Statistical Modelling for Cell Reprogramming

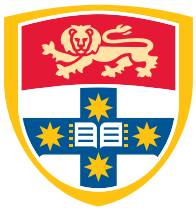
Andy Tran

A thesis submitted in partial fulfillment of

the requirements for the degree of

Master of Philosophy

School of Mathematics and Statistics



THE UNIVERSITY OF SYDNEY

September 2021

Statement of originality

This is to certify that to the best of my knowledge, the content of this thesis is my own work. This thesis has not been submitted for any degree or other purposes.

I certify that the intellectual content of this thesis is the product of my own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Andy Tran

Abstract

Cells generally begin their lives as a pluripotent stem cell that gradually differentiates into specialised cell fates over time. However, recent advances in cell reprogramming have successfully converted differentiated cells to other cell types, by overexpressing a combination of transcription factors, fundamentally altering our view of cell identity. This could have large implications for the field of regenerative medicine as cell reprogramming offers the potential to regrow, repair or replace tissues and organs which have been damaged from age or disease. Despite much attention, combinations of transcription factors to drive reprogramming were mostly determined through trial and error, taking up considerable time and resources. To this end, computational methods have been developed to simulate the reprogramming process in silico with the goal of guiding the hypotheses to be experimentally validated. These models have provided a variety of perspectives to the process of cell reprogramming, leading to the discovery of novel cell conversions. However, they all suffer from limitations in terms of generalisability and scalability.

Here we categorise the existing computational approaches for cell reprogramming and critically evaluate their applicability in a broader context. We then propose a novel method which leverages emerging multimodal single cell data. By integrating these different modes of data, we are able to create a more holistic model of a cell's regulatory systems which provides a more accurate and scalable model for reprogramming. We demonstrate the applicability our method by recapitulating known properties of cell differentiation and reprogramming in both simulated and experimental data.

We hope that this thesis will contribute to our understanding of the role of gene regulation in cell reprogramming by synthesising the existing computational models. Furthermore, our novel method may be a starting point for future computational models to integrate data from multiple modalities to create more comprehensive models for cell regulation.

Acknowledgements

The last couple of years has been an incredible journey of not only intellectual growth, but also mental and emotional growth. I want to sincerely thank everyone who has accompanied me and supported me through these challenging years.

Firstly, I want to thank everyone in the Sydney Precision Bioinformatics Alliance for warmly welcoming me into the group. In particular, Kevin, Yingxin, Hani, Tom, Dario, Yue, Yunwei, Taiyun and Xiangnan who helped get me on my feet after I made the daunting transition into bioinformatics. I would also like to thank the other friends I made along the way, Peng, Di, Amarinder, Hao, Lijia and Carissa, you've really made me enjoy being a part of this group.

I extend this thanks to the data science honours cohort of 2020, who shared my suffering of the year 2020 and the coursework component of my MPhil. Especially to Michael, Anne and Thomas for carrying me through the assessments, and also to Jesenia, Chris, Victor and Priscilla for the emotional support and banter.

I would like to thank my three supervisors. Pengyi, your expertise has been invaluable to my learning and my project. I will always appreciate your swift replies to my messages, even at ungodly hours. Jean, I have learnt and grown so much from your guidance and mentorship. I am so grateful for you providing me with so many collaborative opportunities and supporting me through these. And John, I cannot thank you enough for encouraging me to pursue my MPhil and supporting me through all the challenges. I clearly remember the day where I was a lost graduate and sought advice from you about further study. You offered to take me in for an MPhil and I can confidently say that your decision has pushed me onto a trajectory that completely aligns with my passions and values.

Finally, I would like to thank my friends Victor, Vaish and Daz, and my family: my mother, father, brother, grandparents, and cousins. The last couple years has been full of adversity and I could not have overcome it without your support behind the scenes.

Contents

Chapter 1

Introduction

Throughout all of history, breakthroughs in our understanding of biology was instigated by developments in technology. For example, the invention of the microscope led to the revelation of the cell as the fundamental unit of life. And the development of genetic sequencing techniques led to the discovery of DNA as the blueprint for all of life. Now in today's information age, we are constantly developing new technologies, leading to scientific breakthroughs at an unprecedented scale, helping us battle countless diseases, and improve our quality of life.

In a world where data is constantly becoming cheaper and more accessible, a major roadblock to our progress is the ability to develop mathematics and statistics to model these new types of data. In this thesis, we will critically assess the successes and limitations of the existing mathematical models for cell reprogramming, and develop a novel model to address some of these limitations.

This thesis is structured as:

- **Chapter 1:** a brief overview of the biological and technological background behind cell reprogramming. We focus on the significant value of potential clinical applications and also the roadblocks that currently limit its feasibility.

- **Chapter 2:** a synthesis of the variety of statistical models that have currently been developed for cell reprogramming. We analyse the perspectives that they offer to understanding the regulatory dynamics governing cell reprogramming, and also assess their applicability and generalisability.

- **Chapter 3:** a presentation of the innovative components of new computational models, and the corresponding strategy of how they could be measured in practice. We define a regulation potential which measures the ability for a transcription factor to

regulate a gene, and we propose and evaluate a number of possible models that use this regulation potential to simulate transcription factor perturbations.

- **Chapter 4:** a critical evaluation of the final model derived in Chapter 3 using both simulated and real experimental data. We demonstrate that our model is able to recapitulate known properties of cell differentiation, and cell reprogramming via a series of simulation studies. Furthermore, using real experimental data on mouse hair follicle development, we find that our model successfully predicts the result of overexpressing two key transcription factors.

- **Chapter 5:** a conclusion to the thesis discussing the features and limitations of our work as well as future directions that could be taken.

1.1 Cell reprogramming

Multicellular organisms generally start their lives as an individual pluripotent cell that must differentiate into more specialised cell types (Waddington, 1966). This process of cellular differentiation allows the creation of a spectrum of cell types that perform all the necessary functions for an organism's survival. These specialised cells often have biological mechanisms that help to maintain the cell's identity, ensuring it can reliably perform its required task. However, it has recently been shown that transitioning between these specialised cell types (Takahashi and Yamanaka, 2006), called cell reprogramming, is indeed possible, and has been demonstrated in a variety of cell types and species. This can be performed by many techniques where the current standard is to use lentiviruses to integrate a desired gene into the host cell's genome. This can be used to ectopically overexpress some transcription factors (TFs) in cells (Aydin and Mazzoni, 2019) with the hope of changing the cell's identity.

Recent developments in cell reprogramming has been of great interest to the field of regenerative medicine, as it opens up the possibility to regenerate cells that our body has lost, and cannot normally reproduce. For example, type 1 diabetes occurs when the insulin-producing beta cells in the pancreas are attacked by the immune system. However, recent experiments have shown that overexpressing *Pdx1* and *MafA* in pancreatic alpha cells have

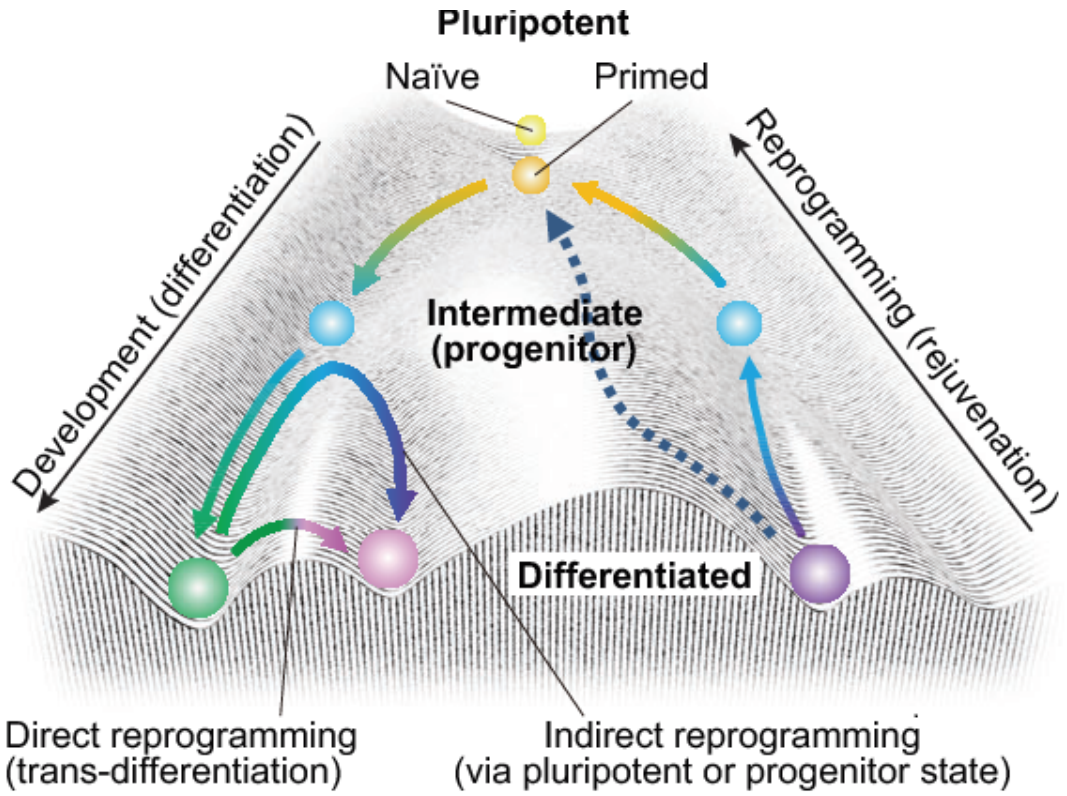


Figure 1.1: Illustration of Waddington's Epigenetic Landscape Model.

Source: Kazutoshi Takahashi, Shinya Yamanaka; A developmental framework for induced pluripotency. Development 1 October 2015; 142 (19): 3274–3285. doi: https://doi.org/10.1242/dev.114249

been able to convert them into insulin-producing beta cells, reversing type 1 diabetes in mice (Xiao et al., 2018; Furuyama et al., 2019). Additionally, Parkinson's disease is attributed to the degeneration of dopamine neurons in the brain, which have been successfully replenished using reprogrammed cells in mice (Kim et al., 2002). This has led to human trials to treat Parkinson's disease using reprogrammed cells which are expected to start soon (Barker et al., 2017; Parmar et al., 2020). Cell reprogramming has been considered as a potential cure for a variety of other diseases including heart disease (Aguirre et al., 2013), spinal cord injury (Khazaei et al., 2017), macular degeneration (Chichagova et al., 2018), hearing loss (Bermingham-McDonogh and Reh, 2011), and aplastic anemia (Melguizo-Sanchis et al., 2018), among others. Furthermore, reprogramming cells from the same patient offers an additional advantage of minimising the risk of an acute rejection, a common problem with other transplant-based therapies (Shaik et al., 2015).

Despite the vast potential for therapeutic applications of cell reprogramming, a major

roadblock is the challenge of determining the combination and quantity of transcription factors to overexpress for a desired cell conversion. Historically, this was done experimentally by trial and error which is time consuming and expensive (Aydin and Mazzoni, 2019; Grath and Dai, 2019). For reference, there are an estimated 1600 human transcription factors (Lambert et al., 2018), which means there are more than 600 million combinations of three transcription factors to overexpress. Furthermore, current reprogramming experiments are very inefficient, often with $< 1\%$ yield of the target cell (Takahashi and Yamanaka, 2016), making it difficult to use for therapeutic purposes. Finding more efficient combinations of transcription factors and better understanding the process of cell reprogramming could significantly increase this yield, and its feasibility for clinical use.

The field of cell reprogramming would greatly benefit from a computational method to predict the effect of perturbing a transcription factor. This could allow many TF combinations to be quickly tested, guiding the combinations and quantities to be experimentally validated, reducing the time and cost of discovering new cell conversions. However, developing such a computational method has been a challenge because of the sheer complexity and number of components involved in gene regulation (Spitz and Furlong, 2012).

1.2 Gene regulation

According to the Central Dogma of Molecular Biology, a cell performs its functions by using its DNA to create RNA which in turn makes useful proteins (CRICK, 1958). However, different cells need to perform different functions, and so regulating the extent to which genes are expressed is a crucial process for a cell's function. This process is managed by a range of mechanisms including DNA methylation, micro-RNAs and transcription factors.

In our work, we focus on the gene regulation through transcription factors, which begins with the transcription factors binding to enhancers, that is, regions along the genome that can upregulate or downregulate a gene's expression. This requires that the enhancer sequence contains a transcription factor binding site, that is, a part of the genomic sequence that matches the transcription factor's motif. Many models assume that the more TF binding sites there are, the more likely a transcription factor can bind to it, leading to upregulation/downregulation of a target gene.

Transcription factor binding is further complicated by the fact that the DNA is wound up in many layers to ensure that it can compacted into the nucleus of a cell. This means that only accessible regions of the DNA can be bound by the transcription factor, regardless of whether it has a matching motif. Furthermore, this accessibility can vary between cells, even if they are of the same cell type which contributes to the heterogeneity within cell populations. However, a sub-category of TFs, called pioneer transcription factors, are believed to open up the local chromatin structure which allows for other transcription factors to bind to the enhancer. This facilitates the upregulation or downregulation of genes, and so it is believed that they may play a key role in cell reprogramming (Zaret and Carroll, 2011; Iwafuchi-Doi and Zaret, 2016).

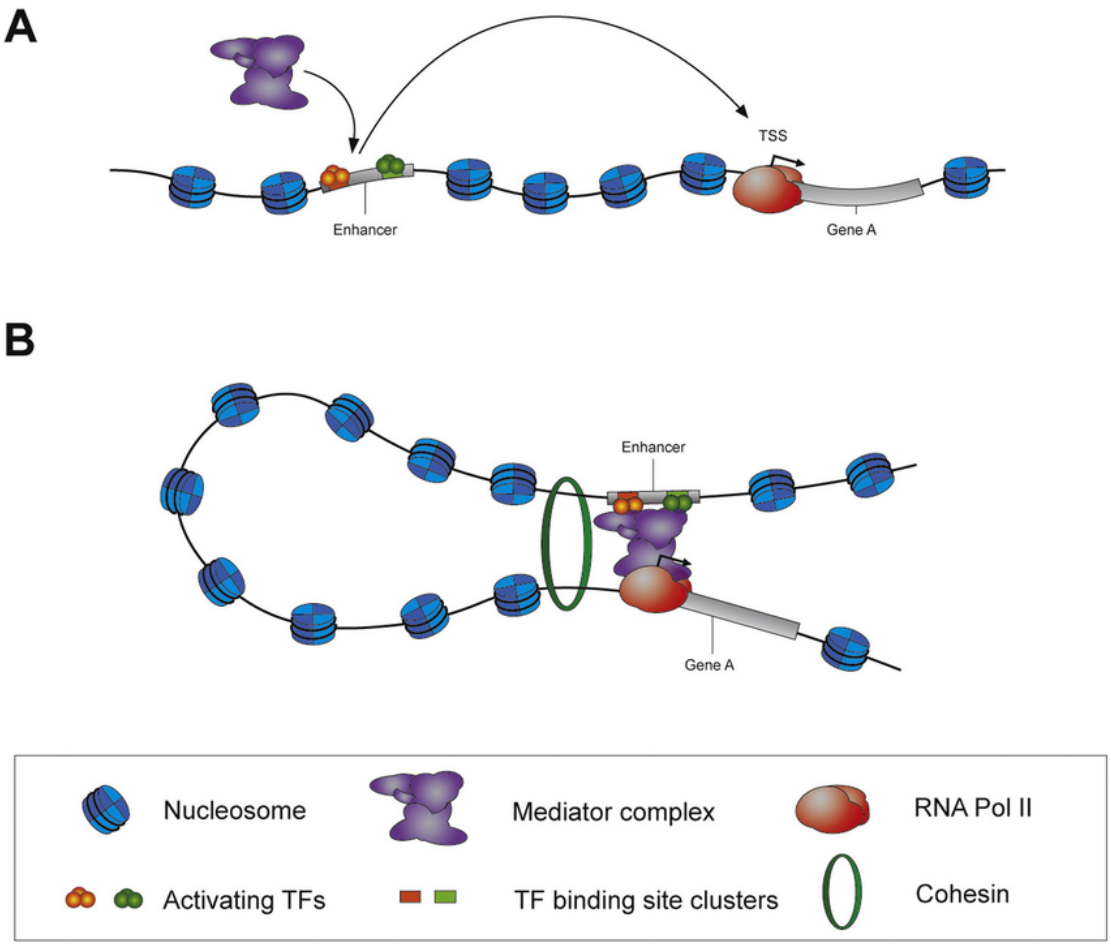


Figure 1.2: Gene regulation through enhancers.

Source: Alexandre Marand, Tao Zhang, Bo Zhu and Jiming Jiang. (2016). Towards genome-wide prediction and characterization of enhancers in plants. Biochimica et biophysica acta. 1860. 10.1016/j.bbagrm.2016.06.006.

When a transcription factor binds to an enhancer, it can form a complex with other

proteins that attaches to the promoter of a gene. This is a chromatin interaction called a DNA loop that occurs between two parts of the DNA sequence. This process is difficult to model as an enhancer can regulate many genes (Fukaya et al., 2016; Qin et al., 2020), and a gene can be regulated by many enhancers (Osterwalder et al., 2018; Qin et al., 2020).

Transcription factors are coded by genes, which themselves are in turn regulated by other transcription factors. This leads to a network of many interconnected genes that regulate each other, often forming complex feedback loops and cascades, allowing cells to remain in stable states. These are called Gene Regulatory Networks (GRNs), and as an indicator to the complexity of these networks, it is estimated that humans have more than 20,000 genes. Understanding the GRN is an important task for understanding the effect of perturbing a transcription factor, as it would provide insight to the downstream result on gene expression, and cell identity. However, inferring the structure of GRNs is a very challenging task, and is an active area of research further elaborated upon in Chapter 2.

1.3 Technology

Many recent advances in biology have been supported by the development of sequencing technologies which act as a lens into biological processes at the molecular level. Historically, sequencing technologies could only process bulk samples, that provided a single measurement for a mixture of cells. Although very useful, it is limited in that it only provides an overall average measurement, and does not provide insight into the heterogeneity within the cell population. In recent years however, we have seen the rise of single-cell sequencing techniques, which allow us to overcome this problem and obtain measurements on individual cells, revealing the heterogeneity of the cell population (Kulkarni et al., 2019). This is crucial for cell reprogramming where we observe vastly varying results when cell populations are exposed to the same perturbations (Weinreb et al., 2020).

Many modes of data can now be collected at the single cell level, which give insight into the different regulatory components of cells. A few important examples include:

- scRNA-seq which measures the level at which a gene is being expressed. After preprocessing, this would result in a $gene \times cell$ matrix of count data. Current technologies

are able to sequence on the order of 10^2-10^5 cells for ∼20000 genes with a total read depth (i.e., count) ranging from 10^4-10^6 per cell (Haque et al., 2017). Here, higher throughput protocols, that is methods which can sequence a large number of cells, often have lower read depth. This means that high throughput technologies tend to be very sparse, often attributed to "drop-out" events where lowly expressed genes are not captured, and will appear as a zero in the data.

- scATAC-seq which detects accessible regions (often called peaks) along the genome. After preprocessing, this would result in a binary $peak \times cell$ matrix of accessibility data. However, regions along the genome can grouped into bins giving a $bin \times cell$ matrix with count data. Current technologies are able to sequence on the order of 10^2-10^4 cells for a read depth ranging from 10^3-10^4 per cell (Fang et al., 2021). scATAC-seq protocols often have a very low detection rate, leading to data that is sparser than scRNA-seq.

- Hi-C which detects the spatial chromatin interactions along the DNA. After preprocessing, this generates a $region \times region$ matrix of count data with the detected contacts between each pair of genomic regions. However, these matrices can be extremely sparse, so they are often binned into large intervals (often 10kb-25kb) (Cameron et al., 2020).

Most techniques are only able to capture one of these modes. While it is still insightful, such single-layered modality provides us with only one perspective of the cell's intracellular dynamics. Recently, new multimodal single cell techniques are able to sequence multiple modes of data from the same cell. This includes sci-CAR (Cao et al., 2018), SNARE-seq (Chen et al., 2019) and SHARE-seq (Ma et al., 2020) which can perform both scRNA-seq and scATAC-seq within the same cell. These multimodal sequencing techniques will be able to reveal much more about the intracellular dynamics governing a cell's function. Furthermore, as these multimodal sequencing techniques are still in their infancy, there is a need for novel mathematical and statistical models to use this data to describe a cell's intracellular dynamics.

In general, a constant challenge is that omics data is generally both quite sparse and

noisy which may not be appropriate for standard statistical analyses. In addition to this, scRNA-seq can often fail to capture lowly expressed genes. Consequently, we may have insufficient data on transcription factors that are known to play an important role, but are simply lowly expressed (Martin and Sung, 2018; Haque et al., 2017). This is further complicated by the limited data of this type, as sequencing experiments can be expensive, and require specialised equipment and expertise. This motivates the need for simulation frameworks to generate synthetic data which has been a powerful tool to evaluate and benchmark bioinformatics tools.

1.4 Data and simulations

In our thesis, we leverage data from multimodal sequencing technologies as this provides a more holistic view of the regulatory mechanisms occuring at the cellular level. In particular, we focus on the SHARE-seq data collected by Ma et al. (2020) on adult mouse skin cells. We choose this data set as it is the highest quality matched scRNA-seq and scATAC-seq data available at this time (Ma et al., 2020) and a subset of these cells form a natural binary cell fate decision. In this case, Transit Amplifying Cells (TACs) differentiate into either an Inner Root Sheath (IRS) lineage or a Hair Shaft lineage (containing Medulla, Cuticle and Cortex Cells). In particular, this example of directed differentiation of hair cells is of great interest in regenerative medicine, as hearing loss can occur when auditory hair cells die in the ear. The ability to reprogram surrounding cells into auditory hair cells could potentially be a treatment for hearing loss (Walters et al., 2017; Luo et al., 2013; Duncan and Fritzsch, 2013).

When developing mathematical models, it is important to benchmark them to assess the efficiency and effectiveness of each model. This requires an extensive range of data sets to see the models' performance under different situations. However, experimental data is expensive and time-consuming to collect, so there is a demand for simulation frameworks which can produce large artificial data sets that resemble the real data. Many simulation frameworks have been developed, such as Splatter (Zappia et al., 2017) for scRNA-seq data and simATAC (Ghaziani et al., 2020) for scATAC-seq data. However, many of these methods only simulate one mode of data, which can't model the emerging multimodal data,

and these methods are usually designed to model the distribution of the data, rather than the biological process of how the data is generated.

A more recent framework, dyngen (Cannoodt et al., 2020), addresses this issue by modelling the amount of pre-mRNAs, mature mRNAs and proteins for each gene. This has many useful applications, in particular modelling the amount of pre-mRNA and mature mRNAs allows the use of RNA-velocity (discussed in Section 2.2). However, this doesn't incorporate or use scATAC-seq information. This means that there are limited simulation frameworks available for newer types of multimodal data being generated, despite the growing demand from the mathematical models being developed.

1.5 Summary

There are still many gaps in our understanding of gene regulation and the mechanisms behind cell reprogramming. Modelling the gene regulatory process and the effect of TF perturbation has proven to be a mathematical and computational challenge. However, with the availability of new technology that has rapidly developed in recent years, we now have access to a vast amounts of data that can allow us to resolve these unanswered questions. In this thesis, I will close this gap by leveraging an emerging type of data to bring new insights with a statistical model for cell reprogramming and developing a corresponding simulation framework.

