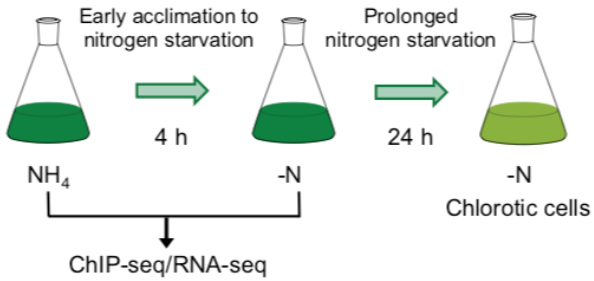
Practice 2. RNAseq Analysis.

Joaquín Giner Lamia

In this practice we will analyze 4 samples from the cyanobacterium *Synechocystis* SP. PCC 6803 grown in ammonium (NH4; 2 samples) and after transition to low nitrogen media (N2; 2 samples) previously published in this work (*Identification of the direct regulon of NtcA during early acclimation to nitrogen starvation in the cyanobacterium Synechocystis sp*. PCC6803. 2017. *Nucleic Acids Research*).



In this practice we are going to align the fastq files to *Synechocystis* genome, normalize and visualize the aligned reads in IGV and perform the Differential Expression analysis of NH4 and N2 samples.

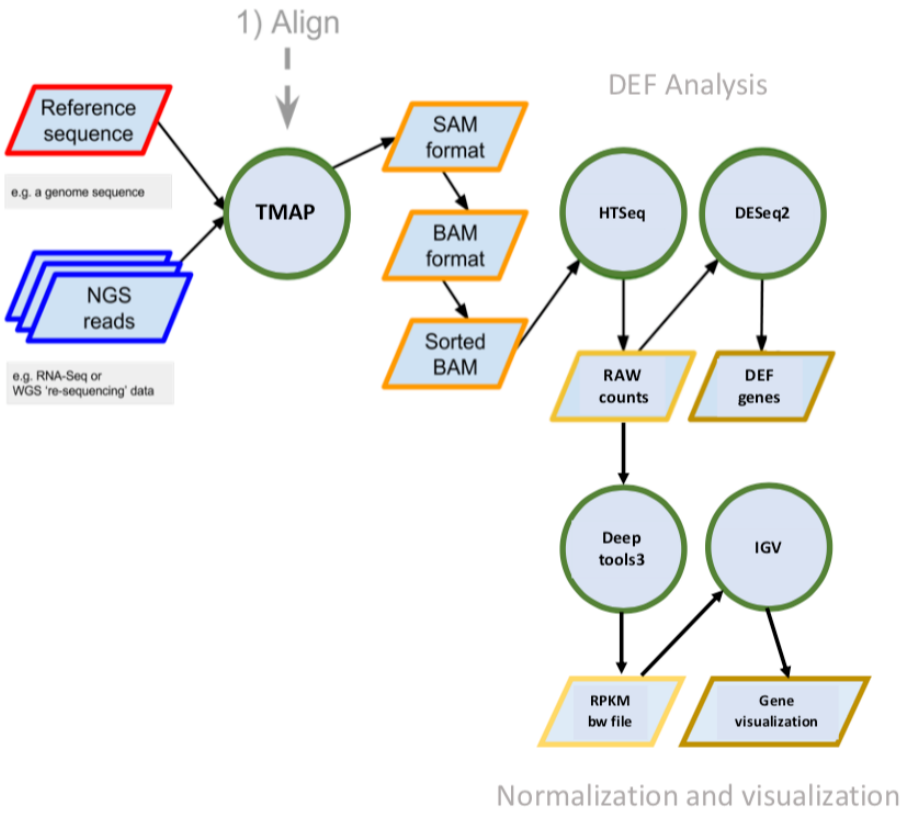


Figure 1. **Pipeline overview of this practice**

General RNA-seq analysis pipeline

This is a general RNA-seq analysis workflow. In our practices we are going to use the same steps for our analysis excepting we changed TopHat2 by BBmap to align our raw reads (fastq files).

TopHat2 is an excellent mapper for Eukaryotic samples, since it analyzes the mapping results from Bowtie2 (used in the TopHat2 pipeline) to identify splice junctions between exons. In our case, we have to use BBmap instead of Bowtie2 cause our reads come from Ion Torrent sequence technology.

