1 The cell

The cell is the smallest unit of life, of which all known living organisms are composed. Every cell houses a plethora of biomolecular processes that allows it to continuously adapt to changes in its environment. Due to the dynamic nature of these processes, it can be very challenging to comprehend the cellular response to a signal. A reductionist approach to understanding a complex biological system is to study the biochemical components of which it is comprised[1].

Recent advances in experimental technologies are playing a crucial role in reductionist biology, allowing to measure the abundance of thousands of different biochemical molecules in tens of thousands of individual cells. With it comes the challenge of analysing large amounts of data that are not easily interpretable by hand. The sheer volume of the data generated from such highly-integrative and high-throughput experiments are not the only reason why they are so challenging to interpret. For instance, the generated data contains high levels of noise arising from inherent biomolecular stochasticity in the cells and from the experimental profiling techniques used, as well as batch effects arising from differences between donors and labs[2]. Biologists thus turn to computer scientists to develop new tools to tackle these problems and help them to extract meaningful biological insights from the data. In this work, incremental contributions were made to the field in order to be able to address the aforementioned problems in a more comprehensive context.

Observing the biomolecular insides of cells can ultimately provide fundamental understanding into the processes that govern these cells and help uncover novel approaches for disease diagnosis, prognosis, and treatment. For example, the Human Cell Atlas (HCA) consortium[3] has set out to develop a comprehensive reference map of all the different types of cells in the human body. Experts in the field often metaphorically describe the HCA initiative as aiming to develop a 'Google Maps' of the human body. Even in its infancy, the HCA has profiled 3.8 million cells from 248 donors across 42 labs[4], and this number is likely to increase well above one hundred million.

The next part of the chapter highlights several key concepts in both cell biology and computer science, upon which the remainder of this work relies.

1.1 The origin of life and the RNA world

The discovery of the double helix shape of deoxyribonucleic acid (DNA)[5] is often considered the pivot point in our understanding of the origin of life and evolution. By now, it is well known that DNA serves as a medium for storing the genetic information required to reproduce a whole organism. With other words, the DNA of an organism contains the complete set of instructions required to build all of the biomolecular machinery present in its body.

Life (or cells) did not originate from DNA, however. A widely-accepted hypothesis states that life originates from its lesser-known cousin, ribonucleic acid (RNA). According to the RNA world hypothesis[6], the very first primitive cells used RNA both to store genetic information and to perform the chemical reactions required to sustain themselves (Figure 1). Only later

did cells develop the ability to use the more chemically stable DNA molecules to self-sustain in a process commonly referred to as the central dogma.

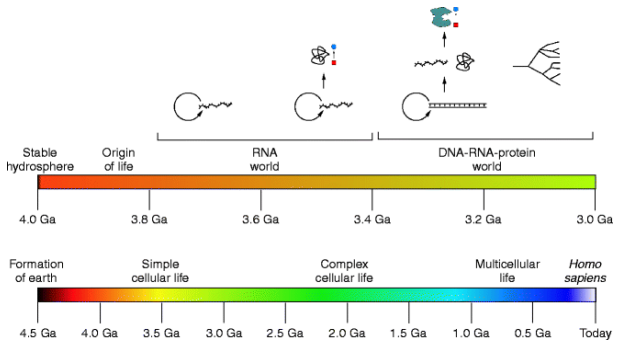

Figure 1: RNA world. The postulated rise and fall of the RNA world during the evolution of life, from early self-replicating RNA to complex, RNA-controlled metabolism, to the invention of translation, followed by diversification of all modern branches of life. Image from Horning (2011)[7]. TODO: combine images, rewrite the description.

1.2 Central dogma

The central dogma describes the general flow of genetic information in almost all existing living cells: DNA is decoded to RNA, which in turn encodes proteins[8]. Main processes involved in the central dogma are **transcription**, **splicing**, and **translation** (Figure 2).

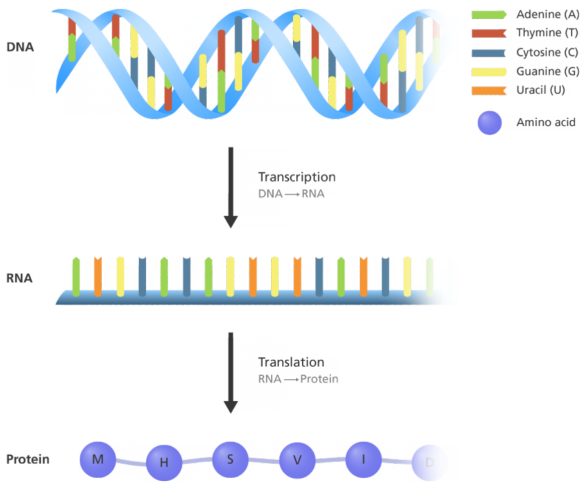

Figure 2: Central Dogma. TODO: combine images, rewrite the description.

During the process of **transcription** that takes place in the cell nucleus, a complementary RNA copy is transcribed from the template DNA. The initial RNA transcript is a precursor messenger RNA (pre-mRNA) that needs to undergo series of maturation steps to ultimately form the mature messenger RNA (mRNA). This maturation includes pre-mRNA **splicing** to remove non-protein coding intervening sequences (the introns) and to join the neighbouring protein-coding sequences (the exons). A single pre-mRNA can be alternatively spliced to generate multiple forms of mRNAs that will result in the production of multiple protein isoforms. This

process of alternative splicing is essential to generate more than 100'000 different proteins starting from just 20'000 genes[9].

The mature mRNA is then transported to the cytoplasm, where it engages with ribosomes to initiate **translation**. During this highly evolutionary conserved process, a chain of amino acids, known as the protein building blocks, is being synthesised. Each amino acid is specified by three nucleotides (a codon) in the mRNA, according to a nearly universal genetic code. After being released by the ribosomes, the translation product undergoes a variety of chemical modifications to form the final folded protein, the structure of which is determined by the sequence of different amino acids in the chain. In addition, polypeptides may be cleaved to yield more than one active polypeptide product. The structure of a protein determines its functionality, which includes catalysing biochemical reactions, providing structure, and transportation of molecules.

1.3 Cell types

Ever since Robert Hook first described the different structures of cells in 1665, biologists have been classifying cells by form and function. The human body is said to contain more than 210 different cell types that are classified into four groups: epithelial, connective, muscle, and nervous. This however, is a major underestimation of the real number of cell types. Neurons, for instance, that are known to be extremely diverse, are estimated to reach numbers above 10,000 different types[8].

The concept of cell types eases reasoning and our understanding about many aspects of biology (e.g. the process of cell differentiation, cell-cell communication, cellular response to certain stimuli). Some cells are known to be highly specialised toward performing a particular function (e.g. memory B cells accelerate immune response by remembering previously encountered pathogens), or they can maintain a strong ability to differentiate into other cell types.

One common approach for understanding the functionality of a particular cell is to observe which molecules are present in the cell and to associate those set of molecules with functionality. Taking a snap shot of the protein or RNA transcript content in a particular cell, might already provide us with major insights into its functionality. However, in order to fulfil a particular task, the biochemical machinery of the cell gradually changes over time. Therefore it is highly informative to also consider the transition states between cell types and the dynamic processes involved therein.

1.4 Cell dynamics and gene regulation

Cells are dynamic entities that can gradually produce the molecules needed to acquire new functionality. The naturally occurring cell-to-cell variability happens at the level of gene expression. Gene expression itself can be controlled at different levels (Figure 3), one of which is gene regulation by transcription.

