UNIVERSITY OF MODENA AND REGGIO EMILIA

PhD in Molecular and Regenerative Medicine (XXXV Cycle)

Reconstruction of Dynamic Trajectories from the Differentiation Potency of Single Cells

Candidate: Luca Calderoni

Tutor: Prof. Mattia Forcato

Co-Tutor: Dr. Oriana Romano

PhD Coordinator: Prof. Michele De Luca

Index of contents

Abstract

Since the introduction of single cell RNA sequencing (scRNA-seq), numerous computational methods have been developed to infer the progression of single cells along “developmental” paths from changes in their transcriptional programs and to model the differentiation processes of adult cells from their pluripotent progenitors. Most of these methods reconstruct dynamic cellular processes from static transcriptional profiles by ordering cells on continuous trajectories based on their similarity in the gene expression space. This strategy relies on assumptions that may not be a-priori guaranteed in every cellular system and requires some prior knowledge on the topology and direction of the expected genealogy. Since scRNA-seq technology is being increasingly used to study complex and unexplored systems characterized by still unclear developmental patterns, new approaches are required that are able to infer cell lineages directly from the dynamic evolution of the differentiation potency of single cells without the need of prior assumptions. Although several methods have been devised to estimate the potency of single cells by measuring the entropy level of their transcriptomes, a clear mathematical model for the inference of the dynamics of such potency across time is still lacking. To tackle this issue, we developed FIERCE (Framework for InfERence of veloCity of the Entropy), a novel composite computational pipeline that employs the mathematical framework of RNA velocity to predict the temporal evolution of the signaling entropy of single cells during dynamic processes. As such, FIERCE allows inferring cell lineages through a fully unsupervised and cell-centered approach that does not need the prior specification of evolutionary parameters. When tested on scRNA-seq data from three well-known mouse differentiation systems, our method correctly reconstructed the developmental genealogy of the adult specialized cell subpopulations from their respective pluripotent progenitors. We envisage that our tool will be a valuable computational resource for the inference of cell trajectories in the absence of solid prior biological knowledge.

1. Introduction

1.1. Reconstructing cell phylogenies from sequencing data: a new perspective from statistical mechanics

1.1.1. The single cell RNA-seq revolution

As Frederick Sanger invented his new sequencing technology in 1977, the scientific community gained the unprecedented opportunity to read the genetic information of living organisms at single nucleotide detail. Since then, a cascade of new methodological advances progressively increased the power and the resolution of this invaluable technology and made it possible to achieve foremost goals such as the publication of the first complete human genome in 2003 (https://www.genome.gov/10001772/all-about-the–human-genome-project-hgp/). However, the earliest sequencing technologies were very expensive and labor-intensive, and this prevented them from being commonly employed by the scientific community in individual research projects[1].

In this respect, the introduction of high throughput Next Generation Sequencing (NGS) technologies in 2005 represented a ground-breaking milestone that allowed scientists to fully integrate efficient and standardized sequencing pipelines into their routine experiments, and to obtain the complete genomic, transcriptomic, or epigenomic landscapes of entire cell populations at an affordable expense of time and money[1]. In particular, RNA sequencing (RNA-seq) technologies allow analyzing the total transcriptomic information of entire biological samples, and consequently unveiling the changes that such information undergoes under different physiological and developmental conditions.

Such huge amount of information provided by “bulk” RNA-seq keeps fueling discovery and innovation in biomedical research and continues to be an invaluable resource to address relevant biological questions. However, this technology inevitably suffers from a remarkable downside: it is designed to provide the average transcriptional state of the whole cell population, given by the sum of the individual states of all the single cells. Consequently, it fails to capture the true complexity of the biological system, represented by the transcriptional diversity of the various subpopulations it is composed of. This hampers the discovery of subtler transcriptional changes that arise in different physiological contexts, as well as the reconstruction of the evolution of the transcriptional landscape of tissues and organs during development. This crucial hindrance was overcome in 2009 with the invention of the first single cell RNA sequencing (scRNA-seq) technologies[2], that were specifically designed to provide the complete transcriptional portrait of each individual cell in a biological sample.

Unlike traditional bulk RNA-seq, this new technology can assess the gene expression landscape of a tissue at the single cell resolution, and thus can unveil the finest transcriptional changes during relevant biological processes. The thorough reconstruction of the gene expression states of all individual cells is indeed fundamental to gain a detailed insight on a wide range of complex biological processes, such as the differential transcription of key genes under particular conditions, the complex network of regulatory interactions between transcription factors and their targets, the

equally complex communication exchanges between different cells within the same tissue, and even the dynamic process that governs the developmental evolution of entire organs[1,3].

The possibility to obtain tens or even hundreds of thousands of transcriptional profiles with a single experiment is the pivotal innovation brought by scRNA-seq, and the main reason that elevated this method as the leading technology for the study of the gene expression dynamics of a wide variety of biological systems.

1.1.2. Reinterpreting cell dynamics in the light of statistical mechanics

Single cell sequencing brought about a profound and unprecedented change of perspective, shifting the focus of biological studies from whole cell populations to single cells. Such radical change requires the development of new theoretic principles for a correct approach to the huge and complex amount of data provided by these technologies, and above all for its correct interpretation.

To extract meaningful biological information from sequencing data it is necessary to perform a bioinformatic analysis, that mainly consists of fitting mathematical models that can identify clear and interpretable patterns from raw nucleotide sequences. Although the models used for the analysis of bulk NGS data have reached remarkable levels of complexity, the advent of single cell sequencing has made them rapidly obsolete. New computational methods are necessary that can describe cell populations not as a single macroscopic entity, but rather as the sum of thousands of complex microscopic components that interact with each other, and above all change in time. In other words, new mathematical frameworks are needed to describe the large-scale dynamics of macroscopic biological systems, like tissues and organs, in terms of the small-scale dynamics of their microscopic components, i.e., the single cells[4].

Statistical mechanics is a discipline that is exactly inspired by such principle[5,6]. Born in the nineteenth century, it is based on the assumption that, if we can describe the behavior of each single microscopic component of a complex dynamic system through appropriate metrics, then we can ultimately both explain the current structure of the whole system and predict its future states. For instance, if we know the distribution of the velocities of the single molecules of a gas, we can both explain its current features, such as pressure and temperature, and predict how these features will change in the near future[4]. As this core principle can be easily applied to a wide variety of complex systems, several diverse scientific disciplines (from chemistry to physics, from ecology to population genetics) have benefited from the concepts of statistical mechanics[4].

In the field of molecular and cell biology, the availability of single cell technologies made it possible to extend the scope of statistical mechanics to the bioinformatic analysis of high-throughput sequencing data. If we can describe the behavior of single cells with appropriate metrics that summarize their genetic, transcriptional, or epigenetic content, we can sum up all this small-scale information into statistical models that can efficiently describe the large-scale structure of whole tissues and organs, as well as predict their future development[4]. This reductionist approach is not applicable with bulk sequencing and represents the most innovative and promising contribution that single cell sequencing technologies can bring to our understanding of the structure and function of biological systems.

In the case of scRNA-seq, the new interpretative lens offered by the principles of statistical mechanics finds its most straightforward application in what is commonly known as “trajectory

inference", i.e., the reconstruction of the temporal dynamics of very complex processes, like development and differentiation sequences, or even disease progression, exclusively from the gene expression data of single cells[7–9]. As cells progress through a dynamic process of any kind, they need to communicate with each other at precise time points, and, most importantly, they need to change their morphology and even their function; this can be only achieved through the expression of specific proteins at specific times, and consequently through well-orchestrated transcriptional changes. This implies that all the macroscopic changes that can be observed as biological processes unfold can be interpreted as emerging properties of all the concerted dynamic changes in the transcriptomes of single cells, that represent the microscopic components of the evolving system.

Although the field of trajectory reconstruction from scRNA-seq data is currently very florid[9], the development of computational methods inspired by the principles of statistical mechanics is still in its infancy. Several attempts have been made to elaborate appropriate metrics to describe the dynamic properties of cell transcriptomes, but a solid mathematical framework that describes how such properties evolve in time has not been proposed yet[4]. Consequently, most current computational methods rely on an opposite holistic approach that is not suited to fully exploit the huge potential brought by scRNA-seq data.

The following paragraphs provide a brief introduction to the basic experimental workflow of scRNA-seq experiments, as well as an introduction to the standard computational analysis of resulting transcriptional data. The aim is to present the appropriate context for explaining the solutions adopted by current trajectory reconstruction tools, as well as for advancing our proposal of a novel approach that takes direct inspiration from the concepts of statistical mechanics.

1.2. The single cell RNA-seq technologies

1.2.1. The basic workflow of scRNA-seq technologies

To date, several different scRNAseq platforms have been developed by different producers[1,3,10]. Although based on different core technologies, they all share a common basic workflow (**Figure 1**). First, cells are isolated from the tissue and lysed, then their mRNA transcripts are extracted and retro-transcribed, and finally the resulting cDNA molecules are amplified and sequenced by an NGS platform[1,3,10]. scRNA-seq technologies mainly differ in the specific methods used for cell isolation, mRNA retro-transcription, and cDNA amplification.

Cell isolation is a particularly delicate passage since cells must be separated from the extracellular matrix and from their neighbors without causing artificial gene expression changes[1]. Micropipetting, micromanipulation and laser capture microdissection are the most precise methods, but they are very labor-intensive and low throughput[3]. Fluorescence-activated cell sorting (FACS) allows quickly isolating a much higher number of labelled cells, but suffers from several downsides, such as the need for specific antibodies and for a large volume of input material[3]. The best techniques for isolating a large quantity of cells in a short amount of time are based on microfluidic devices[3]; in particular, droplet-based methods are usually preferred when the objective is to sequence as many cells as possible, albeit with a relatively low coverage[3,10,11]. They are designed to encapsulate individual cells into micro-droplets that contain all the necessary reagents for the subsequent steps of cell lysis, mRNA retro-transcription and cDNA amplification[11].

Retro-transcription is necessary due to the instability of RNA molecules[3], and is usually performed using oligo-dT primers to capture exclusively polyadenylated transcripts. The following amplification of resulting cDNA molecules is performed either through polymerase chain reaction (PCR), or, alternatively, through *in vitro* transcription (IVT) followed by a final passage of retro-transcription[1,3,10]. Depending on the specific procedures used for these passages, the amplified cDNA molecules that will be sequenced by the NGS platform can be either full length transcripts, or just the 5' or 3' ends.

Full length methods are highly sensitive, and particularly fit for studying splicing variants and differential allelic expression, though they bear the downside of a reduced throughput. Tag-based methods, on the contrary, are specifically designed to maximize the number of sequenced transcripts from each cell by sequencing just their 3' or 5' ends, at the cost of a reduced sensitivity[1,10]. Due to their very high throughput, they are particularly suited for exploratory experiments whose main goal is the description of the complete transcriptional diversity of a biological system, the determination of its cell type composition, or the reconstruction of its developmental history.

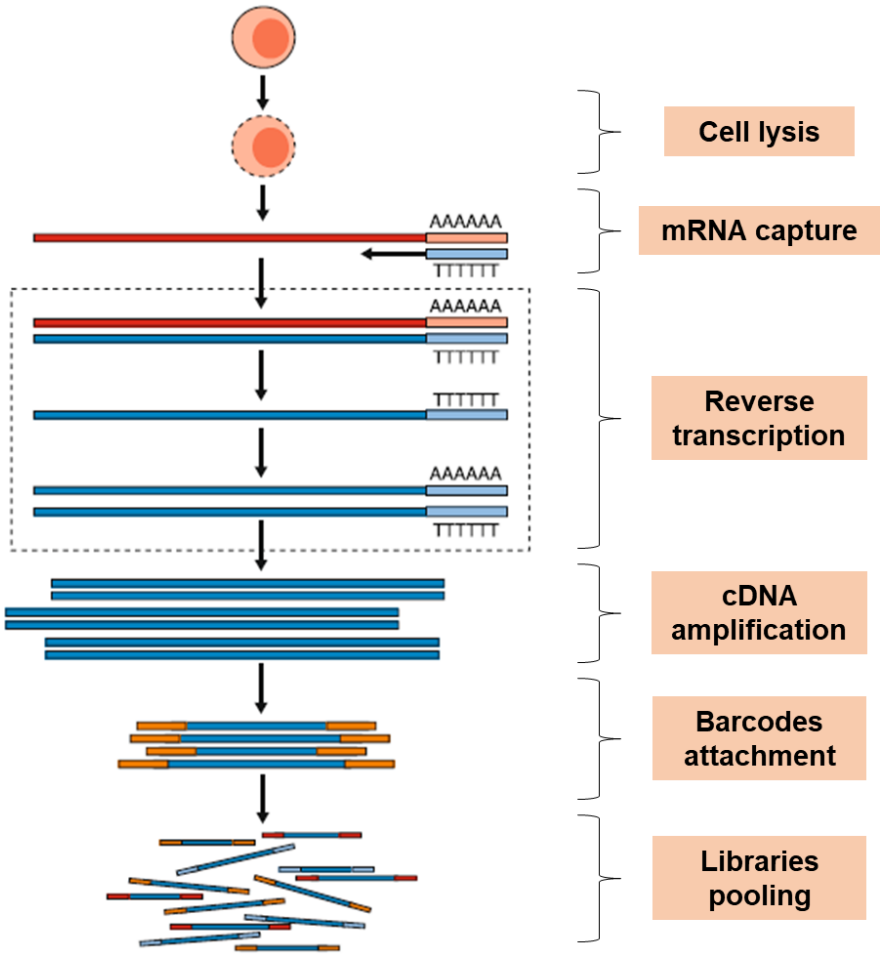

Figure 1: diagram showing the common workflow shared by all scRNA-seq technologies. Adapted from Haque et al 2017[10].

1.2.2. The 10X Genomics Chromium technology

Amongst tag-based droplet technologies, 10X Genomics Chromium is one of the most widely used due to its versatility. It is capable of sequencing up to eight different samples in a single run and provides tailored tools for all the steps of the pipeline, as well as tailored software for the first steps of the bioinformatic analysis of the sequencing data[10,11].

The core of 10X technology is Gel bead in EMulsion[11] (GEM): upon an eight-channel microfluidic chip (**Figure 2**), hundreds of thousands of single cells are individually encapsulated into GEMs with approximately 50% capture efficiency. Each gel bead is functionalized with several different types of oligonucleotides as sequencing adapters and primers, oligo-dTs to specifically prime polyadenylated mRNA, and barcodes to index both individual cells and individual transcripts within each cell.

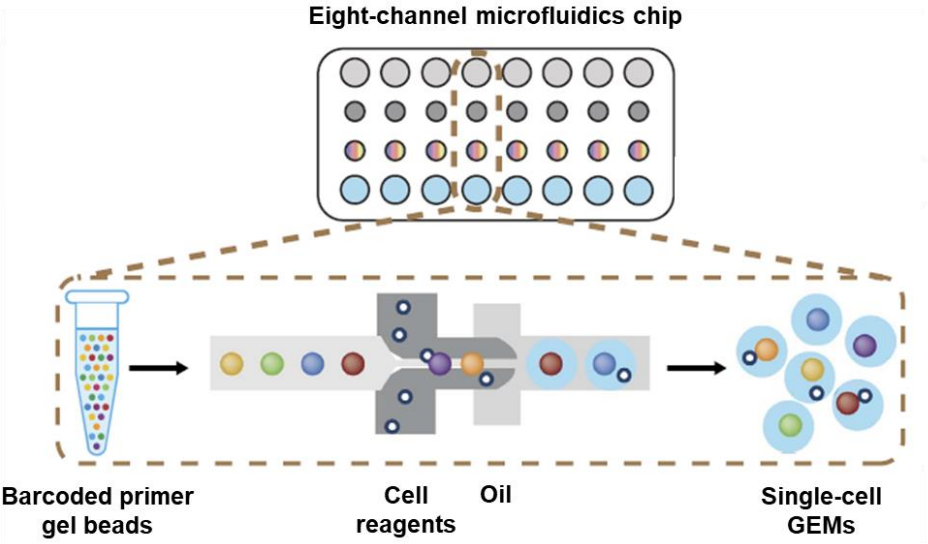

Figure 2: scheme showing the working principle of the 10X microfluidic chip. Adapted from Zheng et al 2017[11].

Cell barcodes are short nucleotide sequences that are unique for each gel bead and thus for each captured cell; they allow sorting the sequences according to their respective cell of origin once all the cells from all the samples are pooled together for the sequencing step. Transcript barcodes, known as Unique Molecular Identifiers (UMIs), are short nucleotide sequences that uniquely tag each single transcript within each GEM, and thus all its copies generated during the PCR step. They are designed to avoid the bias caused by PCR, that preferentially amplifies the most abundant transcripts, thus leading to their overestimation.

The amplified cDNA molecules are sequenced by an Illumina NGS platform, whose output are raw base call (BCL) files that contain all the sequencing reads from all the cells. These files are converted into fastq format and processed by Cellranger[11], a software that has been specifically designed to operate on 10X data. Cellranger automatically aligns the reads to the reference genome with the STAR[12] algorithm, and then counts the number of unique UMI sequences aligned to each gene. All the reads sharing the same UMI tag are considered as a single count, thus the true gene abundances are correctly estimated even in presence of a very skewed expression signal. The output of Cellranger is a gene per cell matrix containing the UMI counts of all genes in all cells; this matrix can

be processed by dedicated bioinformatic pipelines to extract valuable biological information, such as the number of cell types present in the sample, their respective abundance, and their genealogical relationships.

1.3. The computational analysis of scRNA-seq data

As with traditional bulk RNA-seq, the computational analysis of scRNA-seq data starts from a matrix of transcript counts of each gene in each sample[13], the key difference being that such samples represent individual cells rather than entire cell populations. Precisely, the samples of the matrix are the barcodes associated to the GEMs inside which the single cells have been encapsulated[11]. Given this crucial difference, scRNA-seq data need to be analyzed with dedicated pipelines, that are specifically designed to achieve two challenging aims, i.e., to remove or compensate the sources of noise present in the data, and to parse the huge amount of information to extract significant biological insight (**Figure 3**).

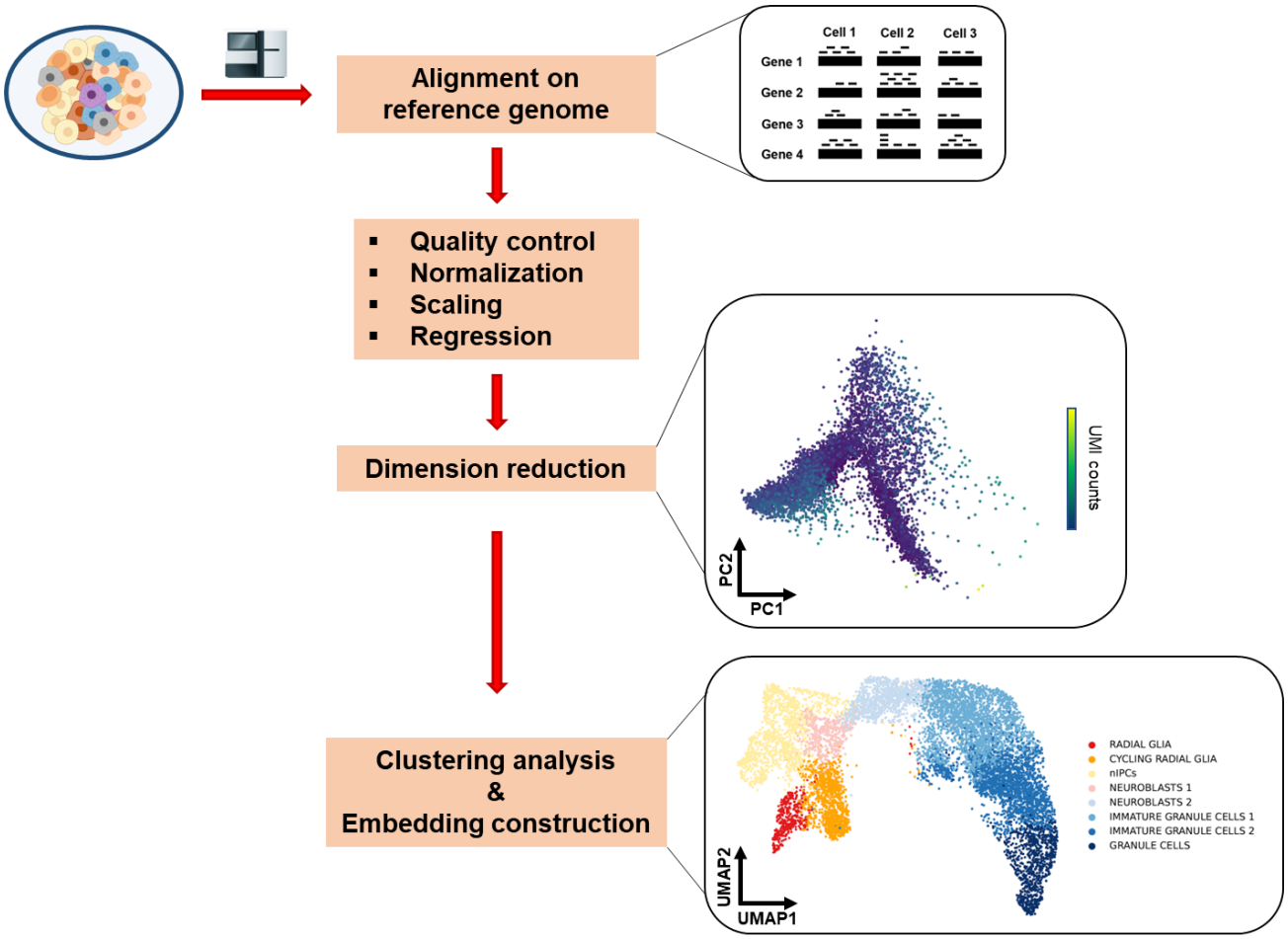

Figure 3: diagram showing the typical workflow of a standard bioinformatic analysis of scRNA-seq data.

Several different pipelines have been developed for scRNA-seq data analysis, but two software packages are largely used in common practice[13], namely Seurat[14–17] and Scanpy[18], working in R and python environments, respectively. They both include tailored functions for all the passages of a standard computational analysis, and, most importantly, they both offer practical plotting tools for the visualization of results. The choice of the package to use primarily depends on the preferred programming environment or on the specific software tools that will be used for more advanced analyses, since most tools are designed to exclusively operate either in R, or in python.

1.3.1. Quality control and noise removal

Due to the small amount of RNA present in a single cell, and to the low capture efficiency of current technologies, scRNA-seq data are unavoidably affected by a considerable amount of technical noise[13]. A prominent percentage of the gene expression differences between cells are indeed random fluctuations caused by differential transcript capture, that can even result in less expressed genes not being sampled merely by chance, a phenomenon that is commonly known as "gene dropout"[19]. In some cases, the expression signal associated to some barcodes may not even come from a viable cell, but instead from free environmental RNA or from broken dead cells[13,19]. Moreover, droplet-based technologies are known to be affected by the "doublets" problem, that occurs when two cells are erroneously encapsulated into a single droplet, and thus assigned the same barcode[13,19]. All these sources of technical noise can be either removed by filtering out low quality barcodes or compensated through appropriate mathematical procedures that are commonly applied as the first steps of a typical bioinformatic analysis.

The aim of quality filters is to increase the fraction of true positives, i.e., true viable cells that have been individually encapsulated inside a droplet and sequenced. They are based on several quality indices that are computed for each barcode, such as the total number of sequenced UMIs, the total number of expressed genes, or the percentage of UMIs mapping to mitochondrial genes (i.e., the mitochondrial fraction)[13,19,20]. A very low number of UMIs and of expressed genes, often coupled with a very high mitochondrial fraction, is indicative of a low-quality barcode, that most likely represents a dead or broken cell retaining just the mitochondrial transcripts[13,19]. On the contrary, a very high number of UMIs and of expressed genes is most likely indicative of two cells sequenced together, and thus of a doublet[13,19].

Quality filters are very useful for polishing the dataset, however, if considered singularly, they may also be misleading. For instance, a very high mitochondrial fraction may be involved in respiratory processes, a low expression level may represent quiescent cells rather than dead cells, and a very high expression level may indicate cells that are simply larger in size[13]. To avoid the exclusion of valuable information, and to maximize the probability to filter out true low-quality barcodes, it is good practice to examine all the quality metrics together, since they tend to correlate with each other. The best strategy is to choose thresholds by visualizing cells on several scatter plots showing every possible combination of metrics, since low quality cells tend to present outlying values for multiple quality metrics simultaneously, and thus to form separate clusters that can be easily identified on a low dimensional space.

1.3.2. Data normalization

After the dataset has been polished and low-quality barcodes have been removed, an important source of noise still remains as, due to low capture efficiency, different cells may have been sequenced at different depths simply by sheer chance[13,19,21]. To compensate this unavoidable bias, and to make the UMI counts of different cells truly comparable, a normalization procedure on library

size is recommended. The default procedure of both Seurat and Scanpy involves the division of the UMI counts of each cell by the total library size, and the multiplication of the resulting ratio by a scale factor that can be either a fixed number (usually 10^4 or 10^6) or a cell-wise quantity (such as the median of the counts)[13,15,18]. Additionally, normalized counts can be log-transformed to mitigate their skewness and their mean-variance relationship, and to allow measuring the gene expression changes as log-fold changes[13]; in this case, a pseudo-count (usually equal to 1) is added to all the expression values to avoid the logarithm of 0.

Optionally, a normalization procedure can be also applied to genes, that are scaled to have zero mean and unit variance[13,19]; the application of this additional normalization depends on whether genes should be equally weighted for the downstream analysis, or whether, on the contrary, the differential magnitude of their expression is of interest and should thus be preserved[13].

1.3.3. Regression on biological variables

Normalization alone does not guarantee the complete removal of the effect of differential transcript capture; depending on the dataset, a further correction may be necessary through a regression procedure on cell-wise measures of sequencing depth, such as the total UMI counts or the total number of expressed genes, to limit the influence of such technical noise on downstream clustering analysis[13,20]. Cells might in fact form separate clusters simply on the basis of their sequencing depth, thus masking important biological effects.

Regression is usually performed on all covariates at once, to account for their interdependence, so additional covariates can be added to the model to remove the effect of additional factors[13]. Such factors can be related to biological processes whose effect is not of interest for the researcher, but that might strongly influence the downstream analysis and hide the effect of more interesting biological signals[21]. The most prominent example is represented by cell cycle[13,21], that can have a dramatic impact on both the clustering analysis and the reconstruction of the development of the tissue. The cell cycle phase of cells is represented in their transcriptome by the oscillatory expression of known genes, so several algorithms have been developed to compute the cell cycle score of each cell based on the expression level of such genes[13]; the resulting scores can be subsequently added to the regression model to remove their effect from the data.

1.3.4. Integration and batch correction

Besides technical and biological factors, cells tend to form separate groups also based on the sample of origin. This is due to a systematic discrepancy in the gene expression data of different samples, that may strongly influence the results of all downstream analyses[13]. This phenomenon is called “batch effect” and can be corrected through dedicated algorithms; these are based either on linear or non-linear approaches, depending on the type of correction they are designed to perform. Linear approaches are more suited in the case of different batches within the same experimental setting, while non-linear approaches are preferable in the case of samples coming from different experiments. The first type of correction is simply called “batch correction”, while the second is commonly known as “data integration”[13]. ComBat[22,23] is an example of an algorithm for batch correction, based on a linear model where the batch factor is considered both in the mean and the variance of the data. The most popular algorithm for data integration is implemented in Seurat and is based on the projection of a transformation calculated on one of the samples, chosen as reference, into the other samples, with the aim of finding a set of mutual nearest neighbors on such reduced dimension space that will serve as “anchors” to transfer discrete or continuous data[16]. The result of

such procedure is a new corrected gene expression matrix where the effect of the different experimental conditions has been removed.

1.3.5. Dimension reduction

Regression models and batch correction methods are very efficient in removing unwanted sources of variation, and thus in enhancing the effect of crucial biological factors. However, even in datasets that have undergone all pre-processing procedures, such biological signal is hidden within a huge excess of unused information, given by the enormous number of variables that are measured in each cell. In a typical scRNA-seq experiment on human samples, the expression level of more than 20,000 genes is measured, but most of these genes are poorly expressed, or are expressed at a very similar level in most cells, and thus are completely uninformative. Typically, during the initial filtering step, only the genes that are expressed in at least a certain number of cells are retained, but even after such skimming procedure, the filtered matrix, in most cases, still includes from 10,000 to 15,000 genes[13]. Detecting meaningful biological signal from a dataset of this size is very difficult, as well as very impractical; therefore, before performing downstream analyses, it is good practice to drastically reduce the size of the expression matrix, to obtain a much smaller set of variables that summarizes only the most relevant information of the original dataset.

A simple but efficient approach is to “cut” the expression matrix to keep only the top variable genes, i.e., the genes that show the highest expression variability across cells, and thus best represent the overall transcriptional variance of the biological system[13]. The method implemented in both Seurat and Scanpy first divides the genes into bins according to their mean expression, and then, within each bin, selects the genes with the highest variance-to-mean ratio[16,18]. This operation allows reducing the total number of genes to just the few thousands that carry the most useful information.

Another very efficient strategy is to compute a new, very small set of variables that recapitulate, in a “compressed” form, the whole transcriptional landscape of the original population. This “dimension reduction” procedure involves the encoding of all the major sources of variance of the data into just a few components that are sufficient to efficiently describe the transcriptional profile of each cell[13]. It can be applied either to the complete gene expression matrix, or to the sub-matrix containing just the top variable genes. The most used techniques for dimension reduction are Principal Component Analysis (PCA) and diffusion maps. PCA[24] is a mathematical method that involves the computation of a small set of components through a linear combination of the original variables; such components are completely uncorrelated to one another and explain a progressively lower percentage of the total variance of the data[8,24]. Diffusion maps[25] are a non-linear approach that captures the probability of each cell to diffuse in the gene expression space and to transition to its neighbors, thus they are particularly suited for describing the transcriptional dynamics of the tissue[8,25].

Regardless of the specific method employed, just a few significant components are typically sufficient to explain most of the biologically relevant information[19]. In particular, the first two principal or diffusion components define a bidimensional space on which cells can be embedded to catch a first glimpse of the structure of the data. This is very useful for checking whether the cells form separate clusters because of undesired factors, such as the differential sequencing depth or the cell cycle score, and thus for checking the efficacy of the quality filters and regression procedures. Moreover, the visualization of cells on the reduced dimension space allows the detection of eventual batch effects, that can be subsequently removed with dedicated algorithms.

1.3.6. Clustering analysis

Albeit much smaller than the original dataset, a reduced dimension space, defined for instance by the principal or the diffusion components, detains all the most determinant transcriptional information in a computationally accessible format. Thus, it lays the foundation for more advanced types of algorithms used to extract interpretable insight on the cellular composition or on the dynamic evolution of the original tissue.

Clustering algorithms are designed to reconstruct the subpopulation structure of biological samples by subdividing the total cell population into transcriptionally homogeneous clusters based on cell-cell distances on the reduced dimension space[13,26]. Under the assumption that, in this space, cells belonging to the same cell type are located closer to each other than cells belonging to different cell types, the resulting clusters can be interpreted as representations of the cytological components of the sample.

Different clustering algorithms employ different strategies for finding the set of clusters that best reflect the real biological heterogeneity of the tissue. A classic example of distance-based method is k-means clustering[13,27,28], that involves the subdivision of cells into k clusters by assigning them to the nearest cluster centroid on the reduced dimension space; the positions of the centroids are iteratively optimized until the distances of the cells from their respective nearest centroids are minimized. Another very popular approach is the Louvain algorithm[29], a graph-based method implemented in both Seurat and Scanpy. The Louvain algorithm is based on community detection: first, a k-nearest neighbor graph, that connects each cell to its k nearest neighbors according to a defined distance measure (usually Euclidean distance), is built on the reduced dimension space; then, the graph is subdivided into communities, i.e., densely connected regions that include more cell-cell links than expected on the basis of the total number of links across the whole dataset.

Regardless of the specific method, clustering algorithms have the downside of being totally unsupervised: they are designed to yield an optimal grouping of cells based solely on their transcriptional distances, but this does not imply that the resulting clusters have a true biological relevance. To ensure that the cell clusters effectively coincide with true cell types, they must be annotated based on the expression level of known marker genes. The annotation procedure is very important for ascertaining that the tissue has been properly sampled, and, eventually, for discovering rare or previously unknown cell populations.

Cluster annotation can be performed either automatically or manually. Automatic annotation is performed by algorithms that compare the expression level of cell type markers in each cluster with the expression level of the same genes in a reference annotated dataset, with the aim of transferring the cell type labels from the reference to the query[13]. One of the most popular tools for automatic annotation is singleR[30], that includes several reference annotated datasets generated through bulk RNA-seq, for both human and mouse. Automatic annotation is particularly useful for exploratory studies where the purpose is to unravel the composition of less known tissues. However, results from the automatic annotation can be difficult to interpret or even misleading if the biological system includes transient cell populations that cannot be assigned to a defined cell type, such as those found in developing systems, or in altered conditions caused by a pathology.

On the contrary, manual annotation is best suited for samples extracted from well-known tissues. Clusters are annotated by the researcher through the direct visualization of the expression level of reliable cell type markers in each cluster, thus some previous knowledge on the expected cell types

is required. This procedure is more time-consuming compared to automatic annotation, but the direct application of the expertise of the researcher can greatly help in clarifying the identity of rare or ambiguous cell subpopulations. However, to be feasible, this approach requires the visualization of cell clusters in a bidimensional space that faithfully depicts the transcriptional topology of the population. Since the low dimensional space defined solely by the first two principal or diffusion components is not representative of the whole variance of the data, a further dimensional reduction is recommended.

1.3.7. Data visualization

The projection, or “embedding”, of cells onto a bidimensional space enables an easy visualization and interpretation of the structure and heterogeneity of the biological sample. Although convenient, this operation poses a challenge, since all the major sources of transcriptional variation within the cell population must be efficiently summarized in just two dimensions. This requires a further reduction of the multidimensional space defined by the principal or diffusion components. Several computational methods have been developed for this purpose, primarily based on non-linear approaches.

One of the most popular methods is the t-distributed Stochastic Neighbor Embedding (t-SNE)[31]. The t-SNE algorithm is designed to emphasize the local similarity of cells, thus allowing a clear visualization of each distinct cluster, but this comes at the expense of the global structure of the dataset. Indeed, t-SNE is not suited to reconstruct the topological relationships between the clusters, nor the dynamic evolution of the population; it is most useful when the aim is to simply visualize the variety of clusters and cell types that have been identified.

Another popular and very efficient method is the Uniform Approximation and Projection Method (UMAP)[32], an embedding technique that, similarly to the Louvain algorithm, is based on the construction of a nearest neighbor graph on the reduced dimension space defined by the principal components. Unlike t-SNE, UMAP can preserve the global topology of the transcriptional landscape and reconstruct the dynamic transitions between different clusters. Thus, it represents the best choice for catching a first “rough” glimpse of the developmental sequence of the original tissue before using dedicated trajectory reconstruction methods. As a downside, in some cases, a clear visual distinction of closely related clusters in the UMAP projection might be difficult, and this is especially evident in the presence of small transient subpopulations, that appear as stretched “smears” linking neighboring clusters to one another.

The choice of the most convenient embedding technique entirely depends on the purpose of the study. For exploratory analyses, t-SNE offers a nice visualization of the different clusters and cell types. For more complex analyses, e.g., to reconstruct the transcriptional evolution of the tissue across time, UMAP allows better appreciating the dynamic relationships between the cell subpopulations. However, UMAP provides just a static screenshot of the underlying dynamic process and, to disentangle the genealogy of the biological system, dedicated developmental trajectory reconstruction methods must be applied.

1.4. Trajectory reconstruction from scRNA-seq data

Amongst the numerous new horizons disclosed by single cell sequencing technologies, one of the most exciting and promising is the possibility to reconstruct the dynamic evolution of cell populations during key biological processes, like ontological development, or during disease onset and progression.

Traditional bulk DNA sequencing technologies, like, for instance, whole exome sequencing (WES), are designed to provide the catalogue of all the genetic variants that are present in a cell population[33]; however, without a clear knowledge on how exactly these variants are distributed across the different subpopulations it is not possible to reconstruct the path of each cell lineage. The advent of single cell sequencing brought the possibility to profile genetic variants at single cell level, thus providing the ideal tool to infer the developmental history of a tissue. By assuming explicit evolutionary models, it is possible to order the mutational profiles of cells along phylogenetic genealogies that explain how tissues and organs evolve during development, or how tumors expand in surrounding healthy tissues[34,35]. Just like the methods that are largely used in phylogenetics for the reconstruction of evolutionary trees of species, this operation is founded on the principle that, if mutations occur only once at the same site, they will progressively accumulate as cell populations diverge and distinct lineages originate from common progenitors[34].

This mutation-driven approach to single cell phylogenetics is still in its infancy, especially for single nucleotide variants[35]; moreover, it cannot be applied to scRNA-seq, that simply quantifies the expression level of every gene in each cell, regardless of its specific variants. However, phylogenetic reconstruction is still possible even with scRNA-seq data, albeit based on a different principle. Given the assumption that progenitor cell subpopulations progressively diverge and transform into their descendants according to a gradual and continuous process, if a static screenshot of the global transcriptional landscape of the whole population is taken at a precise moment of this process, the resulting portrait will be composed of an ensemble of cell states that are positioned at different levels on the continuum[8,36]. Precisely, these cell states will include pluripotent progenitors, terminally differentiated cell types, and, most importantly, transient subpopulations characterized by intermediate transcriptional profiles. The goal of the computational methods for trajectory reconstruction from scRNA-seq data is the reassembly of the original continuum from all these “fragments”, that must be ordered in the correct sequence entirely based on the transcriptional information of single cells[36]. This is achieved through the elaboration of mathematical models that describe how the gene expression profiles of cells evolve in time[7]. Dozens of different mathematical frameworks have been developed for this purpose[9], thus making trajectory reconstruction one of the most active fields in bioinformatics.

