# Bayesian Methods in Transcriptomics

**Inauguraldissertation**

zur

Erlangung der Würde eines Doktors der Philosophie

vorgelegt der

Philosophisch-Naturwissenschaftlichen Fakultät

der Universität Basel

von

**Pascal Grobecker**

2024

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät auf Antrag von

Erstbetreuer: Prof. Dr. Erik van Nimwegen
Zweitbetreuerin: Prof. Dr. Mihaela Zavolan
Exerner Experte: Prof. Dr. Wolfgang Huber

Basel, 14.12.2021

Prof. Dr. Marcel Mayor
Dekan

আমার প্রিয়তমার জন্য, মিমি।

# Abstract

Transcriptomics techniques provide expression measurements across all genes and are therefore crucial for characterising and understanding cellular states in multicellular organisms. The dominant technique in the last decade has been RNA-seq, which can either be applied in bulk or in single cells. For the former, researchers are often interested in identifying marker genes that can be used in subsequent studies to differentiate between two or more classes of samples (e.g. cell types). We developed a novel statistical model for identifying such marker genes from RNA-seq data. Our model is based on a conditional entropy score that works well even when the number of gene expression measurements per class is small and when more than two groups were compared. Single-cell RNA-seq has become a popular experimental method to study variation of gene expression within a population of cells. A main application of scRNA-seq is to obtain an exhaustive picture of the variation in cell types that exist within a given tissue by clustering cells into subsets with distinct gene expression patterns. One challenge to such analysis is that the measured gene expression states of single cells are subject to a large amount of unwanted noise from inherent stochastic fluctuations due to the small mRNA numbers as well as technical noise from the experiment. Existing computational pipelines often try to disentangle these unwanted sources of noise from genuine biological signals by applying several layers of ad hoc steps including feature selection, normalisation, and dimensionality reduction, before clustering cells into subtypes. However, such pre-processing can dramatically distort the measurements by erroneously filtering true biological variability and introducing artefactual correlations. Here we propose a new computational method, called cellstates, that takes raw UMI counts of an scRNA-seq experiment as input and rigorously models the structure of both biological and experimental noise to find maximally resolved clusters of cells, i.e. groups of cells whose gene expression states are statistically indistinguishable. The cellstates method has no tuneable parameters, automatically optimises the number of clusters and returns directly interpretable results, thereby overcoming many issues of other available tools. In addition, cellstates also provides a data analysis toolbox that allows to place the cellstates within a hierarchy and identify differentially expressed genes at each level of this hierarchy, and several novel data visualizations.

# Table of Contents

# 1 Introduction

## 1.1 General Introduction and Outline of Thesis[1]

**What is transcriptomics?** The human body is made up of an estimated number of $3 \times 10^{13}$ cells [56]. Every one of these cells has evolved to perform a highly specialised function in the organism. As a result, there is an immense variety of shapes, sizes and molecular make-up of these different cells. For example, skeletal muscle cells are elongated and contain large arrays of myofilament protein fibres that allow the muscle to contract. Red blood cells are small, round and flexible and contain large amounts of haemoglobin that is used to transport oxygen through the body. Nerve cells have a compact core with many branched protrusions and use electrical potentials across their membrane to transmit information across the body. While it is clear that cells can be grouped into such cell types by shared characteristic features, there is no agreed on definition (see discussion in Chapter 4). What is clear is that in order to systematically study what cell types exist, what functions they perform, or how they are affected by diseases, we need to first measure their molecular make-up. Fundamentally, proteins are the biological molecules performing most cellular functions. For example, haemoglobin is a protein that red blood cells use to transport oxygen; actin and myosin are two of the proteins that form the fibres in muscle cells and help it generate movement. Each one of these proteins is encoded as a gene in the genome, which is a DNA sequence shared by all cells of an organism. To make proteins, this gene sequence is first copied from DNA to one or several molecules of messenger RNA (mRNA) in the process of *transcription*. These mRNA copies are then used in turn to make many copies of the proteins they encode in the process of *translation*. Thus, if a cell needs a certain amount of a specific protein to perform its function, it needs to regulate both transcription and translation of a gene. To fully characterize a cell type, we would ideally want to measure the entire protein content in order to understand this regulation. The field of *proteomics* tries to address such measurements and has been making rapid advances in recent years. However, the chemical properties of proteins are very diverse, and it is therefore still challenging and expensive to obtain such data. In contrast, mRNAs are chemically all

[1] This section is intended as an introduction to the thesis for a general audience. It will therefore necessarily contain many simplifications. To every rule in biology, there is an exception.

very similar and next generation sequencing technologies have made it possible to study the mRNA content across all genes of samples at a large scale. The technique, called RNA-seq, works by extracting mRNAs from a sample, determining the sequence of each of these mRNAs and then matching them to the known gene sequences. A more detailed description of the protocol is presented in Section 1.2. Thus, the result is a list of gene expression levels across all genes. The fundamental premise is that the number of mRNA copies for a gene is predictive for the amount of the corresponding protein in the cell and therefore tells us most of what we would want to know. This field of studying the entire RNA content of cells is called *transcriptomics*.

**What are the Bayesian methods for?** Astonishingly, only approximately 20'000 protein coding genes [15] are needed to generate the huge variety of cells present in our bodies and RNA-seq now allows us to quantitatively measure the expression of all of these genes at the same time. In contrast, traditional methods for studying gene expression could only look at a handful of genes at a time. On the one hand, this means that we can now get a much more unbiased view of gene expression in cell types, as we do not have to choose in advance which genes to study. On the other hand, this number of genes is much larger than what previous data analysis methods had to deal with. As a consequence, we need new statistical methods for correctly analysing such large, complex datasets. The kind of models we developed are based on what is called Bayesian statistics[2]. Furthermore, there is a definite need for implementing these analyses computationally. Such bioinformatics tools help its users to gain biologically relevant insights into their data and allow them to make predictions for further studies. There exist two main kinds of RNA-seq experiments: in *bulk* and in *single cells*. We will address them in the next two paragraphs, along with a statistical model that we developed for their analysis.

**Finding marker genes from bulk RNA-seq data** In a bulk RNA-seq experiment, mRNA is extracted from all cells in a sample, pooled and then sequenced. We can then compare these results across different samples. One question of interest is often what genes have different expression levels between two conditions. If we have replicate samples for each condition, we can estimate the level of gene expression for each gene in each condition. This estimate will have some associated statistical uncertainty. Then, we can calculate for each gene the difference in expression levels between the conditions – and how likely it is that there is no significant difference at all. Many statistical models exist already for finding such differentially expressed genes (DEG) and have been well-established. Here, we developed a model that solves the slightly different, but related, problem of finding the most predictive *marker genes*. These are genes that can be used in experiments to identify a certain type of

[2] What makes a statistical model Bayesian would be beyond this introduction and is not important for understanding the topics of this thesis. Briefly, in Bayesian statistics, probabilities are viewed as representing a state of knowledge about a system. Therefore, we can make statements like: "From our data, we conclude that the probability of this cell being a neuron is 99%." The opposing view, frequentist statistics, interprets probabilities as frequencies of events. Thus, it only allows statements like "Assuming that our cell is a neuron, we will see data like the one we measured in 99% of cases." We use Bayesian statistics as it is less confusing and has a more logical theoretical foundation.

cell. If high expression of a protein is associated with a certain type of neurons, for example, one could use a dye that recognizes this protein to stain all of these neurons in a brain sample and identify their locations and how they are connected. The difference between predicting good marker genes and DEGs is subtle: For marker genes, we care about how much uncertainty there is about a single future *measurement* of the expression level. In contrast, for DEGs, we care about the uncertainty of the *true* gene expression levels. The model is described in detail in Chapter 2. One other feature that distinguishes our model from established statistical models for DEG prediction is that ours works for comparing more than two conditions. Thus, if we have a gene that has high, medium and low expression levels in three experimental conditions, we can find it very easily.

**Finding cell states in single-cell RNA-seq data** If we want to study cell types with bulk RNA-seq, we need to experimentally isolate cells of this type. However, this might not always be possible, or it might be very laborious. In addition, only measuring the average expression levels could hide subpopulations of these cells. Finally, it only allows us to study known cell types, but not to discover new ones. To overcome all these limitations, single-cell RNA-seq (scRNA-seq) was developed, where gene expression is measured across individual cells and genes. That is, the result of such an experiment is a large data table where each row is a gene, each column a single cell and each entry a count for how many mRNAs of a given gene in a given cell were found. As the number of cells typically ranges from 1'000 to 100'000, this is a table with millions of entries. Analysing such data is challenging for another reason: in contrast to bulk RNA-seq measurements, the data is very noisy and not a precise measurement of the true gene expression levels. One common data analysis done on this data is *clustering*: the aim is to find groups of cells with similar expression patterns. These groups (clusters) can then be studied further and identified as known cell types, novel cell types or subpopulations of known cell types, etc. Many computational tools already exist to do such a clustering of scRNA-seq data, but they have several drawbacks. Firstly, they deal with the measurement noise in an *ad hoc* way. That is, they use a series of poorly motivated, complicated steps to pre-process the data to remove the noise before the clustering. However, this can easily introduce unwanted biases. Secondly, as a result, many of the available tools have a lot of parameters that need to be set correctly. To justify these parameter settings, researchers often simply check if the final results are plausible based on their expectations and biological knowledge. Thus, the clustering results are not an objective statement about the data, but can easily be made to fit these expectations. Finally, there is usually also no well-defined criterium for how similar cells should be within a cluster – this is also left to the judgment of the researcher. Starting from the premise the measurement noise in scRNA-seq data has a well-studied and well-understood mathematical form, we want to address these problems with our model that is described in Chapter 3. The main point of this model is that we define clusters as being groups of cells that are statistically indistinguishable. That is, the measured mRNA counts for these cells across all genes can be fully explained by assuming that they are identical and that any differences between them are just random noise. As a consequence, the clusters are as fine-grained as possible and follow directly from the given data. We call those clusters cell states. So the

results naturally tell us something about how diverse cells in a tissue are. For example, the brain is very complex and almost every neuron has a unique function – thus we find a lot of different cell states. Furthermore, the analysis is much simpler as there is no need to set *any* parameters. And importantly, we get results that are not biased by user inputs. One drawback of this method is that predicting many cell states can make it hard for a researcher to understand the results. Therefore, we also implemented a way to group similar cell states, which will help to understand the broader patterns in the data before the details can be studied.

## 1.2 A brief description of RNA-seq

While experimental methods for studying the transcriptome have been around since the early 1990s, RNA-seq technology was first developed in the mid 2000s as high-throughput DNA sequencing methods became available[25, 69]. In contrast to previous technologies, the sequencing of transcripts allows the detection of transcript variants and not just expression levels for known genes. This additional information can be used in many ways such as detection of splicing, genetic variants, transcript isoforms or novel genes. In this thesis, however, we are focussing only on the measurement of gene expression levels. In that regard, improvements such as having lower background noise and a better detection limit for RNAs with low abundance were crucial for the development of single-cell RNA-seq. Here, I will briefly describe the main aspects of RNA-seq and scRNA-seq that are relevant for our statistical methods. The main steps of scRNA-seq are also summarized in Figure 1.1. The first step is to dissociate the cells in a tissue and isolate them. A common method for cell isolation is the use of microfluidic droplets, as is done in the popular commercial 10X Chromium platform [77]. In such a system, each cell gets encapsulated in a tiny droplet that contains all the reagents needed for the steps before the mRNAs get pooled. In each cell, the RNA is extracted and reverse-transcribed to complimentary DNA (cDNA). By using poly-T primers for reverse-transcription, mRNAs that have a poly-A tail get selected while the far more abundant ribosomal RNAs are skipped. Also, established methods for next-generation sequencing of DNA can be used on these cDNA strands. The next step is to add two random DNA sequences to each cDNA: One will act as a unique molecule identifier (UMI) that is different for each cDNA molecule. The second one is a cell barcode that is the same for all transcripts in a cell. Some more recent methods such as SPLiT-seq [52] use direct combinatorial cell barcoding of transcripts and can therefore work entirely without isolation of cells. Next, the cDNA pool is amplified, i.e. duplicated several times, so that every strand is likely to be sequenced at least once. Then, the cDNAs are sequences. Each sequence contains 3 sections: A part of the original mRNA transcript, a cell barcode and a UMI. The transcript is mapped to the known genome to identify the gene it came from. The cell barcode identifies transcripts from the same cell. When several cDNAs with the same UMI, cell barcode and transcript sequence are found, we recognize that they came from one original molecule and only count them once, a process knows as de-duplication [63]. This shows the importance of the UMI tags: They allow us to count how many mRNAs of a given gene were captured in a given cell by the experiment. The amplification step

therefore should not introduce any additional noise to the data [26]. Experimental noise is introduced only through the random loss of molecules at each step. Overall, captured UMIs account for roughly 10-20% of cellular mRNAs, but that number can vary a lot depending on the experimental protocol used [60]. The final result of a scRNA-seq experiment is hence a UMI count table. Each row in this table corresponds to a cell, each column to a gene. The entries are counts for the number of distinct UMIs that were found, associated with the corresponding gene sequence and cell barcode. In other words, these counts correspond to a random fraction of cellular mRNAs that were captured. The resulting table only contains integers and is sparse (i.e. contains many zeros) as the typical total number of UMIs per cell is $\mathcal{O}(1 \times 10^3 - 1 \times 10^4)$ (see Figure 3.9), much lower than the number of genes. As only a small fraction of a cell's transcripts are captured, zero counts are often found for low expressed genes.

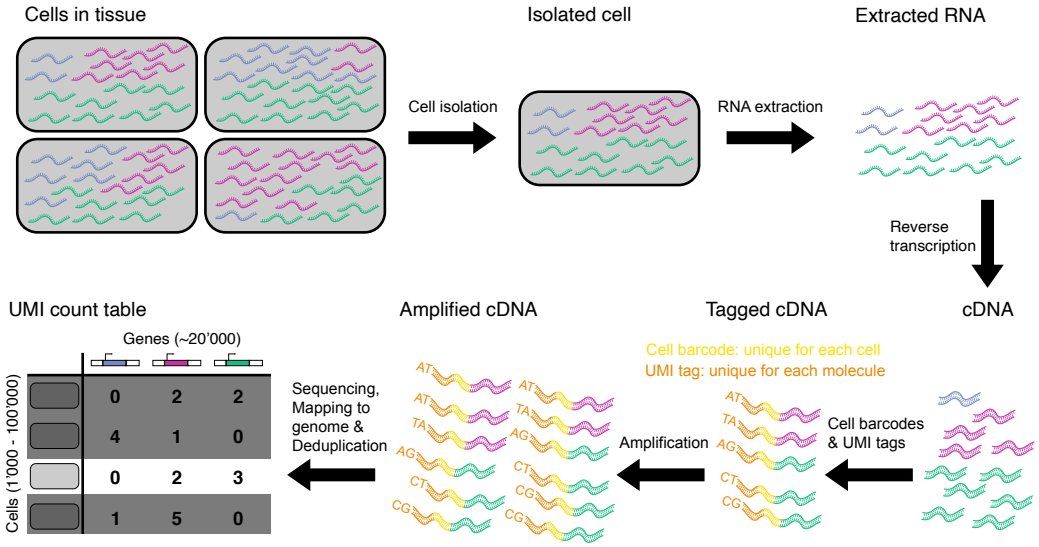



Figure 1.1: Summary of main steps in single-cell RNA sequencing. Cells in a tissue are first dissociated and then isolated. Then, their RNA content is extracted. The reverse transcription to cDNA is usually done with poly-T primers, which select mRNAs but not the far more abundant rRNA molecules. Furthermore, DNA sequencing technology is well established and can be used for the cDNAs. Each cDNA molecule gets tagged with two random sequences: a cell barcode sequence that is unique for each cell and a unique molecular identifier (UMI) which is different for each individual cDNA molecule. After tagging, all sequences get pooled, amplified and sequenced. The UMI sequences are then used to identify multiple sequenced copies of the same original cDNA strand and deduplicate those reads. The cell barcode assigns each sequence to a cell. And the transcript sequence is mapped to the genome to identify the gene from which it originates. Thus, the data can be summarized in a UMI count table, where each row is a cell, each column a gene and each entry a count of the corresponding UMIs found.

There are two important differences in bulk RNA-seq. Firstly, isolation of cells is skipped and mRNA content from all cells is pooled together. Secondly, tagging cDNAs with barcodes is not necessary as noise from cDNA amplification is not as much of a problem compared to single-cells as the initial pool of sequences would be several orders of magnitude larger in bulk. The result of a bulk RNA-seq experiment will be a vector of the number of transcript

sequences mapped to each gene. Unlike for scRNA-seq data, this vector will contain few zeros, and all zeros are likely to correspond to unexpressed genes. Such vectors from several sequenced samples can be combined into a table. To account for differences in the total number of transcripts per sample, this data is usually normalised to units of transcripts per million (TPM). Such a normalisation should not be done for single-cell data as the total number of transcripts per cell is much smaller and therefore differences between 0 and 1 UMI counts would be artificially inflated [61]. To conclude, the resulting normalised data table for bulk RNA-seq is dense and contains rational numbers.

# 2

# Identifying Marker Genes in Bulk RNA-seq

## 2.1 Introduction

To measure the transcriptome of a cell type of interest, single cell RNA-seq has, in recent years, become the state-of-the-art method, as it gives results not just for the mean gene expression, but also its distribution and can be used to characterise subtypes in the sample. However, bulk methods are still widely in use and provide benefits such as lower cost and a much better ability to measure low abundance transcripts. Because of the high number of reads and the low number of samples, data analysis is relatively straightforward with well-established tools.

Here, we establish a statistical model for identifying marker genes. These are genes, that show a characteristic expression in a class of samples of interest and can therefore be used in experiments, for example by antibody staining, to mark them. These classes could be, for example, different cell types or stages of development. To identify marker genes, we need to answer the following question: If we could only measure the expression of one gene in a sample, which one would be most predictive for its class? The mathematical measure to answer this question is the conditional entropy of the class $C$ given a gene expression measurement $X$, $H(C|X)$. This entropy is a measure of the uncertainty about which class a sample belongs after the gene expression has been measured. A low entropy corresponds to low uncertainty.

To make it concrete, let's look at an example from [43]. In this paper, they are studying the differentiation of neuronal stem cells (NSC) into basal progenitors (BP) and newborn neurons (NBN) from embryonic day 10.5 (E10.5) to birth (P0). These cell types were experimentally separated using fluorescence activated cell sorting (FACS) by expression of Hes5-GFP and separately bulk sequenced. The entire dataset consists of 2-4 replicate samples for a given cell type on a given day. Now, a researcher might be interested in being able to distinguish cell types on a sample from day E15.5, so that the three classes would be {BP(E15.5)}, {BP15.5}, and {NBN15.5}. This kind of problem cannot be addressed by looking at differentially expressed genes, as it involves more than two classes.

As our model was developed for RNA-seq data, we will discuss the model in this context and refer to $X$ as gene expression. However, there is nothing about this model specific to such data, and it could therefore potentially be applied in many other contexts. We will

discuss the assumptions and requirements to the data later.

## 2.2 Methods

### 2.2.1 Estimating gene expression variance

We want to model the log-expression of a gene $g$ in a class $c$ with a Gaussian distribution with mean $\mu_{gc}$ and variance $\sigma^2_{gc}$ (for now we will only look at data within one class, so we can drop the subscript $c$). The data $D_g$ consist of the $n_c$ log-expression measurements for that gene and are summarized by their empirical mean $m_g$ and variance $v_g$. The likelihood of the data under such a model is therefore:

$$P(D_g|\mu_g, \sigma^2_g) \propto \sigma^{-n_c} \exp\left[-n_c \frac{(m_g - \mu_g)^2 + v_g}{2\sigma^2_g}\right] \tag{2.1}$$

One problem of bulk RNA-seq data is that there are potentially very few samples in a given class, which makes it hard to directly estimate the variance of gene expression. Therefore, we need to estimate the variance using a Bayesian model, which makes marker gene inference more robust – a process called *variance stabilization*. In contrast to the low number of samples per class $n_c$ ($\mathcal{O}(10^0 - 10^1)$), the number of genes $n_g$ is high ($\mathcal{O}(10^4)$). Thus, we can directly estimate the distribution of variances from the data. To simplify the maths, we work with the inverse variance, or precision, $w_g = 1/\sigma^2_g$. Our first goal is to find the posterior distribution $P(w_g|D_g)$. We start by integrating over $\mu_g$ in Equation 2.1, we get:

$$P(D_g|w_g) = w_g^{(n_c-1)/2} \exp(-n_c v_g w_g/2). \tag{2.2}$$

We assume that the prior probability to have precision $w_g$ is given by a gamma distribution:

$$P(w_g) = \beta^\alpha w_g^{\alpha-1} \exp(-\beta w_g)/\Gamma(\alpha). \tag{2.3}$$

We chose the gamma distribution as it is the conjugate prior to the exponential distribution of Equation 2.2, which allows deriving a closed-form solutions of the posterior. The gamma distribution is also the maximum entropy distribution, given the mean of $w$ and the mean of $\log(w)$. Performing the integral $P(D_g|\alpha, \beta) = \int_0^\infty P(D_g|w_g)P(w_g)dw_g$, we obtain

$$P(D_g|\alpha, \beta) = \frac{\beta^\alpha}{(\beta + n_c v_g/2)^{\alpha+(n_c-1)/2}} \frac{\Gamma(\alpha + (n_c-1)/2)}{\Gamma(\alpha)}. \tag{2.4}$$

It is generally easier to work with the log-likelihood:

$$L_g(\alpha, \beta) = \alpha \log(\beta) - (\alpha + \frac{n_c - 1}{2}) \log(\beta + \frac{n_c v_g}{2}) + \log(\Gamma(\alpha + \frac{n_c - 1}{2})) - \log(\Gamma(\alpha)). \tag{2.5}$$

Note that the same values of $\alpha$ and $\beta$ are used for all genes, as they all share the same prior distribution over $w$. Hence, we fit the data by finding the values of $\alpha$ and $\beta$ that maximize the total log-likelihood over all genes $L(\alpha, \beta) = \sum_g L_g(\alpha, \beta)$. This leads to the following two equations that need to be satisfied at the optimum:

$$\frac{\alpha}{\beta} = \langle \frac{\alpha + (n_c - 1)/2}{\beta + n_c v_g/2} \rangle \tag{2.6}$$

$$\langle \psi(\alpha + \frac{n_c - 1}{2}) - \psi(\alpha) \rangle = \langle \log(1 + \frac{n_c v_g}{2\beta}) \rangle, \tag{2.7}$$

where the averages are over all genes and $\psi(x)$ is the digamma function. Equation 2.6 is for optimal $\beta$ given fixed $\alpha$, and vice versa for Equation 2.7. These two equations can be used to numerically compute the optimum in an iterative way: Fix $\alpha$, then set $\beta$ from Equation 2.6, then fix $\beta$ and set $\alpha$ from Equation 2.7, repeat until conversion. Having fixed the optimal values $\alpha^*$ and $\beta^*$, we can finally obtain the posterior probability of $w_g$ by combining Equations 2.2, 2.3, and 2.4 in Bayes' theorem $P(w_g|D_g, \alpha^*, \beta^*) = P(w_g|\alpha^*, \beta^*)P(D_g|w_g)/P(D_g|\alpha^*, \beta^*)$.

### 2.2.2 How to assess marker quality

Given this posterior distribution $P(w_{gc}|D_{gc})$, how do we best assign marker quality to a gene? If this distribution had a sharp peak, we could estimate the most likely true variance $\sigma_c^2$ of the gene in each class and use a Z-statistic $Z = (m_1 - m_2)^2/(\sigma_1^2 + \sigma_2^2)$ (we will now compare a single gene across classes and drop the subscript $g$). However, the whole point of the variance stabilization is that the posterior distribution may be quite broad when the number of samples $n_c$ is small, precisely when we need the stabilization of the variance in the first place.

Thus, we will assess how well a measurement of gene expression separates different classes $c$ of samples with the conditional entropy $H(C|X)$. That is, given a measurement of the log-expression $X$, how high is the entropy of the class $C$. The lower this entropy, the higher the level of certainty that the class is $C = c$, given a measurement of the expression level $X = x$. This conditional entropy $H(X|C)$ can be written in terms of the joint entropy $H(X, C)$ of the joint distribution $P(x, c)$, and the entropy $H(X)$ of the marginal distribution $P(x)$, i.e. $H(C|X) = H(X, C) - H(X)$.

We are going to estimate $P(x|c)$ based on our data $D_c$ for class $c$, i.e. the gene expression measurements of the samples in class $c$. So we will get an expression of the form $P(x|c) = P(x|D_c)$. This distribution, most formally, is given by

$$P(x|D_c) = \int d\mu_c dw_c \sqrt{\frac{w_c}{2\pi}} \exp\left(-\frac{w_c}{2}(x - \mu_c)^2\right) P(\mu_c, w_c|D_c). \tag{2.8}$$

Using $P(\mu_c, w_c|D_c) \propto P(D_c|\mu_c, w_c)P(w_c)P(\mu_c)$, and using the gamma-distribution prior that we fitted $P(w_c|\alpha_c^*, \beta_c^*)$ and the uniform prior for $\mu_c$, these integrals can be performed and we finally find

$$P(x|D_c) = \left(1 + \frac{(x - m_c)^2}{(n_c + 1)(v_c + \frac{2\beta^*}{n_c})}\right)^{-(\alpha^* + \frac{n_c - 1}{2})}. \tag{2.9}$$

As an aside, while this is not a Gaussian, it can be approximated by a Gaussian with mean $\mu_{\text{eff}} = m_c$ and effective variance

$$\sigma_{\text{eff}}^2 = \frac{(n_c + 1)(v_c + 2\beta/n_c)}{n_c - 1 + 2\alpha}. \tag{2.10}$$

This equation can be used to estimate the uncertainty in gene expression measurements, if we take the class $c$ to be all replicates of a certain condition. However, for the remainder of this derivation, the Gaussian approximation will not be used.
Back to our calculation of the joint entropy $H(X, C)$: for the prior $P(c)$, the obvious choice would be the uniform distribution $P(c) = \frac{1}{|C|}$, where $|C|$ is the number of classes. However, one might also want to include prior information, e.g. when one cell type is known to be much more abundant than another, so we will not specify it further. The expression for the joint entropy is

$$H(X, C) = -\sum_c \int dx P(x|c)P(c) \log\left[P(x|c)P(c)\right]. \tag{2.11}$$

This can be rewritten as

$$H(X, C) = H(C) - \sum_c P(c) \int dx P(x|c) \log[P(x|c)] = H(C) + \langle H(x|c)\rangle, \tag{2.12}$$

where $H(C) = \sum_c P(c) \log[P(c)]$ is the entropy of the classes, $H(X|C) = -\int dx P(x|c) \log[P(x|c)]$ is the entropy of the distribution $P(x|c)$, and the average indicated by the angle brackets is over all classes.