Chapter 2

Statistical perspectives on cell reprogramming methods

Quantifying the gene regulatory network, how gene expression changes over time, and under perturbations has always been and remains a key challenge in cell biology, and is an important prerequisite for cellular reprogramming. The computational challenges have been the interpreted translation of these biological questions into a mathematical or statistical problem. However, the translation of problems from biology to statistics have remains a one-to-many mapping, which has resulted in a wide variety of approaches and models that address similar but different aspects of these problems. My novel contribution in this chapter is the synthesis of current approaches into four main categories, and critically evaluating some of the existing models for these biological processes and examining their successes and limitations. The current computational literature on cellular reprogramming can be broadly classified into four main categories:

- **RNA Velocity:** A method that measures how a cell's gene expression is changing at the time it is sequenced. This method shifts the perspective of gene expression from a static value, to one that changes over time.
- **Gene regulatory network (GRN) inference:** Many methods exist to infer the GRN, allowing us to understand how transcription factors and target genes are interacting with one another. Here will explore SCENIC as it integrates scRNA-seq and ATAC-seq data in its estimation.
- **TF identification for cell reprogramming:** A variety of methods that predict the key transcription factors that could drive cell reprogramming. These methods often perform a GRN inference step, to guide the selection of candidate TFs.

- **Modelling transcription factor perturbations:** Many methods to quantitatively model the GRN, and predict how gene expression changes in response to a perturbation in transcription factor concentration.

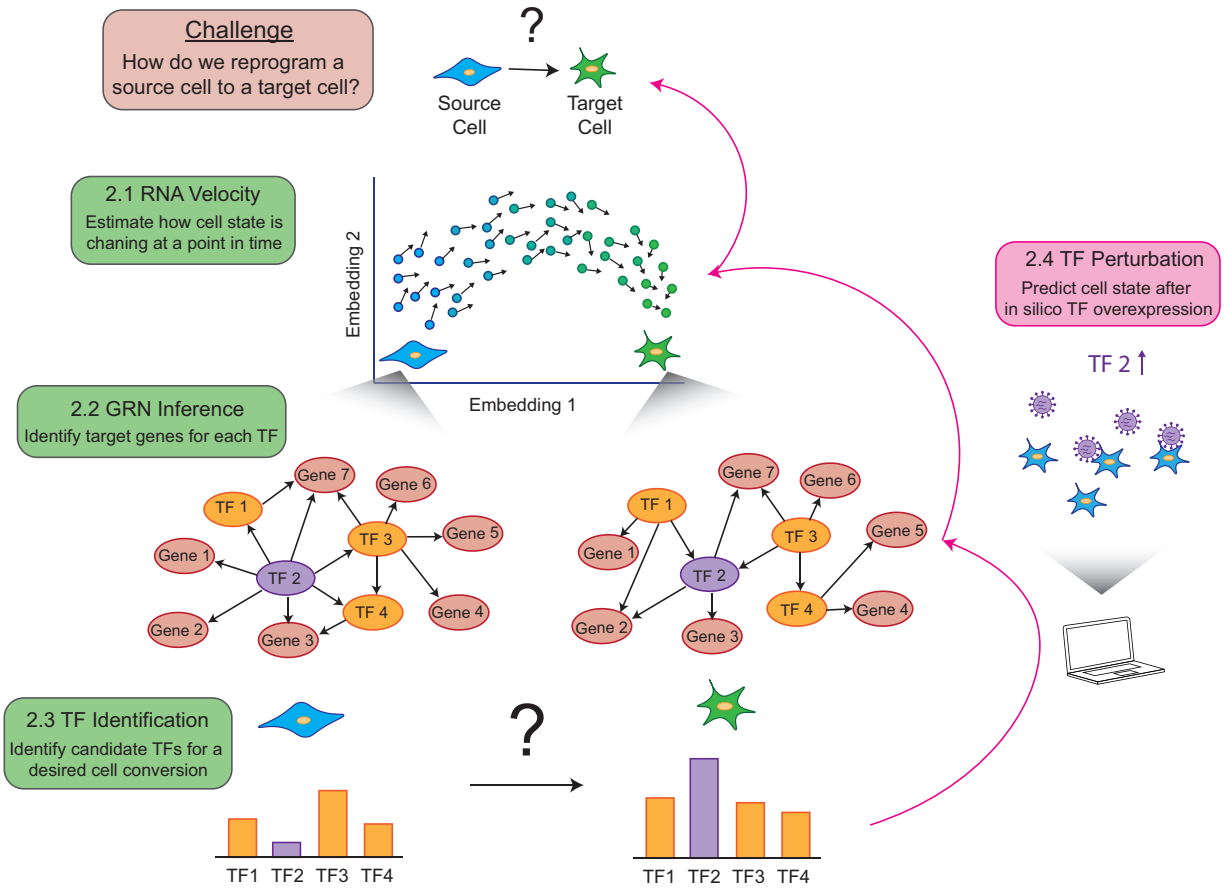


Figure 2.1: Overview of varying problems in systems biology and their respective goals.

2.1 RNA Velocity

A significant limitation of current RNA sequencing technologies is that the sequencing protocol results in the destruction of the cell. This means that any measurements only represent a static snapshot of the cell's state in time, which in reality would be constantly changing.

RNA Velocity is a method to partially resolve this problem, by modelling the rate of change of RNA in a cell at the time of sequencing (La Manno et al., 2018). Knowing this velocity effectively predicts the future state of the cell in the short-term. To calculate this,

it is assumed that the cell should be in a steady state equilibrium where the production of spliced mRNA from unspliced mRNA should balance out with the degradation of spliced mRNA. This means that any excess or deficit in mRNA splicing should manifest as a subsequent change in gene expression.

Application

RNA velocity was originally applied to the developing mouse and human brain, to predict the future cell state on the timescale of hours. It has since been applied to many other cell types and has been generalised to model transient cell states (Bergen et al., 2020).

Limitations

Unfortunately, using this method doesn't predict what happens after a perturbation for which we do not have the data for, which is a necessary goal in modelling cell reprogramming.

2.2 Gene regulatory network inference

Deciphering the gene regulatory network is an important task in understanding how a cell chooses and maintains its identity. As a result, many algorithms have been developed to infer the GRN using a variety of techniques such as ordinary differential equations (ODEs), correlation and partial correlation, mutual information, Boolean models, and regression (Pratapa et al., 2020). In this section, we will explore the algorithm behind SCENIC (Aibar et al., 2017), a widely used method which performs well in a high profile benchmarking study (Pratapa et al., 2020). Furthermore, it infers the GRN by incorporating different biological components of gene regulation which will be a key motivation for our novel model introduced in Chapter 3.

SCENIC

The key idea behind this technique is to identify the target genes that each transcription factor regulates which defines the GRN. This is done in two main steps: identifying

coexpressed genes, and then narrowing them down to direct targets.

Step 1: identifying coexpressed genes

For a gene to be a target of a TF, coexpression is generally considered a necessary condition so that when a TF expression varies, its target genes should vary accordingly. SCENIC does this by using a random forest regression for each TF to find any coexpressed genes. The choice of a random forest regression is motivated by the desire to capture nonlinear associations, as opposed to correlation based methods (Pouyan and Kostka, 2018).

Step 2: narrowing down to target genes

Although coexpression is considered a necessary condition for a gene to be a target of a TF, it is not always a sufficient condition. This is because many TFs can regulate each other, forming a complex GRN where a perturbation of one TF can cause large downstream effects through intermediary TFs. Thus, it is an important task to identify the direct targets amongst the coexpressed genes.

SCENIC address this by filtering the list of target genes to those where the TF-binding motif is enriched in the area around the transcription start site (TSS). This can provide a more accurate picture of the direct targets of each transcription factor by removing downstream targets for which the TF cannot bind to its TSS.

Once it has narrowed down the list of target genes for each transcription factor, they are grouped together into a single module, called a regulon, all of which forms the GRN.

Application

Now that SCENIC has identified each TF's targets, it is able to measure the activity of each regulon in a cell, based on the enrichment of that gene set. The regulon activity can then provide insight on the key processes taking place inside the cell, which is used to cluster cells based on cell type, and also reveals cellular subtypes. SCENIC has demonstrated its utility in data sets from different species (human and mouse), and different complex tissues (oligodendroglioma and melanoma) (Aibar et al., 2017).

Limitations

Although SCENIC has been shown to be quite successful (Pratapa et al., 2020), there are a few key limitations. Firstly, coexpressed genes are narrowed down by only using TF motif enrichment in the area around the TSS. This is not necessarily representative of the regulation that occurs via enhancers which can be up to 1Mb away from the TSS of the gene (Pennacchio et al., 2013). Furthermore, SCENIC is only able to identify gene targets of each TF, and does not infer GRN dynamics. This is an important task if we wish to model the effect of perturbing transcription factors, which will be explored in Section 2.4.

2.3 Transcription factor identification for cell reprogramming

The potential for cell reprogramming to be applied in regenerative medicine has received much attention in recent years. However, one of the greatest roadblocks to this progress is the challenge of identifying the transcription factor combinations that will yield a desired conversion.

Fortunately, the development of microarray and RNA sequencing technologies gave rise to a more holistic view of the gene expression in a cell and led to the compilation of extensive gene expression databases. This facilitated a new wave of computational tools to identify transcription factor combinations for cell reprogramming, even some of which have become commercial products (Rackham et al., 2016).

The general motivation behind all of these methods is that to reprogram a source cell type to a target cell type, a key TF should be relatively highly expressed in the target cell type, with its target genes also highly expressed.

2.3.1 D'Alessio *et al.*

One of the earlier TF identification methods uses the idea that key TFs for reprogramming to a target cell type should not only have the TF relatively **highly expressed** in the target cell type, but also expressed in a **cell type specific fashion** (D'Alessio et al., 2015).

Algorithm

To quantify this, D'Alessio and colleagues define a **specificity score** for each TF in each target cell type, that is how well the TF is highly expressed in the target cell type, but not other cell types. This considers the true expression of the TF in all cell types from a database, and compares it to an "ideal" case. In this ideal case, the TF will be highly expressed in the target cell type but not expressed at all in the "background" cell types (all other cell types in the database).

The Jensen-Shannon Divergence is used to measure the difference between these two distributions. Let $P_x(c)$ represent the **normalised true expression** of TF x in each cell type c, and let $Q_x(c)$ represent the **normalised ideal expression** of TF x in each cell type c. Notice that this means $Q_x(c) = 1$ if c is the target cell type and $Q_x(c) = 0$ if c is any other cell type. Then the Jensen-Shannon Divergence is given by

$$JSD(\,P\,||\,Q\,) = \tfrac{1}{2}D_{KL}(\,P\,||\,M\,) + \tfrac{1}{2}D_{KL}(\,Q\,||\,M\,),$$

where $M = \frac{1}{2}(P + Q)$ and D_{KL} refers to the Kullback–Leibler divergence so that

$$D_{KL}(\,P\,||\,M\,) = \sum_{c \in \mathcal{C}} P_x(c) \log\left(\frac{P_x(c)}{M_x(c)}\right).$$

The Jensen-Shannon Divergence can be interpreted as a smoothed and symmetrised version of the Kullback–Leibler Divergence. This can be calculated for all TFs giving each of them a score, then allowing them be ranked on their importance for each cell type. This ranked list is taken to be the key TFs to drive cell reprogramming.

Application

D'Alessio and colleagues were able to use this algorithm to determine a combination of key TFs to reprogram human fibroblasts to Retinal Pigment Epithelium (RPE) cells, a previously unperformed cell conversion (D'Alessio et al., 2015). This combination of TFs was experimentally validated which opens up the potential for cell reprogramming to be a treatment for vision loss as age-related macular degeneration is often caused by the loss of RPE cells over time.

Limitations

However, a limitation of this approach is that it only looks at the gene enrichment of the TFs rather than all genes. This means that it doesn't consider the entire GRN which may play an important role in cell fate decisions during cell reprogramming. Furthermore, D'Alessio and colleagues. don't consider the identity of the source cell, which would already have its own regulatory system in place, potentially influencing the effect of perturbing certain TFs.

2.3.2 CellNet

Another early TF identification method is CellNet (Cahan et al., 2014) which aims to incorporate information about the GRN in the TF identification. Its algorithm can be split into three main steps.

Step 1: inferring the GRN

CellNet starts by constructing a general GRN that is common to all cell types. This is done by taking a database of different cell types[1] and using the Context Likelihood of Relatedness method, a mutual information based algorithm for GRN inference.

Next, the general GRN is broken down into cell type specific GRNs to account for different cell types having different regulatory systems. This is achieved by performing community detection on the general GRN to identify core sub-networks. Then to allocate these to different cell types, gene set enrichment analysis is performed on these sub-networks to quantify its importance in one cell type compared to all other cell types.

Step 2: measuring the GRN activity

In a similar way to SCENIC's regulon activity, CellNet calculates a GRN activity[2] to measure how active a GRN is for a particular cell type. The Raw GRN Activity ($RG_{\mathcal{G}}^{s,c}$)[3]

[1] called "cell/tissue types" by Cahan and colleagues

[2] called "GRN status" by Cahan and colleagues

[3] called "RGS status" by Cahan and colleagues

for a GRN $\mathcal{G}$ in a sample s of cell type c is given by

$$RG_{\mathcal{G}}^{s,c} = \sum_{y \in \mathcal{G}} z_y^c w_y^s$$

where

- z_y^c is the z-score of the expression of gene y in cell type c relative to all other cell types from the database. This term favours genes that are relatively **highly expressed in the cell type of interest**.

- w_y^s is a weight proportional to the expression of gene y in the sample s. This term favours genes that are relatively **highly expressed in the sample**. Weights can be assigned to prioritise genes that are found to be important from the GRN inference step (e.g., from a random forest classifier).

However, the Raw GRN Activity will be strongly dependent on the size of the GRN, as larger GRNs will sum over more genes and naturally have a larger $RG_{\mathcal{G}}^{s,c}$. So to allow for comparison between GRNs, it is rescaled to the GRN Activity ($G_{s,\mathcal{G}}^c$) by dividing it by the average $RG_{\mathcal{G}}^{s,c}$ across all cell types.

$$G_{\mathcal{G}}^{s,c} = 1000 - \frac{RG_{\mathcal{G}}^{s,c}}{(\sum_{k \in \mathcal{C}} RG_{\mathcal{G}}^{s,k})/|\mathcal{C}|}$$

where $\mathcal{C}$ denotes the set of cell types and so that $(\sum_{k \in \mathcal{C}} RG_{\mathcal{G}}^{s,k})/|\mathcal{C}|$ denotes the average $RG_{\mathcal{G}}^{s,c}$ of GRN $\mathcal{G}$ in sample s across all cell types.

Step 3: calculating network influence score for a TF

Now that the activity of each GRN can be calculated, a cell reprogramming experiment can be interpreted as trying to change the activity of some GRNs in a sample (i.e., the source cell type) to match the desired target cell type.

To this end, Cahan and colleagues define a Network Influence Score[4], denoted $N_{x,\mathcal{G}}^{s,c}$, to estimate the importance of a TF x to convert the GRN $\mathcal{G}$ in sample s to match the target cell type c defined by

$$N_{x,\mathcal{G}}^{s,c} = \frac{\sum_{y \in \mathcal{G}} z_y^c w_y^s}{|\mathcal{G}|} + z_x^c w_x^s \tag{2.1}$$

where

[4] called N by Cahan and colleagues

- $|\mathcal{G}|$ is the number of target genes of x in $\mathcal{G}$.
- z_y^c (similarly z_x^c) is the z-score of the expression of gene y (TF x) in cell type c relative to all other cell types from the database. This term favours genes/TFs that are relatively **highly expressed** in the cell type of interest.
- $w_y^{s,c}$ (similarly $w_x^{s,c}$) is the difference in expression of gene y (TF x) for the sample s and target cell type c in the database. This term favours genes (or TFs) that are **differentially expressed** in source cell type and the target cell type.

This way, the first term in (2.1) represents the dysregulation of the **target genes** of TF x between the source and target cell type[5]. The second term in (2.1) represents the dysregulation of the **transcription factor** x between the source and target cell type. This score can then be used to rank different TFs on how likely they may drive a desired cell conversion.

Application

Cahan and colleagues showed that this algorithm could successfully identify transcription factors to reprogram mouse neurons and cardiomyocytes. This method has since been generalised to SingleCellNet which uses scRNA-seq data to classify cells into different cell types (Tan and Cahan, 2019).

Limitations

A limitation of this approach is that it only considered gene expression data, which means that regulatory links are established between coexpressed TFs and genes, even though they may not be direct targets.

2.3.3 Mogrify

Extending on the models produced so far, Rackham and colleagues took advantage of the expanding omics databases to incorporate the downstream effect of overexpressing

[5]Cahan and colleagues calculate the sum whereas here we are showing the mean.

transcription factors (Rackham et al., 2016). Their algorithm, Mogrify, comprises of three main steps[6].

Step 1: differential expression calculation

Firstly, a score is calculated representing how differentially expressed each gene is for each cell type[7]. To do this, a database of gene expression profiles is used, and in this case the FANTOM5 data set was chosen.

To compute a score for differential expression, the expression of a gene for a cell type needs to be compared to some background. The background is chosen to be the cell types that are not too related to the cell type of interest, but also not too distant. This set of cell types is determined by creating a tree based on cell ontology and by choosing breaking points near the top of the tree. This can be contrasted to the work of D'Alessio and colleagues who use all other cell types in their database as the background.

Now that the background has been created, Mogrify calculates a score for differential expression for each gene y for a cell type c as

$$G_y^c = |L_y^c|(-\log_{10} P_y^c) \tag{2.2}$$

where L_y^c is the log-transformed fold change in the expression of gene y in the cell type c compared to the background. Here, P_y^c is the adjusted P-value for gene y in cell type c compared to the background. This way, a high score will correspond to a gene that is both highly expressed and differentially expressed.

Step 2: calculating network influence score for a TF

Like CellNet, Mogrify aims to account for the effect of overexpressing TFs on their gene targets by calculating a Network Influence Score. However, Mogrify uses a couple extra databases to incorporate additional omics information to estimate the GRN:

- the MARA database, providing protein-DNA interactions, identifying TFs that bind to the promoters of genes.

[6] Rackham and colleagues use 7 steps in their original paper which we group into 3 main steps

[7] called sample by Rackham and colleagues

- the STRING database, providing many other types of interactions like protein-protein and biological pathways.

Then, to quantify the influence of a TF x onto its gene targets y in the GRN $\mathcal{G}$, a weighted sum of gene scores is calculated by

$$N^c_{x,\mathcal{G}} = \sum_{y \in \mathcal{G}_x} G^c_y \cdot \frac{1}{L_{y,\mathcal{G}_x}} \cdot \frac{1}{O_{y,\mathcal{G}_x}} \tag{2.3}$$

where

- $\mathcal{G}$ denotes the GRN that is being used, in this case it will be the GRN derived from either the MARA or the STRING database.
- $\mathcal{G}_x$ denotes the local subnetwork of $\mathcal{G}$ around the TF x. By default, it is chosen to be the set of genes that are at most 3 nodes away from x. This way, the summation in Equation (2.3) is over all genes y in this subnetwork $\mathcal{G}_x$.
- $L_{y,\mathcal{G}_x}$ is the number of steps away gene y is from TF x in network $\mathcal{G}_x$. This way, more weight can be allocated to the direct targets of a TF as opposed to its downstream targets.
- $O_{y,\mathcal{G}_x}$ is the outdegree of the parent of y in subnetwork $\mathcal{G}_x$. This weighting prioritises the specificity of a TF regulating important target genes. This way, TFs that regulate a very large number of genes do not appear overly important.

Step 3: rank TFs and filter

The scores from each of G^c_x, $N^c_{x,MARA}$ and $N^c_{x,STRING}$ are used to rank the transcription factors in 3 different lists, and their ranks are summed together to give a final rank of the importance of each transcription factor in each cell type.

However, for a desired cell reprogramming experiment, the source cell may already have some key TFs highly ranked, indicating that the TF and its targets are already highly expressed. Thus, the list of TFs is filtered down to remove those which are highly ranked in the list for the source cell type.

Furthermore, many TFs would share similar gene targets, and so it may not be necessary to include all such TFs in the final list of key TFs. To account for this, a TF is removed

from the list if there is a higher ranking TF that regulates over 98% of its targets. This produces a final ranked list of TFs which are deemed to be the key TFs for the desired cell conversion.

Application

Mogrify was successful in recapitulating known TFs for reprogramming experiments, to a higher accuracy compared to previous methods. Rackham and colleagues also used the algorithm to discover two new human cell conversions that were experimentally validated, namely fibroblast to keratinocyte and keratinocyte to microvascular endothelial cells. The list of key TFs for any conversion are available on a web based tool, and is also commercially used under the same name.

Limitations

Mogrify relies on information from databases like MARA and STRING to infer the GRN, which means that it can only incorporate experimentally validated regulatory relationships. However, this may only be a snapshot of the true underlying regulatory relationships.

2.3.4 Lisa

In a similar way to SCENIC, Lisa (Qin et al., 2020) is an algorithm that aims to predict the TFs that regulate a set of genes (in this case, we are interested in a set of differentially expressed genes). However, as opposed to using gene coexpression and TF binding motifs, it uses data from ChIP-seq experiments. These experiments can determine the affinity to which a TF binds to each part of the genome. These regions can be interpreted as enhancers which will regulate some nearby genes.

Step 1: calculate regulatory potential

Firstly, Lisa uses a database of ChIP-seq experiments to calculate the regulatory potential R_{xy} of each TF x to regulate each gene y. This is given by

$$R_{xy} = \sum_{k \in \mathcal{E}_y} w_{yk} s_{xk} \tag{2.4}$$

where $\mathcal{E}_y$ is the set of enhancers within 100kb of the transcription start site (TSS) of gene y, the value s_{xk} is the signal strength from the ChIP-seq or DNase-seq experiment for TF x on enhancer k, and w_{yk} is a weight that decays exponentially with the linear genomic distance between enhancer k and the TSS of gene y. This way, the regulatory potential R_{xy} sums up the regulatory potential from each enhancer to obtain an overall measure of how strongly a TF can regulate each gene. However, to allow for comparison between TFs, the regulatory potential is normalised as

$$R'_{xy} = \log(R_{xy} + 1) - \frac{\sum_{j \in \mathcal{Y}} \log(R_{xj} + 1)}{|\mathcal{Y}|} \tag{2.5}$$

where $\mathcal{Y}$ represents the set of genes so that R'_{xy} log transforms R_{xy} and compares it to the mean log-transformed values for R_{xj} over all genes j.

Step 2: identify TFs to predict query gene set

To identify a list of key TFs, Lisa aims to discriminate the query gene set from 3000 background genes, chosen to represent a variety of different genes. This is done by an L1-regularised logistic regression using the normalised regulatory potential which finds an optimal set of TFs to discriminate the query gene set.

Step 3: rank TFs

The motivation behind the ranking procedure is that key TFs should be binding to the enhancers close to many of the query genes, so the removal of these TFs should significantly lower the predictive power of the TF set calculated in Step 2. For each TF x and gene y, the enhancers around the TSS of gene y are removed and the regulatory potentials are re-calculated using Equation (2.4)[8]. After normalisation using Equation (2.5), the differences in regulatory potential, $\Delta R'_{xy}$ are calculated, weighting the results with the coefficients from the logistic regression. The $\Delta R'_{xy}$ of each gene in the query set are compared to the background gene set by a Wilcoxon rank-sum test. The TFs are then ranked by P-value, giving the list of key TFs.

[8]This step is called "in silico deletion" by Qin and colleagues

Application

Qin and colleagues show that Lisa is able to recapitulate the key TF for multiple knockdown and overexpression experiments (Qin et al., 2020). This included a *GATA6* knowckdown experiment on a stomach cancer cell line, upregulation of *AR* on a prostate cancer cell line and upregulation of *GR* on a lung cancer cell line among others. They demonstrated that their ranking system generally ranks the ground truth more accurately than earlier methods that perform the same task using similar data.