1.4.1. The Waddington differentiation landscape

The current vision of how cell populations evolve during dynamic processes is inspired by a conceptual model that dates to 1957, when the embryologist Conrad Waddington formulated the idea of differentiation landscape[37]. According to this concept, any dynamic process that involves the transition of cells through a series of stable states (such as the cell types that compose a tissue) can be described as a multidimensional landscape formed by stationary regions that are linked by low energy paths. For convenience, this multidimensional system can be imagined as a 3D mountain

landscape, characterized by peaks and valleys that are linked by slopes of varying steepness. On this landscape, the sections with a lower curvature (the peaks and the valleys) represent stable cell states, while the sections with a higher curvature (the slopes) represent the transition paths that are preferentially taken by the cells as they move from one state to another, guided by the energy gradient.

This model represents a very intuitive metaphor of the differentiation process that determines the gradual transformation of pluripotent stem cells into fully specialized cell types during tissue development (**Figure 4**). In this specific interpretation of the Waddington landscape, the elevation of the stable states represents the differentiation potency of their constituent cells, i.e., the potential of these cells to transfer to another state[4,38]. Higher peaks represent pluripotent cell states, lower peaks represent intermediate cell states, and valleys represent fully differentiated cell states. The slopes connecting all these states to one another represent transitional cell populations that are gradually moving from one state to the following through low energy paths[4]. Such paths correspond to decreasing differentiation potency gradients, as differentiating cells prevalently move from higher states to lower states, from peaks to valleys, like marbles moved by gravity[7]. As they proceed through the differentiation process, and thus as they descend through the landscape starting from peaks, pluripotent cells progressively lose all their differentiation potency, until they fully specialize into specific functions and reach a permanent differentiated state with very low potency, located in one of the valleys.

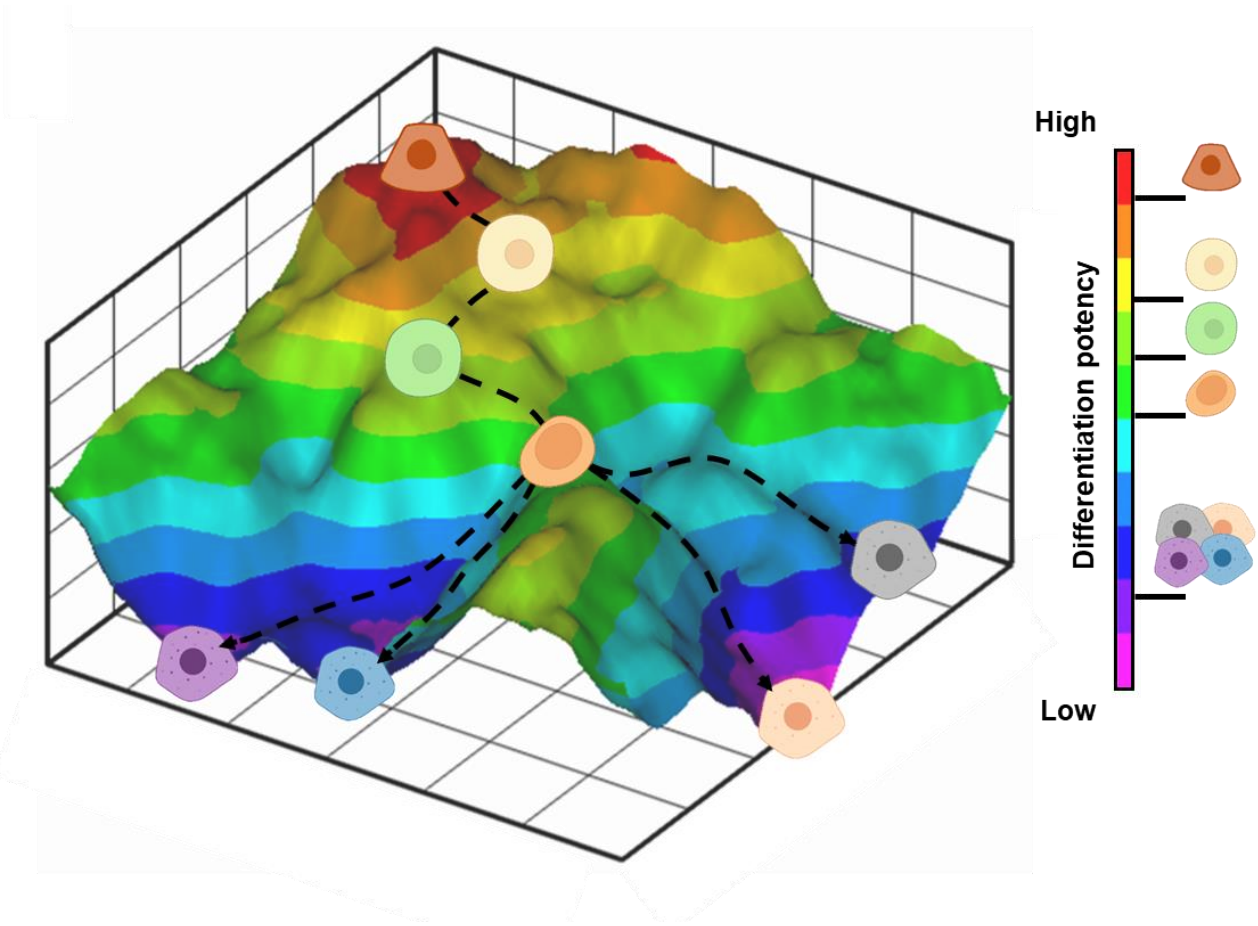

Figure 4: schematic representation of the Waddington differentiation landscape of a hypothetic cell population that evolves according to the laws of classic developmental processes.

This conceptualization works very well for ordinary differentiation processes in healthy tissues, but it is not yet clear to what extent its assumptions can be extended to other types of dynamic processes. For instance, in particular experimental settings, like the reprogramming of fully differentiated cells into induced pluripotent stem cells (iPSCs), the direction of the differentiation potency change should be ideally inverted, i.e., the potency of cells should increase during their regression to a more “primitive” state. In other scenarios, e.g., the onset and progression of cancer and other pathologies, the expected trend is even more unclear; in these situations, it is completely unknown whether the potency should increase or decrease in time, or whether it should even follow an entirely different and probably much more complex pattern. In such cases, the expected direction of the movement of cells on the Waddington landscape cannot be defined *a priori*.

Nonetheless, the interpretation of the Waddington landscape in the light of the differentiation potency of cell populations remains a reliable model for most normal developmental processes, and thus represents a solid background on which most computational methods for trajectory reconstruction have been designed and developed. The primary objective of these methods is to reconstruct, as faithfully as possible, the trajectories followed by cells on the Waddington landscape as they progressively differentiate into all the mature cell types of the tissue. This translates into the reconstruction of the sequence of all the dynamic transcriptional changes that the cells undergo during their descent from the peaks to the valleys. The difference between the various methods basically resides in how exactly they tackle this specific challenge (**Table 1**).

	Top-down	Bottom-up	RNA velocity
Operating principle	▪ Cells are embedded on a state manofild that recapitulates their reciprocal distances in the gene expression space ▪ The pseudotime score of each cell is computed as its distance from a root state chosen by the user	▪ The differentiation potency of each cell is directly inferred from the distribution of its transcriptional data across different functional gene categories ▪ To order the cells on a trajectory, the differentiation potency is assumed to continuously decrease	▪ The velocity of each gene in each cell is computed as the rate of change of its spliced counts over time ▪ The future position of each cell on an embedding of choice is computed as the most probable given the velocities of all its genes and the transcriptional states of its neighbors
Advantages	▪ Clear built-in method for the reconstruction of cell trajectories ▪ Robustness to the noise of scRNA-seq data ▪ Fast computational time	▪ The potency of cells is directly inferred from their transcriptional data ▪ No need for prior knowledge of the root and direction of the trajectory	▪ The dynamic trajectories of cells are directly inferred from their transcriptional data ▪ No need for prior knowledge of the root and direction of the trajectory
Disadvantages	▪ The pseudotime score is computed after the construction of the state manifold, thus it is not representative of the potency of each single cell ▪ Requirement of prior knowledge of the root and direction of the trajectory	▪ No clear built-in method for the reconstruction of cell trajectories ▪ Long computational time	▪ Strong dependence on the reliability of the phase portrait computed for each gene ▪ No robust method to compute the potencies of cells before the construction of the velocity vector field
Example tools	▪ Monocle ▪ Slingshot ▪ Palantir	▪ CytoTRACE ▪ SLICE ▪ SCENT	▪ velocyto ▪ scVelo ▪ veloAE

Table 1: summary of the different computational approaches to trajectory reconstruction from scRNA-seq data discussed in this thesis.

1.4.2. The “top-down” approach

Trajectory inference from scRNA-seq data is characterized by a particularly wide variety of available computational tools. As stated by Saelens et al[9] in a recent benchmark study, it is one of the largest categories in several repositories listing single-cell tools, such as omictools.org[39], the “awesome-single-cell” list[40], and scRNA-tools.org[41]. This huge variety is reflected by the equally wide diversity of the mathematical frameworks that are adopted to address the difficult task of reconstructing the dynamics of cells on the Waddington landscape uniquely from their gene expression data.

Despite their heterogeneity, a common thread is shared by the vast majority of tools, i.e., just like the methods that are commonly employed in standard analyses of scRNA-seq data, most pipelines for trajectory inference are designed to learn the structure and topology of the dynamic process that guides the evolution of a cell population from the distances between the single cells on a reduced dimension space that recapitulates the essential sources of variation of the gene expression data. The progress of each cell along the dynamic process, and thus its position on the Waddington landscape, is deduced from its location in such structure[4,7,8].

The intuition behind this approach comes from the observation that in evolving biological systems the major sources of transcriptional variance are indeed explained by the advancement of cells along a dynamic trajectory[8]. In 2010, Guo et al[42] observed that the first few principal components computed from qPCR data of mouse embryos were sufficient to order the cells on a sequence that efficiently recapitulated the whole process of preimplantation development. Similarly, in 2014 Shalek et al[43] observed that the process of dendritic cells activation after stimulation was accurately reconstructed by simply ordering the cells along the first principal component computed from scRNA-seq data. These observations suggested that the dynamic transcriptional changes that cells undergo as they progressively move across the differentiation landscape are the main factor determining their reciprocal distances on the gene expression space.

Given this premise, the aim of current trajectory reconstruction methods is to learn a “state manifold” from scRNA-seq data, i.e., a low dimension latent space that describes as faithfully as possible the genealogical relationships between the various cell subpopulations[4]. It is constructed based on distance measures that are computed either directly on the original gene expression space, or on a reduced dimension space derived from gene expression. Each algorithm employs a specific computational framework for building its own manifold, but in general the procedure is conceptually similar to the computation of embedding coordinates from the principal or diffusion components. The simplest example is the procedure implemented in the algorithm of Palantir[44], that computes a nearest neighbor graph on the top diffusion components. A prominent example of a more complex procedure is the reversed graph embedding (RGE)[45] algorithm implemented in Monocle[46], one of the most popular trajectory reconstruction methods. The RGE algorithm is designed to learn a principal graph from the gene expression data, i.e., a state manifold that groups the cells based on their transcriptional similarity, and simultaneously minimizes the displacement between the new coordinates of each cell and the respective position on the original gene expression space. The principal graph can be learned either directly from the gene expression matrix, or from the UMAP coordinates (in Monocle 3[47]). Another example of a very efficient and popular algorithm is Slingshot[48], that learns its state manifold by first constructing a minimum spanning tree on pre-computed cell clusters directly on the gene expression space (or, alternatively, on a reduced dimension space), and then by learning a separate principal curve for each lineage on the resulting tree.

Regardless of the specific mathematical details, all state manifolds share a common characteristic: since they are based on distance measures, their construction requires to parse the transcriptional

information of all the cells of the dataset at once[4]. This is very useful for compensating the typical noise of scRNA-seq data, but also rises the risk of overfitting, due to the high number of parameters that need to be optimized and that limit the possibility to extrapolate the results. Moreover, an important related downside is the requirement of a significant amount of prior information on the biological system at hand. Indeed, most tools assume that the user already has a clear idea of both the topology and the direction of the expected trajectory[8].

The topology of the manifold strongly depends on the specific computational method used for its construction, since each algorithm inevitably introduces several constraints on the resulting structure. As evidenced by the benchmark study of Saelens et al[9], who thoroughly tested 45 trajectory tools on both simulated and real datasets, each algorithm is optimized to reconstruct trajectories with a specific topology. For example, Monocle performs better on complex, multi-branched topologies, while Slingshot performs better on simple topologies with few branches. In most cases, when an algorithm is used to reconstruct a topology that is not best suited for, the number of branches is either underestimated or overestimated, to the point that a simple dimension reduction technique like PCA often achieves more plausible results[9]. If the topology of the expected trajectory is not clear *a priori*, the accidental use of a sub-optimal algorithm may result in a strongly biased or even incorrect outcome.

Moreover, even if the correct topology is successfully recovered, the state manifold does not provide any information on the direction of the dynamic transcriptional change. The user must set a specific cell, or a specific cluster of cells, as the root of the genealogy. This allows the computation of a "pseudotime" score that indicates the progress of each individual cell along the trajectory described by the manifold, and thus along the underlying dynamic process[7]. This score is typically computed by directly projecting the cells on the manifold, and then by simply calculating their respective distances from the root. For example, Monocle projects the cells on the principal graph[46] (**Figure 5**), Slingshot projects them on the principal curves[48], and Palantir computes the shortest path of each cell from the root on the neighbor graph[44]. The term "pseudotime" underlines the main purpose of the score as an approximation of the temporal coordinates of the cells on the trajectory, and thus, indirectly, of their respective differentiation potencies. By definition, the accuracy of the pseudotime score entirely depends on the accuracy of the manifold the cells are projected on, and, above all, on the accuracy of the root that has been chosen by the user. If the manifold is a faithful representation of the dynamic process that determines the evolution in time of the cell population, and if the root is correctly set, then the resulting pseudotime scores will most likely be a reliable proxy of the real positions of the cells on the differentiation landscape.

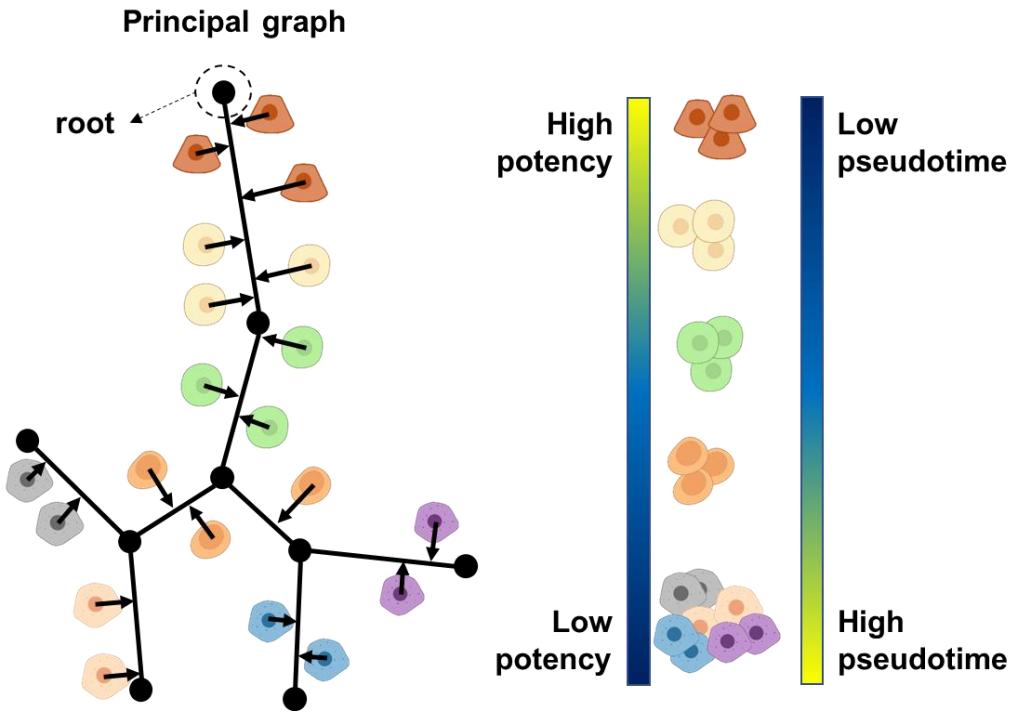

Figure 5: scheme showing the computation of the pseudotime scores of cells from their projection on the principal graph built by Monocle.

The workflow described above, that involves first the construction of a state manifold on the whole population, and then the computation of a pseudotime score for each individual cell, was defined in a recent review of Teschendorff and Feinberg[4] as “top-down” approach (**Table 1**), to highlight the holistic aspect of such implementation. Basically, the dynamic process that drives the evolution of the system is not directly reconstructed from the differentiation potencies of the single cells, but, on the contrary, it is inferred from the state manifold, and then employed to compute an indirect measure of the individual cell potencies.

Due to their robustness against transcriptional noise, and to their easy integration with standard analysis pipelines of scRNA-seq data, “top-down” methods have achieved a great popularity and have proven to be very efficient in reconstructing the differentiation processes of a wide variety of developing systems. However, as scRNA-seq technologies are being increasingly used to investigate more and more complex systems, the need is arising for a new completely unsupervised approach to trajectory reconstruction that does not necessarily require prior hypotheses on the topology and direction of the expected genealogy.

“Top-down” methods are particularly suited for ordinary differentiation processes that take place during tissue development in healthy conditions, since, in this scenario, the gradual change of the transcriptional programs makes it possible to employ classic distance metrics to recover the progression of cells along their respective lineages. But it is not yet clear to what extent this same assumption can be extended to lesser-known systems that evolve according to still obscure laws. In particular, severe diseases like cancer can totally subvert both the normal functioning of biological systems and their evolution in time, making it impossible to formulate reliable hypotheses on the expected trends of the gene expression change. In such situations, both the topology and the direction of the genealogy that has to be inferred are totally left to speculation. Hence, there is the pressing need of a different class of algorithms capable to reconstruct the trajectories of cells on the

Waddington landscape directly from their own differentiation potencies, with minimal input from the user.

1.4.3. The “bottom-up” approach

According to Waddington’s conceptualization of dynamic cellular processes, the differentiation potency of cells gradually changes as they move across the landscape; thus, the potency gradient is the force that moves the cells from one stable state to another, i.e., it is the source of the paths that cells preferentially take during the dynamic evolution of the system[4,7,37,38]. Consequently, to reconstruct the movements of cells across the Waddington landscape without any prior theoretical assumption on the expected trajectories, it is necessary to i) quantify the potency of each cell through an appropriate descriptive metric, and ii) infer the change of such metric across time through a mathematical model. This kind of approach has been defined by Teschendorff and Feinberg[4] as “bottom-up” (**Table 1**), to highlight the crucial difference with the traditional “top-down” approach: the computation of the potency of each single cell is the very first step of the workflow, while the inference of the evolution of the whole population across time is performed as the last step, by combining the dynamics of the potencies of all the cells (**Figure 6**).

The “bottom-up” approach is directly inspired by the principles of statistical mechanics[4], and represents the first real application of such principles to the analysis and interpretation of scRNA-seq data. Both the structure and the direction of the macroscopic dynamic process are modelled as natural emerging properties of the collective behavior of all its microscopic components, whose individual dynamics are empirically measured from the data. In other words, the “bottom-up” approach implies the construction of the differentiation landscape from its smallest components, rather than the decomposition of its overall structure. This totally unsupervised and data-driven approach aims to unleash the full potential of the level of detail comprised in scRNA-seq data by treating tissues and organs as complex dynamic systems that can be described by the theoretical foundations of statistical mechanics just like any other complex system present in nature.

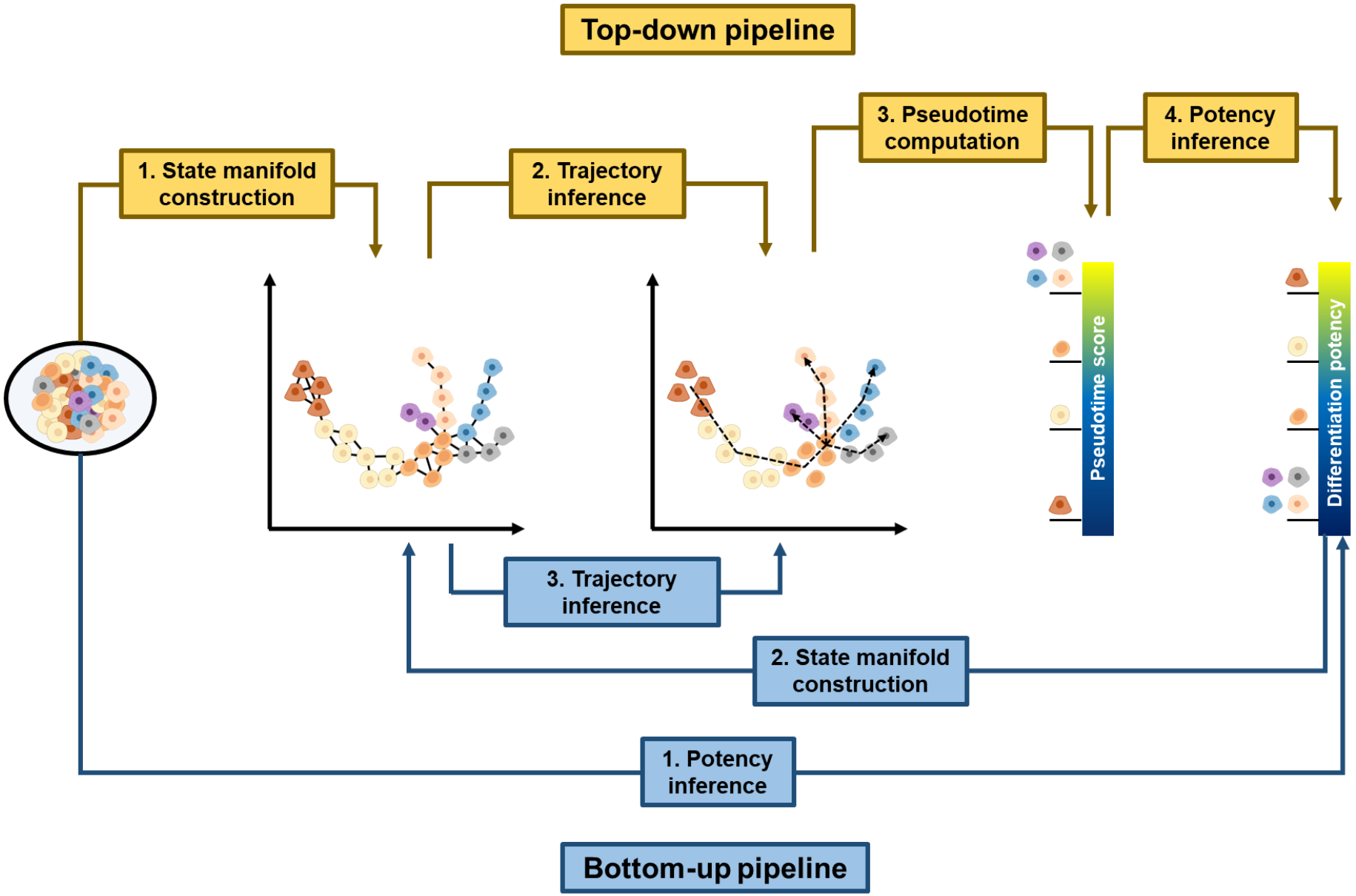

Figure 6: comparison between the workflows of top-down and bottom-up methods for trajectory reconstruction.

This radical change of perspective brings about several challenges, first of which is the definition of the basic building block of the process described above, i.e., of a reproducible descriptive metric that can efficiently summarize the differentiation potency of each single cell from its transcriptional data. To date, several candidate measures have been proposed to play this role[4]. Albeit their mathematical details are different, they all rely on a common rationale. First, since pluripotent cells have the potential to transform into multiple different cell types, they maintain their transcriptional program into a "generalized" state, i.e., they keep in an active state a wide range of molecular pathways that can potentially determine their differentiation into an equally wide range of descendants. Then, as they take a specific differentiation path and progressively move towards more and more specialized states, they gradually lose all their differentiation potency, and at the same time they concentrate their gene expression signal on very specific molecular pathways, i.e., those pathways that are strictly necessary for the specific functions of the mature cell type they are destined to become[49].

This rationale translates into a corollary that can be mathematically tested: while pluripotent cells are expected to express many genes, but at a relatively low level, terminally differentiated cells are expected to express few genes, but at a relatively high level. Cells that are located in intermediate stable states of the differentiation landscape, or that are moving from one state to another, are expected to show an intermediate behavior, both in terms of the number of active genes and in terms of their expression level[49]. Under this perspective, a reliable measure of the single cell differentiation potency must efficiently summarize the distribution of the gene expression signal of cells over different molecular pathways into a simple numeric score. To be effective, this score must ultimately be able to subdivide any cell population into all the differentiation states it is composed of, including pluripotent, intermediate, and terminally differentiated states.

Among several metrics that have been devised to date, a simple yet very efficient example is the total number of expressed genes. In a recent benchmark study performed on nine gold standard scRNA-seq datasets with experimentally confirmed trajectories, Gulati et al[50] demonstrated that the simple count of detectably expressed genes is one of the top-performing correlates of the differentiation status of single cells. This result strongly supports the idea of differentiating cells that progressively concentrate their expression signal on specific gene pathways, thus simultaneously decreasing the total number of transcriptionally active genes. Based on this finding, the authors of the study proposed a new measure of the potency of cells that is obtained by averaging the UMI counts of the genes whose expression is mostly correlated with the total number of expressed genes per cell, to reduce the noise and include only the most "dynamic" genes. The computation of this score is implemented in the tool CytoTRACE[50].

Although very practical and intuitive, the CytoTRACE score exploits just a minimum percentage of the transcriptional information contained in the scRNA-seq data, due to its strong filters. Moreover, being completely based on the raw count of expressed genes, the CytoTRACE score does not consider the interactions between gene products. The advancement of cells along dynamic processes is indeed determined by a complex network of protein-protein interactions, including not only the regulatory interplay between transcription factors and their targets, but also more transitory interactions, such as those involved in signal transduction and paracrine signaling[51]. This means that, as the dynamic process unfolds, entire gene pathways, rather than single genes, are activated or deactivated, and consequently the distribution of the expression signal over the network of protein-protein interactions completely changes. This is precisely the assumption at the core of the concept of "transcriptional entropy"[4,38,49,52,53], another very efficient measure of the differentiation potency of single cells that, unlike the CytoTRACE score, considers each gene as part of a complex interaction network, rather than an isolated entity.

The definition of “entropy” of a complex system is linked to the level of its “internal disorder”, that, in turn, strictly depends on the distribution of all the single microscopic components of the system. As these components transition from a sparse and unstructured distribution to a highly organized state characterized by precise patterns, the total entropy of the system drastically decreases. This concept is applicable to all systems characterized by a high structural complexity, including the transcriptomes of living cells. From the perspective of the distribution of the gene expression signal on different molecular pathways, pluripotent cells can be seen as highly disordered systems, since their signal is widely spread over a multitude of different pathways that could potentially oversee a wide range of different functions. Conversely, fully differentiated cells can be seen as highly ordered systems, since their expression signal is selectively concentrated on a few pathways that collectively oversee just a few very specific functions. Thus, the transcriptional entropy measures the “promiscuity” of the expression signal of a cell and is expected to be directly proportional to its respective differentiation potency[38,49,52,53].

To date, several metrics have been devised to measure the transcriptional entropy of single cells. One of the simplest and most intuitive is the single cell entropy (scEntropy) score computed by the SLICE[52] (Single Cell Lineage Inference using Cell expression similarity and Entropy) R package, that is based on the raw count of the detectably expressed genes within each functional gene category for each cell. According to the rationale behind this score, in pluripotent cells the total active genes are expected to be uniformly distributed across several different functional categories, thus leading to a high entropy score, while in terminally differentiated cells the total active genes are expected to be mainly concentrated in just few categories, thus leading to a more “skewed” distribution of gene expression and to a low entropy score (**Figure 7a**). Albeit the scEntropy score represents just a little step up in complexity compared to the CytoTRACE score, the developers of SLICE demonstrated its efficacy in four independent scRNA-seq datasets, highlighting its significant correlation with the differentiation stages of the respective cell subpopulations[52]. Given this result, the authors managed to reconstruct the expected cell genealogy of each analyzed dataset by simply connecting the entropy minima of pre-computed cell clusters with a minimum spanning tree. This method was implemented in the SLICE package as a graphical tool for reconstructing differentiation trajectories on the basis of the entropy data. It has the advantage of being computationally fast and fully cluster-based (thus less vulnerable to the high noise of scRNA-seq data), but relies on the very strong assumption that, since the direction of the trajectory is not implicit in the scEntropy score, the entropy is assumed to constantly decrease along the dynamic process. If this assumption holds true, then the edges of the minimum spanning tree can be “artificially” directed from the nodes with highest entropy to the nodes with lowest entropy. While this assumption is almost always satisfied in classic developmental systems, it is not necessarily met in different types of dynamic systems whose laws are still obscure.

Another weakness of SLICE scEntropy score is that, similarly to the CytoTRACE score, it is still based on the simple count of active genes, and thus does not take into account the complexity of all the protein-protein interactions that take place within each active molecular pathway. To maximize the discriminative power of transcriptional entropy, it is necessary to catch the finest details of the differential distribution of the gene expression signal of different cell subpopulations within the same biological system. The algorithm implemented in the SCENT[38,53] (Single-Cell ENTropy) R package satisfies this requirement through the concept of signaling entropy. SCENT is designed to operate on a cell-shared genome-wide protein-protein interaction (PPI) network[49], that includes all known interactions between all the gene products of the genome. The aim of the signaling entropy score is to measure the level of activation of each single interaction of the network for each given cell, and subsequently to summarize in a single number the pattern described by the distribution of active interactions on the entire network (**Figure 7b**). A high signaling entropy score is indicative of a large number of active interactions distributed across a large variety of different molecular pathways; this

means that the expression signal is uniformly distributed across a large portion of the PPI network, and that the cell is keeping in an active state all the molecular pathways that may be necessary for all its potential future states. This scenario is typical of pluripotent cells, whose future fate has yet to be determined. Conversely, a low signaling entropy score is indicative of a low number of active interactions distributed across just a few different molecular pathways; this means that the expression signal is mostly concentrated into a small portion of the PPI network, and that the cell is keeping in an active state just those few molecular pathways that are strictly necessary for a particular set of molecular functions. This scenario is typical of terminally differentiated cells, whose fate has already been determined, and are thus fully focused on a specific function. The key difference between SCENT signaling entropy score and SLICE scEntropy score resides in the finer scale of the former, that measures the distribution of the gene expression flow across all active interactions of all active pathways. Nevertheless, SCENT shares the main downside of SLICE, i.e., the lack of a proper mathematical model for the direct reconstruction of a dynamic trajectory from the computed scores without any prior assumption. Indeed, to order cells along a genealogy, the user is still forced to assume that the signaling entropy scores must necessarily follow a specific trend during the whole dynamic process, thus imposing external constraints on the results obtained by the algorithm.

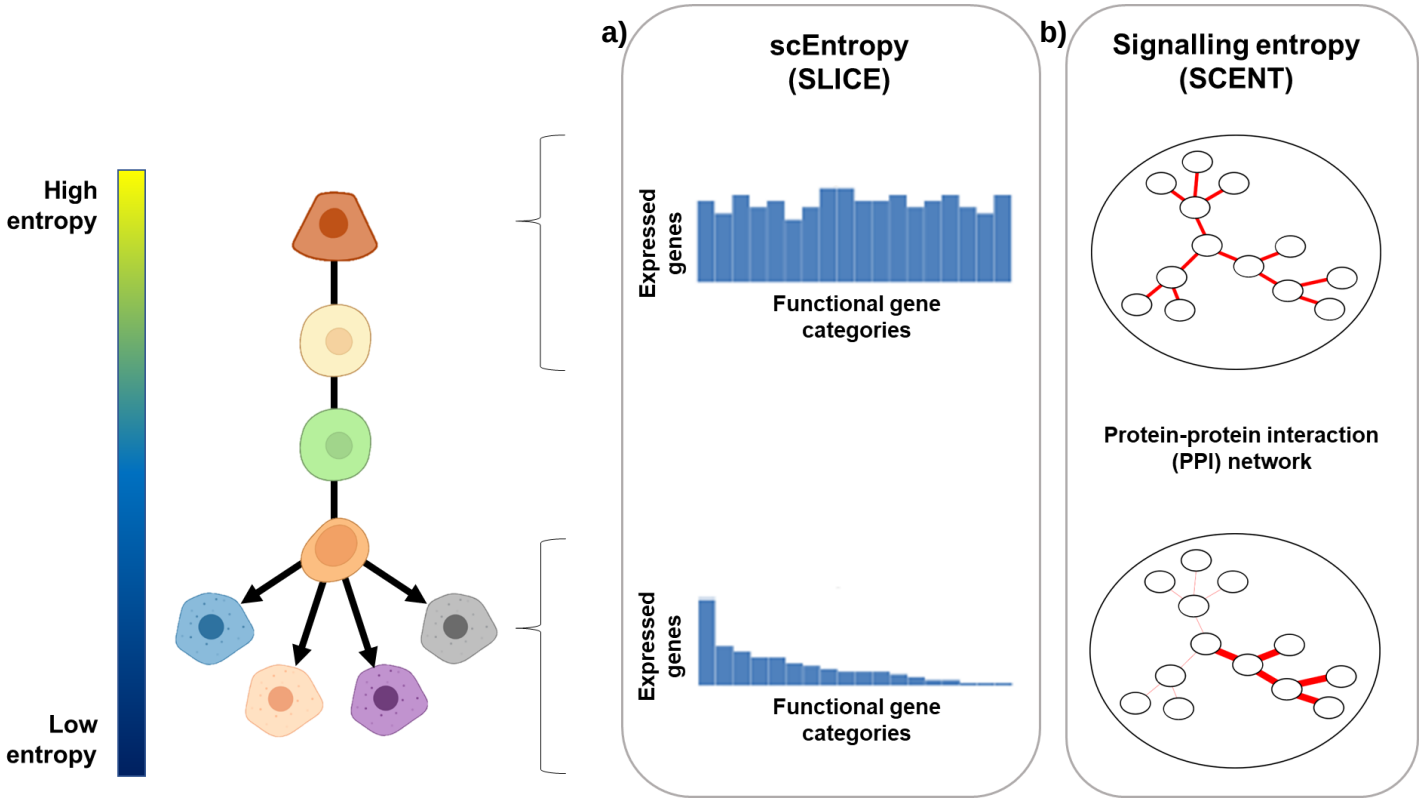

Figure 7: schematic representation of the rationale behind the entropy scores of the a) SLICE and b) SCENT R packages.

Under the assumption that the biological system evolves according to the laws of normal tissue development, any entropy score, regardless of the underlying mathematical details, can be seen as a reversed pseudotime score that, rather than increasing, progressively decreases along the trajectory. The crucial advantage of transcriptional entropy compared to pseudotime is that it directly estimates the differentiation potency of each individual cell from its own expression data, independently from the other cells of the dataset, thus reducing the risk of overfitting and allowing the full exploitation of the transcriptional information of all the cells, without the need to apply any

dimension reduction procedure. However, this total independence also brings about an important limitation: the entropy scores provide no indication about the direction towards which the cells are currently moving in the gene expression space. As discussed for the entropy scores of SLICE and SCENT, the user must impose a personal interpretation of the nature of the dynamic process at hand to be able to make predictions on the expected entropy trend through time. Alternatively, in the complete absence of any clue, the cell or the group of cells characterized by the lowest entropy score can be set as the root state of the system and then another method, like a traditional "top-down" tool, must be necessarily applied to reconstruct the trajectory from the designated starting point[4].

All in all, there is the need of a novel method that is able to both i) predict the potency change of each cell through time, and ii) subsequently relate the resulting measure to the observed potencies of the other cells of the same dataset, with the aim of predicting the direction towards which each cell is currently moving directly in the potency space. The expected result is the reconstruction of the dynamic trajectories of cell populations directly on the Waddington differentiation landscape, without the need to previously build a state manifold from raw gene expression data. According to the concepts of statistical mechanics[4], the entire process must exclusively depend on the prediction of the future change of the differentiation potencies of single cells. To date, a coherent method inspired by such strategy has not been devised yet; however, a very popular algorithm, called RNA velocity[54,55], has been designed to perform a similar operation on raw gene expression data. This method is designed to predict the rate of transcriptional change of each individual cell and to subsequently use this information to reconstruct dynamic trajectories without the superimposition of any direction. As such, it represents the ideal pillar to construct a novel unsupervised "bottom-up" method for the inference of the temporal dynamics of single cell potency measures.

1.4.4. The RNA velocity method

The concept of RNA velocity (**Table 1**) was first defined in 2018, when La Manno et al[54], in a ground-breaking study, proposed an innovative algorithm for predicting the trajectories of single cells in the gene expression space on the sole basis of their transcriptional dynamics, without the need to manually define the expected direction of the underlying process. The intuition at the base of this method comes from a parallelism with physics: given an object that moves on a bidimensional space, if we know the direction and magnitude of its velocity vector, we can easily predict its future spatial coordinates. Similarly, given a cell that moves on a bidimensional space constructed from gene expression data (such as, for instance, a UMAP or t-SNE embedding), we can predict its future coordinates if we know the direction and magnitude of the vector defining the rate of change of its transcriptional activity through time, i.e., the vector of its RNA velocity. This means that if we have a population of cells projected on a bidimensional embedding, we can draw, for each cell, an arrow that starts from the cell itself and points towards its predicted future position. The patterns formed by all these arrows will collectively form a fine-grained representation of the dynamic evolution of the entire cell population. Briefly, the whole trajectory is constructed by the cells themselves, thanks to a mathematical framework that is based on the molecular dynamics that genes typically undergo during differentiation processes.

As cells move across the Waddington landscape, the transcriptional activity of their genes will dynamically change according to patterns that depend on their specific roles. In general, each gene will undergo a phase of transcriptional induction when its product is most needed, and then will undergo a phase of transcriptional repression as soon as its product is not necessary anymore. The main idea behind the concept of RNA velocity is that this oscillatory behavior strongly affects the balance between the spliced and unspliced transcripts of genes. During transcriptional induction, pre-mRNA molecules are produced very quickly to the point that the splicing machinery is not able

to keep up, thus causing a temporary accumulation of unspliced transcripts. On the contrary, during transcriptional repression, the production of pre-mRNA molecules progressively slows down, until it eventually stops, thus the splicing machinery is able to catch up and process the excess of unspliced transcripts, until all of them are finally converted into spliced transcripts. Consequently, the spliced/unspliced ratio is expected to be very low during transcriptional induction and very high during transcriptional repression. During steady states, i.e., states characterized by constant transcriptional activity (regardless of the expression level), the ratio is expected to be close to 1, since transcription and splicing work at a similar regime.

