Understanding mammalian gene expression noise as a function of cell growth

Philip Rhys Davies, BSc (Hons)

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Biology

School of Life Sciences
University of Warwick

May 2020

Contents

List of Figures

Acknowledgements

I would like to thank my supervisors for their invaluable help and guidance, and WISB for the wonderful community, excellent technical facilities and support. My thanks also go to the Synbio CDT directors, who offered me this unique opportunity. Finally, I would like to thank my family and friends for their enduring support the past four years.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree.

The work presented (including data generated and data analysis) was carried out by the author except in the cases outlined below:

- 4sU sequencing library preparation and sequence processing was carried out by Dr Mark Walsh
- Density peak identification script was written by Dr Massimo Cavallaro

Abstract

Mammalian cell growth is a complex process encompassing genome replication, cell mass accumulation and drastic reorganization of the intracellular structure. Each of these processes contributes to gene expression noise, making the design of robust genetic circuits difficult. We use flow cytometry to collect measurements on multiple cell-cycle reporters, which we analyse using probability state modelling. This approach enables us to position each cell into its most likely cell-cycle state. Combining this methodology with measurements of gene expression kinetics such as metabolic labelling of transcription and translation provides a high resolution view into how such activities change during the cell-cycle. Changes in cell-size are another important source of gene expression variation. Cells of different sizes are known to grow at different rates, further confounding our measurements of noise. Using ergodic rate analysis, we correlate our measurements of gene expression kinetics with those of cell-size growth rate as a function of cell-cycle progression. This way, we aim to elucidate the homeostatic mechanisms linking cell growth and gene expression, in order to better understand gene expression noise.

Chapter 1

Introduction

Gene expression is inherently a noisy process as the biochemical reactions underlying it depend on low numbers of molecules, which leads to thermal noise playing a large role. Even at its simplest definition, whereby the RNA polymerase transcribes DNA into RNA, the interaction of the RNA polymerase with the DNA template at the gene locus is inherently stochastic and governed by Brownian motion (Eldar and Elowitz, 2010).

Many genes show a higher degree of noise than what the above model predicts, resulting in over-dispersed mRNA and protein distributions when measured in single cells. It has previously been proposed that this phenomenon is due to transcription bursting, defined by sudden, short-lived increases in transcriptional output, which can be shown analytically to lead to super-Poissonian distributions (Paulsson, 2005). Such effects have been observed for certain genes, including in live cells (Golding et al., 2005).

These observations have been suggested to be due to changes in the state of the gene's promoter, and can be modeled as a random telegraph process Paulsson (2005). Specifically, the promoter is modeled as having two states (on-off), the switching between which results in a bursty transcriptional output. The magnitude and the frequency of these bursts depends on the stochastic rates associated with the transitions between the two states, as well as the transcription rate itself. This analytical framework has been used extensively to describe and analyse the distributions of proteins and mRNA seen in cells (Sun et al., 2019a).

On the mechanistic level, the two promoter states are suggested to reflect the stochastic binding of transcription factors, which increase the affinity of the promoter to the transcriptional machinery, and by extension increase the probability of transcription initiation and re-initiation. The exact mechanism of promoter state-switching is not clear (Hebenstreit, 2013), and, importantly, its relative noise contribution is not fully understood. On the other hand, the relevance of gene expression noise in biology has become increasingly clear over the past two decades, with many examples of organisms exploiting noise arising at both the single cell and multicellular level, as well as how noise can in many cases become an obstacle that organisms have evolved to overcome. An overview of recent examples is given in the following sections.

1.1 Beneficial noise in microorganisms

Noise can be beneficial in the survival of organisms. At the single cell level, it allows clonal populations to form phenotypically distinct subpopulations, either as a form of division of labour, which can be beneficial to the colony, or as a bet hedging strategy, which can prove useful in fluctuating environmental conditions (Losick and Desplan, 2008). Several examples exist where such strategies are implemented.

While bistability can be encoded in a genetic circuit, the existence of noise makes such a design redundant. Specifically, To and Maheshri (2010) showed that bimodality in a population can be generated using bursty expression of an unstable autoregulatory transcription factor, without bistability explicitly encoded. In *Bacillus subtilis*, stochastic expression of the autoregulatory protein ComK enables a small fraction of the population (between 10 and 20%) to become competent, in other words to take up DNA from the environment (Maamar et al., 2007). Although this DNA is not necessarily advantageous, the chance of it providing a competitive advantage makes sacrificing a fraction of the population in this way an effective mechanism for increasing the chances of survival of the population.

More recently, Mugler et al. (2016) showed that noise can expand the dynamic range of the competence stress response, further highlighting the use of noise. In a similar vein, the differentiation of *B. subtilis* into motile and non-motile cells is governed by the stochastic expression of *sigD*. Interestingly, Cozy and Kearns (2010) found that the relative location of *sigD* on the motility operon determined the fraction of the population developing motility. This indicates how stochasticity can be used by biology as an evolvable trait.

A similar strategy is employed in yeast sporulation, a survival mechanism triggered by nutrient deprivation. As timing the initiation of sporulation can have important consequences on the survival of each cell, a degree of variation in this timing would be advantageous on the population level. Indeed, the initiation timing of meiosis prior to sporulation has been shown to be highly variable, and dependent on the stochastic expression of master regulator Ime1 (Nachman et al., 2007). Zhao et al. (2019) verified the high temporal variation in the sporulation response using a microfluidic device, enabling them to study the expression of multiple key genes of the sporulation pathway.

Other such examples exist in the activation of alternate metabolic pathways, in the face of fluctuations in the availability of nutrients. Stochastic switching of the *lac* operon has been observed in *Escherichia coli* (Mettetal et al., 2006; Ozbudak et al., 2004), as well as the galactose utilisation pathway in yeast (Acar et al., 2005). More recently, Ge et al. (2018) demonstrated analytically that the bursty activation of the *lac* promoter extends the range of lactose concentration at which bimodality exists, allowing cells which do not express the *lac* operon to persist in the presence of lactose. While the fraction of the population expressing the wrong metabolic pathway will have sub-optimal growth, such a strategy avoids the necessity to constantly sense changes in the environment, and has been shown to be a viable survival method in fluctuating environmental conditions (Thattai and van Oudenaarden, 2004). Recently, an analytical framework has been presented to study the effects of stochastic gene expression on metabolism (Tonn et al., 2019).

Another interesting observation linked to stochastic gene expression has been the specialisation of cells within a population to different nutrients. Using nanometer-scale secondary ion mass spectrometry (NanoSIMS) to measure the uptake of two different isotope-labelled sugars (arabinose and glucose) in single cells, Nikolic et al. (2017) found that there was a strong variation between single cells in their consumption of the two different sugars. As suggested, such specialisation in the metabolic pathway could enable certain metabolic reactions to be performed more efficiently. Stochastic expression of key metabolic enzymes can have profound effects on cell growth. Kiviet et al. (2014a) showed that in the case of growth limiting genes, stochasticity in the expression of even a single catabolic enzyme can propagate to the cell's growth rate. Counter-intuitively, heterogeneity within a population's growth rate has been shown to increase the average growth rate overall, even in the absence of environmental stresses (Hashimoto et al., 2016).

Finally, stochastic gene expression has been suggested to be implemented in the survival of pathogens. Small numbers of persister cells exist even in untreated populations of *E. coli*, and are generated continuously during growth (Balaban et al., 2004). HIV latency on the other hand is caused by stable integration of the virus into a small population of CD4 T cells, with burst of expression arising from positive feedback of Tat protein stochastic expression (Weinberger et al., 2008).

1.2 Beneficial noise in multicellular organisms

Noise can be a driving force of cellular diversity in multicellular organisms too. There are different ways in which gene expression noise can lead to intracellular phenotypic diversity in multicellular organisms. First, noise in gene expression can aid in the differentiation of stem cells into different tissues. Some of the first discovered examples include the development of the olfactory sensory system in mice, whereby thousands of neurons are involved, each of which is required to have a distinct odour receptor (Tsuboi et al., 1999). While a genetic circuit enabling this would be prohibitively complex, relying on gene expression stochasticity is sufficient. Differentiation of the various muscle fiber types constituting the muscles of vertebrates has also been suggested to rely on expression noise in a similar way (Hughes and Salinas, 1999).

Another example can be drawn from human vision. Trichromacy is a common trait among primates which refers to the presence of three distinct opsins (photo-pigment proteins) for detecting red, green and blue colours, only one of which is expressed in each cone cell of the retina (Johnston Jr. and Desplan, 2010). A random distribution of cells expressing each opsin is required in order to confer full colour vision (Roorda and Williams, 1999). Selection of which opsin each cell produces is a result of a two stage mechanism, both of which steps rely on stochastic gene expression. In Old World primates specifically (including humans), the first step determines the fate between blue and red/green opsins, while the second determines the expression of either the red or green opsins via random selection of one of the two respective alleles. Due to the presence of two copies of each gene per cell in diploid organisms, ensuring that only a single opsin is expressed via the above mechanism would not be possible without inactivating the second allele of each opsin gene. As these genes are located on the X chromosome (Nathans, 1999), this is achieved in females by X chromosome inactivation, and in males by virtue of having only one X chromosome. This suggests that the location of genes can evolve in order to take advantage of noise, leading to beneficial phenotypic mosaicism.

A similar strategy is followed during the development of the compound eye in *Drosophila*, where the appropriate ratio of blue versus yellow-sensitive photoreceptors is determined by the stochastic expression of the gene *spineless* (Wernet et al., 2006). Such expression of *spineless* results in a random distribution of ‘pale’ and ‘yellow’ omatidia, resulting in an uneven ratio of 35:65 on the developed compound eye, a ratio known to be conserved amongst other flies (Franceschini et al., 1981), suggesting it is of functional importance. These examples demonstrate how stochasticity has been exquisitely harnessed by evolution for the development of complex traits such as vision.

Gene expression stochasticity may explain aspects of development more generally, with cell-to-cell variability within a population of stem cells affecting their response to differentiation stimuli, as shown for neuron differentiation (Shah et al., 1996). In mammalian blood differentiation, variability in the expression of the Sca-1 protein is correlated with the probability of choosing either the eryhtroid or myeloid lineage (Chang et al., 2008). Similarly, during the development of the mouse embryo, differentiation of the inner cell mass into the epiblast and primitive endoderm appears to be dependent on the stochastic expression of either Nanog or Gata6 (Dietrich and Hiiragi, 2007). These examples illustrate how stochastic variability can be an effective mechanism for driving differentiation programs during development.

More recent examples include the development of the amoeba *Dictyostelium discoideum*, in which variability has been demonstrated to increase throughout development, and then decrease once cells become terminally differentiated (Antolovic et al., 2017). Similar observations have been made in studies of haematopoeitic progenitor differentiation (Richard et al., 2016), suggesting that noise is harnessed as a mechanism during development and differentiation. While the direction of causality is not yet entirely clear in the relationship between noise and cell-fate decision making, the role of noise has been more clearly demonstrated in the immune system.

As well as providing a mechanism for differentiation of cells into different tissues, noise can lead to phenotypic diversity between cells of the same tissue, a phenomenon termed 'mosaic physiology' (Woods, 2014). This effect can expand the dynamic range of homeostatic responses, thus making the organism more resilient to environmental challenges. For example, as well as enabling greater diversity in the type of targets the immune system can respond to (Schrom and Graham, 2017), variability in gene expression further enables tuning the magnitude of the immune response. Specifically, it has been shown that stochastic on/off switching of IL-2 expression in T_h cells upon immunisation is required in order to ensure a wide-range linear response to the antigen strength (Fuhrmann et al., 2016).

1.3 Detrimental noise in microorganisms

While noisy expression of stress response genes has been shown to be beneficial, one would expect that variability in the expression of genes that are essential to the function of the cell would be more tightly controlled. This was found to be the case in two early genomic studies in yeast (Newman et al., 2006; Bar-Even et al., 2006), both of which found that essential genes (lethal when deleted) such as those related to protein synthesis and degradation were significanlty more precisely controlled than non-essential genes. While these findings are not direct evidence for the existence of noise minimisation mechanisms in yeast, Lehner (2008) showed that an independent set of dosage sensitive genes (lethal when over-expressed), also showed significantly lower noise, thus corroborating the findings of Newman et al. (2006).

Furthermore, Lehner (2008) hypothesised that the existence of noise minimisation mechanisms for certain genes would also have a knock on effect on the capacity of mutations to alter the expression level of these genes on the genetic level. In turn, this could slow down the rate at which the expression level of such genes evolves. By comparing the divergence of genes with varying noise levels between different yeast species (Tirosh et al., 2006), this was found to be the case (Lehner, 2008). These studies further support that gene expression noise has been widely minimised in yeast by natural selection for dosage-sensitive genes. Furthermore, it illustrates that the pressure to minimise gene expression noise of vital genes can limit the rate at which the expression level of these genes can evolve in a population, thus limiting the latter's capacity to adapt to changing environments over time.

Further to these empirical findings, recent analytical treatments have found a theoretical basis for how gene expression noise of essential genes can lead to a reduction in the fitness of an organism. Specifically, Wang and Zhang (2011) highlighted three mechanisms by which we can expect gene expression noise to limit the growth rate of a microbial population. Firstly, fluctuations in the relative concentrations of enzymes in a given metabolic pathway leads to sub-optimal metabolic flux and thus a reduced rate of biomass production. Second, unnecessary production of protein due to stochastic fluctuations in transcription and translation can be costly in terms of the cellular energy budget (Wagner, 2005), thus limiting the overall growth rate. Thirdly, the correct assembly of certain protein complexes depends on the concentrations of their constituent sub-units being in the appropriate ratios (Lehner, 2008). Noisy expression of these sub-units can affect these ratios in potentially detrimental ways (Fraser et al., 2004). Using mathematical modelling, Wang and Zhang (2011) showed that via the above three mechanisms gene expression noise can decrease the fitness of cells by over 25%, thus playing a fundamental role in evolution.

In a similar vein, van Dyken (2017) highlighted that biochemical reactions are predicted, based on Jensen's inequality, to show a decreased rate of conversion from substrate to product in the presence of noise. Briefly, as the Michaelis-Menten (MM) equation describing the relation between substrate concentration and rate of product formation is a convex curve, mapping a distribution of substrate concentrations on this curve results in a positively skewed distribution of reaction rates. This means that in the presence of noise, the mean steady state reaction rate is lower than the noise free equivalent, an effect known as “stochastic slow-down” of MM reactions. Variation in substrate concentrations among a population can result from the intrinsic noise in the expression of upstream metabolic enzymes. Stochastic slow-down can therefore decrease the rate at which a population of cells can metabolise a substrate in order to grow and proliferate.

There are different mechanisms that could potentially be employed for controlling noise in gene expression. Evidence for natural selection of such systems further supports that gene expression noise can be detrimental. Thattai and van Oudenaarden (2001) noted that there are three qualitatively different modes of controlling protein levels, each of which is associated with different levels of noise: high transcription coupled with low translation is shown to minimise noise, while either low transcription coupled with high translation and intermediate levels of transcription and translation are both associated with higher levels of noise. Fraser et al. (2004) used measurements of transcription and translation for all genes in yeast (Blake et al., 2003) to show that essential genes are biased towards the first strategy, which could reveal one of the mechanisms by which such genes have evolved to minimise noise. More recently, Chen and Zhang (2016) integrated a GFP gene cassette at 482 different locations of the yeast genome, and found that positioning can have a 15 fold effect on gene expression noise. Furthermore, they found that regions associated with lower noise were enriched for essential genes, suggesting that chromosome organisation is another way natural selection has acted to minimise unwanted gene expression noise.

Finally, there is evidence implicating the evolution of certain regulatory networks in order to minimise noise. The bacterial chemotaxis regulatory network appears to have evolved to minimise noise perturbations (Kollmann et al., 2005). Specifically, multiple regulation models were proposed which varied in their tolerance to noise, while exhibiting the same adaptive capacity to local food signals. After testing their tolerance to noise, the authors found that the most resilient to perturbations model was the one which most closely matched the natural chemotaxis topology in *E. coli*. This suggests that excessive noise in an essential function such as chemotaxis is detrimental, forcing regulatory networks to evolve in such a way that minimises it. Similar noise mitigation mechanisms have been discovered in the yeast mating pheromone response pathway (Colman-Lerner et al., 2005). The authors found that, although cells vary in their capacity to transmit a signal through the pheromone response pathway, the target gene of this pathway also varied in its capacity to be expressed. Furthermore, these two sources of variability were negatively correlated, allowing cells to respond robustly to pheromone signals in spite of noise variability in the transduction of the signal.

1.4 Detrimental noise in multicellular organisms

Both direct and indirect evidence for the detrimental effects of noise in multicellular organisms exists. Direct evidence can be found in cases where noise suppression mechanisms have been destabilised, leading to pathogenic phenotypes. Indirect evidence can be found in the presence of systems likely to have evolved to suppress noise.

Battich et al. (2015) used a combination of single molecule RNA counting and computer vision to analyse hundreds of single mammalian cells grown in culture. They found that cytoplasmic variation in mRNA counts was largely suppressed, leading to count distributions far narrower than the bursty kinetics of transcription would predict. Using mathematical modelling, the authors showed that a time delay of ~15 minutes between transcription and export of mRNA from the nucleus can reduce the transcriptional noise in the cytoplasm by ~57%, thus posing nuclear retention as an effective mechanism for buffering noise. These results were validated experimentally by overexpressing NUP153, a protein known to decrease the rate of mRNA nuclear export, which produced a reduction in cytoplasmic RNA noise. Bahar Halpern et al. (2015) made similar conclusions in measurements done in cells from a variety of mouse tissues.

Other mechanisms which have been suggested to play a role in mitigating the negative effects of gene expression are regulatory network topologies, such as negative feedback loops. Transcription factors, fluctuations of which are likely to propagate downstream regulatory pathways, are a prime example, many of which are known to repress their own expression via negative regulation of their own genes (Rao et al., 2002). Many non-transcription factor proteins are also implicated in negative feedback loops via co-transcribed microRNAs, which regulate the steady state expression level of the mRNAs they target, as well as suppress fluctuations in mRNA counts (Tsang et al., 2007). Another mechanism which has been implicated in noise reduction is the evolution of polyploidy, as an increase in the number of genes leads to an averaging effect over the bursty effects of transcription (Pires and Conant, 2016).

While noisy gene expression can be useful during multicellular differentiation, development and immunity, it has been shown that the magnitude of the variability as well as the timing during these processes are tightly regulated. For example, while stochastic transcription causes large differences in transcript levels between cells during early zebrafish embryogenesis, changes in cell cycle duration and mRNA half-lives lead to a decrease in gene expression noise at later stages due to increased temporal averaging (Stapel et al., 2017). Similarly, the presence of feedback loops in genetic network in *C. elegans* buffer expression variability of master regulator *elt-2*, which controls key developmental genes (Ji et al., 2013), and removal of this feedback loops leads to bimodal expression of *elt-2* (Raj et al., 2010). The existence of such mechanisms suggests that noise can also be deleterious during the execution of developmental programs in multicellular organisms.

Indeed, in many cases loss of control in gene expression variability has been linked to ageing, as well as the onset of cancer. Specifically, increased transcriptional variability during ageing has been observed in the human pancreas (Enge et al., 2017), cells of the peripheral immune system and lung tissue (Cheung et al., 2018; Angelidis et al., 2019), to the extent that it is now considered a biomarker for ageing (Lu et al., 2016). Barkai and Leibler (2000) used simulations to demonstrate that, while circadian clocks can function robustly in the presence of gene expression noise, the presence of the latter poses strict limitations on the underlying regulatory gene networks. As noise suppression mechanisms weaken with age, noise can compromise the robustness of circadian rhythms, which in turn can degrade many vital rhythms such as sleep/wake patterns (Hood and Amir, 2017).

Noisy gene expression has also been shown to facilitate the transition of a cell from a healthy state to a cancerous one (Jia et al., 2017). Furthermore, once cells have entered a malignant state, noise in epigenetic reprogramming can act to reinforce that state (Shaffer et al., 2017). Finally, phenotypic variability in cancers arising from noisy expression has been shown analytically to enable resistance to chemotherapy (Schuh et al., 2019). Similar effects have also been shown experimentally, with upregulation of p53 upon treatment with anticancer drug cisplatin leading to apoptosis depending on both the level and timing of the upregulation. Heterogeneity in either of these leads to a heterogeneous response to the drug (Paek et al., 2016).

1.5 Noise and cell growth

The contribution of extrinisic sources to gene expression noise has been studied closely in the past, starting with the famous dual-reporter system presented by Elowitz et al. (2002). In brief, the levels of two genomically integrated proteins are measured in single cells using fluorescent protein fusions. Correlated fluctuations in the levels of the two proteins represent the effects of extrinsic sources of noise, while uncorrelated fluctuations reflect the effects of intrinsic sources, as formalised by Swain et al. (2002). Intuitively, the two types of sources can be measured using a covariance plot of the two fluorescent proteins as the spread along the diagonal (extrinsic), and perpedicular to it (intrinsic).

An alternative approach to measuring intrinsic gene expression noise is by obtaining a morphologically homogenous cell population, in order to minimise the contribution of extrinsic sources. Such a method was suggested by Newman et al. (2006), who used flow cytometry to filter cells based on morphological properties (cell size and granularity) revealed by flow cytometry light scatter. By varying the stringency of the filter, the authors were able to identify a threshold which minimises the contribution of noise that can be accounted for by variability in cell morphology, as demonstrated more recently by Meng et al. (2017).

Evidence in animal cells suggests that the majority of cytoplasmic mRNA variation can be predicted by taking into account various aspects of the state of each single cell (Battich et al., 2015). Specifically, Battich et al. (2015) noted that when taking every cell's micro-environment, cell size, cell-cycle state, as well as 180 other image-based variables relating to the cell state, gene expression was found to be minimally stochastic in the cytoplasm. Similar suggestions have been made in earlier studies (Raser and Shea, 2006; Snijder and Pelkmans, 2011). This suggests that cell growth may play a larger role in the gene expression noise of mammals than previously considered, as proposed earlier on (Maheshri and O 'shea, 2007).

Gene expression noise and organism growth are intimately linked, both at the single cell (Kiviet et al., 2014b) and multicellular level (Paldi, 2003). At the single cell level, fluctuations in the concentration of metabolic enzymes due to stochasticity in gene expression affects the growth rate of individual cells. Furthermore, the random partitioning of molecules upon cell division (Huh and Paulsson, 2011b) leads daughter cells to embark on the next cell cycle with different starting conditions, further contributing to this effect.

Variation in cell growth rate in turn is strongly correlated with variation in cell size. In the case of microorganisms, cells are generally larger at faster growth rates, while mammalian cells in culture appear to have an optimal cell size nearer the middle of the size range (Miettinen and Björklund, 2016). In all cases, increases in cell size and division rate are accompanied by proportional increases in gene product generation (Kempe et al., 2015, Padovan-Merhar et al. (2015)), a requirement for maintaining the working concentration of the intracellular machinery responsible for cellular functions (Pérez-Ortín et al., 2019a).

Not taking these effects into account when counting individual transcripts or protein molecules in single cells has been shown to lead to an overestimation of gene expression noise (Battich et al., 2015; Kempe et al., 2015), especially in mammalian cells grown in culture, which have been shown to vary up to six-fold in volume (Tzur et al., 2009). On the other hand, the above homeostatic mechanisms can lead to actual changes in gene expression noise. For example, the higher division rates mentioned above increase the contribution random partitioning of molecules has on noise. Conversely, theoretical analysis has shown that preservation of noise homeostasis could be another driving force explaining changes in both cell size and growth rate (Bertaux et al., 2018).

Due to the complex circular causalities underpinning these effects, it is not possible to understand gene expression noise outside the context of cell growth, and vice versa. To disentangle these relationships, multiple simultaneous measurements have to be performed on single cells, such as measurements of cell size, growth rate, and gene expression dynamics, combined with novel mathematical frameworks for analysing the resulting data.

Such experiments have proven challenging to perform. This is particularly true in mammalian cells, due to the higher cell-cycle complexity and longer generation times. Recent advances in imaging, computer vision and live cell cycle tracking have enabled studies in cell size vs cell cycle and revealed many aspects of cell size regulation by cell cycle (Son et al., 2012; Miettinen et al., 2019). Others have looked at how cell cycle correlates with gene expression (Skinner et al., 2016; Hausnerová and Lanctôt, 2017). Other studies have revealed correlations of cell size with gene expression (Kempe et al., 2015; Schmidt and Schibler, 1995), but very few have looked at all three simultaneously (Padovan-Merhar et al., 2015), and none which correlate all these measurements together, alongside cell growth rate.

This is understandable, as the best available methods for measuring gene expression currently require that cells are either fixed (single molecule *in situ* fluorescent hybridisation, smFISH) or lysed (single cell RNA sequencing, scRNA-seq), making them difficult to correlate with kinetic experiments of cell growth. Recent analytical methods have enabled the extraction of growth dynamics from fixed data (Kafri et al., 2013), based on the fact that cells in an exponentially growing population are at a quasi-steady state, meaning that the relative fraction of cells at each cell-cycle state is constant, even though the overall number of cells is increasing exponentially. This allows one to infer the rate of transition between different cell states, such as cell size, by measuring the fraction of cells at each state. Coupled with independent cell-cycle measurements, the authors showed it was possible to measure the growth dynamics of cells with respect to cell-cycle progression by imaging fixed cells.

In more recent years, cell-growth in mammalian cells has been directly measured using time-lapse microscopy and highly sensitive size measurements, such as the suspended microchannel resonator (Son et al., 2012). Of note is the MS2 method (Golding and Cox, 2004), which allows the real-time measurement of transcript generation, as well as the more recent translation equivalent (Yan et al., 2016). Nevertheless, the analytical method presented in (Kafri et al., 2013) remains valuable, as it enables the measurement of cell growth kinetics in concurrence with other measurements which cannot be currently made in live cells, such as gene expression by smFISH or global transcription and translation rates by metabolic labelling (Larsen et al., 2001; Marciano et al., 2018).

Furthermore, the ability to extract the relevant information from fixed cells enables the use of flow cytometry for data acquisition, which is a well established, high-throughput method capable of analysing millions of single cells. This level of sampling depth is currently not possible using time-lapse microscopy, which can be a limiting factor when using the inherent heterogeneity of a population for comparing the effect different cell sizes have on gene expression, as a large overall sample is required to obtain robust statistics on rare cell states. A limitation of flow cytometry is that the resulting information is not time resolved, so fluctuations in gene expression cannot be tracked in real time (Simpson et al., 2009). Progress in imaging analysis automation, combined with microfluidics, has shown that the advantages of imaging can be made high throughput, especially for prokaryotes and yeast which can be easily grown in suspension (Groisman et al., 2005).

1.6 Outline

Here, we develop experimental, computational and mathematical methods for uncovering the relationships between cell cycle, cell size and RNA dynamics. Specifically, a model of gene expression and cell growth is formulated in Chapter 3, and an algorithm is developed for resolving the cell cycle in experimental data in Chapter 4. In Chapter 5, the same algorithm is used to observe the effects of cell cycle and cell size has on the rate of transcription, while the foundation is laid for a high throughout transcriptomic metabolic analysis of mRNA in Chapter 6.

Chapter 2

Theory, materials and methods

2.1 Flow Cytometry

Flow cytometry is a fluidics technique which relies on lasers exciting the sample and measuring its emission, enabling the high throughput single cell measurement of multiple variables simultaneously. Specifically, flow cytometers use hydrodynamic focusing to generate a narrow stream of single cells directed past a combination of lasers of different wavelengths, each of which is followed by a combination of bandpass filters and photonmultiplier tubes (PMTs). The PMTs measure the fluorescent content of cells by converting the detected photons to electrons, while the bandpass filters define the spectral detection window. As each cell crosses the laser beam, any fluorescently active molecules inside the cell or on the cell surface become excited and emit light at a wavelength specific to the type of fluorophore, the intensity of which is measured by the relevant PMTs.

Flow cytometry has been used extensively in immunology for characterising cell types in immunology, by targeting cell-type specific epitopes with fluorescently conjugated antibodies (Adan et al., 2017), while an array of different chemical dyes exist for measuring different aspects of cellular physiology, such as total DNA, RNA or protein content, intracellular calcium levels, pH and others. One of the main advantages of flow cytometry is the ability to measure multiple fluorophores in parallel, with modern cytometers having >15 different channels. This allows for investigating the correlations between multiple biological factors, such as cell cycle phase, which can be measured by a DNA stain such as Hoechst, and RNA kinetics, which can be measured using metabolic labelling. Furthermore, it allows studying other cellular parameters such as cell size and inner structure, based on light scatter patterns (Zuleta et al., 2014). This makes it an ideal platform for investigating intrinsic and extrinsic contributions to gene expression, as demonstrated by Newman et al. (2006).

2.2 Cell Cycle Analysis

In Chapters 4 and 5 we use the genetically engineered *fucci* cells (*f*luorescent *u*biquitination-based *c*ell *c*ycle *i*ndicator), a Hela cell line introduced by Sakaue-Sawano et al. (2008) to make the study of cell cycle related processes more straightforward.

These cell lines contain red (mCherry) and yellow (Venus) fluorescent protein reporters fused with degron tags derived from the cell cycle regulator proteins Cdt1 and geminin respectively, stably incorporated into their genome. These degron tags render the stability of the fluorescent proteins cell cycle dependent. As the level of the resulting protein fusions can be readily quantified by measuring their fluorescence intensity, cells can be broadly categorised into three different cell cycle phases, namely early G1 (eG1), the G1/S transition and S/G2/M phases according to the presence of Cdt1, Cdt1 and geminin or geminin only, respectively. These reporters in combination with mathematical modelling have proven very useful in cell cycle research (Saitou and Imamura, 2016).

The *fucci* reporters are complementary to the classic cell cycle analysis conferred by DNA stains such as Hoechst. When combined, they can resolve the cell cycle into at least four phases, namely eG1, the G1/S transition, S phase and G2 phase.

Although the cell cycle can be subdivided in many more stages using antibodies (Avva et al., 2012), the progression of fucci cells can be monitored while the cells are alive, which makes it possible to study cell cycle related events at real time using time-lapse fluorescent microscopy. Furthermore, not needing to fix and permeabilise the cells, which are requirements for immunostaining, make *fucci* cells a superior alternative for cell cycle analysis, especially when maintaining the integrity of the cell morphology is important.