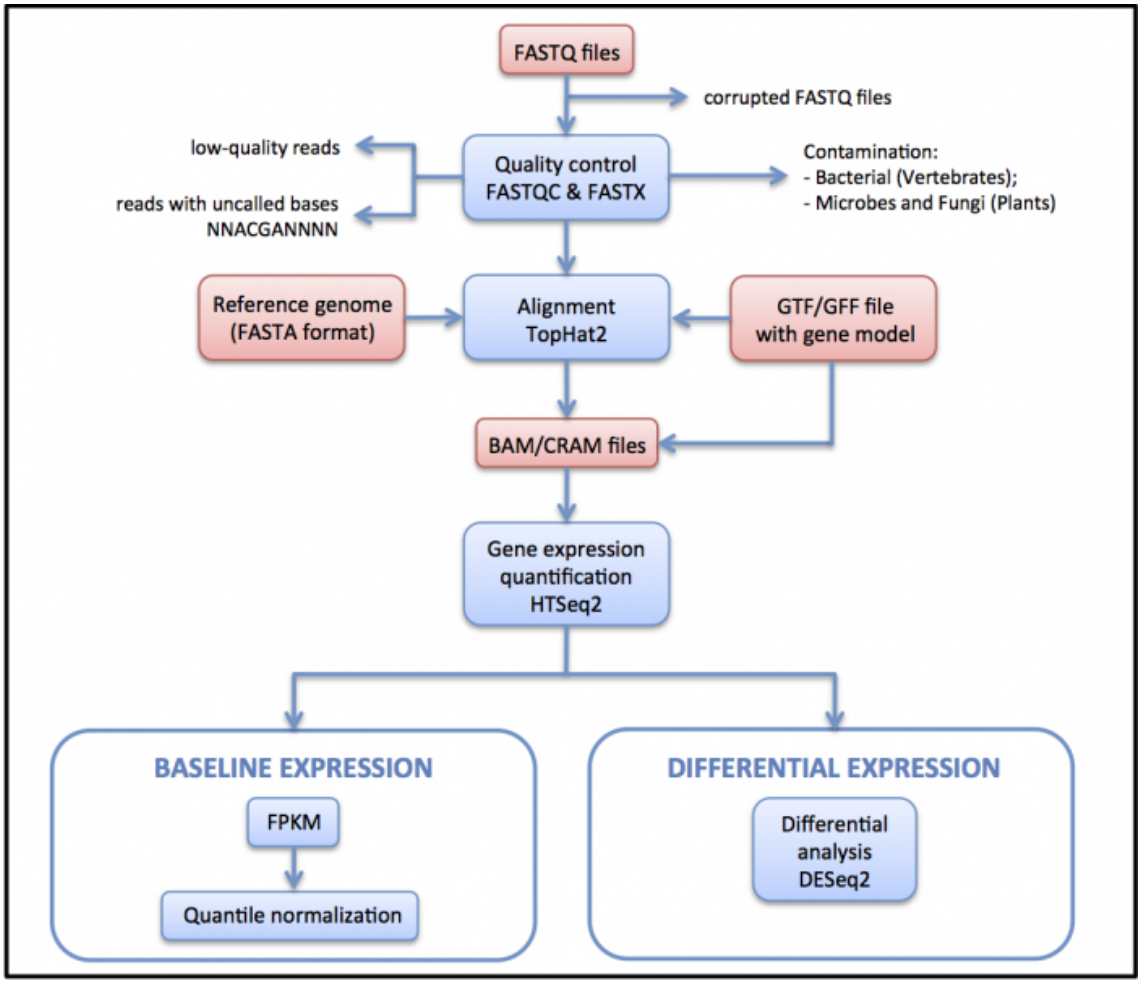


Figure 2. RNA-seq processing pipeline used to generate gene expression data

A. Preparing the practice.

This section is not necessary in case of you have already installed ubuntu 20.04 VM.

1. The data necessary for this practice are available at Zenodo repository

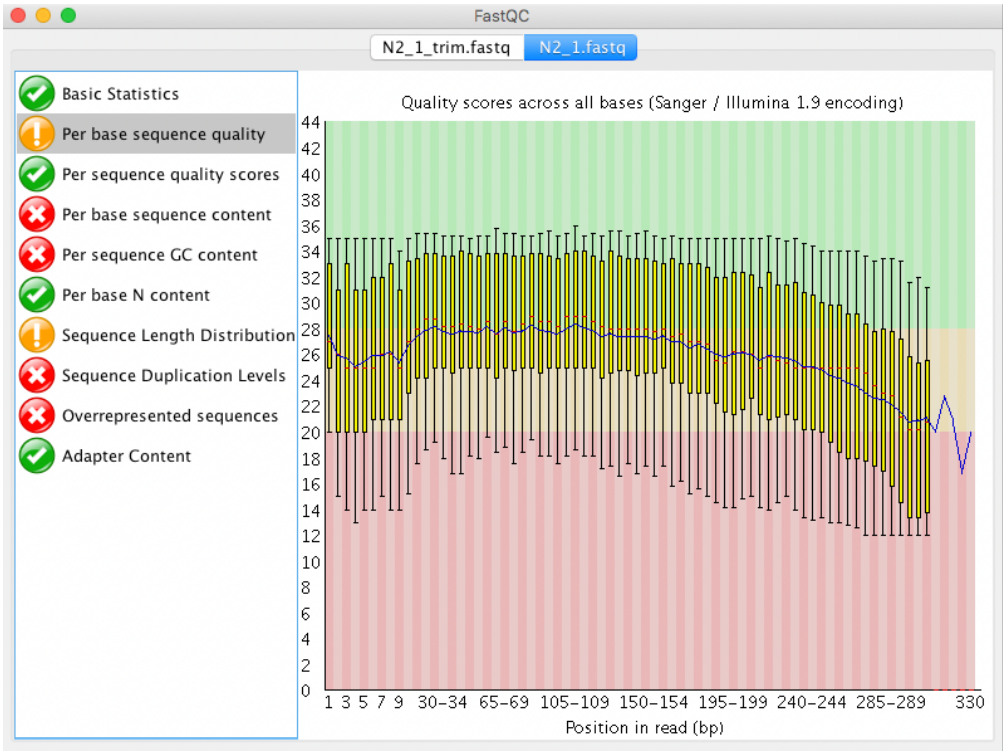
Practica_2_RNAseq:

