Toxicogenomics analysis

Giulia Callegaro, Hugo van Kessel, Steven Kunnen & Bob van de Water

EUROTOX 2023

CEC01 - A hands-on introduction to applied artificial intelligence in toxicology (CAAIT)

RISK HUNT3R

ASPIS

Outline

- Network analysis on omics data
- Small introduction on TempOseq and the data pipeline
- Practice!

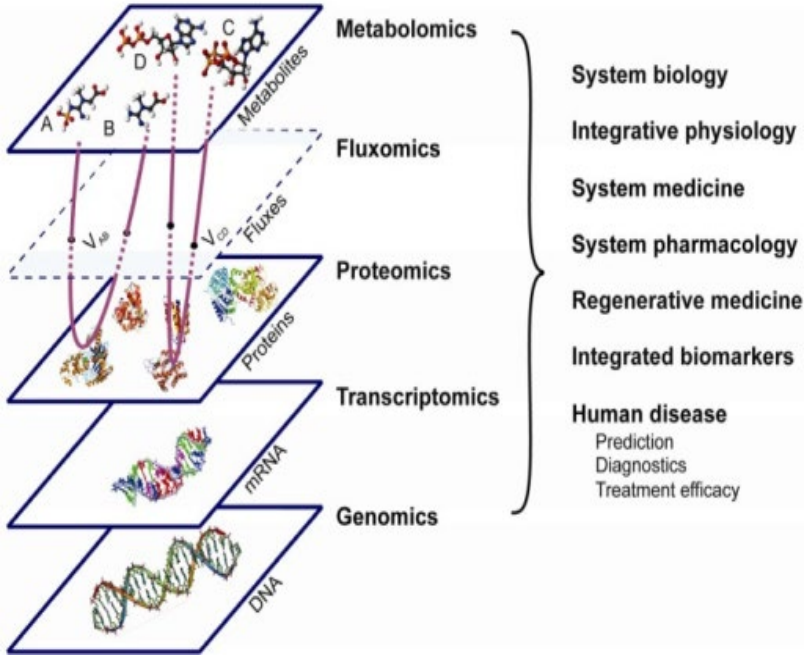
Gene network analysis

The existing model you will be using

Toxicogenomics

Systemic responses to substances

- **Genomics**: structure & function of the genome (e.g. allele frequencies/ types and drug sensitivity/ metabolism)
- **Epigenomics**: Reversible heritable changes in genome functions (often methylome status of genome or histone modifications)
- **Transcriptomics**: mRNA & ncRNA expression (e.g. gene activity/ involved pathways)
- **Proteomics**: Protein expression & activity (e.g. phosphorylation states, specific pathways)
- **Metabolomics**: Metabolic profiling (e.g. formed drug metabolites, or changes in glytolitic/ oxidative phosphorylation metabolites after drug exposure)

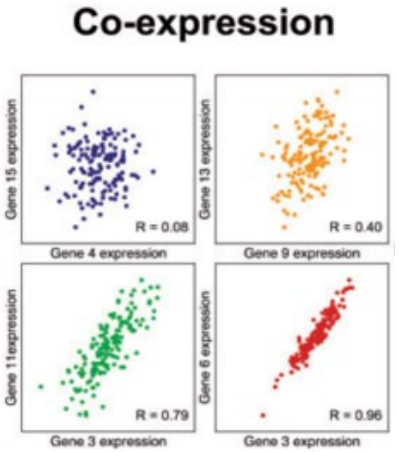


"Toxicogenomics was identified in TT21C as a transformative approach that was expected to play a pivotal role in **identifying** the toxicity pathways and cellular responses associated with exposure to environmental agents."

Krewski, D., et al. (2019). Archives of Toxicology. https://doi.org/10.1007/s00204-019-02613-4

