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

Michielsen, L.C.M.

# chapter 8

Discussion

Single-cell RNA sequencing (scRNA-seq) has massively increased our understanding of tissue compositions, cellular interactions, and developmental processes. Especially in heterogeneous tissues such as the brain, this single-cell resolution led to many newly discovered cell types, insights into the specificity of cell types for particular brain regions or layers, and the proportions of cell types across the brain [1–5]. Besides generating massive datasets, smaller publicly available datasets are combined into tissue-specific reference atlases, such as the Human Lung Cell Atlas [6]. However, analyzing individual datasets or creating these atlases is still mainly done using unsupervised methods.

In this thesis, we introduced several supervised methods to solve two broad tasks: 1) automatic cell-type identification in scRNA-seq data, and 2) understanding cell-type-specific (post-)transcriptional regulation. In part I, we benchmarked different cell-type classification methods for scRNA-seq data (chapter 2), developed scHPL (chapter 3) and treeArches (chapter 4) to automatically match cell types across datasets to construct a reference atlas with corresponding cellular hierarchy, and developed TACTiCS to match cell types across species (chapter 5). In part II, we showed how scRNA-seq with the corresponding cell-type labels can improve our understanding of transcriptional regulation (chapter 6) and alternative splicing (chapter 7) by developing cell-type-specific feature-prediction models. However, for both tasks, several challenges remain that we will discuss in the sections below.

## 8.1 What is a cell type?

In simple eukaryotic organisms, such as C. Elegans, every adult consists of the same amount of cells - 959 in hermaphrodites and 1031 in males [7,8]. This low and consistent number of cells allows researchers to study every cell individually. Studying more complex organisms, such as humans, similarly is challenging since we consist of approximately 37 trillion cells, and this number varies across individuals due to, for instance, differences in height [9]. Categorizing all these cells into cell types enhances our understanding of cells and facilitates effective communication and comparison of results across studies.

Is this discrete grouping that we use repeatedly throughout this thesis optimal, or would a continuous spectrum be beneficial? At least at a high level, cell types seem separate categories. For example, a muscle fiber differs from a neuron regarding its function, morphology, and the genes expressed. Still, both arise from the same stem cell and become continuously more specialized. At what stage during development would one consider these cells differentiated enough to call them different cell types?

Furthermore, due to perturbations, such as stimulations or pathogens, cells can transition to another cell type or state. Should these possible responses be considered in our definition as well [10]? In the pancreas, some alpha cells can, for instance, change into beta cells, which can occur naturally in persons with diabetes [11]. Also in the immune system, naive T-cells transition into memory cells after activation [12]. Both are considered different cell types with a gradient containing the transitioning cells in between.

Despite this evidence for a more continuous spectrum, we still focus on cell-type classification since most downstream methods require cells from the same cell type or cell-type labels as input. This downstream analysis can be a relatively simple task, such as testing for differentially expressed genes between healthy and diseased cells of the same cell type. But for more complex tasks, such as detecting expression quantitative trait loci (eQTLs), the cell-type labels may be beneficial as well. A cell-type-specific eQTL analysis can reveal the effect of variants that were previously hidden when analyzing the complete sample [13]. Also in Chapters 6 and 7, we rely on the cell-type labels to improve our understanding of transcriptional regulation and alternative splicing. Especially for heterogeneous tissues, using this increased cell-type-specific resolution improved the performance.

A potential alternative could be to redevelop current downstream methods such that they produce similar results, but do not rely on cell-type labels. An example is Milo, which tests for differential abundance between two samples [14]. First, cells are assigned to neighborhoods and afterwards, Milo tests whether cells from a certain condition are enriched or depleted within each neighborhood. Cell-type labels are unneeded during this analysis and will thus not bias the results. For the sequence-based models, this problem could be overcome by predicting the features at the cell instead of cell-type resolution as is done by scBasset [15] and seq2cells [16]. As datasets grow bigger and bigger, this might become computationally too expensive at some point. However, the results of cell-type-agnostic methods might be harder to interpret. As a solution, cells could be aggregated into cell types again solely for interpretation. Then, at least the cell-type labels do not bias the analysis itself.

