Differential gene expression analysis using RNA-seq

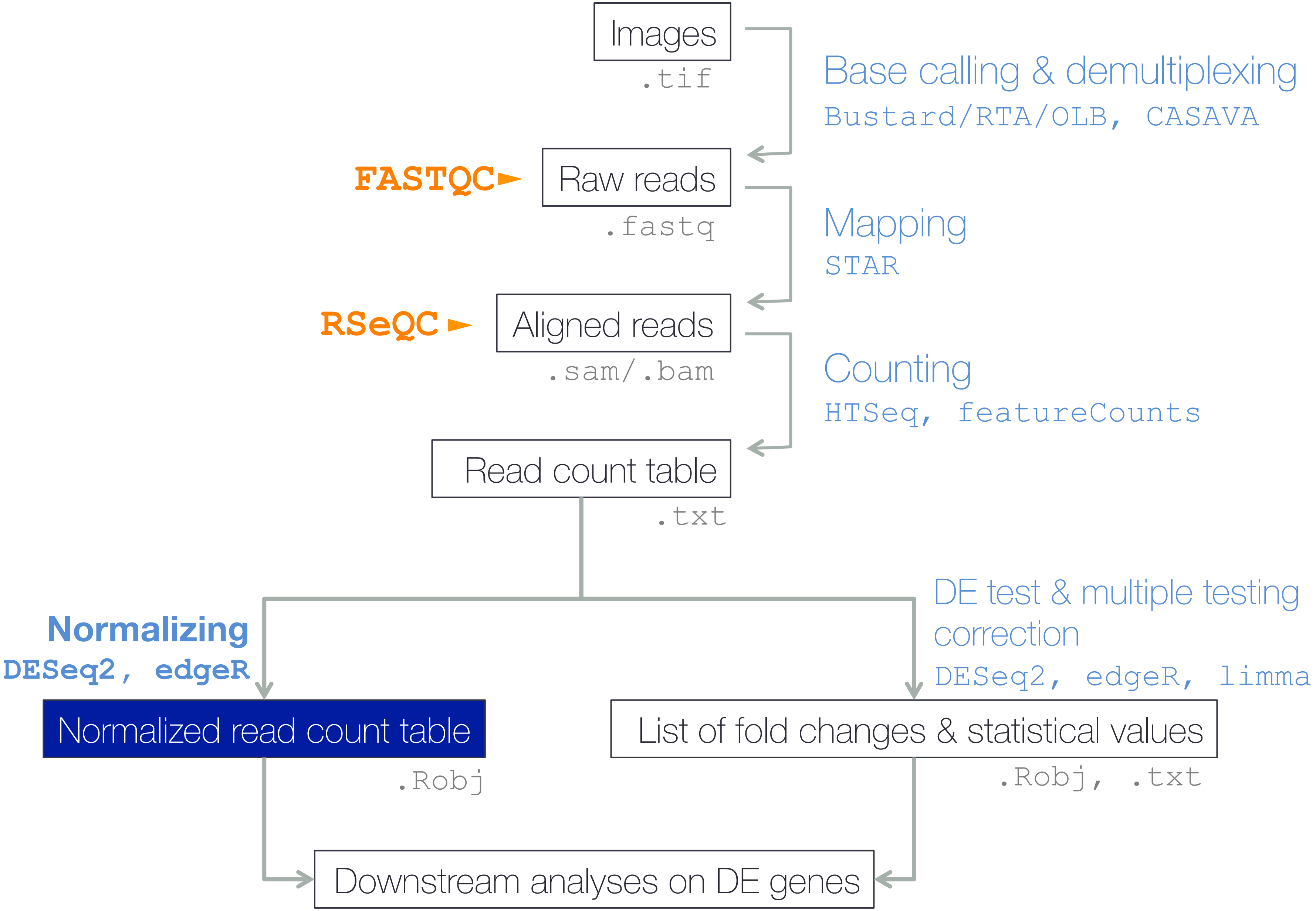
Applied Bioinformatics Core, November 2019

Friederike Dündar with Luce Skrabanek & Paul Zumbo

Day 4 overview

- exploring read counts
  - rlog transformation
  - hierarchical clustering
  - PCA
- (brief) theoretical background for DE analysis
- DE analysis using DESeq2
- exploring the results

Bioinformatics workflow of RNA-seq analysis



Images

.tif

Base calling & demultiplexing

Bustard/RTA/OLB, CASAVA

FASTQC

Raw reads

.fastq

Mapping

STAR

RSeQC

Aligned reads

.sam/.bam

Counting

HTSeq, featureCounts

Read count table

.txt

Normalizing

DESeq2, edgeR

Normalized read count table

.Robj

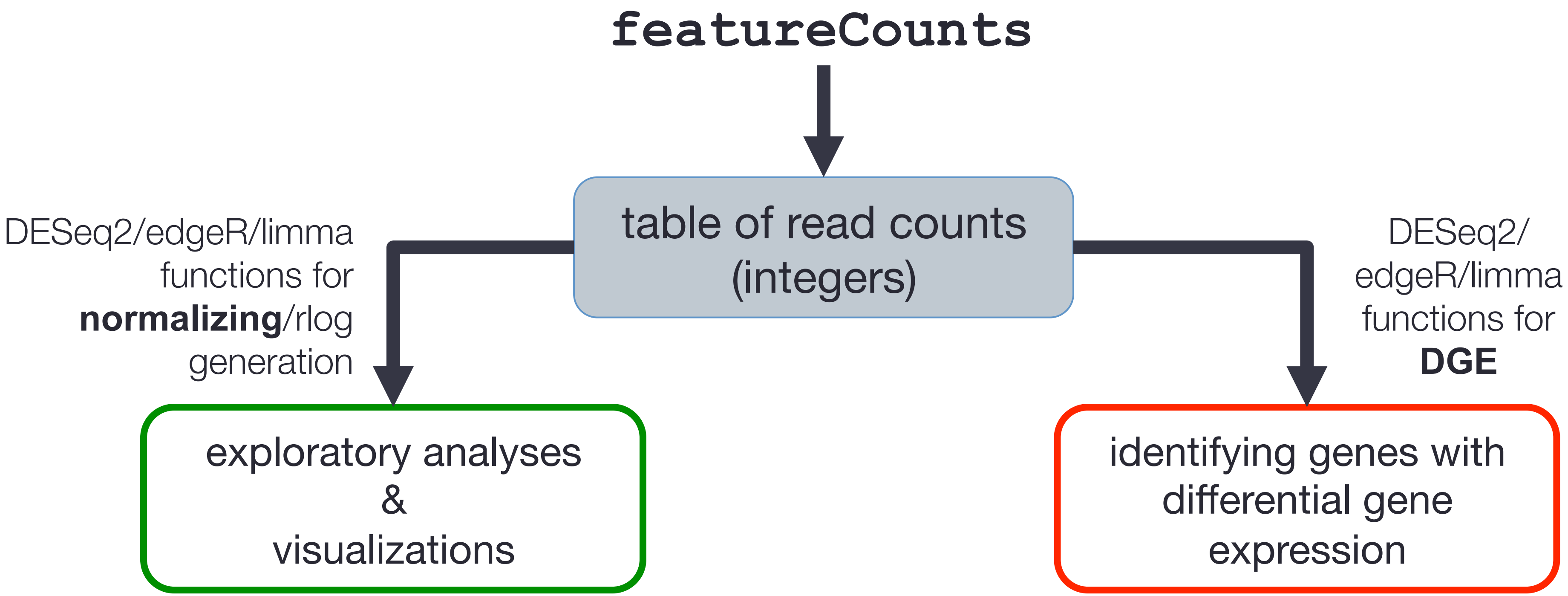
DE test & multiple testing correction

DESeq2, edgeR, limma

List of fold changes & statistical values

.Robj, .txt

Downstream analyses on DE genes



featureCounts

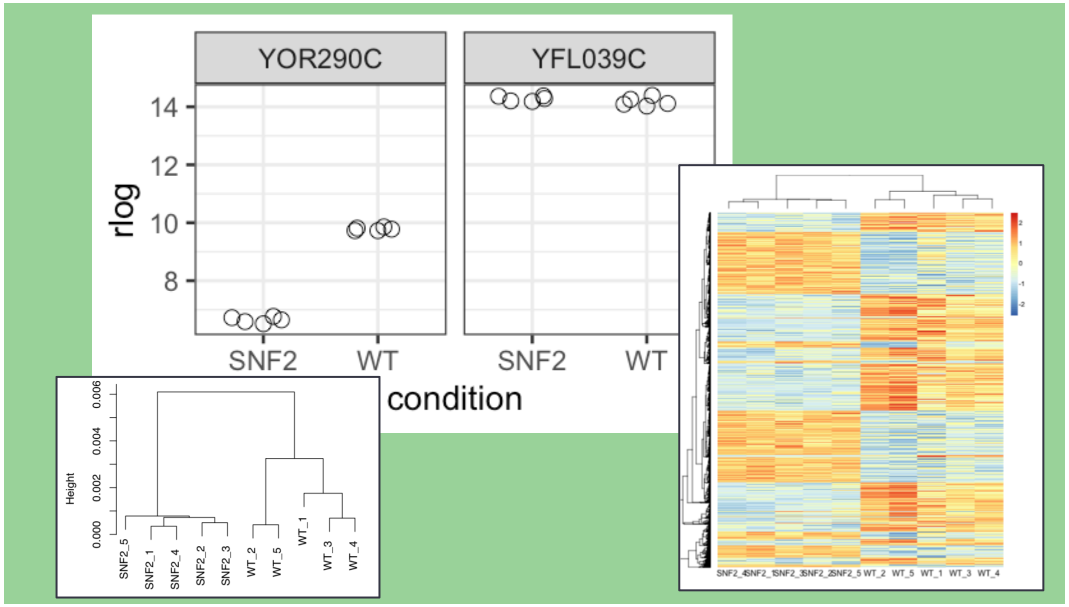
table of read counts (integers)