According to the needs of a cell, different genes are being transcribed. Housekeeping

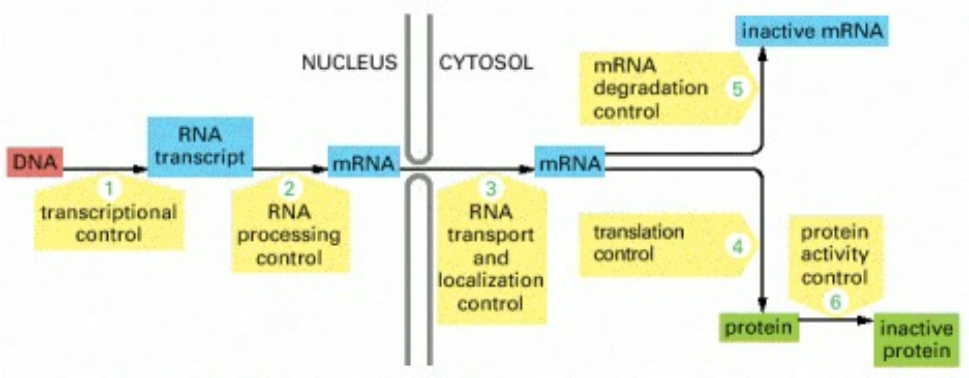

Figure 3: Levels of controlling gene expression can happen at the level of transcription, RNA processing (i.e. splicing), RNA transport and localization, mRNA translation, mRNA degradation and protein activity [10].

genes are being expressed in essentially every cell, while other genes are cell type or tissue specific or may be expressed in response to developmental and environmental signals[8].

Transcription factors (TFs) modulate the rate of gene transcription by binding and recruiting the transcriptional machinery to *cis*-regulatory regions (enhancers, and silencers) that are typically located in the promotor region of target genes. These bindings may result in increased or decreased gene expression. There are several TF families of which members share structural characteristics (e.g. zinc finger, helix-loop-helix).

Many TFs are commonly present in virtually all cell types (e.g. NF-κB), while others are specific for cells and developmental stages[11]. Typically, the same TF can regulate the rate of transcription of many target genes in different cell types, indicating that these gene regulatory networks (GRNs) are dynamic. Moreover, the production of a specific molecule might require several gene regulatory cascades. Studying the active parts of a cell's GRN can thus reveal which dynamic processes are taking place within a cell.

1.5 Profiling single cells

Several technologies are now available to profile (i.e. observe) biomolecular components, allowing us to gain better understanding in the biological processes that take place within a cell. The single-cell "omics" technologies originated from the convergence of two different fields, *"single-cell"* and *"omics"*.

1.5.1 Single-cell

The earliest approach for measuring the abundance of a particular molecule in *single cells* is the microscope. Since its development by Coons et al. (1941), immunohistochemistry (IHC) has been instrumental in visualising proteins.[12]. A cell can present a particular type of protein, also called an antigen, on its cell surface. In many multicellular organisms, antigens can stimulate the immune system to produce antibodies. IHC realises the visualisation of proteins by exploiting the principle of antibodies binding to specific antigens.

IHC (and many other biotechnologies) visualises antigen-antibody reactions by attaching

particular molecules to the antibody, such as an enzyme that catalyses a colour-producing reaction, or a fluorescent chemical compound that can re-emit light upon excitation. The use of several colours (wavelengths) allows measuring expression levels of different antibodies simultaneously. Characterising cells in a semi-quantifiable way is labour intensive, however; since it involves acquiring an image of many cells and drawing a contour around each cell (called cell segmentation). Modern implementations of IHC improve the throughput drastically by using robots and computer software to provide semi-automated image acquisition and cell segmentation[13].

Flow cytometry[14] circumvents imaging and segmentation issues by measuring fluorescently labelled proteins as cells pass through a fluidic system. Since cells need to be suspended in a buffer, flow cytometry is particularly useful for analysing non-adherent cells such as the many different immune cells in blood. However, many protocols already exist to extract viable single cells from tissues and tumours[15]. Conventional flow cytometry devices enable to measure protein expression levels of millions of cells using up to eight different antibody fluorochromes simultaneously, while state-of-the-art instrumentation allows detection of up to 27 biomarkers simultaneously[16].

Besides IHC and flow cytometry, many new technologies have been developed which allow quantifying expression levels of molecules in single cells (e.g. mass cytometry, single-cell quantitative polymerase chain reaction, fluorescence *in situ* hybridization). All of these single-cell (non-omics) technologies are limited by the number of different molecules they measure, however. Selecting molecules of interest prior to analysis, makes the experiment biased towards the preconceptions of the experimenter.

1.5.2 Omics

On the other side of the spectrum are the so-called "omics" technologies. "Omics"[1] is a collective term for profiling all molecules of a particular type in a high-throughput manner. There are at least ten types of "omics". In this work, we mostly consider genomics, transcriptomics, proteomics, and regulomics. Genomics studies the complete DNA sequence of an organism's genome, while transcriptomics and proteomics study the RNA transcripts and proteins, respectively. Regulomics studies the regulatory molecules (e.g. genes, RNAs, proteins) which play a role in determining gene regulation.

Specific examples of omics technologies are whole genome sequencing to determine the DNA sequence of an organism, and RNA sequencing to profiles the sequence of RNA transcripts, both using next-generation sequencing technologies. A gene expression profile can be obtained by mapping the sequences of RNA transcripts to the genome.

Several high-throughput technologies have been developed to investigate proteomes in depth. The most commonly applied are mass spectrometry-based and gel-based techniques (e.g. differential in-gel electrophoresis).

Typically for these methods, to capture enough material to generate a profile, numerous cells need to be pooled and lysed together, thereby granting the technology's name "bulk"

[1]The etymology of "omics" is quite interesting[17].

omics. Bulk omics is a major workhorse in molecular genetics and has applications in cancer research and in diagnostic screening of inheritable disorders.

Increasing evidence shows that cells are biomoleculary heterogeneous, even in very similar cell types[18] (Figure 4A). Since a bulk profile is a population average (or rather, a summation), important cell-to-cell variability is not discernible (Figure 4B).

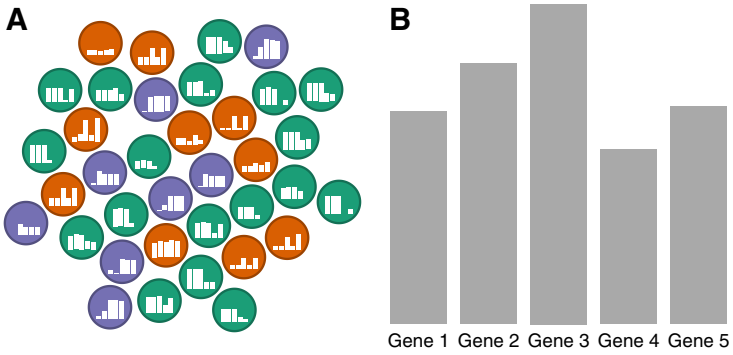

Figure 4: The 'masking' effect of bulk omics. A: Cells from several subpopulations are incorrectly assumed to be homogeneous and are profiled with a single bulk omics experiment. **B:** The signals from the different subpopulations are masked. The resulting profile is dissimilar from the majority of cells it is supposed to represent.

1.5.3 Single-cell omics

Comparing single-cell technologies with omics technologies shows that they have both clear advantages but also significant drawbacks (Figure 5A). Single-cell biology allows profiling thousands or even millions of cells, but only for a select number of genes. On the other hand, omics biology provides a broader view – since genes do not need to be selected beforehand – but is a profile of ensemble of cells and thus masks important cellular heterogeneity.

Advances in microvolume sequencing allowed profiling the transcriptome at single-cell resolution, thereby bringing single-cell biology and omics together to create single-cell omics. During the decade that followed, the number of single-cell omics technologies has skyrocketed, allowing to profile >100'000 cells[19] and measuring other levels of information (e.g. protein abundance and spatial location) [20].