Since cells exist in a continuous spectrum, a second alternative is moving from binary to fuzzy cell-type labels. Using fuzzy labels, a cell can belong to multiple cell types simultaneously with different probabilities. A probability above zero for two cell types can indicate that a cell is transitioning between these two. For scRNA-seq data, this approach has been explored for clustering methods [17,18], but not yet for classification methods. During classification, the posterior probability could easily indicate which cell types a cell belongs to.

## 8.2 Consistent cell-type classification



Since most downstream methods rely on discrete cell-type labels, cells must be labeled consistently to enable combining or comparing information from different datasets. For instance, the sc-eQTL consortium aims to find how variants affect gene expression in immune cell types by combining datasets from multiple labs containing hundreds of individuals [19]. In every individual, the cell types should thus be defined similarly. A high precision in cell-type annotations might be even more important than a high accuracy. Since unsupervised methods are subjective and time-consuming, an automatic supervised approach is needed here.

Ideally, such a classifier is trained on a reference atlas that combines data from enough individuals so that inter-individual variation and rare cell types are captured. The cell types in such a reference atlas should not be characterized as in a periodic table but in a hierarchy [20]. A hierarchical classifier divides the classification problem into smaller subproblems which improves the classification performance. We showed that a hierarchical linear SVM

outperforms a flat linear SVM in Chapter 3. Besides, when using a hierarchical classifier, users can easily choose their resolution of interest. Using Azimuth [21], an easy-to-use web portal, cells can also be annotated at different resolutions. However, these resolutions are not connected in a hierarchy. Consequently, a cell can, for instance, be labeled as a CD8+ T-cell and CD4+ memory T-cell, which is impossible and therefore inconsistent.

Reference atlases exist for many human and mouse tissues and can be downloaded from platforms, such as Azimuth [21] or CELLxGENE [22,23]. For these reference atlases, either 1) one big dataset is used (e.g. the human PBMC reference containing eight individuals [21]), 2) multiple datasets are combined and re-annotated manually (e.g. the human lung cell atlas containing 107 individuals from 14 datasets [6]), or 3) multiple annotated datasets are combined using scHPL and their labels are manually refined (e.g. the mouse kidney atlas combining data from 59 mice from 8 datasets [24]).

However, many datasets are still annotated using unsupervised methods even though a reference atlas for that specific tissue is available [25–27]. Why is this the case? Researchers might not trust supervised methods since their performance is not perfect yet. In Chapter 2, however, we showed that cell-type classification is a relatively easy problem at a low resolution since almost all methods perform (nearly) perfectly. The performance of most methods drops when increasing the resolution or complexity of the data. For most reference atlases, however, the performance is not benchmarked per resolution, making it hard to know how consistent label transfer will be.

Another complicating factor is the batch effects between the reference atlas and the unlabeled dataset. Batch effects are technical variations between datasets due to variations in labs, protocols, sequencing depths, etc. This technical variation has to be removed while preserving the biological variation. This is a complex problem since the effects are usually non-linear and the ground truth is unknown. Benchmark studies showed that methods including scVI [28] and Harmony [29] perform well for this task. For most methods, however, parameters have to be tuned for optimal performance, which might decrease the usability.

Interestingly, researchers are imperfect when annotating a scRNA-seq dataset manually as well. In Chapter 3, we applied scHPL to multiple annotated PBMC datasets, which resulted in a hierarchy with unexpected edges. Visualizing marker genes in the individual datasets indicated that cells had been wrongly annotated in the original datasets. Amongst others, the authors had swapped two cell-type labels, which explained the incorrect hierarchy. We experienced that scHPL is a great tool for discovering such misannotations. Cells can be relabeled based on this unexpected hierarchy.