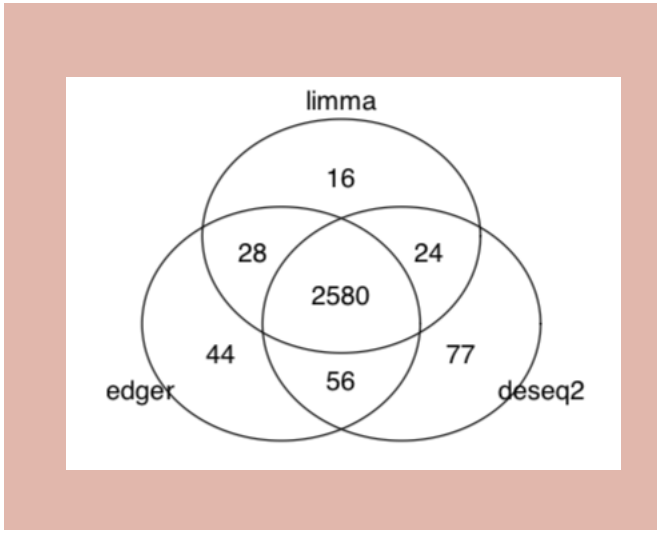
DESeq2/edgeR/limma functions for **normalizing**/rlog generation

exploratory analyses & visualizations

DESeq2/ edgeR/limma functions for **DGE**

identifying genes with differential gene expression





Expression units

- strongly influenced by
  - gene length
  - sequencing depth
  - expression of all other genes in the same sample

DESeq’s size factor normalization

- annoying mathematical properties of read counts
  - large dynamic range
  - discrete values

hetero-skedasticity

log transformation and variance stabilization (DESeq’s rlog())

Use normalized and transformed expression units for exploratory analyses!

EXPLORATORY ANALYSES

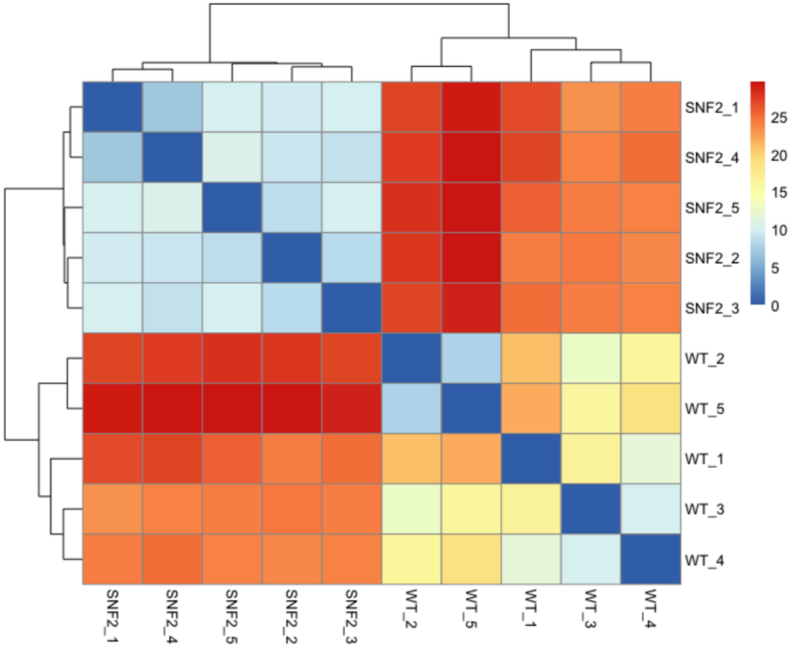
assessing sample similarities & sources of variation

Exploratory analyses

- **do not test a null hypothesis**!
- meant to familiarize yourself with the data at hand and to discover biases and unexpected variability

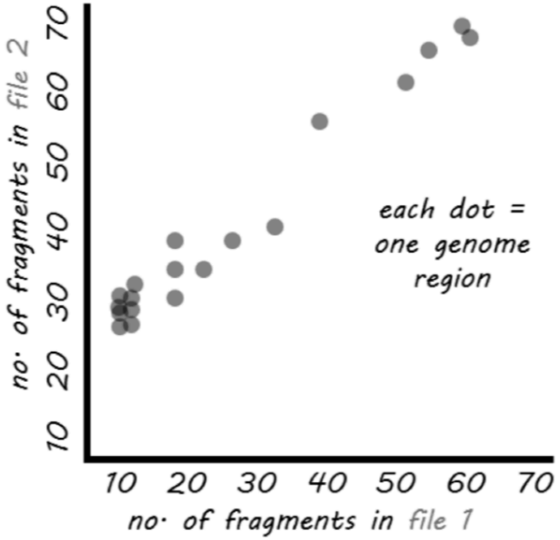
Typical exploratory analyses:

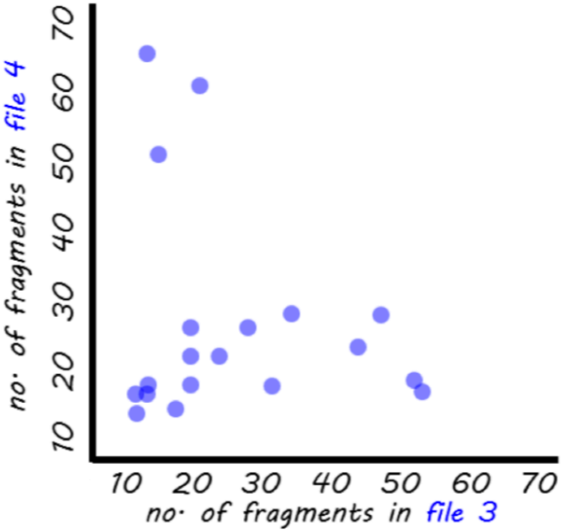
- **correlation** of gene expression between different samples
- (hierarchical) **clustering**
- **dimensionality reduction** (e.g. PCA)
- dot plots/box plots/violin plots of **individual genes**



Pairwise correlation of gene expression values

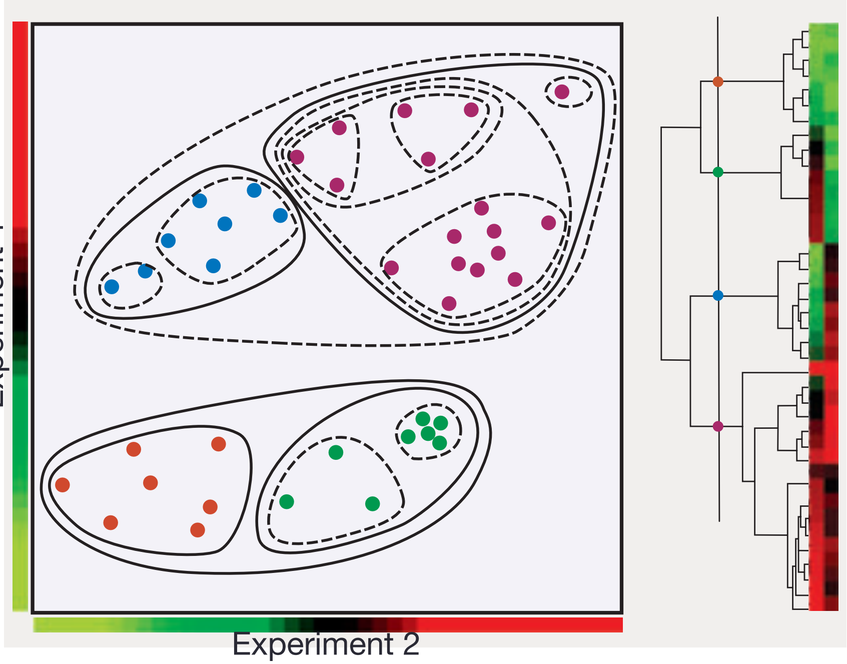
- replicates of the same condition should show high correlations (> 0.9)
- **Pearson** method: *metric* differences between samples
  - influenced by outliers
- **Spearman** method: based on *rankings*
  - less sensitive
  - less driven by outliers
- R function: `cor()`





Clustering gene expression values

Goal: partition the samples into homogeneous groups such that the within-group similarities are large.



single-sample (or single-gene) clusters are successively joined

+ "unbiased"
- not very robust

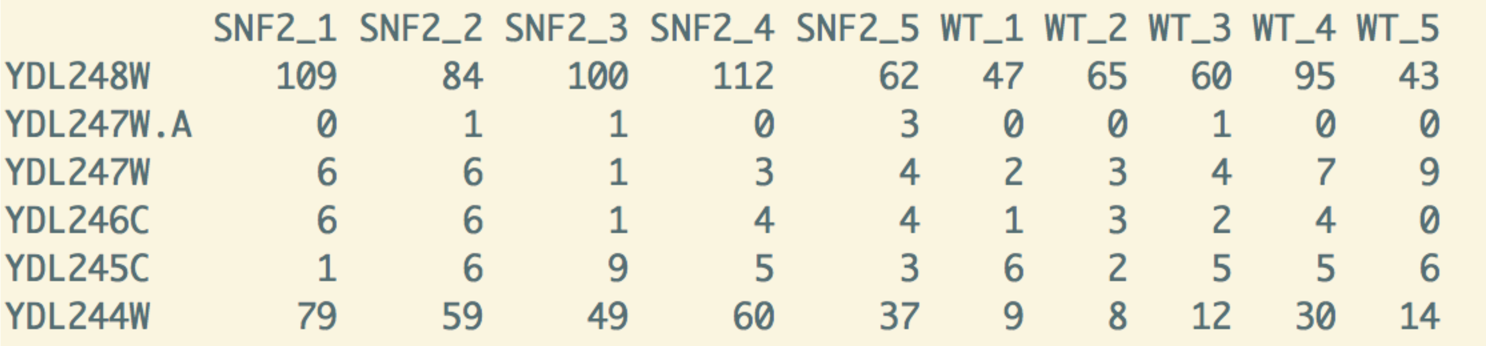
- Result: dendrogram
  - clustering obtained by cutting the dendrogram at the desired level
- Similarity measures
  - Euclidean
  - Pearson correlation
- Distance measures
  - Complete: largest distance
  - Average: average distance