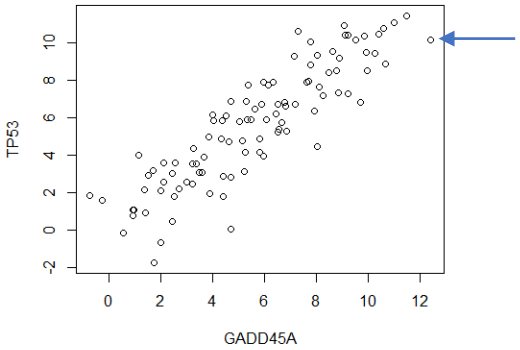
Gene co-expression networks

Group genes with a tendency to co-activate across a group of samples



Co-expression == Pairwise gene correlation across all the conditions

Example



DNA damaging compound

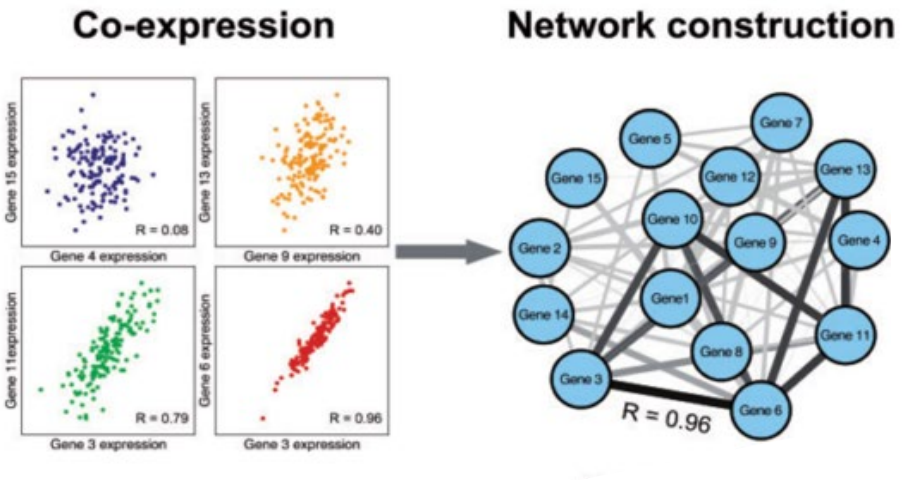
When TP53 is highly expresses, so is GADD45.

When TP53 is lowly expressed, so is GADD45

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017)

Gene co-expression networks

Group genes with a tendency to co-activate across a group of samples



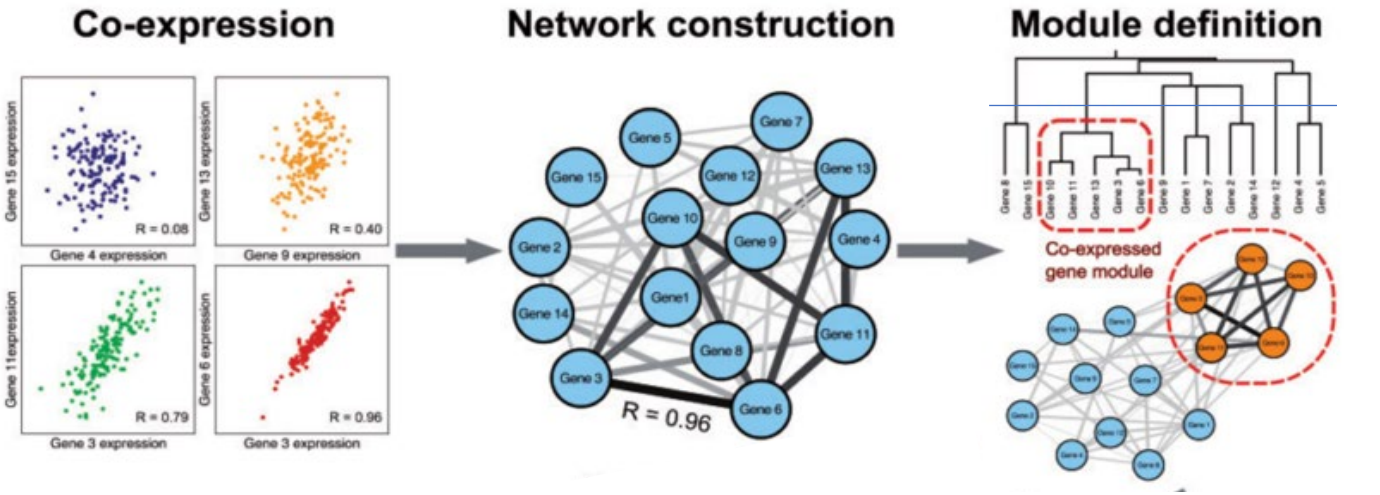
Co-expression == Pairwise gene correlation across all the conditions

Co-expression as distance between genes. Highly correlated genes are closer

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017)