Besides being subjective and time-consuming, another problem with manual annotation is a missing naming convention for cell types. CELLxGENE resolves this problem by forcing users to use Cell Ontology terminology when uploading their datasets. A downside of the Cell Ontology is that this hierarchy only consists of names but lacks information about the cell type, such as its function, morphology, or transcriptomic profile. Consequently, cell types from different datasets with the same name could have a different underlying expression pattern. The most straightforward solution might seem to add marker genes to Cell Ontology,

which can be used to identify cell types. In the benchmark in Chapter 2, however, we noticed that methods relying on marker genes perform worse during cell-type identification, most likely because of the sparsity of scRNA-seq data.

Ideally, all datasets from a similar tissue in CELLxGENE are not harmonized based on the names but based on the expression profile in a data-driven way using tools such as scHPL and treeArches. These tools can be enhanced by reflecting (inter-individual) variation in the width of a branch and allowing for fuzzy labels at the leaf nodes. The growing amount of data poses a challenge and as such both methods must become computationally more efficient. The resulting reference atlases should be updated continuously with newly generated data.

## 8.3 Automatically detecting new cell types

Even though many reference atlases are being constructed [6,21,24], these will never be complete since rare and diseased cell types might be missing. In the human lung cell atlas, for instance, six rare cell types were not defined in any of the individual datasets and had not been defined in the lung before, but could be discovered when combining multiple datasets [6]. Besides, new viruses, such as SARS-Cov-2, can infect cells from different tissues and perturb these cells [30,31]. Identifying such diseased cell types is important for drug or therapy development. Adding such data to a reference atlas leads to new insights in both healthy and diseased samples.

To detect rare or diseased cell types automatically, a classifier needs a rejection option. In Chapter 2, we benchmarked the rejection options of scRNA-seq cell-type identification methods by removing a cell type completely from the data. Here, we noticed that the linear SVM, which had the highest classification performance, performed poorly since it relied on the posterior probability. In Chapters 3 and 4, we introduced scHPL and improved the rejection option by incorporating distance metrics. This improved the detection of unknown cells but still did not perform perfectly. Diseased cells, such as inflamed monocyte-derived macrophages, are immediately rejected (labeled “unknown”) instead of labeled as internal node (e.g. macrophages), which would be preferred.

A hierarchical classifier that can return internal nodes of the hierarchy, so-called “partial rejection”, is beneficial according to a recent benchmark [32]. Here, they only evaluated how a full or partial rejection option affected the classification performance and not whether new cell types could be detected. Detecting new cell types using reference atlases should be benchmarked properly in upcoming benchmarks. An example experiment would be to remove one cell type from the training data and test whether the classifier correctly rejects cells from that cell type in the test dataset.

Ideally, cell-type identification and data integration methods should be benchmarked simultaneously. Data integration considerably influences whether these new cell types can be detected. During data integration biological variation should be preserved and technical variation should be removed. If the difference between a diseased and healthy cell type of two samples is seen as a technical artifact, this difference can be removed as well. Regardless

of the cell-type identification method used afterwards, the cell type will never be detected as a new cell type. Ideally, a diseased and healthy sample are sequenced together, so there are no batch effects. As such the difference between biological and technical variations between the reference atlas and these new samples can be detected more easily [33].

## 8.4 Towards cell-type-specific sequence-based models

Studying tissues untargeted and at a high resolution using scRNA-seq has led to the discovery of many new cell types. Since these cell types are defined based on their transcriptional profile, the underlying transcriptional regulation must be unique for every cell type. In Chapter 6, we aimed to unravel these cell-type-specific mechanisms by training sequence-based models using scRNA-seq data with the corresponding cell-type labels to predict gene expression. In Chapter 7, we focused on alternative splicing mechanisms by training models to predict cell-type-specific exon inclusion in the brain. Interpreting which motifs guide the model to make certain predictions, increases our understanding of the biological mechanisms underlying transcriptional regulation and alternative splicing.

