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 (NGS) 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 has 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 for 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, M. Chen, and X. Peng 2024), exploring cell differentiation and developmental processes (Skinner, Asad, and Haque 2024), among 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, followed by data processing and analysis.

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, which contain hidrogel primers and lysis mix. Primers usually share a common structure, including barcode sequenes, unique molecular identifiers (UMIs), PCR handlers and poly-T sequences (X. Zhang et al. 2019). Cell barcodes are sequences used for determining the cell from which a particular read was sequenced (during sequencing, the 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 sequences are used for capturing RNAs. An example of primer design can be seen in figure 4.1). Once cells are in the droplets, cell lysis takes place, RNAs escape the cells and are captured by the primers. Depending on method, reverse transcription either takes place directly in the droplets (inDrops, 10X) or after demulsification (Drop-seq). The 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 when 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, 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,

Figure 4.1: Example of barcode (inDrops). The top image contains schematic view, while the bottom one shows example sequence. T7RNAP, PE1 and W1 sites here are used for primer assembly, while photo-cleavable spacer is used to release primers from the gel beads. Image taken from (Klein et al. 2015).

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 scRNAseq data offers high resolution insights in the cellular systems, however, it comes with unique (and non-unique) challenges that need to be overcame to have clean, good-quelity data. In this section I will briefly overview noise sources and data scale challenges.

4.3.1 Noise

The noise in the scRNAseq data can have either biological or technical origin. The biological noise is always present in the cellular systems due to the stochasticity of all biological processes (Vázquez-Jiménez, Santillán, and Rodríguez-González 2017). While it is always in the data, more important is to eliminate technical noise. Technical noise is the artefact of sequencing procedures. The typical challenge in scRNAseq data is large number of dropout values (and consequently sparse data matrices). I.e., if there is a zero entry in the matrix, is not clear whether the gene was not expressed in the particular cell, or was expressed but not captured. This problem is particularly important if one is interested in rare transcripts.

Additionally, droplet-based methods introduce droplets and ambient RNA. These two will be discussed in section 4.4.1. In general, when the origin of noise is clear (as in droplet case), it is easier to make computational tools to eliminate it. For dropout values, it is harder to say whether it is artefact or biological variability, hence eliminating such noise is not easy.

4.3.2 Data scale and dimentionality

Another challenge is data scale itself – typical scRNAseq datasets constains thousands of cells and tens of thousands of genes. Such scales requires very efficiant analysis algorithms (some analysis tools are just too slow for using (McCalla et al. 2023) and makes interpretability of data harder. Many

