UNIVERSITÁ DEGLI STUDI DI SALERNO

DOTTORATO IN MANAGEMENT & INFORMATION TECHNOLOGY

CURRICULUM: INFORMATION SECURITY & INNOVATION SYSTEMS

COORDINATORE: Ch.mo. Prof. Antonelli Valerio

Ciclo XVII N.S.

Novel tools for reproducible Next Generation Sequencing data analysis and integration

Relatori
Ch.mo. Prof. Tagliaferri Roberto
Ch.mo. Prof. Angelini Claudia

Candidato
Righelli Dario
Matr. 8800800010

ANNO ACCADEMICO 2017/2018

How to reach a goal?
Without haste but without rest
Goethe

Add acknowledgements here

Write your abstract here

Contents

Chapter 1

Introduction

1.1 Biological Background

1.2 Sequencing Techniques

1.2.1 RNA-Seq

1.2.2 Atac-Seq

1.3 Computational Aspects

Chapter 2

TiCoRSe - Time Course RNA-Seq data analysis

2.1 Introduction

2.1.1 Time Course RNA-Seq

2.2 Methods

2.2.1 General Approach

2.2.2 Time Course Methods

2.2.3 Other Methods

2.2.4 Additional Features

2.3 Results

Chapter 3

DEScan2 - Differential Enriched Scan 2

few words on integration of epigenomic with transcriptomic

To investigate and to answer a epigenetic biological questions we decided to create an instrument useful for analysing the epigenomic data. Very often the biological questions, as for the RNA-Seq, to be answered, need the comparison of two or more different biological conditions. On the basis of a set of already published [1] scripts, we designed *Differential Enriched Scan 2* (*DEScan2*), a software for helping the analysis of epigenomic data.

3.1 Introduction

The *Differential Enriched Scan 2* is an R [2] tool developed for detecting epigenomic signal in order to facilitate the Differential Enrichment of the signal between two or more biological conditions.

The package uses Bioconductor [3] data structures and methods every time it was possible, and it is available on Bioconductor since the version 3.7.

It's organized in three main steps. A peak caller, which is a standard moving window scan that compares the counts within a sliding window to the counts in a larger region outside the window, using a simple Poisson likelihood (no overdispersion estimation) and providing a final score for each detected peak.

The filtering step is aimed to determine if a peak is a "true peak" on the basis of its replicability in other samples. To do so, this step is based on a double user-defined threshold, one on the peak score and one on the number of samples.

Finally, the third step produces a counts matrix where each column is a sample and each row a filtered peak computed in the filtering step. The value of the matrix cell is the number of reads for the peak in the sample.