Furthermore, these models aid in understanding how variants affect a cell type. Approximately 95% of the GWAS variants fall in non-coding regions [34]. Usually, only an association between a group of variants and a trait is discovered, but it remains unclear which variant causes a trait due to linkage disequilibrium, through which mechanism a variant acts, and which cell type is most disrupted. Models that use the genome to predict, for instance, transcription or splicing in a cell-type-specific way can address these problems.

In Chapter 6, we showed that cell-type-specific models always outperformed the tissue-specific models when predicting cell-type-specific gene expression levels. The difference in performance becomes most apparent if a tissue and cell type are dissimilar. Even though this increase was significant, we were unable to pinpoint what caused this increase such as cell-type-specific transcription factor binding sites.

To reliably predict the cell-type-specific effect of variants, our models, as well as other state-of-the-art sequence-based models, such as Enformer [35] and SpliceAI [36], have to overcome several limitations: 1) missing cell-type-specificity, 2) ignoring distal regulatory elements, and 3) incorrectly predicting personalized gene expression. I will discuss these limitations and potential solutions in the coming sections.

## 8.5 Missing cell-type-specificity of sequence-based models

The cell-type-specificity or tissue-specificity of sequence-based models is not thoroughly evaluated. Enformer is trained on 5,313 genomic tracks including different tissues and measurement techniques such as CAGE and DNase-seq reads, and predicts different values for very dissimilar cell types, such as keratinocytes and monocytes. However, an evaluation for more similar tracks, such as 77 CAGE tracks related to the brain, is missing. We noticed the same for Pangolin [37], a model to predict tissue-specific splicing. Pangolin outperformed

SpliceAI, the tissue-agnostic model, but no tissue-specific regulatory elements were discussed. Ideally, the models should be evaluated using cell-type-specific variants, but the ground truth for most variants is missing. A missing ground truth makes proper benchmarking impossible. A feasible alternative is to evaluate the models' performance on marker genes for specific cell types or whether the models correctly learn in which cell type a gene is higher expressed using for instance the log-fold change or the difference between two cell types.

Exploiting the current models as pre-trained models could be beneficial for learning cell-type-specific mechanisms. Some cell types are so similar that it is challenging to train a complete model with millions of parameters from scratch to learn these subtle differences. Seq2cells, for instance, extracts an embedding from Enformer and trains a simple model, a multi-layer perceptron, to predict the cell-type-specific gene expression [16]. Seq2cells assumes that all regulatory features are stored in the embeddings and the simpler model only needs to learn how to combine these during the fine-tuning step.

## 8.6 Limited context of sequence-based models

In our models, the region around the transcription start site and splice sites contributed most to the predictions of gene expression and exon inclusion. These regions are most important for transcription and splicing since RNA polymerase and the spliceosome bind there respectively. However, this signal dominates the predictions entirely, and as such the predicted effect of mutations further away is negligible. While mutations in enhancers far away or deep intronic variants can cause a disease [38–40]. A recent benchmark showed that other models do not capture distal regulatory elements either [41]. Even though Enformer inputs a sequence of 196kb, it incorrectly predicts the effect of variants in distal regulatory elements.

For splicing models, this has yet to be investigated, but since the model architectures and training strategies are similar, we can assume the models suffer here as well. Interestingly, SpliceAI, which inputs 10kb around the splice sites, was recently outperformed by Splam [42], a model that only uses 400 bp, indicating that regions further away might not be needed to predict splicing accurately. However, SpliceAI and Splam are both classification methods that predict whether a certain site is a splice junction instead of how often the junction is used. Distal variants may affect the latter more.

## 8.7 Sequence-based models are data-hungry

Current sequence-based models still suffer from limited training data. For instance, only a few genes are cell-type-specific or regulated by distal regulatory elements. Few examples in the training data make it difficult for models to learn the patterns. However, the number of genes or exons in the human genome limits the size of the training data, so this cannot be easily increased. To overcome this, several models, including Enformer, are trained on human and mouse data simultaneously to increase the size of the training data [35,43]. The weights of the first layers in the model are shared across the species exploiting that regulatory elements are partially conserved. The final fully connected layer is species-specific to allow learning of

species-specific mechanisms as well. In Chapter 7, we applied this trick when training exon-inclusion models, which improved their performance. In general, the current models only combine human and mouse, while data from more closely related species is available. For instance, for a cell-type-specific model predicting gene expression in the brain, scRNA-seq data from five primates could be combined [44].

