THE UNIVERSITY OF
WARWICK

AUTHOR: **Anas A. Rana** DEGREE: **Ph.D.**

TITLE: **Stochastic models for cell populations undergoing transitions**

DATE OF DEPOSIT: .....

I agree that this thesis shall be available in accordance with the regulations governing the University of Warwick theses.

I agree that the summary of this thesis may be submitted for publication.

I **agree** that the thesis may be photocopied (single copies for study purposes only).

Theses with no restriction on photocopying will also be made available to the British Library for microfilming. The British Library may supply copies to individuals or libraries. subject to a statement from them that the copy is supplied for non-publishing purposes. All copies supplied by the British Library will carry the following statement:

> "Attention is drawn to the fact that the copyright of this thesis rests with its author. This copy of the thesis has been supplied on the condition that anyone who consults it is understood to recognise that its copyright rests with its author and that no quotation from the thesis and no information derived from it may be published without the author's written consent."

AUTHOR'S SIGNATURE: .....

USER'S DECLARATION

1. I undertake not to quote or make use of any information from this thesis without making acknowledgement to the author.
2. I further undertake to allow no-one else to use this thesis while it is in my care.

DATE SIGNATURE ADDRESS

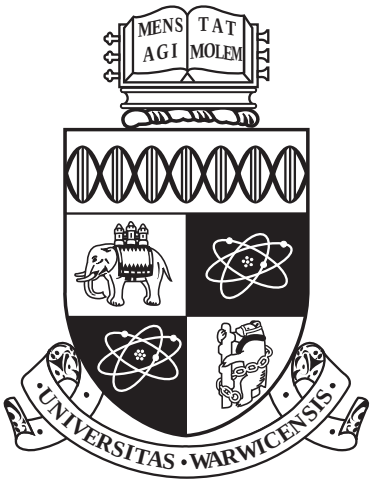
.....

.....

.....

.....

.....



Stochastic models for cell populations undergoing transitions

by

Anas A. Rana

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Physics and Complexity Science

... .

THE UNIVERSITY OF
WARWICK

Contents

Acknowledgments

Declarations

The work contained here is my own except when otherwise stated. Any work based on collaborative efforts has been indicated including the extent of my contribution. The thesis has been written by myself has not been submitted for any other degree at another University or institution.

- All experimental data analysed in this thesis was obtained by others and excluding Chapter 4 is all separately published work.
- The work in Chapter 3 and the analysis of data in Chapter 4 has been submitted for publication.
- The work in Chapter 5 has been published in *Scientific Reports* and my contributions to the work with some additional information are included here.

Abstract

Transformations on a cellular level caused by changes in genes, proteins or the epigenetic material present in cells play a key role in differentiation, reprogramming and disease. Such transformations are frequently stochastic on a single-cell level. The result is a heterogeneous cell population with an ever changing mixture. Often cells undergo transformation via intermediate stages which further convolute the transformation process. Reliable high-throughput data is commonly obtained on a cell population level therefore elucidating the underlying single-cell process is challenging. In this thesis we present and analyse models that probe population level data to answer questions about the transformation process and to distinguish between states.

We investigate a recently proposed stochastic model for transition processes called STAMM which is based on a latent Markov chain at the single-cell level. We present a computationally efficient unbiased approach to estimation, model selection and setting of tuning parameters. To complement our understanding of properties and behaviour of the model we implement a single-cell simulation setup. This not only allows us to investigate parameter estimation but we can also explore behaviour under violations of model assumptions. We also empirically investigate identifiability of the model. We implement application of the model to oncogenic transformation where the data time-course consists of genome-wide RNA-seq measurements. We also compare results from application of STAMM to a stem cell reprogramming microarray time-course to single-cell measurements carried out independently. Results show that not only is the model robust under mild violations of assumptions but state specific results can be compared to single-cell measurements. Under stronger violation of assumptions transitions between states are not estimated well. The model is therefore especially useful to steer further experiments in the right direction.

We then present a model that examines the response of cells in the cell cycle to incident radiation at different doses. Cells can either undergo programmed cell death or reenter the cell cycle after an interruption. A genome-wide RNA-seq measurement is made at the initial time point and subsequently fractions of cells with contrasting cell-fates can be distinguished and counted. The model assigns a score to each gene in a cell corresponding to its importance in determining cell fate. We implement a single-cell level simulation procedure and carry out illustrative simulations for one gene and for four genes. Parameter estimation in this model allows to distinguish genes that are important from genes that are not. This is only possible as long as the noise level is not too high.

Chapter 1

Introduction

The fascination to study and understand biological systems can be easily understand as this subject was initially popularised in the guise of medicine. Mathematics other than basic principles have only been applied to biology in recent years considering the long history of coexistence of the two fields. Maybe one of the first applications of mathematics in biology was the study of the current flow through the squid axon by Hodgkin & Huxley [1952]. Since then the field of neuroscience has broken away from mainstream biology and a number of advancements have been made which would have been unthinkable without the influence of mathematics in the field. In the last two decades due to an increase in high throughput data techniques available and the decrease in cost has lead to an increase in its availability. This in turn has lead to the need of mathematical techniques to analyse the data. Technique previously used of just observing data and drawing conclusions from it is not effective any more as the amount of data to compared and evaluated to draw conclusions is in the thousands. Additionally it has been determined that an approach of studying each component of the system independently is not sufficient any more. Genes interact with each other and so do proteins and any other component, in fact they interact and regulate each other as well. This has increased the need for an approach including biological experiments as well as expertise in mathematics, statistics, physics and engineering. In physics the use of mathematics alongside experiments has allowed for a much deeper understanding of physical phenomena. Classical mechanics allows a relatively simple description of macro phenomena even though we know that some of the assumptions are incorrect. On the other hand we have the description of micro phenomena described by quantum mechanics and even though it is possible to describe macro phenomena using quantum mechanics there is no need to include that detail. In a similar fashion interaction in cells and

between cells that lead to disease can be understood without a full understanding of all the elements involved in its description. Starting from the other direction it is possible to study interactions of simple genes and proteins, as these interactions are not yet fully understood; it is therefore not possible to say if such an approach will eventually allow the description of cell level behaviour. Unlike physics in biological systems there is still work needed for both approaches.

An extremely important process universal in many areas of biology is the transformation of cells from one type to another: cell differentiation [Tang et al., 2010; Vierbuchen et al., 2010], stem cell reprogramming [Takahashi & Yamanaka, 2006; Hanna et al., 2010], and disease formation [Hanahan & Weinberg, 2000; Vogel, 2010] to name but a few. These changes can be on the genetic level; on a protein level; or even on an epigenetic level. The transitions can be driven or initiated by very small perturbations in the form of induced genes. For such a system single-cell stochasticity is a very powerful concept and has be observed in a variety of experimental settings, such as E. coli [Elowitz et al., 2002] and stem cell reprogramming [Hanna et al., 2009]. When stochastic transitions occur on a single-cell level as a consequence at any time during a transition the cell population as a whole exists as a mixture of states. Moreover the exact composition of this mixture changes over time. Any measurements performed on homogenates[1] from this population results in population averages with unknown composition. Trivially the best strategy would be to measure at the single-cell level. Despite advancements in genome-wide single-cell measurements [de Souza, 2012; Tang et al., 2011] there remain challenges including limited availability and limits to the genome-wide measurements. Furthermore such measurements do not allow live tracking of cells and in fact the measurement process itself destroys cells thereby breaking continuity (i.e. the next measurement is on a different set of cells). These are the reasons for an incomplete understanding of transition processes.

Inevitably an important question arises here about our definition of states in the transformation. There are many ways to approach this and an obvious approach is to define states in a biological sense, but there is no consensus on a biologically motivated definition of a state. In a biological sense a complete definition of a state would include all possible information of state, if we don't include the different stages of the cell cycle as distinct states we have to define an area in this complex space as some changes in the cell will be due to inherent noise or the cell cycle. In most case not all information is available or is limited due to cost. In this work when we talk about states we are referring to changes in gene expression space to a

[1] mixture obtained from mechanically broken down cells

number of genes across the genome.

The study of stem cell reprogramming plays an important role in the development of personalised medicine, which in its extreme would allow the regrowth of whole organs to replace damaged ones. This could circumvent any issues arising from treatments derived using foreign cells, as treatments would originate from the patients own cells. Work carried out in this field has yielded the development of techniques that allow the development of neurons from embryonic stem cells [Vierbuchen et al., 2010; Pang et al., 2011] or creating muscle cells [Ieda et al., 2010; Efe et al., 2011]. These processes could become even more powerful when the starting point is a differentiated cell harvested from the patient [Takahashi & Yamanaka, 2006]. There are still unanswered questions in this field about the differences of cells derived from differentiated cells to cells derived directly from embryonic stem cells [Carey et al., 2011; Bock et al., 2011]. A better understanding of the transformation process would help identifying issues and could propose potential ways of improving the transformation process.

Cancer is a disease so prevalent in modern society that in the UK the life time chance of contracting cancer is more than one in three [Sasieni et al., 2011]. The disease has its source in multiple genetic mutations causing changes to natural cell functions such as cell proliferation and apoptosis which transforms cells from a healthy state to a cancerous state. These mutations code for proteins that are implicated in the eight hallmarks of cancer [Hanahan & Weinberg, 2011], which are the circumvention for the need of external growth signals, they are unaffected by external growth-inhibition signals, evasion of apoptosis, unlimited replication, the potential to create additional blood vessels from existing ones, invasion of other tissue types, energy management of the cell and avoiding the bodies immune response. Understanding the transformation process that changes a healthy cell population to a cancerous one would allow targeted intervention.

The cell cycle is central to the proliferation of cells and hence plays a key role in both transformations mentioned above. In fact mutations that can lead to cancer can be acquired during the cell cycle as this is the process during which DNA is replicated and the process can sometimes lead to errors. In a normal cell there are multiple checks that prevent such errors from propagating, but in an unhealthy cell genes central to these processes are mutated leading to malfunctions. Radiation plays an important role in the cell as it can be cause of mutations and is also used as treatment to kill unhealthy cells; damage to DNA can lead to apoptosis during the cell cycle. Radiation effects on the cell includes changes in gene expression and is also related to radiation strength [Gentile et al., 2003]. Some cells will arrest in

the cell cycle after radiation damage leading to apoptosis other will return to the cell cycle [Pawlik & Keyomarsi, 2004]. Understanding the mechanism that leads to the different responses is key in treatment as well as prevention of cancer.

In this thesis we attempt to learn single-cell level information from homogenate population averaged data of various types. We especially focus on gene expression and the role of genes in transformation. We take a data driven approach where model parameters are estimated by comparison to data. The first model we use is based on latent Markov processes aggregated over cells using a least squares estimation; it takes inspiration from the success in application of latent variable models such as hidden Markov models (HMMs) to hidden transitions in biology and genomics. The second one uses simple statistical principles for the derivation of the model and Monte Carlo integration to approximate an integral. These are simple models that allow probe of complicated biological processes.

Chapter 3 starts with the description of a model called *State Transitions using Aggregated Markov Models* (STAMM) based on previous work by Armond et al. [2013]; a latent stochastic process that obtains state-specific gene expression data as well as number of intermediate states from homogenate population time courses. Previous models exist that endeavour to achieve this such as deconvolving the cell cycle [Bar-Joseph et al., 2004]; dissecting expression data with known mixtures [Lähdesmäki et al., 2005]; or a hidden Markov model to determine expression levels and fractions of cells in each population [Roy et al., 2006]. Even though all such methods have strengths they also contain weaknesses addressed by STAMM. First, it provides single-cell level description of the transformations process and just like in the real system this process is hidden from observation due to averaging over multiple realisations (or cells). Second, in our model the latent process is in continuous time and therefore it does not have any need for special techniques to deal with missing data and unevenly distributed measurements. Third, the model relies on very few assumptions such as fractions of mixtures, it estimates all parameters from data. We also discuss in this chapter a single-cell level simulation process which is used to test the model properties and assumptions. Then we also outline a computationally efficient model selection procedure following and expanding on previous work by Armond et al. [2013]. Results show that parameter estimation works well even when violating assumptions only strong violations make estimations difficult and especially transition rates are not estimated well.

Chapter 4 describes an application of STAMM to RNA-seq time course of an oncogenic transformation using a healthy breast cell line (MCF-10A) as the initial population. We outline the pre-processing steps needed to apply the model

to RNA-seq data. Since observation are made as counts it is it is often considered that a Poisson distribution or a negative binomial distribution is the most fitting to such data; we show that once data has been pre-processed to allow for comparison between independent samples the data does not consist of integer counts any more but in fact it becomes continuous. The we show application of STAMM and show that it can be applied to large data sets in a relatively short time.

Chapter 5 briefly discusses results from applying STAMM to a microarray time course. This is obtained by reprogramming of secondary MEF cells to induced pluripotent stem cells. Then we show a possible next step once parameters have been obtained by STAMM. This step includes comparison of estimated parameters to new singlecell measurements which in this case were carried out on a different reprogramming system [Buganim et al., 2012]. We show that results are comparable despite measurements being made on different systems and using different methods.

Chapter 6 derives a model to investigate a slightly different system where less data is available and serves as a proof of concept since data was not available in time for this thesis. The gene expression is measured at the initial time point and the cell population is subject to a stimulus of various strength, at a later time fractions of cells in two distinct states are obtained by counting individual cells. An example of such a system is a population of cells radiated during the cell cycle upon which some go into arrest leading to apoptosis and some re-enter the cell cycle at a later time. This process is believed to be dependant on heterogeneity at the initial time point. This model attempts to assign a weight to each gene signifying its importance to the transformation process. Then we outline a single-cell simulation procedure and apply the model to a single gene simulation and a four gene simulation. Results show that high noise level makes it difficult to estimate parameters but at low noise levels parameters are estimated reasonably well allowing us to at a least distinguish between genes that are important for transformation to ones that are not.

Novel contributions of this thesis are listed below:

- **Chapter 3**
  - Single-cell simulation study. We present a simulation frame work imitating the biological single-cell processes, single-cells undergo random transition between states and observed expressions are an average over cells. The strength of this approach is that single-cell state specific parameters for data generation are known. Therefore we can empirically test estimation of parameters as well as the selection of correct number

of states. It also allows us to check estimation under violation of model assumptions additionally we can empirically investigate identifiability of the model.

  - Full investigation of estimation, including tuning parameters. We discuss and verify with simulations parameter estimation including setting tuning parameters using an unbiased approach. This is followed by sensitivity analysis performed for tuning parameters.
  - Computationally efficient model selection. For STAMM to be useful an unbiased estimation of model parameters especially the number of states is important. We put forward a simple approach which uses a form of cross-validation to determine number of states and other model parameters. We show that this method is effective during simulation and computationally efficient.

- **Chapter 4**

  - Application to RNA-seq time-course data. We show using an example how STAMM is able to analyse sequencing time-course data. The example we chose is using RNA-seq data from an *in vitro* study of oncogenic transformation of a healthy breast cancer cell line (MCF-10A) under induction of the oncogene *src*.

- **Chapter 5**

  - Comparing estimation to single-cell measurements for stem cell reprogramming. In this chapter the main contribution is computational and the comparison of estimated parameters from STAMM to single-cell level measurements taken at different time points.

- **Chapter 6**

  - Proof of concept of a novel model. Here we introduce a new model to understand the importance of genes for radiation response. The gene expression is only measured for the initial population subsequent measurements are fractions of cells transforming at different radiation doses.

Chapter 2

Background

This thesis is multidisciplinary and as such requires an introduction to multiple distinct areas. In this chapter we set out a description of background material which should prove useful to a reader with expertise in only one of the ares. The current chapter is therefore split into three self-contained sections covering the individual areas. First Section 2.1 includes mathematical background for the main techniques used in the thesis and introduces additional ideas that might place the work in broader context. Section 2.2 contains some background to the main biological ideas discussed. Finally in Section 2.3 the experimental techniques to obtain the data used in this work are outlined.

2.1 Mathematical background

This section contains the important basic principles we use to investigate biological systems. We start off by examining Markov chains and introduce basic principles for discrete time and continuous time Markov chains. They are an essential part of the model introduced in Chapter 3 and applied in later chapters. Then in Section 2.1.2 we introduce the slightly more involved hidden Markov models which are often applied to biological systems and have similarities to the aggregate Markov chains introduced in Chapter 3. Section 2.1.3 contains a brief presentation to the estimation procedure used in our investigation followed by a discussion on regularisation during estimation in Section 2.1.4 and outline two types of penalties commonly used to regularise estimated parameters. The concept of identifiability can of importance in estimation, in Section 2.1.5 identifiability is defined. Then we move on to a discussion on model selection and various techniques for distinguishing between models. Finally in Section 2.1.7 a very useful numerical tool is present for integrating

functions where a closed form solution is not possible. This method is employed in the second model put forward in this thesis, see Section 6.

2.1.1 Markov Chains

In Physics prior to the advent of Statistical Physics and Quantum Mechanics in the early 20th century the world was modelled as deterministic. Of course we now know that despite many aspects of the observable world being deterministic there is an even larger set of objects which does not lend itself to a deterministic description. Objects or ideas that can be described using deterministic principles do at times derive from non deterministic affects cancelling out or being only important at a different scale. Stochastic processes are used to describe systems where deterministic principles fail. A concept shared by many such systems is that they are evolving with a time-dependant stochastic part. One of the first attempts to describe such a system was the modelling of Brownian motion by Einstein [1905] which paved the way for further research on this topic. Here we start by defining some variables and simple principles governing one simple model that has found widespread application.

Let $X(t)$ be a time dependant random variable and $x_1, x_2 \ldots$ observations at $t = 1, t = 2, \ldots$ with joint probability $p(x_1, t = 1; x_2, t = 2; \ldots)$. The conditional probability for such a system given observations $y_1, y_2, \ldots$ at $\tau_1, \tau_2, \ldots$ is written as:

$$p(x_1, t = 1; x_2, t = 2; \ldots | y_1, \tau = 1; y_2, \tau = 2, \ldots). \tag{2.1}$$

The model we wish to consider is a special case of a stochastic process: a *Markov chain*. The most important principle underlying a Markov chain is the *Markov assumption* and it can be written in terms of the conditional probability. Say the current state at $t = n$ of a system is x_n; if we now write the probability of this measurement conditional on all preceding measurements $x_{n-1}, x_{n-2}, \ldots x_1$ the following principle must hold:

$$p(x_n, t = n | x_{n-1}, t = n-1; x_{n-2} t = n-2, \ldots x_1 t = 1 \ldots) = p(x_n, t = n | x_{n-1}, t = n-1), \tag{2.2}$$

where $t = 1 \leq t = 2 \leq \ldots t = n$. This means that the Markov property states; an observation is only conditionally dependant on the observation immediately preceding it. Further, the Markov property eqn. (2.2) and an initial distribution $\pi_1 = \pi(1) = p(x_1, t = 1)$ uniquely determines a Markov chain in discrete time and with a discrete state space. This only holds because any joint probability can be

written as a product subsequent transition probability starting from the initial distribution. If we now write the transition probability as $p_{ij}(t)$ as the transition from state i to state j we can write the distribution of a Markov chain at time t, $\pi(t)$ where $t \in \mathbb{N}$ as:

$$\pi(t) = \pi(1)P^t. \tag{2.3}$$

Continous time

If we extend the Markov chain to continuous time but still with a discrete state space, we have to introduce the generator matrix G. It uniquely defines a continuous time Markov chain together with the initial distribution similar to a Markov chain discrete time. The entries in the generator matrix are the transition rates from state i to state j such that $g_{i,j} \geq 0$. Diagonal entries of the matrix are $g_{i,i} = -\sum_{j \neq i} g_{i,j}$ and G is a stochastic matrix. We can now write the time evolution as:

$$\frac{d}{dt}P(t) = G\,P(t) = P(t)\,G, \tag{2.4}$$

which are known as the backward and forward equation respectively. If we are now interested in the state occupation as a function of time $\pi(t)$ and define set $\pi(t=0)$ as the initial state distribution we can use eqn. (2.4) and write

$$\frac{d}{dt}\pi(t) = \pi(0)\frac{d}{dt}P(t) = \pi(0)P(t)G = \pi(t)G. \tag{2.5}$$

It is often useful to rewrite eqn. (3.5) in parametric form using the definition of the generator matrix such that:

$$\frac{d}{dt} = \sum_{j \neq i} \left(\pi_j(t)g_{j,i} - \pi_i(t)g_{i,j}\right), \tag{2.6}$$

which is know as the Master equation and is useful in allowing us to derive many results for Markov chains.

2.1.2 HMM

An extension to Markov chains that has found widespread application is the Hidden Markov Model (HMM) initially developed by Baum & Petrie [1966]. The big difference to a classical Markov chain is that in a HMM the states of the Markov chain are not directly observed instead observations are made on outputs dependant on the hidden states. Each hidden state of the Markov chain has a probability distribution

over all possible outputs. The possible observable outputs can be continuous while the hidden Markov has a discrete state space.

More specifically the hidden Markov chains has transition probabilities (as described above) as well as emission probabilities. We can write the hidden state process at time t as S_t in discrete time and discrete states i.e. $S_t \in \{1, \dots K\}$ with transition probabilities $p_{i,j} = p(S_{t+1} = j | S_t = i)$ and initial distribution $\pi_1 = p(S_1 = k)$. Now we write the observation from this hidden Markov chain at time t as X_t; here we have to distinguish between two types of outputs they can either be discrete or continuous. If the observation is discrete i.e. $X_t \in \{1, \dots M\}$ we have emission probabilities $b_j(m) = p(X_t | S_t = k)$ with $j = 1, \dots M$. If the observation is continuous i.e. $X_t \in \mathbb{R}^M$ we have to use a continuous probability density function which is usually a weighted sum of Gaussian distributions $b_j(X_t) = \sum_{m=1}^{M} c_{jm} \mathcal{N}(\mu_{jm}, \Sigma_{jm})$, where c_{jm} is a weighting coefficient.

More information on HMMs can be found in MacDonald & Zucchini [1997] and in Zucchini & MacDonald [2009] including sample applications.

2.1.3 Least squares

Once the model to be used to analyse data is established the question of parameter estimation arises. The most widely used method in statistics for such a parameter estimation from data is to write down and maximise the likelihood function $\mathcal{L}(\theta) = f(X; \theta)$, where X is a random variable and θ is a set of parameters. Often it is more convenient to work with the log-likelihood function with $l(\theta) = \log \mathcal{L}(\theta)$. Since the logarithm is a monotone function the maximum of $l(\theta)$ is the same as the maximum of $\mathcal{L}(\theta)$. The advantage of the log-likelihood is that it can be easier to work with as the log transformation can simplify the likelihood. In some cases it is possible to obtain the maximum likelihood estimator (MLE) $\hat{\theta}$ that maximises $\mathcal{L}(\theta)$ and $\ell(\theta)$ in closed form. Especially in real world applications, this can be difficult or there might not exist a closed form solution; in such cases we need to use a more numerical approach.

To show one such numerical approach we present a simple application of the principles and choose a common statistical model to illustrate the idea. Say there exists a model which predicts the response variable Y from a set of input variables $X_1 \dots X_n$. The illustrative model we choose as an example system is the simple linear regression:

$$Y = \beta_0 + \beta_1 X_i + \epsilon_1, \tag{2.7}$$

where $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ and is independent of observations, β_0 and β_1 are parameters. It is often beneficial to write eqn. (2.7) in a vector form and include a 1 in the $\mathbf{X}$ vector and the β_0 in the β vector to write $y = \mathbf{X}^T\beta$. The least squares method proposes finding estimates of parameters by determining the parameter set that minimises

$$RSS(\beta) = (Y - \mathbf{X}^T\beta)^2, \tag{2.8}$$

also known as the residual sum of squares. Of course this can be extended to multiple input and response variables by computing the sum over all RSS, written in vector notiation it is:

$$RSS(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|_2^2, \tag{2.9}$$

where Y is now a vector over all responses and has length n and X is now a matrix with dimensions $n \times p$. This is a very simple and general way of finding parameters of a model that best describes observations.

2.1.4 Regularisation

It can be of benefit to penalise complex models that contain too many parameters and also to regularise estimation; this will also prevent overfitting. One way of achieving this regularisation to minimising the log-likelihood subject to a constraint on model parameters. An analogous but easier to implement solution is to minimise the log-likelihood with an additional penalty term on parameters. Choosing the example mentioned in Section 2.1.3 we write

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\, g(\beta) \right\}, \tag{2.10}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, $g(\beta)$ is the penalty function on β parameters and is chosen depending on application and λ is a parameter controlling the strength of the penalty; a larger value corresponds to a stronger penalty. The strength parameter λ is generally set by cross-validation, or any other model selection criteria (see below for a discussion of model selection). The penalty term itself can take several forms each favourable in different applications, with the same primary aim of reducing model complexity but different properties. Here we present two possibilities, Ridge regression and Lasso.

In Ridge regression the added penalty term takes the form of a ℓ_2 norm over the parameters $g(\beta) = \|\beta\|_2^2$. It penalises the magnitude of parameters and

also shrinks them towards zero but they are never exactly zero. Another method is least absolute shrinkage and selection operator (Lasso) which has been used before but was reintroduced to the statistics community by Tibshirani [1996]. It uses a ℓ_1 penalty i.e. $g(\beta) = \|\beta\|_1$. Just like ridge regression Lasso penalises the magnitude of model parameters and therefore larger positive or negative values shrinking parameters. The advantage of Lasso over ridge regression is that in addition to penalising higher values of parameters increasing the strength of the penalty forces more and more parameters to be exactly zero reducing model complexity, see Hastie et al. [2001] for more details.

2.1.5 Identifiability

An important concept for models where parameters are of physical importance identifiability plays a crucial role. It is important to remember that the concept of identifiability is a theoretical concept. It does not relate to application of the model to data but rather refers to idealised potentially infinite noise free data. Stating the definition formally, if we have a model $f(\theta_1)$ where θ_1 are parameters from the possible set of parameters Θ and θ_2 are also parameters from the same set. Then we say the model is identifiable when $f(\theta_1) = f(\theta_2)$ holds if and only if $\theta_1 = \theta_2$ for all possible $\theta_1, \theta_2 \in \Theta$. In other words if the model returns the same output for two different data sets it is not identifiable.

More information on identifiability and some applications can be found in Saccomani et al. [2003], Saccomani et al. [2010] and Jacquez & Greif [1985]. It is a widely studied subject especially in the context of linear models, but results for nonlinear models are more difficult to obtain.

