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.

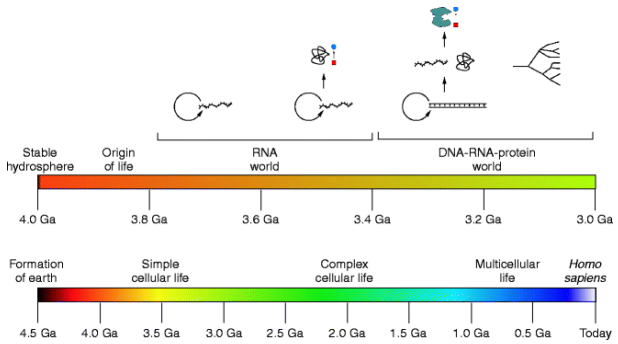

Figure 1: RNA world. The postulated rise and fall of the RNA world during the evolution of life, from early self-replicating RNA to complex, RNA-controlled metabolism, to the invention of translation, followed by diversification of all modern branches of life. Image from Horning (2011)[7]. TODO: combine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

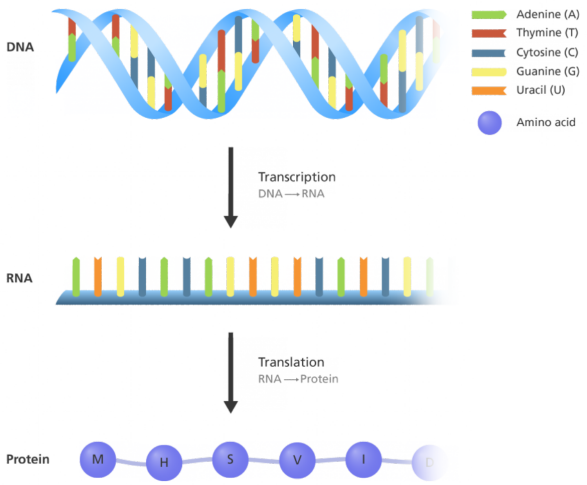

Figure 2: Central Dogma. TODO: combine images, rewrite the description.

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

Homo sapiens like to categorise everything they encounter, and so too have they conceptualised groups of cells called "cell types". The human body contains more than 200 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

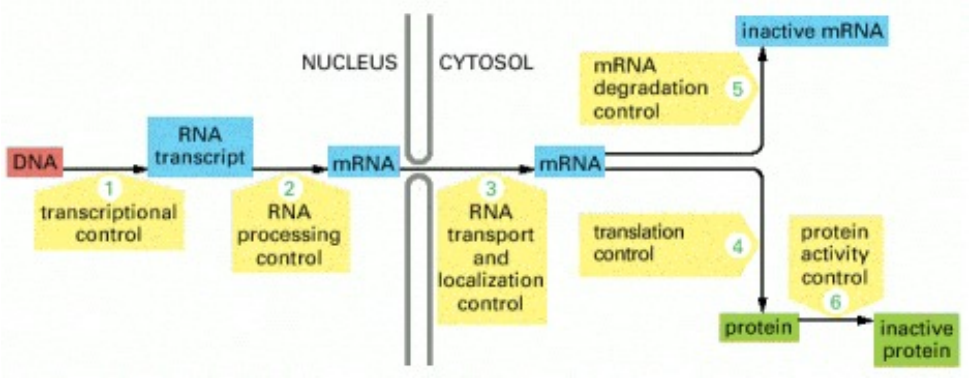

Figure 3: Levels of controlling gene expression can happen at the level of transcription, RNA processing (i.e. splicing), RNA transport and localization, mRNA translation, mRNA degradation and protein activity [10].

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 antigen-antibody proteins[12]. In many multicellular organisms, antibodies and antigens serve as crucial communication tools as part of the organism's immune system. A cell can present a particular type of antigen on its cell surface, which allows a particular type of antibody to bind to it.

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 light excitation. The use of different 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). While modern implementations of IHC improve the throughput drastically by using robots to automate the image acquisition and computer software to automate cell segmentation, the procedure is still labour intensive as the robots and computer software still needs to be kept in check.

Flow cytometry[13] is a technique which circumvents imaging and segmentation issues by having a steady stream of cells run through a laser and measuring the amount of light scattered from those cells. Flow cytometry technology enables to measure protein expression levels for millions of cells and tens of different antibodies.