**Calculating the conditional entropy $H(X|C)$:** Note that the distribution $P(x|c)$ takes the mathematical form

$$P(x|c) = Z_c \left(1 - \frac{(x - m_c)^2}{V_c}\right)^{-\gamma_c}, \tag{2.13}$$

with $V_c = (n_c + 1)(v_c + 2\beta^*/n_c)$, $\gamma_c = \alpha^* + (n_c - 1)/2$, and the normalization constant is

$$Z_c = \frac{\Gamma(\gamma_c)}{\sqrt{\pi V_c}\Gamma(\gamma_c - 1/2)}. \tag{2.14}$$

So the integrals that we have to perform are

$$H(X|C) = -\log(Z_c) + \gamma_c Z_c \int dx \frac{\log\left[1 + \frac{(x-m_c)^2}{V_c}\right]}{\left(1 + \frac{(x-m_c)^2}{V_c}\right)^{\gamma_c}}. \tag{2.15}$$

If we introduce the change of variables $y = (x - m_c)/\sqrt{V_c}$, we find

$$H(X|C) = -\log(Z_c) + \gamma_c \sqrt{V_c} Z_c \int dy \frac{\log[1 + y^2]}{(1 + y^2)^{\gamma_c}}. \tag{2.16}$$

Now, finally, we note that if we defined

$$F(\gamma) = \int dy \frac{1}{(1 + y^2)^{\gamma}}, \tag{2.17}$$

this can be written as

$$H(X|C) = -\log(Z_c) - \gamma_c \sqrt{V_c} Z_c \frac{d}{d\gamma_c} F(\gamma_c). \tag{2.18}$$

Equation (2.17) can be analytically solved by $F(\gamma) = \frac{\sqrt{\pi}\Gamma(\gamma-1/2)}{\Gamma(\gamma)}$. Combining all this, we finally get for the entropy

$$H(X|C) = -\log(Z_c) + \gamma_c \left(\psi(\gamma_c) - \psi(\gamma_c - 1/2)\right), \tag{2.19}$$

where $\psi(x)$ is the digamma function, the logarithmic derivative of the gamma function.

**Calculating the marginal entropy $H(X)$:** The marginal distribution $P(x)$ is given by a mixture over all classes, i.e. $P(x) = \sum_c P(x|c)P(c)$. So the marginal entropy is given by $H(X) = -\int dx P(x) \log[P(x)]$. This can be written as

$$H(X) = \sum_c P(c) \int dx P(x|c) \log[\sum_c P(c)P(x|c)]. \tag{2.20}$$

There is no analytical solution to this integral, so it will have to be calculated numerically.

### 2.2.3 Misclassification error in binary case

If we have two classes (say one cell type vs the other cell types) and, given our data $x$ (a gene expression measurement), we define the probability that we assign this to the right class as $1-p$, and the probability that we assign this to the wrong class as $p$. So $p$ is the probability of making an error. The entropy of this distribution is $H(p) = -p\log(p) - (1-p)\log(1-p)$. As we know the entropy of our distribution $H(C|X)$, we want to find the inverse of the entropy function $H(p)$, i.e. we want to find the function $p(H)$ that calculates the error-probability from the entropy. While the entropy $H$ is the more general measure that works in all cases, in the common binary case, we can associate an error probability with the entropy in order to give a more intuitive meaning to its value. We are particularly interested in the regime where $p$ (the error probability) is small. For small $p$, we have $H(p) \approx -p(\log(p) - 1)$. In terms of the lower branch of the Lambert W function, this is solved by

$$p(H) \approx -H/W_{-1}(-H/e). \tag{2.21}$$

## 2.3 Results

### 2.3.1 Data

We tested our algorithm on a dataset, that at the time of writing has not been fully published yet. A small part of the data is published in [43] and the paper also contains details how the experiments were conducted. Briefly, transgenic mouse lines *Hes5::GFP* and *Tbr2::GFP* (*Tbr2* is also known as *Eomes*) were used to isolate pure populations of neural stem cells (NSC), basal progenitors (BP) and newborn neurons (NBN) from the developing cerebral cortex between embryonic day 10.5 (E10.5) and birth (PN). The separation of cell types was done through fluorescence activated cell sorting (FACS) based on the expression of the fluorescent reporter genes. NSC were identified as cells with *Hes5::GFP* expression, BP as *Tbr2::GFP*$^{high}$, and NBN as *Tbr2::GFP*$^{low}$. NSC were present from E10.5 to PN, BP from E12.5 to PN, and NBN from E15.5 to PN. For each cell type and time point, 2-4 biological

replicate samples were produced, resulting in a total of 70 samples that were each RNA sequenced.

### 2.3.2 Marker genes

To illustrate that the algorithm works, we are showing some example results for different comparisons across cell types and embryonic days. The first two are between cell types which have been selected by a distinct expression pattern of a gene (*Hes5* is a marker for NSC, low *Tbr2* expression distinguishes NBN from BP). Thus, the scores of these genes in the corresponding comparisons can be used to validate our approach. In the last comparison, we look at marker genes for NSC during gliogenesis for which we only have 9 data points, so we can check if the results look reasonable even when the observed variance might not reflect the real ones.

**Marker genes for neural stem cells.** To get marker genes for NSC, we are comparing data from NSC at all embryonic days against data from the other cell types. That is, the classes are $c_1 = \{\text{NSC E10.5 - PN}\}, c_2 = \{\text{BP E12.5-PN, NBN E15.5-PN}\}$. The top marker gene is *Hes5*, which is the gene that was used to experimentally isolate NSC. As can be seen in Figure 2.1, the top 12 marker genes all clearly separate NSC from other cell types.

**Marker genes for differentiating newborn neurons from basal progenitors.** Here, the classes are $c_1 = \{\text{BP E12.5 - PN}\}$ and $c_2 = \{\text{NBN E15.5-PN}\}$. The marker gene that was used to experimentally isolate NBN, *Tbr2*, appears at rank 38, with $H(C|X) = 1.60 \times 10^{-2}$ and $p = 2.25 \times 10^{-3}$ and is hence predicted to also be a significant marker gene. As can be seen in Figure 2.2(a), the top 12 marker genes all clearly separate NSC from other cell types. As comparison, *Tbr2* expression is shown in Figure 2.2(b) – it scores lower because of the relatively high variance in expression over time among NBN.

**Marker genes for neural stem cells during gliogenesis.** Neural stem cells go through three main developmental phases: expansion (E10.5-11.5) where the cells are replicating to expand the stem cell pool, neurogenesis (E12.5-16.5) where some of the NSC differentiate into BP to become neurons, and gliogenesis (E17.5-PN) where NSC generate non-neuronal brain cells, including astrocytes, oligodendrocytes, and ependymal cells [42]. Here, we want to look for the top markers of gliogenic NSC, so our classes are $c_1 = \{\text{NSC E10.5 - 16.5}\}$ and $c_2 = \{\text{NSC E17.5-PN}\}$. The top 12 marker genes are shown in Figure 2.3(a) and are all clearly genes that separate data taken during gliogenesis from those taken on other days.

**The importance of variance stabilisation.** The last example was chosen as the small number of gliogenic samples (9) makes variance stabilization important in order to correctly assess the true uncertainty in a new measurement of the gene expression. To illustrate this, we calculated the t-statistic without variance stabilization:

$$t^2 = \frac{m_1 - m_2}{v_1/n_1 + v_2/n_2}, \tag{2.22}$$

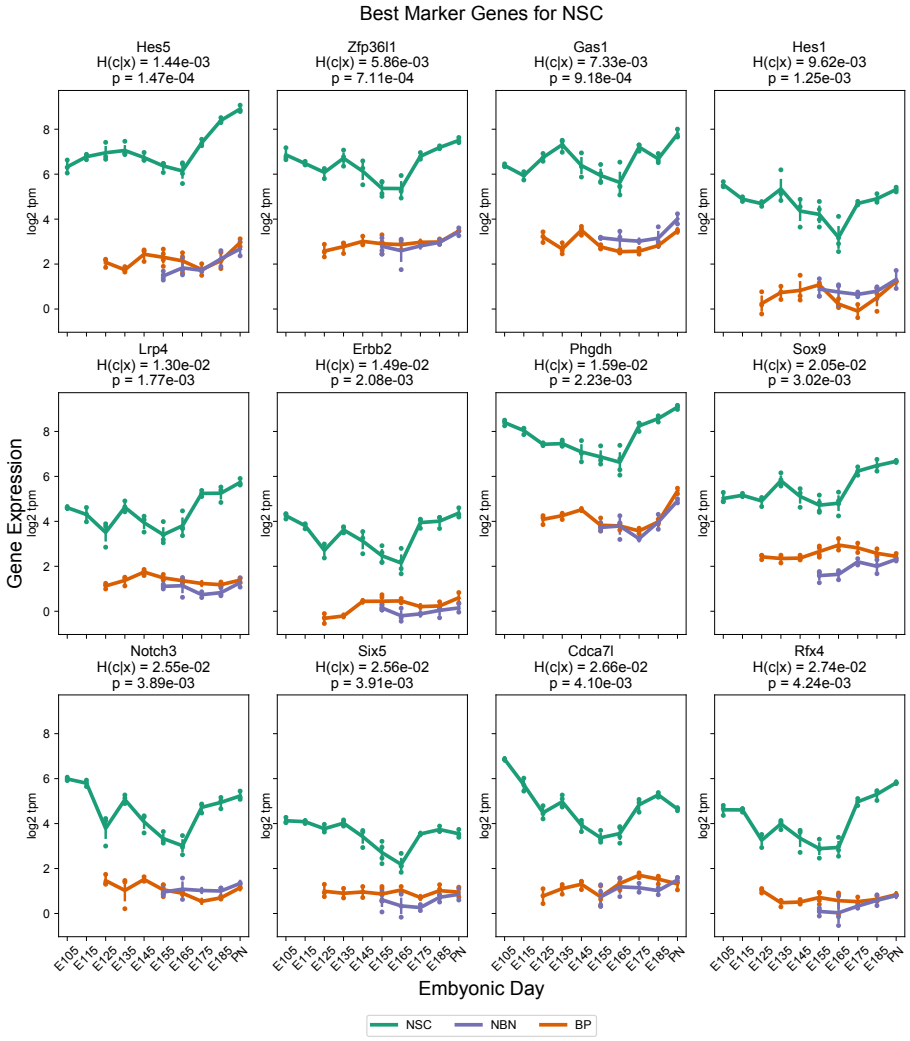


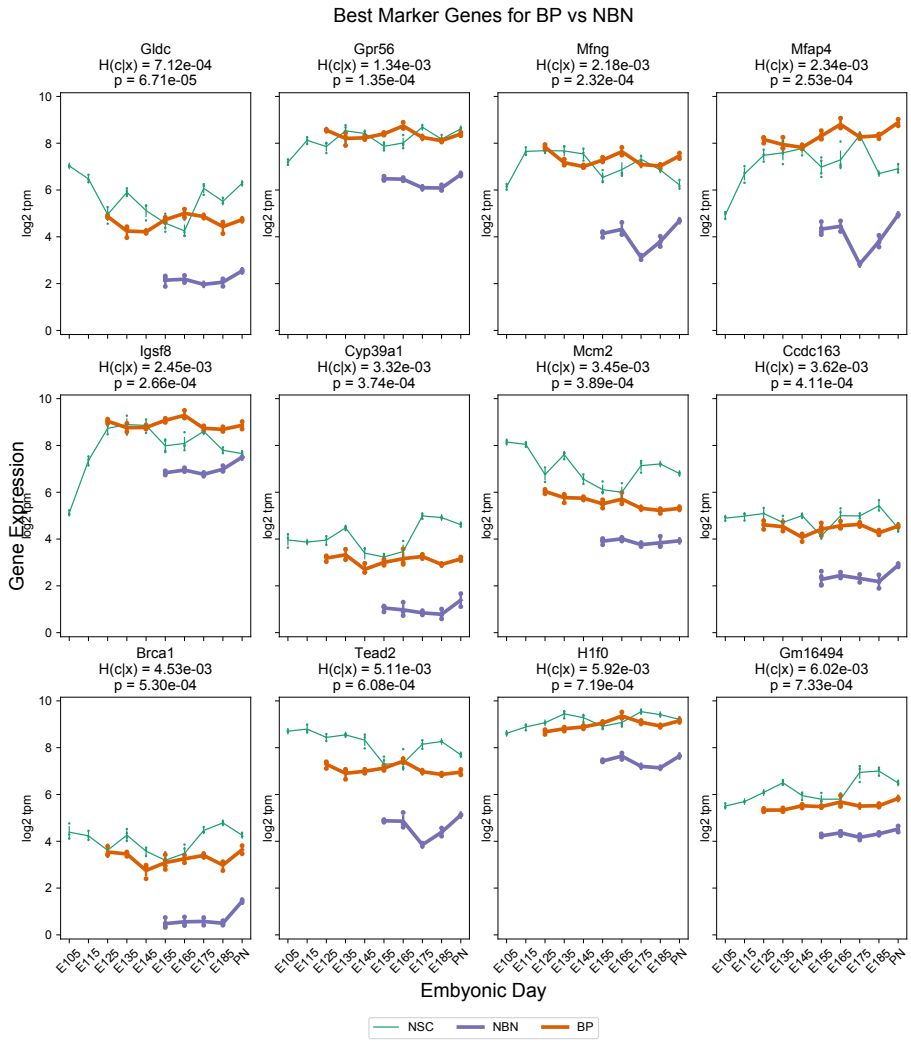

Figure 2.1: Top 12 marker genes for NSC, comparing gene expression measurements from NSC against BP and NBN from all time points. Measurements for replicate samples from different time points and cell types are shown as dots. The lines show the time evolution of the estimated measurement mean, with error bars indicating the estimated true standard deviations $\sigma_{\text{eff}}$, across replicates. For each marker gene, the conditional entropy $H(C|X)$ and the error of misclassification $p$ are noted.

where the subscripts refer to the two classes. Figure 2.3(b) shows the expression patterns for *Hdhd2* and *Vcam1*, which both have high and very similar t-scores of 264 and 259, respectively. For *Hdhd2* there is only a 1.6-fold change in expression between classes. As the variation in replicate gene expression measurements is smaller than expected, the estimated standard deviation $\sigma_{\text{eff}}$ indicated by the error bars is often larger than the spread of the data. In contrast, the expression of *Vcam1* changes 6.5-fold and the spread of replicate measurements tends to be larger than $\sigma_{\text{eff}}$. Therefore, it makes sense that *Vcam1* would be a much better marker gene for gliogenic NSC than *Hdhd2*.

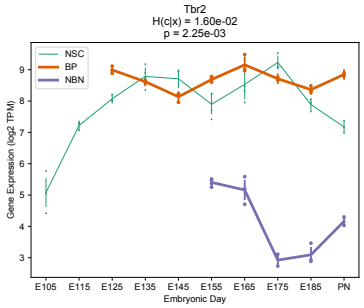
## 2.4 Discussion

In the most literal sense, marker genes are those that are used in an experiment to mark cells according to the expression of those genes. This is what we are trying to achieve here: we find the genes whose expression is most predictive for a condition of interest. But, we could also have defined marker genes as being expressed in our condition of interest but not a reference condition. Or if we have time course data like in the example above, we could look for genes that are predictive at any given time point rather than averaging over time. How one defines marker genes should determine the method used to find them. A related but different problem is that of finding differentially expressed genes. These are genes for which we have the most evidence that their *average* expression is different between conditions. Many different algorithms already exist for this task [17, 57], such as DESeq [2], edgeR [51] or limma [50].

The method presented here has several advantages: Firstly, it is robust even when the number of samples per class is small, as it uses a Bayesian method to account for the likely underlying variation in the data rather than the observed one. Secondly, it could be used to find genes that can differentiate between more than two conditions. And thirdly, the method is not specific to RNA-seq measurements, but can be applied to any kind of data where the number of measurements per condition is small, the number of features is high and the measurements are approximately Gaussian distributed.

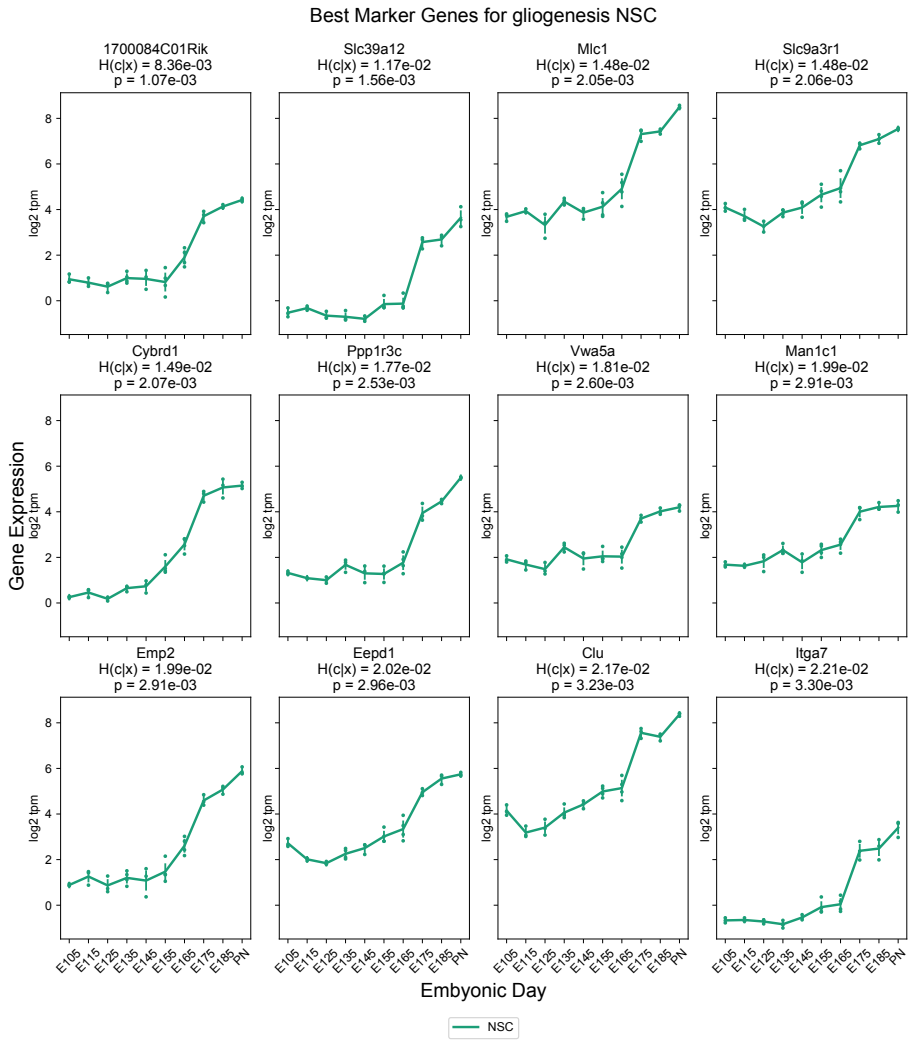



(a) Top 12 marker genes

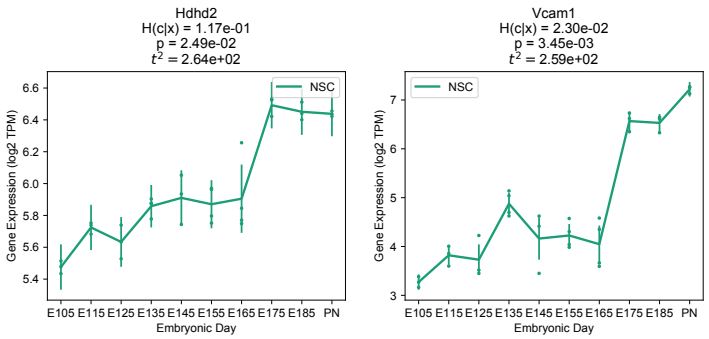



(b) Tbr2 expression

Figure 2.2: (a) Top 12 marker genes comparing gene expression measurements from NBN against BP. Measurements for replicate samples from different time points and cell types are shown as dots. The lines show the time evolution of the estimated measurement mean, with error bars indicating the estimated true standard deviations $\sigma_{\text{eff}}$, across replicates. The bold lines illustrate which samples are being compared against each other. For each marker gene, the conditional entropy $H(C|X)$ and the error of misclassification $p$ are noted. (b) The same expression plot for *Tbr2*, the gene used as a marker in the experiment.




(a) Top 12 marker genes




(b) Example of two genes with similar t-statistic

Figure 2.3: (a) Top 12 marker genes comparing gliogenesis NSC (E17.5-PN) against NSC from other embryonic days (E10.5-E16.5). Measurements for replicate samples from different time points and cell types are shown as dots. The lines show the time evolution of the estimated measurement mean, with error bars indicating the estimated true standard deviations $\sigma_{\text{eff}}$, across replicates. For each marker gene, the conditional entropy $H(C|X)$ and the error of misclassification $p$ are noted. (b) Expression plots for two genes that have very similar t-statistics, but different conditional entropy. Note the difference in the y-axis scale.

# 3

# Identifying cell states in single-cell RNA-seq data at statistically maximal resolution


## Abstract

Single-cell RNA sequencing (scRNA-seq) has become a popular experimental method to study variation of gene expression within a population of cells. However, obtaining an accurate picture of the diversity of distinct gene expression states that are present in a given dataset is highly challenging because of the sparsity of the scRNA-seq data and its inhomogeneous measurement noise properties. Although a vast number of different methods are applied in the literature for clustering cells into subsets with 'similar' expression profiles, these methods generally lack rigorously specified objectives, involve multiple complex layers of normalisation, filtering, feature selection, dimensionality-reduction, employ *ad hoc* measures of distance or similarity between cells, often ignore the known measurement noise properties of scRNA-seq measurements, and include a large number of tunable parameters. Consequently, it is virtually impossible to assign concrete biophysical meaning to the clusterings that result from these methods.

Here we address the following problem: Given raw unique molecule identifier (UMI) counts of an scRNA-seq dataset, partition the cells into subsets such that the gene expression states of the cells in each subset are statistically indistinguishable, and each subset corresponds to a distinct gene expression state. That is, we aim to partition cells so as to maximally reduce the complexity of the dataset without removing any of its meaningful structure. We show that, given the known measurement noise structure of scRNA-seq data, this problem is mathematically well-defined and derive its unique solution from first principles. We have implemented this solution in a tool called Cellstates which operates directly on the raw data and automatically determines the optimal partition and cluster number, with zero tunable parameters.

We show that, on synthetic datasets, Cellstates almost perfectly recovers optimal partitions. On real data, Cellstates robustly identifies subtle substructure within groups of cells that are traditionally annotated as a common cell type. Moreover, we show that the diversity of gene expression states that Cellstates identifies systematically depends on the tissue of origin and not on technical features of the experiments, such as the total

number of cells and total UMI count per cell. In addition to the CELLSTATES tool, we also provide a small toolbox of software to place the identified cellstates into a hierarchical tree of higher-order clusters, to identify the most important marker genes at each branch of this hierarchy, and to visualize these results.

## 3.1 Introduction

All cells in multicellular organisms contain the same genome with typically around 20,000 genes, but are able to take on a wide variety of phenotypes and perform specialized functions by selective expression of these genes. Therefore, it is one of the fundamental problems of cell biology to characterize the gene expression states cells take on in a multicellular organism. Addressing this requires investigating gene expression states in single cells, which has become possible through progress in the development of single-cell technologies, and single-cell RNA sequencing (scRNA-seq) in particular, over the last years. Numerous cell atlas projects [20, 22, 48, 49, 55] using this approach are already published or in progress. It is often assumed that cells can be divided into discrete cell types which have characteristic molecular profiles and perform specific functions, but despite all the available experimental data, it is still debated how such discrete types should be defined [16, 40, 41, 72]. Indeed, it is also often proposed that gene expression states are not discrete but rather occupy a continuous subspace of gene expression space, which is typically assumed to be of much lower dimensionality than the full gene expression space [35, 67].

It is thus currently not clear to what extent the assumption that cells can be grouped into discrete states is appropriate. Arguably, during cellular differentiation cells must be traversing an approximately continuous space of gene expression states, but for fully differentiated tissues it may not be unreasonable to approximate cells as deriving from a set of discrete states. However, even if we take for granted the assumption that cells take on discrete states or 'types', there is currently also no agreement regarding how such types should be defined or identified. That is, although intuitively cells of the same type should have 'similar' expression profiles, there is currently no agreed upon metric of closeness of gene expression states and no agreement on how close cells need to be in order for them to be considered the same type. Furthermore, even if a distance metric is chosen, for example Euclidean distance in log mRNA fractions, the sparseness and inhomogeneous noise properties of scRNA-seq data make it very challenging to accurately estimate the true distances between cells [11].

In spite of these problems, the current practice in the field is to simply apply *ad hoc* clustering approaches to scRNA-seq data, typically inspired by unsupervised machine learning methods, with the aim of grouping cells of the same 'type', e.g. [3, 28, 46]. These clustering approaches generally include several complex layers of data pre-processing, such as normalisation and imputation, feature selection, and dimensionality reduction, before the clustering algorithm is applied. These pre-processing steps not only include many fairly arbitrary choices but, as we have recently shown [11], such pre-processing can also severely distort the data by erroneously filtering true biological variability and introducing artefactual correlations. Furthermore, for the clustering itself many different approaches are available, and these typically additionally have many tunable parameters whose values in practice seem to be mostly set by trial-and-error. Given the many layers of *ad hoc* choices involved in these approaches, the resulting clusters lack any biophysical or even methodological interpretation. Instead, the approach taken to confirm the 'biological validity' of the clusters, is to show that the cluster exhibit some features that match known biological information, e.g. that certain 'marker' genes of a particular cell type are on average higher expressed in a given cluster. However, given that there are combinatorially many different

clusterings that exhibit such partial matches with prior biological knowledge, it seems problematic to us to take such partial matches to prior biological knowledge as a validation of the clusters that happened to result from the complex layers of analysis that were applied to the data.

We strongly feel that, instead of applying *ad hoc* clustering methods and attempt to validate these retrospectively by comparison with prior biological knowledge, it is more constructive to first rigorously specify the aims of the analysis, and then *derive* the appropriate algorithm that accomplishes these aims from first principles. This is the approach we take here. We are not going to attempt to solve the general problem of how to define cell types and how to identify them, for reasons laid out above. Instead, our aim is to use clustering so as to maximally reduce the complexity of the dataset without losing *any* of the structure in the data. In particular, we aim to partition the cells of an scRNA-seq dataset into subsets such that the gene expression states of all cells within each subset are statistically indistinguishable. We thus aim to cluster cells at the highest possible level of resolution that is statistically meaningful, i.e. within each cluster all cells are within measurement noise in expression state, and between clusters the expression states are all distinct.

Because the nature of biological and measurement noise in scRNA-seq experiments is known, as characterized in previous studies [61], this task has a uniquely defined solution determined by first principles, as we show below. The resulting method, CELLSTATES, directly clusters the unnormalised data so that any pre-processing steps are avoided, measurement noise is properly taken into account, and there are no free parameters to tune. For example, the number of clusters is determined by the data, in contrast to most approaches in which the number of clusters is tuned by the user. Moreover, the resulting clusters have a clear and simple interpretation.

