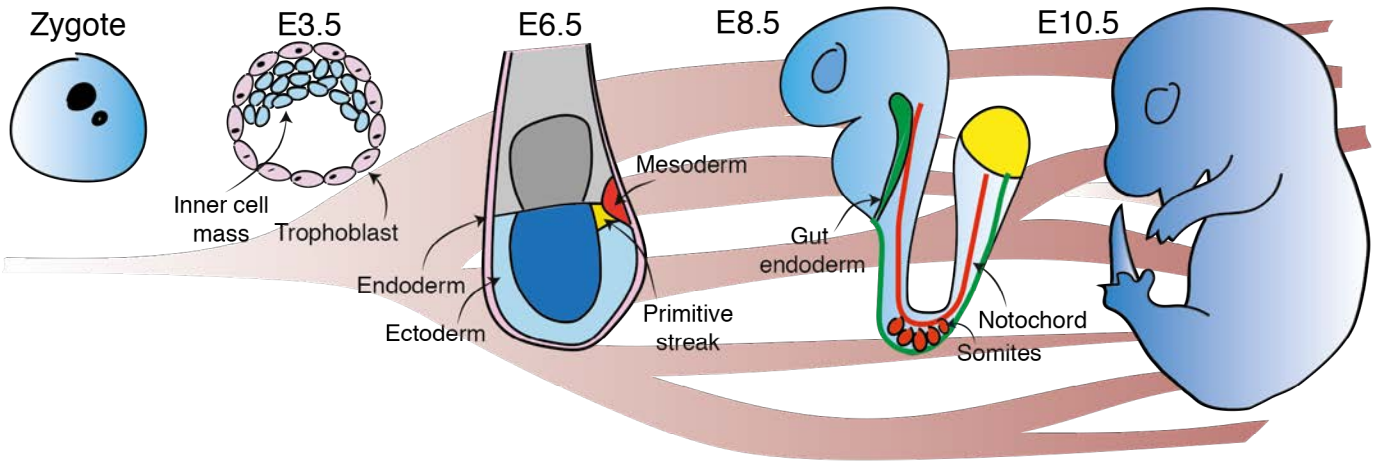


CELL DIFFERENTIATION UNRAVELLED BY SINGLE-CELL RNA SEQUENCING

Anna Alemany – https://doi.org/10.1051/epn/2020505
Oncode Institute, Hubrecht Institute-KNAW (Royal Netherlands Academy of Arts and Sciences) and University Medical Center Utrecht, Utrecht, The Netherlands.

All of us originate from a single cell, known as the zygote. Nevertheless, we are made of thousands of cells with different functionalities and morphologies: a skin cell is not the same as a neuron, yet they share the same genetic information. It is during embryo development that, through multiple cell divisions, the zygote gives rise to each of the cell types present in the different organs of each organism. One main challenge of developmental biology is to understand how, when, and where lineage commitment to each cell type takes place.

The differentiation of a cell from an initial to a final state can be viewed as a dynamical process in which the system (*i.e.*, the cell) transits through a sequence of states. This transformation occurs out of equilibrium, typically irreversibly. To characterise how the cell changes during differentiation we can explore modifications in genome architecture, cell morphology or protein content, among others. Here, we describe cellular states using the transcriptome[1], which is the collection of all RNA molecules present in a cell and can be measured using single-cell RNA sequencing (scRNA-seq). These molecules are transcribed from regions of the DNA known as genes, and encode the sequence of amino acids required to make proteins. By measuring the transcriptome of a cell we can infer which genes it is using and hence what is its main function.