R function: `hclust()`

Figure from D'haeseleer P. Nat Biotechnol. 2005 Dec;23(12):1499-501. 10.1038/nbt1205-1499

PCA

starting point: matrix with expression values per gene and sample,
e.g. 7,100 genes x 10 samples

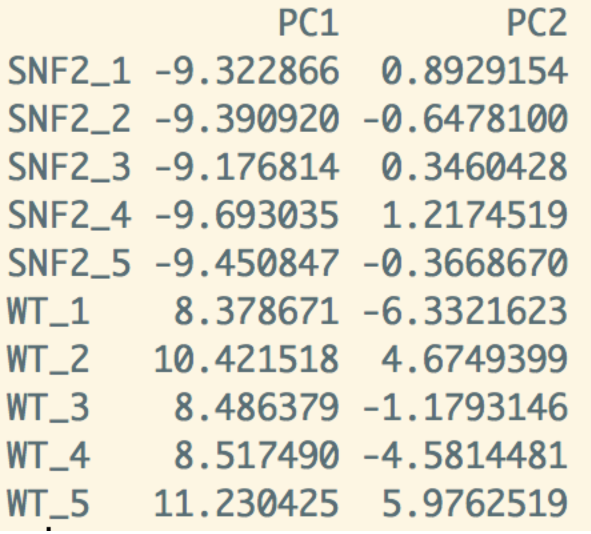


```
          SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5 WT_1 WT_2 WT_3 WT_4 WT_5
YDL248W      109     84    100    112     62   47   65   60   95   43
YDL247W.A      0      1      1      0      3    0    0    1    0    0
YDL247W        6      6      1      3      4    2    3    4    7    9
YDL246C        6      6      1      4      4    1    3    2    4    0
YDL245C        1      6      9      5      3    6    2    5    5    6
YDL244W       79     59     49     60     37    9    8   12   30   14
```

If we want to understand the main differences between SNF2 and WT samples, the most detailed view (with the most "dimensions") would entail all 7,100 genes.

However, it is probably enough to focus on the genes that are actually different.
In fact, it'll be even better if we could somehow identify entire groups of genes that capture the majority of the differences.

PCA does exactly that ("grouping genes") using the correlation amongst each other.



```
             PC1        PC2
SNF2_1 -9.322866  0.8929154
SNF2_2 -9.390920 -0.6478100
SNF2_3 -9.176814  0.3460428
SNF2_4 -9.693035  1.2174519
SNF2_5 -9.450847 -0.3668670
WT_1    8.378671 -6.3321623
WT_2   10.421518  4.6749399
WT_3    8.486379 -1.1793146
WT_4    8.517490 -4.5814481
WT_5   11.230425  5.9762519
```

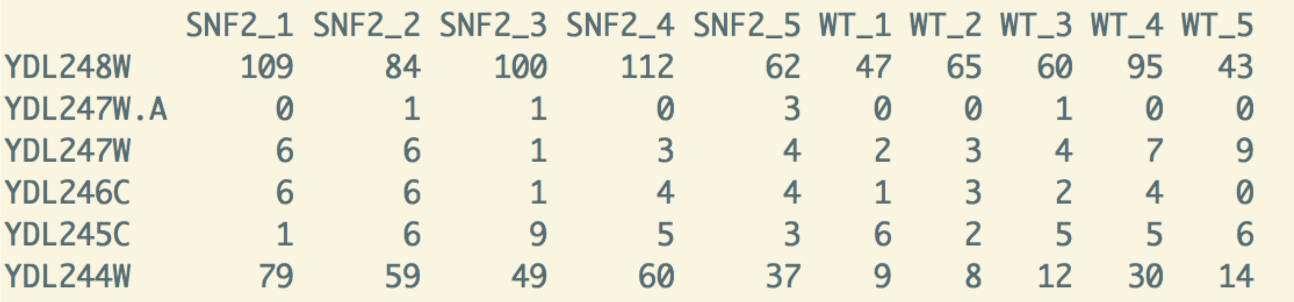
2 PCs (or more) x 10 samples

http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Principal component analysis

Goal: Reduce the dataset to fewer dimensions yet approx. preserve the distance between the individual samples

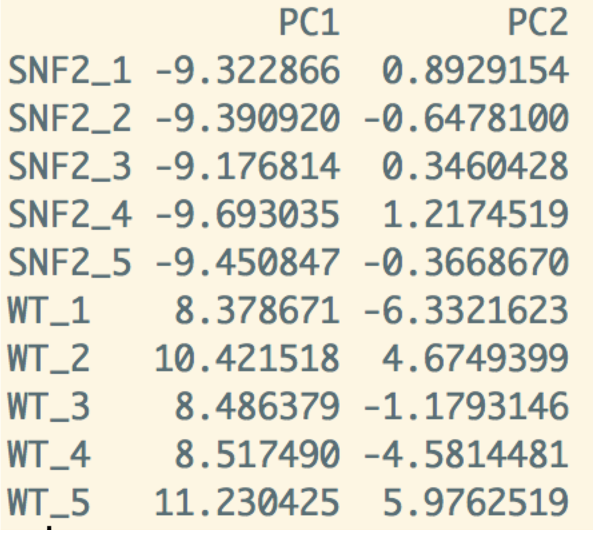
starting point: matrix with expression values per gene and sample, e.g. 7,100 genes x 10 samples



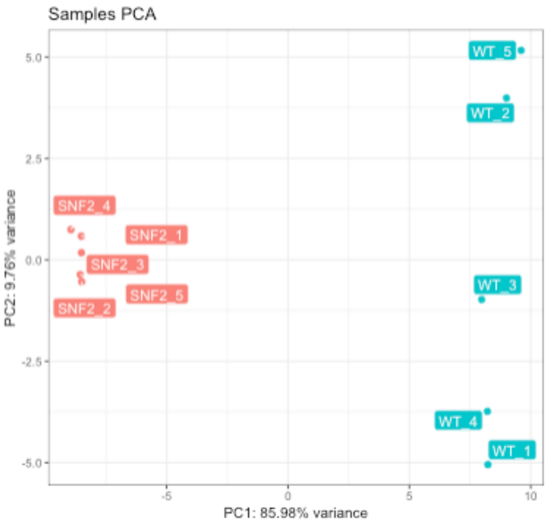
```
          SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5 WT_1 WT_2 WT_3 WT_4 WT_5
YDL248W      109     84    100    112     62   47   65   60   95   43
YDL247W.A      0      1      1      0      3    0    0    1    0    0
YDL247W        6      6      1      3      4    2    3    4    7    9
YDL246C        6      6      1      4      4    1    3    2    4    0
YDL245C        1      6      9      5      3    6    2    5    5    6
YDL244W       79     59     49     60     37    9    8   12   30   14
```

7,100 principal components x 10 samples

- vectors along which the variation between samples is maximal
- PC1-3 usually sufficient to capture the major trends!



```
             PC1        PC2
SNF2_1 -9.322866  0.8929154
SNF2_2 -9.390920 -0.6478100
SNF2_3 -9.176814  0.3460428
SNF2_4 -9.693035  1.2174519
SNF2_5 -9.450847 -0.3668670
WT_1    8.378671 -6.3321623
WT_2   10.421518  4.6749399
WT_3    8.486379 -1.1793146
WT_4    8.517490 -4.5814481
WT_5   11.230425  5.9762519
```