Limitations

The use of this algorithm relies on ChIP-seq data for all potential TF regulators which may not be practically possible due to the limited number of antibodies that can be used. A workaround proposed by Qin and colleagues is to use chromatin accessibility data from DNase-seq or ATAC-seq paired with TF binding motifs as a proxy for the affinity of a TF to regulate a gene.

2.3.5 ANANSE

Using a similar idea to Lisa, ANANSE incorporates enhancer information to build a more accurate estimate of the GRN (Xu et al., 2020), extending the Mogrify algorithm. There are two key steps in this algorithm.

Step 1: inferring the GRN

Firstly, a ChIP-seq database is used to identify putative locations of enhancers and an intensity score for each enhancer. The probability of a transcription factor binding to an enhancer is then predicted with a logistic regression, using the motif enrichment Z-score and enhancer intensity as predictors.

Similar to Lisa, a **regulation potential**[9] is then calculated for each TF x on gene y as a weighted sum of TF binding intensities onto each enhancer k of the gene.

$$R_{xy} = \sum_{k \in \mathcal{E}_y} w_{yk} s_{xk} \tag{2.6}$$

[9] called "TF-gene binding score" by Xu and colleagues

where $\mathcal{E}_y$ is the set of enhancers within 100kb of the transcription start site (TSS) of gene y, s_{xk} is the predicted binding intensity of TF x onto enhancer k, and w_{yk} is a weight that decays exponentially with the linear genomic distance between enhancer k and the TSS of gene y. This way, the regulatory potential R_{xy} sums up the regulatory potential from each enhancer to obtain an overall measure of how strongly a TF can regulate each gene.

This quantifies the overall potential for a transcription factor to regulate a gene via enhancers. However, for a true regulatory relationship, one would expect that both the transcription factor and target gene should be highly expressed. Thus, a TF-gene **interaction score** is then calculated to be the mean of the TF-gene binding score, the TF expression and gene expression.

$$P_{xy} = \frac{R_{xy} + g(x) + g(y)}{3}$$

where $g(x)$ and $g(y)$ denote the gene expression of TF x and gene y respectively. This way, a high TF-gene interaction score means that the TF and target gene are highly expressed, with the possibility for regulation via enhancers. This score can then be used to identify the strongest TF-gene interactions which can be used as an estimate of the GRN.

Step 2: calculating network influence score for a TF

In a similar way to Mogrify, a network influence score can be calculated for each TF measuring how important it will be for a desired conversion. To do this, a differential GRN is obtained by taking the TF-gene interactions of the target cell type and removing any interactions present in the source cell type. This differential GRN can be interpreted as the required regulatory changes for the desired cell conversion. The network influence score $N^{s,c}_{x,\mathcal{G}}$ of a TF x on the differential GRN $\mathcal{G}$ of source cell type s and target cell type x is calculated by

$$N^{s,c}_{x,\mathcal{G}} = \sum_{y \in \mathcal{G}_x} |G^{s,c}_y| \frac{P_{x,y}}{L_{y,\mathcal{G}_x}}$$

where

- $\mathcal{G}$ denotes the differential GRN calculated above.
- $\mathcal{G}_x$ denotes the local subnetwork of $\mathcal{G}$ around the TF x. By default, it is chosen to be

the set of genes that are at most 3 nodes away from x. This way, we are summing over all genes y in this subnetwork.

- $|G_y^{s,c}|$ is the log-transformed fold change of the expression of gene y between the source cell type s and target cell type c.
- $P_{x,y}$ is the TF-gene interaction score of TF x onto gene y as calculated in step 1.
- $L_{y,\mathcal{G}_x}$ is the number of steps away gene y is from TF x in subnetwork $\mathcal{G}_x$. This way, more weight can be allocated to the direct targets of a TF as opposed to its downstream targets.

Essentially, $N_{x,\mathcal{G}}^{s,c}$ sums the expression of each gene target, weighted by the TF-gene interaction score, and the distance of regulation. This means that a high scoring TF will have target genes that are differentially expressed, and have a strong potential to bind to the enhancers near the gene. All TFs are then ranked by their network influence score, which forms the final list of key TFs.

Application

ANANSE was shown to successfully recapitulate many key TFs for known cell reprogramming experiments to a higher accuracy than all earlier approaches. Xu and colleagues also compile a list of key TFs in different human tissues.

Limitations

However, a major limitation of ANANSE (and Lisa) is that they assume that the regulatory effect of an enhancer decays with linear genomic distance. This does not account for the three dimensional spatial chromatin structure where distal enhancers can form DNA loops. Furthermore, ANANSE doesn't consider the accessibility of the enhancers which is required for a TF to be able to bind to the enhancer.

2.3.6 Transcription factor identification summary

All of these transcription factor identification methods have led to the discovery of many novel cell conversions, with exciting applications in regenerative medicine.

Table 2.1: **Summary of TF identification methods for cell reprogramming.**

Method	Year	Data	GRN Estimation	Ranking Algorithm
D'Alessio *et al.*	2015	RNA-seq	N/A	Jensen-Shannon Divergence to identify TFs that are highly expressed in a cell-type specific way.
CellNet (Cahan *et al.*)	2014	RNA-seq	Mutual information to build main network, community detection to identify sub-networks	GRN activity estimated for each subnetwork. Network Influence Score calculated to estimate impact on GRN activity.
Mogrify (Rackham *et al.*)	2016	RNA-seq	Known interactions from protein-protein interaction databases (STRING and MARA)	Network Influence Score calculated to estimate importance for TF to regulate required genes.
Lisa (Qin *et al.*)	2020	ChIP-seq, DNase-seq	Regulatory potential estimated from ChIP-seq	Identifying key regulators of differentially expressed genes based on regulatory potential.
ANANSE (Xu *et al.*)	2021	RNA-seq, ChIP-seq	Regulatory potential estimated from ChIP-seq	Network Influence Score calculated to estimate importance for TF to regulate required genes.

However, they are limited in that they are only able to rank transcription factors, failing to identify the combinations or quantities that are needed for a desired conversion. They also do not determine the effect of overexpressing these TFs and so does not account for the

efficiency of such perturbations. Furthermore, by only considering the source and target state, these methods would not identify transcription factors that need to be expressed transiently in the reprogramming process.

In addition, most of these methods were validated and benchmarked against each other by recapitulating key TFs of experimentally validated reprogramming experiments. However, this may be a source of bias as this does not account for possible TF combinations that have not yet been discovered. This motivates the need for a model that can predict the effect of perturbing a transcription factor.

2.4 Modelling transcription factor perturbations

To model the process of cell reprogramming, one must model the underlying gene regulatory network and quantitatively predict the effect of perturbing a transcription factors. Several types of models have been proposed, which each bring a different perspective to the complex GRN dynamics. The models that will be discussed include Boolean network models, dynamical systems models and regression models.

2.4.1 Boolean network models

The first approaches to model a GRN came in the form of Boolean Networks (Bornholdt, 2008), where the activity of each gene is simplified into an "on" or "off" state. This way, every gene is represented as a node in a network, and regulatory relationships between TFs and genes are represented as a directed edge. The network then follows fixed rules which govern how each state of the network leads to a future state. As there are only finitely many possible states, one can find steady states or cycles which reflect stable cell states, and the downstream effect of perturbing TFs on these states can then be modelled.

Although the model is simple, it has been successfully used to model the GRN of differentiation and reprogramming processes like hematopoiesis, cardiac development and myeloid differentiation (Mandon et al., 2019).

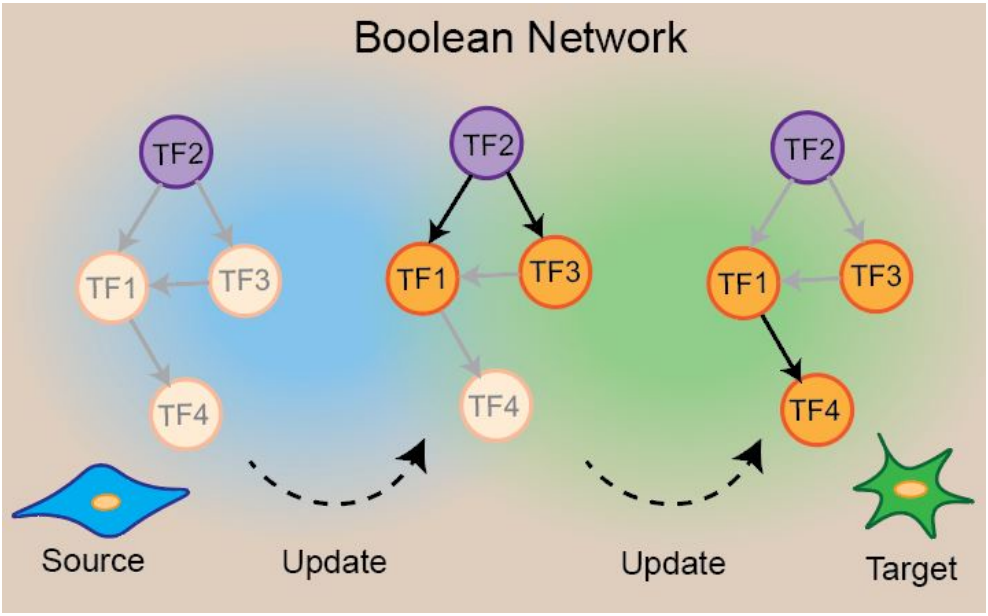


Figure 2.2: Identifying key TFs in a Boolean network model.

Hopfield neural networks

An innovative approach to model the evolution between cell states was developed by Lang and colleagues, who used Hopfield neural networks (Lang et al., 2014) to model the entire epigenetic landscape, as described by Waddington (Waddington, 1966). This landscape H is constructed by using known experimental results:

$$H = H_{basin} + H_{bias} + H_{culture} + H_{switch}$$

where

- H_{basin} ensures that observed cell states are valleys in the landscape, that is stable steady states.
- H_{bias} adds bias towards the TFs that will be introduced in the reprogramming experiment.
- $H_{culture}$ incorporates the culture conditions which may add environmental signals that are favourable to certain cell fates.
- H_{switch} adds the effect of regular cellular development where the cell state naturally evolves over time.

This assigns an "energy" to each possible cell state (a binarised gene expression space) which determines how cells transition from high to low energy states. Lang and colleagues successfully used this model to recapitulate some known key TFs for a variety of reprogramming protocols, such as pluripotent stem cells, cardiomyocytes, and neurons.

Toggle switches

Okawa and colleagues introduced another powerful application of Boolean networks, which can infer properties of the GRN, creating a more biologically meaningful model (Okawa et al., 2016). This model is motivated by the theory that binary cell fate decisions are directed by antagonistic TF pairs, often called **toggle switches** and these TF pairs may be key candidates for cell reprogramming (Heinäniemi et al., 2013).

In this scenario, a stem/progenitor cell state would have a balanced expression of these two TFs, but the overexpression of either TF would cause the other to be suppressed and eventually lead the cell into a distinct daughter cell fate. Thus, identifying such pairs could reveal the key TFs that drive cell fate, and possibly cell reprogramming.

This was done by calculating a Normalised Ratio (NRD) Difference for any TF pair between a progenitor and daughter cell type defined by

$$NRD = \frac{\frac{C_1^P}{C_2^P} - \frac{C_1^D}{C_2^D}}{\frac{C_1^P}{C_2^P}}$$

where C_1^P and C_2^P are the expression values of TFs 1 and 2 in the progenitor cell type and C_1^D and C_2^D are the expression values of TFs 1 and 2 in the daughter cell type.

For a binary cell fate decision, the NRD can be calculated for each TF pair in both lineages. Then, the TF pairs with significant NRD in both lineages in opposite directions are taken to be the lineage determinants.

Using this, Okawa and colleagues were able to reconstruct the GRN in a Boolean Network only using gene expression data. This allowed them to recapitulate known toggle switches, and also discover a novel toggle switch for the mouse neuronal stem cell system, which they experimentally validated.

However, a limitation of this approach is that it may only find markers of the cell fate, and not necessarily the drivers of the lineage specification.

IQCELL

A recent approach, IQCELL (Heydari et al., 2021), aims to take advantage of single cell data which provides a greater resolution into the process behind cell differentiation. Firstly, they arrange the cells into a pseudo-time ordering which can be used to infer the causality of gene interactions. They then score gene-gene interactions using mutual information between gene pairs, and then correlation is used to determine the sign of the regulation (activation or repression). The gene expression is then binarised and the pseudo-time ordering is then used to form a gene interaction hierarchy. A modified network inference strategy, the *Z3* engine, is used to find the logical rules representing how the cell states can evolve, by assigning up to four activators, and two repressors per gene.

This method has the advantage of using a data-driven approach to determine the regulatory relationships in the Boolean Network in an unbiased way. In contrast, the earlier methods required extensive experimental literature to assign model parameters which would fail to incorporate regulatory relationships that have not been experimentally validated.

Limitations

However, the Boolean aspect of the model requires the simplification of genes into an "on" or "off" state. This may not be representative of the true gene activity, for example some cell fate decisions may require genes to be expressed at specific intermediary levels, which not be able to be captured in a Boolean model.

2.4.2 Dynamical systems model

Another natural model for cell reprogramming would be to use dynamical systems, which can model how variables change over time. This has made dynamical systems fundamental to developments in many fields of science, which study how phenomena change over time. In our case, we wish to model the gene expression as a function of transcription factor concentration over time. These dynamical systems can be solved, numerically or analytically, which can reveal steady states and their stability, corresponding to stable cell states.

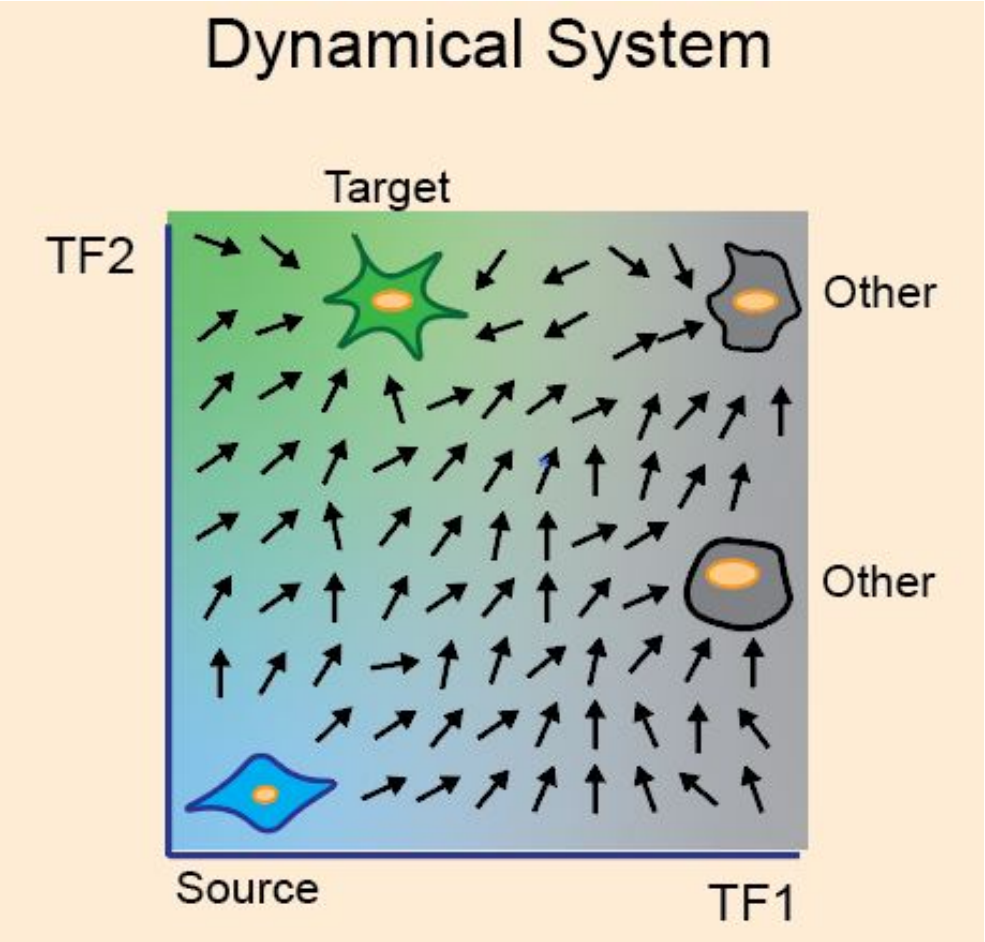


Figure 2.3: Identifying key TFs in a dynamical systems model.

Genetic feedback controllers

The time can be treated as continuous, which can be modelled with ordinary differential equations (ODEs). For example, Del Vecchio and colleagues (Del Vecchio et al., 2017) devise a blueprint for a genetic feedback controller, and model the cell reprogramming process using ODEs. Here, the regulatory effect of n TFs $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ on each other is modelled as a set of ordinary differential equations, for $i = 1, 2, \ldots, n$:

$$\frac{dx_i}{dt} = H_i(\mathbf{x}) - \gamma_i x_i + u_i$$

where $H_i(\mathbf{x})$ is the Hill function capturing the GRN regulation of TF x_i, γ_i is the constant decay rate of TF x_i, and u_i is the input of the perturbation corresponding to TF x_i.

This model was used to show that simply overexpressing TFs to some preset value may not be able to reprogram a source cell type to the desired cell fate. This is significant

because the current protocols for transcription factor overexpression forces the key TFs to be expressed at some high preset level. For a desired cell conversion, the key TFs may need to be maintained at some intermediary level, so naïvely overexpressing the TFs to a high level may force the cell to bypass the desired cell state, ending up in an undesired terminal fate.

Del Vecchio and colleagues provide the blueprints for an alternative method for TF overexpression that resolves this issue. This takes the form of a genetic feedback controller that is able to continuously update the value of u_i so that it is equal to $G_i(x_i^* - x_i)$ where x_i^* is the desired target concentration and x_i is the current concentration. This allows the concentration of the desired TF to reach the desired state, and at this point, the input is set to 0 and the endogenous system takes over and maintains the stability of the reprogrammed state.

This has been theoretically applied to model the process of reprogramming to pluripotency using two transcription factors *Nanog* and *Oct4*.

Cell-cell communication

In a similar way, Franke and MacLean (2021) use ODEs to model how cell-cell communication influences cell fate decisions during differentiation. They use the following model

$$\begin{aligned}
\frac{dG}{dt} &= -\beta_1 G + \frac{\alpha_1 A + \alpha_2 G}{1 + \gamma_1 A + \gamma_2 G + \gamma_3 GP} \\
\frac{dP}{dt} &= -\beta_2 P + \frac{\alpha_3 B + \alpha_4 P}{1 + \gamma_4 B + \gamma_5 P + \gamma_g GP + \gamma_7 GX} \\
\frac{dX}{dt} &= -\beta_3 X + \frac{\alpha_5 G + \alpha_2 G}{1 + \gamma_8 G + \gamma_9 C}
\end{aligned}$$

where G, P and X are the transcription factor concentrations, β_i are the degradation rates, α_i are the activation rates, γ_i are the inhibition rates, and A, B and C are external signals, which arise from cell-cell communication. This allows them to model the heterogeneity that is observed in cell reprogramming experiments (Weinreb et al., 2020) and provides a biologically meaningful explanation for it.

They successfully apply this to model the antagonistic TF pair *GATA1* and *PU.1* in the commitment of a myeloid progenitor cell to the erythroid/megakaryocyte lineage or granulocyte/monocyte lineage.

Difference equations

In a dynamical system, the time can be considered in discrete intervals, which can then be modelled with difference equations. For example, Ronquist et al. (2017) model $\mathbf{x}_k$, the gene expression of the cell at time k, based on the following difference equation

$$\mathbf{x}_{k+1} = \mathbf{A}_k\mathbf{x}_k + \mathbf{B}\mathbf{u}_k$$

where

- $\mathbf{A}_k$ is the transition matrix, representing the time-varying changes to the cell state at time k. It is calculated from time series data of the source cell type, representing the natural cell cycle dynamics.
- $\mathbf{B}$ is the regulatory matrix indicating the regulatory effect of each TF on each gene. This is estimated using the number of transcription factor binding sites within 5kb of the transcription start site of the gene. It is also weighted by its activity (activator or repressor) using information from a literature search, and also weighted by gene accessibility from publicly available DNase-seq data.
- $\mathbf{u}_k$ is to be estimated, representing the binary input vector indicating which TFs to introduce. The optimal value of $\mathbf{u}_k$ is found by minimising the Euclidean distance between the final state $\mathbf{x}_T$ and the target state $\mathbf{z}$.

Ronquist and colleagues use this algorithm to demonstrate that some TFs have a preference for being added at the start or end of the cell cycle. This suggests the importance of considering the timing for overexpressing TFs in cell reprogramming, especially as a cell's state is constantly changing. Furthermore, they successfully recapitulate combinations of key TFs for known cell reprogramming experiments with fibroblast as the source cell type and embryonic stem cell, myotube and heart cell as the target cell types.

Limitations

However, estimating the parameters of large dynamical systems models can be rather difficult, making these methods difficult to generalise to more complex systems. This is because they are often determined from results in experimental literature, or they need to be modelled with time course data which may not be readily available in public databases. Even

if one were to generate their own data, this would take additional time and resources, and narrows the options for source cell types to those used in the experiment.

2.4.3 Regression models

A major limitation of many of the methods discussed so far is that they assume that the initial cell population is homogeneous and can be modelled with a single GRN. This is mostly due to the fact that these methods were mostly developed when only bulk sequencing data was available, only providing a general average of the cell population.

This can be a major issue for cell reprogramming as the outcome can be highly dependent on the individual cell states. Recent single cell technologies have shown that there is usually large heterogeneity within cell populations (Weinreb et al., 2020; Gam et al., 2019), which may explain the inefficiency of cell reprogramming experiments, and can often lead to several clusters of undesired cell fates.

However, regression models are well suited to take advantage of the increasingly available single cell data that provides a lens into the dynamics of individual cells.

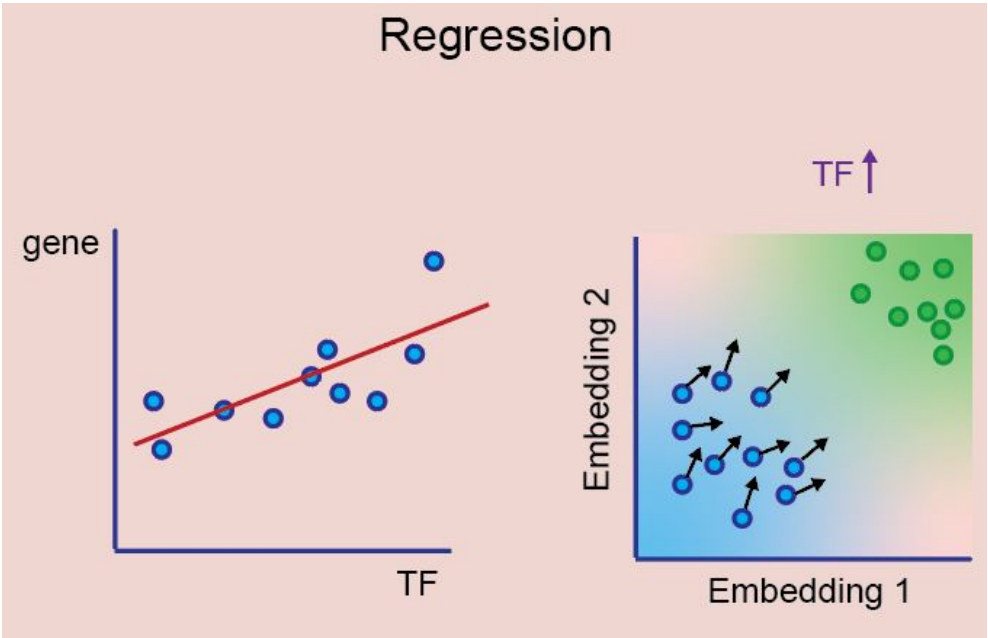


Figure 2.4: Identifying key TFs in a regression model.

