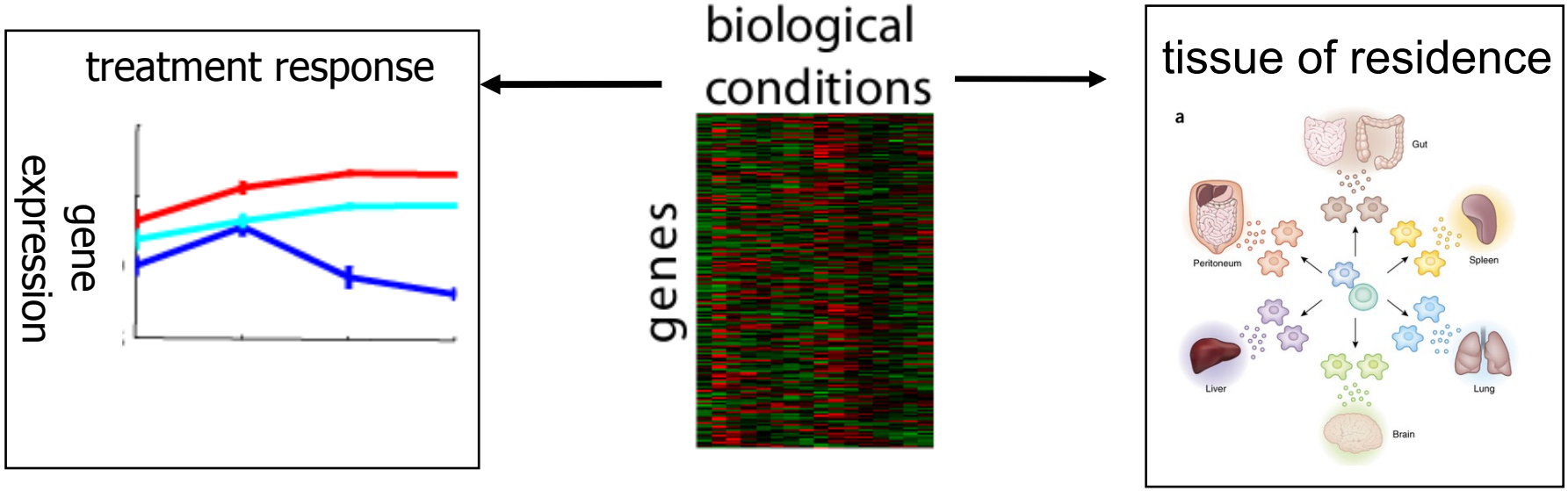
INTRICARE Course
Bioinformatics for gene expression analysis

Ivan G. Costa & Tiago Maie
Institute for Computational Genomics
RWTH Aachen
www.costalab.org

Institute for Computational Genomics
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

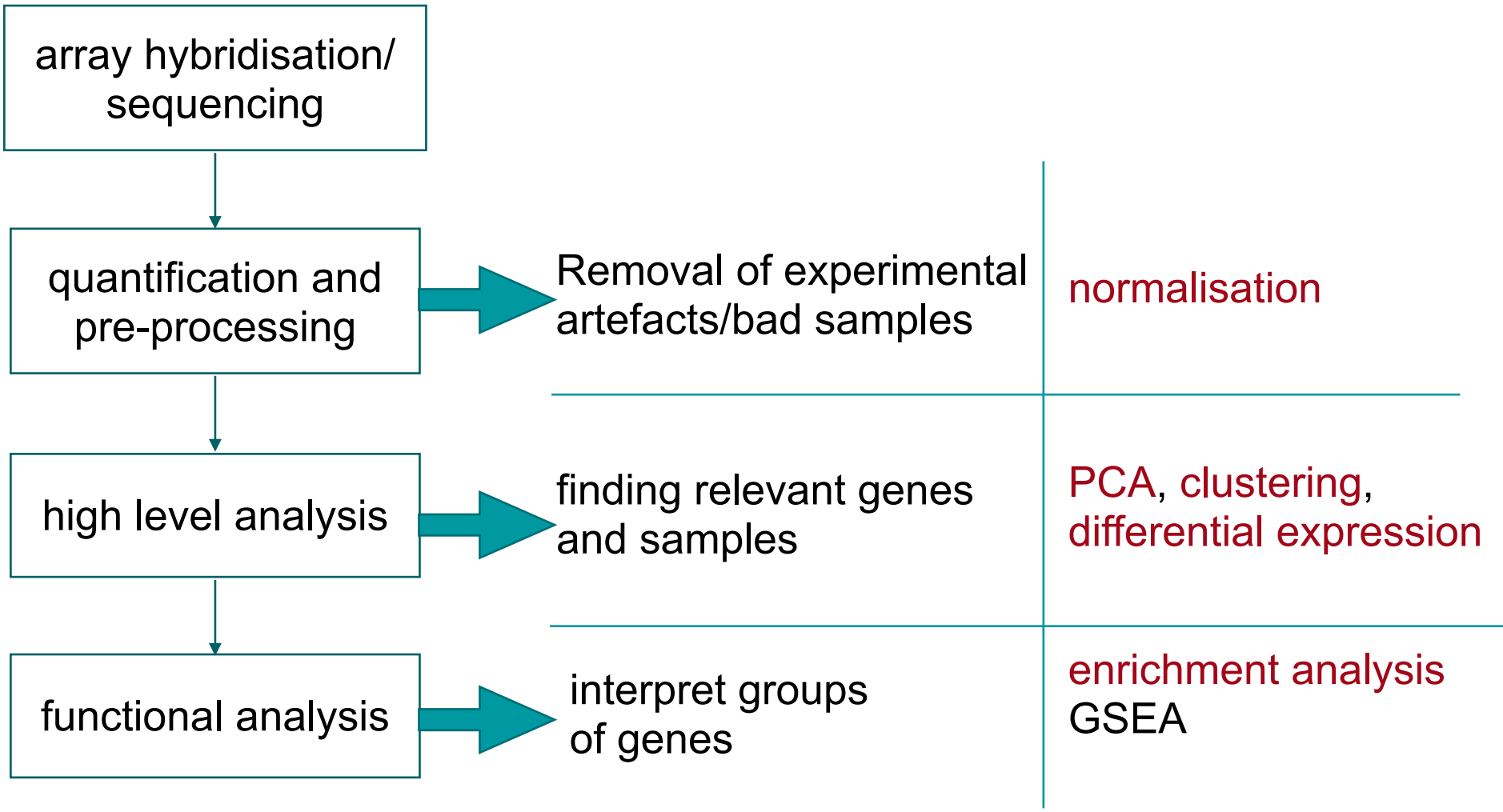
1 - Give you a (very) brief overview on the use of R/ bioconductor tools

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

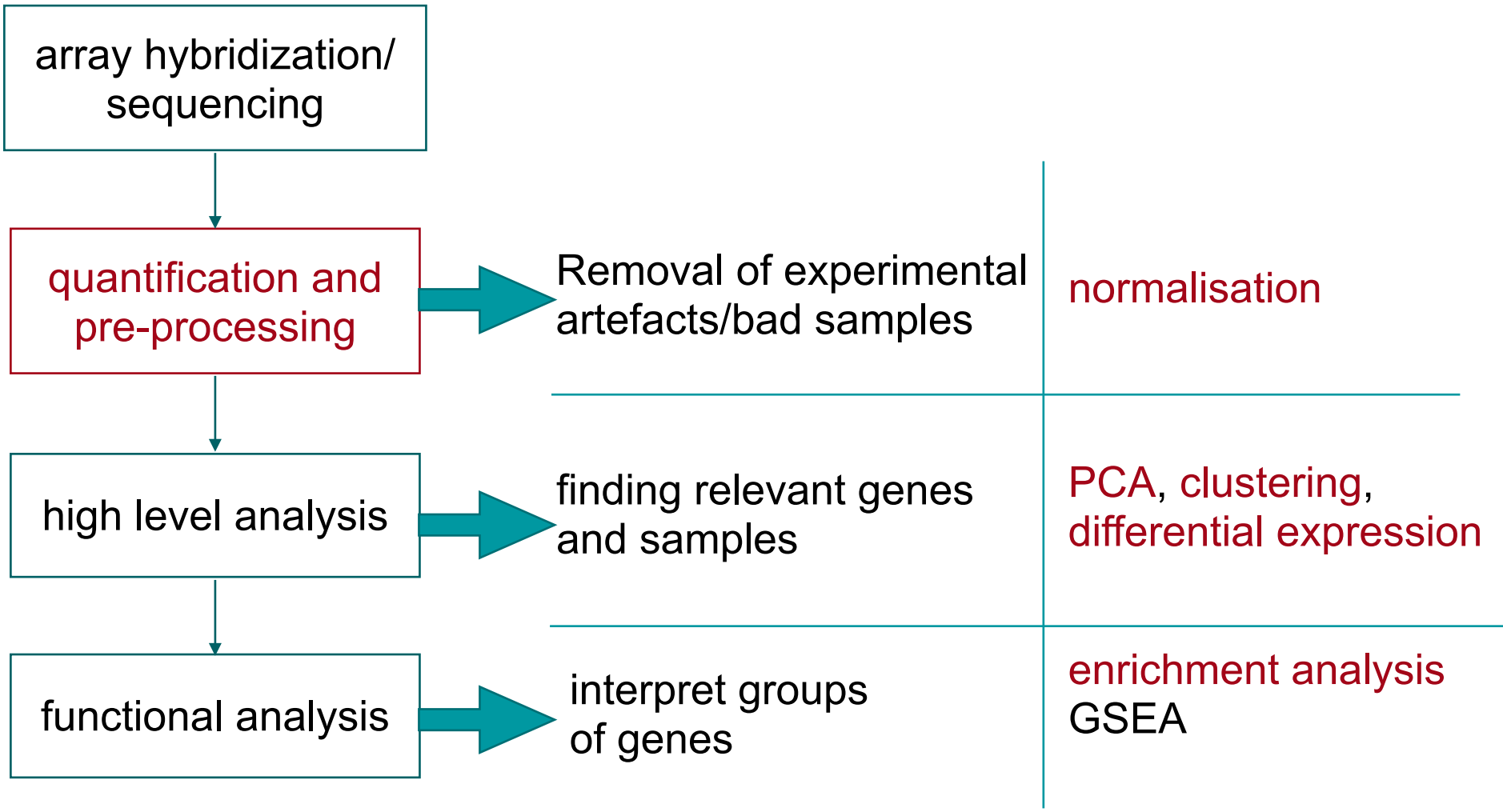
3 - Why arrays? Analysis of sequencing data is still complex / requires command line “programming”.

However, high level analysis are the same!

Bioinformatics - Gene Expression Analysis

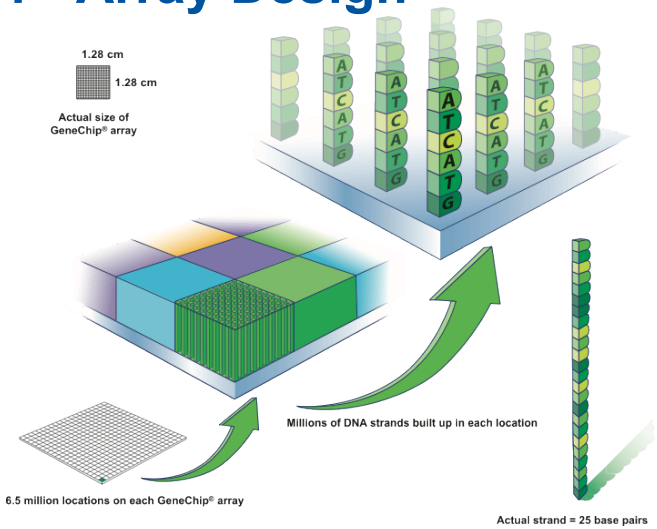


Bioinformatics - Gene Expression Analysis

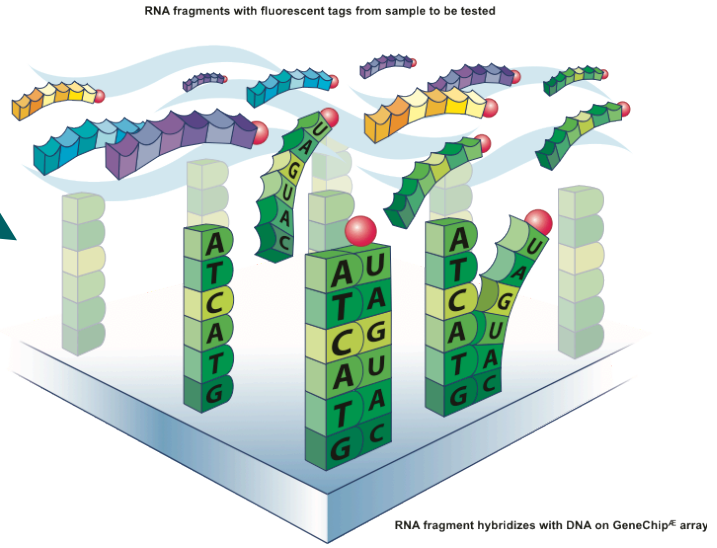


Affymetrix Arrays - Example

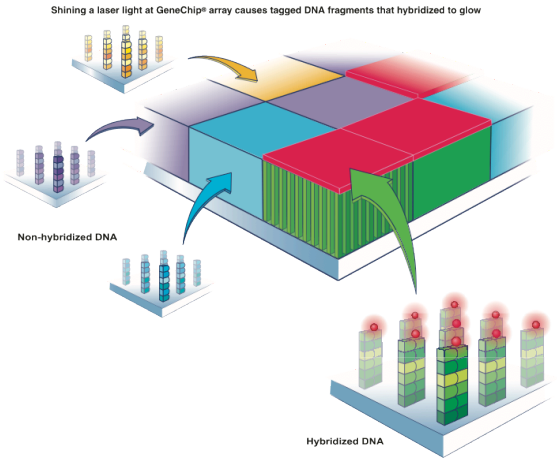
1 - Array Design



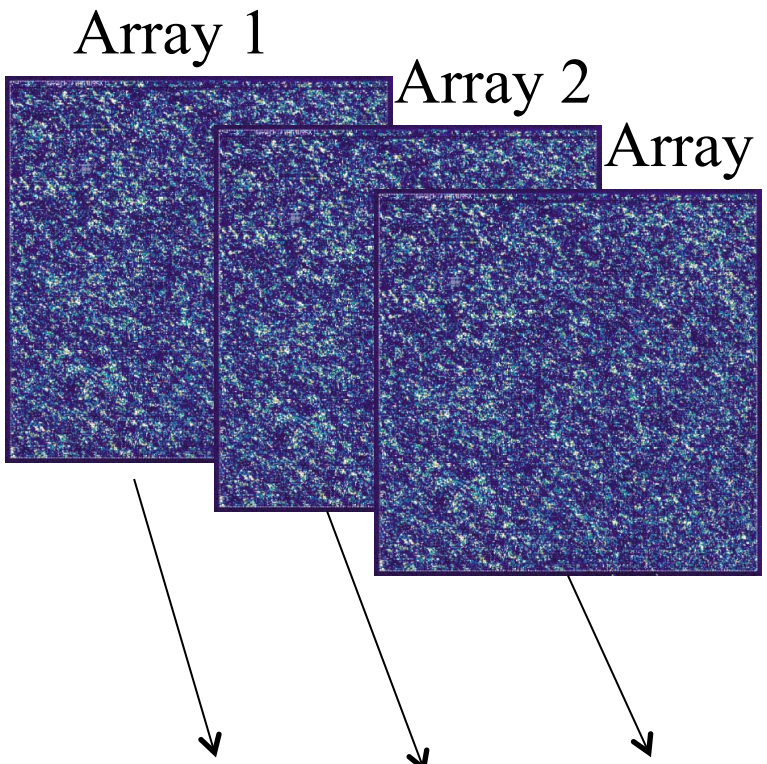
2 - cDNA Hybridization



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

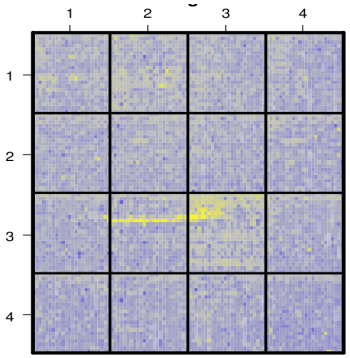
2 - Quality Control
- remove bad samples

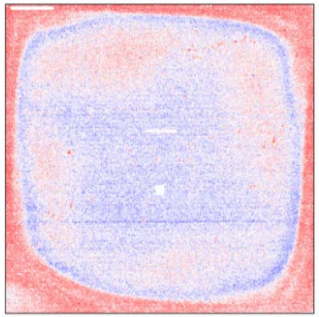
3 - Correct for Experimental artefacts
- normalisation

Why is QC / Normalisation important?