Because CELLSTATES only groups cells whose expression states are statistically indistinguishable, it typically divides the data into many more subsets than other clustering algorithms. To allow comparison with the more coarse clusterings provided by other methods, we additionally provide methods for hierarchically merging CELLSTATES's clusters into coarser clusters and to identify marker genes associated with each branching in this hierarchy. As we show below, marker genes of conventionally annotated biological cell types typically correspond to coarser clusters in this hierarchy, allowing us to interpret CELLSTATES's clusters as subtypes of conventionally annotated cell types.

## 3.2 Review of current clustering tools

To illustrate how currently available clustering algorithms work, we will summarise the typical steps involved. Then, we will give a detailed description for some selected tools that are being used later in Section 3.4.3 for benchmarking against our model. A summary of the tools is also given in Table A.3. The aim of these section is to illustrate how complex and intransparent current methods are. They are rarely motivated by specific theoretical considerations about UMI count data, but rather follow a number of ad-hoc steps used in machine learning or data science for clustering various kinds of data. For all of these steps, methods and parameter values have to be chosen, that will influence the final clusters. As

a result, these clusters lack a clear interpretation. Researchers could be tempted to deal with this problem by tweaking the parameters until they get results that *they* think look plausible and agree with biological knowledge. But that also means, that the clusters are not objective statements about the data, but can be largely biased by preconceived ideas about what the data should show.

### 3.2.1 Common steps of a clustering analysis

**Normalisation** The total number of reads in a cell can vary due to random fluctuations in the capture efficiency of the scRNA-seq protocol between cells. Therefore, the number of UMI counts for a gene will depend on the specific capture efficiency in that cell. Thus, one needs to normalise the counts by a *size factor* for each cell. Typically, this size factor is simply the total number of UMIs in that cell multiplied by a constant, so that gene expression levels are measured in *counts per million* (CPM)[3]. Typically, this normalisation for size factors is also combined with a non-linear feature transformation. Without such a transformation, differences in gene expression are proportional to their absolute levels. For example, a change from 100 to 120 CPM would be as large as a change from 10 to 30 CPM. Relative expression levels of genes cover several orders of magnitude, so only changes in highly expressed genes would be relevant. As regulatory interactions often have multiplicative effects [6], it is standard to measure expression levels as log-transformed CPM. As zero counts are frequent, but incompatible with the logarithm, the data are normalised as $\log(\text{CPM} + 1)$. Many other normalisation techniques exist [65], and they typically try to address three points: differences in mRNA capture efficiencies between cells, rescaling of expression levels, and removal of zero measurements. Many different kinds of normalisation algorithms exist [65], but it has recently been shown that most normalisation algorithms introduce artificial biases into the data [11]. These biases would therefore also affect clustering results. Furthermore, zero measurements are an expected outcome of count data and should not need to be treated as a special case [12, 54, 59, 70]. While some normalisation techniques exist that take into account the count statistics [11, 34, 61], we will show that there is no need for this step in the first place and that clustering can be done directly on the raw UMI counts.

**Feature Selection and Dimensionality Reduction** As clustering algorithms generally try to find groups of "similar" cells, they need to define a measure of this similarity. This is usually a distance in gene expression space. However, as this space has ca. 20'000 dimensions, it suffers from the so-called *curse of dimensionality*, the phenomenon that distances between points in space become evenly distributed as the number of dimensions goes up. To remove this effect, *feature selection* and *dimensionality reduction* are both used to lower the number of dimensions in which distances are calculated and also to speed up computations. Many genes do not carry any relevant signal, for example if they are low expressed and therefore noisy or if they are housekeeping genes that have similar expression in all cells. Feature selection is the process of removing such uninformative genes and only retaining

[3] For scRNA-seq other constant scaling factors than 1 million are often used, such as the median total UMI count per cell or 10'000.

those that are most relevant to the underlying structure of the data [30]. As many pairs of genes have correlated expressions, the effective dimensionality of the biological information is much lower than that of the full gene expression space. Thus, dimensionality reduction techniques, such as principal component analysis, are applied that try to preserve the relevant information while removing any noise in the data.

**Similarity Metrics and Clustering** After these transformations of the raw UMI count data, a vast corpus of clustering algorithms from the machine learning literature can be applied [1]. The general aim of such an algorithm is to group cells into clusters such that cell within a cluster share similar gene expression characteristics, while cells in different clusters have clear differences. Thus, most methods will require the definition of a metric to quantify similarity between cells, e.g. Euclidean distance in gene expression space or the correlation of gene expression vectors. Some clustering algorithms work on graphs, and a popular choice to construct these from the data is to make a k-nearest-neighbours graph.

### 3.2.2 Description of published clustering tools used in this thesis

**BackSPIN** BackSPIN [75] is a biclustering method, that clusters both cells and genes simultaneously. Raw counts are pre-processed through feature selection and log(CPM) normalisation, cell-to-cell similarities are measured by correlation. The idea is to reorder the rows and columns of the expression matrix to get a block-diagonal forms, where each block corresponds to a cluster of cells and its associated genes. The basis for BackSPIN is SPIN [62], which is a method for optimally sorting the correlation matrix. After sorting, the optimal split in the correlation matrix is found such that cells on the same side of the split have a higher correlation than cells on different sides. Genes are associated with each of the two parts to generate two sub-matrices. Such splits can iteratively be performed on the sub-matrices until some conversion criterium. There are a number of parameters that can be set (most will have a default value): the maximal number of splits $d$ (the maximum number of clusters will be $2^d$); the number of genes retained in feature selection; several parameters related to how large sub-matrices can be after each iteration, and when a split is deemed meaningful; computational parameters for the optimization performed in SPIN.

**RaceID3** RaceID3 [21, 24] is a clustering algorithm that specialises in detecting rare cell types that do not fit into the main clusters it infers. The processing of the data uses feature selection, normalisation, imputation of zero counts and removal of unwanted variation (such as cell cycle or batch effects). The first clustering step is done by k-medoids clustering on a correlation-based distance matrix, where the optimal k can be inferred directly from the data. After initial clustering, outlier cells are defined as cells which have at least 2 (or as chosen by a user) genes which are significantly differentially expressed with respect to the other cells in their cluster. Outlier cells are merged into clusters or rare cell types in the final step. The tool gives a large choice in methods and parameters values for each of the preprocessing steps, but also the clustering algorithm and the number of clusters. There are also several parameters and thresholds related to the outlier finding steps that can be set.

**SC3** SC3 [29] is a clustering tool that combines predictions from several parallel clustering runs with different parameter sets into a consensus solution. The algorithm starts with normalised data as an input, but does include feature selection. Using PCA, the dimensionality is reduced to $d$, Euclidean distances are calculated, and $k$-means clustering is performed on the cells. This is done for a range of values of $d$, and a similarity matrix between cells is created based on how often cells appear in the same cluster. Finally, this similarity matrix is used for hierarchical clustering with complete agglomeration to $k$ clusters. If the number of cells is very large, there is also the option to do this clustering on a subset of cells and train a support vector machine (SVM)[7] that assigns each cell to a cluster. To summarise, there are parameters that need to be set are related to feature selection, the number of clusters $k$, the range of values $d$, and potentially parameters related to the SVM model.

**SNN-Cliq** SNN-Cliq [74] is a clustering algorithm based on shared nearest neighbour (SNN) graphs. Preprocessing is not done by the tool, but suggested normalisation is with $\log(\text{RPKM}+1)$[4] and suggested feature selection is to take genes with RPKM $> 20$. Based on this normalised data and a distance metric, a list of $k$ nearest neighbours is generated for each cell. The SNN graph connects pairs of cells that have at least one shared nearest neighbour. Finally, a novel graph clustering algorithm is applied to this SNN graph. The parameters to be chosen after normalisation and feature selection are: the number of nearest neighbours $k$, the distance metric used, and two parameters that define the granularity of the graph-based clustering algorithm.

**DIMM-SC** DIMM-SC [58] is a clustering algorithm based on a Dirichlet mixture model. Unlike the other methods presented here, DIMM-SC directly models UMI counts using a Bayesian model. The data is assumed to fall into one of $K$ clusters. UMI counts $\vec{n}$ of a cell are drawn from a multinomial distribution with parameter vector $\vec{\alpha}$, which obeys $\alpha_g \geq 0$ and $\sum_g \alpha_g = 1$:

$$P(\vec{n}|\vec{\alpha}) \propto \prod_g (\alpha_g)^{n_g}, \tag{3.1}$$

where the product is over the genes $g$. The parameter vector $\vec{\alpha}$ has a prior given by the Dirichlet distribution

$$P(\vec{\alpha}|\vec{\theta_j}) \propto \prod_g (\alpha_g)^{\theta_{gj}}, \tag{3.2}$$

where the concentration parameters $\vec{\theta_j}$ are different for each cluster $j$. Then, we find the marginal likelihood of $\vec{n}$:

$$P(\vec{n}|\{\vec{\theta_j}, \pi_j\}) = \sum_{j=1}^{K} \pi_j \int P(\vec{n}|\vec{\alpha}) P(\vec{\alpha}|\vec{\theta_j}) d\vec{\alpha}, \tag{3.3}$$

[4] Reads Per Kilobase of transcript, per Million mapped reads – a unit that is not applicable to UMI based data.

where $\pi_j$ is the prior probability to be in cluster $j$. Hence, the model parameters $\{\vec{\theta_j}, \pi_j\}$ are optimised to maximise the total likelihood for all cells under this model, using am expectation maximisation (E-M) algorithm. To conclude, this here is only one parameter that needs to be set by the user, the number of clusters $K$. Other parameters are related to E-M algorithm and do not change the optimal solution, only how close the program gets to finding it. Our model described below is actually very similar to DIMM-SC, but has a few differences. Firstly, we do not have a fixed number of clusters, but infer it from the data. And secondly, DIMM-SC allows for potentially large variance in gene expression within a cluster by adjusting the relevant values of $\theta_{gj}$. In contrast, in our model the allowed variance within a cluster, set by the equivalent variable $\theta_g$, is fitted globally for all clusters and is proportional to the average expression of that gene across all cells.

## 3.3 Methods

### 3.3.1 Multinomial noise in scRNA-seq data implies a parameter-free solution for probabilities of partitions of cells into states

The internal gene expression state (GES) of a cell $c$, which we will also refer to as a cellstate, is determined by a multitude of biological processes that influence the transcription rates $\lambda_{gc}(t)$ and degradation rates $\mu_{gc}(t)$ of mRNAs across genes $g$ and time $t$ in the history of the cell.

These rates determine the probabilities for the mRNA counts in the cell, which in turn ultimately determine the probabilities of the number of reads captured in a UMI-based scRNA-seq measurement. The probability distribution for the number of mRNAs in a cell $m_{gc}$ follows a Poisson distribution with mean $a_{gc}$ given by

$$a_{gc} \equiv \langle m_{gc} \rangle = \int_0^\infty dt \lambda_{gc}(t) \exp\left[ - \int_0^t \mu_{gc}(s) ds \right], \tag{3.4}$$

where the time is measured backwards from the present ($t = 0$) to the distant past ($t = \infty$) in the history of the cell [11]. Thus, notably, for each gene $g$ in each cell $c$, the entire complex history of transcription rate and mRNA decay rate can be summarised into a single parameter $a_{gc}$ that fully determines the probability distribution of its current mRNA count for gene $g$. The scRNA-seq measurement process is noisy, and typically only a small fraction ($\sim 20\%$ or less) of cellular mRNAs are captured. As this capture rate can vary substantially between cells, information about absolute gene expression levels is lost, at least to some extent. Therefore, more accurate inferences can be made regarding the expected *fractions* of total cellular mRNA that mRNAs of each gene $g$ represent. Following [11], we denote these fractions by transcription quotients $\alpha_{gc}$, which we define by

$$\alpha_{gc} = \frac{a_{gc}}{\sum_g a_{gc}}. \tag{3.5}$$

We now define the GES of a cell as the vector $\vec{\alpha}_c$ of transcription quotients across all $G$ genes. Thus, a GES is a point in the $G$-dimensional simplex $\alpha_{gc} \geq 0 \; \forall g$ with $\sum_g \alpha_{gc} = 1$.
Given an scRNA-seq dataset with $N$ cells, we will assume that the GESs of the cells derive from an unknown set $S$ of GESs, where each GES $s \in S$ is characterized by a distinct vector

of transcription quotients $\vec{\alpha}_s$. That is, we assume that there are somewhere between 1 (all cells having the same GES) and $N$ (all cells having a distinct GES) cellstates represented in the dataset. Our goal is to derive which cells are in the same state and thus separate differences in UMI counts due to biological and measurement noise from differences in the underlying biological state. Thus, the space of hypotheses for this problem is the space of possible partitions of the $N$ cells into non-empty non-overlapping subsets. In particular, we aim to calculate a likelihood for each possible partition that quantifies how probable the data is under the assumption that all cells in each subset of the partition are in the same GES.

The first step is to derive the relationship between the GES of a cell $c$ characterized by $\vec{\alpha_c}$ and the vector of its measured UMI counts $\vec{n_c}$, which is summarised in Fig. 3.1A. Given the transcription activities $a_{gc}$ of a cell $c$, the mRNA counts are not uniquely determined, but due to inherent biochemical noise in the gene expression process, the mRNA counts $m_{gc}$ are given by Poisson samples with means $a_{gc}$. Defining the total transcription activity $A_c = \sum_g a_{gc}$, the expected mRNA count for gene $g$ can be expressed as the product of $A_c$ and the transcription quotient $\alpha_{gc}$:

$$m_{gc}|\alpha_{gS}, A_c \sim \text{Poisson}(\alpha_{gc}A_c) \tag{3.6}$$

Assuming that, for cell $c$, each transcript was captured and sequenced with a probability $p_c$, the distribution of UMI counts $n_{gc}$ will also be Poisson distributed with mean $\alpha_{gc}A_cp_c$ for each gene $g$. If we marginalize over the unknown capture probability $p_c$ and condition on the total number of mRNAs $N_c$ that were captured for cell $c$, the counts $n_{gc}$ are simply distributed as a multinomial sample of the transcription quotients $\vec{\alpha}_{gc}$ (see the supplementary methods of [11] for a more extensive derivation):

$$\vec{n_c}|\vec{\alpha_c}, N_c \sim \text{Multinomial}(\vec{\alpha_c}, N_c) \propto \prod_g (\alpha_{gc})^{n_{gc}}. \tag{3.7}$$

Thus, the probability of the observed mRNA counts $\vec{n}_c$ of a cell $c$ conditioned on its GES $\vec{\alpha}_c$ is simply a multinomial sample of size $N_c$ of the expression state $\vec{\alpha}_c$. As an aside, we note that the vector of observed UMI counts $\vec{n}_c$ is the unique sufficient statistic for the GES $\vec{\alpha}_c$ of cell $c$.

Given a partition $\rho$ of the cells into non-overlapping subsets, we now use the above results to calculate a probability $P(n|\rho)$ of the observed UMI counts $n$ across all genes and cells, given the assumed partition $\rho$. The derivation of our model follows [68], is explained in detail in the Supplementary Information section 3.8.1.1, and the general approach is illustrated in Fig. 3.1. Briefly, a partition $\rho$ contains subsets of cells $s$, with one GES $\vec{\alpha}_s$ for each subset $s \in \rho$ (i.e. each subset $s$ corresponds to a cluster of cells), and all cells $c \in s$ are assumed to have the same GES $\vec{\alpha}_s$ for each subset $s \in \rho$. The probability for the counts $n_{gc}$ of all cells in the subset $s$ is simply the product over multinomial distributions for each of the cells. Thus, if we define the cluster UMI counts $n_{gs} = \sum_{c \in s} n_{gc}$ and $N_s = \sum_{c \in s} N_c$, then these cluster counts also simply derive from a multinomial distribution

$$\vec{n_s}|\vec{\alpha_s}, N_s \sim \text{Multinomial}(\vec{\alpha_s}, N_s) \propto \prod_g (\alpha_{gs})^{n_{gs}}. \tag{3.8}$$

Next, because we do not know the transcription quotients $\vec{\alpha}_s$ we marginalize over these parameters using a Dirichlet prior. The family of Dirichlet priors is the unique set of priors that is invariant under rescaling of the unknown transcription quotients $\alpha_{gc}$ and is parametrized by a vector of concentrations $\Theta$. This marginalization can be done analytically, leading to a ratio of products of Gamma functions of the counts $n_{gs}$ (see Supplementary Information section 3.8.1.1). In this way, a likelihood of the UMI counts $\vec{n}_s$ is obtained for each cluster $s$ in the partition $\rho$. By taking the product of these likelihoods over all subsets in $\rho$, we arrive at an expression for the likelihood $P(D|\rho, \Theta)$ of the entire dataset $D = \{\vec{n_c}\}$ as a function of the partition $\rho$ and the parameters of the Dirichlet prior $\Theta$. Taking a uniform prior over both partitions $\rho$ and the parameter $\Theta$, the posterior $P(\rho, \Theta|D)$ is simply proportional to the likelihood $P(D|\rho, \Theta)$, which we have obtained in analytical form. The aim of our algorithm is to now find the partition $\rho$ and prior parameters $\Theta$ that jointly maximize this likelihood.

Importantly, this approach uses only the assumptions that both the inherent biochemical noise in gene expression and the scRNA sequencing introduce Poisson sampling noise, and from first principles derives a parameter-free solution for the most likely partition $\rho^*$ of cells into cellstates that is entirely determined by the raw data $D$. Note that defining cellstates in this way also determines how many distinct cellstates there are, and how many cells there are in each state, directly from the data. The total likelihood of a partition $\rho$ simply quantifies how consistent the cells' measured UMI counts are with the assumption that all cells in each cluster share a common (but unknown) GES $\vec{\alpha}$. Importantly, over-clustering of the cells into too many cellstates is avoided through the Bayesian framework, where increasing the number of GES $\vec{\alpha}$ that are marginalized over will lower the likelihood if not supported by the data. To summarise, we partition all cells into subsets such that it is most likely that within each subset the remaining variation between the captured UMI counts is due to random fluctuations.

### 3.3.2 The likelihood function is optimized using a Markov-Chain Monte-Carlo algorithm

The number of possible partitions of $N$ cells grows faster than $e^N$, and we have confirmed that simple greedy searches, such as iteratively fusing clusters of cells to maximally increase the likelihood of the partition, tend to get stuck in local optima of the likelihood function. This makes maximization of the likelihood function challenging. To search for the optimal partition, we start from the partition in which each cell forms a cluster by itself and use a stochastic Markov-Chain Monte-Carlo (MCMC) scheme as previously developed in [68]. In each step, a randomly selected cell is proposed to move into a randomly selected different cluster – and accepted if this move increases the likelihood of the partition. If the move decreases the likelihood by a factor $p < 1$, the move is accepted with a probability $\tilde{p}$ that is adjusted to ensure uniform sampling of partitions (see Supplementary Information section 3.8.2.1). Although theoretically, this MCMC scheme samples partitions in proportion to their likelihood in the long run, we have observed that, in practice, either $p \gg 1$ or $p \ll 1$ for most of the proposed moves, most likely due to the fact that the total number of UMI per cell is generally large. Therefore, in practice, the optimization essentially performs a

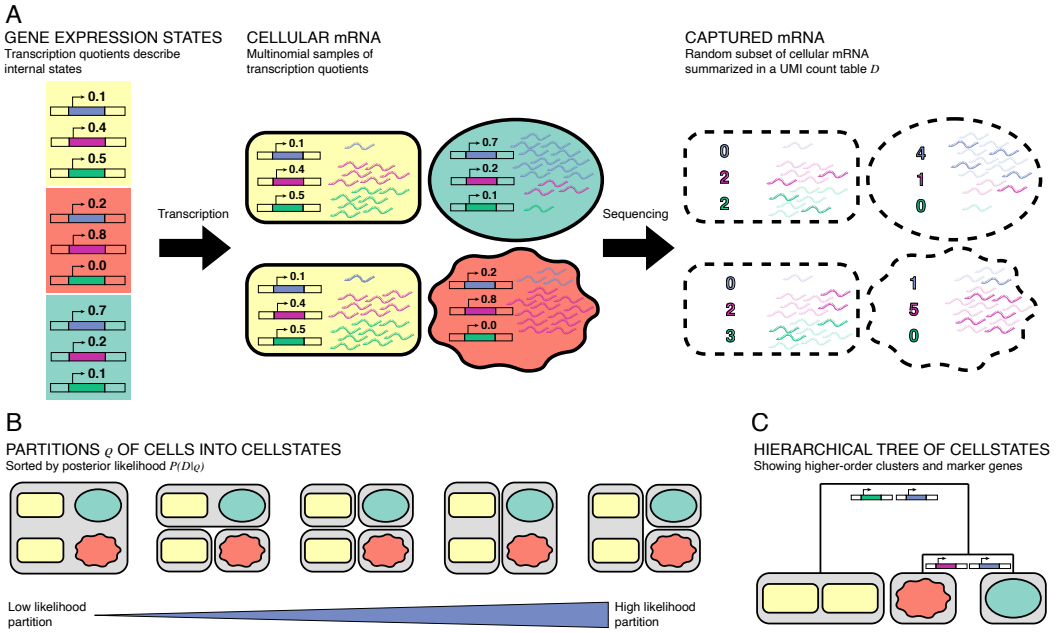



Figure 3.1: (A) Summary of model assumptions. Each cell is in a gene expression state (indicated by shape and colour) characterized by the transcription quotients across genes. The relative numbers of mRNAs in the cell follow a multinomial distribution of these rates. The counts obtained from sequencing reflect a random subset of captured cellular mRNAs and follow the same multinomial. (B) Summary of clustering algorithm. Each partition $\rho$ of cells into clusters gives a likelihood of the data under the noise model. By optimizing the partition, we find groups of cells with shared gene expression states. (C) Cellstates can be hierarchically merged into higher-order cell types. For each merging step, we indicate which genes most contribute to distinguishing the transcription quotients to the left and right below the merger.

random uphill walk to a local optimum rather than sampling the full probability distribution of partitions. We have experimented with a number of different search schemes, including simulated annealing and Gibbs' sampling schemes, but found that these random uphill walks provide the best balance between total run time and optimality of the final partition. After the MCMC converges, the optimization is followed by some deterministic steps, as described in detail in Supplementary Information section 3.8.2.1. Multiple runs of CELLSTATES on the same data can yield slightly different partitions, and we simply select the best-scoring partition from the partitions obtained in different runs.

### 3.3.3 Merging cellstates hierarchically into higher-order clusters

As the optimal cellstate partition gives a very fine-grained view of the data, it makes sense to relate the obtained cellstates to each other in a structured manner. To examine the higher-order structure between the cellstates of the optimal partition $\rho^*$, we devised a scheme to hierarchically merge them into higher-order clusters. We define a pairwise cluster similarity as the ratio of the likelihoods of the partition in which the two clusters are merged and the partition in which the two are separated. By construction of the optimal partition $\rho^*$, the similarity will be $< 1$ for any pair of clusters of this partition, and we define a 'distance'

between two clusters as minus the logarithm of this similarity. The most similar clusters are iteratively merged, resulting in a hierarchical tree of higher-order clusters, see Supplementary Information section 3.8.1.3 and Fig. 3.1C. As discussed below, we find that these higher-order clusters are often similar to the cell type annotations given in the publications of the datasets on which we ran our algorithm. Additionally, by approximating the multinomial as the product of independent binomials for each gene, we can calculate the contribution of each gene to the cluster similarity score, thus quantifying which genes drive differences in GES between cellstates or higher-order clusters (see Supplementary Information section 3.8.1.4). This allows users of our method to explore the types of cellstates present in the hierarchical tree more easily, i.e. by identifying which genes are associated with particular branchings.

## 3.4 Results

### 3.4.1 CELLSTATES accurately finds optimal partitions in simulated data

As discussed below, we tested CELLSTATES by running it on a number of published experimental scRNA-seq datasets. However, since there is no ground-truth information available for the GESs of cells in real scRNA-seq datasets, we decided to validate our likelihood maximization algorithm on synthetic datasets that were generated so as to be in agreement with the noise model described above. To get realistic simulated data, we modelled the simulations after results obtained by CELLSTATES from 18 of the analysed real experimental datasets as follows. For each of the 18 real datasets we took the optimal partition inferred by CELLSTATES and then, for each of the clusters $s$ in this partition, sampled the UMI counts of its cells from a multinomial distribution with mean equal to the inferred GES $\vec{\alpha}_s$ of the cluster. We generated three independent simulated datasets for each of the 18 scRNA-seq datasets, for a total of 54 simulated datasets (See Supplementary Information section 3.8.2.2 for details). We then ran the CELLSTATES algorithm three times on each simulated dataset. For the large majority of runs, CELLSTATES found the exact same partition as the one that generated the data (Figure 3.5), which is remarkable since only a tiny fraction of the total space of partitions is sampled during the MCMC likelihood optimization. Moreover, when the partition that CELLSTATES found differed from the partition that generated the data, this was because CELLSTATES found a partition with even higher likelihood than the one that generated the data. In fact, for each of the 54 datasets our MCMC likelihood maximization procedure found a partition with likelihood at least as large as the partition that generated the data, and in $\approx 91\%$ of all runs overall (Figure 3.5).

To compare the similarity of partitions more quantitatively, we will use the two complementary measures of homogeneity and completeness [53] throughout this paper. These measures quantify how much information (as quantified by the Gibbs/Shannon entropy function) a given partition $\rho$ contains about a reference partition $\rho_f$, and are both normalised to lie between 0 and 1. If we imagine that we colour cells by their cluster in the reference partition $\rho_f$, i.e. so that all cells within each cluster of $\rho_f$ are given the same colour, then homogeneity measures how much information the cluster membership in $\rho$ provides about the colour of the cells (i.e. cluster membership in $\rho_f$). Homogeneity is 1 when, for each cluster of $\rho$, all cells have the same colour. Completeness, vice versa, measures how much information the

colour of a cell provides about its cluster in $\rho$. Completeness is 1 when, for each colour from $\rho_f$, all cells of a common colour occur in only one cluster of $\rho$. Note that two measures are necessary because if $\rho$ is the partition in which each cell is its own cluster, homogeneity is 1 by definition (but completeness is 0). Vice versa, if $\rho$ is the partition in which all cells are in one cluster, then completeness is 1 (but homogeneity is 0).
Comparing the partitions inferred by CELLSTATES on the simulated data to the corresponding reference partitions used to generate the data, we find that they overlap very well, with completeness and homogeneity larger than 0.95 for all runs, and larger than 0.9975 for 118/162 (73%) of the runs, as shown in Figure 3.2A. As discussed in Supplementary Information section 3.8.2.2, most 'errors' occur when the maximum likelihood partition found by CELLSTATES is higher than that used to generate the data. This happens in particular when the simulated datasets are too noisy to resolve all ground-truth states because the total UMI counts of cells in the ground-truth states are smaller than in the original data.
In summary, our tests with simulated datasets show that on datasets that mimic real data, CELLSTATES performs extremely well on recovering the ground truth used to generate the data, most often recovering the exact partition. And when there is a difference in the partition found, this is most often because CELLSTATES found an even better partition, which is always very close to the ground-truth partition, as measured by completeness and homogeneity.