## 8.8 Personalized sequence-based models

The third limitation is that current sequence-based models cannot predict variation of gene expression across individuals yet [45,46]. Ideally, for every individual genome, these models would predict the correct expression level, i.e. making personalized predictions. However, when evaluating models using variants found across individual genomes, Enformer predicted the wrong direction of effect for one-third of the tested variants. We did not evaluate making personalized predictions in our models, but since our models rely on models that were evaluated in the benchmark, we assume they incorrectly predict this as well.

State-of-the-art expression and splicing prediction models are all trained on the reference genome. However, the predicted genomic features were measured in individuals with specific variants in their genomes. Recent benchmarks suggested that training on individual genomes could improve personalized gene expression predictions [45,46]. Training on individual genomes might enhance learning of the effect of distal regulatory elements as well because of the increased variance in the training data. Recently, BigRNA [47] was released which predicts gene expression in 51 tissues for 70 individuals. For each individual, both haplotypes are input to identical instances of the model and the output is combined. Their results look promising, but the personalized gene expression task has not been evaluated for this model yet.

## 8.9 What should sequence-based models predict?

One might also question whether predicting gene expression or exon inclusion directly from the sequence is the most optimal approach to reach the goal of predicting the effect of mutations. Measurement techniques are noisy and the measured gene expression does not directly reflect how often a gene is transcribed in a cell. A gene can be highly transcribed but rapidly degraded as well due to (aberrant) splicing isoforms. Also in healthy tissues or cell types, alternative splicing is a way to control gene expression levels [48,49]. If the inclusion of an exon activates nonsense-mediated decay, this exon might not be measured or only in low levels even though it was originally highly included. An alternative would be to train the models on RNA-sequencing data of samples where nonsense-mediated decay was blocked, but this data is scarce.

Instead of predicting gene expression directly, it might be beneficial to predict intermediate layers, such as chromatin accessibility. Models trained to predict cell-type-specific chromatin accessibility in the drosophila brain [50] or for human melanoma [51] could be used to design cell-type-specific enhancers [52]. These models are not limited by the number of genes in the genome but are trained on differentially accessible regions between cell types.

This increased the size of the training data and might explain the cell-type-specificity of the models. However, the designed enhancers are only 500 bp. The effect of these enhancers was tested using a luciferase assay which means that these enhancers are inserted before the transcription start site of the luciferase gene. The effect of distal enhancers is thus not tested during the design. ExPecto [53] and their recent successor ExPectoSC [54] try to overcome this by first predicting 2002 regulatory features for the 40kb region around the transcription start site and using this to train a simpler model to predict gene expression.

An alternative could be to input chromatin accessibility measurements, or similar regulatory features, to the models [55]. This improves the cell-type-specificity since the input data is different now for every cell type or tissue. Another advantage is that these models can extrapolate to new cell types as long as chromatin accessibility data is available for that cell type. Evaluating the effect of variants or model interpretation becomes more complicated though since the input sequence cannot be *in-silico* mutated anymore as it is unknown how a mutation will affect the chromatin accessibility input track.

ENCODE-rE2G [56] combines a cell-type-specific input with an interesting training strategy: instead of training on healthy data, the model is trained on perturbation data. This logistic regressor predicts whether an element, a part of the DNA sequence, regulates a gene based on extracted features from the cell-type-specific DNase and cell-type-agnostic features, such as the distance between the element and the gene of interest. Since the model learns the relation between an element and the gene, it is not biased towards features close to the transcription start site and learns distal regulatory elements as well. However, they assume that a variant that falls in an element is always linked to the gene, and the direction of effect is not predicted. Instead of using the extracted features, a sequence-based model with a similar training strategy might be beneficial here.