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 qPCR, FISH). All of these single-cell (non-omics) technologies are limited by the number of different molecules they could 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 many types of "omics", but the most commonly used are the following. In genomics, all of an organism's genes are studied – its whole genome. Transcriptomics and proteomics study the organisms RNA transcripts and proteins, respectively.

Specific examples of omics technologies are next-generation sequencing, which can be used to determine the DNA sequence of an organism, and RNA sequencing, which profiles the sequences of RNA transcripts. By mapping the sequences of RNA transcripts to genes in the organisms DNA, a gene expression profile can be obtained.

Traditionally, profiling nano-litre volumes was infeasible. To capture enough material to generate a profile, an ensemble of cells needs to be pooled and lysed together, thereby granting the technology the name "bulk" omics. Bulk omics is a major workhorse in <...> and has applications in <...>. To capture enough material, an ensemble of cells is pooled and profiled all together.

TODO: fill in

A notable downside of bulk omics arises when the cells are biomolecularly heterogeneous (Figure 4A). Since the profile is a population average (or rather, a summation), information pertaining the individual profiles of cell subpopulations cannot be discerned (Figure 4B).

[1] The etymology of "omics" is quite interesting[14].

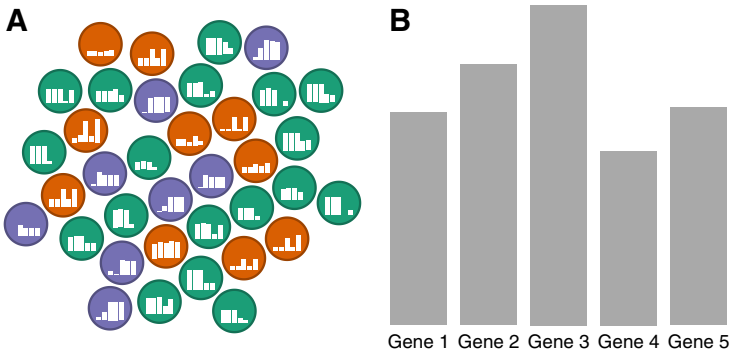

Figure 4: The 'masking' effect of bulk omics. A: Cells from several subpopulations are incorrectly assumed to be homogeneous and are profiled with a single bulk omics experiment. **B:** The signals from the different subpopulations are masked. The resulting profile is dissimilar from the majority of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

1.5.3 Single-cell omics

Transformative technological advances in microvolume sequencing allowed Tang et al. to analyse the transcriptome at single-cell resolution[15], thereby bringing single-cell biology and omics together to create single-cell omics (Figure 5A). During the decade that followed, the number of single-cell omics technologies has skyrocketed, allowing to profile tens of thousands of cells and measuring other levels of information (e.g. proteomic expression levels and spatial location).