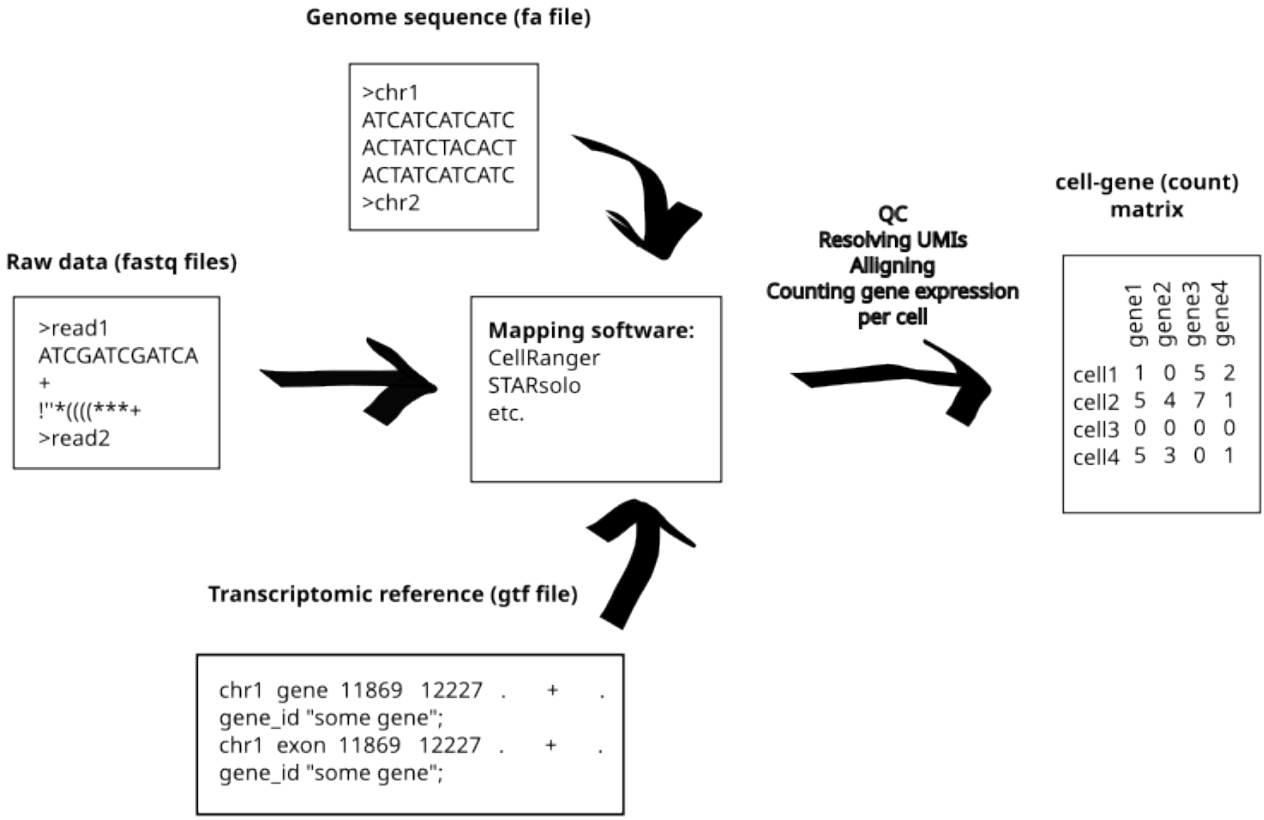


Figure 4.2: Pipeline of processing raw data.

analysis pipelines uses dimensionality reduction to reduce the dimensions of the data (more on it in 4.4.3).

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 (based on quality scores), filtering dublicate reads (using UMIs), mapping reads to the genome sequence, assigning the reads to the genes, and finally, counting gene expression per cell (barcode) (Heumos et al. 2023) (see figure 4.2). Usually, all these steps are performed with a single piece of dedicated software, such as STARsolo (Kaminow, Yunusov, and Dobin 2021), CellRanger (Zheng et al. 2017) or others. It should be noted, that there are variations in the pipeline described above, depending on many experiment-related (e.g., whether the genome sequence or transcriptone of the study organism is known), or method-related (e.g., whether UMIs are used in the protocol) factors. The typical result of such processing is cell-gene matrix (i.e., a matrix where rows represent cells, columns represent genes, and each entry indicates the number of captured RNAs for a given gene in a specific cell).

4.4.2 Preprocessing of count matrices

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

The quality of indidual cells can be evaluated based on several factors, such as mitochondrial gene content (apoptotic cells tend to have a higher proportion of mitochondrial genes (Heumos et al. 2023)) or total number of captured genes (very low numbers can be produced by empty droplets). In some cases, two cells can end up in one droplet, resulting in count matrix row 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). Another source of noise in scRNAseq data is ambient RNA, which consists of RNA that escapes individual droplets and spreads into the medium or other droplets, leading to background noise. Even though the amount of such RNA is not high (in good quality datasets it can be around 2% (Young and Behjati 2020)), removing these RNAs from the count matrix can improve data quality. This can be achieved by identifying the background noise profile from empty droplets and adjusting the 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 goal of normalization is to transform the data so that the variation in gene expression levels is comparable, making subsequent analysis more efficient (Ahlmann-Eltze and Huber 2023). Normalization can also help eliminate biases, such as differences in sequencing depth when combining data from multiple samples (Lingen, Suarez-Diez, and Saccenti 2024). There are numerous normalization methods, based on different approaches (e.g., delta-method-based, residual-based, latent gene expression-based, count-based (Ahlmann-Eltze and Huber 2023)). Thus, selecting a normalization method should be done carefully, depending on the experimental design. General recommendations for normalization suggest comparing several methods, and if the results are similar, opting for the simpler method (Lingen, Suarez-Diez, and Saccenti 2024). Sophisticated methods do not necessarily show better results, and a recent benchmarking study by Ahlmann-Eltze and Huber (2023) has shown that simpler 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.

