Learning cell identities and (post-)transcriptional regulation using single-cell data

Michielsen, L.C.M.

chapter 1

Introduction

In the 17th century, Robert Hooke discovered something fascinating when analyzing a piece of cork under a microscope: the cork consists of tiny pores. This reminded him of the cells in a monastery and therefore he called these pores 'cells' [1]. Almost two centuries later, Matthias Jakob Schleiden and Theodor Schwann formulated the first concept of cell theory: every organism consists of either one or multiple cells, and cells are the building blocks of life [2,3]. We estimate that the human body consists of ~3.7e13 cells [4].

Looking at our own human body, we know that cells have different functions. For example, immune cells fight against pathogens, skeletal muscle cells help us move, and sensory nerve cells receive information from the outside world. How is it possible that all these cells share the same DNA yet execute such a variety of functions? To explain this, we must understand the central dogma of molecular biology that describes the genetic flow of information in a cell (Figure 1) [5,6]. In every cell, there are chromosomes, very long DNA molecules, that provide the genetic code for an organism. Some parts of the DNA sequence, called genes, are transcribed into RNA molecules. Even though many different types of RNA exist with all important functions, we will focus on messenger RNA (mRNA) here. As the name already suggests, these mRNA molecules come from protein-coding genes and are translated into proteins. The resulting protein has a specific function in a cell.

Except for some somatic mutations, however, every cell in an organism has the same DNA. How could a cell know which genes have to be transcribed? Different control mechanisms tightly regulate transcription and translation to ensure the expression of the correct genes and proteins in a cell. For instance, transcription of protein-coding genes starts when RNA polymerase II and auxiliary factors bind the promotor region, the DNA sequence around the transcription start site (TSS) (Figure 2). A group of proteins, transcription factors (TFs), can bind parts of the DNA sequence, called enhancers and silencers, and either activate or repress the binding of RNA polymerase II or the auxiliary factors. This way, transcription factors control

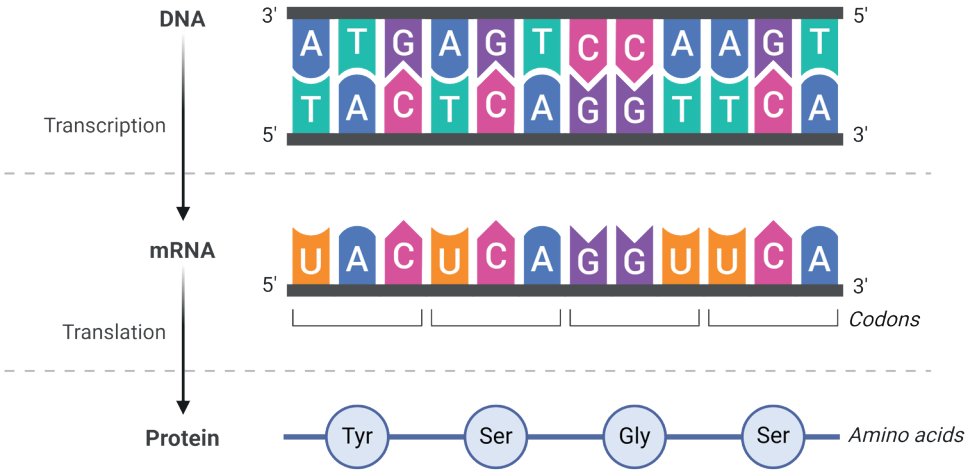


Figure 1. The central dogma of molecular biology. DNA is transcribed into mRNA, which is translated into proteins. [7]

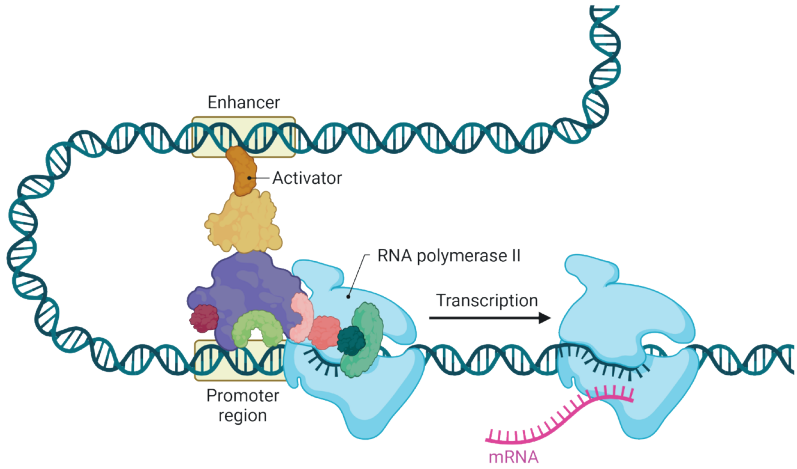


Figure 2. Transcriptional regulation. RNA polymerase II and co-factors must bind to the promoter region to start transcription. Other proteins, called activators, can bind enhancer regions and stimulate this process. The opposite can happen as well. Repressors can bind a silencer region and prevent the RNA polymerase II complex from binding and thus inhibit transcription. [8]

which genes are transcribed in a cell and to which extent. Since humans have approximately 1,400 TFs [9] that can also act combinatorially, the exact regulation mechanisms for each gene are incompletely understood. Understanding transcriptional regulation is important since a mutation in a TF, aberrant expression of a TF, or a mutation in a TF binding site can cause diseases and disorders ranging from cancer, autoimmune diseases, and diabetes to neurological disorders [10-14].

Humans have approximately 20,000 protein-coding genes [15,16]. Some genes, however, can produce different proteins with different functions [17,18]. How can the same mRNA molecule encode different proteins? After transcription, the resulting mRNA molecule has to be processed and spliced before the mature mRNA is transported to the nucleus and translated into a protein (Figure 3A). During the processing, the head and tail are modified to promote stability and export to the nucleus. Splicing, on the other hand, can lead to different proteins. The pre-mRNA molecule consists of exons, the coding regions, and introns. During splicing, the spliceosome, an RNA-protein complex, binds the RNA and catalyzes the removal of the introns. Exons from the same gene can be joined in different combinations, which we call alternative splicing (Figure 3B). Multiple forms of alternative splicing are recognized (Figure 3C). For instance, exons can be included or skipped completely, but alternative splice sites can be used as well. In humans, approximately 90-95% of the genes are alternatively spliced [19,20], which occurs most often in the brain [21].

We can draw a parallel between the regulation of (alternative) splicing and transcription. Where TFs binding the DNA sequence regulate transcription, RNA binding proteins (RBPs) regulate splicing. RBPs can either activate or repress the binding of the spliceosome and thereby control the splicing of exons or introns [22]. Aberrant splicing, for instance, caused by mutations in RBP binding sites, is a hallmark of many neurological diseases [23,24].

