LEEDS
OMICS

genomics
transcriptomics
proteomics
metabolomics

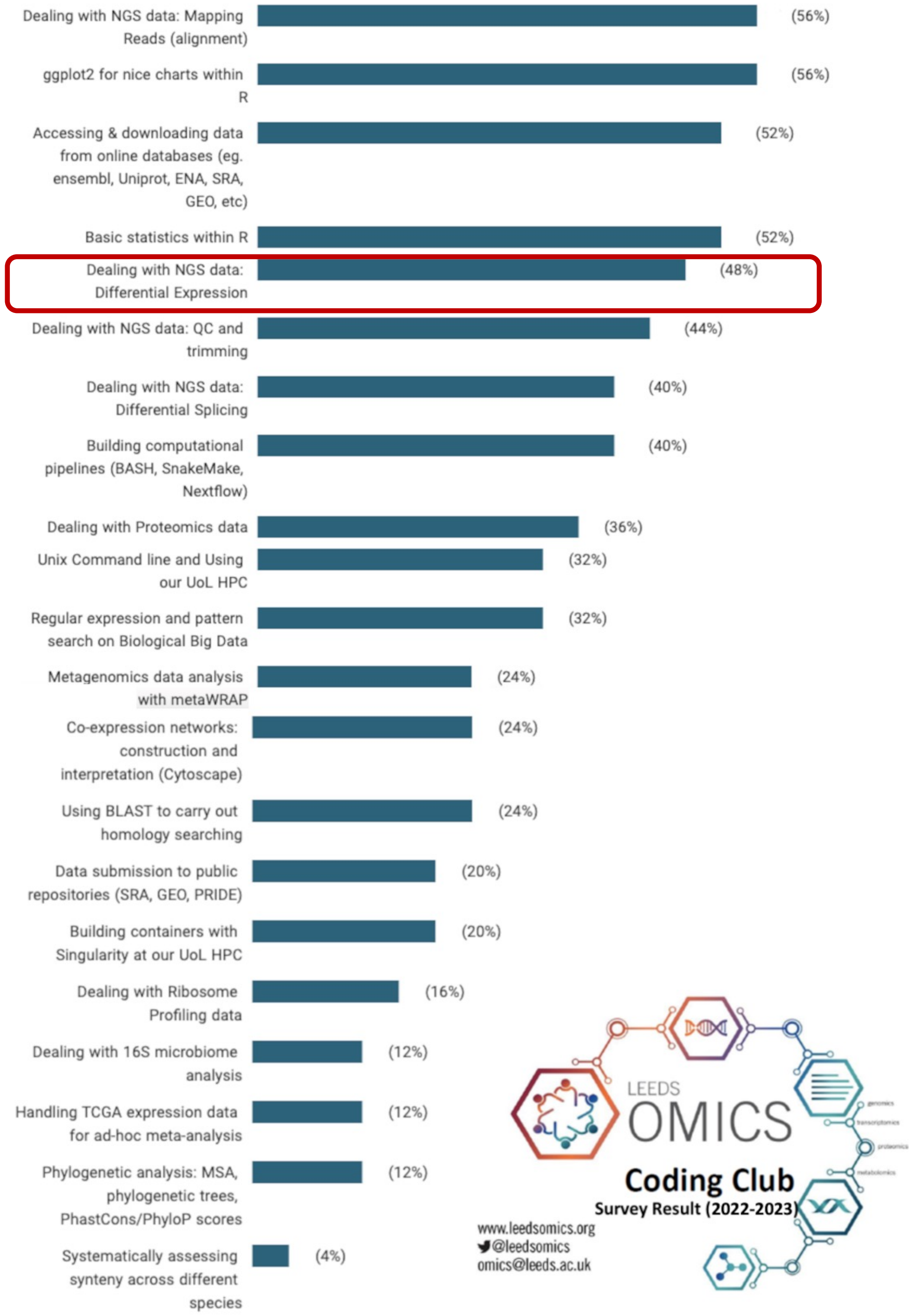
Coding Club

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Dealing with NGS data:
Differential Expression