Once the data is normalized and cleaned, one can filter out non-informative genes. Initially, count matrices contain all the genes that are 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 three cells). Moreover, some genes might be expressed in all the cells more or less evenly (housekeeping genes), which do not provide useful 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, genes that are outside the scope of the specific study can also be filtered out.

4.4.3 Dimensionality reduction

Even after filtering and selecting only highly variable genes, several thousand genes usually remain. It is not feasible 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 little information as possible. There are number dimensionality reduction 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). Although 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 report different findings. The study of Koch et al. (2021) suggested that such overlooked methods as latent Dirichlet alloacation (LDA) and PHATE show best performance. Meanwhile 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 on the top choices. Thus, while UMAP and t-SNE remain the most popular methods in the field, it is worth considering alternative methods as well.

4.4.4 Clustering and other analyses

One of the most popular tasks of scRNAseq data analysis is to identify and classify cell populations (Andrews and Hemberg 2018). This task requires to assign cells to different groups (clusters), such that cells in the same clusters are similar and distinct from cells in other clusters. There is a great variety of clustering algorithms available, including k-means, hierarchical and consensus clustering (L. Peng et al. 2020). Benchmarking studies suggest that "no individual scRNA-seq clustering algorithm can capture true clusters and achieve optimal performance in all situations" (L. Peng et al. 2020).

Clustering is usually followed by cell typing (i.e., assigning cell type to the identified clusters), which is done by finding cell type specific markers or using automatic (machine learning) tools such as CellTypist (Domínguez Conde et al. 2022). The subsequent steps in the analysis depend on the focus of the particular study and can include analysis of the dynamics of cellular systems (RNA velocity, pseudotime), inferring gene regulatory networks (GRNs), and more.

4.5 Enhancing scRNAseq data

Given the challenges associated with scRNAseq data, there have been attempts to improve the quality of such data. In this section, I will provide an overview of two methods: data imputation and enhancing the transcriptomic reference.

4.5.1 Data imputation

One of the challenges present in scRNAseq data is the large number of dropout values. Dropout values refer to instances where gene expression is present in a cell but is missed in the scRNAseq data. This problem can mask important relationships between genes and complicate downstream analysis (M. Wang et al. 2022). To impute dropout values, many tools have been suggested. These methods can be divided into four categories: model-based methods (bayNorm (Tang et al. 2019), BISCUIT

(Azizi et al. 2017), SAVER (Huang et al. 2018) etc.), low-ranked matrix-based (ALRA (Linderman et al. 2022), ENHANCE (Wagner, Barkley, and Yanai 2019), scRMD (C. Chen et al. 2020) etc.), data smoothing methods (e.g. KNN-smoothing (Wagner, Yan, and Yanai 2017), MAGIC (Dijk et al. 2018) etc.) and deep learning methods (e.g. DCA (Eraslan et al. 2019), DeepImpute (Arisdakessian et al. 2019) etc.) (M. Wang et al. 2022).

The study by Dai et al. (2022) has shown that imputation methods are advantageous for recovering gene expression, and among these methods, deep learning-based ones, such as DCA, DeepImpute, scIGANs (Xu et al. 2020) show the best performance. However, it was also shown that imputation methods can introduce false positives. In the study by Andrews and Hemberg (2019), it was shown that data smoothing methods (e.g. MAGIC, KNN-smoothing) generate most false positives among the different types of methods, but other methods can generate relatively large number of false positives as well, depending on the dataset. Data imputation does not necessarily improve downstream analysis (e.g, it was shown that imputation doesn't improve inference of gene regulatory networks (McCalla et al. 2023)), therefore one should carefully choose whether to impute data and which method to use.