In this work, unless noted otherwise, we will be working with transcriptomics data resulting from a single-cell RNA-sequencing experiment (scRNA-seq). The workflow of generating scRNA-seq profiles is as follows. Same as other single-cell (non-omics) profiling methods, cells first need to be isolated (Figure 5B). Different sampling techniques yield different levels of information about cellular state. By now, many protocols for extracting and tagging RNA from single cells have been developed[19], the most popular of which are based on microfluidics or droplets (Figure 5C). By sequencing the trancripts and the attached unique cell identifier tags, each read can be mapped and tallied up. scRNA-seq data can thus be summarised in a matrix, where each column represent a single cell, each row a gene, and each value represents the number of transcripts that were sequenced for that gene and cell.

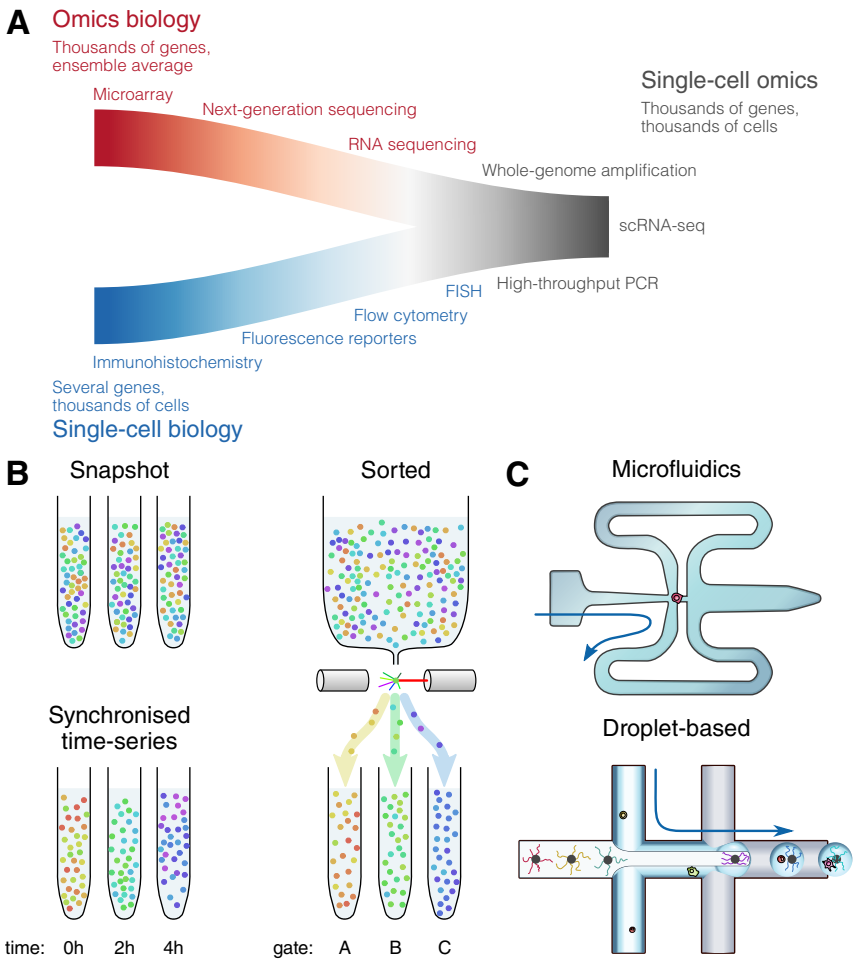

Figure 5: A: Convergence of single-cell and omics biology. **B:** Different approaches for sampling cells with decreasing levels of cellular heterogeneity within the different sub-populations: snapshot, time-series, sorted. **C:** Two common single-cell RNA sequencing technologies. Microfluidics systems let cells travel through nanometer scale tubing, capturing individual cells at intersections. Droplet-based systems encapsulate individual cells in droplets.

The rapidly advancing field of single-cell omics harbours exceptional opportunities to discover new aspects of biology and redefine existing knowledge. Some of these opportunities lie in efforts like the HCA consortium[3]. They have set out to redefine all human cell types in both terms of their gene expression and location, and as well as the developmental trajectories connecting the different cell types. As part of this endeavour, the consortium will likely profile the whole transcriptomes of tens or even hundreds of millions of cells[4].

2 Computational tools

Whole-genome profiling at single-cell level allows new types of analyses with which to study cellular heterogeneity at a hitherto unseen throughput. The new types of analyses permitted by single-cell omics present several computational challenges[21, 22, 23]. This necessitates the development of novel computational tools, either because the problem statement of the performed analysis is completely novel, or to adapt existing methodology to new data characteristics – dimensionality and noise.

scRNA-seq data is typically very sparse – while the human genome has more than 20'000 genes, they only contain non-zero values of a few thousand genes (typically <4'000). This is partially due to cells being specialised in particular functions and thus they do not need proteins of every time, but also due to RNA transcription occurring in bursts rather than continuously[24]. This contributes to the high levels of noise seen in scRNA-seq data: no two cells have the same set of non-zero genes.

Over the past five years, already 450 new tools have been developed to perform various analyses of single-cell omics data[25], taking into account the specific noise characteristics. The most frequent types of analyses are detailed in the following subsections.

2.1 Dimensionality reduction

Single-cell omics datasets are usually one or more high-dimensional matrices, containing between M = 10^3 to 10^5 cells and typically about N = 10^3 to 10^4 genes (Figure 6A). The dimensionality of such datasets is typically too high for humans to interpret manually and for most modelling algorithms to tackle directly. Moreover, in reality, the intrinsic dimensionality of biological systems is probably much lower. For example, a differentiating hematopoietic cell could be described by just three dimensions: the first two dimensions lays out the hematopoietic lineage tree, and a third dimension allows for reprogramming between branches to occur.

Dimensionality reduction (DR) methods transform high-dimensional data into a meaningful low-dimensional representation. DR methods have two main target audiences; computers – to construct a K-dimensional (with K $\ll$ M) representation of the data such that pairwise distances between different samples are retained as well as possible (Figure 6B); and humans – to obtain a visual overview of the data (Figure 6C).

Each dimension frequently corresponds to one or more modules of co-expressed genes. The reduced space can be interpreted in way analogous to Waddington's epigenetic landscapes[26, 27, 28], where the landscape dictates the possible states a cell can reside in, and transitions between states correspond to dynamic cellular processes, such as cell differentiation.

DR methods can be classified into two main categories, feature projection-based and manifold learning[30].

Projection-based DR methods aim to perform a transformation of the data while preserving the pairwise distances between samples as much as possible. Examples of commonly used projection-based DR methods in single-cell omics are Principal Component Analysis[31] (PCA) and Multi-Dimensional Scaling[32] (MDS).

Manifold learning DR methods reconstruct a higher-order structure in the original space (e.g. a graph or a grid), visualising the structure in a lower-dimensional space, and mapping the original samples to the lower-dimensional space. Manifold learning can be an iterative optimisation process using a predefined criterion. Examples of manifold learning techniques are t-distributed Stochastic Neighbor Embedding[33] (t-SNE), Diffusion Maps[34, 35] and Uniform Manifold Approximation and Projection[36] (UMAP).