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

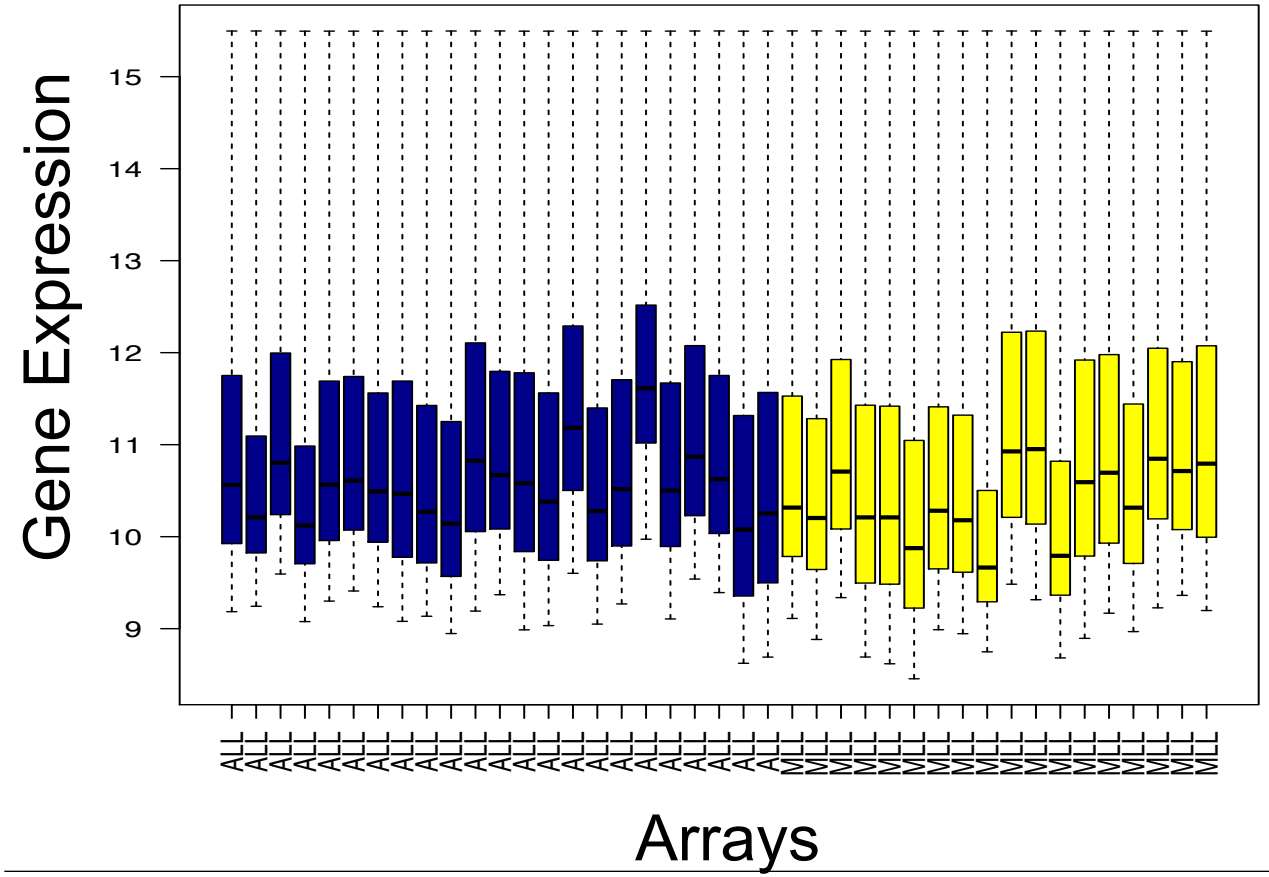
Example of Hybridisation Problems

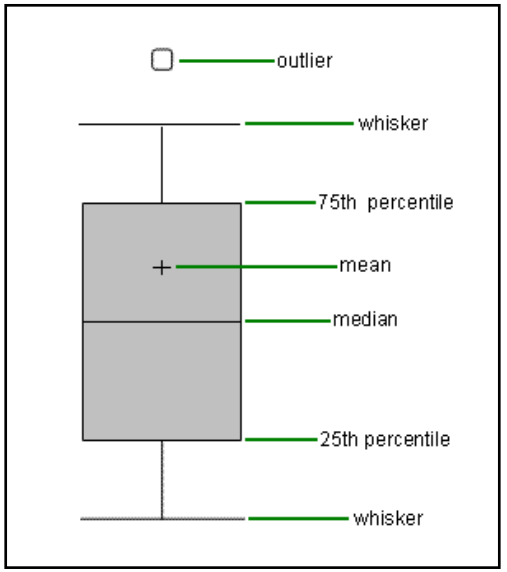




Normalisation Principles

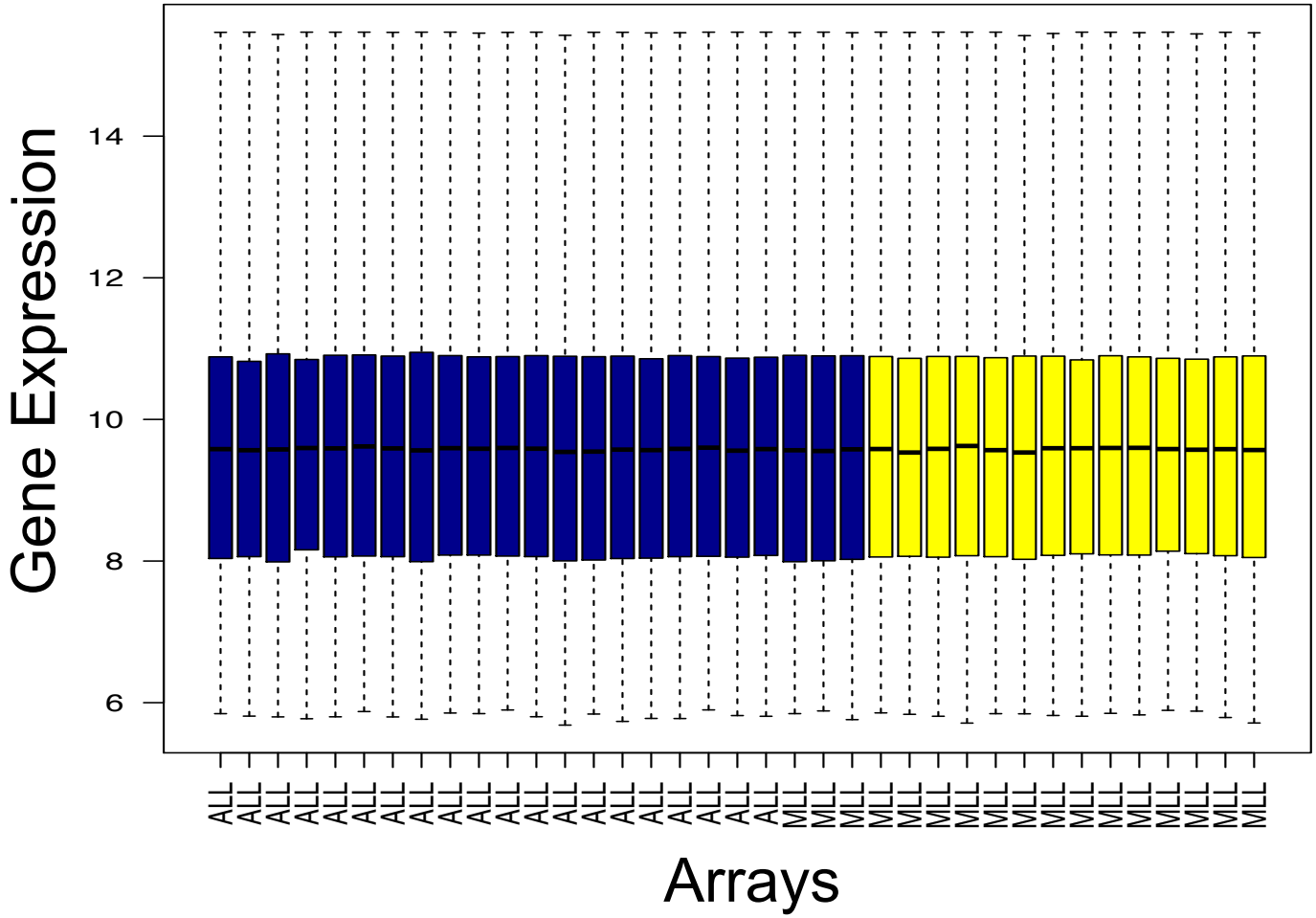
1 - Most genes don't change expression -> small/same variance
2 - Arrays are hybridised with the same amount of DNA -> same mean





Normalisation Results

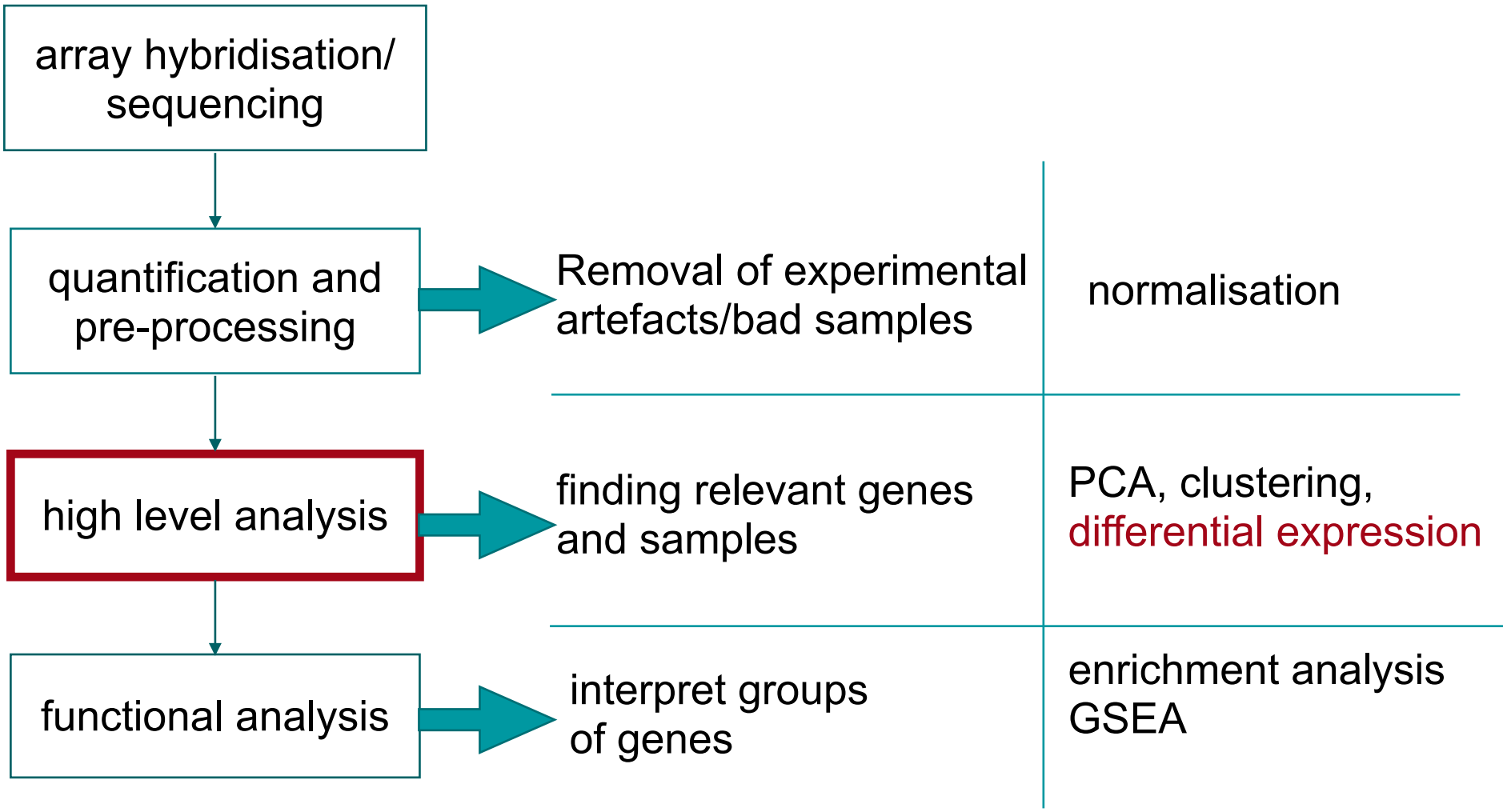
Application of BetweenArray normalisation from limma package



Quantification/Pre-processing - Resume

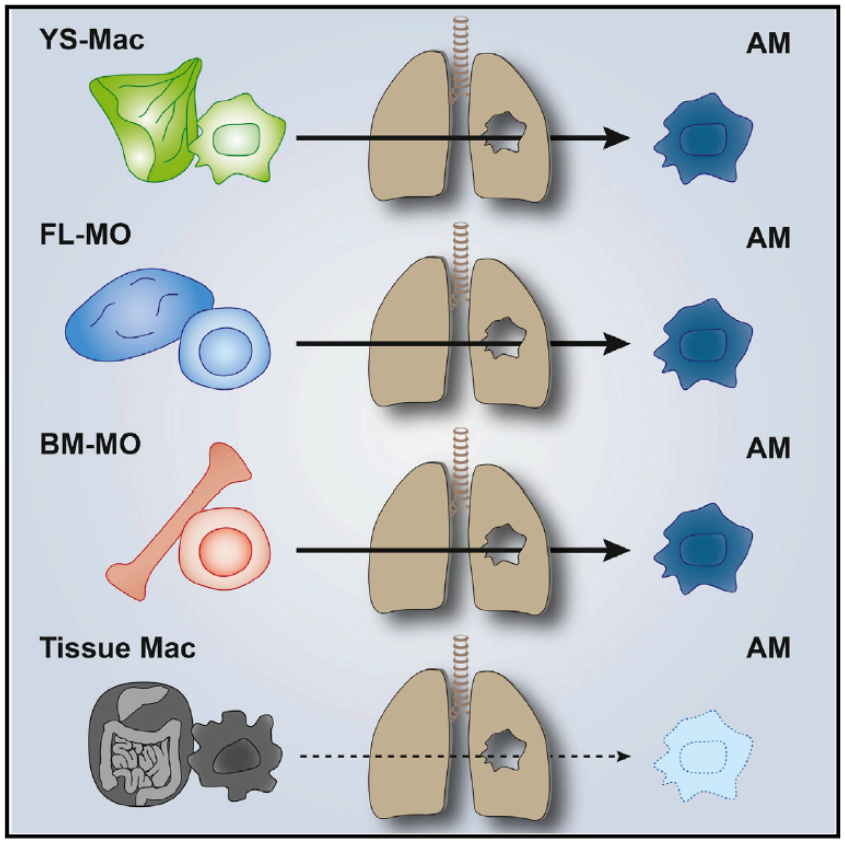
- Normalisation is important to confirm the quality and consistency of data
- Boxplots should also be performed after all steps to assure data standards
- Exclusion of "bad samples" has positive effect on downstream analysis
- **When in doubt, consult a bioinformatician!**

