ENHANCING SINGLE-CELL RNA-SEQ ANALYSIS THROUGH THE INTEGRATION OF UNACCOUNTED INTERGENIC REGIONS

Master Thesis

Systems biology master program

Vilnius university

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

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

Contents

1. LIST OF ABBREVIATIONS

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

2. INTRODUCTION

Single-cell RNA sequencing (scRNA-seq) is becoming an increasingly popular tool for analyzing cellular systems. This technology enables the sequencing of thousands of cells in a single experiment, generating a vast amount of data. The typical workflow for analyzing such data involves mapping reads to a known transcriptome and constructing cell-gene matrices, which are then used in downstream analyses. However, some reads are always mapped to the genome but remain unassigned to any known gene. Such reads are typically excluded from downstream analysis.

Since these unassigned reads can constitute a significant fraction of the total reads (often up to 30%), understanding the reasons behind them could help improve scRNA-seq technologies and data analysis.

There are two potential sources of unassigned reads: either they are sequencing artifacts, or they originate from real transcripts that remain unassigned due to issues related to transcriptomic references.

The two main challenges with transcriptomic references are:

1. Incomplete transcriptome annotations – even though there are given great efforts to annotate all genes, it is very likelly that not all genes are annotated, and many remains to be found. Such yet undefined genes are missed in the typical scRNAseq analysis.

2. Complexity and overlaps in transcriptomic features – the human (and many other species) transcriptome is very complex, with many overlapping features. This prevents mapping algorithms to assign some short reads to a single feature, usually resulting in discarding such reads from analysis.

This project focuses on these unassigned reads, aiming to uncover their origins and, if possible, enhance scRNA-seq data analysis by incorporating biologically meaningful data that is typically disregarded.

3. AIM AND TASKS

Aim The aim of this project is to investigate whether it is possible to improve scRNA-seq analysis by including unassigned reads in the downstream analysis.

Tasks

1. Identify genomic regions containing unassigned reads, and classiffy them as either intersecting with genes or intergenic.
2. For intergenic regions, check if there is evidence that would confirm them being noise or unannotated genes.
3. For intersecting regions, check if it possible to resolve problems related to the transcriptomic reference, which lead to reads being unassigned.

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, such as 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, Chen, and X. Peng 2024), exploring cell differentiation and developmental processes (Skinner, Asad, and Haque 2024), among many other applications.

4.2 scRNAseq data generation and analysis

To better understand this project, it is important to first understand how scRNA-seq and its data analysis work, which will be introduced in this section

4.2.1 Overview of scRNAseq protocols

The generation of scRNA-seq data is a complex, multi-step process that varies across different protocols. These protocols can be grouped based on several criteria, such as the type of RNA capture (e.g., 3' end, 5' end, or full-length) or the method of cell isolation (e.g., droplet-based methods such as inDrops (Klein et al. 2015), Drop-seq (Macosko et al. 2015), and Chromium by 10X Genomics (Zheng et al. 2017), or plate-based methods such as CEL-Seq2 (Hashimshony et al. 2016) and Smart-seq2 (Picelli et al. 2013)).

In this project, datasets generated using droplet-based 3' end sequencing methods were analyzed. Therefore, these methods will be the focus of the following literature review.

3' End Sequencing and Polyadenylation 3' end sequencing captures RNA molecules using primers complementary to poly(A) tails. Polyadenylation at the 3' end is a post-transcriptional modification in which non-templated adenosines are added to the 3' end of mRNA molecules. Although the poly(A) tail is present in almost all mRNAs, its length varies and can influence mRNA fate, stability, and translation efficiency (Brouze et al. 2022).

3' end sequencing protocols exploit this feature to selectively capture RNA molecules, enabling high-throughput and cost-effective sequencing. However, some RNA molecules lack poly(A) tails and are therefore not captured by these methods, such as replication-dependent histone mRNAs (Brouze et al. 2022).

Common Steps in scRNA-seq Protocols Despite differences between scRNA-seq protocols, they share key steps: single-cell isolation, library preparation, and sequencing (Andrews and Hemberg 2018). In droplet-based methods, single cells are encapsulated in individual droplets containing hydrogel primers and a lysis mix (an example of a droplet generation device is shown in Figure 4.3). Primers used in these protocols typically share a common structure, including:

- **Cell barcodes:** Unique sequences that identify the cell from which a particular read originates (since all droplets are pooled and sequenced together).
- **Unique molecular identifiers (UMIs):** Short sequences used to quantify the original number of RNA molecules, helping to eliminate amplification bias.
- **PCR handles:** Sequences that facilitate amplification.
- **Poly-T sequences**: In 3' end sequencing methods are used to selectively capture polyadenylated RNAs (X. Zhang et al. 2019).

An example of primer design is shown in Figure 4.1. Once cells are encapsulated in droplets, lysis occurs, releasing RNA, which is then captured by the primers. A schematic overview of library preparation is provided in Figure 4.2.

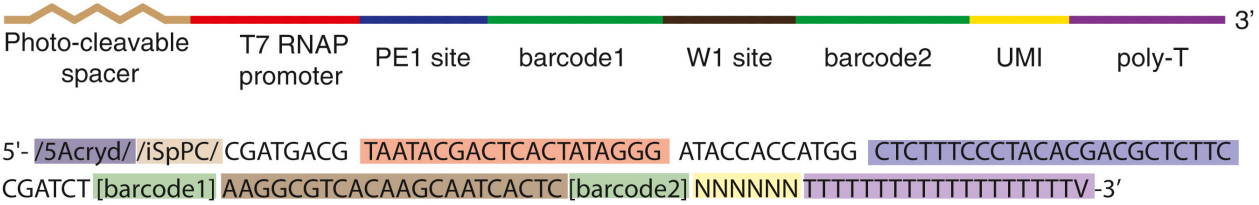


Figure 4.1: Example of primer design (inDrops). The image above shows schematic view, while below is given example with sequences. The geometry of primers varies between the protocols, the main parts (UMI site, barcodes, poly-T's) are found in most of them. Here also present are promoter region for RNA polymerase (red), sequencing primer (blue), synthesis adaptor (dark brown). Figure taken from Klein et al. (2015).

Depending on the method, reverse transcription may occur within the droplets (as in inDrops and 10X Genomics) or after demulsification (as in Drop-seq). Subsequent steps typically include RNA fragmentation, PCR amplification, and next-generation sequencing (NGS).

4.2.2 scRNAseq data analysis

Raw data processing The output from NGS is typically 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 4.4). 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 other. 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).

Cell-gene matrix processing The next steps in the analysis of the scRNAseq data involves following steps:

- **Quality control** The quality of indidual cells (barcodes) 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

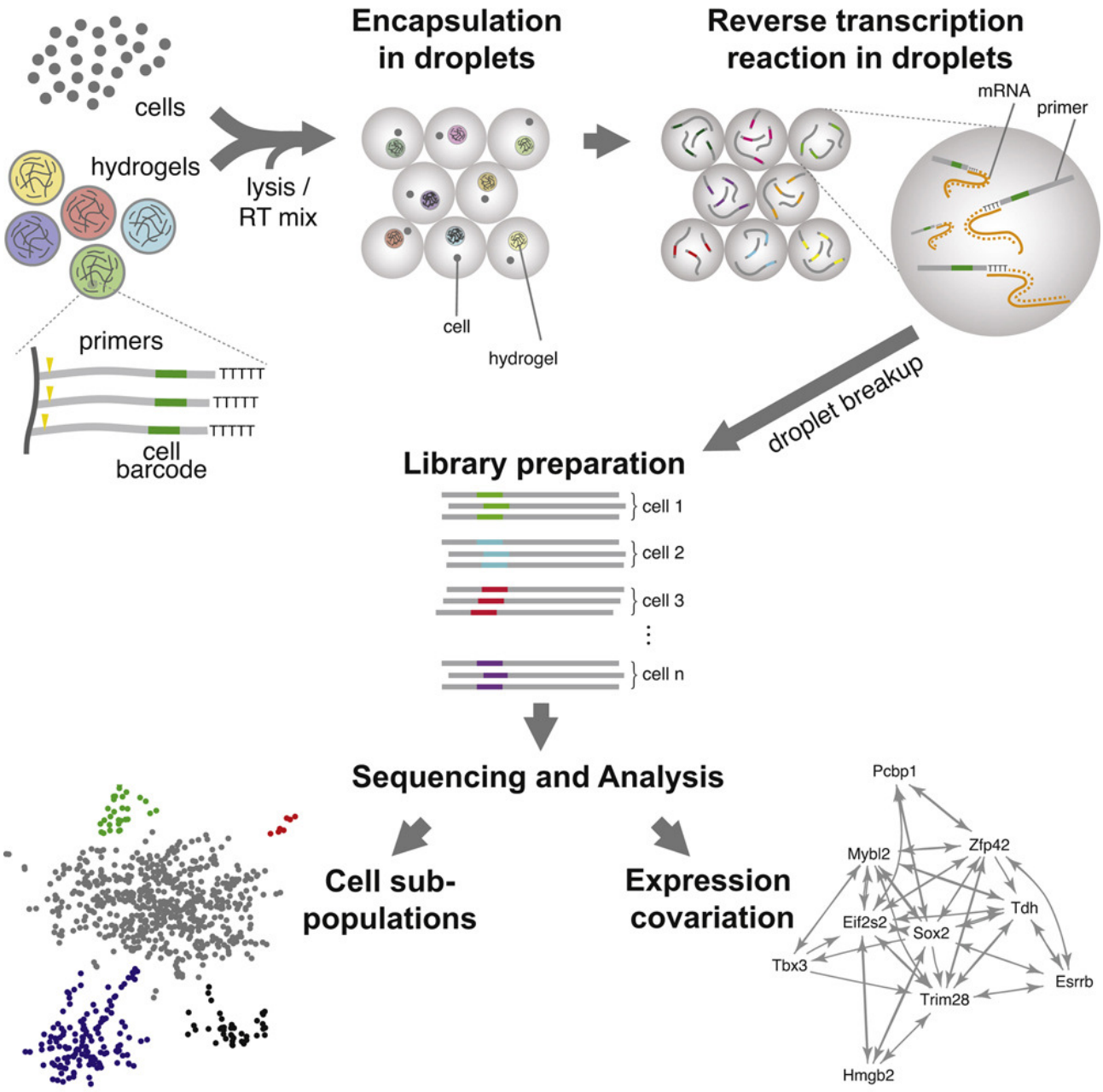


Figure 4.2: The schematic presentation of droplet-based scRNAseq particularly inDrops, however main steps are shared between most of the droplet based protocols. Figure taken from Klein et al. (2015).