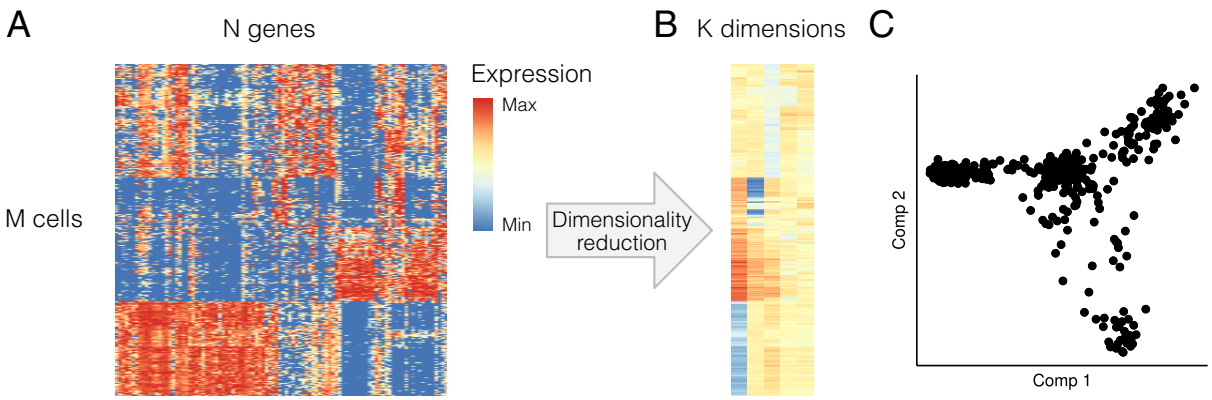

Figure 6: Dimensionality reduction for single-cell omics data. A: A heatmap visualisation of an scRNA-seq expression dataset of fibroblasts being reprogrammed to neuron cells[29]. Only the most variable **B:** The reduced space is a M×K-dimensional matrix which attempts to conserve the cellular heterogeneity of the original space as well as possible. **C:** A dot plot of the first two components of the reduced space provides a global overview of the cells in the dataset. Colouring the dots according to prior information (e.g. cell type) or gene expression provides insight into the cellular heterogeneity within the dataset.

For scalability reasons, this work mostly makes use of Landmark MDS[37, 38] (LMDS) with a Spearman rank correlation distance metric. LMDS is an extension of classical MDS, but rather than calculating a complete distance matrix between all pairs of cells, a set of landmark cells is sampled, only the distances between a set of landmarks and the samples are calculated.

2.2 Clustering

To learn about the different cellular states within a population of cells, clustering methods divide up the cells into separate groups of highly similar cells (Figure 7A). By visualising gene expression of known genes involved in the cell types of interest, the clusters can be annotated (Figure 7B).

Usually, the number of clusters is determined by the user, either as a direct parameter (e.g. k-means[39]) or an indirect parameter (e.g. a height cutoff in hierarchical clustering). In some exceptional cases, the number of clusters is strictly determined by the data itself and cannot be altered with a parameter (e.g. Louvain clustering[40]).

Clustering methods used in this work are mostly restricted to k-Means for clustering low-dimensional spaces and Louvain for clustering networks, since both are highly scalable with respect to the number of cells.

2.3 Trajectory inference

While clustering methods divide cells into distinct groups, trajectory inference (TI) methods acknowledge that cells are dynamic entities which transition from once cellular state to another via various dynamic processes. Rather than making distinct groups, TI methods allow studying dynamic processes by reconstructing the topology of a dynamic process as a trajectory, and map the cells onto that trajectory. A trajectory is a graph where the nodes represent

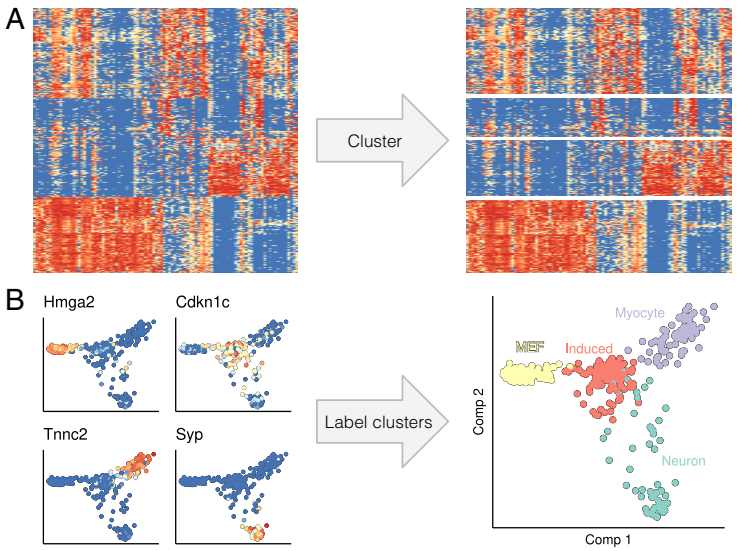

Figure 7: Clustering for single-cell omics data. **A:** Clustering methods group cells with similar omics profiles together. **B:** By overlaying gene expression levels on a dimensionality reduction, the clusters can be annotated to allow better interpretation of the cellular heterogeneity.

noteworthy cellular states, and each cell is predicted to be progressing along transitions between the different states (Figure 8A).

A trajectory can be visualised as a graph to better highlight the topology of the trajectory(Figure 8A middle), as a heatmap to better depict the changes in gene expression along the different transitions (Figure 8A right), or may be embedded in a dimensionality reduction of the cells to better demonstrate cellular heterogeneity along the trajectory (Figure 8B right). Similar to clustering, by colouring the cells according to the expression of genes known to be involved in the dynamic process of interest, the milestones in the trajectory can be annotated (Figure 7B).

A lot of TI methods use similar algorithms to be able to infer a trajectory. By breaking down each method into its set of core algorithms, we can create a map of TI methodology[42] (Figure 9A), which is divided into two major classes. In the first step, dimensionality reduction techniques such as manifold learning, clustering, or graph-based methods are used to convert the dataset to a more simplified representation. This representation of the data then allows the trajectory itself to be more easily modelled in a second step. In this second step, the trajectory is modelled within the data using graph-based or curve-based approaches, after which the cells themselves can be ordered using a variety of methods.

A common way to classify TI methods is by the types of trajectories they can infer[43] (Figure 9B). About half of TI methods specialise in inferring linear or cyclic trajectories (i.e. they order the cells). Others model the trajectory as a rooted tree, allowing for one or more bifurcations to occur. Only a few methods are able to infer more generalised trajectories containing disconnected subgraphs or cycles.

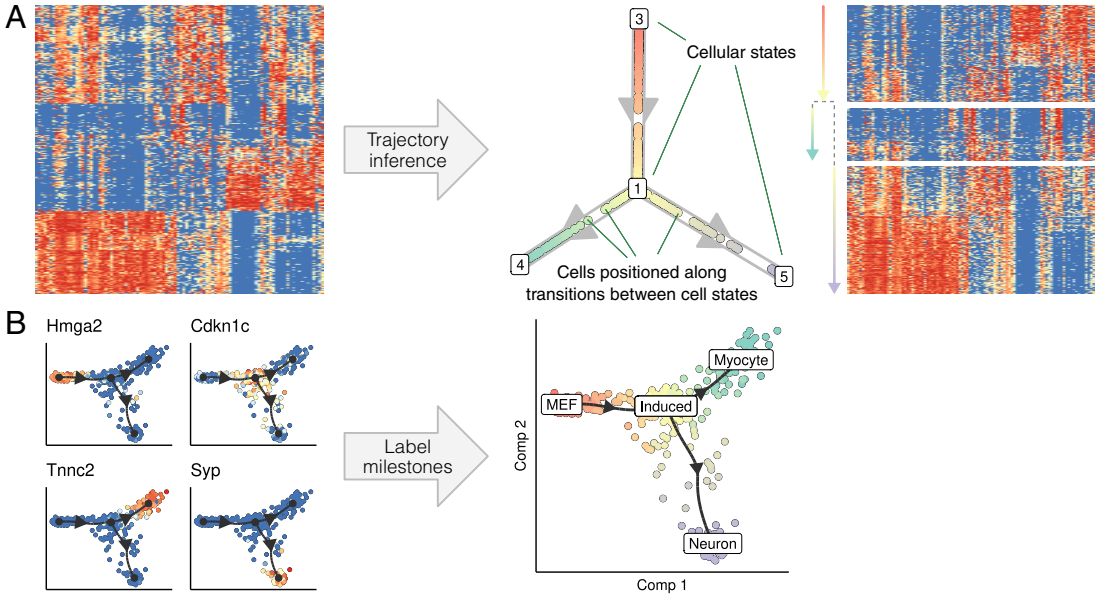

Figure 8: Trajectory inference for single-cell omics data. A: During a dynamic process cells pass through several transitional states, characterized by different waves of transcriptional, morphological, epigenomic and/or surface marker changes[41]. TI methods provide an unbiased approach to identifying and correctly ordering different transitional stages. **B:** By overlaying gene expression levels on a dimensionality reduction, the milestones can be annotated to allow better interpretation of the cellular heterogeneity.