Bioinformatics - Gene Expression Analysis



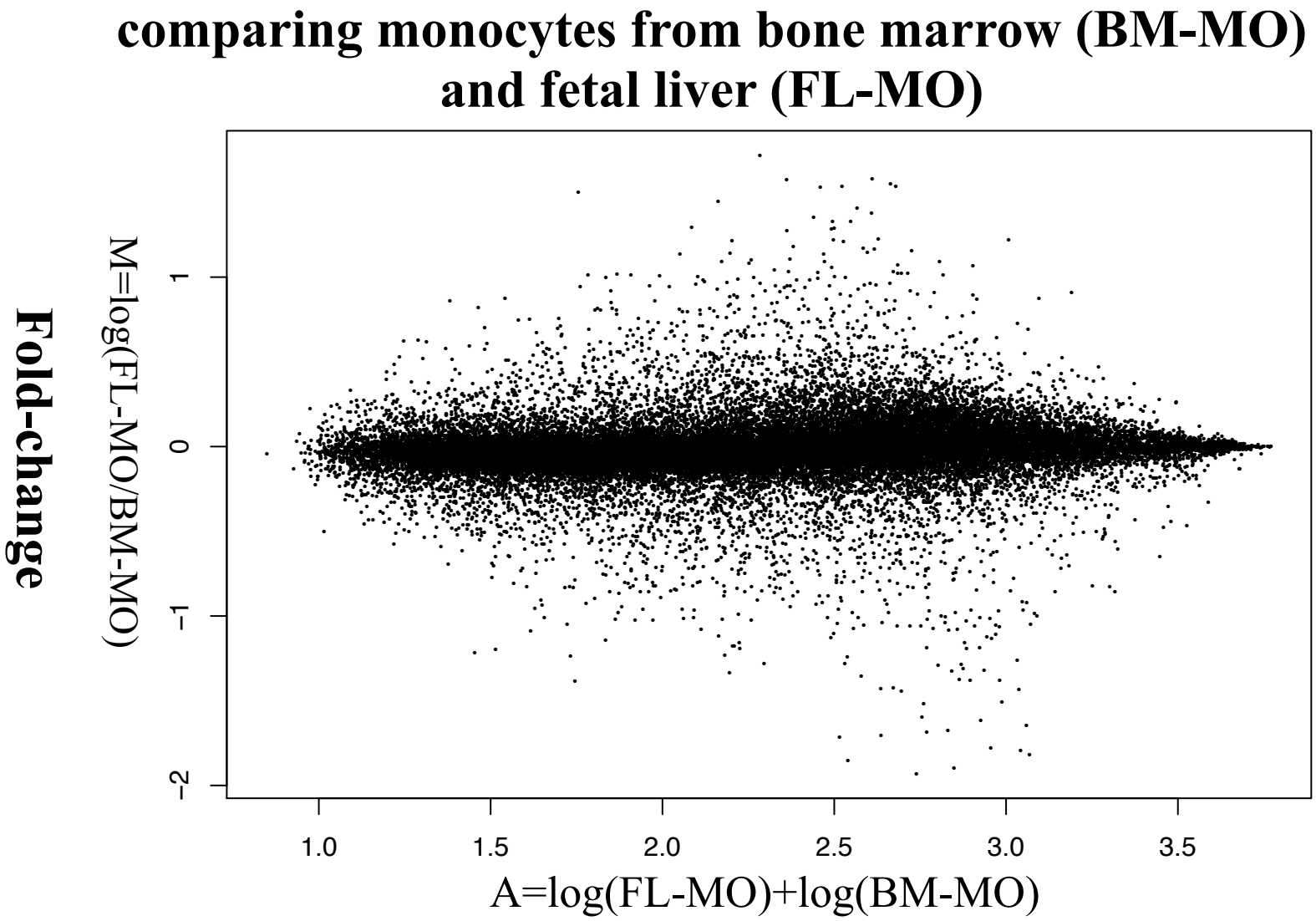
Differential Expression Analysis

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



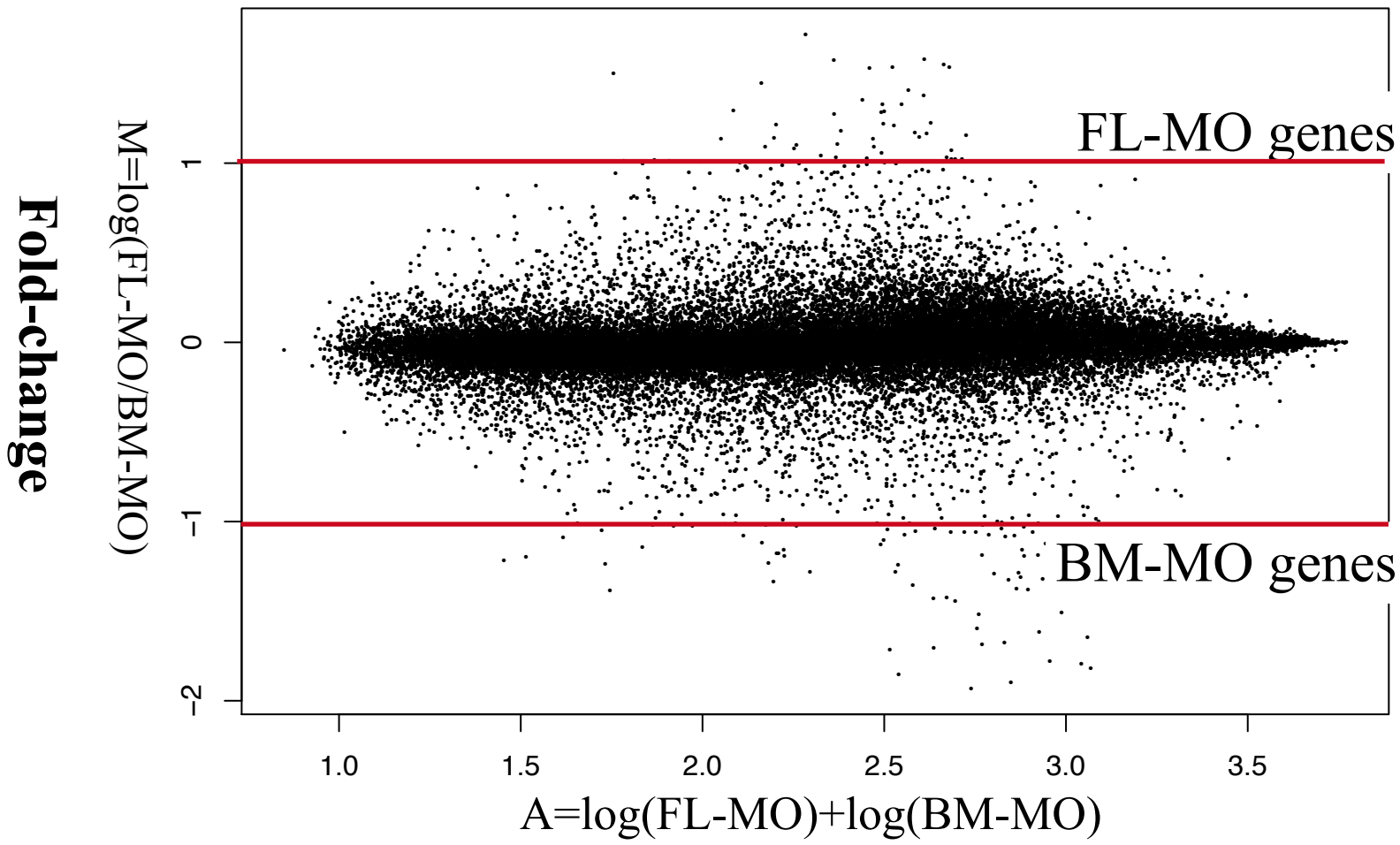
Source: van de Laar, et al. 2016, Immunity, 2016.

Differential Expression - Example



Differential Expression - Example

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

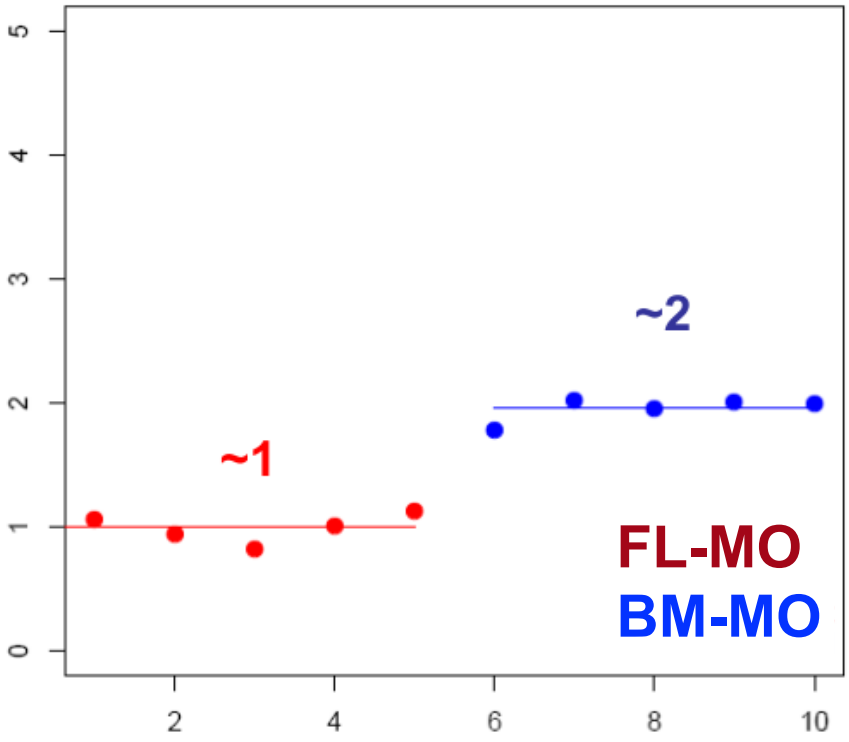


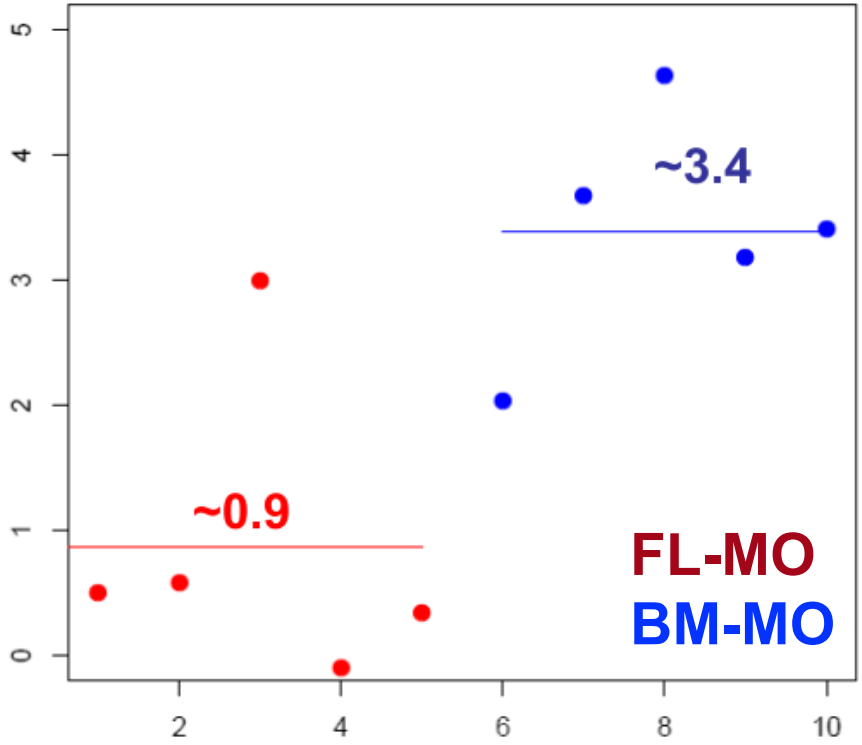
Problems - Fold change

- 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

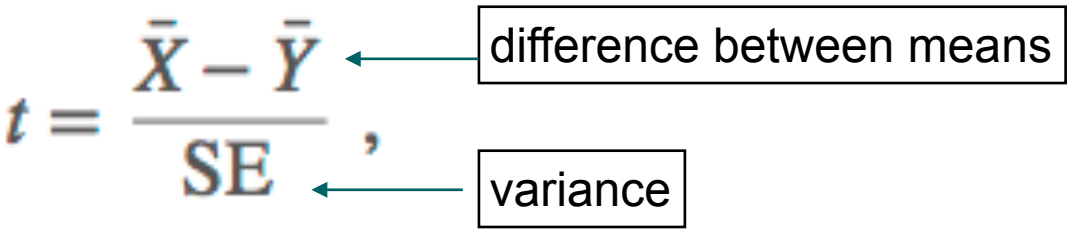
Mean vs. variability





Student T-test

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

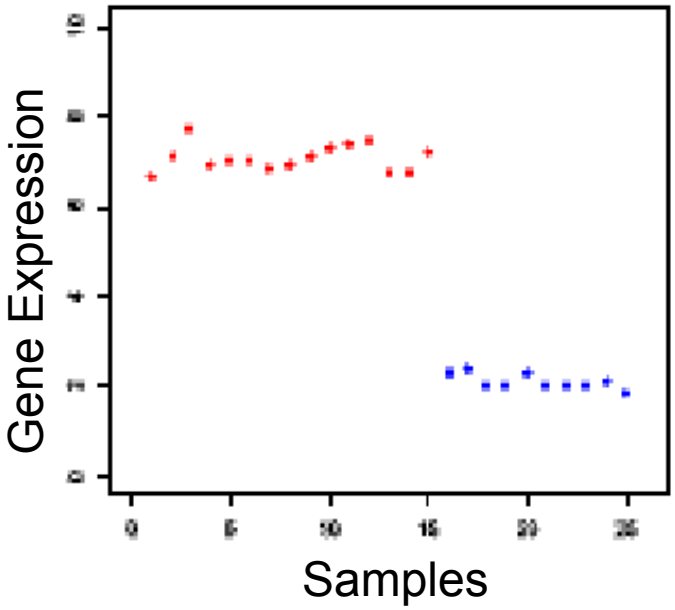


Test the hypothesis $H_0 : X - Y = 0$ No diference

$H_1 : X - Y \neq 0$ Difference

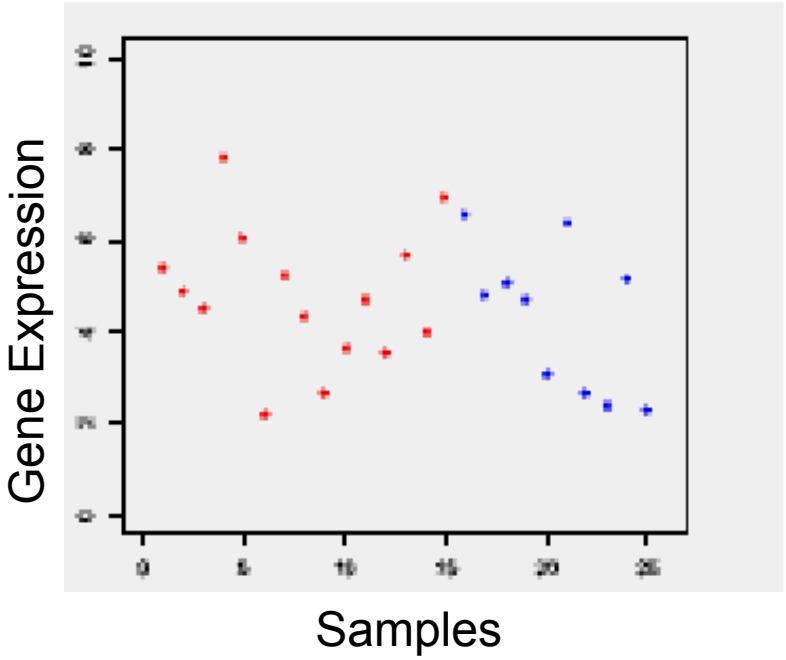
Examples

Change: HIGH
Variance: SMALL



T huge

Change: SMALL
Variance: HIGH

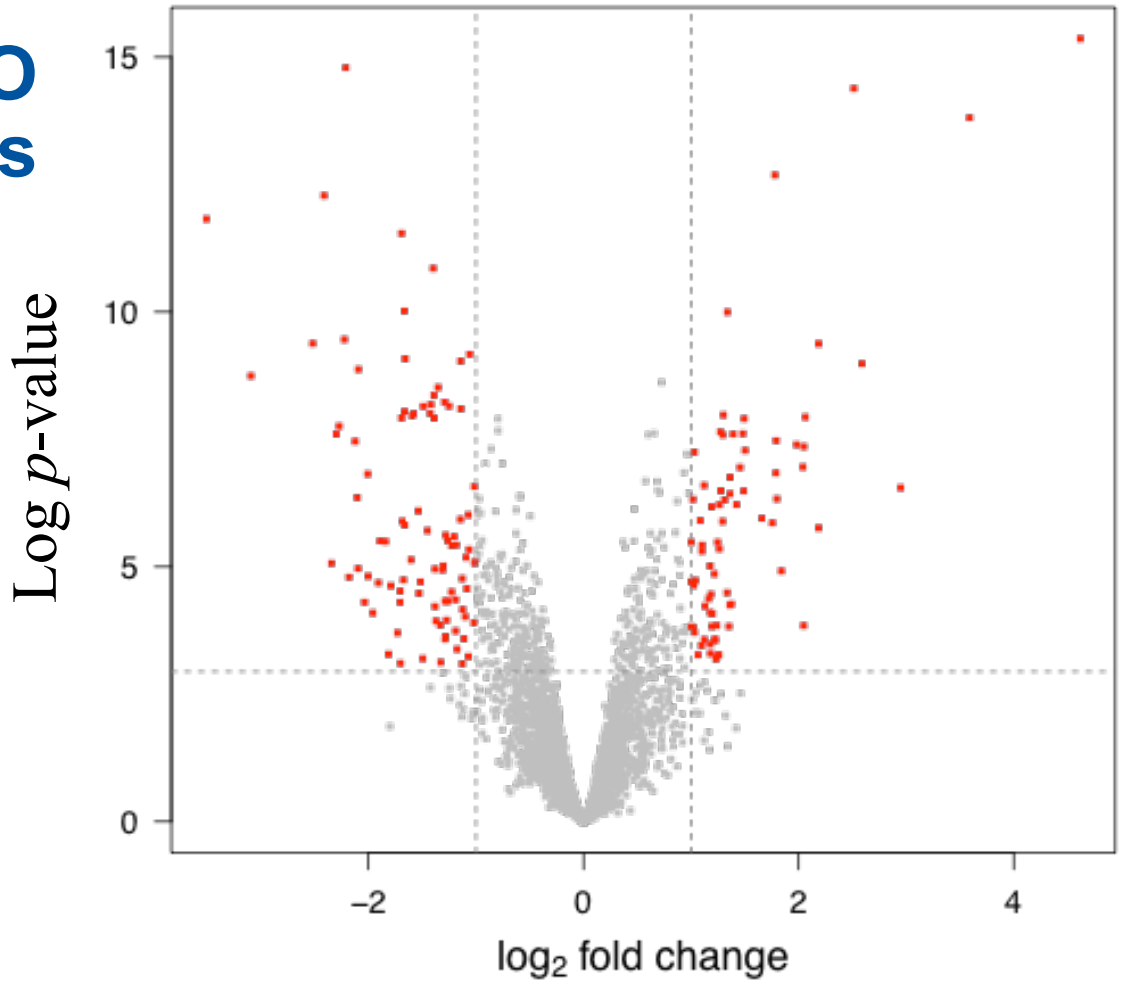


T ~ 0

Results - FL-MO vs. BM-MO

Volcano Plot - combine p-value and fold change

FL-MO genes



BM-MO genes

Multiple Test Correction

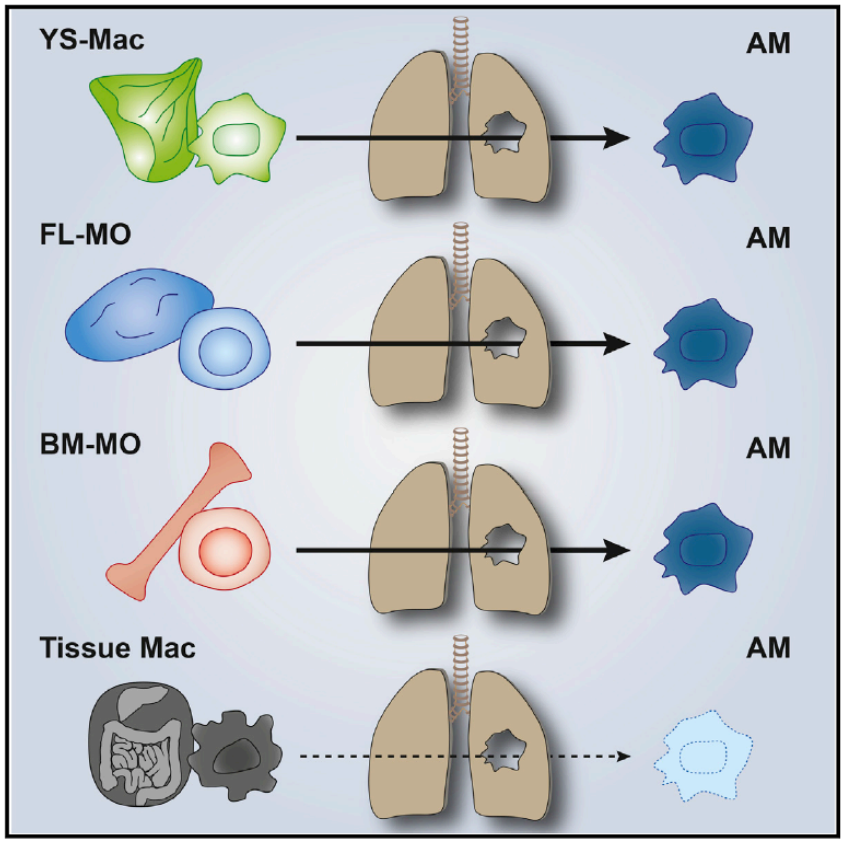
- With a p-value of 0.01, we expect to make one mistake every 100 tests
  - If 20.000 genes are tested then errors might happen for 200 genes !!!
- To solve this, a multiple test correction method is necessary (i.e. Benjamini-Hochberg)
  - It is based on the false discovery rate, i.e. the proportion 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!