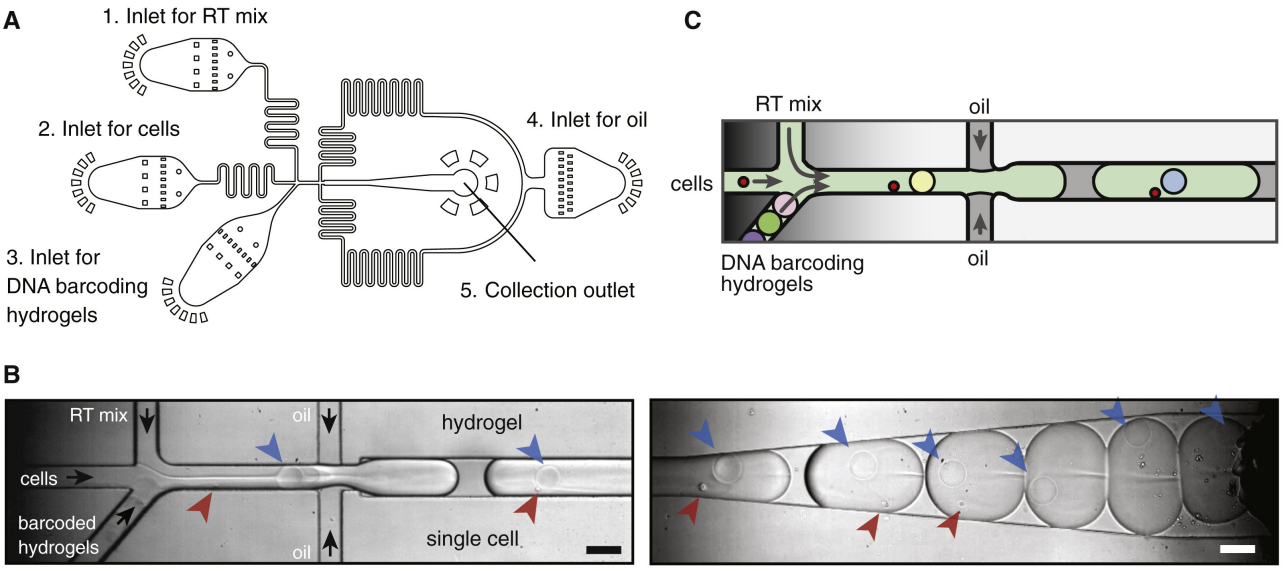


Figure 4.3: An example of microfluidics device for droplet generation (inDrops). A) Schematic view of the device. B) Snapshots of droplet generation and colection. C) Scheme of droplet generation. Figure taken from Klein et al. (2015).

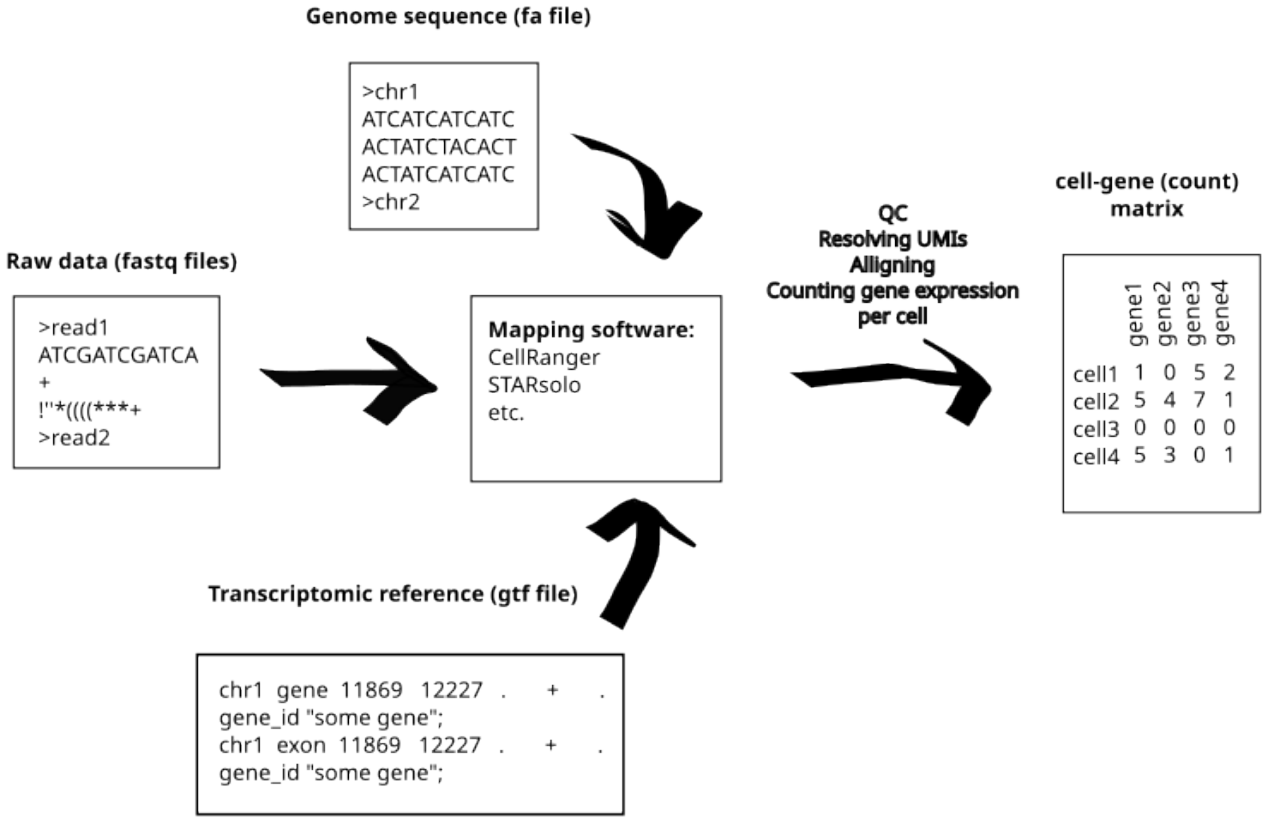


Figure 4.4: Inputs and outputs of mapping software.

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.

- **Normalization** 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.

- **Filtering genes** 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.

- **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.

- **Clustering and other analyses** One of the most common 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.

The described analysis pipeline enables insights into multicellular systems. It is evident that all downstream analyses are directly influenced by the initial steps of raw data processing, with mapping being particularly crucial in the context of this project. While mapping algorithms and tools differ, they all rely on the genome and its annotation (also refered as 'transcriptomic reference'), which directly impact their results. In the next section, I will further expand on this topic.

4.3 Genome annotation

The genome annotation process typically refers to the identification and mapping of genes within a given genome sequence (Guigó 2023). While the definition itself is straightforward, the process is highly complex. This is evident from the fact that, even more than 20 years after the first human genome assembly, human genome annotations are continuously updated with new transcripts and are expected to evolve further (Mudge et al. 2024). Figure 4.5 illustrates the changes in GENCODE annotation over time.

It is important to note that many medical and scientific research efforts rely on an accurate human gene list. Examples include genome-wide association studies (GWAS), which attempt to link genomic variants to nearby genes; RNA-seq analysis; and exome sequencing projects that use capture kits targeting most known exons (Pertea et al. 2018).

Currently, the two most widely used genome annotations are GENCODE (Mudge et al. 2024) and RefSeq (O'Leary et al. 2015) (maintained by NCBI). Despite their status as mature genome references, they report different numbers of genes. For instance, RefSeq includes 20,078 protein-coding genes (NCBI 2025), whereas GENCODE contains 19,868 (Ensembl 2025). This discrepancy highlights the ongoing challenge of accurately annotating genomes. To better understand these difficulties, the genome annotation process is first reviewed.

Genome annotation proccess RNA sequencing (RNA-seq) is the primary tool used in genome annotation. Full-length RNA sequencing allows for the capture of RNA molecules, which, when

aligned to a reference genome, help construct the genome annotation of a given species (Salzberg 2019). However, RNA-seq has limitations, the most significant being its inability to capture all RNA molecules. This poses particular challenges for detecting rare transcripts, which may either be treated as noise or not captured at all (Salzberg 2019). Therefore, bioinformatics tools are necessary to complement RNA-seq data (Guigó 2023), and most modern genome annotation pipelines integrate computational methods alongside sequencing data.

Computational genome annotation methods can be broadly categorized into two types: comparative annotation, which leverages the fact that protein-coding sequences tend to be more evolutionarily conserved, and *ab initio* annotation, which uses known sequence biases to predict genes (Guigó 2023). Despite significant advancements, genome annotation remains imperfect.

While the number of protein-coding genes is reaching a consensus — major databases report around 19,000 to 20,000 protein-coding genes — the number of non-coding genes is expected to increase in the future (Amaral et al. 2023). This is largely due to the structured nature of protein-coding genes (e.g., open reading frames, codon biases that skew nucleotide distributions), which makes them easier to detect using computational tools (Guigó 2023). Additionally, protein-coding genes have historically received greater attention because of their direct links to phenotype, leading to more experimental validation.

Current challenges in human genome annotation Despite extensive efforts, a fully comprehensive human genome annotation has not yet been achieved. Several factors contribute to this challenge.

First, the complexity of the human genome itself presents significant obstacles. Compared to the total genome length, the number of genes is relatively low, and their sequences are interrupted by introns (Salzberg 2019). Additionally, bulk RNA sequencing often has relatively low sequencing depth, making it difficult to detect rare transcripts or those expressed in a cell-type-specific manner (Guigó 2023).