Club Moderators: Elton Vasconcelos and Euan McDonell

Topics to be addressed on the 2022-23 season - **Survey Result**



5th session

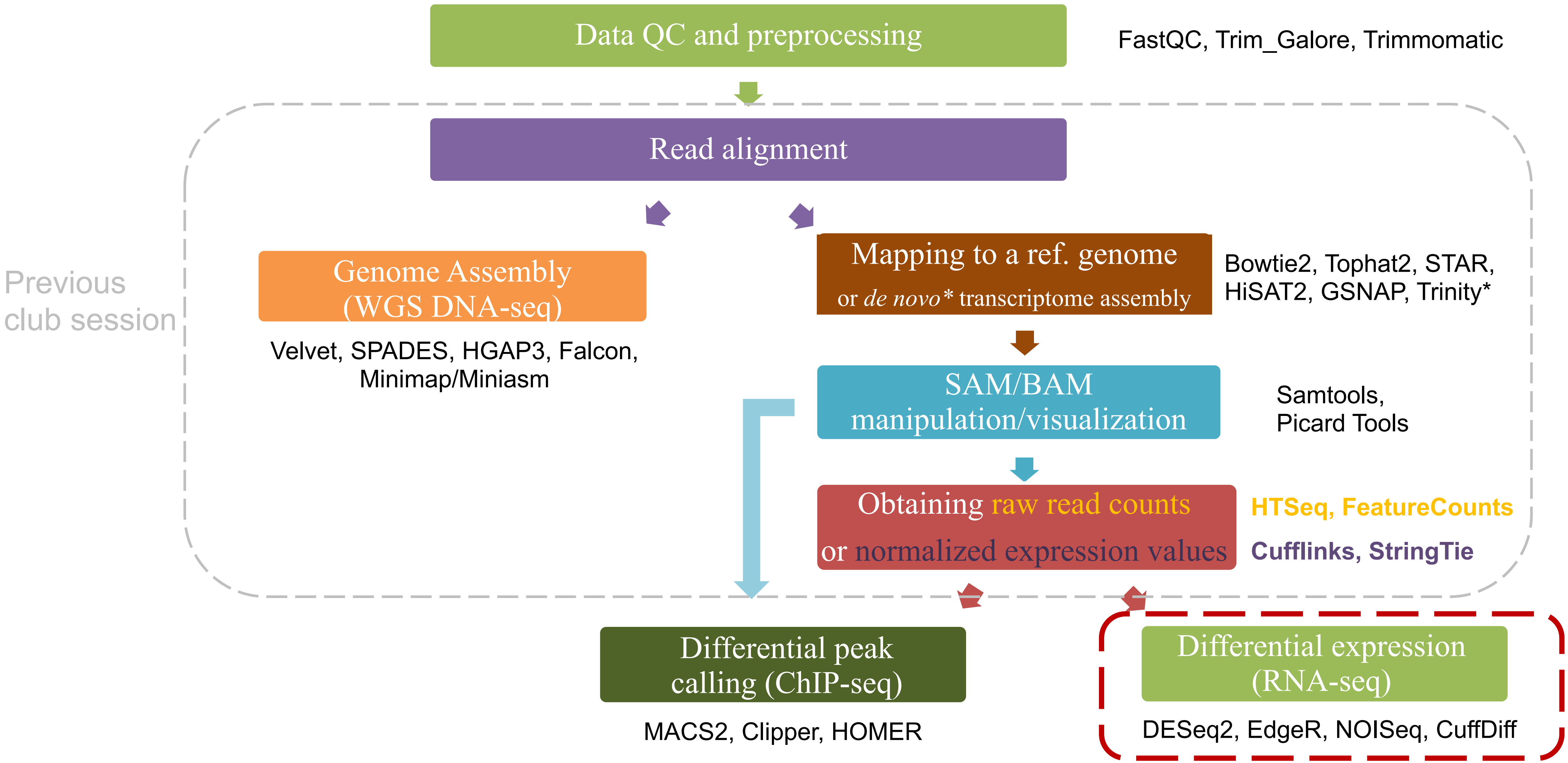
LEEDS OMICS

Coding Club

Survey Result (2022-2023)

www.leedsomics.org
@leedsomics
omics@leeds.ac.uk

Important steps on NGS data analysis workflow



Data QC and preprocessing

FastQC, Trim_Galore, Trimmomatic

Read alignment

Previous club session

Genome Assembly (WGS DNA-seq)