https://zenodo.org/record/1466050#.W8nF_y-B0Wo

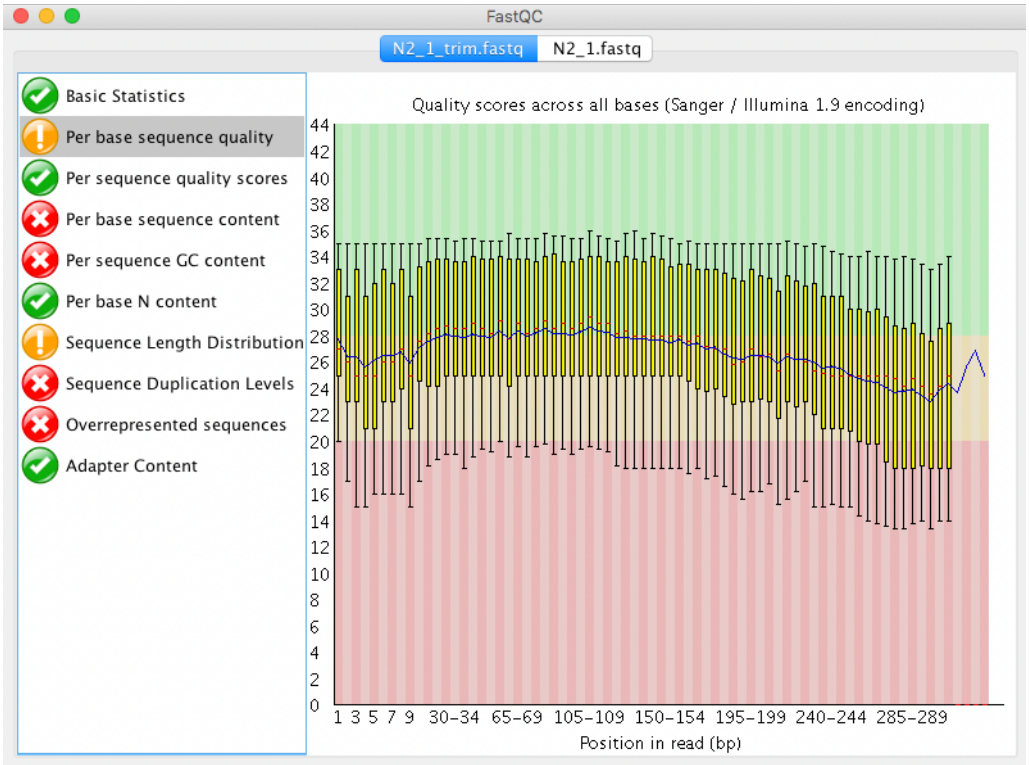
B. Quality control analysis of the sequencing reads using FastQC.

Before analyzing your sequences, you should always carry out quality control of the raw sequence data to identify potential artifacts. The FastQC (Figure 3) software contains different analysis modules including: (i) Per base sequencing quality (the higher the score the better the base call; in any case the lower quartile for any base should be higher than 10); (ii) Per base sequence content (this should show a non-random distribution of the nucleotide at each base; differences between A and T, or G and C should not be greater than 10% for any position); and (iii) Duplicate sequences (non-unique sequences should not constitute more than 20% of the total sequences). More information on FastQC modules is available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/.

