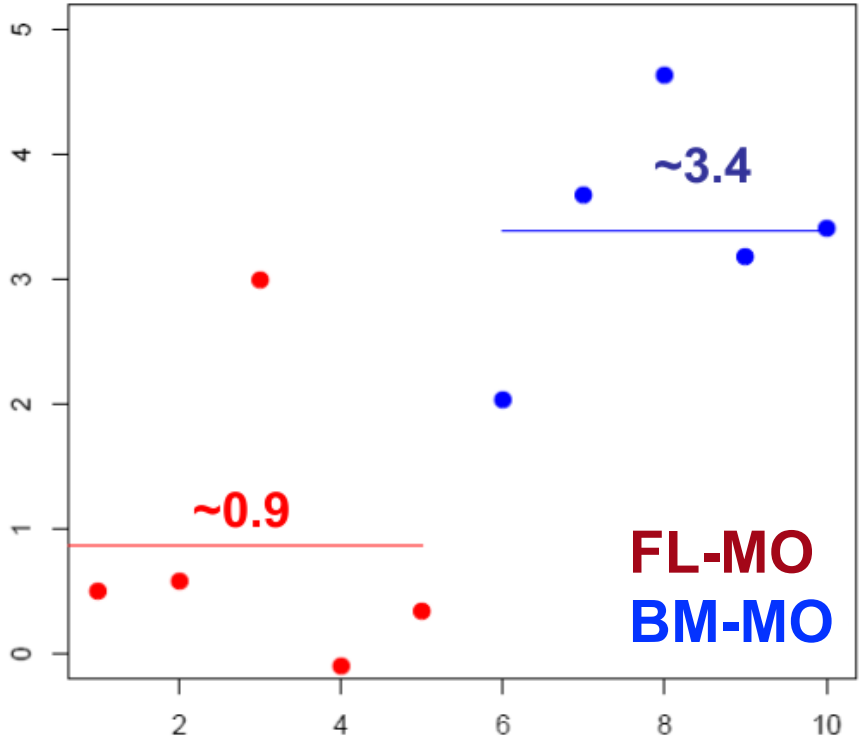
Problems - Fold change

- Low expression genes are treated equally as high expression genes
- We lose information about the variance from genes
- No statistical significance
- Is the only alternative when no replicate samples are available (not recommended!)

Basic Concepts

Mean vs. variability





T-test

We can use the t-statistic as an indication of differential expression

$$t = \frac{\bar{X} - \bar{Y}}{\mathrm{SE}},$$

difference between means

variance

$$\mathrm{SE} = \sqrt{\frac{s_X^2}{n_X} + \frac{x_Y^2}{n_Y}} \quad \text{and} \quad s_X^2 = \frac{1}{n_X - 1} \sum_{i=1}^{n_X} (X_i - \bar{X})^2 .$$

where $\overline{X}$ and $\overline{Y}$ are the mean (log) expression values of a gene in each group sample and n_x and n_y are the number of samples on these groups

Student T-test

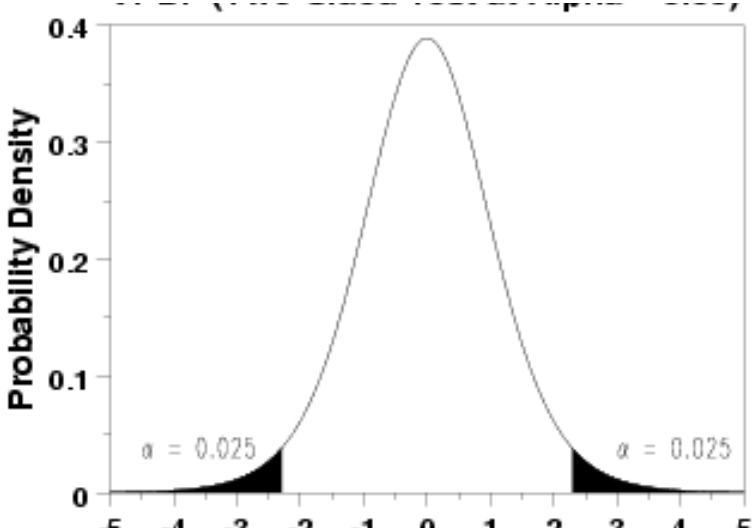
Test the hypothesis $H_0 : X - Y = 0$

$H_1 : X - Y \neq 0$

We can use the t-student distribution to estimate for which t-statistic values the null hypothesis is rejected.

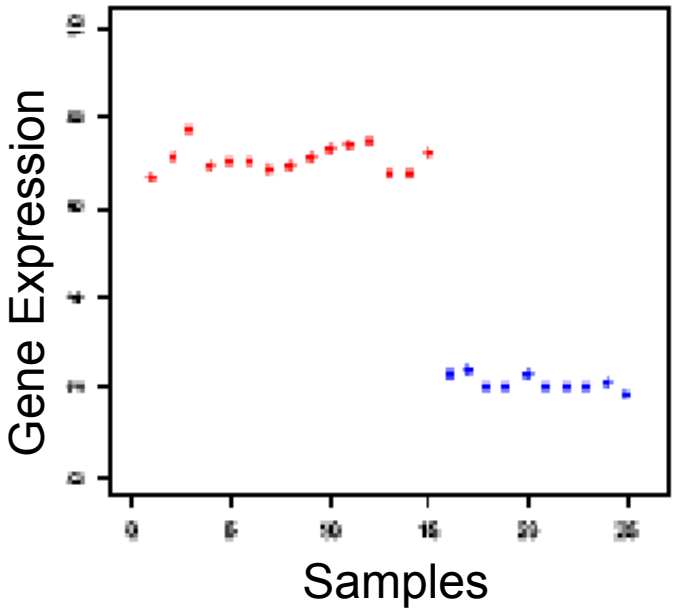
$$P\text{-value} = \Pr(t \text{ as extreme or more} | H_0),$$