During the entire course of a dynamic process, only a very small fraction of the genes of any given cell is expected to show a significant oscillatory behavior. These genes will most likely go through all the phases of induction, repression and steady state, therefore the static snapshot of the dynamic process provided by a typical scRNA-seq experiment will likely comprehend a sample of cells for all such phases. Consequently, to reconstruct the full dynamic transcriptional cycle of each single gene, i.e., its phase portrait, cells can be projected on a bidimensional space defined by spliced and unspliced UMI counts (**Figure 8a**). On this portrait, the steady state is represented by a straight line, whose inclination corresponds to the constant spliced/unspliced ratio, while the phases of transcriptional induction and repression are represented by curves that depart from the straight line upwards and downwards, respectively.

The RNA velocity of a given gene in each cell is defined as the rate of change of its transcriptional activity, measured as the rate of change of its spliced counts. This rate will be equal to 0 in steady states and will assume a positive or a negative value in the induction and repression phases, respectively. Graphically, it corresponds to the vertical deviation of the cell from the straight line on the phase portrait of the gene (**Figure 8a**). Computationally, it corresponds to the difference between the unspliced transcripts of the cell and the product between the spliced transcripts and the inclination of the straight line, that equals the degradation rate of the spliced mRNA molecules. The combined velocities of all genes allow predicting the future position of the cell in the gene expression space: the new coordinates are given by the correlation between the vector containing the velocities of all genes and the vector containing the transcriptional distance of the cell from its nearest neighbors for all genes. In other words, the cell moves towards the neighbors that are transcriptionally similar to its future “self” given by the combined velocities of all its genes (**Figure 8b**). As previously explained, each cell is linked to its predicted future coordinates by an arrow, and all the arrows of all cells collectively constitute the velocity vector field of the cell population. The velocity vectors can be drawn either on each single cell, or on grid points representing the average of a certain number of neighboring cells.

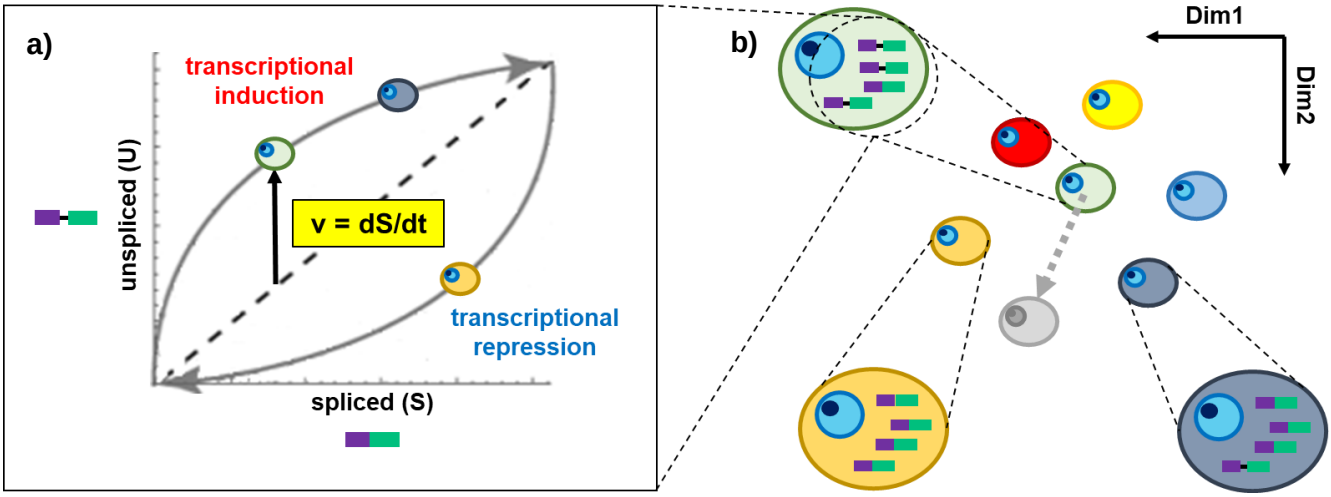

Figure 8: overview of a) the computation of RNA velocity from the phase portraits of genes, and b) the construction of the RNA velocity vector field on a cell embedding.

The key step of the RNA velocity method, i.e., the computation of the velocity of each gene in each cell, requires the construction of reliable phase portraits, especially for key genes that are directly involved in the biological process. In turn, the construction of the phase portraits requires the reliable estimation of several different parameters for each single gene, such as the transcription rate (both for the transcriptional induction and repression phases), the splicing rate, the mRNA degradation rate, and the latent time of all the cells, i.e., a numeric value that represents their "advancement" along the transcriptional cycle defined by the phase portrait. The original RNA velocity algorithm designed by La Manno et al is implemented in the velocyto[54] tool, that circumvents the difficult problem of estimating the latent times of all cells by assuming a constant splicing rate of 1 for all genes. By measuring all velocities in units of splicing rate, this expedient drastically reduces the parameters to be estimated to the sole mRNA degradation rate, but, as a downside, requires an adequate sample of cells located in the steady state portion of each phase portrait.

To overcome the limits posed by the assumption of a unitary splicing rate for all genes, as well as by the requirement of a minimum number of steady state cells, the model at the core of velocyto was expanded by Bergen et al into the fully dynamical implementation of the scVelo[55] algorithm. This upgraded version is designed to infer all the parameters of the phase portraits, including the latent times of all cells, by fitting a dynamical model to spliced and unspliced count data that exploits an expectation maximization framework iterating on maximum likelihood estimates. This implementation allows the computation of a separate set of transcription, splicing and degradation rates for each single gene, and does not necessarily require a fair sample of cells in steady state. Thanks to its capacity to efficiently fit detailed transcriptional models, scVelo has rapidly become the most popular method for the quantification of RNA velocity.

The clear advantage of RNA velocity methods is that the cell themselves dictate their own transcriptional fate, independently from each other. The most important downside is the necessity of the presence, in the scRNA-seq sample, of the whole spectrum of transcriptional states that are expected in the complete genealogy for each gene that shows a dynamic oscillatory behavior. Even if scVelo does not technically require that a fair number of cells have been sampled for all the phases of the biological process, to ensure the robustness of the estimated velocities the phase portraits of the key genes should encompass their whole dynamic cycle of transcriptional induction, repression, and steady state[56,57]. RNA velocity analyses conducted on samples that include only the initial or

terminal portion of the cell genealogy often yield unclear or even incorrect results. As such, an accurate sampling procedure is a key requirement of this method.

From the computational point of view, the most important characteristic of the RNA velocity algorithm is its flexibility. In the original implementation[54], the velocity vectors of cells are computed from their raw gene expression data, and the vector field is drawn on an embedding that has been previously constructed from the same data. However, the core mathematical framework can be easily adapted to other types of data, as has been demonstrated in a recent study where Tedesco et al[58] devised a new method, called Chromatin Velocity, that applies the velocity framework to chromatin accessibility data. As the RNA velocity method reconstructs the transcriptional dynamics of genes from the ratio between their spliced and unspliced transcript counts, the Chromatin Velocity method reconstructs the dynamics of chromatin folding from the ratio between the euchromatin and heterochromatin of cells. This alternative use of the velocity framework is made possible by the replacement of the count matrices provided by scRNA-seq with the count matrices provided by a single cell assay of chromatin accessibility called scGET-seq[58] (single-cell genome and epigenome by transposases sequencing), that is specifically designed to discriminate between open euchromatin and condensed heterochromatin. The Chromatin Velocity corresponds to the rate of change of the condensed portions of genes across time, and, similarly to RNA velocity, is used to predict the future positions of cells in an embedding that has been previously constructed from scGET-seq data. The Chromatin Velocity method proved to be very efficient in uncovering paths of epigenetic reorganization both during the reprogramming of cultured fibroblasts into induced pluripotent stem cells, and during the differentiation of these stem cells into neuronal progenitors[58].

The efficacy of Chromatin Velocity demonstrates that the mathematical framework of RNA velocity can be adapted to work in radically different contexts, provided that a single requirement is met, i.e., the definition of a numerical variable whose rate of change can be used to predict the future fates of cells on a bidimensional space that has been constructed from the same type of data. Regardless of the nature of the chosen variable, this framework allows each cell to autonomously define its own future trajectory according to its own internal dynamics. For this reason, the velocity algorithm represents the ideal component of a computational procedure for the reconstruction of dynamic trajectories from the differentiation potencies of single cells. Indeed, adapting the velocity implementation to the concept of transcriptional entropy, it is possible to devise a new fully “bottom-up” strategy for trajectory reconstruction that faithfully follows all the theoretical steps of statistical mechanics, from the independent measure of the dynamic properties of all the microscopic components of a biological system, to their combination into a coherent model that describes their collective emergent properties.

2. Aim of the study

2.1. FIERCE: a new “bottom-up” computational method for trajectory inference

The work of this thesis is focused on the design, implementation, and application of a “bottom-up” computational approach to reconstruct dynamic trajectories from scRNA-seq data. Taking inspiration from the principles of statistical mechanics, we developed FIERCE (Framework for InfERence of veloCity of the Entropy), a novel method, implemented into an R package, that is designed to estimate the differentiation potencies of single cells directly from their transcriptional data, and to subsequently use the resulting estimates to build a velocity vector field that predicts the future dynamics of cell populations directly on the potency space.

The algorithm of FIERCE is centered on the concept of “velocity of the entropy”, a novel notion that we specifically devised to provide a solid and coherent method for inferring the trajectory of each cell on the differentiation landscape on the sole basis of its own potency score. To achieve this task, we adapted the velocity mathematical framework to transcriptional entropy data, to predict the movements of cells on the entropy space based on the comparison between the rate of change of their entropy scores through time and the current entropy scores of their respective neighbors. The final aim of the method is the reconstruction of a velocity vector field that describes the dynamics of the whole cell population. Since transcriptional entropy is one of the most reliable measures of differentiation potency[4,38,50], this vector field offers the best representation of the dynamic evolution of the Waddington landscape given our current knowledge and computational tools.

FIERCE is designed to operate on raw scRNA-seq data, and thus does not require any prior explorative analysis. Although a prior independent cell annotation might be useful for visualizing the patterns of the vector field on specific subpopulations of interest, the computational step of the algorithm is fully based on a “bottom-up” strategy, with no prior assumptions needed neither on the structure of the population, nor on the direction of the dynamic process that must be reconstructed. As such, FIERCE is particularly suited for revealing the finest details of the structure and evolution of cell differentiation landscapes and possibly for uncovering new and unexpected patterns. We envisage that FIERCE will constitute a valuable addition to the current scenario of trajectory reconstruction methods from scRNA-seq data and will prove to be particularly useful for disentangling the dynamics of biological systems whose developmental details are still not completely clear, as well as for discerning previously unnoticed details of well-known processes.

2.2. Application of FIERCE on murine developmental processes

To evaluate the capability of FIERCE to build reliable and informative velocity vector fields, and to prove its utility in revealing the intrinsic patterns of the potencies of cells, we tested its performance by reconstructing three well known murine developmental processes from three scRNA-seq datasets. The resulting vector fields were compared with both the corresponding vector fields built by a widely used tool based on RNA velocity, and the corresponding trajectories built by a traditional "top-down" method. Our aim is to highlight the important advantages brought by our "bottom-up" approach, as well as the additional information that the velocity of the entropy can provide compared to classic RNA velocity. Below a brief introduction is provided for each dataset included in the study.

2.2.1. The pancreas endocrinogenesis dataset

The first dataset was sequenced by Bastidas-Ponce et al[59] to study mouse pancreas endocrinogenesis, a developmental process that leads to the differentiation of endocrine cells from multiple pluripotent cell subpopulations. In particular, the authors were interested in studying the maturation of functional alpha, beta, delta, and epsilon endocrine cells from pluripotent ductal cells, passing through two intermediate subpopulations, the endocrine progenitors and the pre-endocrine cells. The pivotal point of this linear genealogy is located at the level of endocrine progenitors: upon lineage priming, these cells go through a short mitotic phase, after which they are specifically committed to the endocrine fate by a transient peak of expression of the neurogenin 3 (Ngn3) transcription factor, that maintains a high expression level until the insurgence of pre-endocrine cells. To isolate the rare endocrine progenitors undertaking this crucial passage, the authors generated a novel reporter mouse line whose endogenous Ngn3 protein was fused with the Venus fluorescence reporter protein[59]. Additionally, they exploited the Epcam marker to isolate the nearby pancreatic

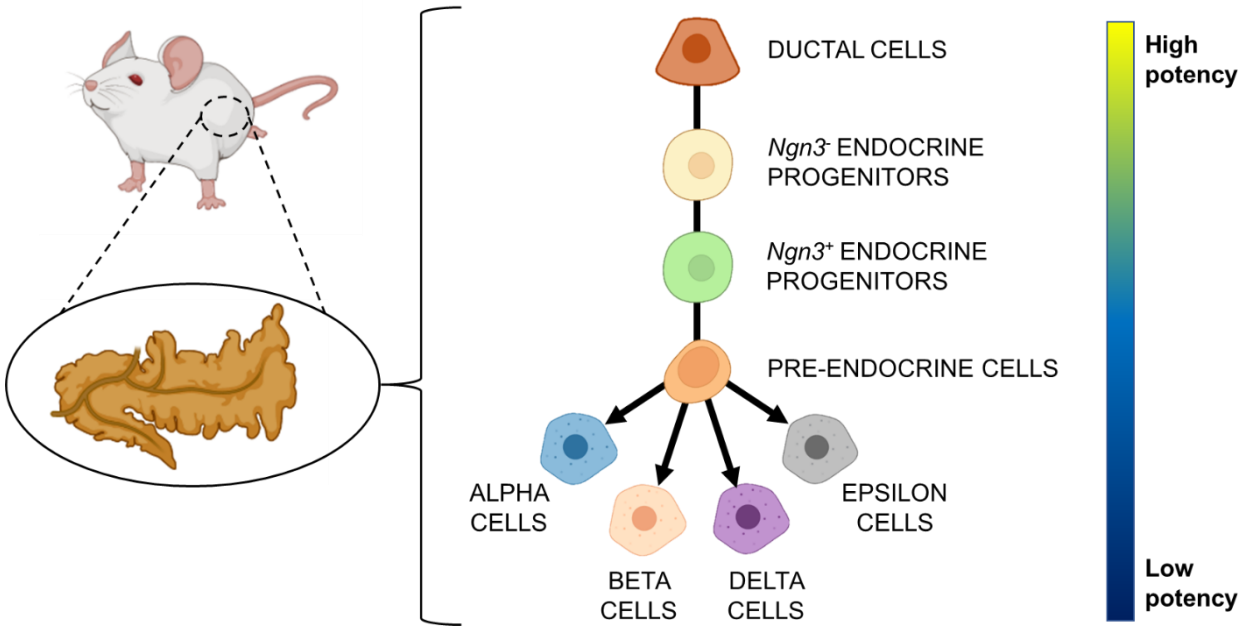

Figure 9: scheme showing the cell-type composition of the pancreas endocrinogenesis dataset from Beastidas-Ponce et al[59] and its expected genealogy. Figured created with BioRender.com.

epithelial fraction, to include the ancestors and the descendants of endocrine progenitors. From the total of 36,351 sequenced cells, we selected a subset of 3,696 cells from embryonic day 15.5, that comprehend ductal cells, endocrine progenitors (both *Ngn3*$^{-}$ and *Ngn3*$^{+}$), pre-endocrine cells, and endocrine cells of the alpha, beta, delta, and epsilon subtypes (**Figure 9**). We chose this specific subset since it was used by Bergen et al[55] to test the efficacy of their RNA velocity tool scVelo and represented the ideal dataset to perform a fair comparison between the vector fields constructed with classic RNA velocity and our velocity of the entropy. Hereafter, we will refer to this dataset as the "pancreas endocrinogenesis dataset".

2.2.2. The dentate gyrus neurogenesis dataset

The second dataset was sequenced by Hochgerner et al[60] to study the neurogenesis of mouse dentate gyrus, a developmental process leading to the differentiation of various neuronal cell types from radial glia and neuronal progenitor cells. In their original study, the authors assembled three separate datasets, that differ in the sampled time points and in the sequencing technology. For our study, we selected a subset of 9,505 cells from dataset C, that was sequenced with the 10X Chromium technology. These cells, coming from juvenile mice of postnatal days 0 and 5, cover the complete differentiation process of a specific neuronal cell type, the granule cells, from their pluripotent ancestors of the radial glia and neuronal progenitor subpopulation. Specifically, the included cell types, ordered according to the expected genealogical sequence, are quiescent radial glia cells (referred to as "radial glia" in the figures), cycling radial glia cells, intermediate neuronal progenitor cells (referred to as "nIPCs" in the figures), early neuroblasts (marked by the expression of the *Eomes* gene), neuroblasts, immature granule cells, and mature granule cells (**Figure 10**). We specifically chose this subset to concentrate on a single developmental sequence, and, similarly to the previous dataset, to perform a fair comparison with the RNA velocity vector field constructed by La Manno et al[54] with their tool velocyto. Hereafter, we will refer to this dataset as the "dentate gyrus neurogenesis dataset".

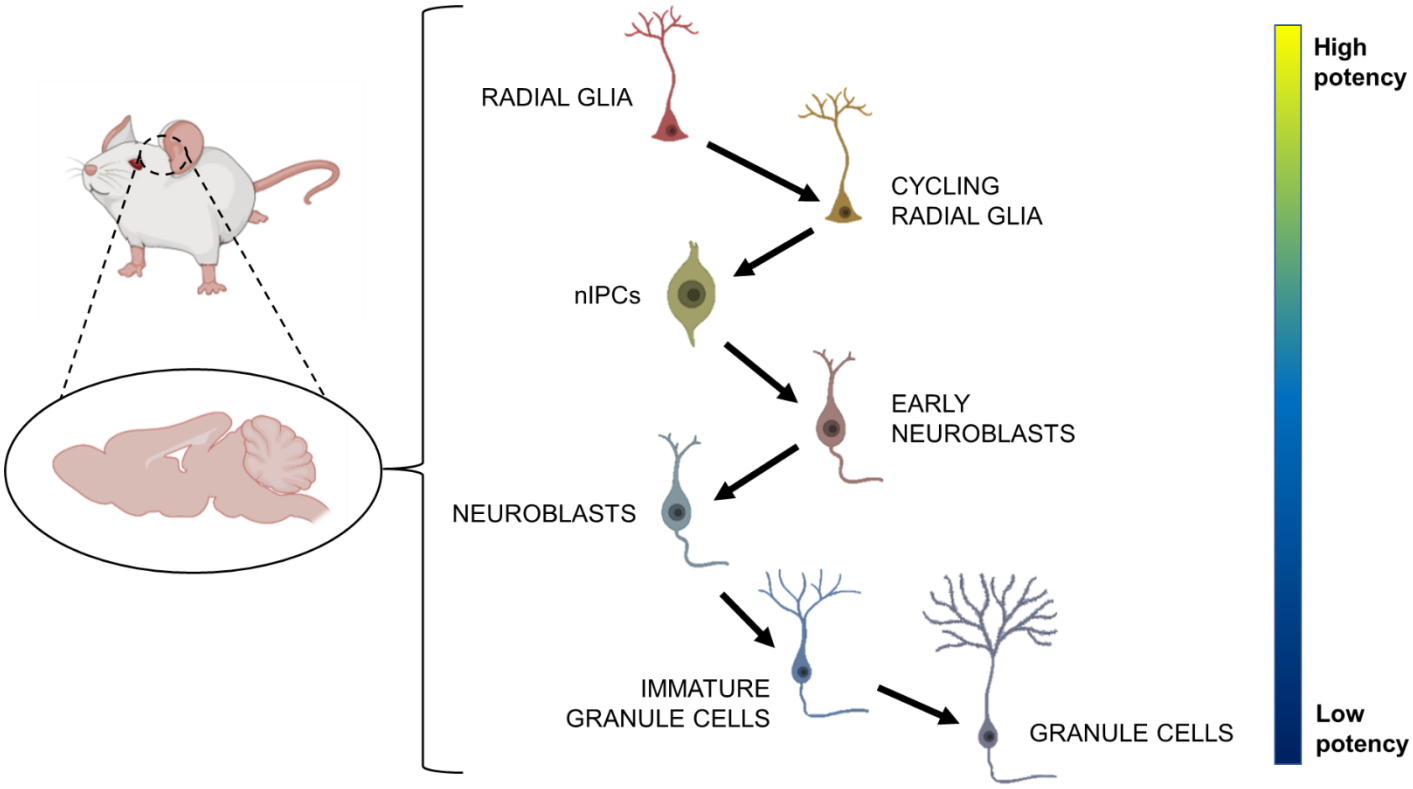

Figure 10: scheme showing the cell-type composition of the dentate gyrus neurogenesis dataset from Hochgerner et al[60] and its expected genealogy. Figured created with BioRender.com.

2.2.3. The mammary gland development dataset

The third dataset was sequenced by Giraddi et al[61] to study the biological programs that distinguish the fetal stem cells of murine mammary gland from their differentiating descendants. To cover the full developmental process, the authors sequenced a total of 4 samples from 4 different time points, specifically: embryonic day 16, embryonic day 18, postanal day 4, and adult week 12. We selected a total of 5,168 epithelial cells from all time points. The selected cells include the embryonic epithelial cells, that are supposed to be the root of the genealogy of this system, and the cells belonging to their two direct descendant lineages, i.e., basal cells and luminal progenitors; in addition, alveolar precursors and luminal cells, i.e., the direct descendants of the luminal progenitors, are included (**Figure 11**). Differently from the previous datasets, here the expected genealogy includes two branching points, which constitute an exemplary situation to test the efficacy of FIERCE in handling complex developmental landscapes. Hereafter, we will refer to this dataset as the "mammary gland development dataset".

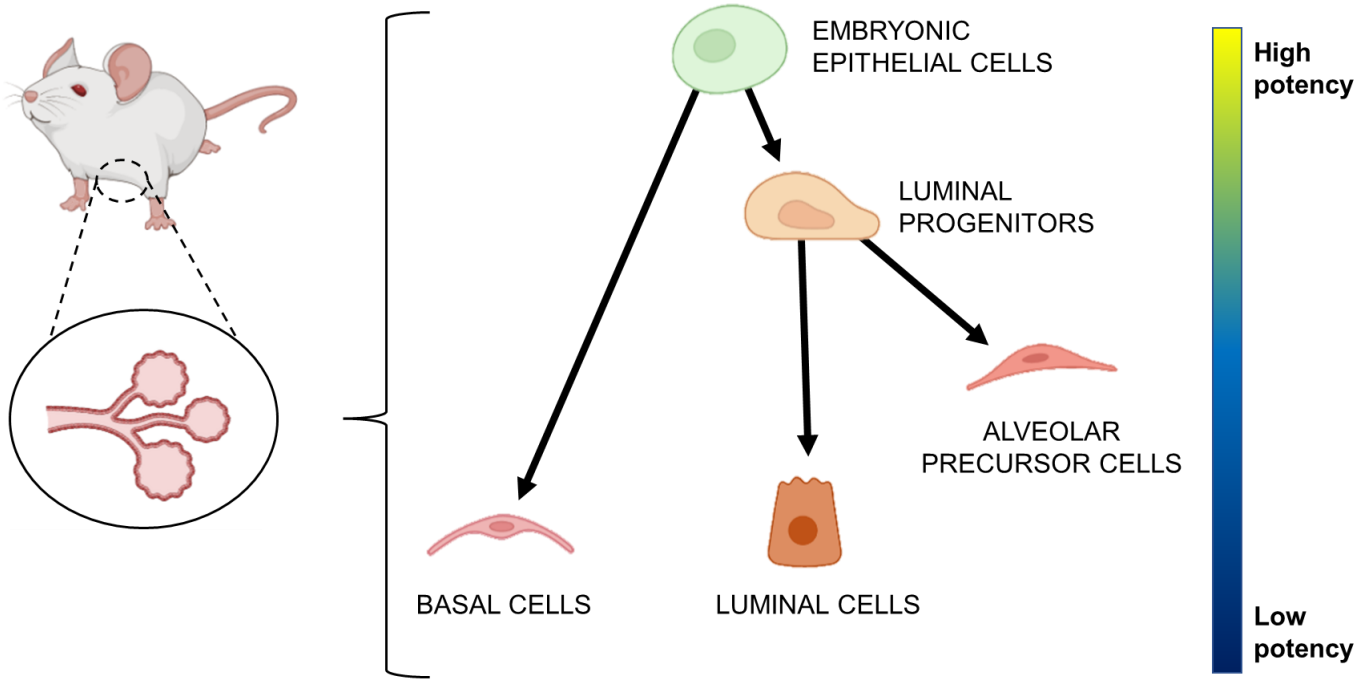

Figure 11: scheme showing the cell-type composition of the mammary gland development dataset from Giraddi et al[61] and its expected genealogy. Figured created with BioRender.com.

3. Materials and methods

3.1. FIERCE

FIERCE (Framework for InfERence of veloCity of the Entropy) is an R package that is designed to operate on an object of type anndata. Because anndata is a python package, commonly used for handling annotated data matrices, we structured FIERCE to mainly operate in the R environment and to occasionally interface with python through the reticulate[62] package, mainly for performing velocity-related computations and for handling the anndata object. We chose this solution to optimize the different steps of the computational analysis, that are best performed in different programming environments.

The anndata object, similarly to its Seurat equivalent, is organized in layers containing different types of data for the same cells and the same genes, and can store several additional information in dedicated slots, such as multiple annotations for both cells and genes, principal components and embedding coordinates. In the specific case of FIERCE, the anndata object must include 3 different layers: the main expression matrix, the spliced counts matrix and the unspliced counts matrix. Additional layers are subsequently added during the course of the pipeline, as will be explained below.

3.1.1. Computational details

3.1.1.1. The concept of velocity of the entropy

The computational pipeline of FIERCE is centered on the concept of “velocity of the entropy” and on the construction of the respective vector field. The velocity of the entropy is a measure that defines the rate of change of the transcriptional entropy of cells in time, and its vector field describes the dynamic evolution of a cell population directly on a representation of its differentiation landscape. To implement this concept, we devised a computational strategy that is founded on the following intuition: if we can predict the future transcriptional entropy of cells based on the dynamics of their genes, then we can calculate the difference between this predicted future entropy and the observed entropy of the same cells to obtain their entropy variation during a specific time lapse. To achieve this task, we resorted to the RNA velocity algorithm: as the first step of the analysis, the RNA velocity of all genes is computed to estimate their transcriptional change through a unitary time lapse; then, such rate of change is summed to the observed expression values of genes to predict their future transcriptional states; then, transcriptional entropies are computed from both the future and the observed transcriptional states; finally, the observed entropies are subtracted from the future entropies to obtain the change of signaling entropy during a unitary time lapse.

Because the velocity algorithm works on matrices to calculate a velocity value for each single gene in each single cell, we needed an entropy measure that could be decomposed to yield a single entropy value for each gene in each cell. Thus, we decided to use the signaling entropy computed by the SCENT[38,53] R package (v1.0.3), that is designed to compute both a total signaling entropy score for all the cells, and a specific local entropy score for each gene within each cell. This allowed us to relate the signaling entropy with RNA velocity by computing the local entropy scores on both

the observed and future expression matrices, making it possible to subtract the observed local entropies from the future local entropies to obtain a specific velocity of the entropy value for each gene in each cell.

Once the velocity of the entropy has been calculated, the construction of its vector field is straightforward: as we can predict the future position of cells on a UMAP embedding based on their RNA velocity and on the transcriptional states of their neighbors, we can translate this operation into the entropy language and predict the future position of cells on an entropy-derived UMAP embedding on the basis of their velocity of the entropy and of the local entropies of their neighbors. Each cell will be given an arrow that points towards the neighbors that are most similar to its "future self" given its rate of change of signaling entropy.

Figure 12 shows a schematic overview of the whole computational process. In the paragraphs below, more details are provided for the algorithms involved in each single step.

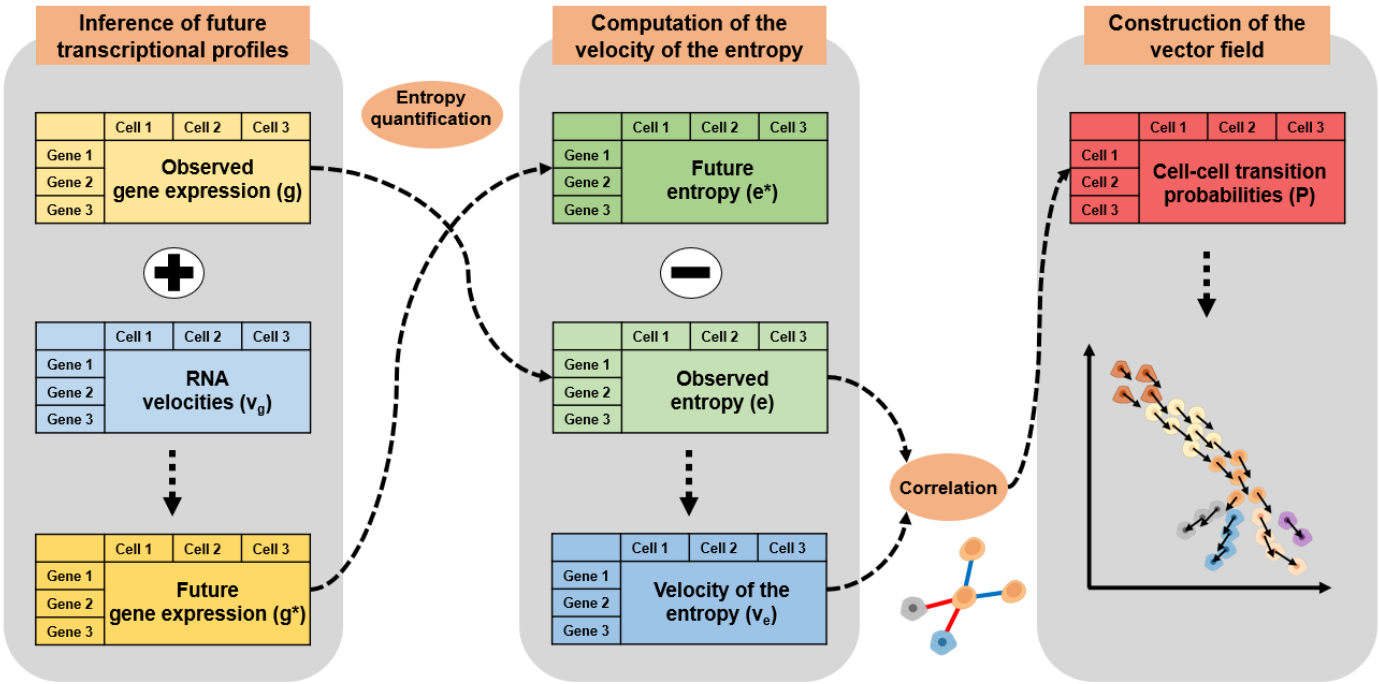

Figure 12: schematic overview of the core computational process of FIERCE.

3.1.1.2. Inference of future transcriptional states

Albeit the inference of future transcriptional states of cells is not the aim of the RNA velocity method, it is one of its most interesting implications: since RNA velocity is defined as the rate of change of the spliced transcripts of a particular gene[54], if we sum such rate to the current expression value of said gene, we can predict its future expression value after a unitary time lapse. If we perform the same operation for all the genes in all the cells, we obtain the full future gene expression matrix of the entire dataset. To achieve this task, we need robust RNA velocity estimates that can efficiently recapitulate the complex dynamic changes undertaken by all the genes during the course of the dynamic process. We chose the dynamical model implemented in the scVelo[55] python package (v0.2.4), that is currently the most popular tool for RNA velocity computation.

scVelo was developed by Bergen et al[55] as an upgrade to the original velocyto[54] algorithm that could overcome the two most important limits of the latter, i.e., the cumbersome assumption of a single unitary splicing rate for all genes, and the need for a fair sample of cells in steady transcriptional state. The dynamical model of scVelo circumvents these requirements thanks to its expectation maximization algorithm that operates on maximum likelihood values computed for all the dynamic parameters that need to be estimated to reconstruct robust phase portraits for all the key genes. These portraits are basically statistical models that describe the whole transcriptional cycle of each gene during the course of the dynamic process and are obtained by finding the optimal set of parameters that minimize the distance between the real cells and their "projections" on the model itself.

Let $x_i^{obs} = (u_i^{obs}; s_i^{obs})$ be the observed transcriptional state of the cell i for a particular gene, measured as its unspliced and spliced counts, and let $\hat{x}(t;\theta) = (\hat{u}(t), \hat{s}(t))$ be the expected transcriptional state defined by the gene-specific dynamical model (described by parameters θ) at latent time t. The latent time t of a given transcriptional state describes its "advancement" on the transcriptional cycle of the gene; it defines the coupling between a given observation and the corresponding estimate of the model. θ represents the ensemble of gene-specific parameters that define the model, and includes the splicing rate β, the mRNA degradation rate γ, and four section-specific transcription rates α. The four sections of the phase portrait include transcriptional induction, transcriptional repression, and the two respective steady states.

The aim of the algorithm is to find the model whose estimates best describe the observations at corresponding latent times. The residuals of the observations to the estimates of the model are defined as signed Euclidean distances (Deming residuals) given by:

$$e_i = \text{sign}(s_i^{obs} - \hat{s}(t)) \cdot ||x_i^{obs} - \hat{x}(t_i;\theta)||$$

If we assume that the residuals follow a normal distribution within each section of the portrait, then the overall likelihood of the gene, given by the contribution of all cells, is:

$$L(\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2n}\sum_i^n \frac{||x_i^{obs} - \hat{x}(t_i;\theta)||}{\sigma^2}\right)$$

Thus, the log-likelihood to be minimized is:

$$l(\theta) = \frac{1}{2}\log(2\pi\sigma^2) + \frac{1}{2n\sigma^2}\sum_i^n ||x_i^{obs} - x_{t_i}(\theta)||^2$$

To minimize this log-likelihood score for each single gene, the latent times t of all the cells in all the gene-specific models must be inferred, and this is a task of extreme complexity. The problem is solved by an expectation maximization algorithm that iterates between computing the latent times given the maximum likelihood estimates of parameters θ, and updating the parameters θ to maximize the likelihood based on the computed latent times.

Once the optimal dynamical model has been fit for all the genes given the observed transcriptional states, the computation of RNA velocity itself is very simple. For each gene, the velocity of cell i is given by:

$$v_i = u_i - \hat{\gamma} s_i$$

$\hat{\gamma}$ is the maximum likelihood estimate of the gene-specific mRNA degradation rate given by the optimal dynamical model. This is precisely the key innovation that makes scVelo much more precise than its predecessor velocyto: instead of inferring γ from the cells in steady state and measuring it in units of splicing rate, scVelo computes its maximum likelihood value given the observed expression data, thus allowing its accurate measure even for those genes that lack enough cells in steady state. In this scenario, the accuracy of a given velocity value depends on the accuracy of the γ estimate, that is indicated by the "final" likelihood of the optimal dynamical model.

Another key innovation of the scVelo algorithm is its reliance on the first order moments of genes. Both the dynamical model fitting and the velocity computation are not performed on raw spliced and unspliced counts, but on their first order moments, a particular transformation that aims to "smooth" the expression values of genes on the basis of their dynamics in the surrounding cells. Briefly, the raw spliced and unspliced counts are first normalized on the median of each cell; then, given a pre-computed neighborhood graph, the resulting normalized values are either increased or decreased depending on whether the expression value of the same gene in the neighboring cells is increasing or decreasing. Computationally, this is performed through a matrix multiplication between the spliced and unspliced normalized counts matrices and the cell-cell adjacency matrix that contains the neighbor connectivities. In practice, the advantage of this operation is to provide two new expression matrices where the dynamic trend of each gene across cells is drastically accentuated. This allows greatly improving the dynamical model fitting of all genes, and consequently to further improve the accuracy of their estimated parameters.

As the first step of its pipeline, FIERCE employs the refined algorithm of scVelo to calculate the RNA velocities of all genes in all cells. Because in any biological process only a minor subset of genes shows a true dynamic behavior across cells, the dynamical model will achieve a successful fitting in a minority of cases; for all the other genes, the RNA velocity is set to 0 in all cells. Afterwards, FIERCE sums the matrix containing the RNA velocities of all genes in all cells to the matrix containing the correspondent spliced first order moments, to obtain the matrix of the predicted future first order moments. This operation is performed on moments rather than on spliced counts because the dynamical model is fit on the former, and thus the RNA velocities effectively represent the rate of change of the spliced first order moments of cells.

3.1.1.3. Computation of the velocity of the entropy

Once the future transcriptional states of cells have been predicted, FIERCE computes the local signaling entropy scores both on the observed and on the future first order moments matrices. For this step, it employs the SCENT[38,53] R package (v1.0.3), that is designed to assess the distribution of the gene expression signal of cells on a cell-shared genome-wide protein-protein interaction (PPI) network.

The SCENT algorithm can ideally use any user-provided network, but obviously the robustness of the results directly depends on the completeness of such network. The more interactions are included, the more precise the results. The network may contain all known kinds of protein-protein

interactions for a given species, including the activation or repression of gene products by their regulators, signal transduction cascades, or even extracellular interactions like paracrine signaling. The only mandatory requirement is that all interactions must be coded into a gene-gene binary matrix, where 1 stands for an interaction between two gene products, and 0 stands for no interaction. We decided to program FIERCE to leave total liberty to the user regarding the choice of their own customized network; nonetheless, we also provide the original built-in human PPI network of SCENT[49] (derived from the integration of several databases from the PathwayCommons[63] platform), and, most importantly, four more complex networks that we built *de novo* for both human and mouse. For each of the two organisms, we built two binary matrices on the bases of all the interactions stored in STRING[64,65], a web database that can be used to retrieve all known and predicted protein-protein functional associations for multiple species. STRING is based on a specific score that indicates the level of confidence of a particular interaction, obtained through the integration of multiple different kinds of evidence, such as the conserved neighborhood of genes in different prokaryotic species, the co-occurrence of linked proteins across species, the co-expression of genes in the same or in other species, and evidence derived from other databases or from text-mining in literature. According to the guidelines provided in the STRING website (https://string-db.org/), we chose, for both human and mouse, two different thresholds for the significance of each interaction: a medium threshold of 0.4, and a high threshold of 0.7. Thus, we built two "medium confidence" networks where a protein-protein interaction is deemed significant, and thus is assigned a value of 1, if the correspondent STRING score is higher or equal to 0.4, otherwise it is assigned a value of 0. Similarly, we built two "high confidence" networks on the basis of the same principle, but with a higher threshold of 0.7. In absence of a user-provided network, the high confidence network of the selected species is used by default by FIERCE.