Gene co-expression networks

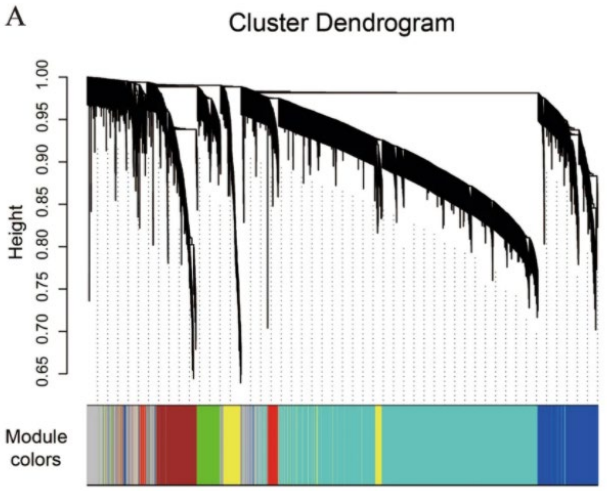
Group genes with a tendency to co-activate across a group of samples



Co-expression == Pairwise gene correlation across all the conditions

Co-expression as distance between genes. Highly correlated genes are closer

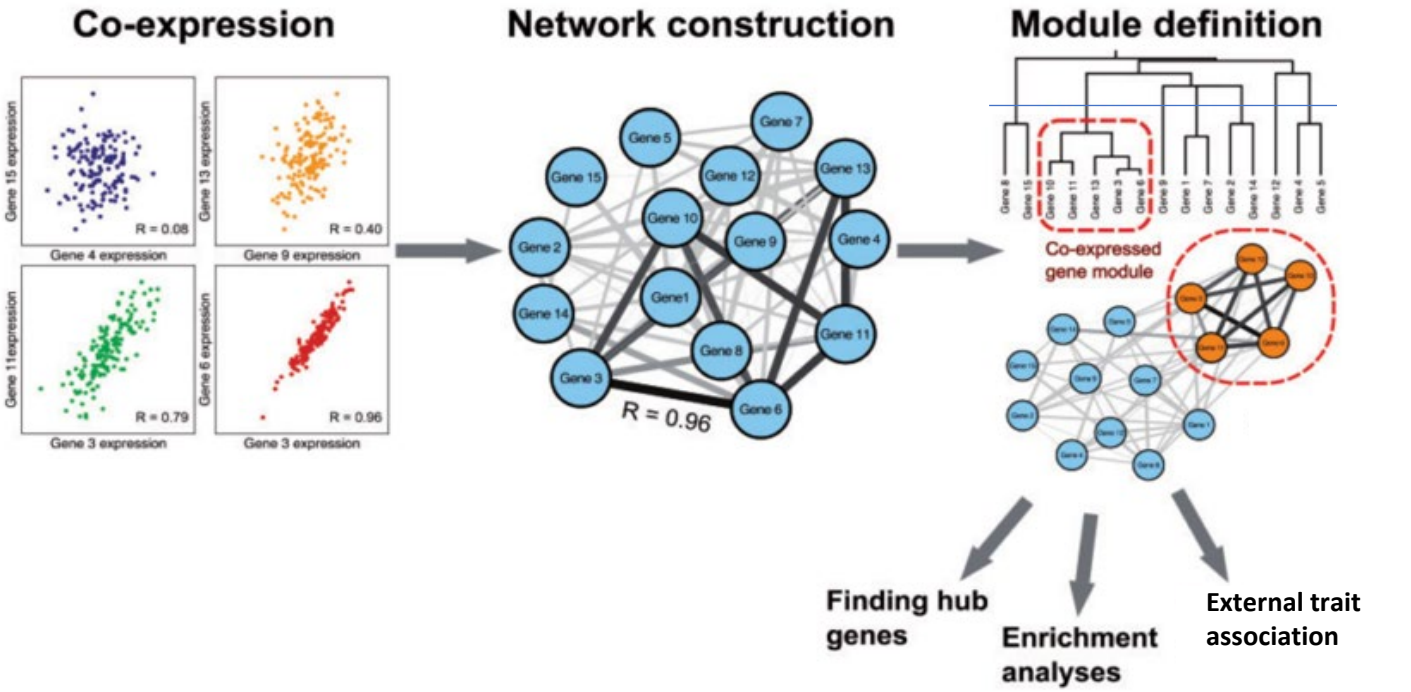
Isolate highly connected genes: the modules