t student pdf – p-value =0.05



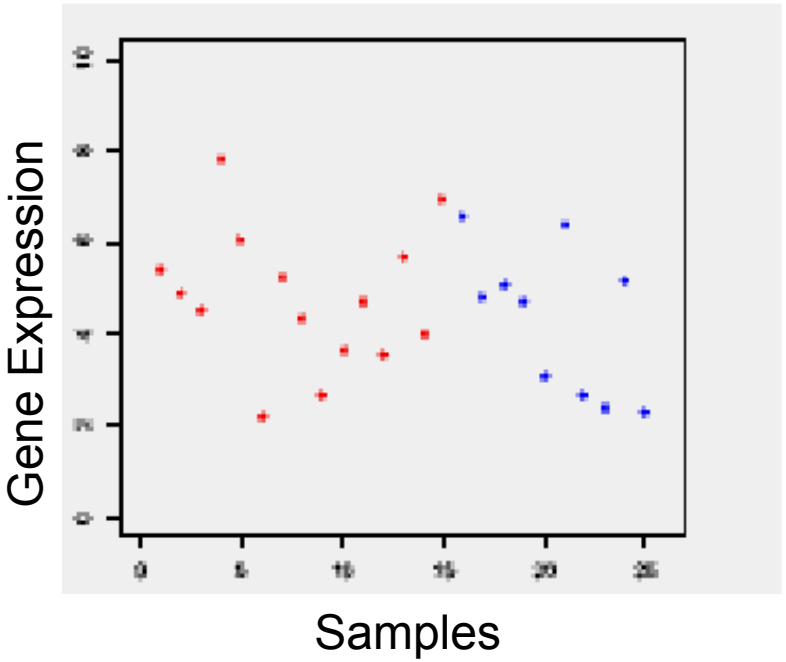
Examples

Change: HIGH
Variance: SMALL



T huge

Change: SMALL
Variance: HIGH



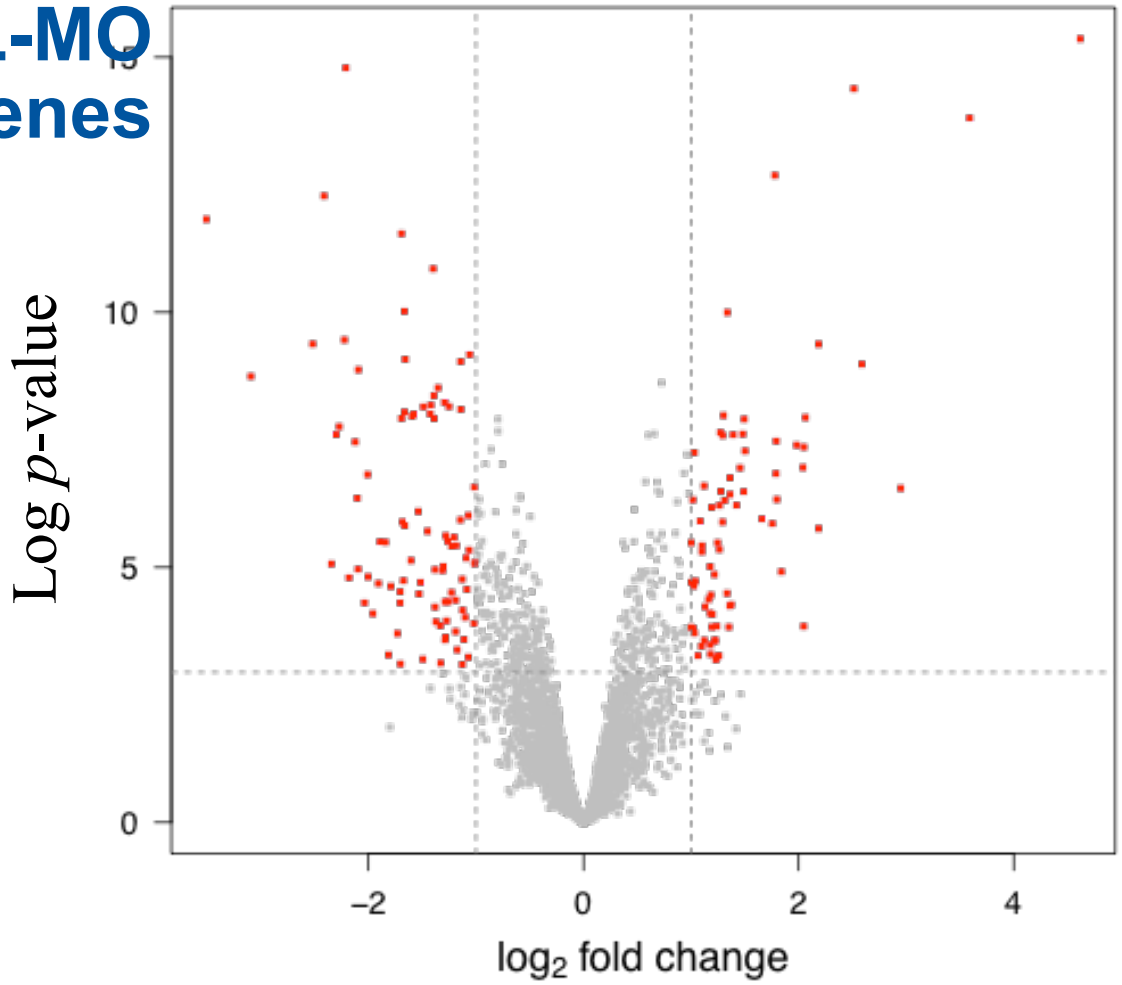
T ~ 0

Results - FL-MO vs. BM-MO

Volcano Plot - combine p-value and fold change

FL-MO genes

BM-MO genes



Multiple Test Correction

- With a p-value of 0.01, we expect to make one mistake every 100 tests
- We have 12.626 genes, therefore 126 mistaken from 1046 DE genes.
- To solve this, a multiple test correction method is necessary (i.e. Benjamini-Hochberg)
  - It is based on the false discovery rate, i.e. the proportion of false DE genes in your list of DE genes

Differential Analysis - Conclusions

- Fold-change (alone) -> should be avoided
- For patient samples
  - high number of replicates are necessary (>30)
  - otherwise - low DE genes replicability
- For model (mouse) experiments
  - at least 3 samples (and moderated t-test)
  - we can not tell the variance without measuring it!
- All correct for multiple testing!

Hanc

Series GSE76999 Query DataSets for GSE76999

Status	Public on Mar 01, 2016
Title	Capacity of yolk sac macrophages, fetal liver and adult monocytes to colonize an empty niche and develop into functional tissue resident macrophages

sion Omnibus

Platforms (1) GPL6246 [MoGene-1_0-st] Affymetrix Mouse Gene 1.0 ST Array [transcript (gene) version]

Samples (36)
More...

GSM2042244 Monocyte extracted from adult (wk6-12) Bone Marrow, biological replicate 1

GSM2042245 Monocyte extracted from adult (wk6-12) Bone Marrow, biological replicate 2

GSM2042246 Monocyte extracted from adult (wk6-12) Bone Marrow, biological replicate 3

Relations

BioProject PRJN 9234

Analyze with GEO2R

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Citation(s) Hofmann E, Beyaert R, Saeys Y, Lambrecht BN, Guilliams M

van de Laar L, Saelens W, De Prijck S, Martens L et al. Yolk Sac Macrophages, Fetal Liver, and Adult Monocytes Can Colonize an Empty Niche and Develop into Functional Tissue-Resident Macrophages. *Immunity* 2016 Apr 19;44(4):755-68. PMID: 26992565

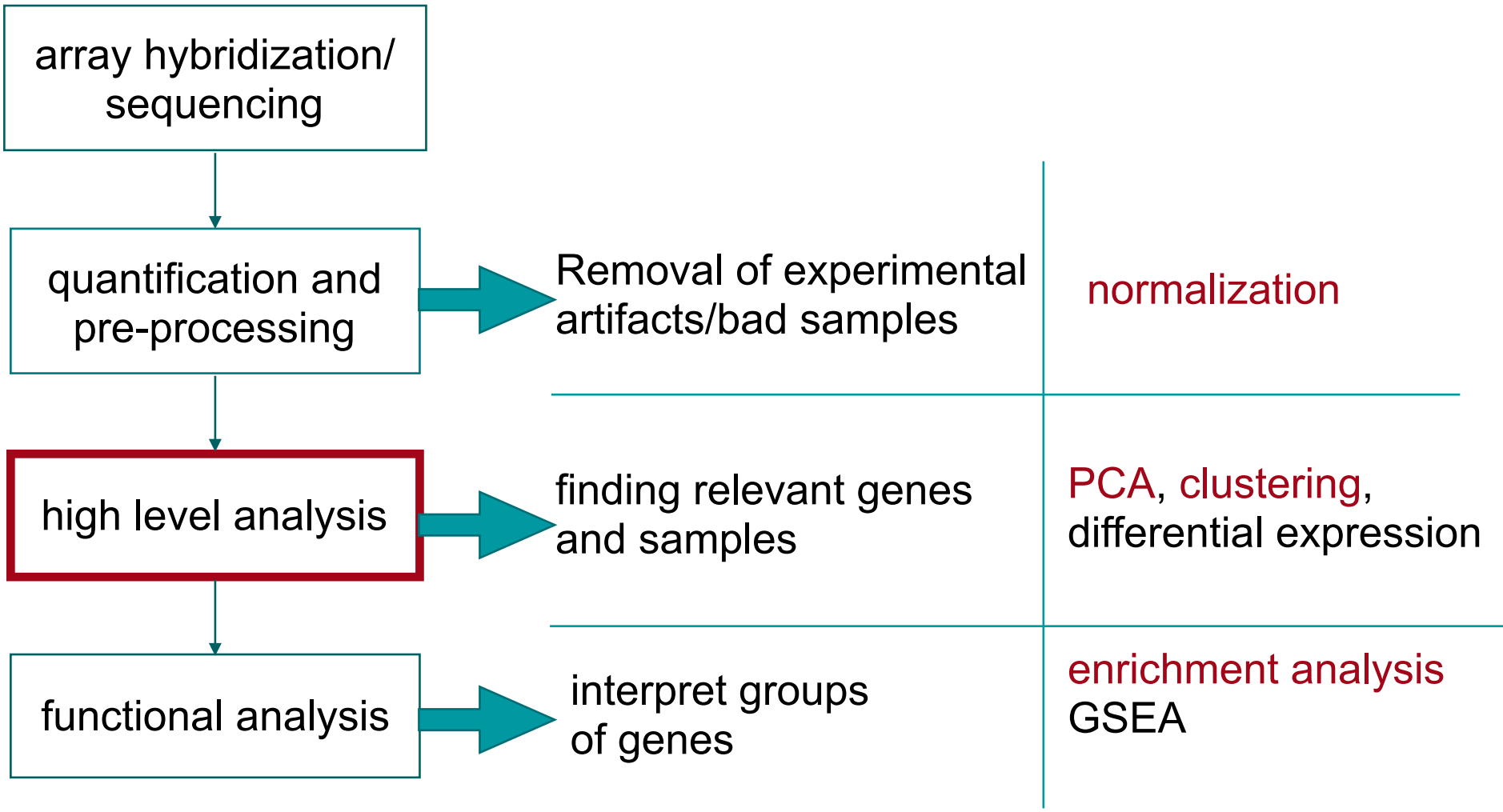
NTH AACHEN UNIVERSITY

Using GEO2R

- Select the interested data:
  - Monocyte extracted from adult Bone Marrow (BM)
  - Monocyte extracted from E15.5 Fetal Liver (FL)
  - Macrophage extracted from E12.5 Yolk Sac (YS)
- Define three groups
- Get top 250 DE genes
- See the R script

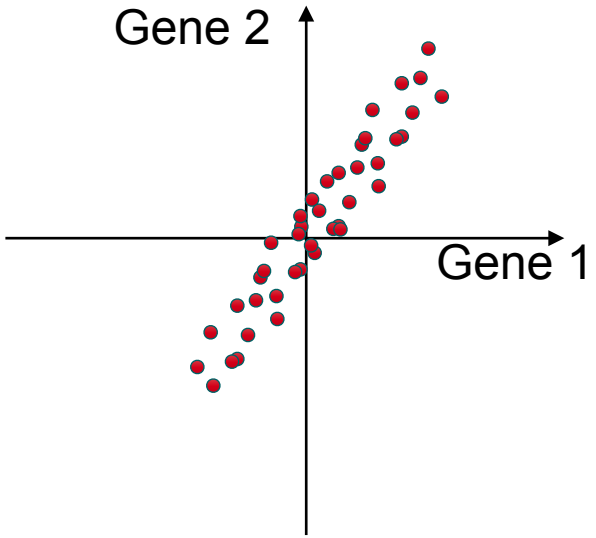
- Handout step 1 to 3

Bioinformatics - Gene Expression Analysis



Principal Component Analysis

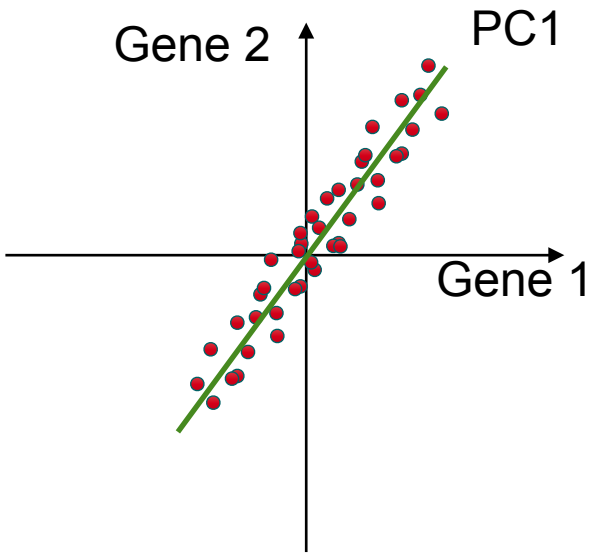
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