Given a reliable genome-wide PPI network, the computation of the signaling entropy of a given cell involves the assignment of weights to the edges of the network based on the expression values of the genes located at the respective nodes. Given two linked genes i and j, the weight of the edge is directly proportional to the expression of both:

$$w_{ij} \sim x_i x_j$$

Thus, the weights are interpreted as interaction probabilities, based on the assumption that in a sample where both genes are highly expressed, their products are more likely to interact with each other than in a sample where the same genes are expressed at low levels. Given this general interpretation of edges as signaling interactions, a random walk on the network is defined, provided that the outgoing weights of each node are normalized to sum up to 1. The result of the random walk is a stochastic matrix P with entries:

$$P_{ij} = \frac{x_j}{\sum_{k \in N(i)} x_k} = \frac{x_j}{(Ax)_i}$$

$N(i)$ denotes the neighbors of protein i, and A is the adjacency matrix of the PPI network. The signaling entropy is defined as the entropy rate (Sr) over the weighted network:

$$Sr(\vec{x}) = -\sum_{i=1}^{n} \iota_i \sum_{j \epsilon N(i)} p_{ij} \log(p_{ij})$$

ι is the invariant measure, that represents the relative probability of finding the random walker at a given node in the network in steady state conditions, i.e., long after the walk is initiated. The entropy rate is finally normalized to the maximum possible value among all compatible stochastic matrices:

$$SR(\vec{x}) = \frac{Sr(\vec{x})}{maxSr}$$

The equation above defines the total normalized entropy score over the whole PPI network, thus we will have a single such value for every single cell.

In a biological sample that includes all the stable cell states of a dynamic process, the distribution of the total entropy scores is expected to be multimodal, because each stable state will be characterized by a specific mean score and a specific variance. These distinct sub-distributions are called “potency states”, since they represent distinct subset of cells characterized by a specific differentiation potential. They can be identified by fitting a mixture of Gaussian curves to the total entropy scores in logit scale. Each successfully fit curve represents a potency state, characterized by a specific mean entropy score. The potency states are numbered on the basis of such mean score in decreasing order, meaning that the first potency state is characterized by the highest entropy, and thus is the most undifferentiated, while the last is characterized by the lowest entropy, and thus is the most differentiated.

The total entropy score of each cell can be decomposed into the local entropy scores of all the genes that participated into signaling entropy computation, i.e., those genes that are present both in the gene expression matrix, and in the PPI network. The local entropy score describes the distribution of the expression signal exclusively over the outgoing edges of each node of the network, i.e., over the interactions in which each gene is involved. Specifically, since the total entropy score can be seen as a weighted average of all the local scores, these are defined as:

$$S_i = -\sum_{j \epsilon N(i)} p_{ij} \log(p_{ij})$$

To allow the comparison of genes of different connectivity, each local score is normalized on the total number k of outgoing edges of the respective node i:

$$NS_i = -\frac{1}{\log(k_i)} \sum_{j \epsilon N(i)} p_{ij} \log(p_{ij})$$

This normalized local entropy score is bounded between 0 and 1, and is directly involved into the computation of the velocity of the entropy.

As the second step of its pipeline, FIERCE computes the signaling entropy scores of cells twice, first from the expression matrix containing the observed spliced first order moments, and then from the matrix containing the future spliced first order moments. The results of this operation are two separate entropy datasets, each one consisting of a vector of total entropy scores for all cells, a vector containing the potency state assignment of each cell, and a matrix containing the local entropy score of each gene in each cell. To calculate the velocity of the entropy, FIERCE subtracts the observed local entropies matrix from the predicted local entropies matrix. The result is a matrix containing, for each gene in each cell, the change of the local entropy score across a unitary time

lapse (i.e., the velocity of the entropy). Of course, the two local entropies matrices and the velocity of the entropy matrix include only the subset of genes that are involved in the computation of signaling entropy. For this reason, at this point of the pipeline FIERCE generates a subset of the original anndata object containing all the cells, but only the genes used for entropy computation. All the downstream operations described below are executed on such object.

3.1.1.4. Construction of the vector field

The velocity of the entropy matrix is finally used to predict the future positions of cells on the signaling entropy space. This operation involves two key steps: the construction of a cell embedding from signaling entropy values, and the computation of cell-cell transition probabilities on such embedding.

The first step is achieved by performing a PCA on the observed local entropies matrix, followed by the construction of a neighborhood graph and a UMAP embedding from the resulting principal components. Both operations are performed with the Scanpy[18] python package (v1.9.1). The construction of a good embedding is fundamental for drawing a reliable vector field, thus the user is free to specify any number of significant components for the neighborhood graph computation, as well as any number of neighbors for each cell on such graph (a very impactful parameter for UMAP construction); the UMAP coordinates deriving from the different combinations of these two parameters are saved into separate slots of the anndata object, and the correspondent UMAP plots are saved into different directories, so that afterwards the user is free to decide which particular embedding to use for the downstream analyses.

The second key step mentioned above, the computation of cell-cell transition probabilities, is performed by calculating, for each cell i, the cosine correlation between the vector containing the velocities of the entropy (v_e) of all its genes, and each one of the vectors containing, for the same genes, the difference (δ) between the local entropy of the cell and the local entropy of all its neighbors j on the entropy-based embedding:

$$\pi_{ij} = cos\angle(\delta_{ij}; v_{ei}) = \frac{\delta_{ij}^T v_{ei}}{\left|\left|\delta_{ij}\right|\right| ||v_{ei}||}$$

The resulting similarity matrix π_{ij} is called “velocity graph”, and its entries measure the similarity in direction, not in magnitude. They range from -1, that indicates opposite direction, to 1, that indicates identical direction, with the middle point 0 indicating orthogonal direction (and thus no correlation). By default, FIERCE also performs an additional variance-stabilizing transformation, such that the cosine correlation is computed between $sign(\delta_{ij})\sqrt{\left|\delta_{ij}\right|}$ and $sign(v_{ei})\sqrt{|v_{ei}|}$; in practice, this transformation greatly helps in drawing clearer and smoother vector fields. In any case, the velocity graph, either transformed or not, is converted into the final cell-cell transition probability matrix through an exponential kernel:

$$\tilde{\pi}_{ij} = \frac{1}{z_i}\exp(\frac{\pi_{ij}}{\sigma_i^2})$$

$z_i = \sum_j \exp(\frac{\pi_{ij}}{\sigma_i^2})$ is a row normalization factor, and σ_i is the kernel width parameter.

The cell-cell transition probability matrix $\tilde{\pi}_{ij}$ defines the probability of each cell to move towards each of its neighbors on the entropy embedding based on the velocity of the entropy of its genes. Thus, to predict the future position of the cell on the embedding, it is necessary to take into consideration all these probabilities at once. The positions of cells on the embedding are defined by a set of coordinate vectors $\tilde{s}_1, \ldots, \tilde{s}_n$, thus the vectors of two cells i and j can be subtracted from one another to obtain the distance between the respective positions on the embedding:

$$\tilde{\delta}_{ij} = \frac{\tilde{s}_j - \tilde{s}_i}{||\tilde{s}_j - \tilde{s}_i||}$$

The future position of cell i on the embedding is defined as its displacement from its current position, also called "embedded velocity" ($\tilde{v}_i$), and is computed by weighting the distances $\tilde{\delta}_{ij}$ from all its neighbors j on the basis of the transition probabilities $\tilde{\pi}_{ij}$:

$$\tilde{v}_i = \sum_{j \neq i} (\tilde{\pi}_{ij} - \frac{1}{n}) \tilde{\delta}_{ij}$$

Subtracting $\frac{1}{n}$ corrects for the non uniform density of points on the embedding. Once the displacements have been computed for all the cells, an arrow is drawn to connect the current position of each cell to its predicted future position. To improve the clarity of the vector field, the arrows can be also drawn on grid points that represent a weighted average of the surrounding cells. Specifically, FIERCE is designed to draw the vector field on such a grid, and then to apply a smoothing procedure to connect the arrows of nearby grid points into continuous curves that allow a much better visualization of the overall dynamics of the cell population. The resulting plot is commonly known as "streamplot", and represents the final output provided by FIERCE. The computation of both the velocity graph and the embedded velocities is performed with the scVelo[55] python package (v0.2.4).

3.1.2. Overview of the functions

The three essential computational steps of the workflow of FIERCE described above are performed by four required functions, that are briefly illustrated below.

compute_velocity: the task of this function is to compute the first order moments from spliced and unspliced counts and to fit the dynamical model for RNA velocity computation. Moreover, it uses the results to predict the future expression profiles of cells. The user can customize the RNA velocity computation in several ways: they can choose how many principal components to use for the neighborhood graph computation (necessary for converting normalized counts into first order moments), how many neighbors to compute for each cell, and which distance measure to use for the smoothing procedure of first order moments computation (the choice is between the connectivities or the distances of the neighborhood graph). The user can also filter the genes based on their spliced and unspliced counts, or even decide to directly use the normalized counts for RNA velocity computation rather than the first order moments.

compute_signaling_entropy: this is the core function of the computational pipeline of FIERCE. It computes the signaling entropy on both the current and the future expression profiles of cells, and subtracts the observed local entropies from the future local entropies to compute the velocity of the entropy. Here the user can decide whether to use the built-in PPI networks of FIERCE for signaling

entropy computation, or to provide their own personalized network, as long as it is formatted as a binary gene-gene matrix. This is the most time-consuming and computation-intensive function. The running time directly depends on the size of the dataset, but can be roughly estimated in the order of a few hours.

compute_entropy_UMAP: this function performs the PCA and the UMAP embedding computation on the entropy values calculated by the *compute_signaling_entropy* function. It prints several diagnostic plots into a separate sub-directory, including an "elbow plot" that shows the variance explained by each principal component, a separate heatmap for each single component that shows the local entropy values of the most correlated genes, and the embedding of cells on both the first two principal components and the UMAP coordinates. The embeddings can be colored according to any desired cell annotation, and both the order and colors of labels can be personalized. A key feature of this function is the possibility to build the neighborhood graph for UMAP computation from any number of principal components, and with any number of neighbors for each cell. Multiple values for both these parameters can be specified at once; in this case, the UMAP embeddings corresponding to each combination of principal components and cell neighbors are stored into separate slots within the anndata object, a feature that will be very important for the following function.

compute_graph_and_stream: this function performs the final step of the analysis, i.e., the velocity of the entropy vector field construction. The vector field can be drawn on any embedding stored in the anndata object, thus on any entropy-based UMAP embedding computed by the *compute_entropy_UMAP* function. The resulting streamplot can be colored according to any cell annotation, and, as before, the order and colors of the labels can be freely specified. Additionally, the user can decide whether to perform the variance-stabilizing transformation during the velocity graph computation, and whether to include, in such computation, only the genes that achieved a good fitting of the RNA velocity dynamical model performed by the *compute_velocity* function. The threshold for a good fitting is a final likelihood value of at least 0.01. By default, all the genes of the anndata object are used for velocity graph computation.

FIERCE also includes a few additional functions that are not strictly necessary for the main computation, but can be very useful for preprocessing the dataset, or for visualizing particular features of interest. Below we provide a brief illustration of each of these optional functions.

load_test_dataset: this function loads in the current working environment the anndata object of one of the two built-in tutorial datasets that can be used to familiarize with the package. These datasets are a subset (499 cells each) of the pancreas endocrinogenesis and the dentate gyrus neurogenesis datasets analysed here.

save_h5ad: this function is used to save an anndata object into an h5ad file.

load_h5ad: this function is used to load an anndata object from an h5ad file.

build_adata_object: this utility function is necessary for the creation of a suitable anndata object from sequencing data if it is not already available. Its only requirement is the availability of a loom file created from one of the dedicated functions of velocyto[54]. The loom file contains the raw sequencing data of a given sample subdivided into spliced and unspliced counts. The *build_adata_object* function creates a brand new anndata object with a dedicated layer for both spliced and unspliced counts. By default, the slot of the main gene expression matrix is filled with a copy of the spliced

counts matrix. An important side-feature of this function is the possibility to retrieve additional useful information from a Seurat[14,16,17] object and store it into dedicated slots within the new anndata object. We included this feature because, before proceeding with trajectory analysis, samples are often analysed separately with Seurat, so the user might want to include in the anndata object some cell or gene annotations, or some pre-computed principal components or UMAP embeddings. It is also possible to retrieve all this information from merged or integrated datasets, and to replace the spliced counts in the main expression matrix with the expression data of the Seurat object. In absence of the latter, the user can also provide separate data-frames containing cell annotations or embeddings of interest.

perform_preprocessing: this is another utility function that can be very useful in case the user wishes to perform a quick standard bioinformatic analysis on an anndata object that contains still unprocessed sequencing data. This standard analysis is performed with the Scanpy[18] python package (v1.9.1), and includes all the early steps of filtering, normalization and scaling, but, among the later advanced steps, only includes PCA and UMAP computation, that are strictly necessary for the FIERCE analysis. The plots that illustrate the results of each step of the analysis are saved into a separate sub-directory, and the PCA and UMAP plots can be colored according to any cell annotation stored in the anndata object. Additionally, the user can specify the desired order by which the labels will be printed in the legends, and the color to assign to each label.

plot_velocity: this function is simply used to draw an RNA velocity vector field from the results of the *compute_velocity* function. The vector field can be drawn on any embedding of choice, that can be colored according to any cell annotation stored in the anndata object. The order and the colors of the labels can be freely specified. The user can also choose whether to apply the variance-stabilizing transformation during the computation of the velocity graph.

plot_entropy_results: this function prints the results of the *compute_signaling_entropy* function into several diagnostic plots that are saved into a separate sub-directory. Among these, the most important are a boxplot that shows the distribution of total entropy scores across any cell annotation of interest, and a dotplot that shows the distribution of the potency states identified by the previous function across the same cell annotations. The user can freely specify the order of the labels on both the boxplot and the dotplot, as well as the colors of the labels on the boxplot.

plot_signature_statistics: this function prints the phase portraits of the top 20 likelihood genes of the dynamical model fit by the *compute_velocity* function, together with a cell embedding chosen by the user and colored according to the expression values (either the normalized spliced counts, or the spliced first order moments), the local entropy scores, the RNA velocity and the velocity of the entropy of the same genes. Alternatively, the user can also provide their own customized gene signature; of course, in this case only the genes that successfully fit the RNA velocity dynamical model will be shown. The cells on the phase portraits can be colored according to any annotation.

3.2. Preprocessing of scRNA-seq datasets

3.2.1. Pancreas endocrinogenesis

The anndata object containing the sequencing data of the pancreas endocrinogenesis dataset of Bastidas-Ponce at al[59] was directly retrieved from the scVelo[55] python package (v0.2.4). The object contained a total of 27,998 features profiled for 3,696 cells, subdivided as in **Table 2.** The sequencing data included both spliced and unspliced raw counts.

Cell type	Number of cells
DUCTAL CELLS	916
Ngn3⁻ ENDOCRINE PROGENITORS	262
Ngn3⁺ ENDOCRINE PROGENITORS	642
PRE-ENDOCRINE CELLS	592
ALPHA CELLS	481
BETA CELLS	591
DELTA CELLS	70
EPSILON CELLS	142

Table 2: cell type composition of the pancreas endocrinogenesis dataset.

A total of 50 principal components, computed from 4,004 highly variable genes, were already included in the object, as well as the UMAP coordinates computed from all the principal components and derived from a neighborhood graph with 15 neighbors for each cell. The first 30 principal components were used by the *compute_velocity* function to initiate the FIERCE analysis, through the construction of a new neighborhood graph (with 30 neighbors for each cell) for the computation of first order moments. The rest of the FIERCE analysis was executed according to the default pipeline (in particular, the built-in high confidence murine PPI network was used).

The vector field built by FIERCE was compared with both the corresponding vector field built by scVelo[55] (v0.2.4) from classic RNA velocity, and the principal graph reconstructed by Monocle 3[47] (v1.3.1). The scVelo analysis was directly executed during the pipeline of FIERCE, with the default procedure described above. In the Monocle 3 analysis, we used the dedicated functions of the tool to pre-process the raw gene expression data, to perform a PCA, and to construct a UMAP embedding on 50 principal components. Then we followed the default procedure to build the principal graph and to compute the pseudotime score of cells based on their projections on such graph. For the latter step, we used the *get_earliest_principal_node* helper function to manually set the ductal cells as the root of the genealogy.

3.2.2. Dentate gyrus neurogenesis

The anndata object containing the sequencing data of the dentate gyrus neurogenesis dataset of Hochgerner et al[60] was retrieved from the url http://pklab.med.harvard.edu/velocyto/DentateGyrus/DentateGyrus.loom, as indicated in the tutorial of velocyto[54] dedicated to this dataset (https://github.com/velocyto-team/velocyto-notebooks/blob/master/python/DentateGyrus.ipynb). The object contained a total of 27,998 features profiled for 18,213 cells, from which we selected the 9,505 cells involved in the granule cells developmental trajectory, subdivided as in **Table 3**. The sequencing data included both spliced and unspliced raw counts.

Cell type	Number of cells
RADIAL GLIA	388
CYCLING RADIAL GLIA	1,043
nIPCs	1,230
EARLY NEUROBLASTS	419
NEUROBLASTS	1,003
IMMATURE GRANULE CELLS 1	2,460
IMMATURE GRANULE CELLS 2	2,099
GRANULE CELLS	863

Table 3: cell type composition of the dentate gyrus neurogenesis dataset.

Since the anndata object did not contain any principal components or UMAP coordinates, necessary for the FIERCE analysis, we used the *perform_preprocessing* function to perform a standard bioinformatic analysis with Scanpy[18] (v1.9.1) directly on the matrix of spliced counts. We normalized the count data with a scale factor of 10^4, and log-transformed the normalized counts. To identify the highly variable genes, we used the default procedure of Scanpy, that subdivides the genes into bins based on their mean expression, and then obtains the normalized dispersion within each given bin by scaling with the mean and standard deviation of the dispersions of the included genes; we obtained 1,795 highly variable genes with lower and upper cut-offs of respectively 0.0125 and 3 for the mean expression, and a lower cut-off of 0.5 for the normalized dispersion. We scaled the highly variable genes to 0 mean and unit variance and regressed out the effect of total UMI counts and total expressed genes. We then performed a PCA on the scaled matrix and constructed a neighborhood graph on the first 30 principal components, with 30 computed neighbors for each cell. We finally computed the UMAP coordinates on the resulting graph, that was also used by the *compute_velocity* function to compute the first order moments, thus initiating the FIERCE analysis. The latter was then executed according to the default procedure (in particular, the built-in high confidence murine PPI network was used).

The vector field built by FIERCE was compared with both the corresponding vector field built by scVelo[55] (v0.2.4) from classic RNA velocity, and the principal graph reconstructed by Monocle 3[47] (v1.3.1). The scVelo analysis was directly executed during the pipeline of FIERCE, with the default procedure described above. In the Monocle 3 analysis, we used the dedicated functions of the tool

to pre-process the raw gene expression data, to perform a PCA, and to construct a UMAP embedding on 50 principal components. Then we followed the default procedure to build the principal graph and to compute the pseudotime score of cells based on their projections on such graph. For the latter step, we used the *get_earliest_principal_node* helper function to manually set the quiescent radial glia as the root of the genealogy.

3.2.3. Mammary gland development

The mammary gland development dataset of Giraddi et al[61] was first analysed with Seurat[16] (v3.1.5) to obtain clusters and embedding coordinates, then the results were combined with the loom files containing the spliced and unspliced counts into a new anndata object, through the dedicated *build_adata_object* function of FIERCE.

The raw UMI count data of the four samples were downloaded from Gene Expression Omnibus (GEO), and the corresponding fastq files were downloaded from the Sequence Read Archive (SRA). **Table 4** reports the GEO and SRA accession codes of all samples.

We first analysed the four raw count matrices with Seurat. Based on visual inspection of several quality metrics, we applied a series of filters on both the genes and the cells of all samples, reported in **Table 4** alongside the number of cells of each sample before and after the filtering procedure. For doublet score computation, we used the dedicated Scrublet[66] python package (v0.2.1). After quality filters, we log-normalized the data with a scale factor of 10^4, and used the dedicated function of Seurat to compute the cell cycle scores of each cell. We computed the highly variable genes of each sample with the default *vst* procedure of Seurat, that, similarly to the default procedure of Scanpy, first subdivides genes into bins based on their mean expression, and then computes a separate dispersion within each bin. We applied the default thresholds for mean expression and dispersion, i.e., a lower mean cut-off of 0.1, an upper mean cut-off of 8, and a lower dispersion cut-off of 1. We scaled the 2,000 identified highly variable genes of each sample to 0 mean and unit variance, and then performed a PCA and a clustering analysis (resolution 0.4) on the first 10 principal components for all samples. We annotated the resulting clusters with the scMCA[67–69] R package (v0.2.0), an automatic annotation tool that correlates the expression of the genes of each query cell to the expression of the same genes in the various cell types of the Mouse Cell Atlas[67–69] (MCA) reference database (http://bis.zju.edu.cn/MCA/). We finally filtered each sample to retain only the cells that were assigned to mammary epithelial cell types. **Table 5** and **Table 6** show, respectively, the sampling age and cell type subdivision of the remaining 5,168 total cells.

Sample	GEO accession	SRA accession	Pre-QC number of cells	QC filters genes	QC filters cells	Post-QC number of cells
EMBRYONIC DAY 16	GSM3022283	SRX3742051	821	▪ expressed in >0.1% of cells	▪ expressed genes >1,000 ▪ mitochondrial fraction <10% ▪ 2,000< UMI counts <75,000 ▪ doublet score <0.4	731
EMBRYONIC DAY 18	GSM3022284	SRX3742052	1,698	▪ expressed in >0.1% of cells	▪ expressed genes >1,000 ▪ mitochondrial fraction <10% ▪ 2,000< UMI counts <50,000 ▪ doublet score <0.4	1,274
POSTNATAL DAY 4	GSM3022285	SRX3742053	1,542	▪ expressed in >0.1% of cells	▪ expressed genes >1,000 ▪ mitochondrial fraction <10% ▪ 2,000< UMI counts <40,000 ▪ doublet score <0.4	1,408
ADULT	GSM3022286	SRX3742054	2,723	▪ expressed in >0.1% of cells	▪ expressed genes >500 ▪ mitochondrial fraction <10% ▪ 1,000< UMI counts <40,000 ▪ doublet score <0.4	2,256

Table 4: summary of the four samples of the mammary gland development dataset, with GEO and SRA accessions, quality filters, and number of cells before and after the filtering procedure.

Sampling age	Number of cells
EMBRYONIC DAY 16	698
EMBRYONIC DAY 18	1,115
POSTNATAL DAY 4	1,284
ADULT	2,071

Table 5: sampling ages of the cells included in the mammary gland development dataset.

Cell type	Number of cells
EMBRYONIC EPITHELIAL CELLS	1,305
LUMINAL PROGENITORS	536
ALVEOLAR PRECURSOR CELLS	1,149
LUMINAL CELLS	1,347
BASAL CELLS	831

Table 6: cell types included in the mammary gland development dataset.

To merge the four samples in a single dataset and simultaneously exclude any batch effect, we applied the default data integration procedure of Seurat v3.1.5[16]. Then, we scaled the 2,000 shared highly variable genes used for integration to 0 mean and unit variance, and we simultaneously regressed out the effect of cell cycle score. Finally, we performed another PCA on the integrated dataset, and computed a UMAP embedding on the first 18 principal components.

The cell annotations and embedding coordinates of the integrated dataset were successively transferred to the new anndata object, that was created as follows. We first used Cellranger[11] (v3.1.0) to align the reads contained in the fastq files of each single sample to the reference mouse genome. Then we used the *run10X* function of velocyto[54] (v0.17.17) to subdivide the UMI counts into spliced and unspliced, and to generate a loom file containing the resulting matrices for each sample. Finally, we used the *build_anndata_object* function of FIERCE to generate a new anndata object containing, in the dedicated layers, the spliced and unspliced counts of all the samples merged together; with the same function, we added to the new object the cell annotations and embedding coordinates contained in the integrated Seurat object. The first 18 principal components were successively used by the *compute_velocity* function to initiate the FIERCE analysis, through the construction of a new neighborhood graph (with 30 neighbors for each cell) for the computation of first order moments. The rest of the FIERCE analysis was executed according to the default pipeline (in particular, the built-in high confidence murine PPI network was used).

The vector field built by FIERCE was compared with both the corresponding vector field built by scVelo[55] (v0.2.4) from classic RNA velocity, and the principal graph reconstructed by Monocle 3[47] (v1.3.1). The scVelo analysis was directly executed during the pipeline of FIERCE, with the default procedure described above. In the Monocle 3 analysis, we used the dedicated functions of the tool

to pre-process the raw gene expression data, to perform a PCA, and to construct a UMAP embedding on 50 principal components. During the pre-processing we applied a batch correction procedure and performed a regression on cell cycle scores through the batchelor[70] R package (v1.10.0), one of Monocle's dependencies. Then we followed the default procedure to build the principal graph and to compute the pseudotime score of cells based on their projections on such graph. For the latter step, we used the *get_earliest_principal_node* helper function to manually set the embryonic epithelial cells as the root of the genealogy.

4. Results

4.1. Signaling entropy recapitulates the differentiation potency of single cells

We first tested the efficacy of signaling entropy as a measure to recapitulate the differentiation potency of single cells. The *plot_entropy_results* function of FIERCE allows the user to examine the distribution of the total entropy scores computed by the *compute_signaling_entropy* function across the categories of any previously defined cell annotation. To evaluate the accordance between entropy scores and expected developmental sequence, we examined the distribution of the entropy scores across the pre-defined cell type annotations of each dataset. Moreover, we examined the distribution of cells of different cell types across the potency states computed by FIERCE from the total entropy scores.

4.1.1. Pancreas endocrinogenesis

Figure 13 shows the distribution of total signaling entropy scores across the cell types of the pancreas dataset. The ductal cells, that are assumed to be the root of the genealogy, present the highest signaling entropy; subsequently, the signaling entropy starts decreasing in endocrine progenitors in a fashion that is inversely proportional to the expression of *Ngn3*, a transcription factor known to drive the commitment of progenitors to the endocrine fate[60]. Indeed, the entropy of early progenitors, characterized by low *Ngn3* expression, is lower than the entropy of ductal cells, but sensibly higher than the entropy of late progenitors, characterized by high *Ngn3* expression. When the progenitors start taking the route that will lead them to the transformation into the various endocrine cell types, these cells undergo a transcriptional “bottleneck” that involves the deactivation of all those gene pathways that will not be necessary for their future function. This results in the drastic entropy drop of pre-endocrine cells, that also present the widest range of entropy values. These cells clearly represent the pivotal section of the genealogy, linking the high-entropy progenitor populations to the fully differentiated alpha, beta, delta and epsilon endocrine cells, whose entropy distributions are centered at much lower levels.

As reported also by previous studies[38], the distribution of entropy values of most cells lay within a very narrow range, e.g., between 0.87 and 0.9. However, despite such reduced variance, the descending trend of signaling entropy across the cell types nicely recapitulates the developmental sequence of this biological system. Indeed, the ANOVA test performed by FIERCE confirmed the statistical significance (P-value < 1e-10) of the effect of cell type on the distribution of entropy values, with a sum of squares of 0.24.

Figure 14 shows the distribution of cells of different cell types across the four potency states identified by FIERCE in this dataset. As described in paragraph **3.1.1.3**, the potency states are numbered according to decreasing mean entropy score, meaning that potency state 1 is

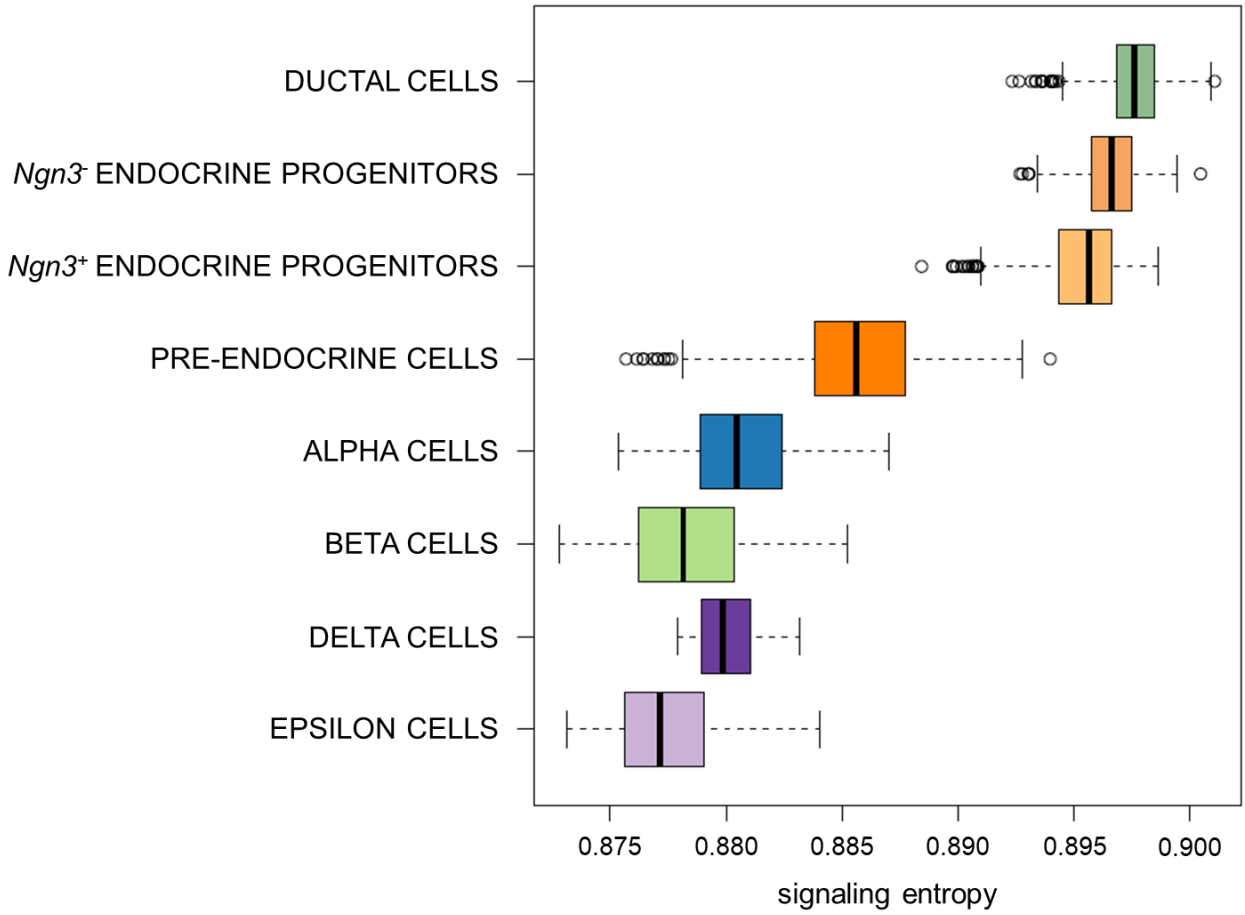

Figure 13: boxplot showing the distribution of total signaling entropy scores across the cell types of the pancreas endocrinogenesis dataset.

characterized by the highest mean entropy, while potency state 4 is characterized by the lowest. The distribution of cell types shows a very clear pattern that reflects the progressive decrease of signaling entropy across the dataset: the ductal cells, characterized by the highest entropy, are almost entirely localized in potency state 1, while conversely the various endocrine cells are distributed across potency states 3 and 4. Endocrine progenitors, albeit still largely localized in potency state 1, begin to progressively "fill" potency state 2, while the pre-endocrine cells, i.e., the pivotal cell state of the differentiation process, present an intermediate distribution, with the vast majority of cells localized in potency state 3 and two minor groups localized in potency states 2 and 4.

Overall, the pattern of signaling entropy and potency states is in full accordance with the expectance, i.e., a classic developmental scenario where the differentiation potency of cells progressively decreases as their transcriptomes become more and more focused on a specific function.

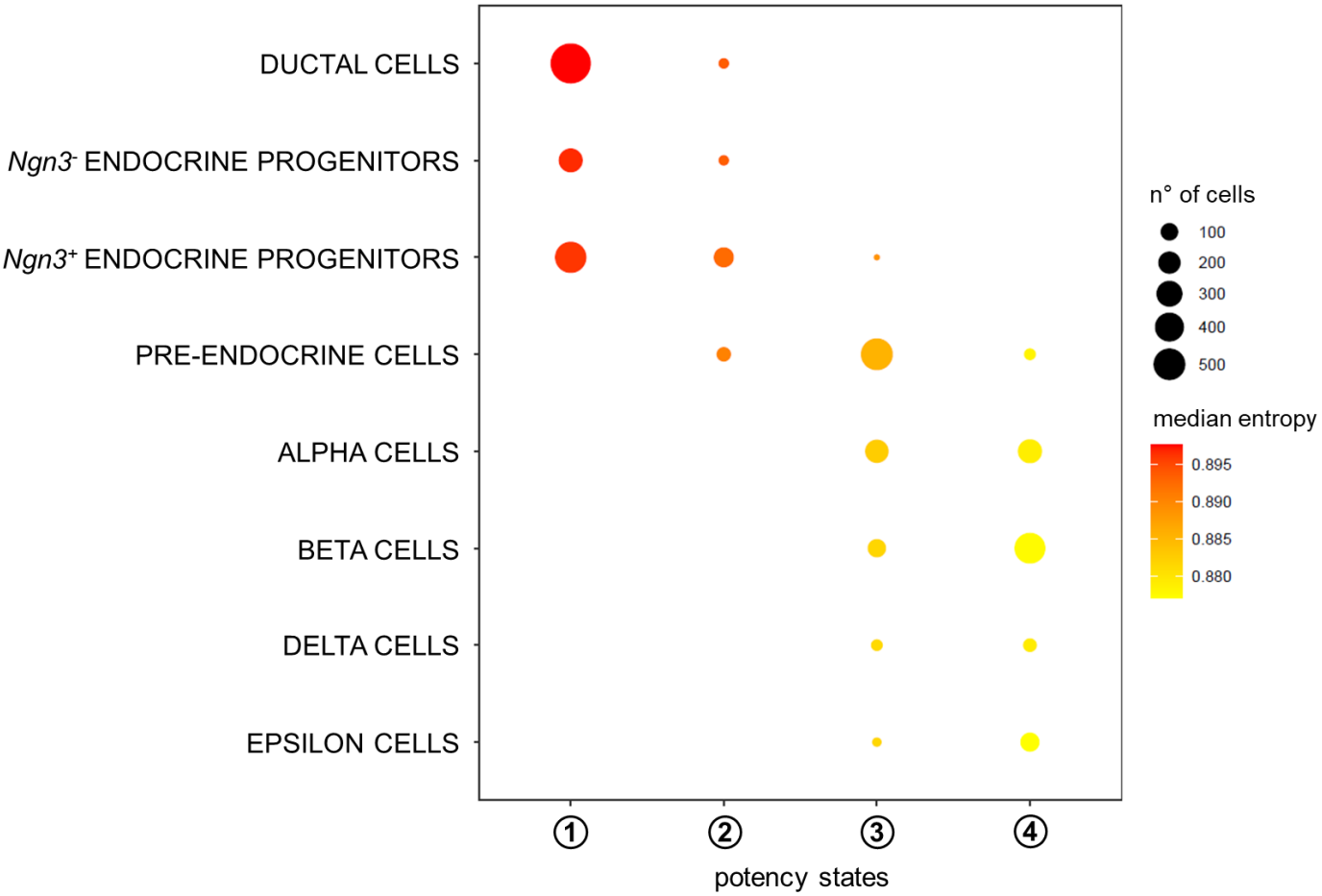

Figure 14: distribution of pancreatic cells across both the cell types and the four potency states identified by FIERCE. The size of the dots represents the number of cells, while their color represents their median signaling entropy.

4.1.2. Dentate gyrus neurogenesis

Figure 15 shows the distribution of total signaling entropy scores across the cell types of the dentate gyrus dataset. The overall pattern reveals a two-phases process where the signaling entropy first gradually increases, and then gradually decreases. In the ascending phase, signaling entropy increases from quiescent radial glia to cycling radial glia, keeps increasing in neuronal progenitors (nIPCs), and finally reaches its maximum in early neuroblasts. Afterwards, in the descending phase, the entropy starts decreasing in late neuroblasts, keeps decreasing in immature granule cells, and finally drops to its minimum in mature granule cells. Notably, the gap between immature and mature granule cells is very large, suggesting that the final step of this differentiation process implies a drastic redistribution of the gene expression signal on the PPI network of differentiating cells.

The unexpectedly low scores observed in radial glia and nIPCs, that are supposed to be the root of this differentiation process, find a possible explanation in the inherent heterogeneity of these pluripotent subpopulations, that are known[60] to include both cycling proliferative subgroups entirely dedicated to self-renewal and already committed subgroups that will differentiate into several different neuronal cell types. Indeed, the astrocytes lineage starts from radial glia, and the pyramidal cells descend from nIPCs[60]. All these more specialized subgroups are expected to progressively concentrate their expression signals into more and more specific branches of the PPI network, a process that, in turn, is expected to lead to a signaling entropy decrease. The overall distribution of

signaling entropy in early pluripotent subpopulations could have been dragged down by the presence of these subgroups. The subsequent drastic entropy drop that we observe from early neuroblasts to mature granule cells is explained by the drastic rearrangement of the distribution of the gene expression signal of cells as they gradually deactivate the proliferation-related pathways and fully commit to the neuronal fate.

The scenario described above implies a strong dependence of signaling entropy from cell type, that was fully confirmed by the ANOVA test performed by FIERCE, that yielded a highly significant (P-value < 1e-10) sum of squares of 0.29.