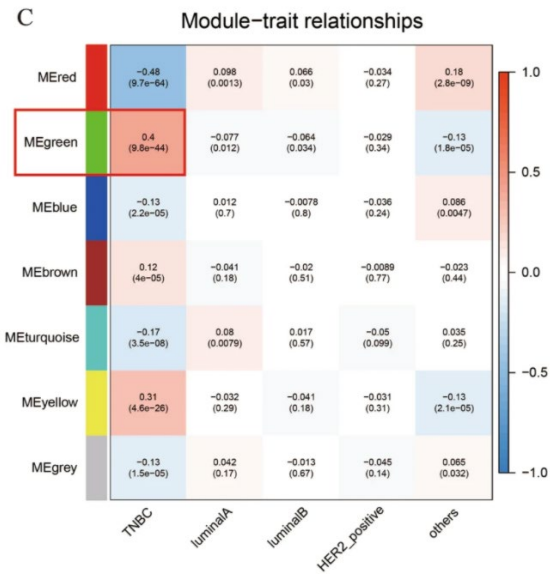


van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017)

Gene co-expression networks

Group genes with a tendency to co-activate across a group of samples





Annotate modules for their function (using known information)

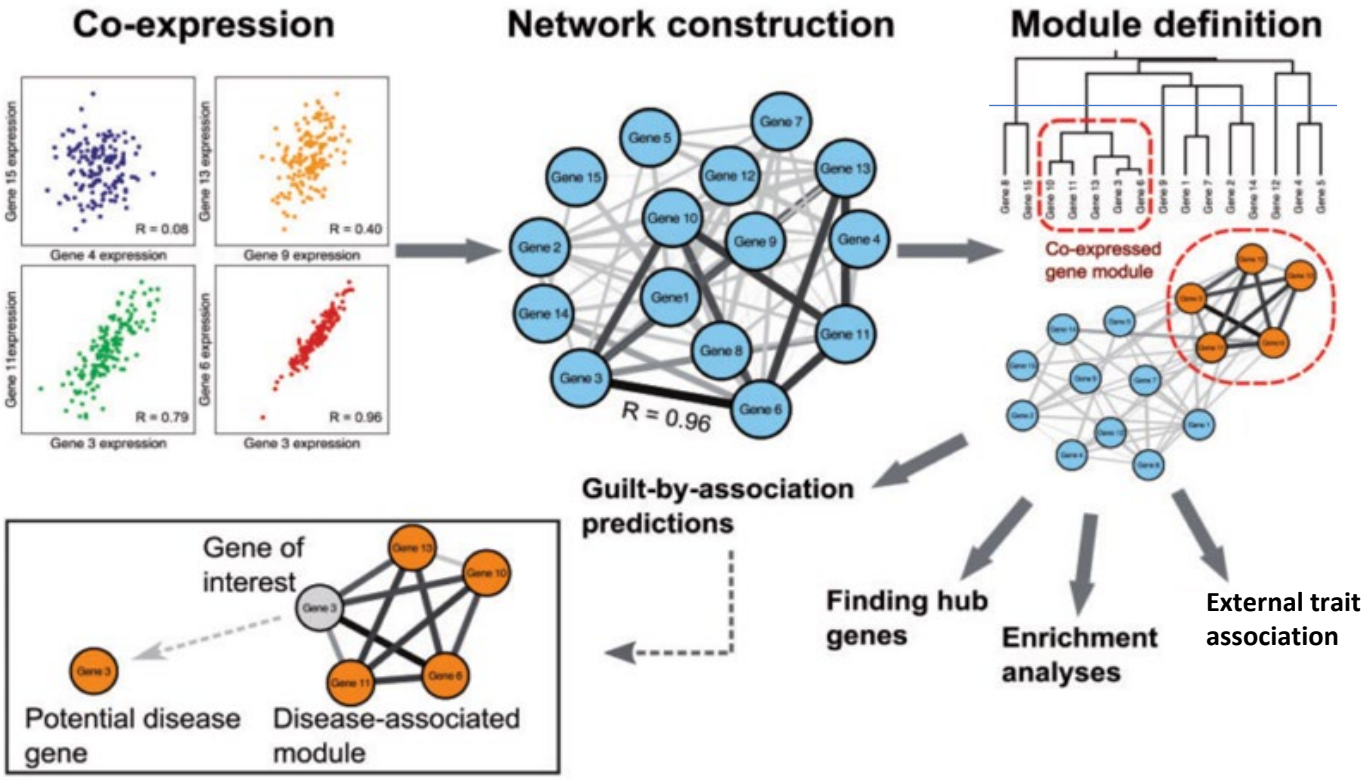
but not all genes are annotated..