2.1.6 Model selection

The aim of a model is to enable description of a complex (sometimes not even fully understood) phenomenon in way tractable by mathematics. Statistical or even mechanistic models include in their core assumptions and simplification of the real world problem they are attempting to describe. In some cases this can be the only way to describe properties of the system. Often experiments are sufficient to distinguish between models and identify the one closest to the real world problem. There are also cases where due to insufficient data or the type of data available two distinct models appear feasible. This problem is encountered especially when employing statistical models and comparing observed data with predictions from such models. The universal problem then becomes the comparison of model predictions

to a set of noisy data. Even in cases where the problem itself is identifiable (see discussion above) the existence of noise in observations poses a real difficulty. In such cases the question that one is really trying to answer is one of prediction. Model fit to data is not a sufficient measure, after all the error between model prediction and a specific data set will not carry over to a different data set; but it is still an important indicator and cannot be discarded. Another issue one wishes to avoid is overfitting as this ensures that prediction only works for fitted data.

Cross-validation

One method that uses this idea in a data driven fashion is cross-validation. The basic principle is quite simple, data is split into two independent subsets (the training set and the validation set) and model parameters are estimated on the training set and prediction using these parameters are compared with the validation set resulting in a performance score. A practical approach is called k-fold cross-validation. Here the data set is split into k randomly chosen equally sized subsets, one subset is retained as the validation set and the remaining $k-1$ subsets are used as training data. This step is repeated for each of the k subsets and the performance score is combined giving one score each model. In some applications as the ones discussed in later chapters of this work it due to limitation in data is only feasible to leave out one data point at a time. This procedure is then repeated for every model that is considered and the optimal model is chosen based on the best score. It is clear that such an approach has drawbacks; When dealing with large data sets computation times can quickly become unfeasible since estimations is repeated for k subsets, for all data and for all possible models. An additional problem can be that due to the random splits in data the choice of the split influences results significantly. Therefore it is often advisable to try multiple splits and compare results.

AIC and BIC

To avoid lengthy computation time there are many methods to approximate model selection results based on information obtained when estimating parameters, reducing the number of computations to one per model and data. This is an obvious advantages but it is a further approximation hence one has to be careful interpreting results and the choice of approximation used. In all such approaches the goodness of fit is juxtaposed to model complexity i.e. since more complicated models will perform better during estimation for a given data set we want to penalise models dependant on the complexity of the model. Such methods are due to historical nam-

ing convention referred to as information criteria. Here we will briefly introduce two such models; One of the most widespread is the Akaike information criterion (AIC) formulated by Akaike [1974] and the other is the Bayesian information criterion (BIC) presented by Schwarz [1978]. AIC is computed using the log-likelihood $l(\theta)$ for model parameters θ and the degrees of freedom d:

$$AIC = -2\, l(\theta) + 2d. \tag{2.11}$$

It is derived using ideas in information theory namely the Kullback-Leibler divergence between the true model for a data set and the used model. When dealing with small sample sizes this measure is not satisfactory and is replaced by the corrected version the AICc. It adds an additional term dependant on sample size n: $(2 * d(d+1))/(n-d-1)$.

The other information criterion, the BIC is derived using a Bayesian approach. It is also applied for a log-likelihood approach just like AIC. The derivation of BIC starts with the assumption that there exists a posterior distribution for models M_i given some data $\mathbf{Y}$ written $P(M_i|\mathbf{Y})$. Using Bayes' theorem we can write the odds of two models as:

$$\frac{P(M_i|\mathbf{Y})}{P(M_j|\mathbf{Y})} = \frac{P(M_i)}{P(M_j)} \cdot \frac{p(\mathbf{Y}|M_i)}{p(\mathbf{Y}|M_j)}. \tag{2.12}$$

The final term in eqn. (2.12) the ratio of marginal likelihoods is also known as the Bayes factor. The prior over models is typically assumed to be flat, then $P(M_i) = P(M_j)$ for all i, j and the only important in eqn. (2.12) here is the Bayes factor is the only important term. Approximating the log marginal likelihood $p(\mathbf{Y}|M)$ using the Laplace approximation yields the BIC score for a given model with parameters θ:

$$BIC = -2\, l(\theta) + \log(n)\, d. \tag{2.13}$$

The penalty for complex models in BIC is larger than in AIC hence it will select for simpler models. Additionally as sample size $n \to \infty$ and the model space includes the true model BIC will select the correct model, i.e. it is asymptotically consistent unlike AIC. Though AIC will sometimes perform better for smaller sample sizes [Hastie et al., 2001]. Hence the decision of information criterion will be application dependant.

2.1.7 Monte Carlo integration

In many applications of mathematical models to real world problems one encounters integrals without closed form solutions. For such cases numerical methods which have become particularly wide spread since advancements in computational power allow for large scale computations in relatively short time. Monte Carlo integration is one such example, generally Monte Carlo techniques are ubiquitous for classes of problems which include random numbers.

If we have a function $f(x)$ of a random variable x which is distributed between a and b, we can write the expectation of $f(x)$ as the integral

$$E(f(x)) = \frac{1}{b-a}\int_a^b f(x)dx, \tag{2.14}$$

The law of large numbers states that the sum of n random variables divided by n converges to the expected value of the random variable as $n \to \infty$. This is a very powerful idea and since we know that the function of a random variable is also a random variable we can extend this to the function of a random variable $f(x)$. As the number of samples n taken from random variable x approaches infinity

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \to \frac{1}{b-a}\int_a^b f(x)dx \qquad \text{as } n \to \infty. \tag{2.15}$$

The left hand side of eqn. (2.15) is therefore an asymptotically consistent estimator i.e. it converges to the right hand side as $n \to \infty$. Eqn. (2.15) is the Monte Carlo estimate of an integral and an important issue to always bear in mind for this method is convergence as without the result having converged the results are meaningless. Therefore in applications the shape of the probability distribution of x plays a central role in determining how well the Monte Carlo integral will converge. This means if function $f(x)$ has a large weight for a value of x which is unlikely to be sampled then for fine n the estimate will be bad. More information on Monte Carlo integration can be found in the review article by James [1980] and in [Norris, 1998, Chapter 5].

2.2 Biological background

In biological systems, distinct states and especially states that are indicative of disease can often be characterised by changes in gene expression levels [DeRisi et al., 1997; Spellman et al., 1998; Eisen & Brown, 1999; Brown & Botstein, 1999]. It is important to first understand the role played by genes in cells and also what it is

we mean by expression of a gene; therefore below we present a brief introduction to these topics as well as specific examples where changes in gene expression have been found to play a central role in changes of cell states.

2.2.1 The cell

If we define reproduction as a basic principle of life, the cell is the smallest unit that autonomously allows for this process. Independent of the process of reproduction, to ensure a faithful reproduction of organisms, information is passed on from one generation to the next. This information takes the form of a molecule, Deoxyribonucleic acid (DNA), organised in a double helix structure made up of two DNA strands. Information on such a strand is stored as a sequence of four distinguishable subunits. These four subunits are adenine (A), guanine (G), cytosine (C) and thymine (T), where C-G and A-T base pairs held together by hydrogen bonds make up the DNA double helix. Some sequences of the DNA code for proteins (coding region) and are known as genes. There is still debate about the extent to which DNA is made up of coding regions and noncoding regions; the difference between organisms can be very large. In humans only about 2% of DNA is considered to specifically contain genes [Elgar & Vavouri, 2008]. The initial idea that a major part of noncoding DNA is junk has been refuted recently as part of the international project of the ENCODE consortium [Pennisi, 2012]. Some of these areas have been shown to contain areas for proteins to bind and influence gene activity, or stretches where chemical modifications can switch off part of the DNA. It has also been established that regulation of genes is much more involved than previously thought and areas of DNA far away from a given position can influence gene expression at that position. The first step of converting coding regions of the DNA to proteins is known as transcription. In this process, information from the DNA is read off and converted into a single stranded molecule of ribonucleic acid (RNA); more specifically the type of RNA used is known as messenger RNA (mRNA). The following step involves some post-transcriptional modifications, where the mRNA is modified to mature mRNA. There are many such modifications the cell performs, but an important one is splicing. This process removes regions of mRNA called introns[1] that do not code for proteins after which the remaining exons are *spliced* together. There are multiple ways to splice a set of exons which can lead to many different proteins being read from the same sequence on DNA. Finally proteins are synthesised from mature mRNA during the process known as translation. During translation mRNA information is read in groups of three subunits called codons,

[1] a region inside a gene sequence that bind together the gene and are subsequently removed

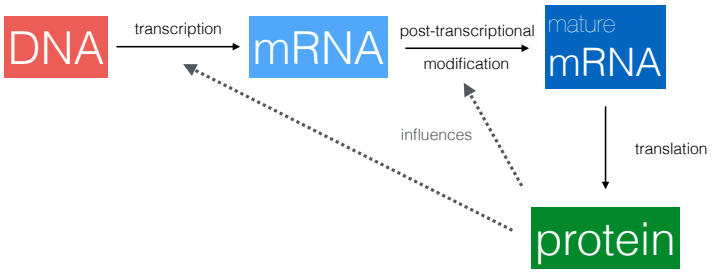


Figure 2.1: Gene expression. The central dogma of molecular biology states that genes on DNA are transcribed to mRNA which is transformed to mature mRNA by post-transcriptional modifications such as splicing. The mature mRNA is then translated to proteins. This final product which plays a central role in the functions of cells in turn also influences the transcription step as well as the post-transcriptional modification step.

hence there are $4^3 = 64$ codons which are mapped to 20 amino acids used to make proteins. These final DNA products then perform numerous functions in the cell including steps in their own synthesis. These three steps together are known as gene expression and form the central dogma of molecular biology, for a Summary see Figure 2.1. The DNA molecules contained in cells is more than two meters long and to economically pack them inside cells, which also contain vast amounts of other material, they are normally in packed structures known as chromatin. This is quite simply a combination of the DNA proteins which allow DNA be wrapped and packed very tightly in the nucleus of the cell. In addition to packing DNA, it also serves a functional purpose in that the topology of the packing hides and exposes certain genes on DNA. In a system where gene expression occurs via several protein dependant steps hidden genes cannot be expressed since regulatory proteins cannot interact with them.

Each step during gene expression outlined above happens at a different time scale therefore when interpreting biological data it is important to keep this in mind; as the effect one sees on gene expression at a given time might be due to a protein interaction process which took place at an earlier time point. The time scale can vary from seconds for proteins [Herce et al., 2013] up to 16 hours for the largest gene [Tennyson et al., 1995]. This of course becomes even more complicated once interactions between the different players are accounted for.

The control of gene expression is responsible for characteristics of a cell. As indeed a lung cell and a brain cell have identical genetic information, but gene expression is responsible for defining a cells purpose. This is in part controlled

by material outside of the coding region. Daughter cells in addition to inheriting genetic information from parent cells, also inherit information that determines the characteristics of the cell unrelated to DNA, i.e. daughter cells of a cell in lungs will remain lung cells. The umbrella term often used to describe all types of material that could pass on this information not directly part of DNA is *epigenetic material*. The final word on epigenetics has yet to be spoken but two types of information passed on in this manner are: Firstly the chromatin structure which determines active and inactive genes and to some extent the way they are read off. Active sections of chromatin are called euchromatin and inactive parts are called heterochromatin. Secondly DNA methylation which is the addition of a methyl group to certain bases in DNA, altering expression of genes.

Biological changes in state, which are the focus of this work, can be described in many different ways. Some changes in state are due to changes in gene expression. Other changes in state can also be epigenetic without direct changes in gene expression. Although both of these types of state changes are deeply interlinked they can occur at vastly different time scales. Here we will focus on state changes directly due to changes in gene expression.

Further information on the cell in general as well as epigenetics can be found in [Alberts et al., 2007, Chapters 1,7]. A specific overview of epigenetics has been attempted in Goldberg et al. [2007] and its effect on gene expression in Gibney & Nolan [2010].

2.2.2 Cancer Biology

An area of biology that is of great importance due to its impact on large populations and where state changes play a critical role is the study of cancer. Cancer is an extremely complex disease and attempts have been made to determine basic underlying principles, which the authors refer to as the 'Hallmarks of cancer' [Hanahan & Weinberg, 2000, 2011]. Genetic aberrations play a central role in this disease as many carcinogens (agents that can cause cancers) directly influence DNA sequences or are themselves mutations of genes naturally occurring in the cell. These defects range from point mutations on single base pairs all the way to deletions of large sections of DNA. Especially the process of cell division is highly susceptible to such attacks; in healthy cells there is a DNA repair mechanism in place that prevents changes from becoming permanent or leads to apoptosis (programmed cell death) if repair is impossible.

One important principle shared among cancer types is the unbound proliferation of cells leading to the build up of concentrated cell masses, also known as

tumours. Note that many such tumours are benign since they do not transform into cancers. The bodies defence mechanisms against such unchecked proliferation are circumvented either by introduction of oncogenes or mutations in tumour suppressor genes. Oncogenes are mutations in genes that result in a protein which drives uncontrolled cellular proliferation, resulting in tumour. These proteins do not respond to the natural signals that inhibit cell division, hence the proliferation is controlled and cannot be kept in check by the cells defence mechanisms against such growth. Alternatively mutations can occur in genes responsible for onset of apoptosis or DNA repair (also known as anti-oncogenes). Genetic mutations are especially problematic if they occur in the germline[2], as exemplified by LiFraumeni syndrome [Li & Fraumeni, 1969], where the mutation in an essential tumour suppressor gene is passed on to offspring and leads to hereditary predisposes to a large number of cancer types.

It is important to note that one mutation or defect is not sufficient to lead to the development of cancers; in fact several processes need to be affected. Additional processes developed by cells to defend against invading cancers include a limit on the number of times a cell can differentiate. Cells also rely on external growth signals to start differentiation as well as external growth inhibition signals to stop differentiation. Cancers are known for an uninhibited expansion, for which they need nutrients and therefore develop the ability to initiate the deployment of additional blood vessels a process known as angiogenesis. Tumorous cell become especially dangerous if they develop the ability to brake off from the main tumour and invade surrounding organs, tissues or even distant parts of the body known as metastasis.

As mentioned above an oncogene is a gene that can cause cancer and the first such oncogene (v-Src) was discovered in the late 1970s and early 1980s Martin [2001] after the initial discovery almost 70 years earlier; hinting at the possibility to induce solid tumours in chicken using a filtered agent by Rous [1911]. Investigating the Rous sarcoma virus in chicken the v-Src was discovered. Further research determined that a variant of v-Src, called c-Src is also contained in normal chicken. This discovery fundamentally changed the understanding of cancer which until then had been ascribed to viral causes. Later this gene was also found in humans and since this discovery it is probably the most widely studied oncogene; despite which there remain many unknowns. The protein from this oncogene has many downstream interactions with numerous other proteins. Hence it is not surprising that in almost 50% of tumours originating in breast, colon, liver, lung and the pancreas the c-Src interaction pathway is activated Dehm & Bonham [2013]. Due to mutations,

[2]cells from which egg or sperm cells are derived

c-Src is overexpressed and activated leading to the constant activation of downstream signalling pathways that ensure survival, proliferation and invasion leading to development of cancers.

Further details on cancer biology and the role of genes can be found in Weinberg [2013].

2.2.3 Stem cells

Central to cellular development is the creation of distinct differentiated cell types from a small collection of undifferentiated cells in the embryo referred to as embryonic stem cells. Clearly this transformation involves state changes on the genetic and epigenetic level. In recent years there has been an increase in research in these areas especially due to its potential in applications to personalised medicine; the next big frontier in medicine. The idea is to enable induction of alternative cell fates from embryonic cells or to enable development of cells that allow for a change in cell fates of tissues or blood samples. Examples include the development of neurons from cells that are responsible for creating extra cellular matrix known as fibroblasts [Vierbuchen et al., 2010; Pang et al., 2011] or the development of muscle cells from fibroblasts [Ieda et al., 2010; Efe et al., 2011]. All types of undifferentiated cells that can produce differentiated cell types are comprehensively known as stem cells (SC). Another property shared by all stem cells that they can differentiate to produce more stem cells multiple times. Broadly speaking there are two types of stem cells. Embryonic stem cells (ES cells) are cells derived from an embryo in its early development and adult (or somatic) stem cells which are found in fully developed organs. The main difference between the two types is how many types of cells they can differentiate into. ES cells are pluripotent i.e. they can differentiate into many possible cell types. Adult SCs can only differentiate into limited cell types often serving the function of replenishing damaged cells of a single organ also known as multi-potent stem cells. For medical applications of course ES cells are more useful, but for some people there are ethical concerns associated with their usage and their harvesting; independently of the rationale behind these concerns, it does create issues in research if such cells are used.

A new approach, proposed by Takahashi & Yamanaka [2006], is to use differentiated somatic cells and derive induced pluripotent stem cells (iPS cells) which have distinct advantages if successful. These iPS cells are able to differentiate to various cell types and could in future allow for personalised medicine. The process in creating iPS cells involves artificially inducing 4 genes (reprogramming factors) for several days and they do indeed find cells which have properties comparable to ES

cells. More detailed studies show that iPS are influenced by the used reprogramming factors and there are epigenetic differences between ES and iPS [Carey et al., 2011; Bock et al., 2011]. One important concern is the difference in DNA methylation of iPS cells and ES cells in terms of epigenetic material that would make their use difficult, this problem is now being addressed [Bagci & Fisher, 2013] as well as other safety concerns.

Further information on Stem cells, ES cells, somatic cells as well iPS cells, can be found in Lanza et al. [2009] and in Lanza & Atala [2013].

2.2.4 Cell cycle

An essential step governing the division, differentiation and maturation of all cells is the cell cycle. In simple terms it is the process by which two daughter cells are produced from one mother cell by duplication of the cell contents; most importantly the DNA. Details of the process can vary between organisms as well as between different stages of development. Most cells in the human body are not taking part in the cell cycle, but are in a resting phase. The most basic principles of the mammalian cell cycle can be summarised into four phases:

- **G_1 phase** The first gap phase during which cells increase in size. It also includes a restriction point up to which the cell is driven by external stimuli and after which it can progress through G_1 independently.
- **S phase** The transition from the G_1 to the S (Synthesis) phase contains a powerful checkpoint after which the cell is committed to duplication. During the S phase itself DNA is replicated.
- **G_2 phase** The second gap phase is not present in all organisms. In short the cell keeps increasing in size, synthesises proteins and prepares for mitosis. It contains a checkpoint to determine DNA damage and stops the process.
- **M phase** The final step in the cell cycle is the mitotic (M) phase. The duplicated chromosomes are separated into two cells and a new nucleus is created. The M phase also contains a checkpoint to ensure the cell is ready for division.

Despite all these checks and balances in place during the cell cycle, uncontrolled cell division still occurs in tumourigenesis as mentioned above. Intervention on the cell cycle plays a central role in unbound growth of cells. In many cases proteins essential during check points are mutated, inhibited or overexpressed [Williams

& Stoeber, 2012]. Understanding and perturbing elements in the cell cycle could be a good approach for potential cancer treatments since the mammalian cell cycle is conserved across a variety of cell types; at the same time it plays a central role in cancers.

The cell cycle is also sensitive to UV radiation which is a well known carcinogen. UV radiation incident on a cell can lead to damage to DNA, and if the damage is too extensive, cells can undergo apoptosis. If the damage is not too large some cells will arrest and re-enter the cell cycle at a later time. Radiation has an effect on gene expression which is also related to the intensity of the radiation as different genes become active to respond to the stimulus, but more importantly the type of response is driven by expression of certain genes. [Gentile et al., 2003].

More information about the cell cycle can be found in [Alberts et al., 2007, Chapter 17] and its relationship to cancer in [Weinberg, 2013, Chapter 8].

2.3 Experimental background

In this section we explore two techniques to obtaining time-course assays for genome wide gene expression data. Current techniques for such measurement are only possible on a population level. There exist techniques for single cell measurements [Buganim et al., 2012], but these techniques are not yet fully developed and only allow measurement of a limited number of genes. Initially to be able to obtain these measurements is to create a homogenate from the sample which is done by mechanically breaking down cells using a variety of different procedures. After which different subsets are filtered out of the mixture to perform microarray experiments and mRNA for RNA sequencing (RNA-seq).

2.3.1 Microarray

Different stages of gene expression can be measured using microarrays, here we present the one most commonly used the DNA microarray; the aim is to measure mRNA levels. There are two types of DNA microarray, the cDNA [Hughes et al., 2001] and the highdensity oligonucleotide chips [Lockhart et al., 1996]. The basic principle of cDNA[3] microarrays is based on high-density array with DNA sequences printed on them. The sample mRNA reverse-transcribed to cDNA labelled in two different colours. Equal proportions are mixed together and hybridised to the array and using a scanner fluorescence measurements are made for each colour. The resulting expression is obtained by the ratio of the measurements in each colour, see

[3]Complementary DNA, DNA reverse-transcribed from mRNA

Phimister [1999] for further information. To ensure measurements are comparable between different samples and even different experiments it is important to normalise each sample. The most robust normalisation method is to based on subtracting a position and intensity (A) dependant constant from the log ratios of the intensity measurements in green G and red R, where the colours are achieved by staining the two cDNA populations:

$$\log_2 \frac{R}{G} - l(A, j), \tag{2.16}$$

where $l(A, j)$ is the lowess fit [Cleveland, 1979] to the plot of $\log_2 R$ against $\log_2 G$ rotated counterclockwise by 45°. More detail on normalisation and cDNA microarray measurements can also be found in Dudoit, Sandrine and Yang, Yee Hwa and Callow, Matthew J and Speed, Terence P [2002].

2.3.2 RNA sequencing

Until recently the most prevalent method for obtaining gene expression data which is essential to understanding disease states has been DNA microarray measurement. One drawback is that observations are relative and indirect i.e. measurements via fluorescence intensity and ratio of two colours. Additionally microarray measurements are limited by prior knowledge of genes, since arrays can only be constructed to include sequences of known genes. A new contender that attempts to address some of the shortcomings of these is RNA-seq developed roughly 5 years ago [Mortazavi et al., 2008; Nagalakshmi et al., 2008]. RNA-seq measurements are integer count data and cover the whole genome.

Experimental protocol

RNA-seq experiments also attempt to measure mRNA obtained from homogenate. The first step is a random fragmentation of the sample mRNA. The next step is a reverse-transcription of the fragmented samples to cDNA. Next comes a Polymerase chain reaction (PCR) step which is a method of amplification to obtain more copies of DNA from few copies of a slice of DNA and is performed multiple times. This step is the source of one type of systematic error since different sections of DNA have a different susceptibility to a PCR amplification. In the next step each fragment is sequenced in high throughput machine; resulting sequences are referred to as *reads*. These reads can now be mapped to a known genome or transcriptome[4] resulting for one count for each fragment of gene found; alternatively reads can also

[4]collection of all RNA

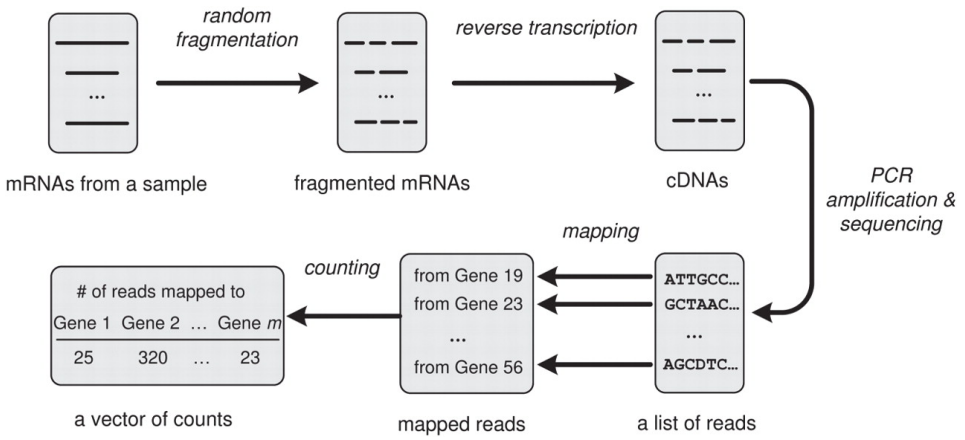


Figure 2.2: RNA-sequencing (RNA-seq). The most commonly used modern genome wide assay. From a homogenate cell sample mRNA is filtered out and passes through this pipeline to give integer count expression values for genes. Figure from Li et al., Normalization, testing, and false discovery rate estimation for RNA-sequencing data, Biostatistics, 2012, 13(3), by permission of Oxford University Press.

be used to construct a transcriptome without mapping it to a known genome (de novo assembly). Figure 2.2 from Li et al. [2012] summarises this protocol.

Normalisation

Though RNA-seq avoids some of the issues associated with microarray measurements it still has it's own difficulties that need to be overcome before analysing the data. The first as, already mentioned above, is the systematic error due to the PCR step. Another is due to the random fragmentation step; since larger genes will have more fragments resulting in a higher per gene count. Due to these reason the total number of counts is also not conserved across different samples.

Therefore there is a real need for a normalisation step prior to comparison of data, to ensure the affect of these problem are minimised. The issue of different gene lengths can be removed by simply normalising for gene length which is a known quantity when mapping counts to a genome. Additionally this is only an issue when comparing genes in the same sample. For comparisons between samples this step is unnecessary since gene length is constant across samples. The question normalisation between samples especially arises in differential expression. In RNA-seq data unlike for microarray data the question of normalisation has yet to be settled. The normalisation method most commonly used is to normalise all samples to a fixed number of reads, but this leads to issues as can be illustrated using a

simple example.

Say N_{ij} is the total number of reads in experiment i for gene j. In a simple case $N_{1j} \simeq 2\,N_{2j}$ for all genes and hence sequencing depth of experiment 1 is twice the sequencing depth of experiment 2. In a slightly more complex example the issue becomes clear. If we now consider $j = 1, \ldots, 100$ to have $N_{1j} = 1$ and $N_{2j} = 2$ and for $j = 101$ $N_{1j} = 100$ and $N_{2j} = 0$ both experiments have the same sequencing depth which would suggest all genes are differentially expressed. But it would be more realistic to consider that the sequencing depth of experiment 2 is twice that of experiment 1 and that only gene $j = 101$ is differentially expressed. Therefore a good strategy would be identify genes that are not differentially expressed and calculate a sequencing depth just for those and use it for the whole sample. One idea would be to identify housekeeping genes[5] for the particular application at hand and calculate a sequencing depth; but this is an unsatisfactory solution since it does not generalise (each application would have different housekeeping genes) and requires further experiments or prior knowledge. Bullard et al. [2010] propose a method using quantiles instead of total counts. Robinson & Oshlack [2010] propose a method based on total count normalisation called trimmed mean of M values (TMM). The idea is to impose a cutoff on log-fold change M-value and the absolute expression level A-value and calculate the sequencing depth using what remains. The M-value and A-value is calculated as follows:

$$M_j = \log_2 \frac{N_{ij}/N_i}{N_{i'j}/N_{i'}} \tag{2.17}$$

$$A_j = \tfrac{1}{2} \log_2 \left(N_{ij}/N_i \bullet N_{i'j}/N_{i'} \right). \tag{2.18}$$

In the R package `edgeR` this method as well as a few other have been implemented for application to RNA-seq data Robinson et al. [2010]. A summary comparing different normalisation methods can be found in Oshlack et al. [2010].