2.3 Metabolic Labelling

2.3.1 Total RNA

Metabolic labelling of transcription relies on the incorporation of a chemically modified nucleotide into elongating RNA, which can subsequently be detected by flow cytometry either using fluorescently labelled antibodies (Larsen et al., 2001) or by covalently attaching a fluorescent group directly. We choose to measure RNA transcription by metabolic labelling with 5-ethynyl-uridine (5EU), as it does not require antibodies and has been demonstrated to be superior to the original bromouridine (BrU) analogue in terms of detection (Jao and Salic, 2008). Instead, the alkene group on 5EU readily reacts with azide groups ('click chemistry'), and this can be used to fluorescently label 5EU with the fluorophore Cy5 after incorporation.

Metabolic labelling can be used for measuring the *in vivo* rates of transcription and RNA degradation. This is a technique whereby cells are grown for a set amount of time in the presence of chemically modified ribonucleotides, which become incorporated in the elongating RNA molecules (Rabani et al., 2011). Newly transcribed RNA can thus be labelled and subsequently detected and quantified, providing a more direct method for measuring the rate of transcription and RNA degradation than the steady-state RNA number can provide, via classical RNA sequencing.

As the vast majority of RNA in a cell consists of ribosomal RNA (rRNA), one notable limitation of the above approach is that measurements of global transcription rate are likely dominated by rRNA kinetics. A modern alternative method of metabolic labelling based on sequencing has been proposed, which addresses this issue and permits quantification of mRNA kinetics for every species individually (Herzog et al., 2017), as discussed in Section 2.3.2.

For the total RNA metabolic labelling, the method is outlined here. The experiment was performed in triplicate. Cells were seeded overnight at a density of 500,000 per well in a six well dish. The next day, cells were labelled for 1 hour with 1mM 5EU in DMSO, detached by trypsinisation, washed twice (10 seconds at 10,000 rpm at 4C) in cold PBS, fixed in 4% paraformaldehyde for 15 minutes at room temperature and then o/n at 4C. The next day, cells were permeabilised with 0.2% Triton X for 20 minutes and 0.5 % for 10min. Cells were then labelled for 1 hour in Cy5 click reaction solution (Jena Bioscience), followed by 2min at 37 shaking. Cells were washed 2x in 0.05% Triton X, and once in 3% BSA for 20min. Cells were spun down and resuspended in 10mg/ml Hoechst DNA stain for 1 hour. Cells were spun down and resuspended in PBS, and left overnight at 4 degrees C. Cells were analysed the following morning by flow cytometry.

2.3.2 Gene specific

The slam-seq protocol relies on the fundamental principle that the uracil analogue 4-thiouridine (4sU) gets incorporated during the synthesis of nascent RNA, and then can be subsequently converted to cytidine following a chemical step, which can ultimately be detected by RNA sequencing, thus enabling us to identify newly transcribed RNA (Herzog et al., 2017; Schofield et al., 2018). By incubating the cells in 4sU for a specified length of time, we can obtain the change in nascent RNA with respect to time, and thus get the rates of RNA synthesis and decay. The advantage of using a sequencing based method is that the rates can be obtained for every single gene individually, allowing for a more complete picture of RNA kinetics to be achieved.

As the rate of 4sU incorporation in newly transcribed mRNA is quite low (Russo et al., 2017), and there is a background level of T to C mutations which varies according to the sequencing method, inferring the true fraction of new to old mRNA is not trivial. The efficacy of the inference will depend on the size of the labelled fraction, the associated error rates and the read count associated with each transcript. In order to maximise the number of genes for which we can robustly estimate the labelled RNA fraction, we need to understand these relationships and plan the experiment accordingly. A simple simulation outlined by Baptista and Dölken (2018) can help us interpret the effect that these parameters have on the quality of the data. Such an approach is employed in Section 6.2.

For the metabolic sequencing experiment in this thesis, the method is oultined here. Cells were seeded in 10cm dishes at a 50% confluence over night. The next day, cells were treated with either 0.5mM or 1mM of 4sU dissolved in DMSO, for times ranging between 10 minutes and 1 hour. Negative controls were treated with DMSO only. Cells were detached by trypsinisation using TrypLE (ThermoFisher) for 5min at 37C. Once in suspension, cells were transferred to a 15ml Falcon tube and fixed directly by adding 10 microliters of 50mg/ml reversible crosslinker dithio-bis(succinimidyl propionate) (DSP) in DMSO, drop-wise while vortexing, similar to (Attar et al., 2018). The cells were incubated in fixative for 15 minutes at 37C, then pelleted gently at 500g for 5 minutes, followed by washing in PBS to remove traces of trypsin and fixative. Cells were subsequently sorted using FACS (±UV laser). Crosslinking was reversed using 50mM DTT prior to nuclear fractionation, which was carried out according to Nabbi and Riabowol (2015). RNA was purified from isolated nuclei, which was subsequently treated with IAA (Herzog et al., 2017) prior to library preparation (SMART-Seq Stranded) with ribodepletion. Libraries were sequenced using Next Generation Sequencing with paired end, 150 length reads, on a HiseqX10 lane (Omega Bioservices). Alignement and T to C conversion counting were performed by Dr Mark Walsh using STAR (Dobin et al., 2012).

2.4 Ergodic Rate Analysis

Ergodic rate analysis (ERA) is a method introduced by Kafri et al. (2013) for extracting the rate of change of any measurement from static data. It relies on two assumptions. First, that a growing population of cells is in a quasi-steady state, whereby the fraction of cells at any cell cycle stage remains constant, even though the total size of the population is growing exponentially. There is therefore a balance between the rates of cells entering and leaving a given state and the number of cells in that state. Second, the determinants of the cell cycle position accurately describe both individual and collective cell behaviour. By relying on these assumptions, it is possible to calculate the rate at which different biological traits change. Kafri et al. (2013) use this method to measure the difference in cell growth between cells of different sizes. Specifically using the above principles, they developed the mathematical framework

$$v(s, \Delta_l) = \frac{\alpha(2-\lambda_A)F(s|l-w) - (2-\lambda_A-\lambda_B)F(s|l+w) - F(s|\Delta_l)\lambda_B}{f(s|\Delta_l)\lambda_B}, \tag{2.1}$$

where s is the cell size, l is the stage in the cell cycle, $F(s|l-w)$ is the cumulative distribution function (CDF) of cell sizes among cells in the state $l-w$ at the entrance to the interval $\Delta_l = (l-w, l+w)$, shown in Figure 2.1 in red (leftmost). The width $\Delta_l = 2w$ represents the resolution limit of the calculation, were w needs to be appropriately chosen, shown in Figure 2.1 in blue. For the calculations in Section 5.4, 2w is chosen to be 10% of the length of the cell cycle, which represents the mean resolution obtained by PSM. $F(s|\Delta_l) = F(s|l-w, l+w)$ is the size distribution within the interval Δl, λ_A is the fraction of cells occupying all cell cycle stages preceding Δl, and λ_B is the fraction of cells in Δl. $\alpha = \frac{\ln(2)}{\tau}$, where τ is the period of the cell cycle. $f(s|\Delta l)$ is the density of cells with size s in the interval Δl.

Once these statistics have been obtained, Equation (2.1) is used to calculate the rate of cell size change at a given cell cycle stage. By applying this calculation throughout the cell cycle, Kafri et al. (2013) were able to compare the growth rate at different phases. To verify the utility of the method presented in Kafri et al. (2013), I tested it on synthetic data based on an imaginary function of cell size. The results of the analysis is seen in Figure 2.2. Although the calculated rates (red arrows) match well with the *true* trajectories (blue lines), the accuracy of the result depends on the size of the sample and the choice of the calculation interval w, which in turn depends on the resolution of the cell cycle. It would be useful in the future to perform further tests on the limitations of the method.

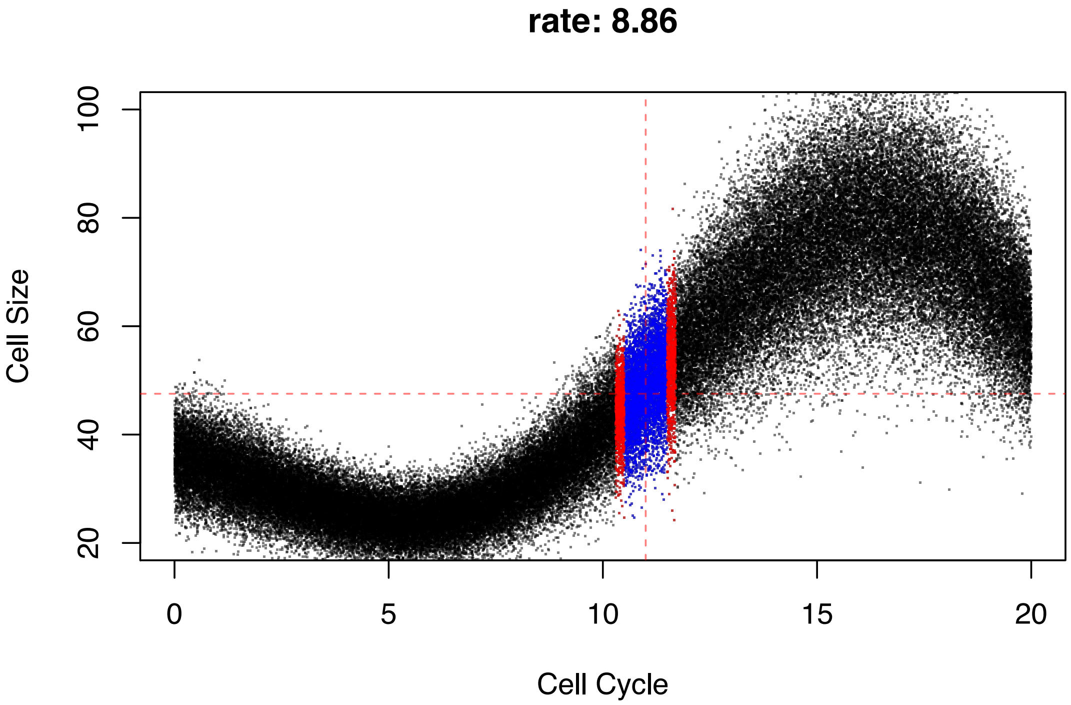

Figure 2.1: **ERA stepwise calculation** Points in black indicate cells transitioning through the cell cycle, simulated based on an imaginary function of cell size. Dashed lines indicate the coordinates for which the rate is calculated. Points in blue and red indicate the subpopulations required for calculating the rate of change, as shown in Equation (2.1).

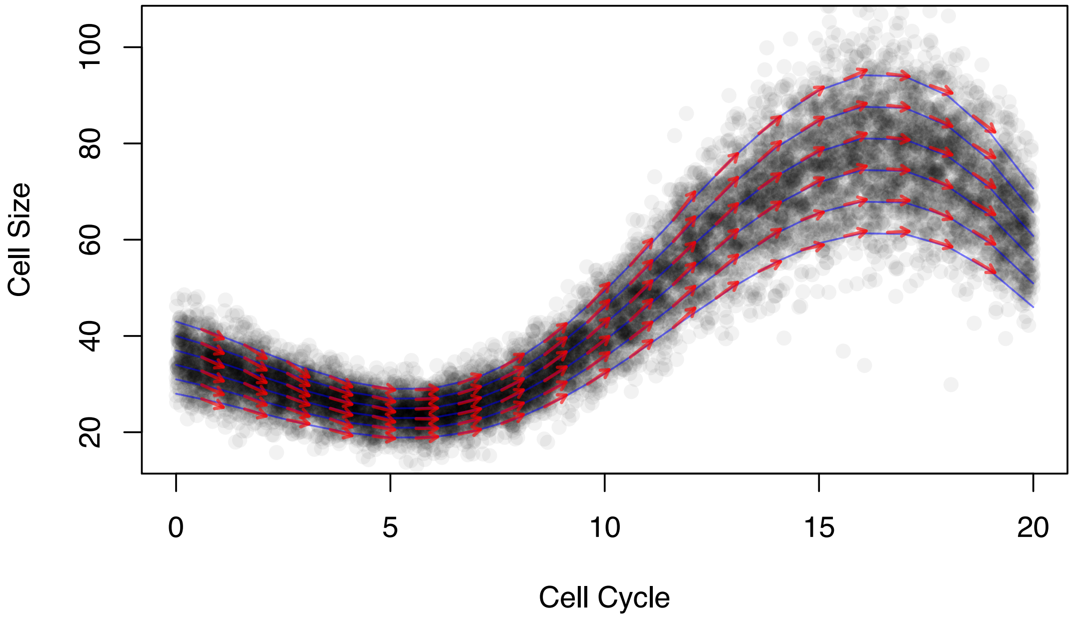

Figure 2.2: **ERA validation by simulation** Points in black indicate cells transitioning through the cell cycle, arrows indicate the predicted trajectory of cells of different sizes based on the result of Equation (2.1), and cyan lines indicate the *true* trajectories based on the function used to simulate the data.

2.5 Probability State Modelling

CB Bagwell, who introduced fluorescence spectral overlap compensation for flow cytometry (Bagwell and Adams, 1993) has more recently introduced the theory of PSM for studying cell differentiation based on multiparametric flow cytometry immunological data (Bagwell et al., 2015b). In their companion article (Bagwell et al., 2015a), the method is applied to elucidate the stages of B-cell CD19 upregulation. The method relies on using prior knowledge of how a given biomarker changes over relative time to inform a model, the parameters for which can be fitted to cytometry data, given reasonable boundary constraints. Although time in (Bagwell et al., 2015a) is relative to differentiation, there is in principle no reason why it cannot reflect other time dependent progressions such as the cell cycle. Here, we adapt the principles of PSM for the purpose of cell cycle analysis.

2.5.1 Model Proposal

The overall modelling process can be described in five simple steps. First, a suitably simple model is devised, describing our prior understanding of the biomarker. This can be a simple progression such as the low-to-high transition of the DNA stain Hoechst, or a more complex transition such as that followed by the *fucci* Cdt1 marker, as will be seen in Section 4.2. Bagwell et al. (2015b) suggest using piecewise linear models to describe the changes in the mean and error of the intensity measurement, but in principle any type of function can be used. These models describe how a given biomarker's measurement level and variance changes over the course of the progression, in our case the cell cycle. The progression is defined from 0 to 100, which reflects the cumulative density percentiles.

Here, we start with the cell cycle reporter with the simplest transition, namely DNA quantity, as measured by Hoechst staining. The proposed model contains all three phases that can be distinguished by this stain; G1, S and G2/M. This model is defined by two levels of intensity, indicated by the horizontal lines in Figure 2.3, and two change points indicated by the two vertical lines. Together these define two 'Control Definition Points' (red circles), between which the intensity level and error are linearly interpolated. Further interpolations extend the progression to start and end of the cell cycle. The error of the measurement, illustrated by the height of the polygon, we assume to be normally distributed.

2.5.2 Calculation of Probabilities

The proposed model can be formally described by a pair of functions, giving the change in measurement intensity level and spread. These are defined as $Q(\tau, C)$ and $\sigma(\tau, C)$, where τ defines the state of the progression and C the piecewise linear model, in accordance to the notation in (Bagwell et al., 2015b). By extension, the model in Figure 2.3 can be seen as two dimensional density, defined by

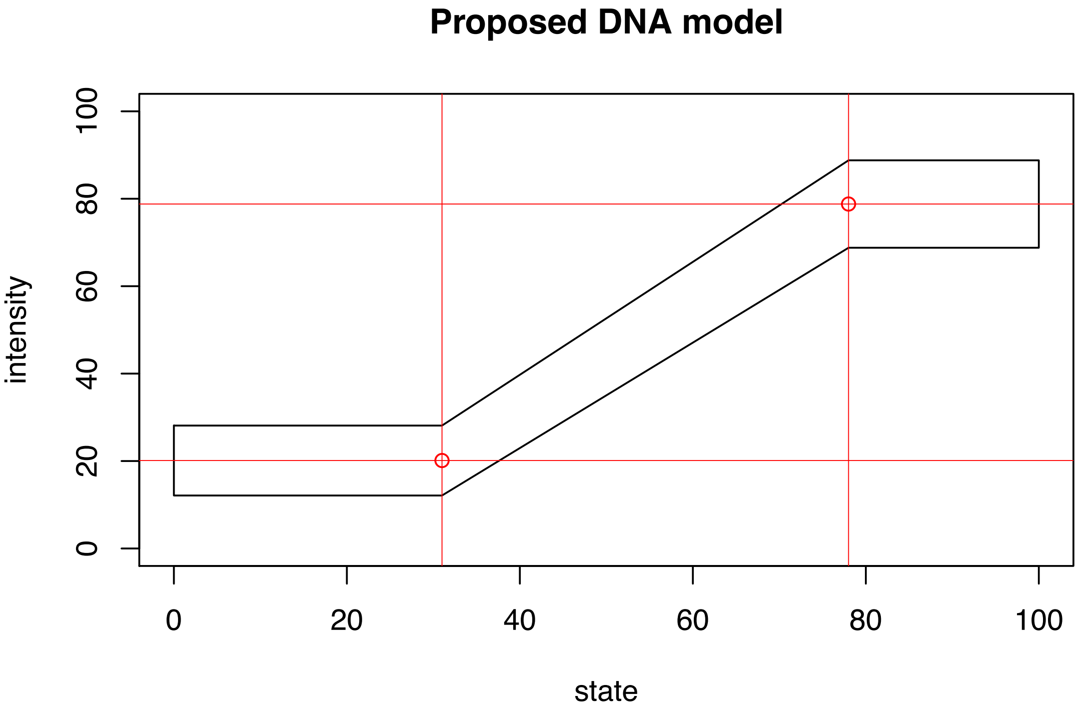

Figure 2.3: **Proposed DNA cell cycle transition.** Intensity is the normalised measurement level of the DNA stain Hoechst. State refers to the cell cycle progression. Red lines and circles define the control definition points, between which the level is interpolated linearly. Black polygon lines show the interpolated level of DNA stain ±2 standard deviations.

$$P(x|\tau, C) = N(x, Q(\tau, C), \sigma(\tau, C)),$$

where x is the observed measurement level. $P(x|\tau, C)$ allows us to calculate the probability of a cell being at any stage in the cell cycle, based on the measured DNA intensity level and the proposed model. Next, we discretise the above probability density function,

$$E_{\nu,s} = \frac{r}{100} \int_{\frac{100}{r}(s-1)}^{\frac{100}{r}s} \int_{\frac{100}{w}(\nu-1.5)}^{\frac{100}{w}(\nu+0.5)} N(x, Q(\tau, C), \sigma(\tau, C)) dx d\tau, \qquad (2.2)$$

where $\nu \in 1, 2, ..., w$ and $s \in 1, 2, ..., r$ are the discrete intensity levels and cell cycle states a cell can be associated with. A probability matrix termed E-matrix, defined by Equation (2.2), is calculated by numerically integrating the proposed model over cell cycle time and measurement level at discrete intervals in this way. The width of the intervals defines the resolution of the matrix, which is a rate limiting step for the overall PSM algorithm and thus should be chosen according to available resources. The result of numerically solving Equation (2.2) leads to the matrix seen in Figure 2.4. The colour intensity indicates the relative density distribution.

2.5.3 State prediction

Once a probability matrix has been obtained, the next step is to bin the cytometry data into the same number of bins as the number of rows in the respective E-matrix. This can be seen in Figures 2.5 and 2.6, where we use 10,000 data points simulated using the proposed DNA model shown in Figure 2.3 as an example.

The next step is critical, as it uses the above construction to determine the state of each data point along the transition. Specifically, for every data point, we use the *corresponding row* in the E-matrix, defined by the data point's intensity level, as a probability vector with values corresponding to each of the different columns

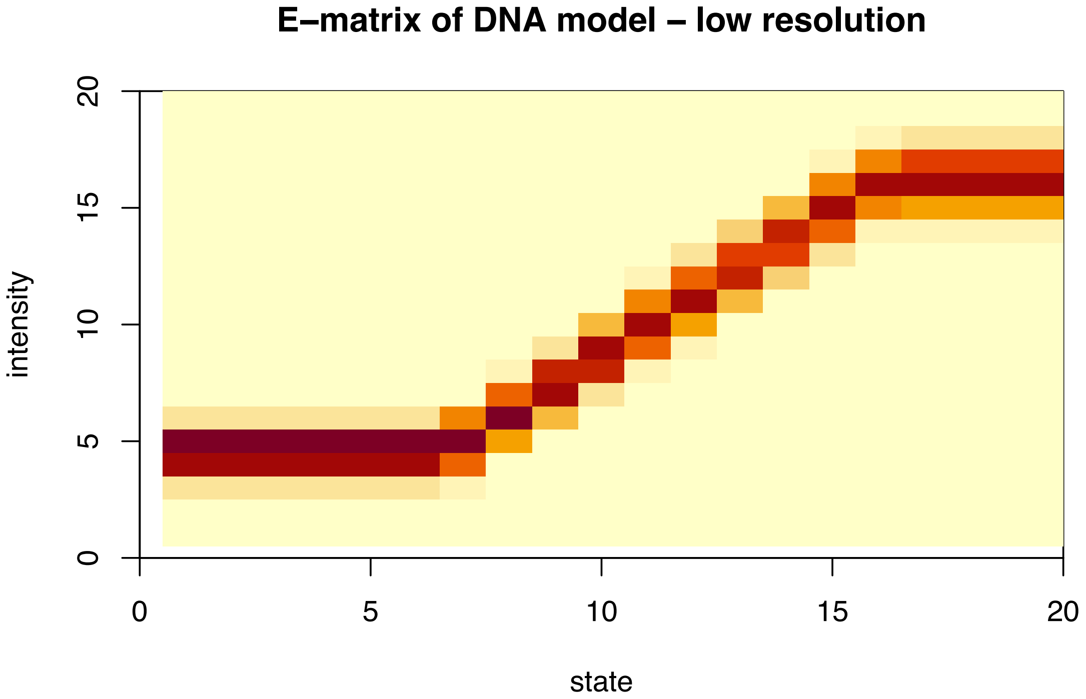

Figure 2.4: **E-matrix of proposed DNA cell cycle transition**. Darker colour equals higher density. The matrix was constructed using a resolution of 20 by 20.

of the matrix (cell cycle states). This vector is then used to perform weighted sampling across the columns of the matrix, and thus predict the cell cycle state of the cell.

This way we can stochastically assign each data-point to a state τ along the progression, noting that the number of possible states is defined by the resolution of the matrix. Once a state has been sampled for every single data point, the process is complete. Uniform noise of appropriate bandwidth can be added to reflect the uncertainty due to the resolution limit. The result is seen in Figure 2.7. Figure 2.8 shows the result of the same process but using an E-matrix of higher resolution.

2.5.4 Goodness of fit

The fit of the proposed model is evaluated by comparing each E-matrix to its empirical equivalent, ES-matrix. To construct the ES-matrix, we bin the

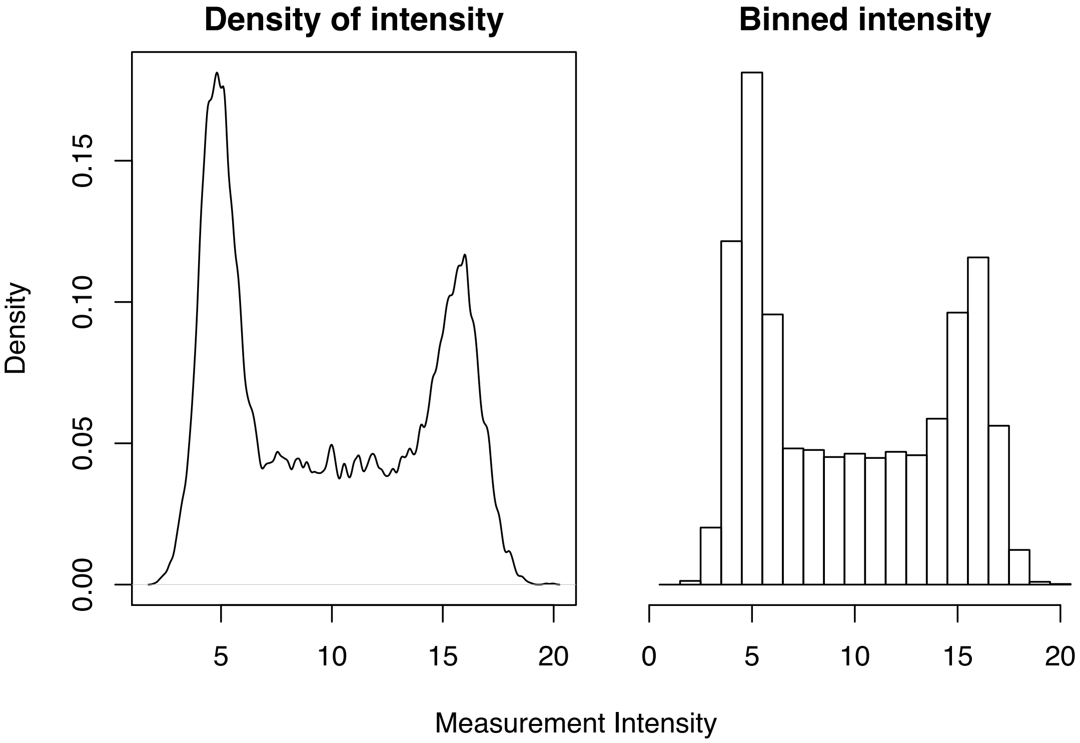

Figure 2.5: **Binning of data points**. 10,000 data points simulated according to the proposed DNA model, binned in a one-to-one correspondance with the rows of the E-matrix. Left: kde of simulated data. Right: histogram of the same data binned (n = 20) as described. Two peaks correspond to G1 and G2 respectively, and are due to the fact that the duration of these phases is longer, resulting a larger fraction of the population to found in these phases at any one time.

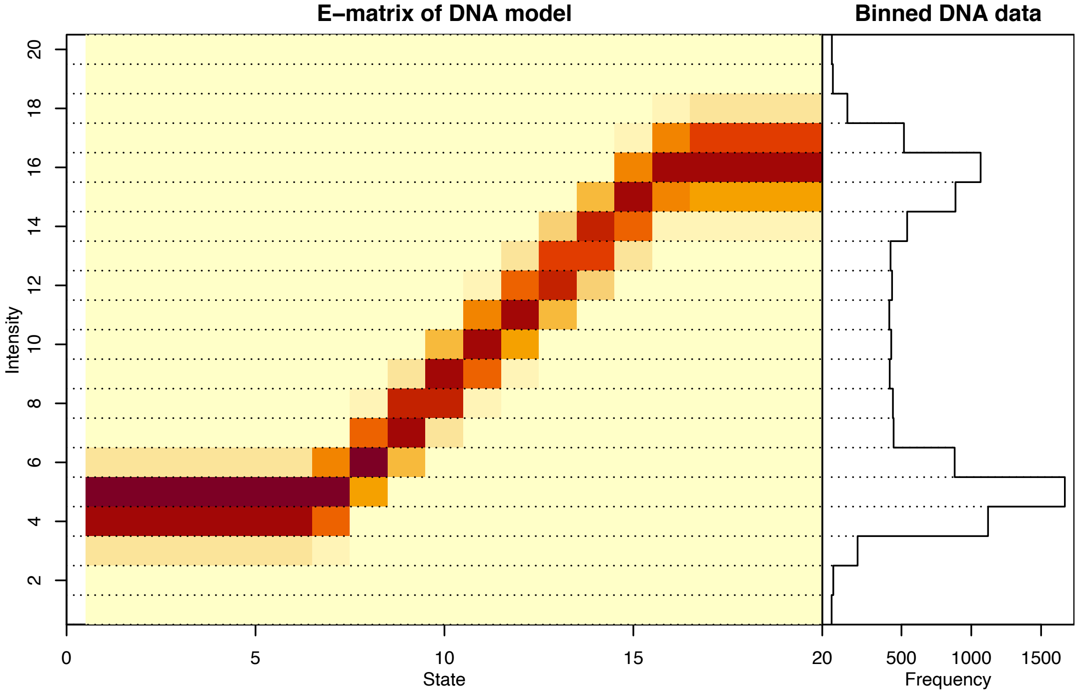

Figure 2.6: **Assigning of E-matrix row to each data point**. The data are binned in such as way so that each bin corresponds to a row in the E-matrix. Here, 20 bins were used to match the 20 rows of the E-matrix.

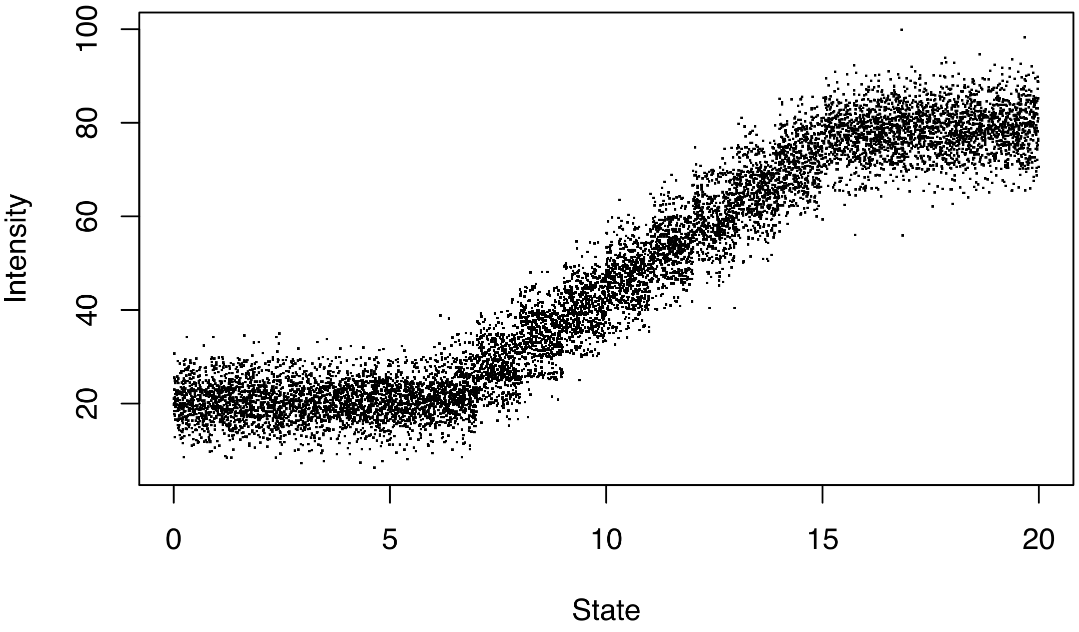

Figure 2.7: **Result of PSM using 20x20 resolution *E*-matrix** 10,000 data points simulated according to the proposed DNA model, analysed by PSM using the same model.

