Gene regulation

Master Thesis in Biostatistics (STA495)

by

Joël Meili
14-679-393

supervised by

Prof. Dr. Mark D. Robinson
Dr. Simone Tiberi

Zurich, October 2022

Gene regulation

Joël Meili

Version September 28, 2022

Contents

Preface

Joël Meili
October 2022

Abstract

Chapter 1

Introduction

1.1 RNA sequencing

RNA sequencing (RNA-seq) is a technology for detecting and quantifying the mRNA molecules of a biological sample (Stark *et al.*, 2019). The invention of RNA-seq was a major breakthrough in the field of bioinformatics that replaced the use of microarray technology in the late 00's. In comparison to microarrays, RNA-seq allows for full sequencing of the whole transcriptome whereas microarrays only profile predefined transcripts through hybridization (Rao *et al.*, 2019). Further, various protocols have since been derived from the standard RNA-seq protocol, e.g. single-cell RNA sequencing (Stark *et al.*, 2019).

1.1.1 Bulk RNA-seq

Bulk RNA sequencing allows detecting an aggregated signal across a mixture of cells. There are many applications for bulk RNA-seq. For example, it can be used to study the differences of expression profiles between tissues in healthy vs disease or across treatments (Stark *et al.*, 2019). However, with bulk RNA-seq one can only estimate the average expression of each gene across a population of cells without regard for the differences between cell types. RNA-seq has several use cases. It can be used to study which genes are turned on in a cell and what their level of transcription is. This allows researchers to understand the biology of a cell at a deeper level. Further, RNA-seq allows the identification of variants and allele specific expression. It is also possible to study the patterns of alternative splicing, which are important to understand their contribution to cell differentiation and human disease.

1.1.2 Single-cell RNA-seq

Single-cell RNA sequencing was developed to overcome some of the limitations of bulk RNA sequencing. With scRNA-seq it is possible to estimate the distribution of expression levels for each gene across a population of cells. This allows answering new biological questions where cell-specific characteristics are important. However, there are some caveats with scRNA-seq (Haque *et al.*, 2017). scRNA-seq data in general is much more variable than bulk RNA-seq data due to both higher biological and technical variability at single-cell level (Haque *et al.*, 2017). Figure 1.1 shows the typical workflow of a scRNA-seq experiment. Said workflow is broadly summarized by the following steps:

1. RNA extraction
2. Reverse transcription into cDNA
3. Adapted ligation

4. Amplification

5. Sequencing

6. Downstream analysis using bioinformatics tools

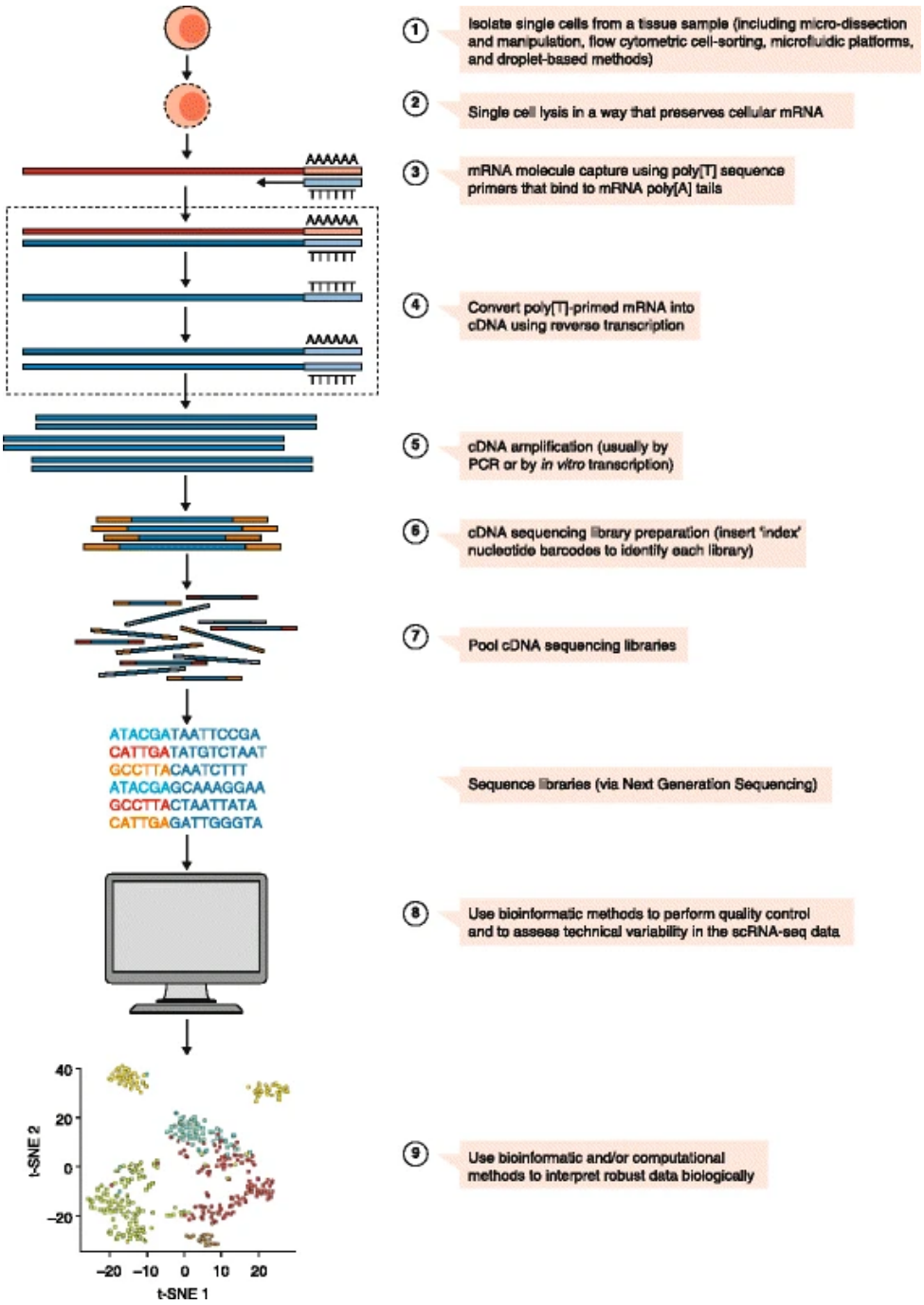

Figure 1.1: General workflow of single-cell RNA-sequencing experiments (Haque *et al.*, 2017)

1.1.3 Quantification of single-cell RNA-seq data

scRNA-seq data has distinct characteristics that prevent it from being processed by widely used tools developed for bulk RNA-seq data (He *et al.*, 2022). In general, quantification works by aligning the reads generated from the RNA-seq to the reference genome. There are several tools that allow to do that, notably: STAR (Dobin *et al.*, 2013), kallisto | bustools (Melsted *et al.*, 2021) and alevin (Srivastava *et al.*, 2019). However, there is a difference between the first tool and the other two. STAR is an aligner, whereas the other two tools are mapping tools (pseudo-aligners). The difference between an full-aligner and a mapping tool is that the latter does not look for the exact location of the read, as a consequence pseudo-alignment is much faster than full-alignment. Here we focus on alevin-fry, and the method we have developed, which will be introduced later, has been built to work on the output of alevin-fry. Alevin was developed to tackle the computational challenges that come with scRNA-seq data and to provide a tool that supports technologies other than 10x Genomics. Alevin works in two steps. First, it parses a read file which contains the cellular barcode and a unique molecule identifier to generate a frequency distribution of observed barcodes. Second, it maps the reads to the transcriptome and generates a cell-by-gene count matrix. Alevin-fry (He *et al.*, 2022) was designed to be the successor to alevin and achieves similar accuracy at significantly lower computational costs. It generates a permit list for cellular barcodes that will be quantified in subsequent steps. By using a multi-thread approach, alevin-fry filters and collates the mapping records for permitted cellular barcodes to produce a representation optimized for quantification (He *et al.*, 2022).

1.2 Objective

1.2.1 RNA velocity

We investigate spliced and unspliced reads from single-cell RNA-sequencing data. During transcription, DNA is decoded into precursor messenger RNA (pre-mRNA). Pre-mRNA contains both coding (exons) and non-coding regions (introns). In a next step, introns are removed from the pre-mRNA which leaves only the mature mRNA. Figure 1.2 shows the process from DNA to mature mRNA, where α is the transcription rate, β is the splicing rate and γ is the degradation rate.

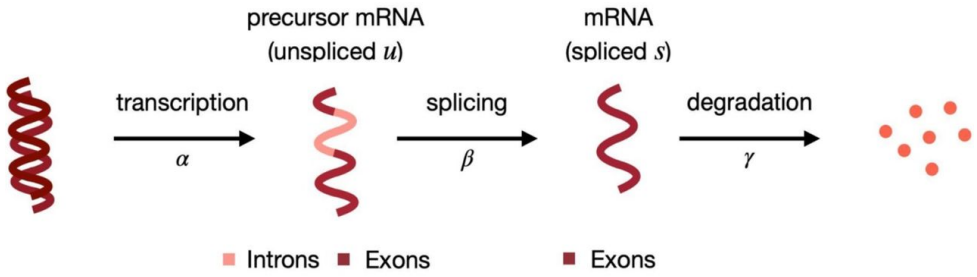

Figure 1.2: The transcription process from DNA to mature mRNA (Weiler *et al.*, 2021)