CellOracle

Unlike earlier approaches, CellOracle is the first method to explicitly predict the effect of overexpressing a transcription factor at a single-cell level by inferring the GRN. This can then be used to compare the effect of overexpressing different transcription factors, helping to determine which ones should be experimentally validated (Kamimoto et al., 2020).

The first step in the algorithm is to build a base GRN structure, representing a list of potential TF-gene regulations. This is done by using scATAC-seq data, and identifying co-accessible promoters/enhancers, and then scanning the DNA sequences for TF binding motifs. This means that they only consider TF-gene pairs for which the TF has the potential to regulate the gene.

The remaining TF-gene pairs are fitted to a linear model in each cell type cluster, as it often believed that different cell types have different GRNs. This linear model is fit using Bayesian Ridge or Bagging Ridge as the regularisation helps to remove unnecessary variables, and reduces overfitting. The base regression equations are of the form

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j \tag{2.7}$$

where $\mathbf{y}_j$ is the gene expression to be predicted from TF expression $\mathbf{X}$, and $\alpha_j, \boldsymbol{\beta}_j, \boldsymbol{\varepsilon}_j$ are the fitted intercept, coefficients and residuals respectively. The coefficients of the linear model $\boldsymbol{\beta}_j$ can be used to predict the effect of perturbing a transcription factor, and the following downstream effect can be predicted by propagating this signal through the model multiple times.

By comparing the simulated gene expression values with that of the local neighbours, the identity of the reprogrammed cell can be estimated by the cell type of these local neighbours. This can be used to calculate a cell identity/state transition probability. Combining this together for all cells creates a transition trajectory graph, essentially an RNA velocity after the perturbation of a TF.

Application

CellOracle was used to computationally recapitulate known changes in hematopoiesis under TF perturbations. It also predicts *Fos* to increase the efficiency of reprogramming mouse

embryonic fibroblasts to induced endoderm progenitors which was validated experimentally.

The success of the CellOracle algorithm paves the way to computationally test the effect of perturbing transcription factor expression. This has the potential to significantly reduce the time and resources spent to experimentally reprogram cells.

Limitations

However, the regression used in CellOracle does not use the motif/ATAC-seq information, and so assumes that the effect of TF perturbations is felt evenly throughout the population. Furthermore, it only makes linear predictions, which may not be suitable for long term predictions. This could be necessary for cell reprogramming which results in large complex changes in the cell state.

2.5 Summary

All of the different types of models discussed so far have brought different perspectives to the complex challenge of modelling a GRN under a perturbation. However, each model was limited by the data that could be collected with the available technology at the time. In particular, as most methods were developed when there were only bulk sequencing technologies, we see that many of them are unable to account for cellular heterogeneity. Furthermore, more recent methods (Kamimoto et al., 2020; Franke and MacLean, 2021) are still only able to make short-term predictions about the future cell state, which would not be able to model the vast changes that occur during cell reprogramming.

However, we saw that as sequencing technologies developed, more detailed and insightful models could be developed. Given the pace at which sequencing technologies are continuing to improve, creating larger data sets at a single cell resolution, we should expect to see more methods in the future, draw greater insight from these data. In the remainder of our thesis, we will introduce our model for cell reprogramming which will leverage data from emerging single cell multimodal sequencing techniques to create more accurate and longer term predictions for cell reprogramming.

Table 2.2: **Summary of TF identification methods for cell reprogramming.**

Method	Year	Category	Data	GRN Inference
Hopfield Neural Networks (Lang *et al.*)	2014	Boolean Networks	RNA-seq	Waddington's epigenetic landscape estimated from experimental data.
Toggle Switches (Okawa *et al.*)	2016	Boolean Networks	RNA-seq	Normalised Ratio Difference of TF pairs to identify toggle switches.
IQCELL (Heydari *et al.*)	2021	Boolean Networks	scRNA-seq	Mutual information of gene expression from cells sorted by pseudo-time.
Del Vecchio *et al.*	2017	Dynamical Systems (ODEs)	RNA-seq	ODE parameters estimated from experimental literature.
Ronquist *et al.*	2017	Dynamical Systems (Difference Equations)	RNA-seq	Transition matrix estimated from time series data. Regulatory effect estimated by the number of TFBSs.
Franke *et al.*	2021	Dynamical Systems (ODEs)	scRNA-seq	ODE parameters estimated from experimental literature.
Kamimoto *et al.*	2020	Regression	scRNA-seq, scATAC-seq, TF motifs	All possible regulations filtered from scATAC-seq and TF motif data, then further filtered from regression with scRNA-seq data.

Chapter 3

scREMOTE: single cell reprogramming model through enhancers

Cell reprogramming offers a potential treatment to many diseases, by regenerating specialised somatic cells. Despite decades of research, discovering the transcription factors that promote cell reprogramming has largely been accomplished through trial and error, a time-consuming and costly method. A computational model for cell reprogramming, however, could guide the hypothesis formulation and experimental validation, to efficiently utilise time and resources. Current methods (described in Chapter 2) are unable to account for the heterogeneity observed in cell reprogramming or capture complex and long-term predictions.

In this chapter we present scREMOTE (single cell REprogramming MOdel Through Enhancers), a novel computational model for cell reprogramming that leverages data from emerging multimodal single cell sequencing techniques. Using data at single cell resolution will allow us to capture a greater extent of the regulatory systems at the cellular level, creating a more holistic model for cell reprogramming. Through our development, we propose and assess a variety of regression-based models for cell reprogramming and their respective components. These models will be motivated by the goal of modelling the regulation that occurs through enhancers. This will require the integration of several modalities of data, which capture the different mechanisms behind gene regulation.

This chapter is arranged as follows:

1. Model components;

2. Deconvolution of the chromatin conformation;

3. Regulation potential; and

4. Modelling TF perturbations.

3.1 Model components

The motivation for this model is to incorporate four key components of gene regulation (Figure 3.1) which will be the inputs into the model. If we have sequenced n cells (of k cell types) for p genes (t of which are transcription factors) with q enhancers[1], we have:

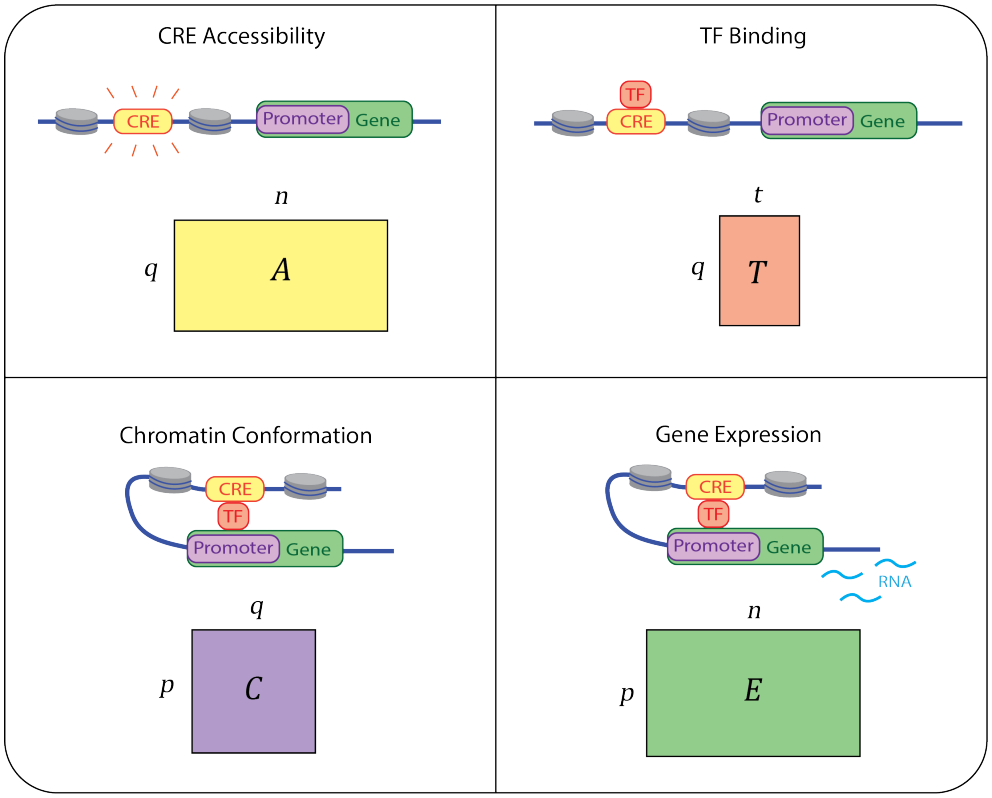


Figure 3.1: The components of gene regulation.

- **E** is a $p \times n$ matrix measuring **Gene Expression**.

  This component measures the activity of each gene in each cell and this data can be easily obtained through standard scRNA-seq protocols, for which there are extensive public databases.

[1]This model can generalise to all cis-regulatory elements, including silencers. We will simply refer to these as enhancers for this thesis.

- **A** is a $q \times n$ matrix measuring **Enhancer Accessibility**.

  This component measures the accessibility of each enhancer in each cell. Standard scATAC-seq protocols can detect accessible regions along the entire genome, sorted into bins. Using the locations of enhancers from an external database, this can correspond to the accessibility of each enhancer in each cell.

- **T** is a $q \times t$ matrix measuring **Transcription Factor Binding Probability**

  This component measures the probability that each TF will be able to bind to each enhancer. This can be quantified in a variety of ways including:

  - TF motif enrichment analysis, where it is assumed that enhancers that are highly enriched for a TF's motif have a higher chance for the TF to bind to the enhancer.
  - ChIP-seq analysis, which detects protein-DNA interactions, measuring how much a TF is binding to an enhancer (as in Lisa and ANASE).

  ChIP-seq analysis of TFs would provide the most accurate representation of TFs binding to enhancers, however there is very minimal data of this form due to limited availability of antibodies (Qin et al., 2020). Thus for our analysis, we choose to quantify **T** using TF motif enrichment analysis as the motif information and genome sequence are readily available and consistent across cell types.

- **C** is a $p \times q$ matrix measuring **Chromatin Conformation**

  This component measures the possible three dimensional chromatin interactions across the DNA for each cell type. This is used as a measure of how likely each enhancer can form a DNA loop with the promoter of each gene which may activate or repress the gene expression. This data can be collected with the Hi-C sequencing protocol, detecting chromatin interactions across the entire genome. However, current Hi-C protocols have a rather low resolution, often detecting interactions in 10kb bins (Teng et al., 2015).

  For cases where we distinguish chromatin interactions between different cell types, we will denote this as $\mathbf{C}^K$ which would give us a $p \times q \times k$ array, as there would be a **C** for each cell type.

To account for cellular heterogeneity, we ideally would want to take all measurements within the same cell. However, as of the publication of this thesis, the most recent sequencing technologies can only measure **E** and **A** in the same cell (Ma et al., 2020). Fortunately, **T** should not change between different cells as it depends on the motifs that a TF can recognise, which is constant across all cells of the same organism. This means that we can still reliably measure **T** from a database instead of each individual cell.

On the other hand, the chromatin conformation **C** is known to vary between cell types, and we are not yet able to simultaneously measure this with gene expression and enhancer accessibility. Further complicating this is that current Hi-C protocols operate at the bulk level and it is rather difficult to perform, requiring specialised equipment and expertise (Yardımcı et al., 2019). This means that there will often be limited data of this type available in online databases.

Thus, in order to obtain an estimate for **C**, we could perform Hi-C on the same tissue of interest that would give an overall indication of the chromatin conformation of the corresponding cell type. We could also consider estimating a general background chromatin conformation across all cell types, using a database of DNA-DNA interactions, coming from a variety of tissue types and sequencing protocols, including Hi-C among others like 3C, 4C, 5C, Capture-C, ChIA-PET and IM-PET (Teng et al., 2015). In both cases, this data comes from a bulk tissue sample, which would include multiple cell types.

We aim to refine our estimate for chromatin conformation by deconvoluting the bulk Chromatin Conformation data into individual cell types for a more accurate estimate of **C** for each cell type. To do this, we use a similar approach to DC3 (Zeng et al., 2019), originally designed to deconvolute bulk Chromatin Conformation data, given unmatched single cell gene expression and genome accessibility data for the same tissue (Figure 3.2). Our approach is slightly different in that it uses matched single cell gene expression and genome accessibility data, and the chromatin capture data comes from a different sample. This is described in the next section.

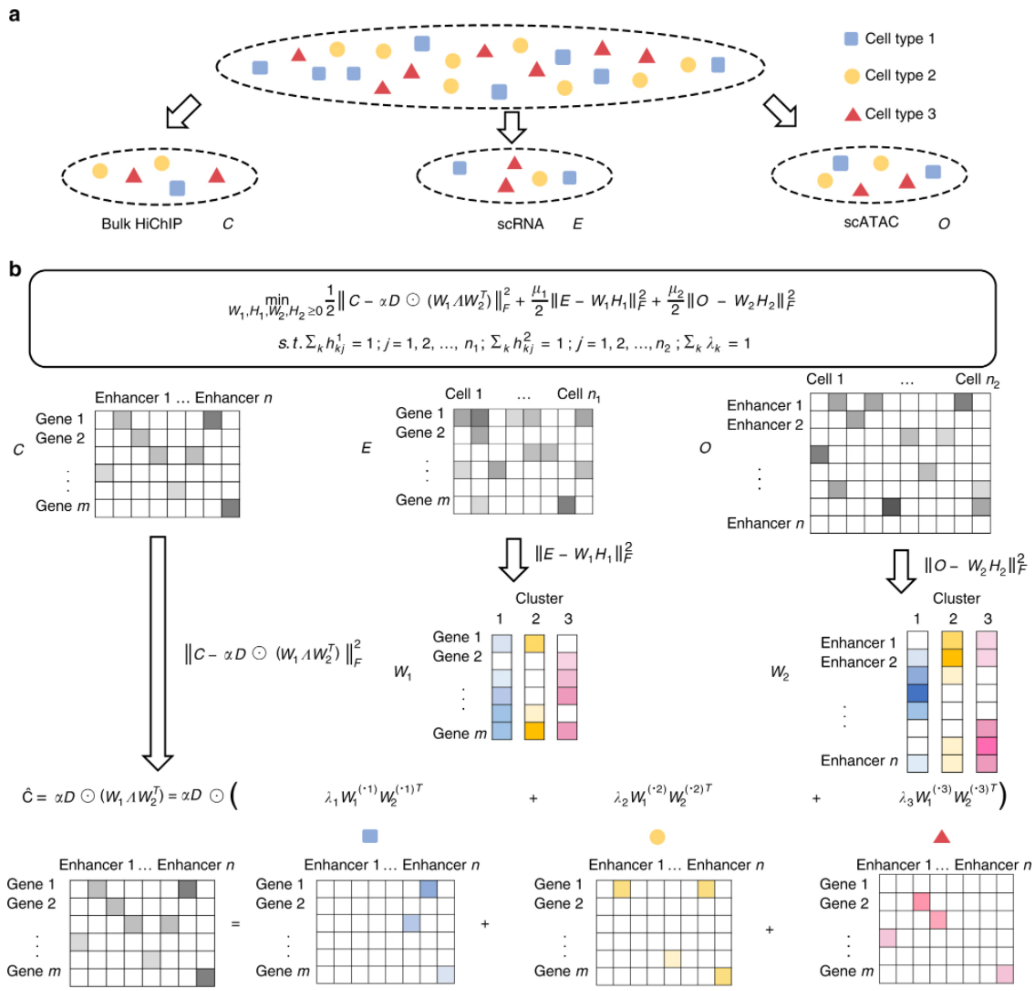


Figure 3.2: Schematic of DC3 algorithm. Bulk Hi-C data is deconvoluted using scRNA-seq and scATAC-seq data.

Source: Zeng, W., Chen, X., Duren, Z. et al. DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nat Commun 10, 4613 (2019). https://doi.org/10.1038/s41467-019-12547-1

3.2 Deconvoluting the chromatin conformation

Here, we attempted to perform a modified version of DC3, with the goal of deconvoluting bulk Chromatin Conformation data given matched scRNA-seq and scATAC-seq data of the same tissue with cell type labels. The data that we are given is:

- $\mathbf{E}, \mathbf{A}, \mathbf{C}$, as defined in the previous section.
- $\mathbf{H}$ is a $k \times n$ binary matrix representing the cell type assignments. Here, there are k cell types, and each column i contains one 1 indicating the cell type of cell i and the remaining entries are 0. This can be generalised to a non-binary matrix which

represents transient cell types which may have properties of multiple cell types.

- $\mathbf{\Lambda}$ is a $k \times k$ diagonal matrix whose ith entry is proportional to the size of cluster i and whose terms sum to 1. This term will be used to rescale values to match the size of each cell type cluster.

To deconvolute the bulk data into individual cell types, we just need to estimate the average gene expression and enhancer accessibility for each cell type:

- $\mathbf{W}_1$ is a $p \times k$ matrix representing the mean gene expressions for each cell type cluster.
- $\mathbf{W}_2 =$ a $q \times k$ matrix representing the mean enhancer accessibility for each cell type cluster.

In this way, we can obtain approximate expressions for each of $\mathbf{E}, \mathbf{A}$ and $\mathbf{C}$ in terms of $\mathbf{W}_1$ and $\mathbf{W}_2$:

- $\mathbf{E} \approx \mathbf{W}_1\mathbf{H}$;
- $\mathbf{A} \approx \mathbf{W}_2\mathbf{H}$; and
- $\mathbf{C} \approx \alpha\mathbf{W}_1\mathbf{\Lambda}\mathbf{W}_2^T$ where α is a scaling constant.

Writing $\mathbf{C}$ in this way splits each value into a contribution from each cell type cluster, effectively deconvoluting the bulk data. We achieve this by solving the following optimisation problem:

$$\min_{\mathbf{W}_1,\mathbf{W}_2>0} \frac{\mu_1}{2}\|\mathbf{E} - \mathbf{W}_1\mathbf{H}\|_F^2 + \frac{\mu_2}{2}\|\mathbf{A} - \mathbf{W}_2\mathbf{H}\|_F^2 + \frac{\mu_3}{2}\|\mathbf{C} - \alpha\mathbf{W}_1\mathbf{\Lambda}\mathbf{W}_2^T|_F^2$$

where μ_1, μ_2, μ_3 are weights and $\|\mathbf{X}\|_F$ refers to the Frobenius norm of the matrix $\mathbf{X}$.

For simplicity, we assume that each entry of $\mathbf{E}, \mathbf{A}$ and $\mathbf{C}$ is normally distributed with means as the corresponding entry in $\mathbf{W}_1\mathbf{H}, \mathbf{W}_2\mathbf{H}$ and $\alpha\mathbf{W}_1\mathbf{\Lambda}\mathbf{W}_2^T$ and with variances of σ_1^2, σ_2^2 and σ_3^2 respectively. We can choose the weights in the loss function to satisfy $\frac{\mu_i}{2} = \frac{1}{2\sigma_i^2}$ for $i = 1, 2, 3$ which balances the contribution of each component to the loss function and reduces the number of parameters to be estimated.

Note that most observed omics data is often very skewed and sparse (de Torrenté et al., 2020), meaning it may not be appropriate to assume that the gene expressions, enhancer

accessibility and chromatin conformation are normally distributed. However, we proceed with these assumptions for mathematical simplicity.

We can then take an initial estimate for $\mathbf{W}_1$ and $\mathbf{W}_2$ as the corresponding means of each cell type in the observed data, and iteratively estimate $(\mathbf{W}_1, \mathbf{W}_2)$ and $(\alpha, \sigma_1^2, \sigma_2^2, \sigma_3^2)$ using maximum likelihood estimation.

We attempted to perform this using matched scRNA-seq and scATAC-seq taken from the adult mouse brain from the SNARE-seq protocol (Chen et al., 2019) and Hi-C data of the developing mouse brain (Bonev et al., 2017). However, we found that our deconvolution was not meaningful, as it often failed to incorporate the information in $\mathbf{C}$. This could be because our model assumes that all values should be normally distributed. But it could also be that the Hi-C data was not taken from the same tissue as the scRNA-seq and scATAC-seq, as we could not find an experiment performing Hi-C on the adult mouse brain.

Perhaps this method could be more useful for matched scRNA-seq and scATAC-seq data with Hi-C data from the same tissue. However, due to the recentness of these technologies, such data is not available yet.

In what follows, we take the bulk chromatin conformation to be our estimate for $\mathbf{C}$. Although we could not deconvolute this into individual cell types, it would still provide a crude estimation of the chromatin interactions that can occur in a cell.

3.3 Regulation potential

Before we can create our model for perturbing transcription factors, we need to create some measure for the potential for a TF to regulate a gene, in a similar way to Lisa and ANANSE. We assume that for a TF to regulate a gene via an enhancer, it is required that:

- The enhancer is enriched with the **TF motif**, allowing the TF to bind to it.
- The enhancer is **accessible**, allowing the TF to bind to it.
- There can be a **chromatin interaction** between the enhancer and promoter of the gene, allowing the formation of a DNA loop.

Each of these components can be quantified using our data, as explained in Section 3.1. Then we can then consider the product of three scores to be the regulation potential of a TF to a gene via an individual enhancer.

To find the total potential for a TF to regulate a gene via all enhancers, we can sum up the regulation potential for all enhancers. We can represent this for a particular cell i for each TF-gene pair by

$$\mathbf{R}_i = \mathbf{C}\mathbf{A}_i\mathbf{T}$$

where $\mathbf{A}_i$ is a $q \times q$ matrix with the enhancer accessibility scores for the ith cell along the diagonal, and zeroes elsewhere. This gives us a $p \times t$ matrix which represents the potential for all t transcription factors to regulate all p genes via all enhancers. Calculating this for all n cells, we end up with a $p \times t \times n$ array which we call $\mathbf{R}$ containing the regulation potential of all transcription factors to each gene in each cell.

To verify that our proposed regulation potential is capturing true regulatory relationships, we calculate the regulation potential using the SHARE-seq data (details in Section 4.2.1) and compare it to known TF-gene regulations. There are a variety of databases that record known and predicted regulations, such as TRRUST (Han et al., 2018), hTFtarget (Zhang et al., 2020), TFBSDB (Plaisier et al., 2016), RegNetwork (Liu et al., 2015) and MSigDB (Subramanian et al., 2005; Kolmykov et al., 2021). As our regulation potential is at a single cell resolution, we took the average over all cells to obtain a $gene \times TF$ matrix to compare it to these databases. If the regulation potential is accurate, we expect that the TF-gene regulations from the databases should have a greater regulation potential than a random subset of TF-gene regulations.

By resampling 1 million random subsets of the same size as each database, we compute an empirical p-value as the probability that the mean regulation potential from a random sample is greater than the mean regulation potential of the database interactions. From Table 3.1, we see that all databases are significantly enriched with interactions containing a high regulation potential, implying that our regulation potential captures true regulatory TF-gene relationships.

Table 3.1: **Results from testing regulation potential for TF-gene interactions in databases.**

Database	Number of Regulations	Empirical p-value
TRRUST	43	0.03
hTFtarget	10294	$< 1 \times 10^{-6}$
TFBSDB	5253	$< 1 \times 10^{-6}$
RegNetwork	729	3.9×10^{-5}
MSigDB	695	$< 1 \times 10^{-6}$
Combined	14112	$< 1 \times 10^{-6}$

Here, the number of regulations is those that remained after filtering. The Combined database takes the union of interactions from the 5 other databases.

3.4 Predicting the effect of transcription factor perturbations

Now that we can measure the potential of each transcription factor to regulate each gene, we can combine this with the gene expression information to estimate the effect of perturbing transcription factors.

To state this in mathematical terms, for a gene j, we wish to predict

- $\mathbf{y}_j$ is a vector of length n, representing the expression of gene j in all n cells.

given

- $\mathbf{X}$ is a $n \times t$ matrix, representing the gene expression of the t transcription factors in all n cells.
- $\mathbf{R}_j$ is a $n \times t$ matrix, representing the regulation potential of all transcription factors onto gene j in all n cells. This can be interpreted as a slice of the array R that corresponds to the gene j.