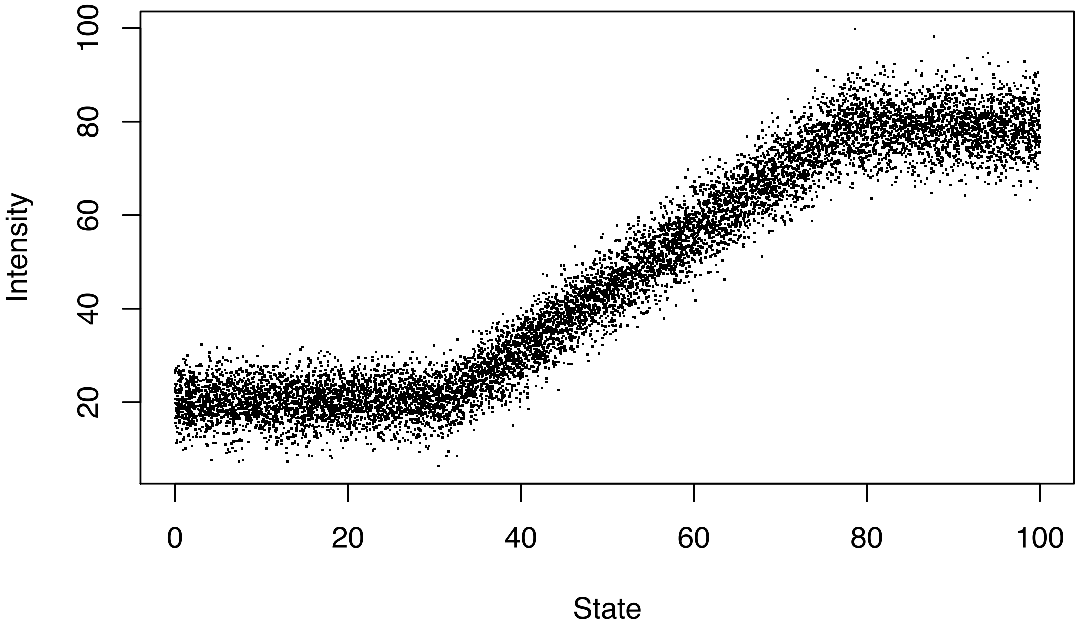

Figure 2.8: **Result of PSM using 100x100 resolution *E*-matrix** 10,000 data points simulated according to the proposed DNA model, analysed by PSM using the same model.

resulting two-dimensional distribution of points seen in Figure 2.8, along both axes, and count the points in each bin (see Figure 2.9). The resulting matrix is normalised appropriately and then compared to the *E*-matrix by means of a reduced *chi*-square test, a common test for comparing binned data between groups (McHugh, 2013). In this case, the data are in good agreement with the model (not shown), which is expected, as it is the same model that was used to simulate these data.

2.5.5 Model parameter inference

Once we have set up the above algorithm, we can use it to solve the inverse problem of inferring the model parameters that best describe the experimental data. We do so by iteratively proposing new values for the parameters describing the model (change-point timings, intensity level means and variances), and testing their agreement to the data, based on the scoring calculation in Section 2.5.4. This process can be formulated into an objective function which in turn can subjected

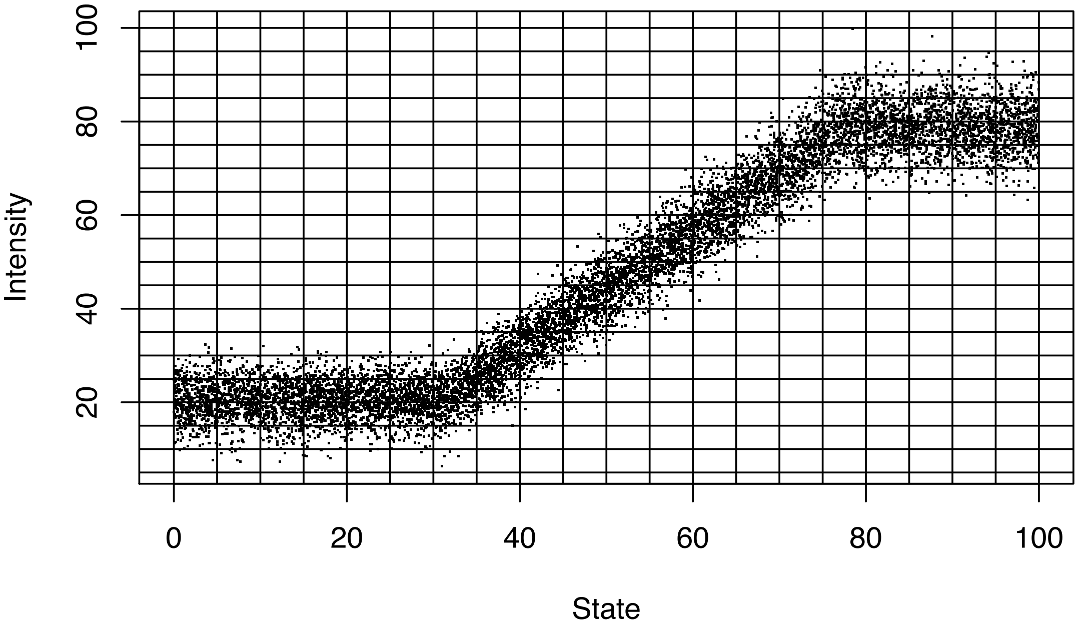

Figure 2.9: **Construction of *ES*-matrix.** Two dimensional distribution of points is binned as shown to construct the empirical equivalent to the probability E-matrix. Comparing the two matrices returns a score which can be used to assess the fit.

to optimisation.

A suitable global optimisation algorithm needs to be chosen which takes the characteristics of such an objective function into account. Specifically, the optimiser needs to be able to cope with expensive and noisy evaluations, potentially multiple false minima, as well as inequality constraints in the parameter space.

Here, we use the Bayesian Optimisation package mlrMBO developed by Bischl et al. (2017) to fit a probability state model to our DNA cytometry data. The specific implementation was preferred as it is well developed and can be interfaced with the rest of our analysis pipeline, also written in R.

To set the initial priors, we use a script written for this purpose by Dr Massimo Cavallaro, which identifies the peaks in a bimodal distribution and fits a Gaussian function to either peak. This way we obtain an initial estimation of the means and errors of the two stationary states. Specifically, to overcome the convoluting effect of the transitory state between the two stationary states we used a half-Gaussian, fitted on the corresponding extreme ends of the two clusters, as shown in Figure 2.10.

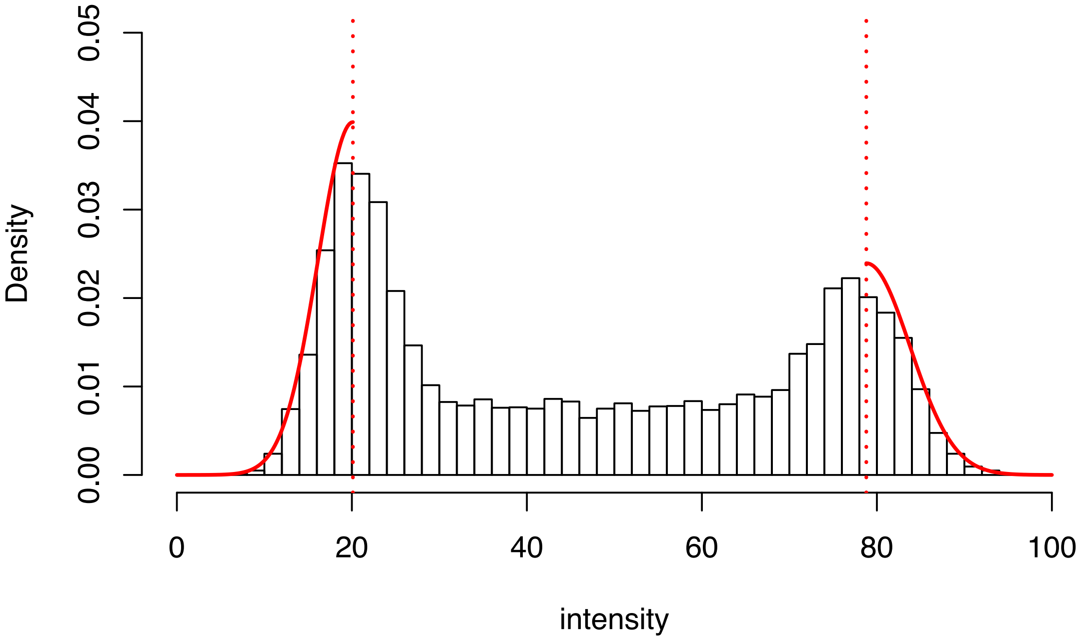

Figure 2.10: **Peak identification - half gaussian fit** Histogram of DNA intensity. Dashed red lines correspond to peaks identified, curved lines correspond to the fitted half-Gaussian density. These are used as initial values for the mean intensity and spread of the stationary phases in the DNA progression, between which the values are linearly interpolated.

The second task is to pick suitable boundaries within which each parameter can vary. Such boundaries, alongside the proposed model, form a type of prior and are based on our current understanding of the process. Here, we use a two step optimisation approach, employing broad boundaries to get an initial fit of the parameters, which inform a set of secondary, narrower boundaries for the final fitting. This way we benefit from the freedom of weak initial priors, but also the precision that can be achieved from the subsequent stronger priors.

2.5.6 Converting cumulative percent to time

Once a satisfactory model has been identified, PSM can be used to predict the state of each cell in the cell cycle, which is expressed in terms of cumulative percent. To convert this output to cell cycle time, we rely on the definition of flux, according to which the longer a certain stage is in duration, the higher the expected density of cells in that stage is. Using this principle we can convert this relative time to absolute cell cycle time, using the duration of the average cell cycle, which can be easily measured or obtained from the literature. This method has been used before for correlating density of static data to relative time (Kafri et al., 2013).

One important consideration to make is that due to cells dividing at the end of each cell cycle period, there is a disproportionate number of cells at the start of the cell cycle compared to the end. This can be accounted for using a simple transform (Kafri et al., 2013):

$$t = \frac{1}{\alpha}\ln\left(\frac{2}{2 - F(l)}\right),$$

where $F(l)$ is the cumulative distribution describing the frequency of cells that are either at cell cycle stage l or at earlier cell cycle stages, and

$$\alpha = \frac{\ln(2)}{\tau},$$

where τ is the doubling time of the population, which for Hela cells is about 20 hours (Sherwood et al., 1994). PSM is based on quantile modelling (Gilchrist, 2000) and as a result the predicted state of each cell is expressed in terms of cumulative percent. In order to obtain the predicted timeline in terms of cell cycle time, the above transform can therefore be directly applied to each cell by substituting $F(l)$ above with its predicted state from PSM.

2.6 Bayesian Optimisation

It is often useful to be able to computationally fit a function to a set of data, either in order to evaluate how well a model describes the data or in order to obtain certain parameters which are difficult to extract directly. For example, using a simple model such as that of gene expression developed in Chapter 3 it is possible to estimate the rates of transcription and mRNA degradation using a single snapshot of the mRNA distribution in a population of cells. Numerical algorithms exist which can perform this procedure very efficiently, especially when the analytical formula of the model being fitted (or its likelihood function) is available.

In cases where this is not possible, an alternative approach is to use simulations which describe the model, alongside a scoring method for evaluating the fit to the data. As these simulations are often expensive to evaluate and the results noisy, a different class of algorithms is required for performing the fitting. Sequential model-based optimisation (SBMO) is such a class of algorithms, and has become the state-of-the-art optimisation strategy in recent years (Jones et al., 1998).

Briefly, such models work by constructing a surrogate regression model which approximates the objective function, and is much cheaper to evaluate than the objective function itself. The surrogate function is initialised using a series of points randomly sampled, and is then used to identify future candidate points based on a certain criterion, called the *infill* criterion. The proposed points are evaluated by the objective function, and the resulting score is used to update the surrogate function.

By iterating between these two steps, an increasingly accurate surrogate function is constructed, resulting in improved parameter sets being proposed with every iteration. Ultimately, the algorithm is stopped when a maximum number of iterations is reached. Furthermore, using a suitable criterion such as “expected improvement”, a balance is struck between the exploration of underrepresented areas and the targeting of the most promising areas of the parameter space, ensuring that the process does not become stuck in a local optimum.

Here, the Bayesian Optimisation package by Bischl et al. (2017) was identified as a suitable SBMO implementation for fitting cell cycle parameters using PSM in

Chapter 3

Modelling gene expression and cell growth

Gene expression is considered to be a noisy process (Raser and Shea, 2006). This observation can be attributed in part to intrinsic sources, such as the bursty nature of transcription (Raj et al., 2006), which result in broad, super-Poissonian distributions of transcript numbers. As mentioned in Section 1, this burst-like behaviour has been observed directly in certain genes, using live cell microscopy (Golding et al., 2005). Ever since, the 'random telegram' model has become the prevalent model for transcription (Raj et al., 2006, Lenstra et al. (2016)), though recent evidence has questioned the universality of this model. Specifically, careful consideration of extrinsic sources of noise (Swain et al., 2002), such as the variation of cell size and cell cycle, has shown that, when taken into account, non-bursty kinetics are sufficient in describing transcriptional dynamics (see findings in plants (Ietswaart et al., 2017), yeast (Zopf et al., 2013) and mammals (Battich et al., 2015, Klein et al. (2015))). To investigate this, we formulate a simple, non-bursty model which takes these extrinsic sources into account, and determine whether such as model can sufficiently describe the observed gene RNA distributions.

A similar approach was followed by Soltani et al. (2016), who studied the effects of molecule partitioning during cell division and the oscillation in active alleles due to DNA replication, demonstrating the importance of the timing of these events on gene expression noise. Specifically, variability in the period of the cell cycle and timing of replication and division were considered explicitly, and the moments of the resulting distribution were derived. Here, we simplify by assuming deterministic timings of gene replication and division. This allows us to derive the analytical form of the probability mass function (PMF) describing the RNA distribution in a population of cells, which can be used directly to test the likelihood of different parameter sets with respect to single-cell gene expression data. We further extend the work of Soltani et al. (2016) by incorporating the change in cell size during the cell cycle, a requirement towards understanding the effects of cell growth on gene expression.

3.1 Two phase model

As a starting point, we consider a simple two-step mass action kinetics model of RNA metabolism. To simulate the doubling and halving of the DNA substrate due to the periodic replication of the genome and subsequent cell division, we start by making the simplifying assumption that the transcription rate periodically doubles and halves as a consequence. This constraint is relaxed in a later section. Furthermore, we do not consider the partitioning of molecules at the end of the division cycle at this stage. Instead, we proceed without it and account for it in Section 3.3.

3.1.1 Deriving the RNA PMF at steady state

We begin by considering the model

$$\begin{aligned} \emptyset &\xrightarrow{\lambda} RNA \\ RNA &\xrightarrow{\mu} \emptyset, \end{aligned} \tag{3.1}$$

where λ and μ represent the transcription and degradation rates respectively, $\emptyset$ represents void. In order to simulate the changes in gene dosage during the cell cycle, λ is replaced by 2λ during the G_2 phase, thus constituting a null model of transcription dynamics during the cell cycle in the absence of dosage compensation mechanisms (Voichek et al., 2016b, Padovan-Merhar et al. (2015)). This leads to two distinct Chemical Master Equations (CMEs),

$$\frac{dP_k}{dt} = -(\lambda + k\mu)P_k + \lambda P_{k-1} + \mu(k+1)P_{k+1}, \text{during G1} \tag{3.2}$$

$$\frac{dP_k}{dt} = -(2\lambda + k\mu)P_k + 2\lambda P_{k-1} + \mu(k+1)P_{k+1}, \text{during G2} \tag{3.3}$$

where P_k is the probability of having k molecules at time t. The probability rate equation is derived by considering the change in RNA molecules due to rates λ and μ. We start using Equation (3.2) to derive the time varying probability distribution at time t during the first phase of the cell cycle, which will serve as a stepping stone towards deriving the equivalent distribution for the second phase, and ultimately the whole cell cycle at steady state. To do so, we employ the probability generating function (p.g.f.),

$$G(z,t) = \sum_{k=0}^{\infty} P_k z^k,$$

where G is a power series representation of the PMF, from which the PMF is recovered by taking derivatives of G with respect to z. Here, when differentiated k times G returns the probability of having k number of RNA molecules at time t. The PMF is recovered by taking derivatives of G with respect to z To obtain the relevant expression for the generating function, we differentiate *G(z, t)* with respect to t (Peccoud and Ycart, 1995). This gives us

$$\frac{\partial G}{\partial t} = \mu(1-z)\frac{\partial G}{\partial z} - \lambda(1-z)G. \tag{3.4}$$

To solve this equation for G, we employ the method of characteristic lines. First we parameterise G using τ, and differentiate using the chain rule, which gives

$$\frac{dG}{d\tau} = G_z\frac{dz}{d\tau} + G_t\frac{dt}{d\tau}. \tag{3.5}$$

Equations (3.4) and (3.5) give us the system of ODEs below, which we can solve to obtain an expression for G.

$$\begin{aligned} \frac{dt}{d\tau} &= 1 && (3.6)\\ \frac{dz}{d\tau} &= -\mu(1-z) && (3.7)\\ \frac{dG}{d\tau} &= -\lambda(1-z)G, && (3.8) \end{aligned}$$

where τ is what we use to parameterise z and t. By setting $P_0(0) = 1$, and by the way we define the characteristic lines we get the initial conditions $z = s$, $t = 0$, $G(s, 0) = 1$, at $\tau = 0$. The above equations together with the initial conditions give us the below expressions,

$$\frac{dG}{dt} = -\lambda(1-z)G, \text{and} \tag{3.9}$$

$$z = 1 + (s-1)e^{\mu t}. \tag{3.10}$$

We solve for G by plugging (3.10) into (3.9) and integrating by sides,

$$\begin{aligned} \int \frac{1}{G}dG &= \int \lambda(s-1)e^{\mu t}dt \\ \ln G &= \lambda(s-1)(\frac{1}{\mu}e^{\mu t}) + c. \end{aligned}$$

Using the initial conditions, we find *c*,

$$c = -\frac{\lambda}{\mu}(s-1), \tag{3.11}$$

which we substitute in (3.11),

$$G = e^{\frac{\lambda}{\mu}(1-s)(1-e^{\mu t})}. \tag{3.12}$$

Finally, we need to express *G(z, t)* in terms of *z* and *t*. We rearrange Equation (3.10) and plug into (3.12) to get

$$(3.10) \rightarrow s = (z-1)e^{-\mu t} + 1,$$

$$G(z,t) = e^{\frac{\lambda}{\mu}(z-1)(1-e^{-\mu t})}. \tag{3.13}$$

Equation (3.13) is the full expression of the p.g.f.. It can be shown that by differentiating *G(z,t) k* times with respect to *z*, we get the generalised formula below for deriving the probabilities at time *t*,

$$P_k(t) = \frac{1}{k!}\frac{\partial^k G}{\partial z^k}\bigg|_{z=0}. \tag{3.14}$$

Using equation (3.12), it can be shown that

$$\frac{\partial^k G}{\partial z^k} = \left(\frac{\lambda}{\mu}(1-e^{-\mu t})\right)^k e^{-\frac{\lambda}{\mu}(1-e^{-\mu t})}. \tag{3.15}$$

By substituting Equation (3.15) into (3.14), we get

$$P_k(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k (1-e^{-\mu t})^k e^{-\frac{\lambda}{\mu}(1-e^{-\mu t})}, \tag{3.16}$$

which describes the time varying RNA distribution. We use the same procedure to obtain the PMF for the second phase of the cell cycle, where the transcription rate changes from λ to 2λ.

3.1.2 Probability distribution in G_2, following G_1.

To derive the RNA distribution in the second phase, we use (3.3) alongside the relevant initial conditions, which we obtain from Equation (3.16). Specifically, by solving (3.16) at $t = t_1$, where t_1 is the duration of the first phase, we obtain the initial conditions of the second phase. As such, we obtain the p.g.f. for the second phase,

$$G(z,t) = e^{\frac{\lambda}{\mu}(z-1)(2-e^{-\mu(t-t_1)}-e^{-\mu t})}. \tag{3.17}$$

Following the same steps as before, Equation (3.17) in turn gives the PMF of RNA numbers during the G_2 phase,

$$P_k(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k (2-e^{-\mu(t-t_1)}-e^{-\mu t})^k e^{-\frac{\lambda}{\mu}(2-e^{-\mu(t-t_1)}-e^{-\mu t})} \tag{3.18}$$

3.1.3 Probability Generating Functions at steady state

So far we have obtained the time varying distributions of RNA numbers during the first two phases of the cell cycle. These distributions though do not reflect the steady state distributions. To obtain the steady state PMFs, we repeat the above process for the 2^{nd} cell cycle, by solving Equation (3.2) as shown in Section 3.1.1, though this time using initial conditions obtained by solving Equation (3.18) at $t = t_1 + t_2$, where t_2 is defined as the duration of the second phase. This way we obtain the PMFs for the 3^{rd} cell cycle, then the 4^{th}, and so on. The resulting sequence is captured by the closed forms

$$G_1(z,t) = e^{\frac{\lambda}{\mu}e^{-\mu t}(z-1)\left(e^{\mu t}-1+\sum_{n=1}^{m} e^{n\mu(t_1+t_2)}-\sum_{n=1}^{m} e^{\mu(nt_1+(n-1)t_2)}\right)} \quad \text{for phase 1,} \tag{3.19}$$

$$G_2(z,t) = e^{\frac{\lambda}{\mu}e^{-\mu t}(z-1)\left(2e^{\mu t}-1+\sum_{n=1}^{m-1} e^{n\mu(t_1+t_2)}-\sum_{n=1}^{m} e^{\mu(nt_1+(n-1)t_2)}\right)} \quad \text{for phase 2,} \tag{3.20}$$

where m represents the number of cell cycles. As $m \to \infty$, Equations (3.19) and (3.20) converge to the steady state functions

$$G_1(z,t) = e^{\frac{\lambda}{\mu}(z-1)\left(1+e^{-\mu(t-t_1)}\frac{1-e^{t_2\mu}}{1-e^{(t_1+t_2)\mu}}\right)} \quad \text{and} \tag{3.21}$$

$$G_2(z,t) = e^{\frac{\lambda}{\mu}(z-1)\left(2-e^{-\mu(t-(t_1+t_2))}\frac{1-e^{t_1\mu}}{1-e^{(t_1+t_2)\mu}}\right)}. \tag{3.22}$$

The p.g.f. Equations (3.21) and (3.22) are used to derive the equivalent probability mass functions,

$$P_{k,1}(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k (e^{-\mu(t-t_1)}\frac{1-e^{\mu t_2}}{1-e^{\mu(t_1+t_2)}}+1)^k e^{-\frac{\lambda}{\mu}(e^{-\mu(t-t_1)}\frac{1-e^{\mu t_2}}{1-e^{\mu(t_1+t_2)}}+1)}, \tag{3.23}$$

$$P_{k,2}(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k (2-e^{-\mu(t-t_1-t_2)}\frac{1-e^{\mu t_1}}{1-e^{\mu(t_1+t_2)}})^k e^{-\frac{\lambda}{\mu}(2-e^{-\mu(t-t_1-t_2)}\frac{1-e^{\mu t_1}}{1-e^{\mu(t_1+t_2)}})}, \tag{3.24}$$

which match well with simulations (see Figure 3.1). It is worth noting however that while the durations t_1 and t_2 of phases G1 and G2 respectively are modelled as constants here, this is a simplification used to compute the closed forms of the

PMFs. In reality, the durations of all eukaryotic cell cycle phases are known to be variable (Charvin et al., 2008, Brooks et al. (1980)), and incorporation of this source of variability into the model should be considered in the future, as used by Soltani et al. (2016) for calculating the first two moments of the probability functions.

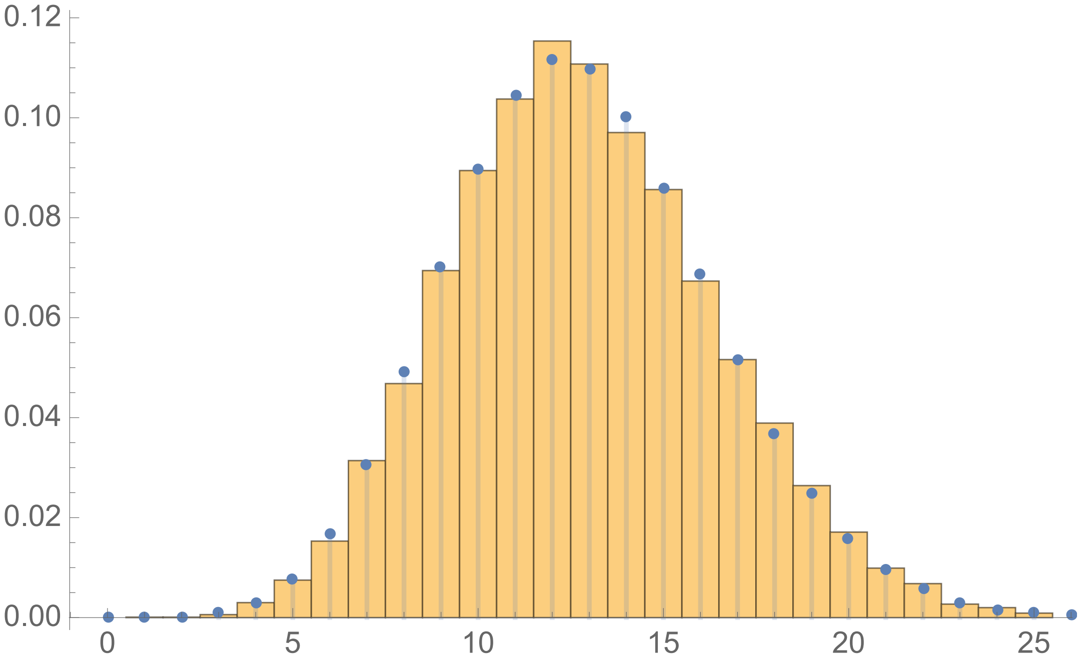

Figure 3.1: **Steady state mRNA distribution at end of G1.** Histogram represents 10,000 Gillespie simulations ran with arbitrary parameters. Dots represent the solution of the PMF from Equation (3.23) using the same parameters.

3.2 Phase integration

Equations (3.23) and (3.24) describe the time-dependent RNA distributions within a synchronised population of cells with respect to the cell cycle. However, synchronisation achieved either by chemical cell cycle inhibition, or *in silico* synchronisation based on cell cycle reporters, are not perfect. In order to compare the above PMFs to the experimental data, we thus need to consider the resolution limit of the cell cycle in the synchronised cells. As we will see in Chapter 4, in the case of *in silico* synchronisation, this limit depends on the information that can be obtained from the available cell cycle reporters. In the simplest case, a DNA stain such as Hoechst can be used to differentiate between G1, S and G2 phases.

To compare with population data resolved in these phases we require the PMF for distributions over the whole of each phase, which we obtain by integrating over each phase's time-span. To get the PMF over the whole G1 phase, we integrate Equation (3.23) from zero to t_1, where zero is defined as the start of the G_1 phase at steady state, and t_1 the end. Alternatively, we can integrate the p.g.f. (3.19) over the same time span,

$$\begin{aligned} G_1(z,t) &= e^{\frac{\lambda}{\mu}(z-1)\left(1+e^{-\mu(t-t_1)}\frac{1-e^{t_2\mu}}{1-e^{(t_1+t_2)\mu}}\right)}, \\ \int_0^{t_1} G_1(z,t)dt &= e^{\frac{\lambda}{\mu}(z-1)} \sum_{n=0}^{\infty} \left(\frac{A^n(-1+z)^n \int_0^{t_1} e^{(-t+t_1)\mu n} dt}{n!} \right), \\ &= \frac{1}{\mu} e^{\frac{\lambda}{\mu}(z-1)} \Gamma \left(0, -\frac{\left(1-e^{\mu t_2}\right)(z-1)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}, -\frac{e^{t_1\mu}\left(1-e^{\mu t_2}\right)(z-1)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu} \right), \end{aligned} \tag{3.25}$$

and then derive the PMF as before,

$$\begin{aligned} P_{1_k} = \frac{1}{\mu k! t_1} &\left(e^{-\frac{\lambda}{\mu}} \left(\frac{\lambda}{\mu}\right)^k \Gamma \left(0, \frac{\left(1-e^{\mu t_2}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu} \frac{e^{\mu t_1}\left(1-e^{\mu t_2}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu} \right) + \right. \\ &\left. \sum_{n=0}^{k-1} \left(\frac{\lambda}{\mu}\right)^{k-(n+1)} \Gamma \left(1+n, \frac{\lambda}{\mu} + \frac{\left(1-e^{\mu t_2}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}, \frac{\lambda}{\mu} + \frac{e^{\mu t_1}\left(1-e^{\mu t_2}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu} \right) \right), \end{aligned} \tag{3.26}$$

where, as before, k is the number of RNA molecules. The same approach can be employed to derive the RNA number PMF during the G_2 phase. By integrating the p.g.f. (3.20) over the t_1 to $t_1 + t_2$ time span, we can derive the PMF for the

G2 phase

$$P_{2_k} = \frac{1}{\mu k! t_2} \left(e^{-\frac{2\lambda}{\mu}} \left(\frac{2\lambda}{\mu}\right)^k \Gamma\left(0, -\frac{\left(1-e^{\mu t_1}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}, -\frac{e^{\mu t_2}\left(1-e^{\mu t_1}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}\right) + \right.$$
$$\left. \sum_{n=0}^{k-1} \left(\frac{2\lambda}{\mu}\right)^{k-(n+1)} \Gamma\left(1+n, \frac{2\lambda}{\mu} - \frac{\left(1-e^{\mu t_1}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}, \frac{2\lambda}{\mu} - \frac{e^{\mu t_2}\left(1-e^{\mu t_1}\right)\lambda}{\left(1-e^{\mu(t_1+t_2)}\right)\mu}\right)\right). \tag{3.27}$$

Both steady state PMFs compare well to simulations (see Figure 3.2), noting however that the integration and respective simulations reflect the popular ‘mother machine’ experimental setting (Wang et al., 2010), and thus cannot be directly compared to snapshot-type data such as that acquired by flow cytometry, due to the non-uniform distribution of cells throughout each phase caused by the continuous influx of newborn cells at the start of the cell cycle. These equations should therefore be adjusted accordingly before comparing to the data.