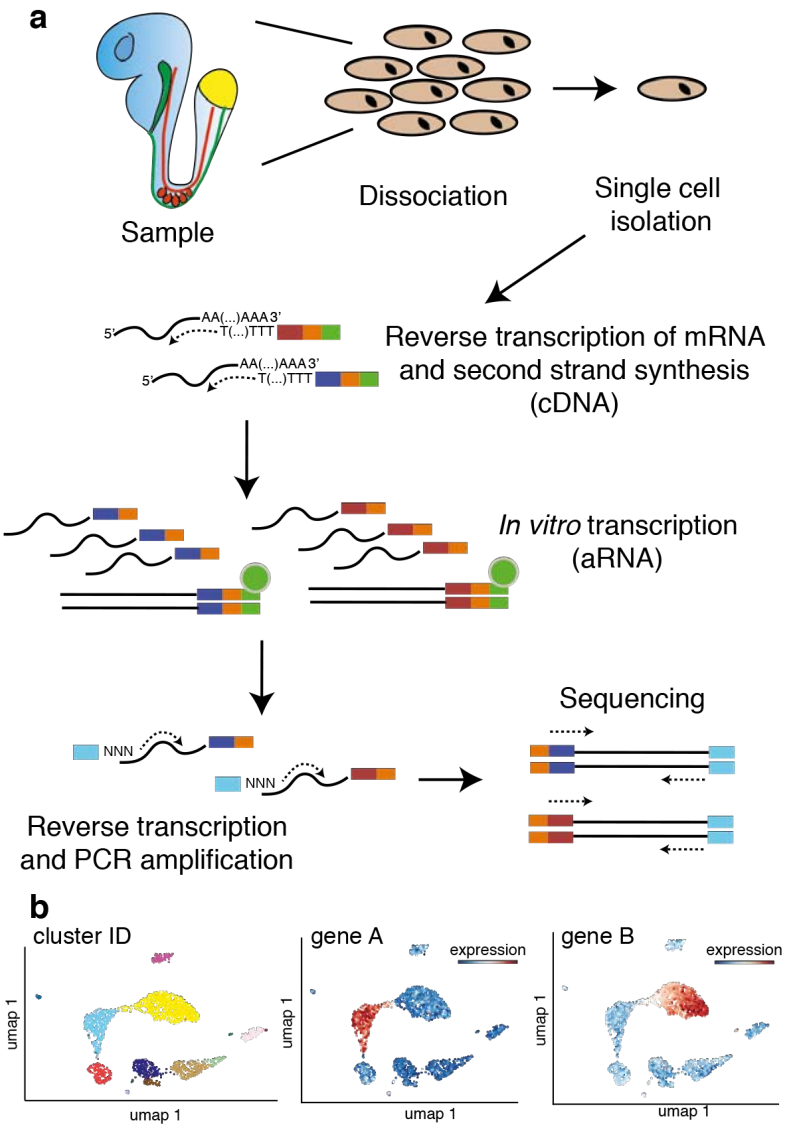
The potential landscape for cell-fate commitment

Seminal work in the field of systems biology has set the ground work to mathematically model the molecular mechanisms underlying cellular lineage commitment [2]. A cell is viewed as a Brownian particle whose state is characterised by a vector of RNA expression, $\boldsymbol{T}=\{m_1,m_2,...,m_N\}$, where each dimension is a gene. The probability $P(\boldsymbol{T},t)$ of a cell to have a transcriptional profile $\boldsymbol{T}$ at time t time satisfies the Fokker-Planck equation:

$$\frac{\partial P(\boldsymbol{T},t)}{\partial t} = \Sigma_i \frac{\partial^2}{\partial m_i^2}[D(\boldsymbol{T})P(\boldsymbol{T},t)] - \frac{\partial}{\partial m_i}[F_i(\boldsymbol{T})P(\boldsymbol{T},t)] \quad (1)$$

where $F(\boldsymbol{T})=\{F_1(\boldsymbol{T}),F_2(\boldsymbol{T}),...,F_N(\boldsymbol{T})\}$ describes the interactions between genes (known as gene regulatory network, Box), and $D(\boldsymbol{T})$ represents the stochastic diffusion between states. The solution of Eq. (1) is a high-dimensional

By measuring the transcriptome of a cell using single-cell RNA sequencing, we can infer which genes it is using and hence what is its main function