## 8.10 Final remarks

Single-cell RNA sequencing has revolutionized our understanding of heterogeneous tissues. In this thesis, we presented several methods to automatically identify cell types in scRNA-seq data and use scRNA-seq data to increase the resolution of current sequence-based models. However, when analyzing scRNA-seq data, or using this data to train sequence-based models, we should remember that cells or cell types are not isolated compartments, but that they interact and communicate with each other. Many spatial transcriptomics datasets are now generated to focus on this. Ideally, we integrate this spatial information into the sequence-based models.

Not only do neighboring cells influence which genes are expressed, but the expression of other genes in a cell can influence the gene of interest as well. A more holistic view might be needed instead of predicting the expression of one gene at a time. Also when predicting splicing, we know that exons are very often coordinated. Using a different transcription start site might determine the complete isoform used. Predicting the inclusion of individual exons might be very difficult or near impossible in such a case. Ideally, sequence-based models would predict the expression of multiple isoforms simultaneously in the future.

# Bibliography


1. Zeisel A, Hochgerner H, Lönnerberg P, Johnsson A, Memic F, van der Zwan J, et al. Molecular Architecture of the Mouse Nervous System. Cell. 2018;174: 999–1014.e22. doi:10.1016/j.cell.2018.06.021
2. Hodge RD, Bakken TE, Miller JA, Smith KA, Barkan ER, Graybuck LT, et al. Conserved cell types with divergent features in human versus mouse cortex. Nature. 2019; 1–8. doi:10.1038/s41586-019-1506-7
3. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, de Rivera H, et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell. 2018;174: 1015–1030.e16. doi:10.1016/j.cell.2018.07.028
4. Bakken TE, Jorstad NL, Hu Q, Lake BB, Tian W, Kalmbach BE, et al. Comparative cellular analysis of motor cortex in human, marmoset and mouse. Nature. 2021;598: 111–119. doi:10.1038/s41586-021-03465-8
5. Siletti K, Hodge R, Mossi Albiach A, Lee KW, Ding S-L, Hu L, et al. Transcriptomic diversity of cell types across the adult human brain. Science. 2023;382: eadd7046. doi:10.1126/science.add7046
6. 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
7. Kimble J, Hirsh D. The postembryonic cell lineages of the hermaphrodite and male gonads in Caenorhabditis elegans. Dev Biol. 1979;70: 396–417. doi:10.1016/0012-1606(79)90035-6
8. Sulston JE, Horvitz HR. Post-embryonic cell lineages of the nematode, Caenorhabditis elegans. Dev Biol. 1977;56: 110–156. doi:10.1016/0012-1606(77)90158-0
9. 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
10. Fleck JS, Camp JG, Treutlein B. What is a cell type? Science. 2023;381: 733–734. doi:10.1126/science.adf6162
11. Chakravarthy H, Gu X, Enge M, Dai X, Wang Y, Damond N, et al. Converting Adult Pancreatic Islet α Cells into β Cells by Targeting Both Dnmt1 and Arx. Cell Metab. 2017;25: 622–634. doi:10.1016/j.cmet.2017.01.009
12. Raphael I, Joern RR, Forsthuber TG. Memory CD4+ T Cells in Immunity and Autoimmune Diseases. Cells. 2020;9. doi:10.3390/cells9030531
13. Van Der Wijst MGP, Brugge H, De Vries DH, Deelen P, Swertz MA, Franke L. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018;50: 493–497. doi:10.1038/s41588-018-0089-9
14. Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022;40: 245–253. doi:10.1038/s41587-021-01033-z
15. Yuan H, Kelley DR. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. Nat Methods. 2022;19: 1088–1096. doi:10.1038/s41592-022-01562-8
16. Schwessinger R, Deasy J, Woodruff RT, Young S, Branson KM. Single-cell gene expression prediction from DNA sequence at large contexts. bioRxiv. 2023. p. 2023.07.26.550634. doi:10.1101/2023.07.26.550634
17. Mallik S, Zhao Z. Multi-Objective Optimized Fuzzy Clustering for Detecting Cell Clusters from Single-Cell Expression Profiles. Genes . 2019;10. doi:10.3390/genes10080611
18. Wang J, Xia J, Tan D, Lin R, Su Y, Zheng C-H. scHFC: a hybrid fuzzy clustering method for single-cell RNA-seq data optimized by natural computation. Brief Bioinform. 2022;23. doi:10.1093/bib/bbab588
19. van der Wijst MG, de Vries DH, Groot HE, Trynka G, Hon C-C, Bonder M-J, et al. The single-cell eQTLGen consortium. Elife. 2020;9. doi:10.7554/eLife.52155
20. Domcke S, Shendure J. A reference cell tree will serve science better than a reference cell atlas. Cell. 2023;186: 1103–1114. doi:10.1016/j.cell.2023.02.016
21. 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
22. Megill C, Martin B, Weaver C, Bell S, Prins L, Badajoz S, et al. cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. bioRxiv. 2021. p. 2021.04.05.438318. doi:10.1101/2021.04.05.438318
23. CZI Single-Cell Biology Program, Abdulla S, Aevermann B, Assis P, Badajoz S, Bell SM, et al. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. bioRxiv. 2023. p. 2023.10.30.563174. doi:10.1101/2023.10.30.563174
24. Novella-Rausell C, Grudniewska M, Peters DJM, Mahfouz A. A comprehensive mouse kidney atlas enables rare cell population characterization and robust marker discovery. iScience. 2023;26: 106877. doi:10.1016/j.isci.2023.106877
25. Yazar S, Alquicira-Hernandez J, Wing K, Senabouth A, Gordon MG, Andersen S, et al. Single-cell eQTL mapping identifies cell type-specific genetic control of autoimmune disease. Science. 2022;376: eabf3041. doi:10.1126/science.abf3041
26. Sirkis DW, Warly Solsberg C, Johnson TP, Bonham LW, Sturm VE, Lee SE, et al. Single-cell RNA-seq reveals alterations in peripheral CX3CR1 and nonclassical monocytes in familial tauopathy. Genome Med. 2023;15: 53. doi:10.1186/s13073-023-01205-3
27. Zhu H, Chen J, Liu K, Gao L, Wu H, Ma L, et al. Human PBMC scRNA-seq-based aging clocks reveal ribosome to inflammation balance as a single-cell aging hallmark and super longevity. Sci Adv. 2023;9: eabq7599. doi:10.1126/sciadv.abq7599
28. 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

29. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with Harmony. Nat Methods. 2019;16: 1289–1296. doi:10.1038/s41592-019-0619-0

30. Liao M, Liu Y, Yuan J, Wen Y, Xu G, Zhao J, et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. Nat Med. 2020;26: 842–844. doi:10.1038/s41591-020-0901-9

31. 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

32. Theunissen L, Mortier T, Saeys Y, Waegeman W. Uncertainty-aware single-cell annotation with a hierarchical reject option. bioRxiv. 2023. p. 2023.09.25.559294. doi:10.1101/2023.09.25.559294

33. Dann E, Cujba A-M, Oliver AJ, Meyer KB, Teichmann SA, Marioni JC. Precise identification of cell states altered in disease using healthy single-cell references. Nat Genet. 2023;55: 1998–2008. doi:10.1038/s41588-023-01523-7

34. Leslie R, O'Donnell CJ, Johnson AD. GRASP: analysis of genotype-phenotype results from 1390 genome-wide association studies and corresponding open access database. Bioinformatics. 2014;30: i185–94. doi:10.1093/bioinformatics/btu273

35. 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

36. 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

37. Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biol. 2022;23: 103. doi:10.1186/s13059-022-02664-4

38. Vaz-Drago R, Custódio N, Carmo-Fonseca M. Deep intronic mutations and human disease. Hum Genet. 2017;136: 1093–1111. doi:10.1007/s00439-017-1809-4