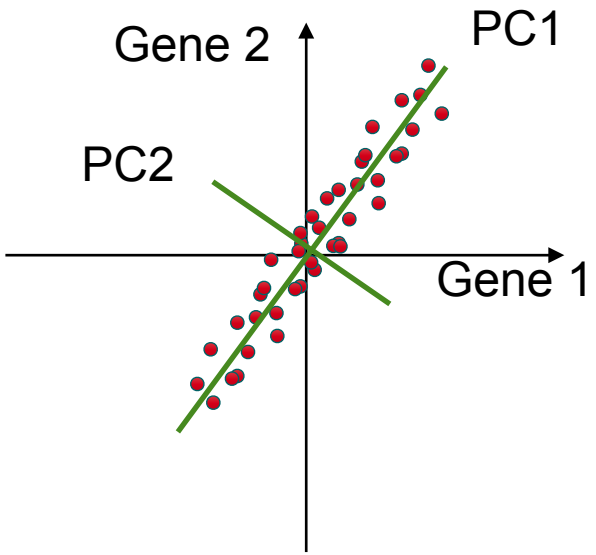
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

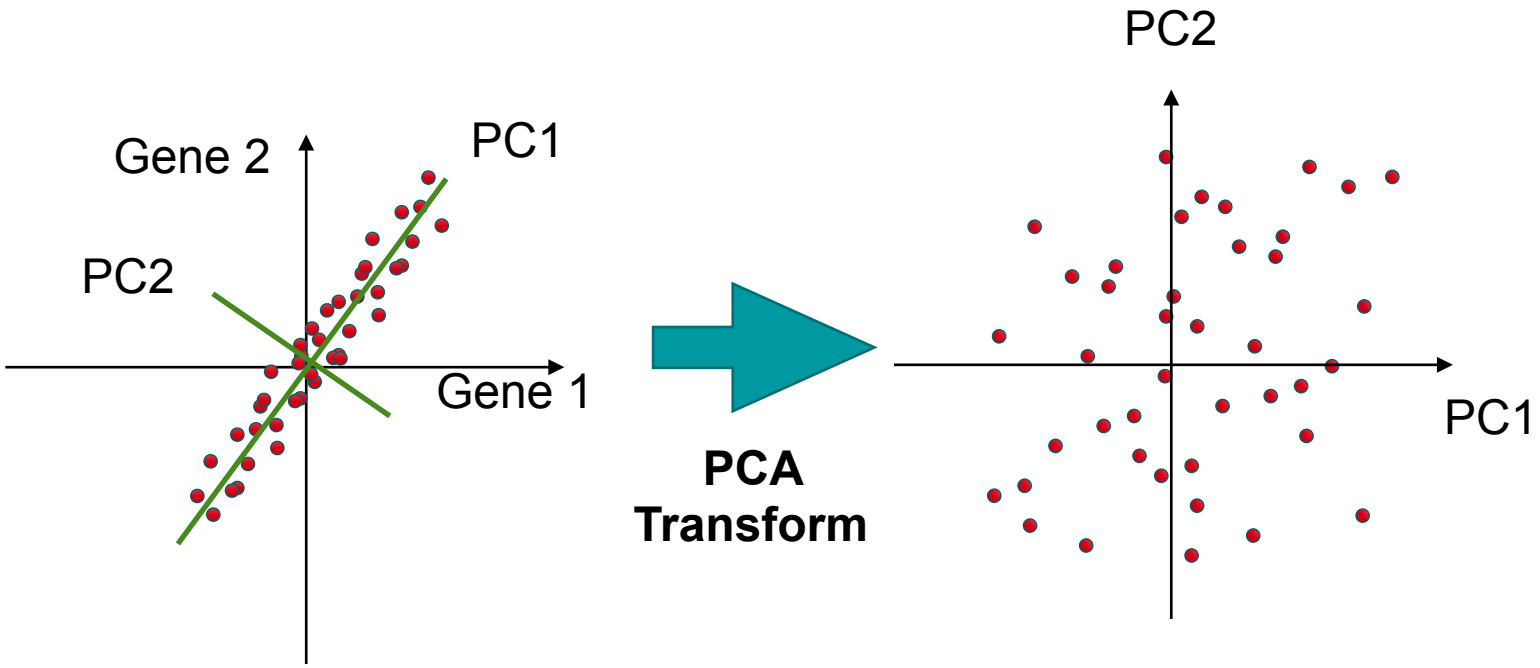
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

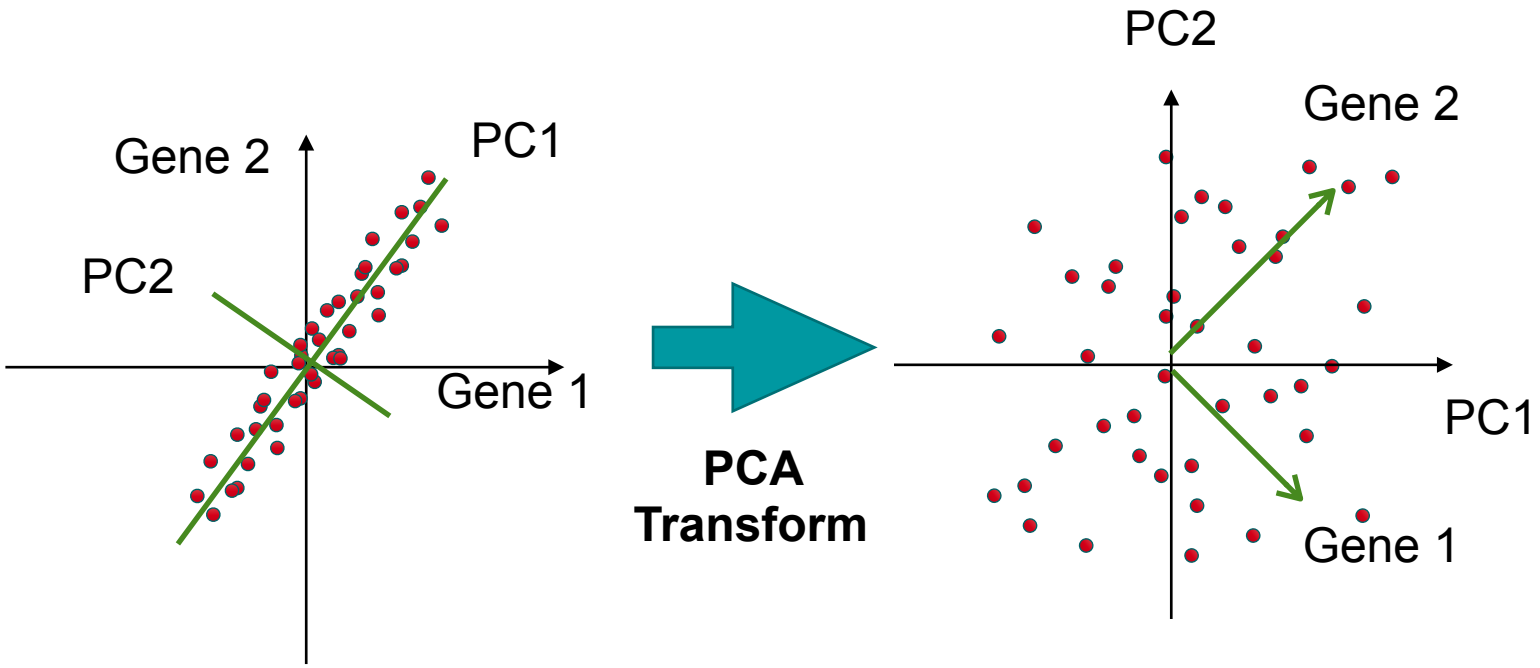
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Principal Component Analysis