▲ **FIG. 1:** Single cell RNA sequencing provides a snapshot of the cellular heterogeneity in a sample. a) Single-cell RNA sequencing workflow [1]: the sample (tissue, organ, embryo, etc) is dissociated and single cells are isolated. Protein-coding RNA molecules, which have a poly-A tail, are uniquely labeled with an oligo-dT primer that contains a unique molecular identifier (UMI) and a cell-specific sequence (red/blue boxes), a molecular adaptor (orange) and the T7-promoter sequence (green). Complementary DNA (cDNA) is produced by reverse transcription and second strand synthesis. Next, the cDNA is linearly amplified by in vitro transcription, producing aRNA (amplified RNA). Reverse transcription is performed on the aRNA with a random primer that introduces another molecular adaptor (cyan). Molecules with the two adaptors (orange and cyan) are enriched by a polymerase chain reaction (PCR). Paired-end sequencing, in which both ends of the resulting DNA molecules are sequenced (dashed arrows) is usually performed. b) Two-dimensional projection of the generic scRNA-seq data obtained by uniform manifold approximation and projection (UMAP). Each point is a cell that has been coloured according to cluster identity assigned by hierarchical clustering (left), or by detected gene expression levels (gene A, gene B). Each cell can be assigned to a cell type based on its gene expression patterns.

manifold that connects to the Waddington landscape model (Box 1), in which a differentiating cell is represented by a marble that rolls down a landscape made of hills (transition states) and valleys (intermediate states) until reaching the final state (attractor state). We can also describe the cell state with a set of deterministic dynamic equations to describe RNA expression as a function of $F(\boldsymbol{T})$ and t:

$$\frac{\partial m_i}{\partial t} = F_i(\boldsymbol{T}), i = 1,...,N \qquad (2)$$

A minimalistic regulatory network comprising two interacting molecules can already produce systems displaying bi-/tri-stability depending on the strength of the interactions (see Box). There, a bi-potent progenitor cell can be seen as an attractor state (red solid dot) whose stability is compromised by a change in its self-activation strength. The bi-potent progenitor then becomes a metastable state (red empty dot) that can differentiate towards two distinct terminal fates (cyan dots). This approach has successfully been applied to qualitatively describe lineage commitment in the blood cell types [2]. However, most models have only been experimentally tested using well-defined *in vitro* differentiation systems where the state of a cell was inferred based on population-averaged ensemble measurements. This reduced the temporal resolution and obscured rare populations and transition states. Hence, a goal in developmental biology is to extract information about the underlying regulatory networks (*i.e.*, $F(\boldsymbol{T})$) during cell fate commitment *in vivo* with single-cell resolution and genome-wide.

Cell type identity and RNA dynamics in single cells

In multicellular organisms, cells are specialised and perform a limited set of tasks. The identity of a cell (*i.e.*, the cell type) is commonly defined by its function, and the function can be partly determined by its RNA content. Therefore, scRNA-seq experiments provide an excellent platform to identify cell types present in heterogeneous tissues, since they allow measuring the RNA content in thousands of single cells independently (Fig. 1a) [1]. Overall, scRNA-seq provides a snapshot of the cellular heterogeneity in a sample, and can reveal rare or transient subpopulations usually undetectable by bulk sequencing approaches. Because of the huge number of genes in an organism (both the human and the mouse genome have ~20,000 protein-coding genes), datasets are typically large and sparse. It is a priority to develop analytical methods that exploit these high-dimensional data (each dimension being a gene) to compare transcriptome profiles between different cells. At the moment, there are several pipelines to normalise the data, reduce dimensions, cluster similar cells together, extract differential gene expression profiles, visualise results, and ultimately call cell types (Fig. 1b) [1].

In scRNA-seq, the sample is destroyed when the cells are prepared for sequencing. Therefore, these experiments

are not suitable to measure temporal changes in the transcriptome of the same cell. To circumvent this, experiments are performed on samples where several differentiating cells coexist at different stages (such as embryos), or in biological replicates at different time points, to ensure that all stages of the cellular differentiation trajectory are represented [3].