3.3 Cell division

Cell division has been shown experimentally to lead to a random partitioning of RNA molecules, in a binomial fashion similar to Golding et al. (2005). Huh and Paulsson (2011a) demonstrated mathematically that random partitioning of molecules at division can explain the observed RNA variation in a population of cells just as well as bursty transcription dynamics can. Here, we take into account the effect of cell division by adapting our model accordingly. Specifically, in a similar way to Section 3.1.2, we use the PMF solution at the end of the cell cycle to obtain the initial conditions for solving the next cell cycle, though in this case we add another step, whereby we take the binomial distribution of the initial conditions. This is shown using a Gillespie simulation in Figure 3.3. The derivation of the resulting PMF is outlined below.

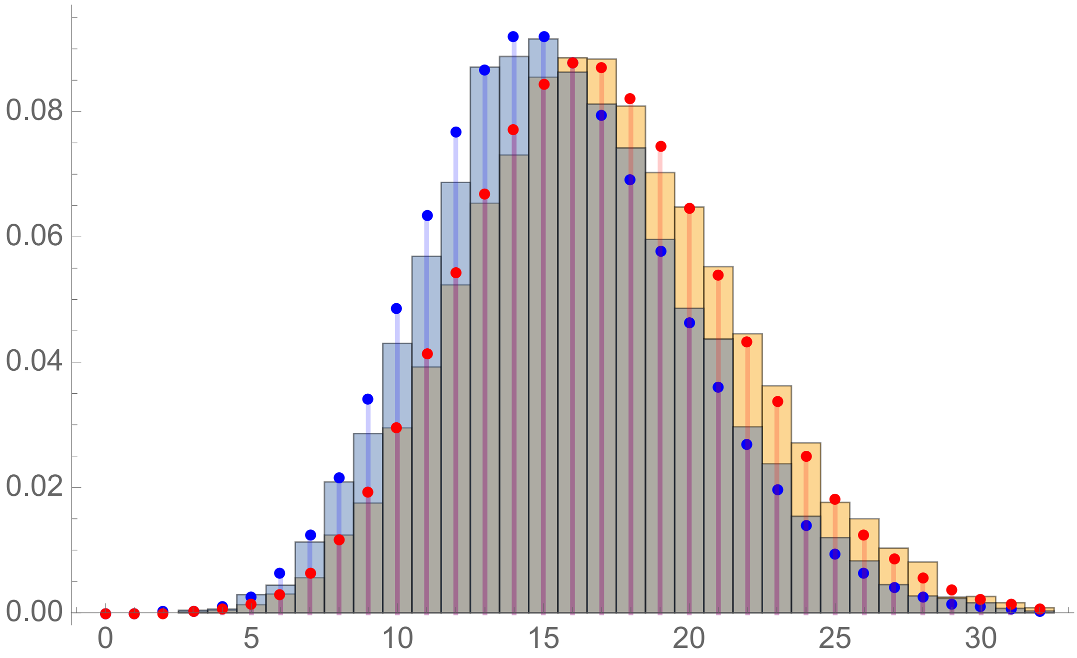

Figure 3.2: **Integrated G1 and G2 phase mRNA distributions.** Histograms represent 10,000 Gillespie simulations ran with arbitrary parameters. Dots represent the solutions of the PMFs from Equations (3.26) and (3.27), using the same parameters. Blue indicates the G1 phase, red the G2.

3.3.1 Probability distribution in G_1, following division

Let x_1 be the RNA number right before division and x_2 be the RNA number right after division. x_2 is given by a binomial $B(x_1, p)$, where p is the ratio of the size of the daughter cell in question over that of the mother cell, assuming homogeneous distribution of transcripts in the cytoplasm.

Then

$$\begin{aligned} P\left(X_2 = x_2\right) &= \sum_{x_1=0}^{\infty} P\left(X_1 = x_1, X_2 = x_2\right) \\ &= \sum_{x_1=0}^{\infty} P\left(X_2 = x_2 | X_1 = x_1\right) P\left(X_1 = x_1\right). \end{aligned} \tag{3.28}$$

To get the p.g.f., we transform both sides,

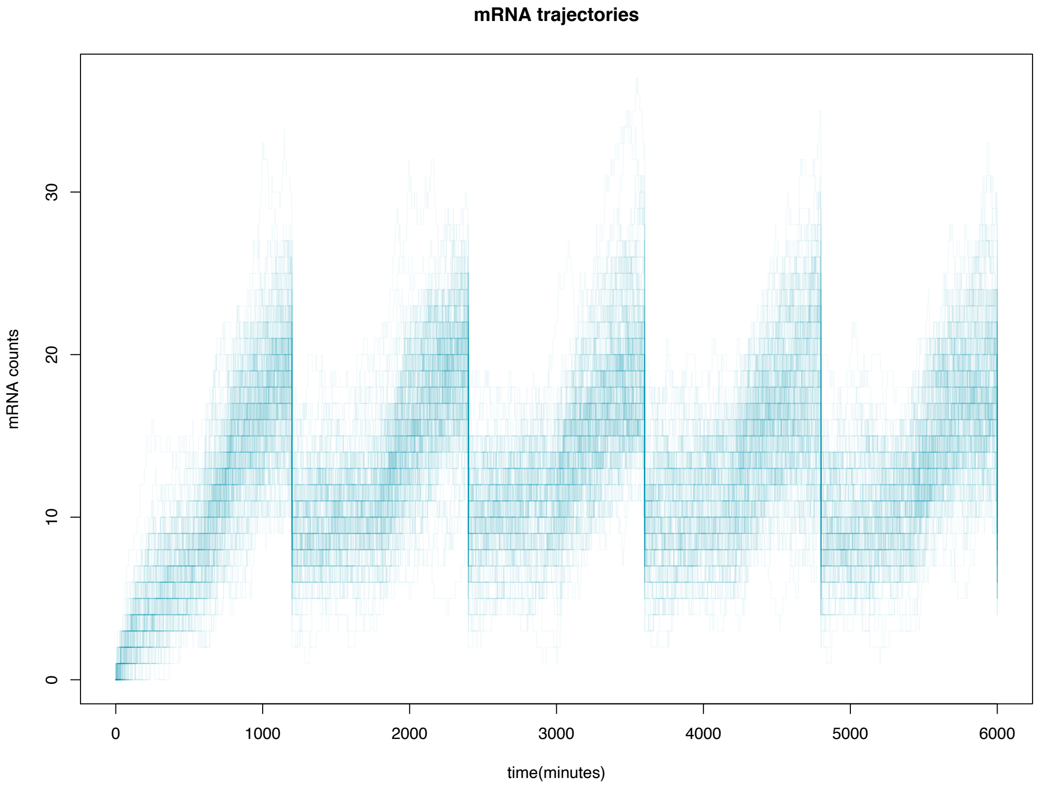

Figure 3.3: **Gillespie simulation of dividing cell cycle model.** 1,000 Gillespie simulations ran with arbitrary parameters based on dividing 2-phase cell cycle model.

$$\begin{aligned}\sum_{x_2=0}^{\infty} P\left(X_2 = x_2\right) z^{x_2} &= \sum_{x_2=0}^{\infty}\sum_{x_1=0}^{\infty} P\left(X_2 = x_2 | X_1 = x_1\right) P\left(X_1 = x_1\right) z^{x_2} \\ &= \sum_{x_1=0}^{\infty}\sum_{x_2=0}^{\infty} P\left(X_2 = x_2 | X_1 = x_1\right) P\left(X_1 = x_1\right) z^{x_2} \\ &= \sum_{x_1=0}^{\infty} \left(P\left(X_1 = x_1\right) \sum_{x_2=0}^{\infty} P\left(X_2 = x_2 | X_1 = x_1\right) z^{x_2} \right). \end{aligned} \tag{3.29}$$

The x_2 sum on the RHS is the p.g.f. for the binomial distribution $B(x_1, p)$. Thus the above equation becomes

$$G_{x_2}(z) = \sum_{x_1=0}^{\infty} P\left(X_1 = x_1\right)\left((1-p)+pz\right)^{x_1}, \tag{3.30}$$

which is the p.g.f. of $P\left(X_1 = x_1\right)$ with parameter ((1 - p) + p z), or in other words,

$$G_{x_2}(z) = G_{x_1}(w), \text{ for } w = p(z-1)+1, \tag{3.31}$$

where G_{x_1} corresponds to the RNA distribution at the end of the first G_2 phase, which we calculated in the first section, see Equation (3.17). Thus

$$G_{x_1}(z) = G_2(z,t) = e^{\frac{\lambda}{\mu}(z-1)(2-e^{-\mu(t-t_1)}-e^{-\mu t})}. \tag{3.32}$$

Hence, equation (3.31) becomes

$$\begin{aligned} G_{x_2}(z,t) &= G_{x_1}(p(z-1)+1) \\ &= e^{\frac{\lambda}{\mu}(p(z-1)+1-1)\left(2-e^{-\mu(t-t_1)}-e^{-\mu t}\right)} \\ &= e^{\frac{\lambda}{\mu}p(z-1)\left(2-e^{-\mu(t-t_1)}-e^{-\mu t}\right)}. \end{aligned} \tag{3.33}$$

Similar to Section 3.1.1, Equation (3.33) can now be used to find the initial conditions for calculating the p.g.f. of the first G_1 phase after the first division. The resulting function is shown below,

$$G_1 = e^{\frac{\lambda}{\mu}(z-1)\left(e^{-\mu(t-(t_1+t_2))}(2p-1)-pe^{-\mu(t-t_1)}-pe^{-t\mu}+1\right)}. \tag{3.34}$$

Then the PMF of RNA numbers during the first G_1 phase after cell division, is given by

$$\begin{aligned} P_{1k}(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \left(1+e^{-\mu(t-(t_1+t_2))}(2p-1)-pe^{-\mu(t-t_1)}-pe^{-t\mu}\right)^k \\ e^{-\frac{\lambda}{\mu}\left(e^{-\mu(t-(t_1+t_2))}(2p-1)-pe^{-\mu(t-t_1)}-pe^{-t\mu}+1\right)}. \end{aligned} \tag{3.35}$$

3.3.2 Probability distribution in G_2, following G_1 after division

Equation (3.34) can be used to calculate the PMF in the next phase. As shown before, we use the conditions at the end of the first phase to calculate the initial conditions for the second phase. This gives us

$$G_2 = e^{\frac{\lambda}{\mu}(z-1)\left(e^{-\mu(t-(t_1+t_2))}(2p-1)-pe^{-\mu(t-t_1)}-pe^{-\mu t}-e^{-\mu(t-(2t_1+t_2))}+2\right)}, \tag{3.36}$$

which in turn gives us

$$P_{1_k}(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \left(e^{-\mu(t-(t_1+t_2))}(p2-1) - pe^{-\mu(t-t_1)} - pe^{-\mu t} - e^{-\mu(t-(2t_1+t_2))} + 2\right)^k \\ e^{-\frac{\lambda}{\mu}\left(e^{-\mu(t-(t_1+t_2))}(2p-1)-pe^{-\mu(t-t_1)}-pe^{-\mu t}-e^{-\mu(t-(2t_1+t_2))}+2\right)}. \tag{3.37}$$

When cross-checked with the simulation results both PMFs are found to agree well.

3.3.3 Cell division model - steady state PMFs

Following the procedure described for the non-dividing model, the time varying solution for the RNA distributions of the division model at steady state can be shown to be

$$P_{1_k}(t) = \frac{1}{k!}\exp\left(\frac{\lambda\left(\frac{e^{\mu(t_1-t)}}{2e^{\mu(t_1+t_2)}-1}-1\right)}{\mu}\right)\left(\frac{\lambda\left(e^{\mu(t_1-t)}-2e^{\mu(t_1+t_2)}+1\right)}{\mu\left(1-2e^{\mu(t_1+t_2)}\right)}\right)^k \tag{3.38}$$

for G1, and

$$P_{2_k}(t) = \frac{1}{k!}\exp\left(\frac{2\lambda\left(\frac{e^{\mu(-t+2t_1+t_2)}}{2e^{\mu(t_1+t_2)}-1}-1\right)}{\mu}\right)\left(\frac{2\lambda\left(1-\frac{e^{\mu(-t+2t_1+t_2)}}{2e^{\mu(t_1+t_2)}-1}\right)}{\mu}\right)^k \tag{3.39}$$

for G2, noting that here we simplify by assuming cell division to be perfect, and thus set the binomial coefficient p to 0.5. We thus have the time varying solution for a model including gene dosage effects and cell division. These results match well when compared to simulated data, see Figure 3.4.

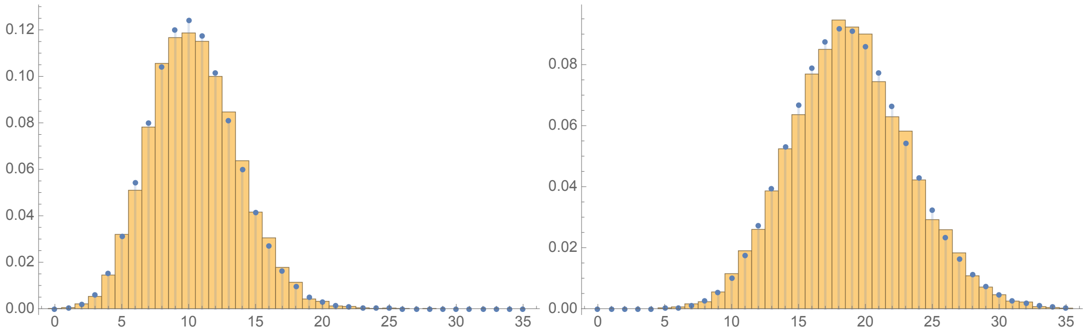

Figure 3.4: **Dividing cell cycle model mRNA distribution.** Histograms represent 10,000 Gillespie simulations ran with arbitrary parameters. Dots represent the solutions of the PMFs from Equations (3.38) and (3.39), using the same parameters.

3.3.4 Cell cycle effects on RNA number mean and variance

At this point we can already ask questions about how the system behaves under different parameter regimes. For instance, how are the mean and variance of RNA molecule numbers affected by changes in cell cycle period? The cell cycle can vary in length depending on culture conditions, but also stochastically, depending on the number of cycle-related proteins inherited by newborn cells (Dowling et al., 2014). Depending on the RNA turnover rate, this can have affect on the resulting distributions, especially for longer living RNA species. Furthermore, we may ask what the effects of variation in the relative duration of the two phases are? This is relevant as the replication timing varies between genes, see 'early' and 'late' replicating genes (Voichek et al., 2016a), which could result in downstream effects on RNA distributions.

To obtain the mean and variance, we use the p.g.f.s for the PMFs (3.38) and (3.39), integrated over their respective time duration. Specifically, it can be shown that

$$E[X] = G'(1)$$

and

$$\mathrm{Var}(X) = G''(1) + G'(1) - [G'(1)]^2 ,$$

where $E[X]$ and $Var[X]$ are the expectation and variance of variable X. We thus have for the first phase

$$E_{G_1}(X) = \frac{\lambda}{\mu t_1}\left(t_1 - \frac{e^{\mu t_1} - 1}{\mu\left(2e^{\mu(t_1+t_2)} - 1\right)}\right)$$

and

$$\mathrm{Var}_{G_1}(X) = \frac{\lambda}{\mu t_1}\left(t_1 - \frac{e^{\mu t_1} - 1}{\mu\left(2e^{\mu(t_1+t_2)} - 1\right)}\right),$$

where E_{G_1} and Var_{G_1} are the expectation and variance over the whole G1 phase. It should be noted that here the expectation and variance are the same. The distribution is thus Poissonian in spite of the changes we have made. In the same way we get the results for the second phase,

$$E_{G_2}(X) = \mathrm{Var}_{G_2}(X) = \frac{\lambda}{\mu}\frac{\left(\frac{2e^{\mu t_1}-1}{2e^{\mu(t_1+t_2)}-1} + 2\mu t_2 - 1\right)}{\mu t_2}.$$

The mean over the whole cell cycle, E_{cc}, is obtained by taking the weighted average over the two expectations, where the weight is proportional to the fraction of the cell cycle spent in each of the two phases,

$$E_{cc}(X) = \frac{E_{G_1}(X)t_1 + E_{G_2}(X)t_2}{t_1 + t_2}.$$

In Figure 3.5, we look at how E_{cc} changes when we vary the cell cycle duration length, as well as the relative durations of the two phases. We further look at how these effects are influenced by RNA stability. In order to make the comparison fair, we adjust the transcription rate for each value of RNA stability in order to keep the overall mean constant. For a range of RNA half life values, changes in cell cycle properties have a considerable effect on the mean RNA number.

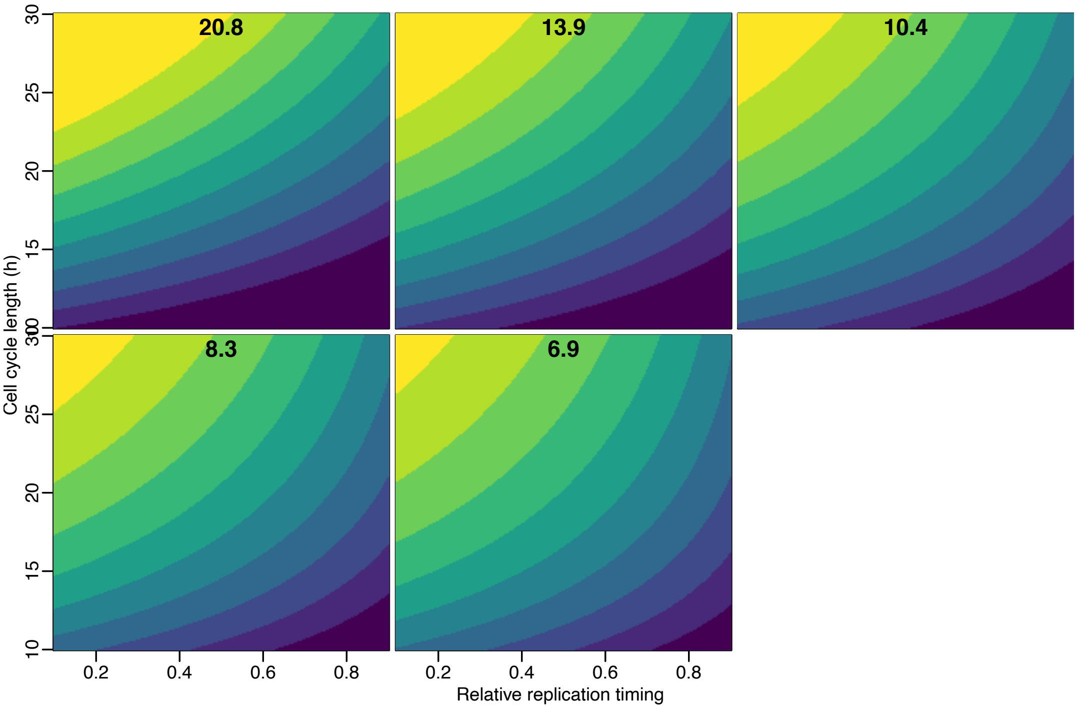

Figure 3.5: **Cell cycle effect on mean RNA number for different RNA turnover rates.** RNA halflives (hours) labelled in black. x-axis refers to the gene replication timing with respect to the duration of the cell cycle period, y-axis refers to the duration of the cell cycle period.

We observe that shorter half lives are generally more resilient to cell cycle effects than longer ones (compare 20.8 hours with 6.9 hours in Figure 3.5). On the other hand, in order to achieve the same level of RNA the transcription rate needs to be increased. This can be shown in Figure 3.7, where the relationship between the strength of the cell cycle effects (measured as the coefficient of variation of means within each panel) and the RNA synthesis rate required to keep the mean RNA level constant can be seen.

3.4 Multi-phase model

So far we have modeled the cell cycle in two phases, defined by the gene copy number before and after DNA replication of the gene locus of interest.

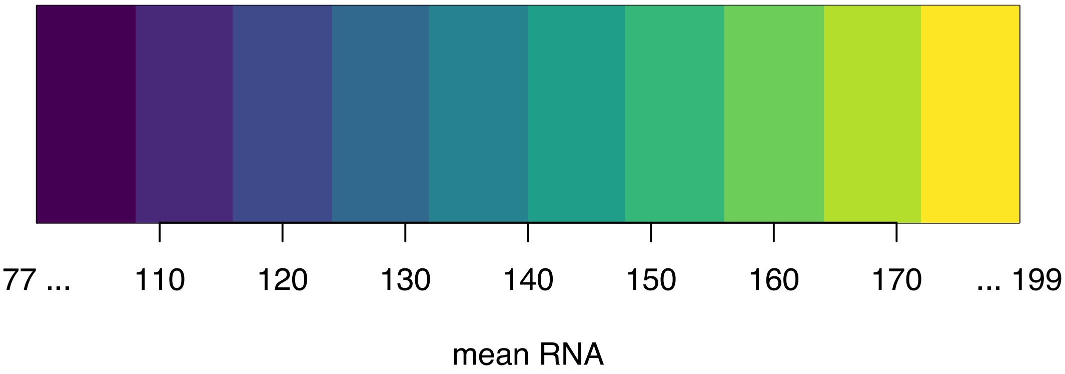

Figure 3.6: **Colour legend for Figure 3.5.** Colours reflect the mean quantity of RNA.

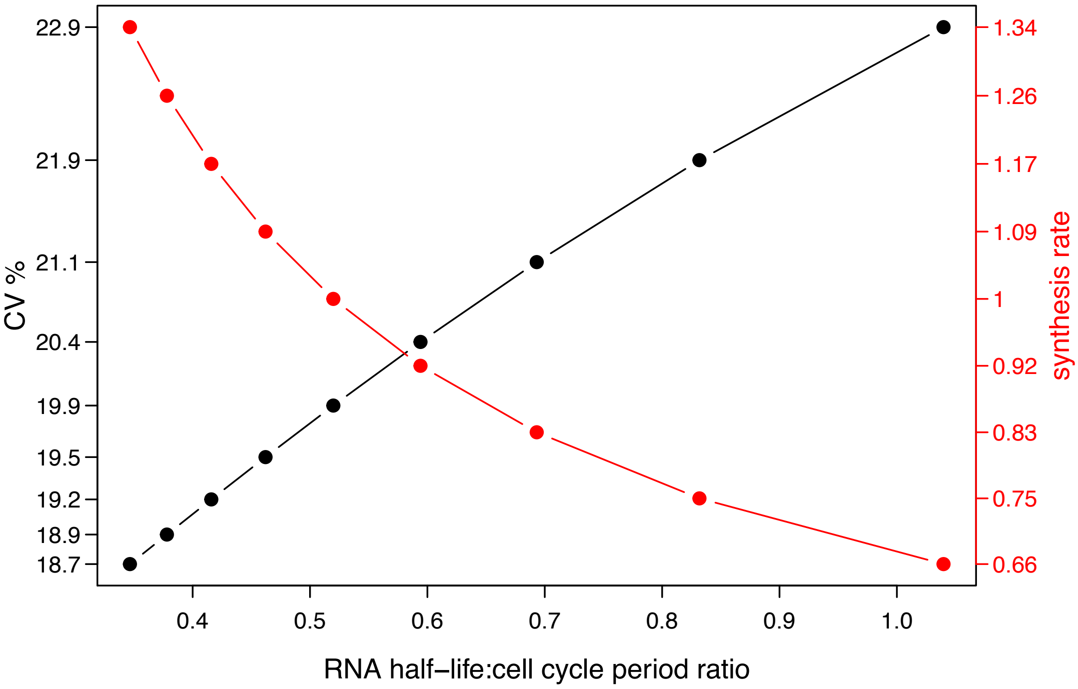

Figure 3.7: **Transcription rate vs cell cycle effects for different RNA turnover rates.** Black dots indicate the increase in variability due to cell cycle effects seen in RNA species with longer half-lives. Red dots indicate the reciprocal synthesis rate required to maintain the same mean RNA numbers between different values of half-lives.

Furthermore, we have made the simplifying assumption that changes in the rate of RNA synthesis during the cell cycle are dependent solely on the copy number of the gene, resulting in a periodic doubling of the rate. Finally, we have assumed that the degradation rate remains constant throughout the cell cycle. In reality, the cell cycle is a complex sequence of events, during which the rates of RNA synthesis and degradation can change for various reasons, such as changes in the concentration of the transcription and degradation machinery, or variations in the state of chromatin.

To account for these changes, here we extend the dividing cell model from two to three cell-cycle phases, in accordance with the easily resolvable experimentally G1, S and G2/M phases, and find the steady state equations for each in the same way as we did in Section 3.1.3. In addition, we let the rate of transcription and degradation constants, λ and μ, vary unconstrained during the cell cycle. We thus end up with a different set of rates for each phase. As before, we make the simplifying assumption that cell division results in equally sized daughter cells, in other words set the RNA binomial partitioning probability to $p = 0.5$. This results in the below set of equations, describing the distribution of RNA molecules in each cell cycle phase:

$$
\begin{aligned}
P_{G1}(t) &= \frac{1}{k!}\left(\frac{\lambda_1}{\mu_1} - \frac{\lambda_1 e^{-\mu_1 t}}{\mu_1} + A_{G1}\right)^k \exp\left(-\left(\frac{\lambda_1}{\mu_1} - \frac{\lambda_1 e^{-\mu_1 t}}{\mu_1} + A_{G1}\right)\right), \\
A_{G1} &= \frac{e^{-\mu_1 t}\left(\frac{\lambda_1 (e^{\mu_1 t_1} - 1)}{\mu_1} + \frac{e^{\mu_1 t_1}(-\lambda_2\mu_3 + \lambda_3\mu_2 e^{\mu_2 t_2 + \mu_3 t_3} + (\lambda_2\mu_3 - \lambda_3\mu_2) e^{\mu_2 t_2})}{\mu_2\mu_3}\right)}{2e^{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3} - 1}
\end{aligned}
\tag{3.40}
$$

$$
\begin{aligned}
P_S(t) &= \frac{1}{k!}\left(\frac{\lambda_2 + \lambda_2\left(-e^{\mu_2(t_1 - t)}\right) + A_S}{\mu_2}\right)^k \exp\left(-\frac{\lambda_2 + \lambda_2\left(-e^{\mu_2(t_1 - t)}\right) + A_S}{\mu_2}\right), \\
A_S &= \frac{e^{\mu_2(t_1 - t)}\left(-\lambda_2\mu_1\mu_3 + 2\lambda_1\mu_2\mu_3 e^{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3} + (\lambda_2\mu_1\mu_3 - \lambda_3\mu_1\mu_2) e^{\mu_2 t_2}\right)}{\mu_1\mu_3\left(2e^{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3} - 1\right)} \\
&+ \frac{e^{\mu_2(t_1 - t)}\left(\mu_2\left(\lambda_3\mu_1 - 2\lambda_1\mu_3\right) e^{\mu_2 t_2 + \mu_3 t_3}\right)}{\mu_1\mu_3\left(2e^{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3} - 1\right)}
\end{aligned}
\tag{3.41}
$$

$$
\begin{aligned}
P_{G2}(t) &= \frac{1}{k!}\left(\frac{\lambda_3 + \lambda_3\left(-e^{\mu_3(-t + t_1 + t_2)}\right) + A_{G2}}{\mu_3}\right)^k \exp\left(-\frac{\lambda_3 + \lambda_3\left(-e^{\mu_3(-t + t_1 + t_2)}\right) + A_{G2}}{\mu_3}\right), \\
A_{G2} &= \frac{e^{\mu_3(-t + t_1 + t_2)}\left(\lambda_3\mu_1\mu_2\left(e^{\mu_3 t_3} - 1\right) + 2\mu_3 e^{\mu_3 t_3}\left(-\lambda_1\mu_2 + \lambda_2\mu_1 e^{\mu_1 t_1 + \mu_2 t_2} + (\lambda_1\mu_2 - \lambda_2\mu_1) e^{\mu_1 t_1}\right)\right)}{\mu_1\mu_2\left(2e^{\mu_1 t_1 + \mu_2 t_2 + \mu_3 t_3} - 1\right)}
\end{aligned}
\tag{3.42}
$$

where t_1, t_2, t_3 are the durations of the G1, S and G2 phases respectively, and $\lambda_1, \lambda_2, \lambda_3$ and μ_1, μ_2, μ_3 their associated transcription and degradation rates, respectively.

Equations (3.40), (3.41) and (3.42) can be generalised to an arbitrary number of phases, and this will be the focus of future progress.

3.5 Cell growth

Finally, we incorporate cell-cycle related cell-growth into the model. As the transcription rate depends on the intracellular concentration of molecules such as RNA polymerase II, ribonucleotides and the DNA template itself, we would expect that an increase in cell volume, which can be due to either intrinsic size variation or cell growth, would lead to a decrease in the stochastic rates of transcription, as the molecules will effectively become diluted. A similar effect could be seen with regards to the degradation machinery, thus affecting RNA turnover rates. It has been shown that RNA concentration homeostasis is actively maintained in mammalian cells of varying sizes, though the underlying mechanisms are not understood (Kempe et al., 2015, Padovan-Merhar et al. (2015)).

Here, we construct a null model whereby cell size has a diminishing effect on the above rates. In order for concentration homeostasis to be preserved, we thus expect the rates to increase sufficiently during the cell cycle in order to compensate for the dilution effects caused by cell size increases. We further assume, as a starting point, that a population of cells will on average follow a linear growth trajectory with respect to the cell cycle period.

We thus model the cell growth as $v(t) = \alpha t + \beta$, where α and β are the growth rate and initial cell volume respectively, and scale the rates λ and μ accordingly. Following the same procedure used to obtain Equation (3.16), we get the time dependent RNA distribution of a growing cell, here without considering DNA doubling or cell division

$$P_k(t) = \frac{1}{k!}\left(\frac{\lambda}{\mu}\right)^k \left(1 - \left(\frac{\alpha t}{\beta} + 1\right)^{-\frac{\mu}{\alpha}}\right)^k e^{-\frac{\lambda\left(1-\left(\frac{\alpha t}{\beta}+1\right)^{-\frac{\mu}{\alpha}}\right)}{\mu}}. \tag{3.43}$$

The resulting equation was compared to simulations and found to agree well (see Figure 3.8. The cell growth model needs to be extended to include changes in cell-cycle phases and cell-division, and this will be the focus of future progress.