The motivation for our model is that a true regulatory relationship between a TF and a gene should have a high regulation potential between the TF and gene, and they should

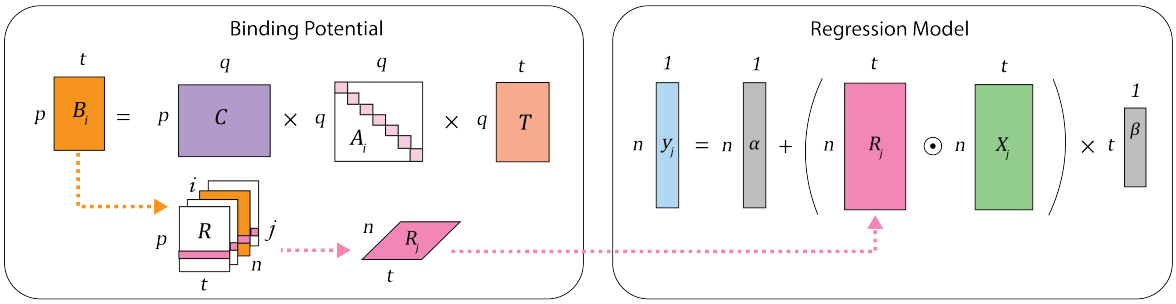


Figure 3.3: Illustration of the inputs to the regression model.

both be coexpressed.

For now, we only consider linear models, like CellOracle, for interpretability and the ability to form predictions. We have proposed a few candidates for a model for transcription factor perturbations.

Model 1

The first model we consider is for benchmarking purposes (Kamimoto et al., 2020). This is a simple model that only uses the gene expression values, and not regulation potential, similar to the model used in CellOracle. For Model 1 we use the regression equation

$$\mathbf{y}_j = \alpha_j \mathbf{1} + \mathbf{X}\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{3.1}$$

where $\mathbf{y}_j$ is the gene expression to be predicted, $\mathbf{X}$ is the TF expression, α_j and $\boldsymbol{\beta}_j$ are the fitted intercept and coefficients respectively, and $\boldsymbol{\varepsilon}_j$ is residual noise. However, a disadvantage of this model is that many genes are correlated in their expression, making it difficult to determine the structure of the regulation network and causing fitted coefficients to have large standard errors.

Model 2

This model, illustrated in Figure 3.3, extends Model 1 to incorporate the regulation potential on top of the gene expression. The sparsity of $\mathbf{X}$ and $\mathbf{R}_j$ should in theory only allow true regulatory TFs to have non-zero coefficients. This is because the TF needs to be coexpressed with its target gene but also needs the potential for gene regulation via

enhancers. For Model 2 we use the regression equation

$$\mathbf{y}_j = \alpha_j \mathbf{1} + (\mathbf{X} \odot \mathbf{R}_j)\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{3.2}$$

where $\mathbf{y}_j$ is the gene expression to be predicted, $\mathbf{X}$ is the TF expression, $\mathbf{R}_j$ is the regulation potential, α_j and $\boldsymbol{\beta}_j$ are the fitted intercept and coefficients respectively, and $\boldsymbol{\varepsilon}_j$ is residual noise. As this model captures the biological processes that govern gene regulation, we choose to use this model for our simulation study, explored in Section 4.1. However, when applied to real data, we found that $\mathbf{X}$ and $\mathbf{R}_j$ were too sparse, and the values in $\mathbf{R}_j$ were often very skewed, as it is a product of 3 sparse matrices. This led the fitted models to use very limited information and ultimately were not insightful.

Model 3

To address the issue of $\mathbf{R}_j$ having skewed values, we considered normalising the values before using it in the model. This can be achieved by raising each term in $\mathbf{R}_j$ to the power of λ, where a value for λ between 0 and 1 would bring all the regulation potential values closer to 1. For Model 3 we use the regression equation

$$\mathbf{y} = \alpha_j \mathbf{1} + [\mathbf{X} \odot (\mathbf{R}_j)^{\lambda}]\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{3.3}$$

where $\mathbf{y}_j$ is the gene expression to be predicted, $\mathbf{X}$ is the TF expression, $\mathbf{R}_j$ is the regulation potential, λ is a weight, α_j and $\boldsymbol{\beta}_j$ are the fitted intercept and coefficients respectively, and $\boldsymbol{\varepsilon}_j$ is residual noise. However, we found that the model still did not have enough information from the data. This could be because this normalisation does not change 0 values, leaving $\mathbf{R}_j$ being too sparse, which is compounded by taking the Hadamard product with $\mathbf{X}$, which is generally sparse as well, as it comes from scRNA-seq.

Model 4

To account for such sparsity in $\mathbf{R}_j$, we consider adding a small constant to all values in $\mathbf{R}_j$. This has the effect of when $\mathbf{R}_j$ values are all not available (i.e., 0) for a particular TF, the regression will rely on the available gene expression values only. However, if $\mathbf{R}_j$ values are available (i.e., some not 0) for a TF, then they will be incorporated into the regression.

For Model 4 we use the regression equation

$$\mathbf{y}_j = \alpha_j \mathbf{1} + [\mathbf{X} \odot (w\mathbf{1}_{n\times t} + (1-w)\mathbf{R}_j)]\,\boldsymbol{\beta}_j + \boldsymbol{\varepsilon}_j, \tag{3.4}$$

where $\mathbf{y}_j$ is the gene expression to be predicted, $\mathbf{X}$ is the TF expression, $\mathbf{R}_j$ is the regulation potential, α_j and $\boldsymbol{\beta}_j$ are the fitted intercept and coefficients respectively, and $\boldsymbol{\varepsilon}_j$ is residual noise. Here, w is a parameter which can be used to weight the influence of $\mathbf{R}_j$ when available, choosing a default value of $w = 0.1$. We found that this model was the most effective when applied to real data, as it was robust to the sparsity.

3.4.1 Model comparison

To test the usefulness of each model, we compare their ability to recapitulate the effect of overexpressing key TFs in the simulation framework under Section 4.1. Here, we start at a source cell type (A) and predict the effect of overexpressing the key TFs for two different cell types (B and C). In our simulation, these key TFs are the drivers of cell identity, so an effective model should direct the source cells towards the target cell type. We test each model under 4 scenarios, with or without dropouts in $\mathbf{E}$ and with or without dropouts in $\mathbf{R}$.

Scenario 1: no dropouts in E, no sparsity in R

In this case, all models have complete information about the data so we would expect them to perform well. We can see that indeed in Figure 3.4, all models generally perturb the source cells in the correct direction.

It should be noted however that the perturbations do not point directly to the target cell type. This would most likely be due to the source cell retaining aspects of its cell identity in the short term. In Chapter 4, we will be modelling the long term result of perturbing a TF, where the identity of the source cell is eventually suppressed.

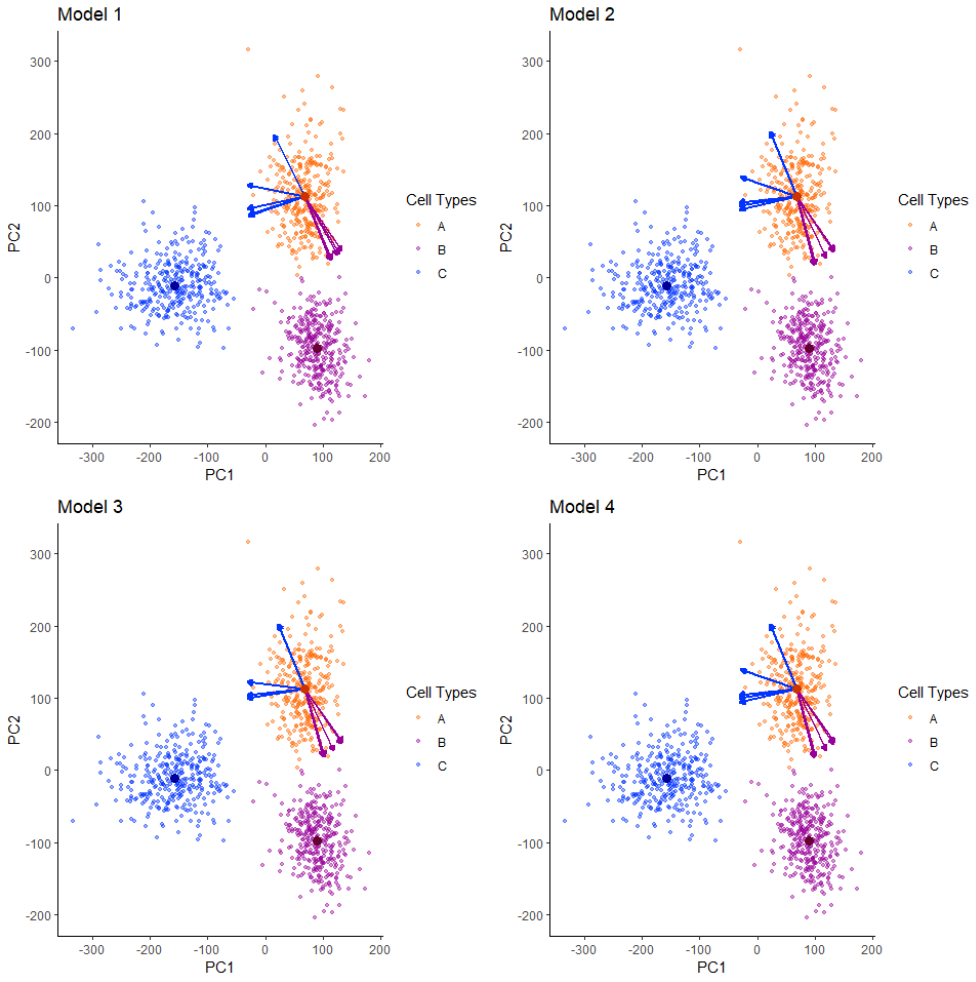


Figure 3.4: Perturbation results for Scenario 1: no dropouts in **E**, no sparsity in **R**. Coloured arrows indicate the predicted change after a perturbation of a key TF (coloured accordingly to the cell type). Centroids of each cell type cluster are drawn in, and arrows are drawn from centroid of cell type A.

Scenario 2: many dropouts in E, no sparsity in R

Here we have replaced up to 60% of the observed gene expression with 0, closer resembling the data from scRNA-seq experiments. Here we can see in Figure 3.5 that models 2 and 3 are barely affected with stable predictions, since they also incorporate the information from the regulation potential. However, models 1 and 4 become much less accurate as their inputs essentially become much noisier.

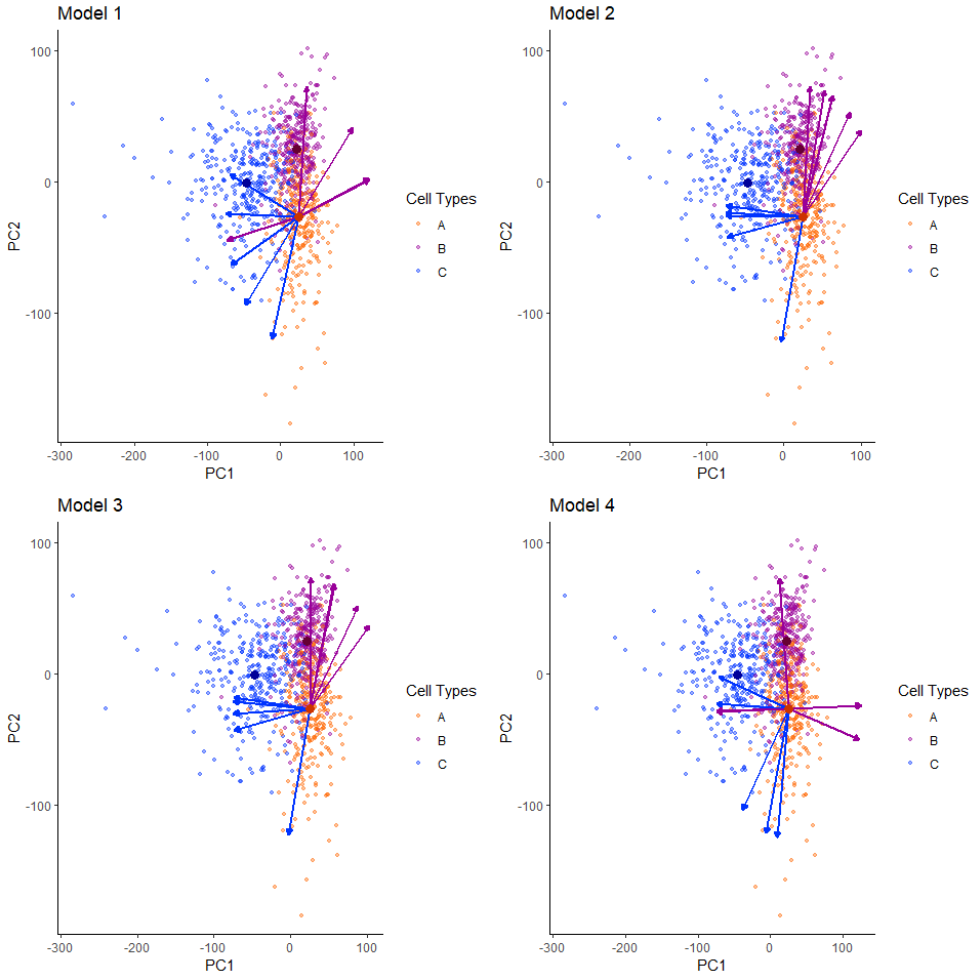


Figure 3.5: Perturbation results for Scenario 2: many dropouts in **E**, no sparsity in **R**. Coloured arrows indicate the predicted change after a perturbation of a key TF (coloured accordingly to the cell type). Centroids of each cell type cluster are drawn in, and arrows are drawn from centroid of cell type A.

Scenario 3: no dropouts in E, large sparsity in R

Here we have forced up to 99.5% of the values in **R** to be 0, closer resembling what it would look like in reality. Here we can see in Figure 3.6 that models 1 and 4 are barely affected with stable predictions, since they are able to use the gene expression data. However, models 2 and 3 become much less accurate as they lose a lot of information from the regulation potential, a necessary input.

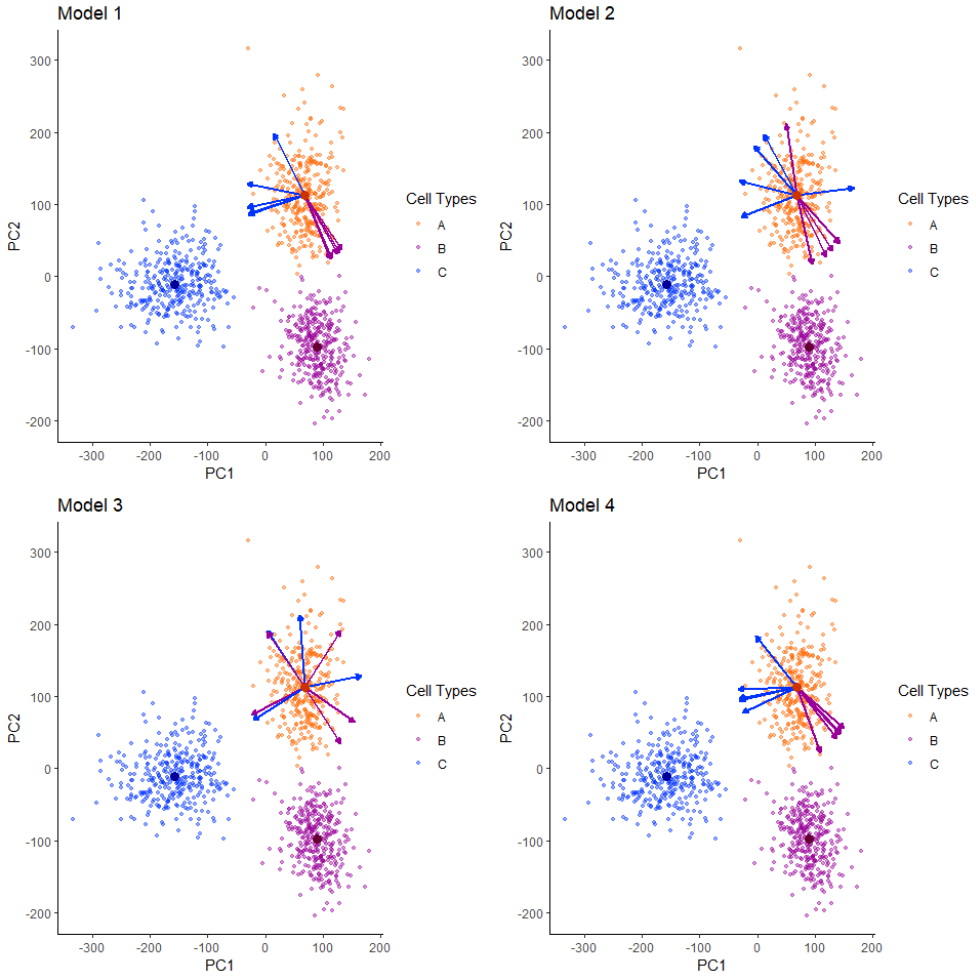


Figure 3.6: Perturbation results for Scenario 3: no dropouts in **E**, large sparsity in **R**. Coloured arrows indicate the predicted change after a perturbation of a key TF (coloured accordingly to the cell type). Centroids of each cell type cluster are drawn in, and arrows are drawn from centroid of cell type A.

Scenario 4: many dropouts in E, large sparsity in R

Now we have set 60% of the values in **E** to be 0 and forced up to 99.5% of the values in **R** to be 0, most closely resembling experimental data. Here we can see in Figure 3.7 that only model 4 is able to retain a reasonable prediction, most likely as it can take advantage of the little signal available in the data. However, all other models become much less accurate as they have lost a lot of information from one of their necessary inputs. This justifies the use of model 4 when we apply this to experimental data in section 4.2.

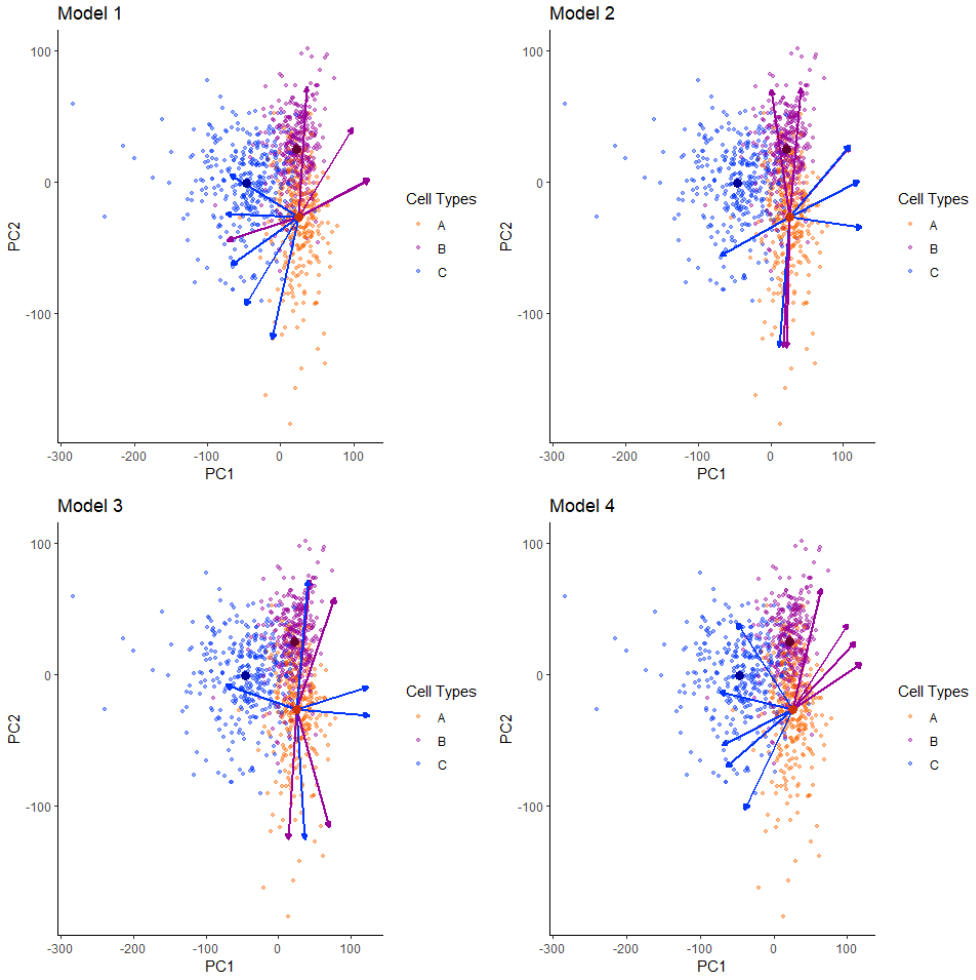


Figure 3.7: Perturbation results for Scenario 4: many dropouts in **E**, large sparsity in **R**. Coloured arrows indicate the predicted change after a perturbation of a key TF (coloured accordingly to the cell type). Centroids of each cell type cluster are drawn in, and arrows are drawn from centroid of cell type A.

3.5 Summary

Here, we have introduced scREMOTE, a framework for predicting the effect of transcription factor perturbations for cell reprogramming. We calculated a regulation potential that captures the regulatory relationships between TFs and target genes, which we demonstrated was able to recapitulate known regulations. We proposed several models for cell reprogramming, and found a model that was robust to dropouts in gene expression and regulation potential.

Chapter 4

Impact of scREMOTE on single cell biology

This chapter's original contribution is to demonstrate the utility of our concept by two innovative applications of our scREMOTE. The first novel application will be in a data simulation, showing that our framework can recapitulate some key properties of cells during development and reprogramming. This will be the first simulation framework for matched scRNA-seq and scATAC-seq data in single cell research that allows for the estimation of transient cell states. The second application will be to simulate a cell reprogramming experiment using experimentally obtained matched scRNA-seq and scATAC-seq data. By recapitulating known results, we show for the first time that a computational model, scREMOTE, has the potential to model the regulatory mechanisms behind cell reprogramming and make long-term predictions at the cellular level.

4.1 Multi-modal simulation

Experimental data is often expensive and time consuming to collect, requiring extensive expertise, equipment and materials. This motivates the need for methods to simulate omics data to supplement experimental data for validating and benchmarking the continuously growing collection of bioinformatics tools.

Multiomics sequencing methods are becoming increasingly popular, which can be expected to be followed by a wave of computational tools to analyse and use this new type of data. While there are extensive benchmark studies (Cao et al., 2021), to my knowledge, there are currently no tools available to simulate matched scRNA-seq and scATAC-seq data. Our framework provides a natural connection between these two modalities opening up the possibility to simulate them simultaneously.

4.1.1 Simulation components

Our simulation framework will take the following parameters:

- p = the number of **genes**. Default value chosen to be $p = 500$.
- q = the number of **enhancers**. Default value chosen to be $q = 100$.
- k = the number of **cell types**. Default value chosen to be $k = 3$.
- n = the number of **cells**. Default value chosen to be $n = 150$.
- t = the number of **transcription factors** (a subset of the genes). Default value chosen to be $t = 20$.
- t' = the number of **key TFs** per cell type. These will be the transcription factors responsible for maintaining cell identity. Default value chosen to be $t' = 3$.
- q' = the number of **enhancer targets** for each TF. That is, the enhancers that each TF will be able to bind to. Default value chosen to be $q' = 2$.
- p' = the number of **gene targets** for each enhancer. That is, the genes that an enhancer can form a DNA loop with. Default value chosen to be $p' = 2$.