39. Smemo S, Campos LC, Moskowitz IP, Krieger JE, Pereira AC, Nobrega MA. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. Hum Mol Genet. 2012;21: 3255–3263. doi:10.1093/hmg/dds165

40. Claringbould A, Zaugg JB. Enhancers in disease: molecular basis and emerging treatment strategies. Trends Mol Med. 2021;27: 1060–1073. doi:10.1016/j.molmed.2021.07.012

41. Karollus A, Mauermeier T, Gagneur J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. Genome Biol. 2023;24: 56. doi:10.1186/s13059-023-02899-9

42. Chao K-H, Mao A, Salzberg SL, Pertea M. Splam: a deep-learning-based splice site predictor that improves spliced alignments. bioRxiv. 2023. p. 2023.07.27.550754. doi:10.1101/2023.07.27.550754

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

44. Jorstad NL, Song JHT, Exposito-Alonso D, Suresh H, Castro-Pacheco N, Krienen FM, et al. Comparative transcriptomics reveals human-specific cortical features. Science. 2023;382: eade9516. doi:10.1126/science.ade9516

45. Huang C, Shuai RW, Baokar P, Chung R, Rastogi R, Kathail P, et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. Nat Genet. 2023;55: 2056–2059. doi:10.1038/s41588-023-01574-w

46. Sasse A, Ng B, Spiro AE, Tasaki S, Bennett DA, Gaiteri C, et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. Nat Genet. 2023;55: 2060–2064. doi:10.1038/s41588-023-01524-6

47. Celaj A, Gao AJ, Lau TTY, Holgersen EM, Lo A, Lodaya V, et al. An RNA foundation model enables discovery of disease mechanisms and candidate therapeutics. bioRxiv. 2023. p. 2023.09.20.558508. doi:10.1101/2023.09.20.558508

48. Zheng S. Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression. Int J Dev Neurosci. 2016;55: 102–108. doi:10.1016/j.ijdevneu.2016.03.003

49. Nickless A, Bailis JM, You Z. Control of gene expression through the nonsense-mediated RNA decay pathway. Cell Biosci. 2017;7: 26. doi:10.1186/s13578-017-0153-7

50. Janssens J, Aibar S, Taskiran II, Ismail JN, Gomez AE, Aughey G, et al. Decoding gene regulation in the fly brain. Nature. 2022;601: 630–636. doi:10.1038/s41586-021-04262-z

51. Atak ZK, Taskiran II, Demeulemeester J, Flerin C, Mauduit D, Minnoye L, et al. Interpretation of allele-specific chromatin accessibility using cell state-aware deep learning. Genome Res. 2021; gr.260851.120. doi:10.1101/gr.260851.120

52. Taskiran II, Spanier KI, Dickmänken H, Kempynck N, Pančíková A, Ekşi EC, et al. Cell type directed design of synthetic enhancers. Nature. 2023. doi:10.1038/s41586-023-06936-2

53. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG. Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat Genet. 2018;50: 1171–1179. doi:10.1038/s41588-018-0160-6

54. Sokolova K, Theesfeld CL, Wong AK, Zhang Z, Dolinski K, Troyanskaya OG. Atlas of primary cell-type-specific sequence models of gene expression and variant effects. Cell Rep Methods. 2023;3: 100580. doi:10.1016/j.crmeth.2023.100580

55. Zhang Z, Feng F, Qiu Y, Liu J. A generalizable framework to comprehensively predict epigenome, chromatin organization, and transcriptome. Nucleic Acids Res. 2023;51: 5931–5947. doi:10.1093/nar/gkad436

56. Gschwind AR, Mualim KS, Karbalayghareh A, Sheth MU, Dey KK, Jagoda E, et al. An encyclopedia of enhancer-gene regulatory interactions in the human genome. bioRxiv. 2023. p. 2023.11.09.563812. doi:10.1101/2023.11.09.563812