Differential Expression Analysis

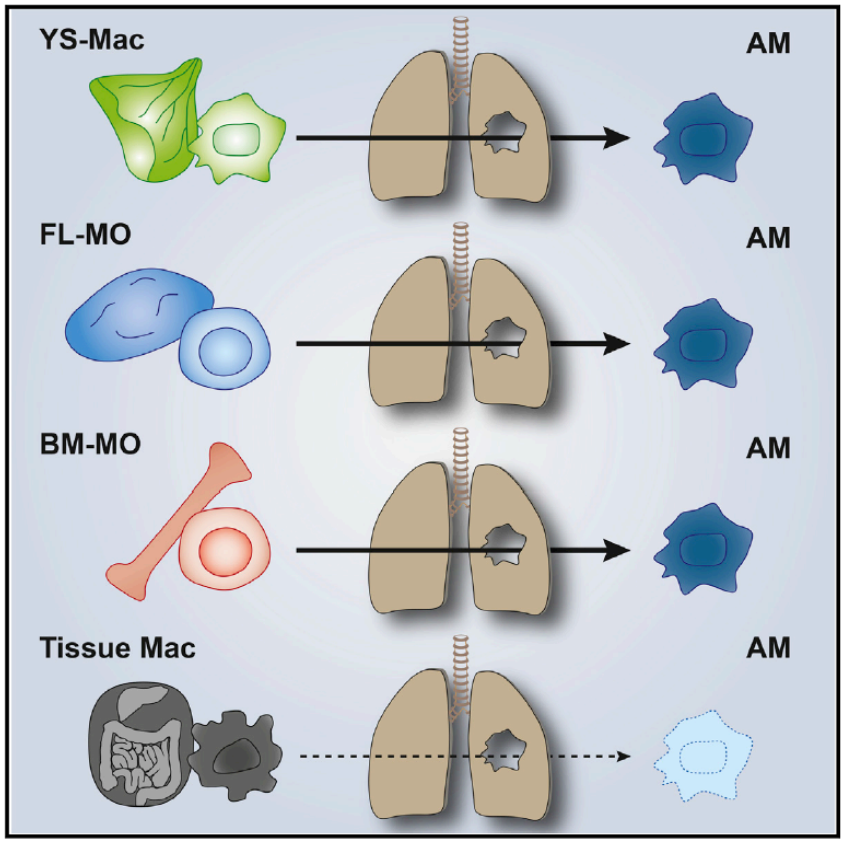
- Identify genes related to a particular condition
  - example - van de Laar, et al. 2016, Immunity, 2016.
- We will consider:
  - You Sac Macrophages (YS-Mac)
  - Fetal Liver Monocytes (FL-MO)
  - Bone Marrow Monocytes (BM-MO)
    - 4 replicates per condition



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

Differential Expression Analysis

- This data is deposited in the public repository GEO under accession number GSE76999
- This can be found at the materials and methods of papers.
- GEO - public database with raw, pre-processed data and experimental details of expression (and other omics) experiments.



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

GEO - van de Laar, et al. 2016

Series GSE76999 Query DataSets for GSE76999

GEO ID

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
Organism	Mus musculus
Experiment type	Expression profiling by array
Summary	Tissue-resident macrophages can derive from yolk sac macrophages, fetal liver monocytes or adult bone marrow monocytes. Whether these precursors can give rise to transcriptionally identical alveolar macrophages is unknown. Here, we transferred traceable yolk sac macrophages, fetal liver monocytes, adult bone marrow monocytes or adult alveolar macrophages as a control, into the empty alveolar macrophage niche of neonatal Csf2rb-/- mice. All precursors efficiently colonized the alveolar niche and generated alveolar macrophages that were transcriptionally almost identical, with only 22 genes that could be linked to their origin. Underlining the physiological relevance of our findings, all transfer-derived alveolar macrophages self-maintained within the lungs for up to 1 year and durably prevented alveolar proteinosis. Thus, precursor origin does not affect the development of functional self-maintaining tissue-resident macrophages.
Overall design	CD45.1+CD45.2+ yolk sac macrophages, fetal liver monocytes, adult bone marrow monocytes or adult alveolar macrophages from the bronchoalveolar lavage were sorted from wild type CD45.1+CD45.2+ mice of indicated ages. From part of these samples RNA was isolated. The other part was transferred intranasally into the lungs of neonate Csf2rb-/- mice. 6 weeks post-transfer, transfer-derived CD45.1+CD45.2+ alveolar macrophages were sorted from the bronchoalveolar lavage. Wild type CD45.1+CD45.2 alveolar macrophages from the bronchoalveolar lavage of 6 week old mice were sorted as control. 36 samples (arrays) in total. RNA was isolated, amplified with Nugene pico kit, converted to cDNA and then hybridised on Affymetrix GeneChip Mouse Gene 1.0 ST Arrays.
Contributor(s)	van de Laar L, Saelens W, De Prijck S, Martens L, Scott CL, Van Isterdael G, Hoffmann E, Beyaert R, Saeys Y, Lambrecht BN, Guilliams M
Citation(s)	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

Information about the study

GEO - van de Laar, et al. 2016

Sample GSM2042244	Query DataSets for GSM2042244
Status	Public on Mar 01, 2016
Title	Monocyte extracted from adult (wk6-12) Bone Marrow, biological replicate 1
Sample type	RNA
Source name	Monocyte, extracted from Bone Marrow (BM)
Organism	Mus musculus
Characteristics	strain: C57BL/6 tissue: Bone Marrow age: wk6-12
Treatment protocol	not applicable
Growth protocol	Tissues were isolated from the mice at the indicated ages.
Extracted molecule	total RNA
Extraction protocol	Single cell suspensions were prepared by organ digestion (yolk sac and fetal liver) with 1 mg/ml collagenase A and 10 U/ml DNA (30 and 5 minutes at 37oC), crushing (bones) or flushing of the lungs (broncholaveolar lavage). 2x104 cells were FACS purified into RLT buffer (Qiagen) containing 10 ml/ml 2-mercaptoethanol. RNA was isolated using the RNA isolation kit micro (Qiagen no74034).
Label	biotin
Label protocol	Affymetrix WT Terminal Labeling Kit
Hybridization protocol	Standard Affymetrix protocol. cDNA was hybrised on Affymetrix GeneChip Mouse Gene 1.0 ST Arrays (GPL6246).
Scan protocol	Affymetrix Gene ChIP Scanner 3000 7G
Description	Monocyte extracted from Bone Marrow
Data processing	Data were processed using Bioconductor. Normalisation was done by RMA. MoGene-1_0-st-v1.r4.pgf MoGene-1_0-st-v1.r4.mps

ID of array

name of condition

details

GEO - van de Laar, et al. 2016

Submission date	Jan 20, 2016
Last update date	Jul 13, 2018
Contact name	Martin Guilliams
Organization name	VIB-University of Ghent
Department	VIB Inflammation Research Center
Street address	Technologiepark 927
City	Ghent
ZIP/Postal code	9000
Country	Belgium

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 PRJNA309234

Analyze with GEO2R

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

Supplementary file	Size	Download	File type/resource
GSE76999_RAW.tar	135.3 Mb	(http)(custom)	TAR (of CEL)

Raw data provided as supplementary file

array used

single experiments

raw data

omics

RWTH AACHEN UNIVERSITY

Using GEO2R

- Select the data of interest:
  - 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

- GEO2R will provide you an **R code** to perform normalisation and DE analysis.

R Programming Language

- Script based Programming language
- Focus on statistical data analysis
- Open source
- Contributing packages
  - Bioconductor (bioinformatics functions)
  - ggplot (plotting functions)
  - …