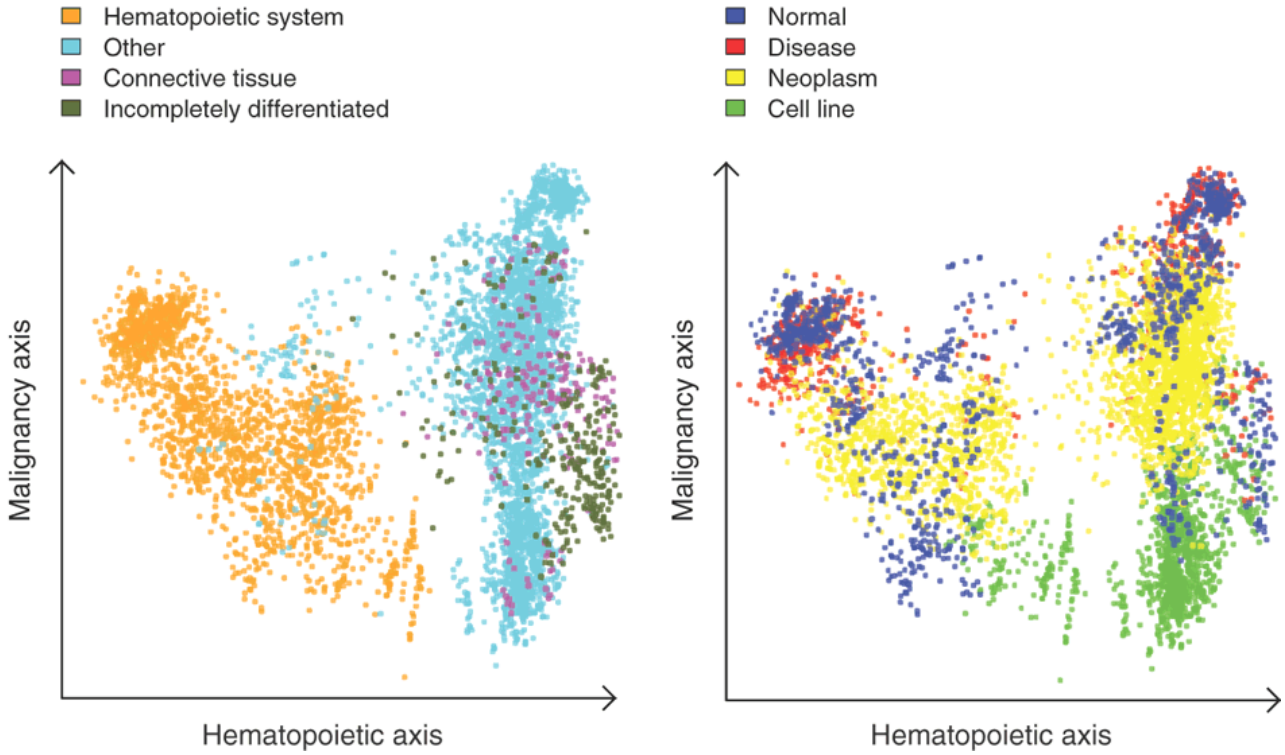
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



Recommended reading:
Ringner M., *Nature Biotechnology* 26, 303 - 304 (2008)

Gene Expression - PCA Example 1

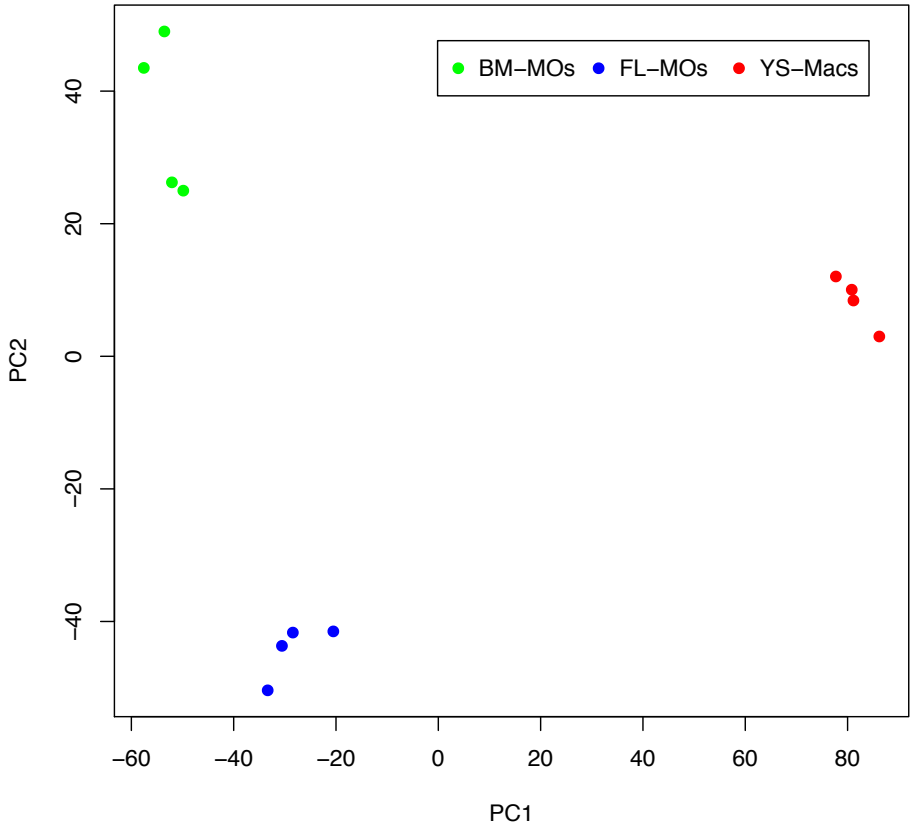
Can be interpreted as a computational FACs sorting (without knowing the markers)