[5] genes that are supposed to remain constant

Chapter 3

State transitions using aggregated Markov models

3.1 Introduction

Diverse biological processes have been observed to undergo transitions under influence of a stimulus. These transitions lead to changes on a cellular level between distinct phenotypic states. The source of these phenotypic changes can be morphological, epigenetic, or even at a protein interaction level.

One big obstacle in understanding the source of such phenotypic changes at a single-cell level are that observations are generally not at the single-cell level although it is only possible to perform observations on a population level. This restriction is experimental in nature and although sometimes it is possible to make observations on a single-cell level there are limitations often on the amount of information that can be obtained as discussed in detail in Chapter 1.

The model we explore is called *State Transitions using Aggregated Markov Models* (STAMM) initially proposed by Armond et al. [2013]. This is a stochastic model identifying state level information for single-cell transformations using population level data. Single-cell level dynamics, latent in the model, is described by a Markov chain which is then aggregated over multiple cells. This model has been applied to cancer cell lines [Casale et al., 2013] and stem cell reprogramming (see Chapter 5) there still remains work to be done in obtaining a better understanding of this model. We use single-cell level simulations to probe model properties and model assumptions.

A type of model that has found widespread application in the description and has shares certain characteristics with STAMM is the Hidden Markov model (HMM).

Many examples of real world applications of HMMs exist including the deconvolution of population level microarray data [Roy et al., 2006]. Other types of models that attempt to study population level heterogeneity exist some using deconvolution algorithms for the cell cycle and for microarray data involving additional information [Bar-Joseph et al., 2004, 2008]. The main advantage of STAMM is that the latent model is in continuous time and therefore application to time-course data with uneven time-points is quite simple. This is discussed in more detail in Section 1. Another type of model that is even closer to STAMM was studied by Kalbfleisch et al. [1983] but this model investigates panel data and often near or at equilibrium.

In this chapter we follow and expand on previous work done starting with a detailed description of STAMM in Section 3.2 including a detailed description of model assumptions. In Section 3.3 we outline parameter estimation including a way to perform model selection as well as an efficient and unbiased estimation pipeline. Then we specify the single-cell simulation setup (Section 3.4) before moving on to the results from single-cell simulation data in Section 3.5. These are split up into results from performing a small scale simulation to probe the model and to test sensitivity to breaking assumptions; and large scale simulation results to demonstrate the whole pipeline.

3.2 Model outline

3.2.1 The STAMM model

STAMM defines a latent stochastic process on the single-cell level that isn't directly observed. Using the latent stochastic processes and aggregating across cells we can obtain a cell-population level likelihood. The latent single-cell process is described using a Markov chain with a discrete and finite state space but it is continuous in time. Biological states in the system are identified with the state space of the underlying biological system, indexed by $k \in \{1, ..., K\}$. Transitions between states k and k' are determined by transition rates between these states denoted by $\mathbf{w} = \{w_{k,k'}\}$. Assuming that cell death and cell doubling compensate each other, i.e. the number of cells is conserved at all time t; the probability for any cell to be in state k at any given time t can be obtained by solving the master equation of the Markov chain. The resulting state occupation probability for the population $p_k(t; \mathbf{w})$ is a function of time and also the state transitions.

This model can be applied to any type of time-course data, including transcript or protein abundance. Here unless otherwise stated we will focus the description, without loss of generality, on gene expression data. Let $x_j(t)$ be the cell-

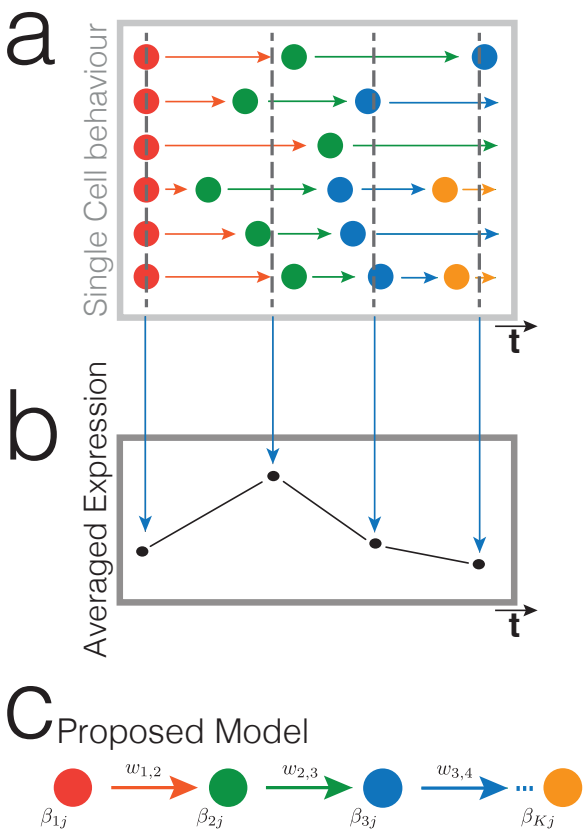


Figure 3.1: Model description. (a) In any biological system undergoing transitions between multiple states where the time of transition is stochastic, cells states are heterogeneous in the population at any given time. (b) Assays performed on homogenates of that cell population will only yield data averaged over sampled sub-populations. (c) We describe this system with 'State Transitions using Aggregated Markov Models' (STAMM) where single-cell level processes are described by a latent continuous-time Markov chain which is aggregated over cells to give a likelihood (see Section 3.2.1). The Markov chain has a discrete state space which corresponds to biological states of the system (shown in different colours). Estimation of parameters in STAMM is performed using population level data. (Figure adapted from Armond et al. [2013].)

population-averaged gene expression of gene j at time t, obtained from a homogenate assay such as RNA-seq or microarray expression. When investigating transitions it is prudent to design experiments with an initial state that is reasonably homogeneous, therefore our model assumes that initially all cells in the population occupy the same state, this is often part of the experimental design of investigating changes

from an initial homogeneous starting population. At any subsequent time point cells exist in a mixture of states, hence any measurement $x_j(t)$ made on a population level is an average over multiple states. We further assume that there is a mean expression level per gene constant across a state. This is denoted by β_{kj}, the gene expression level for gene $j \in \{1 \dots p\}$ in state k.

In the limit of large numbers of cells the fraction of cells in any state k is given by the state occupation probability $p_k(t; \mathbf{w})$. We can now write the observed average gene expression, $x_j(t)$ for gene j at time t, as the sum of all occupation probabilities weighted by their respective gene expression signatures. The resulting model for the average gene expression from a latent Markov chain model is written as:

$$x_j(t) = \sum_k p_k(t; \mathbf{w}) \beta_{kj} = P(t; \mathbf{w})\, \beta_j, \tag{3.1}$$

where the right hand side is the vectorised form of the model, with the row vector $P(t; \mathbf{w}) = [p_1(t, \mathbf{w}), \dots, p_K(t, \mathbf{w})]$ and column vector $\beta_j = [\beta_{1j}, \dots, \beta_{Kj}]^T$. Assuming an additive Gaussian noise model with gene-specific noise variance σ_j^2 we arrive at the likelihood:

$$\mathcal{L}\left(\mathbf{w}, \beta_j, \sigma_j \mid \{x_j(t)\}\right) = \prod_{t=1}^{T} \mathcal{N}\left(g(x_j(t)) | g\left(P(t; \mathbf{w})\, \beta_j\right),\ \sigma_j^2\right), \tag{3.2}$$

where $\mathcal{N}(\cdot|\mu, \sigma^2)$ denotes a Normal density with mean μ and variance σ^2 and the function g denotes a transformation whose choice depends on the data type under investigation.

Applied to microarray experiment which use ratios of fluorescence intensities between measurements in the red spectrum R and green spectrum G. The transformation g used is $\log_2$ [Dudoit, Sandrine and Yang, Yee Hwa and Callow, Matthew J and Speed, Terence P, 2002], for further details see Section 2.3.1. When investigating RNA-seq data we use arcsinh as the transformation [Hoffman et al., 2012; Johnson, 1949], defined as $\text{arcsinh}(x) = \ln(x + \sqrt{(}x^2 + 1))$ (for more details see Section 4.4). RNA-seq data cannot be normalised in the same way as microarray data most importantly because it contains measurements which are exactly zero. The arcsinh normalisation is useful here, because unlike the log transformation it does not have a singularity at zero but has the same variance-normalisation properties.

To use the likelihood it is necessary to compute the state occupation probabilities at any time t observations are made.

Markov chain and the master equation

Until now we have not placed any restrictions on the latent Markov process in this model and we have formulated the likelihood eqn. (3.2) for a general case. To classify the Markov chain we first clarify some notation: let the states of the latent process be $k \in |1 \dots K|$ and denote transitions between states k and k' as $\mathbf{w} = |w_{k,k'}|$, when the value of the transition is equal to zero the transition does not exist for that model. Using these parameters we can record the structure of any Markov chain, the topology of which has implication on identifiability of the model (for further discussion see Section 3.2.2). In this discussion we limit ourselves to a pure birth process where $w_{k,k+1} \neq 0$ for all k and zero otherwise. Such a Markov chain also excludes branches. The resulting master equation is written as:

$$\frac{dp_k(t)}{dt} = w_{k-1,k}\, p_{k-1}(t) - w_{k,k+1}\, p_k(t). \tag{3.3}$$

We can write the master equation in matrix notation using the generator matrix $\mathbf{G}(\mathbf{w})$. It is a $K \times K$ matrix where the only non-zero entries are on the diagonal, $g_{kk} = -w_{k,k+1}$, and the subdiagonal $g_{k\,k-1} = w_{k-1,k}$. The general soluation to the master equation is written as:

$$P(t; \mathbf{w}) \quad = \quad \exp\left(\mathbf{G}(\mathbf{w})\, t\right) P(0) \tag{3.4}$$

In investigating transition processes (such as Chapters 4, 5) generally an experimental design is chosen such that the initial cell population is in the same state. Therefore we can set the initial conditions for the state occupation probability, $P(0) = (1, 0, 0, \dots)$. This means all cells are in state $k = 1$ at $t = 0$ just before the cell population is perturbed. This allows us to write the closed form solution for the state occupation probability as:

$$P(t; \mathbf{w}) \quad = \quad \exp\left(\mathbf{G}(\mathbf{w})\, t\right) P(0). \tag{3.5}$$

This expression is also used to evaluate the likelihood (eqn. (3.2)) of the model for different parameters.

Model Assumptions

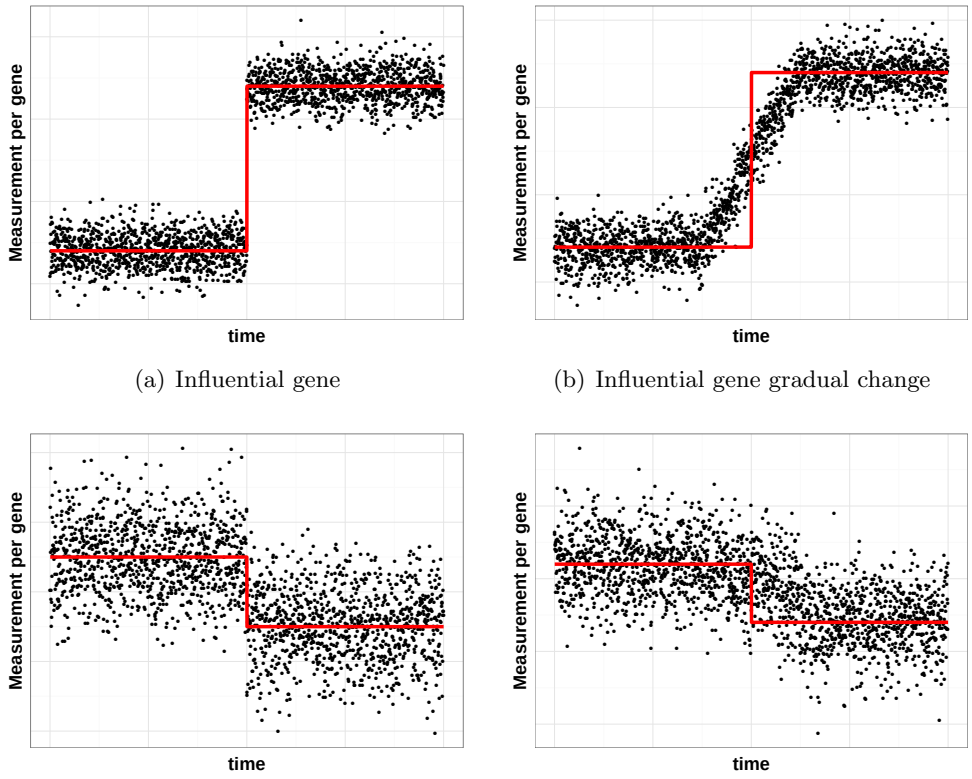
We make a number of assumptions in the above model derivation. Here, we focus on some of the key assumptions made regarding the transition process on a single-cell level and investigate them further. Ensuring an analytically tractable latent

state change model makes these assumptions necessary. In the discussion below we discuss how legitimate these assumptions are, if and how they can be relaxed and how they can be justified.

First, we assume expression of a gene remains constant while it remains inside a given state. The single-cell expression of each gene is modeled by a piecewise flat trajectory where expression changes are instantaneous due to a change in state. It also has the effect that the only time-dependence in the likelihood is due to the state occupation probabilities of the Markov chain. In this simple approximation, interaction between genes are ignored; allowing us to formulate a computationally efficient pipeline to estimate parameters for time courses with many genes, see Section 3.3.3 for further details. It is a very strong assumption and apart from the noise that is prevalent in most biological systems gene expression also changes inside a state due to cell internal mechanisms e.g. the cell-cycle. The slightly more relaxed but sufficient assumption here is: Temporal changes within a state should be much smaller than the difference between biologically distinct states for genes influential in such a transition. This case is illustrated in Figure 3.2(a) and 3.2(b). Therefore this is still a good first approximation in the case of transition processes.

A second assumptions relates to the topology of the Markov chain. To ensure parameter identifiability (see Discussion Section 3.2.2) we have to restrict the latent process to a linear pure birth process. This restricts the topology of the Markov chain quite drastically, but is arguably defensible when applied to externally driven transition processes. The external drive can take many different forms, in the two examples we investigate the system is driven by genetic induction. Of course back transitions are likely, for such cases our model is mis-specified and the forward transitions are only effective values where the back transitions have been absorbed into the model. Consequently estimated forward transitions rates are lower than the real values. On occasion back transitions or topologies of the latent process are of interest. The likelihood eqn. (3.2) is general and does not make any assumptions about the topology of the Markov chain, but additional data or constraints would be required for identifiability of more complex transition topologies. Often the limiting factor is available data hence we focus here on the more useful but special case where only time-course data is available and the latent stochastic process is a linear birth process. In Section 3.4 we include a detailed investigation the impact of breaking this assumption in a simulation.

Finally, we assume rates of cell death and cell duplication cancel each other out and the population therefore remains roughly constant in time. Consequently the fraction of cells in a given state only depends on the transition rates between the



(a) Influential gene

(b) Influential gene gradual change

(c) Gene not influential

(d) Gene not influential gradual change

Figure 3.2: Illustration. First assumption made is that gene expression remains constant for a gene while it remains inside a state. The model Section 3.2.1 describes the single-cell measurement of a gene transitioning between states as an instantaneous step change (red line). In reality the measurement will at least fluctuate and transition won't be instantaneous. For influential genes (a) - (b), this assumptions is reasonable whether or not the transition is instantaneous, the points here are meant to represent single measurements. Genes where within state temporal changes are comparable to between state changes, (c) - (d), the approximation is not good. These genes are not influential for the transition process therefore this is not problematic.

states. Especially in the case of oncogenic transformation (Section 4) this is clearly not the case since tumorous cells in general have a much higher proliferation rate. In Section 3.4 we test how well parameters are estimated when this assumption is violated.

3.2.2 Identifiability

Parameter identifiability is a very important concept in STAMM since parameters represent physical properties of cells. A result for the identifiability of such a model with a discrete time latent process was presented by Clifford [1977]. Unfortunately there does not exist a conclusive results on identifiability for latent stochastic model with a continuous time latent process. To establish if STAMM is identifiable analytically is highly non-trivial therefore we perform tests for empirical identifiability using single-cell simulations in Section 3.5.1.

3.3 Estimation

3.3.1 Parameter estimation

We begin by stating the maximum likelihood estimates (MLEs) based on the likelihood eqn. (3.2):

$$(\{\hat{\beta}_j\}, \hat{\mathbf{w}}) = \underset{\{\beta_j\},\mathbf{w}}{\operatorname{argmin}} \sum_{j=1}^{p} \sum_{t=1}^{T} \|g(x_j(t)) - g\left(P(t;\mathbf{w})\,\beta_j\right)\|_2^2, \tag{3.6}$$

where $\|\cdot\|_q$ denotes the ℓ_q norm with respect to its argument. The transformation g is in general non-linear as discussed above (e.g. log in microarrays or arcsinh in RNA-seq); for such transformations the MLE (3.6) cannot be obtained in closed form. Genome wide measurements yield readings with number of genes, p, of up to 10^4. Directly optimising eqn. (3.6) is not practical for a problem with large p as the parameter space rapidly increases. We adopt a two-step estimation procedure proposed by Armond et al. [2013]. The first step is based on the observation that many genes have similarities in their measured time-courses; this allows us to cluster genes obtaining m clusters describing typical temporal patterns. Details for choosing the parameter m are discussed in Section 3.3.3. The m cluster centroids are used to estimate the transition rates $\mathbf{w}$ via eqn. (3.6), instead of all genes. This approach reduces computation time significantly when $m << p$. The transition rates estimated using cluster centroids, $\hat{\mathbf{w}}$, are fixed and the β values for all remaining genes

are estimated. The MLE is now written as:

$$\hat{\beta}_j = \underset{\beta_j}{\operatorname{argmin}} \sum_{t=1}^{T} \|g(x_j(t)) - g\left(P(t; \hat{\mathbf{w}})\, \beta_j\right)\|_2^2 + \lambda \|\beta_j\|_1 \tag{3.7}$$

where the final term is an (optional) ℓ_1 penalty with tuning parameter λ. It is invoked when potential over-fitting needs to be counteracted (choice of λ is discussed in Section 3.3.3).

The optimisation eqn. (3.7) greatly simplifies estimation (compared to eqn. (3.6)), since estimation for individual β_j for gene j can be performed independently. This is possible because time-courses between individual genes are only coupled by transition rates $\mathbf{w}$; once they are fixed, individual gene trajectories can be examined independently.

3.3.2 Model selection

The estimation steps described in Section 3.3.1 apply to a model with a fixed number of states K. Here we present a procedure to determine the number of states K that best represent a data set under investigation. Depending on the application K itself can be of scientific interest. In general estimated state-specific expression signatures, β, are influenced by the number of states. Underestimating number of states results in distinct states being merged. Overestimating the number of states introduces artificial states in the transformation. Both scenarios lead to poor estimation of parameters.

In general model selection can be performed using a form of cross-validation (CV) by leaving out part of the data as a validation set. In some cases it is also possible to use BIC e.g. when the number of genes is small, but this quickly breaks down as p increases as can be seen in eqn. (2.13) as the second term would completely dominate the BIC term; therefore we restrict ourselves to CV. In applications to time-series, cross-validation is often non-trivial due to discrete and irregularly spaced observations. The STAMM model has an underlying continuous-time latent process, which allows for prediction of any time points from estimated parameters; therefore comparison between predicted time-points, from estimated parameters, and the corresponding held-out data can be performed without any problems. In this application due to poor time resolution it is often not possible to include more than one time point in the validation data; this variant is called leave-one-out cross-validation (LOOCV). If t is the held-out time point let the estimated parameters for the remaining subset be rates $\hat{\mathbf{w}}^{-t}$, state specific expression $\{\hat{\beta}_j^{-t}\}$ and gene-specific

standard deviation $\{\hat{\sigma}_j^{-t}\}$. State occupation probabilities at the held-out time point $P(t; \hat{\mathbf{w}}^{(-t)})$ are obtained by solving the master equation using estimates derived from the training data. There we can now write a prediction for the expression of gene j at the held-out time point $\hat{x}_j^{\mathrm{CV}}(t) = P(t; \hat{\mathbf{w}}^{(-t)})\, \hat{\beta}_j^{(-t)}$ and the cross-validation mean squared error ($\mathrm{MSE}_{\mathrm{CV}}$) is simply

$$\mathrm{MSE}_{\mathrm{CV}} = \sum_{t=2}^{T} \sum_{j=1}^{p} \frac{1}{\hat{\sigma}_j^{(-t)}} \big(g(\hat{x}_j^{\mathrm{CV}}(t)) - g(x_j(t))\big)^2. \tag{3.8}$$

The strength of this type of model selection in comparison to the Bayesian approach presented in Armond et al. [2013] is twofold. Firstly application of the computationally efficient estimation procedure outlined in Section 3.3.1, allows this cross-validation procedure to be applied to the whole data set efficiently. Secondly it doesn't require parameters to be set by the user except those required for estimation. The Bayesian approach requires a computationally demanding Monte Carlo estimation and has several hyper-parameters which have to be set by the user.

3.3.3 Estimation pipeline

We now present a computationally efficient pipeline for setting tuning parameters required for estimation. The pipeline is also summarised in Figure 3.3. The required tuning parameters are:

- The number of clusters, m, used in the first step of the two-step estimation.
- The strength of the penalty term, λ, applied in eqn. (3.7), where $\lambda = 0$ is equivalent to no penalty.
- The number of states in the latent Markov chain, K.

Number of cluster: In the initial estimation step we cluster gene expression trajectories which results in cluster centroids describing typical trajectories; these permit estimation of transition rates. In empirical results (see Section 3.5.2) we see; if the number of clusters is large enough to capture most of the information in typical trajectories changing m does not have a significant impact on parameter estimation. Therefore we set m using a simple k-means algorithm and inspecting the relative decrease of within-cluster sum of squares objective $J(m)$ as a function of m:

$$\Delta J(m) = \frac{J(m-1) - J(m)}{J(m-1)}$$

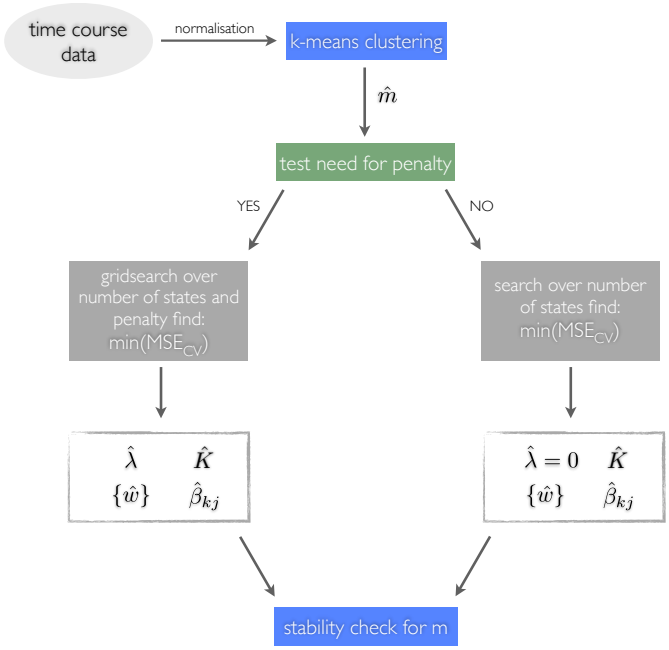


Figure 3.3: Schematic of estimation pipeline. Clustering of normalised gene trajectories provides an optimum number of clusters $\hat{m}$, the centroids of the clusters summarise typical trajectories contained in data. Fitting the model to centroids we can find transition rates w. Stability of estimate test is performed to determine the need for penalisation of model parameters with strength λ. State-specific gene expression signatures β_{kj} are estimated by varying the number of states in the model K and λ (if applicable). Optimum number of states are determined by performing a form of cross-validation. Finally sensitivity of estimation to change in $\hat{m}$ is performed. If necessary the pipeline is rerun with a better choice of $\hat{m}$.

We select $\hat{m}$ such that $\hat{m} = \min\{m : \Delta J(m-1) < 0.1\}$, i.e. if the relative decrease in the objective function is smaller than 0.1 for $m-1$ we choose m as the number of clusters. Small fluctuations in the objective function for higher m lead to instabilities in the relative decrease. Once $\hat{m}$ is set we include a post-estimation sensitivity test for the choice of m. In section 3.5.2 we demonstrate with the help of empirical results, how well behaved and computationally efficient this model is. Though it should be noted that the choice of m can be made with any clustering method and the corresponding objective function.

Penalisation: The penalty term introduced in eqn. (3.7) is useful in high dimensional models; even though the data investigated using this model will often be high dimensional, estimation is carried out separately for each gene. Therefore penalisation may not be required unless the number of time points is large. Consequently we introduce an additional step to test the need for penalisation by comparing estimated expression signatures β with estimates obtained from leaving out individual time points. We specify stability as a Pearson correlation between estimated β values greater than 0.8. If we deem a data set stable under such a test there is no need for penalisation and we choose $\lambda = 0$. If the correlation is smaller than 0.8 a penalty term is required (setting the penalty strength is discussed below). In both the simulated data sets and application to oncogenic transformation penalisation was not required and $\lambda = 0$ in Sections 3.5 and 4.

Number of states: The final parameter to be set is the number of states in the latent Markov chain. This parameter is set by minimising the CV score (MSE_{CV}) for a range of different K. If penalisation is required MSE_{CV} is minimised by performing a grid search over both λ and K.