Velvet, SPADES, HGAP3, Falcon, Minimap/Miniasm

Mapping to a ref. genome or *de novo** transcriptome assembly

Bowtie2, Tophat2, STAR, HiSAT2, GSNAP, Trinity*

SAM/BAM manipulation/visualization

Samtools, Picard Tools

Obtaining raw read counts or normalized expression values

HTSeq, FeatureCounts

Cufflinks, StringTie

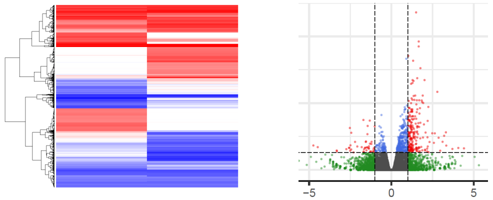
Differential peak calling (ChIP-seq)

MACS2, Clipper, HOMER

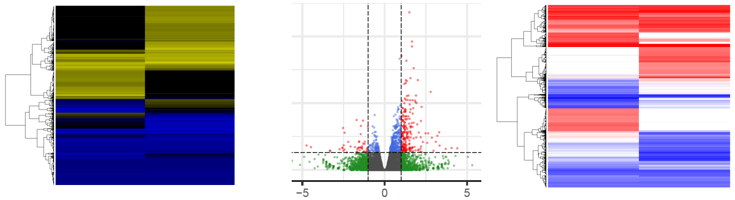
Differential expression (RNA-seq)