4.5.2 Enhancing transcriptomic reference

One of the problems that scRNAseq is facing is the complexity of the genome. The "raw" human transcriptomic reference (a file containing information about genes) contains over 60000 genes (Frankish et al. 2022). Not all of these genes are expected to be captured by scRNAseq data Thus, a simple approach to improving the transcriptomic reference is to filter out the genes that are not expected to appear in scRNAseq data. In this way, events where two genes overlap in the genome, but one is not expected to appear in the scRNAseq data, are resolved, allowing alignment tools to more easily assign reads from these regions to the correct genes. This approach is used in publicly available 10X transcriptomic references (Zheng et al. 2017). Even though it improves mapping performance, it does not address all the issues with the transcriptomc reference.

Pool et al. (2023) has suggested three steps to enhance transcriptomic reference: including reads mapped to intronic sequence to the analysis, extending 3' ends of some genes and resolving overlaps between certain genes. The first suggestion is not new in the field of scRNAseq. There are concepts such as RNA velocity based on spliced an unspiced RNA ratio (La Manno et al. 2018), showing that such including intronic reads in the analysis can provide valuable information. Moreover, most mapping tools (e.g. STARsolo, CellRanger) contains options to use either only exonic parts or full genes for read alignment. The second suggestion is based on the observation, that scRNAseq data often contains peaks of reads just after the 3' end of genes. While the exact biological reasons for this are unclear, it makes sense to associate these reads with the genes they are closest to. The third suggestion focuses on resolving overlaps between genes. Reads from such overlapping regions are often unassigned to any gene, but in some cases, it is more likely that they originate from one gene rather than another. Overlapping gene resolution aims to address this by deleting or shortening some genes in the transcriptomic reference.

Although Pool et al. (2023) proposed the tool for such tasks, the tool is not without limitations: some aspects of it are debatable (such as thresholds used), some seem unnecessarily (e.g. handling exon and intron sequences when most alligning tools provide option for this), and the process still requires a significant amount of manual work. Thus, there remains a need for a more comprehensive

tool for enhancing transcriptomic references, which will be addressed in this thesis.

4.6 Getting insights from scRNAseq data

4.6.1 Trajectory inference

The scRNAseq data is static snapshot due to a destructive nature of sequencing methods, which give raise to the challenges that could be only overcome with modelling approaches. Trajectory inference methods aim to reconstruct the dynamics of cellular processes of interest, such as development, differentiation or immune response (Deconinck et al. 2021). These inference methods assign a numerical value referred as pseudotime for each cell, and based on it, cells can be organized along the pseudotemporal axis and may recapitulate biological dynamical processes (L. Wang et al. 2021). Inference of pseudotime usually firstly reduces dimentionality of the data, and then applies either clustering or graph approaches for placing cells into the trajectory structures (Deconinck et al. 2021). The scRNAseq data contains both spliced and unspliced RNA transcrips, which provide additional temporal information. Based on it, there were proposed RNA velocity models, that aim to find the vectors predicting future state of individual cells (La Manno et al. 2018).

There are plenty of methods both for pseudotime ordering and RNA velocity, and before using them, one should take into account the assumptions and limitations of individual models. Such assumtions often include the type of trajectories (e.g. branching, linear etc.), systems state (e.g. steady state, dynamical) and others.

4.6.2 Inferring gene regulatory networks

Understannding regulatory relationships between genes is one of the main problems in system biology and medicine (Lamoline et al. 2024). High resolution scRNAseq data offers a chance to do this via inference of gene regulatory networks (GRNs). GRN is a graph representing relationships between transcription factors and genes they control, i.e. it is a graph where genes are represented by nodes and their relationships (activating or inhibiting) are represented by edges. Various mathematical concepts are used in GRN inference algorithms, including correlation, mutual information, regression, Baysean networks, boolean networks, differential equations and others (Akers and Murali 2021).

Even though there are plenty of inference models, there are no single methods that would be best in all situations. Moreover, performance of such algorithms often shows poor results, on global metrics similar to randomly creating GRNs. This was shown in the benchmarking study by McCalla et al. (2023). In the study 13 inference algorithms were compared, and none of them were best on different datasets and gold standards used. However, while not showing great performances on global metrics, methods were able to extract some useful local information (e.g. on some specific transcription factors that are major regulators in some cell lineages).