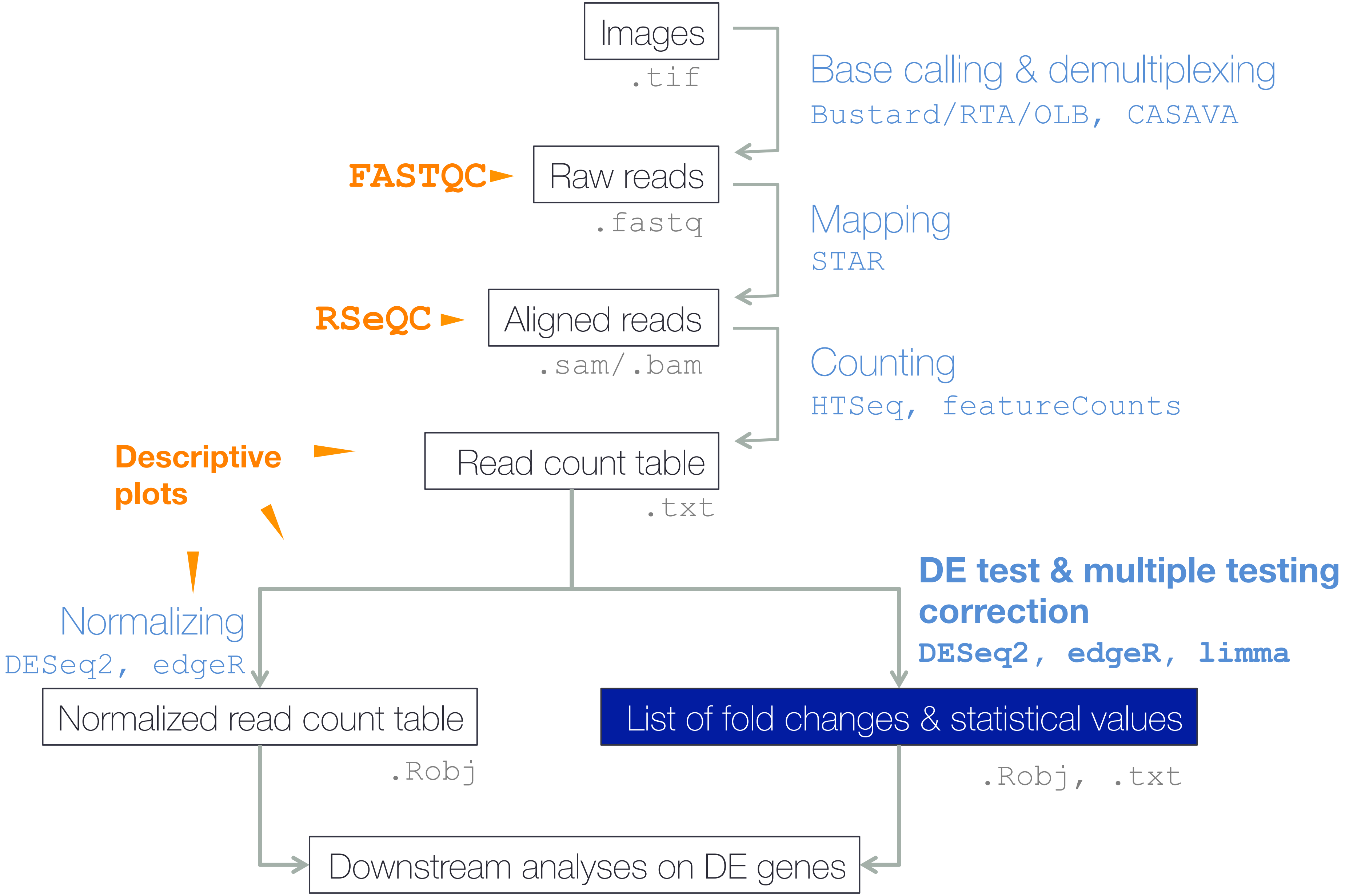


http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

DIFFERENTIAL GENE EXPRESSION

Identifying genes with statistically significant expression differences between samples of different conditions

Bioinformatics workflow of RNA-seq analysis



Images

.tif

Base calling & demultiplexing

Bustard/RTA/OLB, CASAVA

FASTQC

Raw reads

.fastq

Mapping

STAR

RSeQC

Aligned reads

.sam/.bam

Counting

HTSeq, featureCounts

Descriptive plots

Read count table

.txt

Normalizing

DESeq2, edgeR

DE test & multiple testing correction

DESeq2, edgeR, limma

Normalized read count table

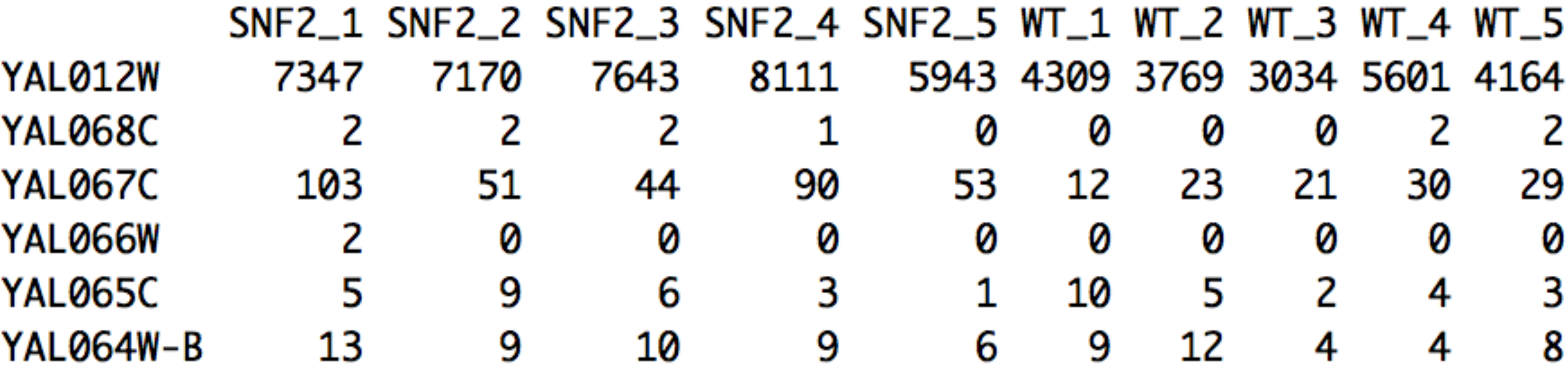
.Robj

List of fold changes & statistical values

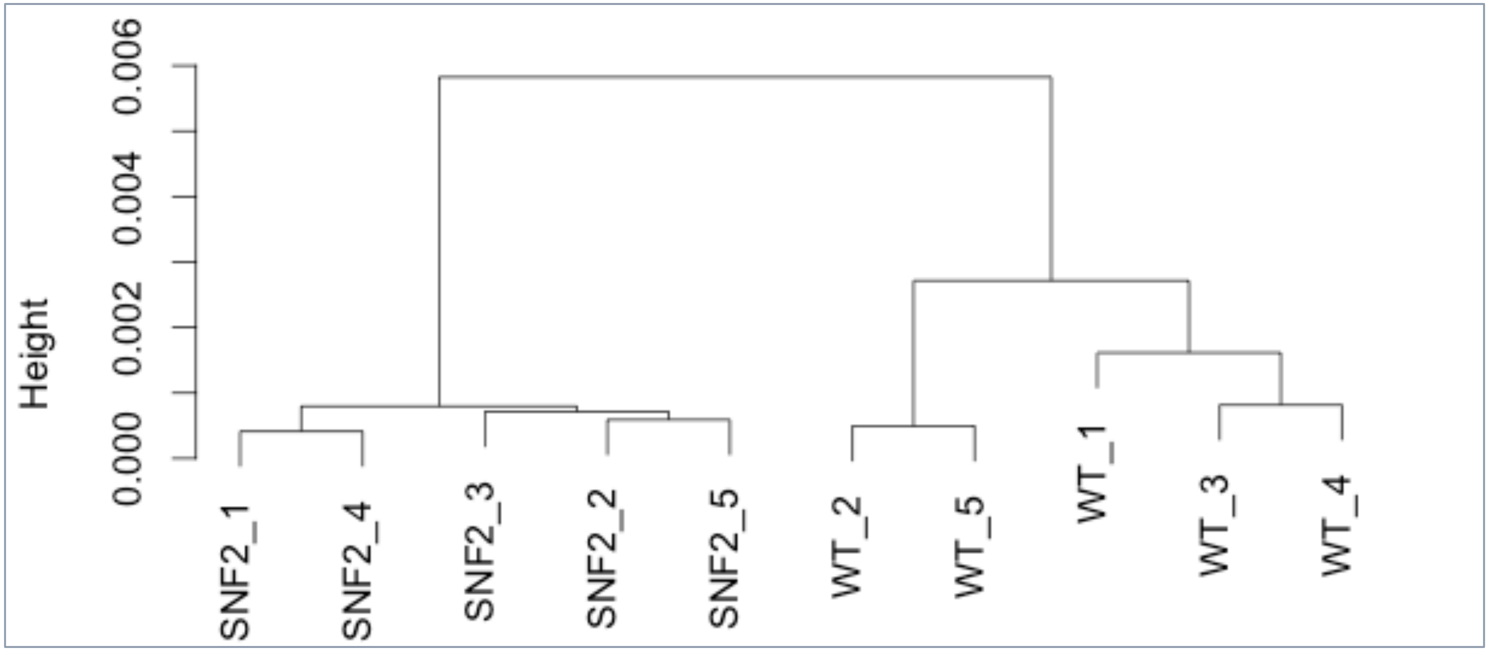
.Robj, .txt

Downstream analyses on DE genes

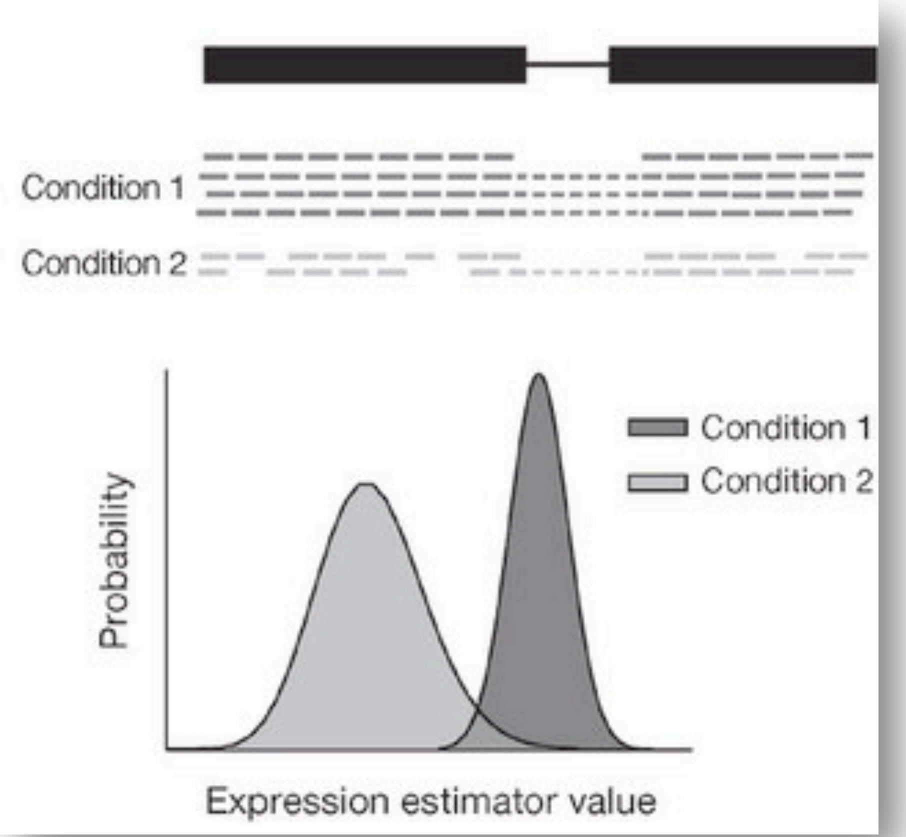
Read count table



```
           SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5 WT_1 WT_2 WT_3 WT_4 WT_5
YAL012W      7347   7170   7643   8111   5943 4309 3769 3034 5601 4164
YAL068C         2      2      2      1      0    0    0    0    2    2
YAL067C       103     51     44     90     53   12   23   21   30   29
YAL066W         2      0      0      0      0    0    0    0    0    0
YAL065C         5      9      6      3      1   10    5    2    4    3
YAL064W-B      13      9     10      9      6    9   12    4    4    8
```



DE basics



1 test per gene!