First 2 PCs on the analysis of 5000 samples from Array Express/EBI

Gene Expression - PCA Example 2

PCA Analysis of van de Leer, 2016 data

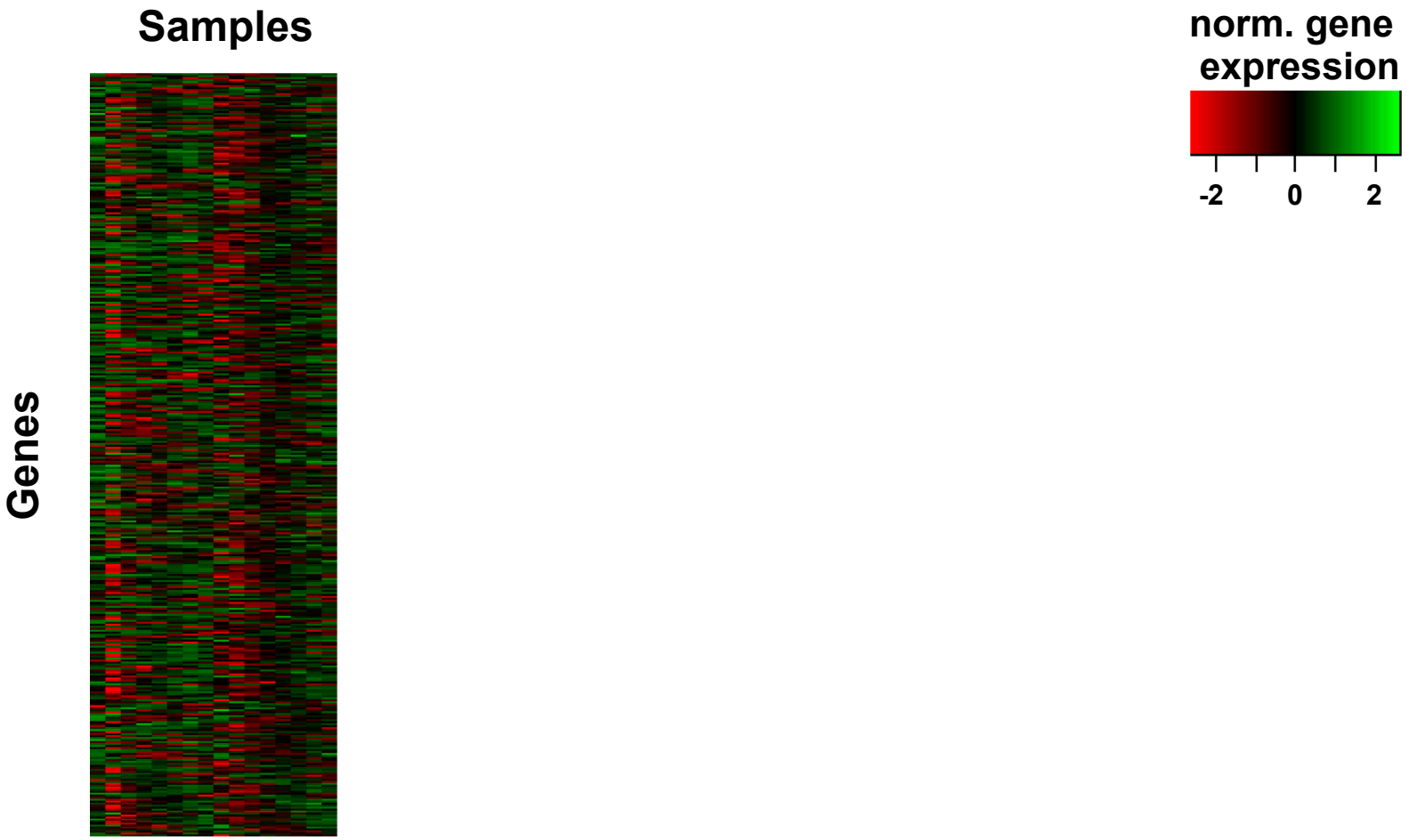


First 2 PCs van de Leer, 2016 data

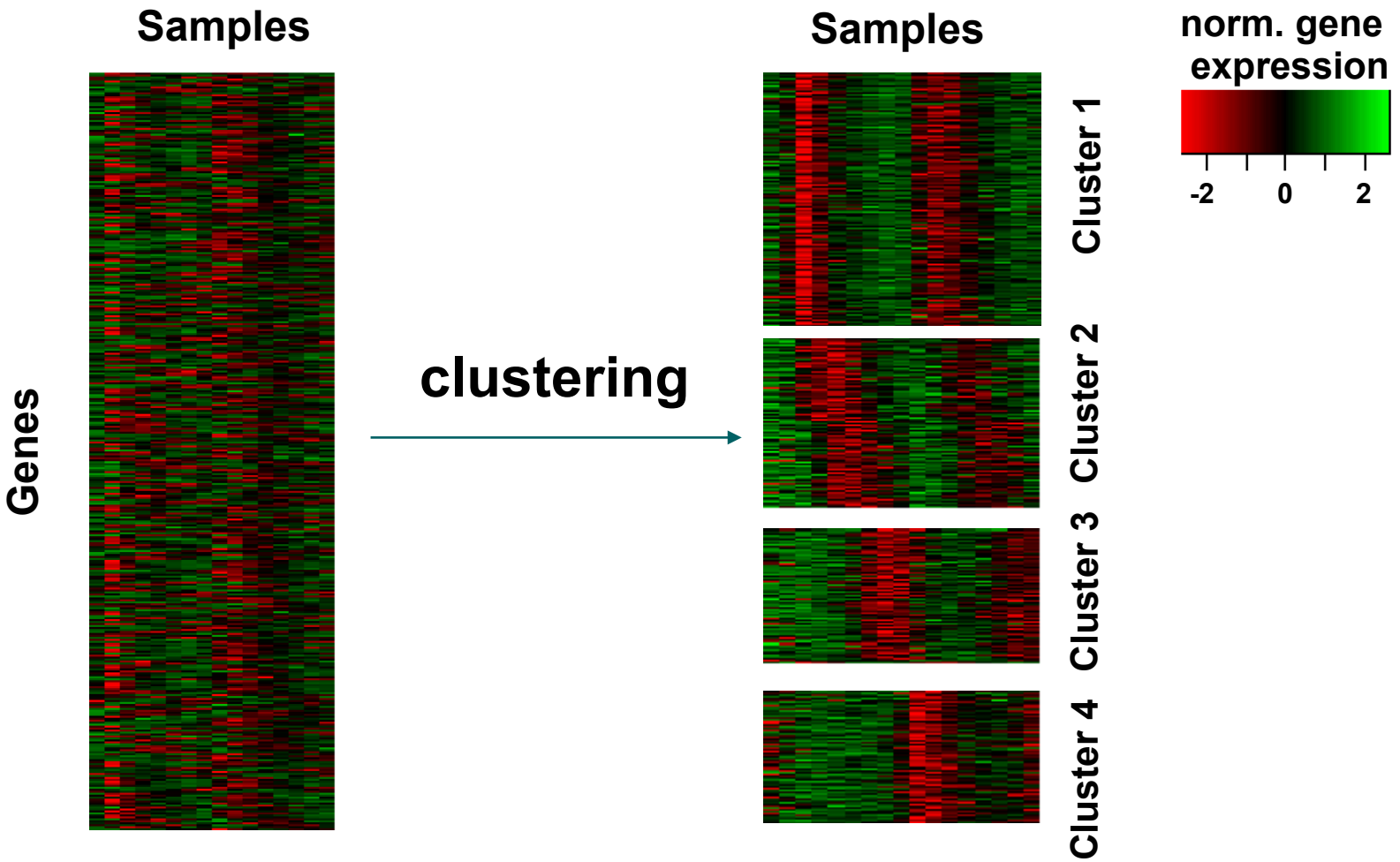
PCA Analysis - Conclusions

- PCA allows an “blind” cell sorting
  - only works if variant directions split the groups
  - is complementary to clustering
- Weights allow interpretation of relevant variables
- Can also be used for quality check
  - samples not fitting to groups
- Alternatives to PCA:
  - tSNE - very commonly used in single cell RNA-seq

Clustering / Heatmaps

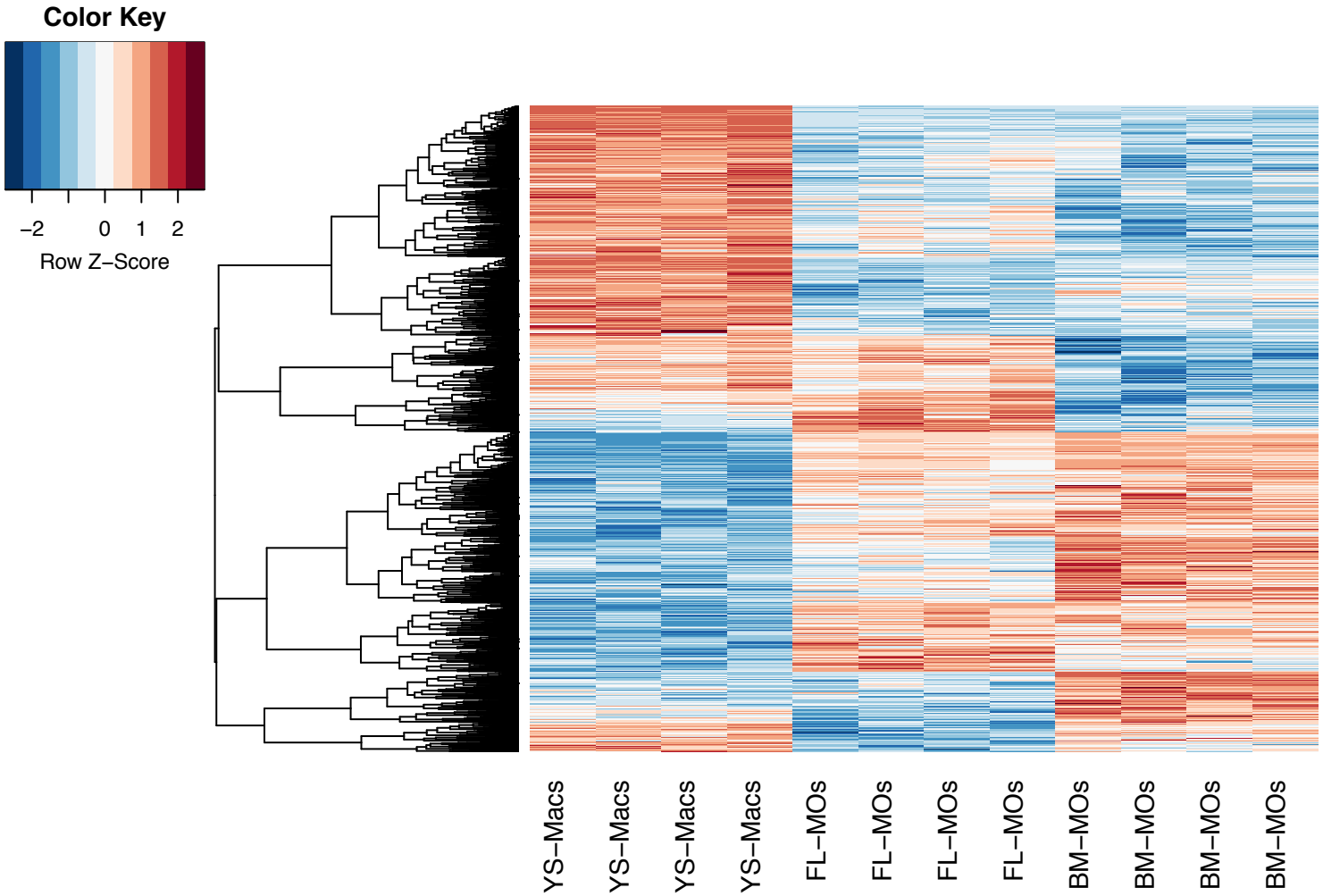


Clustering / Heatmaps



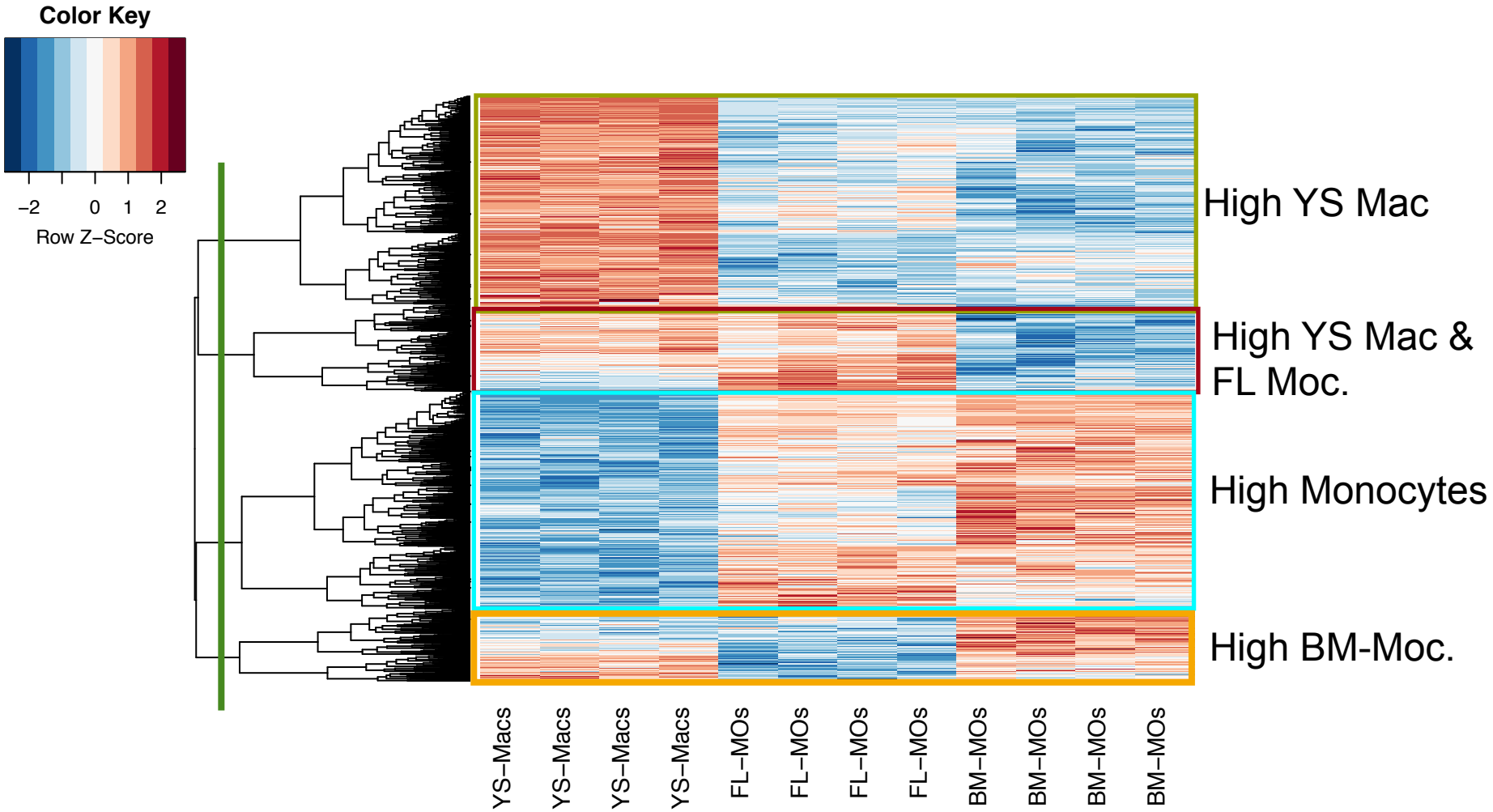
clustering methods: k-means, hierarchical clustering, …