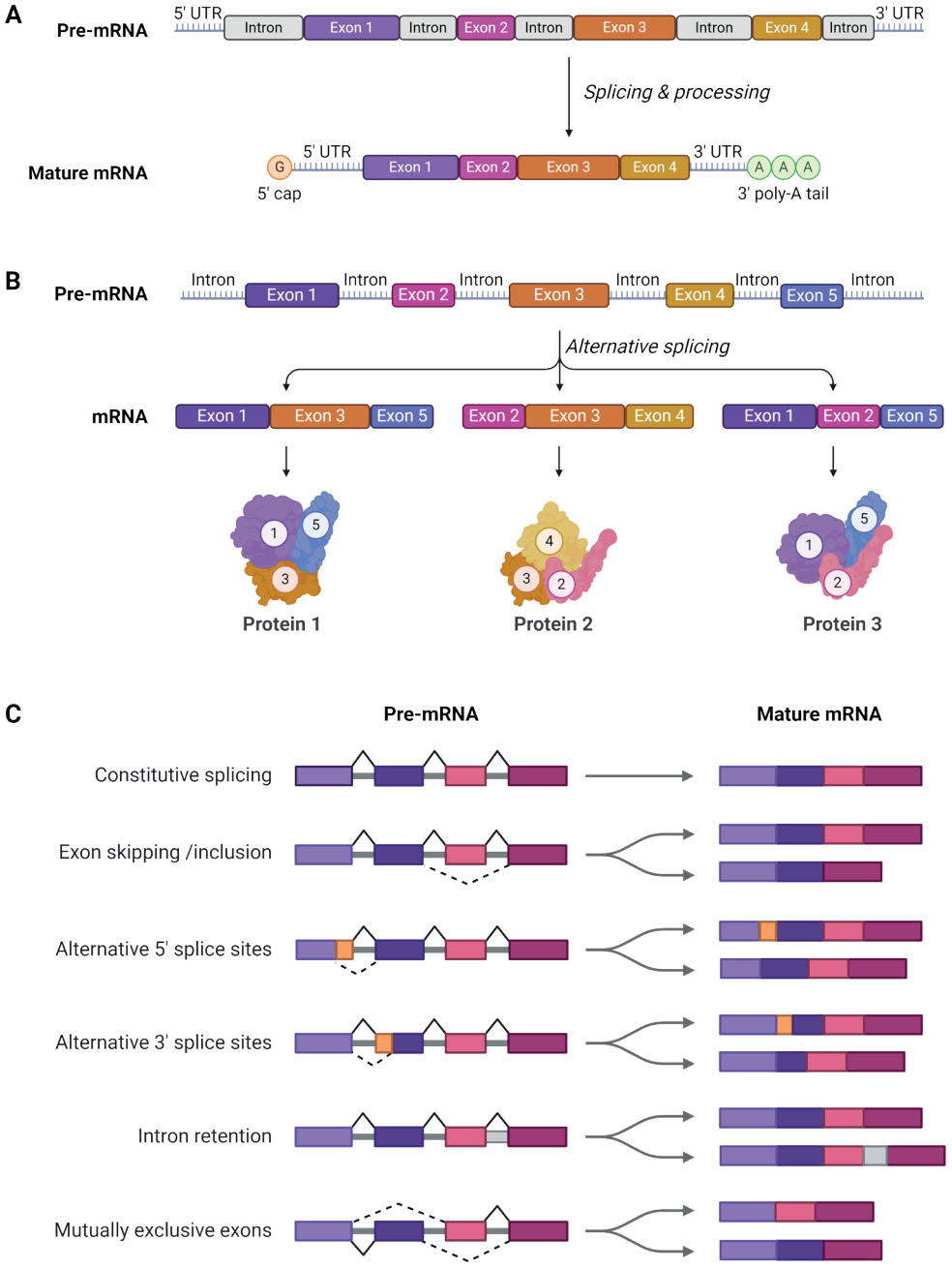


Figure 3. mRNA processing and splicing. **A)** The 5′ cap is added, the tail is polyadenylated, and the introns are spliced out. Afterwards, the mRNA can be transported to the nucleus and translated into a protein. [25] **B)** Alternative splicing. The pre-mRNA can be spliced in different ways. Different combinations of exons can be included in the mRNA molecule which will result in different proteins after translation. [25,26] **C)** Overview of different mRNA splicing types. [27]

1.1 Measuring transcription

To increase our understanding of cells in health and disease, we quantify which genes are expressed. RNA sequencing is a high-throughput technique to measure the number of mRNA molecules in a sample. This is often done using next-generation sequencing (NGS) technologies such as Illumina and Ion Torrent [28]. The general workflow consists of the following steps: 1) isolating the RNA from the cells, 2) fragmenting the RNA, 3) converting the mRNA into cDNA using reverse transcription, 4) ligating sequence adapters, 5) sequencing using a sequencing platform, 6) mapping the reads to the reference transcriptome, 7) constructing a count matrix (Figure 4). The final count matrix indicates how often a gene was measured in a sample.

NGS technologies generate relatively short reads. For instance, the read length is only 150 bp for most Illumina platforms. This short read length makes it impossible to study complete isoforms since the average length of human protein-coding transcripts is approximately 2.8kb [29]. Some reads map to splice junctions, so from such reads, we can extract whether exons are skipped or if alternative 3' or 5' splice sites are used.

1.1.1 Single-cell RNA sequencing

NGS techniques have been developed to measure transcription in groups of cells. This has the downside that the signal is evened out. If a gene's expression differs between two samples, it is impossible to know whether the sample consists of the same cells with altered expression or whether the cell-type composition changed (Figure 5A). This is especially disadvantageous when analyzing heterogeneous tissues, such as the brain.

In 2009, a new revolution began: single-cell RNA sequencing (scRNA-seq) [30]. Using scRNA-seq, the tissue is dissociated and the gene expression of individual cells can be measured

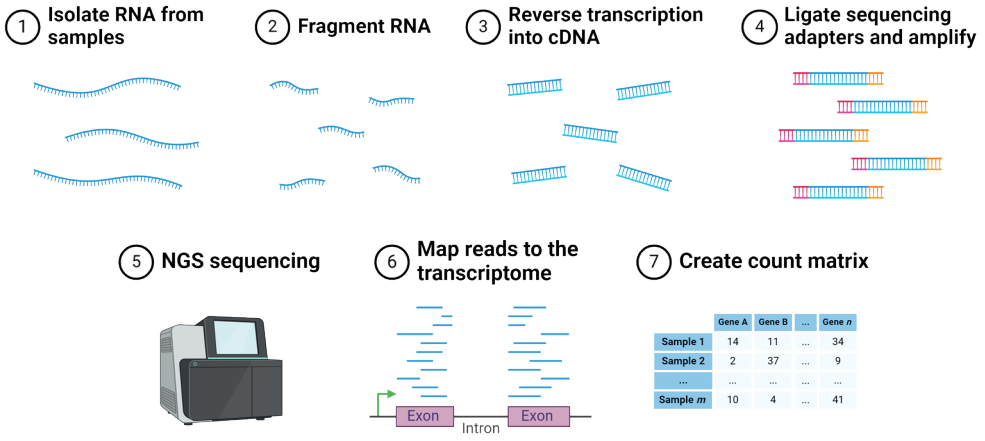


Figure 4. Overview of next-generation sequencing. [36]

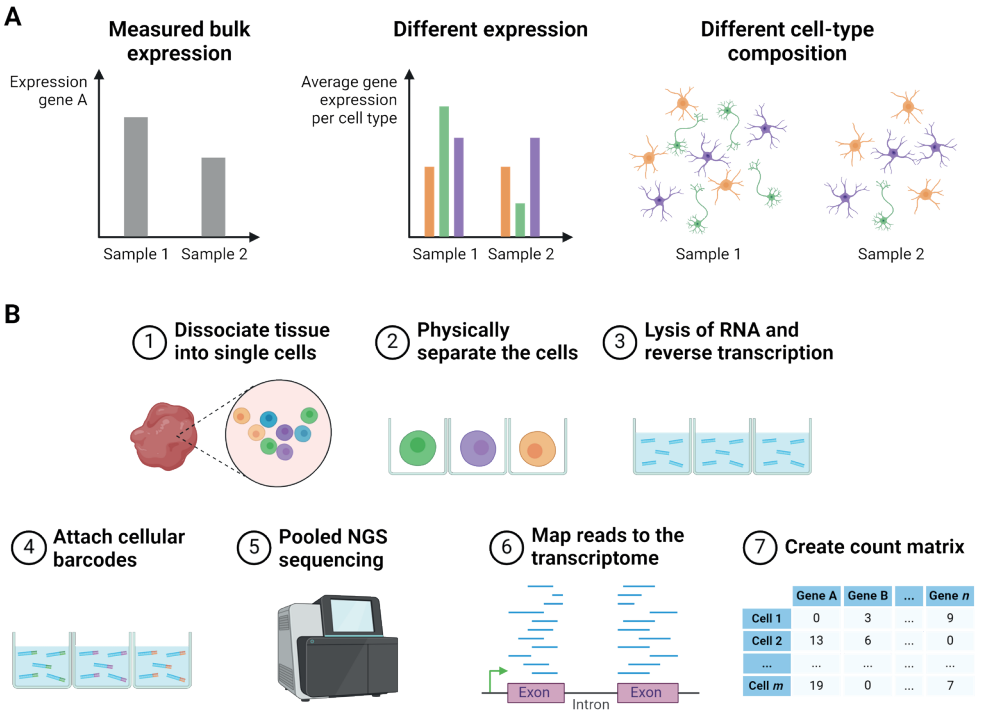


Figure 5. Single-cell RNA sequencing. **A)** The disadvantage of bulk RNA sequencing. Multiple scenarios can explain the decreased expression of gene A in sample 2. For instance, the expression of gene A decreased in the green cell type, or the cell-type composition changed which resulted in fewer green cells in sample 2. **B)** The general pipeline of single-cell RNA sequencing. This is similar to bulk RNA sequencing, except that cells are physically separated and cellular barcodes are attached to the cDNA. [36,37]

instead [31-35] (Figure 5B). The process is quite similar to sequencing in bulk, except that the cells are physically separated from each other and a barcode is attached to every cDNA molecule after reverse transcription. This barcode informs which reads originated from the same cell during the mapping step later. After barcoding, all material is pooled and sequenced together using an NGS platform. During the mapping step, the reads are split into the barcode and cDNA sequence. Based on the barcode, we know which cell the molecule came from, and based on the cDNA we know which gene was expressed.

In general, the data generated by all scRNA-seq protocols is sparse -around 90% of the values in the count matrix are zeros. Furthermore, when more cells are measured during an experiment, the sparser the data becomes [38]. Both biological and technical limitations explain this sparsity. Even essential genes will not always be expressed in a cell. Transcriptional bursting is the phenomenon in which genes are actively transcribed for a short period followed by a longer period of silence, which causes temporal fluctuation in gene levels [33]. Furthermore, since the mRNA content in a cell is low, it is difficult to capture all molecules.