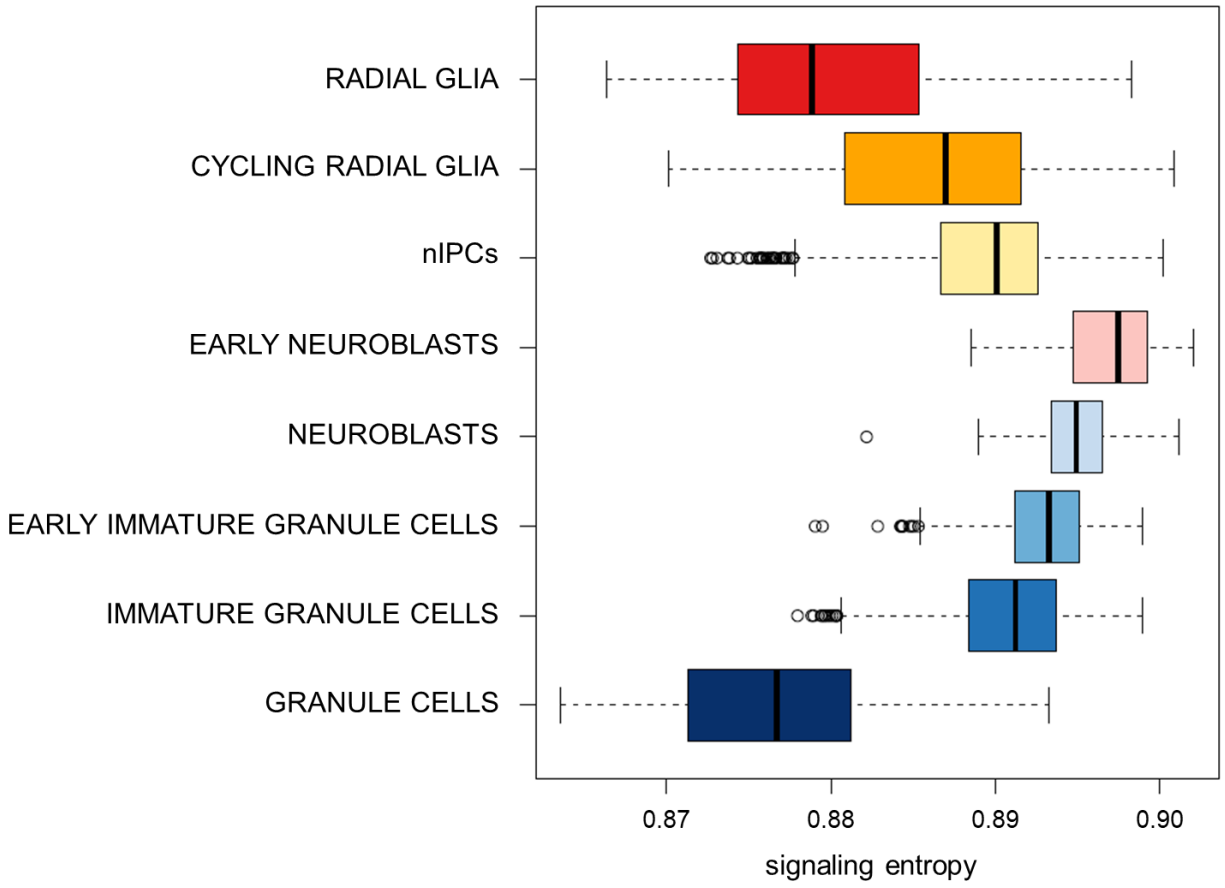

Figure 15: boxplot showing the distribution of total signaling entropy scores across the cell types of the dentate gyrus neurogenesis dataset.

Figure 16 shows the distribution of the cells of different cell types across the three potency states identified by FIERCE for this dataset. In accordance with the distribution of entropy scores, neuroblasts are mainly localized in potency state 1, characterized by the highest mean entropy, while their direct descendants, the immature granule cells, are fairly shared by potency state 1 and potency state 2, the latter being characterized by an intermediate mean entropy. Predictably, mature granule cells are almost entirely localized in potency state 3, characterized by the lowest entropy, thus confirming their fully differentiated state. Interestingly, nIPCs and radial glia cells are fairly distributed across all three potency states. This confirms the high heterogeneity of these pluripotent subpopulations, that include subgroups of cells that are either specializing into the proliferative function, or initiating other differentiation processes. Conversely, the progressive redistribution of cells into potency states 2 and 3 along the sequence of neuroblasts, immature and mature granule

cells confirms that this portion of the genealogy fully fits the characteristics of classic developmental processes, characterized by a progressive loss of differentiation potency along the cell lineage.

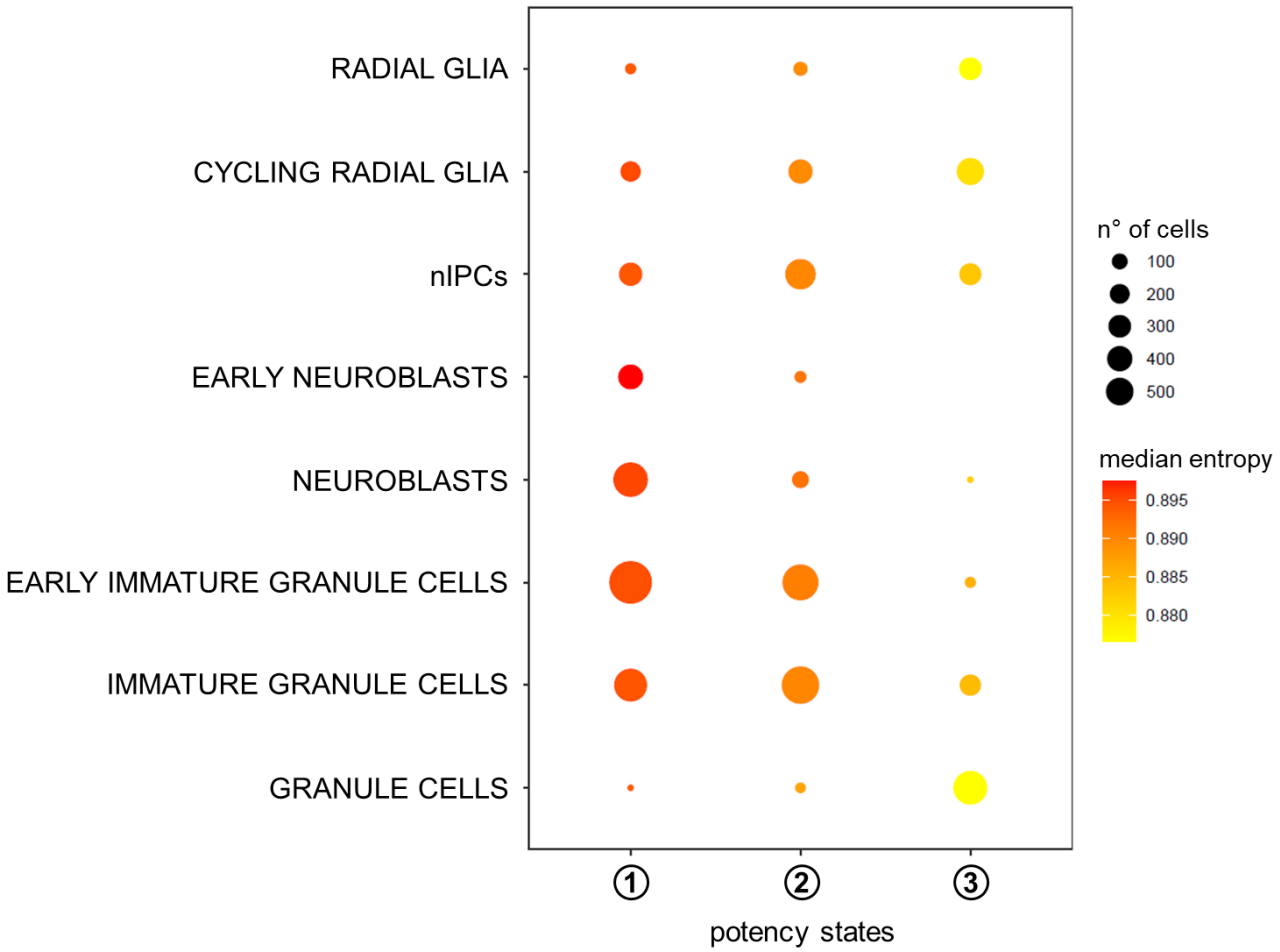

Figure 16: distribution of dentate gyrus cells across both the cell types and the three potency states identified by FIERCE. The size of the dots represents the number of cells, while their color represents their median signaling entropy.

4.1.3. Mammary gland development

Figure 17 shows the distribution of total signaling entropy scores across the cell types of the mammary gland development dataset. Compared to the previous datasets, the gap between subsequent entropy distributions appears much smaller. Nonetheless, a noticeable decreasing pattern can still be observed from embryonic epithelial cells, that are characterized by the highest entropy as expected, to the two fully differentiated cell types, i.e., luminal and basal cells, that are characterized by the lowest entropy. The luminal progenitors and the alveolar precursors are localized midway along the sequence, thus confirming their intermediate state. As observed for the pancreas dataset, this scenario is indicative of a classic developmental process, with the differentiation potency of cells progressively decreasing from multipotent to differentiated states. The reduced difference between the entropy distributions of cell types could be possibly explained by the reduced length of the process itself, that involves just a few developmental steps that are accompanied by a redistribution of the gene expression signal on the PPI network of cells.

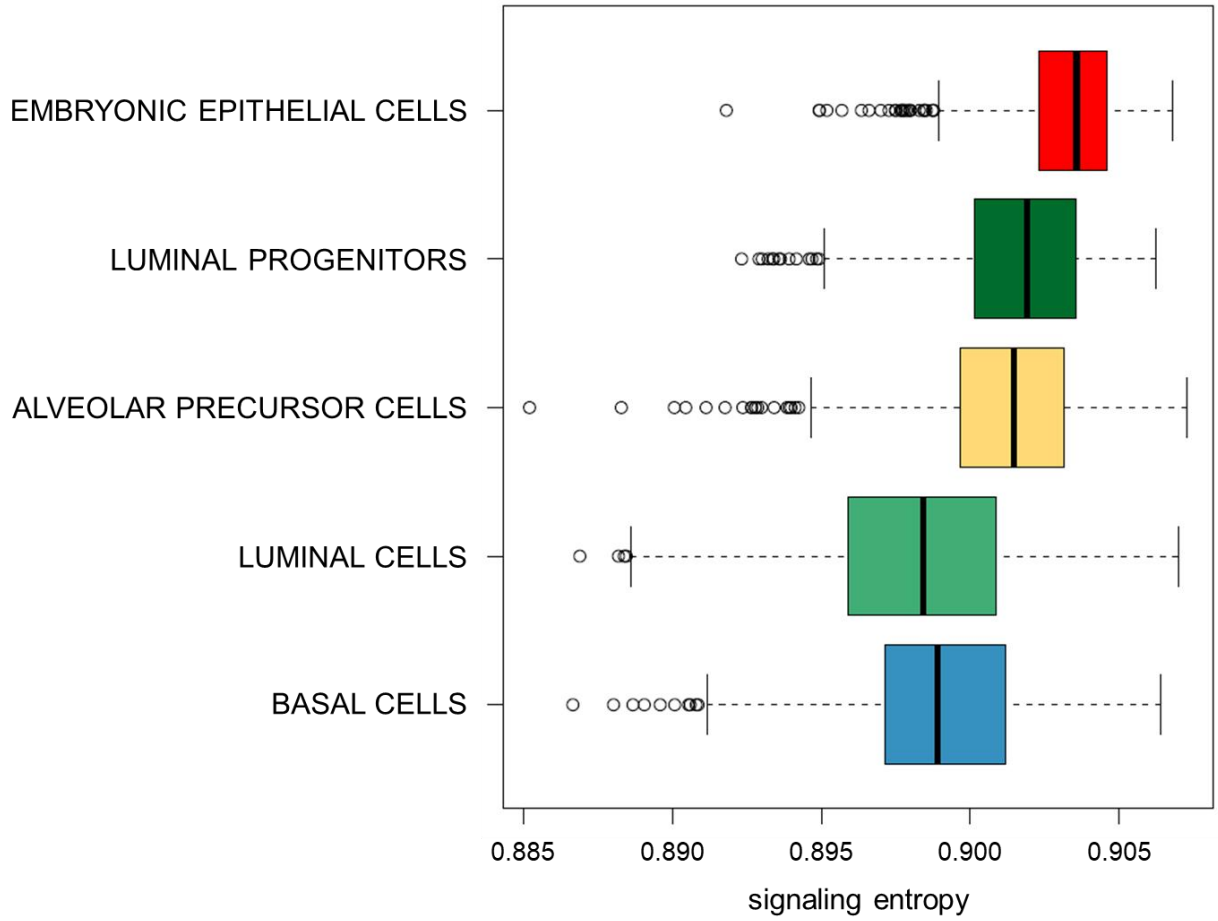

Figure 17: boxplot showing the distribution of total signaling entropy scores across the cell types of the mammary gland development dataset.

The conclusion described above is supported by the distribution of the cells of different cell types across the three potency states identified by FIERCE, shown in **Figure 18**. While embryonic epithelial cells are mainly localized in potency state 1 (characterized by the highest entropy), and luminal and basal cells are mainly localized in potency state 3 (characterized by the lowest entropy), luminal progenitors and alveolar precursors span all three potency states, in full accordance with the distribution of entropy scores.

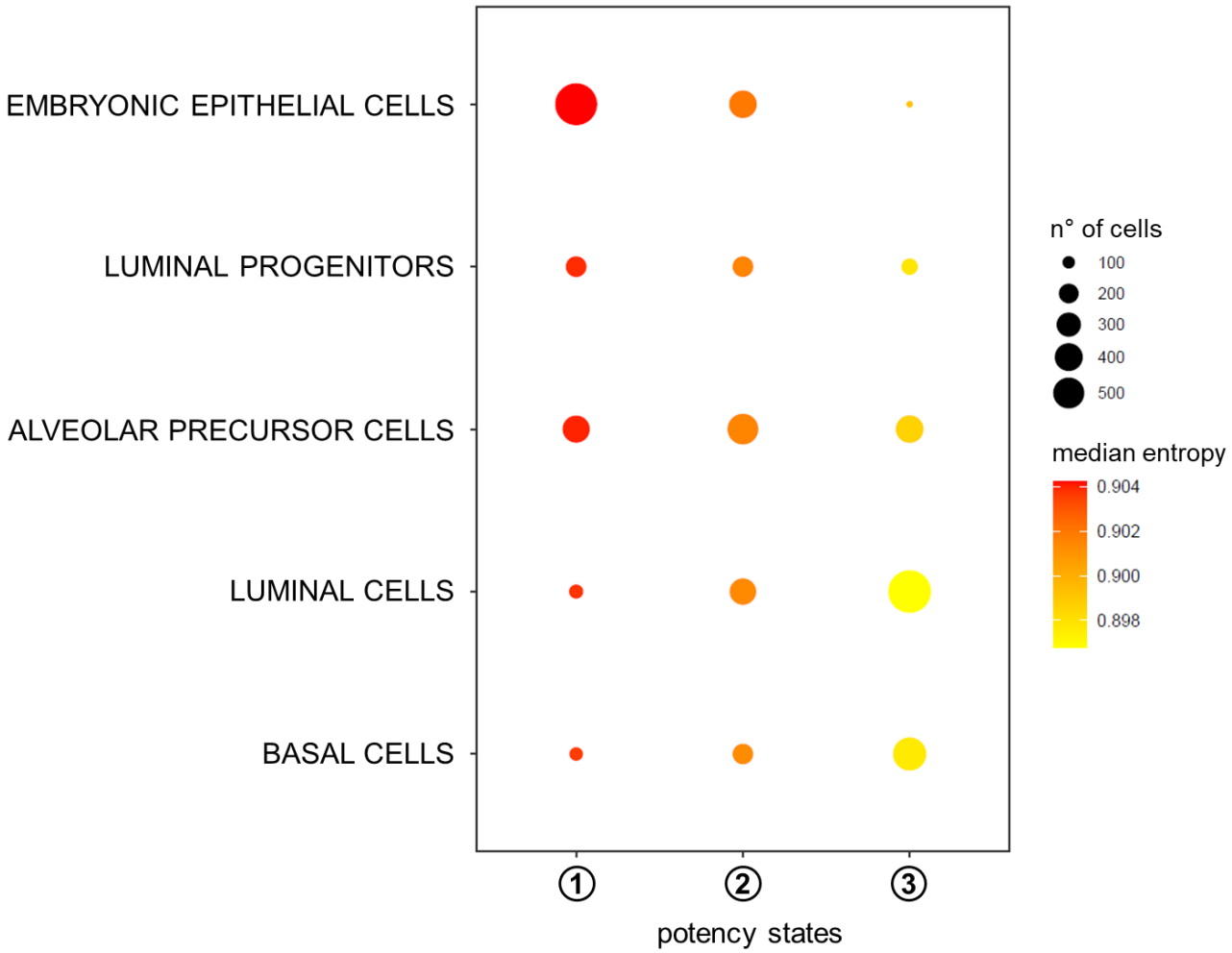

Figure 18: distribution of mammary gland cells across both the cell types and the three potency states identified by FIERCE. The size of the dots represents the number of cells, while their color represents their median signaling entropy.

A similar entropy gradient can also be observed in the distribution of total signaling entropy scores across the sampling ages of the dataset, shown in **Figure 19**, as well as in the distribution of the cells of these ages across the three potency states, shown in **Figure 20**. Although, once again, the distributions of entropy scores are not very different from one another, a decreasing entropy gradient still emerges from older to younger sampling ages, in accordance with the intuitive notion that the percentage of fully differentiated cells tends to increase during the development of tissues and organs.

Despite the reduced entropy differences, the effect of both cell type and sampling age on the distribution of entropy scores are deemed as significant by the ANOVA test performed by FIERCE (P-value < 1e-10), with a sum of squares of 0.02 and 0.01, respectively.

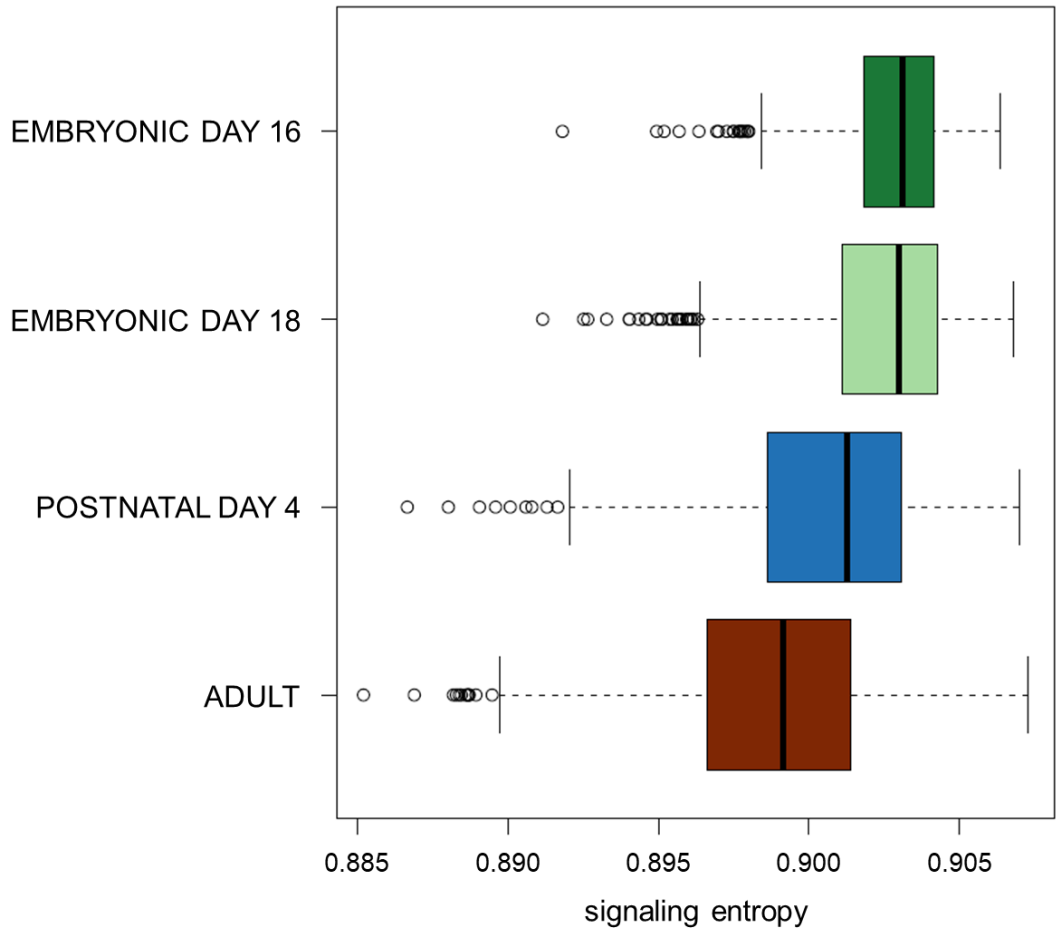

Figure 19: boxplot showing the distribution of total signaling entropy scores across the sampling ages of the mammary gland development dataset.

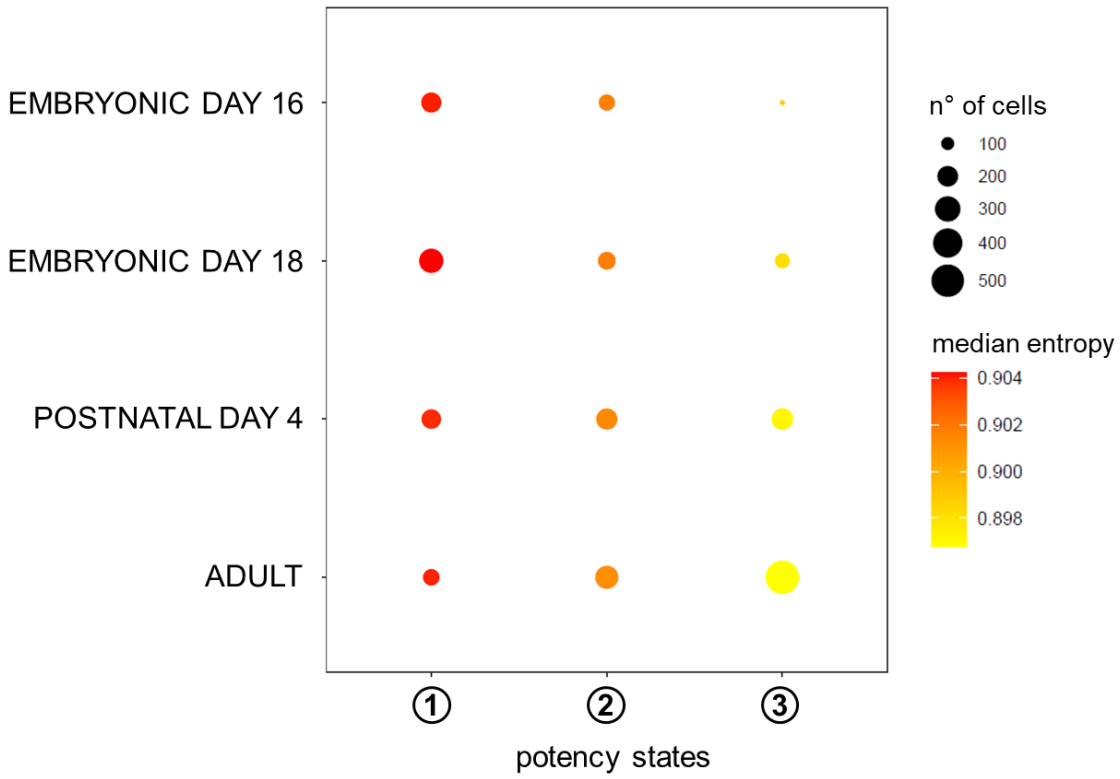

Figure 20: distribution of mammary gland cells across both the sampling ages and the three potency states identified by FIERCE. The size of the dots represents the number of cells, while their color represents their median signaling entropy.

4.2. FIERCE reconstructs the topology of differentiation processes

After investigating the efficacy of signaling entropy as a measure of the differentiation potency of single cells, we tested the capacity of FIERCE to reconstruct reliable representations of cell differentiation landscapes from the gene-wise local entropy scores of cells. Precisely, we used the *compute_entropy_UMAP* function of FIERCE to build UMAP embeddings from the local signaling entropy matrix of each dataset, and then we compared the resulting topologies with the corresponding UMAP embeddings built from the gene expression matrices.

4.2.1. Pancreas endocrinogenesis

Since the construction of cell embeddings from entropy scores is a possibility that has not been explored yet, we decided to verify the stability of the results produced by the *compute_entropy_UMAP* function of FIERCE by testing different combinations of parameters. Specifically, since the UMAP embedding is based on the construction of a cell-cell neighborhood graph[32], we built, for each dataset, three different graphs based on three different values for the number of neighbors of each cell, i.e., 10, 30 and 100 neighbors. Moreover, we computed each of these neighborhood graphs three times, each time retaining a different number of principal components, i.e., 15, 30 and 50 components. **Figure 21** shows the nine resulting UMAP embeddings for the pancreas endocrinogenesis dataset.

The entropy-derived embeddings show a very high stability with respect to the change of both the number of neighbors and the number of retained principal components. The general topology of the reciprocal relationships between the various cell types is largely maintained throughout all the combinations. Given such robustness, for all downstream analyses of the pancreas dataset we decided to use all the computed principal components and a medium number of neighbors, thus we selected the embedding corresponding to 50 principal components and 30 neighbors.

Figure 22 shows a close-up of the chosen UMAP embedding (**a**), in comparison with the correspondent UMAP embedding constructed directly from gene expression data (**b**). The two embeddings share a very similar topology, that clearly recapitulates the well-known genealogy that leads to the differentiation of endocrine cells from their ductal precursors. In both cases the pivotal section of the genealogy (i.e., the differentiation of late *Ngn3*-expressing endocrine progenitors into fully committed pre-endocrine cells) is nicely represented by a prominent “bridge” that links the pluripotent ductal cells to the fully differentiated endocrine cells. Overall, the entropy-based embedding built by FIERCE is in very good agreement with the correspondent embedding built with traditional methods; this suggests that signaling entropy retains the discriminative power of the raw gene expression data it is computed from.

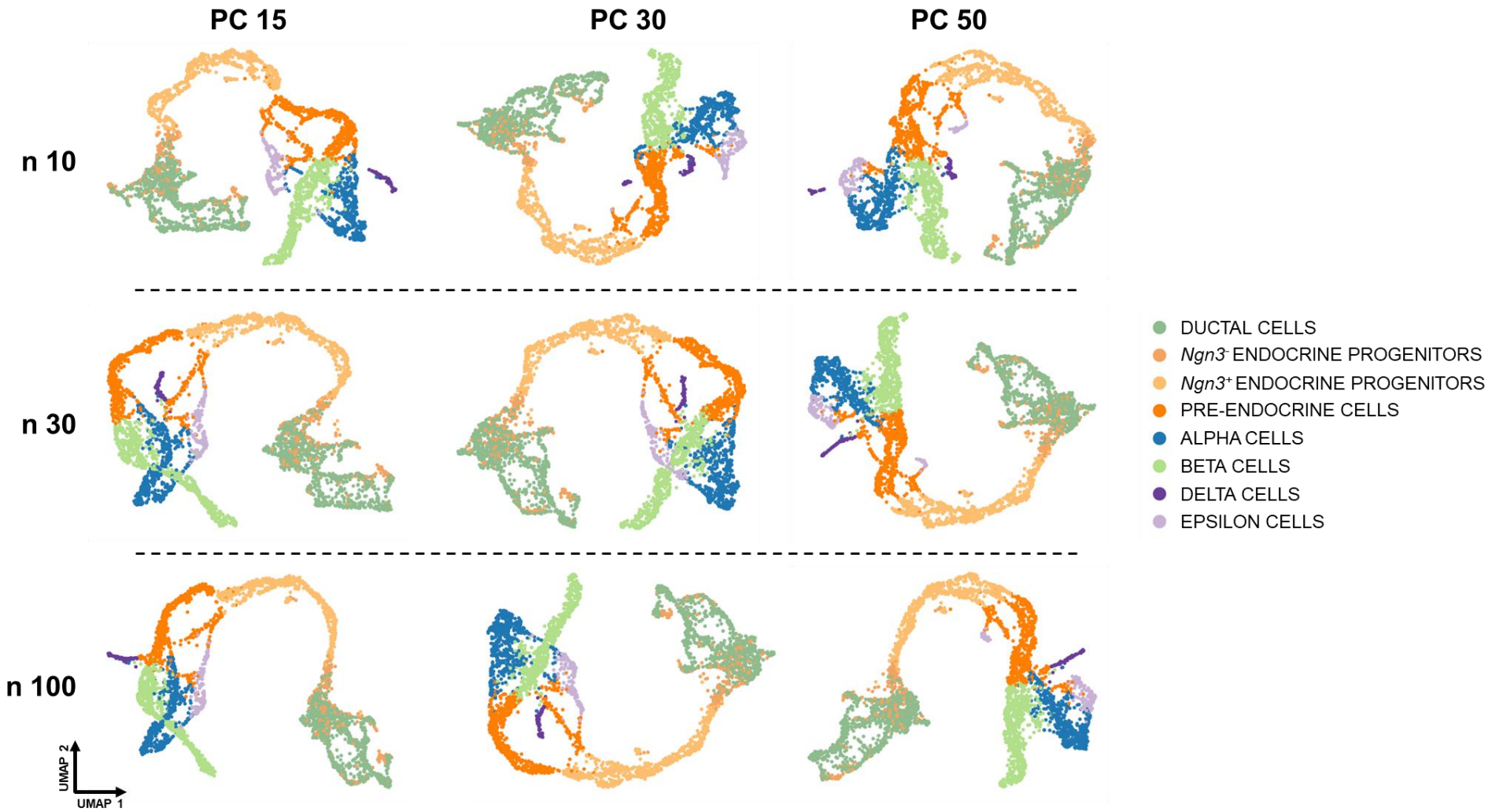

Figure 21: UMAP embeddings constructed by FIERCE from the local signaling entropy matrix of the pancreas endocrinogenesis dataset. The nearest neighbors graph has been computed with 10, 30 and 100 neighbors, and from 15, 30 and 50 retained principal components.

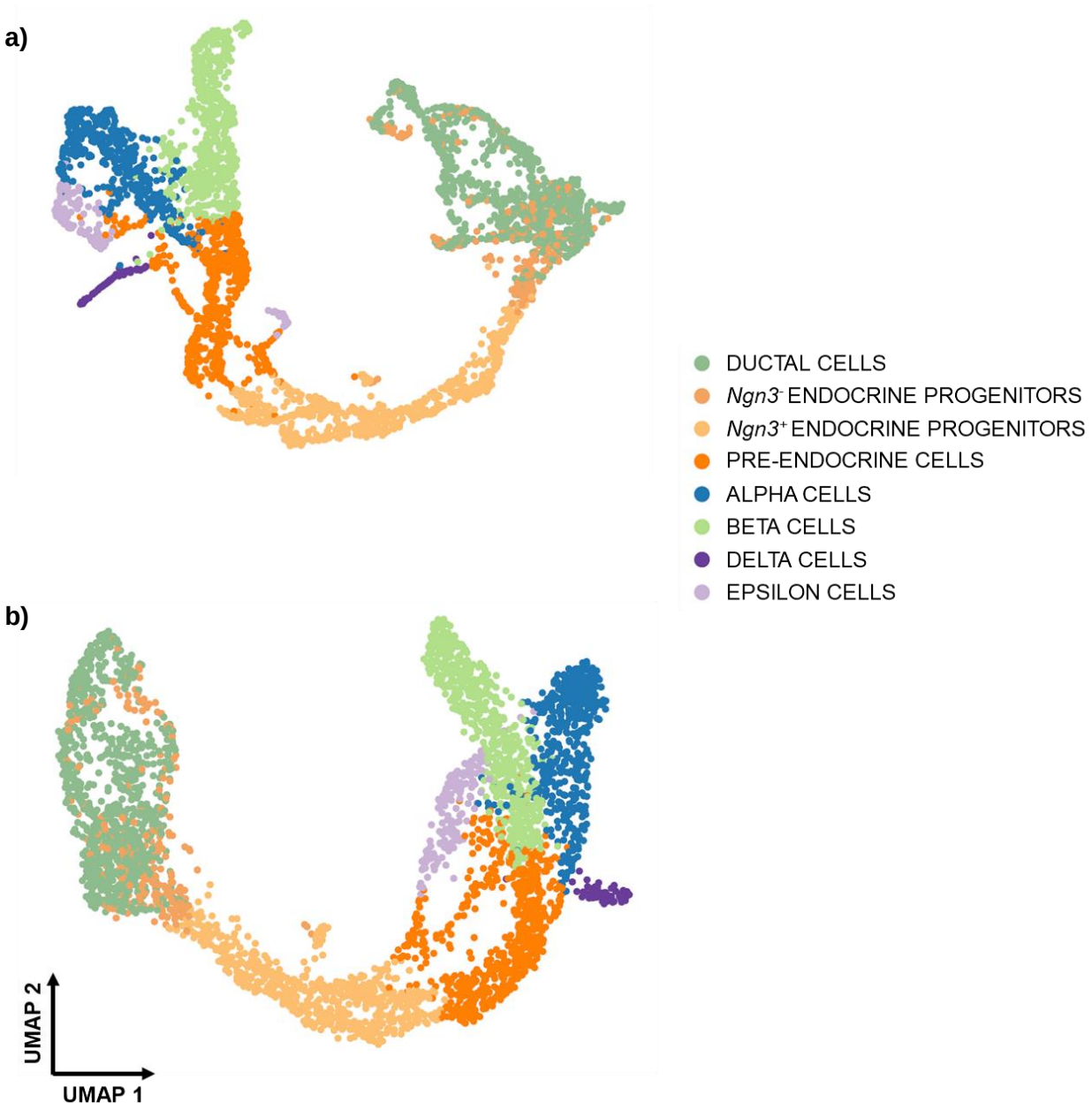

Figure 22: comparison between a) the UMAP embedding constructed from signaling entropy and b) the UMAP embedding constructed from gene expression for the pancreas endocrinogenesis dataset.

Figure 23 shows the entropy based UMAP embedding colored according to the cell types (**a**), the total signaling entropy scores of single cells (**b**), and the potency states identified by FIERCE (**c**). The signaling entropy of cells shows a very clear decreasing pattern along the differentiation process, in complete accordance with the distribution of the potency states. As shown before, precursor populations, i.e., the ductal cells and the endocrine progenitors, are characterized by a very high entropy score, and almost exclusively reside in potency state 1. Interestingly, the section of the embedding showing the passage from late *Ngn3*-expressing endocrine progenitors to committed pre-endocrine cells is characterized by a sudden drop in signaling entropy, and forms a separate potency state on its own, i.e., potency state 2. Afterwards, along the differentiation process, pre-endocrine cells dominate in potency state 3, where signaling entropy keeps constantly dropping. Finally, the various terminally differentiated endocrine cell types dominate in potency state 4, that is characterized by stable low entropy. This pattern represents a faithful depiction of the expected

differentiation potency landscape of the endocrine tissue of murine pancreas, thus fully confirming the efficiency of FIERCE in reconstructing the topology of dynamic processes directly from the signaling entropy values of single cells.

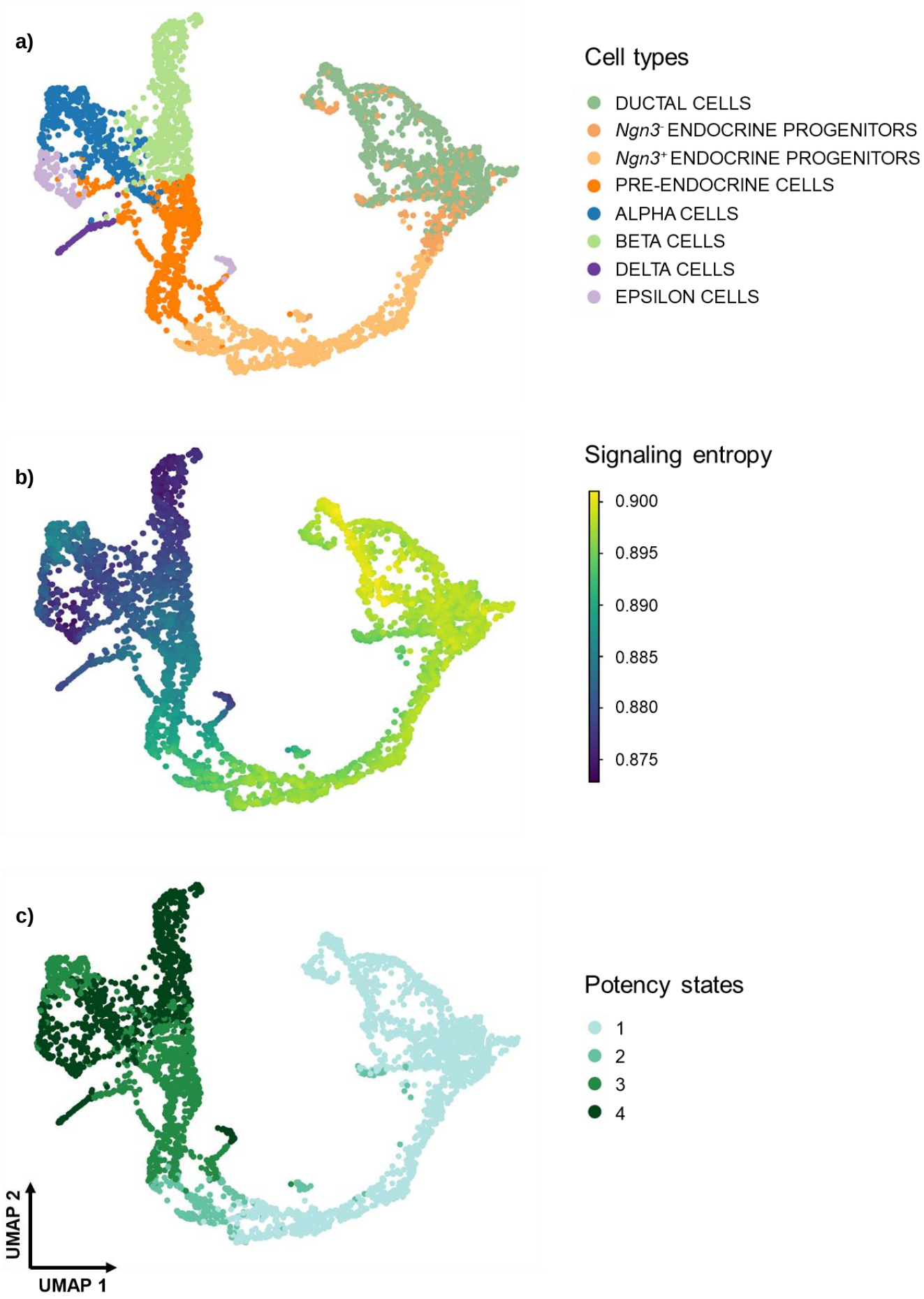

Figure 23: UMAP embedding of the pancreas endocrinogenesis dataset constructed from signaling entropy, colored according to a) cell types, b) signaling entropy score, and c) potency states.