Hierarchical Clustering



distance metric - Pearson correlation recommended

Hierarchical Clustering



distance metric - Pearson correlation recommended

Hands on!

Handout Step 4 and 5

Functional Analysis

Clustering/Differential Expression (DE) returns lists of hundreds of genes How to functionally characterize these?

Solution 1 - Look at each gene individually

Solution 2 - Relate these genes to annotations from databases

- Gene Ontology, pathways, gene sets, disease ontology, …

Databases

Manually or automatic curated annotation of genes

Pathways

KEGG Kyoto Encyclopedia of Genes and Genomes

BIOCARTA

REACTOME

Experimental

MSigDB
Molecular Signatures Database

Ontologies

GO
the Gene Ontology

DO
Disease Ontology

Gene Ontology

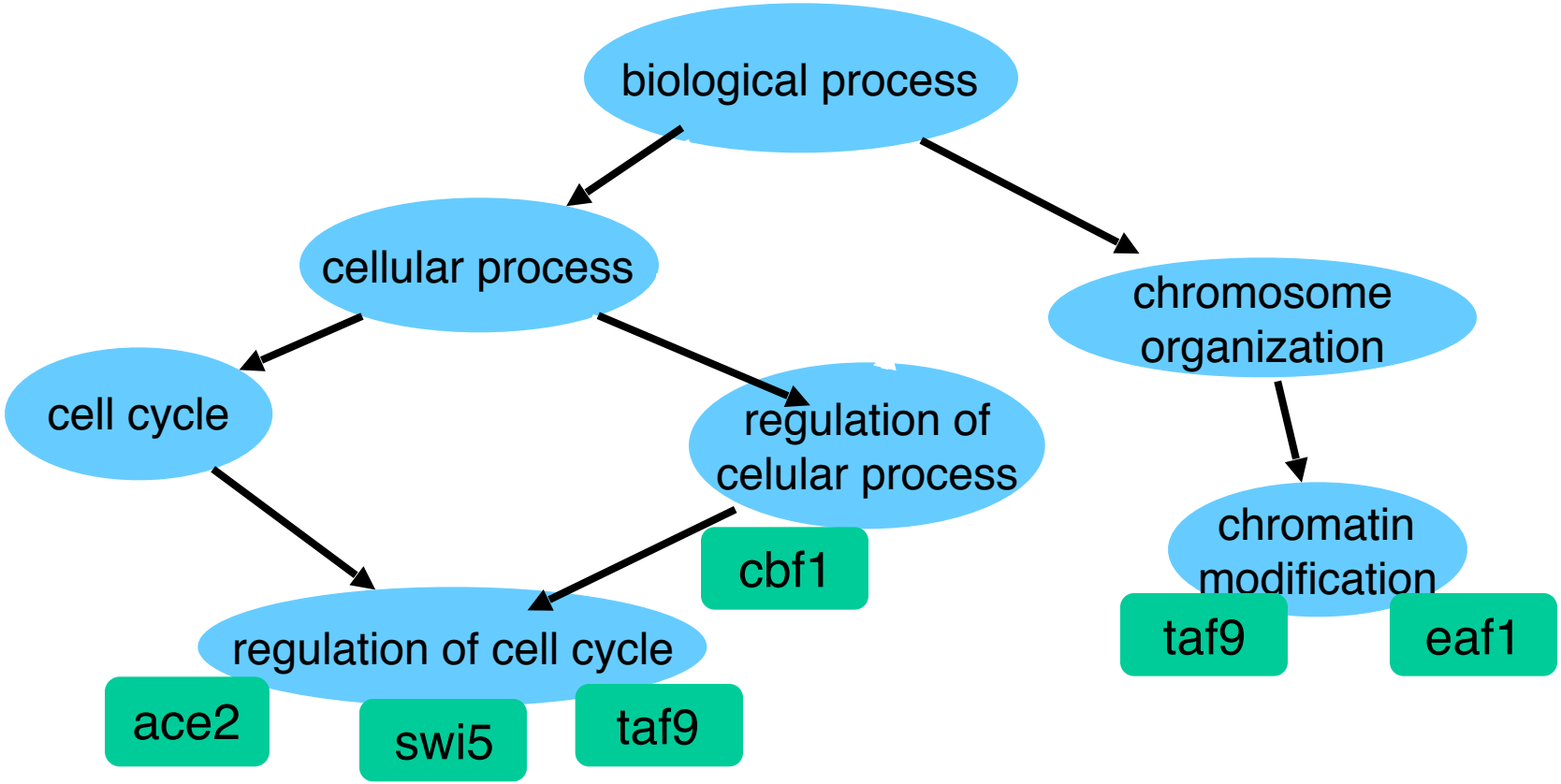
Controlled vocabulary to describe gene and gene product attributes in any organism

Formed by three ontologies

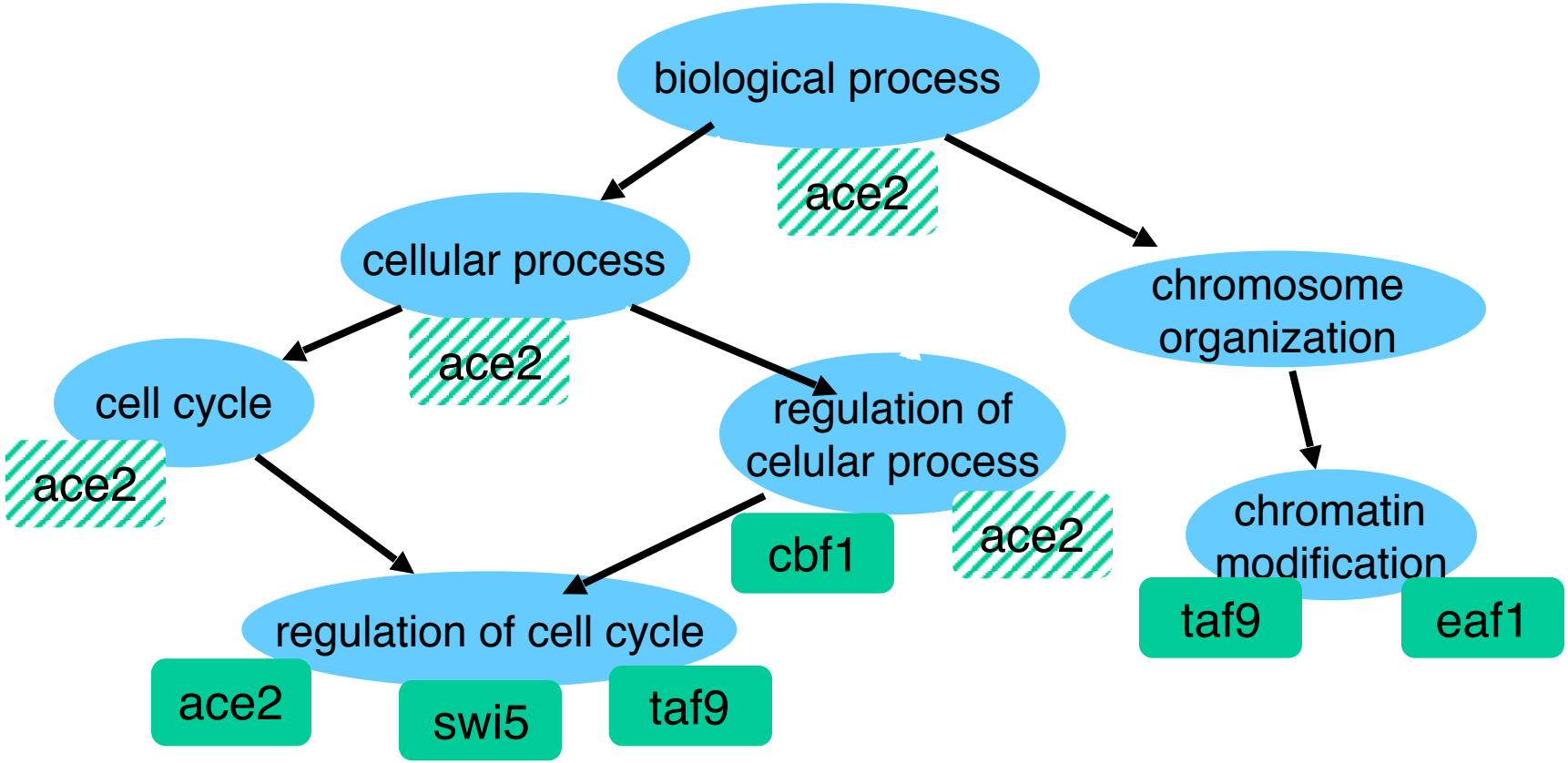
1. Biological Process (BP)
2. Molecular Function (MF)
3. Celular Component (CC)