It was assumed that there is a signal (RNA velocity) detectable in scRNA-seq data that could reveal the rate and direction of change of an entire transcriptome (La Manno *et al.*, 2018). To quantify the relationship between the abundance of pre-mRNA and mature RNA, a simple system of ordinary differential equations was assumed (1.1): The solution of said system at equilibrium can then easily be estimated and used to explore the regulation of genes:

$$\begin{aligned} \frac{du}{dt} &= \alpha - \beta u \\ \frac{ds}{dt} &= \beta u - \gamma s \end{aligned} \tag{1.1}$$

The derivative of the spliced counts is then defined as the RNA velocity of cells. Thus, the balance of spliced and unspliced counts allows estimating whether a gene is up- or downregulated. If a larger fraction of unspliced counts are present than expected at equilibrium, a gene is likely upregulated. This is because within a short time interval, the newly spliced mRNA will exceed the amount of spliced mRNA which is degraded. Contrarily, if more spliced counts are present at equilibrium than expected, a gene is likely downregulated.

1.2.2 Differential regulation

The abundance of spliced and unspliced reads is directly linked to the regulation of genes and RNA velocities. Our idea is to examine how the abundance of spliced and unspliced counts changes between experimental conditions and biological replicates. We translate this intuition into the comparison of two experimental conditions, e.g. healthy vs. disease. Following the same intuitive rationale of RNA velocity, if a gene has a higher abundance of unspliced (spliced) counts in group A compared to group B, then this gene is likely being up-regulated (down-regulated) in group A compared to group B. Thus, we explore the differences in abundance of spliced and unspliced counts to study the differences in regulation between experimental conditions.

If the data contains multiple cell clusters (e.g. cell types), similarly to differential state analyses (Crowell *et al.* (2020) and Tiberi *et al.* (2021)) we will perform differential analyses in each cluster of cells, hence identifying cell-cluster/cell-type specific changes between conditions. The idea of performing differential analyses on the abundance of spliced and unspliced or exonic and intronic reads is not completely novel as there are at least two other methods that achieve that: eisaR and BRIE2.

1.2.3 Existing methods

eisaR (Gaidatzis *et al.*, 2015) is a R package implementation that allows for the split analysis between exons and introns. It allows one to measure changes in mature RNA and pre-mRNA across different experimental conditions. Ultimately, eisaR differential testing is based on edgeR (Robinson *et al.*, 2010). edgeR is a R package that performs differential expression analyses between groups of samples. It implements statistical methods that are based on the negative binomial distribution as a model for count variability.

BRIE2 (Huang and Sanguinetti, 2021) is a Bayesian hierarchical model that is implemented in Python and supports the analysis of splicing processes between spliced and unspliced RNA. There are two modes in which the tool can be used. First, the use of differential alternative splicing (DAS), where the aim is to quantify the proportions of alternative splicing isoforms. Second, the use of differential momentum genes (DMG), where the objective is to quantify the proportions of spliced and unspliced RNA in each gene and each cell. In this thesis we focus on the DMG mode as it performs differential testing on the relative abundance of spliced and unspliced reads.

Originally, eisaR and BRIE2 were developed to analyze all cells, but can easily be adapted to perform cell-type specific differential analyses.

1.2.4 Mapping uncertainty

We can identifiy two main sources of mapping uncertainty concering spliced and unspliced reads: i) multi-mapping reads across spliced and unspliced versions of a gene, and ii) reads compatible with multiple genes. In fact, it has been shown that many reads (5-40%) map to multiple genes

(Dharshini *et al.* (2020), McDermaid *et al.* (2018)). In our real data analyses (see Section 3), we found approximately 20-30% of such multi-mapping reads across genes. We additionally found that a significant fraction of reads (6-19%) are compatible with both S and U versions of a gene. Therefore, the estimated spliced and unspliced counts carry a substantial amount of uncertainty, which should be accounted for in downstream analyses. However, both eisaR and BRIE2 use estimated spliced and unspliced counts and neglect the mapping uncertainty. In this thesis, we propose two approaches that account for said mapping uncertainties.

Chapter 2

Methods

2.1 DEXSeq

DEXSeq (Anders *et al.*, 2012) is a statistical method originally proposed to test for differential exon usage in RNA- seq data, which has been widely adopted in other contexts too, such as differential transcript usage (Love *et al.*, 2018). The model is based on the negative binomial distribution and allows for covariates such as batch effects to be taken into account to offer reliable control of false discoveries (Anders *et al.*, 2012). In its original implementation DEXSeq inputs how many reads map to each exon, but the method has alsob been used on transcript level counts. Equation (2.1) shows that the read counts follow a negative binomial distribution where α is the dispersion parameter. Further, a generalized linear model is used to predict the mean via a log-linear link:

$$K_{ijl} \sim \text{NB}(\text{mean} = s_j \mu_{ijl}, \text{dispersion} = \alpha_{il}) \tag{2.1}$$

$$log(\mu_{ijl}) = \beta_i^G + \beta_{il}^E + \beta_{i\rho_j}^C + \beta_{i\rho_j l}^{EC} \tag{2.2}$$

where NB(a, b) denotes the negative binomial distribution with mean a and dispersion b, s_j is ..