For now, we assume that t', q' and p' are constant across all cell types, TFs and enhancers respectively. In reality, this would not be the case as some TFs are considered as master regulators that may regulate many other genes (Oestreich and Weinmann, 2012), and some enhancers are consider super-enhancers that are targeted by many TFs (Whyte et al., 2013). A future simulation method could sample t', q' and p' from some statistical distribution, perhaps a negative binomial. We can now simulate the model components as explained in section 3.1.

4.1.2 Data initialisation

Step 1: generate T

Firstly, we select the enhancer targets for each TF by randomly selecting q' enhancers out of q for each TF. Next, we generate $\mathbf{T}$ by

$$\mathbf{T} = \begin{bmatrix} T_{11} & T_{12} & \cdots & T_{1t} \\ T_{21} & T_{22} & \cdots & T_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ T_{q1} & T_{q2} & \cdots & T_{qt} \end{bmatrix}$$

where $T_{ij} \sim N(\mu_T, \sigma_T^2)$ if enhancer i is a target of TF j, and $T_{ij} = 0$ otherwise. We choose default values of $\mu_T = 1$ and $\sigma_T = 0.3$

Note that this is quite a simplification because certain TFs would have greater affinity to target certain enhancers. In practice, this selection of target enhancers should guided by experimental data. However, we continue with this purely theoretical example to demonstrate the potential of our model.

Step 2: generate $\mathbf{C}^K$

Here, we will generate a $\mathbf{C}$ matrix representing the chromatin interactions for each cell type h which we denote as $\mathbf{C}^h$. The collection of all $\mathbf{C}^h$ will form $\mathbf{C}^K$. We start by creating a base regulatory structure (common to all cell types) by randomly selecting p' gene targets out of p genes for each enhancer. Then we create a cell type-specific regulatory structure (common to cells of the same cell type) by randomly selecting a further p' gene targets out of p genes for each enhancer. We select the key TFs for each cell type by selecting t' random TFs out of t. Then we generate $\mathbf{C}$ by

$$\mathbf{C}^h = \begin{bmatrix} C_{11}^h & C_{12}^h & \cdots & C_{1q}^h \\ C_{21}^h & C_{22}^h & \cdots & C_{2q}^h \\ \vdots & \vdots & \ddots & \vdots \\ C_{p1}^h & C_{p2}^h & \cdots & C_{pq}^h \end{bmatrix}$$

where

$$C_{ij}^{h} = \begin{cases} w(\mu_C + \mu_{key}^{h}) + \varepsilon_{ij} & \text{if gene } i \text{ is a target of enhancer } j \text{ in cell type } h \\ 0 & \text{otherwise} \end{cases}$$

- w is a weight that scales and determines the sign of the regulatory relationship. We choose $w = -0.3$ if j is a target enhancer of a key TF in cell type h, and i is a key TF of another cell type and $w = 1$ otherwise. This way, key TFs will downregulate the key TFs of other cell types. This simulates antagonistic TF-pairs that determine the cell fate (Okawa et al., 2016; Heinäniemi et al., 2013).

- μ_C is a base chromatin conformation score for which the enhancer can form a DNA loop with the gene's promoter. We choose $\mu_C = 1$.

- μ_{key}^{h} is an additional score that favours key TFs of the same cell type. We choose $\mu_{key}^{h} = 2$ if j is a target enhancer of a key TF in cell type h, and i is a key TF of the same cell type and $\mu_{key}^{h} = 0$ otherwise. This way, key TFs of a cell type will upregulate each other. This simulates feedback loops in GRNs which help maintain cell identity (Burda et al., 2011).

- $\varepsilon_{ij} \sim N(0, \sigma_C^2)$ adds random noise to the data. We choose $\sigma_C = 0.3$.

Again, this simulation is quite a simplification and ideally the selection of target genes and value of additional parameters should be guided by experimental data.

Step 3: generate A

To generate the enhancer accessibility matrix, we use

$$\mathbf{A} = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1q} \\ A_{21} & A_{22} & \cdots & A_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p1} & A_{p2} & \cdots & A_{pq} \end{bmatrix}$$

where

$$A_{ij} = A_{ij}^{base} + A_{ij}^{celltype} + A_{ij}^{noise}$$

and

- $A_{ij}^{base} \sim NB(\mu_{base}, \sigma_{base}^2)$ using the mean and variance to parameterise the negative binomial distribution. This term represents a base enhancer accessibility common to all cells. We choose $\mu_{base} = 0.2$ and $\sigma_{base}^2 = 0.24$.

- $A_{ij}^{celltype} \sim NB(\mu_{celltype}, \sigma_{base}^2)$. This represents a cell type-specific enhancer accessibility common to cells belonging to the same cell type, helping to create cell type structure. We choose $\mu_{celltype} = 0.2$ and $\sigma_{celltype}^2 = 0.24$.

- $A_{ij}^{noise} \sim NB(\mu_{noise}, \sigma_{noise}^2)$. This cell adds noise at the individual cell level, simulating the heterogeneity of the cell population. We choose $\mu_{noise} = 0.2$ and $\sigma_{noise}^2 = 0.24$.

Step 4: generate E

Here, we generate the gene expressions of all the cells in three parts.

Step 4.1: Firstly, we calculate an **initial TF expression**. The cells are randomly allocated into one of k cell types. Then we simulate an initial gene expression of TFs given by

$$\mathbf{E}'_{TF} = \begin{bmatrix} E'_{11} & E'_{12} & \cdots & E'_{1n} \\ E'_{21} & E'_{22} & \cdots & E'_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ E'_{t1} & E'_{t2} & \cdots & E'_{tn} \end{bmatrix}$$

where

$$E'_{ij} = \max\left(\mu_{base} + \mu_{key,ij} + \varepsilon_{ij}, 0\right)$$

and

- μ_{base} is a base TF expression, which we choose to be $\mu_{base} = 0.3$.

- $\mu_{key,ij}$ increases the expression for key TFs. We choose to be $\mu_{key} = 3$ if TF i is a key TF for the cell type of cell j, and $\mu_{key} = 0$ otherwise.

- $\varepsilon_{ij} \sim N(0, \sigma_{tf}^2)$ which adds random noise to the data. We choose $\sigma_{tf} = 0.1$.

However, this initial gene expression may not resemble a biologically possible cell state. To resolve this, we proceed to our next step.

Step 4.2: Secondly, we want our **TF expression to converge to a steady state**, representing a biologically possible state. To do this, we iteratively recalculate the TF expression in a similar form to Equation (3.1), using the previously simulated matrices. That is, we calculate the regulation potential for cell i using

$$\mathbf{R}_i = \mathbf{C}\mathbf{A}_i\mathbf{T}^T$$

where $\mathbf{A}_i$ is a $q \times q$ matrix with the enhancer accessibility scores for the ith cell along the diagonal, and zeroes elsewhere. And then $\mathbf{R}$ is a $p \times t \times n$ array of all $\mathbf{R}_i$'s. Then for TF j, the expression $\mathbf{y}_j$ is given by

$$\mathbf{y}_j = \mathbf{X} \odot \mathbf{R}_j + \boldsymbol{\varepsilon}_j \tag{4.1}$$

where $\mathbf{X}$ is the TF expression, $\mathbf{R}_j$ is the regulation potential corresponding to gene j and $\boldsymbol{\varepsilon}$ is residual noise. Note that Equation (4.1) does not need any coefficients, as the gene expression values can be assumed to be scaled in our simulation. Each value will then converge, giving us our $\mathbf{E}_{TF}$.

Step 4.3: Lastly, we can generate the **gene expression** in all cells. Now that we have the expression of the TFs that represent the biological steady states, we can simulate the expression of any gene j using Equation (4.1). And since the genes do not regulate each other, no reiteration is required.

We have now simulated the $\mathbf{T}$, $\mathbf{C}$, $\mathbf{A}$ and $\mathbf{E}$ matrices.

4.1.3 Estimating transient cell states

One useful application of our simulation method is that it allows for the estimation of transient cell states. That is, we may often have data about distinct cell types but we may be interested in other cell states that have properties of multiple different cell types. Or we may be interested in the differentiation trajectory from a progenitor cell type to a more specialised cell type.

Naively, we could interpolate the gene expression and enhancer accessibility between two cell types to obtain an estimate for an intermediary cell state. However, this state

may not be biologically stable, and so is not an accurate representation of a cell state. To resolve this, we can iteratively apply Equation (4.1) to the interpolated state, and the cell state will eventually converge, representing a biologically stable steady state.

To apply this idea to our data simulation, we first interpolate between our three cell types. This is done by simulating new cells that have gene expression and enhancer accessibility of varying proportions from the three cell types (Figure 4.1) with some added noise.

These cells are then left to converge to a steady state based on their GRN, modelled by Equation (4.1). In this case, we can see that the interpolated cells converge to two distinct paths between the cell types. This could represent possible differentiation trajectories across these cell types.

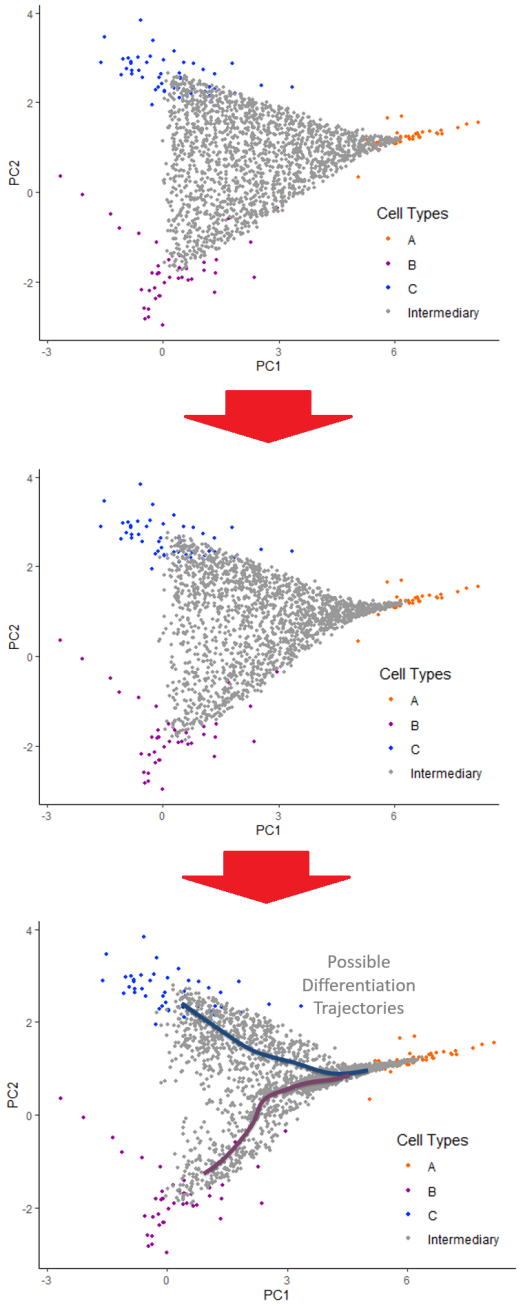


Figure 4.1: The simulation framework reveals possible differentiation trajectories between cell types.

4.1.4 Simulating cell reprogramming

Our simulation approach models the cell state based off the TF expression, we can also use it for simulating transcription factor perturbations. We can simply overexpress a TF, like in a real cell reprogramming experiment, and iteratively apply Equation (4.1) until the

perturbed cells converge to a new cell state.

Applying this to our data simulation, we can overexpress the key TFs for our target cell type in our source cell type and we can see that there is an initial shift that pushes the cells towards the target state (Figure 4.2). Then using the simulation, we allow the cells to converge to a steady state and we can see that only a small fraction of cells successfully reprogram. This is in line with observations in reprogramming experiments where the reprogramming efficiency remains very low (Takahashi and Yamanaka, 2016).

Interestingly, our simulation reveals that the differentiation trajectory after a perturbation is decided early on, where individual cells that are closer to the target state are more likely to successfully reprogram as opposed to cells that are further away. This supports a recent theory that suggests that the initial state of the cell plays a critical role in determining the outcome of cellular differentiation or reprogramming (Weinreb et al., 2020; Kong et al., 2020). This can be contrasted to previous models which often assumed that the heterogeneity of the outcome was purely due to the stochastic nature of these processes (Buganim et al., 2012; Floettmann et al., 2012).

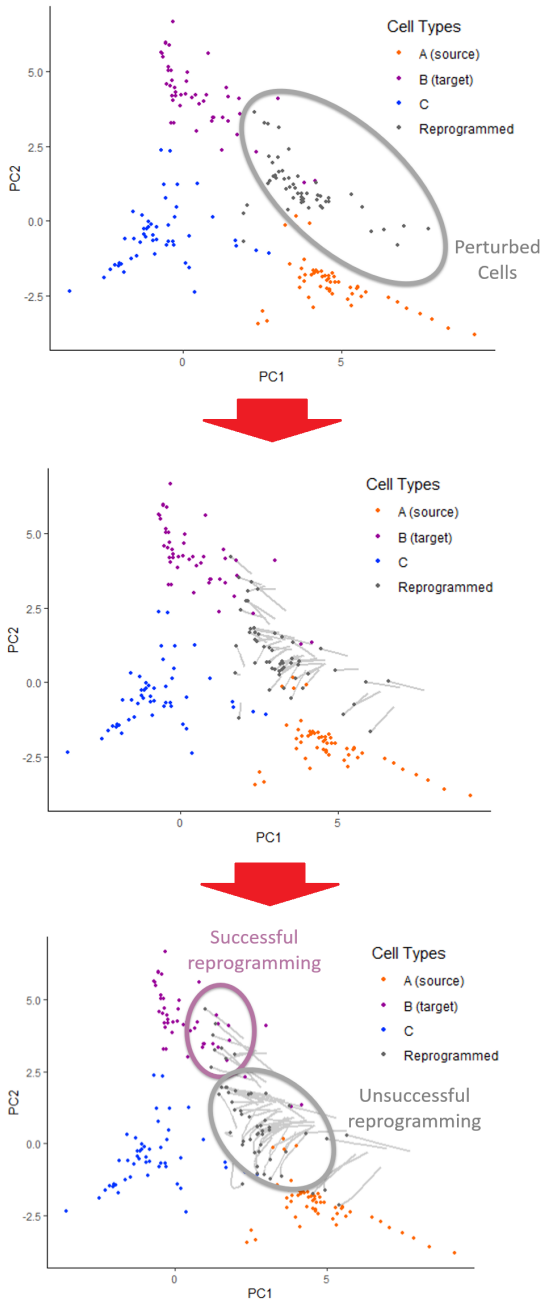


Figure 4.2: The simulation framework reveals different reprogramming trajectories dependent on the initial cell state.

4.2 Application to matched single cell data

To demonstrate the utility of our model on experimental data, we use SHARE-seq data collected by Ma et al. (2020) on adult mouse skin cells. This data measures matched scRNA-seq and scATAC-seq at a single cell resolution, representing the gene expression and enhancer accessibility respectively. Here, Transit Amplifying Cells (TACs) differentiate

into either an Inner Root Sheath (IRS) lineage or a Hair Shaft lineage and we will study the effect of TF perturbations on this system. *GATA3* has been long associated with the IRS fate during lineage determination in skin (Kaufman et al., 2003; Kurek et al., 2007). But more recently, it has been found that *GATA3* is a key reprogramming TF with *ATOH1* whose upregulation is able to reprogram supporting cells into hair cells (Walters et al., 2017). Similarly, *Runx1* has been known to play a key role in the differentiation of the hair follicle (Osorio et al., 2008; Hoi et al., 2010) and has been shown to be a key reprogramming factor for the Hair Shaft cell fate (Raveh et al., 2006).

4.2.1 Methods

We downloaded the SHARE-seq data from GEO (Accession number: GSE140203) and obtained the cell type labels directly from the authors (Ma et al., 2020). Chromatin conformation data was downloaded from the 4D Genome Database (Teng et al., 2015) using the full mouse dataset. TF motifs were downloaded from the JASPAR database (Fornes et al., 2019) using the full vertebrates position frequency matrices.

In the chromatin conformation data, all coordinates were realigned from the mm9 genome to the mm10 genome using the LiftOver tool provided by the Human Genome Browser at UCSC (Kent et al., 2002). This list of chromatin interactions is filtered down to those which include gene promoters, determined as any interactions within 500bp of the TSS of a gene. Gene coordinates were downloaded from the Mouse Genome Informatics (MGI) website (Eppig, 2017).

All chromatin regions which had an interaction with a promoter were taken to be an enhancer. These regions were sorted into bins of length 1000bp which is now taken as our enhancer list. We then construct $\mathbf{C}$ as a binary matrix which indicates a recorded connection between an enhancer and gene. Usually the chromatin conformation would be measured as a score representing the strength of the connection, however we are using a database containing many different experiments which would not be comparable.

The ATAC-seq data from the SHARE-seq protocol is then realigned to match our new enhancer list in 1000bp bins. As the bin cutoffs did not match exactly, any observed ATAC-seq measurement that overlapped with our 1000bp enhancers was considered as a count.

Applying this to all our enhancers gives us our **A** matrix.

And now that we have the coordinates of our enhancers, we obtained the genomic sequence using the BSgenome.Mmusculus.UCSC.mm10 package on Bioconductor. TF Motif enrichment was performed on each sequence using the AME function in the MEME Suite collection (McLeay and Bailey, 2010) with default settings. Only TFs that were highly enriched (marked as true positives) were kept, and their Position Weight Matrix score was normalised by dividing by the maximum value, so all values are between 0 and 1. This is then used as the corresponding value in **T**.

The gene expression matrix was taken directly from the SHARE-seq protocol but due to the sparsity of the data, we only used the 1000 most highly expressed genes which were then $\log(x+1)$ transformed.

4.2.2 Results: *Gata3* overexpression

We then applied Equation (3.4) to see the effect of a constant upregulation of *Gata3*. This was simulated by increasing the log expression of *Gata3* by 1 at each iteration of Equation (3.4) for 25 time steps, after which *Gata3* is no longer added to the system and the cell's natural GRN dynamics are allowed to take over.

The perturbed cells tend towards the IRS cell fate (Figure 4.3) and remain there while *Gata3* is constantly introduced into the system. However, when *Gata3* is no longer ectopically added to the system, the cells quickly fall back to their original cell state (Figure 4.4).

We can further interrogate the identity of these reprogrammed by looking at the effect on marker genes. We see that as *Gata3* is overexpressed (Figure 4.5), average *Mical3* levels increase accordingly, a marker gene for the IRS cell fate (Ma et al., 2020). On the other hand, average *Lef1* levels decrease accordingly, a marker gene for the Hair Shaft cell fate (Ma et al., 2020). And then when *Gata3* is no longer overexpressed at time point 25, we see the expression return to the original values. This suggests that the when *Gata3* is added to the system, the TAC cells adopt an IRS cell identity.

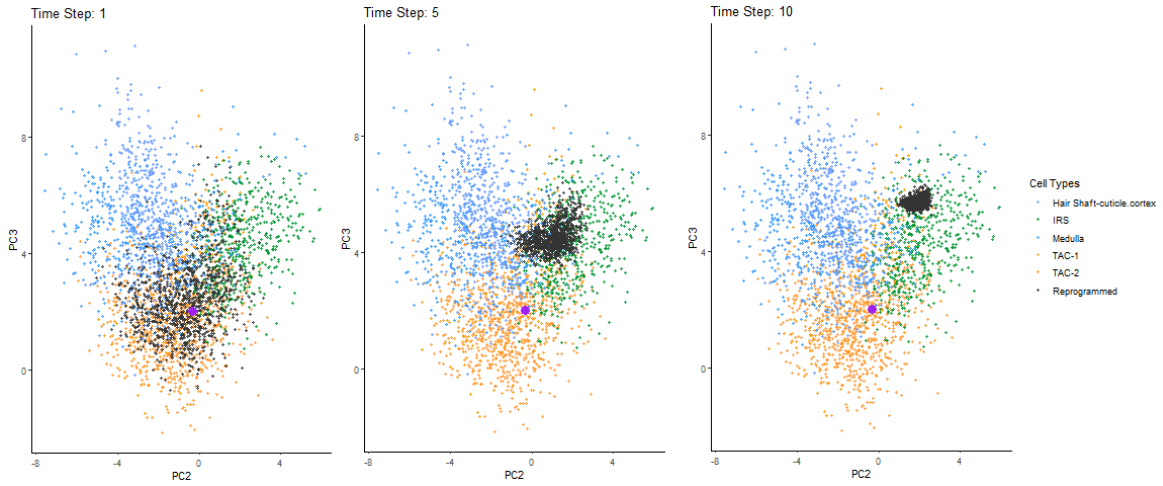


Figure 4.3: Increasing the *Gata3* expression converts the TACs (orange) towards the IRS cell fate (green).

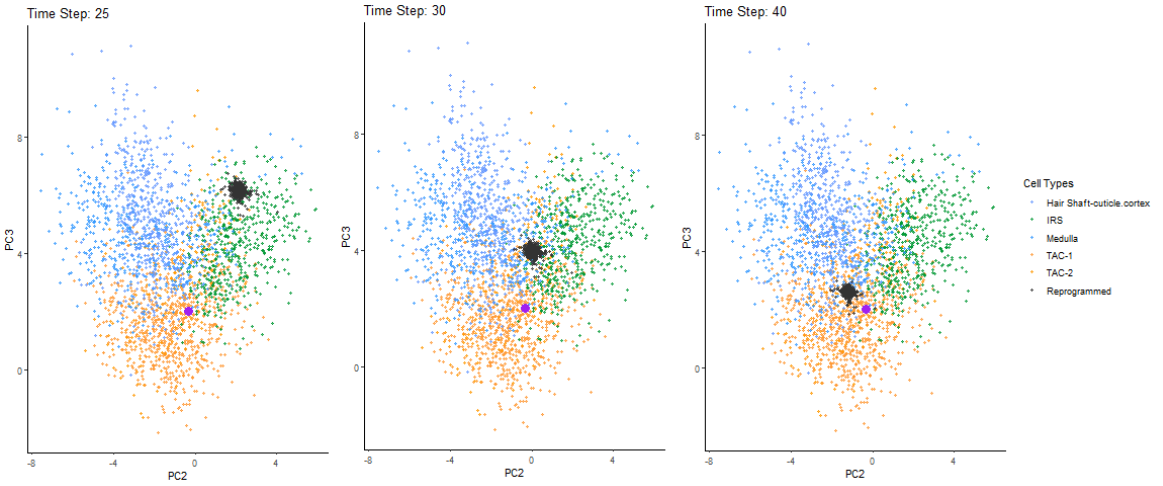


Figure 4.4: Removing the *Gata3* overexpression causes the reprogrammed cells (grey) to lose their identity returning to the starting state (orange).

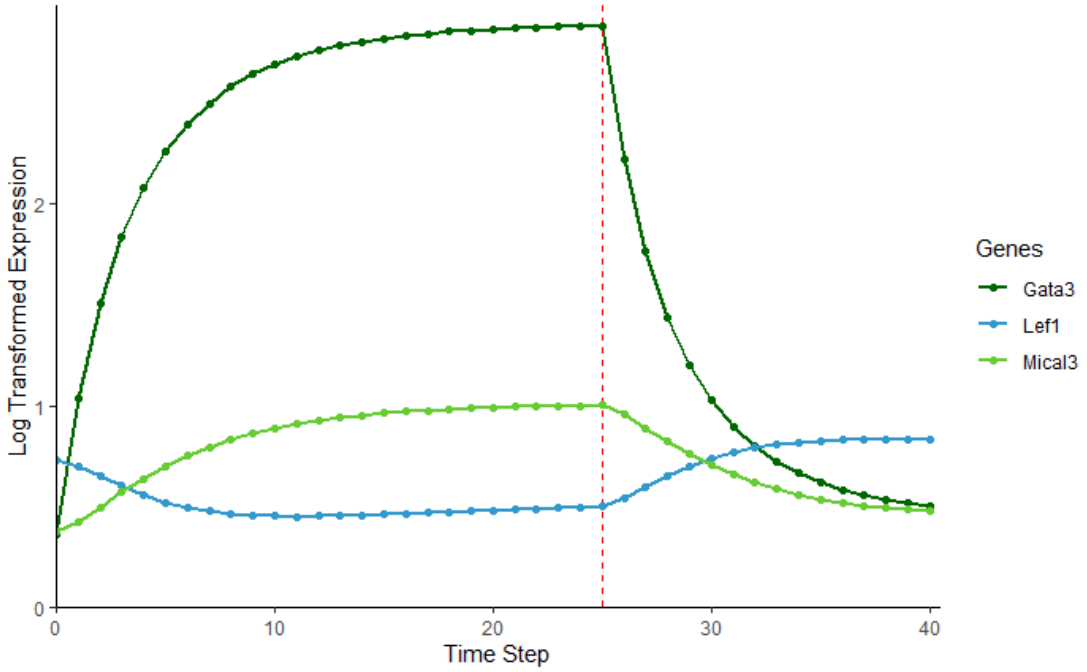


Figure 4.5: Perturbing *Gata3* causes changes in markers for IRS (green) and Hair Shaft (blue) fates. The red dotted line indicates when *Gata3* is no longer added to the system.