In principle, full-length single-cell RNA sequencing (scRNA-seq) could help overcome some of these limitations. However, it still has inherent biases introduced during library preparation, including those related to RNA processing status, post-transcriptional modifications, transcript length, cellular localization, and structural features (Guigó 2023). Short-read scRNA-seq typically provides higher throughput (Heumos et al. 2023), enabling the detection of rare transcripts. However, it does not capture full transcript structures, limiting its utility to supporting evidence, such as validating computationally predicted genes or indicating transcriptional activity at specific genomic locations.

Beyond technical limitations, there are ontological challenges, such as defining what constitutes a gene. While genes are often regarded as well-defined, discrete entities, the reality is more complex. Coding and non-coding transcripts frequently overlap in intricate arrangements with unclear boundaries, suggesting that transcripts may not be discrete, countable units but instead form a transcriptional continuum (Salzberg 2019). Additionally, the classification systems used in genome annotations may be partially artificial. For example, some pseudogenes, generally considered non-functional copies of functional genes, are transcribed and may have biological functions (Pei et al. 2012). Similarly, the distinction between protein-coding and non-coding genes is debated, as many protein-coding loci generate both coding and non-coding transcripts, and numerous long non-coding RNAs (lncRNAs) contain potentially coding open reading frames (ORFs) (Salzberg 2019).

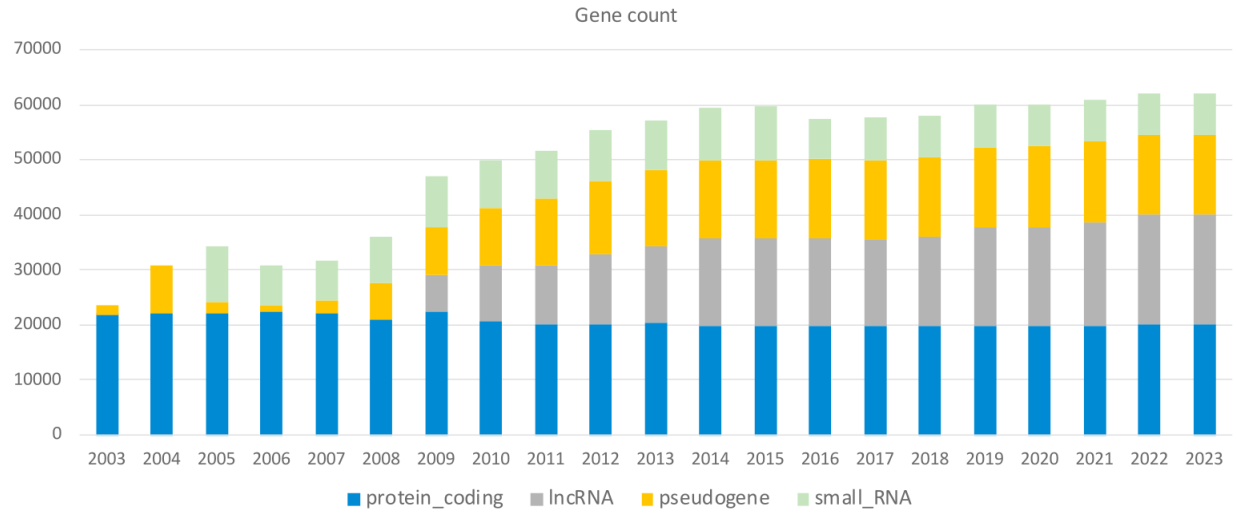


Figure 4.5: Number of different genes in the gencode annotation during time. Figure taken from paper by Guigó (2023)

Future perspectives of annotating genomes The size of eukaryotic genomes necessitates the use of automated methods for genome annotation. The accuracy of such methods depends directly on RNA capture technologies. If highly accurate and sensitive methods are developed, high-quality genome annotations can be expected (Salzberg 2019). However, at present, manual curation of human genome annotations remains essential due to the imperfections of both sequencing technologies and computational tools. Given the critical role of accurate gene annotation in both medical and scientific applications, continued improvements in annotation methods remain a priority.

4.4 Transcriptomic references for scRNAseq

Having a well-annotated genome does not solve all the challenges encountered during the mapping step. In simple terms, mapping algorithms determine the genomic location to which a sequence aligns, and if an annotated gene is present at that location, the read is assigned to that gene. However, there are cases where assigning a read to a gene is not straightforward, such as when a read maps to multiple locations or to a unique location where overlapping genes are annotated.

Multimappers Reads with ambiguous origins (commonly referred to as multimappers) are typically excluded from analysis (Almeida da Paz, Warger, and Taher 2024). However, as demonstrated by (Almeida da Paz, Warger, and Taher (2024)), this approach can introduce biases that affect downstream analyses. Currently, various methods exist to handle such cases (Deschamps-Francoeur, Simoneau, and Scott 2020), yet no entirely satisfactory solution has been established (Almeida da Paz, Warger, and Taher 2024).

The aforementioned methods primarily rely on computational strategies to resolve multimappers, but the issue can also be approached from a transcriptomic reference perspective. For example, in the case of CellRanger (the mapping software from 10x Genomics) the GENCODE reference is filtered based on gene types (e.g., retaining protein-coding genes while filtering out pseudogenes) (10x

Genomics 2025). While this filtering improves the number of reads included in downstream analyses, there is still room for improvement.

Pool et al. (2023) proposed three key steps to enhance transcriptomic references:

1. Including reads mapped to intronic sequences in the analysis.
2. Extending the 3' ends of certain genes.
3. Resolving overlaps between specific genes.

The first suggestion is not new in scRNA-seq research. Concepts such as RNA velocity, which rely on the ratio of spliced to unspliced RNA (La Manno et al. 2018), demonstrate that including intronic reads can provide valuable information. Moreover, most mapping tools (e.g., STARsolo, CellRanger) offer options to align reads either to exonic regions only or to entire genes. The second suggestion is based on the observation that scRNA-seq data often exhibits peaks of reads just beyond the 3' ends of genes. While the exact biological reasons remain unclear — possibly due to imprecise annotations — it is reasonable to associate these reads with the nearest genes. The third suggestion addresses gene overlaps. Reads originating from overlapping regions are often unassigned to any gene, yet in some cases, they are more likely to come from one gene rather than another. Overlapping gene resolution seeks to correct this by modifying the transcriptomic reference, either by shortening or removing certain genes.

Although Pool et al. (2023) proposed a tool to implement these strategies, it has limitations. Some of its aspects remain debatable (e.g., threshold choices), others seem unnecessary (such as handling exon-intron distinctions when most alignment tools already provide this option), and the process still requires significant manual effort. Thus, there is still a need for a more comprehensive tool for enhancing transcriptomic references — a need that will be addressed in this thesis.

Reads mapped to the intergenic regions Some reads map to the genome but not to any annotated gene. While these could simply be noise in scRNA-seq data, there is also the possibility that they contain biologically relevant information. As will be shown in the results section, scRNA-seq datasets indeed contain such reads, and they are not entirely noise.

5. METHODS

5.1 Data Acquisition

scRNAseq datasets We analyzed datasets from four different tissues: brain, blood, lung, and eye. All samples, except for two Peripheral Blood Mononuclear Cell (PBMC) datasets, were generated using the 10x Genomics v3.1 protocol. The two exceptions were prepared using the inDrops2 and inDrops protocols. All protocols capture short reads from the 3' ends of RNA molecules. The datasets are publicly available, with sources and additional details provided in Table 5.1.

sample name	tissue	cell count	donors	protocol	source
PBMC_10x	blood (PBMC)	5000	1	10x v3.1	10x genomics
PBMC_10x_2	blood (PBMC)	10000	1	10x v3.1	10x genomics
PBMC_10x_3	blood (PBMC)	10000	1	10x v3.1	10x genomics
PBMC_indrops	blood (PBMC)	2000	1	indrops2	-
PBMC_indrops_2	blood (PBMC)	9000	1	indrops(?)	-
brain	brain	6000	1	10x v3.1	Siletti et al. 2023
brain_2	brain	7000	1	10x v3.1	Siletti et al. 2023
eye	retina	10000	6	10x v3.1	Menon et al. 2019
eye_2	peripheral retina	2500	1	10x v3.1	Voigt et al. 2019
eye_3	peripheral retina	2500	1	10x v3.1	Voigt et al. 2019
lung_2	lung	5000	1	10x v3.1	Mould et al. 2021
lung_5	lung	5000	1	10x v3.1	Mould et al. 2021
lung_7	lung	4500	1	10x v3.1	Mould et al. 2021
lung_8	lung	5000	1	10x v3.1	Mould et al. 2021

Table 5.1: Datasets summary.

Genome and Transcriptomic references The human genome GRCh38 was used in this project, downloaded from the Ensembl website.

Four transcriptomic references were analyzed (see Table 5.2 for details). To ensure compatibility with the genome FASTA file, chromosome name prefixes ('chr') were removed from references that included them, as the genome file does not use these prefixes. This modification was performed using basic Linux command-line tools. Additionally, for the NCBI reference, chromosome names were converted to Ensembl-style notation using a custom Python script (e.g., 1, 2, 3 instead of NC_000001.11, NC_000002.12, NC_000003.12, etc.).

For references that do not have 'gene' entries, such entries were added (entry that spans all the components of a particular gene).

reference name	source	version
10x	10x website	2024-A
GENCODE	Gencode website	47
RefSec (NCBI)	NCBI website	GCF_000001405.40
lnc	LNCipedia website	5.2

Table 5.2: References used in this project.

Gene Prediction Tracks and Conservation Scores Gene predictions (tracks 'AUGUSTUS', 'Geneid genes', 'Gescan genes', 'SGP genes', 'SIB genes') and genome conservation scores ('phastCons100way' track) were downloaded from UCSC genome browser.

ATAC data The open chromatin regions (bed format) in PBMCs was downloaded from 10x website.

5.2 Computational Tools and Environment

All analyses were performed on a high-performance computing (HPC) cluster running a Linux environment. Software packages and tools were managed using Conda. The exact Conda environment specifications (YAML file) can be found on GitHub repository. Besides tools managed by conda, also basic command line tools were used (e.g. *awk, wc, grep, sed, uniq, sort* and similar). The specific functions and parameters used are described in the following sections, where the general analysis pipeline will be explained.