Explanation: module green is activated always by TNBC, and not much by the other subtypes

van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017)

Gene co-expression networks

Group genes with a tendency to co-activate across a group of samples



Unknown genes are associated to existing information!

`library(WGCNA)`

The TXG-MAPr tool
https://txg-mapr.eu/login/

Credentials:

- username: eurotox2023
- password: txg_mapr01

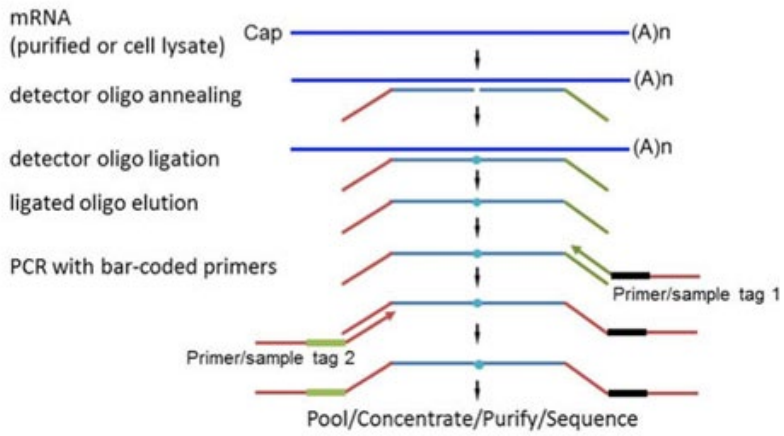
van Dam, S., Võsa, U., van der Graaf, A., Franke, L., & de Magalhães, J. P. (2017)

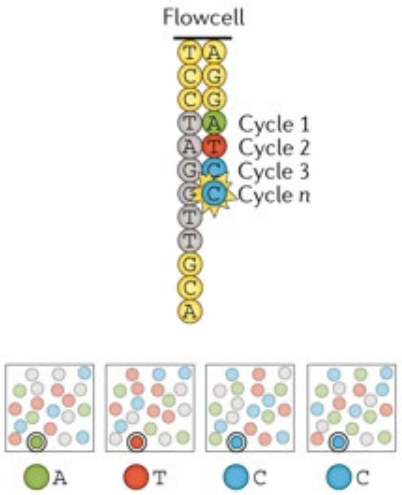
Practice: generate your data to load into the model