The dispersion parameter allows to model over-dispersed data (i.e. higher variance than mean). Here, we propose to use DEXSeq on estimated USA counts, and perform a differential usage test between conditions. This models ambiguous reads separately from spliced and unspliced or exonic and intronic, thus eliminating one of the main sources of mapping uncertainty. However, the uncertainty related to reads mapping to multiple genes is still neglected by this approach.

2.2 Differential regulation

To address both sources of mapping uncertainty we propose our novel method. Similar to the idea above, we implemented a hierarchical Bayesian approach that models ambiguous counts separately from spliced and unspliced. Gene allocation is modeled as a latent state to address the gene-related mapping uncertainty. The model consists of two nested models: First, we use a Dirichlet-multinomial model for the relative abundance of the USA counts in each gene. Second, a multinomial model that models the relative abundance of genes for each sample individually.

Chapter 3

Results

3.1 Exploratory Data Analysis

3.1.1 Mouse kidney cells

The first data set stems from a paper that investigates potential cellular targets of kidney disease in mice (Park *et al.*, 2018). The authors isolated and sequenced a total of 57'979 cells from whole kidney cell suspensions (one kidney per mouse) derived from seven healthy male mice using droplet-based single-cell RNA sequencing. The samples were labelled as: normal1, normal2, normal3, normal4, Ksp-cre-GFP, Scl-cre-GFP and Pod-cre-GFP. For our work, we decided to use the raw data from the four samples that were labelled as normal to ensure the biological reproducibility between the samples.

We used the *alevin-fry* pipeline to quantify the raw single-cell RNA sequencing data for further use in the R programming environment. Quality control is a crucial stage in data pre-processing as low-quality libraries can contribute to misleading results in downstream analyses (Amezquita *et al.*, 2020). Therefore, we filtered lowly abundant genes and low-quality cells to mitigate said problems to improve interpretability of the results.

To identify low-quality cells, cell-specific QC metrics were calculated with the *perCellQC-Metrics* function from the *scater* R package (McCarthy *et al.*, 2017). These metrics include the total number of expressed genes, the overall count across all genes, and the fraction of counts assigned to control genes such as mitochondrial genes. By setting a specific threshold on per-cell QC metrics, high-quality cells can be retained. In our setting, outliers are defined as cells with library sizes more than two median absolute deviations away from the median library size. Figure 3.1 summarizes the process from unprocessed to processed single-cell experiment.

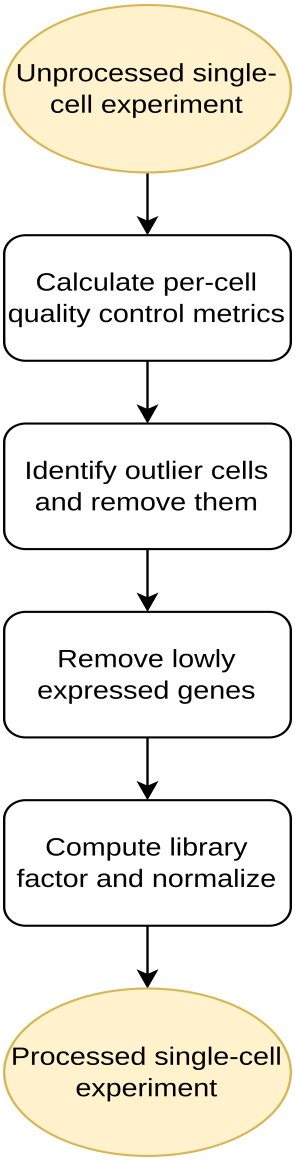

Figure 3.1: Quality control process from unprocessed, raw to processed, filtered single-cell experiment

After filtering, the data set consists of 23'543 cells and 18'537 genes. Next, we used the *singleR* function from the *singleR* R package (Aran *et al.*, 2019) for cell-type annotation. Cell-type annotation is important to determine what biological state is represented by cell clusters which helps the interpretability of the results and their implications (Amezquita *et al.*, 2020). *singleR* is a method that assigns labels based on the reference samples with the highest Spearman rank correlation while only using marker genes between pairs of labels to focus on the relevant differences between cell types (Aran *et al.*, 2019). Figure 3.2 shows the Uniform Manifold Approximation and Projection (UMAP) of the cells coloured by their respective sample id. From Figure 3.2 one can observe that the projection of the cells is very similar across the samples. Further, Figure 3.3 also illustrates that cells from the same cell-type cluster together, as one would expect.