All three parameters (m, λ, K) can in principle be set by performing a grid search with respect to MSE_{CV}, but this is computationally very challenging and would make estimation impractical. Our pipeline make use of heuristic observations to reduce the grid search to one dimension. The observation that estimates are robust to changing the number of clusters used in the first step, allows us to remove m from the grid search. Observing that the penalty term is not always needed enables us to exclude λ from the grid search. When choosing m using a clustering method it could be that transition rates haven't converged. Therefore we carry out an additional diagnostic post estimation. The issue we are trying to address is that an increase in m corresponds to an increase in information; this can lead to changes in estimated parameters. If the choice of parameters is appropriate increasing the

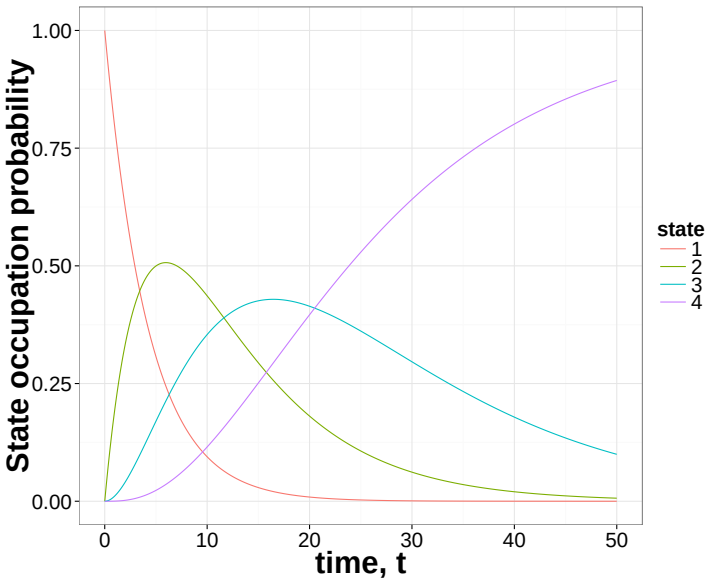


Figure 3.4: Simulation study. State occupation probabilities for a four state model with transition rates $[1/5, 1/8, 1/15]$. The transition rates are set to these values for all simulations, the values are chosen to allow $k = 4$ to have a high occupation probability at the final time-point of the simulation.

number of clusters should not significantly impact estimated parameters. To this end we compute correlation values between expression levels β estimated using $\hat{m}$ clusters and estimated using larger values $m' > \hat{m}$. Provided results have Pearson correlation above 0.8, the choice of $\hat{m}$ was appropriate, if the correlation is below 0.8 we repeat the pipeline with larger m.

3.4 Simulation setup

To test the validity of the model we need to test it with simulated data where true parameters are known. This will allow both evaluation of strengths and weaknesses in parameter estimation and model selection (choosing number of states). Simulations are performed not at the cell population level of the likelihood eqn. (3.2) but at the single-cell level; allowing for extensive testing of model assumptions. The single-cell trajectories are then averaged to obtain homogeneity data analogous to RNA-seq data.

Here we describe the step by step simulation procedure for a K state model independent of the number of genes simulated:

State transitions. When setting transition rates between discrete states of the Markov chain we need to keep a few things in mind. Firstly the smallest sampled (observed in real data) time step needs to be smaller than the transition rates. Just like in typical experimental designs for transition processes. Additionally the model won't be able to extract information about a process taking place on a time-scale smaller than gaps between observations. Secondly we are considering transitions processes driven towards an established final state (e.g. oncogenic transformation, pluripotency); so to mimic this behaviour in simulated data we need to insure the occupation probability for the final state is higher than the others at the final time point. Of course in realistic experiments even at the final time point the cell population will still be heterogeneous In the discussion that follows in Section 3.5 we use the three transition rates $[1/5, 1/8, 1/15]$ for four state model. In Figure 3.4 we show the state occupation probabilities for these parameters and $k = 4$ has an occupation probability of ≈ 0.89 at the final time point. For every cell in the simulation, state transitions are simulated by drawing jump-times from exponential distribution with parameters given by transition rates as defined for a continuous-time Markov chain.

State-specific expression levels. For all cells, each gene j and state k we set gene expression levels β_{kj}; per gene the expression levels are set to zero with probability $1/K$ otherwise they are sampled uniformly from $(0, \gamma_j]$. Parameter γ_j, chosen from the range $[1, \ldots, 12000]$, effectively sets the scale of gene j [1]. This method ensures simulated trajectories for genes on different scales (see Figure 3.5(a) and the corresponding gene expression signatures Figure 3.5(b)), to emulate real RNA-seq data where a range of five order of magnitude was observed [Wang et al., 2009; Mortazavi et al., 2008]. Gene expression trajectories for single-cells are piecewise flat for each gene once β values are sampled. Changes in trajectories only occur at jumps between states and are instantaneous.

Aggregation and time-sampling. For each gene j each cell has an associated gene expression trajectory. Similar to RNA-seq experiments where observations are averages of gene expressions over many cells; these trajectories are averaged over a large number of cells to give an average gene expression trajectory. The occupation probability in the model outlined in Section 3.2.1 is derived in the limit of number of cells $n \to \infty$, of course in practice the number of cells is finite. We set the number of cells to 1000 which serves as a good test of the limiting assumption.

The simulated time-course is obtained by sampling the simulated trajectories

[1] It is always chosen from the following $\gamma_j = \{1, 10, 50, 100, 200, 500, 1000, 2000, 4000, 7000, 10000, 12000\}$

at discrete unevenly distributed time points. Finally Gaussian noise is added to the transformed data (see Section 3.2.1 for details, for RNA-seq arcsinh) with mean zero and standard deviation σ; which we set to $\sigma = 0.2$ unless states otherwise, this provides a reasonable signal to noise ratio for all observations. Similar to the RNA-seq data discussed in Section 4 we choose 15 unevenly spaced time points at $t = \{0, 2, 4, 7, 8, 11, 14, 20, 24, 29, 32, 35, 40, 44, 48\}$. The simulation setup is summarised in pseudocode in Algorithm 1.

Algorithm 1 Pseudocode for single-cell simulations

procedure SIMULATION($n.states, n.genes, n.cells, r, p, \tau, dt$)
  $\beta \leftarrow NB(r; p)$
  $jump.t \leftarrow Exp(1/\tau_k)$
  for all $genes, cells$ **do**
    for $t \leftarrow 0, T$ **do**
      while $t < jump.t_{states}$ **do**
        $sim.traject(t) \leftarrow \beta_{j,k}$
      end while
    end for
  end for
  average $sim.traject$ per gene for all cells
  $sim.data \leftarrow sim.traject$ (sampled at discrete time)
  $sim.data \leftarrow sim.data + \mathcal{N}(0, \sigma)$
end procedure

3.5 Simulation results

We present results from simulations in two separate phases. For both simulation setup the transition rates are fixed at $[1/5, 1/8, 1/15]$ and we simulate from a model with four states in latent Markov chain. First in a small scale simulation with $p = 9$ genes we perform multiple rounds of direct estimation of the whole data without the need for the initial clustering step. This simpler simulations allows investigation of identifiability and an investigation for model selection without considering the two-step estimation procedure outlined above.

Than we consider a larger scale simulation with $p = 120$ genes, where we put the full two-step estimation procedure to the test; including clustering, setting of tuning parameters and finally model selection.

3.5.1 Small scale simulation

Using the small scale simulation we perform three separate tests. One in which we only estimate transition rates and state-specific expression levels; we consider the number of states to be known. Then we consider the model selection problem and finally we investigate estimation under breaking model assumptions.

Number of states known

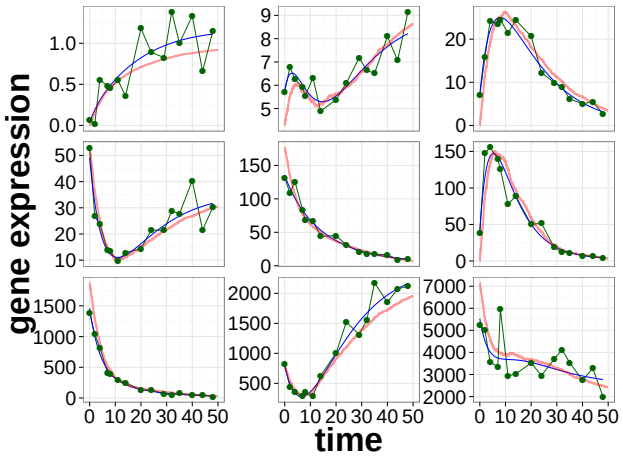
We simulated 9 genes from a 4 state model as described in Section 3.4. In this small simulation we do not use a penalisation term, i.e. $\lambda = 0$. In Figure 3.5(a) we show trajectories for one such realisation, here the thicker line represents trajectories from averaging 1000 cells for each gene. The green dots show sampled data with the addition of Gaussian noise to transformed data. In Figure 3.5(b) we show state-specific gene expression signatures for all 9 simulations. The values are shown in pairs of true and estimated. The value on the right is in each case the true value used in simulating the trajectories. The left-hand value is estimated by fitting the 15 time point of the simulation. The corresponding estimated and true transition rates for this realisation can be seen in Table 3.1. Using the estimated transition rates and the expression signatures we can obtain an estimated trajectory, seen as a blue line in Figure 3.5(a).

Transition rates	$w_{1,2}$	$w_{2,3}$	$w_{3,4}$
true mean	0.200	0.125	0.067
estimated	0.236	0.114	0.068

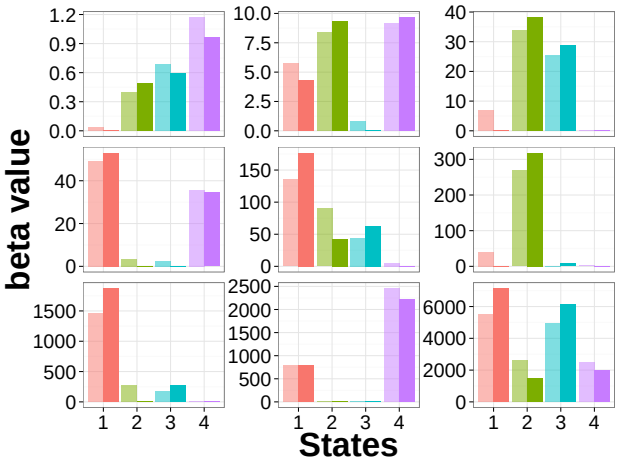
Table 3.1: Transition rates used in the simulation and the estimated values

We repeat fifty such independent simulations at four different noise levels [2], each time β_{kj} are resampled as described above (Section 3.4) while transition rates are shared across simulations. We compute the correlation between estimated and true gene expression signatures for each simulation run $\rho(\beta, \hat{\beta})$. The correlation coefficients for all simulations are summarised in a boxplot, Figure 3.6(a). For all tested noise levels we compute a mean and standard deviation of the correlation coefficient across all fifty runs in Table 3.2; The mean is above 0.9 for all simulations and the highest level for the variance is 0.13. Therefore we can conclude that state-specific gene expression signatures are recovered well in the simulation. We also introduce a new measure, $s_k = |\hat{w}_{k,k+1} - w_{k,k+1}|$ to test recovery of transition rates. For each simulation we use the mean $\bar{s}$ over the three transition rates as measure for

[2] $\sigma = \{0.05, 0.1, 0.15, 0.2\}$d

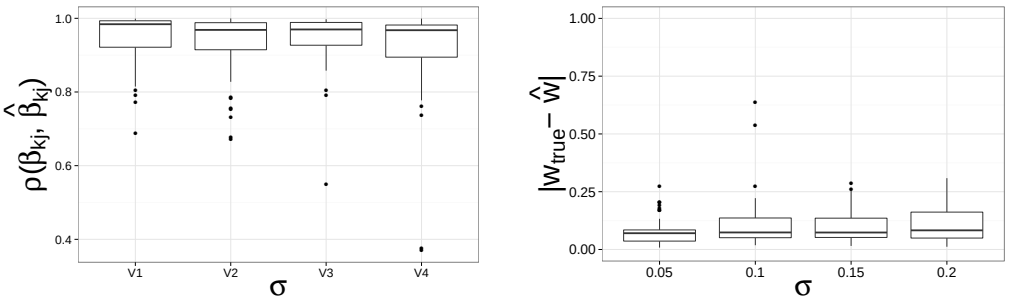


(a) Simulated Trajectories for $p = 9$



(b) Expression signatures for $p = 9$ simulation

Figure 3.5: Simulation study. Small scale simulation for $p = 9$ genes. (a) shows the trajectories for these simulations. The thick red line shows the averaged trajectories over 1000 cells. The green dots show 15 sampled data points with normal noise ($\mathcal{N}(0, \sigma)$, with $\sigma = 0.2$) added to the average data. The blue thin line shows the trajectory from estimated parameters. (b) shows state-specific gene expression signatures for all 9 simulated genes. The true and estimated parameter values are shown next to each other. The lighter colour on the left shows estimated parameter values, the solid colours shows true parameter values.



(a) State-specific gene expression

(b) Transition rates

Figure 3.6: Simulation study. Small scale simulation using $p = 9$ genes with 50 independent repeats. Boxplots show results over all repeated simulations at four different noise level $\sigma = \{0.05, 0.1, 0.15, 0.2\}$. (a) Boxplots for correlations between estimated and true gene expression signatures ($\rho(\beta_{\text{true}}, \hat{\beta})$) at four different noise levels. (b) Boxplots for the mean of absolute differences between the estimated and true transition rates $\bar{s}$ for each simulation at four different noise levels.

how well transition rates are recovered. In Figure 3.6(b) we show boxplots for the fifty simulations for each of the four noise levels. We find that transition rates are also recovered well, though as expected the estimates become worse with increasing noise levels.

σ	0.05	0.1	0.15	0.2
mean	0.95	0.93	0.94	0.91
std. dev.	0.07	0.09	0.08	0.13

Table 3.2: Correlation between true and estimated gene expression signatures. Mean and standard deviation are estimated across 50 independent simulations.

Determine number of states

Next we consider the problem of model selection in this small simulation setup. We simulate data as described above for $p = 9$ genes. In such a model with a latent stochastic process, model selection is a challenging problem especially using noisy data sets. Therefore to test model selection we fifty independent simulations for each of the following noise regimes: $\sigma = \{0.05, 0.1, 0.15, 0.2\}$. We compare models with $K = 1 \ldots 5$ and perform model selection using leave-one-out cross-validation (see Section 3.3.2), for each of the fifty simulations. We determine the minimal MSE_{CV} scores (eqn. (3.8)) for different models and juxtapose a comparison between the

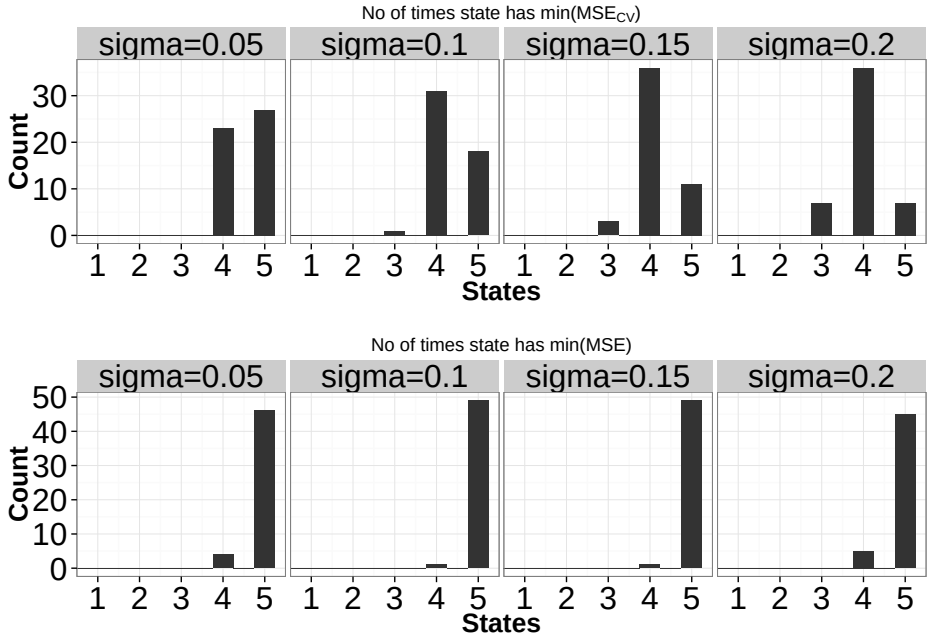


Figure 3.7: Simulation study. We perform fifty independent small scale simulation with $p = 9$ at four different noise levels $\sigma = \{0.05, 0.1, 0.15, 0.2\}$ for $K = 4$. We perform model selection using a form of cross-validation and find the state K that minimises the MSE_{CV} score, top row. As a comparison we also show fit to data and find the state K that minimises MSE. We find that at even at very high noise levels using MSE_{CV} it is possible to identify the right model. An example trajectory and parameters can be seen in Figure 3.5.

different models using a simple normalised MSE score for model fit without held-out time points. In each simulation and for each noise regime we determine the model with lowest MSE_{CV} score and lowest MSE score. Then we show the distribution of these minimal scores over the selected number of states in Figure 3.7; the top row shows the distribution for MSE_{CV} and bottom row show the distribution for MSE in different noise regimes.

Here number of parameters increase with number of states, and as a result model fit improves; therefore as expected at all noise levels the maximum number of states ($K = 5$) results in the best fit.

Violating model assumptions

Until now we have considered simulations with a correctly specified model where assumptions underlying the model are not violated. Breaking these assumptions is

especially easy in the single-cell simulation. We investigate consequences on parameter estimation under violation of a subset of these assumptions. We use three types of plots to investigate parameter estimation for these simulations.

- **Correlation.** For state specific gene expression signatures β_{kj} we compute the Pearson correlation coefficient between true parameters and estimated parameters, $\rho(\beta_{\text{true}}, \hat{\beta})$.
- **Transition times.** We show boxplots of estimated average transition times for 10 simulations and a horizontal dashed line to represent the true value used in the forward simulation.
- **Probability** In the model itself the transition times do not enter directly they are used to calculate probabilities. We compare the values by calculating a mean difference between probabilities:

$$\langle |\hat{p}_k(t) - p_k(t)| \rangle_{k,t} \,, \tag{3.9}$$

  where k is the number of states, $\hat{p}_k(t)$ is the probability calculated from estimated parameters and $p_k(t)$ is the probability calculated from true values. The average is taken over both the states and time.

Cell death and cell doubling An assumption implicit in STAMM is that cell death and cell duplication happens at a constant rate across all states in the transformation process. This is of course not the case in the discussed example of oncogenic transformation, since transformed tumorous cells have a much higher proliferation rate than the initially healthy cells. In the single-cell simulation setup we sample a time of death, t_i^d, and a time for cell doubling, t_i^{dup}, from an exponential distribution. If sampled rates for a cell are outside of the time range of the simulation, the cell remains unchanged. If they are both in the range there are two possible scenarios. Firstly if the death rate is the smaller of the two, cell i is taken out of the simulation $t > t_i^d$. Secondly if $t_i^{dup} < t_i^d$, cell i is taken out of the simulation at $t > t_i^{dup}$ and two new cells are simulated with new sampled state transitions. The simulation and estimate is performed 10 times. Investigating the oncogenic transformation discussed in the paper it was observed that generally cells have a doubling rate of close to 0.05 i.e. doubling of cells is roughly every 20 hours. In Figure 3.8 we fix the doubling rate and since cells in this experiment rarely die, we choose very small death rates. The left panel shows the average as a dark line and the shaded area represents the standard deviation for the 10 repeated simulations. The middle

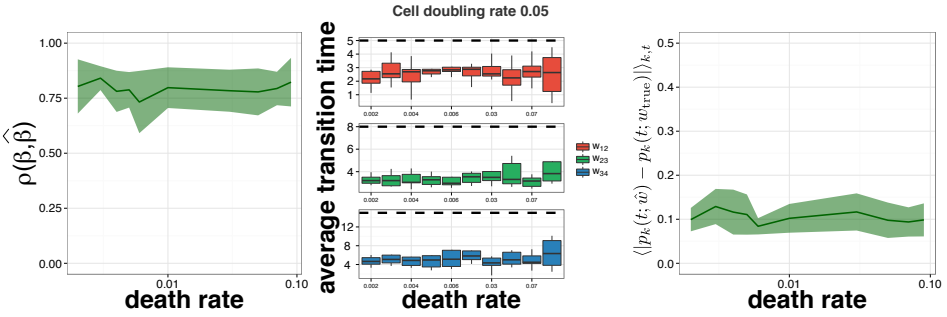


Figure 3.8: Simulation study. Testing assumption about cell death and cell doubling. For each cell a time of death, t_i^d, and a time for cell doubling, t_i^{dup}, is sampled from an exponential distribution with varying average rates. If the sampled rates for a cell are outside of the time range of the simulation, the cell remains unchanged. If they are both in the range there are two options. The first option is that the death rate is the smaller of the two in that case the cell i is taken out of the simulation $t > t_i^d$. If $t_i^{dup} < t_i^d$, cell i is taken out of the simulation at $t > t_i^{dup}$ and two new cells are simulated with new state transitions. The simulation and fit is performed 10 times. In experiments we observe cell doubling time to be roughly 18 hours and very few dead cells. Therefore the simulation with a cell doubling rate of 0.05 and a variety of death rates. The left panel shows the average as a dark line and the shaded area represents the standard deviation for the 10 repeated simulations. The middle panel shows box plots for the estimated average transition times. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

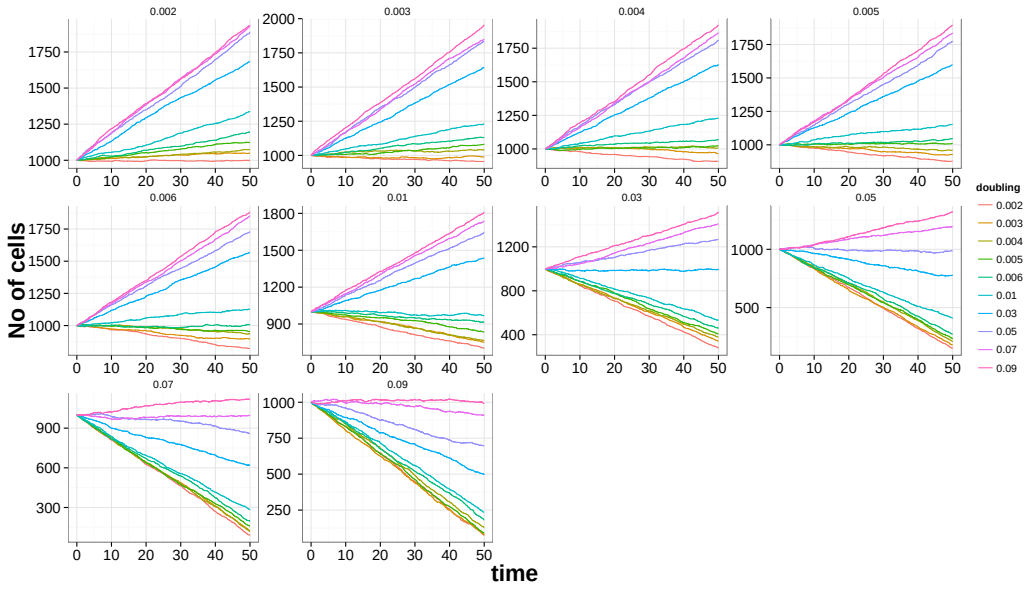


Figure 3.9: Simulation study. Testing assumptions about cell death and cell doubling. Plots show number of cells at different time during the simulation for one of the 10 simulations. Each panel represents different death rates and each colour different doubling rates.

panel shows boxplots for the estimated average transition times. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

In Figure A.1 we include additional cell doubling rates. In general gene expression signatures are estimated well, but the transition rates are not. During estimation transition rates only enter as probabilities hence badly estimated transition rates don't have a significant negative effect on the estimation of expression signatures. To see what effect the different parameters have on the number of cells simulated at any given time Figure 3.9 show the number of cells as a function of time for different cell death and cell doubling rates.

