MY VERY SOPHISTICATED MASTER THESIS

Master Thesis

Systems biology master program

Vilnius university

STUDENT NAME:	Juozapas Ivanauskas
STUDENT NUMBER:	2316457
SUPERVISOR:	dr. Simonas Juzėnas
CONSULTANT:	Justina Žvirblytė

SUPERVISOR DECISION:	.....
FINAL GRADE	.....
DATE OF SUBMISSION:	DD MMMM 20YY

Contents

1. LIST OF ABBREVIATIONS

GRN	gene regulatory network
NGS	next generation sequencing
RT	reverse transcription
scRNAseq	single cell RNA sequencing
UMI	unique molecular identifier

2. INTRODUCTION

3. AIM AND TASKS

4. LITERATURE REVIEW

In this chapter I will provide general review of single cell transcriptomics and related challenges.

4.1 Introduction to single cell transcriptomics

Cells are the fundamental units of life, forming the basis of all living organisms. One of the major goals of biology is to understand cellular systems and the processes occurring within cells. Since the discovery of the DNA structure in 1953 and the development of the conceptual framework for genetic information transfer, scientists have made significant efforts to sequence the genomes of various organisms. This led to the development of the first sequencing methods, such as Sanger sequencing in 1975, which laid the foundation for next-generation sequencing technologies in use today, including the widely used Illumina platform (Heather and Chain 2016). Current sequencing methods allow us to obtain the complete genetic sequence of any organism. However, the genome alone cannot explain the full diversity of cells in multicellular organisms, as all cells share the same genome but exhibit significant variation in shape, size, and function.

RNA sequencing (RNAseq), on the other hand, enables the measurement of gene expression within cells, providing valuable insights into cellular processes. RNAseq methods largely follow DNA sequencing protocols, with the addition of a step where complementary DNA (cDNA) is synthesized from RNA (Heumos et al. 2023). The first RNAseq methods were developed for bulk sequencing, where RNA from entire cell populations is sequenced, providing an average gene expression profile across the population. Although bulk RNAseq provided valuable insights into the dynamics of cellular processes (such as changes in disease states in response to therapeutics, detection of gene isoforms, gene fusions, and various other properties of target cells (Heumos et al. 2023)), this approach masks non-dominant processes and cell-to-cell variability through averaging. This limitation was addressed by the introduction of single-cell RNA sequencing (scRNAseq) methods, which allow the generation of transcriptomic profiles from individual cells, providing high-resolution insights into cellular systems.

Current scRNAseq methods enable the generation of transcriptomic profiles from thousands of cells at unprecedented resolution in a single experiment. These data can be used for constructing cellular atlases (Rozenblatt-Rosen et al. 2017), understanding disease mechanisms (Z. Zhang, Chen, and Peng 2024), exploring cell differentiation and developmental processes (Skinner, Asad, and Haque 2024), and many other applications.

4.2 Key methods and technologies in scRNAseq

4.2.1 Key methods

All scRNAseq protocols share these main three steps: isolation of single cells, library preparation and sequencing (Andrews and Hemberg 2018).

The first step is mainly done in two ways: either by placing cells in separate droplets (microfluidics approach), or by separating cells into different wells (plate-based approach).

The next generation sequencing (NGS) usually requires nanograms or more of DNA, and the RNA content in single cells is far from this amount (Wu et al. 2017). Consequently, before sequencing, reverse transcription (RT) and amplification is needed.

Finally, the prepared library is sequenced using NGS methods. The most popular is .....

4.2.2 Current scRNAseq Platforms

As mentioned before, scRNAseq methods mainly can be grouped in two groups: droplet-based and plate-based.

Droplet-based methods (e.g. inDrops (Klein et al. 2015), Drop-seq (Macosko et al. 2015), Chromium by 10X Genomics (Zheng et al. 2017)) separate cells by placing them into different droplets, containing hidrogel primers and lysis mix. Primers usually share common structure, including barcode sequenes, unique molecular identifiers (UMIs), PCR handlers and poly-T (X. Zhang et al. 2019). Cell barcodes are sequences used for determining the cell from which particular read sequenced (in sequencing step, content from all droplets is mixed and sequenced at once). UMIs are used to quantify real amount of RNA in cells (after amplification, more than one copy of each captured RNA is present). PCR handlers are used for the amplification, while poly-T are used for capturing RNAs. Example of primer design can be seen in figure **??**). Once cells are in the droplets, cell lysis takes place, RNAs escapes cells and are captured by primers. Depending on method, reverse transcription either takes place directly in the droplets (inDrops, 10X) or after demulsification (Drop-seq). Next steps usually include RNA fragmentation and PCR amplification, followed by NGS.