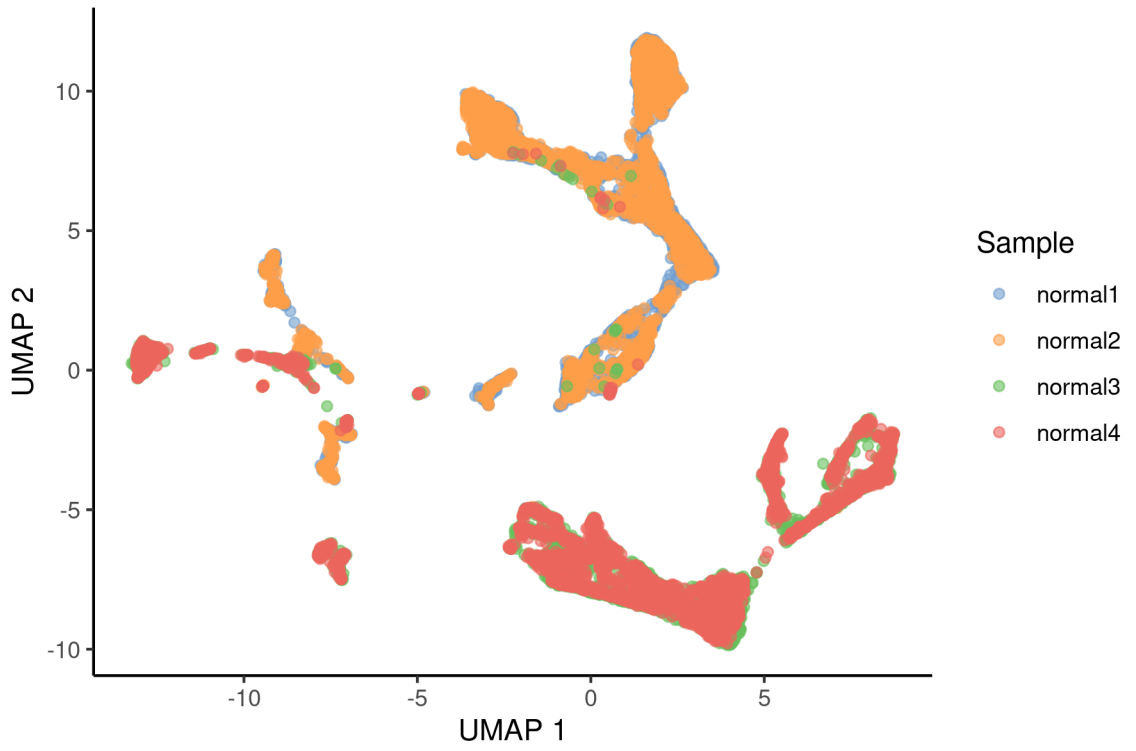

Figure 3.2: UMAP representation of the mouse kidney cells coloured by sample id

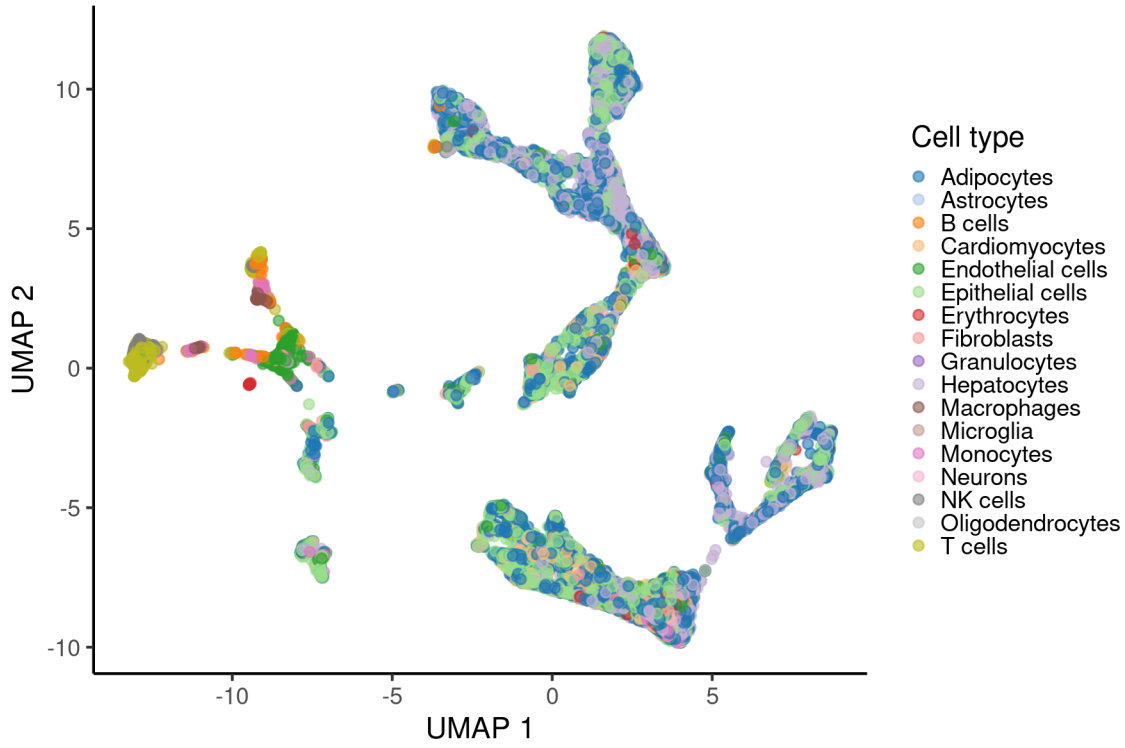

Figure 3.3: UMAP representation of the mouse kidney cells coloured by cell type

Figure 3.4 shows that the annotated cells were largely classified as either: Adipocytes, Epithelial cells and Hepatocytes. Additionally, one can observe that those three cell types are approximately evenly distributed across the four samples. For further analyses we focus on cell types with at least 100 counts to investigate the performance of existing methods for the detection of differentially expressed genes and design our own simulation study.