Back transitions The second assumption we test is the inclusion of back transitions in the single-cell. We simulate trajectories with back transition from $k = 4$ to $k = 3$; they are sampled from exponential distributions with different means. In Figure 3.10 we show comparisons between estimated and true values of parameters as a function of the average back transition time from state $k =$. In the left panel we plot the average correlation for 10 independent runs, between true and estimated β_{kj} parameters as a solid line. The shaded area shows the standard deviation. The verti-

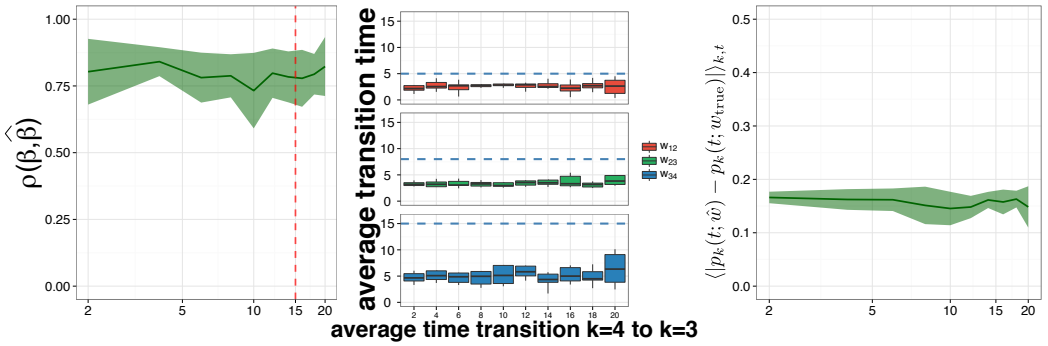


Figure 3.10: Simulation study. To test the affect of back transition on the estimation, we simulate trajectories with back transition at $k = 4$ with different transition times. In the left panel we plot the average correlation for 10 independent runs, between true and estimated β_{kj} parameters for different average time for the back transition as a solid line. The shaded area shows the standard deviation. The vertical dashed red line shows the average forward transition time for $k = 3$. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

cal dashed red line shows the average forward transition time for $k = 3$. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

Markovian assumption Finally in this section we investigate the latent Markov process and consider a case where jumps are non-Markovian. We want to consider the more realistic case that the transition time is fat tailed; therefore we choose a truncated Student t-distribution with a variety since transition rates are positive. We sample using the *tmvtnorm* package in R with degrees of freedom, *df*, as the varied parameter. We perform the simulation as before, but sample transition rates from the t-distribution with means $(1/5, 1/8, 1/15)$ and consider a range of *df* parameters in *tmvtnorm*. The results are shown in Figure 3.11; the left panel shows the mean correlation between true and estimated β_{kj} as a solid line and the shaded area constrains the standard deviation. The middle panel shows boxplots for the average transition rate estimated from the model for a system with $K = 4$ and only

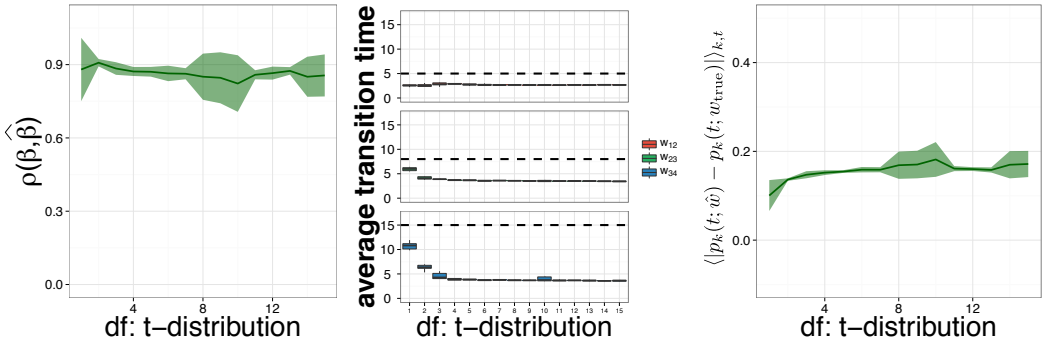
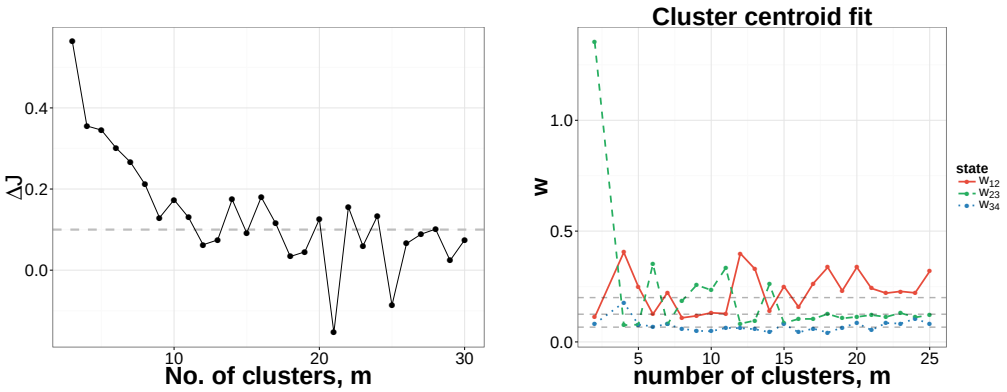


Figure 3.11: Simulation study. We simulate from a non-Markovian system, one where average transition rates are heavy tailed. Here we sample from a truncated Student t-distribution using the *tmvtnorm* package in R. It is truncated at zero since transition rates are always positive. The transition rates are sampled to have means $(1/5, 1/8, 1/15)$ and we vary the degrees of freedom *df* parameter in the package used. In the left panel we show the average correlation between true and estimated β_{kj}, the mean is shown as a solid line and the standard deviation as a shaded area. The middle panel shows boxplots for the for the average transition time estimated from the model for a system with $K = 4$ states. The right panel shows the mean differences between estimated and true state occupation probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.



(a) k-means clustering

(b) transition rates

Figure 3.12: Large scale simulation study. Results from clustering. (a) Initial step in the estimation pipeline is use k-means clustering for $p = 120$ genes. The plot shows relative change in the k-means objective function as a function of the number of cluster: $\Delta J(m) = 1 - J(m)/J(m-1)$. We choose the optimum number of clusters $\hat{m}$ such that $\Delta J(m-1) < 0.1$; here $\hat{m} = 13$. For larger m, $J(m)$ is small therefore relative changes have large fluctuations. (b) Estimated transition rates as a function of the number of cluster. Horizontal dashed lines show true values. After large initial fluctuations the transition rates fluctuate around the true value.

forward transitions. The right panel shows the mean differences between estimated and true probabilities averaged over 10 trajectories as a solid line and the shaded area represents the standard deviation.

3.5.2 Large scale simulation

Lastly we want to check how well the two-step estimation pipeline outlined in Section 3.3.3 works applied to simulated data. We simulate $p = 120$ genes as described above with $K = 4$ states. To mirror real data where genes are on different scales, we sample 12 scale parameters τ. To get to $p = 120$ genes we sample from all scale parameter 10 sets of β values. All other parameters are as set out in Section 3.4. We follow the procedure set out above and start by clustering simulated trajectories into m clusters. Then we estimate transition rates $\mathbf{w}$ from cluster centroids and keep them fixed for the second step. Next we use these transition rates and estimate expression signatures β_j independently for each gene j.

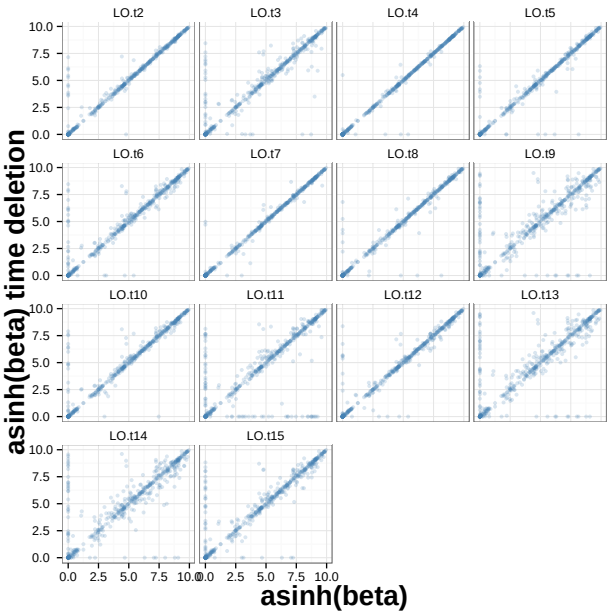
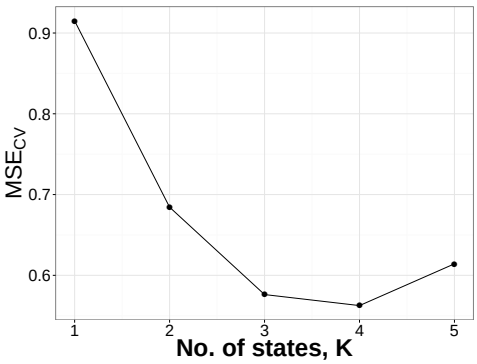


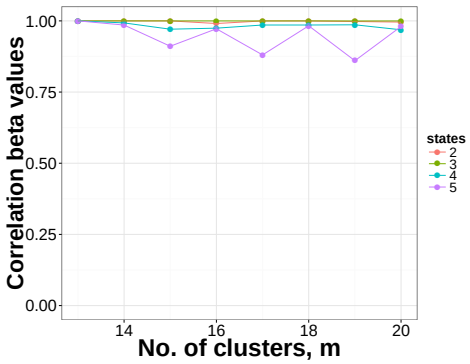
Figure 3.13: Large scale simulation. Stability of estimated expression signatures under time point deletion to determine need for ℓ_1 penalization. We conclude that estimated parameters are stable with a Pearson correlation > 0.8, therefore there is no need for a penalty.

3.5.3 Number of states

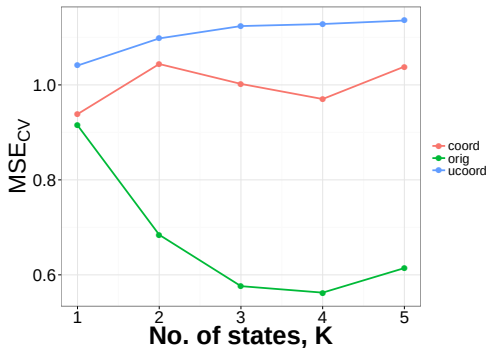
Applying the first step we cluster the simulated data using a k-means clustering algorithm. We use the relative decrease in the objective function $\Delta J(m)$ to determine the number of clusters and vary m in the range $[2, 30]$, see Figure 3.12. The relative decrease is smaller than 0.1 for $m = 12$ therefore we choose $\hat{m} = 13$. Note that for larger m the objective function $J(m)$ is small and we observe that $\Delta J(m)$ has large fluctuations due to slight deviations in the objective function. Then we test if for this set of data penalisation is necessary using stability of estimated gene expression signatures under deletion of time points. Figure 3.13 shows that for all deletions estimated parameters are stable, therefore we conclude that there is no need for a penalty term. Then we perform a model selection step to determine the number of states K of the latent Markov chain. We compute the MSE_{CV} score for $K = \{2, \ldots, 5\}$ states, see Figure 3.14(a); we see a clear minimum for $K = 4$ which is also the correct number of states. In the final step of the pipeline we perform a post-estimation stability test to ensure the number of clusters chosen is not too small (see Section 3.3.3). We compute the correlation for expression parameters estimated with increasing number of clusters, see Figure 3.14(b). We carry out this



(a) Number of states



(b) Stability number of clusters



(c) Permuted time points

Figure 3.14: Large scale simulation. Simulation performed for $p = 150$ genes, 15 time points and $K = 4$. Figures summarise results obtained from applying the estimation pipeline (see Section 3.3.3) to the simulated data. (a) MSE_{CV} score to determine optimal number of states in the model, the minimum is at $K = 4$. (b) Test to determine stability under perturbation of number of clusters $\hat{m}$. Estimation is stable therefore the choice of parameters is reasonable. (c) To determine dependence on structure of data we perform a permutation test, first by permuting time-points of all trajectories in a coordinated fashion and then by permuting trajectories independently. The MSE_{CV} score for permuted data is significantly higher for all K other than $K = 1$ and data permuted in a coordinated fashion perform better than independently permuted data.

test for all models $K = \{2 \dots 5\}$ and the estimated parameters are stable therefore we conclude that the choice of $\hat{m}$ was a reasonable choice.

A question that arises from these results is to what extent they are indicative of estimation or if states are only estimated because the model assumes there are states in data. One good way of addressing this question is to permute data and compare MSE_{CV} estimated for permuted data and the original data. Here we distinguish between two ways of permuting data: first we perform a coordinated permutation where all simulated trajectories are permuted in the same way. Then we perform a permutation for each trajectory independently. In Figure 3.14(c) we show the results and we can see that with both types of permutations the MSE_{CV} values are significantly larger than the original data set.

Estimated parameters

One of the strengths of using simulated data is that we can compare estimated and true parameters. Figure 3.15 shows a scatter plot of estimated β parameters against their true values used in the simulation. The parameters are in general well estimated with a Pearson correlation of 0.95; as the plot (Figure 3.15) shows in certain cases, despite the large correlation, the true value of β is exactly zero but estimates are non-zero.

The first clustering step we take is crucial and has a considerable affect on parameter estimation of transition rates. Hence it is valuable to delve a bit deeper into the first estimation step and its sensitivity to the number of clusters. Figure 3.14(b) shows that estimates expression signatures are very stable when increasing number of clusters. Figure 3.12(b) shows the three transition rates for a model with $K = 4$ as a function of m. The horizontal dashed lines in the plot show true transition rates. We observe that estimated transition rates strongly fluctuate for small m and with increasing number of clusters amplitude of fluctuations decrease.

3.6 Discussion

We present an extension to previous work with a detailed description of the model including its assumptions, an unbiased estimation pipeline and simulation based tests to investigate STAMM recently proposed in Armond et al. [2013]. To address concerns about practical application we have made computation more efficient including the use of parallelisation which due the two-step estimation process allows for much higher efficiency; the exploration of model behaviour under violation of underlying assumptions; and an extension to allow for application to RNA-seq data

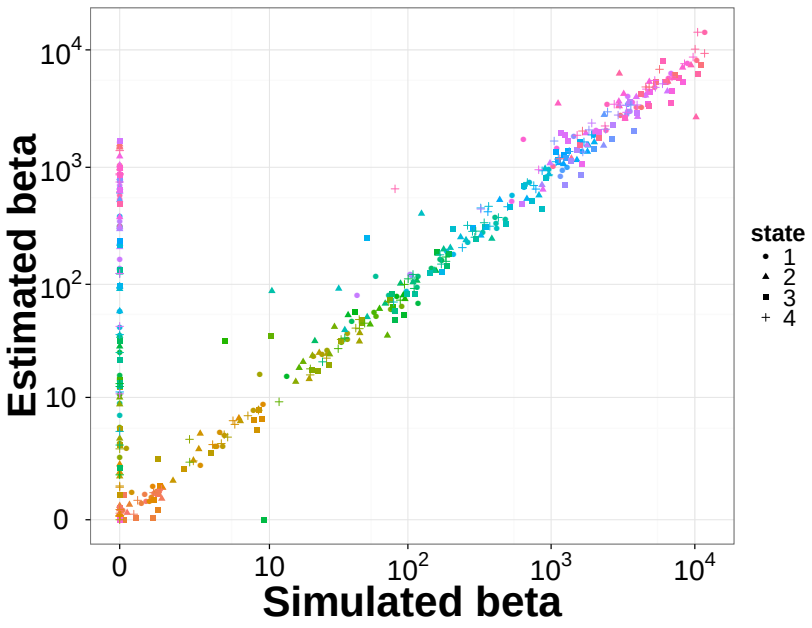


Figure 3.15: Large scale simulation. Estimated gene expression signatures scattered against their true values.

though more on that in Chapter 4.

Establishing and investigating identifiability in a latent stochastic model aggregate over single-cells is highly non-trivial. Some relevant literature exists [Kalbfleisch et al., 1983; Kalbfleisch & Lawless, 1984, 1985], but it is applied to panel data or the latent stochastic model is in discrete time. Identifiability results that can directly be applied to STAMM do not exist yet. Therefore we present empirical results using single-cell simulations and show that true parameters used during simulation can be identified. However it is important to understand the explore restrictions on Markov chain topology. Especially if there is potential for allowing more complex topology (e.g. including branches or back transitions) by including additional single-cell data. This would allow for precise experimental design in applications to ensure more details can be determined about transformation processes.

To allow for efficient estimation and to ensure identifiability of model parameters a number of simplifying assumptions are made in STAMM. Of course these assumptions do not hold in real biological systems. To investigate capabilities and limits of STAMM we simulated data by breaking these assumptions in turn. In summary, we find the model relatively robust under violations of underlying assump-

tions, but especially transition rates under strong violations of the assumptions are estimated badly. This leads to the conclusion that even though state-specific gene expression signatures are often well estimated they should serve only as hypothesis to be confirmed by experiment.

STAMM is versatile in its application as it can be used for a broad range of transition processes and data types. Two data types we investigate in this thesis are microarray data and RNA-seq data but it can also applied to transcript, protein and epigenetic time course assays. New development of cheaper bulk assays means it is a feasible first step to study a transforming biological system. STAMM could be used to identify a subset of genes important in transformation; to pinpoint cell surface markers that distinguish states facilitating a single-cell separation; or discovery of transition to allow for targeted single-cell experiments. Using multiple types of data in iterative stages would allow for a step by step improvement in information gathered about the system. For instance once transition rates are known estimation of expression signatures is much more precise and vice versa. In future it would also be interesting to relax some of the assumptions made and extend the model to allow for back transitions using experimentally verified forward transition rates.

Chapter 4

Oncogenic Transformation

4.1 Introduction

There are a variety of possible applications for the model outlined in Chapter 3; examples include stem cell reprogramming [Armond et al., 2013] (contributions to which are discussed in Section 5) and estrogen response of breast cancer cell lines [Casale et al., 2013]. Another example is oncogenic transformation which is the transition from healthy cells to cancer cells which we discuss in this Chapter. We consider a derivative of the human epithelial MCF10A cell line where the v-Src and estrogen receptor (ER) fusion is integrated; the new cell line is called MCF10A-Er-Src [Hirsch et al., 2010] (for brevity in the discussion that follows we will refere to these as MCF10A). The Src oncogene is activated by addition of tamoxifen resulting in a rapid transformation of this system. Morphological changes on a cellular level are observed as early as $t = 24h$ (between $t = 24h - 36h$) they show the ability to form colonies in soft agar in the final transformed state [Hirsch et al., 2010] . Figure 4.1 shows images taken of one realisation of the experiment using a camera attached to a microscope; they are taken at $t = \{0, 24, 48\}$ hours at two different magnification levels ($10x$ and $20x$ as indicated on the figure). The top rows shows images taken after induction of tamoxifen and morphological changes in the cells can be observed. The bottom row shows a null where no tamoxifen is added and as expected we don't see any morphological changes just an increase in population density. A comparison between the final two images of the two sets is especially revealing as we can see cells elongated and overlapping cells in the induced system compared to tightly packed smaller cells.

In this Chapter we discuss one application of the two-step estimation pipeline of STAMM (see Section 3.3.3). We investigate the oncogenic transformation of an

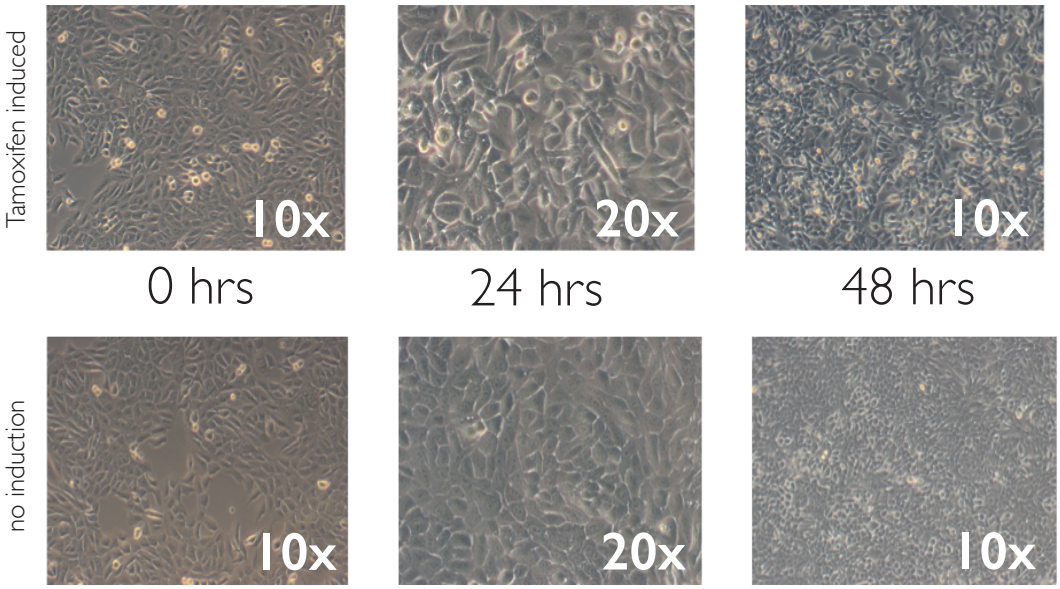


Figure 4.1: During the *in vitro* oncogenic transformation of an MCF10A-Er-Src cell line, morphological transformations can be observed between $24-36$ hours [Hirsch et al., 2010]. Here we show images taken of the experiment at three different time points. The initial measurement at $t = 0$ hours and two subsequent measurements at $t = 24$ and $t = 48$ hours. The magnification level for each image is indicated in the figure. The top row shows images of the experiment where Src is induced by addition of Tamoxifen at $t = 0$; the bottom row shows images taken at the same time points without addition of Tamoxifen. It is possible to see morphological changes at the final time point $t = 48$ where after addition of Tamoxifen cells are elongated and overlapping compared to the tightly packed structure in the system without induction. These images were taken by Sandra de Vries in the van Steensel Lab at the NKI, Amsterdam

MCF10A cell line using data obtained by performing RNA-seq measurements. This experiment was carried out for this work and the experimental design is summarised in Section 4.3. In Section 4.4 we summarise the pre-processing step for RNA-seq measurements starting from integer count data to be used with STAMM. Then we present results from applying the estimation pipeline to RNA-seq data in Section 4.5.

4.2 Relevance

This is a very interesting system to consider mainly because it consists of only a single perturbation (i.e. induction of the classical oncogene v-Src) leading to tumorigenesis and as such is very clean. Additionally it is an exceptionally fast transformation (between $t = 24 - 36$ hours) which makes repeat experiments to probe further properties of the systems or the verification of estimated parameters easy to implement. The transformation also leads to morphological changes which can be observed under a microscope. Moreover the initial state is a stable cell-line therefore tests or follow up experiments are easy to carry out.

4.3 Experimental design

The data we analyse here was obtained by Sandra de Vries at the Netherlands Cancer Institue (NKI). The choice of time points was made considering the following criteria: to facilitate a better understanding of the system using results from previous work in Hirsch et al. [2010]; to have sufficient data points to work well with STAMM; and the cost of each RNA-seq measurement.

The experiments were performed in MCF10A-Er-Src cells, a derivative of mammary epithelial cell line MCF10A containing an integrated fusion of the v-Src oncoprotein with the ligand binding domain of ER [Hirsch et al., 2010; Iliopoulos et al., 2009]. The cells were cultured in DMEM/F12 medium supplemented as described in Debnath et al. [2003]. There are two starting points and for both the medium is refreshed, then a medium containing $1\mu M$ activating tamoxifen is added and the mixture is grown to a $80 - 85\%$ confluency population, and induced and uninduced samples were harvested at the time points $0h$, $0.5h$, $1h$, $2h$, $4h$, $6h$, $8h$, $10h$, $12h$, $24h$, $28h$, $32h$ and $48h$ for RNA isolation. The second series is started when the first one is at $8h$ and samples collected at $0h$, $2h$, $4h$, $16h$, $18h$, $20h$, $24h$, and $40h$. Time points are collected as follows: wash with PBS; add 2ml Trizol resuspend thoroughly and store at $-80°C$. The tamoxifen stock was made by

dissolving 5mg in 12.9ml 96%$EtOH = 1mM$ stock. We did not send all the samples to the microarray facility, only the following (due to cost considerations):

1st series: $0h$, $0.5h$, $1h$, $2h$, $4h$, $6h$, $8h$, $12h$, $16h$, $20h$ and $24h$

2nd series: $0h$, $24h$, $32h$, $40h$ and $48h$

The RNA was isolated by the Trizol method, and prepared for sequencing by the Illumina RNA TruSeq sample protocol.

4.4 Pre-processing data

As described in Section 2.3.2, RNAseq data obtained from samples results in integer counts for genes corresponding to the number of times a sequenced strip belongs to a particular gene. To ensure that we are able to compare samples from different experiments in different setting the first step is to *normalise*. Many methods exist to normalise sequencing experiments. We pre-process the data using the `edgeR` package in R [McCarthy et al., 2012; Robinson et al., 2010]. One important assumption is that most genes are not differentially expressed between samples. To determine genes that are not differentially expressed the procedure uses a robust estimate for the ratio of RNA between samples: the weighted trimmed mean of M values (TMM). Two parameters are employed to filter out genes that are differentially expressed; the M-values, log-fold-changes, and the A-value, absolute expression levels. The cut-off set for both the M-value and the A-value is tuneable and the best way to set the tuning parameters is to select a range of cut-off parameters and determine when they stabilise (see Appendix B). It is important to remember `edgeR` was developed to analyse differential expression [Robinson & Oshlack, 2010], still the assumptions are also applicable to time course such as the data set for the oncogenic transformation. Note without this step it is not possible to compare data from different samples.

After the first pre-processing step we still can't use the data in our model because the likelihood is based on an additive Gaussian noise model (eqn. (3.2)). For RNA-seq data once it has been *normalised* the next pre-processing step is to transform the data such that distorting effects at high expression values are reduced. We use a nonlinear transform proposed by Hoffman et al. [2012] the $\operatorname{arcsinh} x = \ln(x + \sqrt{(}x^2 + 1))$. The advantage of using this transformation compared to the more regularly used $\ln x$ is that it has the same effect of reducing variance at higher values but a much smaller effect at lower values. Now after applying these two pre-processing steps we can use this data in our model.

4.5 Results

The data obtained uses RNA-seq to examine changes in gene expression during this transformation time-course with $T = 13$ time points[1]. According to the assumption in our model all cells are in the initial state at $t = 0$. Here, by experimental design, the initial state consists of the derivative MCF-10A-Er-Src cell line fulfilling this assumption. More specifically the time point $t = 0$ corresponds to the initial treatment with Tamoxifen an anti-oestrogen drug that binds to ER activating v-Src. Of course this is only the case excluding any unidentified epigenetic heterogeneity in the initial cell culture. Therefore it is reasonable to assume that the cell population is approximately homogenous and comprised mainly of untransformed MCF-10A cells.