Broadly, scRNA-seq methods can be split into two groups: either the full transcript is sequenced, which is similar to bulk analysis (e.g., using Smart-Seq2 [34]), or only the 3' or

5' end of the molecule can be captured and counted (e.g., using 10x Chromium [35]). An advantage of Smart-Seq2 is that the cells are sequenced deeper, which results in less sparse data. Furthermore-similar to bulk RNA sequencing- the reads can cover splice junctions. On the other hand, 10x optimized their pipeline for sequencing many cells simultaneously at a low cost. Up to hundreds of thousands of cells can be sequenced per experiment compared to thousands with Smart-Seq2. However, 10x only captures the 3' or 5' end of the mRNA molecule and ~100 nucleotides are measured. This short part of the sequence is enough to differentiate between all genes but lacks information about splice sites.

1.1.2 Long-read single-cell sequencing

To study alternative splicing, one would ideally sequence the whole mRNA molecule instead of looking at short fragments. Two technologies facilitate this nowadays: Oxford Nanopore [39,40] and PacBio [41]. Using Oxford Nanopore either the RNA molecule or the cDNA passes through a pore, which creates a changing electrical current. A base caller deciphers the order of nucleotides that generated these currents. PacBio uses single-molecule real-time (SMRT) sequencing which means that the cDNA molecule of interest is replicated using DNA polymerase. The incorporated new nucleotides are all fluorescent, with the four different bases each having a different fluorescent tag. When a nucleotide is incorporated, the fluorescent tag is cut off and a detector detects the fluorescent signal to decode the order of nucleotides.

Many different human tissues have been sequenced using such long-read protocols, which enhanced the discovery of more than 70.000 new transcripts [42]. This, however, is all in bulk. These protocols have been applied to single cells as well, but initially, only up to a hundred cells could be sequenced [43,44]. Several protocols have been developed to increase the throughput of long-read single-cell sequencing methods [45–47]. For example, some protocols combine short- and long-read sequencing (Figure 6) [48,49]. The single cells are barcoded using the 10x approach. After amplification, the cDNA is split into two pools. One pool is sequenced using Illumina and the other using Oxford Nanopore or PacBio. Due to

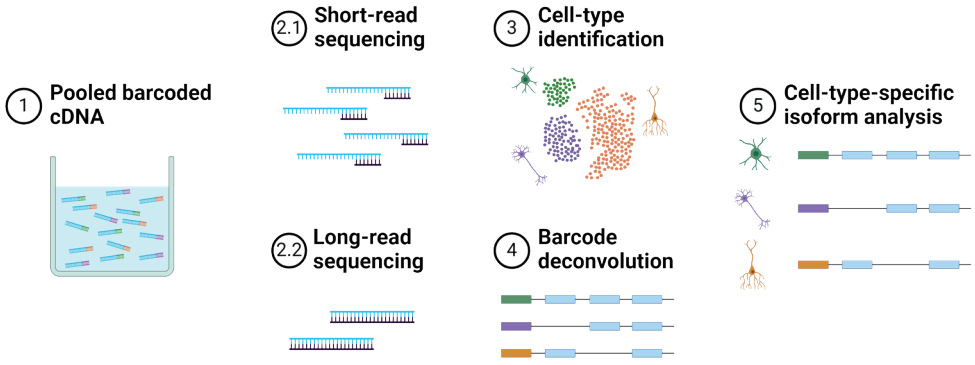


Figure 6. Schematic overview of long-read single-cell sequencing. The pooled barcoded cDNA is split into two pools. The first part is sequenced using short-read technologies, which can be used for cell-type identification. The second part is sequenced using long-read technologies. Since the barcodes of the short- and long-read data are similar, the data can be combined to study cell-type-specific isoforms. Figure adapted from Joglekar et al. (2021) [50].

the high costs of long-read sequencing, the coverage of the short reads generated by Illumina is usually higher, which results in better gene quantification and can be used to group the cells into specific cell types (see Section 1.2.1). The short-read barcodes are also present in the long-read data and can assign a cell and a cell type to every long-read. The long reads can be grouped per cell type and be used to study cell-type-specific isoform usage.

1.2 Cell types

Studying individual cells in scRNA-seq data is challenging since the data is sparse. Therefore, cells are grouped into cell types, which greatly reduces the complexity of the analysis, especially for organisms with as many cells as humans. But what is a cell type? How do we define them? The concept of a cell type might seem intuitive, but a clear definition is still missing.

In the past, cells were mainly studied under the microscope, so cell types were defined based on morphology. Camillo Golgi, for instance, developed a staining technique to visualize neurons that could later be used to classify them based on their dendritic patterns [51]. Nowadays, more and more features are measured, which changes our groupings of cells into cell types. With these new techniques, we can define a cell type based on which genes or proteins are expressed in a cell [52].

Even though the definition of cell types is dynamic, Cell Ontology [53] attempts to structure all identifiable cell types into a hierarchy. Most cells can be classified at different levels. For instance, a cell can be a blood cell, a lymphoid cell, a T cell, and so on (Figure 7). This hierarchical structure is inherent to cell types since all cells develop from the same cell and become gradually more specialized. The hierarchy shows that some cell types are more similar to one another. However, the cell-type hierarchy does not always align with development.

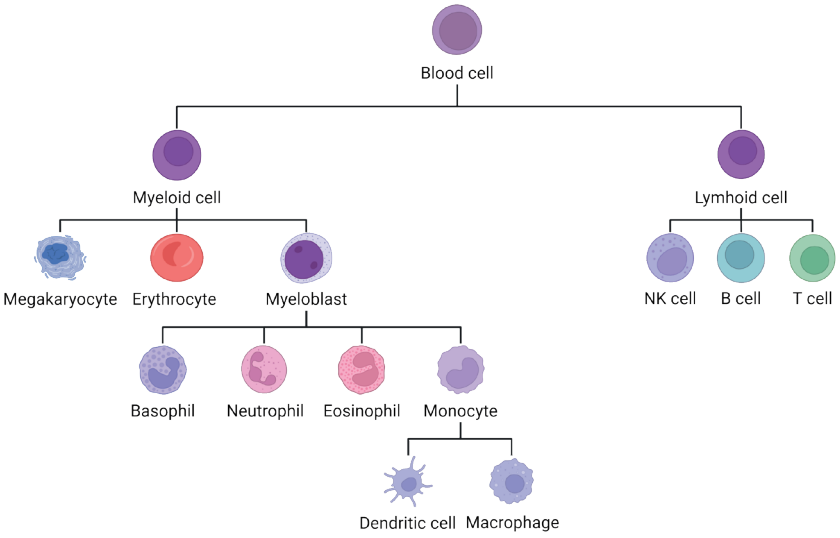


Figure 7. Example of a cell-type hierarchy for blood cells. Figure adapted from Monga et al. (2022) [54]

1.2.1 Discovering cell types in scRNA-seq data

In scRNA-seq data, the cell type of a cell is defined based on which genes are measured in a cell. Because the data is sparse, we cannot determine the cell type of individual cells by looking only at the expression of marker genes. As a solution, we first group cells with a similar gene expression profile and annotate these groups based on the expression of the marker genes (Figure 8A). The standard pipeline from a raw (short-read) scRNA-seq count matrix consists of different preprocessing, clustering, and visualization steps to annotate the clusters, which we will discuss in more detail below [55,56]. Several computational toolkits, such as Scanpy [57] and Seurat [58], have been developed to analyze scRNA-seq data, and all steps discussed below can be performed with these tools. After annotating the cells, other downstream analysis tasks, such as testing for differentially expressed genes between cell types, can be applied.