With the possibility to measure full transcriptomes in single cells undergoing differentiation, new algorithms to investigate the lineage-commitment dynamics have been developed [3]. Usually, it is assumed that cells change their transcriptome continuously as they transit from one state to another. Consequently, cells with similar transcriptomic profiles can be connected together and ordered (Fig. 2a). This gives rise to the concept of pseudo-time as a latent dimension that quantifies the degree of differentiation along a cellular transformation. These approaches are currently challenged by results obtained from lineage tracing studies in which both the transcriptome and the common ancestry of single cells are simultaneously measured (Fig. 2a) [4]. There, it has been observed that transcriptomically identical cells can derive from ancestors with different RNA profiles, suggesting that cell differentiation has a stochastic component in which fluctuations in gene expression play an important role and cells with similar transcriptome can have different lineages.

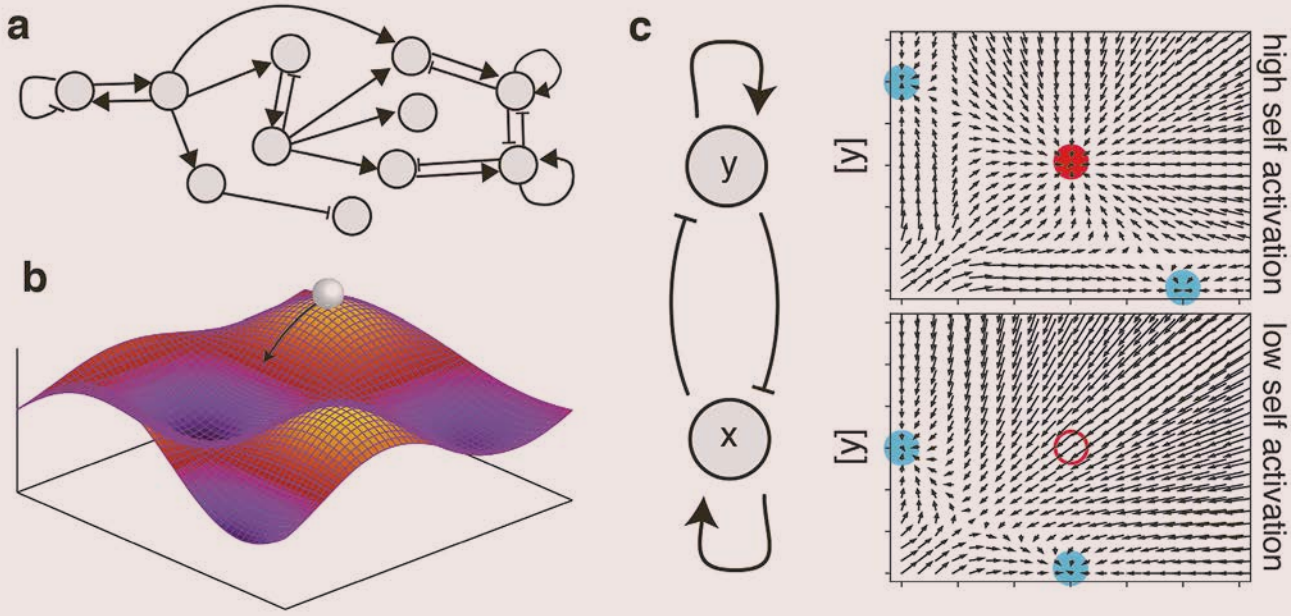
In order to assign differentiating cells to a pseudo-time value more accurately, experimental protocols and analysis pipelines have been improved to distinguish newly synthesised (u) transcripts from mature (m) ones (Fig. 2b) [5,6]. Then, we can extract the transcription, degradation and splicing rates, α, γ and β respectively, for each transcript, by solving the rate equations:

$$\frac{du(t)}{dt} = \alpha - \beta u(t)$$
$$\frac{dm(t)}{dt} = \beta u(t) - \gamma m(t) \qquad (3)$$