We want to focus our investigation only on genes that change during the experiment. To that end we filter out all genes j where the standard deviation σ_j is too small, setting the threshold at $\sigma_j > 20$ (on the linear scale, over time). This still leaves $p = 2809$ gene expression trajectories to which we apply STAMM. We carried out parameter estimation as described in Section 3.3.3 using a two-step process. The initial step is to perform a stability test, determining the need for a penalty term by comparing estimated expression signatures from the whole data set and data sets with left out time points. We conclude from the results in Figure 4.2 that there is no need for a penalisation. The next step is to cluster the data using k-means clustering where we want to use cluster centroids to estimate transition rates for the next step. We conclude from Figure 4.3 that the optimal number of cluster $\hat{m}$ is 13. We use this result to perform model selection using cross-validation for $k = \{1, \ldots, 5\}$. We find that the MSE_{CV} score shows a minimum at $K = 4$ states (see Figure 4.4(a)). This would suggest two distinct intermediate states in the oncogenic transformation of MCF-10A cells. Figure 4.5(a) shows a representative set of trajectories from RNA-seq data in green and the blue lines shows predicted trajectories using estimated parameters. Overall we find that the model performed well and fits even diverse trajectories well. State-specific gene expression signatures corresponding to these trajectories are shown in Figure 4.5(b). These values show distinct patterns that would allow for filtering out surface markers distinguishing states. The final step of the estimation pipeline is to check sensitivity of estimation due to an increase in the number of clusters m. This test is performed by computing correlation between expression signatures estimated using $\hat{m}$ clusters and higher numbers of clusters. Figure 4.4(b) shows the correlations as a function of m for $K = 3$ and $K = 4$ and

[1] $t = \{0, 0.5, 2, 4, 6, 8, 12, 16, 20, 24, 32, 40, 48h\}$, at $t = 0$ and $t = 20$ we have a repeated measurements

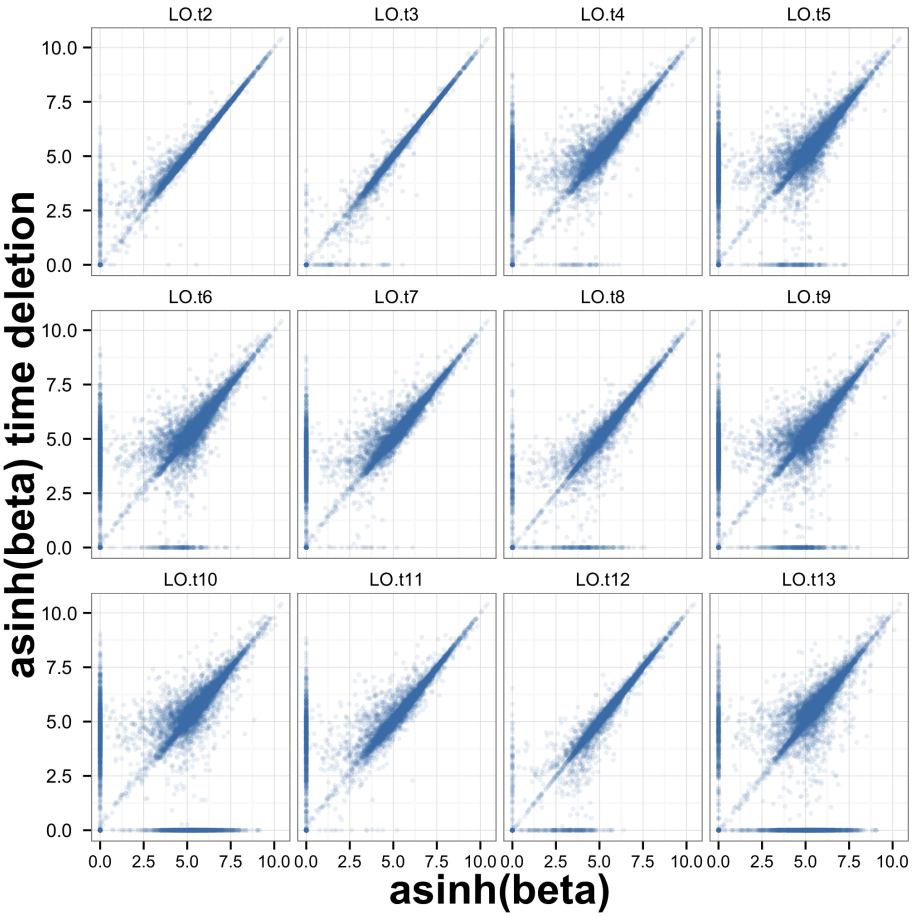


Figure 4.2: The RNA-seq measurements are filtered with respect to standard deviation (we filter out genes with $\sigma_j > 20$ before transformation) because we are interested in genes that change significantly in time. This plot scatters estimated expression signatures using all time points to estimates where one time point is dropped. This tests stability of estimates to determine need for ℓ_1 penalisation under deletion of time points. The estimates are stable (Pearson correlations is > 0.8 therefore we conclude that there is no need for a penalty term.

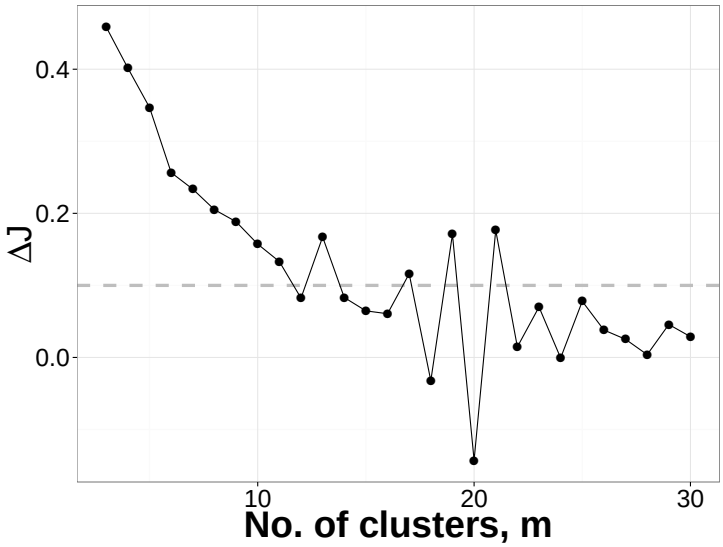
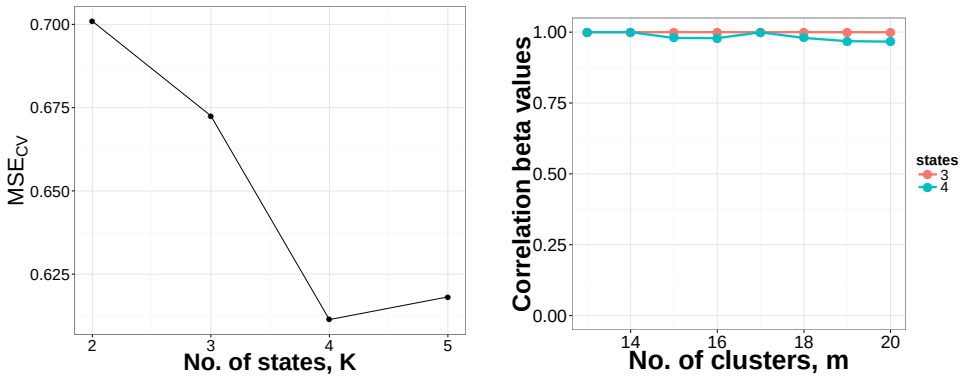


Figure 4.3: K-means clustering of *in vitro* data for the transformation of an MCF10A cell line. Initial k-means clustering is performed to identify representative trajectories. As described in Section 3.3.3 we choose the optimal number of clusters $\hat{m}$ by considering relative changes in the objective function $\Delta J(m) = (J(m-1) - J(m))/J(m-1)$; we set m where $\Delta J(m-1) < 0.1$. The plot shows ΔJ as a function of m and the horizontal dashed line represents our threshold at 0.1. For this example we can see that the $\hat{m} = 13$. Note that fluctuation in ΔJ at higher m are due to small values of $J(m)$.



(a) Number of state K (b) Stability of number of cluster

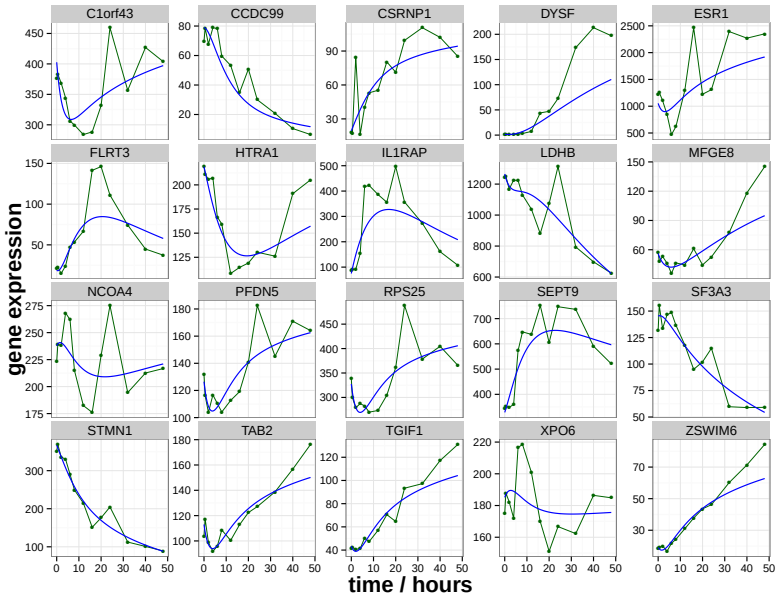
Figure 4.4: MCF-10A data. (a) MSE_{CV} score as a function of K. The minimum at $K = 4$ shows the optimal number of states for this oncogenic transformation. (b) To determine if the choice of $\hat{m}$ in the k-means clustering algorithm was chosen correctly a stability test is performed. Correlations are calculated between expression signatures estimated between the $\hat{m}$ used and larger m. Here we find the clustering is close to one for $K = 3$ and $K = 4$ therefore the choice is reasonable.

we find that the estimates are stable and therefore $\hat{m}$ was chosen well.

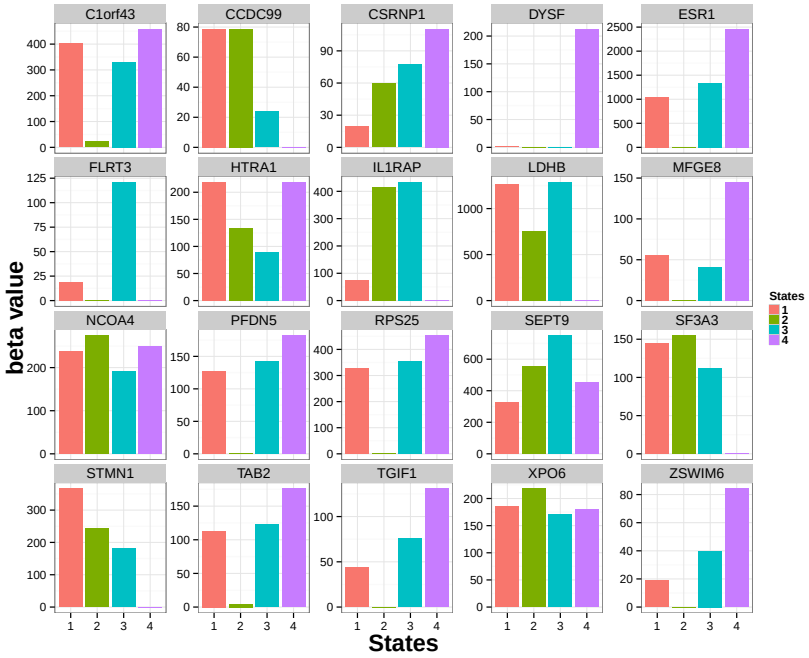
This concludes the application of STAMM to the RNA-seq time course for an oncogenic transformation. The example illustrates the application of the model to novel RNA-seq data can be carried out efficiently. However to investigate and validate of the intermediate states further experiments are needed since the current time-course is biologically not viable for further study (see discussion below) this has to be left as future work.

4.6 Discussion

We applied the estimation pipeline (see Section 3.3.3) to time course data obtained by *in vitro* experiments for oncogenic transformation of a MCF-10A cell line. The systems has many fascinating properties, the fact that it undergoes a rapid state transition between phenotypically distinct states makes it a very useful system to study. Even though the driver for the transformation is one of the most well studied oncogenes there are still many open questions about epigenetic and genetic changes during this process as well as general oncogenic transformations. Note that due to a mycoplasm infection in the cells used for the *in vitro* study substantive biological conclusions can be drawn only with extreme caution. Additional experiments are needed to verify the validity of the time-course data first.



(a) Gene trajectories



(b) Expression signatures

Figure 4.5: MCF-10A data. (a) The green line connects data points from measurements on the oncogenic transformation of MCF-10A cell-line. The blue line shows predicted trajectories from the model using estimated parameters. We can see that the model fits the trajectories well. (b) The estimated β_{kj} values for the trajectories are plotted here as bar charts. We can see that there are diverse expression signatures present in the data.

This application serves only the purpose of illustrating that STAMM can be applied to RNA-seq data yielding useful results. Though in future due to the clean and relatively rapid transformation in the system it is very useful to validate some claims from the model. It would also serve well as an example system to extend the model to include verification steps after an initial application followed by application of a more detailed model. If it is possible to find surface markers that allow us to distinguish specific states it should also be possible to get more detailed information on the latent process i.e. measure transition rates between states and maybe include backward transition to obtain more precise estimates for state-specific expression signatures.

Model selection for $k = \{2 \dots 5\}$ and estimation for all parameters for the RNA-seq dataset was performed after applying threshold on the expression count leaving us with $p = 2809$ genes. Computation took 3.8 hours on 50 cores each with an AMD $2600MHz$ cpu.

Chapter 5

Stem cell reprogramming

5.1 Introduction

Another application of the STAMM model is to the reprogramming of somatic cell to a pluripotent state as investigated in Armond et al. [2013]. In the experimental setup a secondary mouse embryonic fibroblasts (MEFs) is used and transformed to a state of pluripotency [Takahashi & Yamanaka, 2006; Jaenisch & Young, 2008]. More specifically we apply the model to the genome-wide microarray gene expression time-course data obtained by Samavarchi-Tehrani et al. [2010]. This transformation system has received a lot of attention and has been extensively studied in recent years; it has been suggested that the reprogramming process is inherently stochastic [Hanna et al., 2009]. Progress has also been made on single-cell investigations of the biological system [Buganim et al., 2012]. Questions still remain on genome-wide level, including the number of intermediate state between the initial MEF state and the final pluripotent state.

In this Chapter we start by briefly outlining results obtained Armond et al. [2013] when applying STAMM to a microarray data set in Section 5.2. Then (in Section 5.3) we discuss the main contribution in detail which is a comparative study of parameters obtained from STAMM and single-cell experiments performed by Buganim et al. [2012]. The single-cell data was obtained by a new kind of experimental technique called a Fluidigm assay. This also illustrates an example of a possible next step in investigating a biological system once parameters from STAMM have been obtained. The main contributions to this work are: **Computational** including verifying model selection results for both approaches (see below), comparison of multiple hyperparameters for Bayesian model selection and testing underlying structure of data; **Comparison** where estimated parameters from the model are

compared to single-cell measurement.

5.2 Results from model application

5.2.1 Differences in estimation

The initial step before we can make a comparison to single-cell results is to apply STAMM to the microarray time-course; obtaining single-cell level parameters and the number of states K. In Armond et al. [2013] there are differences in the estimation pipeline compared to the one outlined in Section 3.3.3, here we highlight the main differences.

The most important idea of a two-step estimation process is shared is both setups. The first difference is that the optimal number of clusters is chosen when increasing the number of clusters does not significantly improve the k-means objective function. The penalty for estimation used to regularise estimation in eqn. (3.7) is set to a small positive number (in this application set to $\lambda = 0.1$). Estimation of transition rates $\{\mathbf{w}\}$ is performed on genes closest to cluster centroids, instead of the cluster centroids themselves; then transition rates are fixed and estimation of expression signatures β_{kj} is performed in the same way. Finally estimation of the number of states $\hat{K}$ is approached in two separate ways. The heuristic approach is to look at two quantities the model fit, i.e. the residual sum-of-squares (RSS), and the distinctness for individual state signatures quantified by the condition number $C = \max(s_i)/min(s_i)$; where s_i are the singular values of a matrix made up of the expression signatures. The other approach for finding an optimum number of states is employed for genes closest to centroids using ideas from Bayesian model selection. Let $\mathbf{y} = \{y_j\}$ denote observed data and M_k the model with k states. The posterior probability is $P(M_n|\mathbf{y}) \propto p(\mathbf{y}|M_k)$ with a flat prior distribution over models. The marginal likelihood $p(\mathbf{y}|M_k)$ accounts for the fit-to-data and model complexity. Writing all model parameters as $\theta = (\beta_{kj}, \{\mathbf{w}\}, \{\sigma_j\})$ the marginal likelihood is:

$$p(\mathbf{y}|M_k) = \int p((y)|\theta, M_k)\, p(\theta|M_k) d\theta. \tag{5.1}$$

The marginal likelihood of the model eqn. (5.1) is computed using annealed importance sampling (AIS) [Neal, 2001], a MCMC method. Hyperparameters for this model are set by hand to reasonable values, see Supplement of Armond et al. [2013] for details. The normalised score is the required posterior probability over the number of states.

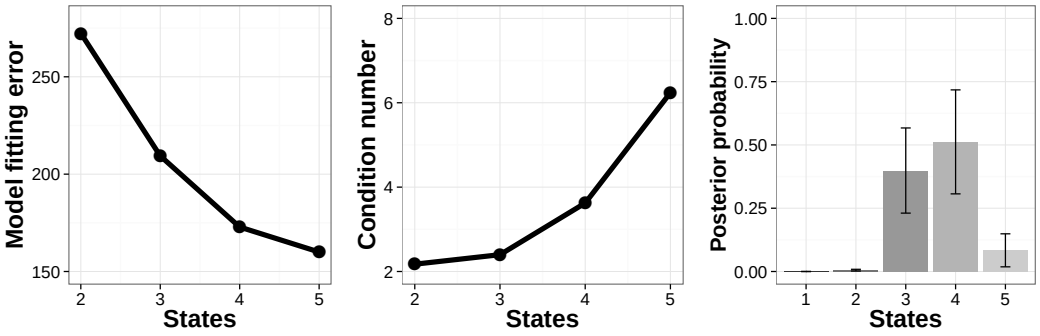
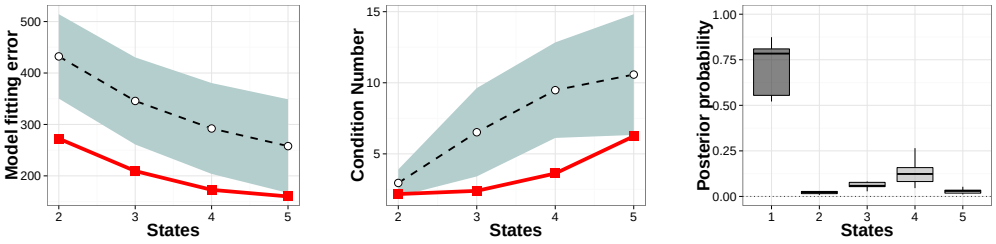


Figure 5.1: Application of STAMM to a microarray time-course (a) Plot of the model fit residual sum of squares (RSS). (b) Plot of the condition number for estimated expression signatures quantifying linear dependence between states. A larger number corresponds to more dependence. (c) Posterior probabilities obtained from Bayesian model selection (see Section 5.2.1 for details).

5.2.2 Estimation results

The primary data used in Armond et al. [2013] is obtained by reprogramming of a secondary mouse embryonic fibroblast (MEF) where Oct4, Sox2, Klf4, and cMyc are expressed under induction in the system for 30 days [Samavarchi-Tehrani et al., 2010]. Microarray measurements were made at $t = \{0, 2, 5, 8, 11, 16, 21, 30\}$ days after induction of expression factors. The microarray data is standardised per gene such that $y_j(t) = (z_j(t) - \mu_j)/\sigma_j$, where $z_j(t)$ is original $\log_2$ transformed data, μ_j is the mean and σ_j is the standard deviation of the time course data for gene j. A total of 4383 genes are retained out of the whole gene list after filtering out genes with small mean and standard deviation. Genes are removed if they are expressed at very low levels and therefore would be dominated by noise.

The number of clusters chosen for this data set of 8 time points is $\hat{m} = 7$. As mentioned above, the penalty used in this application is $\lambda = 0.1$. With these parameters set, the transition rates are estimated from cluster representative genes. Once transition rates are fixed, the expression signatures for the remaining genes are estimated. The analysis is carried out for $K = \{2 \dots 5\}$ and results for model selection are summarised in Figure 5.1. Unsurprisingly the RSS keeps decreasing for increasing K (Figure 5.2(a)) since numbers of model parameters increase. To determine number of states heuristically, we compare these results with the condition number (Figure 5.2(b)). We find that the difference in condition number from $K = 4$ to $K = 5$ is larger than the preceding changes. This suggests that decrease in RSS from 4 to 5 states is mostly due to overfitting and the additional state is not distinct.



(a) RSS (b) Condition number (c) Bayesian model selection

Figure 5.2: Random permutation of time points. The time points of data are randomly permuted ten times and parameter estimation is carried out for each case. (a) shows the RSS as a function of the number of states. (b) shows the condition number a measure for the independence of state-specific expression signatures. In (a) and (b) the solid red line show parameters for the original data and the dashed line represents the average of ten estimates and the dashed area represents a standard deviation. (c) shows posterior probabilities as a function of number of states calculated using Bayesian methods, results from ten estimations are summarised as box plots. For both RSS and condition numbers randomised data performs worse for all states. Bayesian model selection for the permuted data shows no indication that there are intermediate states. The data-set we use is obtained from [Samavarchi-Tehrani et al., 2010].]

The posterior probabilities form Bayesian model selection for 7 genes closest to the centroid (see above for details), are shown in Figure 5.2(c). Combined, these results indicate that a $\hat{K} = 4$ since it strikes a good balance between model fit and distinct expression signatures for states as well as having the highest posterior probability.

5.3 Testing against single-cell data

5.3.1 single-cell experiment

Results in Section 5.2.2 are obtained analysing homogenate time-course data; but the transformation process itself takes place on a single-cell level therefore obtaining data single-cell level and studying the behaviour is tremendously valuable. Comparing results from STAMM to single-cell observations also indicates how well the underlying single-cell process is modelled. For this purpose we investigate the mRNA single-cell expression performed by Buganim et al. [2012]. They also investigate a secondary MEF system reprogrammed by transduction of Oct4, Sox2, Klf4, and cMyc; obtaining data with the Fluidigm assay, resulting in 96 single-cell measure-

ments with gene expression from 48 genes. Observations are made in populations, starting with MEFs, over cells at $2-6$ days during reprogramming, to the final reprogrammed cells.

5.3.2 Comparing results

The single-cell measurements [Buganim et al., 2012] allow for analysis that has not been possible for population average data. Although important questions about the transformation process such as the number of states and transition rates still remain difficult to track down; due to the fact that each time a single-cell measurement is made the cell has to be destroyed and additional work is necessary to determine distinctive marker for known states for purification.

Given available data we can address interesting questions on expression patterns; since we assume that cells belonging to the same states would have a comparable expression patterns across observed genes. This is especially the case since measured genes are deemed important for reprogramming. To this end we cluster the data for all cells in a 48 dimensional gene expression space. To perform the clustering we use a widely available clustering tool in R `mclust`; it employs a variety of multi-variate clustering methods and scores them using the Bayesian Information Criterion (BIC). The best performing method is shown in Figure 5.3. We find that optimal number of clusters is 3 since the BIC score starts decreasing for larger cluster sizes. Buganim et al. [2012] suggest four clusters which they obtain using principle component analysis which is depends heavily on initial normalisation.

Next we try to determine if state specific expression signatures, estimated from microarray data, can be compared with this new single-cell data. Disregarding conditions for each cells measurement we assign each of the single-cell measurements to each of the states in the $K = 4$ model. Since measurements are performed different systems as well as a using different procedures we scale pre-processed data to be in the interval $[0, 1]$. Then we compute the euclidean distance between gene expression on a single-cell level and estimated gene expression signatures and assign each cell to a specific state. The heatmap in Figure 5.4 shows fractions of cells that are assigned to each state. All MEF conditions have a peak at $K = 1$. Measurements obtained between $t = 2$ and $t = 44$ days are spread over state $K = 1$ and $K = 3$ with very few cells also in the second state. No cells from each of these measurement are close to the final state. Measurements for dox-independent and iPS cells occupy only the final two states. These results clearly show a transformation starting with MEF state and undergoing changes across intermediate states before reaching the final reprogrammed state. Of course this is a very small study and a study made on

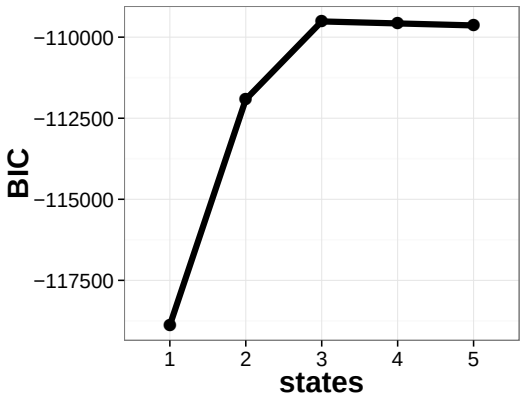


Figure 5.3: single-cell expression levels in different experimental settings from Buganim et al. [2012] are clustering using a standard clustering procedure in R called mclust. We use the Baysian Information Criterion (BIC) to score different cluster sizes. We find the optimal number of clusters to be 3 since the BIC score decreases for larger cluster sizes.

a slightly different system under different conditions, therefore even the approximate similarities we find to our estimated parameters are promising.

5.4 Discussion

We showed how STAMM can be applied to the transformation of differentiated cells from MEF to iPS cells with the help of reprogramming factors. The data used for the first part of the analysis was a population averaged microarray time-course. We outlined the procedure used in Armond et al. [2013] highlighting differences to the estimation pipeline introduced in Section 3. We present results from fitting STAMM to the data and we find that a four state model best describes stem-cell reprogramming, which has also been corroborated by previous experiments Samavarchi-Tehrani et al. [2010]. Additionally we showed that data indeed has underlying structure that can be described by STAMM. To test this we performed estimation for ten samples with randomly permuted time points and compare RSS scores, condition number and Bayesian model selection. We found that RSS and condition number are always lower for the original data than the permuted data. The Bayesian model selection result show a very high concentration at $K = 1$ states.

We also compared model predictions from STAMM to the single-cell data, obtained in a different secondary MEF experiment measuring 48 genes for 96 single-cells [Buganim et al., 2012]. We used a standard clustering tool and determined only 3 states in the data which can either mean that there is not sufficient information

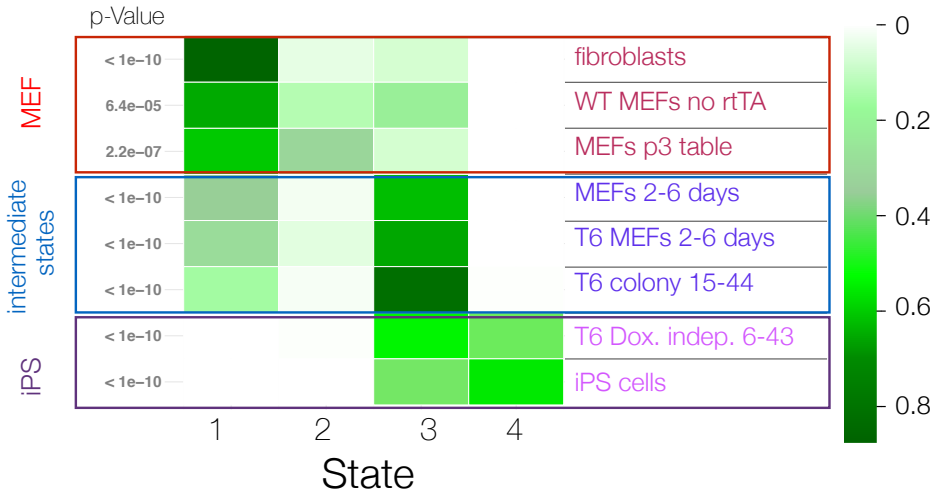


Figure 5.4: Estimated gene expression signatures using STAMM are compared with single-cell measurements, performed by Buganim et al. [2012]. Each single-cell is assigned to a state by finding minimum euclidean distance. The heatmap summarises the fraction of cells from each experimental condition, assigned to specific states. Different conditions show a clear preference for specific states. The prediction of our model are in line with this observation where initial MEF states undergo a transformation via intermediate states to a final reprogrammed state. As an example all MEF populations (top three entries) have a significantly higher fraction of cells in $K = 1$. Cells measured between $t = 2$ and $t = 44$ days have cells that spread across the first and third state, with very few cells occupying the second state. None of these cells are close to the final state. The two measurements which are reprogrammed cells (iPS cells and Dox. indep.) show similarities to $K = 3$ and $K = 4$, but none of these cells are close to the first two states.

available or that intermediate states are characterised by genes not measured in this experiment. If we map single-cell measurements at different time points to gene expression signatures we find that single-cells measured at different times are close to the states predicted by our model at those times. These results are encouraging but for them to be conclusive, as mentioned above, we would need to carry out further experiments.