4.2.2. Dentate gyrus neurogenesis

We tested on the dentate gyrus dataset the same combinations of parameters described above for the pancreas dataset, i.e., 10, 30 and 100 neighbors combined with 15, 30 and 50 retained principal components. **Figure 24** shows the nine resulting UMAP embeddings. Again, we observe a remarkable stability of the overall topology to the change of both parameters, thus confirming the robustness of signaling entropy. As previously done for the pancreas dataset, for all downstream analyses of this data we used the UMAP embedding that corresponds to 30 neighbors and 50 principal components. **Figure 25** shows a close-up of such embedding (**a**) in comparison with the UMAP constructed from gene expression data (**b**). As in the previous dataset, the two topologies are very similar, thus, once again, signaling entropy proved to be as efficient as raw gene expression in discriminating the mutual relationships between the cell subpopulations.

Figure 26 shows the entropy based UMAP embedding colored according to the cell types (**a**), the total signaling entropy scores (**b**), and the three potency states identified by FIERCE (**c**). The pattern described by the distribution of signaling entropy and of the potency states confirms the presence of low-entropy spots in radial glia and neuronal progenitors. These low-entropy spots correspond to sudden transitions from potency state 1 to potency state 2, or even to potency state 3. This is in agreement with the presence of different subgroups at various stages of differentiation within these heterogeneous subpopulations.

The maximum entropy peak corresponds to early neuroblasts, that are also the starting point of a long descending gradient that traverses late neuroblasts and immature granule cells, before finally ending in mature granule cells. This gradient is perfectly recapitulated by a gradual passage from potency state 1 to potency state 3, passing through potency state 2. This very clear entropy drop is explained by the progressive and exclusive specification of the neuronal fate along the genealogy of granule cells.

These results, confirm the ability of FIERCE to construct entropy-based embeddings that are as reliable as classic embeddings constructed from raw gene expression. Furthermore, the unexpected complexity unveiled by FIERCE within radial glia and neuronal progenitors suggests that “bottom-up” methods are capable of detecting interesting emerging patterns within the distribution of the differentiation potential of heterogeneous subpopulations.

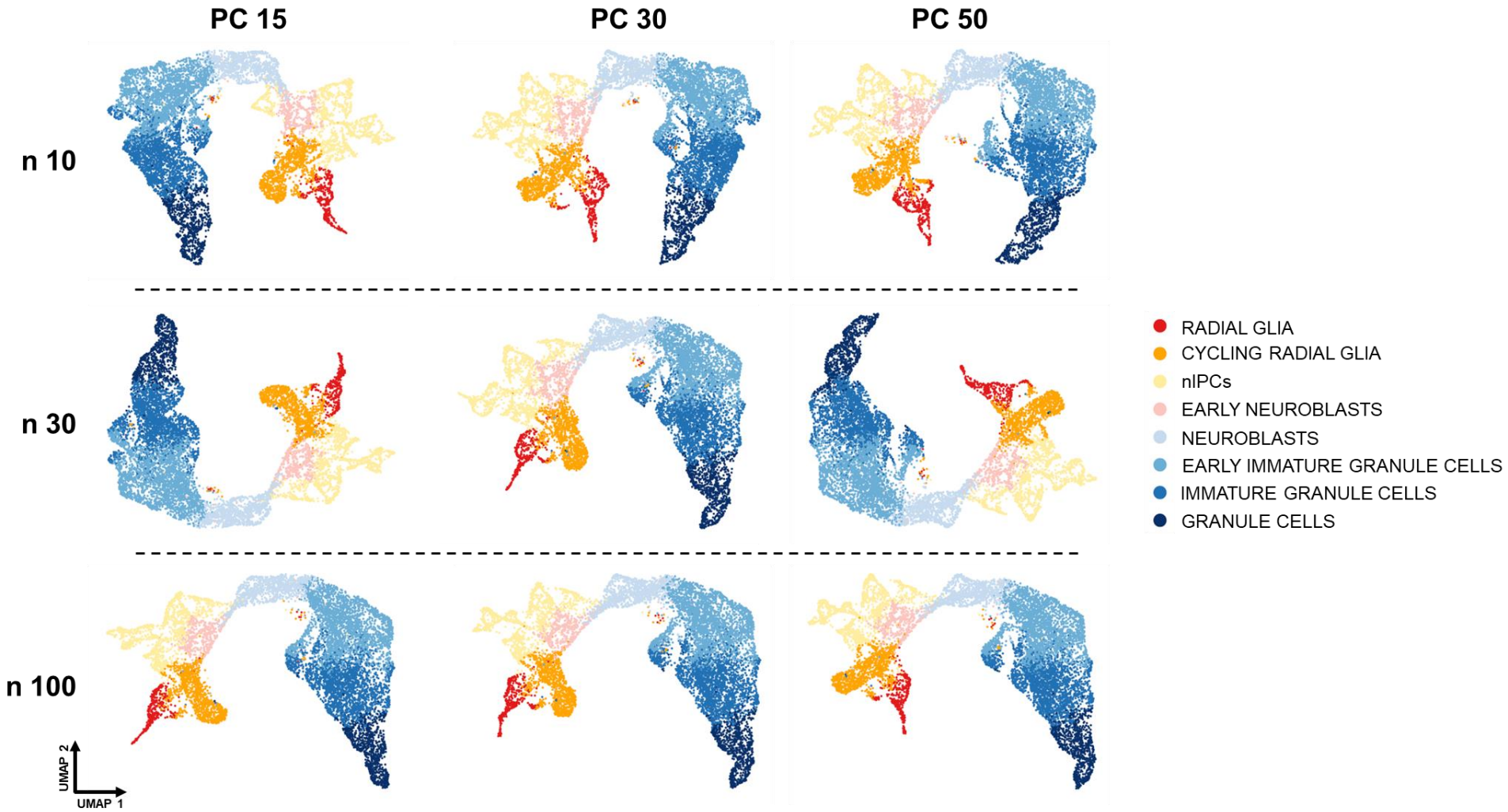

Figure 24: UMAP embeddings constructed by FIERCE from the local signaling entropy matrix of the dentate gyrus neurogenesis dataset. The nearest neighbors graph has been computed with 10, 30 and 100 neighbors, and from 15, 30 and 50 retained principal components.

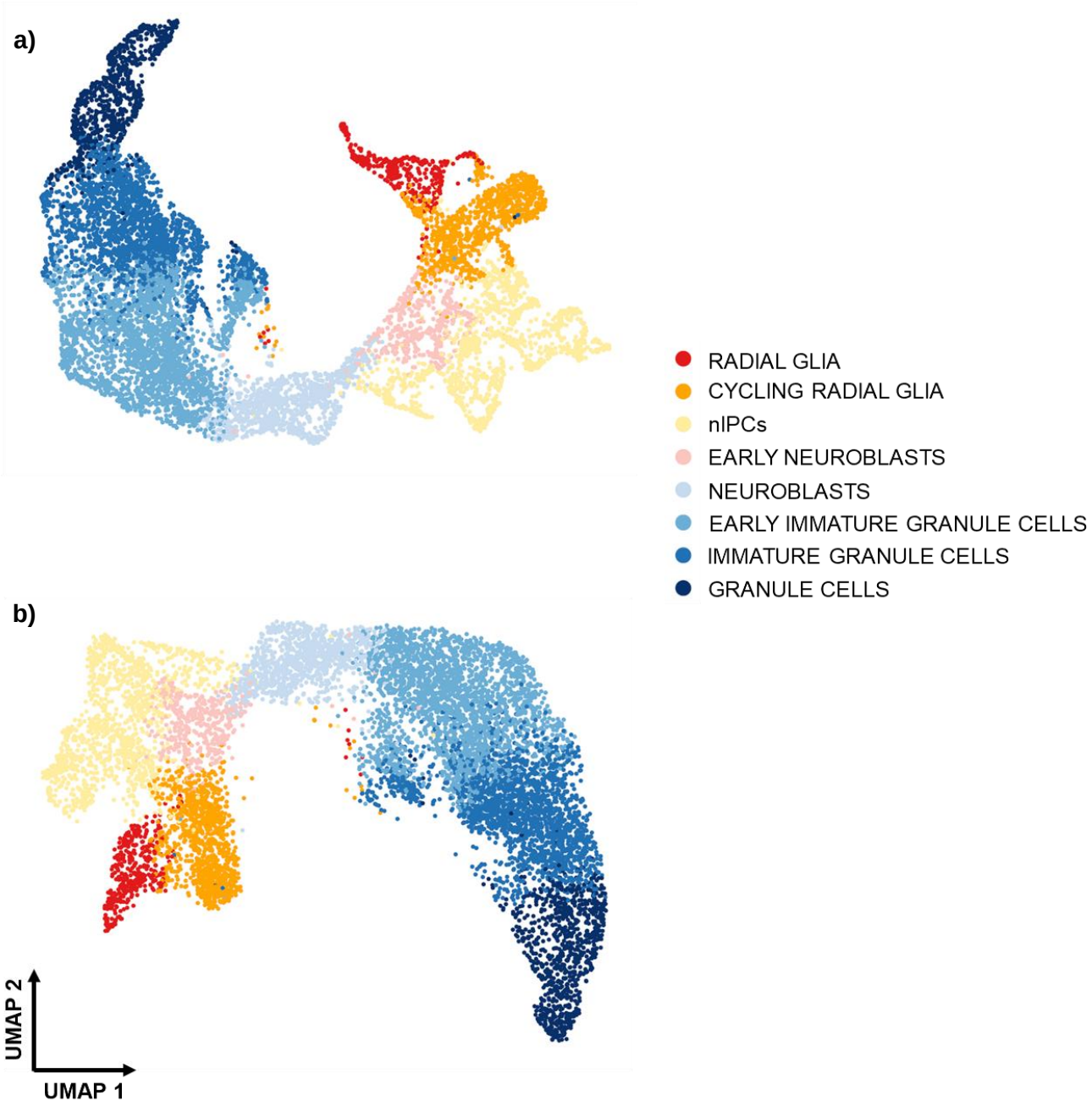

Figure 25: comparison between a) the UMAP embedding constructed from signaling entropy and b) the UMAP embedding constructed from gene expression for the dentate gyrus neurogenesis dataset.

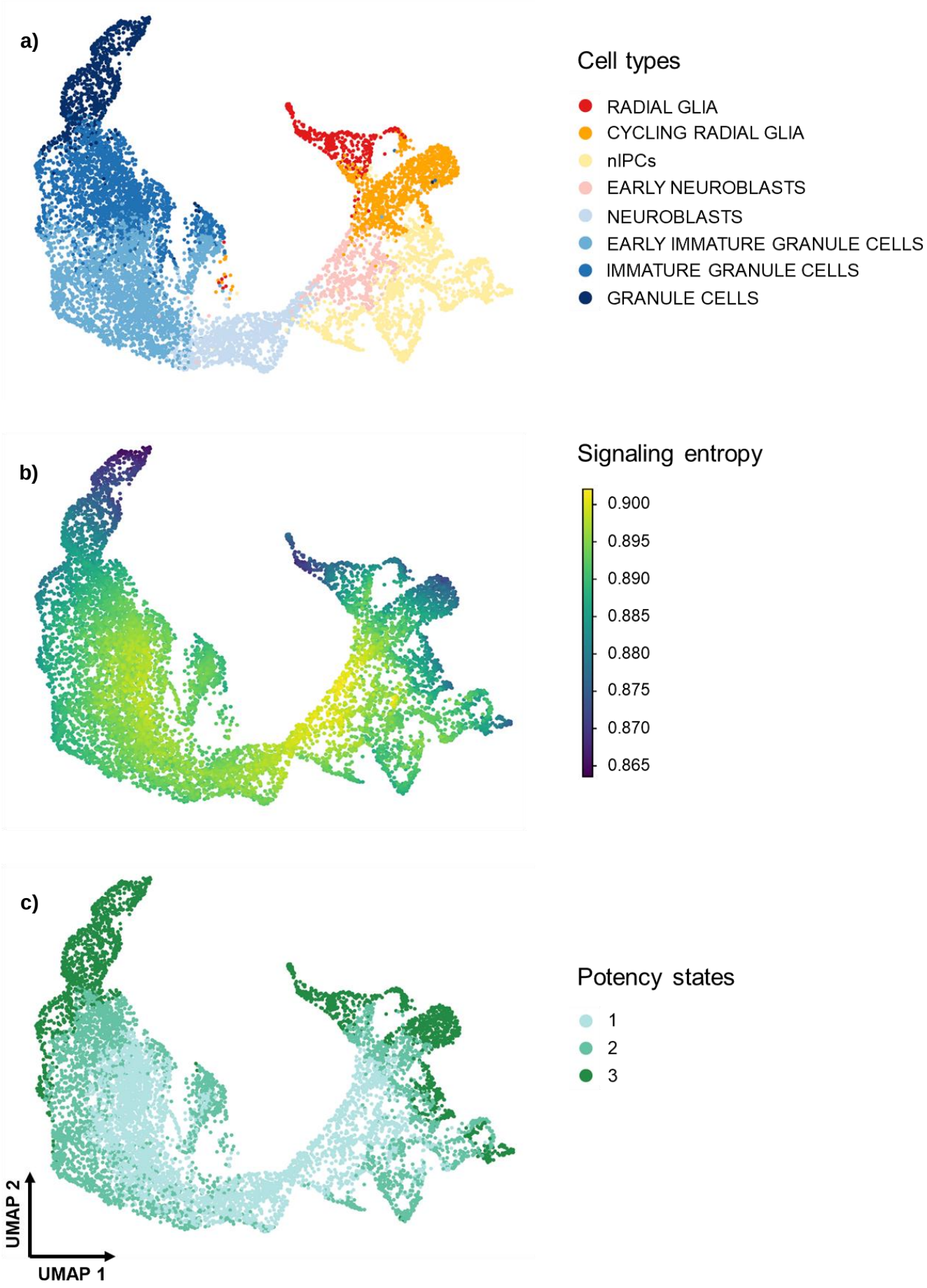

Figure 26: UMAP embedding of the dentate gyrus neurogenesis dataset constructed from signaling entropy, colored according to a) cell types, b) signaling entropy score, and c) potency states.

4.2.3. Mammary gland development

Figure 27 shows the nine embeddings built by FIERCE from the signaling entropies of the mammary gland dataset, according to the same combinations of neighbors and principal components that were tested on the other datasets. Again, both the topology of the embeddings and the mutual relationships between the cell types show a very high stability to the variation of both parameters. In this case, since the difference between the entropy distributions of the various subpopulations is particularly reduced, we chose, for all the downstream analyses of the mammary gland dataset, the UMAP embedding constructed from 100 neighbors and 50 principal components, to maximize the probability to capture the long-term dynamics of the potencies of cells.

Figure 28 shows in more detail the chosen UMAP embedding (**a**), compared with its equivalent constructed from gene expression data (**b**). In this case, the two embeddings appear very different from each other. The topology of the embedding constructed by FIERCE delineates a very clear tripartite structure that nicely recapitulates the known mammary gland development. As expected, from the embryonic epithelial cells depart two big branches: the first branch immediately leads to the basal cells, while the second undergoes a second bipartition in correspondence of the luminal progenitors, leading to alveolar precursors on one side and to luminal cells on the other. This layout is in good agreement with the distribution of the sampling ages of cells: the adult cells are concentrated on the three extremities of the structure, in correspondence of the most differentiated cell types, and are linked to embryonic cells through the postnatal cells, thus confirming the observation that, as development unfolds, the percentage of differentiated cells increases with the age of the organism. Indeed, postnatal cells are particularly enriched among luminal progenitors, in agreement with their role of link between the embryonic epithelial cells and the two end points of the luminal branch, i.e., alveolar precursors and luminal cells.

On the contrary, the embedding constructed from gene expression shows a more compacted structure that is hardly interpretable based on the current knowledge of this biological system. Basically, this embedding is subdivided into two well separated macro-clusters, one containing part of the embryonic epithelial cells, the luminal progenitors, the alveolar precursors and the luminal cells, and the other containing the remaining embryonic epithelial cells and the basal cells. The first macro-cluster corresponds to the branch of the entropy-based embedding that leads to the luminal fate, while the second corresponds to the branch that leads to the basal fate. However, the separation of the embryonic epithelial cells that are destined to the luminal fate from those that are destined to the basal fate is unplausible, given the finding of Giraddi et al[71] that embryonic cells do not show a clear subdivision into separate subclusters. Furthermore, within the luminal macro-cluster, the embryonic cells do not form a separate subpopulation, but rather appear intermingled with luminal progenitors and part of alveolar precursors into a mixed group of cells that connects the main nuclei of luminal cells and alveolar precursors. This confused disposition makes it very difficult to discern the mutual relationships between the cell subpopulations that constitute this macro-cluster, as well as to understand the relationship of the macro-cluster itself with the basal cells. Accordingly, the distribution of sampling ages looks equally noisy and difficult to interpret, with embryonic cells extending within the adult regions and postnatal cells spreading all over the whole embedding. Mouse mammary organogenesis is known to unfold according to a strictly reproducible scheme[61,72]: bipotent stem cells rapidly multiply during the embryonic development[61,73], and then, shortly after birth, they produce both the luminal and the basal compartments of the epithelium[61,73–76]. Given such background knowledge, the UMAP embedding constructed by FIERCE from signaling entropy appears much clearer than its equivalent constructed from gene expression, and above all much more faithful to the expected dynamic process. To support this conclusion, we next examined in detail the distribution of the differentiation potencies of cells.

Figure 29 shows the UMAP embedding of FIERCE colored according to the cell types (**a**), the sampling ages of cells (**b**), the total signaling entropy scores (**c**), and the three potency states identified in this dataset (**d**). The distribution of entropy scores shows the expected pattern of a typical developmental process: from the highest peak, localized on the embryonic epithelial cells, depart two descending gradients that slowly flow into three entropy “sinks”, localized on the basal cells, on the alveolar precursors, and on the luminal cells. Such gradients are reflected in the disposition of the potency states: while potency state 1 spans the embryonic epithelial cells and the luminal progenitors, potency state 3 is mainly localized in correspondence of the three entropy minima, at the extremities of the structure; potency state 2, as expected, is mainly distributed on the transition points, between potency states 1 and 3. Overall, FIERCE reconstructed a very clear and reliable representation of the differentiation landscape of this biological system, with well-defined progenitor and descendant subpopulations and continuous potency gradients linking the central peak to the peripheral valleys.

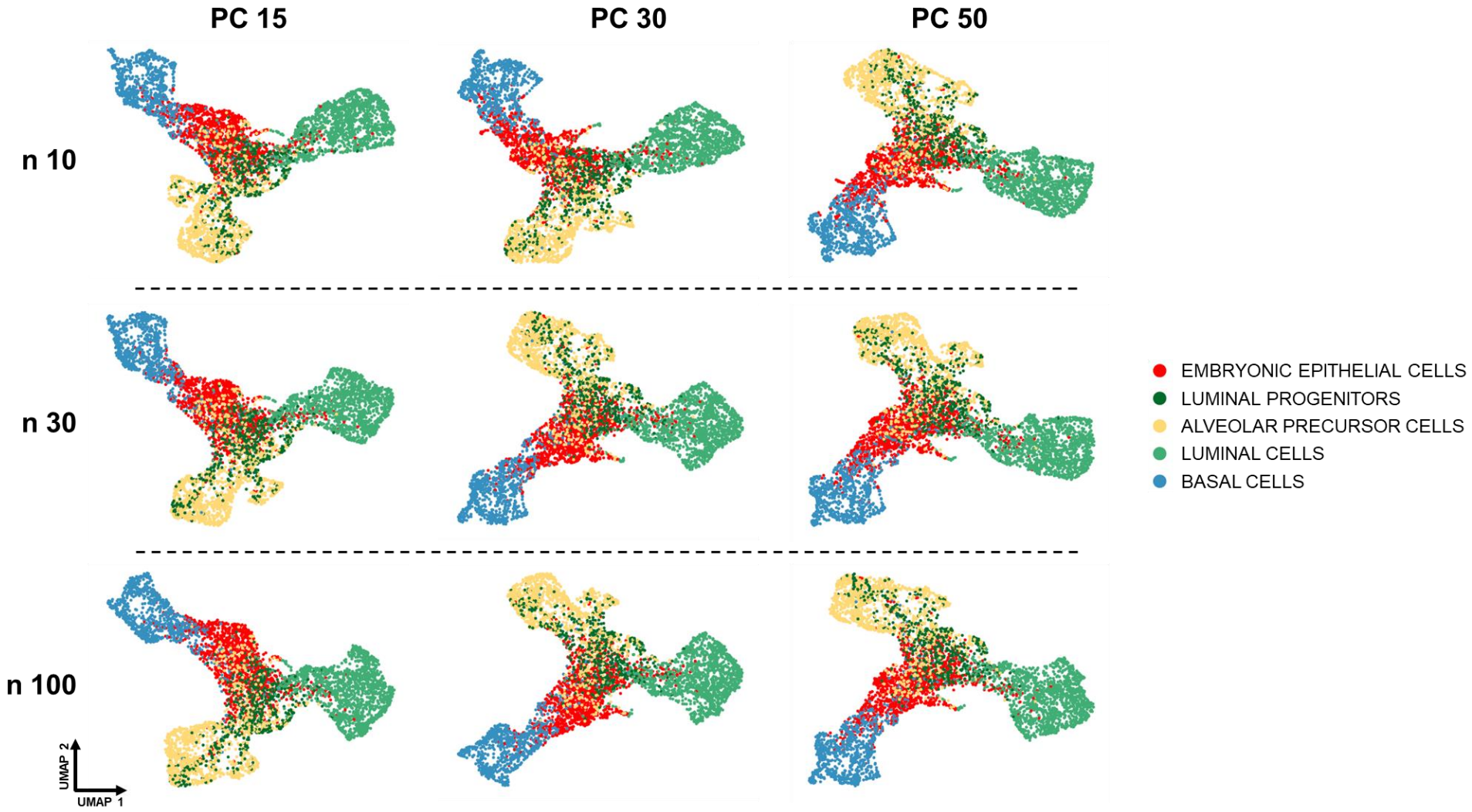

Figure 27: UMAP embeddings constructed by FIERCE from the local signaling entropy matrix of the mammary gland development dataset. The nearest neighbors graph has been computed with 10, 30 and 100 neighbors, and from 15, 30 and 50 retained principal components.

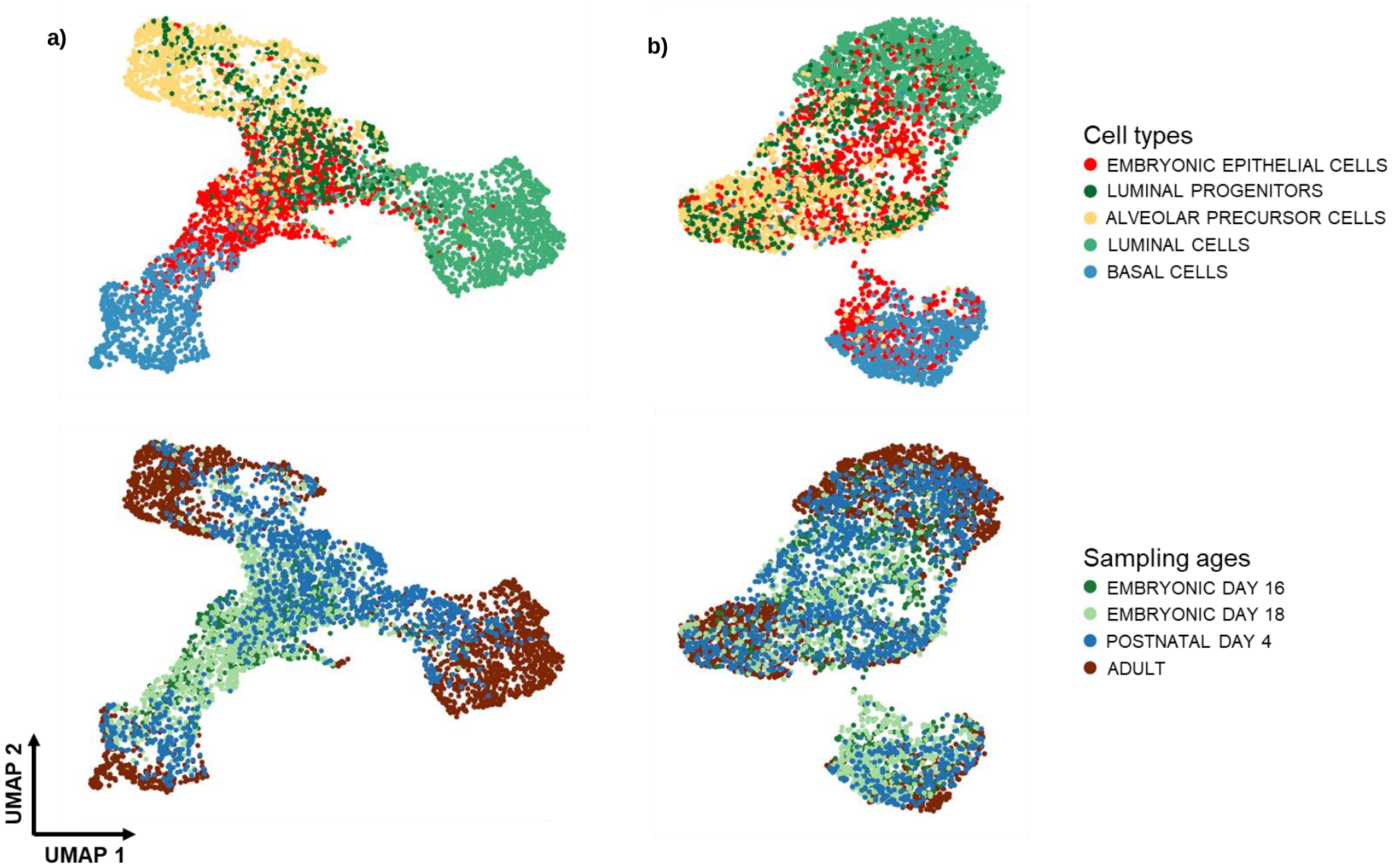

Figure 28: comparison between a) the UMAP embedding constructed from signaling entropy and b) the UMAP embedding constructed from gene expression for the mammary gland development dataset.

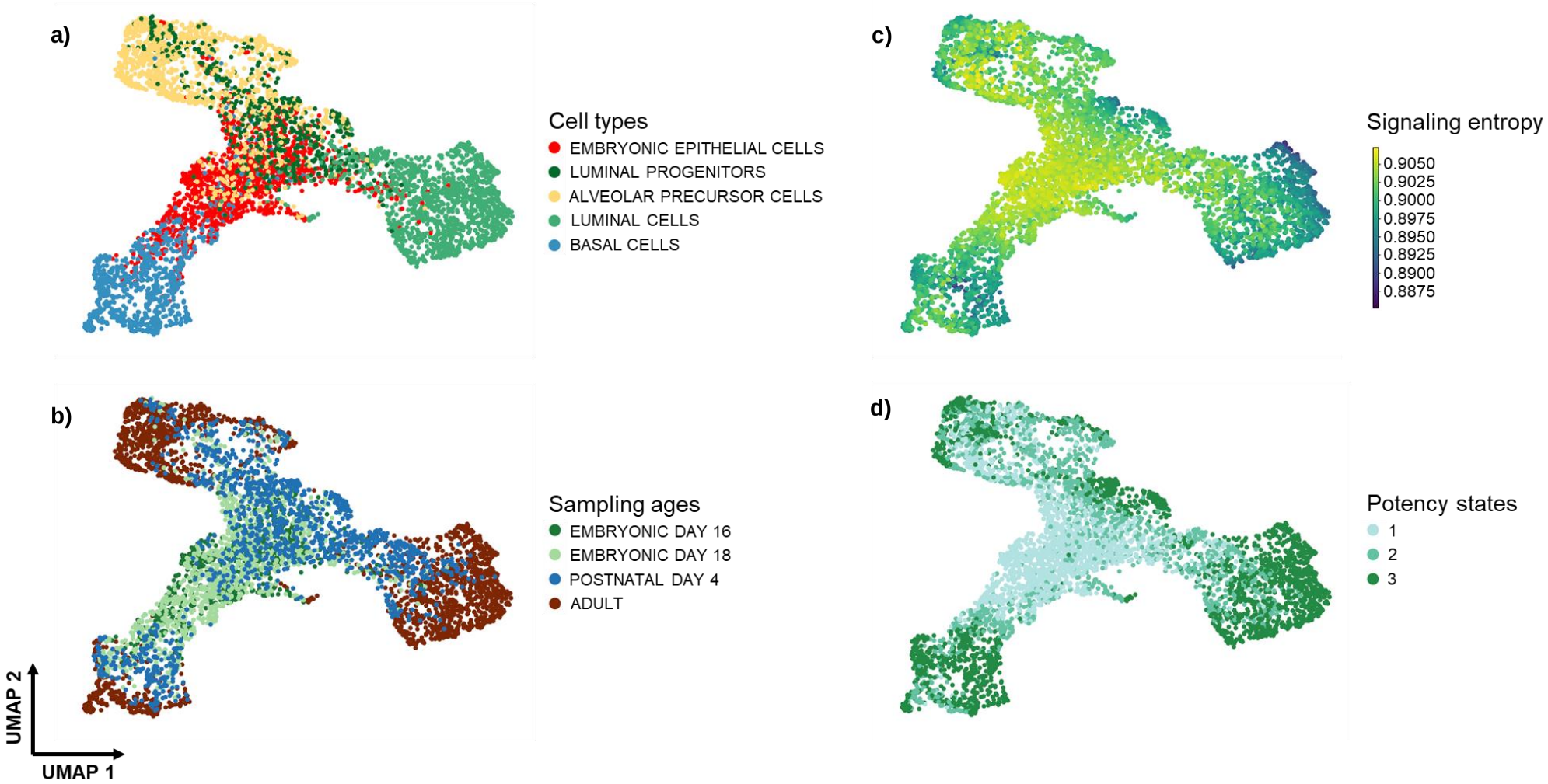

Figure 29: UMAP embedding of the mammary gland development dataset constructed from signaling entropy, colored according to a) cell types, b) sampling ages, c) total signaling entropy score, and d) potency states.

4.3. FIERCE reconstructs the dynamic of differentiation processes in the signaling entropy space

For each dataset, the UMAP embeddings derived from signaling entropy were used as “templates” by the *compute_graph_and_stream* function of FIERCE to draw the vector field of the velocity of the entropy, with the aim to reconstruct the dynamics of the cell population directly in the entropy space. Since both the topology of the UMAP template and the direction of the arrows entirely depend on the distribution of the gene expression signals in the various transcriptional programs of cells, these vector fields represent the most accurate depiction of how the differentiation landscape of the biological system evolves over time. To evaluate such depiction, we compared the streamplots produced by FIERCE to their equivalents built directly from raw gene expression data by scVelo[55], the most popular tool to construct vector fields from classic RNA velocity. Additionally, we compared the streamplots of FIERCE to the principal graphs constructed by Monocle 3[46,47], the most popular “top-down” method for trajectory inference from scRNA-seq data, and thus the best term of comparison for our opposite “bottom-up” strategy.

4.3.1. Pancreas endocrinogenesis

Figure 30 shows the velocity of the entropy vector field constructed by FIERCE on the entropy based UMAP embedding of the pancreas endocrinogenesis dataset, colored according to the cell types (**a**), the total signaling entropy scores (**b**), and the four potency states identified for this dataset (**c**). The dynamic process that governs the development of this biological system is perfectly delineated by a prominent stream of arrows that originates amongst the ductal cells, stretches along the pivotal section constituted by the endocrine progenitors and the pre-endocrine cells, and finally splits into the various endocrine cell types, at the end of the trajectory. These arrows perfectly follow the underlying decreasing entropy gradient, as well as the sequence of potency states. Such clear trend confirms that the pancreas system evolves according to the laws of classic developmental processes, characterized by a continuous decrease of the differentiation potency of cells as they proceed along the sequence of stable states.

Interestingly, the portion of the vector field corresponding to ductal cells, at the beginning of the trajectory, describes a divergent pattern: in correspondence of a marginal group of cells located in the bottom left of this area, two sets of arrows depart from one another, one pointing downwards, towards the main lineage leading to endocrine cells, and the other pointing upwards, in the opposite direction, initiating a second short trajectory that terminates in a small separated group of ductal cells. Such pattern suggests that only a subset of ductal cells is destined to undertake the endocrine fate, while the remaining cells maintain their pluripotent state.

This shorter secondary trajectory of ductal cells does not correspond to any particular trend of the total entropy scores, nor of the potency states. All ductal cells, regardless of their predicted fate, reside within potency state 1, and are characterized by an equally high entropy score. This demonstrates that, although the overall trend of the dynamic trajectory is already suggested by the mere distribution of signaling entropy scores, the construction of the vector field of the velocity of the entropy is essential to detect small-scale patterns that are indicative of different fates for cells belonging to the same cell type.

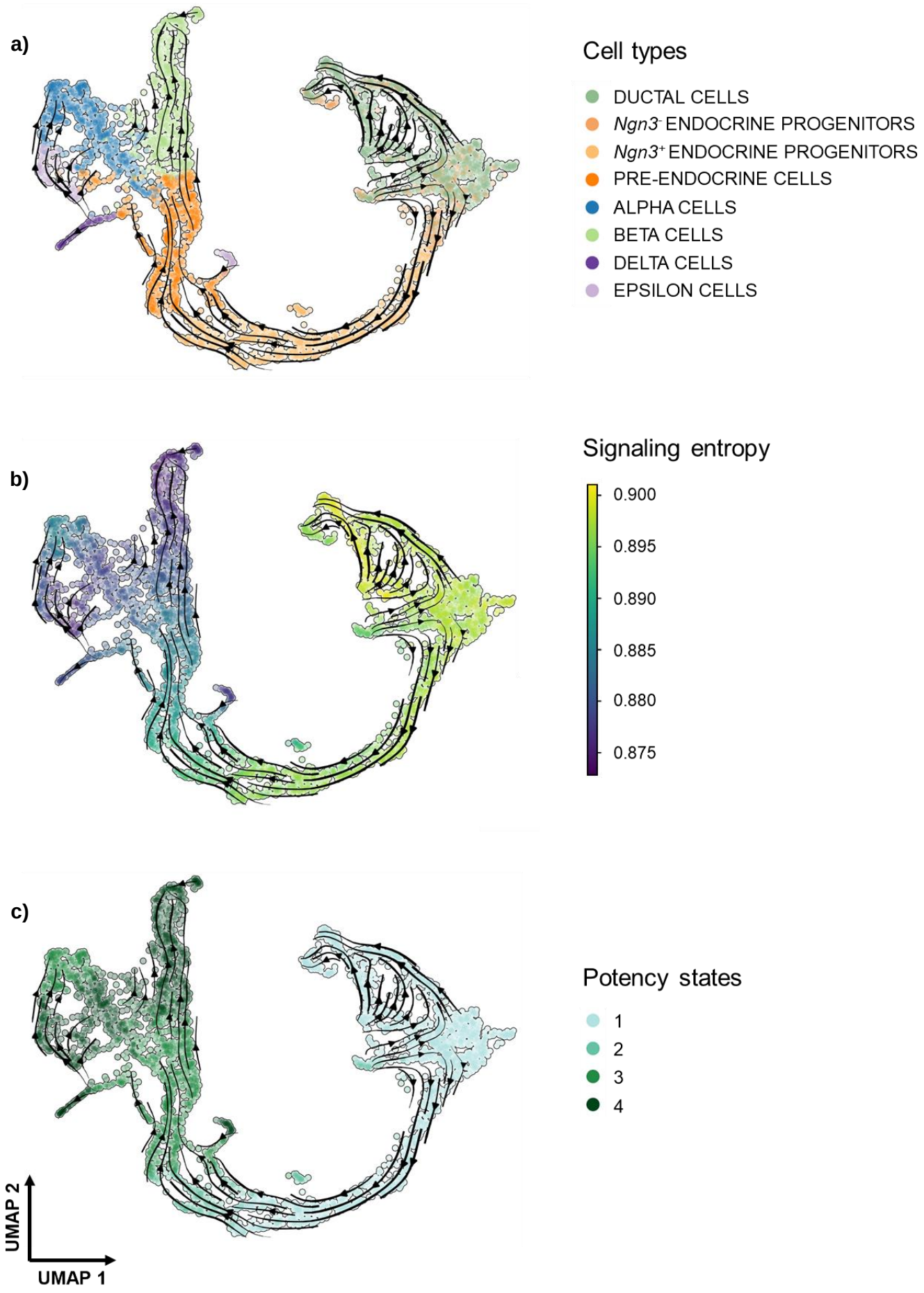

Figure 30: velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the pancreas endocrinogenesis dataset, colored according to a) cell types, b) total signaling entropy score, and c) potency states.