### 3.4.2 CELLSTATES yields highly reproducible partitions on real datasets

We gathered a total of 29 published datasets from UMI-based scRNA-seq experiments, covering a large range of experimental protocols, tissues and two species (mouse and human), as summarised in Supplementary Table A.1. We ran CELLSTATES on five times on all datasets and compared the best-scoring partition from the five runs with the partitions from the other four runs. We find that the agreement between multiple runs of CELLSTATES is high, with 88% of the homogeneity and completeness scores larger than 0.9 (Figure 3.7). These results show that, even though different runs yield different partitions, they do not change substantially between runs.

### 3.4.3 CELLSTATES partitions agree better with published annotations than those of other clustering tools

We next compared CELLSTATES partitions on real datasets with those of a set of previously published methods (BACKSPIN [75], DIMM-SC [58], RaceID [21] and SC3 [29]). A short summary of these methods is provided in Table A.3. Assessing the relative performance of different clustering algorithms on real datasets is challenging because in general the ground truth is not known. Here we consider two tests. First, we selected 3 scRNA-seq datasets for which hand-curated annotations of cell types were provided in the publication [5, 14, 75] and compared the partitions obtained by each of the clustering methods with the published annotation. Although there is of course no guarantee that the published annotations are correct, it is reasonable to assume that a better match with these published annotations generally indicates better performance. Second, we also generated a set of five *in silico*

mixtures of pure cell populations with known identity, as has been done before for similar benchmarks [18, 61], using data from [77], and checked to what extent each algorithm can correctly recover these known mixtures.

We ran each of the methods on each of these 8 datasets, without tweaking any of the default parameters. Only when a method required the cluster number to be set, we set it to the correct number of annotated cell types. For CELLSTATES we obtained both the partition given by the method without specifying any parameters, and the partition obtained when hierarchically merging clusters until the number of clusters matches the number of annotated clusters. As for our tests with the simulated data above, we compared the clusters obtained by each of the methods to the annotated clusters through homogeneity and completeness (Figure 3.2B).

Notably, the partitions found by CELLSTATES always had the highest homogeneity. This supports that CELLSTATES indeed only clusters together cells that are in the same underlying gene regulatory state, and that this works automatically without the need to correctly set model parameters. However, CELLSTATES typically partitions the data in more clusters than the annotation, so that the completeness is generally significantly below one. When cellstates are hierarchically merged until the number of clusters matches the annotation (cellstates_hierarchical in Figure 3.2B), the completeness increases substantially, typically without lowering homogeneity by much.

On the three datasets with published annotations, CELLSTATES obtained partitions that match the published annotations well, especially in comparison to the partitions produced by the other methods, with only SC3 showing a similar performance on these annotated datasets. Importantly, for the *in silico* cell mixtures – which are the closest we have to a ground-truth annotation – CELLSTATES clearly outperforms the other methods, and often by a substantial margin. Overall, these results suggest that CELLSTATES can correctly predict higher-order cell types in scRNA-seq data and that it can do so more accurately than other tools. Moreover, this performance is obtained without any need to pre-process the data or any adjustable model parameters.

### 3.4.4 Cellstate diversity patterns depend on tissue of origin and not on technical features of the experiment

We next investigated the variation in the number and sizes of the clusters that CELLSTATES infers on different real datasets. Because measures such as the absolute number of cells per cluster obviously scale with the total number of cells sequenced, we decided to focus on the distribution of cellstate abundances $f_{\text{cellstate}}$, i.e. the fraction of cells associated with each cellstate. The distribution of $f_{\text{cellstate}}$ reflects the diversity of different GESs present in a given dataset. As an example, Figure 3.3A shows the distribution of $f_{\text{cellstate}}$ for the data from the mouse cortex and hippocampus of [75]. As illustrated by Fig. 3.3A, we find that $f_{\text{cellstate}}$ typically varies over multiple orders of magnitude, and that a substantial fraction of the clusters correspond to singlets, i.e. where GESs were only associated with a single cell. That is, the counts in these cells are statistically different from those of all other cells. To obtain a quantitative measure of diversity we looked at various statistics of the distribution

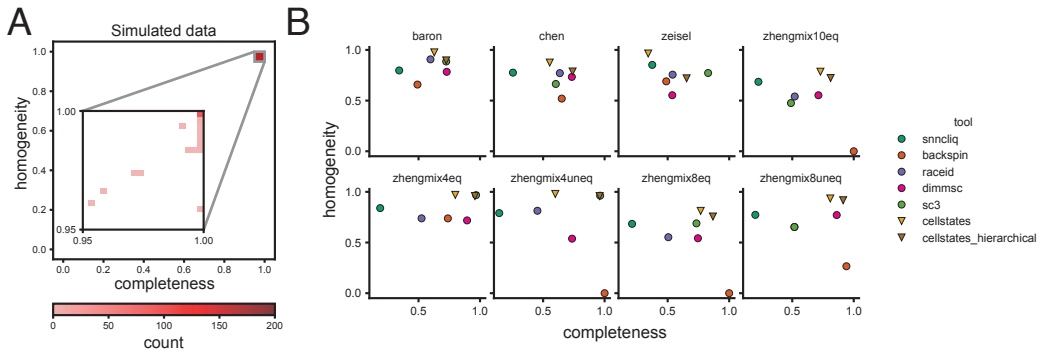



Figure 3.2: Benchmarking of CELLSTATES. (A) 2D-histogram of homogeneity and completeness of inferred cellstate memberships in 162 simulated datasets. The inset shows the distributions in the region $[0.95, 1], [0.95, 1]$ where all results fall. (B) Comparison of the performance of CELLSTATES with those of other clustering tools. For CELLSTATES, we show the results for the full partition into cellstate-groups ("cellstates") and merged to the same number of clusters as the annotation ("cellstates_hierarchical"). In each plot, we show homogeneity and completeness of the partitions obtained by the different methods using the published annotation as the reference partitions. Note that low homogeneity and completeness may indicate under-clustering and over-clustering, respectively.

of $f_{\text{cellstate}}$ including the fraction of cells that are singlets, the average cellstate abundance $\langle f_{\text{cellstate}} \rangle$, its median, and the entropy of the distribution of $f_{\text{cellstate}}$.

Although ideally these diversity measures would directly reflect the underlying biology of the tissue from which the data derives, we expected that these diversity measures might also strongly depend on technical features of the experiment such as the total number of cells and the typical total number of UMI per cell. For example, the higher the total UMI count per cell, the easier it becomes to distinguish subtly different cellstates, so that one would expect the cellstate diversity to increase with total UMI counts. Similarly, one would expect that the more cells are sequenced, singlet clusters should become less common. To investigate systematically to what extent the distributions of $f_{\text{cellstate}}$ reflect underlying biology versus technical features, we collected 29 scRNA-seq datasets from 9 different tissues, from different labs and using different sequencing technologies (see Table A.1), and investigated how the various diversity measures varied across tissues and with technical features such as cell number and total UMI counts.

Remarkably, we find that all these diversity statistics vary over several orders of magnitude across datasets. For example, the fraction of cells that are singlets varies over three orders of magnitude among the analysed datasets, from $5 \times 10^{-4}$ to nearly $7 \times 10^{-1}$ (see Figure 3.3B), and similarly for the other statistics (Fig. 3.3B and Fig. 3.8). Moreover, although there is a lot of variation across datasets, Fig. 3.3B and Fig. 3.8 also show that all the diversity measures systematically depend on the tissue of origin of the sample, despite vastly different experimental protocols used to obtain and sequence them. For example, for datasets stemming from biologically diverse cell populations such as brain or peripheral blood mononuclear cells (PBMC), CELLSTATES correctly and automatically infers few cells per GES, whereas vastly lower diversity is inferred for datasets stemming from Thymus. Moreover, and somewhat to our surprise, the diversity measures show almost no correlation with technical features such as total UMI counts and cell number (Figure 3.9). The only

clear correlation observed is a negative correlation between number of cells and median of the cellstate fractions. This correlation is explained by the fact that the median cellstate fraction often corresponds to singlets, i.e. $f_{\text{median}} = 1/n_{cells}$.
In summary, we find that the diversity of cellstates that is found in different datasets reflects the underlying biology of the system, and does not systematically depend on technical features of the experiment. Given this, the observation that in most of the datasets analysed a large fraction of the clusters are singlets, strongly suggests that the true biological diversity of cellstates is still severely under-sampled in these datasets. That is, many more cellstates exist than are captured at these sampling depths.

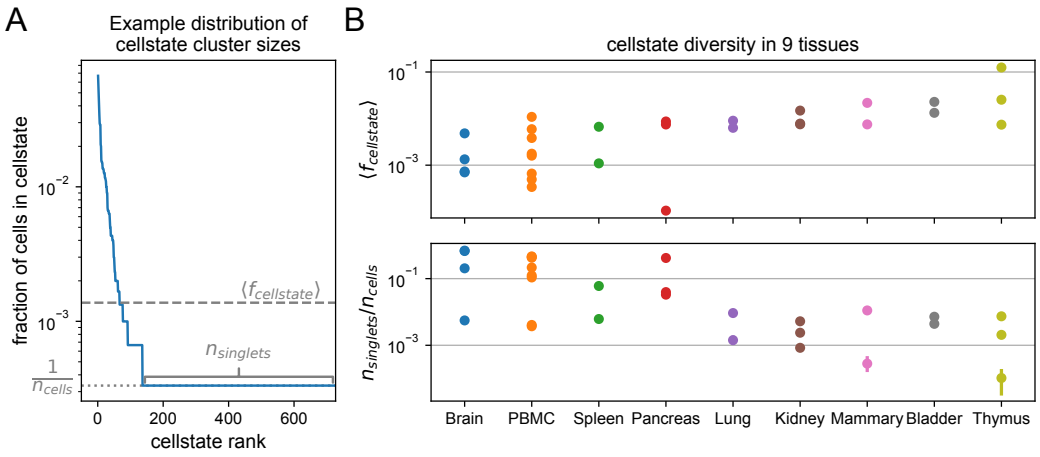



Figure 3.3: Cellstate diversity reflects the tissue of origin of the data. (A) Example rank-abundance curve for the fraction of cells associated with each cellstate in the dataset from [75]. Such curves describe the diversity of gene expression states in a dataset. The length of the horizontal tail gives the number of singlet cellstates $n_{\text{singlets}}$ with only one cell; the average cellstate abundance $\langle f_{\text{cellstates}} \rangle$ is also annotated. (B) For each of the CELLSTATES results for 29 different datasets from 9 tissues, the average abundance $\langle f_{\text{cellstates}} \rangle$ and the fraction of singlets $n_{\text{singlets}}/n_{\text{cells}}$ are plotted by the tissue they originate from. These diversity measures show a clear dependence on the tissue, despite the large variation in experimental set-ups used. Error bars show the standard deviations from 5 independent runs of CELLSTATES and are often so small that they are not visible.

### 3.4.5 CELLSTATES captures diversity of gene expression states in the mouse brain

Finally, to illustrate how the CELLSTATES data analysis pipeline can be used for in-depth analysis of a given dataset, we focused on the dataset of [75] consisting of 3005 cells from the somatosensory cortex and from the CA1 region of the mouse hippocampus. CELLSTATES infers a remarkable diversity in this tissue, with a total of 763 different GESs. Almost a quarter of the cells (727/3005) are in a unique singlet state, but there are also GES with up to 201 cells (7% of all cells), as can be seen in Fig. 3.3A.
Visualizing how this large number of cellstates relates to another is difficult because the GESs are objects in a very high-dimensional space. The approach that is currently by far most popular in the field is to use stochastic embedding methods that attempt to place cells that are close in gene expression space near each other in a 2-dimensional plane, in particular

UMAP [36] and t-SNE [66]. However, it is well appreciated that proper application of these tools is challenging [31], and that beyond approximate conservation of close-neighbourhood relationships, the large scale structure in these visualizations is virtually meaningless. In fact, we share the opinion of some in the field that the current usage of t-SNE and UMAP visualizations may be doing more harm than good [13]. Nonetheless, since such visualizations have become the *de facto* standard in the field, we decided to illustrate how cells with different cellstates are placed within a t-SNE visualization of the data (Fig. 3.4A, left panel).

This visualization confirms that cells that are predicted by Cellstates to have the same underlying GES (indicated by the marker colour, with singlets in gray) tend to be placed more closely in the t-SNE visualization. Cellstates infers that any variation between cells in the same GES is due to noise, which argues that we can collapse cells in each cellstate and replace them with a single disc at the average of their positions in the t-SNE visualization (right panel in Fig. 3.4A, with the area of the disc corresponding to the number of cells in the cellstate). This illustrates Cellstates' ability to reduce the complexity in the data, allowing for a tidier visualization which, for example, highlights that different common cellstates (large coloured discs) have different numbers of singlets (grey dots) in their neighbourhood.

Next, we hierarchically merged the cellstates to determine higher-order clusters in the dataset, and again visualized the results by colouring either the cells or cellstates in the t-SNE visualization (Fig. 3.4B). We see that when cellstates are merged into 8 higher-order clusters, these clusters largely match the structure of the t-SNE visualization.

However, we feel that a more useful visualization of the relations between the cellstates is obtained by displaying the hierarchical tree resulting from iteratively merging the statistically most similar clusters (Fig. 3.4C). The tree indicates which cellstates and higher-order clusters are most similar in expression, although it should be remembered that 'distance' between clusters is here measured in terms of how statistically significant the differences in the expression patterns are, as opposed to in terms of the magnitude of the changes in gene expression. Notably, at 8 higher-order clusters we find good correspondence with the cell type annotation given in the original publication (Fig. 3.10), and this is also confirmed by expression of marker genes for these annotated cell types (Fig. 3.12).

There are however two main differences. Firstly, at this level of resolution in our cluster hierarchy, the annotated clusters of endothelial-mural cells and microglia are merged and they separate only at 15 higher-order clusters (Fig. 3.11). Secondly, one cluster had a mixed annotation at the chosen resolution. As shown in Fig. 3.13, this cluster contains cells that express genes which are considered markers for multiple different cell types. This indicates that either the assumption that these genes are markers for specific cell types is incorrect or, alternatively, a technical artefact in the data, e.g. these 'cells' might correspond to multiple cells getting the same cell-barcode and having their counts combined.

Finally, we also provide software to extract genes that contribute most to separating the GESs on opposite sides of each branch. To illustrate the use of this tool we focus on the cluster of interneurons, which was particularly diverse with 98 out of 290 cells in singlet states. For each node in the subtree corresponding to the interneurons we identified the genes

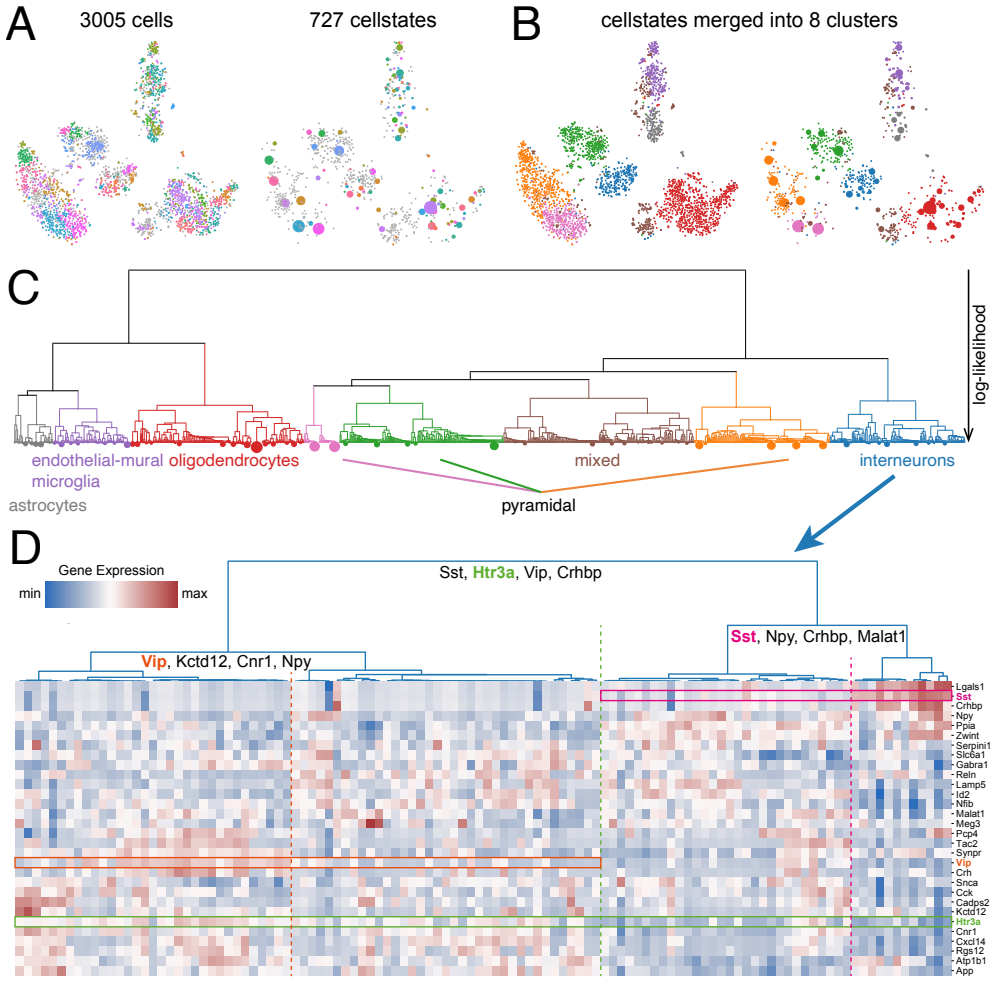



Figure 3.4: Example analysis of a mouse cortex and hippocampus dataset [75] with CELLSTATES. (A) Visualization of the data using t-SNE. The colours represent the inferred cellstates, with all singlets shown in gray. On the left, cells are shown individually while on the right cells in the same state were merged into discs. This plot contrasts regions of large gene expression diversity with many small clusters and singlets with low-diversity regions with fewer large clusters. (B) The eight higher-order clusters shown in the same visualization as in (A), with colours representing the eight higher-order clusters defined in (C). (C) Hierarchical higher-order relations between the cellstates. Leaves of the tree correspond to cellstates with their area proportional to the number of cells in them. The vertical height of the branches indicates the negative log-likelihood of the corresponding partition. This tree allows us to split the data into eight higher-order clusters that correspond well to the cell types annotated in [75]. (D) Heat map of gene expression in the interneuron-cluster. Every column corresponds to one GES and shows the corresponding expression pattern. The hierarchical tree shown on the top corresponds to the interneuron subtree of (C). Rows correspond to selected genes that are predicted to be differentially expressed between these GESs. In particular, for the first three splits in the tree, top genes contributing to their separation are indicated. Three of these are highlighted in the heat map in green, cyclamen, and orange.

that contribute most to the separation between the cellstates at opposite branches below it (Supplementary Information section 3.8.1.4), and in Fig. 3.4D plotted the expression of these genes in a heat map with rows corresponding to genes and columns to individual cellstates. As expected, all columns display unique gene expression patterns, confirming that there are clear differences between the GESs of all cellstates. Furthermore, for three nodes we highlight one example gene whose expression is clearly distinct between the corresponding branches by a rectangle in the heatmap, with the dotted line separating the cellstates on opposite sides of the branch. It should be noted that the most significant genes are those for which the *average* expression on opposite sides of the branch is most significantly different, but the expression of these genes might be quite variable across the cellstates below each branch.

At the highest level, the gene *Htr3a* (green box in Fig. 3.4D) contributes significantly to separating the expression of interneuron cellstates to the left and right of the branch (dotted green line in Fig. 3.4D). Similarly, the genes *Sst* (cyclamen box and dotted line) and *Vip* (orange box and dotted line) separate cellstates at branches lower in the tree of interneuron cellstate clusters. In general, we find many known markers of interneuron subtypes among the list of differentially expressed genes including *Sst*, *Npy*, *Crhbp*, *Cnr1*, *Cck* and *Vip* [23, 71, 75, 76] which supports the biological relevance of the cellstates that we identified.

These results illustrate how CELLSTATES uncovers substantial sub-structure among cells of the interneuron type, with the tree structure illustrating the relationships between these subtypes, and the lists of top differentially expressed genes for each branch in the tree providing information regarding the biological differences between these subtypes.

## 3.5 Discussion

With the popularization of scRNA-seq a vast number of techniques for normalising and post-processing single-cell gene expression profiles have been developed, including a large number of methods for clustering cells into 'cell types', e.g. reviewed in [3, 18, 30, 37]). These methods typically involve several complex layers of analysis steps including the normalisation of the raw data, transformation to logarithmic expression (or other 'variance stabilization' procedures), selection of 'features' to be used in the analysis, mapping of the data to a lower-dimensional representation, often involving abstract latent spaces, selection of a similarity or distance metric, and selection of the final clustering algorithm by which cells are grouped. Moreover, each of these analysis steps typically comes with tunable parameters.

Our impression of the current practice in the field is that the analysis methods are being deployed almost in a trial-and-error manner, i.e. with researchers iteratively trying out different methodologies and tweaking of their tunable parameters, comparing the results with expectations from prior biological knowledge, until results are obtained that look consistent with prior knowledge and, ideally, make some new suggestions that appear biologically plausible.

We believe this kind of approach to analysing complex data is extremely problematic. The layers of *ad hoc* processing steps and tweaking of parameters make it virtually impossible to give any unambiguous interpretation of the results, to rigorously compare results across

different studies, and prohibit direct comparison of the results with those from other experimental approaches. Instead of iterative trial-and-error tweaking of several layers of *ad hoc* methods, we feel that the proper approach to data analysis is to specify the goals of the analysis and the assumptions about the data with enough precision, such that the proper analysis method is unique and transparently follows from these specifications.

This is the approach we have taken in this paper. Instead of attempting to solve the general and very difficult problem of how to determine which cells belong to the same 'type', we aimed to solve the simpler problem of maximally reducing the complexity of a given scRNA-seq dataset without *any* loss of structure, by grouping together all cells whose gene expression states are statistically indistinguishable. We have shown that, once a rigorous specification of the measurement noise relating the gene expression states of the cells to the raw scRNA-seq data is given, the appropriate clustering algorithm solving this problem is uniquely determined from first principles. We derived analytical expressions for the posterior probabilities for partitions of the cells into non-overlapping subsets in terms of the raw UMI counts across all genes and cells in the dataset, without any tunable parameters. Moreover, the clusters in the partition with maximal posterior probability have a clear and unambiguous interpretation: they are the optimal way of splitting cells into subsets with transcriptional profiles that are identical up to measurement noise.

A key assumption that our CELLSTATES algorithm makes is that the cells in a given dataset derive from a finite number of distinct gene expression states, i.e. that in general there will be groups of cells in identical gene expression states, and one might wonder how realistic this assumption is. It is certainly possible that, rather than a discrete set of states, cells could derive from some continuous manifold of gene expression states and it is interesting to ask whether this would manifest itself in the results that we observed with CELLSTATES. If cells were derived from a continuum, then no two cells would ever be in the exact same state, and one would expect the number of cellstates to grow systematically as the number of cells increases. However, as we have seen in Fig. 3.9, we find that the observed cellstate diversity depends on the tissue of origin, and does not show systematic correlations with either number of cells or sequencing depth (i.e. total UMI count per cell). Although these are only fragmentary observations at this point and more in-depth study of this question is required, it hints that perhaps discrete cellstates do exist. However, it should also be noted that most datasets analysed here have a large fraction of singlet cellstates, i.e. clusters with only a single cell per cluster. This suggests that, for these datasets, we are still largely under-sampling the true diversity in cellstates that exist in most tissues and it is conceivable that for some tissues we might observe that the number of cellstates does continue to grow as the total number of cells increases, which might then point to the existence of continuous manifolds of cellstates.

One may also ask to what extent CELLSTATES is vulnerable to batch effects. Our measurement model makes some simplifying assumptions, such as ignoring potential systematic biases that cause transcripts of different genes to be captured with varying efficiency. We note that such gene-dependent capture efficiencies would not affect the distribution over partitions and the optimal partition $\rho^*$, as long as these capture biases are equal in all cells. In fact, as long as systematic biases are the same for the cells within each cellstate, the opti-

mal partition would even remain the same if cells in different cellstates had different capture biases. However, all analysis of gene expression differences between different cellstates of course do rely on the assumption that capture biases are the same across all cells in a given dataset.

One limitation of our approach is that, since the number of partitions increases faster than exponentially with the number of cells and is vast for any realistic scRNA-seq dataset, there is no way to guarantee that our algorithm finds the global optimum even after re-running the program several times. However, our analyses of synthetic datasets with realistic size and structure shows that, in many cases, Cellstates manages to recover the single exact partition that generated the data, and when the generating partition was not recovered exactly, most often this was because a slightly different partition with even higher likelihood was found (Fig. 3.5). On real data, we also found that the partitions obtained in different runs of cellstates are generally very similar (Fig. 3.7). These results suggest that the vast space of partitions can be effectively searched by Cellstates's Monte Carlo Markov Chain procedure.

However, it should be noted that, especially compared to most methods currently used in the field, on larger datasets Cellstates can have long run-times and requires significant computational resources. In the future we intend to improve the speed of the method by using computationally less expensive methods to either first subdivide larger datasets into coarse subsets before running Cellstates or to pre-select neighbourhood relationships, i.e. which pairs of cells are candidates for mergers with each other. Nonetheless, we note that although it may take quite some time to run Cellstates on a dataset, it generally is still considerably less than the time required to perform the experiment. Moreover, since Cellstates has no tunable parameters, the method has to only be applied once. In fact, we believe that the strong importance that is currently assigned in the field to having fast analysis methods derives largely from the fact that most researchers apply these methods in a trial-and-error manner, running many times with different parameters settings and filters until results are obtained that 'look best' by some preconceived notions of what the data should show. As we already discussed above, we think this 'fast analysis' methodology is scientifically unhealthy, and like the movement advancing 'slow food' over 'fast food', we propose that analysis of complex large-scale datasets in biology would strongly benefit from a 'slow analysis' movement that favours slow but rigorously motivated methods over iterative tuning of fast *ad hoc* methods.

Finally, we would like to comment on the way we imagine Cellstates can be applied in practice. The most obvious application of Cellstates, and the one we highlighted here, is to identify subtle substructure among known cell types, the relationships between these subtypes, and the genes that most distinguish these subtypes. However, we feel that an arguable even more important use of cellstates is as a way to significantly reduce the complexity of a dataset without losing *any* structure in the data. That is, after cells have been clustered into cellstates, one can decide to simply treat these clusters as if they were 'super cells' and perform further analysis and processing such as trajectory reconstruction, pseudo-time analysis, visualizations or differential gene expression inference treating these clusters as if they were single cells. A recent study has shown that such an approach can

indeed lead to improvements in downstream analyses [10], even in the absence of a rigorous methodology for clustering. CELLSTATES provides precisely the rigorous methodology for reducing the complexity of the dataset and removing some of the inherent noise in scRNA-seq data, while leaving all underlying biological variation completely intact. We propose that this application of CELLSTATES is an ideal first step in any scRNA-seq data analysis pipeline.