Transcriptomic data processing

Transcriptomics – TempOseq RNA sequencing

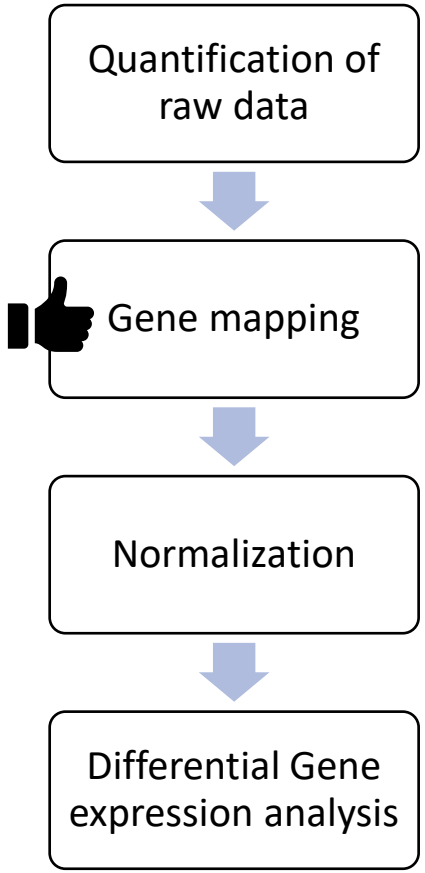
Illumina





Principle: hybridization with complementary probes AND Sequencing by synthesis

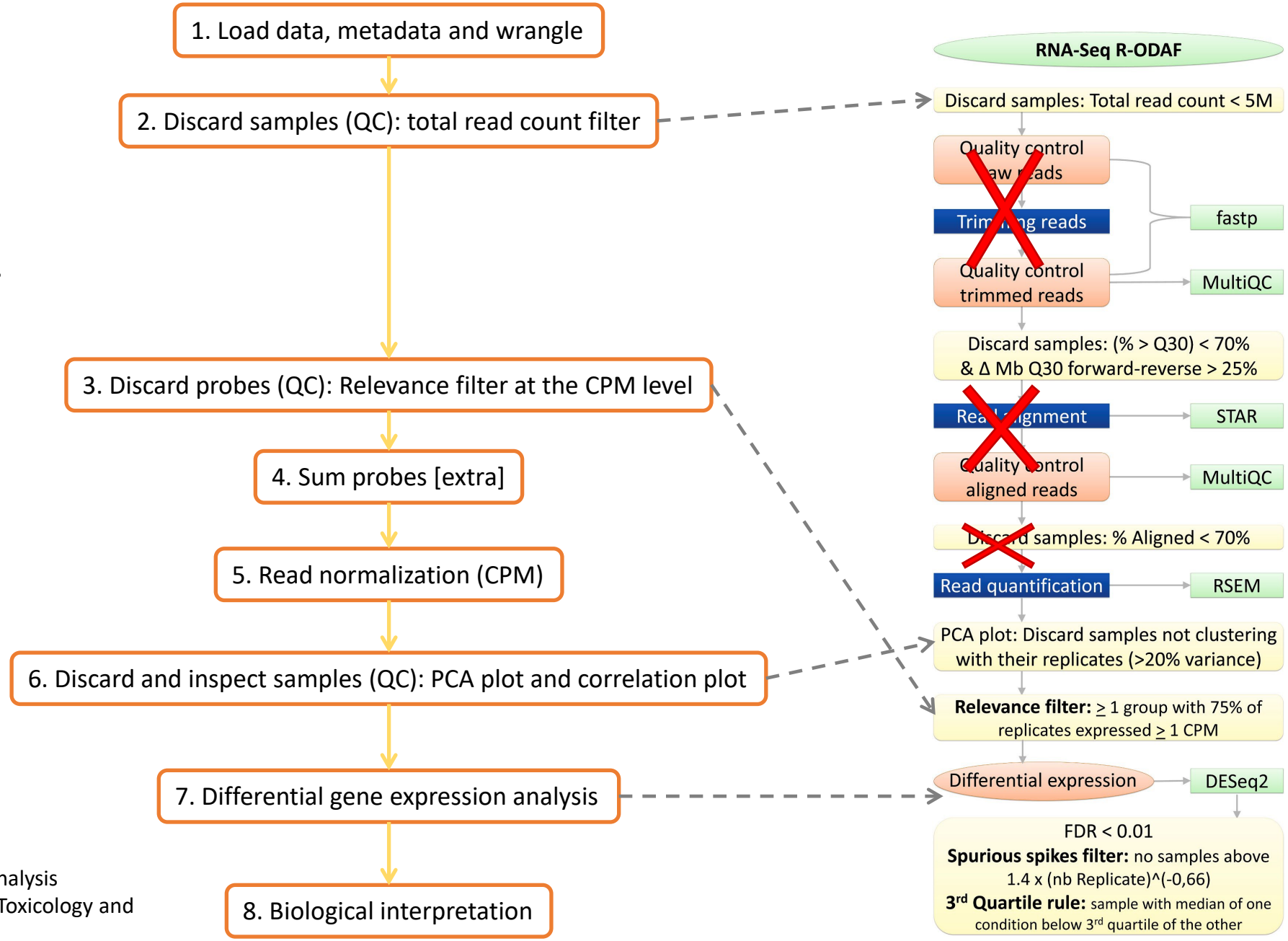
- Prebuilt **set** of oligos:
  - Complementary to mRNA of interest
  - Including a primer
  - Including a sample tag
- Only oligos annealing with mRNA in the samples are retained
- Amplification
- Sequencing of amplified oligos
  - “we know what we sequence” concept
- Easy to define target space: suitable for high- throughput



Definitions

- Read depths: (aimed) number of reads per sample
- Total read count: the actual number of reads per sample
- Count per million (CMP): one way to correct for different sequencing depths → (read/total_reads)*10^6

Operational pipeline and comparison with R-ODAF



Marcha CT. Verheijenet al., R-ODAF: Omics data analysis framework for regulatory application, Regulatory Toxicology and Pharmacology, Volume 131, 2022, 105143, https://doi.org/10.1016/j.yrtph.2022.105143.