Figure 31 shows the comparison between the velocity of the entropy vector field of FIERCE (**a**) and the RNA velocity vector field constructed by scVelo on the UMAP embedding derived from raw gene expression (**b**). Both vector fields are similarly dominated by a prominent stream of arrows that delineates a clear trajectory from ductal cells to their endocrine descendants; however, the streamplot produced by scVelo describes a different pattern for the ductal cells, characterized by just one circular trajectory that originates in the middle of the group, expands outwards, and finally terminates in correspondence of the connection with *Ngn3*$^{-}$ endocrine progenitors. This difference between the two embeddings might suggest that signaling entropy may be particularly sensitive to

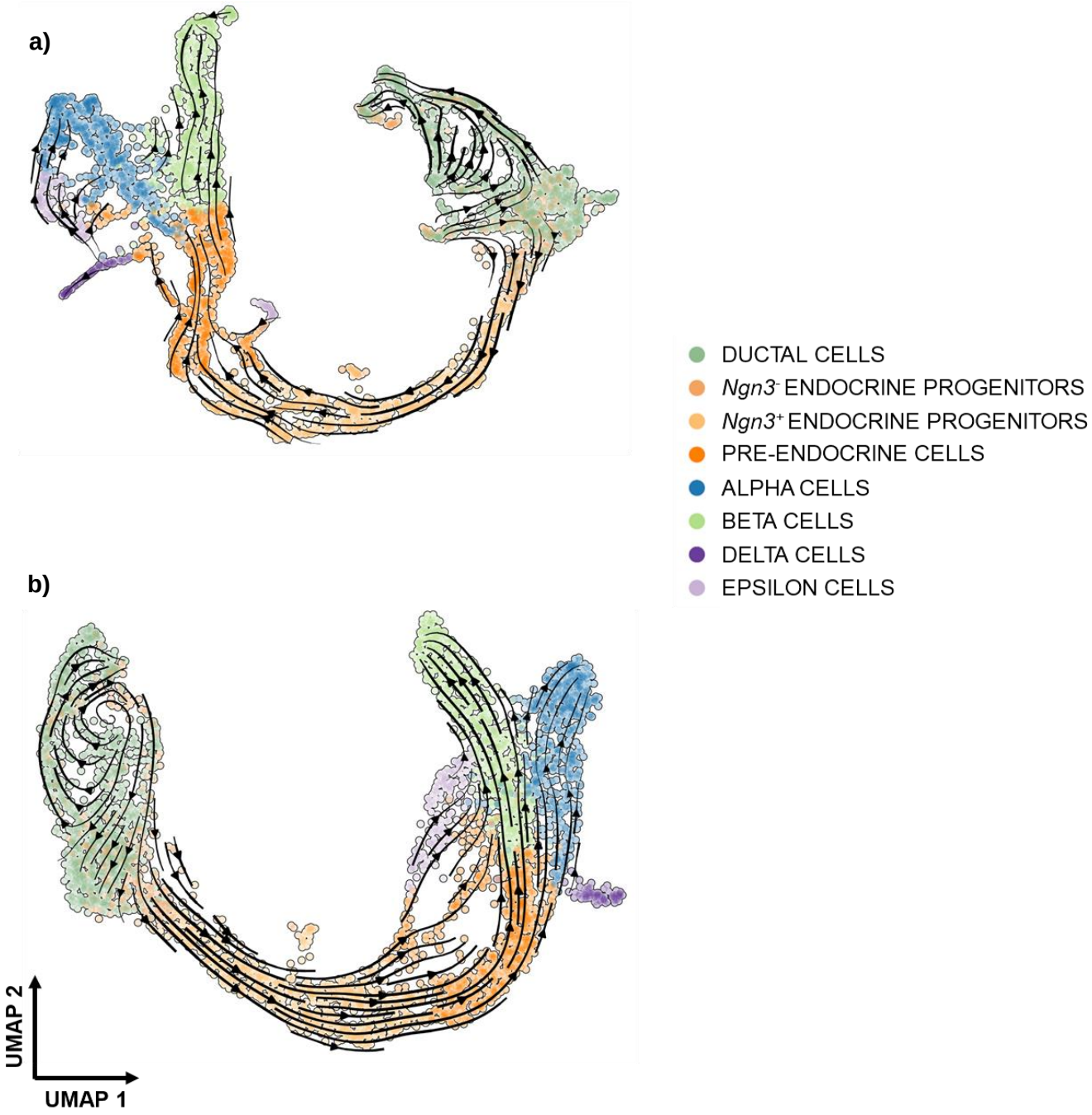

Figure 31: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the pancreas endocrinogenesis dataset and b) the RNA velocity vector field constructed by scVelo on the UMAP embedding built from gene expression.

the reorganization of the gene expression signal of ductal cells as they diverge to different fates, thus detecting detailed patterns that are missed by classic RNA velocity. However, it is important to underline that, given the limited extension of this dataset, it is very difficult to discern which of the two embeddings is more biologically reliable. Thus, the main conclusion that can be drawn from this comparison is that both classic RNA velocity and the velocity of the entropy succeed in detecting the dominant flow of differentiating cells that travel from pluripotent to fully specialized states.

Figure 32 shows the comparison between the vector field of FIERCE (**a**) and the principal graph drawn by Monocle 3 on its own UMAP embedding built from gene expression (**b**). Both are colored according to cell types and to the scores computed by the corresponding tools, i.e., the signaling entropy score of FIERCE and the pseudotime score of Monocle. The topology of the embedding constructed by Monocle is very similar to its counterpart constructed by FIERCE, and the principal graph, just like the vector field of FIERCE, recovers a primary trajectory that follows the differentiation path of endocrine cells, as well as a secondary trajectory that remains inside the ductal cells. As expected, the pseudotime score computed by Monocle shows an opposite trend with respect to the entropy score, as it progressively increases along the main developmental trajectory of endocrine cells.

Despite the general accordance of the results, the “top-down” approach of Monocle shows important downsides. First, since it is designed to summarize the information on the transcriptional similarity of all the cells of the dataset into a single graph structure, it inevitably misses all the small-scale details of each section of the trajectory. For example, the principal graph fails to identify the divergence of the trajectory at the end of the main dynamic process, in correspondence of the origin of the various endocrine cell types. It is well known that, during the embryonic development of murine pancreas, endocrine cells are generated in two consecutive phases[59]: in the first phase, spanning embryonic days 9-12.5, alpha cells are primarily produced, while in the second phase, spanning embryonic days 12.5-15.5, all the other endocrine cell types, including beta, delta, and epsilon cells, are massively produced alongside a few additional alpha cells. Since this dataset comes from embryonic day 15.5, a massive production of all four endocrine cell types is expected. While the vector fields generated by FIERCE and scVelo allow appreciating such divergent pattern at the end of the trajectory, the principal graph generated by Monocle is much more unclear, terminating in a “curly” end localized almost entirely within the alpha cells.

Another very important difference between the vector field of FIERCE and the principal graph of Monocle is related to the direction of the trajectory, that is inherently inferred by FIERCE, but requires to be manually set in the case of Monocle. Without our external imposition of a root state, localized within the ductal cells, the direction of the two trajectories could not have been evinced from the principal graph alone, thus the computation of the pseudotime score could not have been possible. This implies that the pseudotime score entirely depends on our choice of the root state, thus its accuracy in recapitulating the differentiation state of each cell is inevitably affected. On the contrary, the “bottom-up” approach of FIERCE involves first the direct measure of the differentiation potency of the single cells through signaling entropy, and then the direct inference of the direction of the trajectory from such potencies, without the need to specify any root state. This allows increasing both the level of detail of the resulting trajectory, and the precision of its inferred dynamics.

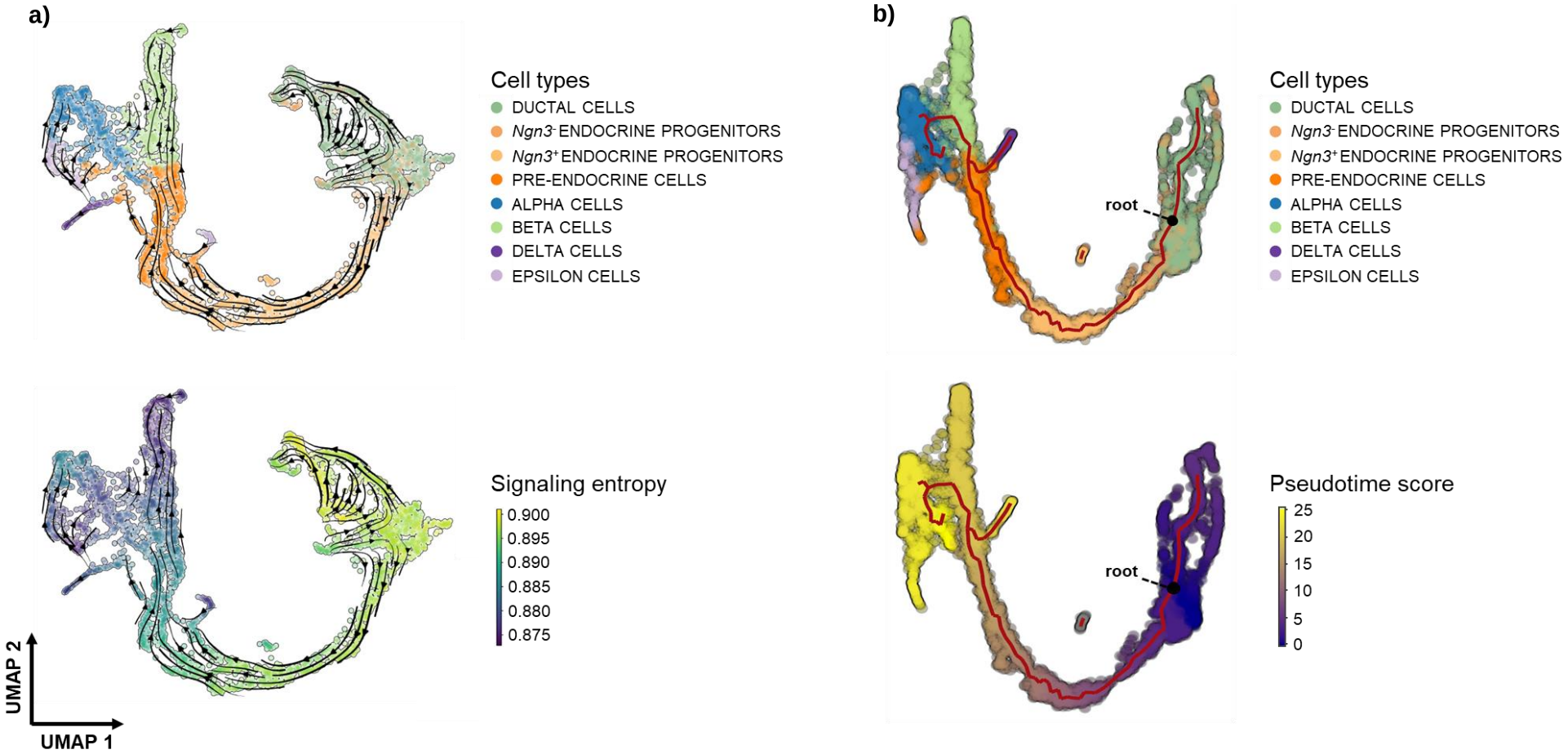

Figure 32: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the pancreas endocrinogenesis dataset and b) the principal graph constructed by Monocle 3 on its own UMAP embedding.

4.3.2. Dentate gyrus neurogenesis

Figure 33 shows the velocity of the entropy vector field constructed by FIERCE on the entropy based UMAP embedding of the dentate gyrus neurogenesis dataset, colored according to the cell types (**a**), the total signaling entropy scores (**b**), and the three potency states identified for this dataset (**c**).

Overall, two very different patterns can be observed in the vector field. The primary and most evident pattern involves the second part of the neurogenesis process, i.e., the differentiation of granule cells from their neuroblasts progenitors. This portion of the vector field nicely recapitulates the expected trajectory, with a massive stream of arrows that originates in early neuroblasts, passes through late neuroblasts and immature granule cells, and finally terminates into mature granule cells. This stream strictly follows the decreasing gradient of signaling entropy, as well as the gradual passage from potency state 1, to potency state 2, and finally to potency state 3, localized on the granule cells. This confirms that this portion of the genealogy follows the rules of classic developmental processes, with cells progressively losing their differentiation potential as they become more and more specialized in the neuronal function.

Another minor pattern shown by the vector field involves the precursor subpopulations, including quiescent radial glia, cycling radial glia, and neuronal progenitors. Instead of a common trajectory that, starting from the quiescent radial glia, points towards the early neuroblasts passing through the neuronal progenitors, several distinct streams of arrows can be observed. These streams spawn together in a precise location at the intersection between cycling radial glia, neuronal progenitors and early neuroblasts, and then diverge to different directions, terminating in separate regions of cycling radial glia and neuronal progenitors that correspond to the low-entropy spots localized within these cell types. This pattern suggests that the different subgroups of the pluripotent subpopulations might share a common ancestry, but to confirm such hypothesis additional analyses on an expanded dataset are necessary. Indeed, even if the diverging pattern observed here suggests that radial glia and nIPCs can initiate multiple cell lineages, this might as well be an artifact due to the velocity of the entropy being unable to recover the full dynamics of the potencies of cells. In the absence of the full lineages, including their respective end points, it is very difficult to reconstruct the complete phase portrait of key genes, thus this early portion of the vector field must be interpreted with caution.

Figure 34 shows the comparison between the velocity of the entropy vector field of FIERCE (**a**) and the RNA velocity vector field constructed by scVelo on the UMAP embedding derived from raw gene expression (**b**). Besides a few marginal differences, the two streamplots show a good agreement both in the prominent stream of arrows that follows the differentiation of neuroblasts into mature granule cells, and in the multiple small streams that, starting from the intersection between early neuroblasts, neuronal progenitors and cycling radial glia, head towards separate isolated spots of these precursor subpopulations. This seems to confirm that these multiple paths are not an artifact of signaling entropy, but rather represent real dynamics of this complex and heterogeneous portion of the cell population. However, without a fair number of representative cells for each lineage, it is not possible to support this hypothesis, that must be validated with further analyses. Nevertheless, it is worth appreciating that the streamplot produced by FIERCE allows interpreting cell dynamics in the light of the underlying pattern of the differentiation potency of cells, and thus to formulate new hypotheses on the evolution of the differentiation landscape.

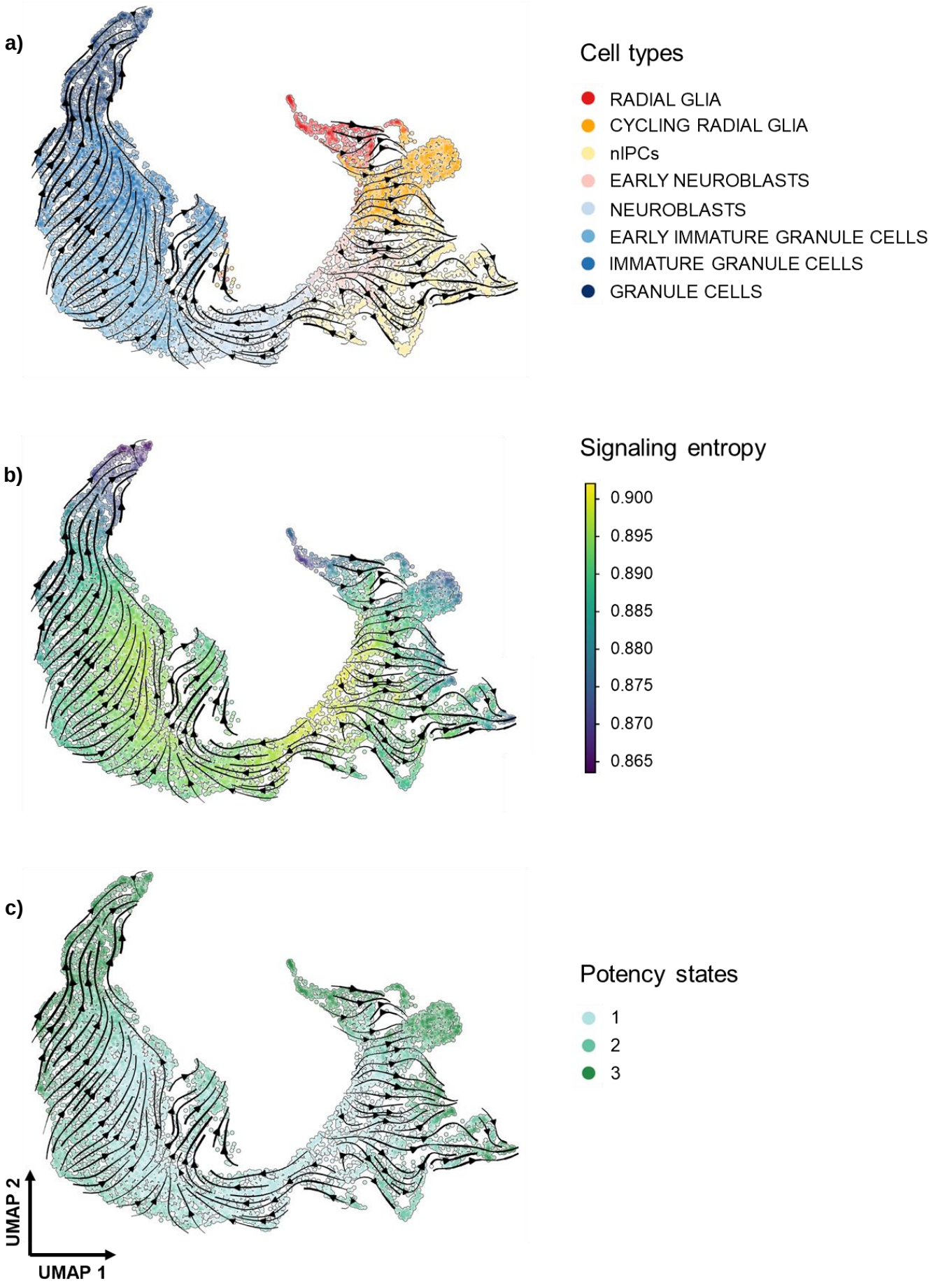

Figure 33: velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the dentate gyrus neurogenesis dataset, colored according to a) cell types, b) total signaling entropy score, and c) potency states.

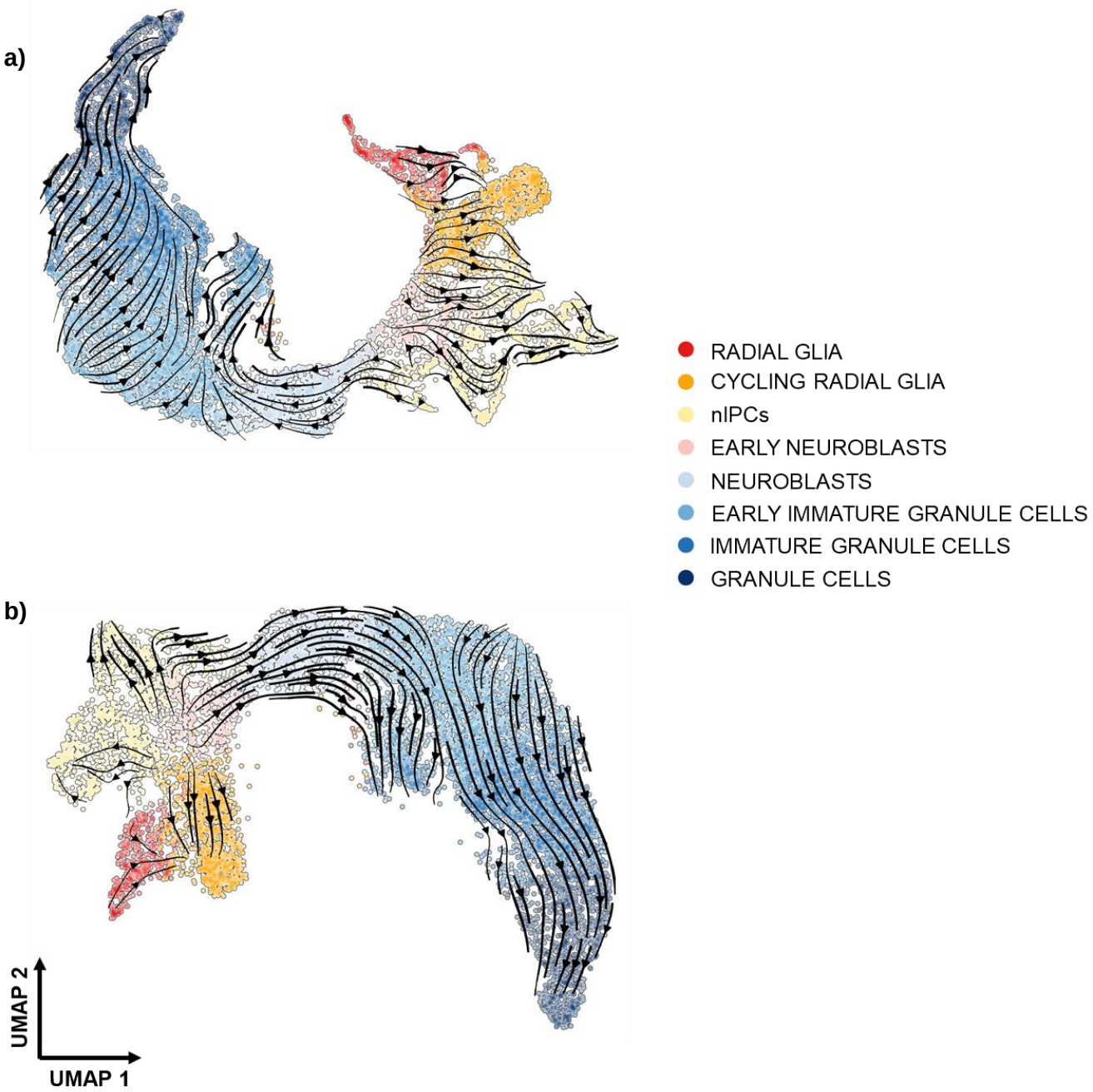

Figure 34: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the dentate gyrus neurogenesis dataset and b) the RNA velocity vector field constructed by scVelo on the UMAP embedding built from gene expression.

Figure 35 shows the comparison between the vector field of FIERCE (**a**) and the principal graph drawn by Monocle 3 on its own UMAP embedding (**b**). The trajectory reconstructed by Monocle appears incoherent and very difficult to interpret in the light of the known biology of murine dentate gyrus. The principal graph depicts an additional separate trajectory that is exclusive of granule cells, and totally disconnected from the main genealogy; due to such disconnection from the root state, the computed pseudotime scores of the cells belonging to this section are unrealistically uniform, and thus completely unreliable. Furthermore, besides this disconnected section, the main genealogy also suffers from a very important issue: although the structure of the principal graph nicely fits the expected sequence of cell states, the biological significance of the pseudotime score completely depends on which of these states is chosen as root. We chose the quiescent radial glia based on

our prior biological knowledge of this system; however, as shown both by the results of this study and by the analysis of Hochgerner et al[60], the heterogeneity and the complexity of the early pluripotent subpopulations are much higher than expected, thus the choice of a confident root state is very difficult. This uncertainty directly impacts the computed pseudotime score, that must be interpreted with extreme caution.

Overall, these results show that the “bottom-up” approach of FIERCE offers important advantages in the absence of solid preconceived biological knowledge on the system at hand. Since our approach is completely independent from any prior expectance, its results are more confident and easier to interpret than the results of a more traditional “top-down” method that first requires the manual specification of the “desired” direction, and then calculates an approximation of the differentiation potency of cells on the basis of such arbitrary setting.

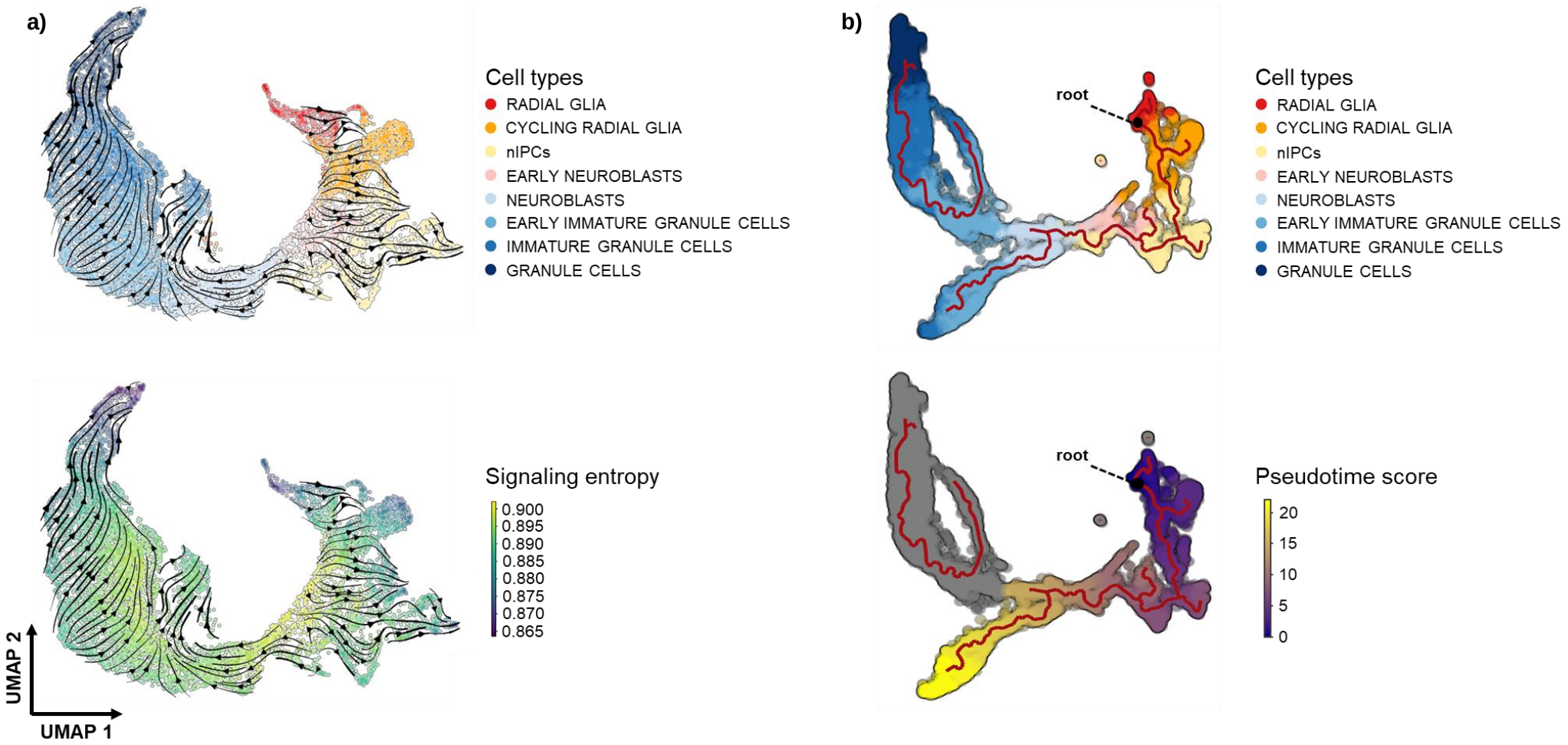

Figure 35: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the dentate gyrus neurogenesis dataset and b) the principal graph constructed by Monocle 3 on its own UMAP embedding.

4.3.3. Mammary gland development

Figure 36 shows the velocity of the entropy vector field constructed by FIERCE on the entropy based UMAP embedding of the mammary gland development dataset, colored according to the cell types (**a**), the sampling ages (**b**), the total signaling entropy scores (**c**), and the three potency states identified for this dataset (**d**). In the streamplot two emerging patterns can be clearly detected: the first is a divergent pattern that originates in a peripheral region of embryonic epithelial cells and then immediately bifurcates, leading to the luminal branch (including luminal cells and alveolar precursors) on one side and to the basal branch on the other; the second is a convergent pattern, that is localized within each of the three extremities of the tripartite structure, i.e., within each of the three most differentiated cell types. This second pattern is particularly evident within basal cells and alveolar precursors, where it forms a "vortex" structure converging in the middle of each group. In luminal cells, the trend appears more unclear and incomplete, to the point that it almost generates an inversion of direction. The reasons of this particular behavior are not immediately clear; however, since a small hint of a convergent pattern can be noticed at the bottom of this cell group, we hypothesize that this extremity of the tripartite structure might be simply more affected by the typical transcriptional noise of scRNA-seq data, that prevents the identification of a precise motif. A marked sensitiveness to transcriptional noise of this subpopulation is not unlikely, given that it represents a terminal stable state where we do not expect to observe particularly strong dynamics that can be easily detected by the algorithm.

The interpretation of the two patterns described above is straightforward in the light of the underlying gradients of signaling entropy and of the distributions of potency states. As embryonic epithelial cells "descend" the entropy peak, they take either the path that leads to basal cells, or the path that leads to luminal progenitors, where they will diverge again to head either towards the alveolar precursors, or towards the luminal cells; in the meantime, while descending the entropy gradient, they progressively lose their differentiation potency, and pass from potency state 1 to potency state 2. Finally, as they reach the entropy minima localized at the three extremities of the overall structure, the cells stably "settle" in potency state 3, thus creating the convergent patterns pointing to the core of these terminal stable states. This scenario is in perfect agreement with our knowledge on the developmental sequence of this biological system and confirms that the velocity of the entropy can accurately reconstruct the dynamics of cells on the differentiation landscape defined by the entropy based UMAP embedding.

On the contrary, the vector field built by scVelo (**Figure 37b**) does not define any clear trajectory, besides a prominent stream of cells originating in embryonic epithelial cells and pointing towards alveolar precursors. In this case, RNA velocity fails to reconstruct the transcriptional dynamics of the system, although it is very likely that the construction of a reliable vector field is affected by the inaccuracy of the underlying UMAP embedding, rather than by an unfitting velocity dynamical model. To ascertain this, we used the *plot_signature_statistics* function of FIERCE to print the phase portraits of two "balancer" gene signatures composed by Giraddi et al[71], i.e., two lists of genes that are highly expressed in embryonic epithelial cells and only in one of the two main branches that depart from them at the start of the trajectory. **Figure 38** shows the phase portraits of a selection of 20 genes from the basal "balancer" signature (genes that are highly expressed in embryonic epithelial cells and in basal cells), and **Figure 39** shows the phase portraits of a selection of 20 genes from the luminal "balancer" signature (genes that are highly expressed in embryonic epithelial cells and in the luminal branch, including alveolar precursors and luminal cells). For each of these signatures, the 20 selected genes are those that are characterized by the highest likelihood in the velocity dynamical model. The phase portraits of the basal signature show that, indeed, embryonic epithelial cells and basal cells (in red and blue, respectively) tend to concentrate on the upper-right half of the curves, that correspond to the high expression steady state; the same pattern can be

observed for embryonic epithelial cells and luminal cells (in green) in the phase portraits of the respective signature. This is the expected outcome of a dynamical model that correctly fits the transcriptional cycle of key regulator genes. Accordingly, the velocity of the entropy, that partly depends on the RNA velocity of the genes of the PPI network, yields a clear vector field that correctly depicts the expected trajectory. The crucial difference is thus represented by the underlying UMAP embeddings, that are topologically very different. The embedding built by FIERCE from signaling entropy shows a much more realistic structure than the embedding built by raw gene expression, suggesting that the additional information provided by the distribution of the gene expression signal on the cell-shared PPI network is crucial for the construction of a solid template where to draw the dynamics of cells on the differentiation landscape of this biological system. These observations demonstrate that the fully entropy-based “bottom-up” implementation of FIERCE is a valuable tool to reconstruct dynamic vector fields when the results obtained by the classic RNA velocity approach are not optimal or even ambiguous.

Figure 40 shows the comparison between the vector field of FIERCE (**a**) and the principal graph drawn by Monocle 3 on its own UMAP embedding (**b**). Although the structure of the principal graph shows several “twists” that partially hinder its visual interpretation, the global topology of the trajectory reconstructed by Monocle is correct, since we can recognize both the initial divergence between the basal and the luminal fate, and the subsequent divergence between the alveolar precursors and the luminal cells, at the level of luminal progenitors. However, the distribution of the pseudotime scores is rather misleading, especially in the basal cells which show unrealistically low values. Given our prior knowledge on this system, there is no biological reason why basal cells should present lower values than luminal cells or alveolar precursors, as they are supposed to be located at the same level on the differentiation landscape. This is most likely an artifact due to the much shorter length of the basal branch of the principal graph: since the pseudotime scores derive from the projection of cells on the graph, shorter branches result in lower scores, independently from the real potency level of cells. In this case, the advantage of a reverse “bottom-up” approach is particularly evident, since the signaling entropy is computed directly from the transcriptional data of cells, and thus reflects their true differentiation potency, rather than their position on a pre-computed sequence. Indeed, all the three end points of the trajectory present an equally low signaling entropy score and reside in the same potency states. These observations represent an effective example of how the “bottom-up” implementation of FIERCE can help in the reconstruction of branching dynamic processes in the presence of diverging cell lineages of unequal length and complexity.

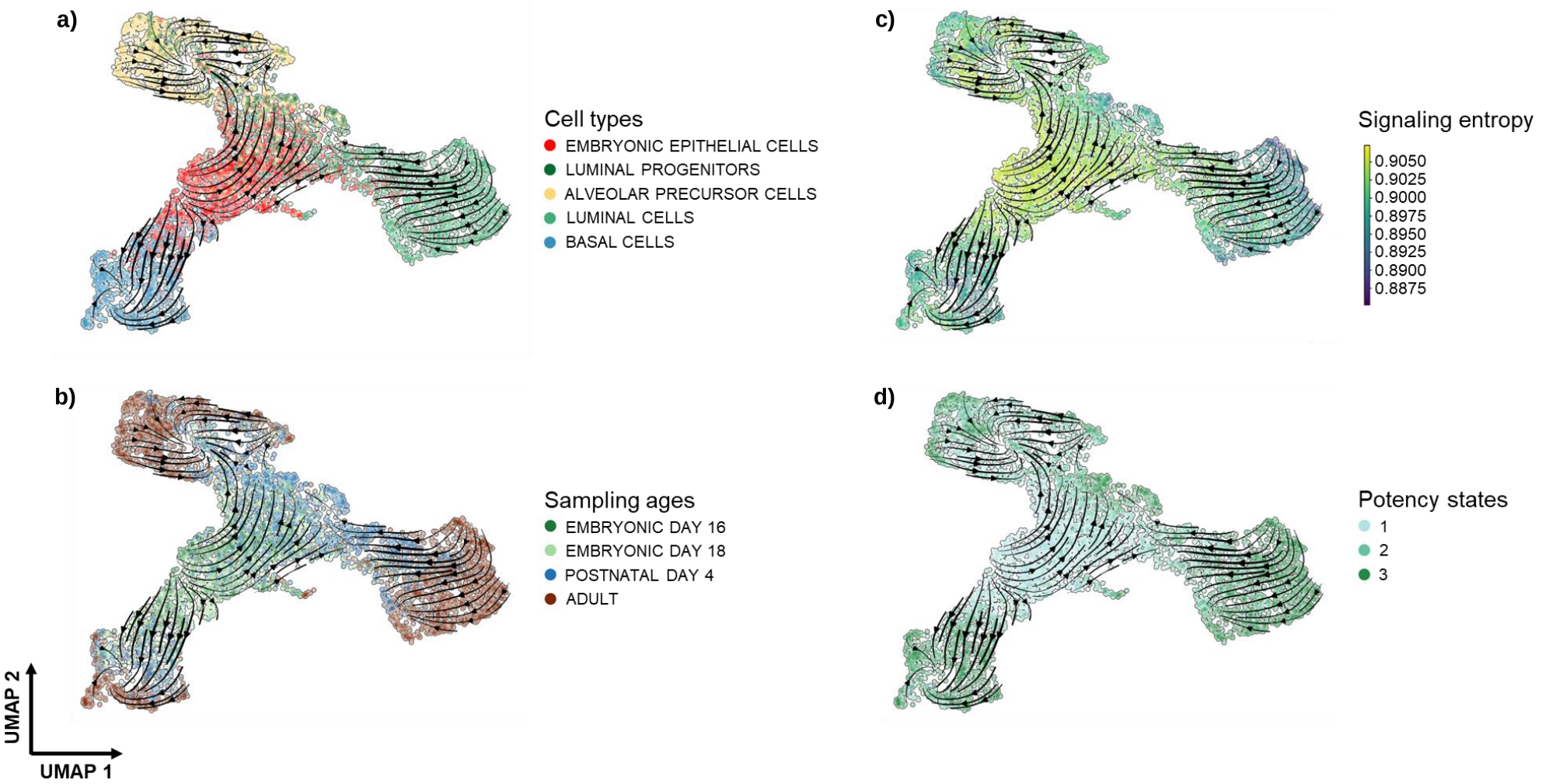

Figure 36: velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the mammary gland development dataset, colored according to a) cell types, b) sampling ages, c) total signaling entropy score, and d) potency states.

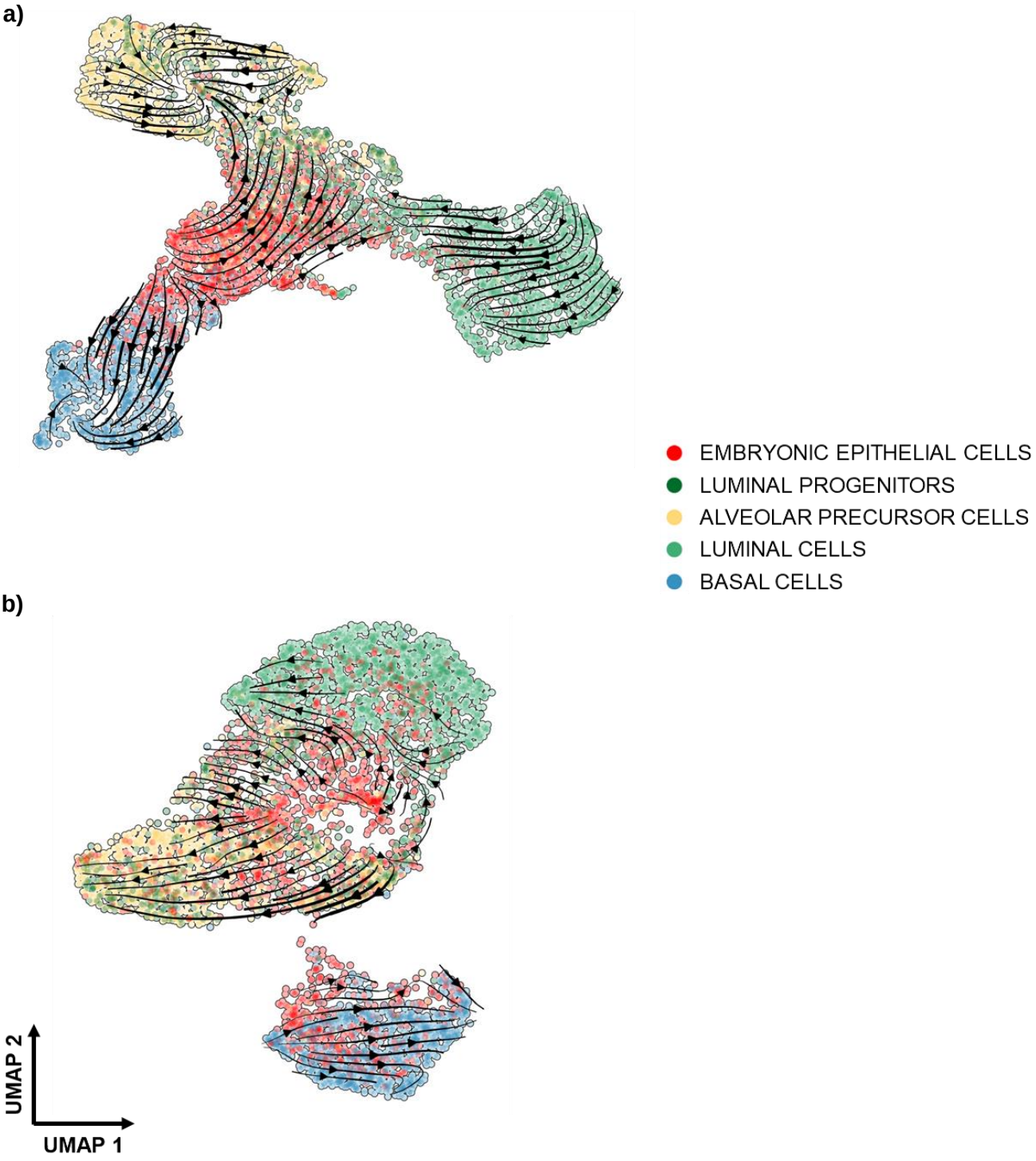

Figure 37: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the mammary gland development dataset and b) the RNA velocity vector field constructed by scVelo on the UMAP embedding built from gene expression.