1.2.1.1 Preprocessing scRNA-seq data

Preprocessing starts with quality control to ensure that only high-quality, viable cells are in the data. Here, for instance, we filter out apoptotic cells based on the high content of mitochondrial genes [59,60]. Next, we normalize the data to remove differences in read depth between the cells. Most often, the data is normalized using library size normalization and log-transformed. After these steps, the dimensionality of the count matrix is still huge since ~20.000 genes are measured. Some of these genes are uniformly expressed across all the cells and uninformative for downstream tasks. We select 1000-5000 genes that show

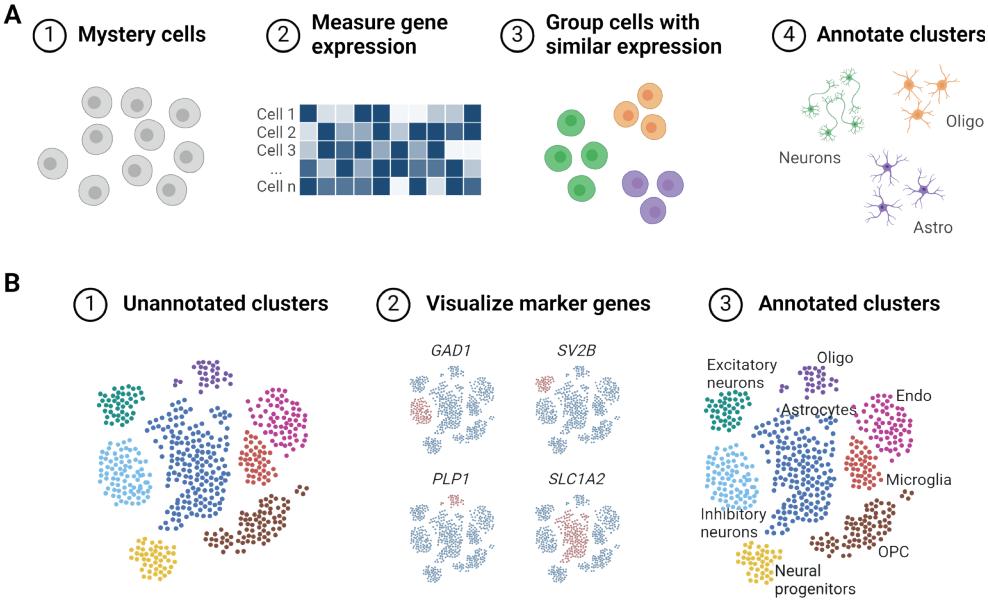


Figure 8. Annotating cell types in single-cell RNA-sequencing data. **A)** Mystery cells are grouped based on their expression pattern and these groups are annotated. **B)** Clusters are annotated by visualizing the expression of marker genes in two dimensions using t-SNE or UMAP.

the most variance in the dataset. Usually, the genes with the highest variance-to-mean ratio are selected. Next, we reduce the dimensions to 30-50 using principal component analysis (PCA). PCA is a linear dimensionality reduction method that reduces the data to a new set of features that is a linear combination of the old features that explain most of the variance. Instead of linear dimensionality reduction methods, non-linear methods can be applied as well. For instance, scVI [61], a variational autoencoder, can map the cells to a latent space of 10-50 dimensions.

1.2.1.2 Identifying cell types in scRNA-seq data

After preprocessing, the data is ready for downstream analysis such as cell-type identification. First, we cluster the data into groups of similar cells. We construct a *k*-nearest-neighbor graph in which every cell is connected to the *k* cells with the most similar gene expression pattern. Next, we detect clusters in this graph using Louvain [62] or Leiden [63] community detection. Here, the resolution parameter influences the number of clusters found. The resulting clusters can be visualized in two dimensions using t-SNE [64] or UMAP [65] (Figure 8B). To annotate the clusters, we visualize the expression of marker genes in, for instance, the two-dimensional space or a dot plot. However, marker genes might be unknown or not clearly expressed in scRNA-seq data, which makes annotating some clusters challenging.

1.3 Supervised learning for scRNA-seq data

In scRNA-seq data, cells are commonly annotated using clustering methods, an example of unsupervised learning. Unsupervised learning means that the data itself is unlabeled (i.e., the cell types are unknown) and the goal is to find groups in the data. However, unsupervised methods have drawbacks: they are subjective and time-consuming. Different parameters yield different clusterings, and the number of clusters or cell types discovered in scRNA-seq data is even correlated with the number of sequenced cells [66-68].

A shift towards supervised methods is needed to overcome this subjectiveness. Supervised models learn the relation between input data (e.g., the measured gene expression) and the label (e.g., the cell type). The trained model can annotate new, unlabeled data automatically. In this example, we predict the cell types that are discrete categories (classification), but supervised models can also be used to predict continuous outcomes (regression).

Many different types of supervised methods exist. Some rely on relatively simple principles and try to find a linear decision boundary between different groups, such as linear discriminant analysis or the linear support vector machine (SVM) (Figure 9A). Other methods, such as a *k*-nearest neighbor (kNN) or nearest mean classifier, look at which samples of the different groups of samples are closest and transfer the closest-group label to the new, unlabeled sample (Figure 9B). Deep learning models, such as neural networks, convolutional neural networks (CNN), and recurrent neural networks (RNN), can learn more complex relationships between the input features and the label (Figure 9C). Deep learning models have the disadvantage that much training data is needed and the models are difficult to interpret. With the linear models, we can easily see which input features guided the decision while

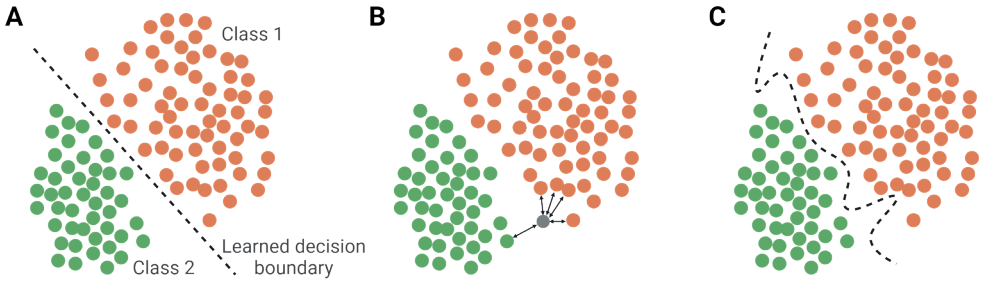


Figure 9. Supervised learning. **A)** Linear classifiers learn a linear decision boundary between the class 1 and class 2 samples. **B)** The *k*-nearest neighbor classifier looks at the neighboring samples and classifies new samples, for example, using a majority vote. In this case, the gray unlabeled sample would be classified as class 1. **C)** Deep learning models can learn complex decision boundaries.

this is impossible to know exactly for deep learning models. Approximation methods, such as Shapley values, exist though [69,70].

Automatic cell-type identification is one example of applying supervised models on scRNA-seq data. In this thesis, we will focus on two types of models: either we use the measured gene expression to predict the cell-type label (Section 1.4), or we know the cell-type label and use a generic input (e.g., the DNA sequence) to predict gene expression or splicing (Section 1.5).

1.4 Part I - Learning cell identities in scRNA-seq data

Ideally, we want to annotate the cells in a new scRNA-seq dataset automatically and consistently by using a classifier trained on an annotated dataset to transfer the labels to this new dataset. Several methods have been developed for this task, each varying considerably in their underlying principles. Some rely on relatively simple machine learning techniques such as a kNN classifier [71,72], SVM [73,74], or random forest (RF) [75–77], while others rely on more complex deep learning architectures [78,79]. We can also categorize methods by whether their approach is flat or hierarchical. Hierarchical methods exploit the inherent hierarchical structure of cell types; instead of learning the differences between all cell types in one go, they split the problem into smaller subproblems. Flat classifiers, on the other hand, do not benefit from this advantage. Another notable example of classifiers is methods that leverage the Cell Ontology [80,81]. Leveraging this ontology might be beneficial in the future, but currently, many newly discovered cell types are still missing in their hierarchy.

1.4.1 Challenges for cell-type identification

Even though many classification methods exist, we still face several challenges when automatically annotating cells.

1.4.1.1 Choosing the training dataset

An enormous amount of scRNA-seq datasets is publicly available, but it remains unclear which one is most optimal to train the classifier. Even datasets from the same tissue will contain different cell types since these datasets are annotated using unsupervised methods. Most research groups are interested in different cell compartments. Their cells of interest might be annotated at a fine-grained resolution, while the other cells are annotated at a low resolution-again relating to the inherent hierarchy of cell types. Comparing the annotations of different datasets can be challenging since a naming convention is missing.

An extra challenge is that most individual studies are incomplete. Rare cell types might be missing completely, or more difficult to discover when looking at one study only. Therefore, multiple datasets should be combined into a reference atlas, as demonstrated by initiatives like the Human Lung Cell Atlas [82]. Here, scRNA-seq data from 14 studies, 107 individuals, and different anatomical locations of the respiratory system is combined into one reference atlas. The cell-type labels of the datasets were manually harmonized using a group of experts, which is very time-consuming. Ideally, annotated datasets from the same tissue could be automatically combined to create a reference atlas.

1.4.1.2 Batch effects between datasets

Unwanted technical variations between datasets pose a second challenge for automatic cell-type identification. These batch effects are caused by variations in sequencing depths, handling of the cells, protocols, laboratories, etc. Consequently, batch effects between datasets should be removed before a classifier can be trained (Figure 10).