Droplet-based methods are high-throughput (current microfluidic devices are able to generate thousands of above described droplets per second (Prakadan, Shalek, and Weitz 2017)), cost-effective, but have low detection rates compared to other methods and captures only 3' (or 5') ends of transcripts (Heumos et al. 2023). Capturing only 3' ends of transcripts might be not a problem when trying to identify cell populations, however, it masks such processes as splicing variants, thus should be considered carefully while planning experiments.

Plate-based methods (e.g. CEL-Seq2 (Hashimshony et al. 2016), Smart-seq2 (Picelli et al. 2013)) separate cells by placing them into different microwells on a plate. Before this, cells can be sorted using ,for example, fluorescent-activated cell sorting (FACS) (Heumos et al. 2023). Similarly to droplets, microwells contain lysis buffers and RT mix, and these processes are followed by amplification and NGS (Hashimshony et al. 2016). Barcodes can be integrated into reverse transcription step similarly as in droplet case.

Overall, plate-based methods have lower throughput, might be more costly and labor-intensive, but offers recovery of many genes per cell, allows prior sorting and (for some protocols) it is possible

to sequence full transcripts (Heumos et al. 2023).

In the next sections, we will focus on the droplet-based approaches, as all the data used in this thesis is generated by droplet-based methods.

4.3 Data quality and challenges in scRNAseq data

The quality of scRNAseq data

4.3.1 Noise

The noise present in the scRNAseq data can be either biological or technical.

4.3.2 Dimentionality

4.4 Computational tools and analytical approaches

4.4.1 Raw data processing

The output of the typical scRNAseq experiment is FASTQ files, containing recorded sequences, as well as (depending on method) barcode and UMI sequences and quality scores. The subsequent processing steps include quality control of FASTQ file (is done based on quality scores), filtering dublicate reads (using UMIs), mapping reads to the genome sequence and assigning the reads to the genes, and, finally, counting gene expression per cell (barcode) (Heumos et al. 2023) (see figure **??**). Usually, all these steps are done with single piece of dedicated software, such as STARsolo (Kaminow, Yunusov, and Dobin 2021), CellRanger (Zheng et al. 2017) or other. It should be noted, that there are variations in the above described pipeline, depending on many experiment-related (e.g., if there is known genome sequence or transcriptone of the study organism), or method-related (e.g., if UMIs are used in the protocol) factors. The typical result of such processing is cell-gene matrix (i.e. a matrix with cells as rows, genes as columns, and the entries being the number of captured RNAs in the particular cell coresponding to particular gene).

4.4.2 Preprocessing of count matrices

Preprocessing of count matrices usually involves such steps: quality control, normalization and feature selection.

Quality of indidual cells can be evaluated based on several factors, such as mitochondrial gene contents (apoptotic cells tend to have high proportion of mitochondrial genes (Heumos et al. 2023)) or total number of captured genes (very low numbers can be produced by empty droplets). Also, in some cases, two cells can end up in one droplet, resulting count matrix entry corresponding to genes from both cells. Such matrix entries (doublets) can be filtered by using specialized software such Scrublet (Wolock, Lopez, and Klein 2019) or scDblFinder (Germain et al. 2022). Ambient RNA is another aspect, increasing noise in the scRNAseq data. It is the RNA that escapes individual droplets and spreads in the medium and other droplets, causing some background noise. Even though the amount of such RNA is not high (give some numbers and citation), cleaning count matrices from such RNAs can increase quality of the data. This can be done by finding out background noise profile from

the empty droplets and correcting count matrix accordingly. There are dedicated softwares, such as SoupX (Young and Behjati 2020), decontX (Yang et al. 2020), CellBender (Fleming et al. 2023) and others.

The next step in preprocessing pipeline is normalization. The aim of it is to transform the data such that variantion in gene expression levels is similar, so that subsequent analysis would be more efficient (Ahlmann-Eltze and Huber 2023). Also, it might help to remove some biases, such as sequencing depth in the case of combining data from several samples (Lingen, Suarez-Diez, and Saccenti 2024). There are plenty of methods for normalization, based on various approaches (e.g. delta-method-based, residual-based, latent gene expression-based, count-based (Ahlmann-Eltze and Huber 2023)). Therefore, one should carefully choose the normalization method, based on the individual experiment design. General recomendations for normalization suggest comparing several methods and if the results are similar, to use the simpler method (Lingen, Suarez-Diez, and Saccenti 2024). Sophisticated methods doesn't necessarily show better results, and recent benchmarking study (Ahlmann-Eltze and Huber 2023) has showed that such simple method (particularly the logarithm normalization, where each element y of count matrix is transformed by formula $y_{trnasformed} = log(y + 1)$) performs as well or better than more advanced methods.