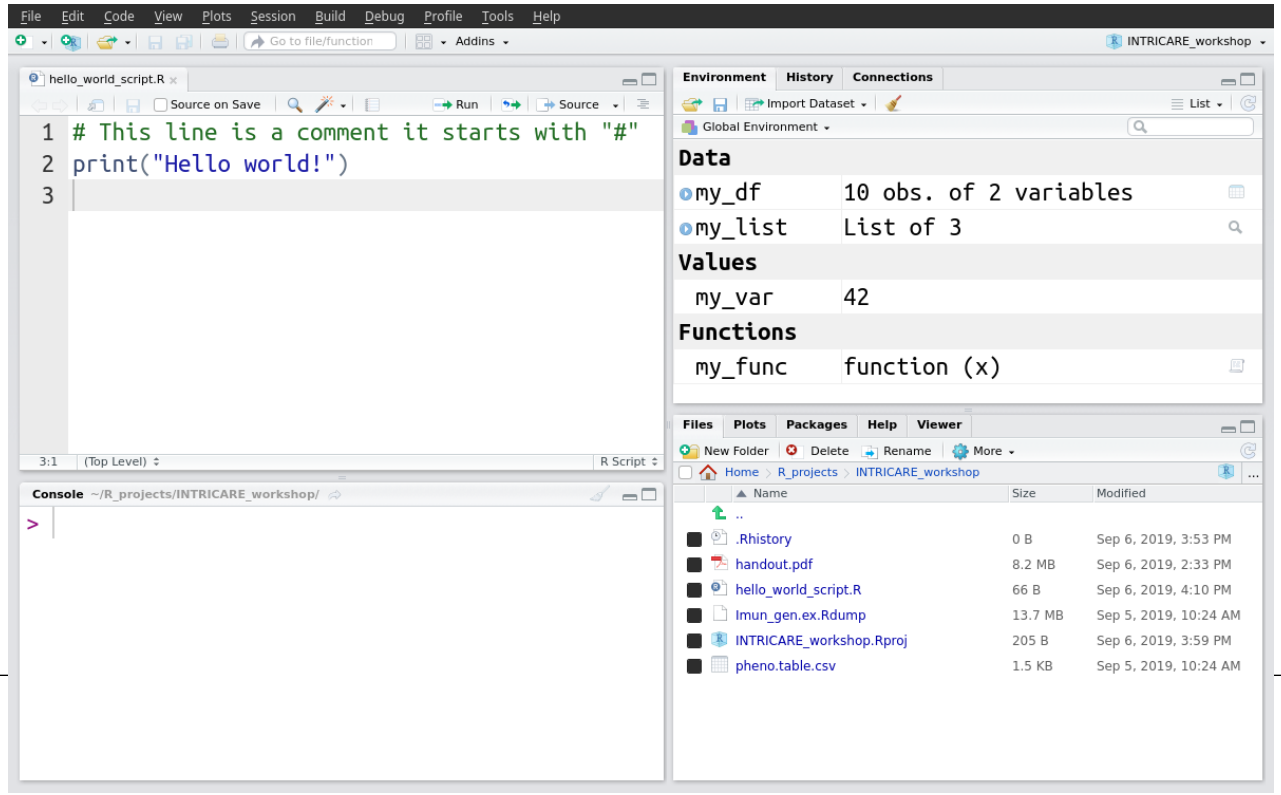
http://www.r-project.org/

RStudio - Getting Started

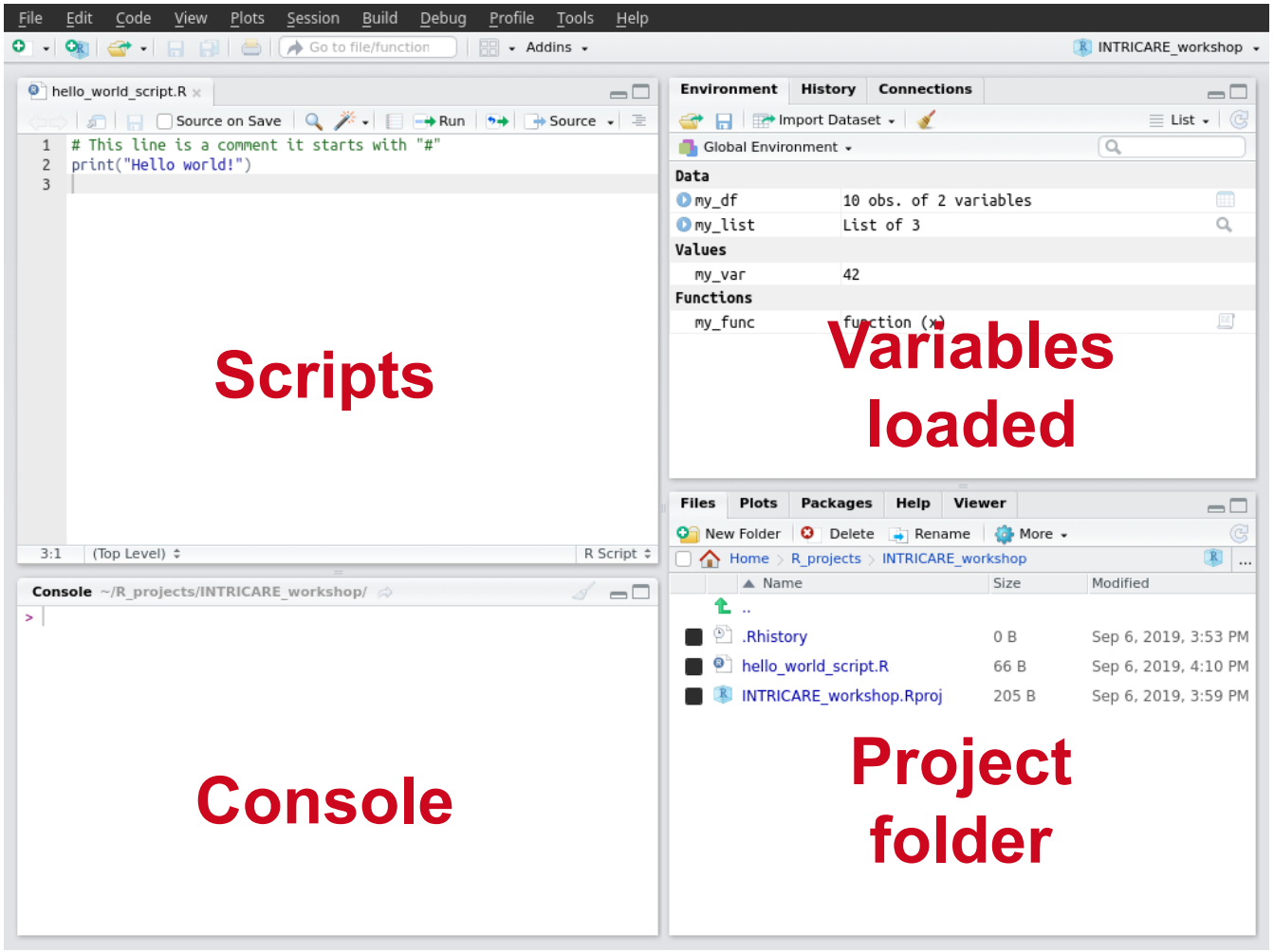
- Install RStudio

  https://www.rstudio.com

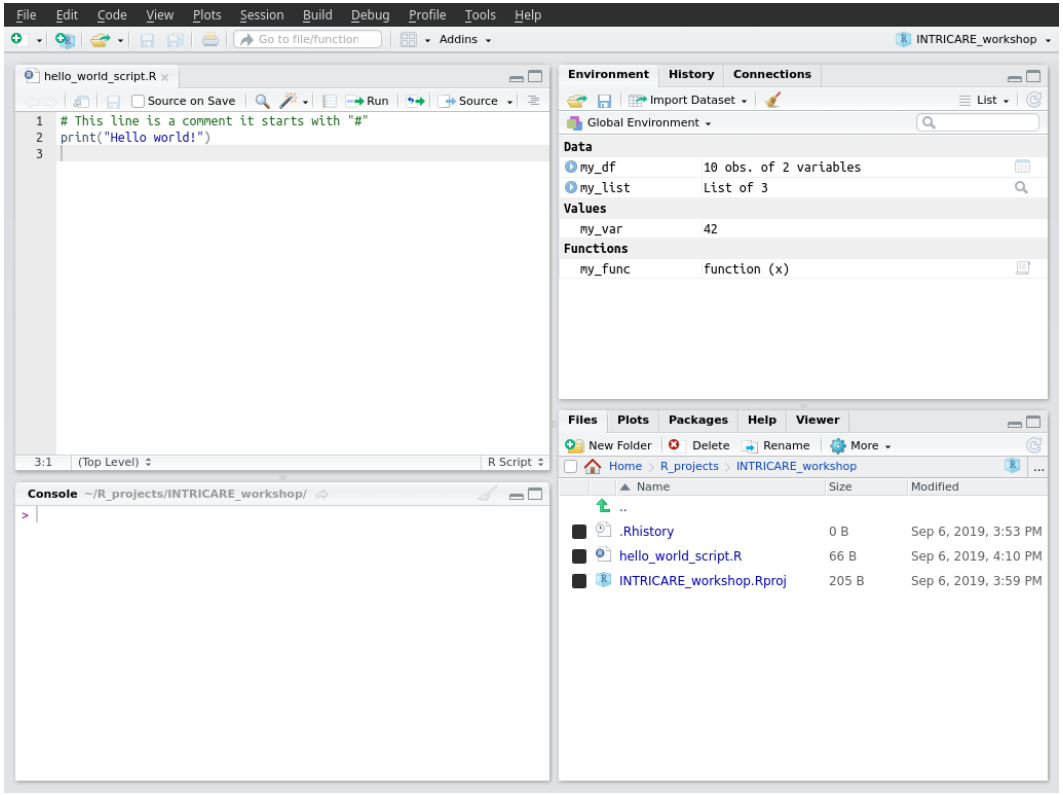
- Run RStudio



RStudio - Organisation

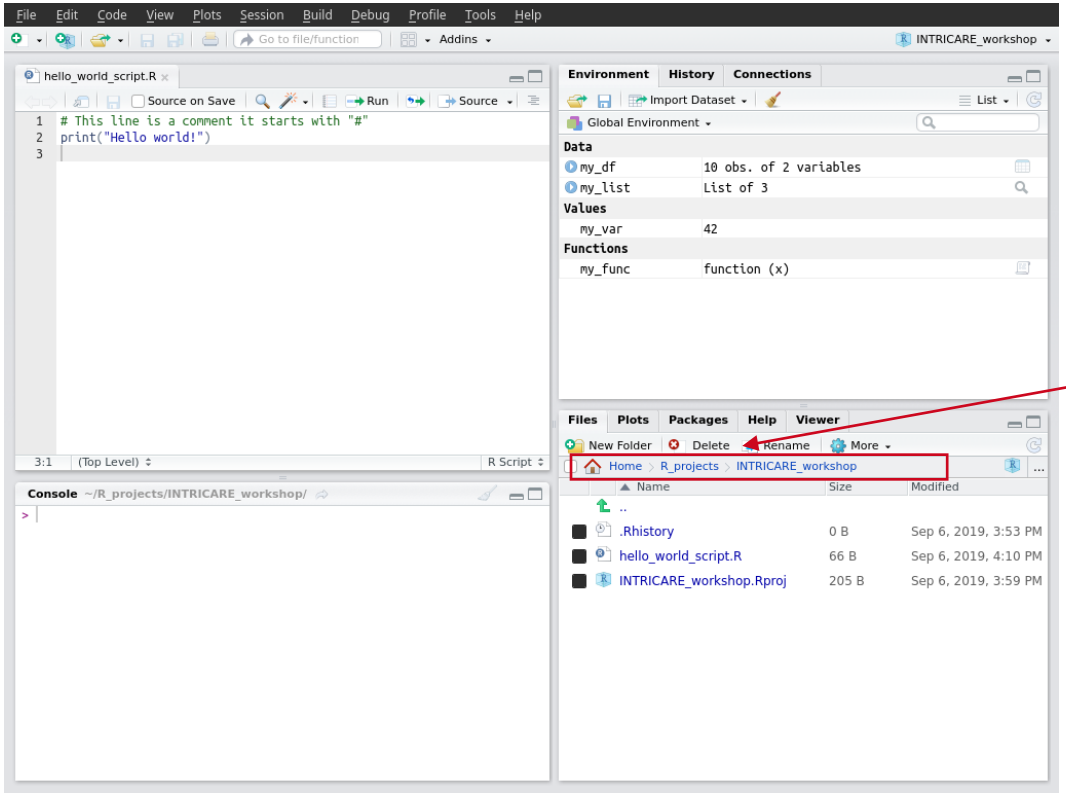


RStudio - Configure Project Directory



We need to configure the project directory:

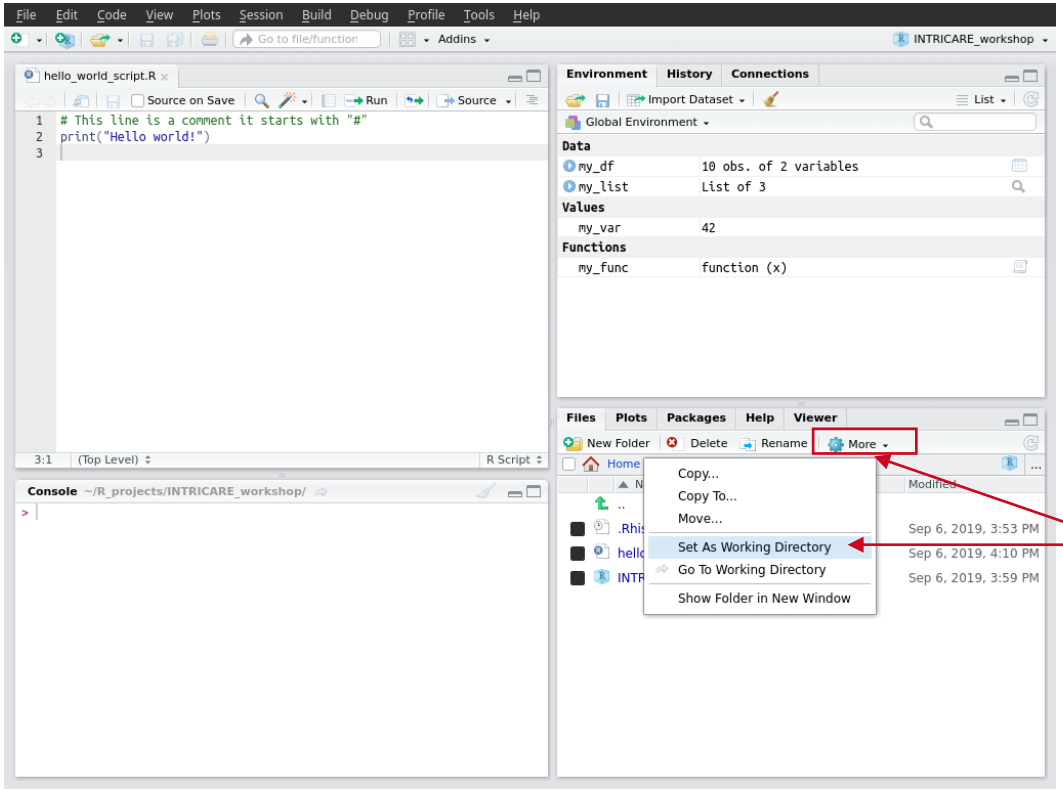
RStudio - Configure Project Directory



We need to configure the project directory:

1 - navigate until folder with course files

RStudio - Configure Project Directory

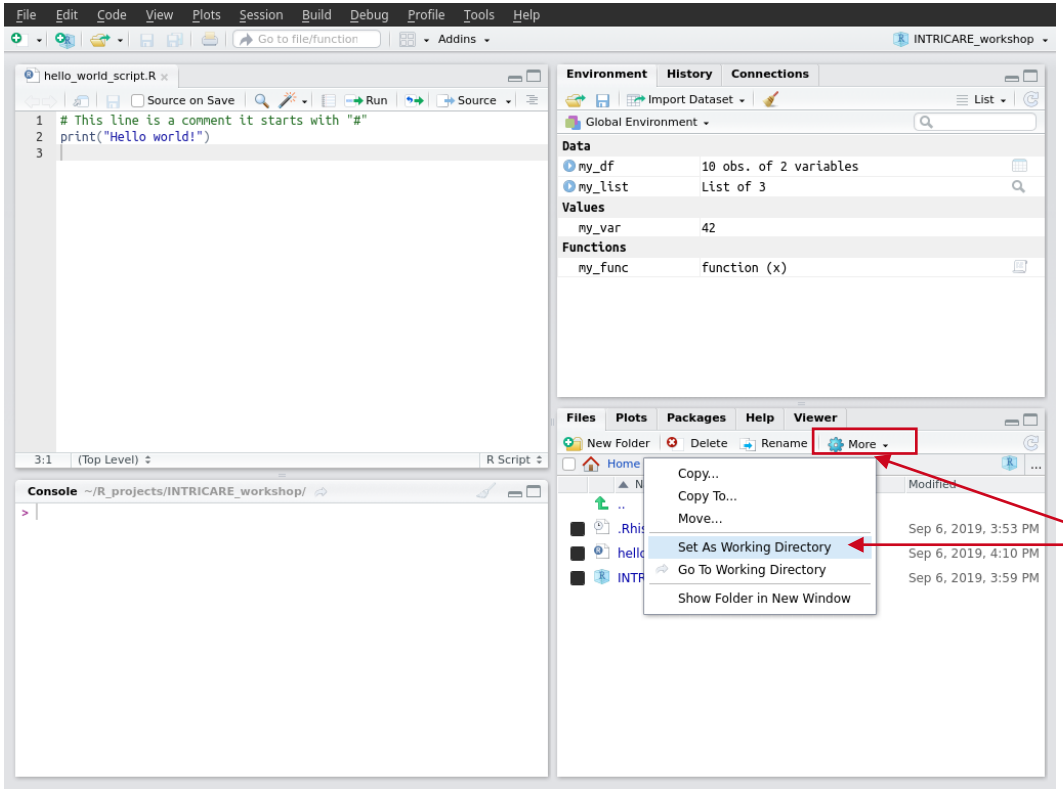


We need to configure the project directory:

1 - navigate until folder with course files

2 - select the "More" option and "Set as Working Directory"

RStudio - Configure Project Directory



We need to configure the project directory:

1 - navigate until folder with course files

2 - select the "More" option and "Set as Working Directory"

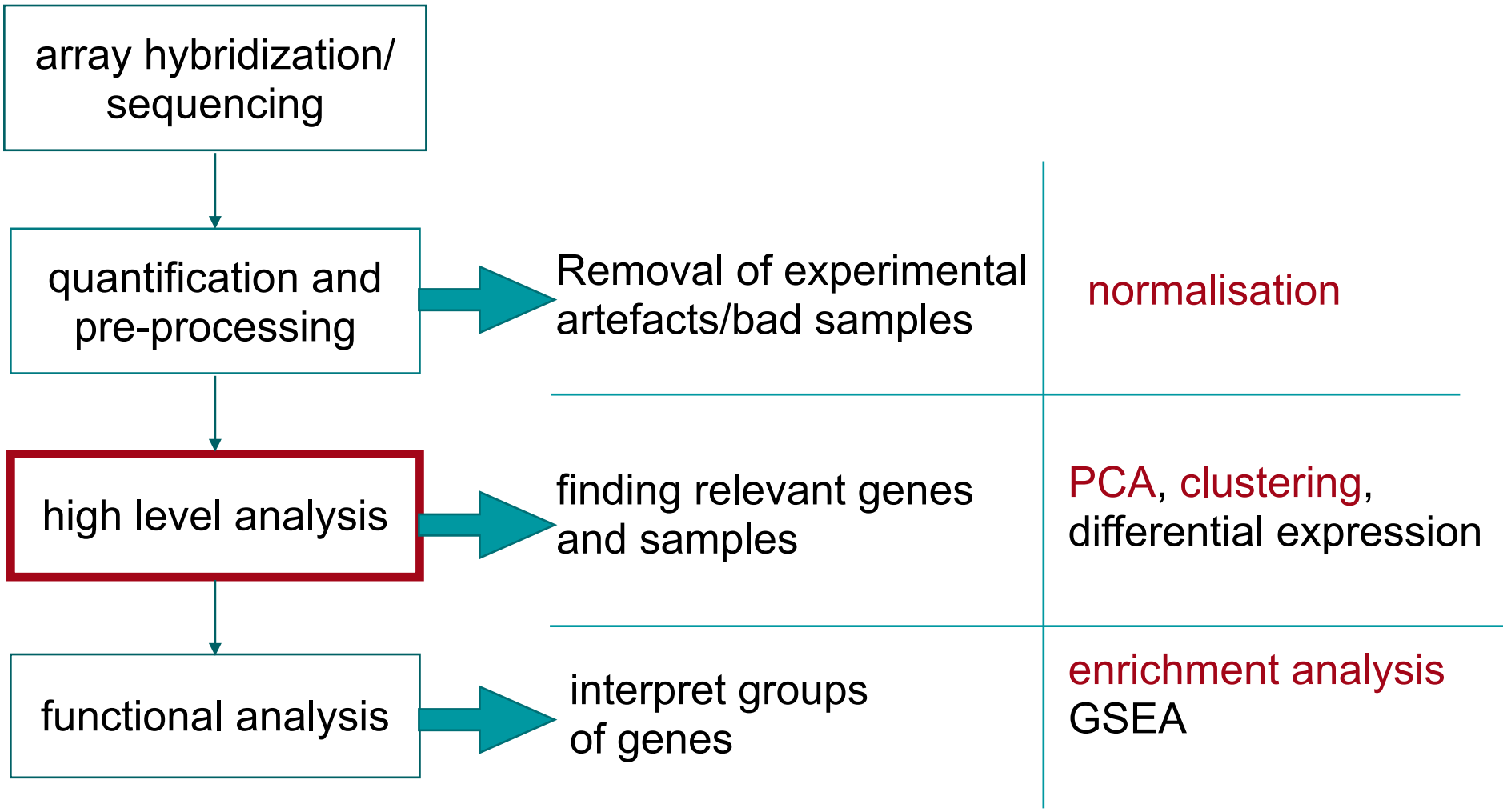
Now R Studio knows where to find files !

Hands on!

1. Download the data from
- https://www.costalab.org/promotionskolleg-intro-to-gene-expression-analysis/

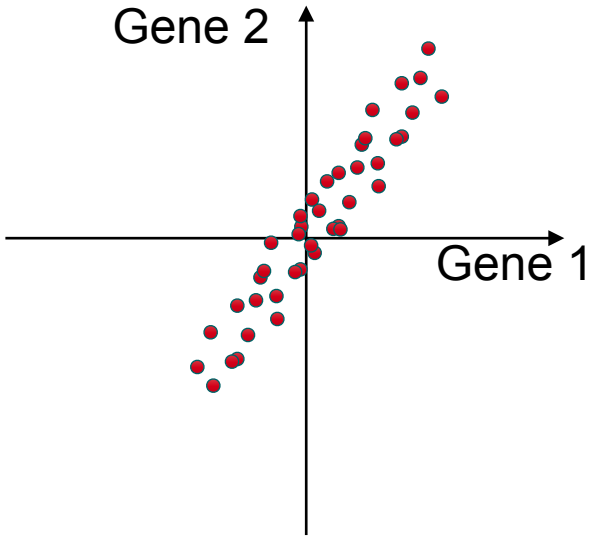
2. Extract files from zip file.
3. Follow instructions from Handout (Step 1 to 3)
- https://www.costalab.org/promotionskolleg-intro-to-gene-expression-analysis/

Bioinformatics - Gene Expression Analysis



Principal Component Analysis

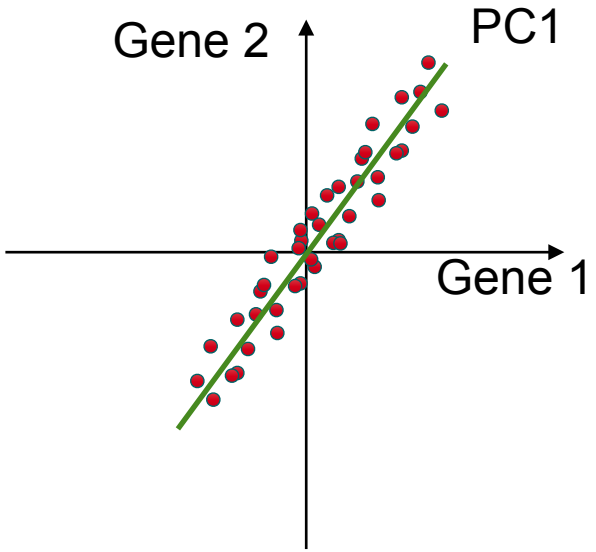
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