Removing batch effects is a trade-off between removing technical variation and preserving biological variation. Methods developed for this task can be categorized into three groups: 1) methods that correct the original gene space, 2) methods that project the data to a corrected latent space, and 3) methods that construct a batch-corrected graph. Methods in the second group usually yield the most optimal performance [83,84]. Another grouping of the current methods depends on whether they adjust all input datasets or allow users to pick one reference and project the query datasets onto it [72,85]. Even though the latter is more difficult, it has the advantage that the reference remains unchanged. As such, a classifier

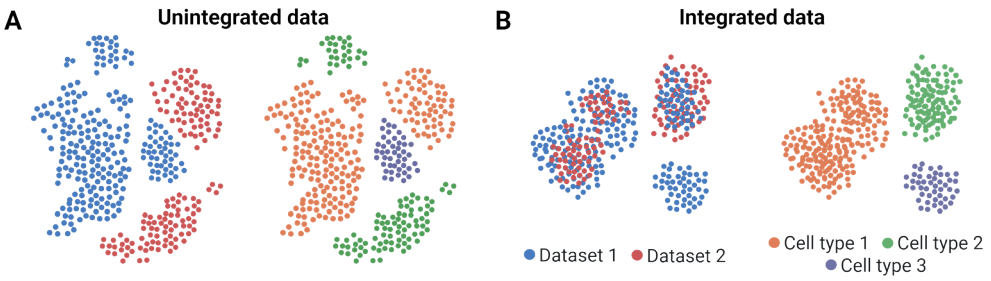


Figure 10. Schematic showing **A)** unintegrated and **B)** integrated scRNA-seq data. In the integrated data, the cells from datasets 1 and 2 overlap.

trained on this reference dataset can be used to annotate any query dataset. Combining these reference mapping methods with an accurate classifier would thus yield a more consistent annotation of the query datasets.

1.4.1.3 Identifying unknown cells

The third challenge for current classifiers is classifying cells as 'unknown' when the label is uncertain. This can be achieved by implementing a rejection option in the classifier. A correctly working rejection option is important for two reasons. First, the border between two cell types might not always be very distinct (Figure 11A). If a cell is close to the decision boundary, the label might be ambiguous and we prefer to keep it unlabeled. A low posterior probability of the classifier is a good indicator of this. Second, some datasets contain new or rare cell types that are not in the training data (Figure 11B). Here, the posterior probability might not work since this only indicates which cell types look most similar to the new cells, but not how similar they are. In this case, a distance metric is required. To correctly identify unknown cells in both scenarios, a classifier needs to use both the posterior probability and a distance metric to reject cells.

1.4.2 Learning cell identities across species

Model organisms, such as mice and rats, are often used to provide insights into biological mechanisms inside a cell or test the effect of new drugs or treatments. Knowing what aspects are similar or different between model organisms and humans is crucial for understanding how results translate. Comparing and matching cell types across species is one fundamental step in this process. Some cell types might be well conserved, while others might be species-specific. Matching cell types is thus interesting from an evolutionary point of view as well and aids in understanding cell-type evolution.

Besides the batch effects described in Section 1.4.1.2, an extra challenge during cross-species comparisons is that the measured gene sets are different. Throughout evolution, genes have been duplicated, deleted, and modified, which results in complex many-to-many relations. Relations between genes of different species are established based on their protein sequence similarity, with the underlying idea that proteins with a similar amino acid sequence will probably execute a similar function [86]. Traditionally, BLAST [87] is used for this task. However, a disadvantage of BLAST is that the whole protein sequence is weighed equally, while certain domains are more important for a specific function. More recently, large language

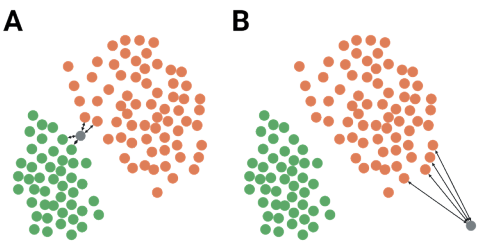


Figure 11. Examples of cells that should remain unlabeled. **A)** The gray cell is close to the decision boundary and therefore it is unclear whether it should be labeled a green or orange. The posterior probability of, for instance, the kNN classifier will be ~0.5, since about half of the neighbors are green and half are orange. **B)** The gray cell is far from the other cell types, which could indicate that it is a new cell type. The closest cells, however, are all orange so the posterior probability will be around one. In this case, a distance metric is needed to reject this cell.

models, such as SeqVec [88] and ProtBERT [89], have been trained to learn a representation of proteins in a lower dimensional space. These embeddings capture functional similarities between proteins and could be used to define homologous genes [90-92].

After matching the genes across species, only one-to-one orthologous genes, which are genes with exactly one match, are commonly used to compare cell types. The scRNA-seq methods developed for same-species data can be applied, which eases downstream analysis. A downside, however, is that much information is ignored. Some methods have been developed for cross-species analysis and use the many-to-many relationships between genes [93,94]. However, these methods currently all rely on the BLAST similarity. Using many-to-many orthologs defined by the protein embeddings would thus greatly enrich the cell type matches made.

1.5 Part II - Using scRNA-seq data to understand (post-) transcriptional regulation

(Post-)transcriptional regulation ensures that every cell expresses the correct genes and isoforms. Since a cell's gene expression level determines its cell type, these regulation mechanisms must be cell-type specific. Which TF or RBP binding sites are used on the DNA or RNA sequence will thus differ per cell type.

Understanding cell-type-specific regulation aids in understanding the underlying fundamental biological processes in a cell, which is, amongst others, essential for drug development. Furthermore, this enables us to predict the effect of mutations in non-coding regions. Mutations in a TF or RBP binding site will only affect gene expression or splicing if that binding site is normally used in that cell type. Knowing which mutations affect which cell types and how, will help to find new targets for drugs or therapies.

1.5.1 Genomic feature prediction models

Training genomic feature prediction models can help to unravel (post-)transcriptional regulation. These models use a generic input, such as the DNA sequence, to predict genomic features, such as gene expression or splicing, that were measured in a sample using RNA-seq. Why is it interesting to train these models though? The model cannot be extrapolated to new genes, as the expression of all genes was measured in the RNA-seq experiment. However, if a model can accurately predict the measured gene expression, interpreting why the model makes a high or low prediction for a gene improves our understanding of regulation. Current genomic feature prediction models can be divided into two groups: 1) feature-extraction-based and 2) sequence-based methods.

1.5.1.1 Feature-extraction-based models

Feature-extraction-based models extract features from the DNA sequence around the TSS of a gene or the RNA sequence around the splice site. These extracted features are used to train a relatively simple model, such as a linear regressor, to predict expression or splicing

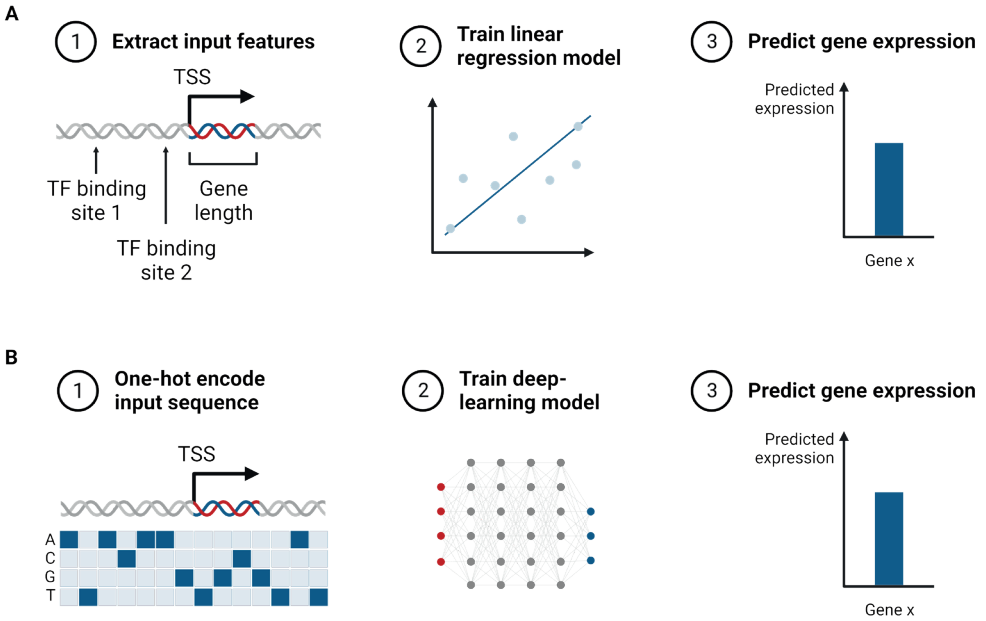


Figure 12. Schematic of **A)** feature-extraction-based and **B)** sequence-based models to predict genomic features. In this example, the DNA sequence is used to predict gene expression, but the RNA sequence could be used to predict splicing similarly.

(Figure 12A) [95–97]. Examples of extracted features are the gene length, GC content of the gene, and measured or predicted TF or RBP binding sites. The coefficients in the linear regressor directly inform us which features were most important for the predictions, which makes these models easy to interpret. However, we need prior knowledge about extracted features to train a model. If the preferred binding motif for a TF or RBP is unknown, we cannot incorporate it into our models either. Furthermore, evaluating how individual variants affect the prediction is more complicated since the sequence is not directly fed into the model.