DESeq2, EdgeR, NOISeq, CuffDiff

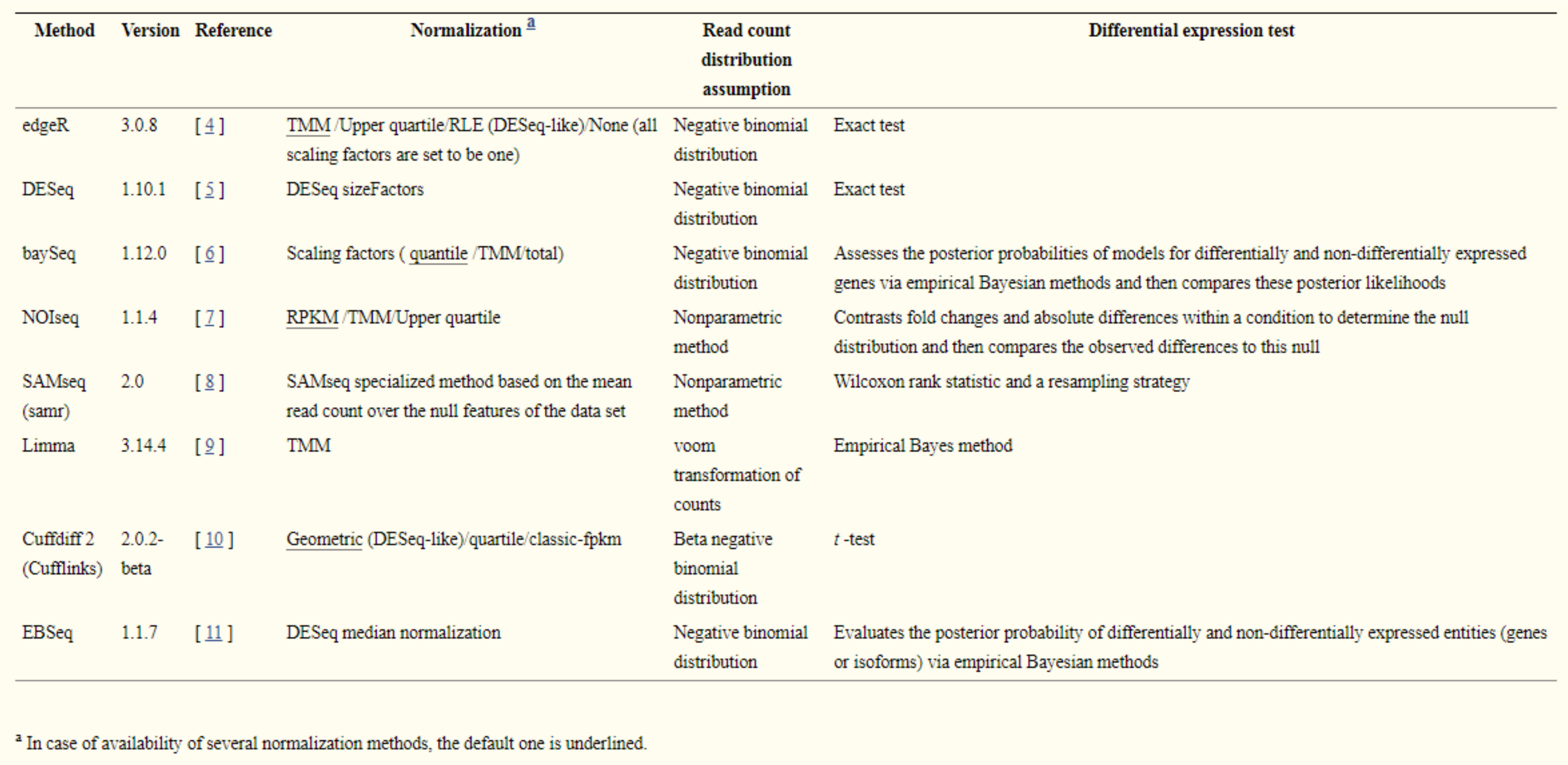
Today's session



Differential Expression



Software packages for DE analysis



Method	Version	Reference	Normalization [a]	Read count distribution assumption	Differential expression test
edgeR	3.0.8	[4]	TMM /Upper quartile/RLE (DESeq-like)/None (all scaling factors are set to be one)	Negative binomial distribution	Exact test
DESeq	1.10.1	[5]	DESeq sizeFactors	Negative binomial distribution	Exact test
baySeq	1.12.0	[6]	Scaling factors (quantile /TMM/total)	Negative binomial distribution	Assesses the posterior probabilities of models for differentially and non-differentially expressed genes via empirical Bayesian methods and then compares these posterior likelihoods
NOIseq	1.1.4	[7]	RPKM /TMM/Upper quartile	Nonparametric method	Contrasts fold changes and absolute differences within a condition to determine the null distribution and then compares the observed differences to this null
SAMseq (samr)	2.0	[8]	SAMseq specialized method based on the mean read count over the null features of the data set	Nonparametric method	Wilcoxon rank statistic and a resampling strategy
Limma	3.14.4	[9]	TMM	voom transformation of counts	Empirical Bayes method
Cuffdiff 2 (Cufflinks)	2.0.2-beta	[10]	Geometric (DESeq-like)/quartile/classic-fpkm	Beta negative binomial distribution	*t* -test
EBSeq	1.1.7	[11]	DESeq median normalization	Negative binomial distribution	Evaluates the posterior probability of differentially and non-differentially expressed entities (genes or isoforms) via empirical Bayesian methods

[a] In case of availability of several normalization methods, the default one is underlined.

→ Important Output Metrics: Log$_2$(FC), p-value and FDR provided in most output files