4.2.3 Results: *Runx1* overexpression

Similarly, we model the overexpression of *Runx1* by increasing the log expression of *Runx1* by 1 at each iteration of Equation (3.4) for 25 time steps, after which *Runx1* is no longer added to the system and the cell's natural GRN dynamics are allowed to take over.

The perturbed cells tend towards the Hair Shaft cell fate (Figure 4.6) and remain there while *Runx1* is constantly introduced into the system. However, when *Runx1* is no longer ectopically added to the system, the cells quickly fall back to their original cell state (Figure 4.7).

Checking the marker genes, we see that as *Runx1* is overexpressed (Figure 4.8), average *Lef1* levels increase accordingly whereas average *Mical3* levels barely change. And then when *Runx1* is no longer overexpressed at time point 25, we see the expression return to the original values.

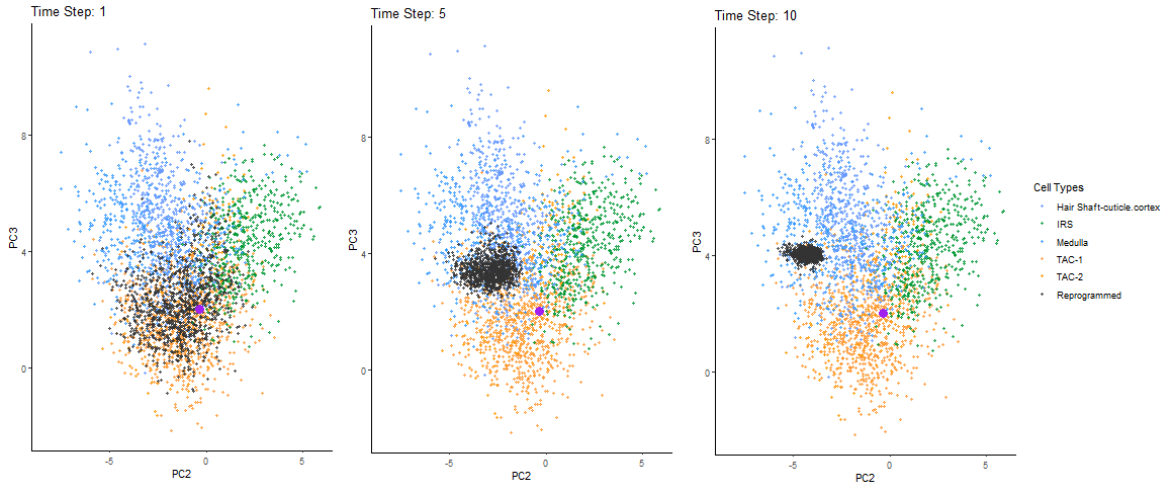


Figure 4.6: Increasing the *Runx1* expression converts the TAC cells (orange) towards the Hair Shaft cell fate (blue).

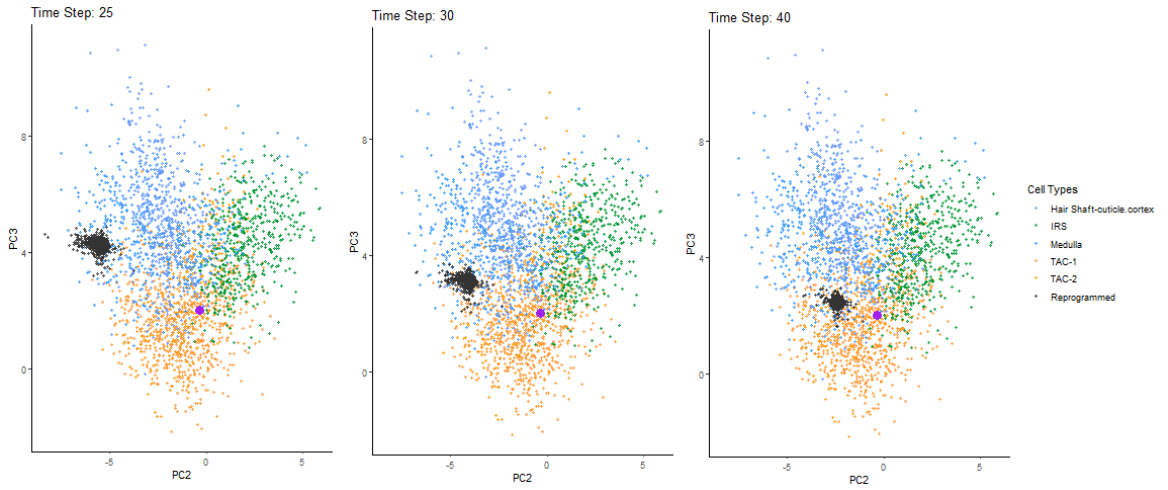


Figure 4.7: Removing the *Runx1* overexpression causes the reprogrammed cells (grey) to lose their identity returning to the starting state (orange).

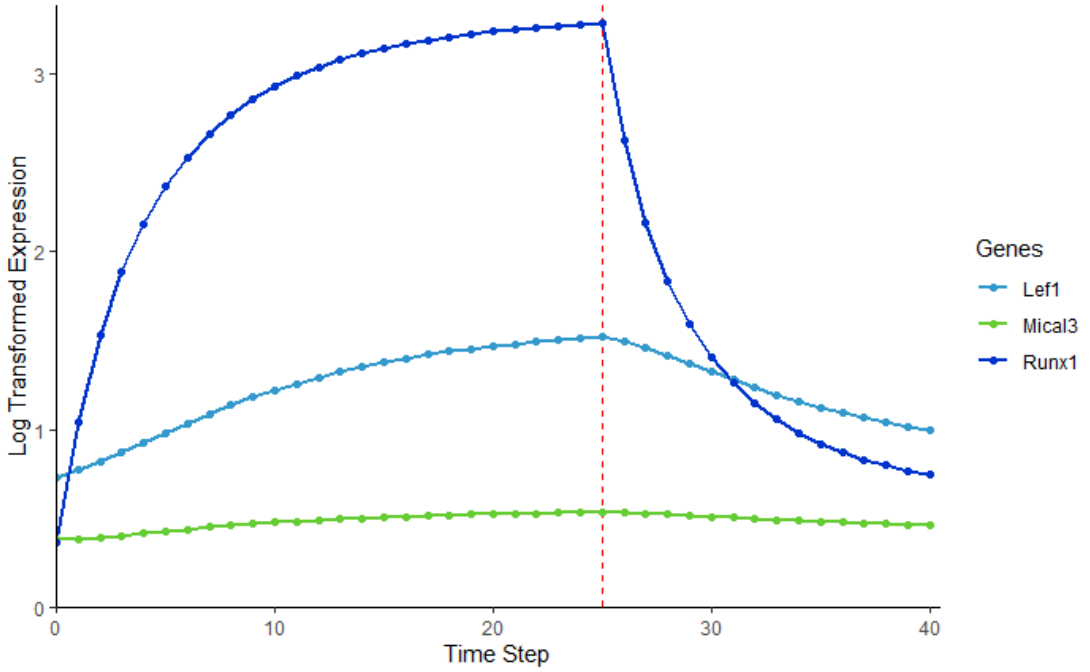


Figure 4.8: Perturbing *Runx1* causes changes in markers for IRS (green) and Hair Shaft (blue) fates.

However, we can see that in both cases, the perturbed cells generally converge to a single steady state, which is not representative of reality which should have a heterogeneous population and possibly multiple terminal fates. This could be a limitation of our modelling framework or the data.

4.3 Summary

Here, we have shown that scREMOTE can successfully integrate data from multiple modalities, creating a holistic model for the regulatory mechanisms occurring in individual cells. The ability to recapitulate known experimental results demonstrates the potential for scREMOTE to discover novel transcription factor combinations to drive cell reprogramming, accelerating research in regenerative medicine.

Chapter 5

Conclusion

5.1 Discussion

Understanding the regulatory dynamics in a cell is a complex yet important challenge, especially in the context of cell reprogramming, which results in large changes to the cell's identity. In this thesis, we synthesise the current mathematical literature in this area and address the challenges by developing scREMOTE, a model for long-term predictions of TF perturbations at the single cell level that extends on existing algorithms (Kamimoto et al., 2020).

We acknowledge several important limitations in our model. As with all previous models for cell reprogramming, the most significant limitation is due to the data that we have available. scRNA-seq data is generally very sparse, with many dropouts due to genes being lowly expressed and the stochasticity of gene expression (Qiu, 2020). This means that many important genes (especially TFs) which are lowly expressed will have an overwhelming proportion of missing data, leading to large bias in the linear regression coefficients. In our case, we only considered genes (and TFs) that were rather highly expressed, which meant that we were very limited in the TFs that we could test for overexpression. Unfortunately, this meant that we could not test any combinations of key TFs for the cell types available. However, as sequencing technologies improve in the coming years, we may be able to sequence all key TFs to sufficient depth allowing our model to test the perturbation of relevant combinations in a variety of cell types.

Furthermore, our model may be oversimplifying the biological processes occurring in the cells. For instance, we assume that the effect of perturbing TFs is additive and linear, which may not resemble reality where TFs work synergically (Cumbo et al., 2017) and non-linearly. We also assume that the GRN dynamics stays constant throughout the reprogramming

process. However, as the cell undergoes reprogramming, it could be possible that as its identity is changing, its GRN is changing as well. Finally, we also took our enhancers to be any part of the genome from our chromatin conformation database, which may not necessarily be reliable. A better alternative could be to use enhancer locations from a database (Wang et al., 2016), however this would be limited by the completeness of the database.

In light of the above mentioned limitations, we propose a few directions for future research in each of the two main applications.

Data simulation - In this thesis, we used a rather crude estimate for the distribution of the scRNA-seq and scATAC-seq data. In the future, we could use a more realistic distribution such as zero-inflated negative binomial which would account for drop outs.

We could also devise some evaluation metrics to compare our simulation to the real data and measure its similarity on data properties (eg. mean and variance) and biological properties (eg. differentially expressed genes). We could also test its computational scalability, which would be a way to demonstrate the utility of our simulation framework.

Transcription factor perturbation - An important task for future work would be to test our model for other TF perturbations and cell conversions. Although there is currently limited high quality data of matched scRNA-seq and scATAC-seq, these technologies are still being developed and improved. For example, a recent study profiled more than 450,000 brain cells in human, marmoset monkey and mouse using SNARE-seq2 (Bakken et al., 2020), another matched scRNA-seq and scATAC-seq protocol.

Our results and predicted conversions could also be validated in more detail. This could take the form of a method to estimate the cell type after conversion and a metric to evaluate the conversion efficiency, similar to that of CellOracle (Kamimoto et al., 2020). Alternatively, any predictions could be experimentally validated, however this will require assistance from a wet lab with access to the required cells, equipment and expertise.

The model could also be extended, by considering non-linear relationships between TFs and target genes. Regularisation (like LASSO) could also be incorporated into the model to reduce the effect of unimportant TFs, however given our small number of reasonably highly expressed TFs this was not necessary. Cells could also be clustered using fuzzy clustering

to account for transient cell states that contain properties of multiple cell types.

5.2 Conclusion

In summary, this thesis provides an in-depth review and synthesis of the current state of the mathematical models for cell reprogramming and related concepts including identifying key TFs for cell reprogramming. Each of these methods have introduced an insightful perspective into these complex biological processes. However, these algorithms were greatly limited by the data available at the time.

Here, we introduced a new method to use emerging matched scRNA-seq and scATAC-seq data to obtain a more accurate model for the intracellular dynamics governing gene regulation. We have shown that our model can be used to simulate matched scRNA-seq and scATAC-seq data, recapitulating known properties of cell differentiation and cell reprogramming. We further demonstrate that this model can be used on real data, simulating the effect of overexpressing key TFs in different hair cell lineages.

We hope that this framework will improve the ability to predict the effect of perturbing TFs in silico, guiding the TF combinations and quantities to be experimentally validated. This should significantly reduce the time and cost of finding new cell conversions, accelerating the development of cell reprogramming therapies, alleviating a significant portion of the world's disease burden.

Bibliography

Aguirre, A., Sancho-Martinez, I., Izpisua Belmonte, J., 2013. Reprogramming toward heart regeneration: Stem cells and beyond. Cell Stem Cell 12, 275–284. URL: https://www.sciencedirect.com/science/article/pii/S1934590913000635.

Aibar, S., González-Blas, C.B., Moerman, T., Huynh-Thu, V.A., Imrichova, H., Hulselmans, G., Rambow, F., Marine, J.C., Geurts, P., Aerts, J., van den Oord, J., Atak, Z.K., Wouters, J., Aerts, S., 2017. Scenic: single-cell regulatory network inference and clustering. Nature Methods 14, 1083–1086. URL: https://doi.org/10.1038/nmeth.4463, doi:10.1038/nmeth.4463.

Aydin, B., Mazzoni, E.O., 2019. Cell Reprogramming: The Many Roads to Success. Annual Review of Cell and Developmental Biology 35, 433–452. URL: https://doi.org/10.1146/annurev-cellbio-100818-125127, doi:10.1146/annurev-cellbio-100818-125127. _eprint: https://doi.org/10.1146/annurev-cellbio-100818-125127.

Bakken, T.E., Jorstad, N.L., Hu, Q., Lake, B.B., Tian, W., Kalmbach, B.E., Crow, M., Hodge, R.D., Krienen, F.M., Sorensen, S.A., Eggermont, J., Yao, Z., Aevermann, B.D., Aldridge, A.I., Bartlett, A., Bertagnolli, D., Casper, T., Castanon, R.G., Crichton, K., Daigle, T.L., Dalley, R., Dee, N., Dembrow, N., Diep, D., Ding, S.L., Dong, W., Fang, R., Fischer, S., Goldman, M., Goldy, J., Graybuck, L.T., Herb, B.R., Hou, X., Kancherla, J., Kroll, M., Lathia, K., van Lew, B., Li, Y.E., Liu, C.S., Liu, H., Lucero, J.D., Mahurkar, A., McMillen, D., Miller, J.A., Moussa, M., Nery, J.R., Nicovich, P.R., Orvis, J., Osteen, J.K., Owen, S., Palmer, C.R., Pham, T., Plongthongkum, N., Poirion, O., Reed, N.M., Rimorin, C., Rivkin, A., Romanow, W.J., Sedeño-Cortés, A.E., Siletti, K., Somasundaram, S., Sulc, J., Tieu, M., Torkelson, A., Tung, H., Wang, X., Xie, F., Yanny, A.M., Zhang, R., Ament, S.A., Behrens, M.M., Bravo, H.C., Chun, J., Dobin, A., Gillis, J., Hertzano, R., Hof, P.R., Höllt, T., Horwitz, G.D., Keene, C.D., Kharchenko, P.V., Ko, A.L., Lelieveldt, B.P., Luo, C., Mukamel, E.A., Preissl, S., Regev, A., Ren, B., Scheuer-

mann, R.H., Smith, K., Spain, W.J., White, O.R., Koch, C., Hawrylycz, M., Tasic, B., Macosko, E.Z., McCarroll, S.A., Ting, J.T., Zeng, H., Zhang, K., Feng, G., Ecker, J.R., Linnarsson, S., Lein, E.S., 2020. Evolution of cellular diversity in primary motor cortex of human, marmoset monkey, and mouse. bioRxiv URL: https://www.biorxiv.org/content/early/2020/04/04/2020.03.31.016972, doi:10.1101/2020.03.31.016972, arXiv:https://www.biorxiv.org/content/early/2020/04/04/2020.03.31.016972.full.pdf.

Barker, R.A., Parmar, M., Studer, L., Takahashi, J., 2017. Human trials of stem cell-derived dopamine neurons for parkinson's disease: Dawn of a new era. Cell Stem Cell 21, 569–573. URL: http://www.sciencedirect.com/science/article/pii/S193459091730382X.

Bergen, V., Lange, M., Peidli, S., Wolf, F.A., Theis, F.J., 2020. Generalizing rna velocity to transient cell states through dynamical modeling. Nature Biotechnology 38, 1408–1414. URL: https://doi.org/10.1038/s41587-020-0591-3, doi:10.1038/s41587-020-0591-3.

Bermingham-McDonogh, O., Reh, T., 2011. Regulated reprogramming in the regeneration of sensory receptor cells. Neuron 71, 389–405. URL: https://www.sciencedirect.com/science/article/pii/S0896627311006428.

Bonev, B., Mendelson Cohen, N., Szabo, Q., Fritsch, L., Papadopoulos, G.L., Lubling, Y., Xu, X., Lv, X., Hugnot, J.P., Tanay, A., Cavalli, G., 2017. Multiscale 3d genome rewiring during mouse neural development. Cell 171, 557–572.e24. URL: https://doi.org/10.1016/j.cell.2017.09.043, doi:10.1016/j.cell.2017.09.043.

Bornholdt, S., 2008. Boolean network models of cellular regulation: prospects and limitations. Journal of The Royal Society Interface 5, S85–S94. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2008.0132.focus, doi:10.1098/rsif.2008.0132.focus, arXiv:https://royalsocietypublishing.org/doi/pdf/10.1098/rsif.2008.0132.focus.

Buganim, Y., Faddah, D., Cheng, A., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S., van Oudenaarden, A., Jaenisch, R., 2012. Single-cell expression analyses during cellular reprogramming reveal an early stochastic and a late hierarchic phase.

Cell 150, 1209–1222. URL: https://www.sciencedirect.com/science/article/pii/S0092867412010215, doi:https://doi.org/10.1016/j.cell.2012.08.023.

Burda, Z., Krzywicki, A., Martin, O.C., Zagorski, M., 2011. Motifs emerge from function in model gene regulatory networks. Proceedings of the National Academy of Sciences 108, 17263–17268. URL: https://www.pnas.org/content/108/42/17263, doi:10.1073/pnas.1109435108, arXiv:https://www.pnas.org/content/108/42/17263.full.pdf.

Cahan, P., Li, H., Morris, S.A., Lummertz da Rocha, E., Daley, G.Q., Collins, J.J., 2014. Cellnet: network biology applied to stem cell engineering. Cell 158, 903–915. URL: https://pubmed.ncbi.nlm.nih.gov/25126793, doi:10.1016/j.cell.2014.07.020. 25126793[pmid].

Cameron, C.J., Dostie, J., Blanchette, M., 2020. Hifi: estimating dna-dna interaction frequency from hi-c data at restriction-fragment resolution. Genome Biology 21, 11. URL: https://doi.org/10.1186/s13059-019-1913-y, doi:10.1186/s13059-019-1913-y.

Cannoodt, R., Saelens, W., Deconinck, L., Saeys, Y., 2020. dyngen: a multi-modal simulator for spearheading new single-cell omics analyses. bioRxiv URL: https://www.biorxiv.org/content/early/2020/09/14/2020.02.06.936971, doi:10.1101/2020.02.06.936971, arXiv:https://www.biorxiv.org/content/early/2020/09/14/2020.02.06.936971.full.pdf.

Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., Daza, R.M., McFaline-Figueroa, J.L., Packer, J.S., Christiansen, L., Steemers, F.J., Adey, A.C., Trapnell, C., Shendure, J., 2018. Joint profiling of chromatin accessibility and gene expression in thousands of single cells. Science 361, 1380–1385. URL: https://science.sciencemag.org/content/361/6409/1380, doi:10.1126/science.aau0730, arXiv:https://science.sciencemag.org/content/361/6409/1380.full.pdf.

Cao, Y., Yang, P., Yang, J.Y.H., 2021. A benchmark study of simulation methods for single-cell rna sequencing data. bioRxiv URL: https://www.biorxiv.org/content/early/2021/06/02/2021.06.01.446157, doi:10.1101/2021.06.01.446157, arXiv:https://www.biorxiv.org/content/early/2021/06/02/2021.06.01.446157.full.pdf.

Chen, S., Lake, B.B., Zhang, K., 2019. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nature Biotechnology 37, 1452–1457. URL: https://doi.org/10.1038/s41587-019-0290-0, doi:10.1038/s41587-019-0290-0.

Chichagova, V., Hallam, D., Collin, J., Zerti, D., Dorgau, B., Felemban, M., Lako, M., Steel, D.H., 2018. Cellular regeneration strategies for macular degeneration: past, present and future. Eye 32, 946–971. URL: https://doi.org/10.1038/s41433-018-0061-z, doi:10.1038/s41433-018-0061-z.

CRICK, F.H., 1958. On protein synthesis. Symp Soc Exp Biol 12, 138–163.

Cumbo, F., Vergni, D., Santoni, D., 2017. Investigating transcription factor synergism in humans. DNA Research 25, 103–112. URL: https://doi.org/10.1093/dnares/dsx041, doi:10.1093/dnares/dsx041, arXiv:https://academic.oup.com/dnaresearch/article-pdf/25/1/103/24151658/dsx041.pdf.

D'Alessio, A.C., Fan, Z.P., Wert, K.J., Baranov, P., Cohen, M.A., Saini, J.S., Cohick, E., Charniga, C., Dadon, D., Hannett, N.M., Young, M.J., Temple, S., Jaenisch, R., Lee, T.I., Young, R.A., 2015. A systematic approach to identify candidate transcription factors that control cell identity. Stem Cell Reports 5, 763–775. URL: https://doi.org/10.1016/j.stemcr.2015.09.016, doi:10.1016/j.stemcr.2015.09.016.

Del Vecchio, D., Abdallah, H., Qian, Y., Collins, J.J., 2017. A blueprint for a synthetic genetic feedback controller to reprogram cell fate. Cell Systems 4, 109–120.e11. URL: https://doi.org/10.1016/j.cels.2016.12.001, doi:10.1016/j.cels.2016.12.001.

Duncan, J.S., Fritzsch, B., 2013. Continued expression of gata3 is necessary for cochlear neurosensory development. PLOS ONE 8, 1–13. URL: https://doi.org/10.1371/journal.pone.0062046, doi:10.1371/journal.pone.0062046.

Eppig, J.T., 2017. Mouse Genome Informatics (MGI) Resource: Genetic, Genomic, and Biological Knowledgebase for the Laboratory Mouse. ILAR Journal 58, 17–41. URL: https://doi.org/10.1093/ilar/ilx013, doi:10.1093/ilar/ilx013, arXiv:https://academic.oup.com/ilarjournal/article-pdf/58/1/17/24325222/ilx013.pdf.

Fang, R., Preissl, S., Li, Y., Hou, X., Lucero, J., Wang, X., Motamedi, A., Shiau, A.K., Zhou, X., Xie, F., Mukamel, E.A., Zhang, K., Zhang, Y., Behrens, M.M., Ecker, J.R., Ren, B., 2021. Comprehensive analysis of single cell atac-seq data with snapatac. Nature Communications 12, 1337. URL: https://doi.org/10.1038/s41467-021-21583-9, doi:10.1038/s41467-021-21583-9.