1. Estimate **magnitude** of DE taking into account differences in sequencing depth, technical, and biological read count variability.

logFC

2. Estimate the **significance** of the difference accounting for performing thousands of tests.

(adjusted) p-value

H0: *no difference in the read distribution between two conditions*

Garber et al. (2011) Nature Methods, 8(6), 469–477. doi:10.1038/nmeth.1613

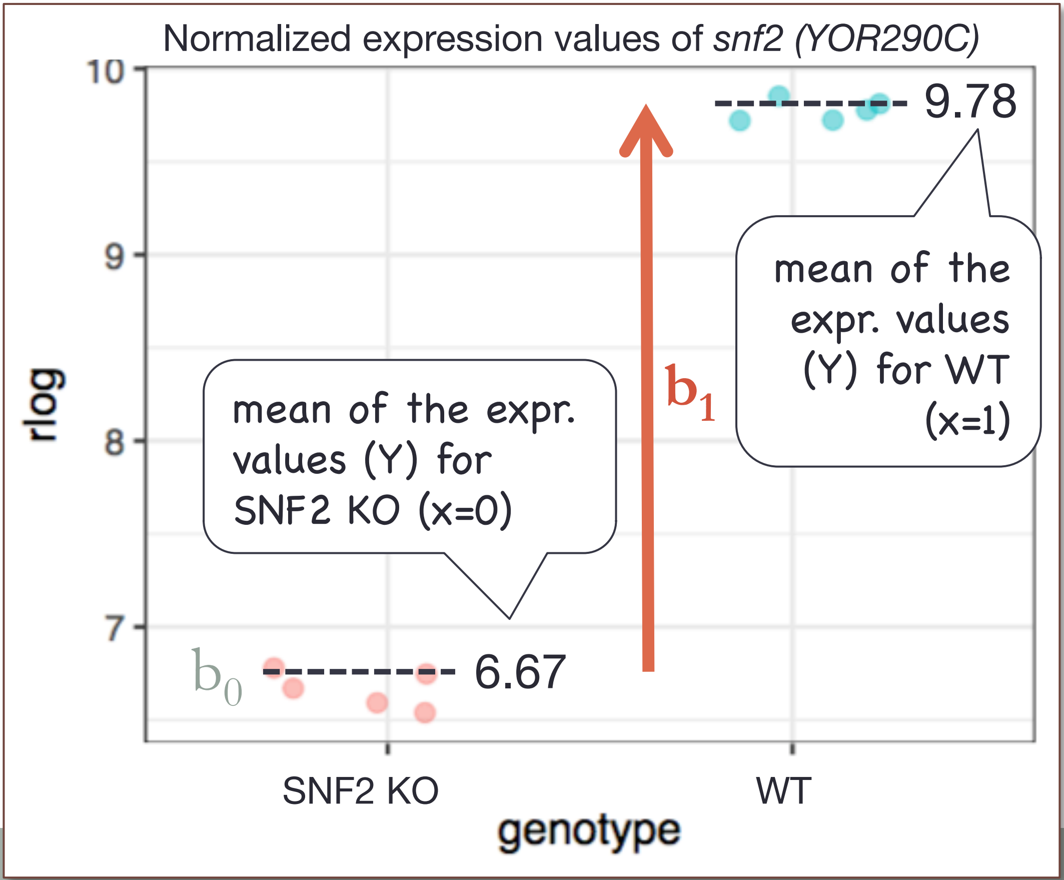
Estimating the difference with regression models

Example: Modeling normalized gene expression values using a linear model

describing all normalized expression values of one example gene using a simple linear model of the following form:

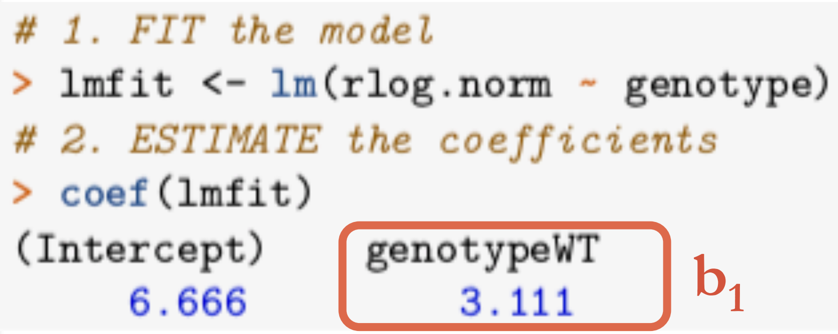
$$\underset{\text{expr. values}}{Y} = \underset{\text{intercept}}{b_0} + \underset{\text{delta}}{\boxed{b_1}} * \underset{\text{genotype}}{x} + e$$

b_0: **intercept**, i.e. average of the baseline group
b_1: **difference** between baseline & non-reference group
x : 0 if genotype == “SNF2”, 1 if genotype == “WT”



mean of the expr. values (Y) for WT (x=1)

mean of the expr. values (Y) for SNF2 KO (x=0)



```
# 1. FIT the model
> lmfit <- lm(rlog.norm ~ genotype)
# 2. ESTIMATE the coefficients
> coef(lmfit)
(Intercept)    genotypeWT
      6.666         3.111
```

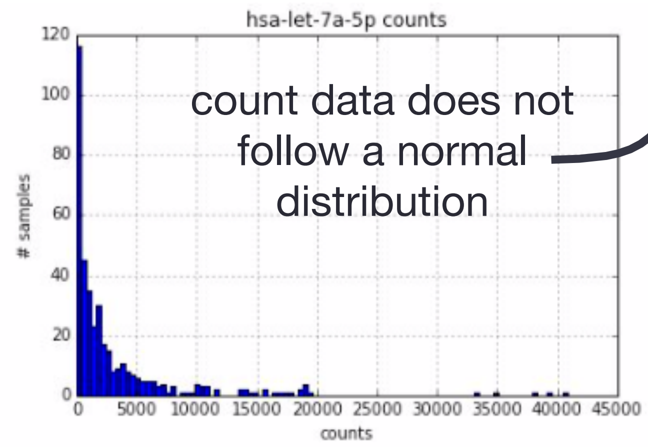
b_1

b_0 b_1

both beta values are **estimates**!
(they’re spot-on because the data is so clear for this example and the model is so simple)

DE analysis: dealing with raw read counts

1. **Fitting** a sophisticated model (not a basic linear model) to get a grip on the read counts (done per gene; includes normalization)
   - library size factor
   - dispersion estimate using information across multiple genes
   - assuming a neg. binomial distribution of read counts



count data does not follow a normal distribution

negative binomial (NB) model

$$K_{ij} \sim \mathrm{NB}(\boxed{\mu_{ij}}, \alpha_i)$$

read counts for gene *i* and sample *j*

gene-specific dispersion parameter (fitted towards the average dispersion)

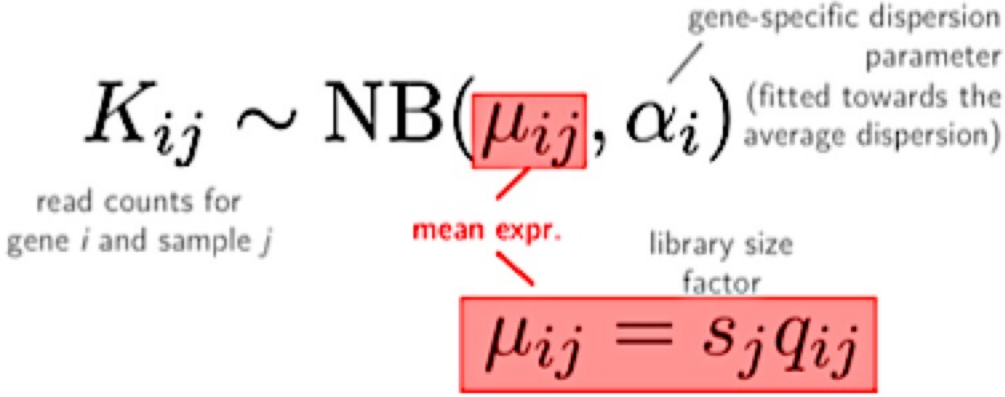
mean expr.