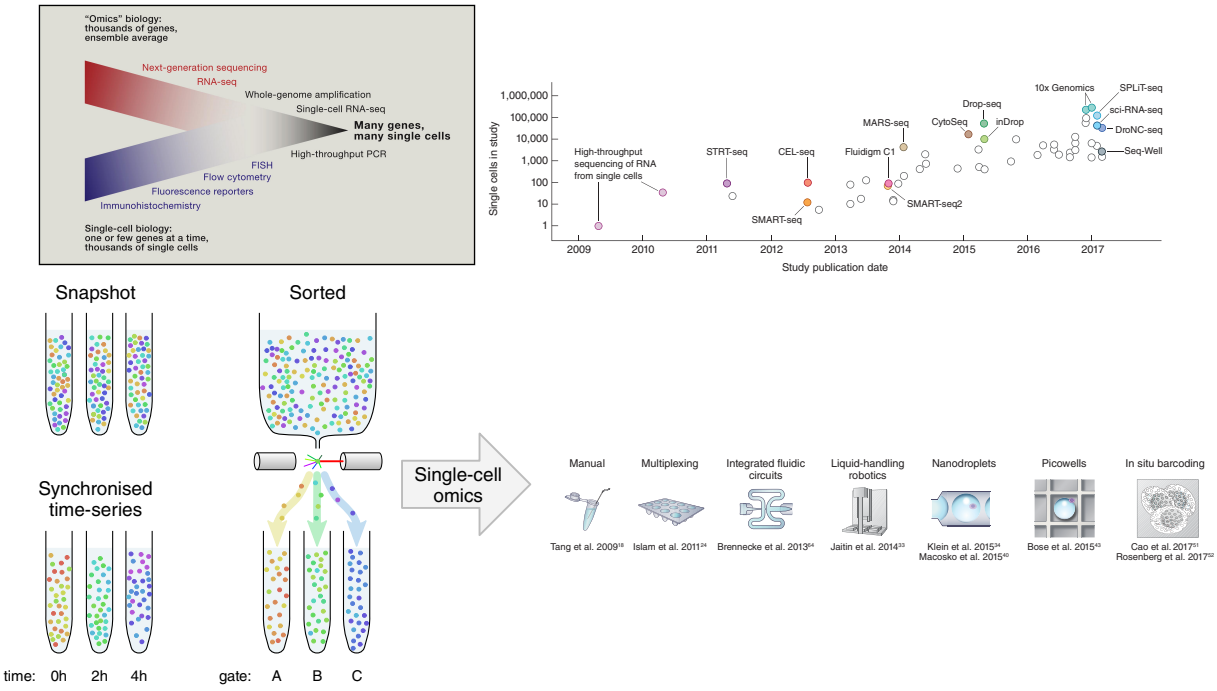

Figure 5:

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 Human Cell Atlas. The HCA consortium has set out to redefine all human cell types in terms of their gene expression and location, and the developmental

trajectories connecting the different cell types. As part of this endeavour, the consortium will likely profile the whole transcriptomes tens or even hundreds of millions of cells.

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[16, 17, 18]. 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 (e.g. dimensionality and noise levels). Over the past five years, already 450 new tools for analysing single-cell omics data have been developed[31]. These tools typically fall within one of nine categories (Figure 6). We discuss each category in the following subsections.

Todo: describe the characteristics of single cell omics data in the subsection on 'single cell omics'. -> counts data. additional information from the cell can be present, such as the time of sampling in a time-series experiment or a cell type if the cells have been sorted using a FACS experiment. information on the features are typically restricted to the gene symbol, or an identifier of a xxx, such as ensembl or entrez. in comparison to bulk data, single-cell omics data is typically very sparse, due to technical and biological stochasticity. if several transcripts of a low abundance gene are not captured, this will be seen as a dropout. dropouts can also be biological, due to transcriptional bursting.

2.1 Normalisation

2.2 Data imputation

2.3 Data integration

2.4 Dimensionality reduction

Single-cell omics datasets typically have too many dimensions (features) in order to be easily interpretable by humans and even by most computational tools. Dimensionality reduction (DR) methods transform high-dimensional data into a meaningful representation with fewer dimensions. It is important to note that its usage depends on the target audience: for humans – to visualise data in a 2-D plane to aid with interpretation by humans, or for computers – to construct a denser representation of the data such that it mostly contains the same information but with fewer dimensions.

There are many ways of classifying DR methods[19], but this work will use the following main categories: feature projection-based and manifold learning. Projection-based DR methods aim to perform a linear 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 PCA and MDS. Manifold learning methods are