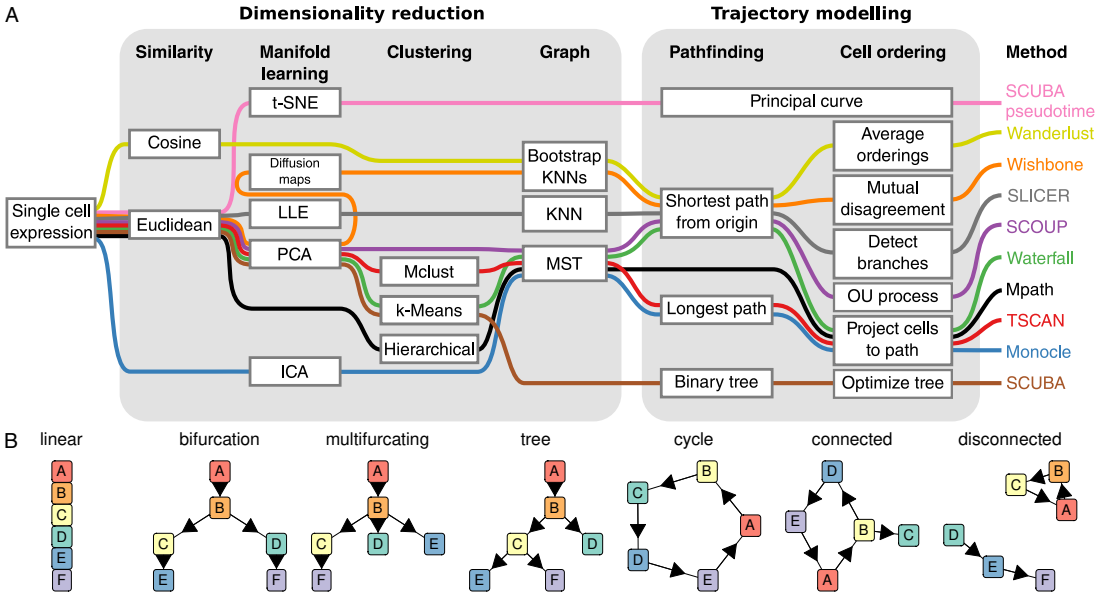

Figure 9: TI methods use several common building blocks and can be organized in a unifying modular framework. A: Most TI methods consist of two major steps, dimensionality reduction and trajectory modelling. TI methods require some form of dimensionality reduction in order to summarise cell heterogeneity in a lower dimensional space. Subsequently, a trajectory modelling step then operates in this reduced space, aiming to identify cell states, constructing a trajectory through the different states, and projecting the cells back on to the trajectory. **B:** TI methods can be classified according to the trajectory topologies they can infer.

2.4 Differential expression

Given that cells are split up into groups differential expression (DE) methods ranks genes based on whether their expression is significantly higher or lower in one group in comparison

to the others. This grouping can be based on prior information or an upstream clustering method. DE methods are useful for summarising the main differences between different groups of cells more compactly (Figure 10A) in comparison to when groups are compared without gene prioritisation (Figure 7A).

Trajectory differential expression (TDE) is an extension of DE where instead genes are prioritised according to whether their gene expression changes smoothly but significantly along a parts of a trajectory (Figure 10).

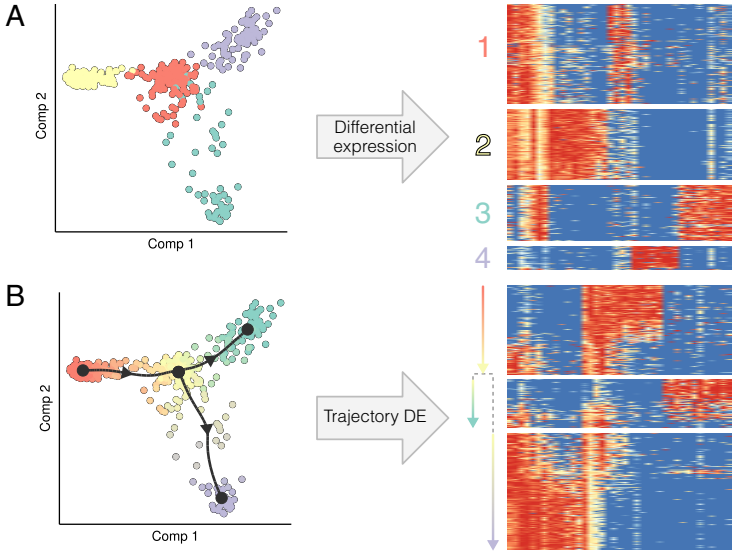

Figure 10: Differential expression for single-cell omics data. A: Differential expression methods prioritises genes that are expressed significantly higher or lower in particular given groups. **B:** Trajectory differential expression prioritises genes that change smoothly but significantly along particular transitions in a trajectory.

2.5 Network inference

One of the central cellular processes underlying development is transcriptional regulation. Modeling the dynamics of gene regulation is therefore essential to better understand why a cellular dynamic processes progresses through several steps, and what goes wrong in the case of disease.

Network inference (NI) methods predict which genes are regulated by which transcription factors (Figure 11). The output of a network inference is thus a graph, where nodes represent genes and edges denote a regulatory interaction between a regulator and a target gene. Interactions typically have two properties: its regulatory strength (a positive real value) and its effect (promoting or repressing).

Before single-cell omics, these methods rely on multiple experiments, amongst which perturbation and time-series experiments, to predict the effect each transcription factor has on the up- or downregulation of a gene. One of the main advantages of single-cell omics is the heterogeneity between cells caused by naturally occurring biological randomness[44] can be exploited to infer regulatory interactions between TFs and their target genes at much lower costs.

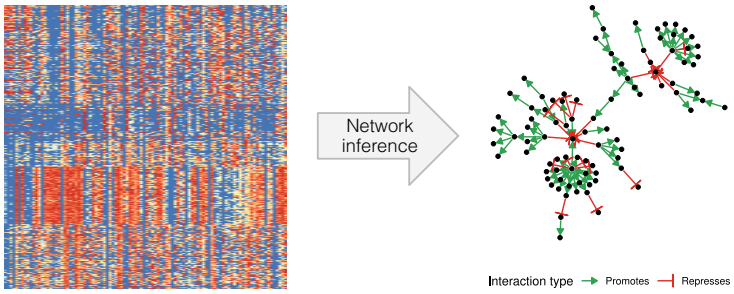

Figure 11: Network inference for single-cell omics data.

3 Research context and objectives

Recent technological advancements in profiling single cells are having significant repercussions in many fields of biology. Profiling thousands of individual cells in a genome-wide manner provides opportunities to study cell heterogeneity and dynamics, for example inferring mechanisms for cellular development or intercellular communication. Hundreds of new software tools were developed[25] to perform these new types of analyses, or to fit existing analytical tools to deal with new data characteristics (e.g. differential expression, dimensionality reduction, normalisation).

One major shortcoming during the advent of single-cell omics was that the majority of newly developed computational tools were not quantitatively and comparatively evaluated. Rather, they relied on anecdotal evidence to demonstrate its usefulness. This issue is not the result of the tool developer's malevolence, but instead of the lack of data required to perform such comprehensive benchmarks.

Uncontrolled development of software tools without comprehensive benchmarking poses serious problems. For one, it slows down scientific progress. Every end-user needs to make a large commitment researching the domain in order to make an informed decision of which tool to use, or risk a higher incidence of false positive discoveries (either way, valuable resources are being wasted). In addition, it also negatively impacts the credibility of the field, thus discouraging potential users or researchers from entering.