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

Principal Component Analysis

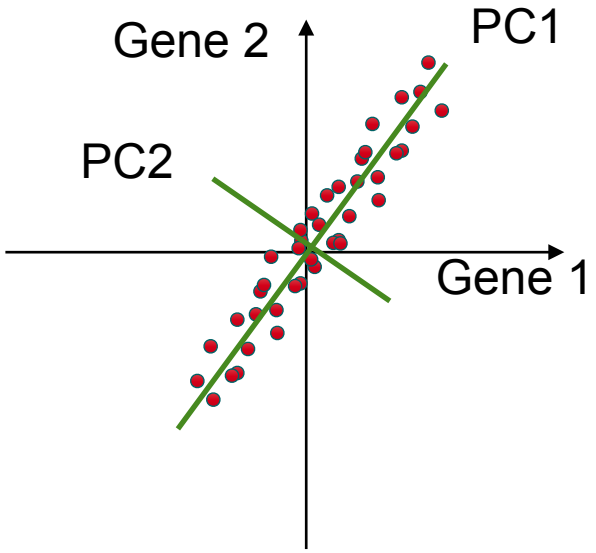
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



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

Principal Component Analysis

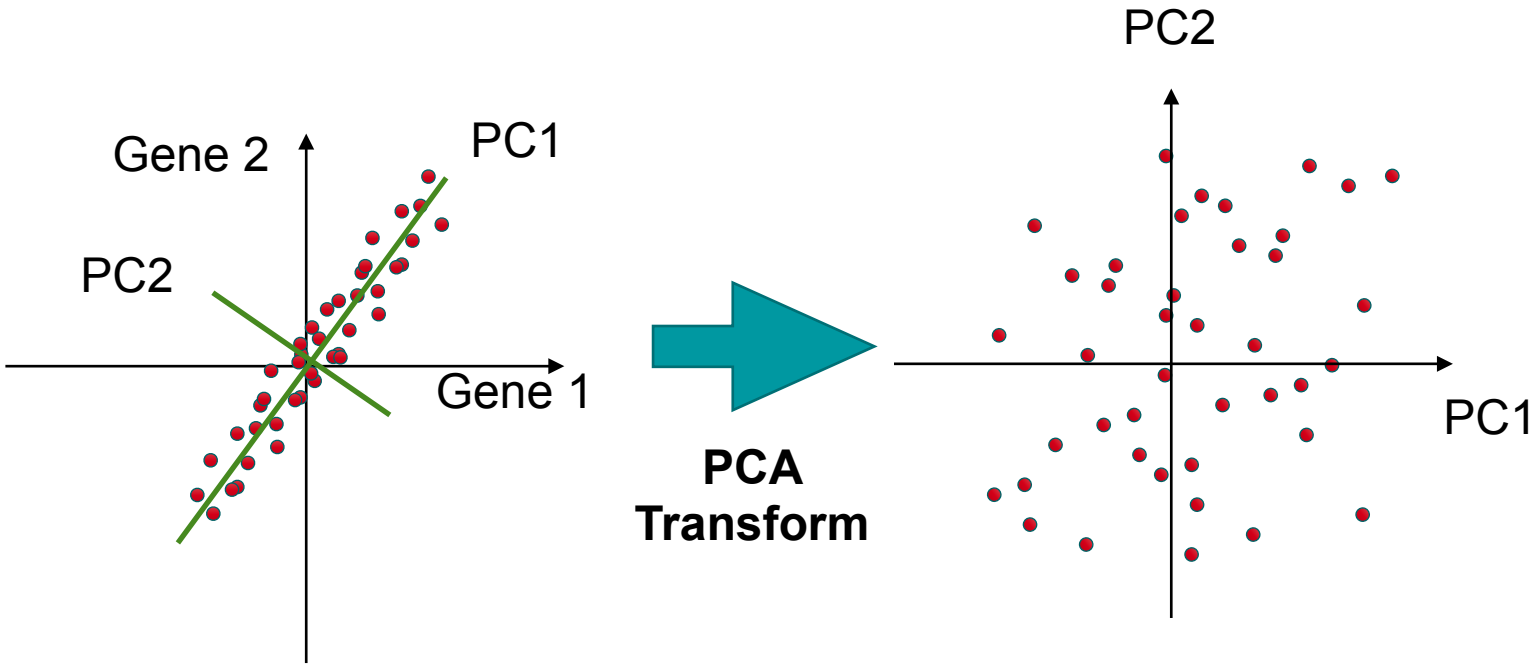
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



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

Principal Component Analysis

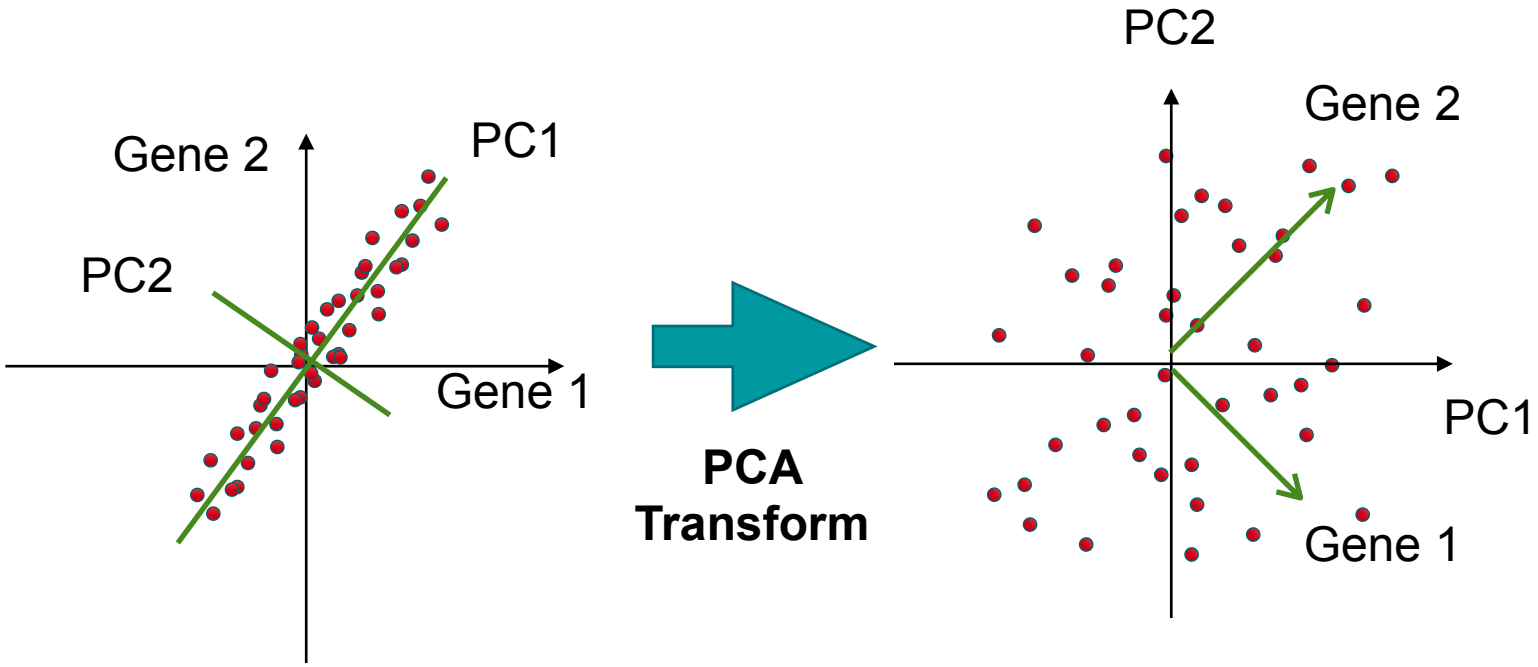
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



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

Principal Component Analysis

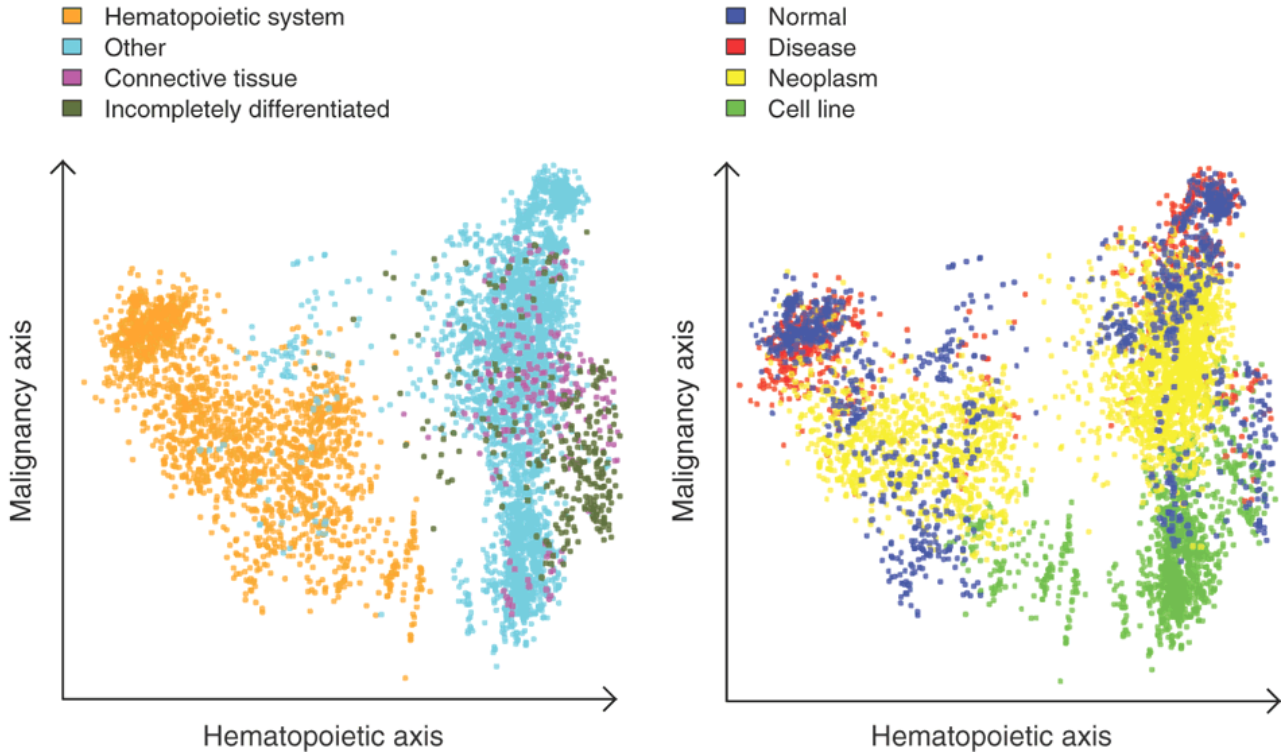
- **method for dimension reduction**
  - **find combination of genes explaining cells with distinct expression**
- **finding directions with highest variance**



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

Gene Expression - PCA Example 1

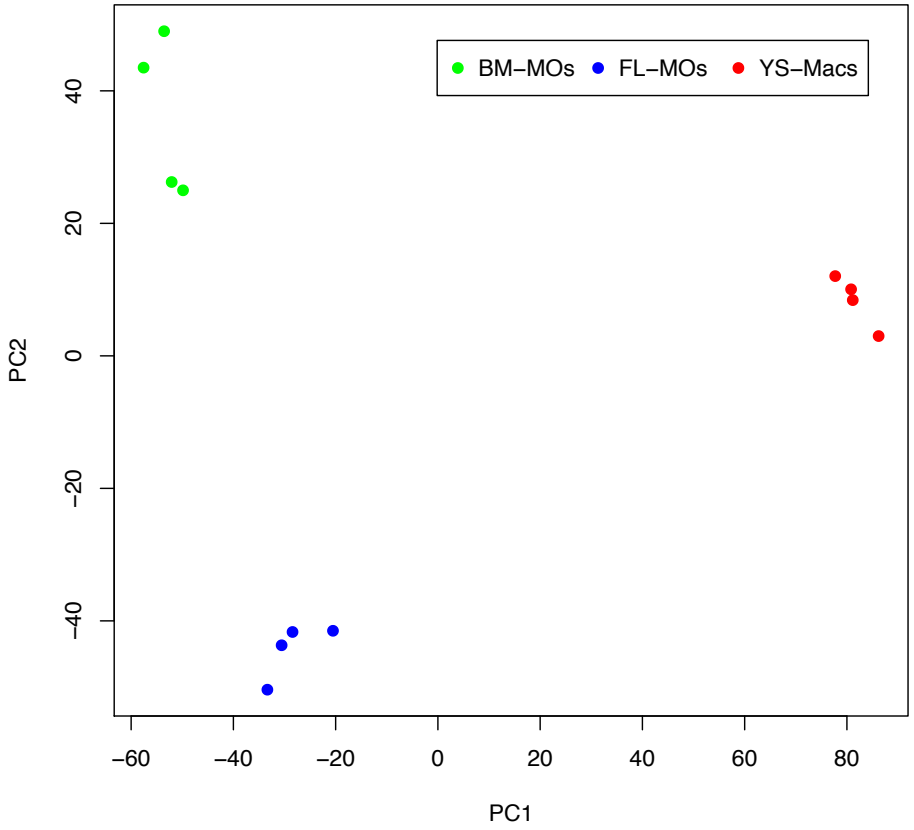
Can be interpreted as a computational FACs sorting (without knowing the markers)