5.3 Data processing pipeline

The full scripts are available on the GitHub, here only general description of the workflow and used tools provided. Below is provided general description of the pipeline, used both for extracting intergenic regions and enhancing trnascriptomic reference:

1. Map reads with initial transcriptomic reference.
2. Take unassigned (and uniquely mapped) reads.
3. Split into intersecting and intergenic reads.
   (a) For intersecting:
      i. Resolve overlapping genes that have unassigned reads (if possible).
      ii. From the second reference and further: add genes to the original GTF that contain unassigned reads and do not overlap with entries from the original one.
   (b) For intergenic:
      i. Cluster.
      ii. Filter-out relatively small clusters (custom threshold).
      iii. For the first reference only: filter-out AT-rich reads.
      iv. For reads that are left, 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. Create list of large (>5CPM) intergenic clusters.

Mapping and filtering bam files Reads were mapped using STARsolo (Kaminow, Yunusov, and Dobin 2021), both in the case of mapping from fastq and bam input files. Parameters were used the same for all samples (see GitHub repository), except the ones regarding barcode geometries. Filtering unassigned reads was done using *awk*, taking those that have valid barcode (i.e. filtering out those reads that have barcode length not equal to the defined length) and were not assigned to any gene (i.e. had 'GN:Z:-' tag). Also, only uniquelly mapped reads were taken (tag 'NH:i:1').

Classifying reads as 'intergenic' or 'intersecting' and reads clustering Intersections of unassigned reads with references were checked using *bedtools intersect* command. Those that intersect with genes were classified as 'intersecting', and those that do not – as 'intergenic'. Intersections were check in strand-specific manner (bedtools '-s' flag). The intergenic reads were clustered using *bedtools merge*, again in a strand-specific manner.

Manipulating transcriptomic references The overlapping gene resolving and construction of enhanced transcriptomic reference was done using custom R script, particularly *rtracklayer* library. The criterions for the resolving of overlapping genes are following:

1. **Gene type**: prefer protein coding genes over other types, lncRNA over remaining (e.g. pseudogenes).

2. **Level**: some annotations have 'level' field, indicating if the the is verified (score 1), manually annotated (score 2) or automatically annotated (score 3). Lower 'level' score was preffered.

3. **Intersection types**: if 5' end gene is overlapping with 3' end of gene, 5' end of gene was shortened, as data we were using is generated using 3' end method, suggesting that reads in the 5' end regions of genes were not expected.

This is rough description of the usage of various tools in the pipeline given above, the pipeline itself was implemented using GNU MAKE.

5.4 Intergenic regions

Extracting intergenic regions The intergenic reads were extracted as described in the previous section (i.e. taked unassigned reads that do not overlap with any reference used) and merged into clusters using *bedtools merge* command. Clusters were filtered based on cluster size, to include at least 5 reads per million (CPM) (computed from the totol number of primary reads in the datasets). The filtering was done using *awk*.

Cluster locations were adjusted using deeptools (to compute coverage) and custom python script, to avoid reads containing long introns making the intergenic clusters very wide. To accomplish this, for each intergenic region the maximum coverage location was found and extended to include neighbouring regions that had at least $max\_value/2$ coverage. In such way, intergenic regions were adjusted to cover only 'peaks' of reads.

The lists acquired from all samples were then merged using *bedtools merge* function. This combined filter then was filtered to contain only regions detected in sufficient number of samples (i.e. in all 'eye', 'lung', 'brain', 'PBMC_10x' or 'PBMC_indrops' samples). Additionally, for each entry it was determined whether it overlaps with predicted genes from UCSC gene prediction archive (using *bedtools merge*) and distances to the closest genes were found (*bedtools closest*).

This filtered combined list were then converted into GTF format and unassigned reads from each sample were mapped using this combined intergenic annotation.

Analysis of cell-gene matrices The matrices produced by STAR were filtered, allowing cells that have sufficient number of reads (thresholds selected manually), and only genes that were expressed in at least 3 cells. Also cells were filtered based on mitochondrial gene count (allowing up to 10% of mitochondrial gene in a cell). Doublets were filtered using *scrublet*. Afterwards, matrices were normalized (using *normalize_total* function from *scanpy* package) and log-transformaed (i.e. for each entry $x = log(x + 1)$, *log1p* function from *scanpy*).

Then only highly-variable genes (min_mean=0.0125, max_mean=3, min_disp=0.5) were selected. For the clustering, principal components (PCs) were computed and optimal number of them were selected based on elbow rule (manually). For blood samples, cells were annotated automatically using CellTypist (for other samples, annotation was skipped, as it was not in the main focus of this project). Then visualizations were made using UMAP embeddings (functions from the same *scanpy* package).

All the scripts with descriptions can be found in the GitHub repository.

Correlations Spearman's rank correlation between intergenic region and gene on the oposite strand was checked using custom python script. Formula for Spearman's correlation is given by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is the difference between the ranks of corresponding values and n is the number of observations.

6. RESULTS

6.1 Intergenic regions

Fourteen datasets from various tissues (blood, brain, eye, and lung) were analyzed. These datasets contained between 11% and 29% unassigned uniquely mapped reads (as shown in Table 6.1) when using the 10x reference. Additionally, 6% to 18% of total reads were intergenic, meaning they did not overlap with any reference used. Given the size of the datasets, such a proportion of reads could potentially contain biologically relevant information.

Sample	Total Reads	Unassigned	Intergenic	Demultiplexed	Unassigned	Intergenic
PBMC_10x	182330834	15.95%	11.37%	70305137	18.51%	12.86%
PBMC_10x_2	496387931	18.52%	14.26%	195948926	21.29%	16.32%
PBMC_10x_3	368640939	18.59%	14.34%	167804193	20.99%	16.17%
PBMC_indrops	112932507	11.12%	6.5%	8176578	22.86%	10.25%
PBMC_indrops_2	471705924	14.87%	8.36%	98063027	25.52%	13.73%
brain	206360627	16.87%	10.67%	143916177	19.02%	12.07%
brain_2	122556503	22.32%	15.98%	95031864	23.77%	17.00%
eye	375397270	28.02%	18.97%	161882445	31.87%	21.60%
eye_2	140981808	29.35%	14.85%	69146995	31.43%	15.99%
eye_3	161261977	29.46%	18.86%	68157696	32.11%	20.58%
lung_2	511080104	15.99%	12.07%	269118747	17.98%	13.33%
lung_5	452105505	14.49%	10.74%	256353198	16.19%	11.83%
lung_7	524095146	18.41%	14.34%	236157103	20.64%	15.79%
lung_8	342092138	14.53%	10.66%	217801055	16.07%	11.67%

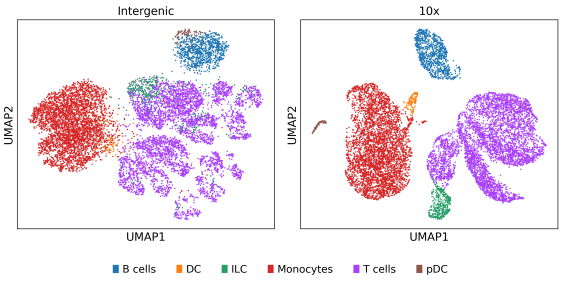
Table 6.1: Statistics of unassigned and intergenic reads per sample. Unassigned reads were counted after mapping with 10x reference. Intergenic reads here are those that do not intersect with any reference used. 'Demultiplexed' column shows number of total reads after demultiplexing using UMIs, and following 'unassigned' and 'intergenic' percentages were computed compared to this number.

Analysing integenic regions For each sample, intergenic regions containing a sufficient number of unassigned reads were identified, as described in the methods section (with a threshold set at 5 CPM). The term "intergenic" is used here in a strand-specific manner, meaning an "intergenic" region may be located on the opposite strand of a known gene.

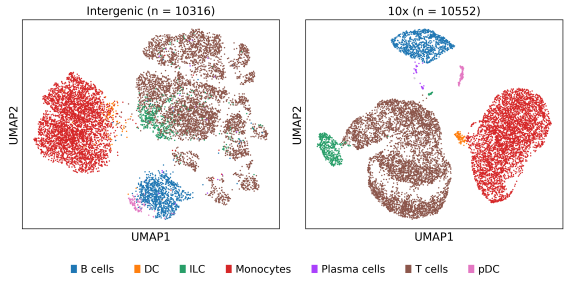
To determine whether these intergenic regions contain biologically meaningful information, cells in each sample were clustered based solely on reads from these regions. As shown in selected examples in Figure 6.1, clustering was observed and roughly corresponded to clustering based on the standard ("10x") annotation. This indicates that biologically meaningful information is indeed present in these

intergenic regions. Consequently, further analyses were performed on these regions.

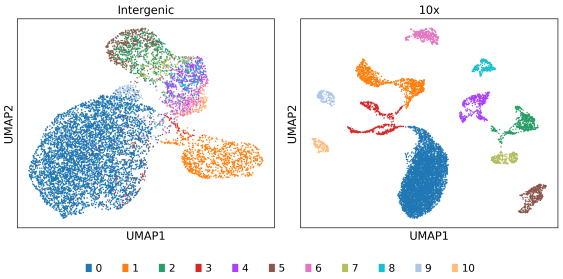
The lists of intergenic regions from all samples were combined and filtered based on the number of samples in which they were detected, resulting in a final set of 2,590 intergenic regions. Of these, 147 did not overlap known genes on the opposite strand and will be referred to as "isolated". Those located on the opposite strand of known genes will be referred to as "antisense".