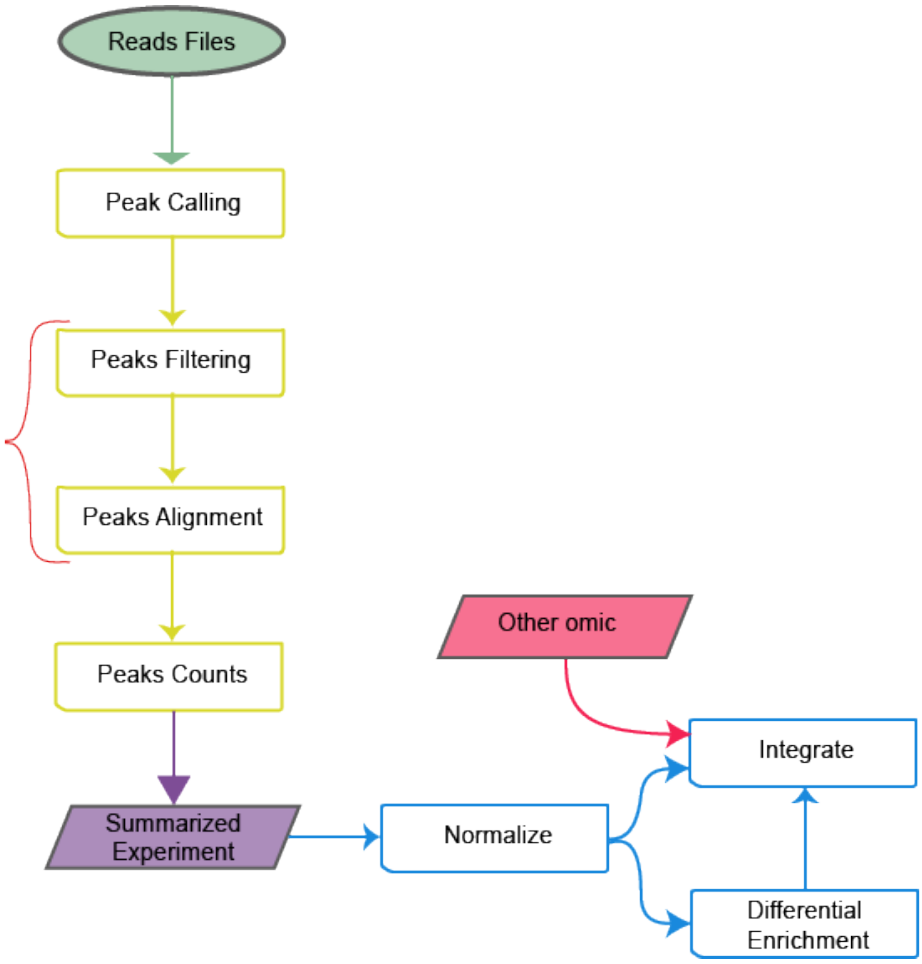

Figure 3.1.1: A differential enrichement flow representation. DEScan2 steps are highlighed in yellow.

The so produced counts matrix, as illustrated in the figure 3.1.1, is useful both for doing differential enrichment between the conditions and for integrating the epigenomic data with other -omic data types.

3.2 Methods

3.2.1 Peak Caller

The Peak Caller takes as input a set of alignment files in BAM [4] or BED format together with several additional parameters, useful for the peak detection setup.

The alignement data are stored as GenomicRangesList [5], where each element represents a file. In order to facilitate the parallelization of the computations over the chromosomes, the list is re-arranged as a list of GenomicRangesList, with a chromosome for each element. Moreover, each element of the GenomicRangesList represents a file containing just the GenomicRanges of the specific chromosome.

On this data structure the algorithm firstly divides each chromosome as bins of user-defined length and then computes the coverage of the reads on the bins with a moving scan window.

In order to be able to catch also spread peaks we compute the coverage also using windows of two different lengths.

Once the coverages are ready the method computes a score for each detected region, applying a poisson likelihood estimation.

[PUT THE POISSON DISTRIBUTION AND THE LIKELIHOOD EXPLAINING THE METHOD]

3.2.2 Peak Filtering and Alignment

In order to detect the most possible "true peaks", we deisgned a filtering step based on two different threshold. A first threshold on the peaks score and a second threshold on the number of samples.

The filtering step is designed to take as input a list of peaks as GenomicRangesList, where each element represents a chromosome. This is the data structure produced by the peak caller, but, we developed a method to load peaks produced also by other software like MACS [6], as described in section 3.2.4.

The filtering step on the peaks score just filters out the peaks with a score

lower than the user-defined threshold value.

While the filtering step on the samples, firstly extends a window in both directions of the detected region, then computes the overlaps between the samples using the *findOverlapsOfPeaks* method (with *connectedPeaks* parameter as *merge*), defined in the ChIPpeakAnno [7] R package.

Basing on this idea, the filtering step is developed to filter out those peaks not present in at least a user-defined number of samples. In the light of this, the user can decide the minimum number of samples where each peak has to be detected. A further threshold can be used over the peak score.

3.2.3 Peak Counts

The counting step is designed to take a *GenomicRanges* data structure as input, where additional informations for the score and the number of samples for each peak are stored. Moreover, it requires also the path of the BAM/BED files within the reads to count for each peak in the *GenomicRanges*.

This step counts the number of reads present for each region in each sample. Indeed, it produces as result a matrix of counts where on the reads there are the regions and on the columns the samples.

In order to keep trace of all the information associated to the regions, it produces a *SummarizedExperiment* [8] data structure, which gives the possibility to access the *GenomicRanges* data structure associated to the peaks with the *rowRanges* method and to access to the count matrix with the *assays* method.

The choice to produce a count matrix is dictated by the versatility of this data structure, useful not only for the differential enrichment of the regions between multiple conditions, but also for integrationg this -omic type with other -omics.

3.2.4 Additional Features

However, the package can work with any external peak caller returning results in terms of bed files, indeed the package provides additional functionalities to load BED files of peaks and handle them as GenomicRanges [5] structures.