Annotation (Organism depend)
- genes are associated to terms manually (literature) or automatically (sequence homology)

Gene Ontology



Gene Ontology



inheritance property

GO Enrichment Analysis

DE analysis results

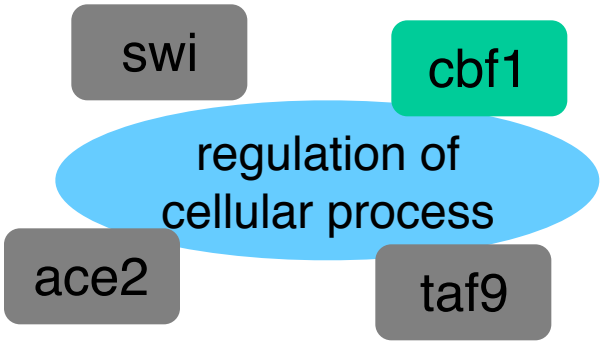
up regulated genes

SWI
ACE2
CBF1
YJL099W
YDL198C
YCR085W
YCR043C
YDR825C

all other genes

YDL093W
YER016W
YNL126W
YKL053W
YJL099W
YDL198C
YCR085W
YBR043C
YDR325W
YCR085W
YBR043C
...

GO Term



How probable is that 3 up regulated genes are annotated to the GO term?

GO Enrichment Analysis

DE analysis results

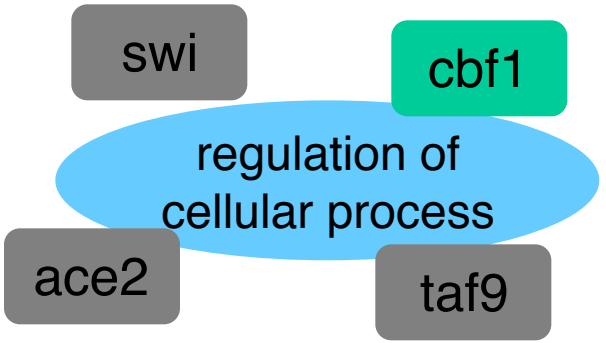
up regulated genes

SWI
ACE2
CBF1
YJL099W
YDL198C
YCR085W
YCR043C
YDR825C

all other genes

YDL093W
YER016W
YNL126W
YKL053W
YJL099W
YDL198C
YCR085W
YBR043C
YDR325W
YCR085W
YBR043C
...

GO Term



Statistics:
Fisher's Exact Test

GO Term Annotation

Up-regulated	YES.	NO
YES	3	1
NO	8	6421

Enrichment Analysis Tools

For a given gene list:

1. evaluate the the overlap of the list vs. all gene sets
   i.e. GO terms, pathways, …

2. Estimate p-value (corrected by multiple testing)

3. Rank gene sets by lowest p-value

G:Profiler

We interface for enrichment analysis with:
Gene Ontology, KEGG Pathway and TF binding

http://biit.cs.ut.ee/gprofiler/index.cgi

Check the results for my favorite genes:

Irf8 Id2 Spi1 Klf4 Runx2 Egr1

Hands on!

Handout Step 6

G:Profiler

We interface for enrichment analysis with:
Gene Ontology, KEGG Pathway and TF binding

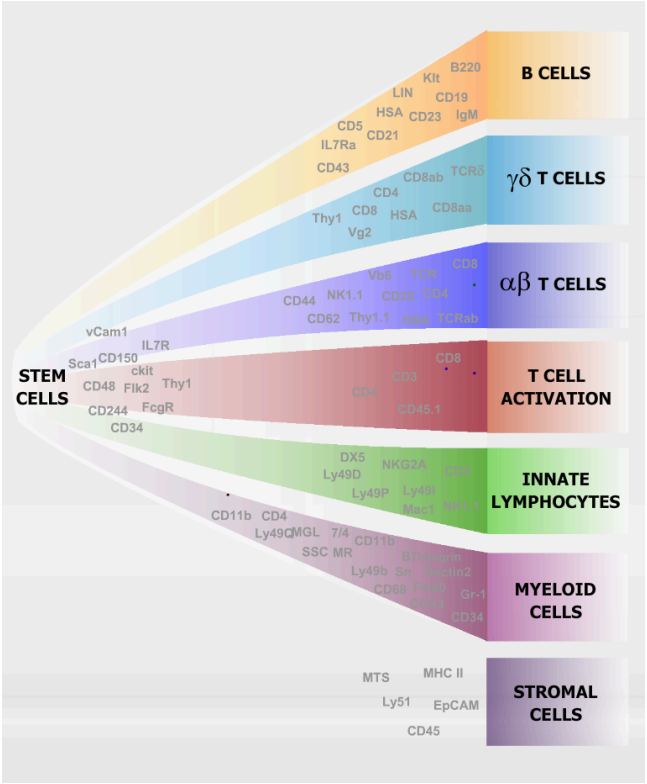
http://biit.cs.ut.ee/gprofiler/index.cgi

Check the results for my favorite genes:

Irf8 Id2 Spi1 Klf4 Runx2 Egr1

Integrative Analysis - ImmGen

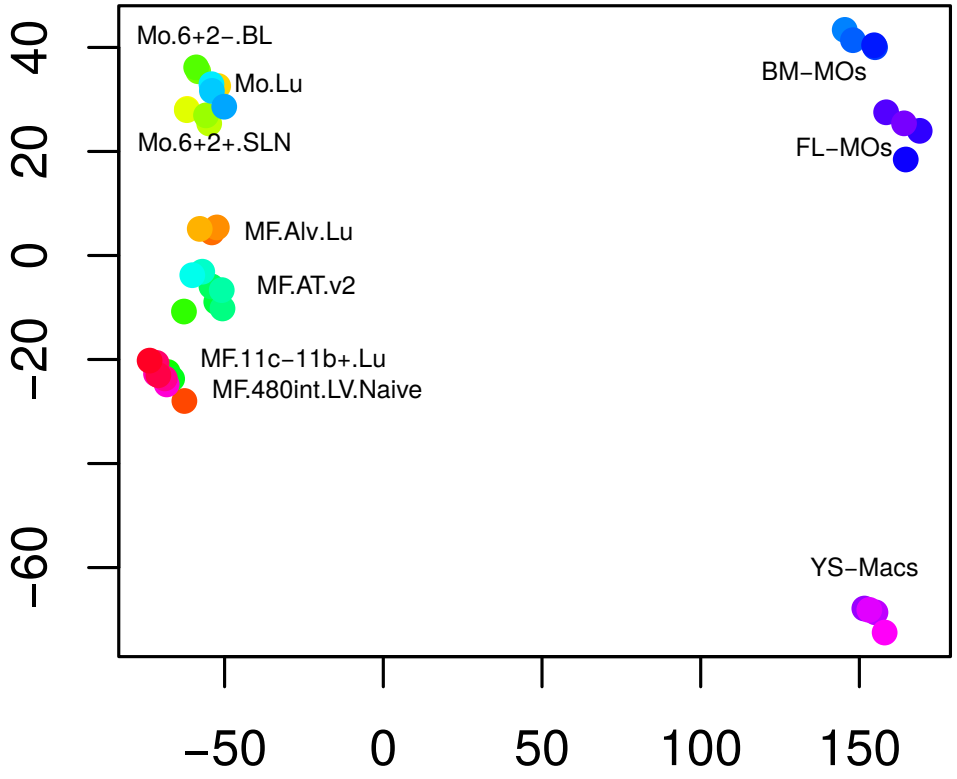
- ImmGen - expression data of immune cells under standardized conditions



- How do cells from **van de Leer, 2016** compares to monocyte/macrophages from ImmGenn?
- we obtained/pre-processed ImmGen data (v1) from GEO (GSE15907)