1.5.1.2 Sequence-based models

Rapid developments in the deep learning field enabled a shift towards sequence-based methods. Sequence-based methods directly use the (one-hot encoded) DNA or RNA sequence as input to predict gene expression or splicing (Figure 12B) [98–100]. Depending on the task, a window varying from 400bp to 100kb around the TSS or the splice site is used as input. This input is unbiased towards known TFs, RBPs, or other extracted features. More complex models, such as CNNs, RNNs, or transformers, are used to learn the relation between the sequence and expression or splicing.

Training these deep learning models can be challenging since they tend to have millions of free parameters, and the sample size of the training data is limited. The training data size cannot be increased since the number of genes per organism is limited. As a solution, models can be trained on multiple species simultaneously, assuming that the regulatory mechanisms are at least partially conserved [101].

A second challenge is interpreting these black-box models. Model interpretation methods can give insights into what the model learned. One example is examining the initial layer of a CNN. The weights learned by these convolutional weight matrices are comparable to position-weight matrices, which indicate which sequences a TF or RBP prefers to bind [102]. Another option is using *in-silico* saturation mutagenesis (ISM) to systematically predict how nucleotide substitutions in the input sequence affect the predicted value [103,104]. Doing this for many input sequences can reveal interesting patterns that can be detected using TF-MoDISco [105]. TF-MoDISco discovers motifs that are predicted to positively or negatively affect the prediction.

1.5.1.3 Tissue-specific models

In the past, these models were trained using data from cell lines and only learned the basic principles of regulation. The models became more specific by training them, for instance, on bulk RNA-seq data from different tissues. In such cases, either a model per tissue or a multitask model can be trained. The regulation mechanisms, however, are cell-type-specific. Thus, there is a need for training these models on scRNA-seq data instead.

1.6 Contributions of this thesis

In this thesis, we address several challenges regarding identifying cell types in scRNA-seq data (Part I, Chapters 2-5) and using scRNA-seq datasets to improve our understanding of (post-) transcriptional regulation (Part II, Chapters 6-7).

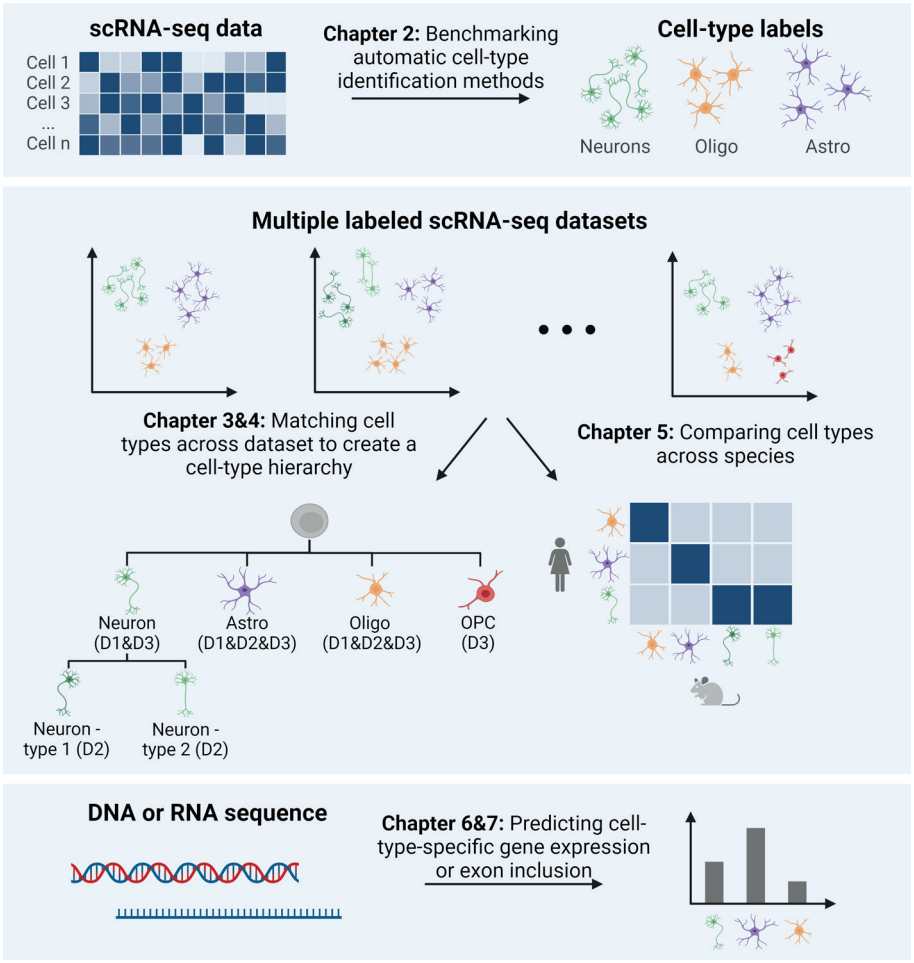
Part I - Learning cell identities in scRNA-seq data

Chapter 2: In Chapter 2, we benchmark sixteen cell-type identification methods designed for scRNA-seq data and six off-the-shelf Python classifiers. We compare their performance on 27 scRNA-seq datasets of different sizes, number of cell types, species, and technologies. Almost all methods perform well on most datasets, but their performance correlates negatively with the complexity of the data. Most classifiers suffer if a dataset contains many or very similar cell types. Overall, the linear SVM, one of the off-the-shelf Python classifiers, outperforms the methods designed for scRNA-seq data. Furthermore, when benchmarking the rejection options of the classifiers, we noticed that designing a proper rejection option is challenging and that relying on the posterior probability alone is not optimal.

Chapter 3: In Chapter 3, we present single-cell Hierarchical Progressive Learning (scHPL). scHPL combines multiple labeled scRNA-seq datasets into one classifier. We exploit the unharmonized labels of the input datasets to automatically create a cell-type hierarchy by matching the cell types of the different datasets. This hierarchy can either be updated progressively using new, labeled datasets or used as a classifier to annotate the cells in an unlabeled dataset. For every node in the hierarchy, we train a linear SVM since this performed best in the benchmark in Chapter 2. Furthermore, we implemented two rejection options using the posterior probability to reject cells between two cell types and the reconstruction error of the PCA to identify new cell types. We show that scHPL can accurately construct the

cell-type hierarchy for PBMC and brain datasets and that scHPL outperforms the flat linear SVM when annotating an unlabeled dataset.

Chapter 4: In Chapter 4, we combine scHPL and scArches [84] into a computational pipeline called treeArches. Before running scHPL, we require datasets to be batch-corrected. A downside of most batch-correction tools is that the complete alignment has to be repeated when adding a new dataset to update the hierarchy. Consequently, the complete hierarchy has to be rebuilt in the new integrated space. Since scArches is a reference-mapping method, it projects a new dataset on top of the reference, which ensures that the reference and corresponding hierarchy do not change. treeArches thus facilitates easy building and extending of reference atlases and the corresponding cell-type hierarchy.



Chapter 5: In Chapter 5, we propose a model to transfer and align cell types in cross-species analysis (TACTiCS). TACTiCS matches genes of different species using protBERT [89], an NLP model, while allowing for many-to-many matches. Next, it employs a neural network to train species-specific cell-type classifiers. Afterwards, it cross-predicts the other species' labels and compares the predicted to the original labels. TACTiCS outperforms state-of-the-art methods when matching human, mouse, and marmoset cell types in the primary motor cortex.

Part II - Using scRNA-seq data to understand (post-)transcriptional regulation

Chapter 6: In Chapter 6, we extend Xpresso, a tool to predict gene expression in bulk RNA-seq samples, to scXpresso which is a multitask model trained on scRNA-seq data to predict cell-type-specific gene expression. We show that cell-type-specific predictions are especially useful in heterogeneous tissues. In all experiments, cell-type-specific models outperform the tissue-specific models. The difference becomes most apparent when the gene expression of a cell type and the corresponding tissue are dissimilar. Furthermore, we show that scXpresso learns TF binding sites and envision that it will be useful for unraveling cell-type-specific transcriptional regulation mechanisms.

Chapter 7: In Chapter 7, we leverage long-read single-cell data to predict exon inclusion in glia and neurons in the human hippocampus and frontal cortex. We show that splicing is more difficult to predict in neurons than glia. Comparing RBP binding sites for exons with high and low exon inclusion between variable and non-variable exons, we found that these differ more in neurons than in glia, indicating that splicing mechanisms in variable exons in neurons diverged more from the standard mechanisms. Furthermore, we could pinpoint interesting RBPs regulating alternative splicing between glia and neurons.

Chapter 8: Finally, we discuss the contribution of our work in both research directions. First, we discuss how consistent cell-type classification can be improved. Next, we discuss the limitations of current genomic feature prediction models and suggest how these could be tackled.