Furthermore, our package provides several functionalities for GenomicRanges data structure handling. One over the others gives the possibility to split a GenomicRanges over the chromosomes to speed-up the computations parallelizing them over the chromosomes.

3.3 Case Study

A few words on epigenomic data

We illustrate the performances of DEScan2 using a dataset [9] describing adult mouse dentate granule neurons in vivo before and after synchronous neuronal activation using Atac-Seq and RNA-Seq technologies (see sections 1.2.2 and 1.2.1).

This dataset is organized in 62 samples of Atac-Seq and RNA-Seq, sampling them at different time points, with four replicates for time point. Of this samples we chose to compare the differences at the first two stages, time 0 (E0) and After 1 hour of neuronal induction (E1), in order to show a possible Atac-Sec workflow for Differential Enrichment, and how to integrate this type of data with RNA-Seq data.

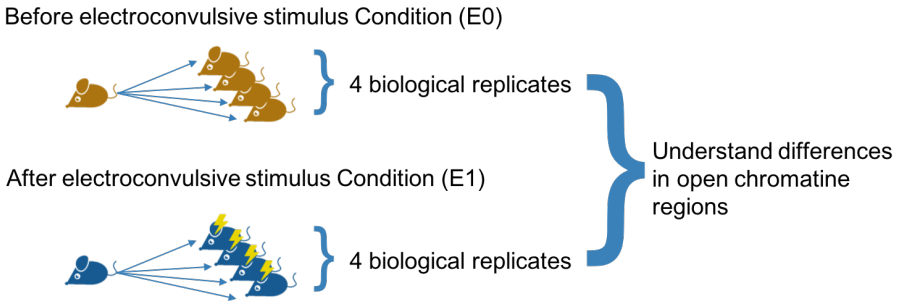

Figure 3.3.1: An illustration of our extraction of the [9] dataset.

We downloaded the data from the *GEO* database [10, 11] with accession number GSE82015[1] and mapped raw data using *STAR* [12] with default parameter on Mus Musculus Genome ver.10 (mm10).

In order to detect the open chromatin regions we run our peak caller, cutting the genome in bins of 50bp and using running windows of minimum 50bp and maximum 1000bp. In such a way we are able to detect not just broad peak, but also smaller peaks.

To be confident with our results we compared the DEScan2 detected peaks on two genes (Arc and Gabrr1) with the same genes validated in [9]. Lower part of figure 3.3.2 shows the detected and validated regions (in blue and red) resulting differentially enriched between the E0 (in pink) and E1 (in green) samples, while the upper part shows DEScan2 peaks (in blue), which is able to catch the same regions of the published ones, but also (gold circles) to be more careful in the detection of smaller peaks.

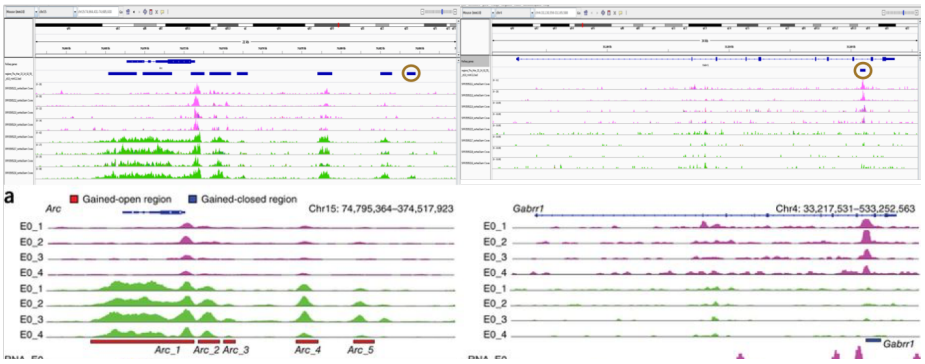

Figure 3.3.2: A comparison of DEScan2 detected peaks with validated peaks in article [9].