In this work, we aim to speed up scientific progress in single-cell omics by providing tools both for end-users and developers alike. For developers of computational approaches, we provide tools and guidelines for benchmarking their method on real and synthetic data. For end-users we develop new tools and guidelines for analysing dynamic processes by inferring trajectories and gene regulatory networks. These contributions are discussed in the following chapters:

- We develop benchmarking strategies for assessing the performance of computational tools constrained by low availability of novel types of real single-cell data (Chapter **??**). *In silico* simulations of individual cells are used to help kick-start emerging domains much more safely and allow anticipation of future technological developments by already developing computational tools.

- We apply this strategy to perform a comparison of TI methods (Chapter **??**). Trajec-

tory inference is one of the largest categories of all the novel single-cell omics tools, yet a comprehensive and quantitative study of the advantages and disadvantages of the numerous tools was hitherto lacking. We provide a set of guidelines for end-users wishing to infer trajectories. We also make our pipeline, datasets, metrics, and containerised wrappers of TI methods publicly available for developers to use.

- We developed dyno, a toolkit to easily infer, visualise and interpret single-cell trajectories using more than 50 different TI methods (Chapter **??**). dyno provides downstream analysis such as: visualising a trajectory in a low-dimensional space or a heatmap, detecting genes differentially expressed at different stages of the tragectory, comparing multiple trajectories in a common dimensionality reduction, and manipulating the trajectory (e.g. adding directionality or adding annotation).

- We introduce a novel TI method specialised in inferring linear trajectories (Chapter **??**). Despite linear TI being the most simple but commonly used form of trajectory inference, the benchmark demonstrated that most TI methods are not capable of producing accurate models of linear datasets.

- We invent a new type of NI method capable of inferring the GRN of individual cells (Chapter **??**). We demonstrate this <yadeyade .. fill in when the chapter is actually written.>

- Every NI method has certain topological biases. We provide a tool for analysing the topological properties of large, evolving networks and use this to iteratively optimise GRN predictions (Chapter **??**).

- We discuss reproducibility problems of TI methods due to low rates of quantitative self-assessment (Chapter **??**). We provide solutions for different causal reasons for this phenomenon in order to spur developers to perform more self-assessments.

- Finally, we summarise our experience in benchmarking computational methods in a list of essential guidelines (Chapter **??**).

4 List of contributions

4.1 First-author publications

- **Cannoodt R**, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, Lambrecht B, De Preter K, Saeys Y. SCORPIUS improves trajectory inference and identifies novel modules in dendritic cell development. bioRxiv 079509. 2016 Oct.

- **Cannoodt R** *, Saelens W *, Saeys Y. Computational methods for trajectory inference from single-cell transcriptomics. European journal of immunology. 2016 Nov;46(11):2496-506.

- **Cannoodt R**, Ruyssinck J, Ramon J, De Preter K, Saeys Y. IncGraph: Incremental graphlet counting for topology optimisation. PloS one. 2018 Apr 26;13(4):e0195997.
- Saelens W *, **Cannoodt R** *, Todorov H, Saeys Y. A comparison of single-cell trajectory inference methods. Nature biotechnology. 2019 May;37(5):547.
- **Cannoodt R**, Saelens W, Sichien D, Tavernier S, Janssens S, Guilliams M, Lambrecht B, De Preter K, Saeys Y. SCORPIUS: Fast, accurate, and robust single-cell pseudotime. In preparation.
- **Cannoodt R** *, Saelens W *, Saeys Y. dyngen: Simulating developing single cells. In preparation.
- **Cannoodt R** *, Saelens W *, Saeys Y. dyno: A toolkit for inferring, visualising, and interpreting trajectories. In preparation.
- **Cannoodt R**, Saelens W, Saeys Y, De Preter K. bred: Inferring single cell regulatory networks. In preparation.
- **Cannoodt R**, Saelens W, Saeys Y. Self-assessment in trajectory inference. In preparation.

*: Equal contribution.

4.2 Co-author publications

- Decock A, Ongenaert M, **Cannoodt R**, Verniers K, De Wilde B, Laureys G, Van Roy N, Berbegall AP, Bienertova-Vasku J, Bown N, Clément N. Methyl-CpG-binding domain sequencing reveals a prognostic methylation signature in neuroblastoma. Oncotarget. 2016 Jan 12;7(2):1960.
- Van Cauwenbergh C, Van Schil K, **Cannoodt R**, Bauwens M, Van Laethem T, De Jaegere S, Steyaert W, Sante T, Menten B, Leroy BP, Coppieters F. arrEYE: a customized platform for high-resolution copy number analysis of coding and noncoding regions of known and candidate retinal dystrophy genes and retinal noncoding RNAs. Genetics in Medicine. 2017 Apr;19(4):457.
- Claeys S, Denecker G, **Cannoodt R**, Kumps C, Durinck K, Speleman F, De Preter K. Early and late effects of pharmacological ALK inhibition on the neuroblastoma transcriptome. Oncotarget. 2017 Dec 5;8(63):106820.
- Depuydt P, Boeva V, Hocking TD, **Cannoodt R**, Ambros IM, Ambros PF, Asgharzadeh S, Attiyeh EF, Combaret V, Defferrari R, Fischer M. Genomic amplifications and distal 6q loss: novel markers for poor survival in high-risk neuroblastoma patients. JNCI: Journal of the National Cancer Institute. 2018 Mar 5;110(10):1084-93.

- Scott CL, T'Jonck W, ..., **Cannoodt R**, Saelens W ..., Guilliams M. The transcription factor ZEB2 is required to maintain the tissue-specific identities of macrophages. Immunity. 2018 Aug 21;49(2):312-25.
- Saelens W, **Cannoodt R**, Saeys Y. A comprehensive evaluation of module detection methods for gene expression data. Nature communications. 2018 Mar 15;9(1):1090.
- Todorov H, **Cannoodt R**, Saelens W, Saeys Y. Network Inference from Single-Cell Transcriptomic Data. In Gene Regulatory Networks 2019 (pp. 235-249). Humana Press, New York, NY..
- Van den Berge K, De Bezieux HR, Street K, Saelens W, **Cannoodt R**, Saeys Y, Dudoit S, Clement L. Trajectory-based differential expression analysis for single-cell sequencing data. BioRxiv. 2019 Jan 1:623397.
- Weber LM, Saelens W, **Cannoodt R**, Soneson C, Hapfelmeier A, Gardner PP, Boulesteix AL, Saeys Y, Robinson MD. Essential guidelines for computational method benchmarking. Genome biology. 2019 Dec;20(1):125.
- Lorenzi L, ..., **Cannoodt R**, ..., Mestdagh P. The RNA-Atlas, a single nucleotide resolution map of the human transcriptome. In preparation.
- Van den Berge K, Roux de Bézieux H, Street K, Saelens W, **Cannoodt R**, Saeys Y, Dudoit S. Trajectory-based differential expression analysis. Submitted to Nature Communications.
- Van de Sande Bram, ..., **Cannoodt R**, ..., Saeys Y, Aerts S. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Submitted to Nature Protocols.

4.3 Open-source software

As part of this work, many open-source software packages were created and many others were contributed to (Table 1).

Packages that were created as part of this work are hosted on Github under the username rcannood[2] or the dynverse organisation[3]. As part of our standard development practices, we automate execution of unit tests and write extensive documentation to ensure the code complies with CRAN policy before submission. We aim to submit all other packages to CRAN as well.

We also helped maintain or extend other packages on Github, CRAN or Bioconductor on which our software depends. This includes speeding up parts of the dependency (slingshot), adding new functionality (devtools, ParamHelpers), fixing bugs (proxyC, rlang, monocle, splatter, slingshot), becoming a maintainer of orphaned packages (diffusionMap, princurve, GillespieSSA), and extending the documentation (devtools, mlr, remotes). Several of these package receive millions of downloads per year (devtools, remotes, rlang).