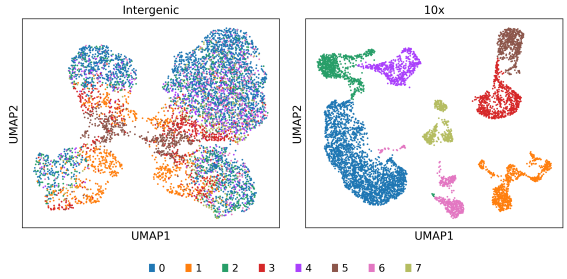
(a) 10x PBMC 3



(b) 10x PBMC 2



(c) eye



(d) 10x Brain 2

Figure 6.1: Comparison of clustering using standard annotation ('10x' reference) and using only defined intergenic regions. As can be seen, for some samples intergenic regions are sufficient for rough clustering, for others (e.g. 'brain_2' sample), noise gives some clustering artefacts. Nevertheless, this rough clustering implies that these unassigned read contain biological information.

A-rich sequences near the intergenic regions 3' end If there is poly-A region near the 3' end of the intergenic region, then the presence of those peaks can be explained by mis-priming, i.e. that would suggest that our intergenic region is not necessarily the actual 3' end of the novel transcript, but may be in the middle of it.

This is indeed the case for the most of reported intergenic regions, see Figure 6.2 for histogram and Table 6.2 for some statistics.

	isolated	antisense	total
mean distance	54.8768	52.3499	52.4867
mean A-rich region length	14.8333	16.755	16.651

Table 6.2: Summary of stats.

Correlations Another aspect to check is whether expression of those intergenic regions correlate with the expression of nearby genes. If it does, it might indicate several things:

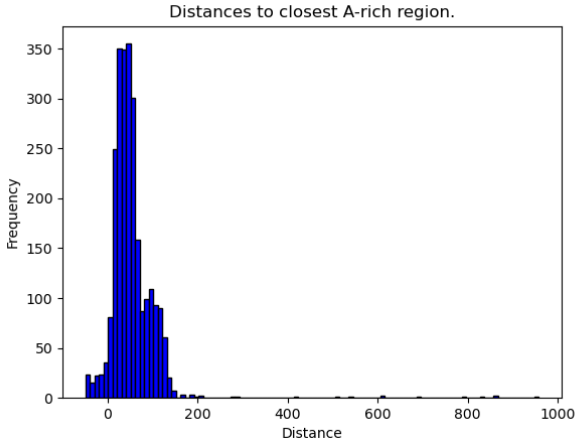


Figure 6.2: Distances from 3' ends to closest A-rich region downstream. From all the intergenic regions, only 40 do not have such region within 1kb from the 3' end.

- If the region correlates with gene on the same strand:
  - There might exist longer unannotated transcripts that involve those intergenic regions.
  - The intergenic genes are involved in the same processes as the gene.
  - They both can be transcribed by the same polymerases.
- If the region correlates with gene on the oposite strand:
  - The intergenic gene is involved in the same processes as the gene on the oposite strand.
  - There happens some transcription errors that cause transcription from the opposite strand.
  - There happens some errors in the library preparation/sequencing steps that cause template switch.

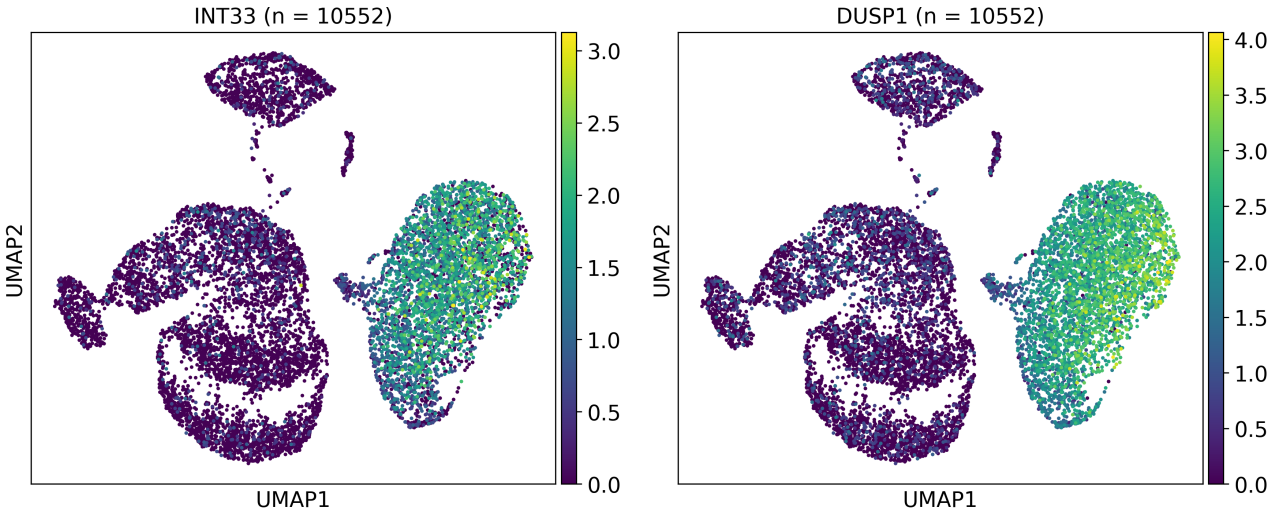
Which of those are the case in our samples, is hard to tell. Out of all defined intergenic regions, 97 showed Spearman correlation >0.5 ($p < 0.05$) in at least 1 sample (83 correlated with genes on the opposite strand, 19 with gene on the same strand, 5 with both). Some examples can be seen in Figure 6.3.

In the following subsections isolated and antisense intergenic regions will be discussed separatelly.

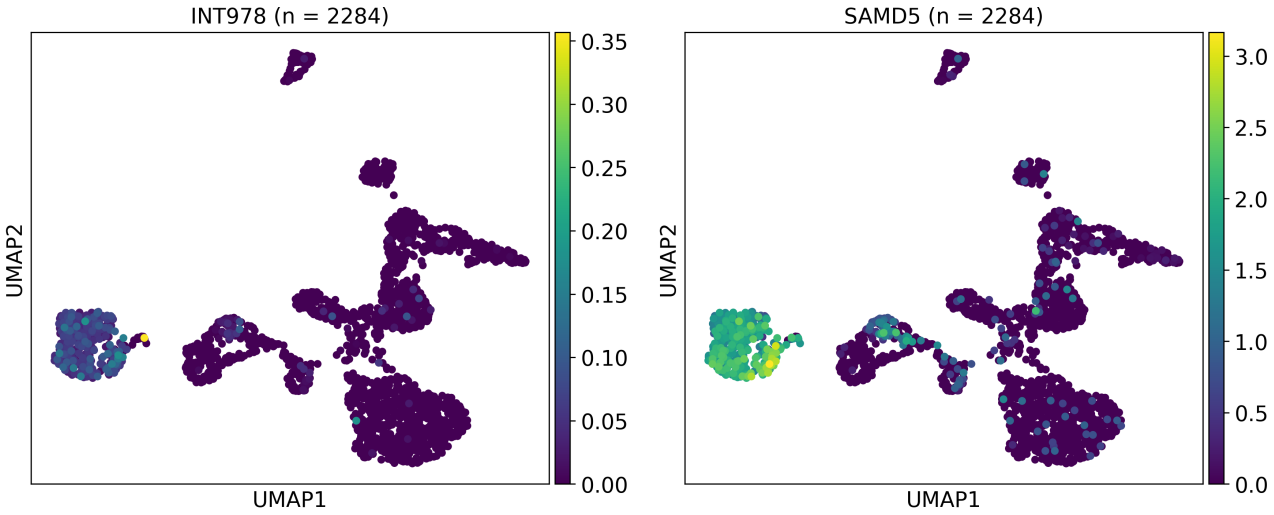
6.1.1 Isolated intergenic regions

The intergenic regions that are distant from known genes and contain reads may indicate novel genes. To determine whether these are genuine signals rather than artifacts, several aspects were considered:

1. Differential expression: Are these regions differentially expressed in any samples, i.e., are they specific to certain cell types? If so, this strongly suggests a biological origin rather than an artifact.
2. Open chromatin: Are open chromatin regions present upstream? Such regions often indicate active transcription.



(a) Umaps of INT33 and DUSP1 normalized expressions (Spearman correlation 0.68) of the PBMC_10x_2 sample, both of them are on the same strand.



(b) Umaps of INT978 and SAMD5 normalized expressions (Spearman correlation 0.85) of the eye_3 sample, intergenic region is on the opposite strand of the gene.

Figure 6.3: Correlation can be seen visually in UMAPs (colored by the normalized expression of the genes), here couple of examples given.

3. Conservation score: Coding regions tend to be more evolutionarily conserved than non-coding regions.

4. Gene predictions: Do these regions overlap with predicted genes identified by computational tools?

Differential expression analysis To assess differential expression, all data samples were first clustered. For *PBMC_10x* samples, clustering and cell annotation were performed using CellTypist. For other samples, the Leiden algorithm was used, with manually chosen parameters to ensure a comparable number of clusters across similar datasets. For instance, all *lung* samples were clustered into five groups, aligning approximately with visible UMAP structures. The UMAP visualizations, colored by clusters, can be found in Appendix 14.1.

After filtering those regions that were differentially expressed (tresholds were used 0.01 for adjusted p-value and 0.5 for 'logfoldchange'), the 29 differentially expressed isolated intergenic regions were found. Some examples can be seen in Figure 6.4.

Open chromatin Open chromatin regions may provide further evidence of transcriptional activity when located upstream. However, distances to the nearest known genes must also be considered, as upstream open chromatin sites may not be associated with the intergenic regions but rather with other genes.

For each intergenic region, distances to the nearest open chromatin sites and upstream genes (strand-unspecific) were calculated. Of the 147 isolated intergenic regions, 66 had open chromatin regions closer than any known upstream genes. To ensure that these sites were not merely associated with distant genes, an additional filter was applied, requiring that the closest open chromatin region be at least 2,000 bp from the nearest gene. This yielded 40 intergenic regions.

It should be noted, that PBMC ATAC data was used for open chromatine regions, thus this feature should be taken with precaution for regions derived from other datasets.

Conservation scores A high conservation score in an intergenic region suggests potential functionality, providing additional evidence for biological relevance. To identify such regions, those with conservation scores above 0.6 that did not overlap with known genes from references used were selected. The resulting regions are presented in Table 6.3.