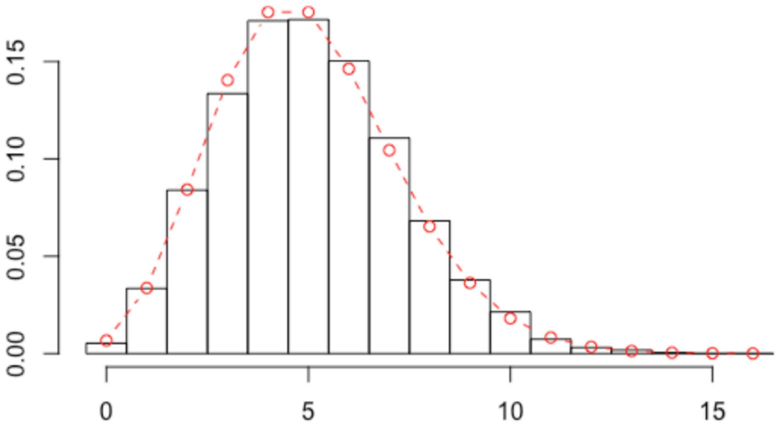

Figure 3.8: **Cell size growth model mRNA distribution.** Histogram represents 10,000 Gillespie simulations ran with arbitrary parameters. Dots and lines represent the solution of the PMFs from Equation (3.43), using the same parameters.

3.6 Discussion

Although the aim of the work in this chapter was to obtain a mathematical framework for analysing experimental data with, some preliminary investigations could be made using the derived models. Specifically, we looked at how properties of the cell cycle such as overall period duration and relative duration of the cell cycle phases affect RNA numbers, and how varying RNA stability changes these effects. Interestingly, for transcripts of the same average abundance, a lower turnover is associated with a greater susceptibility to cell cycle effects, while the opposite is true for more short lived transcripts.

This has likely to do with the fact that the number of transcript molecules associated with a higher turnover rate can more rapidly be adjusted to changes in rates associated with cell cycle progression. The opposite is true for longer living transcripts, which result in transcript numbers from previous phases carrying over to the next, thus skewing the average number. This demonstrates the trade-off that exists between the higher responsiveness associated with increased RNA turnover rate, and the cost of increased synthesis required to maintain a constant RNA level.

Although models based on mass action kinetics such as the ones developed here can be insightful and have been used extensively towards understanding gene expression noise, the implicit assumption made that the cell is a well mixed compartment is arguably an oversimplification. For example, rather than occurring homogeneously in the cell transcription has been shown to occur within specialised locations in the nucleus termed transcription factories, which have recently been suggested to form dynamically by liquid-liquid phase separation (see Cramer (2019) for a recent review). These sub compartments can increase the effective concentration of the transcriptional machinery which would affect transcription noise, going against the assumption that transcription kinetics can be explained by Brownian motion alone. It is therefore important to corroborate any insights made using models such as these presented here with simulations and models that incorporate the above and other effects, such as molecular crowding (Golkaram et al., 2016), and this should be the focus of future research.

Here, we chose to specifically ignore genetic promoter switching in order to understand whether extrinsic factors such as changes in gene dosage and cell size during the cell cycle, as well as partitioning of molecules at cell division are sufficient in describing the super-poissonian distributions of mRNA molecules seen in populations of growing cells, in line with recent results (Ietswaart et al., 2017; Zopf et al., 2013; Battich et al., 2015; Klein et al., 2015). It would be useful to see how well the presented model fits to the experimental measurements, compared with a model that incorporates promoter switching alongside cell cycle effects. To do so, we could follow a similar appoach to the one described in this chapter. Specifically, we could start from the basic formulation of the telegraph model and proceed by incorporating the increase in gene dosage following gene replication. This could be achieved by doubling the frequency of transcriptional bursts, in order to model the existence of twice the source of gene template. In order to model the increase in cell size, either burst frequency or burst size could be scaled, and the most agreeable to the data model could be chosen, as described in (Sun et al., 2019b). Partitioning of molecules at division could be done in the same way as here. As the resulting functions would be harder to derive than the simpler Poisson model, it may be necessary to resort to obtaining just the moments of the PMFs, rather than the full analytical solutions.

To summarise, we have thus far developed four stochastic models of gene expression which include different aspects of cell growth. The first model, described in Section 3.1.3, considers the effects of the change in gene dosage on RNA transcription, due to DNA replication. The second model, described in Section 3.3.3 extends the first by further including the random partitioning of RNA molecules at cell division. Both of these effects have been shown to be important extrinsic sources of gene expression variation, capable of shaping the observed RNA distributions in a population of asynchronously growing cells. In the third model, described in Section 3.4, we further extended the two-phase division model to three phases, and allowed the rates of transcription and degradation to vary in each phase, thus accounting for likely phase-specific changes in rates overlaying the gene dosage effects. All three models were solved analytically using probability generating functions, and their steady states were derived. These models can thus be compared with smFISH data, and used to infer the kinetics of different RNA species. This will be the subject of future work.

A fourth model is considered in Section 3.5, incorporating the effects of cell size increase on transcription and degradation rates, and the time varying solution was found for the first phase. This model can be extended to include multiple cell cycle phases and RNA partitioning at cell division, as in the previous three models, and that will be the focus of future work. However, the increase of cell size during the mammalian cell cycle is not well understood, which is why obtaining more experimental measurements was prioritised here, as opposed to further developing the model. In Chapter 4, we look into how such measurements can be obtained from asynchronous populations of cells, by computationally analysing multiple cell cycle markers, in parallel to measurements of cell size and RNA kinetics.

Chapter 4

Multiparameteric cell cycle analysis using Probability State Modelling

4.1 Background

Multiple markers can be simultaneously employed to resolve the cell cycle at a much higher resolution than DNA content alone can afford. These can be used in combination with other biomarkers such as fluorescent protein tags or antibodies to track the regulation of different molecules with respect to the cell cycle. An example of this is found in (Kafri et al., 2013), where DNA stained with a fluorescent dye, and a fluorescent reporter with a geminin degron which marks the exit of the G1 phase were used in combination to track changes in cell growth rate. We are using a similar method while including a third marker, namely Cdt1 (Sakaue-Sawano et al., 2008), in order to increase the resolution of the cell cycle.

Different sub-populations can be extracted from cytometry data using polygon compartments ("gates") on bivariate plots, a practice used extensively in cell cycle analysis (Jacobberger et al., 2012). Two important limitations became apparent when using this approach. Firstly, it is reliant on manual subsetting ("gating") to be performed on each sample, which is itself subjective and therefore makes the comparison between samples problematic. Secondly, as the number of measured cell cycle reporters increases, so does the complexity and labour intensity of the subsetting process, making it poorly scalable. This issue, also called "the curse of dimentionality" is well known in cytometry data analysis, and has thus led to many gating-free methods being developed in recent years (Mair et al., 2016).

Kafri et al. (2013) use a dimensionality reduction algorithm relying on density peak tracking to define a single trajectory, as a function of two independent cell cycle reporters. Superior methods for exploring the trajectory of cells through time have since been developed, primarily for studying differentiation (see (Saeys et al., 2016) for a review). Such techniques include Probability State Modelling (PSM) (Bagwell et al., 2015b), and nearest-neighbour network exploration (Bendall et al., 2014; Gut et al., 2015; Setty et al., 2016), all of which are better suited for handling data with more than two dimensions.

Unlike non-parametric methods such as density peak tracking (Kafri et al., 2013) and nearest-neighbor network exploration (Bendall et al., 2014; Gut et al., 2015; Setty et al., 2016), PSM (Bagwell et al., 2015b) is a parametric method based on quantile modelling. This allows us to design a suitable model, which is subsequently fitted to the experimental data. In our case, the advantages of using a parametric approach is that prior biological knowledge can be encoded in the constraints of the model and the resulting parameters from fits to different samples can easily be compared. Furthermore, the probability of each data point fitting to the model can be evaluated, enabling the identification of outliers and cells not belonging to the population of interest. Finally, once a model has been fitted, it can be used to generate simulated data, which can be used for estimating the predictive power of the method.

4.2 Building descriptive *fucci* models

In Section 2.5 we introduced PSM (Bagwell et al., 2015b) and described our implementation for analysing the cell cycle, based on DNA quantity as measured by Hoechst staining. Here we look at how our cell cycle model can be expanded to encompass the measurement of two additional cell cycle markers based on he *fucci* reporter system, geminin and Cdt1.

4.2.1 Selecting model priors

Before we can use PSM to predict the state of each cell, we need to have a good understanding of how the levels of the cell cycle markers we are using change as a cell progresses through the cell cycle. In our case, we are using the *fucci* degron markers Cdt1 and geminin, which have a relatively well known cell cycle pattern (Sakaue-Sawano et al., 2008). To confirm that our measurements reflect the theory, we look at how the *fucci* Cdt1 and geminin reporter proteins correlate with a known quantity such as DNA amount, see Figure 4.1.

Our prior biological knowledge of the three available reporters can thus be summarised in table 3.1. What is not known is when exactly in cell cycle time these events take place and at what rate the accumulation ('rising') and degradation ('falling') occur. PSM enables us to test how well different assumptions describe these transitions.

Marker/Phase	eG1	G1	G1-S	S	S-G2	G2/M	M
DNA	n = 1	n = 1	n = 1	rising	rising	n = 2	n = 2
Cdt1	low	rising	max	falling	low	low	low
geminin	low	low	rising	rising	rising	rising	drops

As a first step, we use a DNA model fitted as described in Section 2.6 to analyse the above data based on the DNA measurements alone, in order to observe how the *fucci* markers change with time and roughly characterise their expression (see

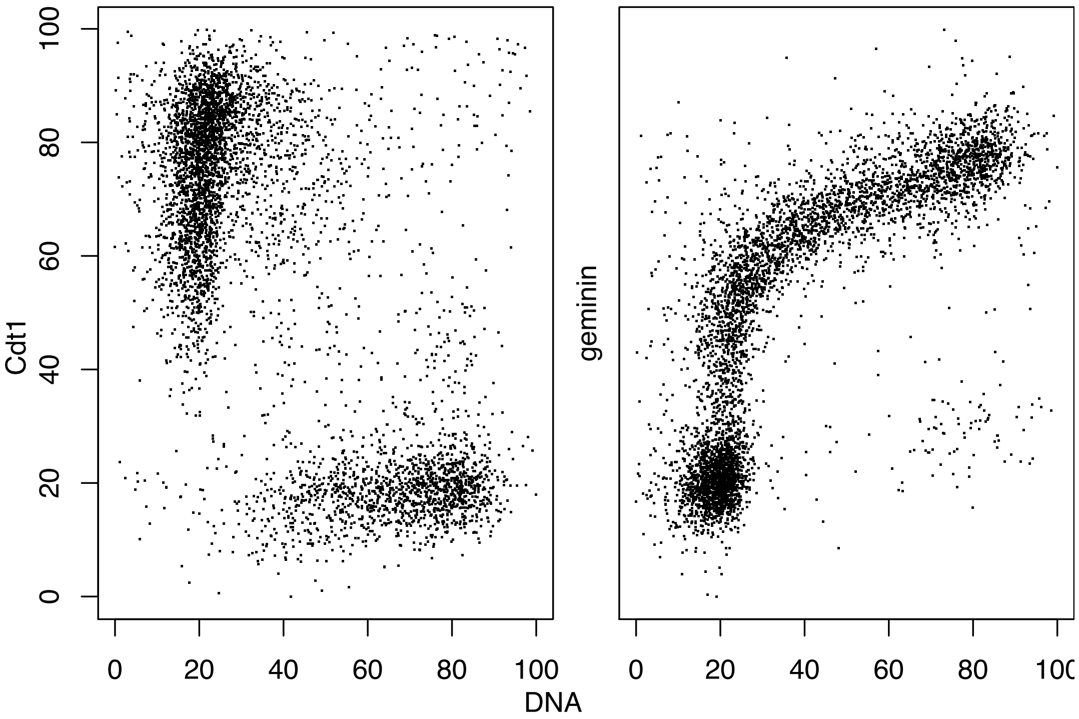

Figure 4.1: **DNA vs *fucci* measurements.** DNA of *fucci* cells was stained with Hoechst for quantification by flow cytometry. Measurements are normalised from 0 to 100, by setting the minimum to 0 and maximum to 100, following outlier exclusion.

Figure 4.2).

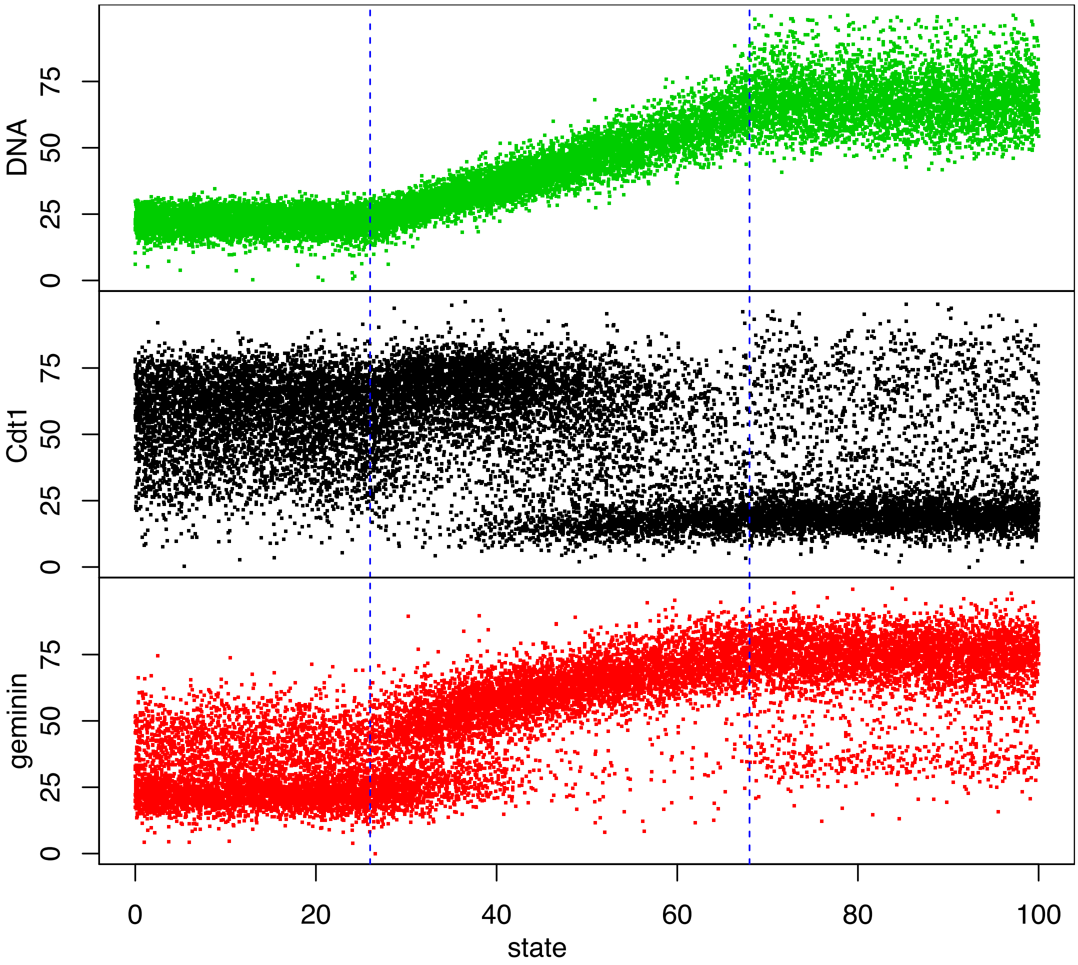

Figure 4.2: **Analysis of *fucci* reporters by DNA staining and PSM.** Data shown is analysed by PSM based on a DNA model fitted as described in the main text. Around 15,000 *fucci* cells were stained by Hoechst DNA dye and analysed by flow cytometry.

This picture is in agreement with what we know from the literature (Sakaue-Sawano et al., 2008). Specifically, Cdt1 appears to be high during G1 and the start of S phase and low during the G2/M phases. For geminin, we can clearly see two populations during the G1 phase, suggesting that there is an upregulation step during that time, followed by a gradual increase until the onset of the G2 phase. What also becomes clear is that the low level of Cdt1 during the G1 phase is not the same as the low level during the G2/M phases, which is important to note when assigning a model to the transition.

Surprisingly, we find a population of highly expressing Cdt1 cells during the G2/M phases (Figure 4.2) that was not described by Sakaue-Sawano et al. (2008). Following up on this observation, we found in the literature that Cdt1 does indeed become upregulated towards the end of the cell cycle (Williams and Stoeber, 2012). Taking together the information from the above table and the observations from Figure 4.2 we propose a set of simple piece-wise linear model to describe the transitions of the Cdt1 and geminin reporters (see Figure 4.3).

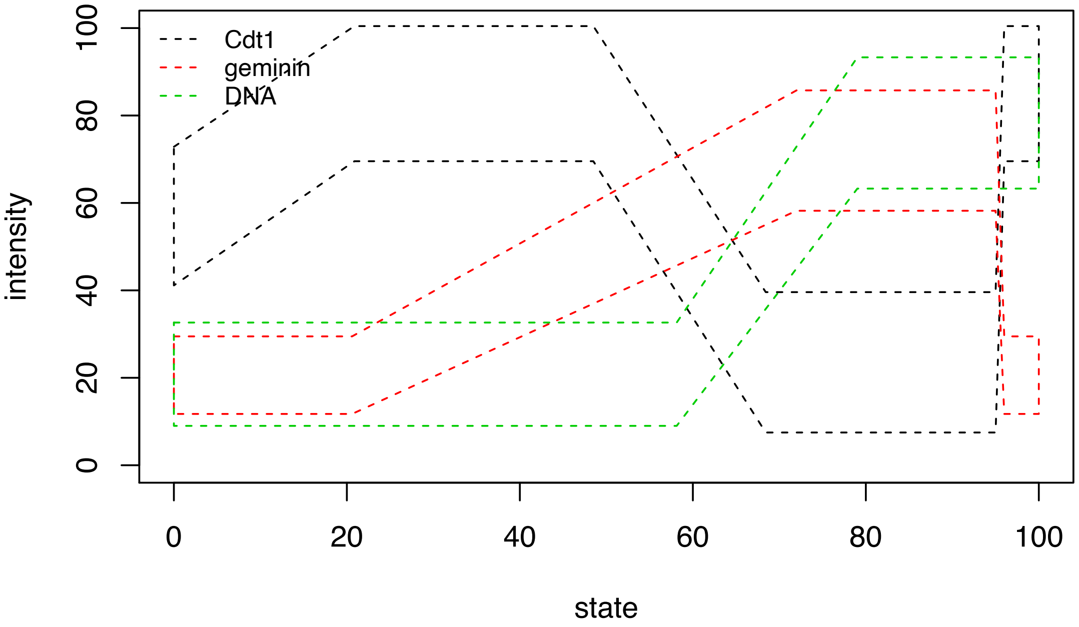

Figure 4.3: **Proposed *fucci* models.**

4.2.2 Multidimensional analysis with PSM

In Section 2.5 we saw how PSM can be used to analyse the cell cycle based on a single measurement - DNA quantity. One of the main advantages of PSM over bivariate gating is that it is specifically designed to handle multiple correlated measurements of a process (Bagwell et al., 2015b), such as the cell cycle. Here, we use PSM to analyse the cell cycle based on all three available measurements; DNA, geminin and Cdt1.

To do so, the steps detailed in Section 2.5 need to be adapted to the multidimensional case, as described in (Bagwell et al., 2015b). First, for each additional measurement we generate a probability E-matrix, as shown in Figure 4.4 for Cdt1 and geminin. Similarly to 2.5.3, for each measurement we assign the data into bins corresponding to the rows of the corresponding measurements (Figure 4.4). This way, for each data point we obtain three probability weight vectors, one for each measurement.

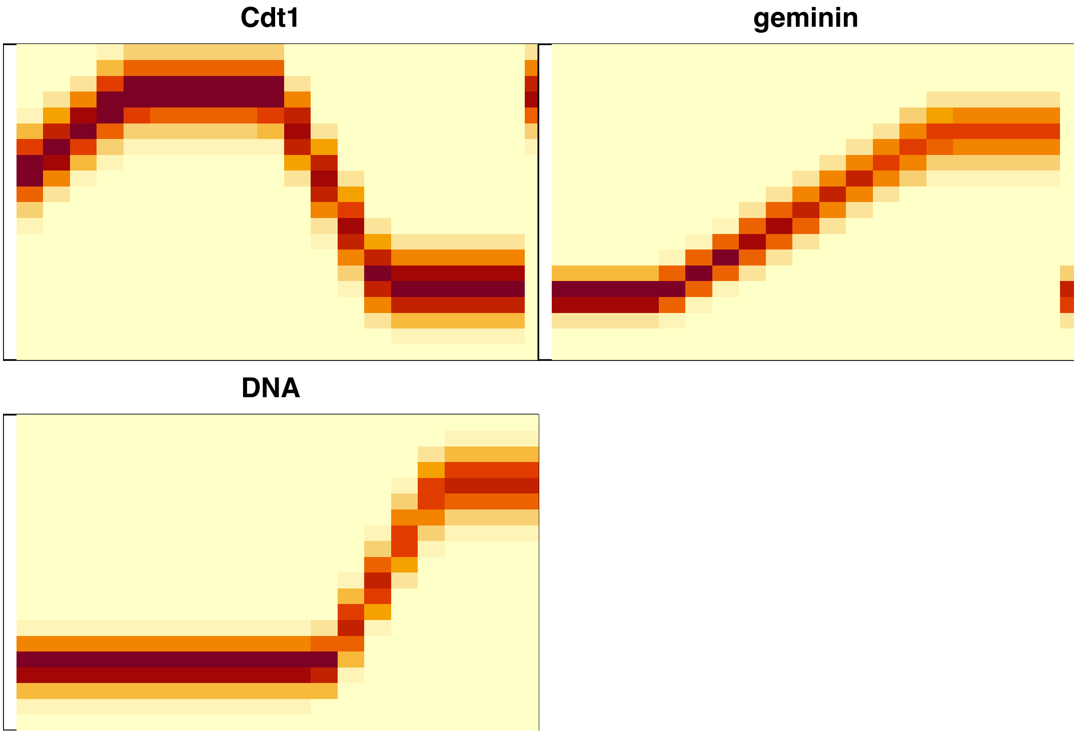

Figure 4.4: ***E*-matrices for *fucci* models.** Matrices corresponding to the proposed models of Cdt1, geminin and DNA.

So far nothing has changed from the unidimentional implementation. In the next step, for each data point we combine the three corresponding weight vectors by element-wise multiplication, in order to obtain a consensus weight vector. The resulting vector describes the probability distribution of finding a point with the given combination of measurement intensities along the cell cycle period. Once a consensus weight vector is obtained for every data point, we use it to assign a cell cycle position to each by performing weighted sampling.

Depending on the number of measurement channels there will be an equivalent number of resulting empirical ES-matrices, computed in the same way as shown in

Section 2.5.4. These can be compared with their respective E-matrices to derive an equivalent number of goodness of fit scores. The mean of these scores can be used as an overall score of how well the three models describe the data. Once again, we use this as the output of our objective function, which we subject to Bayesian Optimisation using mlrMBO (Bischl et al., 2017), as detailed in Section 2.6.

4.3 *fucci* model fitting

Following this method, we run ten instances of the optimisation algorithm in order to estimate the optimisation error. The fitting is performed in three steps, as detailed in Sections 4.3.1 to 4.3.5. This iterative approach enables us to add layers of complexity to the model gradually, as we understand more about the cell cycle markers being used. Furthermore, it allows us to determine which cells can be accounted for by the model, versus cells which may not belong to the process (ie quiescent cells). This way we can chose whether to exlude these cells, or advance the model accordingly. We find that a continuous piecewise linear function is adequate for describing the DNA and most of the geminin transitions, but not for the Cdt1 transition. This is possibly due to a high degree of Cdt1 expression heterogeneity in certain phases of the cell cycle. We make appropriate changes in the model to test whether we can account for this.

4.3.1 Starting conditions

Using peak identification followed by the fitting of a half-Gaussian we first obtain a set of initial estimates for the mean and standard deviation of the steady states in the same way as in Section 2.6. For each measurement, we use the relative density corresponding to the fitted Gaussian on each mode, to obtain an estimate of the proportion of time spent in each.

Following the same approach, we get a rough estimate of the time spent in each of the transitioning phases in order to get starting values for the change-point timings. Specifically, for geminin we assign the remaining density to the transition between the two constant phases, and do the same for the two transitioning phases of Cdt1 by first splitting the density in half. The resulting preliminary models can be seen in Figure 4.3.

4.3.2 Optimisation Round 1:3

The aim of this step is to obtain better starting values for the levels and time points that describe the model of each measurement. We do so using Bayesian Optimisation to fit the preliminary models shown in Figure 4.3 to our data. The resulting values will be used to inform our model and boundary constraints, accordingly, for the next round of optimisation.

Here, we use our initial estimates of the timings of the above cell cycle events from Section 4.3.1 to derive a first set of boundary constraints. At this stage, we keep these constraints very broad as we the uncertainty about the true values is high. Specifically, we use a ± 20 window on either side of each time point estimate, correcting appropriately for any that may have a distance <20 from either extreme of the cell cycle period. The resulting boundary constraints can be seen in Figure 4.5.

Furthermore, we add two additional breakpoints in the middle of each transitory phase in the geminin and DNA models, and set equally broad constraints for each. This way we allow for greater flexibility in the shape of these phases. The mean levels of these new breakpoints, as well as that for the initial G1 phase in Cdt1 for which we have no prior information, are kept free (± 10 window), while the rest are for the time being kept fixed at the initial estimations based on peak identification.

We run the Bayesian Optimisation algorithm ten times for just under 1,500 iterations each. The mean and standard deviations of the resulting optimisation paths can be seen in Figure 4.6, from which we can see that beyond 800 iterations

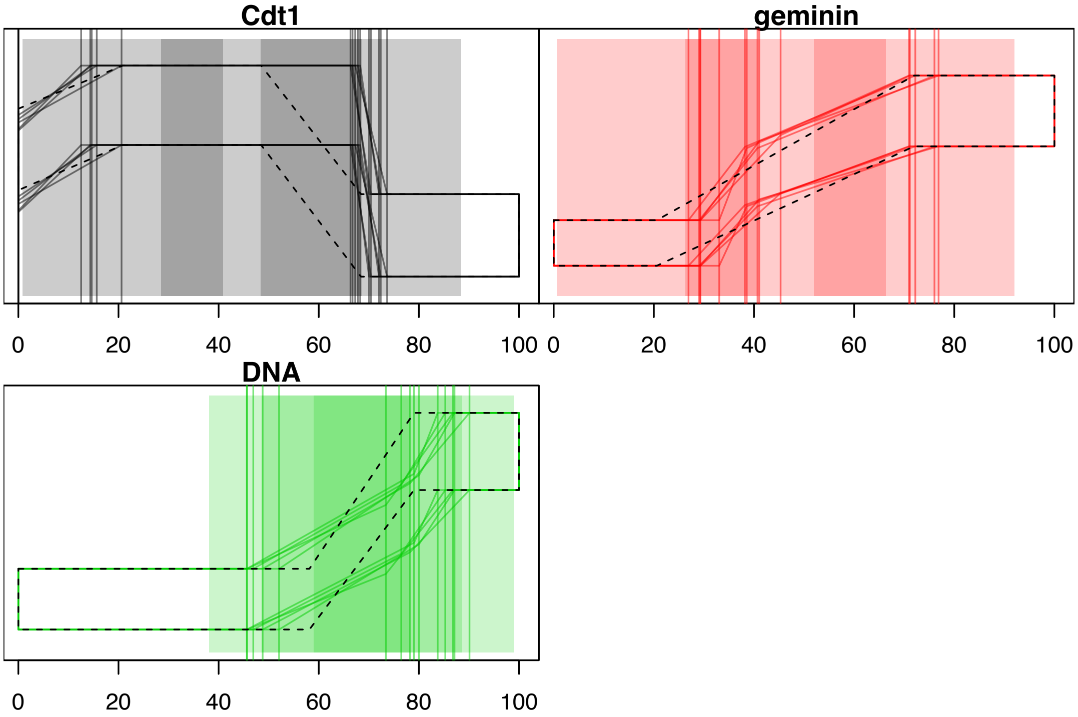

Figure 4.5: **First optimisation step.** Each polygon corresponds to a distrinct optimisation run. Vertial lines correspond to fitted timepoints, translucent rectangles correspond to respective boundaries.

we get diminishing returns.

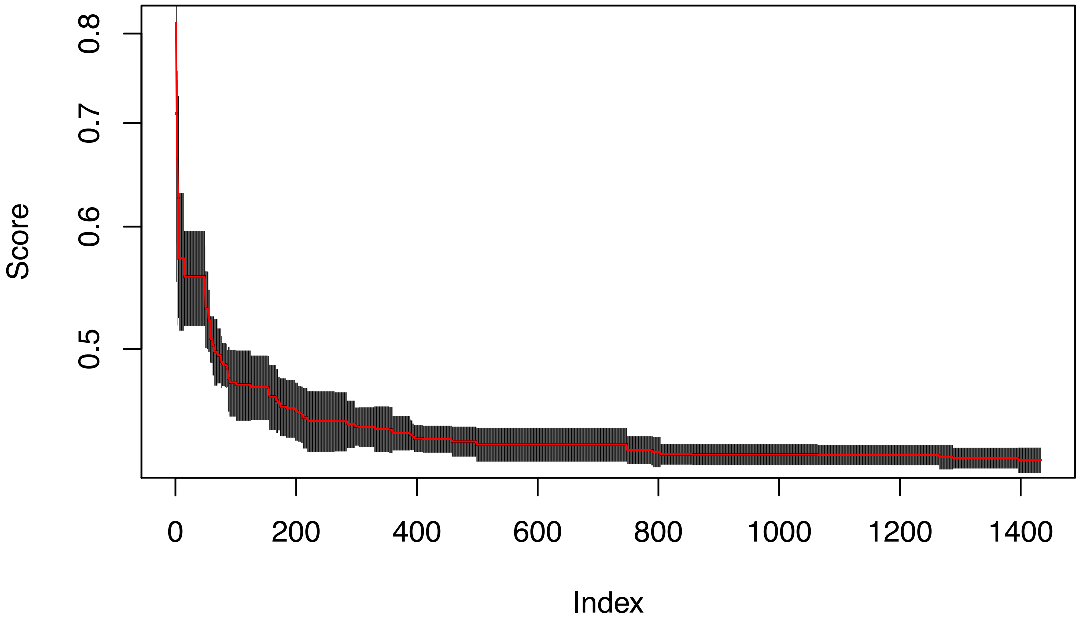

Figure 4.6: **Minimisation path of PSM objective function by Bayesian Optimisation.**

Figure 4.5 shows that the G1 to S transition on the DNA measurement occurs much sooner than our initial estimations, derived from the DNA histograms. This discrepancy is likely due to the contribution of the early S phase to the G1 phase peak which leads to an overestimation of the G1 duration by the half-Gaussian fit. The additional information contributed by the *fucci* reporters via the PSM process is likely to have led to a better estimation of the relative G1 phase duration. Similarly, we can see that the middle Cdt1 time point is near its boundary. These boundaries need to be adjusted accordingly prior to the next optimisation run.