[2] `https://github.com/rcannood?tab=repositories`
[3] `https://github.com/dynverse?tab=repositories`

References

[1] Ingo Brigandt and Alan Love. "Reductionism in Biology". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N Zalta. Spring 201. Metaphysics Research Lab, Stanford University, 2017.

[2] Chung Chau Hon et al. "The Human Cell Atlas: Technical Approaches and Challenges". In: *Briefings in Functional Genomics* 17.4 (July 2018), pp. 283–294. ISSN: 20412657. DOI: 10.1093/bfgp/elx029.

[3] Aviv Regev et al. "The Human Cell Atlas White Paper". In: (Oct. 2018). URL: http://arxiv.org/abs/1810.05192.

[4] Human Cell Atlas consortium. *Human Cell Atlas Data Portal*. 2018. URL: https://data.humancellatlas.org (visited on 08/11/2019).

[5] James D Watson, Francis HC Crick, et al. "Molecular Structure of Nucleic Acids". In: *Nature* 171.4356 (1953), pp. 737–738.

[6] Bruce Alberts et al. "The RNA World and the Origins of Life". In: *Molecular Biology of the Cell. 4th edition* (2002). URL: https://www.ncbi.nlm.nih.gov/books/NBK26876/ (visited on 08/12/2019).

[7] David P. Horning. "RNA World". In: *Encyclopedia of Astrobiology*. Ed. by Muriel Gargaud et al. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1466–1478. ISBN: 978-3-642-11274-4. DOI: 10.1007/978-3-642-11274-4_1740. URL: https://doi.org/10.1007/978-3-642-11274-4_1740.

[8] T. Strachan, A. Read, and T. Strachan. "Human Molecular Genetics. 4th". In: *New York: Garland Science* (2011).

[9] Timothy W. Nilsen and Brenton R. Graveley. "Expansion of the Eukaryotic Proteome by Alternative Splicing". In: *Nature* 463.7280 (Jan. 1, 2010), pp. 457–463. ISSN: 1476-4687. DOI: 10.1038/nature08909.

[10] Bruce Alberts et al. "An Overview of Gene Control". In: *Molecular Biology of the Cell. 4th Edition*. Garland Science, 2002.

[11] Samuel A. Lambert et al. "The Human Transcription Factors". In: *Cell* 172.4 (Feb. 8, 2018), pp. 650–665. ISSN: 0092-8674. DOI: 10.1016/j.cell.2018.01.029.

[12] Albert H Coons, Hugh J Creech, and R Norman Jones. "Immunological Properties of an Antibody Containing a Fluorescent Group." In: *Proceedings of the Society for Experimental Biology and Medicine* 47.2 (1941), pp. 200–202.

[13] Zenonas Theodosiou et al. "Automated Analysis of FISH and Immunohistochemistry Images: A Review". In: *Cytometry Part A* 71A.7 (July 1, 2007), pp. 439–450. ISSN: 1552-4922. DOI: 10.1002/cyto.a.20409.

[14] M. J. Fulwyler. "Electronic Separation of Biological Cells by Volume". In: *Science* 150.3698 (1965), pp. 910–911. ISSN: 0036-8075. DOI: 10.1126/science.150.3698.910.

[15] Nalin Leelatian et al. "Preparing Viable Single Cells from Human Tissue and Tumors for Cytomic Analysis". In: *Current Protocols in Molecular Biology* 118.1 (Apr. 1, 2017), pp. 25C.1.1–25C.1.23. ISSN: 1934-3639. DOI: 10.1002/cpmb.37.

[16] Andrea Cossarizza et al. "Guidelines for the Use of Flow Cytometry and Cell Sorting in Immunological Studies". In: *European Journal of Immunology* 47.10 (Oct. 1, 2017), pp. 1584–1797. ISSN: 0014-2980. DOI: 10.1002/eji.201646632.

[17] Satya P. Yadav. "The Wholeness in Suffix -Omics, -Omes, and the Word Om". In: *Journal of Biomolecular Techniques : JBT* 18.5 (Dec. 2007), p. 277. ISSN: 1524-0215. pmid: 18166670. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2392988/ (visited on 08/15/2019).

[18] Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. "Single-Cell RNA Sequencing Technologies and Bioinformatics Pipelines". In: *Experimental & Molecular Medicine* 50.8 (Aug. 7, 2018), p. 96. ISSN: 2092-6413. DOI: 10.1038/s12276-018-0071-8.

[19] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. "Exponential Scaling of Single-Cell RNA-Seq in the Past Decade". In: *Nature Protocols* 13.4 (Apr. 2018), pp. 599–604. ISSN: 1750-2799. DOI: 10.1038/nprot.2017.149.

[20] Arnav Moudgil. *Multimodal scRNA-Seq*. Feb. 25, 2019. DOI: 10.5281/zenodo.2628012. URL: https://zenodo.org/record/2628012#.XVogbvzRaV4 (visited on 08/19/2019).

[21] Oliver Stegle, Sarah A. Teichmann, and John C. Marioni. "Computational and Analytical Challenges in Single-Cell Transcriptomics". In: *Nature Reviews Genetics* 16.3 (Mar. 2015), pp. 133–145. ISSN: 1471-0064. DOI: 10.1038/nrg3833.

[22] Guo-Cheng Yuan et al. "Challenges and Emerging Directions in Single-Cell Analysis". In: *Genome Biology* 18.1 (May 8, 2017), p. 84. ISSN: 1474-760X. DOI: 10.1186/s13059-017-1218-y.

[23] Geng Chen, Baitang Ning, and Tieliu Shi. "Single-Cell RNA-Seq Technologies and Related Computational Data Analysis". In: *Frontiers in Genetics* 10 (2019). ISSN: 1664-8021. DOI: 10.3389/fgene.2019.00317.

[24] Damien Nicolas, Nick E. Phillips, and Felix Naef. "What Shapes Eukaryotic Transcriptional Bursting?" In: *Molecular BioSystems* 13.7 (2017), pp. 1280–1290. ISSN: 1742-206X. DOI: 10.1039/C7MB00154A.

[25] Luke Zappia, Belinda Phipson, and Alicia Oshlack. "Exploring the Single-Cell RNA-Seq Analysis Landscape with the scRNA-Tools Database". In: *PLOS Computational Biology* 14.6 (June 2018), e1006245. ISSN: 1553-7358. DOI: 10.1371/journal.pcbi.1006245.

[26] Conrad Hal Waddington et al. "The Strategy of the Genes. A Discussion of Some Aspects of Theoretical Biology. With an Appendix by H. Kacser". In: *The strategy of the genes. A discussion of some aspects of theoretical biology. With an appendix by H. Kacser.* (1957), pp. ix+–262.

[27] James E Ferrell. "Bistability, Bifurcations, and Waddington's Epigenetic Landscape". In: *Current Biology* 22.11 (June 2012), R458–R466. ISSN: 0960-9822. DOI: 10.1016/j.cub.2012.03.045.

[28] Jonathan A. Rebhahn et al. "An Animated Landscape Representation of CD4+ T-Cell Differentiation, Variability, and Plasticity: Insights into the Behavior of Populations versus Cells". In: *European Journal of Immunology* 44.8 (Aug. 1, 2014), pp. 2216–2229. ISSN: 0014-2980. DOI: 10.1002/eji.201444645.

[29] Barbara Treutlein et al. "Dissecting Direct Reprogramming from Fibroblast to Neuron Using Single-Cell RNA-Seq". In: *Nature* 534.7607 (2016), pp. 391–395.