While it is very important to detect good peaks with a peak caller, it seems to be more relevant to detect reliable regions. Indeed, during the filtering step, the number of peaks depends not only by the peak score, but also by the number of replicates designed in the experiment. The figure 3.3.3 puts in relation these

[1] `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE82015`

two relevant information. On the x-axis is represented the number of replicates, while on the y-axis the number of peaks is traced, and each line represents a different threshold for the score of the peaks, showing that higher is the threshold on the scores and the number of the replicates, lower is the number of the detected peaks. Highlighting a proportional inversion between the number of the peaks and the number of the samples combined with the score of the detected regions.

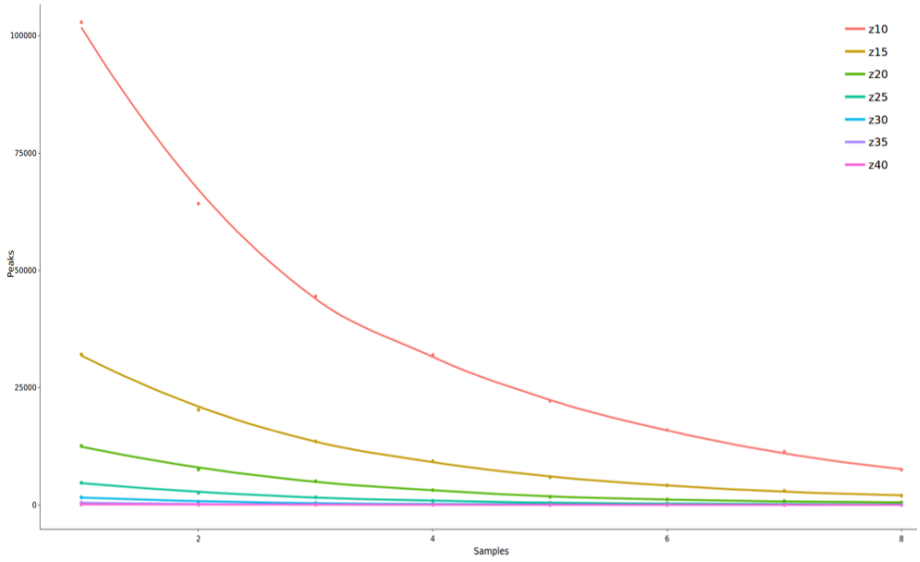

Figure 3.3.3: Filtering the detected regions with different thresholds on peak scores.

The filtered-in regions can be processed by DEScan2 in order to obtain a count matrix with samples on the columns and peaks on the rows. This type of data structure is very versatile, because it enables to perform several operations, like the differential enrichment of regions (DER) and, when possible, the integration with other kind of omics, as RNA-Seq.

In order to preserve the information associated to the peaks, DEScan2 produces as output a *SummarizedExperiment* 3.3.4 data structure, which enables to retrieve the count matrix with *assays* method, and to access the peaks in-

formation in *GenomicRanges* format with the *rowRanges* method.

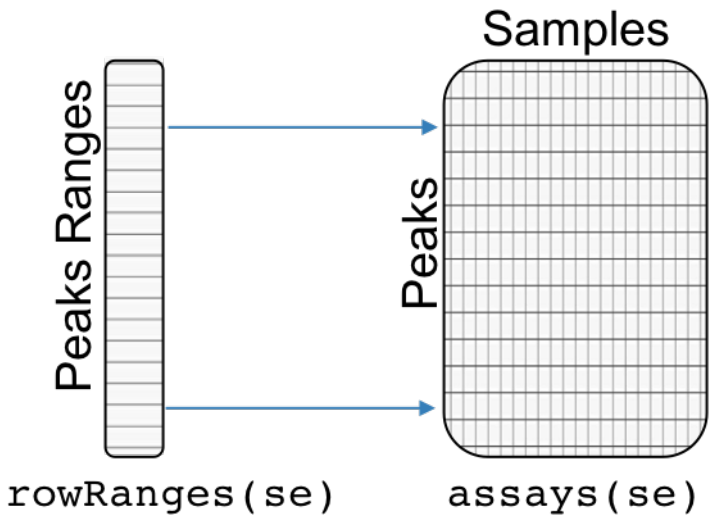

Figure 3.3.4: An illustration of the *SummarizedExperiment* data structure produced by DEScan2.

Before to proceed to detect DERs, it is a good standard to normalize the data, also because without any kind of normalization we are not able to detect any DERs. The nature of the data, in count format, makes it possible to apply several well known RNA-Seq normalizations techniques, as *TMM*, *upper-quartile*, *full-quantile*, *RUV-Seq*, etc [13–15].