Here an example of a trimmed and not trimmed fastq files using:



Not trimmed fastqc



Trimmed fastqc

C. Alignment of the reads to the genome using BBmap.

The reference genome for *Synechocystis* sp. PCC 6803 has the GenBank assembly accession number: GCA_000009725.1; RefSeq: NC_009911.1. In our case, the genome file is *NC_009911.1.fasta*. This genome file is available on the Zenodo tutorial page together the fastq files for the RNAseq analysis. These fastq files were obtained from an Ion Torren PGM sequencer.

Although Bowtie2 don´t work with Ion torrent fastq files we can use other options such as BBmap. Thus, for each sample, we will map the fastq files to the reference genome using the BBmap program.

BBmap index

```
:$ bbmap.sh ref=<genome_file.fasta>
```

#BBmap alignment

```
:$ bbmap.sh in=<read_file.fastq> out=<aligned_file.sam>
```

Generate a shell script.sh (bbmap.sh) containing all the command for index and mapping, to align all the fastq files programmatically.

Explore the CIGAR of the SAM files generated.

D. Convert the SAM formats into BAM format, using SAMTOOLS.

To analyze our alignment reads, we need to transform the format of the SAM file obtained from TMAP to work more efficiently with the aligned reads. SAM format files are very large files and have to be converted into a Binary Alignment Map (.BAM) format.

The generic command lines to transform a SAM file into a sorted BAM file in SAMTOOLS are:

```
:$ samtools view -bS <aligned_file.sam> > <aligned_file.bam>
:$ samtools sort <aligned_file.bam> >  <aligned_file_sorted>
:$ samtools index <aligned_file_sorted.bam>
```

E. Expression. Calculate raw counts with HTSseq-count

Run htseq-count to produce raw counts instead of FPKM/TPM values for differential expression analysis

Refer to the HTSeq documentation for a more detailed explanation:

• http://www-huber.embl.de/users/anders/HTSeq/doc/count.html htseq-count basic usage:

htseq-count [options] <sam_file> <gff_file> extra options specified below:

- '-f --format' specify the input file format one of BAM or SAM. Since we have BAM format files, select 'bam' for this option.
- '-m --mode' determines how to deal with reads that overlap more than one feature. In this case we will use 'intersection-strict' mode. (see htseq-count documentation for more information)
- '-t --type' specifies the feature type (3rd column in GFF file) to be used. (default, suitable for RNA-Seq and Ensembl GTF files: exon)
- '-i --idattr' The feature ID used to identity the counts in the output table. The default, suitable for RNA-SEq and Ensembl GTF files, is gene_id.

Run htseq-count and calculate gene-level counts:

```
htseq-count -m union -i locus_tag -t gene -f bam sorted.bam NC_000911.1.gff >
<bam.count>
```

Once read counts have been calculated for bam files, use the join_HTseq.sh script to generate merged file contained all counts. (Note: take into account the name of count files you used in htseq count script and change it conveniently in join_HTseq.sh)

Normalization of read counts

Several factors preclude raw read counts across different libraries (and across genes within the same library) from being compared directly. The most obvious of these affecting cross-library comparisons is sequencing coverage. Two samples sequenced at 10 and 20 million reads but otherwise identical, for example, will be expected to demonstrate 2-fold differences in gene expression on average, even though there are no actual transcriptional differences. Normalizing read counts by coverage is done in a variety of ways, and in many cases differential expression software will implement a method without input from the user. One strategy, for instance, is to divide each gene's read count for a given library by the total number of mapped reads in that library (Oshlack, Robinson, Young 2010). If a handful of differentially expressed genes are extremely abundant, however, this procedure can result in erroneously calling many lowly expressed genes differentially expressed (Bullard et al. 2010). Other approaches normalize by total read counts only from genes expected to be evenly and/or moderately expressed (Robinson, Oshlack 2010). Perhaps the most popular (and robust) class of normalization procedures uses library read count quantiles (*e.g.* median, 75th percentile, *etc.*) or related values as scaling factors, as is the case in DESeq (Anders, Huber 2010).

If comparison of expression values among different transcripts is of interest, other normalization factors must be considered. Gene length, for example, influences read counts because more reads per single transcript will be observed for longer genes. The normalization procedure known as "RPKM" (reads per kilobase per one million mapped reads) applies a gene length and library size adjustment (Mortazavi et al. 2008).

$$RPKM = \frac{number\ of\ reads\ of\ the\ region}{\frac{total\ reads}{1{,}000{,}000} \quad x \quad \frac{region\ length}{1{,}000}}$$

RPKM facilitates transparent comparison of transcript levels within and between ssamples. Some examples of RPKM calculation regarding to region length and Million reads per library:

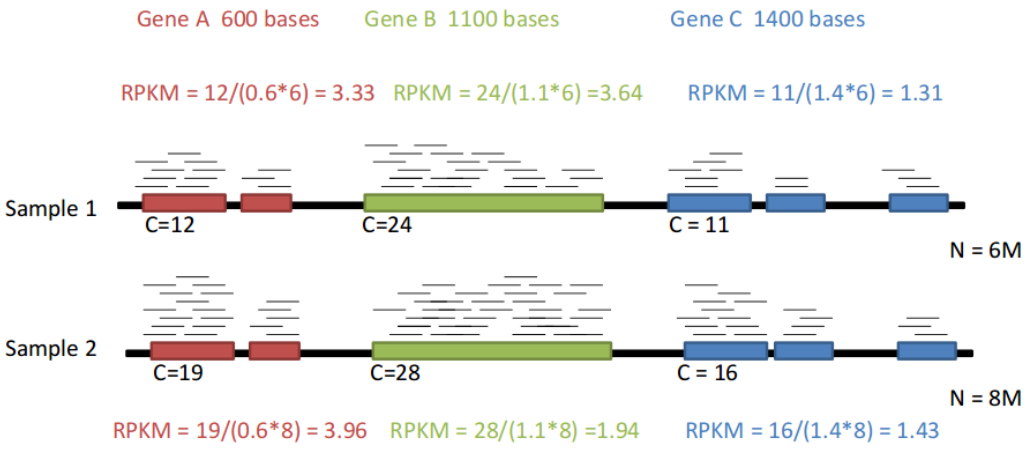


Figure 3. How change RPKM depending on gene count (C) and library size (N)

Although not formally treated in any differential expression software of which we are aware, base content has also been identified as an important among-gene normalization consideration (Oshlack, Robinson, Young 2010). For a detailed, more comprehensive guide to normalization strategies for RNA-seq data, see Dillies et al. 2012 and Bullard et al. 2010.

F. BAM files normalization and transformation in BigWig files.

Bam files are still large files and the inspection of these files using a genome browser like IGV could demand high amount of memory on local machine. To solve this problem, we used the Bamcoverage utility from Deeptools3 suite. This tool takes an alignment of reads or fragments as input (BAM file) and generates a coverage track (bigWig or bedGraph) as output. bigWig files are smaller compare to BAM files facilitating the simultaneous loading of multiple RNA-seq tracks in IGV simultaneously (Figure 3). In addition, Bamcoverage allows to normalize all the RNA-seq files (using different methods, i.e. Reads Per Kilobase per Million mapped reads (RPKM)) necessary to compare the enriched peaks from samples with different sequencing depths (i.e. different number of reads). The bigWig normalized files generated by Bamcoverage can be load

in IGV to inspect and analyze the aligned reads. IGV requires to load the genomes in a special format file (visit http://software.broadinstitute.org/software/igv/LoadGenome to get information about how to create and load genome files in IGV). The *Synechocystis* genome file necessaries for IGV in this tutorial (pcc6803.genome.fasta and pcc6803.genome.fasta.fai) are in the zenodo practice page.

```
:$ bamCoverage -b <aligned_sorted.bam> -o <coverage_file.bw> --normalizeUsing RPKM
```

Arguments:
-b aligned_sorted.bam: BAM file to process (sorted)

-o coverage_file.bw: ouput file in bigWig format.

-normalizeUsing: It is possible to normalize the number of reads per bin using four different methods; CPM= Counts Per Million mapper reads, BPM= Bin Per Million mapped reads, RPGC = reads per genomic content and RPKM = Reads Per Kilobase per Million mapped reads.

Generate a shell script.sh (deeptool.sh) containing all the command to normalize the BAM files programmatically.

G. RPKM calculation using raw counts

Here we are going to use two scripts, the first (normalize_and_corr.sh) is going to merge all .count files, and applicate the RPKM formula using AWK. This script, at the end, call to the R script (samples_corr.R) that uses the R library “CORRPLOT” to calculate the RPKM correlation between the 4 samples and generates as output two files: a table and a graphical representation of the RPKM correlation.

Note: both scripts have to be at the same path.

Differential Expression Analysis

Differential expression analysis means taking the normalized read count data and performing statistical analysis to discover quantitative changes in expression levels between experimental groups. For example, we use statistical testing to decide whether, for a given gene, an observed difference in read counts is significant, that is, whether it is greater than what would be expected just due to natural random variation.

1. Differential Expression Analysis: discrete distribution models

To estimate transcript abundance of a given target in a given sample, a parameter of interest is the probability that a randomly drawn read (from millions in the library) maps uniquely to that target. If thousands of reads in the library map to the particular target, this probability will be higher than if only a few map, but in general these probabilities will be very low owing to the huge size of the library and the complexity of the transcriptome. These sampling properties are characterized by the Poisson probability distribution, which in principle should enable us to draw inferences about, say, whether the probability that a randomly drawn read from a library of type A maps to transcript X is higher than the probability that a randomly drawn read from library type B maps to transcript X (*i.e.* differential expression of transcript X between groups A and B).