When the data is normalized and cleaned, one can get rid of not informative genes. Initially, count matrices contain all the genes that were present in the transcriptome. However, not all of them are expressed in the sequenced data, or are expressed in negligable numbers (Heumos et al. 2023). Therefore, it is common practice to filter such genes (e.g., genes that are expressed in less than 3 cells). Moreover, some genes might be expressed in all the cells more or less evenly (housekeeping genes), and thus don't provide information that could be usefull in, for instance, grouping cells or determining cell types. Therefore, in many applications, it is beneficial to leave only those genes, that are highly variable between cells. In such way, the dimensionality of the count matrix is greatly reduced without loosing significant information. Additionally, one can filter out those genes that are out of scope of indicidual study.

4.4.3 Dimensionality reduction

Even after filtering and selecting only highly variable genes, there are usually left several thousand genes. It is not possible to visualize (and hard to interpret in general) data of such high dimentionality, therefore, dimensionality reduction is essential step of subsequent analysis. The idea of dimentionality reduction is simple: to reduce the dimentions of the data loosing as less information as possible. There are number of such methods based on different mathematical concepts, but the most widely used today include t-SNE (Hinton and Roweis 2002), UMAP (McInnes, Healy, and Melville 2018) and principal component analysis (PCA). While the use of these algorithms are supported by some benchmarking studies (in the study of Xiang et al. 2021, t-SNE was showed best performance, while UMAP showed the highest stability), other benchmarking studies show different results. The study of Koch et al. 2021 suggested that such overlooked methods as latent Dirichlet alloacation (LDA) and PHATE show best performance, while Sun et al. 2019 provided guidelines for choosing dimensionality reduction method depending on downstream analysis tasks, and in their results UMAP and tSNE were not the first choises. Hence, even though UMAP and t-SNE remain the most popular in the field, one should consider using other methods as well.

4.4.4 Clustering and other stuff

4.5 Enhancing scRNAseq data

4.6 Deriving useful information from scRNAseq data

4.7 Current limitations and future perspectives

5. METHODS

6. RESULTS

7. DISCUSSION

8. CONCLUSIONS

9. RECOMMENDATION

10. ACKNOWLEDGEMENTS

11. REFERENCES

(1) Ahlmann-Eltze, Constantin and Wolfgang Huber (Apr. 2023). "Comparison of transformations for single-cell RNA-seq data". In: *Nature Methods* 20.5, pp. 665–672. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01814-1.

(2) Andrews, Tallulah S. and Martin Hemberg (Feb. 2018). "Identifying cell populations with scRNASeq". In: *Molecular Aspects of Medicine* 59, pp. 114–122. ISSN: 0098-2997. DOI: 10.1016/j.mam.2017.07.002.

(3) Fleming, Stephen J. et al. (Aug. 2023). "Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender". In: *Nature Methods* 20.9, pp. 1323–1335. ISSN: 1548-7105. DOI: 10.1038/s41592-023-01943-7.

(4) Germain, Pierre-Luc et al. (May 2022). "Doublet identification in single-cell sequencing data using scDblFinder". In: *F1000Research* 10, p. 979. ISSN: 2046-1402. DOI: 10.12688/f1000research.73600.2.

(5) Hashimshony, Tamar et al. (Apr. 2016). "CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq". In: *Genome Biology* 17.1. ISSN: 1474-760X. DOI: 10.1186/s13059-016-0938-8.

(6) Heather, James M. and Benjamin Chain (Jan. 2016). "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1, pp. 1–8. ISSN: 0888-7543. DOI: 10.1016/j.ygeno.2015.11.003.

(7) Heumos, Lukas et al. (Mar. 2023). "Best practices for single-cell analysis across modalities". In: *Nature Reviews Genetics* 24.8, pp. 550–572. ISSN: 1471-0064. DOI: 10.1038/s41576-023-00586-w.

(8) Hinton, Geoffrey E and Sam Roweis (2002). "Stochastic Neighbor Embedding". In: *Advances in Neural Information Processing Systems*. Ed. by S. Becker, S. Thrun, and K. Obermayer. Vol. 15. MIT Press. URL: https://proceedings.neurips.cc/paper_files/paper/2002/file/6150ccc6069bea6b5716254057a194ef-Paper.pdf.

(9) Kaminow, Benjamin, Dinar Yunusov, and Alexander Dobin (May 2021). "STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data". In: DOI: 10.1101/2021.05.05.442755.

(10) Klein, Allon M. et al. (May 2015). "Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells". In: *Cell* 161.5, pp. 1187–1201. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.04.044.