BOX: CELL DIFFERENTIATION IS REGULATED BY GENE NETWORKS

Mouse embryo development starts right after fertilisation, when the zygote is formed and starts dividing. After ~3.5 days (E3.5) the first commitment event (origination of the inner cell mass and the trophoblast) occurs in the blastocyst. At E6.5 the three germ layers (endoderm, ectoderm and mesoderm) and the three body axes are established via the primitive streak. Around E8.5-E10.5, organ progenitors appear. During these processes, gene expression patterns in single cells are dynamic and interact with each other via gene regulatory networks.

a) Gene regulatory network, where genes are represented as nodes and interactions are represented as edges. Repressive and active interactions are represented with butts and arrows, respectively. **b)** Sketch of the Waddington landscape, in which cell differentiation is illustrated through the position of a ball in a landscape. The ball (*i.e.* the cell) is initially situated at the top of the mountain and it can only roll down via multiple paths. c) (left) Gene regulatory network for two genes x and y, that behave as auto-enhancers but repress each other. The dynamic equations for this network can be written assuming $F_x(x,y) = ax^n/(S^n+x^n)+bS^n/(S^n+y^n)-kx$ and $F_y(x,y)=ay^n/(S^n+y^n)+bS^n/(S^n+x^n)-ky$ (Eq. 2) [2]; (right) Corresponding phase portrait, obtained for S=0.5, b=1, k=1, n=4 and a=1.5 (top; high self-activation) and a=0.5 (bottom; low self-activation). Solid (empty) dots indicate attractor (metastable) states.



It has been proved that the experimental estimation of the time derivative of the mature RNA —termed RNA velocity— can predict the future state of individual cells on a timescale of hours [5]. RNA velocity can be embedded in the two dimensional representation of scRNA-seq data, in the form of a phase portrait from which one can visualise directions of differentiation at the cellular level (Fig. 2c).

A new challenge of developmental biology is to experimentally connect Eq. (2) and Eq. (3), and link RNA velocity to F(**T**) to unravel the regulatory networks that drive cell-fate commitment. This will allow experimentally measuring the Waddington landscape, with the goal to ultimately predict cellular responses to stimuli.

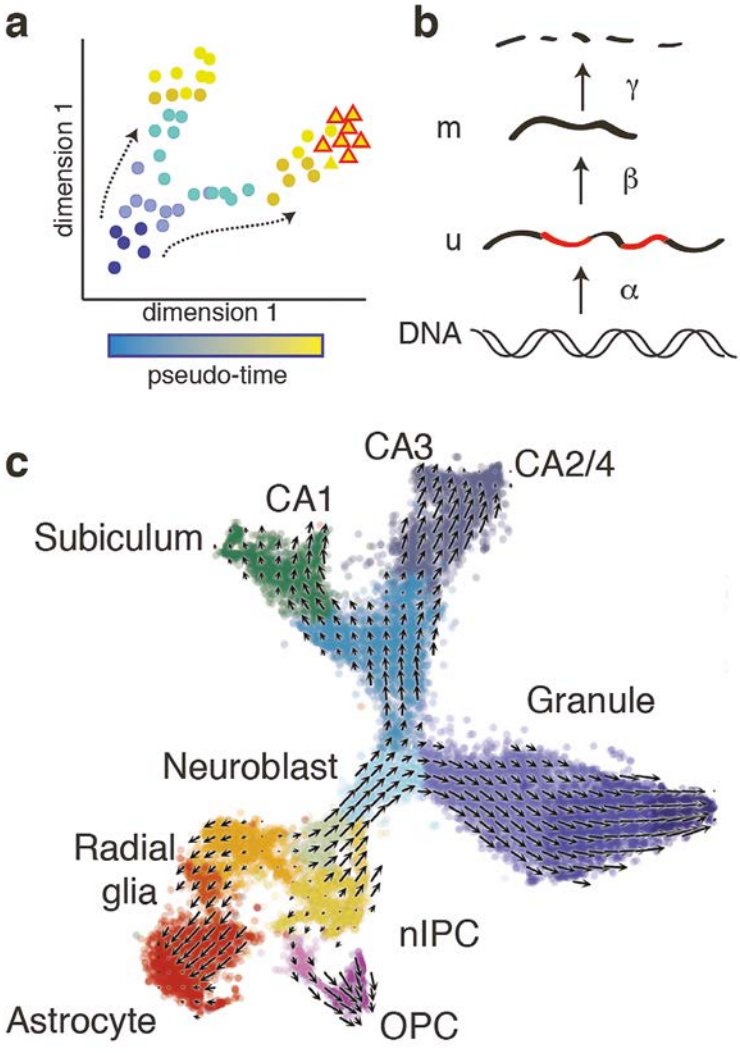
Future perspectives: lineage tracing, spatial transcriptomics and perturbation assays