Before repeating the run with the updated constraints, we look closer into the resulting fits in order to determine whether any further adjustments need to be made to the models. We use the mean of all the optimisation runs for each parameter of our cell cycle models to analyse our data using PSM (see Figure 4.7). We can see that the first Cdt1 stationary phase is in fact an increasing phase, as can been seen by the points with higher expression level than the fixed level between the 28th and 70th quantiles. Similarly, G1 phase DNA measurements are on average lower than the fixed value.

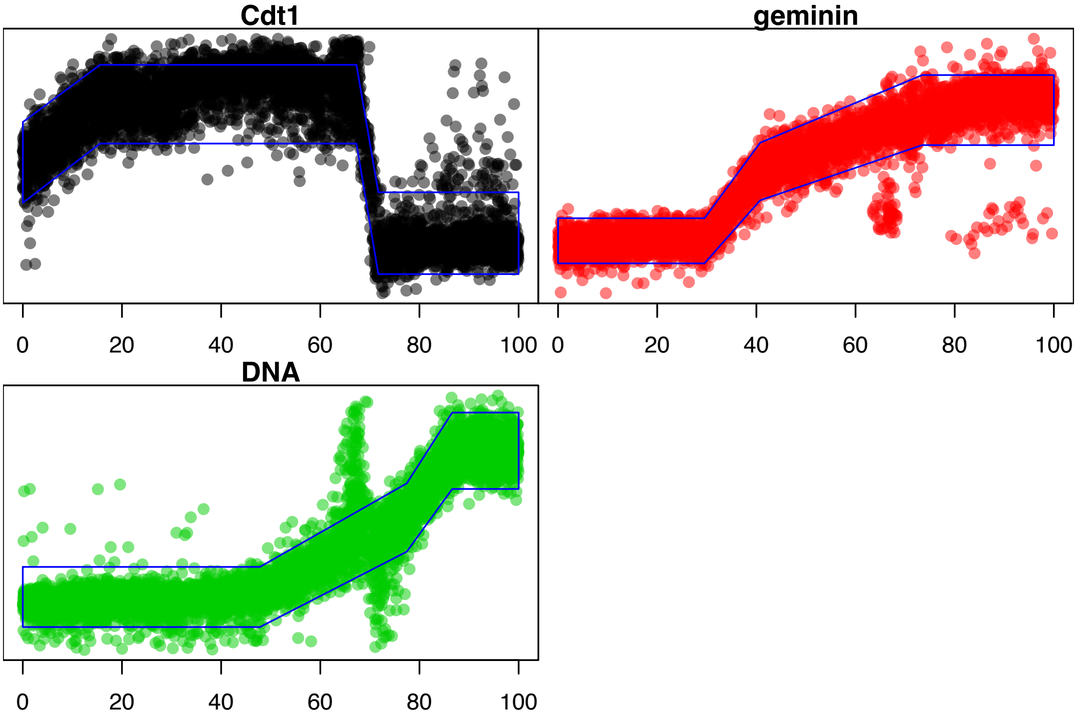

Figure 4.7: **PSM analysis using averaged models from round 1.**

Furthermore, we see that aside from certain density inhomogeneities along the cell cycle period, there are also certain clusters of points lying outside of our modeled trajectories (see 65-70, geminin and DNA; 70 to 75, DNA). These points are clearly not described by our proposed models, either because they correspond to experimental outliers or due to deficiencies of the models themselves.

In the next section, we use the *chi*-square distance of each point's measurements from the proposed models to get a metric of how agreeable the two are, in order to identify outliers (Bagwell et al., 2015b), as well as potential shortcomings of the proposed models.

4.3.3 Outlier Identification

Often in cytometry data analysis, there are data points which do not belong to the population of interest, and thus need to be excluded prior to analysis. Such inhomogeneities can arise from the presence of fractured cells, or cells which happen to be clustered together and as a result are registered as a single event by the flow cytometer. Furthermore, cells which are not actively dividing can be also be considered outliers in our case, and thus confound the interpretation of our results.

Usually the population of interest can be roughly isolated using polygon 'gates' on bivariate plots of select measurements. More sophisticated methods also exist which can identify likely subpopulations within the data set using a modelling approach (Saeys et al., 2016). Although these methods can substantially reduce the number of outliers entering the data analysis pipeline, they are not perfect, and a variable number of outliers is usually still present.

Bagwell et al. (2015b) use a criterion to identify the data points which do not comply with the proposed model, based on the *chi*-square statistic. A threshold 'probability of exclusion' needs to be defined *a priori*, for which the corresponding *chi*-square value can be obtained by solving the inverse CDF of the *chi*-squared distribution with degrees of freedom equal to the number of measurements being modeled. Then the *chi*-square distance of each data point from the proposed model is measured and compared to the threshold value, in order to determine whether it is rejected or not. This is different from the goodness of fit scores described in Section 2.5.4, as it does not measure how well the model describes the *whole* data set. Instead, it measures how likely each *individual* data point is to be part of the proposed model.

Although the above method was proposed by Bagwell et al. (2015b) for identifying outliers, it can also be used to detect deficiencies of the model. Here, we set a probability of exclusion threshold of $p = 10^{-3}$ to identify data points which are not in agreement with the model (see Figure 4.8). We apply k-means to classify these points into clusters, and check whether these can be identified as known populations, which can subsequently either be specifically included in the model or discarded.

At this stage it is important to identify which points correspond to true outliers,

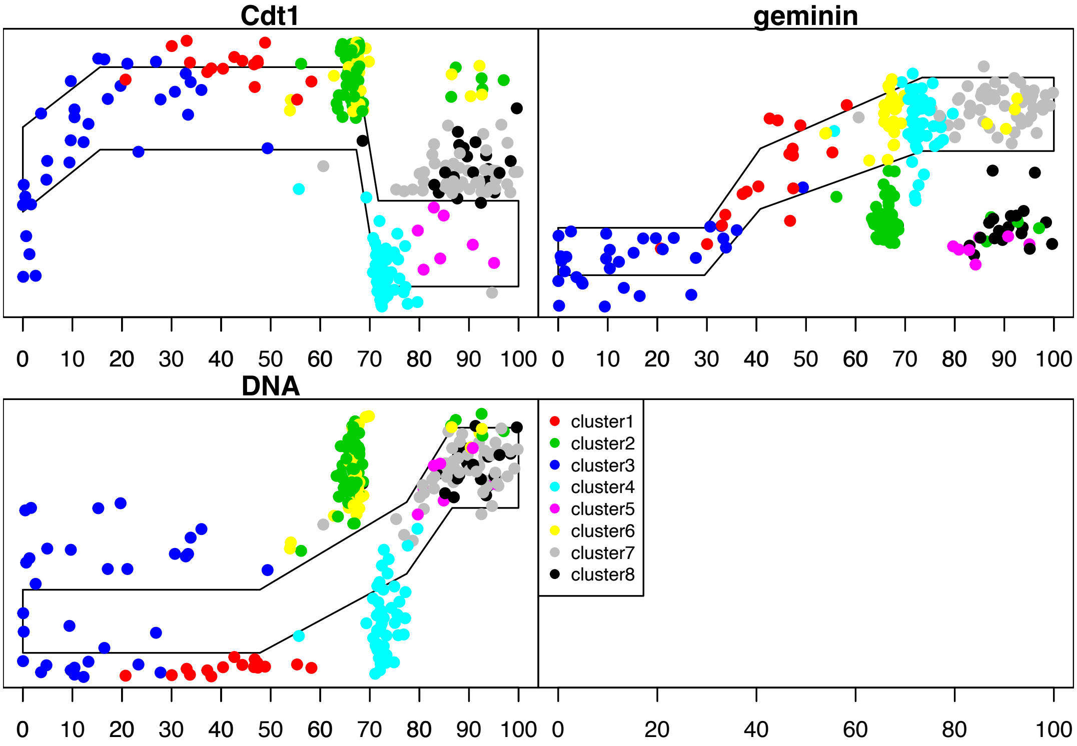

Figure 4.8: **Outliers identified using chi-square test.** Clusters identified using k-means

and which are just flagged due to shortcomings of the proposed models. The former need to be excluded from the analysis as they can impact the fitting process and downstream interpretation, while the latter can inform us on how to improve our proposed cell cycle models to better describe the observed data.

To that end, we use the bivariate plots of the three markers to help us identify these points (see Figure 4.9). Clusters 1 and 3 have a very low DNA amount so are likely to correspond to fractured cells. 5 to 8 and 2 appear in the G2/M transition and M phase, which we have not yet explicitly modeled and could therefore appear as outliers. Cluster 4 cannot be identified as easily. The G1-level DNA amount and low Cdt1 levels suggest these cells may belong to an undividing, quiescent population of cells, though this interpretation is not agreeable with the medium to high geminin expression level.

These results suggest that additional breakpoints need to be inserted in our

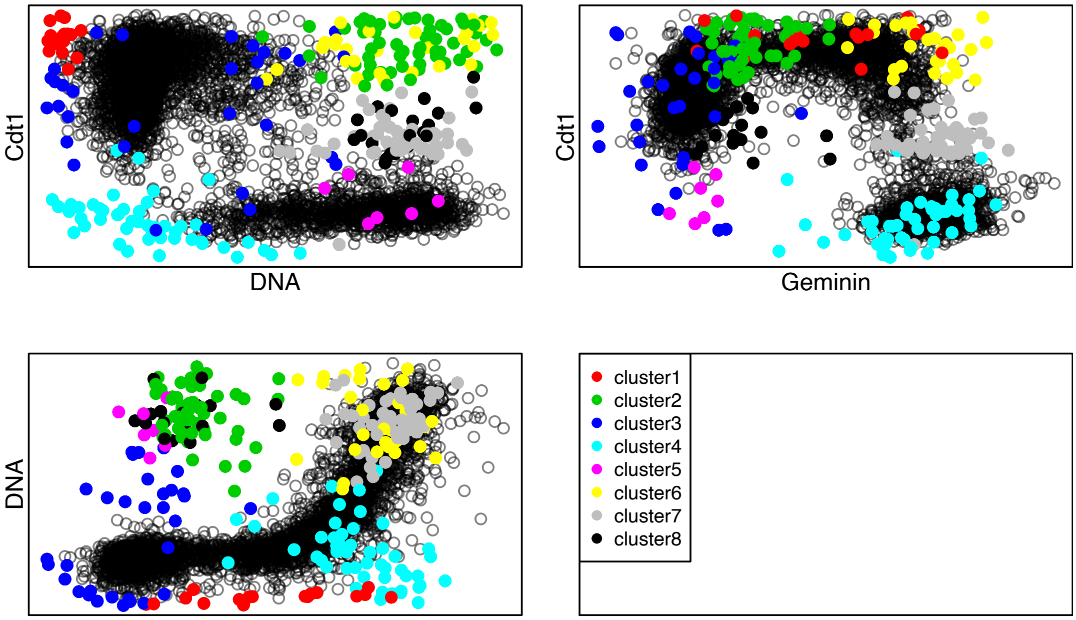

Figure 4.9: **Corresponding locations of outliers on bivariate plots.**

piecewise linear models to account for changes in the expression of geminin and Cdt1 during the M phase. Furthermore, we broaden the boundaries of the Cdt1 middle time point and the DNA G1 phase and manually adjust the relevant levels. The resulting proposed models and constraints can be seen in Figure 4.10.

4.3.4 Optimisation Round 2:3

We run the Bayesian Optimiser again, this time using the improved model from section 4.3.3. This is repeated ten times for roughly around 1,200 iterations each. The resulting fits can be seen in Figure 4.10. The mean and standard deviation of the optimisation paths can be seen in Figure 4.11. We use the mean of the resulting fits for each parameter to analyse our data using PSM, similar to 4.3.2. The resulting trajectories can be seen in Figure 4.12.

We can see in Figure 4.12 that points in the geminin and Cdt1 trajectories have now been shifted to fill the M phases. We can demonstrate this by plotting the clusters of outliers identified in Section 4.3.3 observing their positions (Figure 4.13). We can see that the points in clusters 5 to 8 that we identified as potential M

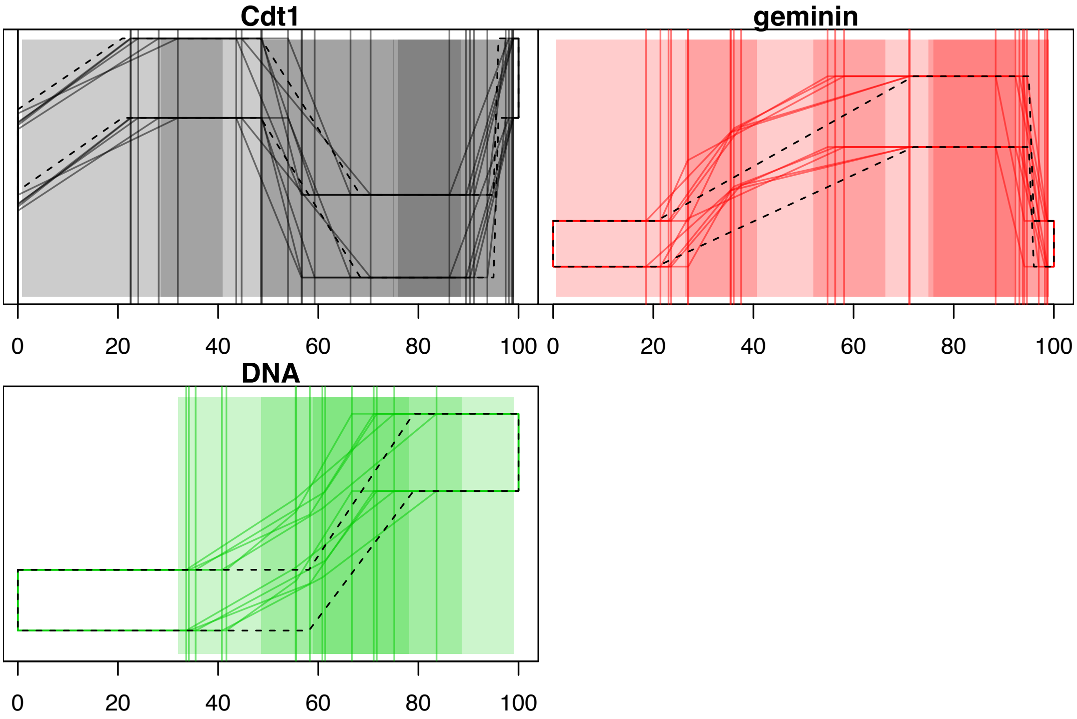

Figure 4.10: **Second optimisation step.** Each polygon corresponds to a distinct optimisation run. Vertical lines correspond to fitted timepoints, translucent rectangles correspond to respective boundaries.

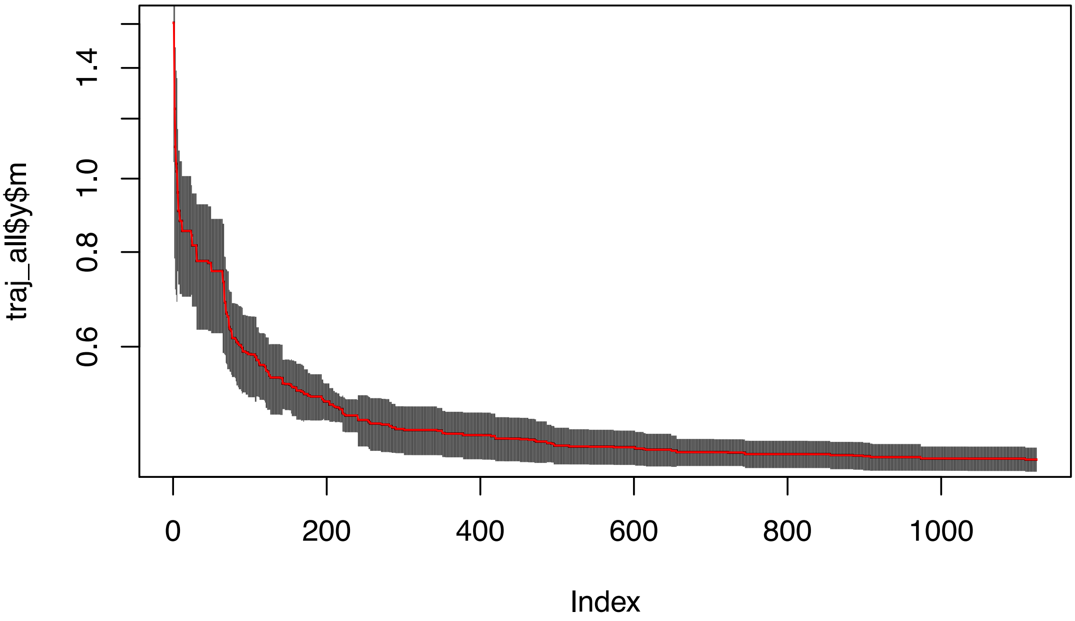

Figure 4.11: **Minimisation path of PSM objective function by Bayesian Optimisation.**

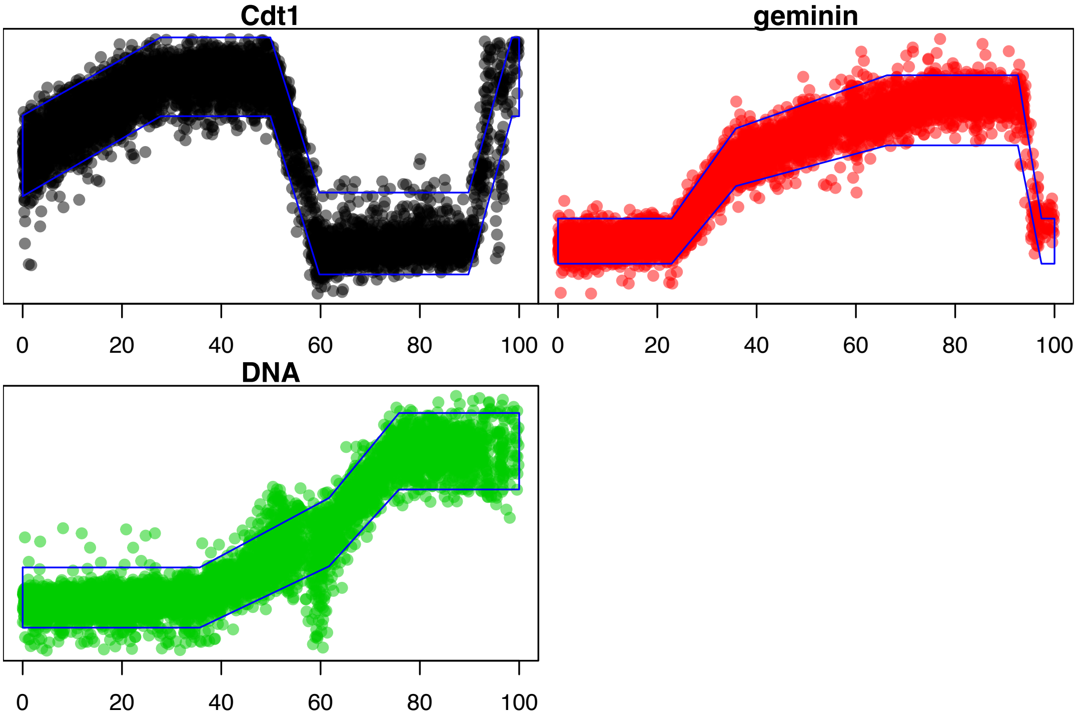

Figure 4.12: **PSM analysis using averaged models.**

phase cells have indeed been allocated in the newly defined M phase. Furthermore, if we count the number of points that fall beyond the cutoff threshold of $p = 10^{-3}$, we find that 43% of these points have been accounted for by the improved model and are no longed deemed outliers (note the reduced number of points in Figure 4.14 compared to Figure 4.13).

Although the advanced model appears to lead to an improved overall fit, there are still clusters of points that remain unaccounted for. Specifically, even though many of the points in clusters 2 and 5 to 8 have been accounted for by the implementation of the M phase, there are still points occupying where clusters 2, 6 and 4 were, and there is a region of low density between them. As these observations coincide with the Cdt1 downregulation step, we ask whether the specific step is adequately described by our proposed model.

There are at least two explanations for why the model may be inadequate in that phase. First, that the Cdt1 downregulation may be so rapid (especially towards the end) that there just are not enough cells present within the sample

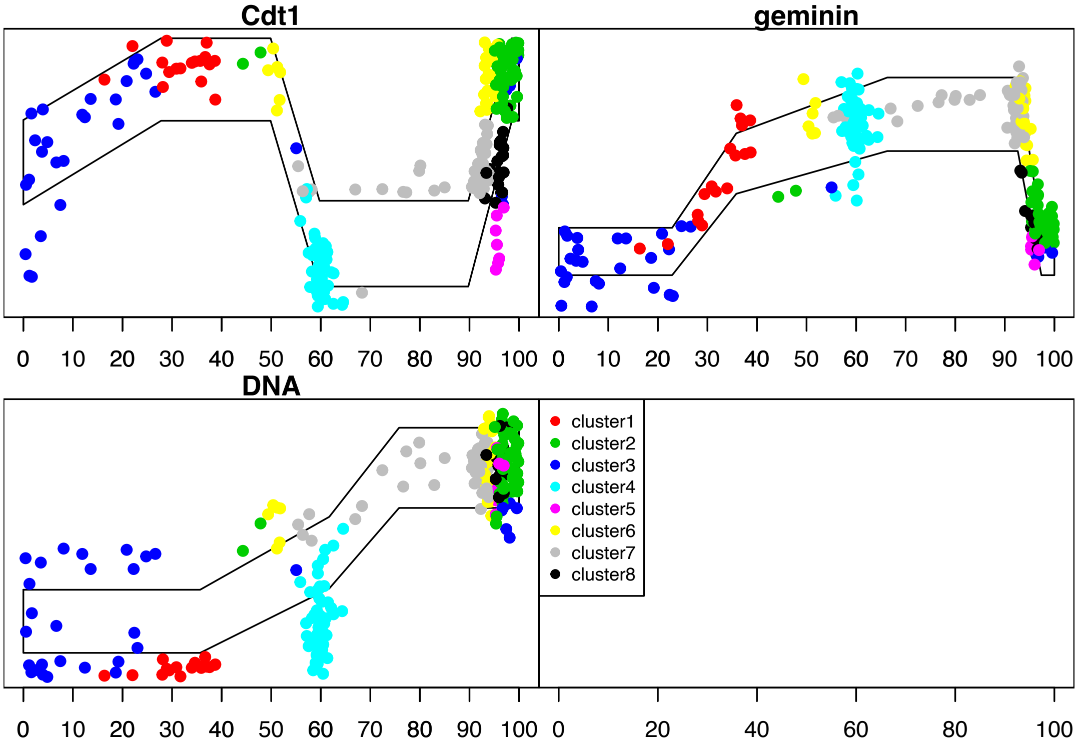

Figure 4.13: **Location of outliers from Round 1 after second round of optimisation.**

with intermediate levels of expression, resulting in a region of low density in that step. Alternatively, there is a high degree of heterogeneity in the expression of Cdt1 between cells at that stage of the cell cycle, which means that a continuous trajectory between the high and low levels of expression would fail to describe the whole population adequately.

If the former explanation is correct, adding additional breakpoints in this step would give sufficient flexibility to allow for a very rapid downregulation, thus removing the low density region. We test this hypothesis by adding a further breakpoint before the final optimisation run. With respect to the rest of the model, we see a similar low density region in the M phase.

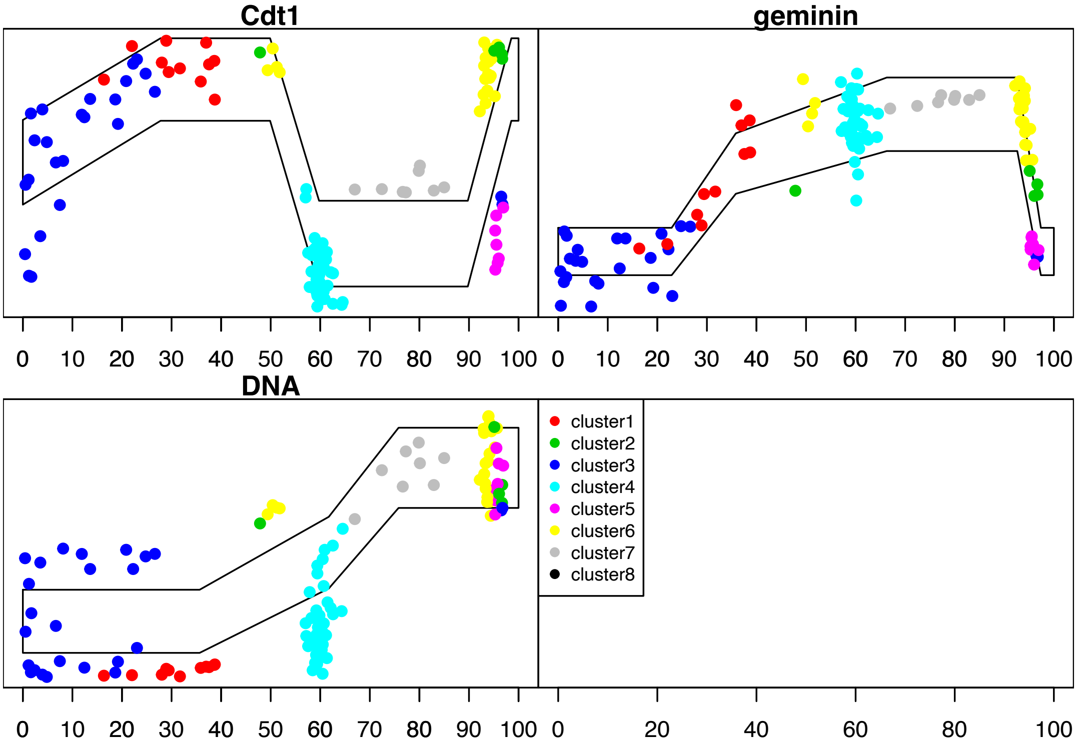

Figure 4.14: **Remaining outliers following addition of M phase to the model.**

4.3.5 Optimisation Round 3:3

In Section 4.3.4 we used the priors from Sections 4.3.2 and 4.3.1 to fit the time points of the three proposed models, as well as the three mean levels that we had no priors for. The initial models were further improved by adding an M phase step in the geminin and Cdt1 transitions. Here, we use the resulting parameters from Section 4.3.4 to design a new set of more constrained boundaries for the time points, while letting the rest of the parameters vary.

Specifically, we fit all time points, mean measurement levels and errors simultaneously, amounting to a total of 44 parameters, including an additional breakpoint in the Cdt1 downregulation step. The resulting optimisation path can be seen in Figure 4.15. The resulting fit can be seen in Figure 4.16. We obtain a consensus fit by averaging over all the runs, and use it to analyse the cell cycle of our data using PSM (Figure 4.17).

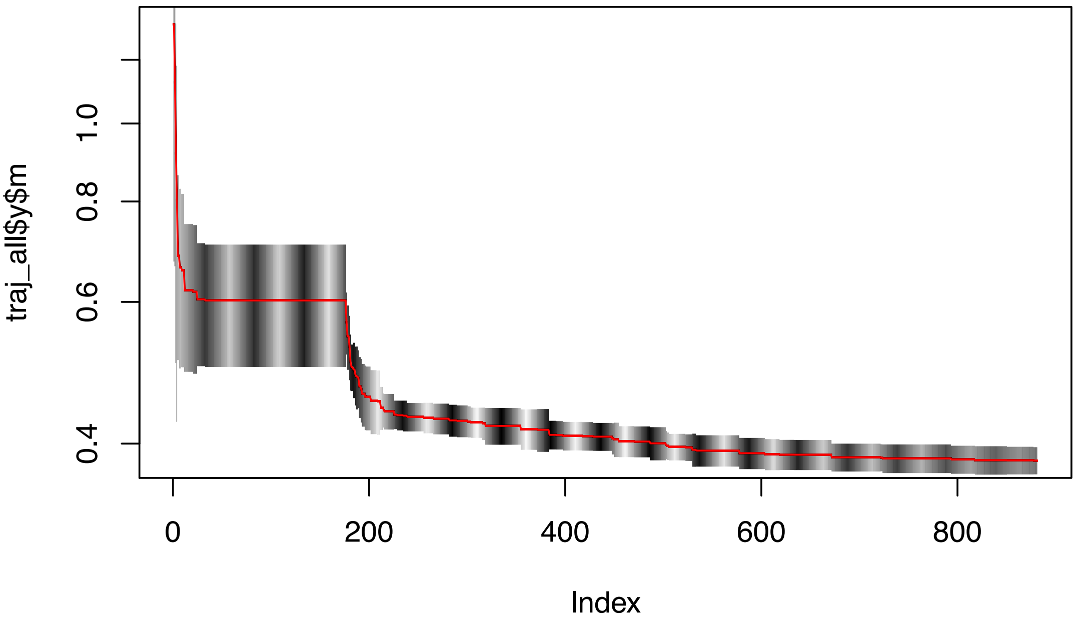

Figure 4.15: **Minimisation path of PSM objective function by Bayesian Optimisation.**

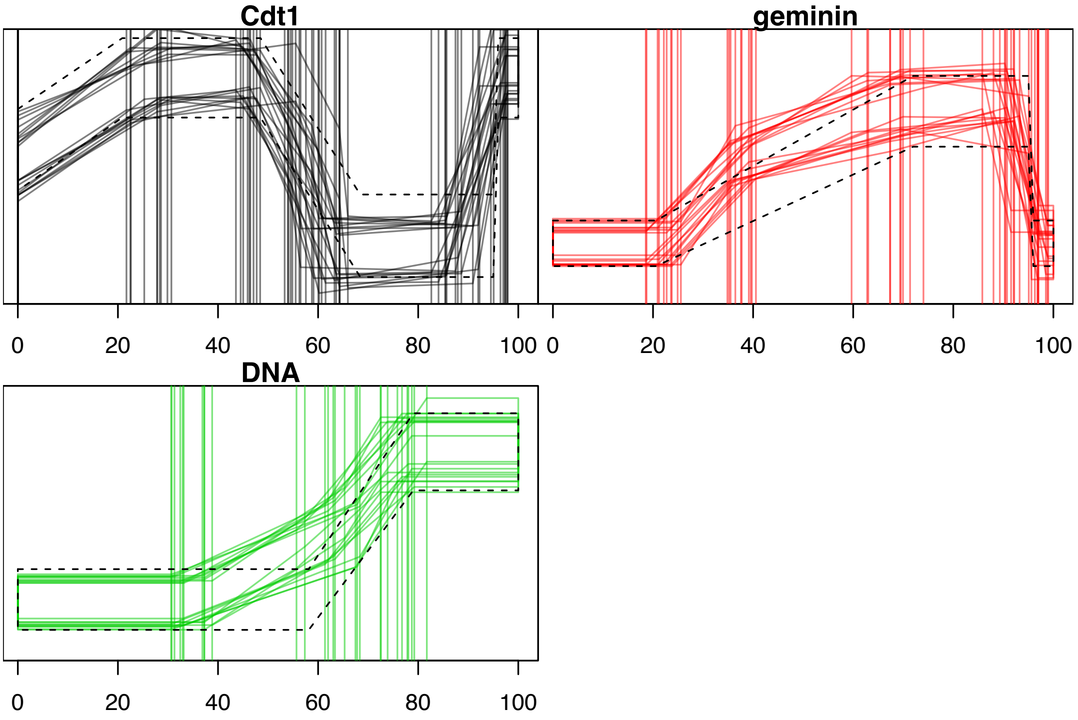

Figure 4.16: **Final optimisation step.** Each polygon corresponds to a distinct optimisation run. Vertical lines correspond to fitted timepoints, translucent rectangles correspond to respective boundaries.