All in all, while there are plenty of inference algorithms, none of them are perfect, but can be used to extract some information about cellular systems.

4.6.3 Integrative approaches

The scRNAseq data alone does not capture all the relevant information of cellular system, therefore good inprovement of analysis is to incorporate other modalities of single cell data (Heumos et al. 2023). For example, CITE-seq allows to simultaneously measure gene expression and surface protein abundance (Mercatelli et al. 2021). Also, it is possible to capture both transcriptomic and epigenomic features of single cells (e.g. scM&T-seq (Angermueller et al. 2016) allows measure transcriptome together with DNA methylation). While such approaches supplies additional information about the cellular systems, they also give rise to additional chalenges when trying to integrate such data. Such challenges comes from high degrees of missing data, noise, and the scale of datasets, which can potentially span millions of cells (Argelaguet et al. 2020). There are plenty of tools designed for such data integration, including MOFA+, totalVI, WNN and multiVI (Heumos et al. 2023).

4.7 Current limitations and future perspectives

While scRNAseq data has attracted significant interest within the scientific community, and numerous tools and methods have been developed for its analysis, substantial challenges remain. Lähnemann et al. (2020) identified four major challenges in the field of scRNAseq data science: addressing data sparsity, defining flexible statistical frameworks for identifying complex differential patterns in gene expression, mapping single cells to reference atlases, and advancing trajectory inference.

The issue of data sparsity, briefly discussed in section 4.5.1, arises when algorithms rely solely on internal data, which can amplify signals artificially. This highlights the need for tools that integrate external information, such as reference atlases. Regarding differential analysis, although scRNAseq datasets capture more detailed information than bulk datasets, methods specifically tailored for single-cell data often do not outperform bulk methods (Soneson and Robinson 2017), indicating room for significant improvement.

The construction of atlases and reference mapping can reduce considerable manual work, and with the continuous growth in available data, the demand for such tools is increasing (Heumos et al. 2023). While some automatic annotation tools are available (Domínguez Conde et al. 2022), most focus on healthy samples, underscoring the need for reference atlases covering a wider range of states, diseases, and organisms (Heumos et al. 2023).

Most current trajectory inference methods are limited to scRNAseq data alone. Incorporating additional data modalities, such as epigenetics or proteomics, would enhance our understanding of dynamic cellular processes at a systems level (Lähnemann et al. 2020).

Two other critical topics in scRNAseq data science are the development of end-to-end pipelines and regular benchmarking (Heumos et al. 2023). The former is essential due to the rapid expansion of scRNAseq data, while the latter would facilitate the selection of appropriate tools—especially as there are now over 1,700 tools for scRNAseq data analysis (scRNA-tools 2024).

5. METHODS

5.1 Datasets

In this thesis following publically available datasets from scRNAseq experiments were used.

5.1.1 PBMC

This dataset is publically available in 10X Genomics website. In this experiment, human peripheral blood mononuclear cells (PBMCs) were extracted from fresh whole peripheral blood samples obtained from StemExpress. PBMCs were isolated using SepMate density centrifugation methods. The library was generated from around 8000 cells (5140 cells recovered) using the Chromium Single Cell 3' v3.1 Reagent Kit, and sequenced on Illumina NovaSeq 6000 to a read depth of approximately 35000 mean reads per cell. The transcript reads have length of 90bp. All this information (and more) is available at 10X Genomics website.

5.1.2 Transcriptomic references

5.2 Enhancing transcriptomic reference

Here is provided general description of the pipeline.