One issue with RNA-seq data, however, is that the variance of this probability among different individuals of a group is substantially higher than the mean, with respect to many genes (Anders, Huber 2010). A Poisson distribution assumes an equal mean and variance and is therefore not a good fit. This issue, known as "overdispersion," has inspired statistical software authors to adopt other models, particularly the negative binomial (NB) distribution, which is characterized by an additional dispersion parameter. Several popular differential expression packages, such as edgeR (Robinson, McCarthy, Smyth 2010) and DESeq (Anders, Huber 2010) are based on the NB distribution, but they differ extensively in how the dispersion parameter is estimated, how normalization is performed, or how the hypothesis test is carried out. For a nice tutorial on how DESeq works in these respects, and its actual usage, see: cgrlucb.wikispaces.com/Spring+2012+DESeq+Tutorial.

2. Differential Expression Analysis: Continuous distribution models

Rather than modeling expression estimates as counts, some statistical approaches treat normalized or transformed count values as continuously distributed variables. If these are distributed approximately normally or log-normally, then those distributions may in principle be used for differential expression inferences. Some authors, for example, have compared normalized count data between two groups via t-test (Busby et al. 2013). Because microarray data is continuously distributed, some have used software originally written for array data (Smyth 2004) to analyze normalized RNA-seq data (Soneson, Delorenzi 2013). One upshot of this approach is the ability to use a general linear model framework for more complex experimental designs. For data sets with larger sample sizes (*e.g.* 5-10), Soneson et al. 2013 found this method to perform especially well.

3. Differential Expression Analysis: Nonparametric models

Nonparametric statistical approaches are especially useful when real data don't conform to specific distributional assumptions. One such approach for RNA-seq data is to calculate a rank-based test statistic (e.g. Mann-Whitney) for differential gene expression between groups. The R software package SAMseq (Li, Tibshirani 2011) adopts this strategy, and uses resampling to get around the issue of unequal library sizes. Nonparametric tests for a variety of experimental design configurations have been derived, so implementing them in differential expression software like SAMseq is convenient. These approaches have proven to be as effective and robust as their parametric counterparts, especially with moderate to high sample sizes (Soneson, Delorenzi 2013)

4. Choice of analysis software

Clearly there are many options when it comes to differential gene expression analysis. Although some packages are more sensitive to some parameters (such as sample size and overdispersion), comparative methods reviews have not determined a "clear leader" in overall performance (Kvam, Liu, Si 2012; Soneson, Delorenzi 2013). It would therefore be inappropriate for us to recommend specific software to readers. Instead, we have compiled an annotated listing (Table 5.2 Diff. Exp. Software) to serve as a decision aid along with published (and future) studies that compare methodologies using real and simulated data. In our opinion, one sensible strategy is to analyze a dataset with several pieces of software that employ different types of models and highlight both the consensus and discrepancies among those analyses when publishing results. This way, a suite of tools is tested every time a group performs a differential expression experiment, and the community will benefit from this cumulative comparative information

Table 5.2 Differential Expression Software

This table summarizes many of the available software packages for differential gene expression analysis using RNA-seq data. All but CuffDiff 2 are implemented in the R statistical language. NB = Negative Binomial. GLM = Generalized Linear Model.

Package	Reference	Model	Experimental Design Flexibility
PoissonSeq	Li et al. 2012	Poisson log-linear	categorical explanatory variable (2+ categories), quantitative explanatory variable
iFad	Chung et al. 2013	Poisson & Gamma	categorical explanatory variable (2 categories, paired design)
TSPM	Auer, Doerge 2011	Quasi-Poisson	categorical explanatory variable (2+ categories), multiple explanatory variables (GLM)
baySeq	Hardcastle et al 2010	NB	categorical explanatory variable (2+ categories)
DESeq2	Love, Huber, Anders 2014	NB	categorical explanatory variable (2+ categories), multiple explanatory variables (GLM)
EBSeq	Leng et al. 2013	NB	categorical explanatory variable (2+ categories)
edgeR	Robinson, McCarthy, Smyth 2010	NB	categorical explanatory variable (2+ categories), multiple explanatory variables (GLM)
NBPSeq	Di et al. 2011	NB	categorical explanatory variable (2 categories), one-dimensional explanatory variable (GLM)
ShrinkSeq	Van De Wiel et al. 2012	NB (or zero-inflated NB)	categorical explanatory variable (2+ categories), multiple explanatory variables (GLM, allows random effects)
sSeq	Yu, Huber, Vitek 2013	NB	categorical explanatory variable (2 categories, paired design optional), multiple explanatory variables (factorial, via multiple paired comparisons)
BBSeq	Zhou, Xia, Wright 2011	NB & Beta	categorical explanatory variable (2+ categories), multiple explanatory variables (GLM)
CuffDiff 2	Trapnell et al. 2013	NB & Beta	categorical explanatory variable (2 categories)
QuasiSeq	Lund et al. 2012		categorical explanatory variable (2+ categories), multiple explanatory variables (GLM)
DEGseq	Wang et al. 2010	MA-plot (normal)	comparison of 2 individual samples (technical replicates optional)
LIMMA	Smyth 2004	general linear model	categorical explanatory variable (2+ categories), multiple explanatory variables (general linear model)
NOISeq	Tarazona et al. 2011	nonparametric	categorical explanatory variable (2 categories)
SAMseq	Li, Tibshirani 2011	nonparametric	categorical or quantitative explanatory variable (2+ categories), paired designs and survival analysis possible