library size factor

$$\boxed{\mu_{ij} = s_j q_{ij}}$$

DE analysis

1. **Fitting** a sophisticated model to get a grip on the read counts (done per gene; includes normalization)



$$K_{ij} \sim \mathrm{NB}(\boxed{\mu_{ij}}, \alpha_i)$$

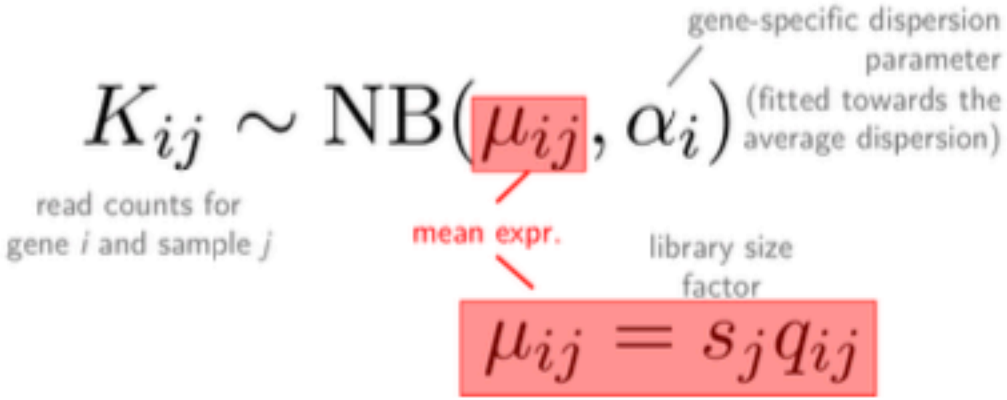
$$\boxed{\mu_{ij} = s_j q_{ij}}$$

gene-specific dispersion parameter (fitted towards the average dispersion)
read counts for gene *i* and sample *j*
mean expr.
library size factor

2. Estimating **coefficients** of the model to obtain the difference between the estimated mean expression of the different groups (log2FC)
   - define the **contrast of interest**, e.g. ~ batchEffect + condition
   - always put **the factor of interest last**
   - order of the factor levels determines the direction of log2FC

DE analysis

1. **Fitting** a sophisticated model to get a grip on the read counts (done per gene; includes normalization)



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

$$\mu_{ij} = s_j q_{ij}$$

gene-specific dispersion parameter (fitted towards the average dispersion)
read counts for gene *i* and sample *j*
mean expr.
library size factor

2. Estimating **coefficients** of the model to obtain the difference between the estimated mean expression of the different groups (log2FC)

3. **Test** whether the log2FC is “far away” from 0
   - log-likelihood test or Wald test are used by DESeq2
   - multiple hypothesis test correction

Modeling read counts and estimating the log2-fold-change (DESeq2)

fitted mean

gene-specific dispersion parameter (fitted towards the average dispersion)

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

read counts for gene *i* and sample *j*

library size factor

expression value estimate

$$\mu_{ij} = s_j q_{ij}$$

moderated log-fold change for gene *i*

$$\log_2(q_{ij}) = x_{j.}\beta_i$$

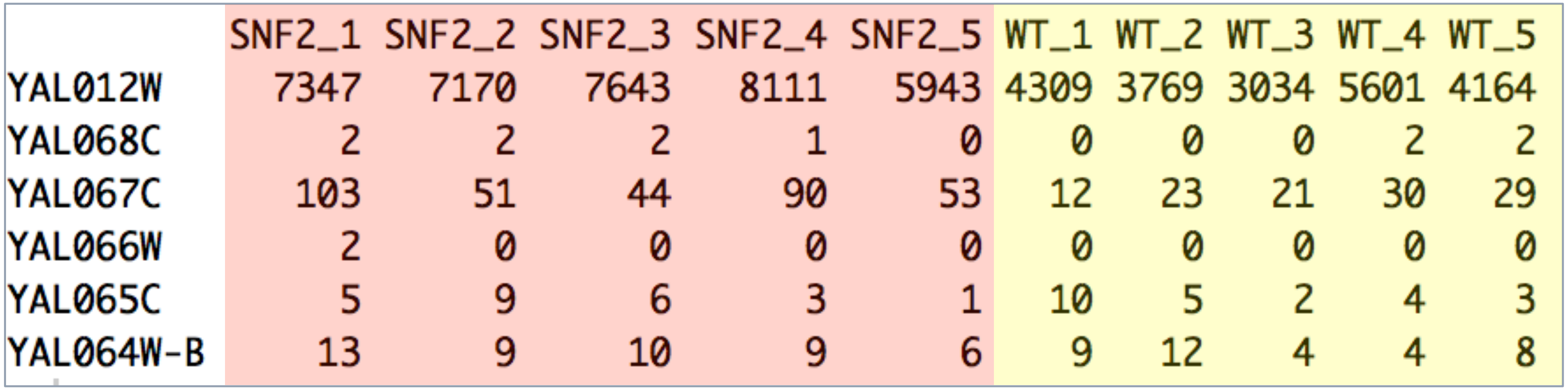
model matrix column for sample *j*

Once the coefficients are estimated, the significance tests need to test how far away from zero they are since zero would mean “no difference”.

H0: *no difference in the read distribution between two conditions*

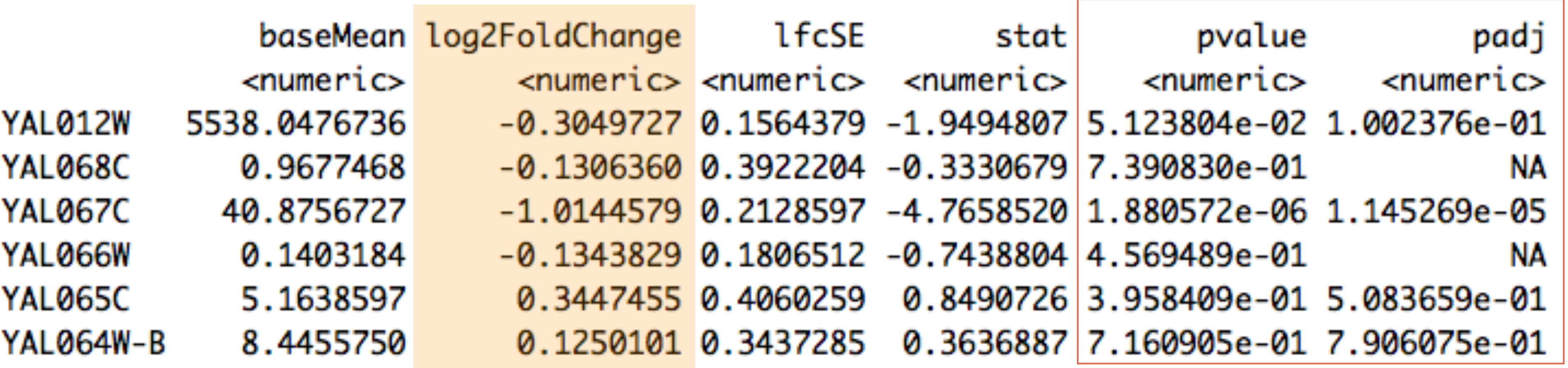
Let’s do this!

From read counts to DE



```
          SNF2_1 SNF2_2 SNF2_3 SNF2_4 SNF2_5 WT_1 WT_2 WT_3 WT_4 WT_5
YAL012W     7347   7170   7643   8111   5943 4309 3769 3034 5601 4164
YAL068C        2      2      2      1      0    0    0    0    2    2
YAL067C      103     51     44     90     53   12   23   21   30   29
YAL066W        2      0      0      0      0    0    0    0    0    0
YAL065C        5      9      6      3      1   10    5    2    4    3
YAL064W-B     13      9     10      9      6    9   12    4    4    8
```

```
DESeq2::DESeq(ds_object)
```



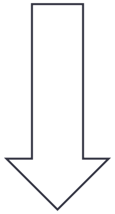
```
              baseMean log2FoldChange     lfcSE       stat       pvalue         padj
             <numeric>      <numeric> <numeric>  <numeric>    <numeric>    <numeric>
YAL012W   5538.0476736     -0.3049727 0.1564379 -1.9494807 5.123804e-02 1.002376e-01
YAL068C      0.9677468     -0.1306360 0.3922204 -0.3330679 7.390830e-01           NA
YAL067C     40.8756727     -1.0144579 0.2128597 -4.7658520 1.880572e-06 1.145269e-05
YAL066W      0.1403184     -0.1343829 0.1806512 -0.7438804 4.569489e-01           NA
YAL065C      5.1638597      0.3447455 0.4060259  0.8490726 3.958409e-01 5.083659e-01
YAL064W-B    8.4455750      0.1250101 0.3437285  0.3636887 7.160905e-01 7.906075e-01
```