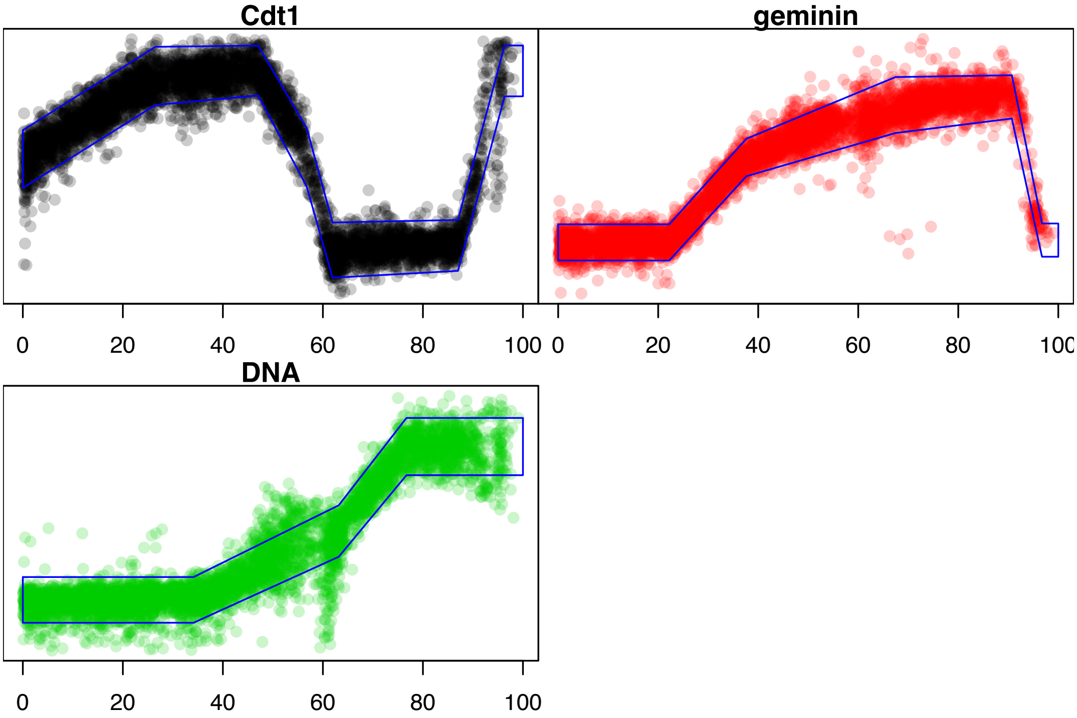

Figure 4.17: **PSM analysis using averaged models.**

In Figure 4.17 we observe that, although the addition of a breakpoint in the Cdt1 downregulation step has resulted in a slightly different profile, the low density region at the 60th quantile persists. This suggests that a continuous piecewise linear model may be insufficient for describing that stage of the Cdt1 transition due to the high degree of expression heterogeneity at that point. Cdt1 level is therefore likely not to be informative in that stage of the cell cycle, as cells with either high or low level of Cdt1 expression can be found there. Furthermore, attempting to describe that stage of the Cdt1 transition using a simple linear model leads to a poor fit. For these reasons, in the next section we exclude Cdt1 from the inference process, specifically during the phase with increased uncertainty.

4.3.6 Accounting for heterogeneity

In Section 4.3.5 we found that a piecewise linear model inadequately described the Cdt1 downregulation step at the G1-S boundary. Here, we discuss a method for selectively ignoring regions within a transition with high uncertainty such as this one from the PSM cell cycle analysis. The method relies on appropriately averaging over the corresponding subset of the probability matrix obtained in Section 2.5.2. This can be seen in Figure 4.18 for the G1-S boundary of the Cdt1 phase, where the selected region has been replaced by a uniformly distributed density with value equal to the mean density corresponding to that region. This has the effect that the precise positioning of data points that fall in that region will only be informed by the alternative measurements, DNA and geminin.

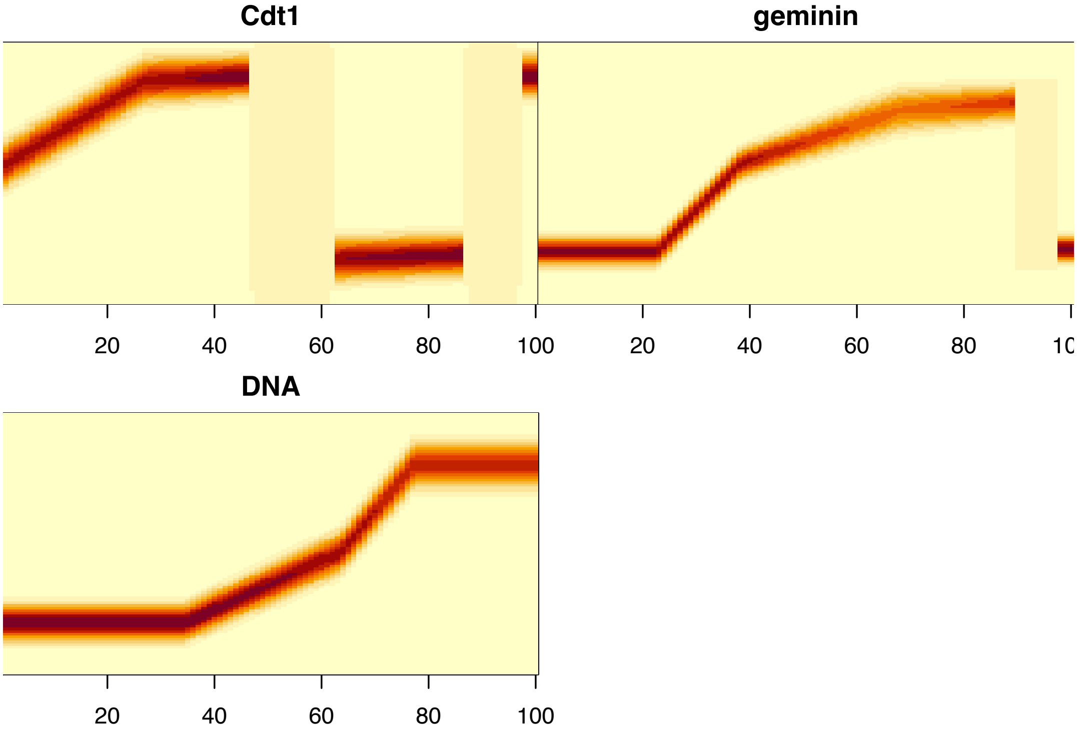

Figure 4.18: **Probability matrices following selective averaging of ambiguous phases.**

This way we can selectively prevent any region of a given measurement from influencing the PSM analysis. Using this approach, we fitted the model from Section 4.3.5 again to the data, employing the same optimisation boundaries. The resulting analysis can be seen in Figure 4.19. As we can see from this plot, there is no longer a region of low density at the G1-S boundary (see DNA and geminin trajectories). Furthermore, we can clearly see the co-existence of two distinct Cdt1 expressing populations in that region, as resolved by DNA and geminin. We keep

Table 4.2: Table of model parameters for Cdt1 transition

	1	2	3	4	5	6
p mean	0.00	24.97	52.36	63.09	93.19	96.93
p sd	0.00	3.70	3.40	2.98	2.03	1.26
l mean	52.43	83.37	86.11	17.92	19.35	87.68
l sd	3.28	2.12	3.29	1.67	1.83	4.50
s mean	6.36	5.00	5.78	5.84	4.35	5.17
s sd	1.27	1.30	0.83	1.43	0.44	1.23

these results for further analysis (Chapter 5).

The parameters resulting from the simulations are summarised in Tables 4.2, 4.3 and 4.4 for Cdt1, geminin and DNA respectively, where columns refer to breakpoints and *p*, *l* and *s* refer to the timing, intesnity level and spread of each breakpoint. Mean and *sd* of each parameter refer to the resulting mean and standard deviations of ten parameter-fitting optimisation runs.

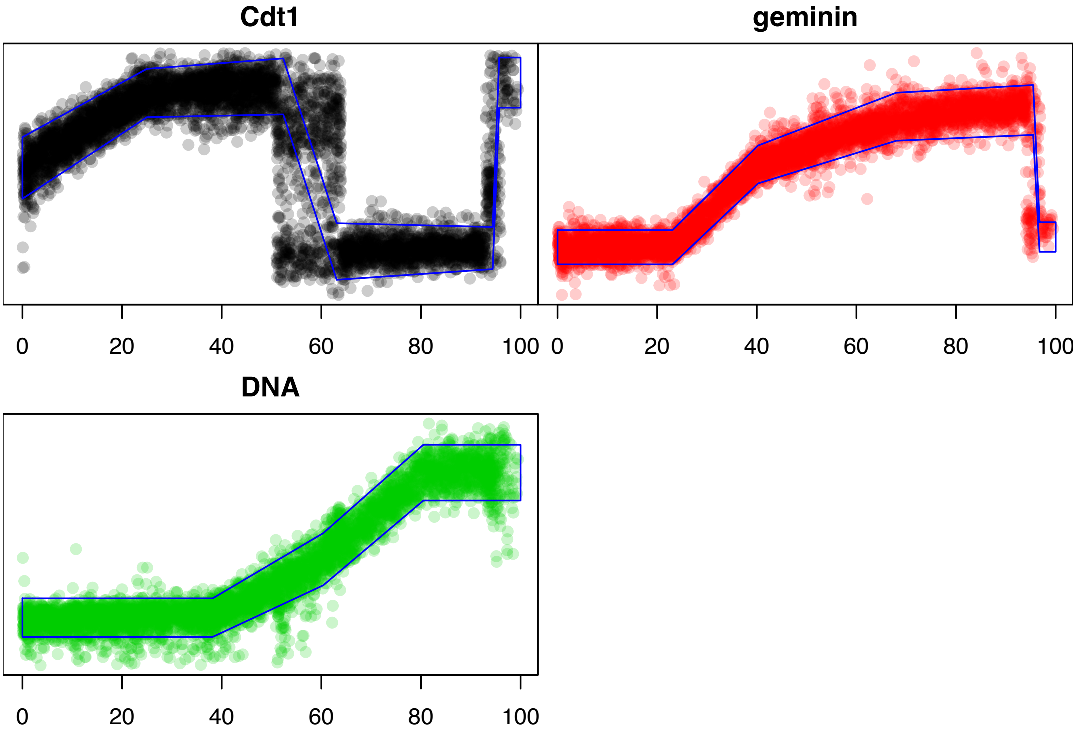

Figure 4.19: **PSM analysis using averaged models.**

Although these results look promising, and potentially greatly improve the utility

Table 4.3: Table of model parameters for geminin transition

	1	2	3	4	5
p mean	23.08	40.26	68.00	94.79	97.46
p sd	2.90	1.70	4.22	2.32	1.25
l mean	19.71	53.96	73.71	76.50	23.95
l sd	0.60	2.91	1.81	1.82	4.51
s mean	3.54	3.92	4.96	5.16	3.06
s sd	0.43	0.81	0.71	0.94	0.49

Table 4.4: Table of model parameters for DNA transition

	1	2	3
p mean	38.18	60.39	80.49
p sd	2.01	3.81	2.81
l mean	19.72	43.88	79.70
l sd	0.67	3.89	3.38
s mean	3.99	5.39	5.77
s sd	0.54	0.71	1.16

of our cell cycle model, we have not looked at the effect of averaging regions of the probability matrix on the resulting scoring of the objective function, which would in turn affect the fitting of the model. However, we have identified the cause underlying the persistent low density region at the G1/S phase, as the heterogeneous expression of Cdt1 at that stage. This effect has also been characterised by Grant et al. (2018), who found that the timing of the Cdt1 reporter fusion degradation depends on its expression level in each cell, during G1. The authors instead proposed an alternative biomarker based on the fusion of the PIP degron tag to a reporter protein, in order to mark the G1/S and S/G2 transitions more accurately.

While ignoring Cdt1 during the G1/S phase allows us to rely on geminin and DNA quantity during that phase, leading to a continuous cell cycle trajectory, it remains to be verified how this affects the fitting procedure. An alternative approach could be to switch the Cdt1 marker for a more precise cell-cycle marker of the G1-S phase transition, such as one based on PIP (Grant et al., 2018). Similarly, an alternative reporter could be used to mark the end of mitosis and onset of G1, such as the replication initiation factor Cdc6-GFP fusions used by Duursma and Agami (2005). This would be especially useful here, as the present set of markers cannot resolve the G2/M phases. An advantage of the *fucci* system is the use of inactive protein fragments of Cdt1 and geminin. A similar protein reporter fusion could be constructed using the Cdc6 C-terminal sequence, which is a target for ubiquitination by Cyclin F in mammalian cells (Walter et al., 2016).

4.3.7 Precision Assessment

Based on the resulting models we can determine how well the cell cycle can be resolved, using simulated data. To do so, we first generate synthetic data using the above models, as seen in Figure 4.20. The benefit of using synthetic data is that the *true* cell cycle state is known for every cell, allowing us to compare it to that *predicted* when analysing the synthetic data via PSM. This gives us an estimation of the resulting resolution of the cell cycle model, as seen in Figure 4.21, which allows us to interpret the data in light of the limitations of the model.

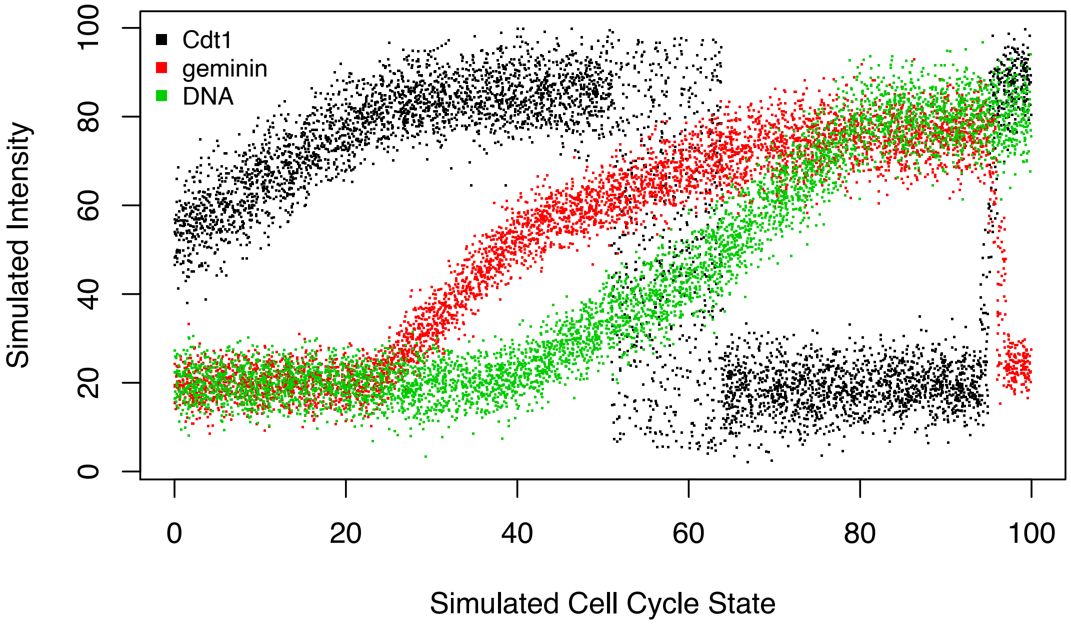

Figure 4.20: **Simulated data based on *fucci* models predicted in Section 4.3.** 10,000 data points simulated using PSM.

As the resolution often varies throughout the cell cycle, this is particularly useful

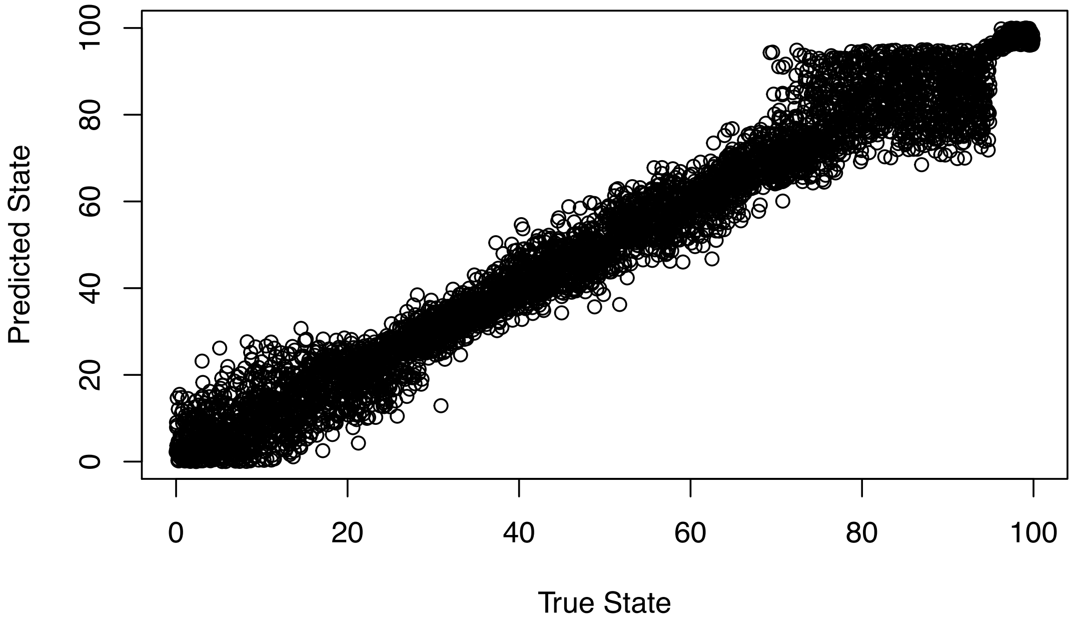

Figure 4.21: **Predicted vesus *true* cell cycle state.** Simulated data generated using the fitted *fucci* models were analysed using the same models. The resulting state (predicted) is plotted against the simulated (true), in order to observe how good the correlation is.

when comparing observations between different phases. In other words certain cell cycle phases are better resolved than others, depending on the combination of markers used. This can be seen in Figure 4.22, where the resolution at each stage in the cell cycle has been measured using the data shown in Figure 4.21. Specifically, a 95% confidence interval can been taken at each simulated cell cycle state, based on the PSM analysis of the simulated data. This can also be seen by taking the absolute distance of the 95% interval 4.23.

Repeating the same procedure for each reporter independently allows us to compare the contribution of each reporter to the cell cycle resolution of the combined model. In Figure 4.24, we can see that analysing the data using all available markers simultaneously (blue line) combines their strengths, thus improving the precision of the predicted cell cycle state. In Figure 4.25 we look at how the different pairwise combinations of cell cycle markers compare with each other. We see that the *fucci* markers alone (black) are the poorest combination when it comes to resolution, in part due to the unidentifiability between the G1 and M phases of the cell cycle, something that is not noted in the manuscript (Sakaue-Sawano et al., 2008). Adding DNA quantitation as an additional dimension resolves this issue due to the change in ploidy between these two phases, leading to a markedly improved resolution.

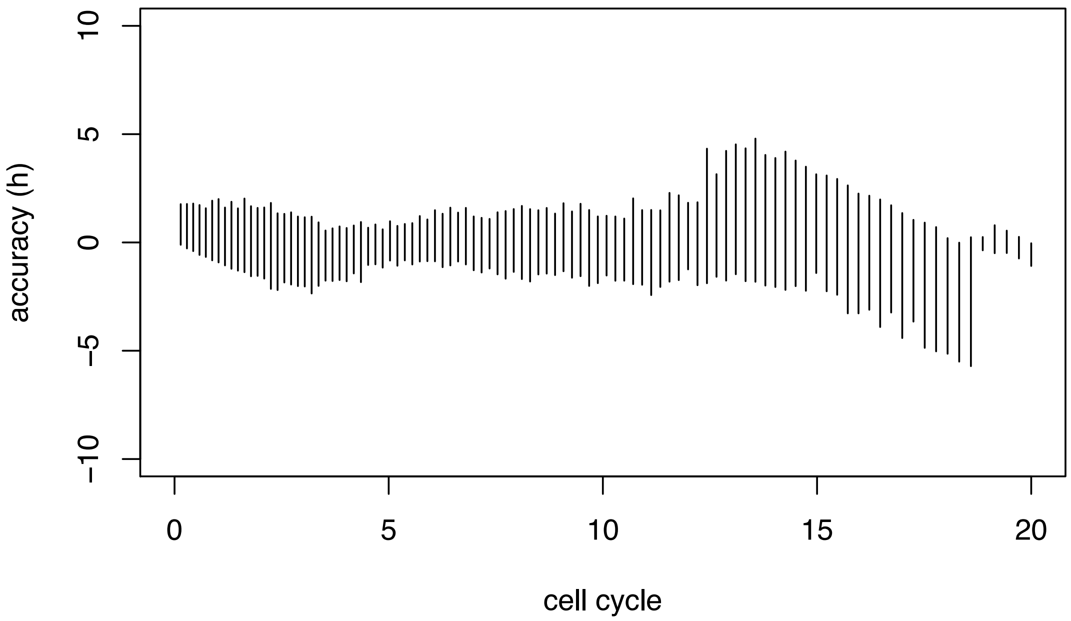

Figure 4.22: **Resolution at each cell cycle stage based on *fucci* markers and DNA.** Ranges obtained using a 95% confidence interval on the results obtained by analysing the simulated data using PSM.

4.4 Discussion

Bagwell et al. (2015b) suggest that transitions are fitted iteratively, in order to make the exploration of unknown markers by the researcher feasible and to prevent the explored state-space from exploding as a result of the increase in the number of fitted parameters. Although fitting well known progressions first can be insightful with respect to the rest of the transitions, the multiple false minima in the objective function often mean that fitting the first progression alone results in a suboptimal set of parameters which may prove disagreeable when trying to fit further measurement channels.

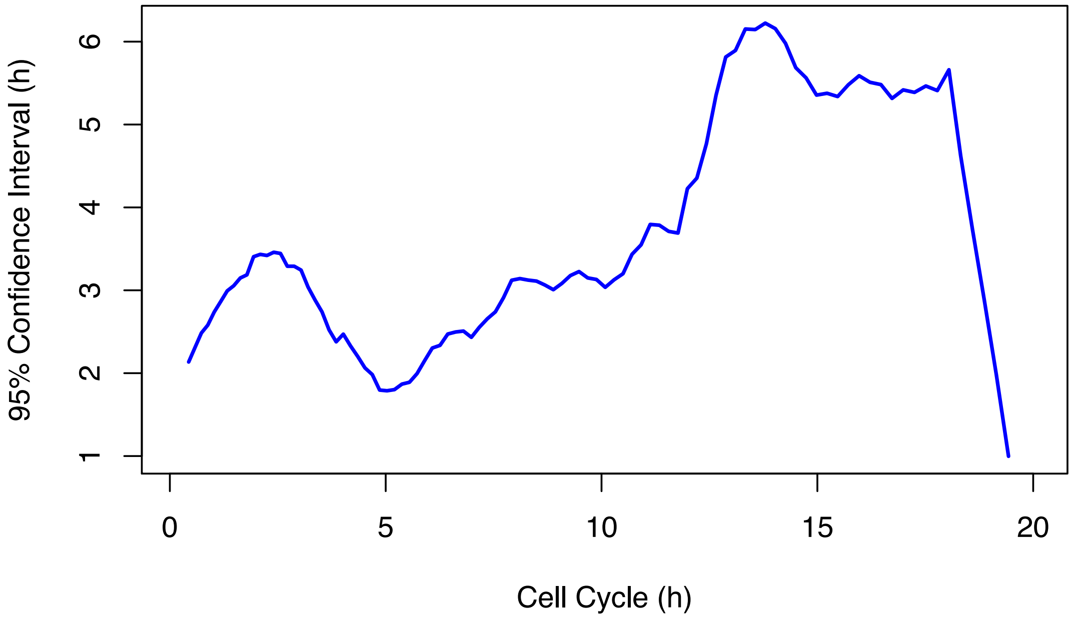

Figure 4.23: **Absolute interval at each cell cycle stage based on *fucci* markers and DNA.** Ranges obtained using a 95% confidence interval on the results obtained by analysing the simulated data using PSM.

Here, a consensus fit between the measurements is achieved by fitting as many progressions at a time as possible. Although fitting that many parameters simultaneously can be computationally demanding, modern global optimisation algorithms exist which can handle such demands. A machine learning-based optimiser based on Baeysian Optimisation (Bischl et al., 2017) is used here, which is specifically suited at handling expensive-to-evaluate, noisy objective functions.

This way descriptive models for three cell cycle markers, namely DNA, geminin and Cdt1, are fitted simultaneously. In the process, we saw that due to heterogeneity in the Cdt1 expression at the G1-S boundary, Cdt1 cannot provide information about this step in the cell cycle, thus trying to fit a continuous piecewise linear model results in a bad fit. In order to avoid this limitation a method is proposed for selectively ignoring ambiguous transitions. This resulted in an improved cell cycle trajectory, which we use in the next Chapter to explore how transcription and cell growth relate to one another.

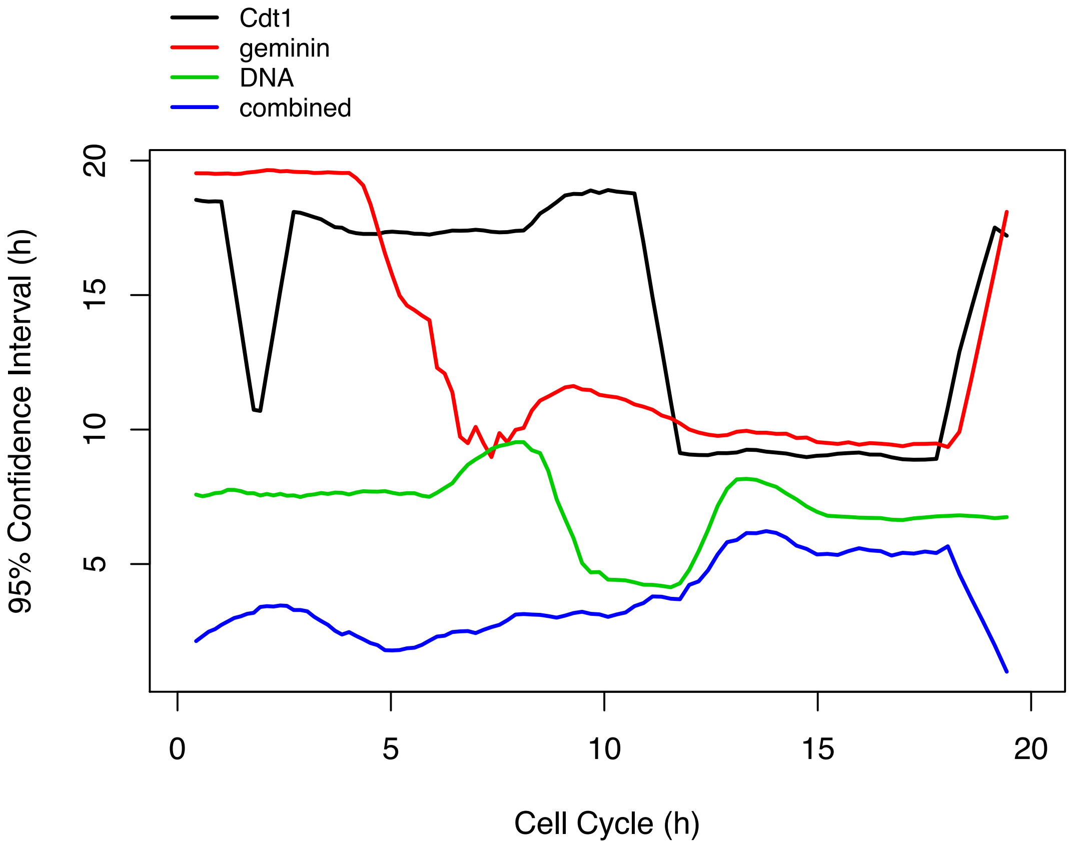

Figure 4.24: **Absolute intervals for each marker independently.** The same procedure described in main text was followed for each marker independently. This allows us to compare the resulting resolution conferred by each individual marker to the that obtained by the combined model.