1. Map reads with transcriptomic reference.
2. Take unassigned (and unique) reads.
3. Split into intersecting and intergenic reads.
   (a) For intersecting:
      i. Cluster.
      ii. Filter-out relatively small clusters (custom threshold).
      iii. Make IGV snapshots.
      iv. Resolve overlapping genes that have some reads.
      v. From the second reference and further: add genes to the original GTF that contain reads and do not overlap with entries from the original.
   (b) For intergenic:
      i. Cluster.
      ii. Filter-out relatively small clusters (custom threshold).

iii. For the first reference only: filter-out AT-rich reads (clusters?).

iv. For reads that have been left unexplained, repeat from the beginning with the next reference.

v. For the last reference only: clusters that start just after 3' ends are assigned to genes (i.e., extend genes).

vi. For the last reference only: add largest intergenic unexplained regions to GTF (INTERGENIC entries).

4. Create final GTF and map initial sequences to it.

5. Compute statistics (reads mapped, genes captured).

6. Check clustering and other steps (in Jupyter notebooks).

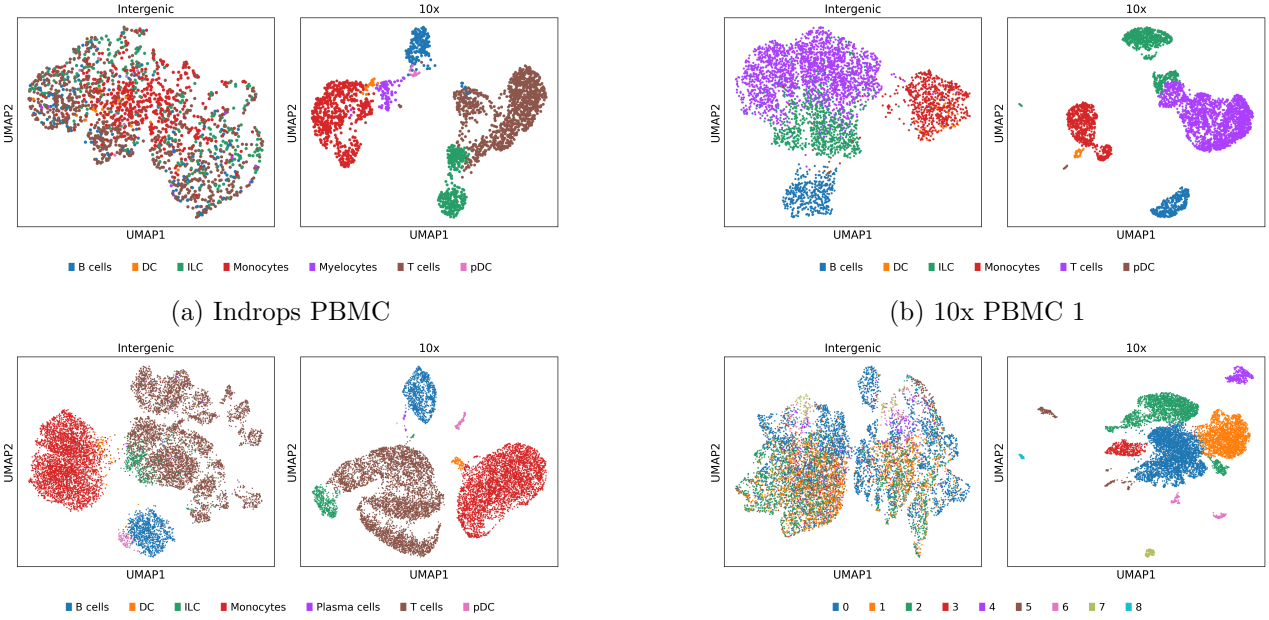
6. RESULTS

6.1 Enhancing transcriptomic reference

The enhanced transcriptomic reference allows to include those reads into downstream analysis that otherwise would be discarded. To achieve this, we have combine data from several different transcriptomic annotations, and additionally included in the annotation intergenic regions that contained relativelly high number of reads.

6.1.1 Intergenic regions

Observed intergenic regions can be either artefacts or be biologically meaningfull. To check this, I have tried to cluster cells based only on the newly defined intergenic regions (see figure 6.1). While for PBMC_indrops sample it looks as noise, for the 10x indrops samples it provides quite good clustering, meaning that at least some of those intergenic regions are not sequencing artefacts.