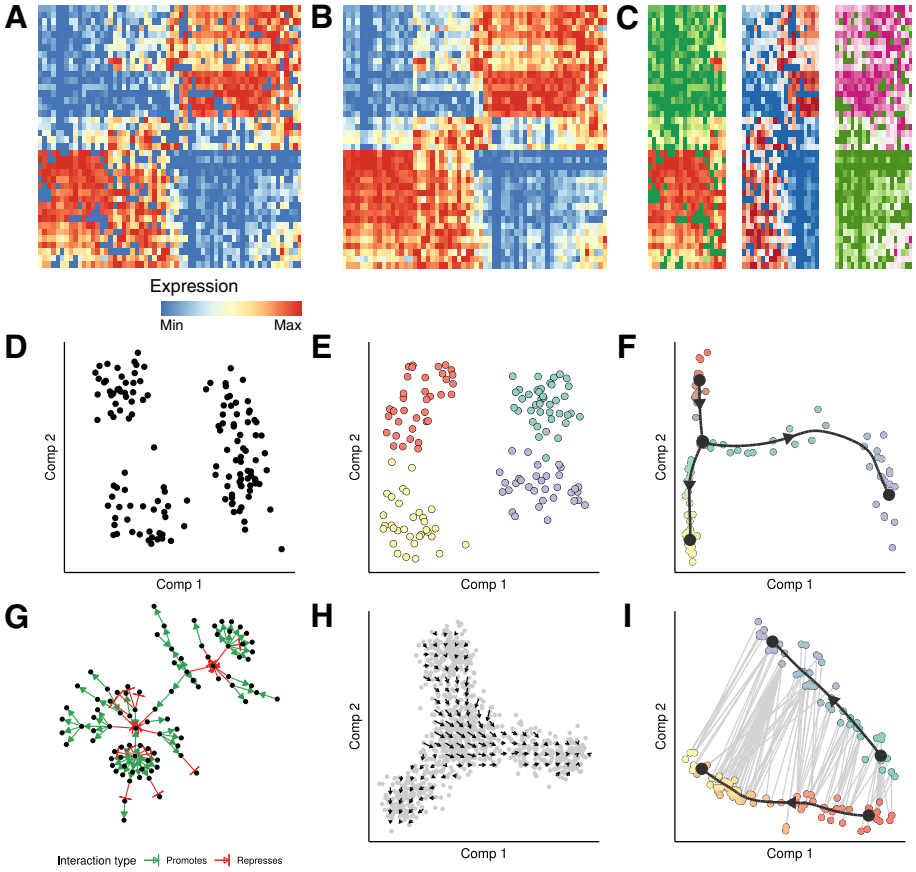

Figure 6: Computational tools developed for processing single-cell omics data usually lie within one of nine categories. **A:** Normalisation, separating meaningful biological signal from technical noise or irrelevant biological information. **B:** Data imputation, masking dropouts. **C:** Data integration, combining multiple single-cell modalities. **D:** Dimensionality reduction, providing a visual overview of the main cellular heterogeneity in the dataset. **E:** Clustering, grouping together similar cells. **F:** Trajectory inference, identifying and characterising transitions between different cellular states. **G:** Network inference, predicting gene regulatory interactions from cell-to-cell heterogeneity. **H:** RNA velocity, inferring the directionality of trajectories by predicting the future state of each cell. **I:** Trajectory alignment, aligning multiple trajectories.

methods which 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-SNE, Diffusion Maps and UMAP.

This section is not very interestingly written

2.5 Clustering

2.6 Trajectory inference

Single-cell omics data provide new opportunities for studying cellular dynamic processes, such as the cell cycle, cell differentiation and cell activation[20, 21]. Trajectory inference (TI) is a new category of computational tools used to offer an unbiased and transcriptome-wide understanding of a dynamic process[20, 22].

Technological advancements in single-cell omics allow studying a dynamic process in

a high-throughput manner. This raises concerns regarding biological fundamentals, such as how to define cell types or transitions between them[21, 20]. Trajectory inference (TI) methods aim to give insight into a dynamic process by inferring a trajectory from omics profiles of cells in which the dynamic process takes place[22]. TI has two objectives: to reconstruct the topology of the dynamic process (e.g. is it linear, cyclical, bifurcating), and to determine the position of each cell along the topology. Some TI methods assumes that the user knows the topology beforehand and only focuses on ordering the cells along a predefined topology.

The dataset can be a single snapshot of a mixture of cells in different stages, or a set of samples collected at different time points (Figure 7A). Typically, TI methods first analyse similarities between cells, optionally infer the topology of the underlying process, and finally order cells along that trajectory (Figure 7B). The second step can be optional, as some methods assume a specific topology beforehand. TI methods allow the identification of new subsets of cells, delineation of a differentiation tree, and characterisation of the main driver genes along a state transition (Figure 7C). Current applications of TI focus on specific subsets of cells, but ongoing efforts to construct transcriptomic catalogs of whole organisms[23, 24, 25] underline the urgency for accurate, scalable[26, 27] and user-friendly TI methods.