Name	Genomic coordinates	Conservation score	Found in samples
INT1216	1:72282700-72282900	0.7534172617	brain, eye
INT1829	1:77772950-77773050	0.6863111111	lung
INT2199	6:148122550-148122750	0.6364459652	brain, eye
INT2208	5:18162950-18163100	0.6138651685	PBMC (indrops)
INT3636	X:129412750-129413050	0.7218984354	brain

Table 6.3: Conserved intergenic regions not overlapping with known genes from ncbi and gencode annotations.

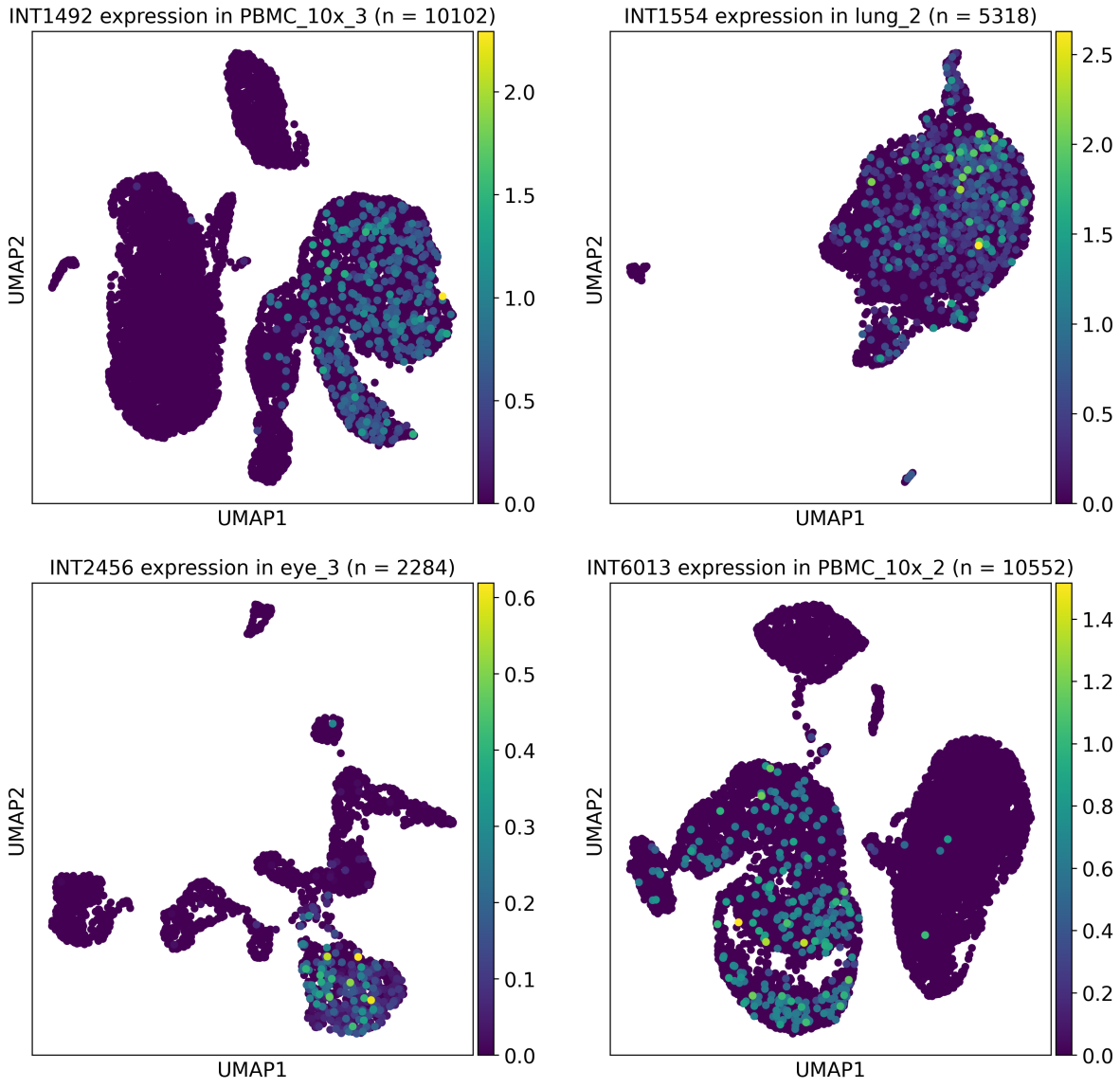


Figure 6.4: Examples of differentially expressed isolated intergenic regions.

Gene predictions To assess whether isolated intergenic regions overlap with predicted genes, UCSC gene prediction archives were examined, focusing on overlaps with 3' ends of predicted genes.

Of the identified regions, 12 overlapped with predicted genes. However, three of these did not overlap with the 3' ends of predicted genes, which would be expected given that the data was generated using 3' end sequencing methods. Five predicted genes were extensions of known genes, while the remaining four corresponded to genes absent from RefSeq and NCBI references but supported by our data. Table 6.4 summarizes these findings.

Name	Genomic coordinates	Prediction tool	Overlaps with 3' region of predicted gene
INT196	6:89086200-89086500	SIB	Yes (PNRC1)
INT827	9:94111950-94112200	SIB	Yes
INT1216	1:72282700-72282900	SIB	No (5' end)
INT1387	11:63570800-63571050	SIB	No
INT1525	8:11842200-11842350	SIB	Yes (CTSB)
INT1801	5:151272550-151272700	SIB	Yes (GM2A)
INT2044	19:4041150-4041400	SIB	Yes (ZBTB7A)
INT4070	4:47430500-47430700	SIB	Yes
INT4147	1:34861500-34861700	SIB	Yes (DLGAP3)
INT4577	5:702050-702300	SIB	Yes
INT4948	5:703250-703450	SIB	Yes
INT5710	4:47428500-47428700	SIB	No

Table 6.4: Isolated intergenic regions overlapping with predicted genes from UCSC gene prediction archive. The gene in the brackets shows if the predicted gene is extended version of already annotated genes.

Combining supporting features In total, five features were examined as supportive evidence for the biological significance of these intergenic regions: conservation, AT-rich/open chromatin presence, differential expression, and gene predictions. Based on these, four sublists of isolated intergenic regions were generated:

- Conserved regions
- Regions associated with open chromatin
- Differentially expressed regions
- Regions overlapping predicted genes

Notably, the list of conserved regions had only one element (INT1216) in common with the other lists (particularly, prediction list). The predicted gene list had two regions in common with the differential expression list (INT196, INT1525) and one in common with the open chromatin list (INT4147). All three of these predicted genes were extended versions of already annotated genes (see example in Figure 6.5). The differential expression and open chromatin lists had 11 regions in common (INT1312, INT1426, INT1492, INT1554, INT1609, INT1993, INT2209, INT349, INT368, INT4216, INT6013). No regions appeared in more than two lists. The full list of isolated regions can be found in Appendix 14.1.

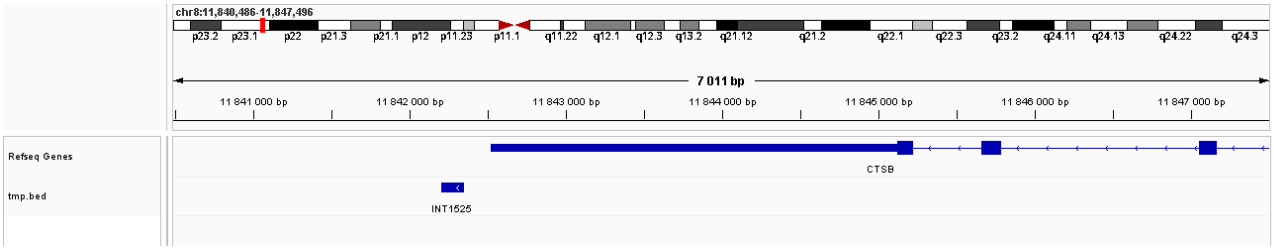


Figure 6.5: Example of gene which is predicted to have some longer transcripts than are present in the gencode and ncbi annotations. The intergenic region is located after the 3' end of gene CTSB, however, SIB prediction tool suggest that longer transcripts should be included in the annotation of this gene.

6.1.2 Antisense intergenic regions

Interpretation of antisense intergenic regions is more complicated. The features such as open chromatin, AT-richness or conservativness are not strand specific, meaning that if they would be present, they could not be attributed with certainty for our intergenic regions. The differential expression can be checked, however, even if the antisense intergenic region is differentially expressed, it still can be an artifact, originating from differentially expressed transcripts. Hence from the features checked for the isolated intergenic regions, only real supporting evidence would be computational gene prediction.

Gene predictions As in the case of isolated intergenic regions, antisense intergenic regions were checked for overlaps with predicted genes from UCSC prediction archive. There were 49 such regions found, out of them 35 were in the 3' ends of predicted genes. 9 of those 35 predicted genes were longer version of already annotated genes in RefSeq or GENCODE references. The all list can be seen in the Table 6.5.

Poly-A segments Having in mind that majority of the antisense intergenic regions have poly-A stretch after 3' end, and some of them strongly correlates with genes on the opposite strands, it is possible that they are artefacts from the library preparation. Particularly, it could happen that poly-T primer binds to poly-A stretch of cDNA produced by other primer, and in such way inversed transcripts are produced, which are later mapped to antisense strands of known genes (see Figure 6.6).