H. RNA-seq differential expression analysis using DESeq2

In our practice we are going to use the Bioconductor R package DESeq2. Bioconductor provides tools for the analysis and comprehension of high-throughput genomic data. Bioconductor uses the R statistical programming language, and is open source and open development (have a look at: https://www.bioconductor.org).

Every package at Bioconductor usually have a pretty nice pdf tutorial named Vignettes and can it downloaded after the package installation directly in R using the command:

```
browseVignettes("DESeq2")
```

In our case you have available de DESeq2 Vignette in the RNA-seq analysis moodle section.

The DESeq.R script contains step-by-step differential expression analysis using DESeq2 and it going to generate several tables and PDF containing all the analysis. To use this script just type in terminal:

```
:$ Rscript DESeq2.R
```

Genome annotation and Gene set enrichment analysis.

Finally to annotate the genes and perform a functional category analysis. We will use the table generated with DESeq2 in the *Synechocystis* web data base Synergy (http://www.synergy.plantgenie.org) to annotate completely our induced and repressed differential expressed genes.

For more information about RNA-seq

A good paper (A survey of best practices for RNA-seq data analysis)
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

A blog/tutorial
https://rnaseq.uoregon.edu

A very complete RNA-seq tutorial in GitHub
https://github.com/griffithlab/rnaseq_tutorial/wiki

Bibliography.

Anders, S, W Huber. (2010). Differential expression analysis for sequence count data. *Genome Biol* 11:R106.

Auer, PL, RW Doerge. (2011) A two-stage poisson model for testing RNA-Seq data.

Blanca, JM, L Pascual, P Ziarsolo, F Nuez, J Canizares. (2011). ngs_backbone: a pipeline for read cleaning, mapping and SNP calling using Next Generation Sequence. *BMC Genomics* 12.

Brautigam, A, T Mullick, S Schliesky, AP Weber. (2011) Critical assessment of assembly strategies for non-model species mRNA-Seq data and application of next-generation sequencing to the comparison of C(3) and C(4) species.*J Exp Bot* 62:3093-3102.

Brown, CT, A Howe, Q Zhang, AB Pyrkosz, TH Brom (2012) A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data *arXiv:*1203.4802v2 [q-bio.GN]

Bullard, JH, E Purdom, KD Hansen, S Dudoit. (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94

Busby, MA, C Stewart, C Miller, K Grzeda, G Marth (2013). Scotty: A Web Tool For Designing RNA-Seq Experiments to Measure Differential Gene Expression. *Bioinformatics*

Camacho, C, G Coulouris, V Avagyan, N Ma, J Papadopoulos, K Bealer, TL Madden. (2009). BLAST+: architecture and applications. *BMC Bioinformatics* 10:421

Catchen, JM, A Amores, P Hohenlohe, W Cresko, JH Postlethwait (2011). Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences. *G3: Genes, Genomes, Genetics* 1:171-182

Chevreux, B, T Pfisterer, B Drescher, AJ Driesel, WE Muller, T Wetter, S Suhai (2004). Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res* 14:1147-1159

Chung, LM, JP Ferguson, W Zheng, F Qian, V Bruno, RR Montgomery, H Zhao (2013). Differential expression analysis for paired RNA-seq data. *BMC Bioinformatics* 14:110

Cock, PJ, CJ Fields, N Goto, ML Heuer, PM Rice(2010). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Nucleic Acids Res* 38:1767-1771

Compeau, PE, PA Pevzner, G Tesler (2011). How to apply de Bruijn graphs to genome assembly. *Nat Biotechnol*29:987-991

Conesa, A, S Gotz, JM Garcia-Gomez, J Terol, M Talon, M Robles (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674-3676

Di, Y, DW Schafer, JS Cumbie, JH Chang (2011). The NBP negative binomial model for assessing differential gene expression from RNA-Seq. *Stat Appl Genet Mol Biol* 10:24

Dillies, MA, A Rau, J Aubert, et al (2012). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Brief Bioinform*