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

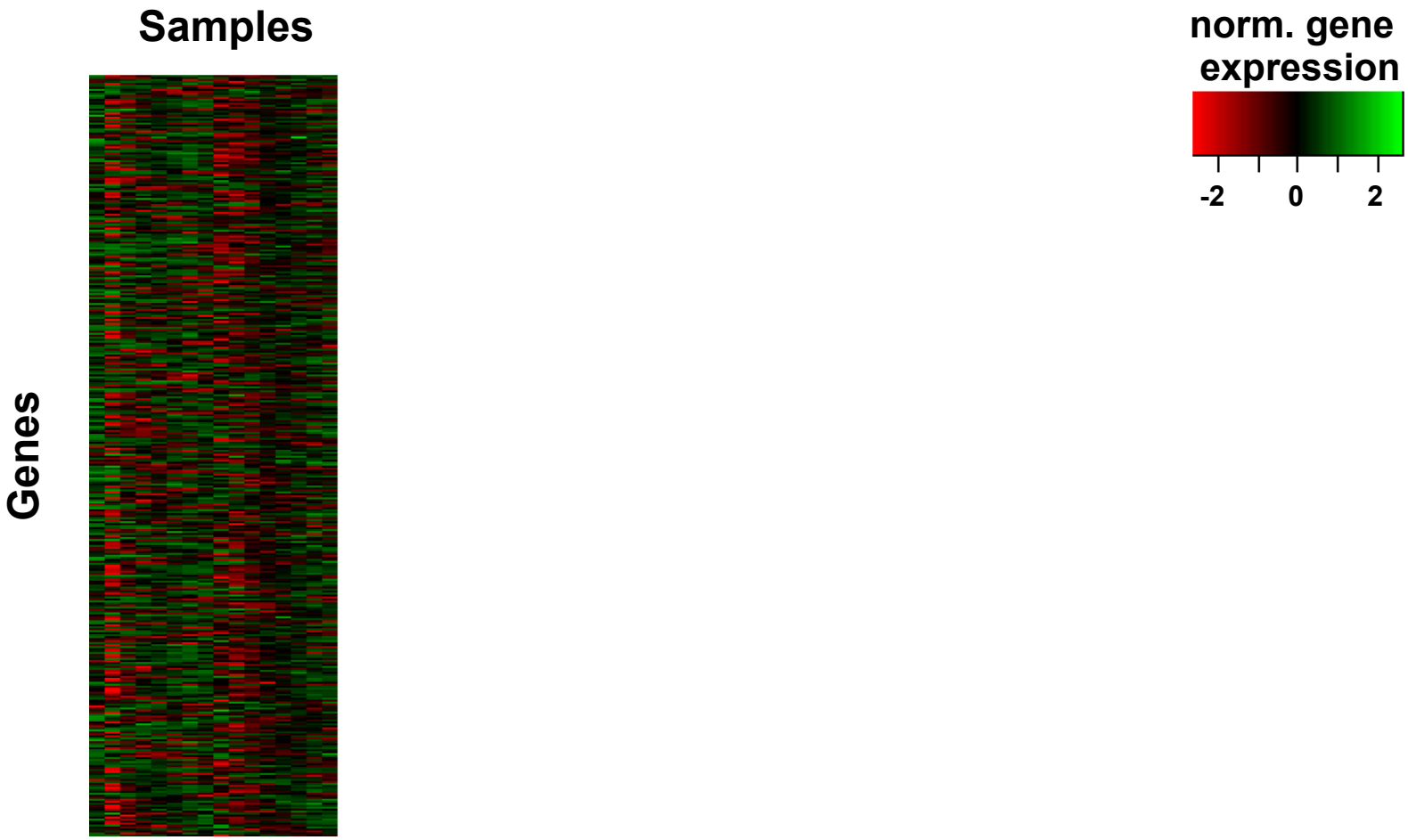


First 2 PCs van de Leer, 2016 data

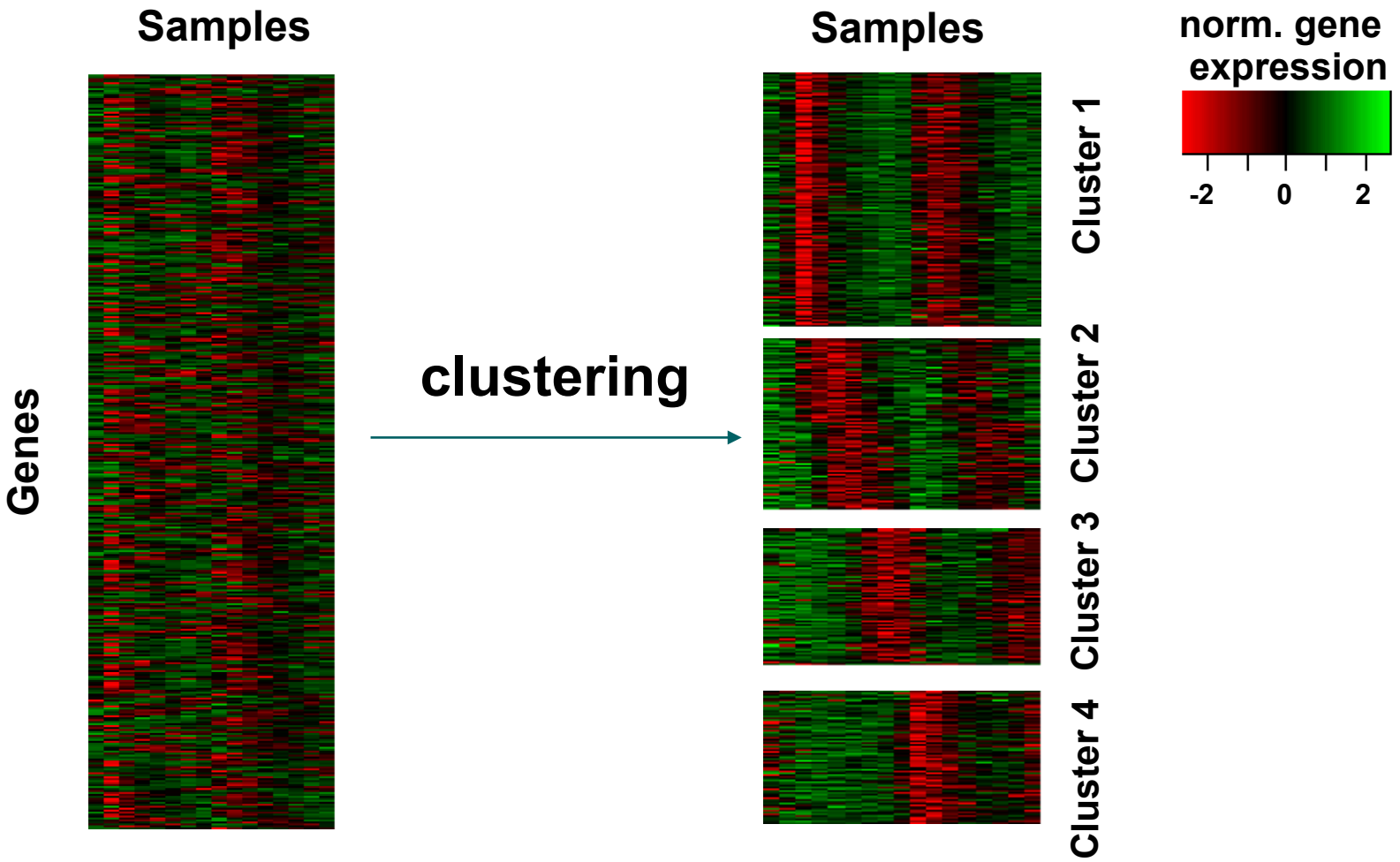
PCA Analysis - Conclusions

- PCA allows a “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

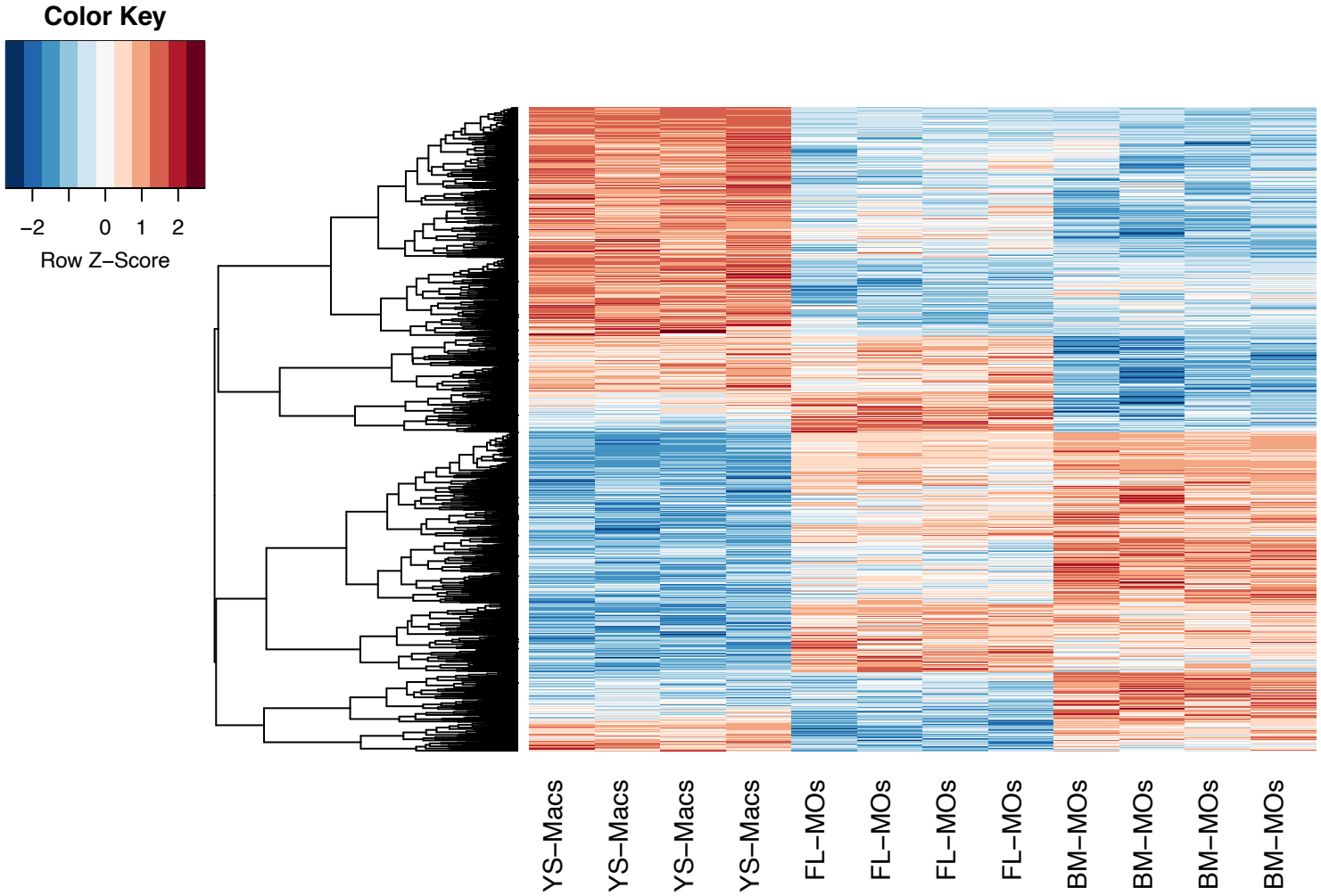


Clustering / Heatmaps



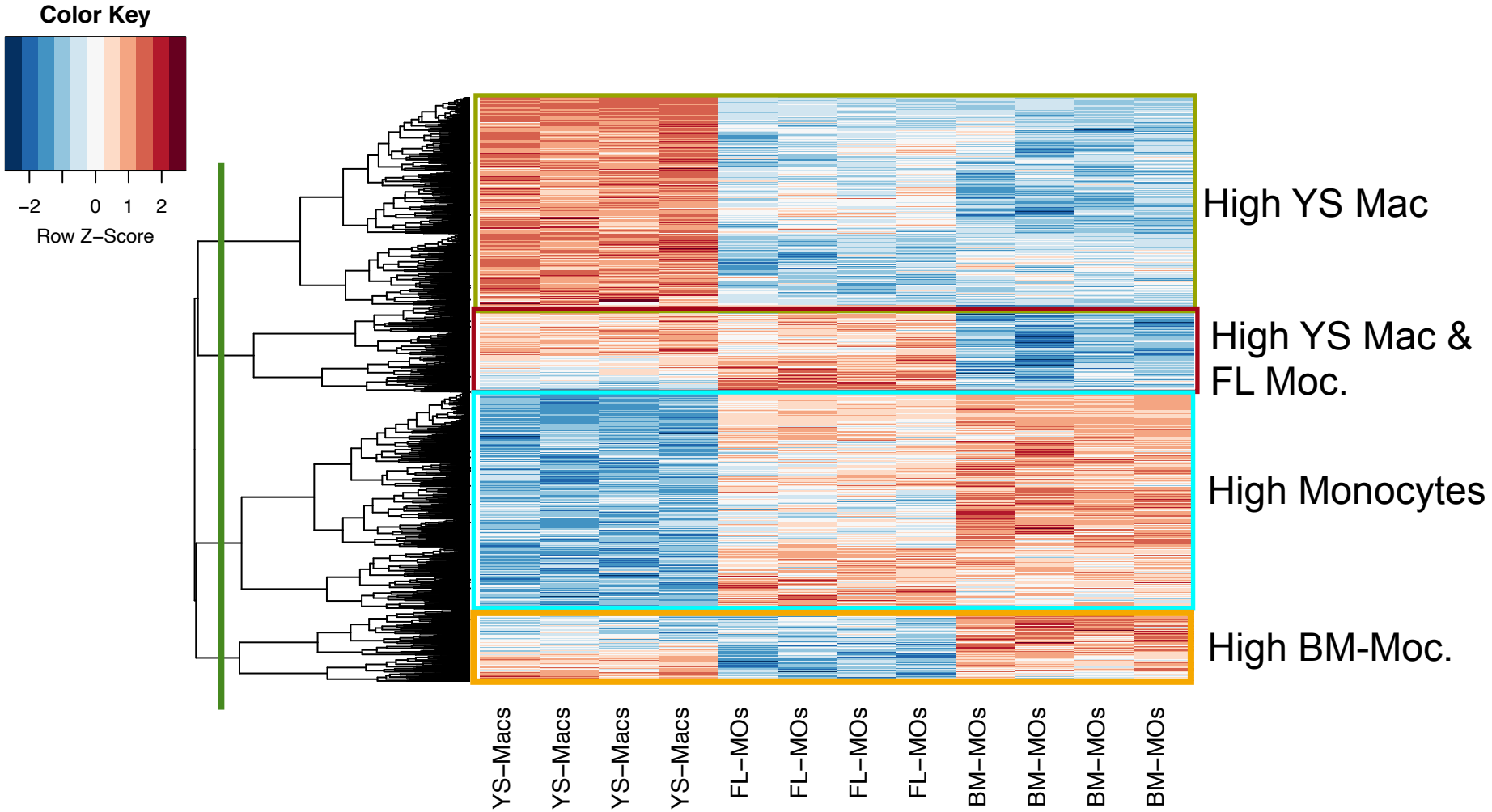
clustering methods: k-means, hierarchical clustering, …

Hierarchical Clustering



distance metric - Pearson correlation recommended

Hierarchical Clustering



distance metric - Pearson correlation recommended

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 automatically 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

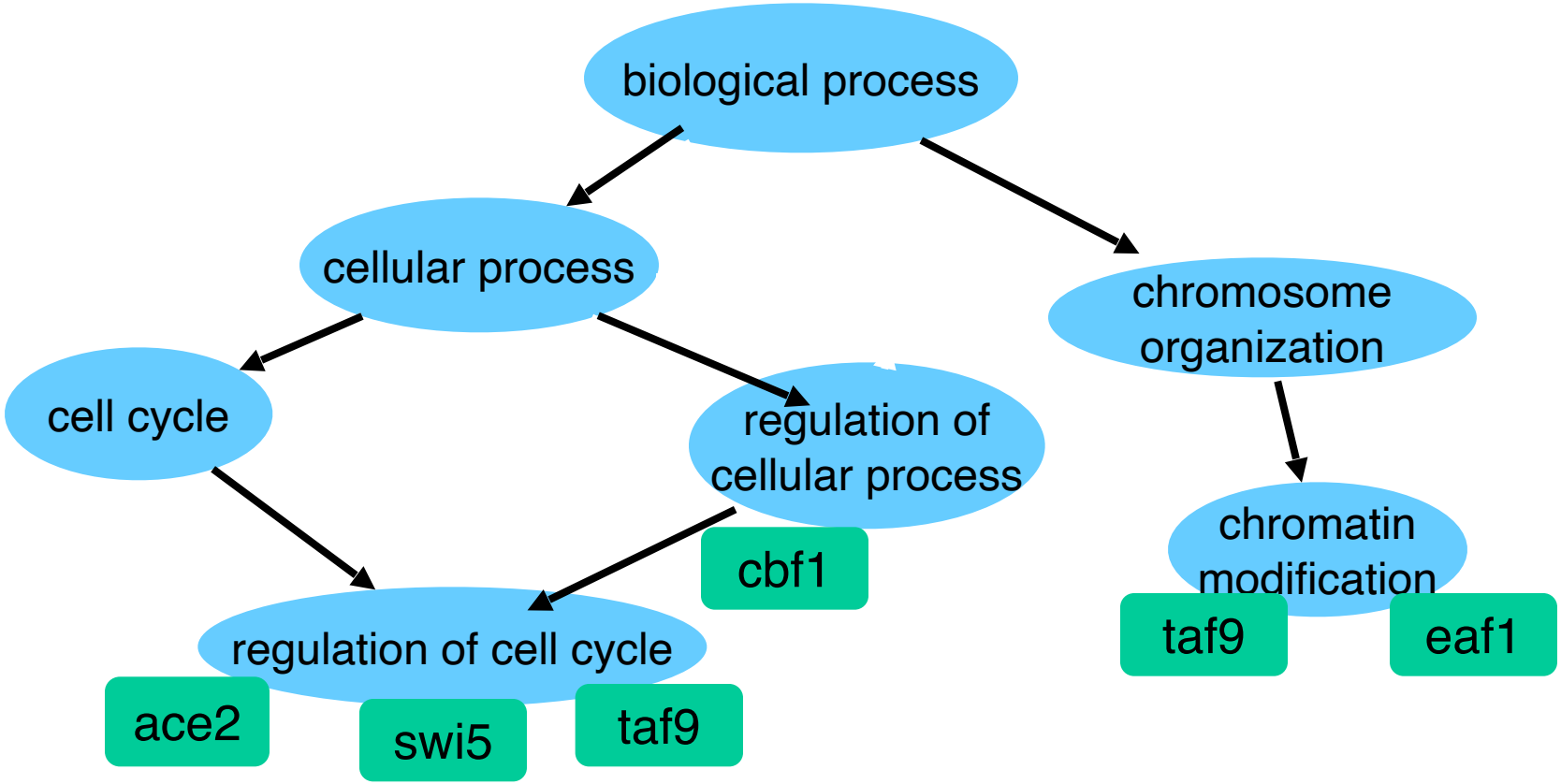
Controlled vocabulary to describe gene and gene product attributes in any organism

Formed by three ontologies

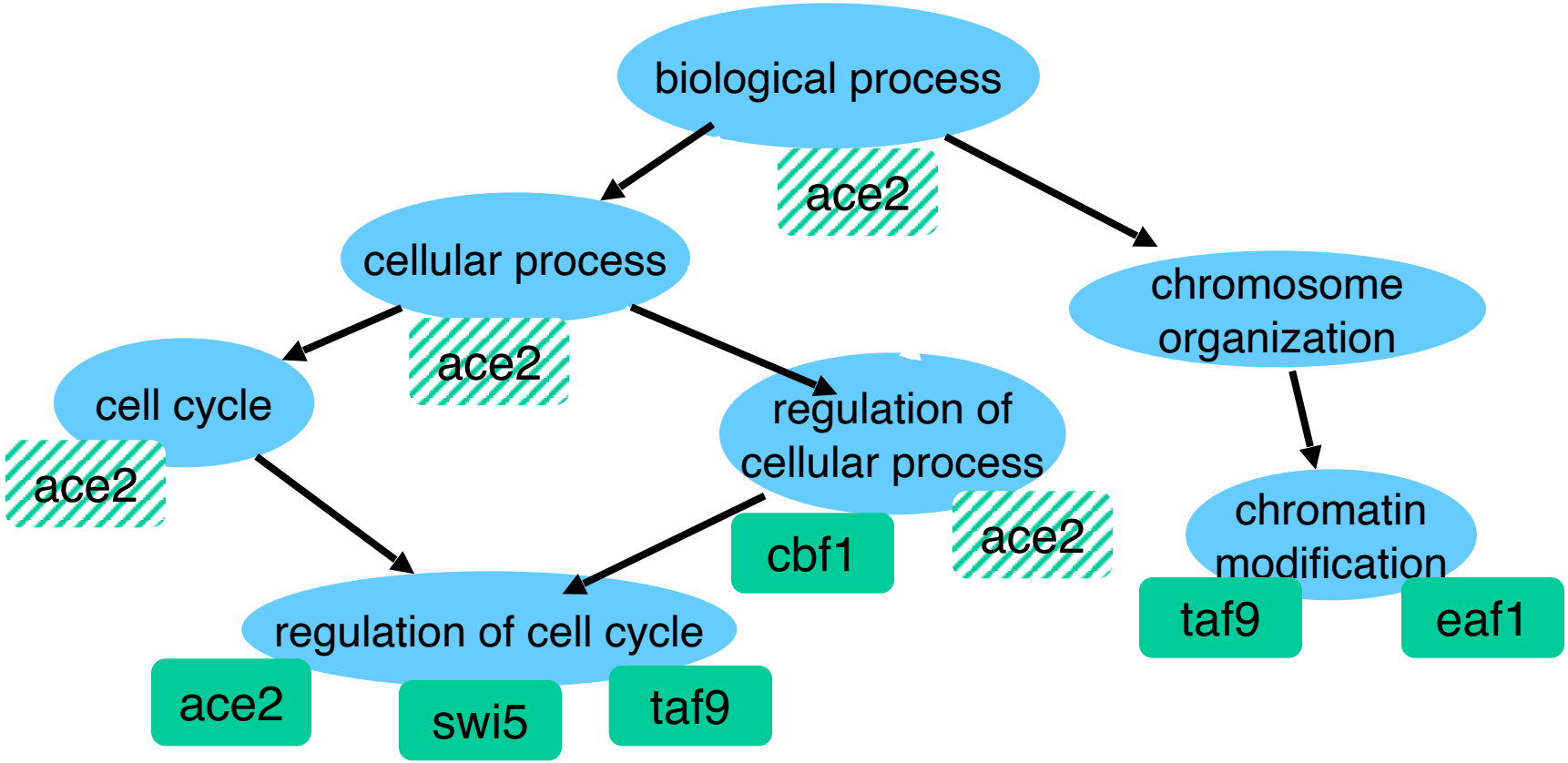
1. Biological Process (BP)
2. Molecular Function (MF)
3. Celular Component (CC)