average norm. count

standard error estimate for the logFC

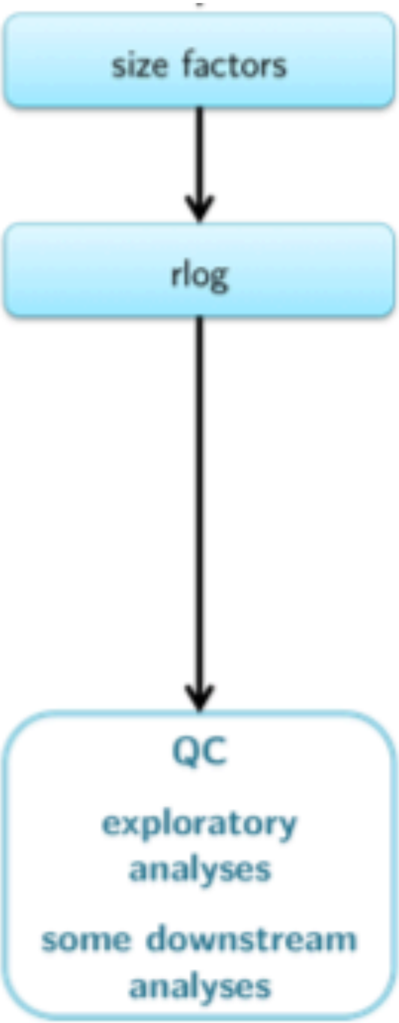
Exploratory vs. DE analysis workflow

raw reads



NORMALIZATION

- lib sizes
- variance
- log-transformation

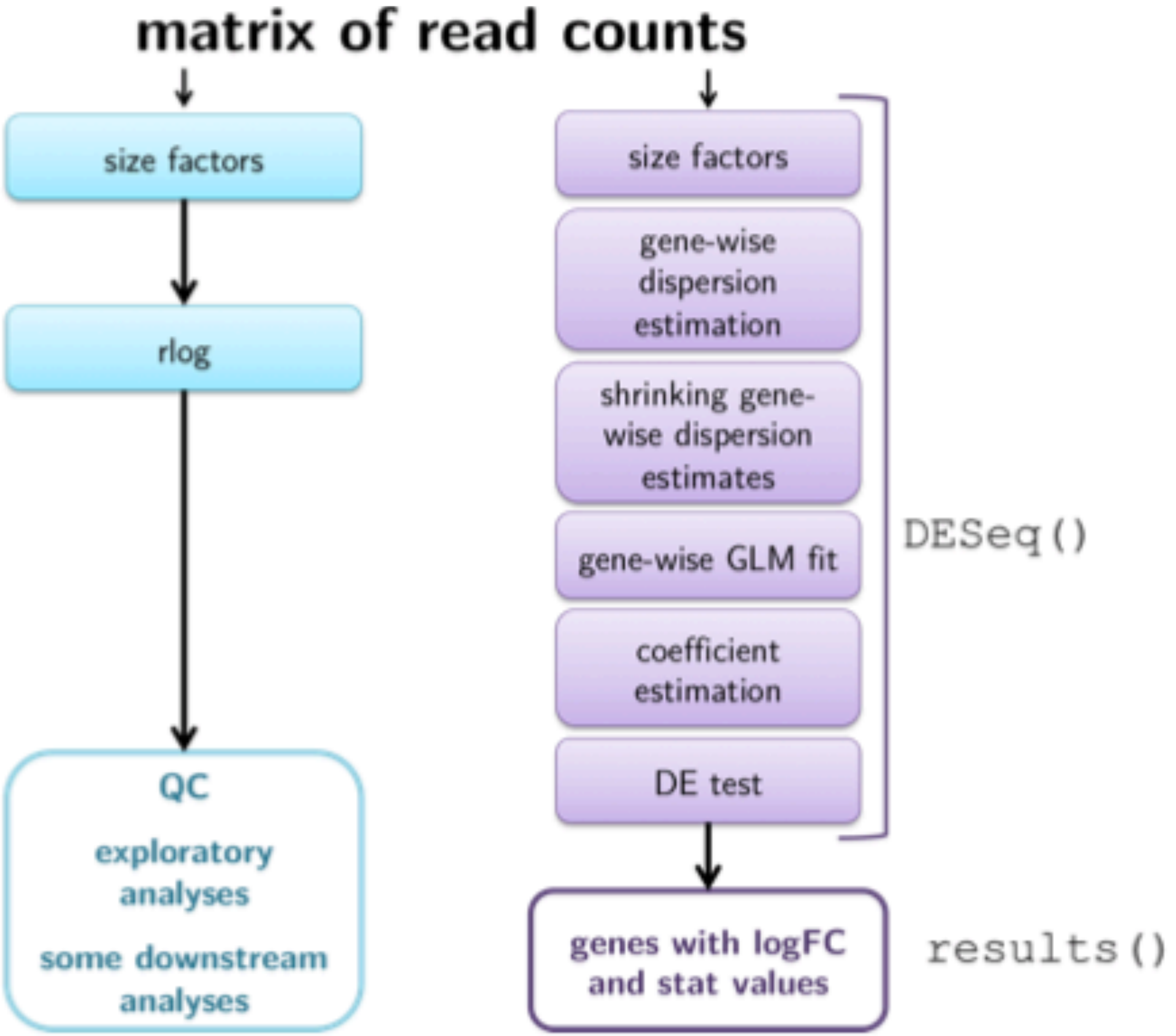


Exploratory vs. DE analysis workflow

raw reads

NORMALIZATION

- lib sizes
- variance
- log-transformation



matrix of read counts

size factors

rlog

QC

exploratory analyses

some downstream analyses

size factors

gene-wise dispersion estimation

shrinking gene-wise dispersion estimates

gene-wise GLM fit

coefficient estimation

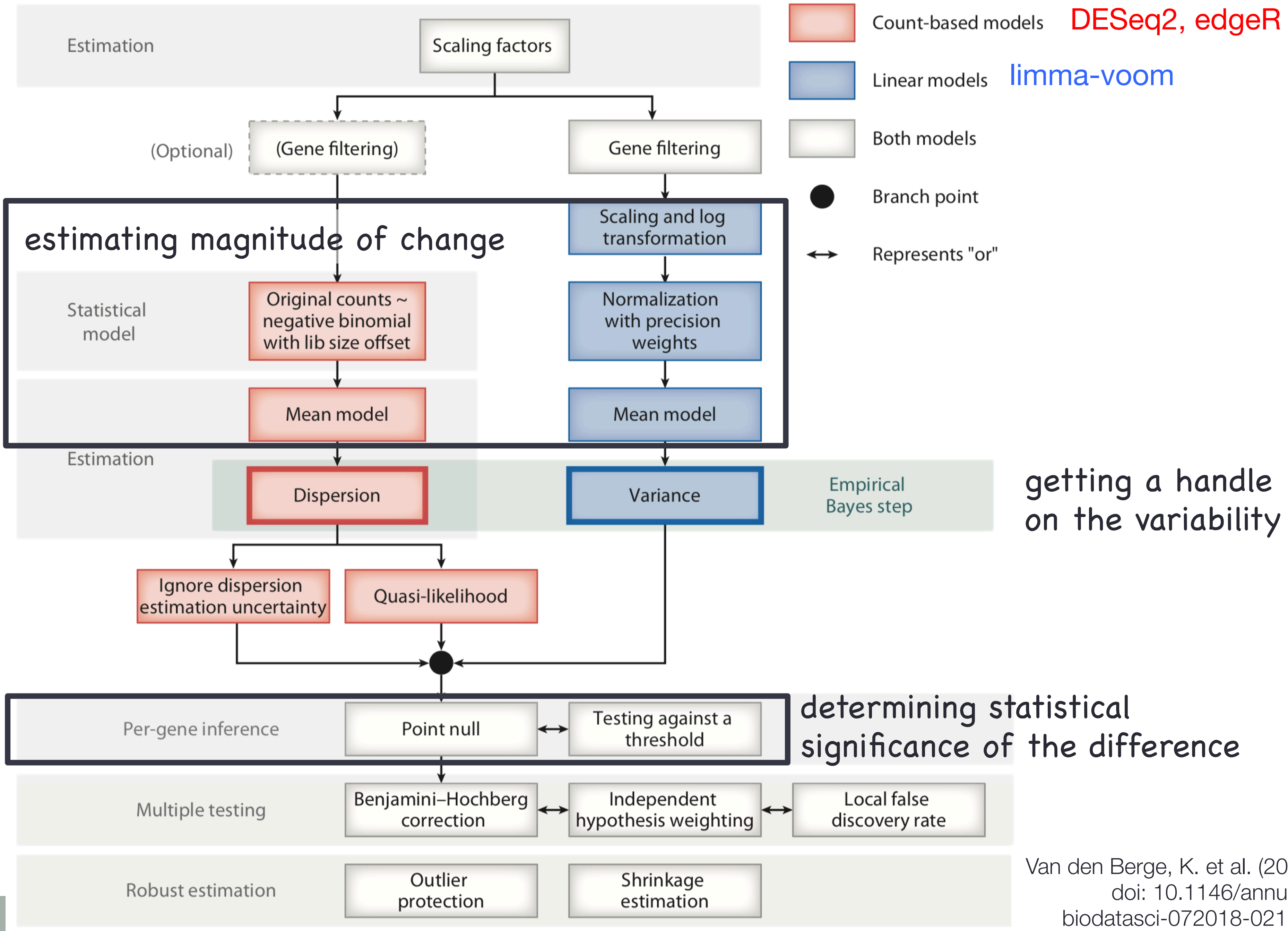
DE test

DESeq()

genes with logFC and stat values

results()

DESeq2 vs. edgeR vs. limma-voom



🟥 Count-based models DESeq2, edgeR

🟦 Linear models limma-voom

⬜ Both models

● Branch point

↔ Represents "or"

Estimation

Scaling factors

(Optional) (Gene filtering)

Gene filtering

estimating magnitude of change

Scaling and log transformation

Statistical model

Original counts ~ negative binomial with lib size offset

Normalization with precision weights

Mean model

Mean model

Estimation

Dispersion

Variance

Empirical Bayes step

getting a handle on the variability

Ignore dispersion estimation uncertainty

Quasi-likelihood

Per-gene inference

Point null ↔ Testing against a threshold

determining statistical significance of the difference

Multiple testing

Benjamini–Hochberg correction ↔ Independent hypothesis weighting ↔ Local false discovery rate