### 3.5.1 Software availability

The CELLSTATES Python package is available online on GitHub (https://github.com/nimwegenLab/cellstates). It can be run through the command-line on files containing the table of unnormalised expression data (UMI counts) in one of several formats including compressed and uncompressed tab-separated values, Matrix Market, and NumPy binary. The outputs of this command-line tool include the optimized partition of cells into cellstates, the optimized prior parameter $\Theta$, the hierarchical tree of cellstates that can be used to find higher-order clusters, and a table of differential expression scores for each gene and node in the tree. Furthermore, we provide python functions for analysing these CELLSTATES outputs, including finding the mean and modal vector of transcription quotients of each cellstate, and visualizations of the hierarchical trees as shown in this paper. We also provide several notebooks with example analyses.

## 3.6 Acknowledgments

This work was supported through grant number 310030_184937 of the Swiss National Science Foundation. We thank Thomas Sakoparnig for helpful discussions, testing of the code, and for contributing example notebooks with cellstates analysis. We thank van Nimwegen lab members for comments on the manuscript and Daan de Groot for identifying bugs in the code.

## 3.7 Tables

- Appendix Section A.1: Table A.1 summarises the datasets used in this chapter. Table A.2 described composition of the zhengmix datasets.
- Appendix Section A.2: Table A.3 summarises the clustering tools that are being compared against CELLSTATES.

## 3.8 Supplementary Information

### 3.8.1 Detailed Derivations

#### 3.8.1.1 Likelihood of partitions

We consider an scRNA-seq dataset $D$ characterized by a matrix of UMI counts $n_{gc}$ corresponding to the number of UMIs for gene $g$ in cell $c$. We denote by $\rho$ partitions of the cells into non-overlapping subsets and want to determine a likelihood function $P(D|\rho)$ under

the assumption that, for each subset $s \in \rho$ of the partition, all cells $c \in s$ have the same gene expression state. To explain how this likelihood function is calculated, we first discuss how the mRNA counts $m_g$ of a cell, and observed UMI counts $n_g$ in an scRNA-seq experiment depend on the gene expression processes in the recent history of the cell, as previously introduced in [11].

Given the inherent thermodynamic fluctuations affecting the molecules inside the cell, and the Brownian motion that they are subject to, even a comprehensive description of the current 'state' of the cell in terms of the number of molecules of each type in each cellular compartment only determines the *rates* with which different molecular reactions occur. For the mRNA levels of a given gene $g$, the relevant rates are the transcription rate $\lambda_g$ and the mRNA decay-rate $\mu_g$. It is well established that for a gene $g$ with constant transcription rate $\lambda_g$ and constant mRNA decay-rate $\mu_g$, the number of mRNA molecules in the cell $m_g$ follows a Poisson distribution with mean $\langle m_g \rangle = a_g = \lambda_g/\mu_g$, e.g. [47]. More generally, when $\mu_g$ and $\lambda_g$ are arbitrary time-dependent functions $\mu_g(t), \lambda_g(t)$, with $\lambda_g(t)$ denoting the transcription rate a time $t$ in the past of the cell, and $\mu_g(t)$ the decay rate of mRNAs for gene $g$ a time $t$ in the past, the probability distribution for the current number of mRNAs $m_g$ in the cell is still a Poisson distribution with mean [11]:

$$a_g = \langle m_g \rangle = \int_0^\infty dt \lambda_g(t) \exp\left[ - \int_0^t \mu_g(s) ds \right], \tag{3.9}$$

which we call the 'transcription activity' of gene $g$. Note that time is measured backwards from the present ($t = 0$) to the distant past ($t = \infty$) in the history of the cell. Thus, a single parameter $a_g$ for each gene $g$ is sufficient to fully characterize the distribution of mRNA numbers in a cell at any given time point. The remaining uncertainty about the actual numbers is due to random thermodynamic fluctuations in events such as RNA polymerase binding or mRNA degradation. To conclude, given the expression state of the cell as defined by the vector of transcription activities $\vec{a}$, the probability of a count vector of cellular mRNAs across all genes $\vec{m}$ is therefore a product of Poisson distributions:

$$P(\vec{m}|\vec{a}) = \prod_g \frac{(a_g)^{m_g}}{m_g!} e^{-a_g}. \tag{3.10}$$

We will assume that the *measured* UMI counts $n_g$ correspond to a random sample of the cell's total mRNA pool $m_g$ with some unknown capture rate $p$ per mRNA. As will be discussed below, our model remains valid if the capture rate $p$ varies between cells or has gene-dependent biases. Given these assumptions, the likelihood for the observed UMI count vector $\vec{n}$ is still a Poisson distribution, albeit with a different mean:

$$P(\vec{n}|\vec{a}, p) = \prod_g \frac{(pa_g)^{n_g}}{n_g!} e^{-pa_g}. \tag{3.11}$$

Following [11], we now define the transcription quotients $\alpha_g = a_g/A$, with $A = \sum_g a_g$ the total transcription activity of the cell. Note that $\alpha_g$ corresponds to the expected fraction of transcripts from gene $g$ among all transcripts in the cell. For a cell with a total UMI count $N = \sum_g n_g$, we have

$$\langle n_g \rangle = \alpha_g p A = \alpha_g N. \tag{3.12}$$

Conditioned on the total count $N$, the distribution of all measured counts $\vec{n}$ is a multinomial in the transcription quotients:

$$P(\vec{n}|\vec{\alpha}, N) = N! \prod_g \frac{1}{n_g!} (\alpha_g)^{n_g} \propto \prod_g (\alpha_g)^{n_g}. \tag{3.13}$$

This is the form of the likelihood of the UMI counts of a single cell as a function of the transcription quotient vector $\vec{\alpha}$.
In our model, we use the transcription quotient vector $\vec{\alpha}$ to represent the 'expression state' of a cell and the key ingredient of our model is that, given a partition $\rho$, all cells within each subset $s$ of the partition have the same transcription quotient vector $\vec{\alpha}$. The likelihood for the counts $D_s$ of a subset of cells $s$ that have equal transcription quotients $\vec{\alpha}$, is given by

$$P\left(D_s|\vec{\alpha}\right) = \prod_{c \in s} P(\vec{n_c}|\vec{\alpha}) \propto \prod_g (\alpha_g)^{n_{gs}} \tag{3.14}$$

where $\vec{n_s} = \sum_{c \in s} \vec{n_c}$ is the vector of total UMI counts among all cells in the subset $s$.
To calculate the likelihood $P(D|\rho)$ of a partition, we need to marginalize over the unknown transcription quotient vector $\vec{\alpha}$ for each of the subsets $s$ in $\rho$ and to do this we have to define a *prior* distribution over possible transcription quotient vectors $\vec{\alpha}$. We will use a Dirichlet prior, which corresponds to a maximal ignorance prior in the sense that it is the unique prior that is invariant under arbitrary rescaling of the transcription quotients $\alpha_g \to \lambda_g \alpha_g$. Moreover, it is the conjugate prior to the multinomial distribution, allowing us to analytically marginalize over the $\vec{\alpha}$. In particular, for each subset $s$ we characterize our prior information regarding its transcription quotient vector $\vec{\alpha}$ by the *same* Dirichlet prior:

$$P(\vec{\alpha}|\vec{\theta}) = \frac{\Gamma(\Theta)}{\prod_g \Gamma(\theta_g)} \prod_g (\alpha_g)^{\theta_g - 1}, \tag{3.15}$$

where the $\theta_g$ are the parameters of the Dirichlet prior and $\Theta = \sum_g \theta_g$. We can now marginalize over $\vec{\alpha}$ and obtain

$$\begin{aligned} P(D_s|\vec{\theta}) &= \int_{\sum_g \alpha_g = 1} P(D_s, \vec{\alpha}|\vec{\theta})\, d\vec{\alpha} && (3.16)\\ &= \int_{\sum_g \alpha_g = 1} P(D_s|\vec{\alpha}) P(\vec{\alpha}|\vec{\theta})\, d\vec{\alpha} && (3.17)\\ &= \frac{\Gamma(\Theta)}{\Gamma(N_s + \Theta)} \prod_g \frac{\Gamma(n_{gs} + \theta_g)}{\Gamma(\theta_g)} && (3.18) \end{aligned}$$

with $N_s = \sum_g n_{gs}$ is the total number of UMI summed over all cells in subset $s$. The likelihood of a partition $\rho$ is now obtained by simple taking the product of this expression over all subsets $s \in \rho$:

$$\begin{aligned} P(D|\rho, \vec{\theta}) &= \prod_{s \in \rho} P(D_s|\vec{\theta}) && (3.19)\\ &= \prod_{s \in \rho} \left[ \frac{\Gamma(\Theta)}{\Gamma(N_s + \Theta)} \prod_g \frac{\Gamma(n_{gs} + \theta_g)}{\Gamma(\theta_g)} \right]. && (3.20) \end{aligned}$$

Note that this expression is very similar to likelihood functions derived previously for clustering DNA sequences [68]. Essentially the only change is that the 4-letter DNA alphabet is here replaced by the 'alphabet' of $G$ genes.

Ideally, we would search for the combination $(\rho, \vec{\theta})$ that jointly maximizes the likelihood, i.e. optimizing both the partition $\rho$ and the parameters $\theta_g$ for each gene individually, but this is computationally intractable. Without loss of generality, we can rewrite the parameters of the prior $\theta_g$ as the product of an overall scale vector $\Theta$ and a normalised vector $\vec{\phi}$ with $\sum_g \phi_g = 1$. Second, we note that for the trivial partition in which all cells are put into a single cluster, the optimal $\phi_g$ are given by

$$\phi_g = \frac{\sum_c n_{gc}}{\sum_{c,g} n_{gc}}, \tag{3.21}$$

i.e. the prior parameter $\phi_g$ simply equals the fraction of UMIs for gene $g$ in the entire dataset. We will simplify the optimization of the prior's parameters by fixing $\vec{\phi}$ to this vector, setting $\vec{\theta} = \Theta\vec{\phi}$, and only optimize the scale factor $\Theta \in \mathbb{R}^+$, while leaving the $\phi_g$ fixed for a given dataset. Setting the prior in this way ensures that, for each subset $s$, the expected direction of the transcription quotient vector $\vec{\alpha}$ matches the overall UMI counts in the entire dataset, while optimizing $\Theta$ allows the tuning of the expected amount of variability around this 'average' vector of transcription quotients.

With this chosen form of the prior, we finally get an expression for the likelihood of the whole dataset $D$ that only depends on the scale factor $\Theta$ and the partition of cells into subsets $\rho$:

$$P(D|\rho, \Theta) = \prod_{s\in\rho} P(D_s|\Theta) \tag{3.22}$$

$$= \prod_{s\in\rho} \left[ \frac{\Gamma(\Theta)}{\Gamma(N_s + \Theta)} \prod_g \frac{\Gamma(n_{gs} + \Theta\phi_g)}{\Gamma(\Theta\phi_g)} \right]. \tag{3.23}$$

Finally, we return to discuss our simplifying assumption that the mRNA capture rate $p$ is constant across genes and cells and show that this assumption can be significantly relaxed without affecting the results. In particular, we can assume that the probability $p_{gc}$ of capturing (and successfully amplifying and sequencing) an mRNA for gene $g$ in cell $c$ can be written as

$$p_{gc} = p_c q_g, \tag{3.24}$$

where $p_c$ is a cell-specific overall capture rate and $q_g$ describes gene-dependent biases that may be specific to the particular experiment, but are assumed constant across the cells in the experiment. With this capture efficiency, the expected UMI count for gene $g$ in cell $c$ is $\langle n_{gc}\rangle = p_c q_g a_g$. Thus, the expected fraction of counts from gene $g$ becomes

$$\frac{\langle n_{gc}\rangle}{N_c} = \frac{a_g p_c q_g}{\sum_g a_g p_c q_g} = \frac{a_g q_g}{\sum_g a_g q_g} = \tilde{\alpha_g}, \tag{3.25}$$

where the last equality defines $\tilde{\alpha}_g$. From here, we can proceed the derivation exactly as before, with $\tilde{\alpha_g}$ replacing $\alpha_g$. As we marginalize out this variable in Equation 3.18, the

final result is invariant. Note that, given that we separately marginalize over the $\vec{\alpha_g}$ of each subset, the result is even invariant when different subsets $s$ have different gene-bias vectors $\vec{q}$, as long as all cells within a subset have the same bias. This suggests that our likelihood over partitions is not only insensitive to fluctuations in overall capture efficiency across cells, but will also be quite robust to fluctuations in gene-dependent capture efficiency as long as cells with equal expression states have equal biases.

#### 3.8.1.2 Posterior of transcription quotients

Although for calculating likelihoods over partitions $\rho$, we marginalize over the GES $\vec{\alpha_s}$ for each subset $s$, we can of course also obtain posterior distributions over these GES for each cluster. Given a subset of cells $s$, the posterior distribution for its GES $\vec{\alpha_s}$ can be obtained from equations 3.14, 3.15 and 3.18 to find:

$$P(\alpha_s|D_s,\Theta) = \frac{P(\alpha_s|\Theta)P(D_s|\alpha_s)}{P(D_s|\Theta)} = \frac{\Gamma(\Theta+N_s)}{\prod_g \Gamma(n_{gs}+\Theta\phi_g)} \prod_g (\alpha_g)^{\Theta\phi_g+n_{gs}-1}. \tag{3.26}$$

From this expression, we can derive expressions for mode, mean and variance of $\vec{\alpha_s}$:

$$\text{Mode}[\alpha_{gs}] = \frac{\max(\Theta\phi_g+n_{gs}-1,0)}{\sum_g \max(\Theta\phi_g+n_{gs}-1,0)} \tag{3.27}$$

$$\langle\alpha_{gs}\rangle = \frac{\Theta\phi_g+n_{gs}}{\Theta+N_s} \tag{3.28}$$

$$\text{Var}[\alpha_{gs}] = \frac{(\Theta\phi_g+n_{gs})(\Theta(1-\phi_g)+N_s-n_{gs})}{(\Theta+N_s)^2(\Theta+N_s+1)} \tag{3.29}$$

Often we are interested in the log-transcription quotients $\delta_{gs} = \log(\alpha_{gs})$ rather than the transcription quotients themselves. For these we find:

$$\text{Mode}[\delta_{gs}] = \log(\Theta\phi_g+n_{gs}) - \log(\Theta+N_s) \tag{3.30}$$

$$\langle\delta_{gs}\rangle = \psi(\Theta\phi_g+n_{gs}) - \psi(\Theta+N_s) \tag{3.31}$$

$$\text{Var}[\delta_{gs}] = \psi_1(\Theta\phi_g+n_{gs}) + \psi_1(\Theta+N_s), \tag{3.32}$$

where $\psi$ is the digamma function (the derivative of the logarithm of the Gamma function) and $\psi_1$ is the first derivative of the digamma function.

#### 3.8.1.3 Cellstate similarities and hierarchical clustering

We define the similarity between two cellstates $S_a$ and $S_b$ as the ratio of the likelihoods of the partitions with both subsets merged and each subset separate:

$$\begin{aligned}\sigma_{ab} =& \frac{P(D|\rho(S_a+S_b),\Theta)}{P(D|\rho(S_a,S_b),\Theta)} \\ =& \frac{P(\vec{n_a}+\vec{n_b}|\Theta\vec{\phi})}{P(\vec{n_a}|\Theta\vec{\phi})P(\vec{n_b}|\Theta\vec{\phi})}\end{aligned} \tag{3.33}$$

Note that this similarity metric does not behave like usual distance metrics, in that similar clusters will have a large $\sigma$ and dissimilar clusters a small $\sigma$ close to 0. To build the hierarchical tree of the cellstates, we start by setting the leaf clusters of the tree to the cellstates of the optimal partition and then calculate all pairwise similarities $\sigma_{ab}$ between all pairs of leaf clusters. We then iteratively merge the pair of clusters with the highest similarity, and recalculate the pairwise similarities with the newly formed cluster until all clusters have merged into a single cluster. For plotting, we save the resulting tree in the Newick format with distances set to positive log-similarities $d_{ab} = \max[-\log(\sigma_{ab}), 0]$.

#### 3.8.1.4 Differentially expressed genes between pairs of clusters

To describe the differences between the cellstates of different clusters, and to help give biological interpretation, it is useful to quantify which genes are most differentially expressed between the clusters. In our framework, we can quite naturally define the extent of differential expression of genes by decomposing Equation 3.33 into contributions of individual genes, i.e. $\sigma_{ab} = \prod_g \sigma_{ab,g}$. A low value of $\sigma_{ab,g}$ for a gene $g$ indicates that the differences in UMI counts are much more different between the two clusters $a$ and $b$ than would be expected from noise. A high value $\sigma_{ab,g}$, in contrast, indicates that counts are within the expected noise levels.

To obtain such a decomposition, we start by decomposing the cluster likelihood of Equation 3.18 into contributions from individual genes $P(\vec{n}|N,\Theta) = \prod_g P(n_g|N,\Theta)$. In the multinomial model, the likelihood is conditioned on the total number of captured mRNA in the cell $N$, so that, formally, the counts $n_{gc}$ are correlated for all pairs of genes. However, since this correlation is generally quite weak, we can make the assumption that the expression noise is independent between genes. Thus,

$$P(\vec{n},\vec{\alpha}|\Theta) \approx \prod_g P(n_g,\alpha_g|\Theta) \tag{3.34}$$

$$= \prod_g P(n_g|\alpha_g)P(\alpha_g|\Theta). \tag{3.35}$$

We can find $P(\vec{n}|\alpha_g)$ by marginalizing over the other variables:

$$P(\vec{n}|\alpha_g) \propto \int_{\sum_{g'\neq g}\alpha_{g'}=1-\alpha_g} P(\vec{n}|\vec{\alpha})P(\{\alpha_{g'\neq g}\}|\alpha_g,\vec{\theta})\,\{d\alpha_{g'\neq g}\} \tag{3.36}$$

$$\propto (\alpha_g)^{n_g}(1-\alpha_g)^{N-n_g} \tag{3.37}$$

where $N = \sum_g n_g$. We can further marginalize over $\{n_{g'\neq g}\}$, requiring that $N$ is constant, and normalise to obtain the binomial distribution:

$$P(n_g|\alpha_g,N) = \binom{N}{n}(\alpha_g)^{n_g}(1-\alpha_g)^{N-n_g}. \tag{3.38}$$

Next, we find $P(\alpha_g|\Theta)$:

$$P(\alpha_g|\Theta) = \int_{\sum_{g'\neq g}\alpha_{g'}=1-\alpha_g} P(\vec{\alpha}|\Theta)\, d\vec{\alpha} \tag{3.39}$$

$$= \frac{1}{B(\Theta\phi_g, \Theta(1-\phi_g))}(\alpha_g)^{\Theta\phi_g-1}(1-\alpha_g)^{\Theta(1-\phi_g)-1}, \tag{3.40}$$

where $B(\alpha,\beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha\beta)$.
Finally, we have all ingredients of Equation 3.35 above. Unlike in Equation 3.18 where we perform the integral subject to the constraint $\sum_g \alpha_g = 1$, we integrate over all $\alpha_g$ separately. With the high dimensionality of $\vec{\alpha}$ and assuming the likelihood function has a sharp peak, the error will be small.

$$P(\vec{n}|\theta) \propto \prod_g \int_0^1 P(n_g|\alpha_g)P(\alpha_g|\vec{\theta})\, d\alpha_g \tag{3.41}$$

$$= \prod_g \int_0^1 \frac{1}{B(\theta_g, \Theta-\theta_g)}(\alpha_g)^{n_g+\theta_g-1}(1-\alpha_g)^{N-n_g+\Theta-\theta_g-1}\, d\alpha_g \tag{3.42}$$

$$= \prod_g \frac{B(n_g+\theta_g, N-n_g+\Theta-\theta_g)}{B(\theta_g, \Theta-\theta_g)} \tag{3.43}$$

This likelihood function clearly has a separate contribution $P(n_g|N,\Theta)$ for each gene:

$$P(n_g|N,\Theta) = \frac{B(n_g+\Theta\phi_g, N-n_g+\Theta(1-\phi_g))}{B(\Theta\phi_g, \Theta(1-\phi_g))} \tag{3.44}$$

$$= \frac{\Gamma(\Theta)}{\Gamma(\Theta(1-\phi_g))\Gamma(\Theta\phi_g)}\frac{\Gamma(n_g+\Theta\phi_g)\Gamma(N-n_g+\Theta(1-\phi_g))}{\Gamma(N+\Theta)} \tag{3.45}$$

Comparing to Equation 3.18, we see that this is equivalent to taking the ratio of the cluster likelihood with gene $g$ and without $g$.
Finally, we can use this expression for $P(n_g|N,\Theta)$ in Equation 3.33 and define a gene-specific score for differential expression

$$\sigma_{ab,g} = \frac{P(n_{a,g}+n_{b,g}|N_a+N_b,\Theta)}{P(n_{a,g}|N_a,\Theta)P(n_{b,g}|N_b,\Theta)}, \tag{3.46}$$

where the subscripts $a, b$ refer to two subsets of cells. Genes with $\sigma_{ab,g} < 1$ have counts that poorly fit a model with a single transcription quotient for both clusters, compared to a model where the two clusters have distinct transcription quotients. In contrast, genes with a score $\sigma_{ab,g} > 1$ favour a model with a single transcription quotient and are not differentially expressed between the clusters.

### 3.8.2 Computational Methods

#### 3.8.2.1 MCMC Algorithm for maximizing the likelihood

Our aim is to identify the combination of a scale factor $\Theta$ and partition $\rho$ that jointly maximize the likelihood $P(D|\rho,\Theta)$ given in Equation 3.22. To do this, we start from an initial guess for $\Theta$ and then iteratively

1. Search for the partition $\rho_*$ that maximizes the likelihood with the current value of $\Theta$,
2. Given $\rho_*$, find the value of $\Theta$ that maximizes the likelihood $P(D|\rho_*, \Theta)$,

until convergence. In order to limit the number of time-costly optimizations of the partition, we only consider values of $\Theta = 2^q, q \in \mathbb{N}$. The initial guess is taken with $q = \lfloor \log_2(\langle N_{UMI} \rangle) + 0.5 \rfloor$ where $\langle N_{UMI} \rangle$ is the average number of total UMI counts per cell. That is, our initial guess for $\Theta$ corresponds to the average total UMI count per cell. This means that the strength of the influence of the prior is about equal to the influence of the data from a single-cell.

To optimize the partition $\rho$ at a given $\Theta$, we start with the partition in which each cell forms its own cluster. To explain how the space of partitions is searched, we conceptualize partitions as putting cells into 'boxes', i.e. such that all cells in the same box form a cluster and different boxes correspond to different clusters. Initially we assign each of the $C$ cells into one of $C$ boxes and we will search the space of partitions by moving cells between these $C$ boxes. Specifically, we use a Markov chain Monte Carlo (MCMC) algorithm, which iterates the following steps:

1. Given the current partition $\rho$, a cell is chosen uniformly at random and taken out of its current box.
2. One of the $C - 1$ other boxes is chosen uniformly at random and we consider the partition $\rho'$ that is created by moving the cell into this box.
3. We calculate the likelihood ratio of the new to old partitions: $P_{\text{move}} = P(D|\rho', \Theta)/P(D|\rho, \Theta)$.
4. If the new partition $\rho'$ has $N_{\text{clus}}$ clusters and $\rho$ had $N_{\text{clus}} + 1$ clusters, we set $P_{\text{bias}} = (C - N_{\text{clus}})$, otherwise $P_{\text{bias}} = 1$.
5. If $P_{\text{accept}} = P_{\text{move}} * P_{\text{bias}} > 1$, the move is accepted, otherwise it is accepted with probability $P_{\text{accept}}$.

Note that, as explained in the supplementary material of [68], the correction factor $P_{\text{bias}}$ ensures detailed balance, i.e. that in the absence of differences in the likelihoods of the partitions $P(D|\rho, \Theta)$, all partitions would be sampled uniformly.

These steps are iterated until the likelihood has stopped increasing for a sufficient number of steps. In the current implementation of the algorithm, the stopping criterion is controlled by two parameters: the number of steps $S$ and the number of tries per step $T$. Each round, a total number of $S \times T$ moves are attempted. This is repeated if at least $S$ of the trials led to an accepted move, i.e. the partition was changed at least $S$ times. If less than $S$ moves were made in the $S \times T$ trials, the value of $S$ is reduced by 10 and the value of $T$ is multiplied by 10, and new rounds of trials are started. This is continued until $S$ falls below 10, after which the MCMC moves are stopped. Note that the algorithm keeps track of the partition with the highest likelihood it has seen, and set the partition to this highest likelihood partition at the end of the rounds of MCMC moves. By default, we set $S = N$, i.e. equal to the number of cells, and $T = 1000$, but these values can be changed by the user.

After these MCMC moves, a final uphill walk is performed as follows. For each pair of clusters existing in the partition, we calculate the likelihood change that would occur if the clusters were merged into one. We then iteratively merge clusters until no more mergers are left that would increase the likelihood. Finally, for each cell we calculate the likelihood change that would occur if the cell were moved into any of the currently existing clusters, and move the cell to its optimal cluster.

#### 3.8.2.2 Simulated Datasets

Simulated datasets were created based on the inferred cellstates of real datasets. Of the 36 datasets which were analysed for this paper, we selected those 18 that have less than 6000 cells, more than 3 cellstates with more than 10 cells in them, and a median number of UMIs per cell greater than 1000. Our aim was to simulate datasets based on the set of transcription quotients inferred to be present in the real datasets. Additionally, we wanted to make sure that the clusters can only be identified by differences in transcription quotients (i.e. relative gene expression levels) and not by differences in total UMI counts. The total number of UMI for each cell $c$, $N_c$, was therefore drawn independently from a log-normal distribution that was fitted to the experimental distribution of $N_c$ in the corresponding dataset.

For each cellstate $s$, we determined the mean expression transcription quotient vector $\langle\vec{\alpha}_s\rangle$ and then sampled the UMI count vectors $\vec{n_c}$ of each cell $c$ in the cellstate $s$ from a multinomial distribution with mean $\langle\vec{\alpha}_s\rangle$. However, we found that the maximal likelihood partition of the simulated dataset often differed from the partition generating the dataset, especially for very small and singlet clusters. In particular, when cells that were in a singlet state were given a lower total UMI count in the simulation, these cells were often no longer statistically significantly different from other states, and this to a lesser extent also affected small clusters. To mitigate this problem, we retained only cellstates with more than 10 cells. The number of cells in each cellstate was kept the same as in the experimental data. With this simulation procedure, we made sure that most simulated cellstates would be statistically distinct, even when total UMI counts were lower than in the original dataset.

In this way, three separate simulated datasets were generated for each experimental dataset. Lastly, an additional three "down-sampled" simulations were carried out with $\langle\log_{10}(N_c)\rangle = 3$ and $\mathrm{var}(\log_{10}(N_c)) = 0.1$ fixed for all datasets.