Floettmann, M., Scharp, T., Klipp, E., 2012. A stochastic model of epigenetic dynamics in somatic cell reprogramming. Frontiers in Physiology 3, 216. URL: https://www.frontiersin.org/article/10.3389/fphys.2012.00216, doi:10.3389/fphys.2012.00216.

Fornes, O., Castro-Mondragon, J.A., Khan, A., van der Lee, R., Zhang, X., Richmond, P.A., Modi, B.P., Correard, S., Gheorghe, M., Baranašić, D., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W.W., Mathelier, A., 2019. JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Research 48, D87–D92. URL: https://doi.org/10.1093/nar/gkz1001, doi:10.1093/nar/gkz1001, arXiv:https://academic.oup.com/nar/article-pdf/48/D1/D87/31697271/gkz1001.pdf.

Franke, M.K., MacLean, A.L., 2021. A single-cell resolved cell-cell communication model explains lineage commitment in hematopoiesis. bioRxiv URL: https://www.biorxiv.org/content/early/2021/04/26/2021.03.31.437948, doi:10.1101/2021.03.31.437948, arXiv:https://www.biorxiv.org/content/early/2021/04/26/2021.03.31.437948.full.pdf.

Fukaya, T., Lim, B., Levine, M., 2016. Enhancer control of transcriptional bursting. Cell 166, 358–368. URL: https://www.sciencedirect.com/science/article/pii/S0092867416305736.

Furuyama, K., Chera, S., van Gurp, L., Oropeza, D., Ghila, L., Damond, N., Vethe, H., Paulo, J.A., Joosten, A.M., Berney, T., Bosco, D., Dorrell, C., Grompe, M., Ræder, H., Roep, B.O., Thorel, F., Herrera, P.L., 2019. Diabetes relief in mice by glucose-sensing insulin-secreting human α-cells. Nature 567, 43–48. URL: https://doi.org/10.1038/s41586-019-0942-8, doi:10.1038/s41586-019-0942-8.

Gam, R., Sung, M., Prasad Pandurangan, A., 2019. Experimental and computational approaches to direct cell reprogramming: Recent advancement and future challenges. Cells 8. URL: https://www.mdpi.com/2073-4409/8/10/1189, doi:10.3390/cells8101189.

Ghaziani, Z.N., Zhang, L., Wang, B., 2020. simatac: A single-cell atac-seq simulation framework. bioRxiv URL: https://www.biorxiv.org/content/early/2020/12/05/2020.08.14.251488, doi:10.1101/2020.08.14.251488, arXiv:https://www.biorxiv.org/content/early/2020/12/05/2020.08.14.251488.full.pdf.

Grath, A., Dai, G., 2019. Direct cell reprogramming for tissue engineering and regenerative medicine. Journal of Biological Engineering 13, 14. URL: https://doi.org/10.1186/s13036-019-0144-9, doi:10.1186/s13036-019-0144-9.

Han, H., Cho, J.W., Lee, S., Yun, A., Kim, H., Bae, D., Yang, S., Kim, C.Y., Lee, M., Kim, E., Lee, S., Kang, B., Jeong, D., Kim, Y., Jeon, H.N., Jung, H., Nam, S., Chung, M., Kim, J.H., Lee, I., 2018. Trrust v2: an expanded reference database of human and mouse transcriptional regulatory interactions. Nucleic Acids Research 46, D380–D386. URL: https://doi.org/10.1093/nar/gkx1013, doi:10.1093/nar/gkx1013.

Haque, A., Engel, J., Teichmann, S.A., Lönnberg, T., 2017. A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. Genome Medicine 9, 75. URL: http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-017-0467-4, doi:10.1186/s13073-017-0467-4.

Heinäniemi, M., Nykter, M., Kramer, R., Wienecke-Baldacchino, A., Sinkkonen, L., Zhou, J.X., Kreisberg, R., Kauffman, S.A., Huang, S., Shmulevich, I., 2013. Gene-pair expression signatures reveal lineage control. Nature Methods 10, 577–583. URL: https://doi.org/10.1038/nmeth.2445, doi:10.1038/nmeth.2445.

Heydari, T., Langley, M.A., Fisher, C., Aguilar-Hidalgo, D., Shukla, S., Yachie-Kinoshita, A., Hughes, M., McNagny, K.M., Zandstra, P.W., 2021. Iqcell: A platform for predicting the effect of gene perturbations on developmental trajectories using single-cell rna-seq data. bioRxiv URL: https://www.biorxiv.org/

content/early/2021/04/03/2021.04.01.438014, doi:10.1101/2021.04.01.438014, arXiv:https://www.biorxiv.org/content/early/2021/04/03/2021.04.01.438014.full.pdf.

Hoi, C.S.L., Lee, S.E., Lu, S.Y., McDermitt, D.J., Osorio, K.M., Piskun, C.M., Peters, R.M., Paus, R., Tumbar, T., 2010. Runx1 directly promotes proliferation of hair follicle stem cells and epithelial tumor formation in mouse skin. Molecular and cellular biology 30, 2518–2536. URL: https://pubmed.ncbi.nlm.nih.gov/20308320, doi:10.1128/MCB.01308-09. 20308320[pmid].

Iwafuchi-Doi, M., Zaret, K.S., 2016. Cell fate control by pioneer transcription factors. Development 143, 1833–1837. URL: https://dev.biologists.org/content/143/11/1833, doi:10.1242/dev.133900, arXiv:https://dev.biologists.org/content/143/11/1833.full.pdf.

Kamimoto, K., Hoffmann, C.M., Morris, S.A., 2020. Celloracle: Dissecting cell identity via network inference and in silico gene perturbation. bioRxiv URL: https://www.biorxiv.org/content/early/2020/04/21/2020.02.17.947416, doi:10.1101/2020.02.17.947416, arXiv:https://www.biorxiv.org/content/early/2020/04/21/2020.02.17.947416.full.pdf.

Kaufman, C.K., Zhou, P., Pasolli, H.A., Rendl, M., Bolotin, D., Lim, K.C., Dai, X., Alegre, M.L., Fuchs, E., 2003. GATA-3: an unexpected regulator of cell lineage determination in skin. Genes Dev 17, 2108–2122.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler, D., 2002. The human genome browser at UCSC. Genome Res 12, 996–1006.

Khazaei, M., Ahuja, C.S., Fehlings, M.G., 2017. Induced pluripotent stem cells for traumatic spinal cord injury. Frontiers in Cell and Developmental Biology 4, 152. URL: https://www.frontiersin.org/article/10.3389/fcell.2016.00152, doi:10.3389/fcell.2016.00152.

Kim, J.H., Auerbach, J.M., Rodríguez-Gómez, J.A., Velasco, I., Gavin, D., Lumelsky, N., Lee, S.H., Nguyen, J., Sánchez-Pernaute, R., Bankiewicz, K., McKay, R., 2002.

Dopamine neurons derived from embryonic stem cells function in an animal model of parkinson's disease. Nature 418, 50–56. URL: https://doi.org/10.1038/nature00900, doi:10.1038/nature00900.

Kolmykov, S., Yevshin, I., Kulyashov, M., Sharipov, R., Kondrakhin, Y., Makeev, V.J., Kulakovskiy, I.V., Kel, A., Kolpakov, F., 2021. Gtrd: an integrated view of transcription regulation. Nucleic Acids Research 49, D104–D111. URL: https://doi.org/10.1093/nar/gkaa1057, doi:10.1093/nar/gkaa1057.

Kong, W., Biddy, B.A., Kamimoto, K., Amrute, J.M., Butka, E.G., Morris, S.A., 2020. Celltagging: combinatorial indexing to simultaneously map lineage and identity at single-cell resolution. Nature Protocols 15, 750–772. URL: https://doi.org/10.1038/s41596-019-0247-2, doi:10.1038/s41596-019-0247-2.

Kulkarni, A., Anderson, A.G., Merullo, D.P., Konopka, G., 2019. Beyond bulk: a review of single cell transcriptomics methodologies and applications. Current Opinion in Biotechnology 58, 129 – 136. URL: http://www.sciencedirect.com/science/article/pii/S0958166918302386, doi:https://doi.org/10.1016/j.copbio.2019.03.001.

Kurek, D., Garinis, G.A., van Doorninck, J.H., van der Wees, J., Grosveld, F.G., 2007. Transcriptome and phenotypic analysis reveals gata3-dependent signalling pathways in murine hair follicles. Development 134, 261–272. URL: https://dev.biologists.org/content/134/2/261, doi:10.1242/dev.02721, arXiv:https://dev.biologists.org/content/134/2/261.full.pdf.

La Manno, G., Soldatov, R., Zeisel, A., Braun, E., Hochgerner, H., Petukhov, V., Lidschreiber, K., Kastriti, M.E., Lönnerberg, P., Furlan, A., Fan, J., Borm, L.E., Liu, Z., van Bruggen, D., Guo, J., He, X., Barker, R., Sundström, E., Castelo-Branco, G., Cramer, P., Adameyko, I., Linnarsson, S., Kharchenko, P.V., 2018. Rna velocity of single cells. Nature 560, 494–498. URL: https://doi.org/10.1038/s41586-018-0414-6, doi:10.1038/s41586-018-0414-6.

Lambert, S.A., Jolma, A., Campitelli, L.F., Das, P.K., Yin, Y., Albu, M., Chen, X., Taipale, J., Hughes, T.R., Weirauch, M.T., 2018. The human transcription factors.

Cell 172, 650 – 665. URL: http://www.sciencedirect.com/science/article/pii/S0092867418301065, doi:https://doi.org/10.1016/j.cell.2018.01.029.

Lang, A.H., Li, H., Collins, J.J., Mehta, P., 2014. Epigenetic landscapes explain partially reprogrammed cells and identify key reprogramming genes. PLOS Computational Biology 10, 1–13. URL: https://doi.org/10.1371/journal.pcbi.1003734, doi:10.1371/journal.pcbi.1003734.

Liu, Z.P., Wu, C., Miao, H., Wu, H., 2015. Regnetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database 2015. URL: https://doi.org/10.1093/database/bav095, doi:10.1093/database/bav095. bav095.

Luo, X.j., Deng, M., Xie, X., Huang, L., Wang, H., Jiang, L., Liang, G., Hu, F., Tieu, R., Chen, R., Gan, L., 2013. GATA3 controls the specification of prosensory domain and neuronal survival in the mouse cochlea. Human Molecular Genetics 22, 3609–3623. URL: https://doi.org/10.1093/hmg/ddt212, doi:10.1093/hmg/ddt212, arXiv:https://academic.oup.com/hmg/article-pdf/22/18/3609/17259258/ddt212.pdf.

Ma, S., Zhang, B., LaFave, L.M., Earl, A.S., Chiang, Z., Hu, Y., Ding, J., Brack, A., Kartha, V.K., Tay, T., Law, T., Lareau, C., Hsu, Y.C., Regev, A., Buenrostro, J.D., 2020. Chromatin potential identified by shared single-cell profiling of rna and chromatin. Cell 183, 1103 – 1116.e20. URL: http://www.sciencedirect.com/science/article/pii/S0092867420312538, doi:https://doi.org/10.1016/j.cell.2020.09.056.

Mandon, H., Su, C., Haar, S., Pang, J., Paulevé, L., 2019. Sequential reprogramming of boolean networks made practical, in: Bortolussi, L., Sanguinetti, G. (Eds.), Computational Methods in Systems Biology, Springer International Publishing, Cham. pp. 3–19.

Martin, E.W., Sung, M.H., 2018. Challenges of decoding transcription factor dynamics in terms of gene regulation. Cells 7, 132. URL: https://pubmed.ncbi.nlm.nih.gov/30205475, doi:10.3390/cells7090132. 30205475[pmid].

McLeay, R.C., Bailey, T.L., 2010. Motif enrichment analysis: a unified framework and an evaluation on chip data. BMC Bioinformatics 11, 165. URL: https://doi.org/10.1186/1471-2105-11-165, doi:10.1186/1471-2105-11-165.

Melguizo-Sanchis, D., Xu, Y., Taheem, D., Yu, M., Tilgner, K., Barta, T., Gassner, K., Anyfantis, G., Wan, T., Elango, R., Alharthi, S., El-Harouni, A.A., Przyborski, S., Adam, S., Saretzki, G., Samarasinghe, S., Armstrong, L., Lako, M., 2018. ipsc modeling of severe aplastic anemia reveals impaired differentiation and telomere shortening in blood progenitors. Cell Death & Disease 9, 128. URL: https://doi.org/10.1038/s41419-017-0141-1, doi:10.1038/s41419-017-0141-1.

Oestreich, K.J., Weinmann, A.S., 2012. Master regulators or lineage-specifying? changing views on cd4+ t cell transcription factors. Nature Reviews Immunology 12, 799–804. URL: https://doi.org/10.1038/nri3321, doi:10.1038/nri3321.

Okawa, S., Nicklas, S., Zickenrott, S., Schwamborn, J., del Sol, A., 2016. A generalized gene-regulatory network model of stem cell differentiation for predicting lineage specifiers. Stem Cell Reports 7, 307–315. URL: https://doi.org/10.1016/j.stemcr.2016.07.014, doi:10.1016/j.stemcr.2016.07.014.

Osorio, K.M., Lee, S.E., McDermitt, D.J., Waghmare, S.K., Zhang, Y.V., Woo, H.N., Tumbar, T., 2008. Runx1 modulates developmental, but not injury-driven, hair follicle stem cell activation. Development 135, 1059–1068. URL: https://dev.biologists.org/content/135/6/1059, doi:10.1242/dev.012799, arXiv:https://dev.biologists.org/content/135/6/1059.full.pdf.

Osterwalder, M., Barozzi, I., Tissières, V., Fukuda-Yuzawa, Y., Mannion, B.J., Afzal, S.Y., Lee, E.A., Zhu, Y., Plajzer-Frick, I., Pickle, C.S., Kato, M., Garvin, T.H., Pham, Q.T., Harrington, A.N., Akiyama, J.A., Afzal, V., Lopez-Rios, J., Dickel, D.E., Visel, A., Pennacchio, L.A., 2018. Enhancer redundancy provides phenotypic robustness in mammalian development. Nature 554, 239–243. URL: https://doi.org/10.1038/nature25461, doi:10.1038/nature25461.

Parmar, M., Grealish, S., Henchcliffe, C., 2020. The future of stem cell therapies for

parkinson disease. Nature Reviews Neuroscience 21, 103–115. URL: https://doi.org/10.1038/s41583-019-0257-7, doi:10.1038/s41583-019-0257-7.

Pennacchio, L.A., Bickmore, W., Dean, A., Nobrega, M.A., Bejerano, G., 2013. Enhancers: five essential questions. Nature Reviews Genetics 14, 288–295. URL: https://doi.org/10.1038/nrg3458, doi:10.1038/nrg3458.

Plaisier, C.L., O'Brien, S., Bernard, B., Reynolds, S., Simon, Z., Toledo, C.M., Ding, Y., Reiss, D.J., Paddison, P.J., Baliga, N.S., 2016. Causal mechanistic regulatory network for glioblastoma deciphered using systems genetics network analysis. Cell Systems 3, 172–186. URL: https://doi.org/10.1016/j.cels.2016.06.006, doi:10.1016/j.cels.2016.06.006.

Pouyan, M.B., Kostka, D., 2018. Random forest based similarity learning for single cell RNA sequencing data. Bioinformatics 34, i79–i88. URL: https://doi.org/10.1093/bioinformatics/bty260, doi:10.1093/bioinformatics/bty260, arXiv:https://academic.oup.com/bioinformatics/article-pdf/34/13/i79/25098554/bty260_k

Pratapa, A., Jalihal, A.P., Law, J.N., Bharadwaj, A., Murali, T.M., 2020. Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. Nature Methods 17, 147–154. URL: https://www.nature.com/articles/s41592-019-0690-6, doi:10.1038/s41592-019-0690-6. number: 2 Publisher: Nature Publishing Group.

Qin, Q., Fan, J., Zheng, R., Wan, C., Mei, S., Wu, Q., Sun, H., Brown, M., Zhang, J., Meyer, C.A., Liu, X.S., 2020. Lisa: inferring transcriptional regulators through integrative modeling of public chromatin accessibility and chip-seq data. Genome Biology 21, 32. URL: https://doi.org/10.1186/s13059-020-1934-6, doi:10.1186/s13059-020-1934-6.

Qiu, P., 2020. Embracing the dropouts in single-cell rna-seq analysis. Nature Communications 11, 1169. URL: https://doi.org/10.1038/s41467-020-14976-9, doi:10.1038/s41467-020-14976-9.

Rackham, O.J.L., Firas, J., Fang, H., Oates, M.E., Holmes, M.L., Knaupp, A.S., Suzuki, H., Nefzger, C.M., Daub, C.O., Shin, J.W., Petretto, E., Forrest, A.R.R., Hayashizaki, Y., Polo, J.M., Gough, J., Consortium, T.F., 2016. A predictive computational framework for direct reprogramming between human cell types. Nature Genetics 48, 331–335. URL: https://doi.org/10.1038/ng.3487, doi:10.1038/ng.3487.

Raveh, E., Cohen, S., Levanon, D., Negreanu, V., Groner, Y., Gat, U., 2006. Dynamic expression of runx1 in skin affects hair structure. Mechanisms of Development 123, 842 – 850. URL: http://www.sciencedirect.com/science/article/pii/S0925477306001201, doi:https://doi.org/10.1016/j.mod.2006.08.002.

Ronquist, S., Patterson, G., Muir, L.A., Lindsly, S., Chen, H., Brown, M., Wicha, M.S., Bloch, A., Brockett, R., Rajapakse, I., 2017. Algorithm for cellular reprogramming. Proceedings of the National Academy of Sciences 114, 11832–11837. URL: https://www.pnas.org/content/114/45/11832, doi:10.1073/pnas.1712350114, arXiv:https://www.pnas.org/content/114/45/11832.full.pdf.

Shaik, I., Carmody, I.C., Chen, P.W., 2015. Chapter 93 - Treatment of Acute and Chronic Rejection. W.B. Saunders, Philadelphia. pp. 1317–1328. URL: https://www.sciencedirect.com/science/article/pii/B9781455702688000932.

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nature Reviews Genetics 13, 613–626. URL: https://doi.org/10.1038/nrg3207, doi:10.1038/nrg3207.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., Mesirov, J.P., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102, 15545. URL: http://www.pnas.org/content/102/43/15545.abstract, doi:10.1073/pnas.0506580102.

Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell 126, 663 – 676.

URL: http://www.sciencedirect.com/science/article/pii/S0092867406009767, doi:https://doi.org/10.1016/j.cell.2006.07.024.

Takahashi, K., Yamanaka, S., 2016. A decade of transcription factor-mediated reprogramming to pluripotency. Nature Reviews Molecular Cell Biology 17, 183–193. URL: https://doi.org/10.1038/nrm.2016.8, doi:10.1038/nrm.2016.8.

Tan, Y., Cahan, P., 2019. Singlecellnet: A computational tool to classify single cell rna-seq data across platforms and across species. Cell Systems 9, 207–213.e2. URL: https://doi.org/10.1016/j.cels.2019.06.004, doi:10.1016/j.cels.2019.06.004.

Teng, L., He, B., Wang, J., Tan, K., 2015. 4DGenome: a comprehensive database of chromatin interactions. Bioinformatics 31, 2560–2564. URL: https://doi.org/10.1093/bioinformatics/btv158, doi:10.1093/bioinformatics/btv158, arXiv:https://academic.oup.com/bioinformatics/article-pdf/31/15/2560/26289446/btv158.

de Torrenté, L., Zimmerman, S., Suzuki, M., Christopeit, M., Greally, J.M., Mar, J.C., 2020. The shape of gene expression distributions matter: how incorporating distribution shape improves the interpretation of cancer transcriptomic data. BMC Bioinformatics 21, 562. URL: https://doi.org/10.1186/s12859-020-03892-w, doi:10.1186/s12859-020-03892-w.

Waddington, C.H., 1966. Principles of development and differentiation. 1 ed., Macmillan, Basingstoke, United Kingdom.

Walters, B.J., Coak, E., Dearman, J., Bailey, G., Yamashita, T., Kuo, B., Zuo, J., 2017. In vivo interplay between p27kip1, gata3, atoh1, and pou4f3 converts non-sensory cells to hair cells in adult mice. Cell Reports 19, 307–320. URL: http://www.sciencedirect.com/science/article/pii/S221112471730390X.

Wang, S., Zang, C., Xiao, T., Fan, J., Mei, S., Qin, Q., Wu, Q., Li, X., Xu, K., He, H.H., Brown, M., Meyer, C.A., Liu, X.S., 2016. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. Genome Res 26, 1417–1429.

Weinreb, C., Rodriguez-Fraticelli, A., Camargo, F.D., Klein, A.M., 2020. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science 367. URL: https://science.sciencemag.org/content/367/6479/eaaw3381, doi:10.1126/science.aaw3381. publisher: American Association for the Advancement of Science Section: Research Article.

Whyte, W., Orlando, D., Hnisz, D., Abraham, B., Lin, C., Kagey, M., Rahl, P., Lee, T., Young, R., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153, 307–319. URL: https://www.sciencedirect.com/science/article/pii/S0092867413003929.

Xiao, X., Guo, P., Shiota, C., Zhang, T., Coudriet, G.M., Fischbach, S., Prasadan, K., Fusco, J., Ramachandran, S., Witkowski, P., Piganelli, J.D., Gittes, G.K., 2018. Endogenous reprogramming of alpha cells into beta cells, induced by viral gene therapy, reverses autoimmune diabetes. Cell Stem Cell 22, 78–90.e4. URL: https://doi.org/10.1016/j.stem.2017.11.020, doi:10.1016/j.stem.2017.11.020.

Xu, Q., Georgiou, G., Veenstra, G.J.C., Zhou, H., van Heeringen, S.J., 2020. Ananse: An enhancer network-based computational approach for predicting key transcription factors in cell fate determination. bioRxiv URL: https://www.biorxiv.org/content/early/2020/11/04/2020.06.05.135798, doi:10.1101/2020.06.05.135798, arXiv:https://www.biorxiv.org/content/early/2020/11/04/2020.06.05.135798.full.pdf.

Yardımcı, G.G., Ozadam, H., Sauria, M.E.G., Ursu, O., Yan, K.K., Yang, T., Chakraborty, A., Kaul, A., Lajoie, B.R., Song, F., Zhan, Y., Ay, F., Gerstein, M., Kundaje, A., Li, Q., Taylor, J., Yue, F., Dekker, J., Noble, W.S., 2019. Measuring the reproducibility and quality of hi-c data. Genome Biology 20, 57. URL: https://doi.org/10.1186/s13059-019-1658-7, doi:10.1186/s13059-019-1658-7.

Zappia, L., Phipson, B., Oshlack, A., 2017. Splatter: simulation of single-cell rna sequencing data. Genome Biology 18, 174. URL: https://doi.org/10.1186/s13059-017-1305-0, doi:10.1186/s13059-017-1305-0.

Zaret, K.S., Carroll, J.S., 2011. Pioneer transcription factors: establishing competence for gene expression. Genes Dev 25, 2227–2241.

Zeng, W., Chen, X., Duren, Z., Wang, Y., Jiang, R., Wong, W.H., 2019. Dc3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. Nature Communications 10, 4613. URL: https://doi.org/10.1038/s41467-019-12547-1, doi:10.1038/s41467-019-12547-1.

Zhang, Q., Liu, W., Zhang, H.M., Xie, G.Y., Miao, Y.R., Xia, M., Guo, A.Y., 2020. htftarget: A comprehensive database for regulations of human transcription factors and their targets. Genomics, Proteomics & Bioinformatics 18, 120–128. URL: https://www.sciencedirect.com/science/article/pii/S1672022920300954.