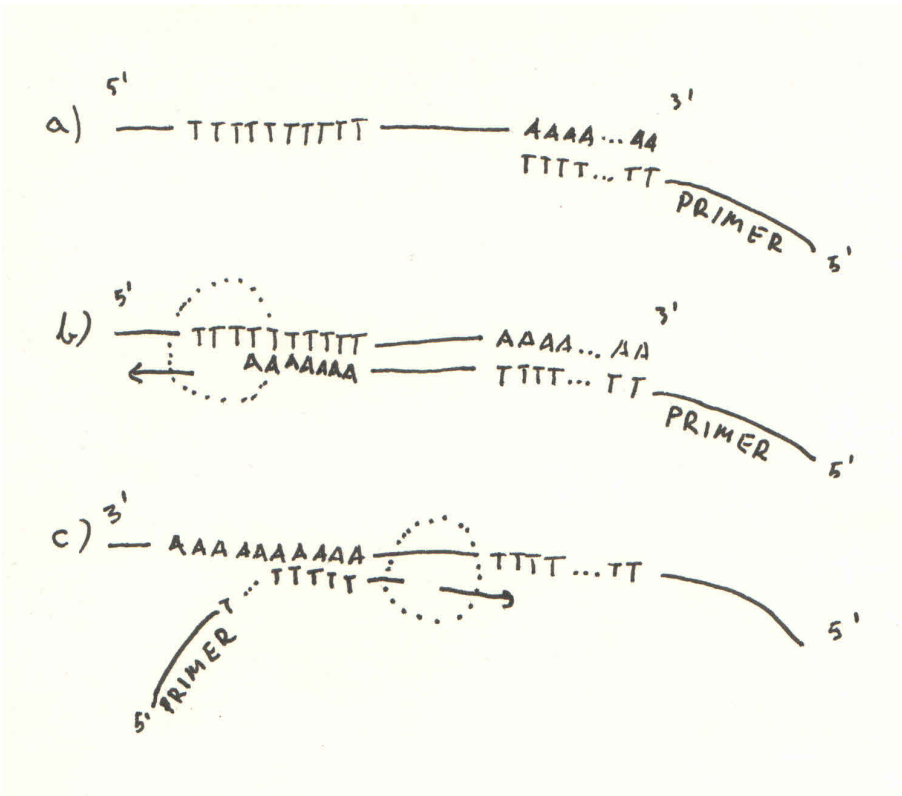


Figure 6.6: Scheme of poly-A mispriming resulting in reads mapped to the opposite strand compared to the original gene. a) RNA is captured by a poly-A tail with a primer containing a poly-T sequence. b) Reverse transcriptase attaches to the primer and synthesizes cDNA from the RNA template. c) If the RNA-cDNA complex becomes unstable and dissociates, another primer may bind to the poly-A stretches on the cDNA, if present. This would result in reversed cDNA, causing the reads to be mapped to the antisense strand.

Name	Genomic coordinates	Prediction tool	Overlaps with 3' region of predicted gene
INT7	6:24718850-24719100	SIB	Yes
INT33	5:172767250-172767500	SIB	Yes (DUSP1)
INT134	15:41280550-41280800	SIB	Yes (extends a bit after)
INT218	5:61409150-61409400	SPG	No (5' end)
INT285	10:110898200-110898550	SIB	Yes (BBIP1)
INT327	1:169132100-169132350	SIB	Yes (NME7)
INT347	14:75277700-75278000	Gescan	No
INT382	12:11892250-11892500	Gescan	Yes
INT386	6:26215200-26215450	SIB	Yes (H2BCB?)
INT438	4:39713300-39713600	SIB	Yes (almost all (short) gene spanned)
INT533	8:130110900-130111150	SIB	Yes
INT617	15:85733950-85734200	SIB	Yes
INT621	9:75148850-75149400	SIB	Yes (OSTF1)
INT651	1:28579100-28579350	SIB	Yes (TRNAU1AP)
INT654	6:34382850-34383050	SIB	No (but short gene)
INT828	8:133035800-133036050	SIB	Yes (SLA)
INT859	2:71373500-71373750	SIB	No (slight overlap on 5' end)
INT898	17:51190400-51190700	SIB	Yes (extends after)
INT903	8:22613400-22613650	SIB	Yes
INT949	9:111693050-111693400	SIB	Yes
INT972	15:64954800-64955350	SIB	Yes
INT984	18:63291900-63292200	Geneid	Yes
INT1028	5:50414200-50414500	SIB	Yes
INT1155	1:36296050-36296300	SIB	No (5' end)
INT1231	10:19890700-19891050	Gescan	Yes
INT1402	5:142786700-142786950	SIB	No (5' end)
INT1440	2:235746450-235746700	SIB	Yes
INT1445	9:40885650-40885900	SIB	Yes
INT1548	7:50342700-50342950	Gescan	No (exon in the middle)
INT1616	17:30895450-30895700	SIB	No (5' end)
INT1749	3:172333550-172333750	SIB	No (5' end)
INT1931	17:82522100-82522350	SIB	Yes
INT2065	10:103606150-103606400	SIB	Yes
INT2158	14:49636100-49636350	SIB	No (5' end, DNAAF2)
INT2194	7:130521250-130521500	SIB	No (but gene is short)
INT2371	1:14632650-14632900	Geneid	No (middle exon)
INT2380	22:33853900-33854150	Geneid	Yes
INT2825	13:45300050-45300300	Geneid	No (middle exon)
INT2979	4:98929350-98929600	SIB	No (5' end, EIF4E)
INT3328	2:184604250-184604500	SIB	Yes
INT3429	17:48172750-48173150	Augustus	Yes
INT3504	15:25332950-25333150	SIB	Yes (UBE3A)
INT3849	3:121432200-121432400	SIB	Yes
INT3868	17:31448500-31448700	SIB	Yes
INT4458	5:44820700-44820900	SIB	Yes (MRP530)
INT4743	20:3188900-3189050	SIB	Yes (DDRGK1)
INT5002	11:92232750-92232950	SIB	Yes
INT5315	20:3183750-3184050	SIB	Yes
INT5329	13:95603000-95603200	SIB	Yes (extends after)

Table 6.5: Antisense intergenic regions overlapping with predicted genes from UCSC gene prediction archive. The gene in the brackets shows if the predicted gene is extended version of already annotated genes.

7. DISCUSSION

8. CONCLUSIONS

9. RECOMMENDATION

10. ACKNOWLEDGEMENTS

11. REFERENCES

(1) 10x Genomics (2025). *10x Transcriptomic References*. URL: https://www.10xgenomics.com/support/software/cell-ranger/downloads/cr-ref-build-steps (visited on 02/19/2025).

(2) 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.

(3) Almeida da Paz, Michelle, Sarah Warger, and Leila Taher (May 2024). "Disregarding multimappers leads to biases in the functional assessment of NGS data". In: *BMC Genomics* 25.1. ISSN: 1471-2164. DOI: 10.1186/s12864-024-10344-9.

(4) Amaral, Paulo et al. (Oct. 2023). "The status of the human gene catalogue". In: *Nature* 622.7981, pp. 41–47. ISSN: 1476-4687. DOI: 10.1038/s41586-023-06490-x.

(5) 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.

(6) Brouze, Aleksandra et al. (May 2022). "Measuring the tail: Methods for poly(A) tail profiling". In: *WIREs RNA* 14.1. ISSN: 1757-7012. DOI: 10.1002/wrna.1737.

(7) Deschamps-Francoeur, Gabrielle, Joël Simoneau, and Michelle S. Scott (2020). "Handling multi-mapped reads in RNA-seq". In: *Computational and Structural Biotechnology Journal* 18, pp. 1569–1576. ISSN: 2001-0370. DOI: 10.1016/j.csbj.2020.06.014.

(8) 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.

(9) Ensembl (2025). *Ensembl website*. URL: https://www.ensembl.org/Homo_sapiens/Info/Annotation (visited on 02/13/2025).

(10) 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.

(11) 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.

(12) Guigó, Roderic (Aug. 2023). "Genome annotation: From human genetics to biodiversity genomics". In: *Cell Genomics* 3.8, p. 100375. ISSN: 2666-979X. DOI: 10.1016/j.xgen.2023.100375.

(13) 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.

(14) 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.

(15) 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.

(16) 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.

(17) 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.

(18) 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.

(19) 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.

(20) 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.

(21) 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.

(22) 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.

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

(24) Menon, Madhvi et al. (Oct. 2019). "Single-cell transcriptomic atlas of the human retina identifies cell types associated with age-related macular degeneration". In: *Nature Communications* 10.1. ISSN: 2041-1723. DOI: 10.1038/s41467-019-12780-8.

(25) Mould, Kara J. et al. (Apr. 2021). "Airspace Macrophages and Monocytes Exist in Transcriptionally Distinct Subsets in Healthy Adults". In: *American Journal of Respiratory and Critical Care Medicine* 203.8, pp. 946–956. ISSN: 1535-4970. DOI: 10.1164/rccm.202005-1989oc.

(26) Mudge, Jonathan M et al. (Nov. 2024). "GENCODE 2025: reference gene annotation for human and mouse". In: *Nucleic Acids Research* 53.D1, pp. D966–D975. ISSN: 1362-4962. DOI: 10.1093/nar/gkae1078.

(27) NCBI (2025). *NCBI website*. URL: https://www.ncbi.nlm.nih.gov/datasets/gene/GCF_000001405.40/ (visited on 02/13/2025).

(28) O'Leary, Nuala A. et al. (Nov. 2015). "Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation". In: *Nucleic Acids Research* 44.D1, pp. D733–D745. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1189.

(29) Pei, Baikang et al. (Sept. 2012). "The GENCODE pseudogene resource". In: *Genome Biology* 13.9. ISSN: 1474-760X. DOI: 10.1186/gb-2012-13-9-r51.

(30) 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.

(31) Pertea, Mihaela et al. (Nov. 2018). "CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise". In: *Genome Biology* 19.1. ISSN: 1474-760X. DOI: 10.1186/s13059-018-1590-2.

(32) 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.

(33) 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.

(34) 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.

(35) Salzberg, Steven L. (May 2019). "Next-generation genome annotation: we still struggle to get it right". In: *Genome Biology* 20.1. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1715-2.

(36) Siletti, Kimberly et al. (Oct. 2023). "Transcriptomic diversity of cell types across the adult human brain". In: *Science* 382.6667. ISSN: 1095-9203. DOI: 10.1126/science.add7046.

(37) 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.

(38) 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.

(39) Voigt, A.P. et al. (July 2019). "Molecular characterization of foveal versus peripheral human retina by single-cell RNA sequencing". In: *Experimental Eye Research* 184, pp. 234–242. ISSN: 0014-4835. DOI: 10.1016/j.exer.2019.05.001.

(40) 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.

(41) 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.

(42) 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.

(43) 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.

(44) 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.

(45) 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.