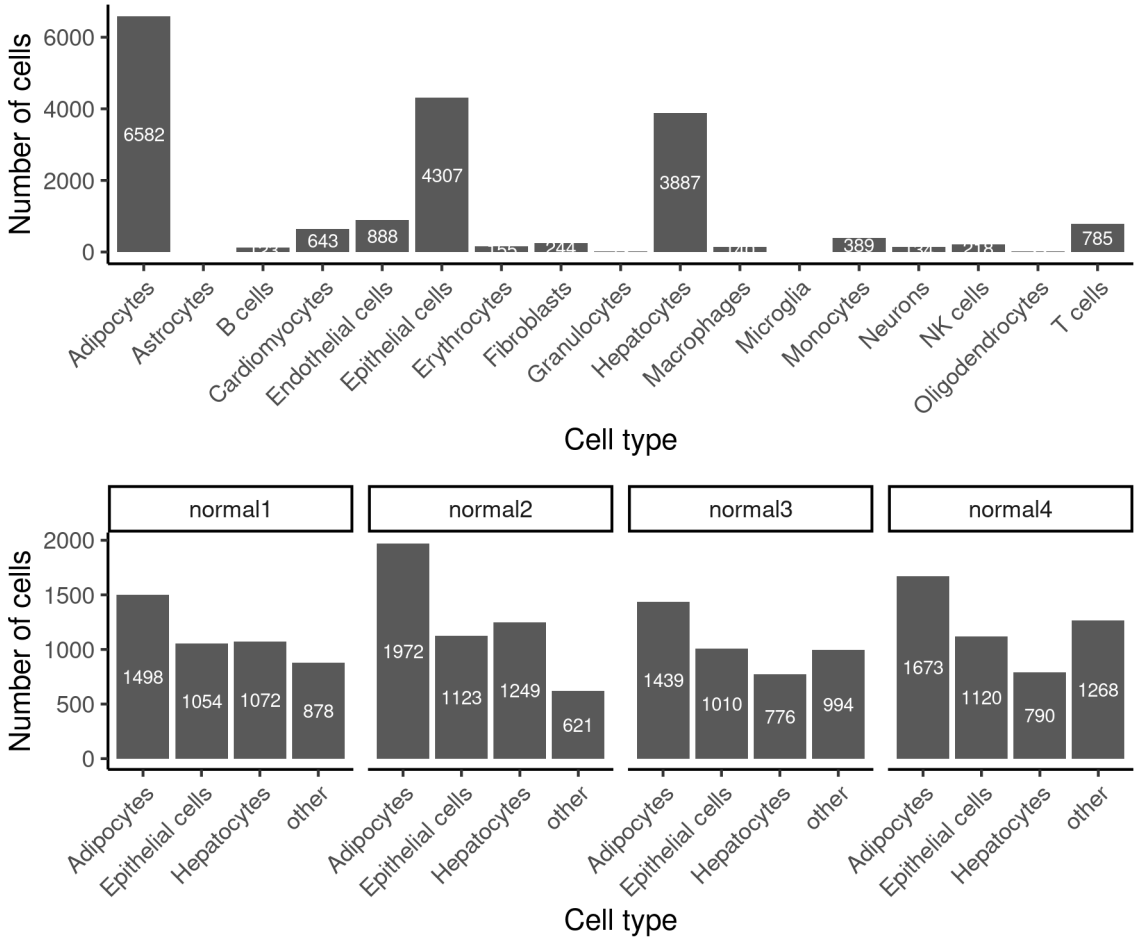

Figure 3.4: Frequency distribution of the cell types after quality control

3.2 Simulation study

3.2.1 Simulation strategy

Initially, the simulation strategy was to invert the spliced and unspliced counts for 10% of genes for all cells that belong to an arbitrary group A. The set of genes whose counts were to be inverted was randomly drawn by a sampling algorithm without replacement (hypergeometric distribution). Additionally, differential gene expression was introduced in 10% of genes in all cells that belong to said arbitrary group A. This was achieved by multiplying the counts (ten-fold gene expression) for 10% randomly drawn genes in group A. Again, the set of genes was randomly drawn by the same sampling algorithm as before. In this manner two datasets were created, one with and one without differential gene expression. Figure 3.5 illustrates the simulation process from the original mouse kidney data set to two the simulated data sets.

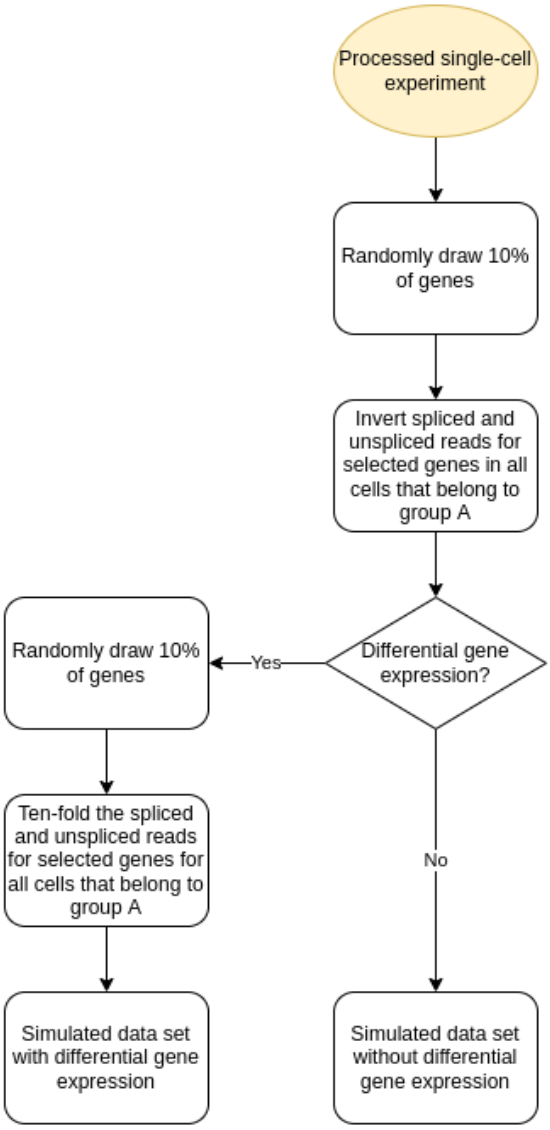

Figure 3.5: Simulation process from original mouse data set to simulated data sets

However, these simulations do not account for mapping uncertainty. Therefore, the simulation process was enhanced with the use of Minnow (Sarkar *et al.*, 2019). (EXPLAIN MINNOW HERE OR IN METHODS SECTION????)

3.2.2 Simulating without Minnow

3.2.3 Simulating with Minnow

3.3 Null data analysis on the mouse kidney data

3.4 Computational benchmark

3.5 Data availability

Kidney mouse cells
The raw data can be downloaded from NCBI GEO (accession number GSE107585).

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE107585

3.6 Code availability

All code for data preprocessing and analysis associated with the thesis is available at https://github.com/joelmeili/DifferentialRegulation. Any updates will also be published on GitHub.

Chapter 4

Discussion

4.1 Conclusion

4.2 Outlook

Bioconductor package *DifferentialRegulation* (Tiberi, 2022) is a method for detecting differentially regulated genes between two groups of samples (e.g., healthy vs. disease, or treated vs. untreated samples), by targeting differences in the balance of spliced and unspliced mRNA abundances, obtained from single-cell RNA-sequencing (scRNA-seq) data. *DifferentialRegulation* accounts for the sample-to-sample variability, and embeds multiple samples in a Bayesian hierarchical model. In particular, when reads are compatible with multiple genes or multiple splicing versions of a gene (unspliced spliced or ambiguous), the method allocates these multi-mapping reads to the gene of origin and their splicing version. Parameters are inferred via Markov chain Monte Carlo (MCMC) techniques (Metropolis-within-Gibbs).