Fonseca, NA, J Rung, A Brazma, JC Marioni (2012). Tools for mapping high-throughput sequencing data.*Bioinformatics* 28:3169-3177

Francis, WR, LM Christianson, R Kiko, ML Powers, NC Shaner, SH Haddock (2013). A comparison across non-model animals suggests an optimal sequencing depth for de novo transcriptome assembly. *BMC Genomics* 14:167

Gillis, J, M Mistry, P Pavlidis (2010). Gene function analysis in complex data sets using ErmineJ. *Nat. Protocols*5:1148-1159

Grabherr, MG, BJ Haas, M Yassour, et al (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol*29:644-652

Hardcastle, TJ, KA Kelly (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics* 11:422

Huang da, W, BT Sherman, RA Lempicki (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc*4:44-57

Huang, X, A Madan (1999). CAP3: A DNA sequence assembly program. *Genome Res* 9:868-877

Kvam, VM, P Liu, Y Si (2012). A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. *Am J Bot*99:248-256

Leng, N, JA Dawson, JA Thomson, V Ruotti, AI Rissman, BMG Smits, JD Haag, MN Gould, RM Stewart, C Kendziorski (2013). EBSeq: An empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*

Li, H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, S Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078-2079

Li, J, R Tibshirani (2011). Finding consistent patterns: A nonparametric approach for identifying differential expression in RNA-Seq data. *Stat Methods Med Res*

Li, J, DM Witten, IM Johnstone, R Tibshirani (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics* 13:523-538

Li, Z, Y Chen, D Mu, et al (2011). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and de-bruijn-graph. *Brief Funct Genomics*

Lohse, M, AM Bolger, A Nagel, AR Fernie, JE Lunn, M Stitt, B Usadel (2012). RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. *Nucleic Acids Res* 40:W622-W627

Love, MI, Huber, W, S Anders (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* 15:550

Lund, SP, D Nettleton, DJ McCarthy, GK Smyth (2012). Detecting Differential Expression in RNA-sequence Data Using Quasi-likelihood with Shrunken Dispersion Estimates. *Stat Appl Genet Mol Biol* 11

Martin, JA, Z Wang (2011). Next-generation transcriptome assembly. *Nat Rev Genet 12:671-682*

Miller, JR, S Koren, G Sutton (2010). Assembly algorithms for next-generation sequencing data. *Genomics* 95:315-327

Mortazavi, A, BA Williams, K Mccue, L Schaeffer, B Wold (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* 5:621-628

Oshlack, A, MD Robinson, MD Young (2010). From RNA-seq reads to differential expression results. *Genome Biol*11:220

Pevzner, PA, H Tang, MS Waterman (2001). An Eulerian path approach to DNA fragment assembly.*PNAS* 98:9748-9753

Robertson, G, J Schein, R Chiu, et al (2010). De novo assembly and analysis of RNA-seq data. *Nat Meth* 7:909-912

Robinson, MD, DJ McCarthy, GK Smyth (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.*Bioinformatics* 26:139-140

Robinson, MD, A Oshlack (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol* 11:R25

Schulz, MH, DR Zerbino, M Vingron, E Birney(2012). Oases: Robust de novo RNA-seq assembly across the dynamic range of expression levels.*Bioinformatics*

Simpson, JT, K Wong, SD Jackman, JE Schein, SJ Jones, I Birol (2009). ABySS: a parallel assembler for short read sequence data. *Genome Res* 19:1117-1123

Smyth, GK (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3:Article3

Soneson, C, M Delorenzi (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics* 14:91

Tarazona, S, F Garcia-Alcalde, J Dopazo, A Ferrer, A Conesa (2011). Differential expression in RNA-seq: a matter of depth. *Genome Res* 21:2213-2223

Trapnell, C, DG Hendrickson, M Sauvageau, L Goff, JL Rinn, L Pachter (2013). Differential analysis of gene regulation at transcript resolution with RNA-seq.*Nat Biotech* 31:46-53

Van De Wiel, MA, GGR Leday, L Pardo, H Rue, AW Van Der Vaart, WN Van Wieringen (2012). Bayesian analysis of RNA sequencing data by estimating multiple shrinkage priors. *Biostatistics*

Wang, LK, ZX Feng, X Wang, XW Wang, XG Zhang(2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data.*Bioinformatics* 26:136-138

Yu, D, W Huber, O Vitek (2013). Shrinkage estimation of dispersion in Negative Binomial models for RNA-seq experiments with small sample size. *Bioinformatics*

Zerbino, DR, E Birney (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs.*Genome Res* 18:821-829

Zhou, Y-H, K Xia, FA Wright (2011). A powerful and flexible approach to the analysis of RNA sequence count data.*Bioinformatics* 27:2672-2678.