[30] Daniel Engel, Lars Hüttenberger, and Bernd Hamann. "A Survey of Dimension Reduction Methods for High-Dimensional Data Analysis and Visualization". In: *Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering - Proceedings of IRTG 1131 Workshop 2011*. Ed. by Christoph Garth, Ariane Middel, and Hans Hagen. Vol. 27. OpenAccess Series in Informatics (OASIcs). Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, pp. 135–149. ISBN: 978-3-939897-46-0. DOI: 10.4230/OASIcs.VLUDS.2011.135.

[31] Karl Pearson. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (Nov. 1, 1901), pp. 559–572. ISSN: 1941-5982. DOI: 10.1080/14786440109462720.

[32] J. B. Kruskal. "Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis". In: *Psychometrika* 29.1 (Mar. 1964), pp. 1–27. ISSN: 0033-3123, 1860-0980. DOI: 10.1007/BF02289565.

[33] Laurens van der Maaten and Geoffrey Hinton. "Visualizing Data Using T-SNE". In: *The Journal of Machine Learning Research* 9.2579-2605 (2008), p. 85.

[34] Boaz Nadler et al. "Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker Planck Operations". In: *Proceedings of the 18th International Conference on Information Processing Systems*. NIPS'05. Cambridge, MA, USA: MIT Press, 2005, pp. 955–962.

[35] Ronald R. Coifman and Stéphane Lafon. "Diffusion Maps". In: *Applied and Computational Harmonic Analysis* 21.1 (July 2006), pp. 5–30. ISSN: 10635203. DOI: 10.1016/j.acha.2006.04.006.

[36] Leland McInnes, John Healy, and James Melville. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction". In: (2018).

[37] Vin D. Silva and Joshua B. Tenenbaum. "Global Versus Local Methods in Nonlinear Dimensionality Reduction". In: *Advances in Neural Information Processing Systems 15*. Ed. by S. Thrun and K. Obermayer. Cambridge, MA: MIT Press, 2002, pp. 705–712.

[38] Seunghak Lee and Seungjin Choi. "Landmark MDS Ensemble". In: *Pattern Recognition* 42.9 (Sept. 2009), pp. 2045–2053. ISSN: 00313203. DOI: 10.1016/j.patcog.2008.11.039.

[39] Stuart Lloyd. "Least Squares Quantization in PCM". In: *IEEE Transactions on Information Theory* 28.2 (Mar. 1982), pp. 129–137. DOI: 10.1109/TIT.1982.1056489.

[40] Vincent D Blondel et al. "Fast Unfolding of Communities in Large Networks". In: *Journal of Statistical Mechanics: Theory and Experiment* 2008.10 (Oct. 9, 2008), P10008. ISSN: 1742-5468. DOI: 10.1088/1742-5468/2008/10/p10008.

[41] Tariq Enver et al. "Stem Cell States, Fates, and the Rules of Attraction". In: *Cell Stem Cell* 4.5 (May 8, 2009), pp. 387–397. ISSN: 1875-9777. DOI: 10.1016/j.stem.2009.04.011. pmid: 19427289.

[42] Robrecht Cannoodt, Wouter Saelens, and Yvan Saeys. "Computational Methods for Trajectory Inference from Single-Cell Transcriptomics". In: *European Journal of Immunology* 46.11 (Nov. 1, 2016), pp. 2496–2506. ISSN: 1521-4141. DOI: 10.1002/eji.201646347.

[43] Wouter Saelens et al. "A Comparison of Single-Cell Trajectory Inference Methods". In: *Nature Biotechnology* 37 (May 2019). ISSN: 15461696. DOI: 10.1038/s41587-019-0071-9.

[44] Olivia Padovan-Merhar and Arjun Raj. "Using Variability in Gene Expression as a Tool for Studying Gene Regulation". In: *Wiley Interdisciplinary Reviews. Systems Biology and Medicine* 5.6 (Nov. 2013), pp. 751–759. ISSN: 1939-005X. DOI: 10.1002/wsbm.1243. pmid: 23996796.

Table 1: Contibutions to open-source software. Following abbreviations denote the relation with respect to the package: *aut* Author, *ctb* Contributor. Yearly download statistics are based on the number of downloads between 2019-08-01 and 2019-09-10. CRAN download statistics are retrieved from the Rstudio CRAN mirror only; other CRAN mirrors do not track download statistics. In addition, many of the dynverse packages have only recently been published on CRAN. For Github repositories, no download statistics could be retrieved.

Name	Role	Host	Downloads per year	Description
babelwhale	aut	CRAN	5810	Interacting with Docker and Singularity containers
diffusionMap	aut	CRAN	30'450	Implements diffusion map method of data parameterization, including creation and visualization of diffusion map
dynbenchmark	aut	Github		Pipeline for benchmarking trajectory inference methods
dyndimred	aut	CRAN	5535	Applying dimensionality reduction methods
dyneval	aut	Github		Evaluating trajectory inference methods
dynfeature	aut	Github		Calculating feature importance scores from trajectories
dyngen	aut	Github		Simulating single-cell data using gene regulatory networks
dynguidelines	aut	Github		User guidelines for trajectory inference
dynmethods	aut	Github		A collection of wrappers for trajectory inference methods
dyno	aut	Github		A pipeline for inferring, visualising and interpreting trajectories
dynparam	aut	CRAN	3265	Creating meta-information for parameters
dynplot	aut	Github		A simple visualisation library for trajectories
dynplot2	aut	Github		A fully customisable visualisation library for trajectories
dyntoy	aut	Github		Generating simple toy data of cellular differentiation
dynutils	aut	CRAN	13'13'	Common functionality for the dynverse packages
dynwrap	aut	CRAN	990	A common format for trajectories
GillespieSSA	aut	CRAN	7880	Gillespie's Stochastic Simulation Algorithm (SSA)
GillespieSSA2	aut	CRAN	4950	Gillespie's Stochastic Simulation Algorithm for Impatient People
gng	aut	Github		An Rcpp implementation of the Growing Neural Gas algorithm
incgraph	aut	CRAN	3565	Incremental graphlet counting for network optimisation
lmds	aut	CRAN	815	Landmark Multi-Dimensional Scaling
princurve	aut	CRAN	29'100	Fits a principal curve in arbitrary dimension
proxyC	aut	CRAN	117'480	Computes proximity in large sparse matrices
qsub	aut	CRAN	3585	Running commands remotely on gridengine clusters
SCORPIUS	aut	CRAN	4520	Inferring developmental chronologies from single-cell RNA sequencing data
badger	ctb	CRAN	6240	Query information and generate badge for using in README and GitHub Pages
ClusterSignificance		Bioc	935	Assess if class clusters in dimensionality reduced data representations have a separation different from permuted data
devtools	ctb	CRAN	5'918'700	Tools to make developing R packages easier
merlot	ctb	Github		A method for reconstructing lineage-tree topologies from scRNA-seq data
mlr	ctb	CRAN	176'330	Machine Learning in R
monocle	ctb	Bioc	34'360	Clustering, differential expression, and trajectory analysis for single-cell RNA-Seq
ParamHelpers	ctb	CRAN	150'775	Helpers for Parameters in Black-Box Optimization, Tuning and Machine Learning
pseudogp	ctb	Github		Probabilistic pseudotime for single-cell RNA-seq
Rdimtools	ctb	CRAN	7367	Dimension Reduction and Estimation Methods
remotes	ctb	CRAN	3'944'090	R package installation from remote repositories, including GitHub
rlang	ctb	CRAN	13'269'115	Functions for base types and core R and tidyverse features
SCope	ctb	Github		Visualization of large-scale and high dimensional single cell data
slingshot	ctb	Bioc	12'085	Tools for ordering single-cell sequencing
splatter	ctb	Bioc	5015	Simple simulation of single-cell RNA sequencing data
URD	ctb	Github		URD reconstructs transcriptional trajectories underlying specification or differentiation processes in the form of a branching tree from single-cell RNAseq data
wishbone	ctb	Github		Identify bifurcating developmental trajectories from single-cell data