#### 3.8.2.3 Further Discussion of the results on simulated data

In Figure 3.5 we show, for each of the three simulations (marked by their colour) generated per experimental dataset (different columns), the detailed outcomes of three independent CELLSTATES runs per simulation. The top panel shows the difference in log-likelihood $\Delta L$ between the inferred partition and the partition used to generate the simulated data. Negative scores mean that the inferred partition had a lower likelihood than the simulated one and can be attributed to a failure of the algorithm to find the optimal partition. As can be seen, such errors are rare and there is always at least one out of the three runs that has $\Delta L \geq 0$, i.e. a partition at least as good as the one used to generate the dataset was always

found. Positive scores indicate that a partition was found with a higher likelihood than the one used to generate the dataset. This can happen for example if there are not enough counts in the simulated cells to statistically distinguish cellstates. To test this hypothesis, we looked at the "down-sampled" simulations with fewer UMI per cell which would make cells with similar, but distinct transcription quotients indistinguishable. Indeed, the results shown in Figure 3.6 confirm this hypothesis: For most down-sampled simulations, the best-scoring partition is different from the ground-truth. Also, they tend to score low on homogeneity but high on completeness - which means that inferred clusters are unions of clusters used to generate the simulated data.

### 3.8.3 Supplementary Figures

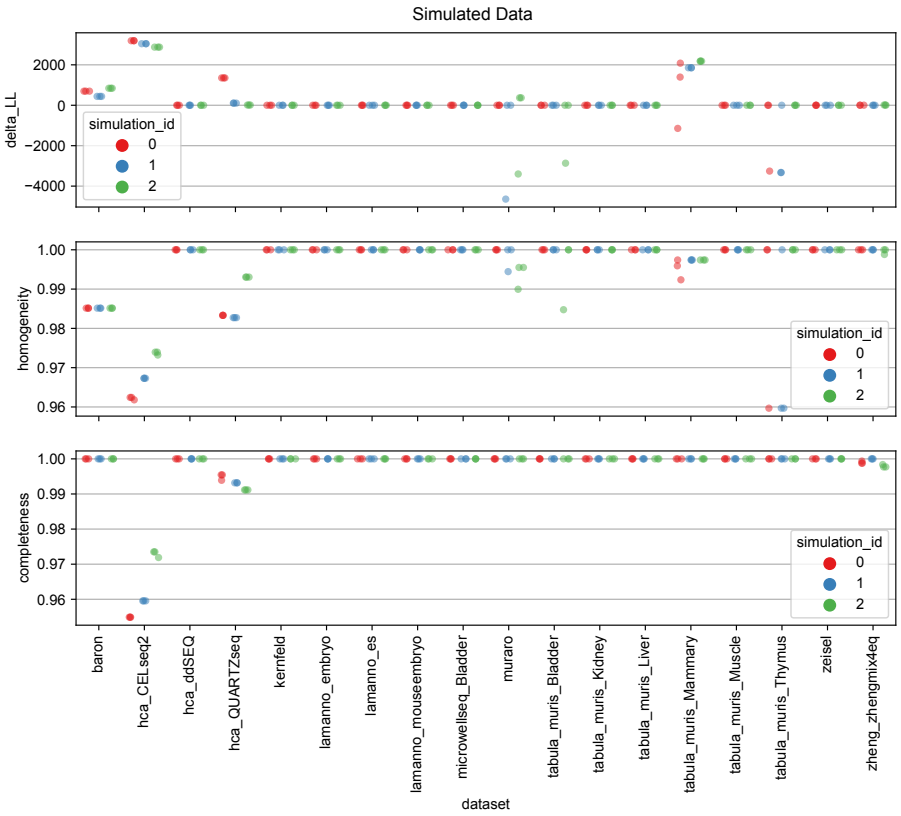



Figure 3.5: Detailed results from CELLSTATES runs on simulated data based on the set of GESs from various indicated datasets. For each of the 18 real datasets, three simulations were generated (red, blue, and green) and CELLSTATES was run three times on each simulated dataset. In the top panel, the difference in log-likelihood delta_LL between the inferred and simulated partitions is shown (with a positive difference meaning that a partition was found with higher log-likelihood than the one used to simulate the data). The corresponding homogeneity and completeness of the inferred compared to the simulated partitions are given in the two lower panels.

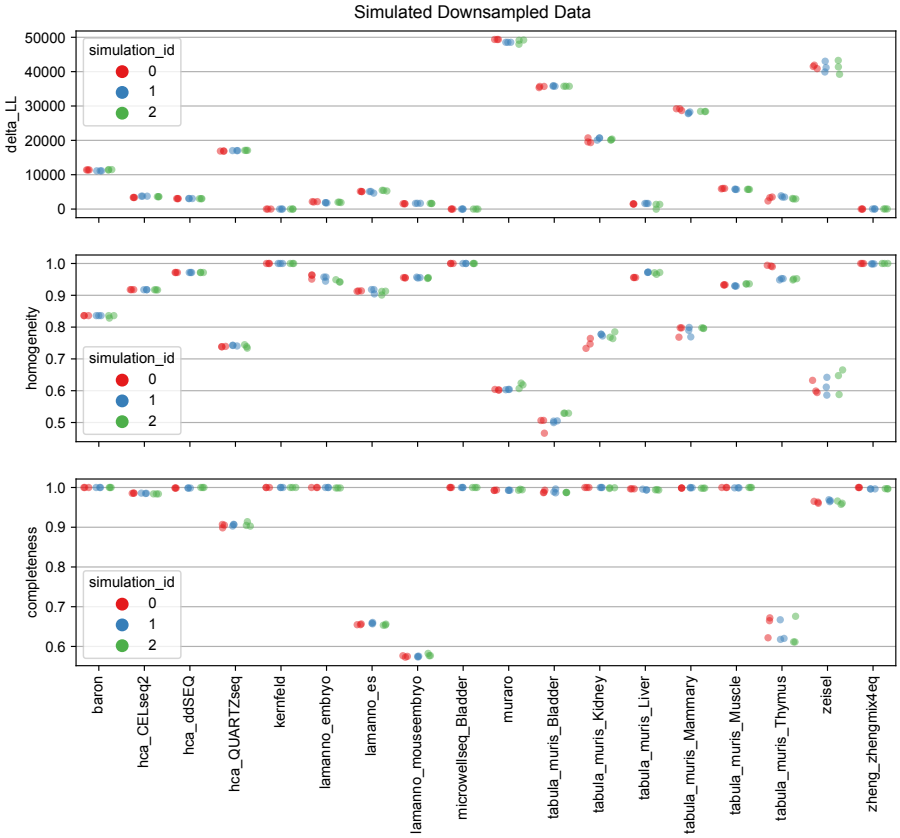



Figure 3.6: Detailed results from CELLSTATES runs on down-sampled simulated data based on the set of GESs from various indicated datasets, but with a median of only 1000 UMI per cell. For each of the 18 GES-sets, three simulations were generated (red, blue, and green) and CELLSTATES was run three times on each simulation. In the top panel, the difference in log-likelihood delta_LL between the inferred partition and the partition used to generate the simulated data is shown (positive meaning that a partition was found with higher log-likelihood than the one used to generate the data). The corresponding homogeneity and completeness of the inferred partitions compared to the partitions used to generate the data are shown in the two lower two panels.

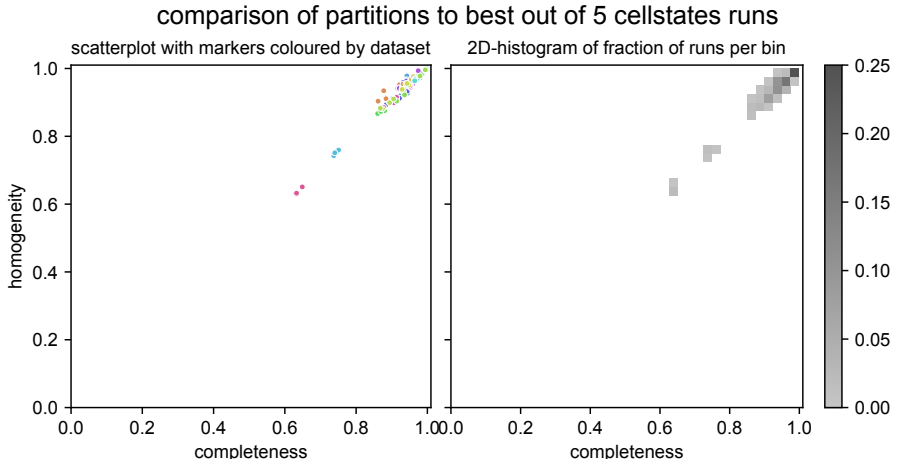



Figure 3.7: Reproducibility of independent CELLSTATES runs. CELLSTATES was run 5 times on each of 34 scRNA-seq datasets, and the highest-likelihood partition that was found was compared with those of the 4 other runs. The resulting homogeneity and completeness scores are shown twice. On the left as a scatter-plot with markers coloured by dataset, illustrating which outliers belong to the same dataset. On the right, a 2D-histogram shows the fraction of runs in bins of size $0.025 \times 0.025$, showing that most fall within the top right corner with homogeneity and completeness $> 0.95$.

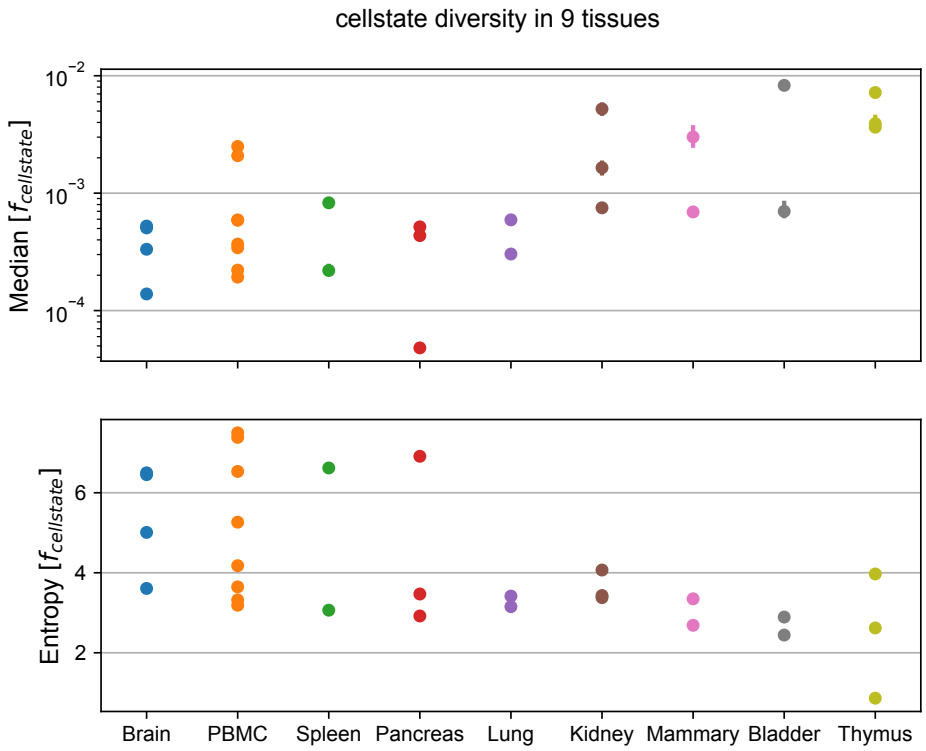



Figure 3.8: Tissue of origin is predictive for cellstate diversity. For 29 different datasets from 9 tissues, the entropy of the distribution of cellstate abundances $f_{\text{cellstate}}$ in each dataset are shown by tissue. Although there is large variability across datasets, datasets from the same tissue tend to have similar cellstate diversity. Tissues are sorted roughly from most to least diverse, from left to right.

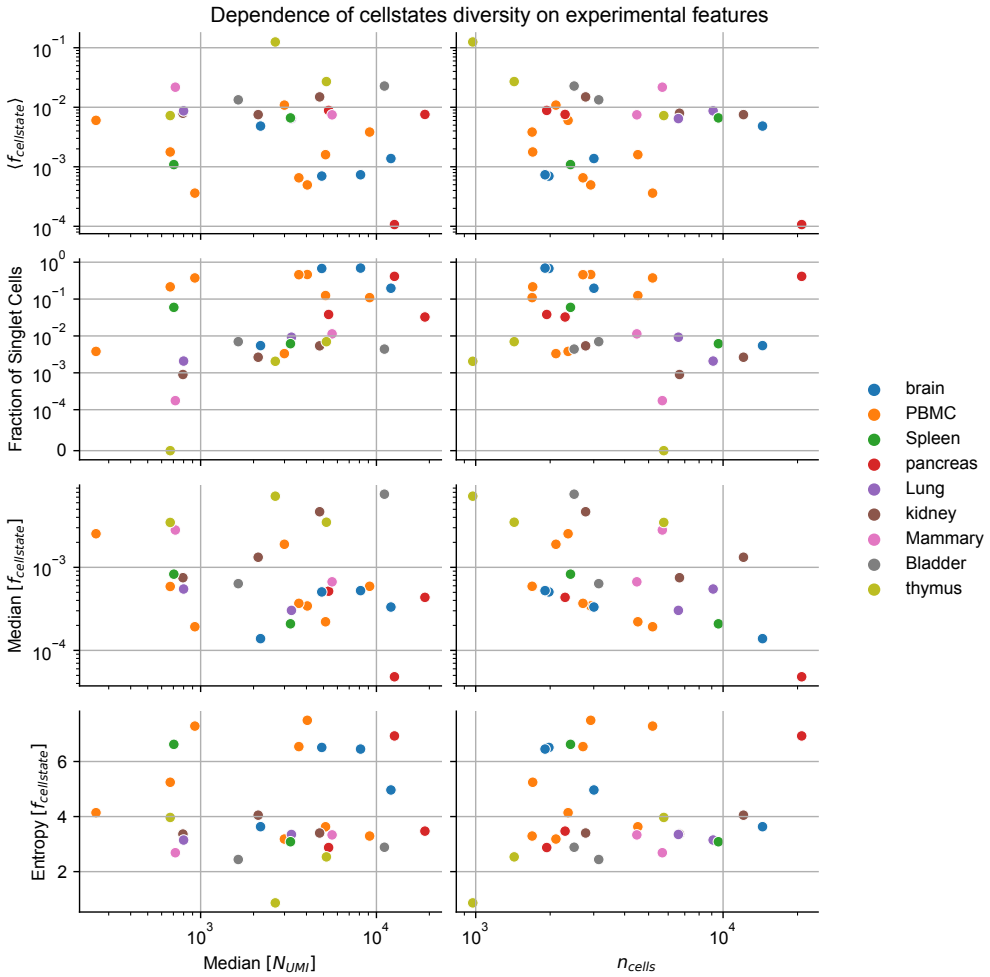



Figure 3.9: Cellstate diversity does not depend on technical features of the experiment. The mean of the distribution of cellstate abundances $f_{\text{cellstate}}$ (top row panels), the fraction of singlet cellstates (second row panels), the median (third row panels) and the entropy (bottom row panels) of $f_{\text{cellstate}}$ are shown as a function of the median total number of UMIs per cell ($N_{UMI}$) (left column panels) and the total number of cells in each dataset (right column panels). Marker colours indicate the tissue from which the samples originate. These results show that there is no correlation between the various diversity measures and either sequencing depth (total UMI count) or the number of cells sequenced, with the exception of a weak negative correlation between the median cellstate abundance Median$[f_{\text{cellstates}}]$ and the number of cells $n_{\text{cells}}$. This weak correlation is explained by the fact that the majority of cellstates are singlets in many experiments. Note that the variation in diversity measures for data from the same tissue (i.e. dots of the same colour) along the $y$-axes is much less than that along the $x$-axes, which means that cellstate diversity is largely driven by biological and not technical experimental features.

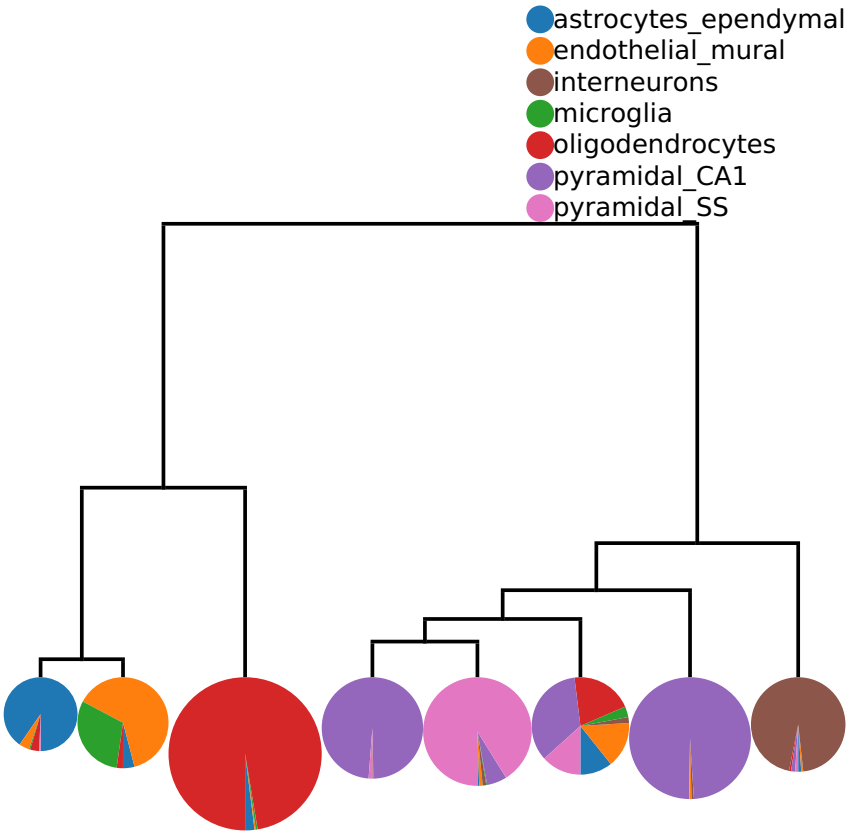



Figure 3.10: Cells in the 8 major cellstate clusters of the Zeisel dataset [75] largely match annotations provided in that publication. colours indicate different annotations from [75] (see legend), each pie chart corresponds to one higher-order cluster, and the area in each pie chart is proportional to the number of cells with the corresponding annotation in that cluster. Note that cells annotated by [75] as microglia and endothelial-mural cells are merged into one cluster at this level of the hierarchy. The tree structure indicates how these clusters are related upon further merging.

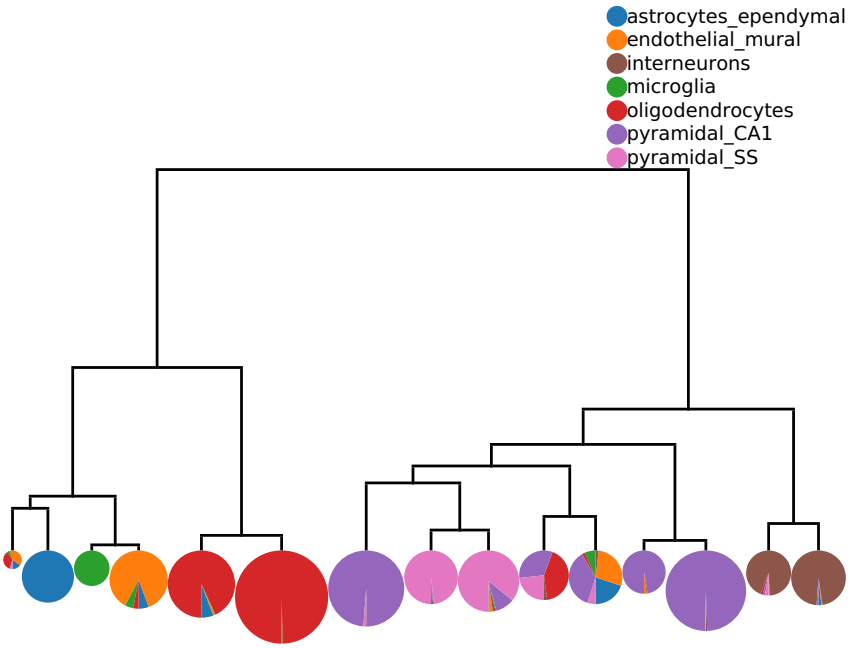



Figure 3.11: At 15 higher-order cellstate clusters, cells annotated in [75] as microglia and endothelial-mural separate. colours indicate different annotations from [75] (see legend) and the area in each pie chart is proportional to the number of cells in the corresponding cluster with the corresponding annotation. Note that at this level of the hierarchy, cells with a common annotation in [75] tend to separate into multiple higher-order cellstate clusters.

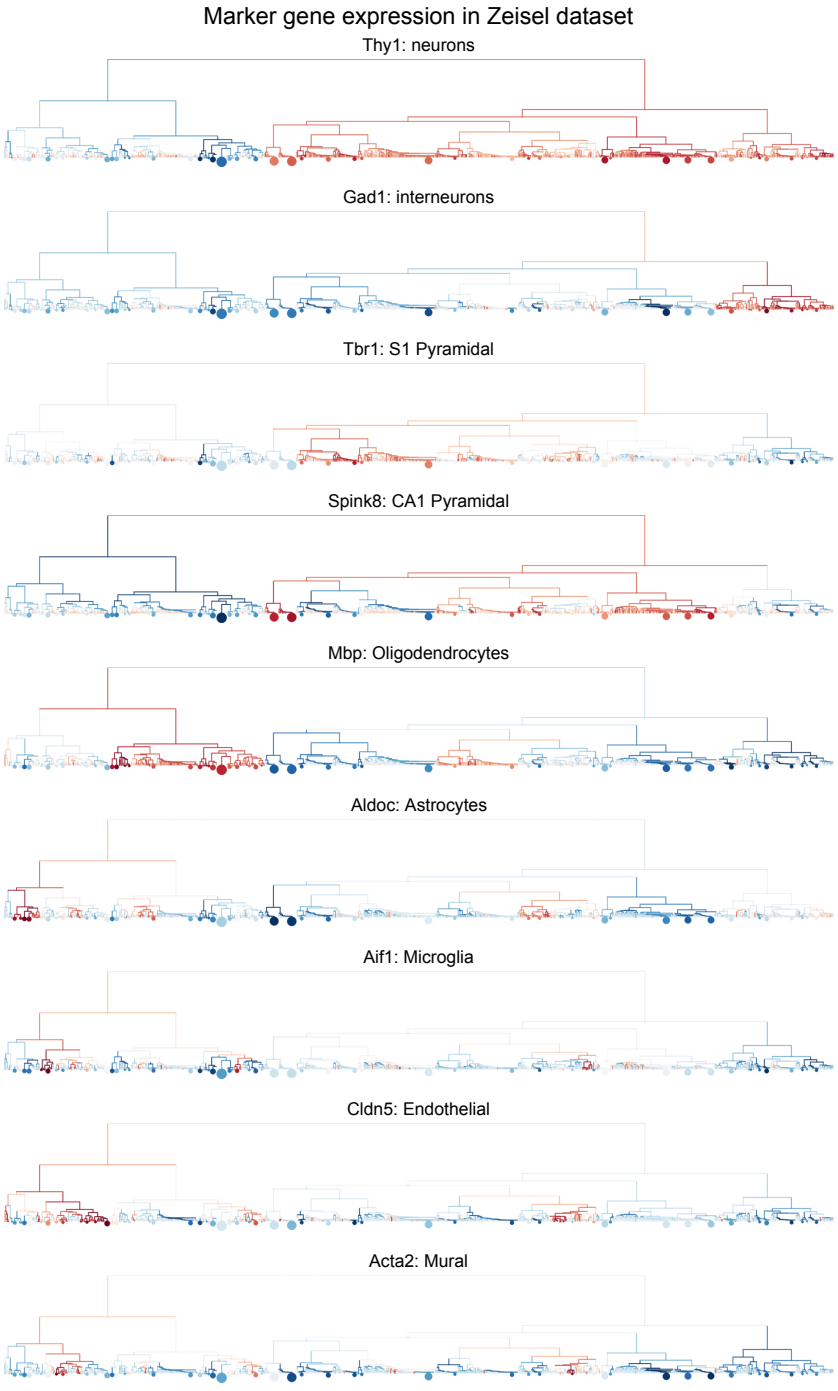



Figure 3.12: Expression of marker genes for annotated cell types visualized in the hierarchical tree of higher-order cellstate clusters. The sizes of the discs at the leaves correspond to the numbers of cells in the corresponding cellstates. See Figure 3.4 to compare with the higher-order clusters of CELLSTATES.


Figure 3.13: Expression of marker genes for annotated cell types visualized in the hierarchical tree of higher-order cellstate clusters. The sizes of the discs at the leaves correspond to the numbers of cells in the corresponding cellstates. See Figure 3.4 to compare with the higher-order clusters of CELLSTATES.

# 4

# Conclusion: Can we formalise the concept of a "cell type" in transcriptomics?

In a 2017 edition, the journal Cell Systems asked 15 scientists for their "conceptual definition of 'Cell Type' in the context of a mature organism" [16]. In the replies, one can probably find more than 15 different answers to that question. Outside of that, there have been several recent attempts at designing systematic frameworks for classifying and defining cell types [40, 41, 72]. There are many attributes with which we can describe cells: their developmental history and potential, their morphology, intracellular organisation, function, location, gene expression pattern, protein content, response to stimuli, etc. Of course, focussing only on gene expression by measuring the transcriptome is a strong simplification of the full complexity of the molecular state of a cell. Nonetheless, it indirectly captures a lot of other information such as active gene regulatory networks [4, 19, 39, 64] and can predict future evolution of the cell state [8, 9, 33]. Currently, transcriptomics is one of the most advanced technologies we have available to study a full set of important biological features on a large number of single cells. In the future, it will be interesting to see if novel technologies can further enrich this data by adding complementary single-cell measurements of epigenomics, proteomics, and metabolomics.

The traditional view has been that we can define a cell type by a small set of characteristic features. Such a view is used in Chapter 2, where we looked at an example dataset which defines three distinct cell types (neural stem cells, basal progenitors and newborn neurons) based on high or low expression of two marker genes *Hes5* and *Tbr2* and their location in the ventricular zone of the developing cerebral cortex [42, 43]. These experimentally separated cell types are then characterised as a uniform group. From the high-dimensional gene expression measurements, we were able to then predict what other genes could be used as marker genes to characterise these three cell types. For this task, we developed a novel statistical model based on a conditional entropy measure that works even when the number of measurements per group is small and when more than two groups were compared. However, even from this bulk sequencing data, it was clear that gene expression is a dynamic process. For many genes, expression in samples of neural stem cells taken on embryonic day 10.5 was very different to that in samples taken on embryonic day 18.5. So it is clear that a

cell type identity is not simply a fixed point in gene expression space. Furthermore, single-cell RNA-seq data from these cells revealed transcriptomic heterogeneity even within cells on the same day [*data not published*]. This result may not actually be surprising, given that parts of the stem cell pool differentiates into basal progenitors while other parts proliferate to maintain the pool.