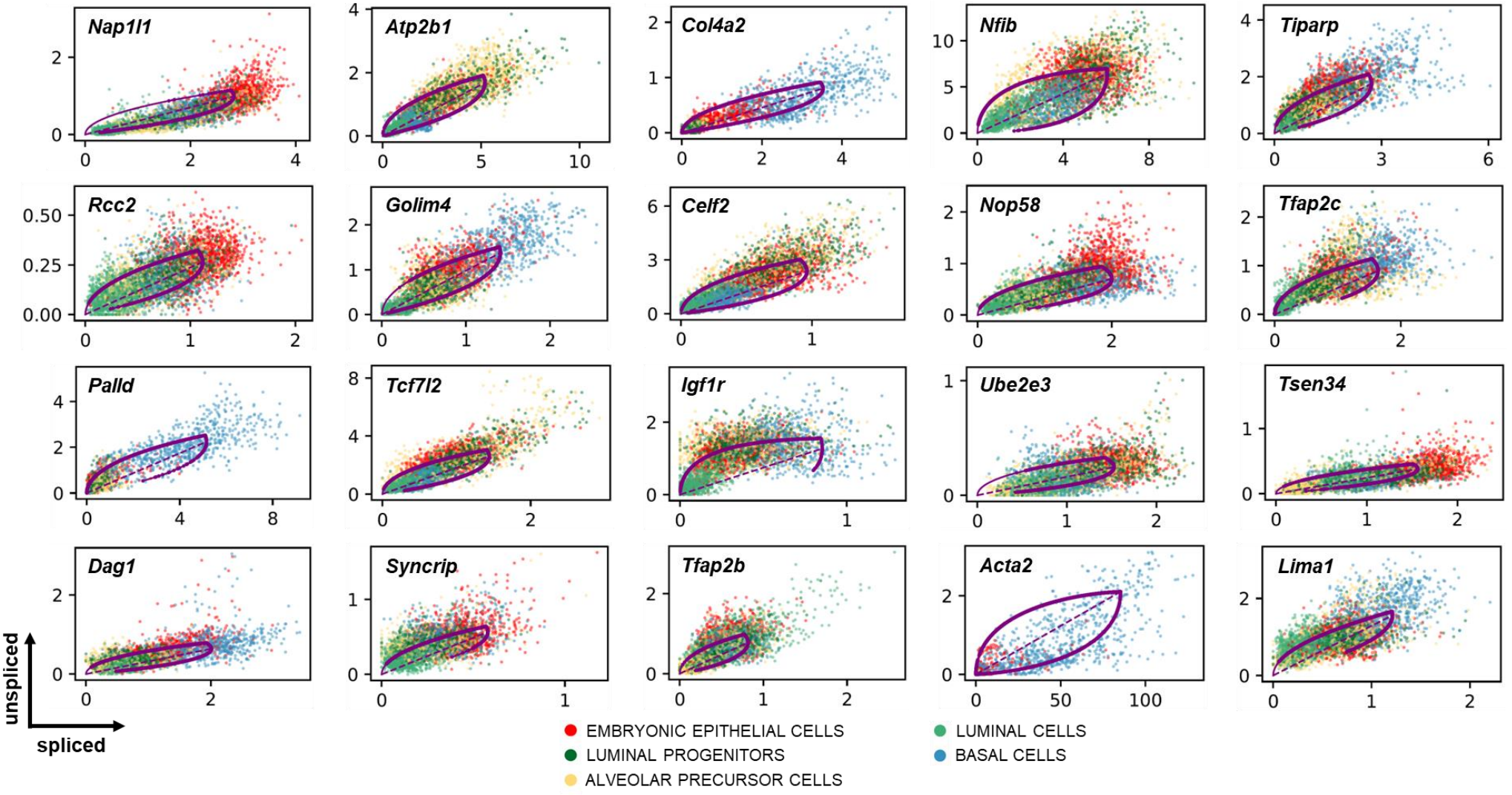

Figure 38: phase portraits of the 20 genes of the basal “balancer” signature that show the highest likelihood in the velocity dynamical model that was fit by FIERCE on the mammary gland development dataset.

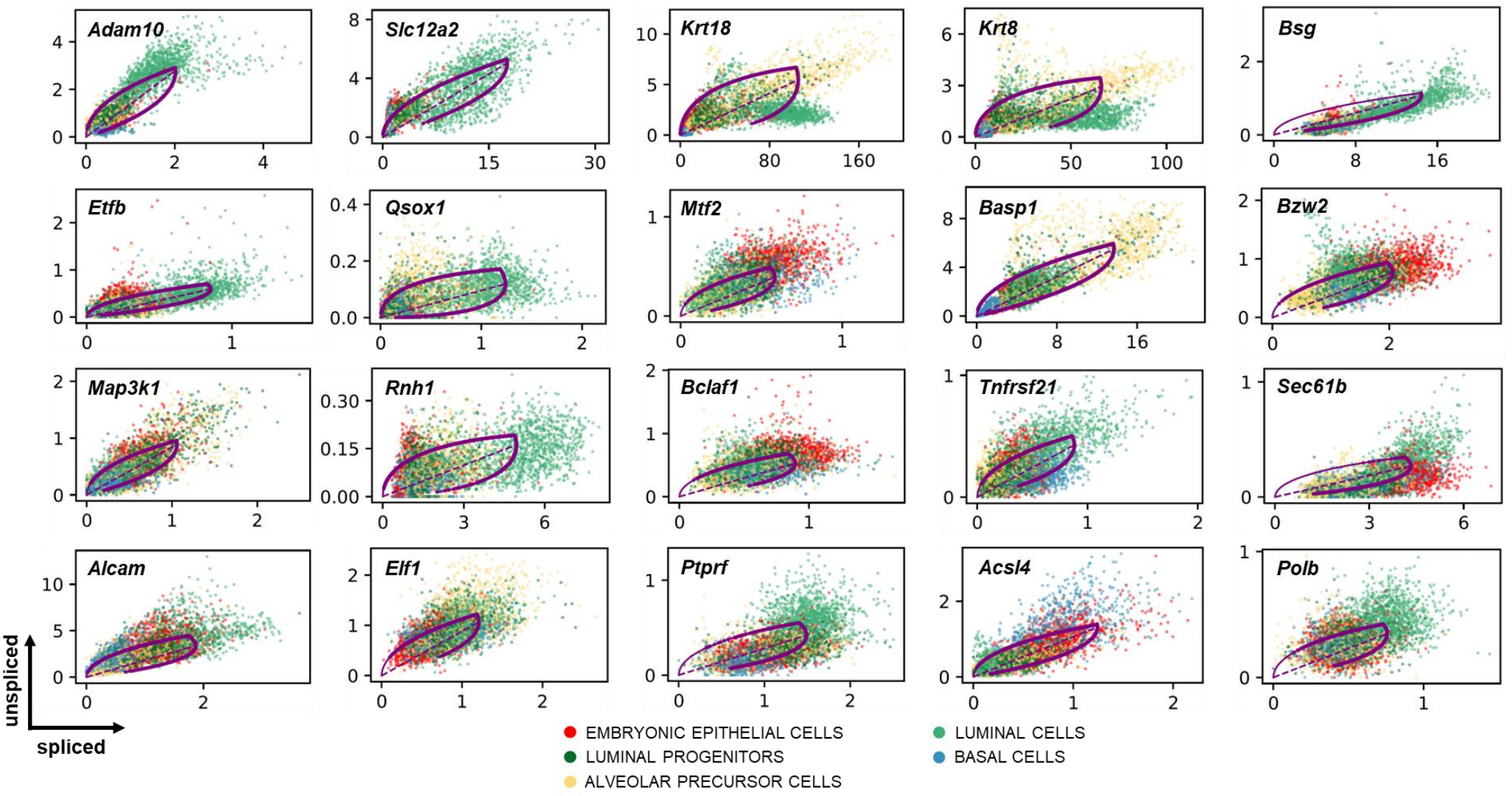

Figure 39: phase portraits of the 20 genes of the luminal “balancer” signature that show the highest likelihood in the velocity dynamical model that was fit by FIERCE on the mammary gland development dataset.

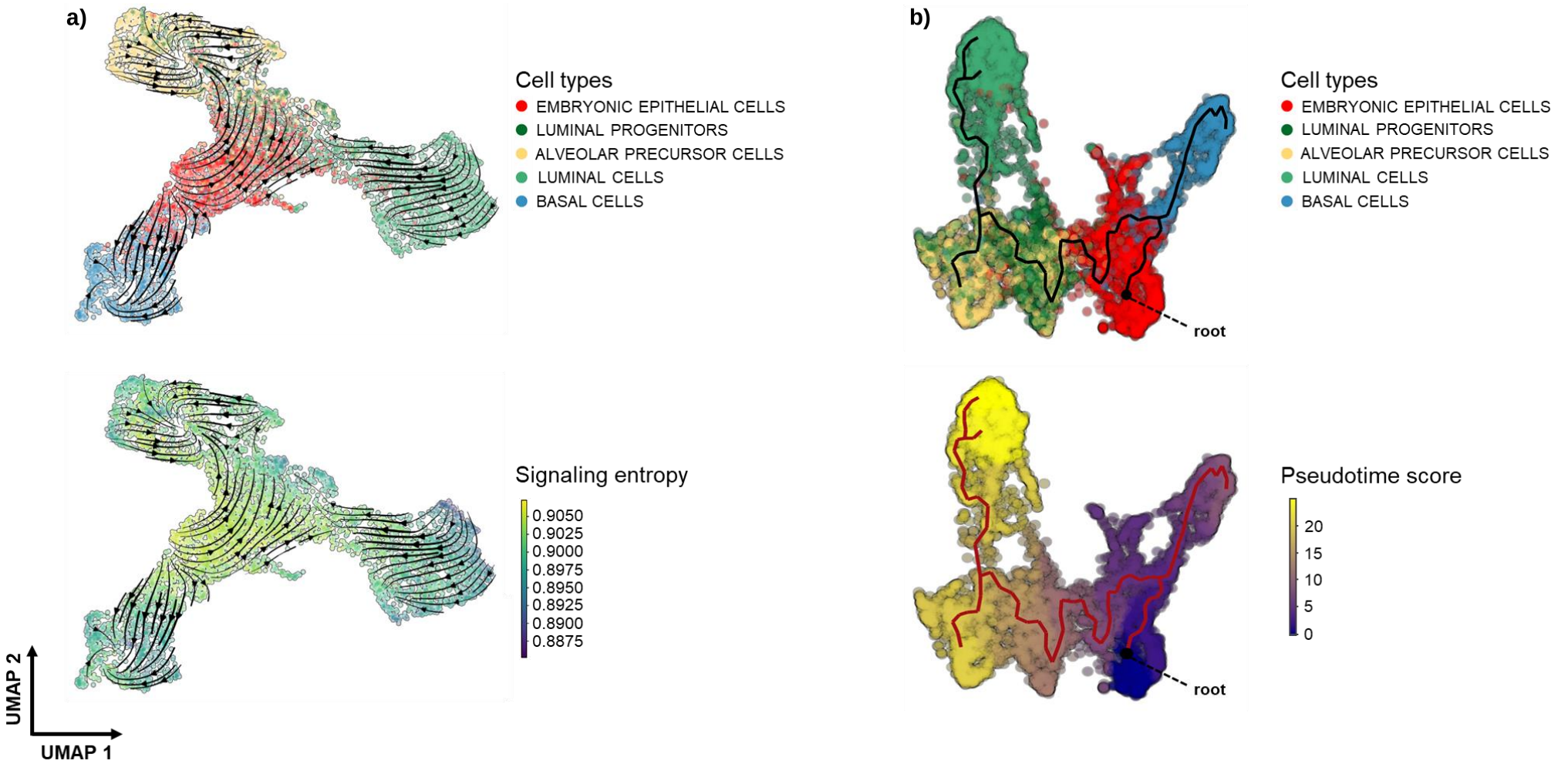

Figure 40: comparison between a) the velocity of the entropy vector field constructed by FIERCE on the entropy-based UMAP embedding of the mammary gland development dataset and b) the principal graph constructed by Monocle 3 on its own UMAP embedding.

5. Discussion

5.1. FIERCE demonstrates the efficacy of the "bottom-up" approach to trajectory reconstruction

We developed the method coded in FIERCE with two main objectives, i.e., i) to build a new tool for trajectory reconstruction from scRNA-seq data that does not require any prior knowledge on the biological system at hand, and ii) to provide a solid and coherent framework for ordering cells on continuous lineages based on the information brought by their own differentiation potencies. Briefly, our aim was to create a computational tool that lets the cell themselves tell their own developmental story.

Based on the results obtained on three scRNA-seq datasets coming from three well-known murine developmental processes, we can assert that FIERCE succeeds both in i) reconstructing reliable representations of the evolution of the Waddington differentiation landscape during a dynamic process, and ii) providing the user all the necessary information to identify and study the complete genealogical history of a cellular system, i.e., origin, direction, and ramifications of all its cellular lineages.

The key advantage offered by FIERCE is the possibility to visualize, on a single embedding, both the current distribution of cell potencies, represented by cell signaling entropy, and the prediction of how this current distribution is going to change in the immediate future. This feature overcomes the burden of formulating assumptions on the expected trend of cell differentiation potency, assumptions that are instead necessary to reconstruct a rough trajectory from entropy scores alone[38,52,53].

This feature allowed us to uncover biologically relevant details in all the analyzed datasets. In the pancreas endocrinogenesis dataset, besides the strong flux of arrows that defines the differentiation process of endocrine cell types from their ductal progenitors, FIERCE was able to identify a second minor trajectory that is entirely located within the ductal cluster. This shorter flux of arrows describes an early bifurcation event of ductal cells, that are either committed to the endocrine progenitor fate, or destined to remain within the ductal basin. This bifurcation was not described by classic RNA velocity, that identifies a single trajectory forming a "vortex" that finally converges into endocrine progenitors. However, it is important to precise once again that, in absence of a clear method to discern which representation is more faithful to the real biology of the system, the secondary trajectory identified by FIERCE must be interpreted with caution.

In the dentate gyrus neurogenesis dataset, FIERCE identified multiple distinct streams of arrows in the pluripotent portion of the cell population, spawning at the intersection between early neuroblasts, neuronal progenitors and cycling radial glia, and terminating into separate peripheral regions of neuronal progenitors and radial glia. These regions were accordingly defined by signaling entropy as sinks that were well separated from the nearby regions, characterized by a higher signaling entropy. This result is in agreement with the known heterogeneity of the pluripotent subpopulations of murine dentate gyrus[60], that are known to give birth to different neuronal cell lineages besides

granule cells, and to host clusters of undifferentiated proliferative cells that are entirely committed to self-renewal. The entropy sinks identified by FIERCE provide fertile ground for several hypotheses that can be tested by analyzing more cells belonging to more lineages of the same system. For example, it can be speculated that the low entropy regions of radial glia and neuronal progenitors represent groups of cells that are fully specializing into a very specific stable state that is entirely committed to the proliferative function. In this scenario, their very low signaling entropy might be explained by the progressive concentration of their gene expression signal into those gene pathways that are specifically dedicated to cell cycle and proliferation. On the other hand, the low signaling entropy of these cells might also find another explanation: since both radial glia and neuronal progenitors are ancestors of multiple different neuronal cell types[60], it is possible that a subset of their cells might already be initiating these multiple lineages, and thus might already be concentrating their gene expression on the necessary functional gene categories. As the observation of FIERCE streamplots suggested the formulation of multiple hypotheses, we are currently performing additional studies on an expanded dataset to shed light on the unfolding of all the cell lineages originating from these heterogeneous pluripotent subpopulations.

In the dataset of the mammary gland development, the streamplots generated by FIERCE are much clearer and much more interpretable compared to their equivalents generated from gene expression and RNA velocity. As suggested by the phase portraits of key genes, this remarkable difference, rather than to the velocity metric itself, is likely to be related to the underlying embedding, that is fully coherent with the expected population structure only when reconstructed from signaling entropy. This suggests that the additional information brought by the distribution of gene expression on the different gene pathways greatly helps in disentangling the topology of the population whereas the intensity of gene expression alone yields ambiguous results. For this dataset, the velocity of the entropy vector field constructed by FIERCE offers a perfect representation of the Waddington differentiation landscape of a typical developmental process: from the central signaling entropy peak, we observe the departure of three distinct streams of arrows corresponding to three distinct potency gradients, that finally terminate into three well separated valleys, where we observe three signaling entropy sinks highlighted by convergent vector field patterns. The “bottom-up” approach of FIERCE was fundamental to recover this structure without the need to manually set the position of the root on the central peak, and to assume a constant drop of signaling entropy towards the three terminal valleys.

5.2. Limitations of FIERCE and future perspectives on the "bottom-up" approach

Despite its efficiency, FIERCE still represents our first attempt to implement a real "bottom-up" tool for trajectory reconstruction, and thus is not devoid of weak points that may affect its performance in specific situations.

The most important downside is its strong dependence on the accuracy of RNA velocity estimates. The prediction of the future transcriptional states of cells is the very first step of the FIERCE pipeline, and the reliability of the results entirely depends on the confidence of the RNA velocity dynamical model. In turn, the confidence of the model entirely depends on the goodness of fit of the phase portraits of genes, and, in particular, of the key genes involved in the process at hand.

It has been demonstrated[57] that, due to its mathematical implementation, the fitting of the RNA velocity dynamical model does not perform well when the key genes of the dynamic process undergo sudden expression "leaps". Indeed, a slow and gradual change, with at least a few representative cells for each major phase of the transcriptional cycle, is a key requirement for an optimal fitting. In the absence of this graduality, cells might even be assigned to the wrong phases, and the entire structure of the whole phase portrait might be totally mistaken. This is a well known weakness of the RNA velocity algorithm[56], that can be compensated either through the direct identification and exclusion of those genes that present abrupt expression changes[57], or through the adoption of more sophisticated velocity based algorithms that can reduce the impact of transcriptional noise (as the recently introduced Velocity AutoEncoder[77] (VeloAE), that uses neural networks to construct a joint representation for both current transcriptional states and velocity vectors).

Another potential weakness of FIERCE is related to the signaling entropy, whose accuracy strongly depends on the completeness of the PPI network. The more complex the network is, the more reliable are the resulting entropy estimates. In particular, the complexity of the network directly affects the precision of entropy estimates, meaning that more complex networks significantly reduce the dispersion of the entropy distributions of the different cell states. Unfortunately, it is not possible to determine *a priori* if the choice of the network has been optimal or not, besides the compatibility of the results with the putative expectations of the user.

Moreover, signaling entropy brings the disadvantage of a relatively long computational time, that is in the order of a few hours, depending on the sample size. In this regard, Monocle[46,47,78] performs much better, with a run time of just a few seconds. However, a much longer run time is an unavoidable collateral effect of a "bottom-up" implementation, since, regardless of the exact metric, the differentiation potency must be computed for each single cell. The use of a very complex measure like signaling entropy, rather than of a simpler measure such as the scEntropy score of SLICE[52], drastically increases the time requirement, but compensates this inconvenience with much more precise and detailed results.

Finally, FIERCE requires in input a loom file created by velocyto[54]. The pipeline of FIERCE starts from RNA velocity computation, that requires that the total UMI counts of the sample are first subdivided into spliced and unspliced counts. This operation can only be performed by one of the dedicated functions of velocyto, that in turn can only process alignment files produced by a dedicated software such as Cellranger[11]. This means that, in the absence of an alignment file or a pre-existent

anndata object, the fastq files of the scRNA-seq experiment are necessary. These files are not always available for public datasets, that are often shared in the form of raw count matrices.

As evidenced by the points discussed above, the reconstruction of dynamic trajectories from the differentiation potency of single cells through a fully "bottom-up" strategy poses several challenges, but the exploration of the potentialities brought by this new approach is still in its infancy. Most currently available trajectory reconstruction tools concentrate on "top-down" strategies that, being based on cell-cell distances in the gene expression space, do not require any specific input format besides a simple gene expression matrix, compensate very well the inherent noise of scRNA-seq data, and are very fast and computationally efficient. However, all these advantages come with a huge cost: the efficacy of the "top-down" strategy depends on the reliability of the prior expectations of the user, who must already know the direction of the dynamic process that has to be reconstructed. If we want to unleash the true potential of scRNA-seq data, and to "read the story" of biological systems without knowing how exactly this story is going to end, we must resort to a fully unsupervised and reductionist approach inspired by the principles of statistical mechanics. As the resolution of sequencing data becomes finer and finer, their computational analysis must become more and more quantitative, and the new computational tools must be able to reconstruct the evolution of macroscopic biological systems from the dynamics of their microscopic components[4].

Although several methods have been devised to infer the differentiation potency of single cells from their transcriptional data[38,50,52], it is not yet clear how to exploit these potencies to reconstruct genealogies without prior knowledge on their expected trend. FIERCE solves this problem with its composite implementation that resorts to the predictive power of the velocity algorithm. Although this solution inevitably brings forth some potential issues, FIERCE unarguably demonstrated both the feasibility and the efficacy of the "bottom-up" strategy. This is the reason why we plan to bring forward this research line, and in particular we aim to accomplish two specific goals.

As the first goal, we aim to quantify the faithfulness of the dynamic trajectories reconstructed by FIERCE to the respective real cell trajectories through a proper correlation index. The results reported in this thesis clearly prove that, in some cases, the genealogies produced by FIERCE appear more biologically plausible than their equivalents produced by classic RNA velocity or by a traditional "top-down" method; however, this judgement derives from a purely visual evaluation, and as such is still subject to our prior knowledge of the biological system, or even to the personal interpretation of the viewer. To demonstrate that the "bottom-up" method brings unquestionable advantages to trajectory reconstruction, the sequences of cell states represented in the streamplots produced by FIERCE must be compared to their correspondent "ground-truth" sequences. The latter can only be recovered by coupling scRNA-seq experiments with lineage tracing experiments, that allow following the fates of cell lineages over time[79]. This is exactly the foundation of a new experimental design devised by Weinreb et al[80], who applied their method to infer the clonal fate of different groups of stem cells during mouse hematopoiesis. This experiment yielded the exact trajectories followed by each cell clone during the developmental process, and the authors were even able to benchmark several trajectory reconstruction methods thanks to the availability of these "ground-truth" cell lineages. In the near future, we plan to test the performance of FIERCE on this same dataset, and in particular to quantitatively measure the biological faithfulness of the resulting trajectories by computing their correlation with the known developmental sequences. The aim of such additional study is to support the results presented in this thesis with solid mathematical evidence of the efficacy of the "bottom-up" approach of FIERCE.

As the second goal, we aim to exploit the insights we gained from the development of FIERCE to investigate novel concepts and computational approaches for the direct reconstruction of tissue

genealogies from the dynamics of the potencies of their cells. In this regard, an interesting starting point will be the investigation of the different kinds of information that can be obtained by genes that, during the dynamic process, change their transcriptional patterns according to different trends. For example, the rapidly evolving genes that currently represent a huge problem for the velocity algorithm, due to the sudden change of their expression patterns with large "leaps", might represent a valuable resource, rather than a hindrance. According to the conceptualization of dynamic processes as Waddington landscapes, cells move from one stable state to another through low energy paths, so it would be interesting to devise a mathematical framework that can decompose the potencies of cells into multiple components whose dynamics account for these different features of the Waddington landscape. In such scenario, the "discretized" transcriptional states of rapidly evolving genes are expected to be most responsible for the component that explains the stable states, while the "continuous" transcriptional states of gradually evolving genes are expected to be most responsible for the component that explains the linking low energy paths. This concept brings forth a strong analogy with the "punctuated equilibrium" theory of biological evolution[81], and appears even more interesting in the light of the ongoing debate on the clonal evolution of cancer, that is questioning the classic view of cancer progression as a gradual "Darwinian" process in favor of an alternative interpretation based on long periods of stasis interleaved by rapid evolutive bursts[82]. A novel "bottom-up" strategy for trajectory reconstruction that aims to model the temporal evolution of the Waddington landscape according to this alternative view of dynamic processes might greatly help in shedding light not only on the patterns of classic differentiation processes, but also on the still obscure patterns that govern the progression of cancer and other diseases.

Despite much work has yet to be done and new mathematical approaches have yet to be devised, still we believe that FIERCE represents a promising starting point to develop a novel set of tools that allow exploiting the potential of scRNA-seq data in new and unexpected ways and that it might reveal new crucial patterns in the transcriptional evolution of biological systems that current "top-down" methods are unable to discern. Therefore, we envision that FIERCE will prove to be a valuable resource for the scientific community, not only to explore the dynamic information hidden within scRNA-seq data from a new perspective, but also to bring forth new ideas and new methods that can help us fully understand the complex laws through which "*endless forms most beautiful and most wonderful have been, and are being, evolved*"[83].

References

1. Hedlund, E. & Deng, Q. Single-cell RNA sequencing: Technical advancements and biological applications. *Molecular Aspects of Medicine* vol. 59 36–46 Preprint at https://doi.org/10.1016/j.mam.2017.07.003 (2018).

2. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* **6**, 377–382 (2009).

3. Dal Molin, A. & di Camillo, B. How to design a single-cell RNA-sequencing experiment: Pitfalls, challenges and perspectives. *Brief Bioinform* **20**, 1384–1394 (2018).

4. Teschendorff, A. E. & Feinberg, A. P. Statistical mechanics meets single-cell biology. *Nature Reviews Genetics* vol. 22 459–476 Preprint at https://doi.org/10.1038/s41576-021-00341-z (2021).

5. Feynman, R. P. *Statistical Mechanics: A Set of Lectures*. (CRC Press, 2018).

6. Landau, D. A. & Lifshitz, E. M. *Statistical Physics*. vol. 5 (Elsevier, 1980).

7. Tritschler, S. *et al.* Concepts and limitations for learning developmental trajectories from single cell genomics. *Development (Cambridge)* vol. 146 Preprint at https://doi.org/10.1242/dev.170506 (2019).

8. Chen, J., Rénia, L. & Ginhoux, F. Constructing cell lineages from single-cell transcriptomes. *Molecular Aspects of Medicine* vol. 59 95–113 Preprint at https://doi.org/10.1016/j.mam.2017.10.004 (2018).

9. Saelens, W., Cannoodt, R., Todorov, H. & Saeys, Y. A comparison of single-cell trajectory inference methods: Towards more accurate and robust tools. *bioRxiv* (2018) doi:10.1101/276907.

10. Haque, A., Engel, J., Teichmann, S. A. & Lönnberg, T. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Medicine* vol. 9 Preprint at https://doi.org/10.1186/s13073-017-0467-4 (2017).

11. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat Commun* **8**, (2017).

12. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

13. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol Syst Biol* **15**, (2019).

14. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F. & Regev, A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol* **33**, 495–502 (2015).

15. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat Biotechnol* **36**, 411–420 (2018).

16. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-1902.e21 (2019).

17. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

18. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* **19**, (2018).

19. AlJanahi, A. A., Danielsen, M. & Dunbar, C. E. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy - Methods and Clinical Development* vol. 10 189–196 Preprint at https://doi.org/10.1016/j.omtm.2018.07.003 (2018).

20. Lun, A. T. L., McCarthy, D. J. & Marioni, J. C. A step-by-step workflow for low-level analysis of single-cell RNA-seq data [version 1; referees: 5 approved with reservations]. *F1000Res* **5**, (2016).

21. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* vol. 16 133–145 Preprint at https://doi.org/10.1038/nrg3833 (2015).

22. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).

23. Büttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A. & Theis, F. J. A test metric for assessing single-cell RNA-seq batch correction. *Nat Methods* **16**, 43–49 (2019).

24. Lever, J., Krzywinski, M. & Altman, N. Points of Significance: Principal component analysis. *Nature Methods* vol. 14 641–642 Preprint at https://doi.org/10.1038/nmeth.4346 (2017).

25. Haghverdi, L., Buettner, F. & Theis, F. J. Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics* **31**, 2989–2998 (2015).

26. Grün, D. & van Oudenaarden, A. Design and Analysis of Single-Cell Sequencing Experiments. *Cell* vol. 163 799–810 Preprint at https://doi.org/10.1016/j.cell.2015.10.039 (2015).

27. Chen, G., Ning, B. & Shi, T. Single-cell RNA-seq technologies and related computational data analysis. *Frontiers in Genetics* vol. 10 Preprint at https://doi.org/10.3389/fgene.2019.00317 (2019).

28. MacQueen, J. Some methods for classification and analysis of multivariate observations. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 281–297 (1967).

29. Blondel, V. D., Guillaume, J. L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J Stat Mech: Theory Exp 2008* (2008).

30. Aran, D. *et al.* Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat Immunol* **20**, 163–172 (2019).

31. van der Maaten, L. & Hinton, G. *Visualizing Data using t-SNE*. *Journal of Machine Learning Research* vol. 9 (2008).

32. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. (2018).

33. Bao, R. *et al.* Review of current methods, Applications, And data management for the bioinformatics analysis of whole exome sequencing. *Cancer Informatics* vol. 13 67–82 Preprint at https://doi.org/10.4137/CIN.S13779 (2014).

34. Navin, N. E. & Hicks, J. Tracing the tumor lineage. *Molecular Oncology* vol. 4 267–283 Preprint at https://doi.org/10.1016/j.molonc.2010.04.010 (2010).

35. Bowes, A. L., Tarabichi, M., Pillay, N. & van Loo, P. Leveraging single-cell sequencing to unravel intratumour heterogeneity and tumour evolution in human cancers. *Journal of Pathology* vol. 257 466–478 Preprint at https://doi.org/10.1002/path.5914 (2022).

36. Herring, C. A., Chen, B., McKinley, E. T. & Lau, K. S. Single-Cell Computational Strategies for Lineage Reconstruction in Tissue Systems. *CMGH* vol. 5 539–548 Preprint at https://doi.org/10.1016/j.jcmgh.2018.01.023 (2018).

37. Waddington, C. H. *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*. (Allen and Unwin, 1957).

38. Teschendorff, A. E. & Enver, T. Single-cell entropy for accurate estimation of differentiation potency from a cell's transcriptome. *Nat Commun* **8**, (2017).

39. Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J. & Desfeux, A. OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford)* **2014**, (2014).

40. Davis, S. *et al.* List of software packages for single-cell data analysis. *https://github.com/seandavi/awesome-single-cell* (2018).

41. Zappia, L., Phipson, B. & Oshlack, A. Exploring the single-cell RNA-seq analysis landscape with the scRNA-tools database. *PLoS Comput Biol* **14**, (2018).

42. Guo, G. *et al.* Resolution of Cell Fate Decisions Revealed by Single-Cell Gene Expression Analysis from Zygote to Blastocyst. *Dev Cell* **18**, 675–685 (2010).

43. Shalek, A. K. *et al.* Single-cell RNA-seq reveals dynamic paracrine control of cellular variation. *Nature* **510**, 363–369 (2014).

44. Setty, M. *et al.* Characterization of cell fate probabilities in single-cell data with Palantir. *Nat Biotechnol* **37**, 451–460 (2019).

45. Mao, Q., Wang, L., Tsang, I. W. & Sun, Y. Principal Graph and Structure Learning Based on Reversed Graph Embedding. *IEEE Trans Pattern Anal Mach Intell* **39**, 2227–2241 (2017).

46. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol* **32**, 381–386 (2014).

47. Cao, J. *et al.* The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).

48. Street, K. *et al.* Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics* **19**, (2018).

49. Banerji, C. R. S. *et al.* Cellular network entropy as the energy potential in Waddington's differentiation landscape. *Sci Rep* **3**, (2013).

50. Gulati, G. S. *et al. Single-cell transcriptional diversity is a hallmark of developmental potential*. https://www.science.org.

51. Ramilowski, J. A. *et al.* A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat Commun* **6**, (2015).

52. Guo, M., Bao, E. L., Wagner, M., Whitsett, J. A. & Xu, Y. SLICE: Determining cell differentiation and lineage based on single cell entropy. *Nucleic Acids Res* **45**, (2017).

53. Chen, W. *et al.* Single-cell landscape in mammary epithelium reveals bipotent-like cells associated with breast cancer risk and outcome. *Commun Biol* **2**, (2019).

54. La Manno, G. *et al.* RNA velocity of single cells. *Nature* **560**, 494–498 (2018).

55. Bergen, V., Lange, M., Peidli, S., Wolf, F. A. & Theis, F. J. Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat Biotechnol* **38**, 1408–1414 (2020).

56. Bergen, V., Soldatov, R. A., Kharchenko, P. v & Theis, F. J. RNA velocity—current challenges and future perspectives. *Mol Syst Biol* **17**, (2021).

57. Barile, M. *et al.* Coordinated changes in gene expression kinetics underlie both mouse and human erythroid maturation. *Genome Biol* **22**, (2021).

58. Tedesco, M. *et al.* Chromatin Velocity reveals epigenetic dynamics by single-cell profiling of heterochromatin and euchromatin. *Nat Biotechnol* **40**, 235–244 (2022).

59. Bastidas-Ponce, A. *et al.* Comprehensive single cell mRNA profiling reveals a detailed roadmap for pancreatic endocrinogenesis. *Development (Cambridge)* **146**, (2019).

60. Hochgerner, H., Zeisel, A., Lönnerberg, P. & Linnarsson, S. Conserved properties of dentate gyrus neurogenesis across postnatal development revealed by single-cell RNA sequencing. *Nat Neurosci* **21**, 290–299 (2018).

61. Giraddi, R. R. *et al.* Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep* **24**, 1653-1666.e7 (2018).

62. Ushey, K., Allaire, J. & Tang, Y. reticulate: Interface to 'Python'. *https://rstudio.github.io/reticulate/, https://github.com/rstudio/reticulate* (2023).

63. Cerami, E. G. *et al.* Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Res* **39**, (2011).

64. Snel, B., Lehmann, G., Bork, P. & Huynen, M. A. *STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene*. *Nucleic Acids Research* vol. 28 http://www.bork.embl-heidelberg.de/STRING (2000).

65. Szklarczyk, D. *et al.* The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res* **49**, D605–D612 (2021).

66. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281-291.e9 (2019).

67. Han, X. *et al.* Mapping the Mouse Cell Atlas by Microwell-Seq. *Cell* **172**, 1091-1107.e17 (2018).

68. Fei, L. *et al.* Systematic identification of cell-fate regulatory programs using a single-cell atlas of mouse development. *Nat Genet* **54**, 1051–1061 (2022).

69. Wang, R. *et al.* Construction of a cross-species cell landscape at single-cell level. *Nucleic Acids Res* **51**, 501–516 (2023).

70. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol* **36**, 421–427 (2018).

71. Giraddi, R. R. *et al.* Single-Cell Transcriptomes Distinguish Stem Cell State Changes and Lineage Specification Programs in Early Mammary Gland Development. *Cell Rep* **24**, 1653-1666.e7 (2018).

72. Veltmaat, J. M., Mailleux, A. A., Thiery, J. P. & Bellusci, S. Mouse embryonic mammogenesis as a model for the molecular regulation of pattern formation. *Differentiation* **71**, 1–17 (2003).

73. Spike, B. T. *et al.* A mammary stem cell population identified and characterized in late embryogenesis reveals similarities to human breast cancer. *Cell Stem Cell* **10**, 183–197 (2012).

74. Prater, M. D. *et al.* Mammary stem cells have myoepithelial cell properties. *Nat Cell Biol* **16**, 942–950 (2014).

75. Giraddi, R. R. *et al.* Stem and progenitor cell division kinetics during postnatal mouse mammary gland development. *Nat Commun* **6**, (2015).

76. Makarem, M. *et al.* Stem cells and the developing mammary gland. *Journal of Mammary Gland Biology and Neoplasia* vol. 18 209–219 Preprint at https://doi.org/10.1007/s10911-013-9284-6 (2013).

77. Qiao, C. & Huang, Y. Representation learning of RNA velocity reveals robust cell transitions. *BIOPHYSICS AND COMPUTATIONAL BIOLOGY COMPUTER SCIENCES* doi:10.1073/pnas.2105859118/-/DCSupplemental.

78. Qiu, X. *et al.* Reversed graph embedding resolves complex single-cell trajectories. *Nat Methods* **14**, 979–982 (2017).

79. Jensen, P. & Dymecki, S. M. Essentials of recombinase-based genetic fate mapping in mice. *Methods in Molecular Biology* **1092**, 437–454 (2014).

80. Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F. D. & Klein, A. M. Lineage tracing on transcriptional landscapes links state to fate during differentiation. *Science (1979)* **367**, (2020).

81. Gould, S. J. & Eldredge, N. Punctuated equilibrium comes of age. *Nature* **366**, 223–227 (1993).

82. Hu, Z., Sun, R. & Curtis, C. A population genetics perspective on the determinants of intra-tumor heterogeneity. *Biochim Biophys Acta Rev Cancer* **1867**, 109–126 (2017).

83. Darwin, C. *On the origin of species by means of natural selection, or preservation of favoured races in the struggle for life*. (1859).