The characterisation of gene regulatory networks and cellular transition states occurring during cell differentiation and their role in cell-fate commitment is a long-standing question in developmental biology. We anticipate that the combination of RNA velocity with theoretical models of molecular phase portraits will help us unravel the fundamental gene regulatory networks that give rise to new cell fates. With the help of lineage tracing and perturbation studies (such as chemical screens or genetic knock-outs), we will be able to validate our findings and test predictions. In combination with spatial transcriptomics [7], we expect to also unravel underlying principles of intracellular communication.

The ongoing revolution in single-cell sequencing technologies is making it possible to extract other epigenetic properties in single cells, and to integrate them into multi-modal measurements from the same cell. With this, we expect to enhance the characterization of the Waddington landscape of development, with all its epigenetic components. ■

About the Author

Anna Alemany received her PhD in the group of Prof. Ritort and afterwards joined Prof. van Oudenaarden laboratory for her postdoc. She will start her own laboratory in the Leiden Medical University Center, in the Netherlands. Her goal is to quantitatively characterise the Waddignton landscape of development using single-cell RNA sequencing experiments.



◂ **FIG. 2: Reconstruction of differentiation trajectories from scRNA-seq. a) Sketch of a two-dimensional embedding of single cells based on their transcriptome. Each point is a cell and has been coloured according to pseudo-time assignment. Cells labeled as a triangle have a high pseudo-time value. However, lineage tracing studies indicate that they have a different clonal origin from the cells they are connected to in the plot. b) The RNA life cycle starts with transcription from the gene. Newly synthesised RNA is made of nucleotide sequences termed as introns (red) and exons (black). Introns are spliced out to yield mature RNA. Finally, RNA gets degraded. c) Two-dimensional projection of cells isolated from the hippocampus (figure adapted from [5]). Each point is a cell, coloured according to cluster identity. Cell types are annotated. The arrows show the local average RNA velocity, evaluated on a regular grid, and indicate directions of differentiation.**

References

[1] A. Kulkarni, A. G. Anderson, D. P. Merullo and G. Konopka, *Current opinion in biotechnology* **58**, 129 (2019).

[2] J. Wang, L. Xu, E. Wang and S. Huang, *Biophysical journal* **99(1)**, 29 (2010).

[3] D. E. Wagner, C. Weinreb, Z. M. Collins, J. A. Briggs, S. G. Megason and A. M. Klein, *Science* **360(6392)**, 981 (2018).

[4] A. Alemany, M. Florescu, C.S. Baron, J. Peterson-Maduro and A. Van Oudenaarden, *Nature* **556(7699)**, 108 (2018).

[5] G. La Manno, R. Soldatov, A. Zeisel, E. Braun, H. Hochgerner, V. Petukhov, K. Lidschreiber, M. E. Kastriti, P. Lönnerberg, A.Furlan, and J. Fan, *Nature* **560(7719)**, 494 (2018).

[6] N. Battich, J. Beumer, B. de Barbanson, L. Krenning, C. S. Baron, M. E. Tanenbaum, H. Clevers and A. van Oudenaarden, *Science* **367(6482)**,1151 (2020).

[7] P. L. Ståhl, F. Salmén, S. Vickovic, A. Lundmark, J.F. Navarro, J. Magnusson, S. Giacomello, M. Asp, J. O. Westholm, M. Huss and A. Mollbrink, *Science* **353(6294)**, 78(2016).