The question of how to define cell types have become particularly important as we try to fully characterise all cells in organisms, as is done in the Human Cell Atlas project [49]. As a result of observing heterogeneities within cells of the same type, their initial working definition of a cell type is “a region or a probability distribution either in the full-dimensional space or in a projection onto a lower-dimensional space that reflects salient features” [49]. As discussed in Section 3.2, this definition is exactly what is done by most established clustering tools for single-cell data in gene expression space, which are already being used to discover novel cell types [21, 46]. It should be noted that these clusters are only meant as a starting point from which to refine cell type definitions into “simpler molecular and morphological signatures”. However, this starting point already is poorly defined, as most established ways to generate the full-dimensional or lower-dimensional gene expression space are not an unbiased reflection of the biological information in scRNA-seq data. There is also no rigorous theory about what constitutes a sufficiently distinct cluster of cells in that space. We tried to address these issues in Chapter 3, where we argue how established models of noise in raw UMI counts from a scRNA-seq experiment can be used to mathematically define a cell state. This description of a cell state contains all the underlying biological information that can be inferred from measurements of the transcriptome. Furthermore, we show how to partition cells into groups that are statistically most likely to have the exact same cell state. This allows us to test in an unbiased manner how well we can describe various tissues as a small set of discrete cell types. We find that, based only on noisy measurements of the transcriptome, in some samples 50% of cells are in statistically distinct gene expression states. This supports the idea, voiced also by several researchers in the article cited at the start of this chapter [16], that in many systems there are no well-defined cell types, but rather a continuum of potential cell states. Another explanation would be that there is a large degree of natural variation in the gene expression that is not functionally important. In this scenario, every cell might be in its own unique gene expression state (by our narrow definition), but there is still a well-defined cell type that could be described by a distribution of cell states around a centre. Further research will be needed to better understand why we observe so many singlet states.

While our tool does provide a method for hierarchically combining the fundamental cell states into higher-order clusters that could correspond to cell types, there is no way to rigorously define how different groups of cells should be before they are considered distinct types. We need to perhaps come to terms with the idea that it is impossible to have *one* definition of what a cell type is. For every rule in biology, there will always be an exception. And just like words, cell type designations have usages and not meanings.

# Bibliography


[1] Charu C Aggarwal and Chandan K Reddy, editors. *DATA CLUSTERING Algorithms and Applications*. CRC Press, Boca Raton, FL, 2014. ISBN 9781466558212. URL http://www.taylorandfrancis.com.

[2] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11(10):R106, oct 2010. ISSN 14747596. doi: 10.1186/GB-2010-11-10-R106. URL /pmc/articles/PMC3218662//pmc/articles/PMC3218662/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3218662/.

[3] Tallulah S. Andrews and Martin Hemberg. Identifying cell populations with scRNASeq. *Molecular Aspects of Medicine*, 59:114–122, feb 2018. ISSN 18729452. doi: 10.1016/j.mam.2017.07.002. URL https://www.sciencedirect.com/science/article/pii/S0098299717300493.

[4] Piotr J. Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J. Gruber, Mihaela Zavolan, and Erik Van Nimwegen. ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs. *Genome research*, 24(5):869–884, 2014. ISSN 1549-5469. doi: 10.1101/GR.169508.113. URL https://pubmed.ncbi.nlm.nih.gov/24515121/.

[5] Maayan Baron, Adrian Veres, Samuel L. Wolock, Aubrey L. Faust, Renaud Gaujoux, Amedeo Vetere, Jennifer Hyoje Ryu, Bridget K. Wagner, Shai S. Shen-Orr, Allon M. Klein, Douglas A. Melton, and Itai Yanai. A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. *Cell systems*, 3(4):346–360.e4, oct 2016. ISSN 2405-4712. doi: 10.1016/J.CELS.2016.08.011. URL https://pubmed.ncbi.nlm.nih.gov/27667365/.

[6] Jacob Beal. Biochemical complexity drives log-normal variation in genetic expression. *Engineering Biology*, 1(1):55–60, jun 2017. ISSN 2398-6182. doi: 10.1049/ENB.2017.0004. URL https://onlinelibrary.wiley.com/doi/full/10.1049/enb.2017.0004https://onlinelibrary.wiley.com/doi/abs/10.1049/enb.2017.0004https://ietresearch.onlinelibrary.wiley.com/doi/10.1049/enb.2017.0004.

[7] Asa Ben-Hur, David Horn, Hava T Siegelmann, Vladimir Vapnik, Nello Critianini, John Shawe-Taylor, and Bob Williamson. Support vector clustering. *jmlr.org*, 2:125–137, 2001. URL https://www.jmlr.org/papers/volume2/horn01a/horn01a.pdf.

[8] Volker Bergen, Marius Lange, Stefan Peidli, F. Alexander Wolf, and Fabian J. Theis. Generalizing RNA velocity to transient cell states through dynamical mod-

eling. *Nature biotechnology*, 38(12):1408–1414, dec 2020. ISSN 1546-1696. doi: 10.1038/S41587-020-0591-3. URL https://pubmed.ncbi.nlm.nih.gov/32747759/.

[9] Volker Bergen, Ruslan A Soldatov, Peter V Kharchenko, and Fabian J Theis. RNA velocity—current challenges and future perspectives. *Molecular Systems Biology*, 17(8):e10282, aug 2021. ISSN 1744-4292. doi: 10.15252/MSB.202110282. URL https://onlinelibrary.wiley.com/doi/full/10.15252/msb.202110282https://onlinelibrary.wiley.com/doi/abs/10.15252/msb.202110282https://www.embopress.org/doi/abs/10.15252/msb.202110282.

[10] Mariia Bilous, Loc Tran, Chiara Cianciaruso, Santiago J. Carmona, Mikael J. Pittet, and David Gfeller. Super-cells untangle large and complex single-cell transcriptome networks. *bioRxiv*, page 2021.06.07.447430, jun 2021. doi: 10.1101/2021.06.07.447430. URL https://www.biorxiv.org/content/10.1101/2021.06.07.447430v1https://www.biorxiv.org/content/10.1101/2021.06.07.447430v1.abstract.

[11] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. Bayesian inference of gene expression states from single-cell RNA-seq data. *Nature Biotechnology*, 2021. ISSN 15461696. doi: 10.1038/s41587-021-00875-x.

[12] Yingying Cao, Simo Kitanovski, Ralf Küppers, and Daniel Hoffmann. UMI or not UMI, that is the question for scRNA-seq zero-inflation. *Nature Biotechnology 2021 39:2*, 39(2):158–159, feb 2021. ISSN 1546-1696. doi: 10.1038/s41587-020-00810-6. URL https://www.nature.com/articles/s41587-020-00810-6.

[13] Tara Chari and Lior Pachter. The specious art of single-cell genomics. *PLOS Computational Biology*, 19(8):1–20, 08 2023. doi: 10.1371/journal.pcbi.1011288. URL https://doi.org/10.1371/journal.pcbi.1011288.

[14] Renchao Chen, Xiaoji Wu, Lan Jiang, and Yi Zhang. Single-Cell RNA-Seq Reveals Hypothalamic Cell Diversity. *Cell reports*, 18(13):3227–3241, mar 2017. ISSN 2211-1247. doi: 10.1016/J.CELREP.2017.03.004. URL https://pubmed.ncbi.nlm.nih.gov/28355573/.

[15] Deanna M. Church, Leo Goodstadt, Ladeana W. Hillier, Michael C. Zody, Steve Goldstein, Xinwe She, Carol J. Bult, Richa Agarwala, Joshua L. Cherry, Michael DiCuccio, Wratko Hlavina, Yuri Kapustin, Peter Meric, Donna Maglott, Zoë Birtle, Ana C. Marques, Tina Graves, Shiguo Zhou, Brian Teague, Konstantinos Potamousis, Christopher Churas, Michael Place, Jill Herschleb, Ron Runnheim, Daniel Forrest, James Amos-Landgraf, David C. Schwartz, Ze Cheng, Kerstin Lindblad-Toh, Evan E. Eichler, Chris P. Ponting, Donna M. Muzny, Shannon Dugan-Rocha, Yan Ding, Steven E. Scherer, Christian J. Buhay, Andrew Cree, Judith Hernandez, Michael Holder, Jennifer Hume, Laronda R. Jackson, Christie Kovar, Sandra L. Lee, Lora R. Lewis, Michael L. Metzker, Lynne V. Nanareth, Aniko Sabo, Erica Sodergren, Richard A. Gibbs, Michael FitzGerald, April Cook, David B. Jaffe, Manuel Garber, Andrew R. Zimmer, Mono Pirun, Lyndsey Russell, Ted Sharpe, Michael Kamal Kabir Chaturvedi, Jane Wilkinson, Kurt LaButti, Xiaoping Yang, Daniel Bessette, Nicole R. Allen, Cindy Nguyen,

Thu Nguyen, Chelsea Dunbar, Rakela Lubonja, Charles Matthews, Xiaohong Liu, Mostafa Benamara, Tamrat Negash, Tashi Lokyitsang, Karin Decktor, Bruno Piqani, Glen Munson, Pema Tenzin, Sabrina Stone, Pendexter Macdonald, Harindra Arachchi, Amr Abouelleil, Annie Lui, Margaret Priest, Gary Gearin, Adam Brown, Lynne Aftuck, Terrance Shea, Sean Sykes, Aaron Berlin, Jeff Chu, Kathleen Dooley, Daniel Hagopian, Jennifer Hall, Nabil Hafez, Cherylyn L. Smith, Peter Olandt, Karen Miller, Vijay Ventkataraman, Anthony Rachupka, Lester Dorris, Laura Ayotte, Richard Mabbitt, Jeffrey Erickson, Andrea Horn, Peter An, Jerome W. Naylor, Sampath Settipalli, Eric S. Lander, Richard K. Wilson, Tina A. Graves, Robert S. Fulton, Susan M. Rock, Asif T. Chinwalla, Kelly Bernard, Laura P. Courtney, Catrina Fronick, Lucinda L. Fulton, Michelle O'Laughlin, Colin L. Kremitzki, Patrick J. Minx, Joanne O. Nelson, Kyriena L. Schatzkamer, Cynthia Strong, Aye M. Wollam, George M. Weinstock, Shiaw Pyng Yang, Jane Rogers, Darren Grafham, Sean Humphray, Christine Nicholson, Christine Bird, Andrew J. Brown, John Burton, Chris Clee, Adrienne Hunt, Matt C. Jones, Christine Lloyd, Lucy Matthews, Karen Mclaren, Stuart Mclaren, Kirsten McLay, Sophie A. Palmer, Robert Plumb, Ratna Shownkeen, Sarah Sims, Mike A. Quail, Siobhan L. Whitehead, David L. Willey, Stephane Deschamps, Steven Kenton, Lin Song, Trang Do, Bruce Roe, Gerard G. Bouffard, Robert W. Blakesley, Eric D. Green, Raju Kucherlapati, George Grills, Li Li, Kate T. Montgomery, Melissa Kramer, Lori Speigel, W. Richard McCombie, Susan Lucas, Astrid Terry, Laurie Gordon, Lisa Stubbs, Paul Denny, Steve D.M. Brown, Anne Marie Mallon, R. Duncan Campbell, Marc R.M. Botherby, Ian J. Jackson, Marc J. Rubenfield, Andrea M. Rogosin, and Douglas R. Smith. Lineage-specific biology revealed by a finished genome assembly of the mouse. *PLoS biology*, 7(5), may 2009. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1000112. URL https://pubmed.ncbi.nlm.nih.gov/19468303/https://pubmed.ncbi.nlm.nih.gov/19468303/?dopt=Abstract.

[16] Hans Clevers, Susanne Rafelski, Michael Elowitz, Allon Klein, Cole Trapnell, Jay Shendure, Ed Lein, Matthias Uhlen, Emma Lundberg, Alfonso Martinez-Arias, Joshua R Sanes, Paul Blainey, James Eberwine, Junhyong Kim, and Christopher J Love. What Is Your Conceptual Definition of "Cell Type" in the Context of a Mature Organism? *Cell Systems*, 4(3):255–259, mar 2017. ISSN 24054720. doi: 10.1016/j.cels.2017.03.006. URL https://www.sciencedirect.com/science/article/pii/S2405471217300911?via%3Dihub.

[17] Juliana Costa-Silva, Douglas Domingues, and Fabricio Martins Lopes. RNA-Seq differential expression analysis: An extended review and a software tool. *PLoS ONE*, 12 (12), 2017. ISSN 19326203. doi: 10.1371/journal.pone.0190152.

[18] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, 2018. ISSN 2046-1402. doi: 10.12688/f1000research.15666.2. URL http://www.ncbi.nlm.nih.gov/pubmed/30271584http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6134335.

[19] Mark W.E.J. Fiers, Liesbeth Minnoye, Sara Aibar, Carmen Bravo González-Blas,

Zeynep Kalender Atak, and Stein Aerts. Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, 17(4):246–254, jul 2018. ISSN 20412657. doi: 10.1093/BFGP/ELX046. URL https://academic.oup.com/bfg/article/17/4/246/4803107.

[20] Christopher T. Fincher, Omri Wurtzel, Thom de Hoog, Kellie M. Kravarik, and Peter W. Reddien. Cell type transcriptome atlas for the planarian Schmidtea mediterranea. *Science*, 360(6391), may 2018. doi: 10.1126/SCIENCE.AAQ1736.

[21] Dominic Grün, Anna Lyubimova, Lennart Kester, Kay Wiebrands, Onur Basak, Nobuo Sasaki, Hans Clevers, and Alexander van Oudenaarden. Single-cell messenger RNA sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, sep 2015. ISSN 0028-0836. doi: 10.1038/nature14966. URL http://www.nature.com/articles/nature14966.

[22] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Zimin Zhou, Haide Chen, Fang Ye, Daosheng Huang, Yang Xu, Wentao Huang, Mengmeng Jiang, Xinyi Jiang, Jie Mao, Yao Chen, Chenyu Lu, Jin Xie, Qun Fang, Yibin Wang, Rui Yue, Tiefeng Li, He Huang, Stuart H. Orkin, Guo Cheng Yuan, Ming Chen, and Guoji Guo. Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell*, 172(5):1091–1107.e17, feb 2018. ISSN 0092-8674. doi: 10.1016/J.CELL.2018.02.001.

[23] Kenneth D. Harris, Hannah Hochgerner, Nathan G. Skene, Lorenza Magno, Linda Katona, Carolina Bengtsson Gonzales, Peter Somogyi, Nicoletta Kessaris, Sten Linnarsson, and Jens Hjerling-Leffler. Classes and continua of hippocampal CA1 inhibitory neurons revealed by single-cell transcriptomics. *PLoS Biology*, 16(6), jun 2018. doi: 10.1371/JOURNAL.PBIO.2006387. URL /pmc/articles/PMC6029811//pmc/articles/PMC6029811/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6029811/.

[24] Josip S. Herman, Sagar, and Dominic Grün. FateID infers cell fate bias in multipotent progenitors from single-cell RNA-seq data. *Nature Methods*, 15(5):379–386, apr 2018. ISSN 15487105. doi: 10.1038/nmeth.4662.

[25] Radmila Hrdlickova, Masoud Toloue, and Bin Tian. RNA-Seq methods for transcriptome analysis. *Wiley Interdisciplinary Reviews: RNA*, 8(1):e1364, jan 2017. ISSN 1757-7012. doi: 10.1002/WRNA.1364. URL https://onlinelibrary.wiley.com/doi/full/10.1002/wrna.1364https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1364https://wires.onlinelibrary.wiley.com/doi/10.1002/wrna.1364.

[26] Saiful Islam, Amit Zeisel, Simon Joost, Gioele La Manno, Pawel Zajac, Maria Kasper, Peter Lönnerberg, and Sten Linnarsson. Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, 11(2):163–166, feb 2014. ISSN 15487091. doi: 10.1038/nmeth.2772. URL http://www.nature.com/articles/nmeth.2772.

[27] Eric M. Kernfeld, Ryan M.J. Genga, Kashfia Neherin, Margaret E. Magaletta, Ping Xu, and René Maehr. A Single-Cell Transcriptomic Atlas of Thymus Organogenesis Resolves Cell Types and Developmental Maturation. *Immunity*, 48(6):1258–1270.e6, jun 2018. ISSN 1097-4180. doi: 10.1016/J.IMMUNI.2018.04.015. URL https://pubmed.ncbi.nlm.nih.gov/29884461/.

[28] Peter V Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7):723–732, 2021. ISSN 15487105. doi: 10.1038/s41592-021-01171-x. URL https://doi.org/10.1038/s41592-021-01171-x.

[29] Vladimir Yu Kiselev, Kristina Kirschner, Michael T. Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N. Natarajan, Wolf Reik, Mauricio Barahona, Anthony R. Green, and Martin Hemberg. SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, 2017. ISSN 15487105. doi: 10.1038/nmeth.4236.

[30] Vladimir Yu Kiselev, Tallulah S. Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics*, 20 (5):273–282, may 2019. ISSN 1471-0056. doi: 10.1038/s41576-018-0088-9. URL http://www.nature.com/articles/s41576-018-0088-9.

[31] Dmitry Kobak and Philipp Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications 2019 10:1*, 10(1):1–14, nov 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-13056-x. URL https://www.nature.com/articles/s41467-019-13056-x.

[32] Gioele La Manno, Daniel Gyllborg, Simone Codeluppi, Kaneyasu Nishimura, Carmen Salto, Amit Zeisel, Lars E. Borm, Simon R.W. Stott, Enrique M. Toledo, J. Carlos Villaescusa, Peter Lönnerberg, Jesper Ryge, Roger A. Barker, Ernest Arenas, and Sten Linnarsson. Molecular Diversity of Midbrain Development in Mouse, Human, and Stem Cells. *Cell*, 167(2):566–580.e19, oct 2016. ISSN 0092-8674. doi: 10.1016/J.CELL.2016.09.027.

[33] Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E. Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Peter V. Kharchenko. RNA velocity of single cells. *Nature*, 560(7719):494–498, aug 2018. ISSN 1476-4687. doi: 10.1038/S41586-018-0414-6. URL https://pubmed.ncbi.nlm.nih.gov/30089906/.

[34] Jan Lause, Philipp Berens, and Dmitry Kobak. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biology 2021 22:1*, 22(1):1–20, sep 2021. ISSN 1474-760X. doi: 10.1186/S13059-021-02451-7. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02451-7.

[35] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods 2018 15:12*, 15

(12):1053–1058, nov 2018. ISSN 1548-7105. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

[36] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861, 2018. doi: 10.21105/joss.00861. URL https://doi.org/10.21105/joss.00861.

[37] Vilas Menon. Clustering single cells: a review of approaches on high-and low-depth single-cell RNA-seq data. *Briefings in Functional Genomics*, 17 (4):240–245, jul 2017. ISSN 2041-2649. doi: 10.1093/bfgp/elx044. URL https://academic.oup.com/bfg/article/17/4/240/4728639http://academic.oup.com/bfg/advance-article/doi/10.1093/bfgp/elx044/4728639.

[38] Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J. McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau, Stéphane C. Boutet, Chad Sanada, Aik Ooi, Robert C. Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Caroline Braeuning, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, Lan T. Nguyen, Aviv Regev, Joshua Z. Levin, Swati Parekh, Aleksandar Janjic, Lucas E. Wange, Johannes W. Bagnoli, Wolfgang Enard, Marta Gut, Rickard Sandberg, Itoshi Nikaido, Ivo Gut, Oliver Stegle, and Holger Heyn. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature biotechnology*, 38(6):747–755, jun 2020. ISSN 1546-1696. doi: 10.1038/S41587-020-0469-4. URL https://pubmed.ncbi.nlm.nih.gov/32518403/.

[39] Thomas Moerman, Sara Aibar Santos, Carmen Bravo González-Blas, Jaak Simm, Yves Moreau, Jan Aerts, and Stein Aerts. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. *Bioinformatics*, 35(12):2159–2161, jun 2019. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTY916. URL https://academic.oup.com/bioinformatics/article/35/12/2159/5184284.

[40] Samantha A Morris. The evolving concept of cell identity in the single cell era. *Development (Cambridge, England)*, 146(12):dev169748, jun 2019. ISSN 1477-9129. doi: 10.1242/dev.169748. URL http://www.ncbi.nlm.nih.gov/pubmed/31249002.

[41] Eran A Mukamel and John Ngai. Perspectives on defining cell types in the brain. *Current Opinion in Neurobiology*, 56:61–68, jun 2019. ISSN 0959-4388. doi: 10.1016/J.CONB.2018.11.007. URL https://www.sciencedirect.com/science/article/pii/S0959438818301818.

[42] Tanzila Mukhtar and Verdon Taylor. Untangling Cortical Complexity During Development. *Journal of experimental neuroscience*, 12:1179069518759332, mar 2018. ISSN 1179-0695. doi: 10.1177/1179069518759332. URL http://www.ncbi.nlm.nih.gov/pubmed/29551911http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5846925.

[43] Tanzila Mukhtar, Jeremie Breda, Alice Grison, Zahra Karimaddini, Pascal Grobecker, Dagmar Iber, Christian Beisel, Erik van Nimwegen, and Verdon Taylor. Tead transcription factors differentially regulate cortical development. *Scientific Reports 2020 10:1*, 10(1):1–19, mar 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-61490-5. URL https://www.nature.com/articles/s41598-020-61490-5.

[44] Mauro J. Muraro, Gitanjali Dharmadhikari, Dominic Grün, Nathalie Groen, Tim Dielen, Erik Jansen, Leon van Gurp, Marten A. Engelse, Francoise Carlotti, Eelco J.P. de Koning, and Alexander van Oudenaarden. A Single-Cell Transcriptome Atlas of the Human Pancreas. *Cell Systems*, 3(4):385–394.e3, 2016. ISSN 24054720. doi: 10.1016/j.cels.2016.09.002.

[45] Jihwan Park, Rojesh Shrestha, Chengxiang Qiu, Ayano Kondo, Shizheng Huang, Max Werth, Mingyao Li, Jonathan Barasch, and Katalin Suszták. Single-cell transcriptomics of the mouse kidney reveals potential cellular targets of kidney disease. *Science (New York, N.Y.)*, 360(6390):758–763, may 2018. ISSN 1095-9203. doi: 10.1126/SCIENCE.AAR2131. URL https://pubmed.ncbi.nlm.nih.gov/29622724/.

[46] Lihong Peng, Xiongfei Tian, Geng Tian, Junlin Xu, Xin Huang, Yanbin Weng, Jialiang Yang, and Liqian Zhou. Single-cell RNA-seq clustering: datasets, models, and algorithms. *https://doi.org/10.1080/15476286.2020.1728961*, 17(6):765–783, jun 2020. ISSN 15558584. doi: 10.1080/15476286.2020.1728961. URL https://www.tandfonline.com/doi/abs/10.1080/15476286.2020.1728961.

[47] Rob Phillips, Jane Kondev, Julie Theriot, Hernan G. Garcia, and Nigel Orme. *Physical Biology of the Cell*. Garland Science, Boca Raton, oct 2012. ISBN 9780429168833. doi: 10.1201/9781134111589.

[48] Mireya Plass, Jordi Solana, F. Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J. Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391), may 2018. doi: 10.1126/SCIENCE.AAQ1723. URL http://dx.doi.

[49] Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C. Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P. Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N. Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W. Shin, Oliver Stegle, Michael Stratton, Michael J.T. T Stubbington, Fabian J. Theis, Matthias Uhlen, Alexander Van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, Dana Pe'er, Anthony Phillipakis, Chris P. Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen,

Joshua Sanes, Rahul Satija, Ton N. Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W. Shin, Oliver Stegle, Michael Stratton, Michael J.T. T Stubbington, Fabian J. Theis, Matthias Uhlen, Alexander Van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, and Nir Yosef. The human cell atlas. *eLife*, 6:1–30[, dec 2017. ISSN 2050084X. doi: 10.7554/eLife.27041. URL https://elifesciences.org/articles/27041.

[50] Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47, jan 2015. ISSN 13624962. doi: 10.1093/NAR/GKV007. URL /pmc/articles/PMC4402510//pmc/articles/PMC4402510/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC4402510/.

[51] Mark D. Robinson, Davis J. McCarthy, and Gordon K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139, nov 2010. ISSN 14602059. doi: 10.1093/BIOINFORMATICS/BTP616. URL /pmc/articles/PMC2796818//pmc/articles/PMC2796818/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC2796818/.

[52] Alexander B. Rosenberg, Charles M. Roco, Richard A. Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T. Graybuck, David J. Peeler, Sumit Mukherjee, Wei Chen, Suzie H. Pun, Drew L. Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, apr 2018. ISSN 10959203. doi: 10.1126/SCIENCE.AAM8999/SUPPL_FILE/PAPV2.PDF. URL https://www.science.org/doi/abs/10.1126/science.aam8999.

[53] Andrew Rosenberg and Julia Hirschberg. V-Measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL 2007 - Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[54] Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nature Genetics 2021 53:6*, 53(6):770–777, may 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00873-4. URL https://www.nature.com/articles/s41588-021-00873-4.

[55] Nicholas Schaum, Jim Karkanias, Norma F. Neff, Andrew P. May, Stephen R. Quake, Tony Wyss-Coray, Spyros Darmanis, Joshua Batson, Olga Botvinnik, Michelle B. Chen, Steven Chen, Foad Green, Robert C. Jones, Ashley Maynard, Lolita Penland, Angela Oliveira Pisco, Rene V. Sit, Geoffrey M. Stanley, James T. Webber, Fabio Zanini, Ankit S. Baghel, Isaac Bakerman, Ishita Bansal, Daniela Berdnik, Biter Bilen, Douglas Brownfield, Corey Cain, Min Cho, Giana Cirolia, Stephanie D. Conley, Aaron Demers, Kubilay Demir, Antoine de Morree, Tessa Divita, Haley du Bois, Laughing

Bear Torrez Dulgeroff, Hamid Ebadi, F. Hernán Espinoza, Matt Fish, Qiang Gan, Benson M. George, Astrid Gillich, Geraldine Genetiano, Xueying Gu, Gunsagar S. Gulati, Yan Hang, Shayan Hosseinzadeh, Albin Huang, Tal Iram, Taichi Isobe, Feather Ives, Kevin S. Kao, Guruswamy Karnam, Aaron M. Kershner, Bernhard M. Kiss, William Kong, Maya E. Kumar, Jonathan Y. Lam, Davis P. Lee, Song E. Lee, Guang Li, Qingyun Li, Ling Liu, Annie Lo, Wan Jin Lu, Anoop Manjunath, Kaia L. May, Oliver L. May, Marina McKay, Ross J. Metzger, Marco Mignardi, Dullei Min, Ahmad N. Nabhan, Katharine M. Ng, Joseph Noh, Rasika Patkar, Weng Chuan Peng, Robert Puccinelli, Eric J. Rulifson, Shaheen S. Sikandar, Rahul Sinha, Krzysztof Szade, Weilun Tan, Cristina Tato, Krissie Tellez, Kyle J. Travaglini, Carolina Tropini, Lucas Waldburger, Linda J. van Weele, Michael N. Wosczyna, Jinyi Xiang, Soso Xue, Justin Youngyunpipatkul, Macy E. Zardeneta, Fan Zhang, Lu Zhou, Paola Castro, Derek Croote, Joseph L. DeRisi, Christin S. Kuo, Benoit Lehallier, Patricia K. Nguyen, Serena Y. Tan, Bruce M. Wang, Hanadie Yousef, Philip A. Beachy, Charles K.F. Chan, Kerwyn Casey Huang, Kenneth Weinberg, Sean M. Wu, Ben A. Barres, Michael F. Clarke, Seung K. Kim, Mark A. Krasnow, Roel Nusse, Thomas A. Rando, Justin Sonnenburg, and Irving L. Weissman. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. *Nature*, 562(7727), 2018. ISSN 14764687. doi: 10.1038/s41586-018-0590-4.