(11) Koch, Forrest C et al. (Aug. 2021). "Supervised application of internal validation measures to benchmark dimensionality reduction methods in scRNA-seq data". In: *Briefings in Bioinformatics* 22.6. ISSN: 1477-4054. DOI: 10.1093/bib/bbab304.

(12) Lingen, Henk J. van, Maria Suarez-Diez, and Edoardo Saccenti (Dec. 2024). "Normalization of gene counts affects principal components-based exploratory analysis of RNA-sequencing data". In:

Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1867.4, p. 195058. ISSN: 1874-9399. DOI: 10.1016/j.bbagrm.2024.195058.

(13) Macosko, Evan Z. et al. (May 2015). "Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets". In: *Cell* 161.5, pp. 1202–1214. ISSN: 0092-8674. DOI: 10.1016/j.cell.2015.05.002.

(14) McInnes, Leland, John Healy, and James Melville (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. DOI: 10.48550/ARXIV.1802.03426.

(15) Picelli, Simone et al. (Sept. 2013). "Smart-seq2 for sensitive full-length transcriptome profiling in single cells". In: *Nature Methods* 10.11, pp. 1096–1098. ISSN: 1548-7105. DOI: 10.1038/nmeth.2639.

(16) Prakadan, Sanjay M., Alex K. Shalek, and David A. Weitz (Apr. 2017). "Scaling by shrinking: empowering single-cell "omics" with microfluidic devices". In: *Nature Reviews Genetics* 18.6, pp. 345–361. ISSN: 1471-0064. DOI: 10.1038/nrg.2017.15.

(17) Rozenblatt-Rosen, Orit et al. (Oct. 2017). "The Human Cell Atlas: from vision to reality". In: *Nature* 550.7677, pp. 451–453. ISSN: 1476-4687. DOI: 10.1038/550451a.

(18) Skinner, Oliver P, Saba Asad, and Ashraful Haque (June 2024). "Advances and challenges in investigating B-cells via single-cell transcriptomics". In: *Current Opinion in Immunology* 88, p. 102443. ISSN: 0952-7915. DOI: 10.1016/j.coi.2024.102443.

(19) Sun, Shiquan et al. (Dec. 2019). "Accuracy, robustness and scalability of dimensionality reduction methods for single-cell RNA-seq analysis". In: *Genome Biology* 20.1. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1898-6.

(20) Wolock, Samuel L., Romain Lopez, and Allon M. Klein (Apr. 2019). "Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data". In: *Cell Systems* 8.4, 281–291.e9. ISSN: 2405-4712. DOI: 10.1016/j.cels.2018.11.005.

(21) Wu, Angela R. et al. (June 2017). "Single-Cell Transcriptional Analysis". In: *Annual Review of Analytical Chemistry* 10.1, pp. 439–462. ISSN: 1936-1335. DOI: 10.1146/annurev-anchem-061516-045228.

(22) Xiang, Ruizhi et al. (Mar. 2021). "A Comparison for Dimensionality Reduction Methods of Single-Cell RNA-seq Data". In: *Frontiers in Genetics* 12. ISSN: 1664-8021. DOI: 10.3389/fgene.2021.646936.

(23) Yang, Shiyi et al. (Mar. 2020). "Decontamination of ambient RNA in single-cell RNA-seq with DecontX". In: *Genome Biology* 21.1. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1950-6.

(24) Young, Matthew D and Sam Behjati (Dec. 2020). "SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data". In: *GigaScience* 9.12. ISSN: 2047-217X. DOI: 10.1093/gigascience/giaa151.

(25) Zhang, Xiannian et al. (Jan. 2019). "Comparative Analysis of Droplet-Based Ultra-High-Throughput Single-Cell RNA-Seq Systems". In: *Molecular Cell* 73.1, 130–142.e5. ISSN: 1097-2765. DOI: 10.1016/j.molcel.2018.10.020.

(26) Zhang, ZhenWei, MianMian Chen, and XiaoLian Peng (July 2024). "Integrated analysis of single-cell and bulk RNA-sequencing identifies a signature based on drug response genes to predict prognosis and therapeutic response in ovarian cancer". In: *Heliyon* 10.13, e33367. ISSN: 2405-8440. DOI: 10.1016/j.heliyon.2024.e33367.

(27) Zheng, Grace X. Y. et al. (Jan. 2017). "Massively parallel digital transcriptional profiling of single cells". In: *Nature Communications* 8.1. ISSN: 2041-1723. DOI: 10.1038/ncomms14049.

12. SUMMARY

13. SUMMARY IN LITHUANIAN

14. APPENDICES