Bibliography

1. Hooke R. Micrographia: or some physiological descriptions of minute bodies made by magnifying glasses. With observations and inquiries thereupon. London : Printed by Jo. Martyn, and Ja. Allestry ... and are to be sold at their shop ..., 1665.; 1665. Available: https://search.library.wisc.edu/catalog/999581426802121
2. Schwann T, Hünseler F. Mikroskopische Untersuchungen über die Ubereinstimmung in der Struktur und dem Wachstume der Tiere und Pflanzen. W. Engelmann; 1910.
3. Schleiden MJ. Arch Anat Physiol. Wiss Med. 1838.
4. Bianconi E, Piovesan A, Facchin F, Beraudi A, Casadei R, Frabetti F, et al. An estimation of the number of cells in the human body. Ann Hum Biol. 2013;40: 463–471. doi:10.3109/03014460.2013.807878
5. Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12: 138–163. Available: https://www.ncbi.nlm.nih.gov/pubmed/13580867
6. Cobb M. 60 years ago, Francis Crick changed the logic of biology. PLoS Biol. 2017;15: e2003243. doi:10.1371/journal.pbio.2003243
7. Reprinted from “Central Dogma”, by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.
8. Adapted from “Eukaryotic Gene Regulation - Transcriptional Initiation”, created by “Shadma Nafis” using BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.
9. Vaquerizas JM, Kummerfeld SK, Teichmann SA, Luscombe NM. A census of human transcription factors: function, expression and evolution. Nat Rev Genet. 2009;10: 252–263. doi:10.1038/nrg2538
10. Lee TI, Young RA. Transcriptional regulation and its misregulation in disease. Cell. 2013;152: 1237–1251. doi:10.1016/j.cell.2013.02.014
11. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science. 2012;337: 1190–1195. doi:10.1126/science.1222794
12. Van Houdt JKJ, Nowakowska BA, Sousa SB, van Schaik BDC, Seuntjens E, Avonce N, et al. Heterozygous missense mutations in SMARCA2 cause Nicolaides-Baraitser syndrome. Nat Genet. 2012;44: 445–9, S1. doi:10.1038/ng.1105
13. Lin CY, Lovén J, Rahl PB, Paranal RM, Burge CB, Bradner JE, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012;151: 56–67. doi:10.1016/j.cell.2012.08.026
14. Nie Z, Hu G, Wei G, Cui K, Yamane A, Resch W, et al. c-Myc is a universal amplifier of expressed genes in lymphocytes and embryonic stem cells. Cell. 2012;151: 68–79. doi:10.1016/j.cell.2012.08.033
15. Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland JE, Mudge JM, et al. GENCODE 2021. Nucleic Acids Res. 2021;49: D916–D923. doi:10.1093/nar/gkaa1087
16. Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic Acids Res. 2023;51: D933–D941. doi:10.1093/nar/gkac958
17. Sabate MI, Stolarsky LS, Polak JM, Bloom SR, Varndell IM, Ghatei MA, et al. Regulation of neuroendocrine gene expression by alternative RNA processing. Colocalization of calcitonin and calcitonin gene-related peptide in thyroid C-cells. J Biol Chem. 1985;260: 2589–2592. doi:10.1016/S0021-9258(18)89396-6
18. Muntoni F, Torelli S, Ferlini A. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. Lancet Neurol. 2003;2: 731–740. doi:10.1016/s1474-4422(03)00585-4
19. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet. 2008;40: 1413–1415. doi:10.1038/ng.259
20. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, et al. Alternative isoform regulation in human tissue transcriptomes. Nature. 2008;456: 470–476. doi:10.1038/nature07509
21. Yeo G, Holste D, Kreiman G, Burge CB. Variation in alternative splicing across human tissues. Genome Biol. 2004;5: R74. doi:10.1186/gb-2004-5-10-r74
22. Fisher E, Feng J. RNA splicing regulators play critical roles in neurogenesis. Wiley Interdiscip Rev RNA. 2022;13: e1728. doi:10.1002/wrna.1728
23. Licatalosi DD, Darnell RB. Splicing regulation in neurologic disease. Neuron. 2006;52: 93–101. doi:10.1016/j.neuron.2006.09.017
24. Dredge BK, Polydorides AD, Darnell RB. The splice of life: alternative splicing and neurological disease. Nat Rev Neurosci. 2001;2: 43–50. doi:10.1038/35049061
25. Adapted from “RNA Processing in Eukaryotes”, using BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.
26. Adapted from “Gene Splicing”, using BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.
27. Reprinted from “mRNA Splicing Types”, by BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

28. Hu T, Chitnis N, Monos D, Dinh A. Next-generation sequencing technologies: An overview. Hum Immunol. 2021;82: 801–811. doi:10.1016/j.humimm.2021.02.012

29. Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database . 2016;2016. doi:10.1093/database/baw153

30. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6: 377–382. doi:10.1038/nmeth.1315

31. Carangelo G, Magi A, Semeraro R. From multitude to singularity: An up-to-date overview of scRNA-seq data generation and analysis. Front Genet. 2022;13: 994069. doi:10.3389/fgene.2022.994069

32. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat Biotechnol. 2020;38: 747–755. doi:10.1038/s41587-020-0469-4

33. Haque A, Engel J, Teichmann SA, Lönnberg T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Med. 2017;9: 75. doi:10.1186/s13073-017-0467-4

34. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10: 1096–1098. doi:10.1038/nmeth.2639

35. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8: 14049. doi:10.1038/ncomms14049

36. Adapted from “RNA Sequencing”, using BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

37. Adapted from “Single-Cell Sequencing”, using BioRender.com (2023). Retrieved from https://app.biorender.com/biorender-templates.

38. Bouland GA, Mahfouz A, Reinders MJT. Consequences and opportunities arising due to sparser single-cell RNA-seq datasets. Genome Biol. 2023;24: 86. doi:10.1186/s13059-023-02933-w

39. Oikonomopoulos S, Wang YC, Djambazian H, Badescu D, Ragoussis J. Benchmarking of the Oxford Nanopore MinION sequencing for quantitative and qualitative assessment of cDNA populations. Sci Rep. 2016;6: 31602. doi:10.1038/srep31602

40. Sereika M, Kirkegaard RH, Karst SM, Michaelsen TY, Sørensen EA, Wollenberg RD, et al. Oxford Nanopore R10.4 long-read sequencing enables the generation of near-finished bacterial genomes from pure cultures and metagenomes without short-read or reference polishing. Nat Methods. 2022;19: 823–826. doi:10.1038/s41592-022-01539-7

41. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, et al. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323: 133–138. doi:10.1126/science.1162986

42. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, et al. Transcriptome variation in human tissues revealed by long-read sequencing. Nature. 2022;608: 353–359. doi:10.1038/s41586-022-05035-y

43. Byrne A, Beaudin AE, Olsen HE, Jain M, Cole C, Palmer T, et al. Nanopore long-read RNAseq reveals widespread transcriptional variation among the surface receptors of individual B cells. Nat Commun. 2017;8: 16027. doi:10.1038/ncomms16027

44. Karlsson K, Linnarsson S. Single-cell mRNA isoform diversity in the mouse brain. BMC Genomics. 2017;18: 126. doi:10.1186/s12864-017-3528-6

45. Lebrigand K, Magnone V, Barbry P, Waldmann R. High throughput error corrected Nanopore single cell transcriptome sequencing. Nat Commun. 2020;11: 4025. doi:10.1038/s41467-020-17800-6

46. Al’Khafaji AM, Smith JT, Garimella KV, Babadi M, Popic V, Sade-Feldman M, et al. High-throughput RNA isoform sequencing using programmed cDNA concatenation. Nat Biotechnol. 2023. doi:10.1038/s41587-023-01815-7

47. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, et al. Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing. Genome Biol. 2021;22: 310. doi:10.1186/s13059-021-02525-6

48. Gupta I, Collier PG, Haase B, Mahfouz A, Joglekar A, Floyd T, et al. Single-cell isoform RNA sequencing characterizes isoforms in thousands of cerebellar cells. Nat Biotechnol. 2018;36: 1197–1202. doi:10.1038/nbt.4259

49. Hardwick SA, Hu W, Joglekar A, Fan L, Collier PG, Foord C, et al. Single-nuclei isoform RNA sequencing unlocks barcoded exon connectivity in frozen brain tissue. Nat Biotechnol. 2022. doi:10.1038/s41587-022-01231-3 50. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021;12: 463. doi:10.1038/s41467-020-20343-5

50. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, et al. A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain. Nat Commun. 2021;12: 463. doi:10.1038/s41467-020-20343-5

51. Abdel-Maguid TE, Bowsher D. Classification of neurons by dendritic branching pattern. A categorisation based on Golgi impregnation of spinal and cranial somatic and visceral afferent and efferent cells in the adult human. J Anat. 1984;138 (Pt 4): 689–702. Available: https://www.ncbi.nlm.nih.gov/pubmed/6204961

52. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. Nat Biotechnol. 2016;34: 1145–1160. doi:10.1038/nbt.3711

53. Diehl AD, Meehan TF, Bradford YM, Brush MH, Dahdul WM, Dougall DS, et al. The Cell Ontology 2016: enhanced content, modularization, and ontology interoperability. J Biomed Semantics. 2016;7: 44. doi:10.1186/s13326-016-0088-7

54. Monga I, Kaur K, Dhanda SK. Revisiting hematopoiesis: applications of the bulk and single-cell transcriptomics dissecting transcriptional heterogeneity in hematopoietic stem cells. Brief Funct Genomics. 2022;21: 159–176. doi:10.1093/bfgp/elac002

55. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15: e8746. doi:10.15252/msb.20188746

56. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc. 2021;16: 2749–2764. doi:10.1038/s41596-021-00534-0

57. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. Genome Biol. 2018;19: 15. doi:10.1186/s13059-017-1382-0

58. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177: 1888–1902.e21. doi:10.1016/j.cell.2019.05.031

59. Griffiths JA, Scialdone A, Marioni JC. Using single-cell genomics to understand developmental processes and cell fate decisions. Mol Syst Biol. 2018;14: e8046. doi:10.15252/msb.20178046

60. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17: 29. doi:10.1186/s13059-016-0888-1

61. Gayoso A, Lopez R, Xing G, Boyeau P, Wu K, Jayasuriya M, et al. scvi-tools: a library for deep probabilistic analysis of single-cell omics data. bioRxiv. 2021. p. 2021.04.28.441833. doi:10.1101/2021.04.28.441833

62. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. J Stat Mech: Theory Exp. 2008. doi:10.1088/1742-5468/2008/10/P10008

63. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9: 5233. doi:10.1038/s41598-019-41695-z

64. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9: 2579–2605. Available: https://jmlr.org/papers/v9/vandermaaten08a.html

65. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018. Available: http://arxiv.org/abs/1802.03426

66. Svensson V, da Veiga Beltrame E, Pachter L. A curated database reveals trends in single-cell transcriptomics. Database. 2020;2020. doi:10.1093/database/baaa073

67. Tasic B, Menon V, Nguyen TN, Kim TK, Jarsky T, Yao Z, et al. Adult mouse cortical cell taxonomy revealed by single cell transcriptomics. Nat Neurosci. 2016;19: 335–346. doi:10.1038/nn.4216

68. Tasic B, Yao Z, Graybuck LT, Smith KA, Nguyen TN, Bertagnolli D, et al. Shared and distinct transcriptomic cell types across neocortical areas. Nature. 2018;563: 72–78. doi:10.1038/s41586-018-0654-5

69. Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2014;41: 647–665. doi:10.1007/s10115-013-0679-x

70. Shrikumar A, Greenside P, Kundaje A. Learning Important Features Through Propagating Activation Differences. arXiv [cs.CV]. 2017. Available: http://arxiv.org/abs/1704.02685

71. Kiselev VY, Yiu A, Hemberg M. scmap: projection of single-cell RNA-seq data across data sets. Nat Methods. 2018;15: 359. Available: https://doi.org/10.1038/nmeth.4644

72. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;0. doi:10.1016/j.cell.2021.04.048

73. Wagner F, Yanai I. Moana: A robust and scalable cell type classification framework for single-cell RNA-Seq data. bioRxiv. 2018; 456129. doi:10.1101/456129

74. Alquicira-Hernandez J, Sathe A, Ji HP, Nguyen Q, Powell JE. ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. Genome Biol. 2019;20: 264. doi:10.1186/s13059-019-1862-5

75. Johnson TS, Wang T, Huang Z, Yu CY, Wu Y, Han Y, et al. LAmbDA: Label Ambiguous Domain Adaptation Dataset Integration Reduces Batch Effects and Improves Subtype Detection. Bioinformatics. 2019. doi:10.1093/bioinformatics/btz295

76. Lieberman Y, Rokach L, Shay T. CaSTLe – Classification of single cells by transfer learning: Harnessing the power of publicly available single cell RNA sequencing experiments to annotate new experiments. Kaderali L, editor. PLoS One. 2018;13: e0205499. doi:10.1371/journal.pone.0205499

77. Tan Y, Cahan P. SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. Cell Syst. 2019;9: 207–213.e2. doi:10.1016/j.cels.2019.06.004

78. Ma F, Pellegrini M. ACTINN: automated identification of cell types in single cell RNA sequencing. Bioinformatics. 2020;36: 533–538. doi:10.1093/bioinformatics/btz592

79. Xu C, Lopez R, Mehlman E, Regier J, Jordan MI, Yosef N. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. Mol Syst Biol. 2021;17: e9620. doi:10.15252/msb.20209620

80. Wang S, Pisco AO, McGeever A, Brbic M, Zitnik M, Darmanis S, et al. Leveraging the Cell Ontology to classify unseen cell types. Nat Commun. 2021;12: 5556. doi:10.1038/s41467-021-25725-x

81. Bernstein MN, Ma Z, Gleicher M, Dewey CN. CellO: comprehensive and hierarchical cell type classification of human cells with the Cell Ontology. iScience. 2021;24: 101913. doi:10.1016/J.ISCI.2020.101913

82. Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An integrated cell atlas of the lung in health and disease. Nat Med. 2023;29: 1563–1577. doi:10.1038/s41591-023-02327-2

83. Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19: 41–50. doi:10.1038/s41592-021-01336-8

84. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21: 1–32. doi:10.1186/s13059-019-1850-9

85. Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2022;40: 121–130. doi:10.1038/s41587-021-01001-7

86. Kimura M, Ohta T. On some principles governing molecular evolution. Proc Natl Acad Sci U S A. 1974;71: 2848–2852. doi:10.1073/pnas.71.7.2848

87. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215: 403–410. doi:10.1016/S0022-2836(05)80360-2

88. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, et al. Modeling aspects of the language of life through transfer-learning protein sequences. BMC Bioinformatics. 2019;20: 723. doi:10.1186/s12859-019-3220-8

89. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. IEEE Trans Pattern Anal Mach Intell. 2022;44: 7112–7127. doi:10.1109/TPAMI.2021.3095381

90. Kilinc M, Jia K, Jernigan RL. Protein Language Model Performs Efficient Homology Detection. bioRxiv. 2022. p. 2022.03.10.483778. doi:10.1101/2022.03.10.483778

91. Villegas-Morcillo A, Makrodimitris S, van Ham RCHJ, Gomez AM, Sanchez V, Reinders MJT. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. Bioinformatics. 2020. doi:10.1093/bioinformatics/btaa701

92. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proc Natl Acad Sci U S A. 2021;118. doi:10.1073/pnas.2016239118

93. Tarashansky AJ, Musser JM, Khariton M, Li P, Arendt D, Quake SR, et al. Mapping single-cell atlases throughout Metazoa unravels cell type evolution. Elife. 2021;10. doi:10.7554/eLife.66747

94. Liu X, Shen Q, Zhang S. Cross-species cell-type assignment from single-cell RNA-seq data by a heterogeneous graph neural network. Genome Res. 2023;33: 96–111. doi:10.1101/gr.276868.122

95. Barash Y, Calarco JA, Gao W, Pan Q, Wang X, Shai O, et al. Deciphering the splicing code. Nature. 2010;465: 53–59. doi:10.1038/nature09000

96. Barash Y, Vaquero-Garcia J, González-Vallinas J, Xiong HY, Gao W, Lee LJ, et al. AVISPA: a web tool for the prediction and analysis of alternative splicing. Genome Biol. 2013;14: R114. doi:10.1186/gb-2013-14-10-r114

97. McLeay RC, Lesluyes T, Cuellar Partida G, Bailey TL. Genome-wide in silico prediction of gene expression. Bioinformatics. 2012;28: 2789–2796. doi:10.1093/bioinformatics/bts529

98. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Rep. 2020;31: 107663. doi:10.1016/j.celrep.2020.107663

99. Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176: 535–548.e24. doi:10.1016/j.cell.2018.12.015

100. Avsec Ž, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18: 1196–1203. doi:10.1038/s41592-021-01252-x

101. Kelley DR. Cross-species regulatory sequence activity prediction. Ma J, editor. PLoS Comput Biol. 2020;16: e1008050. doi:10.1371/journal.pcbi.1008050

102. Novakovsky G, Dexter N, Libbrecht MW, Wasserman WW, Mostafavi S. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24: 125–137. doi:10.1038/s41576-022-00532-2

103. Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Res. 2016;26: 990–999. doi:10.1101/gr.200535.115

104. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12: 931–934. doi:10.1038/nmeth.3547

105. Shrikumar A, Tian K, Avsec Ž, Shcherbina A, Banerjee A, Sharmin M, et al. Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5. arXiv [cs.LG]. 2018. Available: http://arxiv.org/abs/1811.00416