While the *TMM* and *upper-quartile* normalizations modify the data in a way that makes it impossible to detect DERs, other kind of normalizations and combinantions of them give good results.

The figure 3.3.5 sintetizes this concept very well, highlighting a relation between the number of DERs and the minumum number of samples used for filtering the data during the DEScan2 filtering step.

The plot shows that *upper-quantile*, even if combined with *RUV-Seq* normalization, is not able to linearly detect a good amount of DERs, while *full-quantile*,

when combined with *RUV-Seq* seems to affect the data in a way that overdetect the number of DERs. When looking at the *full-quantile* and *RUV-Seq* by themself seem to perform better than the other normalizations. The first one has a downhill almost linear, while the second one has a very fast downhill with a regrowth when the number of samples is higher.

Even if these normalization methods show good performances with this type of epignomic data, our investigations suggest that for sure more testing is required, but maybe an ad-hoc normalization method for these data has to be developed.

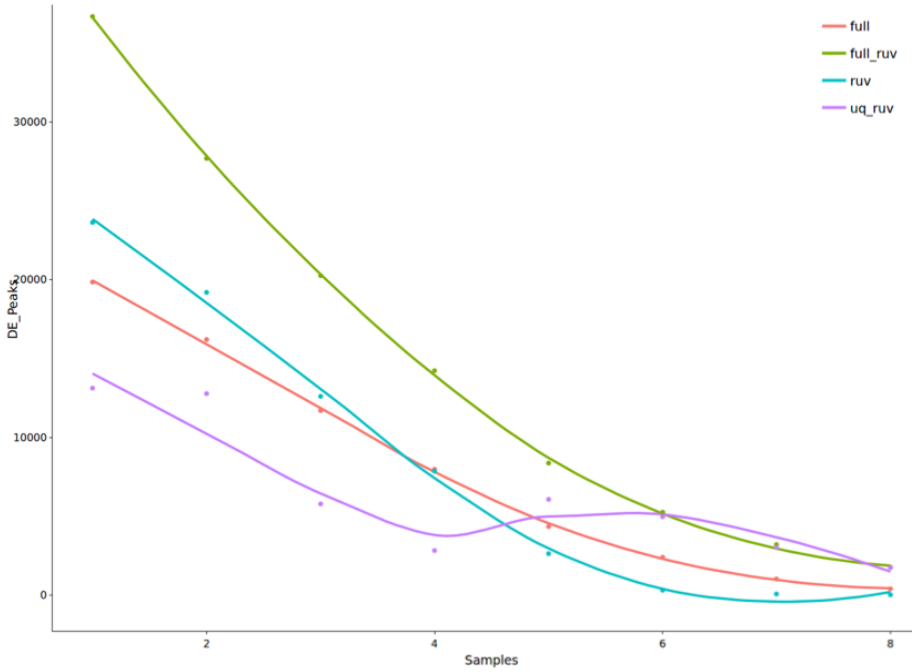

Figure 3.3.5: The figure shows the effects of different normalizations on the epigenomic differentially enriched regions.

To estimate the DERs any of the RNA-Seq methods can be applied, as *DESeq2*, *edgeR*, *NOISeq*, etc [16–18].