(a) Indrops PBMC

(b) 10x PBMC 1

(c) 10x PBMC 2

(d) 10x Brain

Figure 6.1: Comparison of clustering using only intergenic regions versus standard (10x) reference. 'Intergenic' plots are colored according to the '10x' coloring.

6.1.2 Enhanced Reference

Using enhanced reference allowed us to have more captured genes in the data, however, no significant change in clustering can be seen.

6.1.3 Captured genes

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) Akers, Kyle and T.M. Murali (June 2021). "Gene regulatory network inference in single-cell biology". In: *Current Opinion in Systems Biology* 26, pp. 87–97. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2021.04.007.

(3) 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.

(4) — (Mar. 2019). "False signals induced by single-cell imputation". In: *F1000Research* 7, p. 1740. ISSN: 2046-1402. DOI: 10.12688/f1000research.16613.2.

(5) Angermueller, Christof et al. (Jan. 2016). "Parallel single-cell sequencing links transcriptional and epigenetic heterogeneity". In: *Nature Methods* 13.3, pp. 229–232. ISSN: 1548-7105. DOI: 10.1038/nmeth.3728.

(6) Argelaguet, Ricard et al. (May 2020). "MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data". In: *Genome Biology* 21.1. ISSN: 1474-760X. DOI: 10.1186/s13059-020-02015-1.

(7) Arisdakessian, Cédric et al. (Oct. 2019). "DeepImpute: an accurate, fast, and scalable deep neural network method to impute single-cell RNA-seq data". In: *Genome Biology* 20.1. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1837-6.

(8) Azizi, Elham et al. (Jan. 2017). "Bayesian Inference for Single-cell Clustering and Imputing". In: *Genomics and Computational Biology* 3.1, p. 46. ISSN: 2365-7154. DOI: 10.18547/gcb.2017.vol3.iss1.e46.

(9) Chen, Chong et al. (Mar. 2020). "scRMD: imputation for single cell RNA-seq data via robust matrix decomposition". In: *Bioinformatics* 36.10. Ed. by Alfonso Valencia, pp. 3156–3161. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btaa139.

(10) Dai, Chichi et al. (May 2022). "scIMC: a platform for benchmarking comparison and visualization analysis of scRNA-seq data imputation methods". In: *Nucleic Acids Research* 50.9, pp. 4877–4899. ISSN: 1362-4962. DOI: 10.1093/nar/gkac317.

(11) Deconinck, Louise et al. (Sept. 2021). "Recent advances in trajectory inference from single-cell omics data". In: *Current Opinion in Systems Biology* 27, p. 100344. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2021.05.005.

(12) Dijk, David van et al. (July 2018). "Recovering Gene Interactions from Single-Cell Data Using Data Diffusion". In: *Cell* 174.3, 716–729.e27. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.05.061.

(13) Domínguez Conde, C. et al. (May 2022). "Cross-tissue immune cell analysis reveals tissue-specific features in humans". In: *Science* 376.6594. ISSN: 1095-9203. DOI: 10.1126/science.abl5197.

(14) Eraslan, Gökcen et al. (Jan. 2019). "Single-cell RNA-seq denoising using a deep count autoencoder". In: *Nature Communications* 10.1. ISSN: 2041-1723. DOI: 10.1038/s41467-018-07931-2.

(15) 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.

(16) Frankish, Adam et al. (Nov. 2022). "GENCODE: reference annotation for the human and mouse genomes in 2023". In: *Nucleic Acids Research* 51.D1, pp. D942–D949. ISSN: 1362-4962. DOI: 10.1093/nar/gkac1071.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) Huang, Mo et al. (June 2018). "SAVER: gene expression recovery for single-cell RNA sequencing". In: *Nature Methods* 15.7, pp. 539–542. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0033-z.

(23) 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.

(24) 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.

(25) 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.

(26) La Manno, Gioele et al. (Aug. 2018). "RNA velocity of single cells". In: *Nature* 560.7719, pp. 494–498. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0414-6.

(27) Lähnemann, David et al. (Feb. 2020). "Eleven grand challenges in single-cell data science". In: *Genome Biology* 21.1. ISSN: 1474-760X. DOI: 10.1186/s13059-020-1926-6.