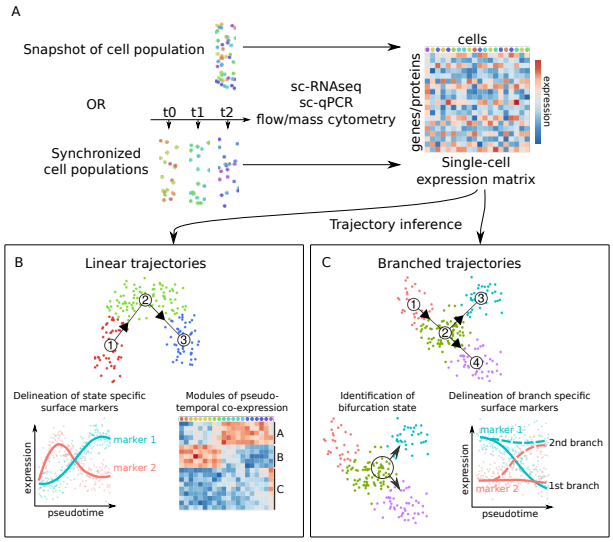

Figure 7: Applications of single-cell trajectory inference methods. (A) Single-cell omics data appropriate for TI can be both obtained from an unsynchronised population of single cells (snapshot data) but also from synchronised cell populations. (B) UPDATE! (C) UPDATE!

Could still expand this section with pieces from the EJI paper, though it needs to be adapted strongly.

2.7 Network inference

Gene regulatory network inference, or network inference (NI) for short, is a type of computational analysis where thousands of transcriptomic profiles are analysed together in order to infer the regulatory interactions between transcription factors and genes. This topic already received much attention with the advent of bulk omics (before single-cell omics). These ef-

forts culminated in several DREAM competitions assessing the performance of 29 different NI methods[28, 29].

After the last DREAM competition, it seemed that interest in NI methodology had declined. After all, NI on bulk omics profiles suffered from several crucial issues. As mentioned previously, bulk profiles are generated by pooling together the RNA transcripts of a supposedly homogeneous population of thousands of cells. Since the expression values are averaged over the whole population, incorrect assumptions on the homogeneity of the pooled cells may lead to the masking of relevant expression patterns in rare cell populations (Figure 8). Besides, NI methods rely on a diverse set of time-series and perturbation experiments in order to reliably identify causal regulatory interactions. Such experiments are expensive and time-consuming, and an inaccurate selection of time points might result in crucial intermediate stages being missed.

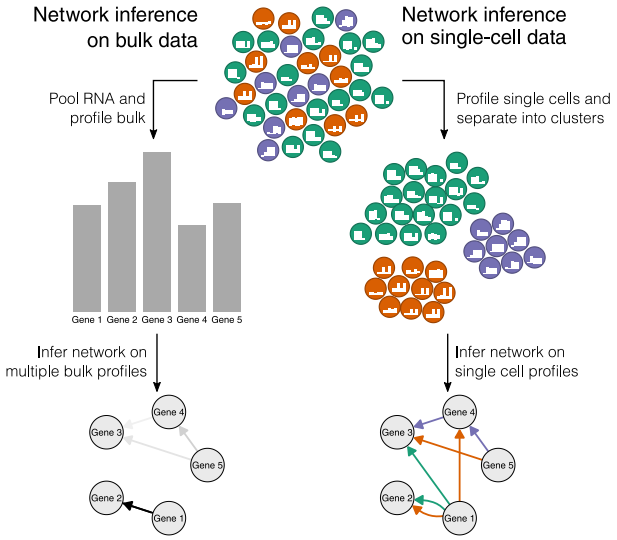

Figure 8: Bulk expression data return the average expressions of genes among large numbers of cells. In order to infer regulatory networks from this type of data, multiple bulk profiles (resulting from time series or perturbation experiments) are required. On the other hand, sequencing the transcriptome at the single-cell level uncovers the high variability among cells, providing the necessary information to infer gene regulatory networks directly.

The advent of single-cell omics has made scientists wonder whether now is the time to revisit network inference[16]. One of the main advantages of single-cell omics is the ability to quantify the exact cellular state of thousands of cells per experiment. The heterogeneity between cells caused by naturally occurring biological randomness[30] can be exploited to infer regulatory interactions between TFs and their target genes at much lower costs (see Figure 8). In this setting, heterogeneity in the cell population eases network inference, rather than mask condition-specific expression patterns in regulatory interactions.

2.8 RNA velocity

2.9 Trajectory alignment

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[31] 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 majority of the 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 **??**). Trajectory 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 *, 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. order authors?
- **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 writing 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).

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.

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

[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] 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.

[14] 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).

[15] Fuchou Tang et al. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell". In: *Nature Methods* 6.5 (May 2009), pp. 377–382. ISSN: 1548-7105. DOI: 10.1038/nmeth.1315.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] Amos Tanay and Aviv Regev. "Scaling Single-Cell Genomics from Phenomenology to Mechanism". In: *Nature* 541.7637 (Jan. 2017), nature21350. ISSN: 1476-4687. DOI: 10.1038/nature21350.