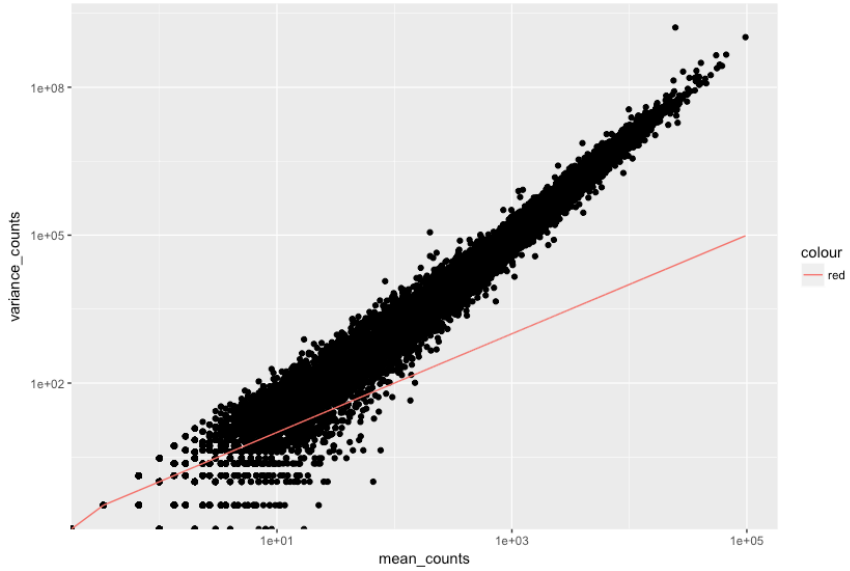
Differential gene expression analysis

Reads counts cannot be approximated with a normal distribution:

- They are integer counts
- low number of counts associated with a large proportion of genes
- long right tail due to the lack of any upper limit for expression

Count data is modeled using the negative binomial distribution (why not Poisson? Because mean < variance)

- For each gene, a generalized linear model (GLM) is fitted, modelling the read counts K ij as following a negative binomial distribution



raw count for gene i, sample j

The mean is taken as "normalized counts" scaled by a normalization factor

one dispersion per gene

$$K_{ij} \sim \mathrm{NB}(s_{ij} q_{ij}, \alpha_i)$$

Love, M. I., Huber, W., & Anders, S. (2014)., https://hbctraining.github.io/DGE_workshop_salmon_online/schedule/links-to lessons.html, https://bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.html#the-deseq2-model adapted slide from courtesy of Dr. Pellegrino Coppola

Deseq2 in practice

- Create DESeqDataSet object

```
deseq_object = DESeqDataSetFromMatrix(countData =  as.data.frame(countdata_raw_fsample_fprobe_sumprobe),
                                      colData = metadata %>% mutate(mean_id = as.factor(mean_id)),
                                      design = ~ mean_id,
                                      tidy = TRUE)
```

- Define size factors (CMP)

```
sizeFactors(deseq_object) = colSums(column_to_rownames(countdata_raw_fsample_fprobe_sumprobe, var = "gene_symbol"))/1E6
```

- Run Deseq model

```
deseq_object = DESeq(deseq_object)
```

- (for each contrast) extract the results

```
results(loop_deseq_object, contrast = c("mean_id", contrast$mean_id_treatment[i], contrast$mean_id_control[i])) %>%
```

We are hiring!

Universiteit Leiden The Netherlands

LACDR

Currently online:

Postdoctoral researcher in Toxicogenomics 2.0: bringing interpretation and FAIRification to the next level

SCAN ME

Soon online:

PhD researcher in Systems Toxicology: connecting gene co-expression networks to cellular injury and clinical pathology

Bioinformatician – (Postdoc or PhD researcher) - in Translational Quantitative Gene Network Analysis for Systems Toxicology-based Human Chemical Safety Assessment

PARC

RISK HUNT3R

ASPIS

efsa EUROPEAN FOOD SAFETY AUTHORITY