(28) Lamoline, François et al. (May 2024). "Gene regulatory network inference from single-cell data using optimal transport". In: DOI: 10.1101/2024.05.24.595731.

(29) Linderman, George C. et al. (Jan. 2022). "Zero-preserving imputation of single-cell RNA-seq data". In: *Nature Communications* 13.1. ISSN: 2041-1723. DOI: 10.1038/s41467-021-27729-z.

(30) 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.

(31) 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.

(32) McCalla, Sunnie Grace et al. (Jan. 2023). "Identifying strengths and weaknesses of methods for computational network inference from single-cell RNA-seq data". In: *G3: Genes, Genomes, Genetics* 13.3. Ed. by L Steinmetz. ISSN: 2160-1836. DOI: 10.1093/g3journal/jkad004.

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

(34) Mercatelli, Daniele et al. (May 2021). "The Transcriptome of SH-SY5Y at Single-Cell Resolution: A CITE-Seq Data Analysis Workflow". In: *Methods and Protocols* 4.2, p. 28. ISSN: 2409-9279. DOI: 10.3390/mps4020028.

(35) Peng, Lihong et al. (Mar. 2020). "Single-cell RNA-seq clustering: datasets, models, and algorithms". In: *RNA Biology* 17.6, pp. 765–783. ISSN: 1555-8584. DOI: 10.1080/15476286.2020.1728961.

(36) 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.

(37) Pool, Allan-Hermann et al. (Sept. 2023). "Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references". In: *Nature Methods* 20.10, pp. 1506–1515. ISSN: 1548-7105. DOI: 10.1038/s41592-023-02003-w.

(38) 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.

(39) 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.

(40) scRNA-tools (2024). *scRNA-tools - scRNA-seq analysis tools database*. URL: https://www.scrna-tools.org/analysis (visited on 11/07/2024).

(41) 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.

(42) Soneson, Charlotte and Mark D. Robinson (May 2017). "Bias, robustness and scalability in differential expression analysis of single-cell RNA-seq data". In: DOI: 10.1101/143289.

(43) 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.

(44) Tang, Wenhao et al. (Oct. 2019). "bayNorm: Bayesian gene expression recovery, imputation and normalization for single-cell RNA-sequencing data". In: *Bioinformatics* 36.4. Ed. by Janet Kelso, pp. 1174–1181. ISSN: 1367-4811. DOI: 10.1093/bioinformatics/btz726.

(45) Vázquez-Jiménez, Aarón, Moisés Santillán, and Jesús Rodríguez-González (July 2017). "How the extrinsic noise in gene expression can be controlled?" In: *IFAC-PapersOnLine* 50.1, pp. 15092–15096. ISSN: 2405-8963. DOI: 10.1016/j.ifacol.2017.08.2236.

(46) Wagner, Florian, Dalia Barkley, and Itai Yanai (June 2019). "Accurate denoising of single-cell RNA-Seq data using unbiased principal component analysis". In: DOI: 10.1101/655365.

(47) Wagner, Florian, Yun Yan, and Itai Yanai (Nov. 2017). "K-nearest neighbor smoothing for high-throughput single-cell RNA-Seq data". In: DOI: 10.1101/217737.

(48) Wang, Lingfei et al. (June 2021). "Current progress and potential opportunities to infer single-cell developmental trajectory and cell fate". In: *Current Opinion in Systems Biology* 26, pp. 1–11. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2021.03.006.

(49) Wang, Mengyuan et al. (Oct. 2022). "Imputation Methods for scRNA Sequencing Data". In: *Applied Sciences* 12.20, p. 10684. ISSN: 2076-3417. DOI: 10.3390/app122010684.

(50) 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.

(51) 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.

(52) 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.

(53) Xu, Yungang et al. (June 2020). "scIGANs: single-cell RNA-seq imputation using generative adversarial networks". In: *Nucleic Acids Research* 48.15, e85–e85. ISSN: 1362-4962. DOI: 10.1093/nar/gkaa506.

(54) 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.

(55) 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.

(56) 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.

(57) 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.

(58) 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