Integrative Analysis - Problem

- Batch Effects - Arrays from distinct lab tends to cluster together

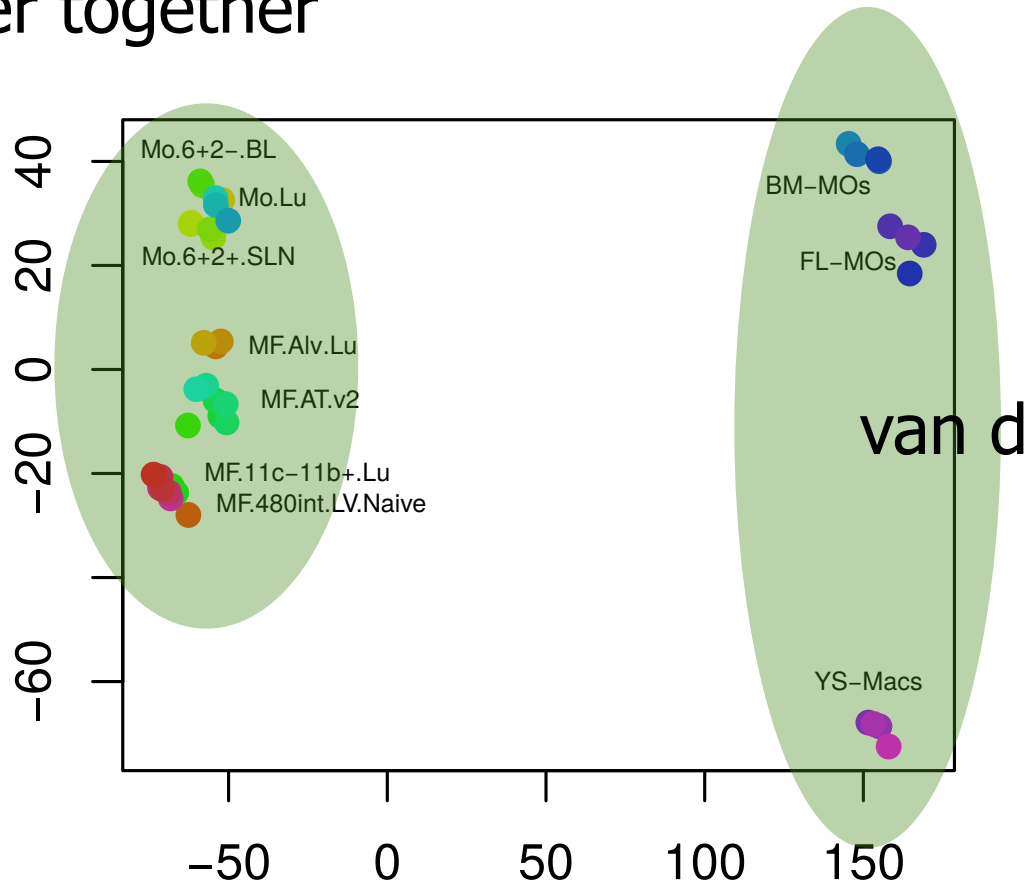


See: Leek JT,…. (2016). sva: Surrogate Variable Analysis. R package version 3.22.0.

Integrative Analysis - Problem

- Batch Effects - Arrays from distinct lab tends to cluster together

ImmGenn

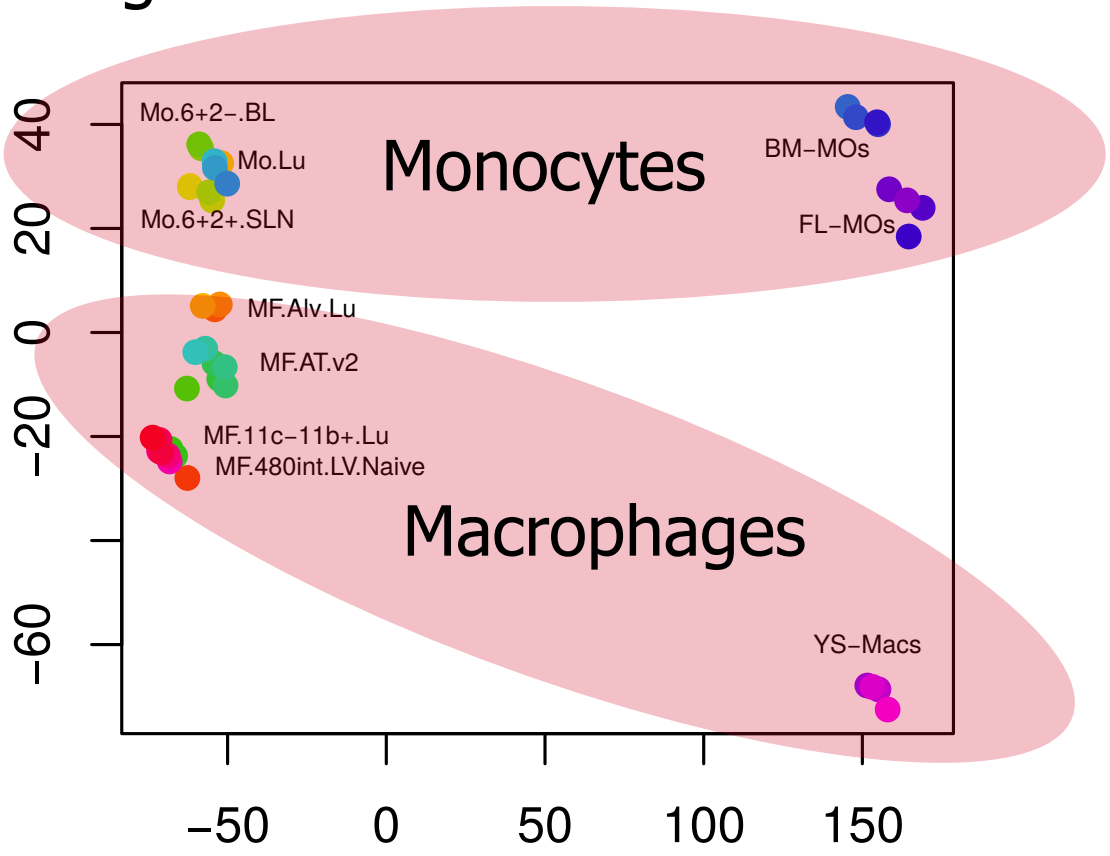


van de Leer, 2016

See: Leek JT,…. (2016). sva: Surrogate Variable Analysis. R package version 3.22.0.

Integrative Analysis - Problem

- Batch Effects - Arrays from distinct lab tends to cluster together



See: Leek JT,…. (2016). sva: Surrogate Variable Analysis. R package version 3.22.0.

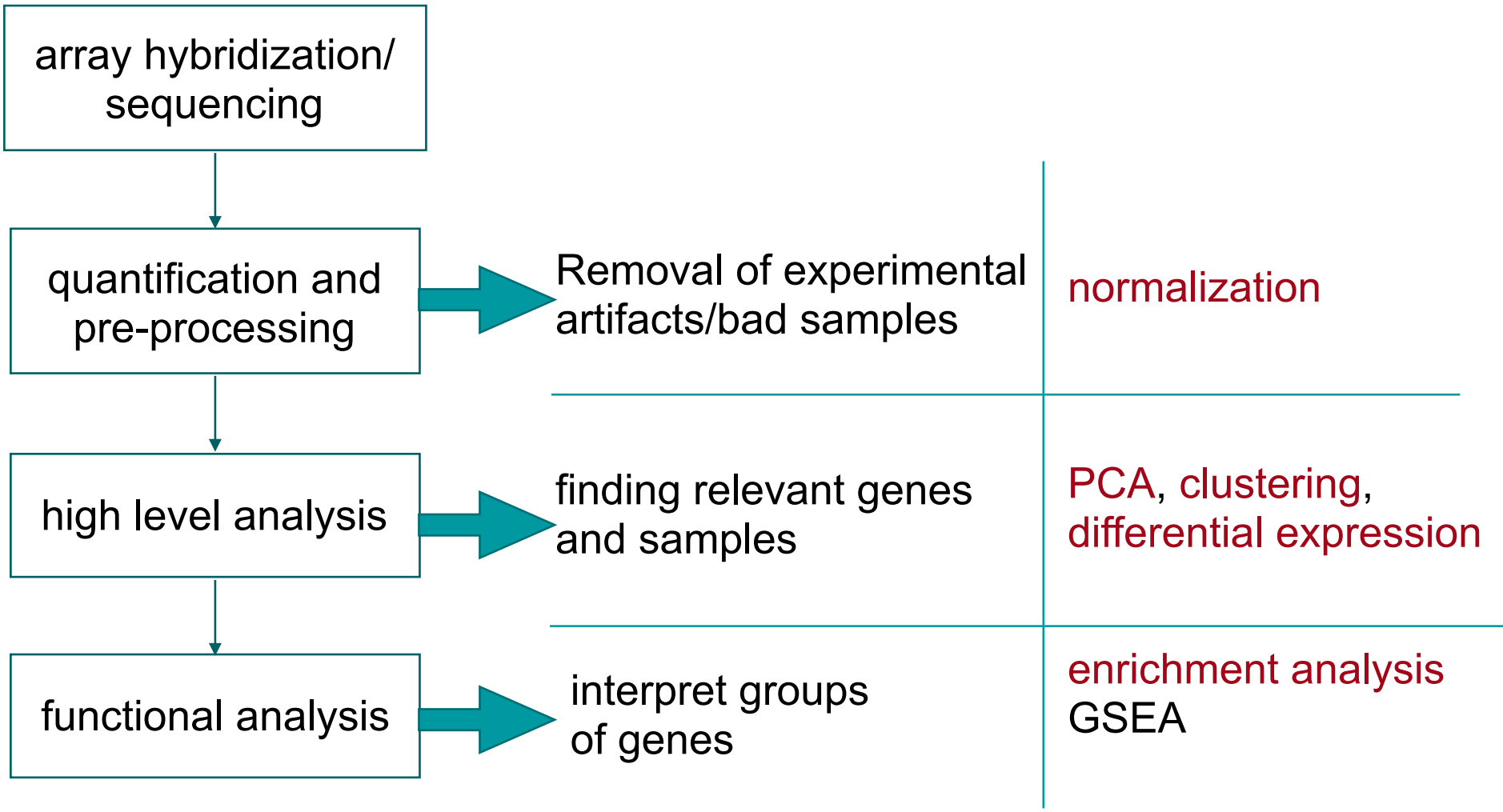
Integrative Analysis - PCA After Combat

- Solution - Batch effect removal with COMBAT
  - annotation of your data: tissue of origin, cell type, experimental batches

Hands on! Handout Step 7

See: Leek JT,…. (2016). sva: Surrogate Variable Analysis. R package version 3.22.0.

Bioinformatics - Gene Expression Analysis



www.costalab.org

IZKF Interdisziplinäres Zentrum für Klinische Forschung

RWTH AACHEN UNIVERSITY