[56] Ron Sender, Shai Fuchs, and Ron Milo. Revised Estimates for the Number of Human and Bacteria Cells in the Body. *PLoS biology*, 14(8), aug 2016. ISSN 1545-7885. doi: 10.1371/JOURNAL.PBIO.1002533. URL https://pubmed.ncbi.nlm.nih.gov/27541692/ https://pubmed.ncbi.nlm.nih.gov/27541692/?dopt=Abstract.

[57] Charlotte Soneson and Mauro Delorenzi. A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, 14, 2013. ISSN 14712105. doi: 10.1186/1471-2105-14-91.

[58] Zhe Sun, Ting Wang, Ke Deng, Xiao Feng Wang, Robert Lafyatis, Ying Ding, Ming Hu, and Wei Chen. DIMM-SC: A Dirichlet mixture model for clustering droplet-based single cell transcriptomic data. *Bioinformatics*, 34(1):139–146, jan 2018. ISSN 14602059. doi: 10.1093/bioinformatics/btx490. URL https://academic.oup.com/bioinformatics/article/34/1/139/4060554.

[59] Valentine Svensson. Droplet scRNA-seq is not zero-inflated. *Nature biotechnology*, 38(2):147–150, feb 2020. ISSN 1546-1696. doi: 10.1038/S41587-019-0379-5. URL https://pubmed.ncbi.nlm.nih.gov/31937974/.

[60] Valentine Svensson, Kedar Nath Natarajan, Lam Ha Ly, Ricardo J. Miragaia, Charlotte Labalette, Iain C. Macaulay, Ana Cvejic, and Sarah A. Teichmann. Power analysis of single-cell RNA-sequencing experiments. *Nature Methods 2017 14:4*, 14(4):381–387, mar 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4220. URL https://www.nature.com/articles/nmeth.4220.

[61] F. William Townes, Stephanie C. Hicks, Martin J. Aryee, and Rafael A. Irizarry. Feature selection and dimension reduction for single-cell RNA-Seq based on a

multinomial model. *Genome biology*, 20(1):295, dec 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1861-6. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1861-6http://www.ncbi.nlm.nih.gov/pubmed/31870412http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6927135.

[62] D. Tsafrir, I. Tsafrir, L. Ein-Dor, O. Zuk, D. A. Notterman, and E. Domany. Sorting points into neighborhoods (SPIN): data analysis and visualization by ordering distance matrices. *Bioinformatics (Oxford, England)*, 21(10):2301–2308, may 2005. ISSN 1367-4803. doi: 10.1093/BIOINFORMATICS/BTI329. URL https://pubmed.ncbi.nlm.nih.gov/15722375/.

[63] Maria Tsagiopoulou, Maria Christina Maniou, Nikolaos Pechlivanis, Anastasis Togkousidis, Michaela Kotrová, Tobias Hutzenlaub, Ilias Kappas, Anastasia Chatzidimitriou, and Fotis Psomopoulos. UMIc: A Preprocessing Method for UMI Deduplication and Reads Correction. *Frontiers in Genetics*, 12, may 2021. ISSN 16648021. doi: 10.3389/FGENE.2021.660366. URL /pmc/articles/PMC8193862//pmc/articles/PMC8193862/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC8193862/.

[64] Turki Turki and Y. h. Taguchi. SCGRNs: Novel supervised inference of single-cell gene regulatory networks of complex diseases. *Computers in Biology and Medicine*, 118:103656, mar 2020. ISSN 0010-4825. doi: 10.1016/J.COMPBIOMED.2020.103656.

[65] Catalina A. Vallejos, Davide Risso, Antonio Scialdone, Sandrine Dudoit, and John C. Marioni. Normalizing single-cell RNA sequencing data: challenges and opportunities. *Nature Methods 2017 14:6*, 14(6):565–571, may 2017. ISSN 1548-7105. doi: 10.1038/nmeth.4292. URL https://www.nature.com/articles/nmeth.4292.

[66] Laurens Van Der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 2008. ISSN 15324435.

[67] David van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J. Carr, Cassandra Burdziak, Kevin R. Moon, Christine L. Chaffer, Diwakar Pattabiraman, Brian Bierie, Linas Mazutis, Guy Wolf, Smita Krishnaswamy, and Dana Pe'er. Recovering Gene Interactions from Single-Cell Data Using Data Diffusion. *Cell*, 174(3), 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.05.061.

[68] Erik van Nimwegen, Mihaela Zavolan, Nikolaus Rajewsky, and Eric D Siggia. Probabilistic clustering of sequences: Inferring new bacterial regulons by comparative genomics. *Proceedings of the National Academy of Sciences*, 99(11):7323–7328, may 2002. ISSN 0027-8424. doi: 10.1073/pnas.112690399. URL http://www.ncbi.nlm.nih.gov/pubmed/12032281http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC124229.

[69] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, jan 2009. ISSN 14710056. doi: 10.1038/nrg2484.

[70] David I. Warton. Why you cannot transform your way out of trouble for small counts. *Biometrics*, 74(1):362–368, mar 2018. ISSN 1541-0420. doi: 10.1111/BIOM.12728. URL https://onlinelibrary.wiley.com/doi/full/10.1111/biom.12728https://onlinelibrary.wiley.com/doi/abs/10.1111/biom.12728https://onlinelibrary.wiley.com/doi/10.1111/biom.12728.

[71] Jochen Winterer, David Lukacsovich, Lin Que, Andrea M. Sartori, Wenshu Luo, and Csaba Földy. Single-cell RNA-Seq characterization of anatomically identified OLM interneurons in different transgenic mouse lines. *The European Journal of Neuroscience*, 50(11):3750, dec 2019. doi: 10.1111/EJN.14549. URL /pmc/articles/PMC6973274//pmc/articles/PMC6973274/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6973274/.

[72] Bo Xia and Itai Yanai. A periodic table of cell types. *Development (Cambridge, England)*, 146(12):dev169854, jun 2019. ISSN 14779129. doi: 10.1242/dev.169854. URL http://www.ncbi.nlm.nih.gov/pubmed/31249003http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC6602355.

[73] Yurong Xin, Giselle Dominguez Gutierrez, Haruka Okamoto, Jinrang Kim, Ann Hwee Lee, Christina Adler, Min Ni, George D. Yancopoulos, Andrew J. Murphy, and Jesper Gromada. Pseudotime ordering of single human B-cells reveals states of insulin production and unfolded protein response. *Diabetes*, 67(9):1783–1794, sep 2018. ISSN 1939327X. doi: 10.2337/DB18-0365/-/DC1. URL https://diabetes.diabetesjournals.org/content/67/9/1783https://diabetes.diabetesjournals.org/content/67/9/1783.abstract.

[74] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, 2015. ISSN 14602059. doi: 10.1093/bioinformatics/btv088.

[75] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, mar 2015. ISSN 10959203. doi: 10.1126/science.aaa1934. URL http://www.ncbi.nlm.nih.gov/pubmed/25700174.

[76] Amit Zeisel, Hannah Hochgerner, Peter Lönnerberg, Anna Johnsson, Fatima Memic, Job van der Zwan, Martin Häring, Emelie Braun, Lars E. Borm, Gioele La Manno, Simone Codeluppi, Alessandro Furlan, Kawai Lee, Nathan Skene, Kenneth D. Harris, Jens Hjerling-Leffler, Ernest Arenas, Patrik Ernfors, Ulrika Marklund, and Sten Linnarsson. Molecular Architecture of the Mouse Nervous System. *Cell*, 174(4):999–1014.e22, aug 2018. ISSN 10974172. doi: 10.1016/j.cell.2018.06.021. URL https://www.sciencedirect.com/science/article/pii/S009286741830789X?via%3Dihub.

[77] Grace X.Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie

Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature communications*, 8, jan 2017. ISSN 2041-1723. doi: 10.1038/NCOMMS14049. URL https://pubmed.ncbi.nlm.nih.gov/28091601/.

# A

# Appendix: Identifying cell states in single-cell RNA-seq data at maximal resolution

## A.1 Summary of datasets used

Table A.1: Summary of all scRNA-seq datasets used in this thesis.

| dataset name | species | tissue | tissue exact | Technique |
|---|---|---|---|---|
| baron | human | Pancreas | Pancreatic islets | InDrops |
| chen | mouse | Brain | Hypothalamus | Drop-seq |
| hca_10X2x5Kcell250Kreads | human | Blood | Periphal blood monocytes | Chromium |
| hca_CELseq2 | human | Blood | Periphal blood monocytes | CEL-seq2 |
| hca_Dropseq | human | Blood | Periphal blood monocytes | Drop-seq |
| hca_MARSseq | human | Blood | Periphal blood monocytes | MARS-seq |
| hca_QUARTZseq | human | Blood | Periphal blood monocytes | Quartz-seq2 |
| hca_SCRBseq | human | Blood | Periphal blood monocytes | gmcSCRB-seq |
| hca_ddSEQ | human | Blood | Periphal blood monocytes | ddSeq |
| hca_inDrop | human | Blood | Periphal blood monocytes | InDrops |
| kernfeld | mouse | Thymus | Thymus P0 | Drop-seq |
| lamanno_embryo | human | Brain | Embryo ventral midbrain | STRT-seq (C1) |
| lamanno_mouseembryo | mouse | Brain | Embryo ventral midbrain | STRT-seq (C1) |
| microwellseq_Bladder | mouse | Bladder | Bladder | Microwell-seq |
| microwellseq_Kidney | mouse | Kidney | Kidney | Microwell-seq |
| microwellseq_Lung | mouse | Lung | Lung | Microwell-seq |
| microwellseq_Mammary | mouse | Mammary | Mammary | Microwell-seq |
| microwellseq_Spleen | mouse | Spleen | Spleen | Microwell-seq |
| microwellseq_Thymus | mouse | Thymus | Thymus | Microwell-seq |
| muraro | human | Pancreas | Pancreatic islets | SORT-seq |
| park | mouse | Kidney | Kidney | Chromium |
| tabula_muris_Bladder | mouse | Bladder | Bladder | Chromium |
| tabula_muris_Kidney | mouse | Kidney | Kidney | Chromium |
| tabula_muris_Lung | mouse | Lung | Lung | Chromium |
| tabula_muris_Mammary | mouse | Mammary | Mammary | Chromium |
| tabula_muris_Spleen | mouse | Spleen | Spleen | Chromium |
| tabula_muris_Thymus | mouse | Thymus | Thymus | Chromium |
| xin | human | Pancreas | Pancreatic islets | Chromium |
| zeisel | mouse | Brain | Somatosensory cortex, CA1 hippocampus | STRT-seq (C1) |
| zhengmix4eq | human | Blood | Periphal blood monocytes | Chromium |
| zhengmix4uneq | human | Blood | Periphal blood monocytes | Chromium |
| zhengmix8eq | human | Blood | Periphal blood monocytes | Chromium |
| zhengmix8uneq | human | Blood | Periphal blood monocytes | Chromium |
| zhengmix10eq | human | Blood | Periphal blood monocytes | Chromium |

Table A.1: Summary of all scRNA-seq datasets used in this thesis. (cont.) References for the datasets are: Baron et al. [5], Chen et al. [14], Han et al. [22], Kernfeld et al. [27], La Manno et al. [32], Mereu et al. [38], Muraro et al. [44], Park et al. [45], Schaum et al. [55], Xin et al. [73], Zeisel et al. [75], Zheng et al. [77]

| dataset name | NCBI GEO Series / Link | Reference | Comments |
|---|---|---|---|
| baron | GSE84133 | Baron et al., 2016 | Only sample 1 (GSM2230757) |
| chen | GSE87544 | Chen et al., 2017 | Only human data |
| hca_10X2x5Kcell250Kreads | GSE133535 | Mereu et al., 2020 | Only human data |
| hca_CELseq2 | GSE133539 | Mereu et al., 2020 | Only human data |
| hca_Dropseq | GSE133540 | Mereu et al., 2020 | Only human data |
| hca_MARSseq | GSE133542 | Mereu et al., 2020 | Only human data |
| hca_QUARTZseq | GSE133543 | Mereu et al., 2020 | Only human data |
| hca_SCRBseq | GSE133544 | Mereu et al., 2020 | Only human data |
| hca_ddSEQ | GSE133547 | Mereu et al., 2020 | Only human data |
| hca_inDrop | GSE133548 | Mereu et al., 2020 | Only human data |
| kernfeld | GSE107910 | Kernfeld et al., 2018 | Only P0 samples (GSM2883201, GSM2883202) |
| lamanno_embryo | GSE76381 | La Manno et al., 2016 | GSE76381_EmbryoMoleculeCounts.cef.txt.gz |
| lamanno_mouseembryo | GSE76382 | La Manno et al., 2016 | GSE76381_MouseEmbryoMoleculeCounts.cef.txt.gz |
| microwellseq_Bladder | GSE108097 | Han et al., 2018 | GSM2889480 |
| microwellseq_Kidney | GSE108098 | Han et al., 2018 | GSM2906425, GSM2906426 |
| microwellseq_Lung | GSE108099 | Han et al., 2018 | GSM2906429-GSM2906431 |
| microwellseq_Mammary | GSE108100 | Han et al., 2018 | GSM2906439-GSM2906442 |
| microwellseq_Spleen | GSE108101 | Han et al., 2018 | GSM2906471 |
| microwellseq_Thymus | GSE108102 | Han et al., 2018 | GSM2906475, GSM2906476 |
| muraro | GSE85241 | Muraro et al., 2016 | |
| park | GSE107585 | Park et al., 2018 | Only 4 WT samples (GSM2871706-GSM2871709) |
| tabula_muris_Bladder | GSE109774 | Schaum et al., 2018 | GSE109774_Bladder.tar.gz |
| tabula_muris_Kidney | GSE109775 | Schaum et al., 2018 | GSE109774_Kidney.tar.gz |
| tabula_muris_Lung | GSE109776 | Schaum et al., 2018 | GSE109774_Lung.tar.gz |
| tabula_muris_Mammary | GSE109777 | Schaum et al., 2018 | GSE109774_Mammary.tar.gz |
| tabula_muris_Spleen | GSE109778 | Schaum et al., 2018 | GSE109774_Spleen.tar.gz |
| tabula_muris_Thymus | GSE109779 | Schaum et al., 2018 | GSE109774_Thymus.tar.gz |
| xin | GSE114297 | Xin et al., 2018 | |
| zeisel | GSE60361 | Zeisel et al., 2015 | |
| zhengmix4eq | https://support.10xgenomics.com/single-cell-gene-expression/datasets | Zheng et al., 2017 | Constructed datasets used for benchmarking; see separate table |
| zhengmix4uneq | https://support.10xgenomics.com/single-cell-gene-expression/datasets | Zheng et al., 2017 | Constructed datasets used for benchmarking; see separate table |
| zhengmix8eq | https://support.10xgenomics.com/single-cell-gene-expression/datasets | Zheng et al., 2017 | Constructed datasets used for benchmarking; see separate table |
| zhengmix8uneq | https://support.10xgenomics.com/single-cell-gene-expression/datasets | Zheng et al., 2017 | Constructed datasets used for benchmarking; see separate table |
| zhengmix10eq | https://support.10xgenomics.com/single-cell-gene-expression/datasets | Zheng et al., 2017 | Constructed datasets used for benchmarking; see separate table |

Table A.1: Summary of all scRNA-seq datasets used in this thesis. (cont.)

| dataset name | Basis for Simulated dataset | Reproducibility of partitions | Benchmarking of Clustering tools | Cellstate diversity |
|---|---|---|---|---|
| baron | Yes | Yes | Yes | Yes |
| chen | No | Yes | Yes | Yes |
| hca_10X2x5Kcell250Kreads | No | Yes | Yes | Yes |
| hca_CELseq2 | Yes | Yes | No | Yes |
| hca_Dropseq | Yes | Yes | No | Yes |
| hca_MARSseq | No | Yes | No | Yes |
| hca_QUARTZseq | Yes | Yes | No | Yes |
| hca_SCRBseq | No | Yes | No | Yes |
| hca_ddSEQ | Yes | Yes | No | Yes |
| hca_inDrop | No | Yes | No | Yes |
| kernfeld | Yes | Yes | No | Yes |
| lamanno_embryo | Yes | Yes | No | Yes |
| lamanno_mouseembryo | Yes | Yes | No | Yes |
| microwellseq_Bladder | Yes | Yes | No | Yes |
| microwellseq_Kidney | No | Yes | No | Yes |
| microwellseq_Lung | No | Yes | No | Yes |
| microwellseq_Mammary | No | Yes | No | Yes |
| microwellseq_Spleen | No | Yes | No | Yes |
| microwellseq_Thymus | No | Yes | No | Yes |
| muraro | Yes | Yes | No | Yes |
| park | No | Yes | No | Yes |
| tabula_muris_Bladder | Yes | Yes | No | Yes |
| tabula_muris_Kidney | Yes | Yes | No | Yes |
| tabula_muris_Lung | Yes | Yes | No | Yes |
| tabula_muris_Mammary | Yes | Yes | No | Yes |
| tabula_muris_Spleen | No | Yes | No | Yes |
| tabula_muris_Thymus | Yes | Yes | No | Yes |
| xin | No | Yes | No | Yes |
| zeisel | Yes | Yes | Yes | Yes |
| zhengmix4eq | Yes | Yes | Yes | No |
| zhengmix4uneq | No | Yes | Yes | No |
| zhengmix8eq | No | Yes | Yes | No |
| zhengmix8uneq | No | Yes | Yes | No |
| zhengmix10eq | No | Yes | Yes | No |

Table A.2: Composition of Zhengmix datasets. The original data [77] consists of 10 pure cell type populations of different kinds of peripheral blood monocytes, which are the rows of this table. Each Zhengmix dataset (columns) consists of a different composition of randomly chosen cells from these pure populations. The entries in the table indicate the number of cells of each cell type in each dataset.

| dataset name | zhengmix_4eq | zhengmix_4uneq | zhengmix_8eq | zhengmix_8uneq | zhengmix_10eq |
|---|---|---|---|---|---|
| CD14+ Monocytes | 1000 | 500 | 1000 | 2600 | 1000 |
| CD19+ B Cells | 1000 | 1000 | 1000 | 1700 | 1000 |
| CD34+ Cells | 0 | 0 | 0 | 30 | 1000 |
| CD4+ Helper T Cells | 0 | 0 | 1000 | 0 | 1000 |
| CD4+/CD25+ Regulatory T Cells | 1000 | 3000 | 1000 | 200 | 1000 |
| CD4+/CD45RA+/CD25- Naive T cells | 0 | 0 | 1000 | 2300 | 1000 |
| CD4+/CD45RO+ Memory T Cells | 0 | 0 | 1000 | 1700 | 1000 |
| CD56+ Natural Killer Cells | 0 | 0 | 1000 | 1500 | 1000 |
| CD8+ Cytotoxic T cells | 0 | 0 | 0 | 0 | 1000 |
| CD8+/CD45RA+ Naive Cytotoxic T Cells | 1000 | 2000 | 1000 | 1700 | 1000 |
| **Total** | **4000** | **6500** | **8000** | **11730** | **10000** |

## A.2 Summary of selected published clustering tools

Table A.3: Summary of selected published clustering algorithms. Mandatory model parameters refers only to those which have no default value; usually there are many that *can* be chosen. Preprocessing refers to steps either done by the tool or that should be done by the user before applying the algorithm.

| **Name & Reference** | **Language** | **Method** | **Mandatory Model Parameters** | **Preprocessing** |
|---|---|---|---|---|
| BackSPIN [75] | Python | Biclustering | Clustering depth $d$: maximum number of clusters is $2^d$ | feature selection, normalization |
| DIMMSC [58] | R | Dirichlet mixture model | $K$: number of clusters | |
| RaceID3 [21, 24] | R | K-medoids clustering for main clusters; afterwards refinement for detection of rare cell types | | feature selection, normalization, dimensionality reduction |
| SC3 [29] | R | combining multiple clustering solutions through a consensus approach | $K$: number of clusters (can give a range) | feature selection, normalization, dimensionality reduction |
| SNN-Cliq [74] | MATLAB/ Python | Based on shared nearest neighbours | | feature selection, normalization |

# Acknowledgements

First, I want to thank Erik for supervising my work of the last four years. You have taught me a lot about how to critically question even established knowledge, about the value of mathematical rigour, and about the persistence to solve challenging problems. Next, I would like to thank Mihaela and Wolfgang for their interest in my work and agreeing to review this thesis. A particular thanks to Mikhail, Thomas S and Jérémie for being the first ones to try out *cellstates* and giving me lots of useful feedback. A huge thank you also to Dorde, Athos, Anne, Arantxa, Luca, Gwendoline, Théo, Dany, Björn, Daan and all the other former and present members of the van Nimwegen lab. You all have made my PhD special, and I will cherish many precious memories from my time with you as colleagues and friends. Thank you as well to the NeurostemX team for the interesting collaboration.

A huge thank you to my parents, Christine, Elena, Ma and Baba. Without you as my family, I would not have come this far so easily. Your continuous emotional support, guidance and advice has helped me more than I can express.

Finally, I want to acknowledge the most important person of all: my wonderful wife Mimi. Your everlasting, unconditional love and support kept me going through the toughest times. Thank you for helping me with proofreading and practising presentations. Thank you for making our home my favourite office. Thank you for your infinite patience, for always having good advice, for being my biggest cheerleader and for never failing to make me smile.

# Curriculum Vitae

# PASCAL GROBECKER

pascal.grobecker@gmail.com • +44 74 75 73 91 47 • London, UK
linkedin.com/in/pascal-grobecker • github.com/pgrobecker

## Experience

### ILLUMINA

Senior Bioinformatics Scientist, Medical Genomics Research
**Since Dec 2022 | Cambridge, UK**

Showing what is possible with next-generation sequencing in oncology. In particular, working on improving diagnosis of cancer patients with whole genome and transcriptome sequencing.

### BIOZENTRUM & SWISS INSTITUTE OF BIOINFORMATICS

Doctoral Researcher in in Genome Systems Biology Group
**Sep 2017 – Jan 2022 | Basel, Switzerland**

#### CELLSTATES – AN INNOVATIVE ML TOOL FOR SINGLE-CELL TRANSCRIPTOMICS DATA

- Derived a rigorous clustering algorithm for sparse, noisy and high-dimensional single-cell RNA-seq data, based on a Bayesian statistical model with minimal assumptions.
- Tests on datasets from public databases show relevant biological signals are automatically found. Inferred clusters are less biased than those from other state-of-the-art machine learning tools.
- Developed a data analysis tool in Python. Published open-source code on GitHub (pgrobecker/cellstates) that is used by researchers for their experimental data. Regularly obtained feedback from users in order to improve usability, add helpful features, and improve relevance of visualisations.
- Optimised performance of code and used parallel programming in Cython to reach 40-fold speed-up. It is run in a high-performance computing (HPC) environment on data tables of dimensions over 20,000 x 20,000.
- Communicated the algorithm and results at conferences and seminars to diverse audiences from industry and academia.

#### NEUROSTEMX – SCIENTIFIC COLLABORATION

- Project in multi-disciplinary scientific team across 6 research groups. Worked closely together with experimental biologists to analyse their laboratory data.
- Developed a novel statistical model for finding salient features in gene expression data using an entropy based scoring function.

#### TEACHING ASSISTANT FOR QUANTITATIVE DATA ANALYSIS IN BIOLOGY

- Supervised practical exercises on statistical analysis of biological data in Python for classes of ~20 Master's students.

### DEPARTMENT OF PHYSICS, UNIVERSITY OF CAMBRIDGE

Master Thesis
**Oct 2016 – May 2017 | Cambridge, UK**

- Modelled the plant meristem as a cell-to-cell contact network. Developed an innovative approach to infer the system's spatial structure from this network alone, but showed that predicting time evolution required a more complex scientific model.
- Collaborated with laboratory at University of Birmingham to obtain real-world datasets from plants.

### CALIFORNIA INSTITUTE OF TECHNOLOGY

Summer Undergraduate Research Fellow

**Jun 2016 – Sep 2016 | Pasadena, USA**

- Studied impact of different drugs on the cardiovascular system.
- Applied wave intensity analysis in MATLAB to experimental time-series data to evaluate how changes in blood fluid dynamics affect the heart workload.

## Education

### UNIVERSITY OF BASEL

PhD in Computational Biology

**Sep 2017 – Jan 2022 | Basel, Switzerland**

- **Result: Magna Cum Laude** (5.5 / 6.0)
- Academic courses: Computational Systems Biology, Bioinformatics Algorithms (including data structures and neural networks), Transcription Regulation in Eukaryotes
- Transferable skills courses: Project Management, Research Ethics and Integrity

### UNIVERSITY OF CAMBRIDGE

Bachelor (BA) and Master (MSci) in Natural Sciences (Physics)

**Oct 2013 – Jun 2017 | Cambridge, UK**

- **Result: Class I (GPA 4.0)** in both Bachelor and Master
- Relevant courses: Biology of Cells (first-year undergraduate level), Advanced Mathematics, Scientific Computing, Biological Physics, Mathematical Biology of the Cell
- Cambridge University Physics Society: active participation in committee for three years, including one year as co-chair.
- Volunteering: Assisted in running Cambridge Hands-on Science and Physics at Work Open Days to communicate science to the general public, particularly children and adolescents.

## Skills

### PROGRAMMING

**Proficient:** Python (NumPy, pandas, Cython, scikit-learn, matplotlib, Jupyter Notebook)
**Experienced:** Linux / Unix shell scripting • Slurm • Git
**Familiar:** MATLAB • R • SQL • C++ • Excel

### LANGUAGES

**Fluent:** English, German, French
**Basic:** Bengali

## Publications

Mukhtar, Tanzila & Breda, Jeremie & Grison, Alice & Karimaddini, Zahra & **Grobecker, Pascal** & Iber, Dagmar & Beisel, Christian & Nimwegen, Erik & Taylor, Verdon. (2020). Tead transcription factors differentially regulate cortical development. Scientific Reports. 10. 4625. 10.1038/s41598-020-61490-5.

Mukhtar, Tanzila* & Breda, Jeremie* & Adam, Manal* & Boareto, Marcelo* & **Grobecker, Pascal*** & Karimaddini, Zahra* & Grison, Alice* & Eschbach, Katja & Chandrasekhar, Ramakrishnan & Vermeul, Swen & Okoniewski, Michal & Pachkov, Mikhail & Harwell, Corey & Atanasoski, Suzana & Beisel, Christian & Iber, Dagmar & Nimwegen, Erik & Taylor, Verdon. (2022). Temporal and sequential transcriptional dynamics define lineage shifts in corticogenesis. The EMBO journal. 41. e111132. 10.15252/embj.2022111132. *These authors contributed equally to this work.

**Grobecker, Pascal** & Nimwegen, Erik. (2023). Identifying cell states in single-cell RNA-seq data at statistically maximal resolution. 10.1101/2023.10.31.564980. (Preprint)