Chapter 6

Cell cycle

6.1 Introduction

In STAMM an important assumption is that the initial cell population is homogeneous. For the applications we discussed in previous chapters this assumption is warranted because experiments are designed to ensure initial homogeneity. In the case of the oncogenic transformation a cell line is used as the initial cell population. In the case of stem cell reprogramming the technique outlined by Hanna et al. [2009] tries to ensure initial homogeneity by using a secondary MEF cells.

Recently it has been shown that even seemingly homogeneous cell populations exist in inherent mixtures, be it at an epigenetic level [Heng et al., 2009; Gerlinger et al., 2012]. Without knowing the initial population it is not possible to apply STAMM to determine individual states. In this Chapter we outline a model that answers the question: What effect does the initial cell population have on cell fate?

An example of such a biological system is one with an initial heterogeneous cell population made up of two types of cells, with an indistinguishable phenotype. At time $t = 0$ the cells receive a stimulus leading to a transformation such that at $t = T$ it is possible to distinguish cells in their final cell fate. Now it is possible to count the fraction of cells that reach each of those final cell fates. The interesting case here is when the strength of the stimulus has an affect on the fraction of cells in each cell fate. A schematic of such a system is shown in Figure 6.1. Individual genes are influential in determining cell fate if their expression level is significantly different between cell populations at $t = 0$.

The possible experimental application for this model was not ready in time for this thesis therefore all simulation parameters are merely chosen to ensure a

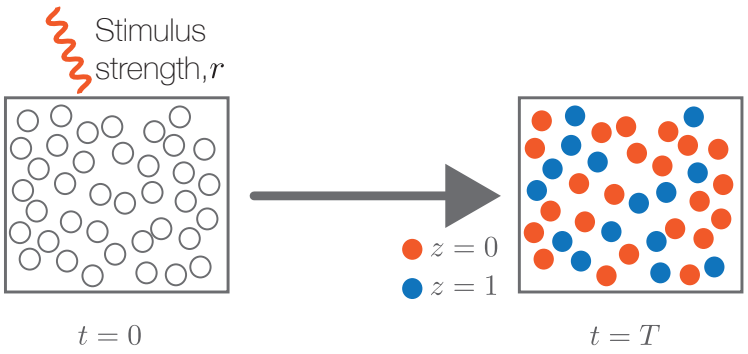


Figure 6.1: Schematic of model. A heterogeneous cell population (the source of heterogeneity is unknown) is at rest at $t = 0$, at which point it is perturbed by an external stimulus of strength r; evoking a transformation such that two distinct populations form at a later time $t = T$. At $t = T$ it is possible to count the number of cells in different states.

full range of stimulus response starting from very few cells responding to a weak stimulus, after which the stimulus is increased until all cells respond and share the same cell fate at $t = T$. The chosen parameter might not bear resemblance to their counterparts in a real biological system; but this model serves as a proof of concept and as long as the basic principles hold it can be applied to a real system.

We start this chapter in Section 6.2 with the basic concepts involved, introducing variables and deriving behaviour of such a system followed by the derivation of a model that can be used to estimate parameters given possible measurements on the system. After that we set up a simulation procedure in Section 6.3. Finally in Section 6.4 we present results using simulated data, first with only one gene and then with four genes.

6.2 Formal system description

6.2.1 Concepts

Suppose at time t cell i (out of N cells) occupies state X_{it}; where state here broadly refers to any aspect of the cell's physical configuration. This can include protein profile, transcription, or its chromatin state. Denote ultimate cell fate at $t = T$ for cell i by Z_i. Cell fate Z_i is determined experimentally by enumerating cells in distinguishable states at $t = T$. In the simplest case cells can have two distinct final states; we label the two states $Z_i = 0$ and $Z_i = 1$. We expect the process that determines cell fate to have a stochastic component such that two physically identical cell at $t = 0$ can end up in distinct final states. Hence we assume the

probability that cell i is in state $Z_i = 1$ at $t = T$ depends on two things:

- The physical state of cell i at $t = 0$, X_{i0}.
- The dose of the stimulus r.

We assume that the fraction of cells that reach the arrested state changes with the strength of the stimulus. The fraction of cells that reach cell fate $Z_i = 1$ is dose dependant and denoted by $\pi(t)$.

6.2.2 Model

Fist we will set up the model in a general sense to give a birds-eye view of the system we are attempting to study. After that we will derive from this picture the specific model one can use with realistic obtainable data. In this more conceptual chapter this is an important step to ensure the actual model can be fully explained.

We start with formalising the variables. Set the expression of gene j in cell i measured at time t under stimulus strength r to $x_{jt}^{(i)}(r)$. In the first instance we are only going to consider one gene therefore we initially drop the subscript j to make the derivation clearer. When we are referring to measurements at $t = 0$ they are independent of the stimulus, therefore we write $X_0^{(i)}(r) = X_0^{(i)}$. The probability of observing a gene expression conditional on the strength of the stimulus can be written as a combination of types of populations present:

$$p(X_t^{(i)}|r) = p(X_t^{(i)}, Z_i = 1|r) + p(X_t^{(i)}, Z_i = 0|r) \tag{6.1}$$

$$= p(X_t^{(i)}|Z_i = 1, r)\, p(Z_i = 1|r) + p(X_t^{(i)}|Z_i = 0, r)\, p(Z_i = 0|r). \tag{6.2}$$

Now to obtain the expected value for the expression conditional on the strength of the stimulus can we calculated using the results of eqn. (6.2) as:

$$\mathbb{E}[X_t^{(i)}|r] = \int X_t^{(i)}\; p(X_t^{(i)}|r) dX_t^{(i)} \tag{6.3}$$

$$\begin{aligned} &= p(Z_i = 1|r) \int X_t^{(i)}\; p(X_t^{(i)}|Z_i = 1, r) dX_t^{(i)} + \\ &+ p(Z_i = 0|r) \int X_t^{(i)}\; p(X_t^{(i)}|Z_i = 0, r) dX_t^{(i)}. \end{aligned} \tag{6.4}$$

Introducing the new variable $\pi(r) = p(Z_i = 1|r)$, and for a two state system naturally $1 - \pi(r) = p(Z_i = 0|r)$ we rewrite eqn. (6.4)

$$\mathbb{E}[X_t^{(i)}|r] = \pi(r)\ \mathbb{E}[X_t^{(i)}|Z_i = 1, r] + (1 - \pi(r))\ \mathbb{E}[X_t^{(i)}|Z_i = 0, r]. \tag{6.5}$$

To further simplify and write a general equation for a system of this type we rewrite the expected value of the expression level for cells in state $Z_i = 1$ and $Z_i = 0$ as $\alpha(r) = \mathbb{E}[X_t^{(i)}|Z_i = 1, r]$ and $\beta(r) = \mathbb{E}[X_t^{(i)}|Z_i = 0, r]$ respectively and write:

$$\mathbb{E}[X_t^{(i)}|r] = \pi(r)\,\alpha(r) + (1 - \pi(r))\ \beta(r). \tag{6.6}$$

In the biological system described above gene expression is measured at $t = 0$ and is independent of the applied stimulus since it is only applied at the initial time point. Therefore we define the expected value of expression measured for gene j $\mathbb{E}[X_{t=0}^{(i)}|r] = X_0$ as:

$$X_{0j} = \pi(r)\,\alpha_j(r) + (1 - \pi(r))\ \beta_j(r). \tag{6.7}$$

The system is now fully described, potential measurements that can be made here are the average expression levels of genes at $t = 0$, X_{0j} and the fraction of cells in state $Z = 1$ at time $t = T$. On account of the lack of information about the variables $\alpha(r)$ and $\beta(r)$ or their distribution, estimation using eqn. (6.7) is not feasible. Therefore below we formulate a procedure to estimate parameters that would allow us to determine genes that significantly influence the transition between states under a stimulus.

Estimation

We start by defining the fraction of cells in state $Z_i = 1$ at $t = T$, note that the derivation in the following few lines is equivalent for cells where $Z_i = 0$. We will perform the detailed calculation for only once case. Here it is also easier to use $\mathbf{X_0}$ as a vector notation for X_{0j}

$$\pi(r) \equiv p(Z_i = 1|r) = \int p(Z_i = 1|\mathbf{X_0}, r)p(\mathbf{X_0}|r)d\mathbf{X_0} \tag{6.8}$$

$$= \int f(\mathbf{X_0}; r, \theta)p(\mathbf{X_0}|r)d\mathbf{X_0}, \tag{6.9}$$

where in eqn. (6.9) we have replaced $p(Z_i = 1|\mathbf{X_0}, r)$ by a general parametric function of the gene expression dependent on the stimulus and a set of parameters θ. Since $p(\mathbf{X_0}|r)$ is the probability distribution of gene expression at time $t = 0$ over all cells a good approximation to this process is the exponential distribution; the parameter λ_j of the exponential distribution can be set to the measured gene expression for gene j at $t = 0$ and will be different for every gene. Therefore we can rewrite eqn. (6.9) as:

$$= \mathbb{E}[f(\mathbf{X}_0; r, \theta)]_{Exp(\lambda_j)}. \tag{6.10}$$

The expression in eqn. (6.10) is general and allows for an expression depending on the problem being investigated. Obtaining a closed form solution of the expectation value is only possible for special cases. For the biological system we present below there is no closed form solution. Hence we propose a numerical approximation using a Monte Carlo integration. We can therefore write:

$$\pi(r) \approx \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{X}_0^{(n)}; r, \theta), \tag{6.11}$$

where each component of the vector $\mathbf{X}_0^{(n)}$ is sampled from the exponential distribution. In each case it has to be considered if a simple application is sufficient, or if there is need for a more involved algorithm. To identify the optimal set of parameters in the estimation we perform a grid search over all parameters and compare the residual sum of squares (RSS) between observation and estimation using eqn. (6.11).

Application cell cycle

A biological system where the model described above could be applied to is cell cycle arrest due to radiation. The system is quite simple a population of cells is exposed to varying levels of radiation starting at $t = 0$. The expression level of these cells is measured using RNA-Seq. At a later time $t = 2h$ cells reach two possible states: (i) arrest leading to cell death, or (ii) the cell still has the ability to enter cell cycle and replicate. The fraction of cells in either state can be counted at $t = 2h$. In this application increasing the dose of radiation of course decreases the number of cells that can re-enter the cell cycle. It is of interest to distinguish discriminatory factors for both cells that determine ultimate fate of a cell.

In this application we have a few simple relationships that our choice of $f(\mathbf{X}_0^{(n)}; r, \theta)$ has to obey. Below a certain radiation threshold the effect of radiation

will be negligible, and above a certain radiation threshold almost all cells will arrest. Under such constraints the best choice is a sigmoid of the form:

$$f(\mathbf{X}_0^{(n)}; r, \theta) = \frac{1}{1 + \exp(-r\,\mathbf{a}^T\mathbf{X}_\mathbf{0}^{(\mathbf{n})} + b)}, \tag{6.12}$$

where $\mathbf{a}$ and b are parameters of the sigmoid to be determined from estimation. A large negative or positive value in $\mathbf{a}$ for a gene means it has a larger influence on transformation.

6.3 Simulation

The simulation setup we choose for this model is based on a single-cell level approach which we employ in Section 3.4 obtaining very useful results. The first step is to simulate gene expression for each gene and a fixed number of cells; this is sampled from an exponential distribution for each gene $Exp(\lambda_j)$. Then the probability of being in state $Z_i = 1$ is determined using eqn. (6.12) based on expression levels of all genes involved and fixed radiation value and parameters $\mathbf{a}$ and b. Drawing a value uniformly at random in the range $[0, 1]$ determines for each cell if it is in the arrest state. Using this final state vector we can determine $\pi(r)$. Measurements for such a system will be limited so we repeat these simulation steps for only 9 values of r. Finally we add Gaussian noise to the fraction of cells observed with zero mean $\mathcal{N}(0, \sigma)$, since in a real experiment measurements of fraction of cells will be noisy.

6.4 Results

6.4.1 Single gene simulation

Initially we perform the simulation and estimation with only a single gene. This is to determine whether or not it is possible to estimate parameters for the simplest possible case. We perform simulations for 1000 cells for one gene with $Exp(\lambda_j = 5)$ at $r = \{0, 0.1, 0.5, 1, 2, 3, 5, 6, 10\}$. The sigmoid parameters are chosen as $a = 2$ and $b = 5$. To visualise the single cell simulation behaviour Figure 6.2 shows the states occupied at $t = 2h$ by 1000 cells at different raditation doses. The y-axis in the plot represents gene expression transformed for convenience to $\text{arcsinh}(X_0)$. Once we have obtained values for $\pi(r)$ from this setup we add two levels of Gaussian noise $\sigma = \{0.025, 0.05, 0.1\}$. Since the measured quantity is in the range $[0, 1]$ this is a reasonable level of noise, ranging from 2.5% to 10% of the simulated trajectory. Finally we choose a grid for b over the interval $[0, 10]$ and a over the interval $[0, 4]$

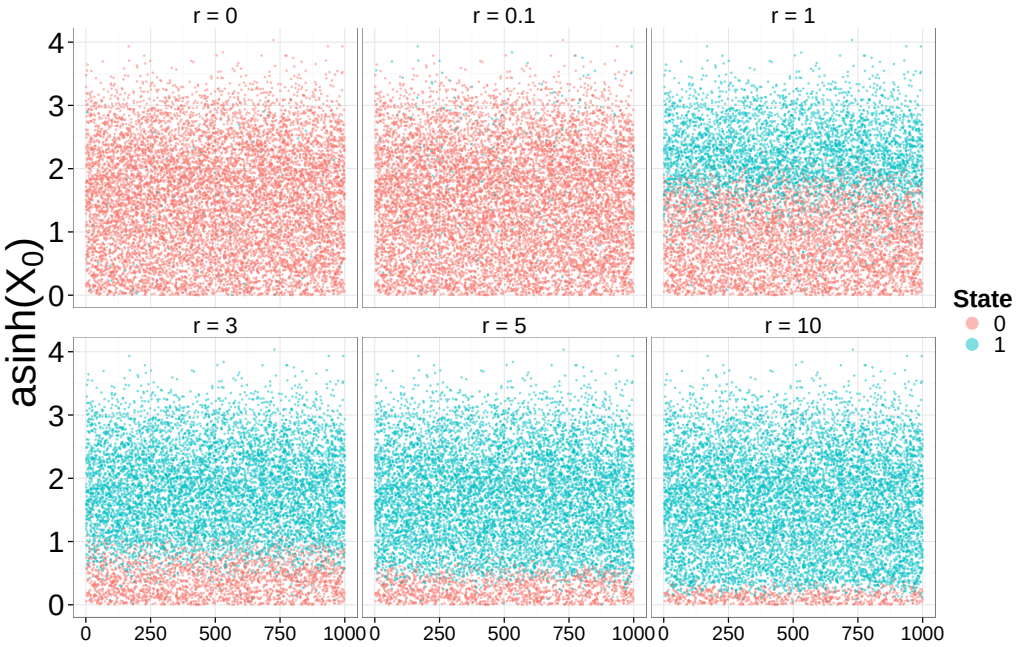
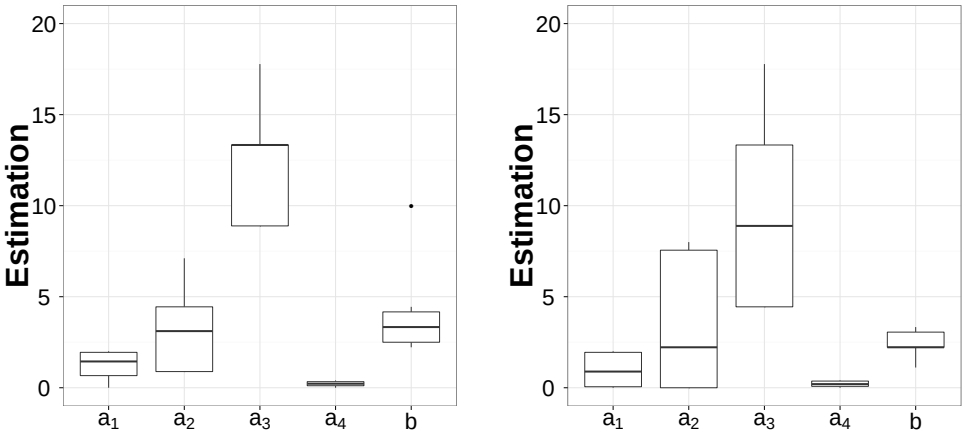


Figure 6.2: Single cell simulation for one gene. Plots show the states of cells at different expression levels after exposure to radiation r for $t = 2h$. Each figure represents a different dose of radiation r. The y-axis shows expression level and the x-axis is just an index over 2000 cells which are used in the simulation. State $Z_i = 0$ is the normal state of the cell and state $Z_i = 1$ is when cells enter the arrest state.

	true value	$\sigma = 0.025$	$\sigma = 0.05$	$\sigma = 0.1$
a	2	1.655	1.124	0.828
b	5	4.483	3.448	2.759

Table 6.1: Estimation of a and b parameters from simulated data. Data is obtained by adding Gaussian noise to the simulated trajectory with differnt σ values.

and find the minimal RSS. The results from these estimations are summarised in Table 6.1. We see that for $\sigma = 0.025$ the estimates are very close to the true values used in simulation. At higher noise the estimates become progressively worse, but the parameter order is still preserved.



(a) $\sigma = 0.05$ (b) $\sigma = 0.1$

Figure 6.3: Simulation study. Ten simulation carried out with four genes with mean chosen uniformly between $[0, 20]$, stimuli used at $r = \{0, 0.1, 0.5, 1, 2, 3, 5, 6, 10\}$. Parameters for the simulation is chosen as $b = 5$ and $a = [1, 5, 10, 0.2]$ for the simulation. (a) is the simulations with Gaussian noise added at $\sigma = 0.05$. The estimated parameters are ranked in the correct order. (b) is the simulation performed with noise added at $\sigma = 0.1$. Parameters estimation is worse, but not unexpected since the noise added is at 10% of the observed data.

6.4.2 Simulation with multiple genes

Just performing simulation using a single gene only serves to show the simulation pipeline and as a first test to determine if the model will work. If we wish to determine how well the proposed model will serve in a realistic application it is necessary to expand the simulation to include multiple genes, otherwise the model is not useful. To this end we implement a simulation with 4 genes. Just like above we perform a simulation with 1000 genes measured at $r = \{0, 0.1, 0.5, 1, 2, 3, 5, 6, 10\}$. We choose the parameters as follows $b = 5$ and $a = [1, 5, 10, 0.2]$ and the mean gene expression for each gene is drawn uniformly between $[0, 20]$. To estimate parameters we apply a grid over all parameters calculating the integral eqn. (6.11) with the Monte carlo approximation and find the parameter set that minimises RSS. This is of course a very inefficient method, and with increase in parameters the grid search quickly becomes unfeasible. A more sophisticated implementation of parameter selection is necessary for a realistic application. We repeat this simulation 10 times with independently drawn over expression and we repeat this for two different standard deviations of noise added to the fraction of transformed cells 0.05 and 0.1. Figure 6.3 shows the results for both noise levels and we can see that for lower noise the estimation works well and parameters are estimated in the right order. For larger noise this is already difficult and some parameters are estimated well and some are not.

6.5 Discussion

We derived the model starting from general biological principles including a description of the system. We introduced one possible application for this model in a radiated cell population either stopping cells in cell cycle or only temporarily halting it. In such an experiment we wanted to find the importance of genes for the transformation. In this model there is one parameter that determines the importance of genes. The higher the value for a gene the more important that gene is in transformation. We showed in simulation that we can obtain parameters even when we add large amounts of noise to the simulated data in single cell scenario. With limited observations made on the system i.e. the initial gene expression and proportions of cells reaching a specific cell fate for different stimuli, we are interested in finding out if a gene influences cell fate or if it does not. It will not be possible to estimate exact parameter values. We show that it is possible to obtain ranks of genes using this model.

For a real application further work is needed. Only with real data comparison

is it possible to determine if the choices made for parameters during simulation are reasonable and if simulated state fractions are realistic. Further work is also needed to implement a more useful parameter search instead of the crude one implemented at the moment. Once it can be compared to data it will also be possible to refine the model if needed.

Chapter 7

Discussion and Outlook

Advancements in measuring population level transcription, proteomic abundance and epigenetic information cheaply has lead to widespread application of these methods. Additional measurements on a single-cell level have also been made in more recent years [Wheeler et al., 2003; Dalerba et al., 2011; Wang & Bodovitz, 2010] but this technique has its own shortcomings, but they suffer from a limitation on the amount of information that can be obtained through such techniques. For transforming biological systems single-cell measurements would be ideal even if there is a limit on the amount of information available, but current techniques require destruction of cells for a measurement of more than a handful of genes or proteins. Therefore techniques that attempt to deconvolve information on the single-cell level from population level data are extremely important for a better understanding of causes and triggers of transformation.

Chapter 3 is focused on state transitions, using aggregated Markov models (STAMM) where we expanded on previous work [Armond et al., 2013] and presented a full description of the model including assumptions. We investigated properties of the model such as (empirical) identifiability, behaviour when assumptions are broken using single-cell simulations and proposed a computationally efficient unbiased approach to parameter estimation and model selection. We showed empirically that the model is identifiable given assumptions. Under breaking model assumptions estimations are stable but if model assumptions are strongly violated, state specific expression are still recovered, but transition rates can be badly estimated. Therefore estimation results should always be considered as motivation for further experiments and further estimation results need to be experimentally verified.

In Chapter 4 we applied STAMM to RNA-seq data for obtained by *in vitro* experiments for oncogenic transformation of a MCF-10A cell line. The system is

based on one of the most well studied oncogene, still there remain open questions about exact steps required for the transformation. Due to a mycoplasm infection in cells used for the *in vitro* study, biological conclusions need to be interpreted with caution. Therefore this application is useful to illustrate that STAMM can also be applied to RNA-seq data with useful results. In future this oncogenic transformation due to its rapid transformation and clean induction can be used to validate the model in more detail and will also enable the extension of the model. Finding unique surface markers combinations for states it will be possible to identify states of cells and to isolate them and verify parameters. We also showed that this implementation of the model is computationally efficient and it takes less than 4 hours to estimate $p = 2809$ genes on fifty cores.

Chapter 5 contains a concise outline of application of STAMM to a microarray time-course obtained from reprogramming differentiated cells to iPS cells. We briefly sketched the procedure used in Armond et al. [2013] and parameters obtained during estimation. To test if estimation results are accidental or if data indeed has underlying structure, we also applied STAMM to randomly permuted data. Results show that there was indeed structure in the data that is modelled by the latent Markov chain. Then we compare model predictions from STAMM to recent single-cell data from a different secondary MEF experiment [Buganim et al., 2012]. We determine that results are consistent with findings from single-cell measurements in terms of the number of intermediate states if we compare single-cells measured at different time points can be mapped well to corresponding state specific expression signatures.

We note that even though we restrict ourselves to applications of STAMM to gene expression data, this is just a consequence of the systems being investigated. The model itself can be applied to any type of data that is considered to be relevant to a transformation process. The future development of this model could include incorporation of additional experiments. For instance once parameters have been estimated from data, a second round of experiments could verify transition rates by filtering out cells in a state and determining transition to the next state for those cells. Once transition rates are fixed estimation of expression signatures becomes more accurate as well. In fact once transition rates and number of states have been experimentally verified the model could be extended to also probe single state dynamics.

In Chapter 6 we proposed a model for an initially heterogeneous population where it is possible to observe ultimate cell-fate subject to a stimulus. The gene expression is only measured for the initial cell population and subsequent cell-fate

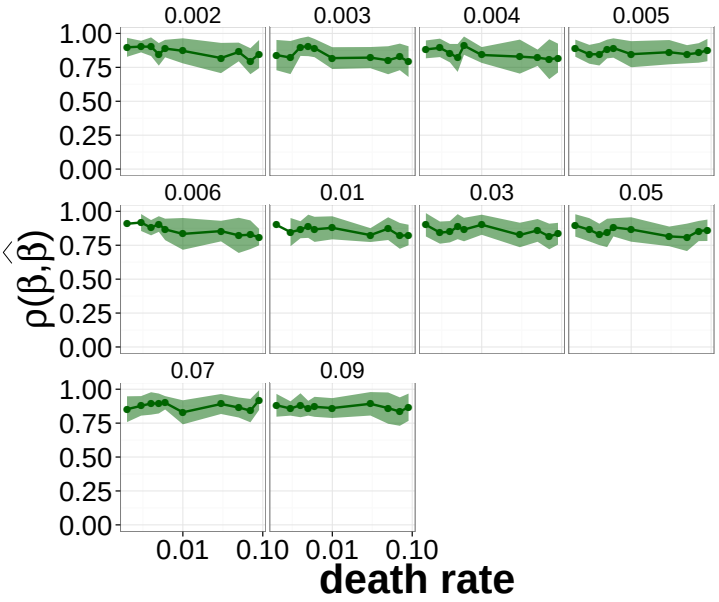
is determineds for different stimulus strengths. Starting from basic principles we set out a description of the biological system followed by an outline of the estimation procedure based on data obtained experimentally. We set out a single-cell simulation procedure for this system. We use a single gene simulation and a four gene simulation to test the model. We show that in single gene simulation it is possible to pick out parameters at a variety of noise levels. In the four gene simulation parameters are picked out well at low noise levels. At higher noise levels this is more problematic and some parameters are not estimated well.

Transformation processes in biology are central to understanding many diseases and potential developments of cures that their study is actively pursued and if anything is being further extended. Due to single-cell stochasticity many studies concentrate on final and initial states of cells because intermediate stages of transformation are more difficult to probe. The types of models we outline in this thesis could prove invaluable for a full understanding of transformation processes. Modern genome-wise experimental techniques that take measurements for single-cells still destroy the cells thus still only giving a snapshot in time de Souza [2012]. These measurements in conjunction with STAMM would allow for a powerful reconstruction of the transformation process. In short, there still remains a lot of work to be done to understand cellular transformations in biology and we believe that such models that take the single-cell level stochasticity into account could provide crucial assistance in this endeavour.