Robust estimation

Outlier protection

Shrinkage estimation

Van den Berge, K. et al. (2019).
doi: 10.1146/annurev-biodatasci-072018-021255

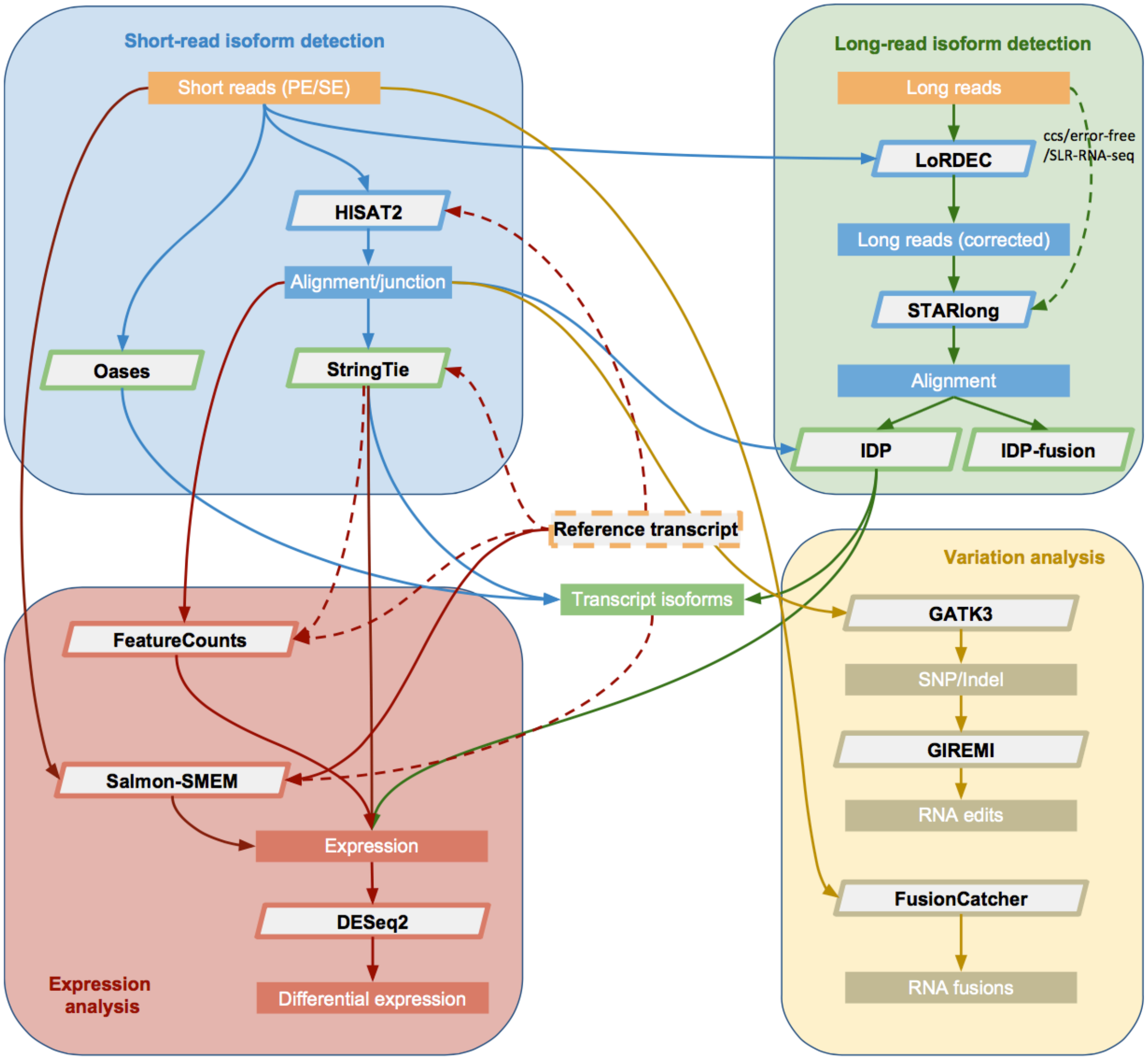
What next?

- Do your results make sense?
- Are the results robust?
  - do **multiple tools** agree on the majority of the genes?
  - are **the fold changes** strong enough to explain the phenotype you are seeing?
  - have **other experiments** yielded similar results?
- Downstream analyses: mostly **exploratory**

How to decide which tool(s) to use?

- function/content of original publication
- code maintained?
- well documented?
- used by others?
- efficient?

RNACocktail tries to implement all (current!) best performers for various RNA-seq analyses



Task	Command
Short-read alignment	`align`
Short-read transcriptome reconstruction	`reconstruct`
Short-read quantification	`quantify`
Short-read differential expression	`diff`
Short-read de novo assembly	`denovo`
Long-read error correction	`long_correct`
Long-read alignment	`long_align`
Long-read transcriptome reconstruction	`long_reconstruct`
Long-read fusion detection	`long_fusion`
Variant calling	`variant`
RNA editing detection	`editing`
RNA Fusion detection	`fusion`
Running all steps	`all`

Sahraeian et al. (2017). Nat Comm, 8(1), 59. https://bioinform.github.io/rnacocktail/

Where to get help and inspiration?

bioconductor.org/help/workflows

F100Research Software Tool Articles

Periodic Table of Bioinformatics:
http://elements.eaglegenomics.com/

mailing lists/github issues of the individual tools

biostars.org
seqanswers.com
stackoverflow.com

WALK-IN CLINICS

@ WCM:
Thursdays, 1:30 – 3 pm,
LC-504 (1300 York Ave)

abc.med.cornell.edu

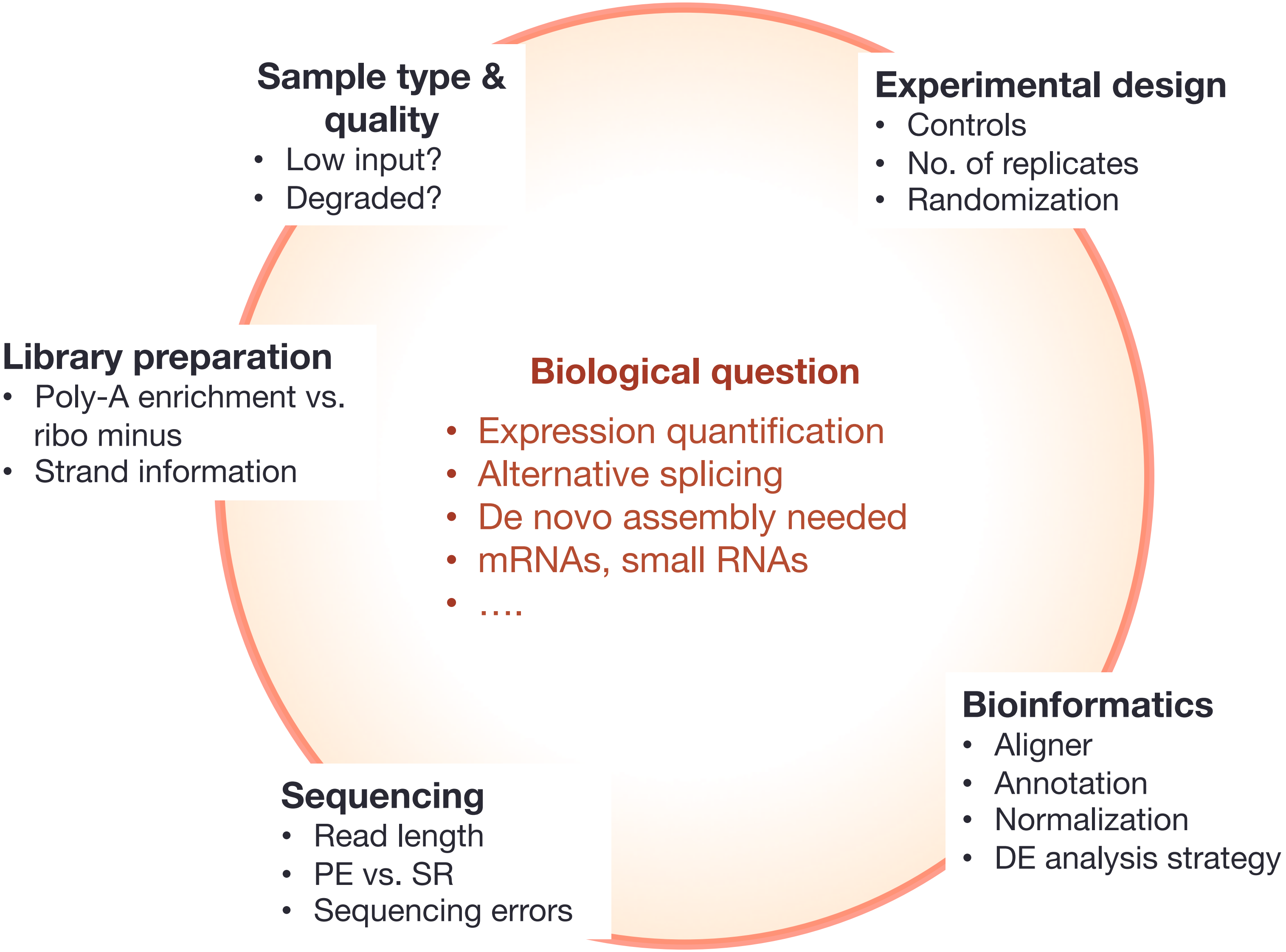
@ MSKCC:
https://www.mskcc.org/research-advantage/core-facilities/bioinformatics

Picardi: RNA Bioinformatics (2015)
https://www.springer.com/us/book/9781493922901

https://github.com/abcdbug/dbug

supplemental material of publications based on HTS data

Everything's connected…



Sample type & quality

- Low input?
- Degraded?

Experimental design

- Controls
- No. of replicates
- Randomization

Library preparation

- Poly-A enrichment vs. ribo minus
- Strand information

Biological question

- Expression quantification
- Alternative splicing
- De novo assembly needed
- mRNAs, small RNAs
- ….

Bioinformatics

- Aligner
- Annotation
- Normalization
- DE analysis strategy

Sequencing

- Read length
- PE vs. SR
- Sequencing errors