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

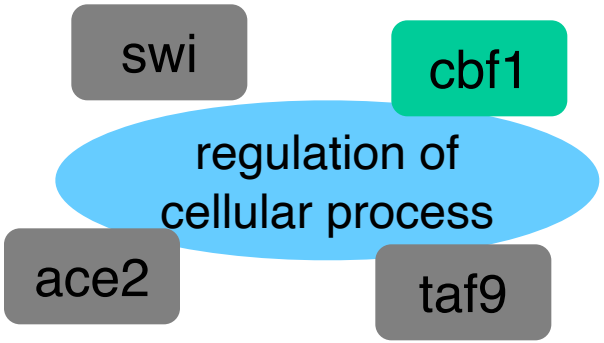
up regulated genes

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

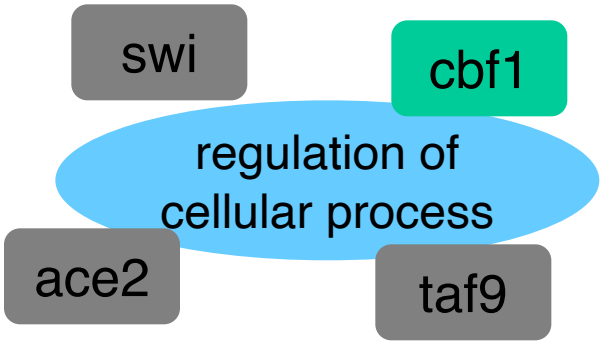
up regulated genes

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 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

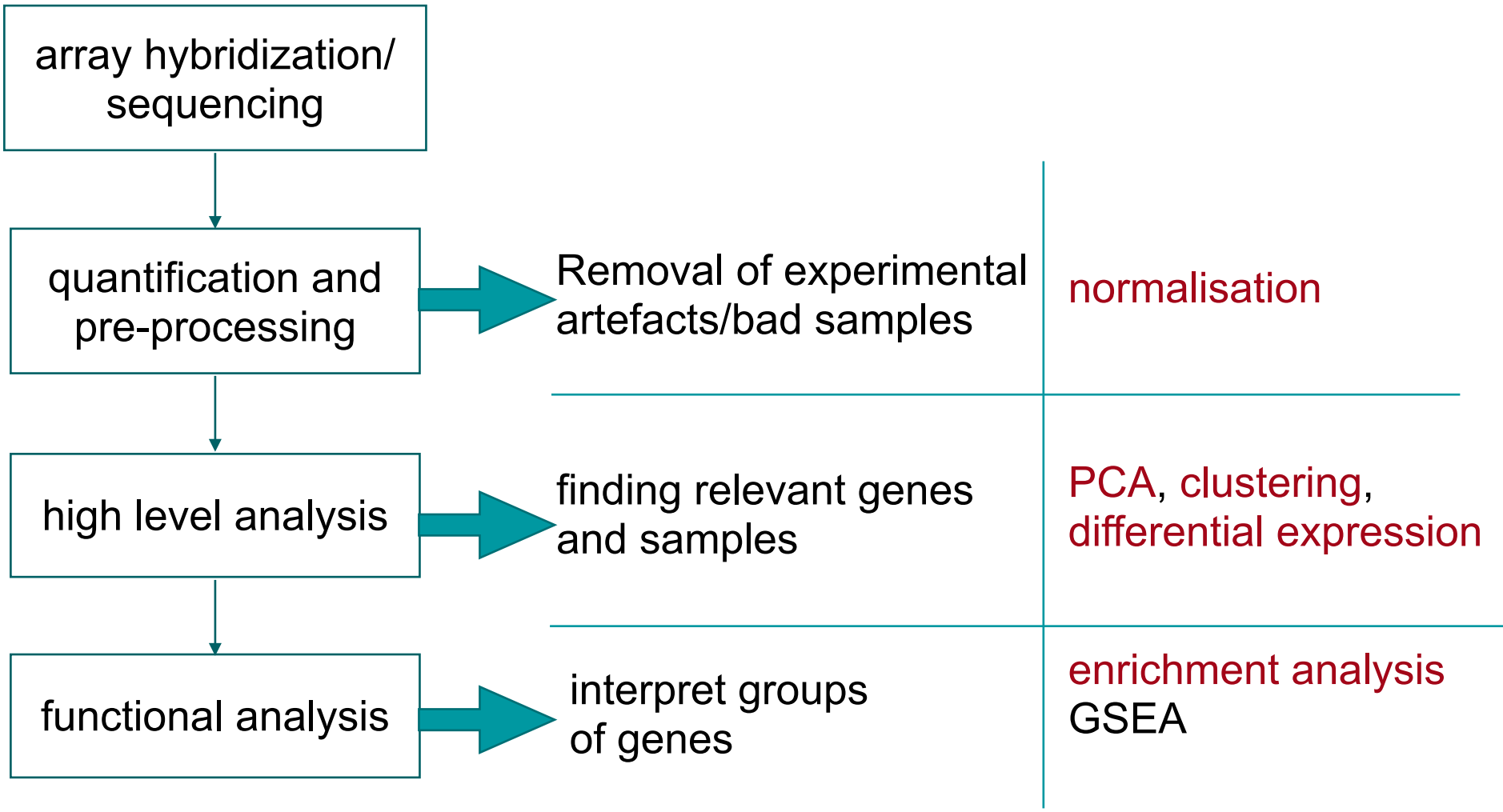
Web interface for enrichment analysis with:
Gene Ontology, KEGG Pathway and TF binding

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

Check the results for my favourite genes:

Irf8 Id2 Spi1 Klf4 Runx2 Egr1

Bioinformatics - Gene Expression Analysis



Hands on!

Handout Step 4,5,6

www.costalab.org

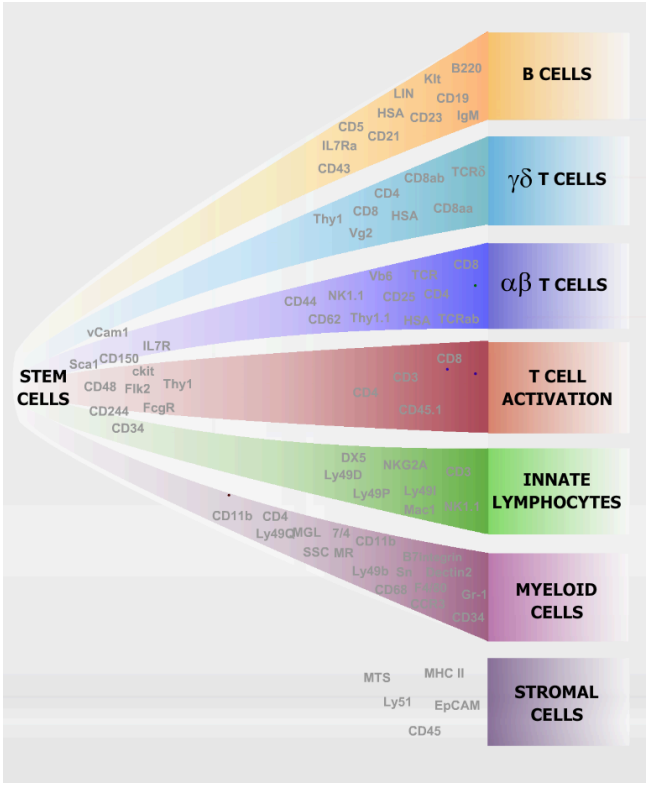
Institute for
Computational Genomics
01011011010
10100100101

RWTH AACHEN
UNIVERSITY

Extra slides!

Integrative Analysis - ImmGen

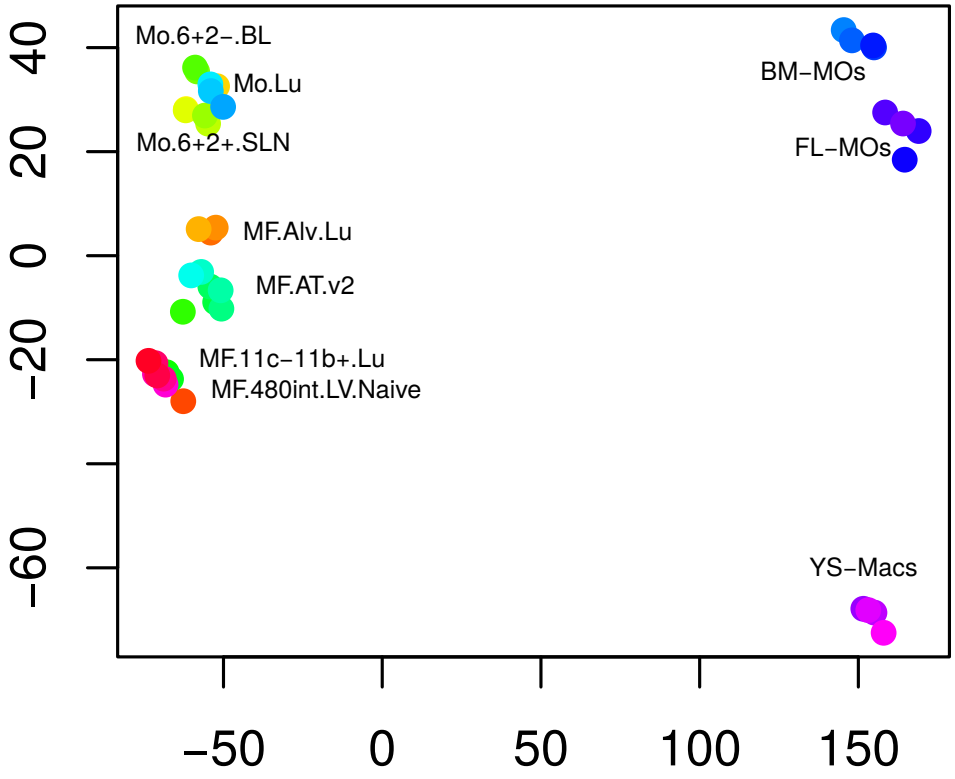
- ImmGen - expression data of immune cells under standardized conditions



- 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 labs tend to cluster together

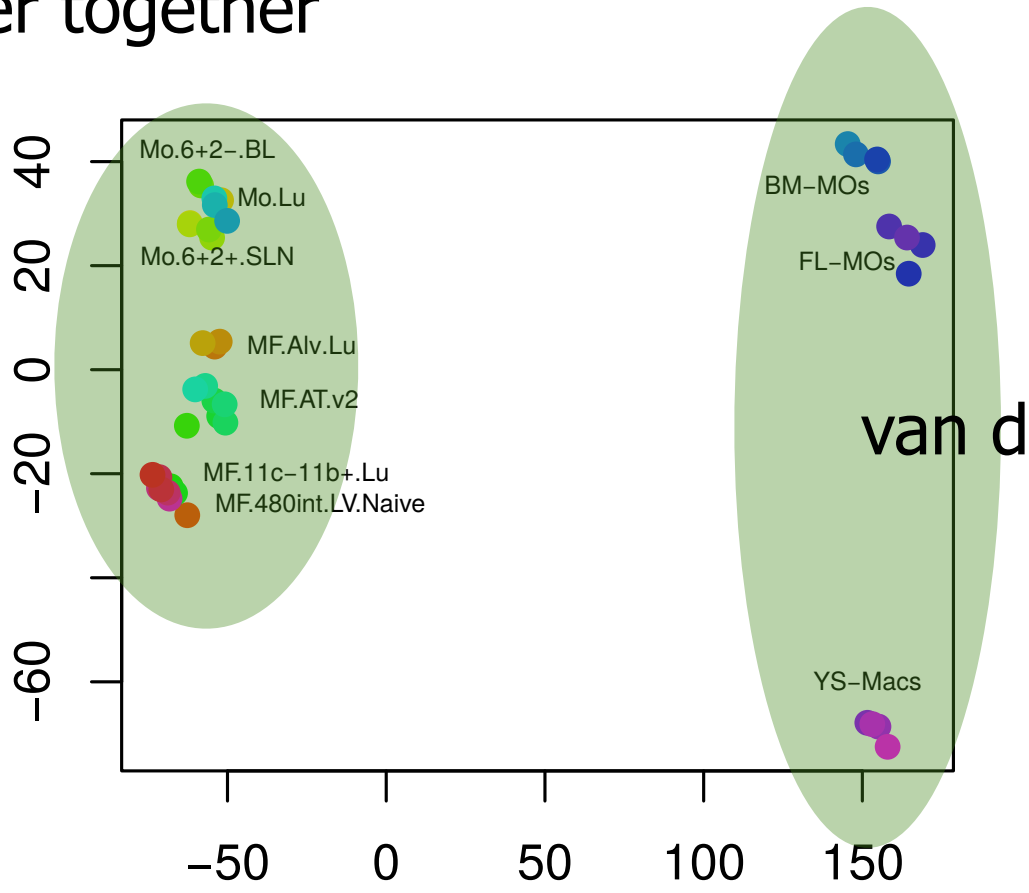


See: Leek JT,…. (2016). sva: Surrogate Variable Analysis. R package version 3.22.0.

Integrative Analysis - Problem

- Batch Effects - Arrays from distinct labs tend to cluster together

ImmGen

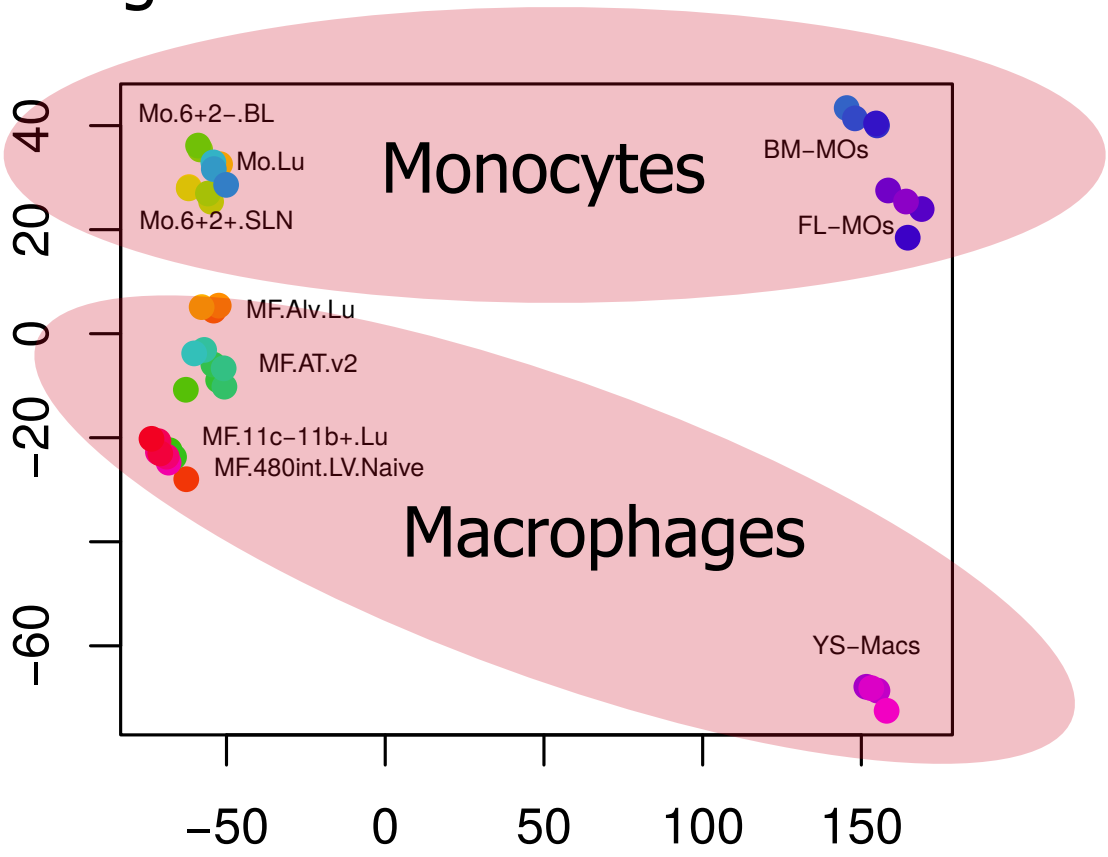


van de Leer, 2016

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.