Appendices

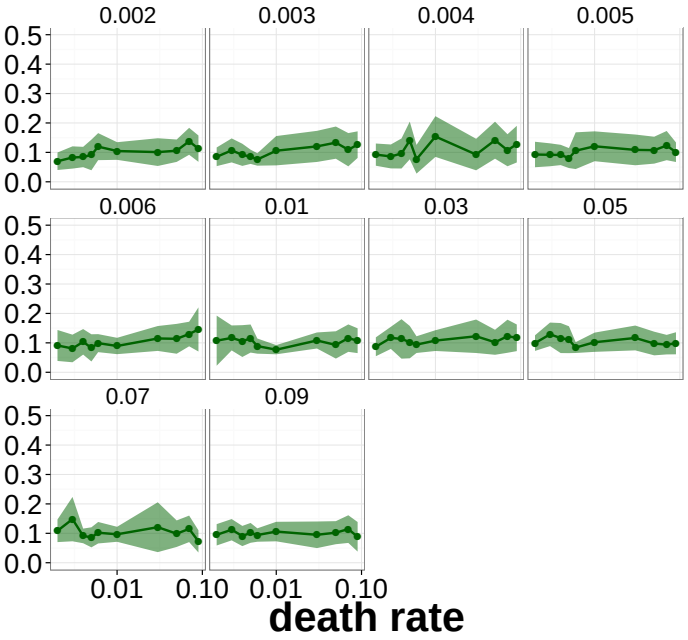
Appendix A

Additional results MAST

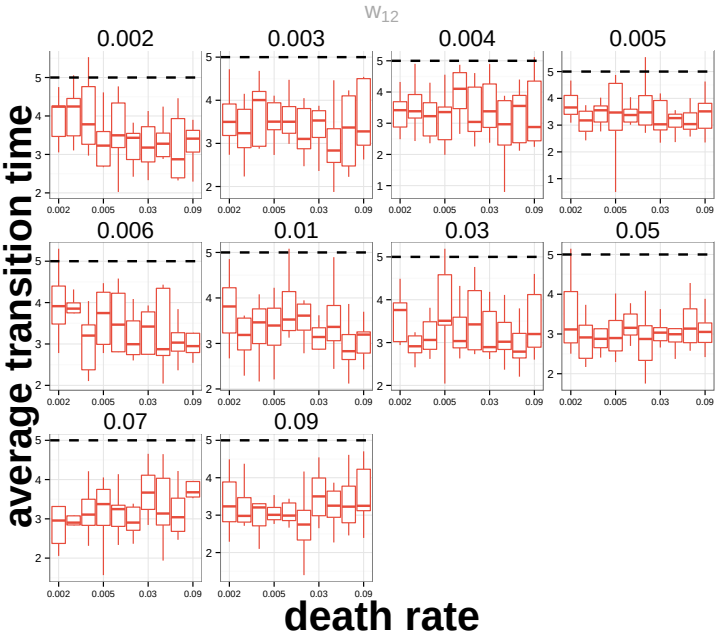


(a) expression signatures

Figure A.1: Figure carried on below.

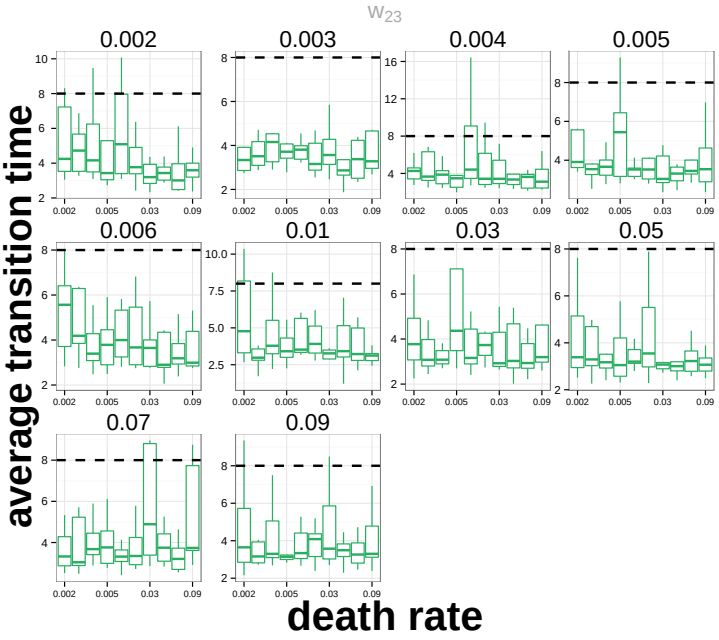


(a) state occupation probabilities

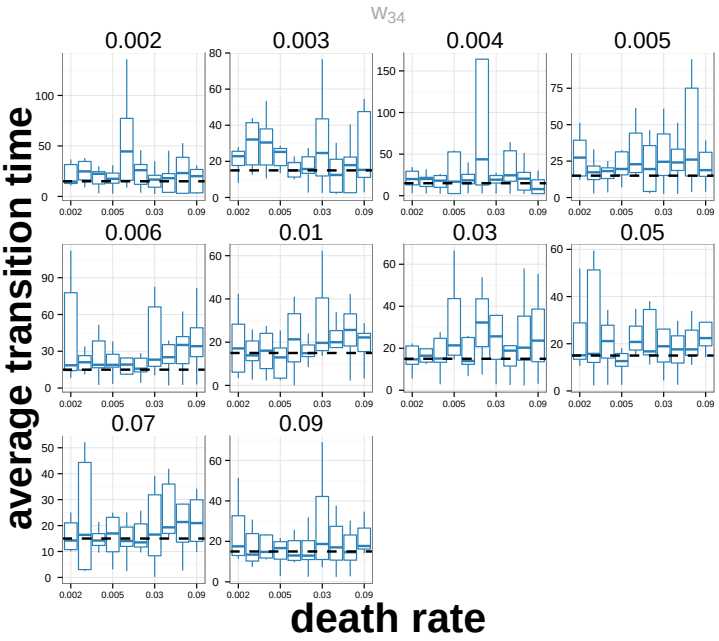


(b) transition rate w_{12}

Figure A.1: Figure carried on below.



(c) transition rate w_{23}



(d) transition rate w_{34}

Figure A.1: Simulation study. This plots extends Figure 3.8 to include addition values for the doubling rate, plots are divided into panels for each cell doubling rate. Simulations are independently repeated 10 times. (a) shows correlation between true and estimated β_{kj} as a function of death rates. (a) shows difference between estimated occupation probabilities and true values (see Section 3.5.1). In (a) and (a) show the mean as a solid line and the shaded area represents the standard deviation. (b) - (d) shows box plots for estimated transition rates with the dashed line showing true values.

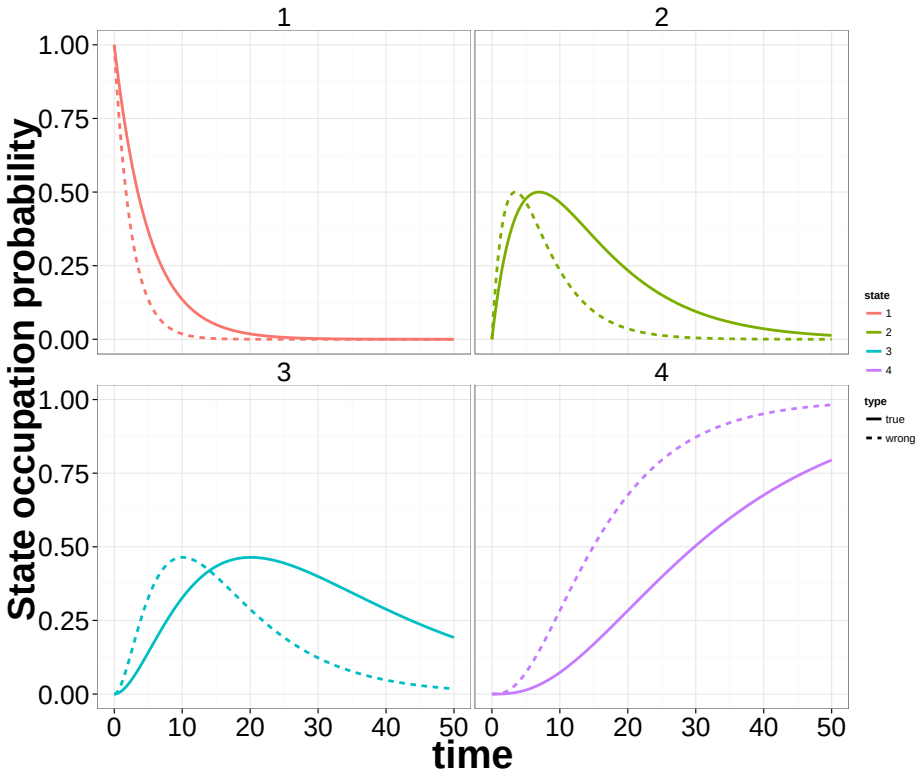


Figure A.2: Simulation study. To highlight the effect of badly estimated transition rates we show state occupation probabilities for each state. The solid line shows probabilities for transition rates $[0.2, 0.1, 0.05]$ the dashed line shows occupation probabilities with transition rates $[0.4, 0.2, 0.1]$. This shows that estimating transition rates is a more difficult problem; as even large deviations in transition rates lead to only small changes in occupation probability which is the only way transition rates enter the estimation.

Appendix B

RNA-seq pre-processing

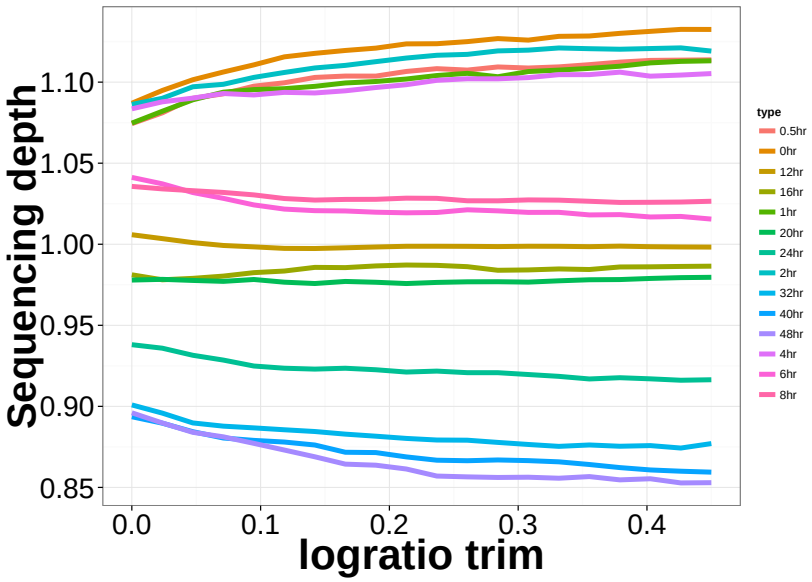


Figure B.1: To determine sequencing depth from RNA-Seq experiments we pre-process data using the *edgeR* package. See package Robinson et al. [2010] for details of exact usage which briefly outlined in Section 2.3.2. We propose a strategy of trimming the log ratio eqn. (2.17) at different values and identifying the cutoff that stabilises sequencing depths for the different samples. For this example we chose 0.4.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2007). *Molecular Biology of the Cell, Fifth Edition*. Garland Science.

Armond, J. W., Saha, K., Rana, A. A., Oates, C. J., Jaenisch, R., Nicodemi, M., & Mukherjee, S. (2013). A stochastic model dissects cell states in biological transition processes. *Scientific Reports (accepted)*.

Bagci, H., & Fisher, A. G. (2013). DNA Demethylation in Pluripotency and Reprogramming: The Role of Tet Proteins and Cell Division. *Stem Cell*, *13*(3), 265–269.

Bar-Joseph, Z., Farkash, S., Gifford, D. K., Simon, I., & Rosenfeld, R. (2004). Deconvolving cell cycle expression data with complementary information. *Bioinformatics*, *20*(suppl 1), i23–i30.

Bar-Joseph, Z., Siegfried, Z., Brandeis, M., Brors, B., Lu, Y., Eils, R., Dynlacht, B. D., & Simon, I. (2008). Genome-wide transcriptional analysis of the human cell cycle identifies genes differentially regulated in normal and cancer cells. *Proceedings of the National Academy of Sciences*, *105*(3), 955–960.

Baum, L. E., & Petrie, T. (1966). Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *The Annals of Mathematical Statistics*, *37*(6), 1554–1563.

Bock, C., Kiskinis, E., Verstappen, G., Gu, H., Boulting, G., Smith, Z. D., Ziller, M., Croft, G. F., Amoroso, M. W., Oakley, D. H., Gnirke, A., Eggan, K., & Meissner, A. (2011). Reference Maps of Human ES and iPS Cell Variation Enable High-Throughput Characterization of Pluripotent Cell Lines. *Cell*, *144*(3), 439–452.

Brown, P. O., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics*, *21*, 33–37.

Buganim, Y., Faddah, D. A., Cheng, A. W., Itskovich, E., Markoulaki, S., Ganz, K., Klemm, S. L., van Oudenaarden, A., & Jaenisch, R. (2012). Single-Cell Expression Analyses during Cellular Reprogramming Reveal an Early Stochastic and a Late Hierarchic Phase. *Cell*, *150*(6), 1209–1222.

Bullard, J. H., Purdom, E., Hansen, K. D., & Dudoit, S. (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, *11*, 94.

Carey, B. W., Markoulaki, S., Hanna, J. H., Faddah, D. A., Buganim, Y., Kim, J., Ganz, K., Steine, E. J., Cassady, J. P., Creyghton, M. P., Welstead, G. G., Gao, Q., & Jaenisch, R. (2011). Reprogramming Factor Stoichiometry Influences the Epigenetic State and Biological Properties of Induced Pluripotent Stem Cells. *Cell Stem Cell*, *9*(6), 588–598.

Casale, F., Giurato, G., Nassa, G., Armond, J. W., Oates, C., Corá, D., Gamba, A., Mukherjee, S., Weisz, A., & Nicodemi, M. (2013). Single-Cell States in the Estrogen Response of Breast Cancer Cell Lines. *submitted*.

Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, *74*(368), 829–836.

Clifford, P. (1977). Nonidentifiability in stochastic models of illness and death. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(4), 1338–1340.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., & Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, *29*(12), 1120–1127.

de Souza, N. (2012). Single-cell methods. *Nature Methods*, *9*(1), 35.

Debnath, J., Muthuswamy, S. K., & Brugge, J. S. (2003). Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods*, *30*(3), 256–268.

Dehm, S. M., & Bonham, K. (2013). SRC gene expression in human cancer: the role of transcriptional activation. *Biochemistry and Cell Biology*, *82*(2), 263–274.

DeRisi, J. L., Iyer, V. R., & Brown, P. O. (1997). Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale. *Science*, *278*(5338), 680–686.

Dudoit, Sandrine and Yang, Yee Hwa and Callow, Matthew J and Speed, Terence P (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statistica sinica*, *12*(1), 111–140.

Efe, J. A., Hilcove, S., Kim, J., Zhou, H., Ouyang, K., Wang, G., Chen, J., & Ding, S. (2011). Conversion of mouse fibroblasts into cardiomyocytes using a direct reprogramming strategy. *Nature Cell Biology*, *13*(3), 215–222.

Einstein, A. (1905). Über die von der molekularkinetischen theorie der wärme gerforderte Bewegung von in ruhenden. *Annalen der Physik*, *322*(8), 549–560.

Eisen, M. B., & Brown, P. O. (1999). DNA arrays for analysis of gene expression. In S. M. Weissman (Ed.) *cDNA Preparation and Characterization*, vol. 303 of *Methods in Enzymology*, (pp. 179 – 205). Academic Press.

Elgar, G., & Vavouri, T. (2008). Tuning in to the signals: noncoding sequence conservation in vertebrate genomes. *Trends in Genetics*, *24*(7), 344–352.

Elowitz, M. B., Levine, A. J., Siggia, E. D., & Swain, P. S. (2002). Stochastic Gene Expression in a Single Cell. *Science*, *297*(5584), 1183–1186.

Gentile, M., Latonen, L., & Laiho, M. (2003). Cell cycle arrest and apoptosis provoked by UV radiationinduced DNA damage are transcriptionally highly divergent responses. *Nucleic Acids Research*, *31*(16), 4779–4790.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., et al. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *New England Journal of Medicine*, *366*(10), 883–892.

Gibney, E. R., & Nolan, C. M. (2010). Epigenetics and gene expression. *Heredity*, *105*(1), 4–13.

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, *128*(4), 635–638.

Hanahan, D., & Weinberg, R. (2000). The Hallmarks of Cancer. *Cell*, *100*(1), 57–70.

Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of Cancer: The Next Generation. *Cell*, *144*(5), 646–674.

Hanna, J., Saha, K., Pando, B., van Zon, J., Lengner, C. J., Creyghton, M. P., van Oudenaarden, A., & Jaenisch, R. (2009). Direct cell reprogramming is a stochastic process amenable to acceleration . *Nature*, *462*(7273), 595–601.

Hanna, J. H., Saha, K., & Jaenisch, R. (2010). Pluripotency and Cellular Reprogramming: Facts, Hypotheses, Unresolved Issues. *Cell*, *143*(4), 508–525.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer series in statistics. Springer.

Heng, H. H. Q., Bremer, S. W., Stevens, J. B., Ye, K. J., Liu, G., & Ye, C. J. (2009). Genetic and epigenetic heterogeneity in cancer: A genomecentric perspective. *Journal of Cellular Physiology*, *220*(3), 538–547.

Herce, H. D., Deng, W., Helma, J., Leonhardt, H., & Cardoso, M. C. (2013). Visualization and targeted disruption of protein interactions in living cells. *Nature Communications*, *4*.

Hirsch, H. A., Iliopoulos, D., Joshi, A., Zhang, Y., Jaeger, S. A., Bulyk, M., Tsichlis, P. N., Liu, X. S., & Struhl, K. (2010). A Transcriptional Signature and Common Gene Networks Link Cancer with Lipid Metabolism and Diverse Human Diseases. *Cancer Cell*, *17*(4), 348–361.

Hodgkin, A. L., & Huxley, A. F. (1952). A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of physiology*, *117*(4), 500–544.

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, *9*(5), 473–476.

Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Schelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephaniants, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H.,

& Linsley, P. S. (2001). Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnology*, *19*(4), 342–347.

Ieda, M., Fu, J.-D., Delgado-Olguin, P., Vedantham, V., Hayashi, Y., Bruneau, B. G., & Srivastava, D. (2010). Direct Reprogramming of Fibroblasts into Functional Cardiomyocytes by Defined Factors. *Cell*, *142*(3), 375–386.

Iliopoulos, D., Hirsch, H. A., & Struhl, K. (2009). An Epigenetic Switch Involving NF-κB, Lin28, Let-7 MicroRNA, and IL6 Links Inflammation to Cell Transformation. *Cell*, *139*(4), 693–706.

Jacquez, J. A., & Greif, P. (1985). Numerical parameter identifiability and estimability: Integrating identifiability, estimability, and optimal sampling design. *Mathematical Biosciences*, *77*(1-2), 201–227.

Jaenisch, R., & Young, R. (2008). Stem Cells, the Molecular Circuitry of Pluripotency and Nuclear Reprogramming. *Cell*, *132*(4), 567–582.

James, F. (1980). Monte Carlo theory and practice. *Reports on Progress in Physics*, *43*(9), 1145.

Johnson, N. L. (1949). Systems of Frequency Curves Generated by Methods of Translation. *Biometrika*, *36*(1/2), 149–176.

Kalbfleisch, J., & Lawless, J. (1984). Least-squares estimation of transition probabilities from aggregate data. *The Canadian Journal of Statistics*, *12*(3), 169–182.

Kalbfleisch, J., & Lawless, J. (1985). The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, *80*(392), 863–871.

Kalbfleisch, J. D., Lawless, J. F., & Vollmer, W. M. (1983). Estimation in Markov models from aggregate data. *Biometrics*, *9*(4), 907–919.

Lähdesmäki, H., Shmulevich, l., Dunmire, V., Yli-Harja, O., & Zhang, W. (2005). In silico microdissection of microarray data from heterogeneous cell populations. *BMC Bioinformatics*, *6*(1), 54.

Lanza, R., & Atala, A. (2013). *Essentials of Stem Cell Biology*. Elsevier Science.

Lanza, R., Gearhart, J., Hogan, B., Melton, D., Pedersen, R., Thomas, E. D., Thomson, J. A., & Wilmut, I. S. (2009). *Essentials of Stem Cell Biology*. Elsevier Science.

Li, F. P., & Fraumeni, J. F., Jr. (1969). Soft-tissue sarcomas, breast cancer, and other neoplasms: A familial syndrome? *Annals of Internal Medicine*, *71*(4), 747–752.

Li, J., Witten, D. M., Johnstone, I. M., & Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for RNA-sequencing data. *Biostatistics*, *13*(3), 523–538.

Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H., & Brown, E. L. (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech*, *14*(13), 1675–1680.

MacDonald, I. L., & Zucchini, W. (1997). *Hidden Markov and Other Models for Discrete- valued Time Series*. Taylor & Francis.

Martin, G. S. (2001). The hunting of the Src. *Nature Reviews Molecular Cell Biology*, *2*(6), 467–475.

McCarthy, D. J., Chen, Y., & Smyth, G. K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research*, *40*(10), 4288–4297.

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., & Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods*, *5*(7), 621–628.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, *320*(5881), 1344–1349.

Neal, R. M. (2001). Annealed importance sampling. *Statistics and Computing*, *11*(2), 125–139.

Norris, J. R. (1998). *Markov Chains*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Oshlack, A., Robinson, M. D., & Young, M. D. (2010). From RNA-seq reads to differential expression results. *Genome Biology*, *11*(12), 1–10.

Pang, Z. P., Yang, N., Vierbuchen, T., Ostermeier, A., Fuentes, D. R., Yang, T. Q., Citri, A., Sebastiano, V., Marro, S., Südhof, T. C., & Wernig, M. (2011). Induc-

tion of human neuronal cells by defined transcription factors. *Nature*, *476*(7359), 220–223.

Pawlik, T. M., & Keyomarsi, K. (2004). Role of cell cycle in mediating sensitivity to radiotherapy. *International Journal of Radiation Oncology Biology Physics*, *59*(4), 928–942.

Pennisi, E. (2012). ENCODE Project Writes Eulogy for Junk DNA. *Science*, *337*(6099), 1159–1161.

Phimister, B. (1999). The Chipping Forecast.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140.

Robinson, M. D., & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, *11*(3), R25.

Rous, P. (1911). A sarcoma of the fowl transmissible by an agent separable from the tumor cells. *The Journal of Experimental Medicine*, *13*(4), 397–411.

Roy, S., Lane, T., Allen, C., Aragon, A. D., & Werner-Washburne, M. (2006). A Hidden-State Markov Model for Cell Population Deconvolution. *Journal of Computational Biology*, *13*(10), 1749–1774.

Saccomani, M., Audoly, S., Bellu, G., & D'Angiò, L. (2003). Parameter Identifiability of Nonlinear Biological Systems. In L. Benvenuti, A. Santis, & L. Farina (Eds.) *Lecture Notes in Control and Information Science*, (pp. 87–93). Springer Berlin Heidelberg.

Saccomani, M. P., Audoly, S., Bellu, G., & D'Angiò, L. (2010). Examples of testing global identifiability of biological and biomedical models with the DAISY software. *Computers in Biology and Medicine*, *40*(4), 402–407.

Samavarchi-Tehrani, P., Golipour, A., David, L., Sung, H.-K., Beyer, T. A., Datti, A., Woltjen, K., Nagy, A., & Wrana, J. L. (2010). Functional Genomics Reveals a BMP-Driven Mesenchymal-to-Epithelial Transition in the Initiation of Somatic Cell Reprogramming. *Stem Cell*, *7*(1), 64–77.

Sasieni, P. D., Shelton, J., Ormiston-Smith, N., Thomson, C. S., & Silcocks, P. B. (2011). What is the lifetime risk of developing cancer[quest]: the effect of adjusting for multiple primaries. *British Journal of Cancer*, *105*(3), 460–465.

Schwarz, G. (1978). Estimating the Dimension of a Model. *Ann. Statist.*, *6*(2), 461–464.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., & Futcher, B. (1998). Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular biology of the cell*, *9*(12), 3273–3297.

Takahashi, K., & Yamanaka, S. (2006). Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors. *Cell*, *126*(4), 663–676.

Tang, F., Barbacioru, C., Bao, S., Lee, C., Nordman, E., Wang, X., Lao, K., & Surani, M. A. (2010). Tracing the Derivation of Embryonic Stem Cells from the Inner Cell Mass by Single-Cell RNA-Seq Analysis. *Stem Cell*, *6*(5), 468–478.

Tang, F., Lao, K., & Surani, M. A. (2011). Development and applications of single-cell transcriptome analysis. *Nature Methods*, *8*.

Tennyson, C. N., Klamut, H. J., & Worton, R. G. (1995). The human dystrophin gene requires 16 hours to be transcribed and is cotranscriptionally spliced. *Nature Genetics*, *9*(2), 184–190.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288.

Vierbuchen, T., Ostermeier, A., Pang, Z. P., Kokubu, Y., Südhof, T. C., & Wernig, M. (2010). Direct conversion of fibroblasts to functional neurons by defined factors. *Nature*, *463*(7284), 1035–1041.

Vogel, G. (2010). Diseases in a Dish Take Off. *Science*, *330*(6008), 1172–1173.

Wang, D., & Bodovitz, S. (2010). Single cell analysis: the new frontier in 'omics'. *Trends in Biotechnology*, *28*(6), 281–290.

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, *10*(1), 57–63.

Weinberg, R. (2013). *The Biology of Cancer, Second Edition*. Garland Science.

Wheeler, A. R., Throndset, W. R., Whelan, R. J., Leach, A. M., Zare, R. N., Liao, Y. H., Farrell, K., Manger, I. D., & Daridon, A. (2003). Microfluidic Device for Single-Cell Analysis. *Analytical Chemistry*, *75*(14), 3581–3586.

Williams, G. H., & Stoeber, K. (2012). The cell cycle and cancer. *The Journal of Pathology*, *226*(2), 352–364.

Zucchini, W., & MacDonald, I. L. (2009). *Hidden Markov Models for Time Series: An Introduction Using R*. Taylor & Francis.