Laiho & Elo., 2015 – *Brief. Bioinform.*

The DESeq2 model

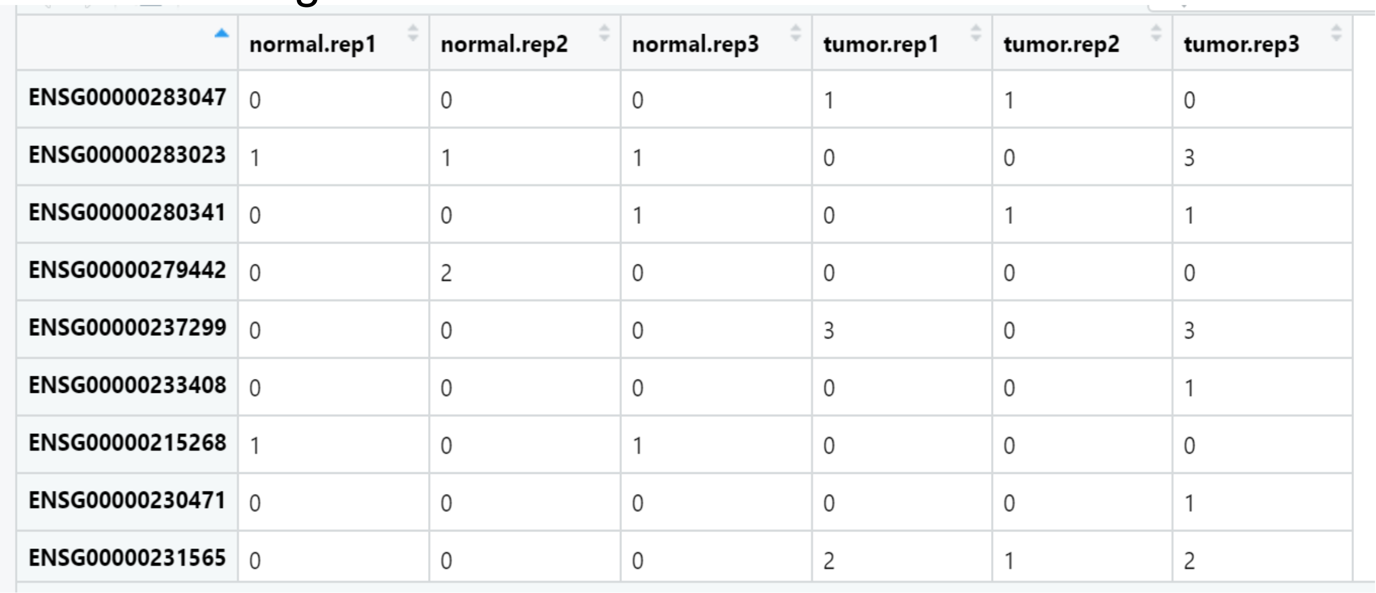
- Perform a “*median of ratios*” normalization to correct for library size and RNA composition bias (counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene);

- Use shrinkage estimation for dispersions and fold changes because small numbers of replicates make it impossible to estimate within-group variance reliably;

- Fit negative binomial generalized linear models for each gene and uses the Wald test for significance testing.

Prepare the data for DESeq2 analysis

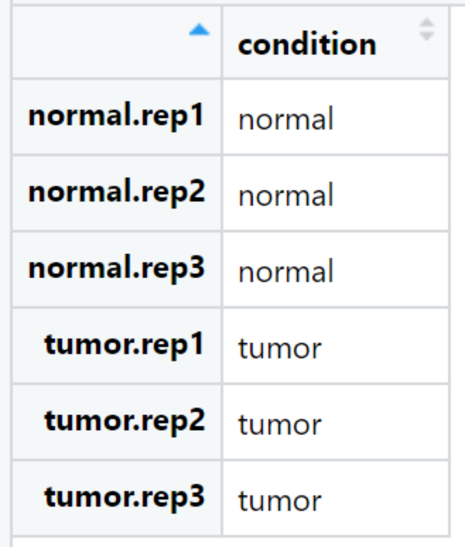
countData: a matrix of non-negative integers



	normal.rep1	normal.rep2	normal.rep3	tumor.rep1	tumor.rep2	tumor.rep3
ENSG00000283047	0	0	0	1	1	0
ENSG00000283023	1	1	1	0	0	3
ENSG00000280341	0	0	1	0	1	1
ENSG00000279442	0	2	0	0	0	0
ENSG00000237299	0	0	0	3	0	3
ENSG00000233408	0	0	0	0	0	1
ENSG00000215268	1	0	1	0	0	0
ENSG00000230471	0	0	0	0	0	1
ENSG00000231565	0	0	0	2	1	2

count.data

colData: a DataFrame with at least a single column. Rows of colData correspond to columns of countData.



	condition
normal.rep1	normal
normal.rep2	normal
normal.rep3	normal
tumor.rep1	tumor
tumor.rep2	tumor
tumor.rep3	tumor

design: a formula expressing the variables which will be used in modelling.

metadata

The main **three steps** of running DESeq2

1. Create a DESeqDataSet object from input. Please note that the colnames of countData must be identical to the rownames of colData.

```
keep = rowSums(count.data) >= 1

count.data.keep = count.data[keep,])

dds <- DESeqDataSetFromMatrix(countData = count.data.keep, colData =
metadata, design = ~ condition)
```

2. Perform the differential expression analysis.

```
dds <- DESeq(dds, fitType = "local")
```

3. Extract a results table.

```
res <- results(dds)

write.table(res[order(res$padj),], file=”resultsDESeq2.tsv", sep = "\t",
quote=F, col.names=NA)
```

Bring your issues on!

Bring your issues on!