(46) 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

Table 14.1: Isolated intergenic regions. The DGE column indicates whether the region is differentially expressed. The Predictions column shows whether the region overlaps with a predicted gene. The Conservation column indicates whether the region has a conservation score greater than 0.6. The TATA column shows the distance to the closest TATA box upstream (a dot indicates no TATA box was not found within 10 kb upstream). The Poly-A column shows the distance to the poly-A stretch downstream (negative values indicate cases where the poly-A stretch slightly overlaps with the defined region, dot indicate that such stretch was not found in the 1kb downstream).

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT1066	16:11540750-11541200	Yes	No	No	.	-10
INT1312	3:141483900-141484100	Yes	No	No	.	25
INT1366	12:14938950-14939300	Yes	No	No	.	33
INT1426	12:10148500-10148700	Yes	No	No	307	21
INT1483	16:11541550-11541700	Yes	No	No	.	74
INT1492	X:56386150-56386400	Yes	No	No	257	42
INT1525	8:11842200-11842350	Yes	Yes	No	.	67
INT1554	14:61119550-61119750	Yes	No	No	.	54
INT1609	1:149879050-149879250	Yes	No	No	.	58
INT1938	6:148107250-148107550	Yes	No	No	.	17
INT196	6:89086200-89086500	Yes	Yes	No	713	-2
INT1993	6:43788000-43788200	Yes	No	No	.	45
INT2008	17:1567900-1568100	Yes	No	No	.	22
INT2141	6:159668300-159668550	Yes	No	No	.	37
INT2209	9:93049900-93050150	Yes	No	No	.	21

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT2328	1:991500-991800	Yes	No	No	.	-44
INT2456	13:113740250-113740500	Yes	No	No	.	-10
INT3259	17:35461800-35462050	Yes	No	No	150	22
INT349	17:43322200-43322700	Yes	No	No	.	97
INT368	19:14511650-14511900	Yes	No	No	231	42
INT404	2:73747700-73748000	Yes	No	No	250	13
INT4044	1:144412800-144413000	Yes	No	No	.	.
INT4216	19:39430250-39430550	Yes	No	No	.	17
INT4287	17:78646500-78646650	Yes	No	No	.	57
INT4872	1:12153700-12153900	Yes	No	No	.	415
INT5677	9:40875600-40875750	Yes	No	No	.	185
INT5882	7:8754100-8754250	Yes	No	No	653	124
INT6013	2:7102300-7102500	Yes	No	No	.	102
INT977	1:145987350-145987600	Yes	No	No	879	37
INT1002	5:82392950-82393200	No	No	No	.	71
INT1020	2:218407950-218408200	No	No	No	881	59
INT1114	21:15897200-15897400	No	No	No	.	42
INT1173	X:81313150-81313400	No	No	No	737	64
INT1216	1:72282700-72282900	No	Yes	Yes	.	-5
INT127	17:32384850-32385100	No	No	No	656	51
INT1385	2:86496700-86496900	No	No	No	89	12
INT1387	11:63570800-63571050	No	Yes	No	231	-26
INT1406	14:21194850-21195150	No	No	No	.	-50
INT1428	3:88001700-88001900	No	No	No	.	47
INT1467	5:112519400-112519600	No	No	No	765	6

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT1470	17:32392900-32393100	No	No	No	670	118
INT1509	2:42067000-42067200	No	No	No	95	27
INT1527	19:7573150-7573300	No	No	No	.	38
INT1542	18:70815200-70815400	No	No	No	.	22
INT1549	1:231017800-231018000	No	No	No	.	48
INT1569	1:157823400-157823600	No	No	No	.	11
INT1571	10:17913700-17913900	No	No	No	.	21
INT1579	3:12435200-12435400	No	No	No	.	30
INT1618	8:16107400-16107550	No	No	No	228	40
INT1619	3:56722050-56722300	No	No	No	.	1
INT1637	4:81423000-81423200	No	No	No	702	40
INT1640	17:7180650-7180850	No	No	No	.	33
INT1675	11:65772750-65772950	No	No	No	.	.
INT1716	9:27320000-27320200	No	No	No	.	49
INT1778	2:222711150-222711300	No	No	No	96	67
INT1779	3:189405250-189405400	No	No	No	141	61
INT1781	20:50205450-50205650	No	No	No	72	36
INT1801	5:151272550-151272700	No	Yes	No	.	43
INT1817	1:42711550-42711800	No	No	No	.	29
INT1829	1:77772950-77773050	No	No	Yes	.	-30
INT1869	13:30764850-30765000	No	No	No	.	.
INT1970	2:55277900-55278150	No	No	No	421	-4
INT2044	19:4041150-4041400	No	Yes	No	.	22
INT2071	15:84812500-84812800	No	No	No	773	39
INT2080	2:218816950-218817250	No	No	No	.	111

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT2088	8:19876500-19876650	No	No	No	699	.
INT212	X:44794800-44794950	No	No	No	.	-43
INT2142	12:54548650-54548850	No	No	No	.	63
INT2162	17:16355850-16356050	No	No	No	.	46
INT2199	6:148122550-148122750	No	No	Yes	.	74
INT2208	5:18162950-18163100	No	No	Yes	.	93
INT225	9:131570800-131571100	No	No	No	400	-23
INT2389	22:17731150-17731400	No	No	No	818	26
INT2420	1:52681350-52681600	No	No	No	980	52
INT2448	15:77040800-77041050	No	No	No	.	47
INT2585	17:75300550-75300800	No	No	No	473	12
INT2716	7:39812450-39812600	No	No	No	.	108
INT2777	12:106347650-106347900	No	No	No	.	50
INT2826	3:156825050-156825300	No	No	No	254	47
INT2970	1:30727150-30727450	No	No	No	.	3
INT312	9:127792200-127792400	No	No	No	.	57
INT313	20:5540550-5540800	No	No	No	.	20
INT316	2:157407250-157407500	No	No	No	237	42
INT3329	12:124911350-124911600	No	No	No	.	19
INT3330	10:26574800-26575050	No	No	No	107	39
INT3432	11:2318900-2319500	No	No	No	.	32
INT3472	1:234336450-234336650	No	No	No	516	80
INT3506	2:94568900-94569100	No	No	No	601	.
INT3519	5:166646400-166646600	No	No	No	273	121
INT3521	1:234330500-234330750	No	No	No	767	97

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT3543	13:83860250-83860700	No	No	No	836	-25
INT3636	X:129412750-129413050	No	No	Yes	.	107
INT3657	10:95078350-95078550	No	No	No	255	110
INT3658	11:29999550-29999750	No	No	No	760	98
INT3668	5:120107300-120107650	No	No	No	.	-21
INT3671	16:56859450-56859650	No	No	No	.	116
INT3704	18:43159050-43159250	No	No	No	70	104
INT3759	X:129426200-129426450	No	No	No	.	108
INT3940	4:157391550-157391750	No	No	No	.	96
INT4032	3:15548800-15549000	No	No	No	843	121
INT4047	21:33361700-33361900	No	No	No	143	39
INT4070	4:47430500-47430700	No	Yes	No	934	90
INT4147	1:34861500-34861700	No	Yes	No	.	609
INT4174	7:128664300-128664550	No	No	No	439	39
INT4224	8:43139250-43139600	No	No	No	.	.
INT4225	6:110097900-110098100	No	No	No	115	95
INT4236	X:121756550-121756750	No	No	No	.	83
INT426	14:24169950-24170200	No	No	No	.	29
INT4328	17:45610750-45610900	No	No	No	835	80
INT4364	Y:20418250-20418500	No	No	No	241	49
INT4394	19:40026900-40027100	No	No	No	.	89
INT4401	9:101413900-101414100	No	No	No	.	47
INT4454	X:57773350-57773500	No	No	No	191	136
INT4488	9:41567350-41567550	No	No	No	304	-23
INT4525	12:11599200-11599400	No	No	No	776	19

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT4577	5:702050-702300	No	Yes	No	893	86
INT4610	2:7218550-7218750	No	No	No	90	53
INT4697	5:166509150-166509400	No	No	No	14	55
INT4706	13:19811450-19811650	No	No	No	22	108
INT480	17:32396150-32396400	No	No	No	137	26
INT4851	X:129410350-129410550	No	No	No	.	94
INT4924	20:33358150-33358400	No	No	No	.	46
INT4948	5:703250-703450	No	Yes	No	.	.
INT4955	8:40521250-40521500	No	No	No	5	76
INT4975	12:47660700-47660950	No	No	No	71	.
INT5017	18:5389700-5389900	No	No	No	.	94
INT5128	16:67010550-67010750	No	No	No	.	81
INT5190	18:5386900-5387050	No	No	No	.	125
INT5456	11:1269150-1269350	No	No	No	.	115
INT5524	19:46024000-46024250	No	No	No	287	73
INT560	6:32862450-32862700	No	No	No	102	34
INT5710	4:47428500-47428700	No	Yes	No	594	91
INT575	11:85955150-85955450	No	No	No	.	-6
INT5801	16:67010700-67010900	No	No	No	949	.
INT5808	5:26875600-26875800	No	No	No	926	109
INT584	2:171768300-171768600	No	No	No	768	13
INT5848	16:3239750-3239950	No	No	No	488	35
INT593	2:32331000-32331250	No	No	No	814	36
INT600	22:47183900-47184150	No	No	No	22	59
INT6019	3:49344550-49344850	No	No	No	.	57

Continued on next page

Name	Coordinates	DGE	Predictions	Conservation	TATA	Poly-A
INT620	21:15883150-15883400	No	No	No	.	20
INT801	15:44718950-44719200	No	No	No	386	10
INT827	9:94111950-94112200	No	Yes	No	336	55
INT896	2:98598550-98598850	No	No	No	.	11
INT904	2:26138650-26138950	No	No	No	.	41
INT906	14:92696800-92697050	No	No	No	.	18
INT925	18:51087400-51087650	No	No	No	855	17

Figure 14.1: Clusterings of datasets. 'n' stands for number of cells in the plot, 'res' for the resolution parameter of leiden algorithm. *PBMC_10x* samples are colored by CellTypist annotations.