Appendix A

Figures

Bibliography

Amezquita, R. A., Lun, A. T., Becht, E., Carey, V. J., Carpp, L. N., Geistlinger, L., Marini, F., Rue-Albrecht, K., Risso, D., Soneson, C.,*et al.* (2020). Orchestrating single-cell analysis with bioconductor. *Nature methods*, **17**, 137–145. 11, 12

Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from rna-seq data. 9

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., Chak, S., Naikawadi, R. P., Wolters, P. J., Abate, A. R., Butte, A. J., and Bhattacharya, M. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.*, **20**, 163–172. 12

Crowell, H. L., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M. D. (2020). Muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature communications*, **11**, 1–12. 6

Dharshini, S. A. P., Taguchi, Y.-H., and Gromiha, M. M. (2020). Identifying suitable tools for variant detection and differential gene expression using rna-seq data. *Genomics*, **112**, 2166–2172. 7

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013). Star: ultrafast universal rna-seq aligner. *Bioinformatics*, **29**, 15–21. 5

Gaidatzis, D., Burger, L., Florescu, M., and Stadler, M. B. (2015). Analysis of intronic and exonic reads in rna-seq data characterizes transcriptional and post-transcriptional regulation. *Nature biotechnology*, **33**, 722–729. 6

Haque, A., Engel, J., Teichmann, S. A., and Lönnberg, T. (2017). A practical guide to single-cell rna-sequencing for biomedical research and clinical applications. *Genome medicine*, **9**, 1–12. 3, 4

He, D., Zakeri, M., Sarkar, H., Soneson, C., Srivastava, A., and Patro, R. (2022). Alevin-fry unlocks rapid, accurate and memory-frugal quantification of single-cell rna-seq data. *Nature Methods*, **19**, 316–322. 5

Huang, Y. and Sanguinetti, G. (2021). Brie2: computational identification of splicing phenotypes from single-cell transcriptomic experiments. *Genome biology*, **22**, 1–15. 6

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M. E., Lönnerberg, P., Furlan, A.,*et al.* (2018). Rna velocity of single cells. *Nature*, **560**, 494–498. 5

Love, M. I., Soneson, C., and Patro, R. (2018). Swimming downstream: statistical analysis of differential transcript usage following salmon quantification. *F1000Research*, **7**, . 9

McCarthy, D. J., Campbell, K. R., Lun, A. T. L., and Willis, Q. F. (2017). Scater: pre-processing, quality control, normalisation and visualisation of single-cell RNA-seq data in R. *Bioinformatics*, **33**, 1179–1186. 11

McDermaid, A., Chen, X., Zhang, Y., Wang, C., Gu, S., Xie, J., and Ma, Q. (2018). A new machine learning-based framework for mapping uncertainty analysis in rna-seq read alignment and gene expression estimation. *Frontiers in genetics*, **9**, 313. 7

Melsted, P., Booeshaghi, A., Liu, L., Gao, F., Lu, L., Min, K. H. J., da Veiga Beltrame, E., Hjörleifsson, K. E., Gehring, J., and Pachter, L. (2021). Modular, efficient and constant-memory single-cell rna-seq preprocessing. *Nature biotechnology*, **39**, 813–818. 5

Park, J., Shrestha, R., Qiu, C., Kondo, A., Huang, S., Werth, M., Li, M., Barasch, J., and Suszták, K. (2018). Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science*, **360**, 758–763. 11

Rao, M. S., Van Vleet, T. R., Ciurlionis, R., Buck, W. R., Mittelstadt, S. W., Blomme, E. A., and Liguori, M. J. (2019). Comparison of rna-seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. *Frontiers in genetics*, **9**, 636. 3

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140. 6

Sarkar, H., Srivastava, A., and Patro, R. (2019). Minnow: a principled framework for rapid simulation of dscrna-seq data at the read level. *Bioinformatics*, **35**, i136–i144. 15

Srivastava, A., Malik, L., Smith, T., Sudbery, I., and Patro, R. (2019). Alevin efficiently estimates accurate gene abundances from dscrna-seq data. *Genome biology*, **20**, 1–16. 5

Stark, R., Grzelak, M., and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, **20**, 631–656. 3

Tiberi, S. (2022). *DifferentialRegulation: Differentially regulated genes from scRNA-seq data*. R package version 1.0.7. 17

Tiberi, S., Crowell, H. L., Weber, L. M., Samartsidis, P., and Robinson, M. D. (2021). distinct: a novel approach to differential distribution analyses. 6

Weiler, P., Van den Berge, K., Street, K., and Tiberi, S. (2021). A guide to trajectory inference and rna velocity. 5