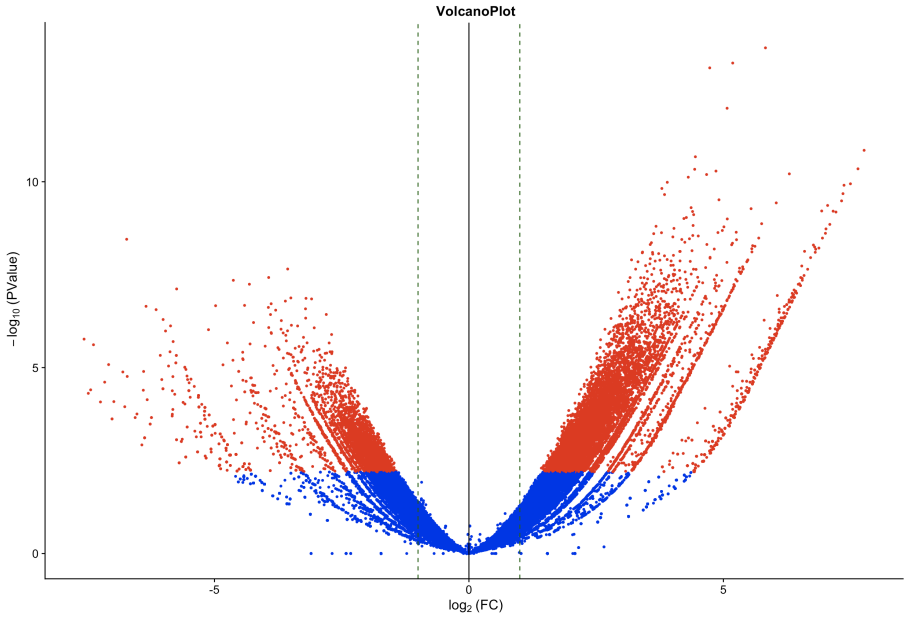

Figure 3.3.6: A volcano plot of Differential Enriched Regions. Blue dots represent the not significant DERs, while the red ones represent the significant DERs.

Chapter 4

IntegrHO - Integration of High-Throughtput Omics data

4.1 Introduction

4.2 Methods

4.2.1 Single Omic Approach

4.2.2 Multi Omic Approach

Low Level Itegration

High Level Itegration

4.3 Implementation Aspects

4.4 Reproducible Computational Research

4.5 Results

Chapter 5

Conclusions & Future Works

Appendices

.1 R Language

.2 R Markdown Language

Chapter 6

Bibliography

1. Koberstein, J. N. *et al.* Learning-dependent chromatin remodeling highlights noncoding regulatory regions linked to autism. *Science Signaling.* ISSN: 19379145. doi:10.1126/scisignal.aan6500 (2018).
2. Ihaka, R. & Gentleman, R. R: A Language for Data Analysis and Graphics. *Journal of Computational and Graphical Statistics.* ISSN: 15372715. doi:10.1080/10618600.1996.10474713. arXiv: arXiv:1011.1669v3 (1996).
3. Gentleman, R. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology.* ISSN: 1465-6914. doi:10.1186/gb-2004-5-10-r80 (2004).
4. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* ISSN: 13674803. doi:10.1093/bioinformatics/btp352. arXiv: 1006.1266v2 (2009).
5. Lawrence, M. *et al.* Software for Computing and Annotating Genomic Ranges. *PLoS Computational Biology* **9** (ed Prlic, A.) e1003118. ISSN: 1553-7358 (2013).
6. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biology.* ISSN: 14747596. doi:10.1186/gb-2008-9-9-r137 (2008).

7. Zhu, L. J. *et al.* ChIPpeakAnno: A Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics.* ISSN: 14712105. doi:10.1186/1471-2105-11-237 (2010).

8. Morgan M, Obenchain V, Hester J, P. H. SummarizedExperiment: SummarizedExperiment container. doi:https://doi.org/doi:10.18129/B9.bioc.SummarizedExperiment (2018).

9. Su, Y. *et al.* Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nature Neuroscience.* ISSN: 15461726. doi:10.1038/nn.4494 (2017).

10. Edgar, R. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research.* ISSN: 13624962. doi:10.1093/nar/30.1.207 (2002).

11. Barrett, T. *et al.* NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research.* ISSN: 03051048. doi:10.1093/nar/gks1193 (2013).

12. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics.* ISSN: 13674803. doi:10.1093/bioinformatics/bts635. arXiv: 1201.0052 (2013).

13. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology.* ISSN: 15461696. doi:10.1038/nbt.2931. arXiv: arXiv:1011.1669v3 (2014).

14. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology.* ISSN: 14747596. doi:10.1186/gb-2010-11-3-r25. arXiv: PMC2864565 (2010).

15. Dillies, M. A. *et al.* A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. *Briefings in Bioinformatics.* ISSN: 14675463. doi:10.1093/bib/bbs046 (2013).

16. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)* **26,** 139–140. ISSN: <null> (2009).

17. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40,** 4288–4297. ISSN: 03051048 (2012).

18. Tarazona, S., García, F., Ferrer, A., Dopazo, J. & Conesa, A. NOIseq: a RNA-seq differential expression method robust for sequencing depth biases. *EMBnet.journal* **17,** 18. ISSN: 2226-6089 (2012).

List of Figures

List of Tables