[21] Martin Etzrodt, Max Endele, and Timm Schroeder. "Quantitative Single-Cell Approaches to Stem Cell Research". In: *Cell Stem Cell* 15.5 (2014), pp. 546–558.

[22] 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.

[23] Aviv Regev et al. "The Human Cell Atlas". In: *eLife* 6 (Dec. 2017). ISSN: 2050084X. DOI: 10.7554/eLife.27041.

[24] Xiaoping Han et al. "Mapping the Mouse Cell Atlas by Microwell-Seq". In: *Cell* 172.5 (Feb. 2018), 1091–1107.e17. ISSN: 1097-4172. DOI: 10.1016/j.cell.2018.02.001. pmid: 29474909.

[25] Nicholas Schaum et al. "Single-Cell Transcriptomics of 20 Mouse Organs Creates a Tabula Muris". In: *Nature* 562.7727 (Oct. 2018), pp. 367–372. ISSN: 1476-4687. DOI: 10.1038/s41586-018-0590-4.

[26] Sara Aibar et al. "SCENIC: Single-Cell Regulatory Network Inference and Clustering". In: *Nature Methods* (Oct. 2017). ISSN: 1548-7091. DOI: 10.1038/nmeth.4463.

[27] Philipp Angerer et al. "Single Cells Make Big Data: New Challenges and Opportunities in Transcriptomics". In: *Current Opinion in Systems Biology*. Big Data Acquisition and Analysis $\backslash$textbullet{} Pharmacology and Drug Discovery 4 (Aug. 2017), pp. 85–91. ISSN: 2452-3100. DOI: 10.1016/j.coisb.2017.07.004.

[28] Daniel Marbach et al. "Revealing Strengths and Weaknesses of Methods for Gene Network Inference". In: *Proceedings of the {N}ational {A}cademy of {S}ciences* 107.14 (Apr. 2010), pp. 6286–6291. ISSN: 1091-6490. DOI: 10.1073/pnas.0913357107. pmid: 20308593.

[29] Daniel Marbach et al. "Wisdom of Crowds for Robust Gene Network Inference". In: *Nature methods* 9.8 (July 2012), pp. 796–804. ISSN: 1548-7091. DOI: 10.1038/nmeth.2016. pmid: 22796662.

[30] 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.

[31] 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.

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. For Github repositories, no download statistics could be retrieved.

Name	Role	Host	Downloads per year	Description
babelwhale	aut	CRAN	3996	Interacting with Docker and Singularity containers
diffusionMap	aut	CRAN	21'361	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	5511	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	3084	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	5657	Common functionality for the dynverse packages
dynwrap	aut	Github		A common format for trajectories
GillespieSSA	aut	CRAN	7546	Gillespie's Stochastic Simulation Algorithm (SSA)
GillespieSSA2	aut	CRAN	6506	Gillespie's Stochastic Simulation Algorithm for Impatient People
gng	aut	Github		An Rcpp implementation of the Growing Neural Gas algorithm
incgraph	aut	CRAN	3175	Incremental graphlet counting for network optimisation
lmds	aut	CRAN		Landmark Multi-Dimensional Scaling
princurve	aut	CRAN	26'991	Fits a principal curve in arbitrary dimension
proxyC	aut	CRAN	117'484	Computes proximity in large sparse matrices
qsub	aut	CRAN	3193	Running commands remotely on gridengine clusters
SCORPIUS	aut	CRAN	4772	Inferring developmental chronologies from single-cell RNA sequencing data
badger	ctb	CRAN	?	Query information and generate badge for using in README and GitHub Pages
ClusterSignificance		Bioc	803	Assess if class clusters in dimensionality reduced data representations have a separation different from permuted data
devtools	ctb	CRAN	3'775'350	Tools to make developing R packages easier
merlot	ctb	Github		A method for reconstructing lineage-tree topologies from scRNA-seq data
mlr	ctb	CRAN	142'605	Machine Learning in R
monocle	ctb	Bioc	35'240	Clustering, differential expression, and trajectory analysis for single-cell RNA-Seq
ParamHelpers	ctb	CRAN	109'408	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	3'704'594	R package installation from remote repositories, including GitHub
rlang	ctb	CRAN	11'470'763	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	11'643	Tools for ordering single-cell sequencing
splatter	ctb	Bioc	3741	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