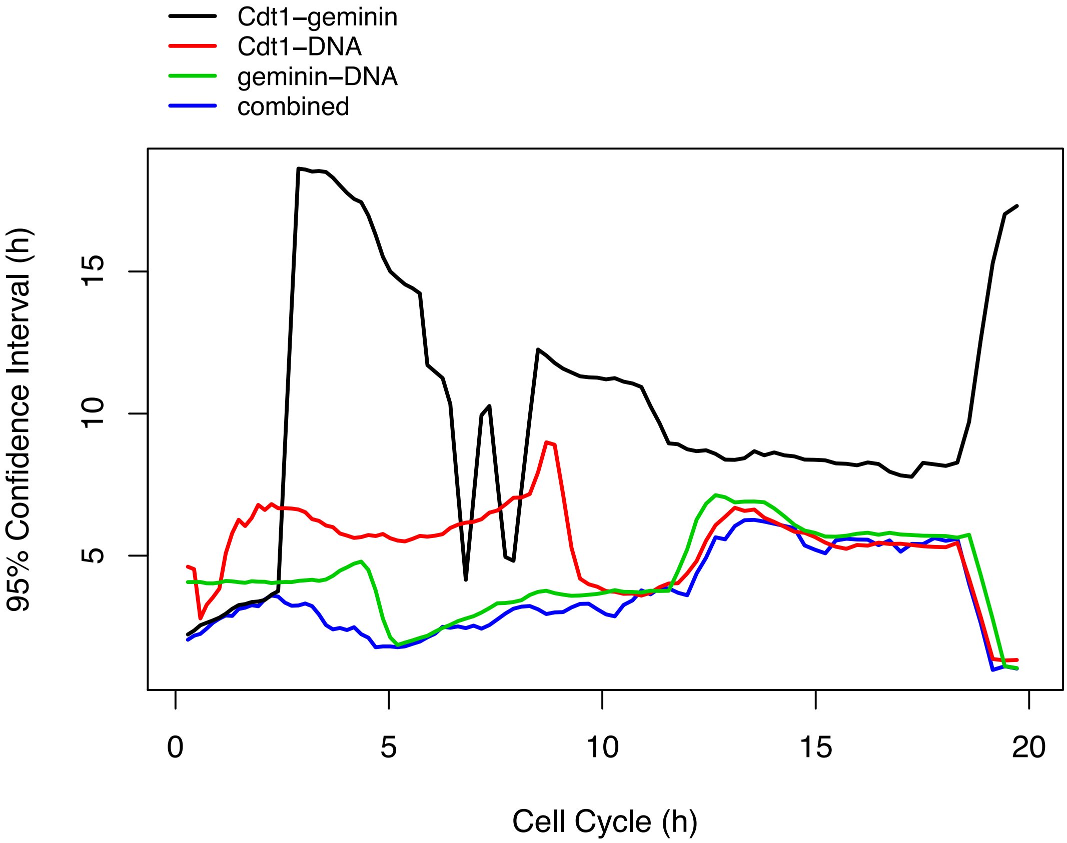

Figure 4.25: **Absolute intervals for each pair of cell cycle markers.** Resolution comparisson of all possible pairwise combinations of available markers.

The result is a phenomenological model consisting of the average cell cycle patterns of the *fucci* markers and measurements of DNA content. This model can be used to calculate the most likely cell cycle state of any *fucci* cell based on these three markers. When applied to a population of growing cells, it lets us study the effects of the cell cycle on various aspects of cell physiology, as will be seen in the next Chapter. It must be noted however that the resulting model is based on the average cell cycle, which means no observations can be made with regards to the heterogeneity of cell cycle phase lengths or period duration between cells of the same population. Furthermore, as the model is based on snaphsot data, no information about the correlations between the lengths of different phases is preserved. As it is known that cell cycle length heterogeneity is prevalent in growing populations of cells (Chiorino et al., 2001), future measurements need to be made to specifically account for these effects.

Chapter 5

Transcription kinetics and cell growth

Mammalian cell growth is a complex process encompassing genome replication, cell mass accumulation and drastic reorganization of the intracellular structure. Each of these processes contributes to gene expression noise, making the design of robust genetic circuits difficult (Huh and Paulsson, 2011a). As seen in Chapter 4, flow cytometry can be used to collect measurements on multiple cell cycle reporters, which can be analysed using Probability State Modelling (Bagwell et al., 2015b) to position each cell into its most likely cell cycle state. Combining this methodology with measurements of gene expression kinetics such as metabolic labelling of transcription and translation provides a high resolution view into how such activities change during the cell cycle.

Changes in cell size are another important source of gene expression variation. Furthermore, cells of different sizes are known to grow at different rates, further confounding our measurements of noise. Using ergodic rate analysis (ERA) (Kafri et al., 2013), we correlate our measurements of gene expression kinetics with those of cell size growth rate as a function of cell cycle progression. This way, we aim to elucidate the homeostatic mechanisms linking cell growth and global transcription, in order to better understand gene expression noise.

The relationship between cell size and RNA abundance in mammalian cells has been investigated in the past by comparing the RNA content of mouse cells from tissues with different characteristic cell sizes (Schmidt and Schibler, 1995). The authors found a strong correlation between cell size and RNA content, which was attributed to an increased rate of RNA synthesis. Although cells taken from different tissues were normalised for DNA quantity, it is likely that other tissue - specific mechanisms contribute to the observed changes in transcript abundance alongside cell volume. In order to understand the effects of cell volume specifically, such measurements would have to be repeated in cells within the same cell type. Using single-cell measurements, it is possible to exploit the intercellular variation within a population of cells for this purpose.

Such a strategy was employed more recently by Padovan-Merhar et al. (2015), who measured mRNA transcript abundance in mouse fibroblasts alongside cell size and cell cycle for a selection of 25 genes, using a combination of single molecule fluorescent *in situ* hybridisation (smFISH) and fluorescent microscopy. The cell cycle was resolved by counting the number of Cyclin A2 transcripts, which are known to accumulate from the start of S to M phase (Gookin et al., 2017). Using this method, the authors showed that the mRNA number of a cell strongly correlates with cell volume, and that this relationship is not affected by cell cycle progression.

Furthermore, using metabolic labeling, the authors conclude that mammalian cells adjust their RNA abundance to maintain a relatively constant concentration between different sizes, by increasing the rate of transcription rather than RNA stability. Here, we repeat this metabolic labelling experiment using a more advanced method of cell cycle analysis (PSM) which combines 3 distinct reporters, and a simpler though well established method for measuring cell size, using flow cytometry.

5.1 Metabolic labelling of transcription in *fucci* cells

As described in Chapter 4, PSM can be used to combine the measurements from multiple cell cycle reporters into a single cell cycle trajectory. This can subsequently be used to track how any other measurable trait changes with respect to the cell cycle. Flow cytometry is an excellent platform in this respect, as it allows multiple traits to be measured in parallel. We utilise this property to simultaneously measure three cell cycle reporters (DNA, Cdt1 and geminin) in *fucci* cells, alongside measurements of global transcriptional activity and cell size, in order to understand to what extent changes in cell size and cell cycle lead to changes in global transcription rates. Metabolic labelling and subsequent quantification was performed according to the method desrcribed in 2.3. Briefly, cells were treated with chemically labelled uridine (5EU), which can subsequently be conjugated to a fluorophore and measured by flow cytometry to quantify the incorporation rate.

The mean fluorescent signal detected in cells that had been administered 5EU was over ten times higher than the background staining, as shown in Figure 5.1. This demonstrates that incorporation of 5EU into the RNA of living *fucci* cells can be readily detected within 1 hour of labeling, as shown before in other cell types (Jao and Salic, 2008).

It is often informative to examine the shape of the resulting distributions. To do so, we look at the density plots of the 5EU measurements without log transforming. In Figure 5.2, we see that although the backgroud stain results in a narrow distribution, the 5EU treated sample has a broad distribution, with two broad peaks identifiable. The large breadth of the distribution suggests that the transcription rate varries substanially between cells within the population, while the two peaks suggest there are two subpopulations with different transcription rates. By controlling for cell cycle and cell size effects, we are able to determine whether these two populations are explained by differences in cell cycle phase, as seen in Figure 5.4.

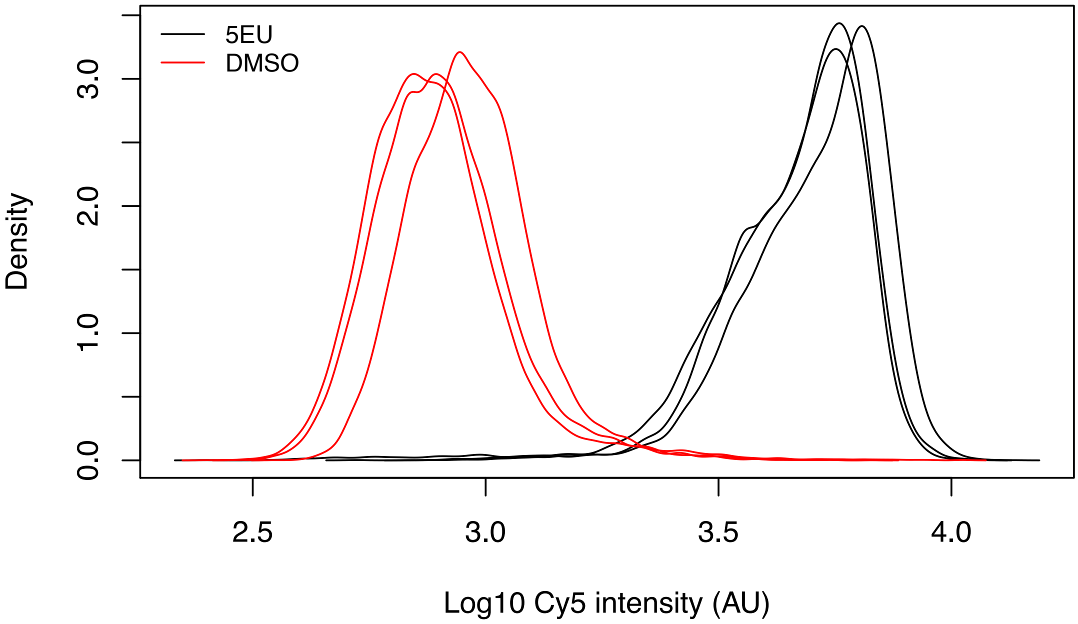

Figure 5.1: **5EU metabolic labelling.** Density plots show the Cy5 labelling intensity distribution of the flow cytometry measurements of roughly 15,000 *fucci* cells. Black lines correspond to samples to which 5EU was administered, red lines are the negative control and indicate the background staining of the Cy5 azide dye.

Cell cycle models were fitted for the three reporters as described in Chapter 4. Five independent runs of the fitting algorithm were performed in order to estimate the associated error. The resulting models (shown in Figure 5.3) were used to analyse the cell cycle trajectories of the 5EU labelled samples, as shown for one of the replicates in Figure 5.4. Although we can see that there is a cell cycle dependent effect on the background staining (in red), this can easily be corrected for, by subtracting the means at each point. The result of this operation is shown in Figure 5.5.

In Figure 5.4 we can see that the incorporation rate is several fold above background levels even at the very start of G1, suggesting that there is significant transcription during the beginning of the cell cycle. It is worth noting, however, that due to the time window of the 5EU pulse (1 hour), it is expected that a

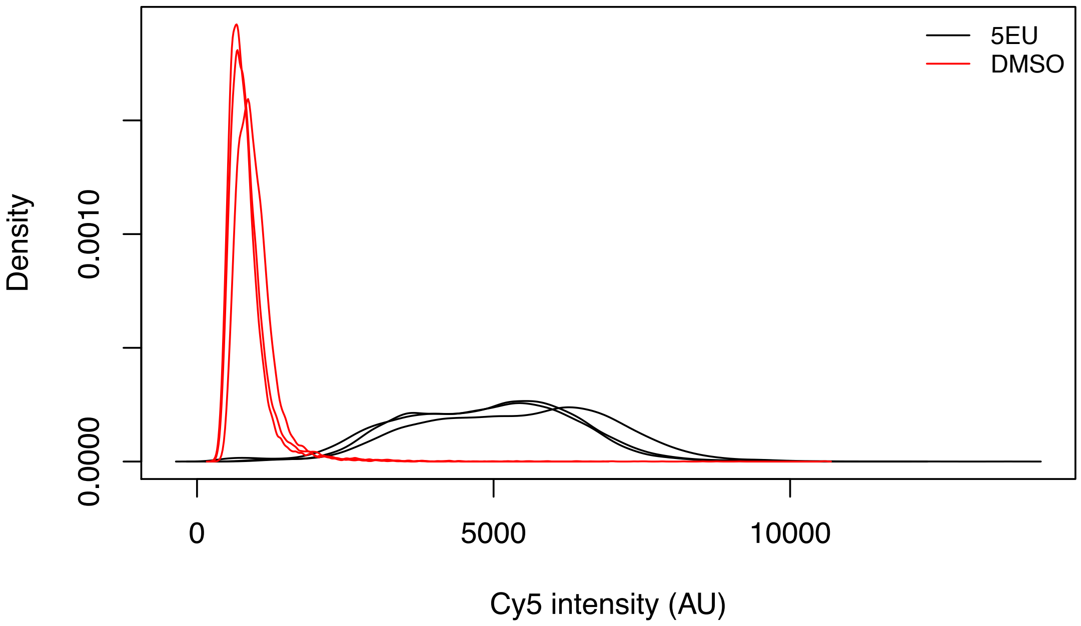

Figure 5.2: **5EU metabolic labelling, untransformed.**

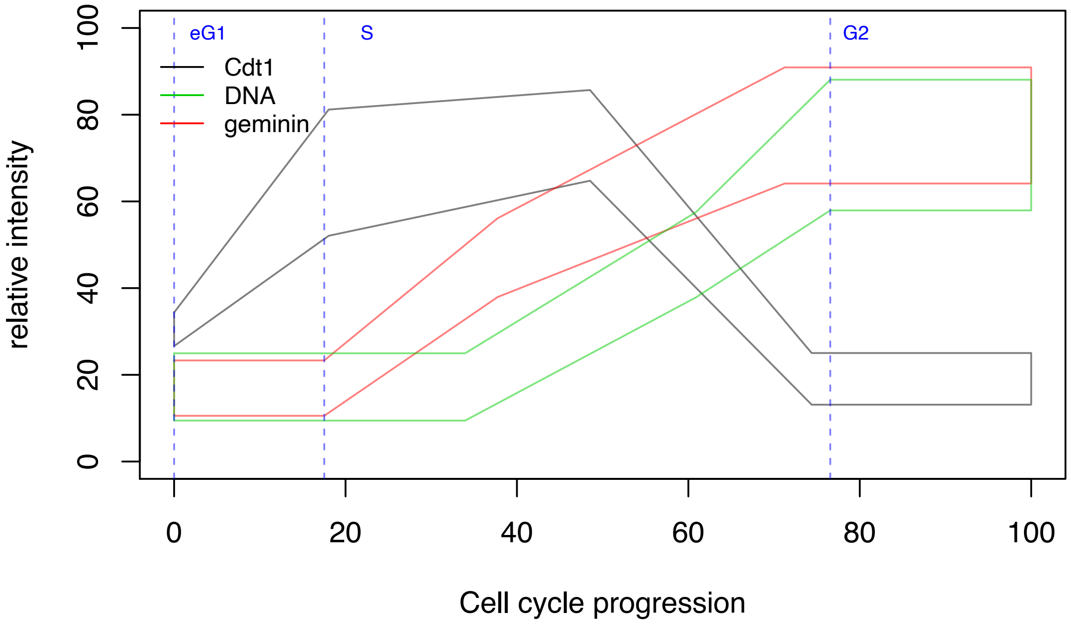

Figure 5.3: **Definition of cell cycle phases**. Models for the three cell cycle reporters fitted by PSM are sectioned into phases as shown by the dashed lines. Early G1 (eG1) is defined from the start of the G1 phase. S is defined as the onset of geminin upregulation (APC activity inhibition)

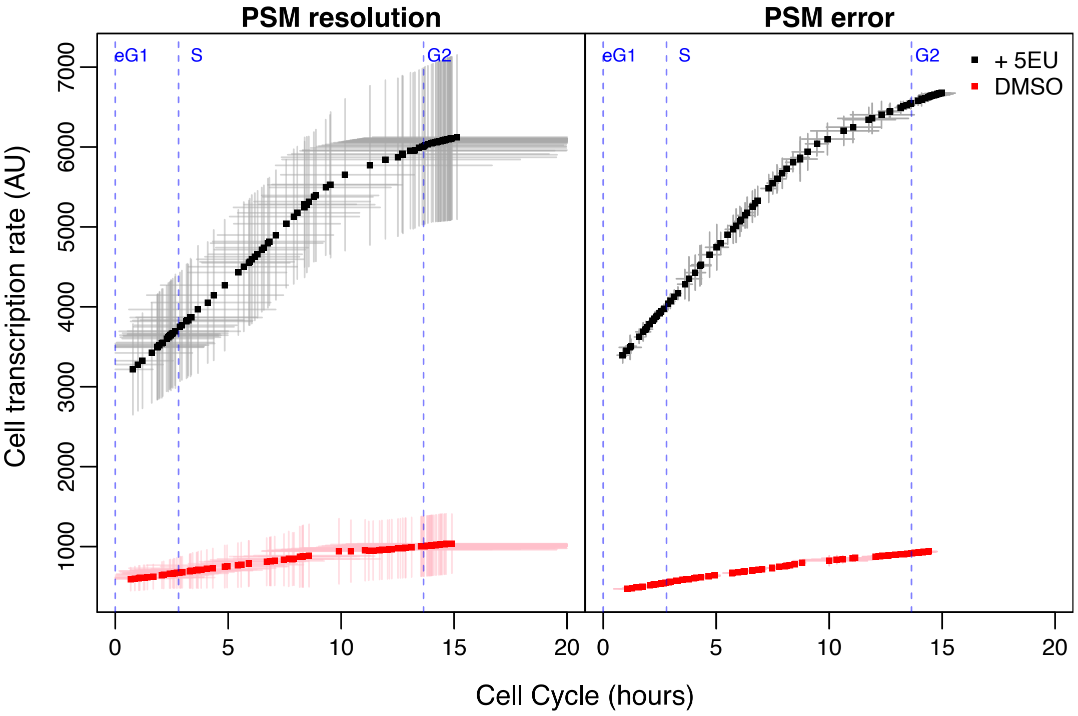

Figure 5.4: **Cell cycle analysis of 5EU incorporation.** Left: points show the mean 5EU incorporation level (black) and background staining (red) at each point in the cell cycle, horizontal lines show the resolution limit at each phase, vertical lines ± 1 standard deviation (SD) of 5EU intensity. Right: points show the means of 5 independent fits of the cell cycle model, horizontal and vertical lines show the respective uncertainty in cell cycle phase and mean 5EU intensity (± 1 SD).

fraction of the 5EU detected in nascent cells within the first hour of the G1 phase will have been incorporated during the M phase of the parent cell. This means that we cannot differentiate between the rates during the first and last hour of the cell cycle. Future experiments with shorter timepoints will enable us to look at these phases more closely.

In other terms, these results are in agreement with earlier RNA kinetics studies in Hela cells, performed in synchronised cultures (Pfeiffer and Tolmach, 1968). Specifically, we see an overall doubling of the global transcription rate during the cell cycle, with the rate heading towards a plateau during the G2, though in our case we cannot resolve the whole of the G2 phase. In contrast to Pfeiffer and Tolmach (1968), we do not observe a constant rate at the start of the cell cycle, though this could also be due to limitations of the cell cycle resolution. Careful optimisation of the flow cytometer lasers according to recent suggestions by Hazen et al. (2018) can lead to a higher resolution cell cycle, as seen in Figure 5.7 for cell size measurements.

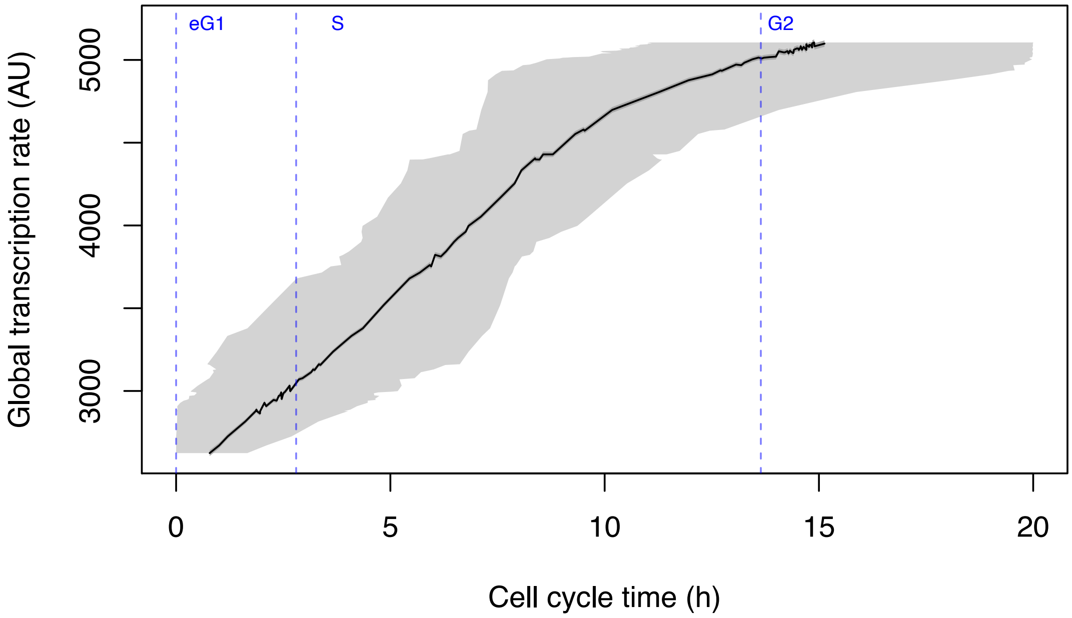

Figure 5.5: **Background subtracted mean 5EU trajectory.** Black line indicates the mean 5EU intensity. Light grey shade indicates the averaging window due to the resolution limit at each stage in the cell cycle.

5.2 Cell volume - cell cycle

Here, we look at the progression of the mean cell size as estimated by the cytometer using forward light scatter intensity (FSC.A), a commonly used measure of cell size which has been shown to correlate best with particle cross sectional area (Hawley and Hawley, 2018). Assuming that mammalian cells are approximately spherical when in suspension, we can convert the arbitrary units of cross sectional area to equivalent units of volume, according to the relation

$$V = \frac{4}{3}\frac{A^{\frac{3}{2}}}{\sqrt{\pi}},$$

where V is the volume and A the cross sectional area of the sphere.

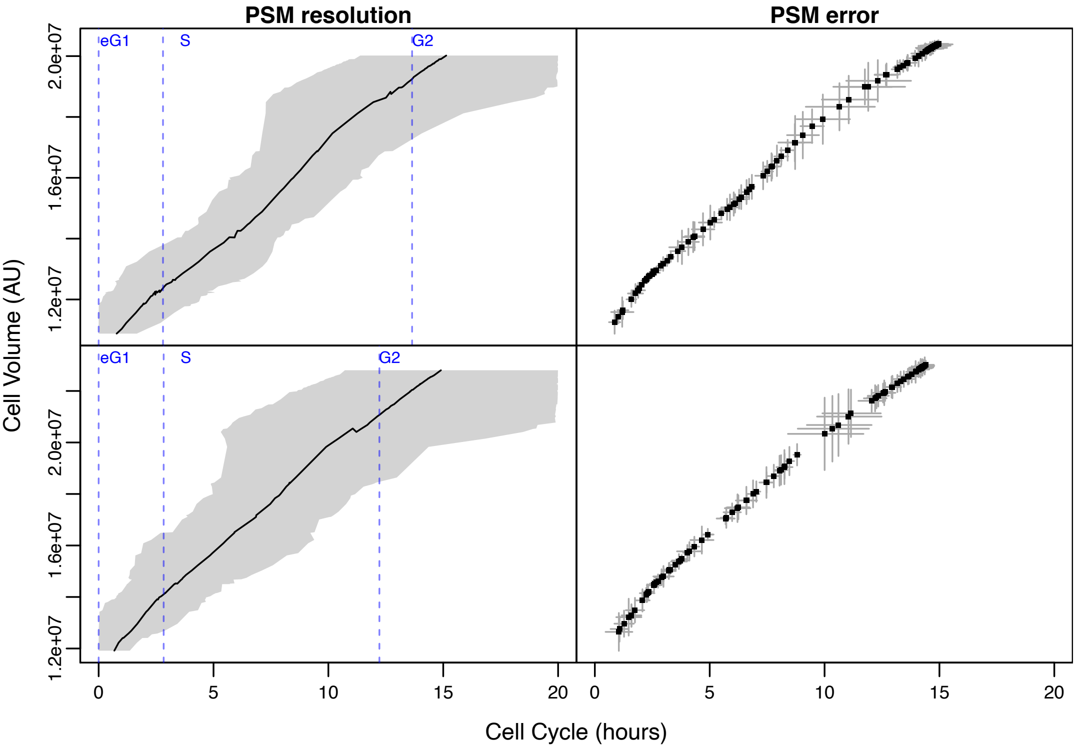

Figure 5.6: **Cell volume increase during the cell cycle.** Rows correspond to two biological replicates. Cell volume units obtained as described in the main text. Left column shows mean (black line) and cell cycle uncertainty (grey shade) from 1 run of cell cycle fitting algorithm. Right column shows mean and error of 5 independent fits.'

In Figure 5.6 we can see that on average, although cells continue to grow throughout most of the cell cycle, the rate of cell growth is not constant. Specifically, cell size appears to be increasing from early G1 to the start of the G2 phase, with at least one obvious decrease in growth rate, during the start of the S phase. This is much more obvious in Figure 5.7, where the suggestions from Hazen et al. (2018) have been implemented to increase resolution. This change in rate has been characterised previously by Kafri et al. (2013), who suggest it is part of a cell size homeostasis mechanism, as will be discussed in later sections.

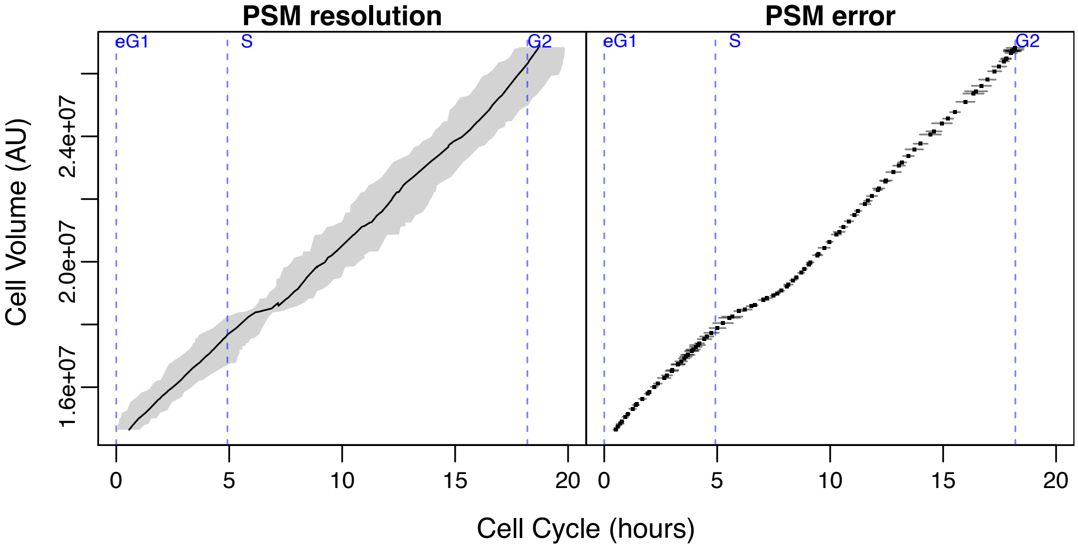

Figure 5.7: **Mean cell volume during the cell cycle.** 30,000 *fucci* cells were analysed by flow cytometry after DNA staining with Hoechst dye. Cell volume was obtained by converting FSC.A measurements of cross sectional area to volume. Cell cycle was analysed using PSM.

5.3 Transcription rate - Volume

Next, we look at how all three variables, namely transcription rate, cell size and cell cycle correlate with each other using a two dimensional heatmap. To that end, we bin cells according to their cell size and cell cycle phase, and colour code each two dimensional bin according to the mean 5EU intensity (Figure 5.8).

Interestingly, although the 5EU incorporation rate more than doubles across the

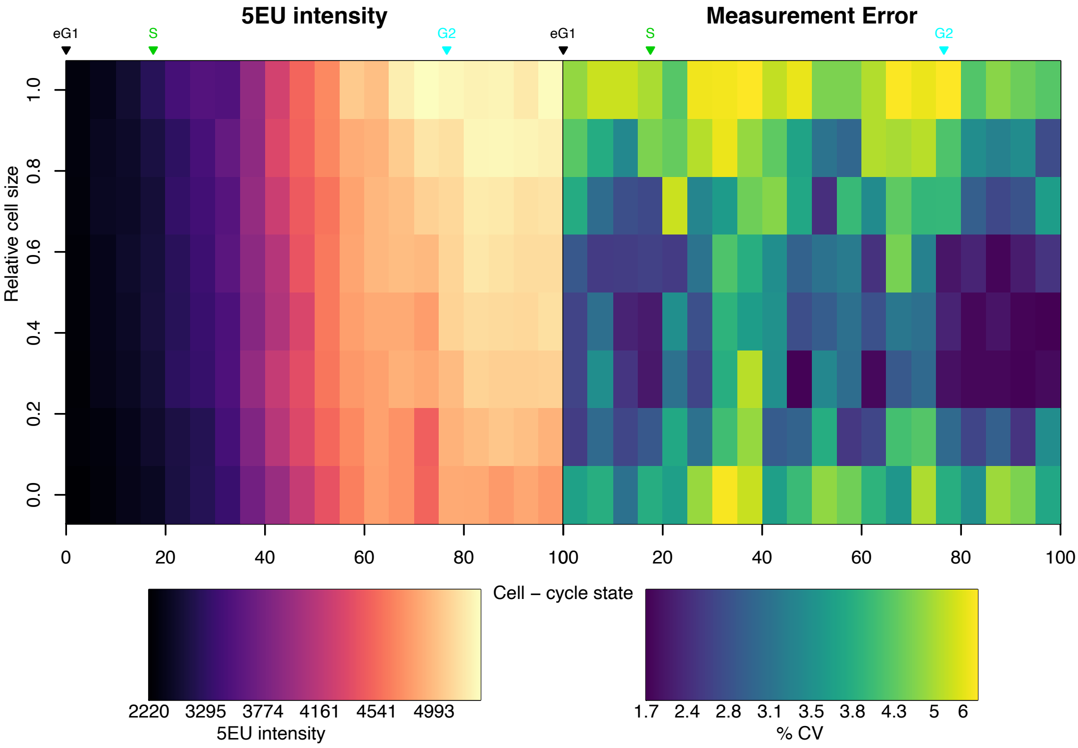

Figure 5.8: **Heatmap of global transcription rate accross cell size and cell cycle.** Global transcription rate measured using metabolic labelling. Cell cycle analysed using PSM. Volume measurements estimated by converting forward light scatter measurements. Bins are defined by sorting the cells first by cell cycle (20 bins), followed by secondary binning by cell volume (8 bins). Mean 5EU incorporation is obtained by averaging the mean within each bin.

cell cycle, it appears to only have a modest correlation with cell size within each phase. The heatmaps in the right hand panel are added to aid interpretation, by allowing us to evaluate the results in light of the error associated with the cell cycle analysis. The error is expressed in percentiles of coefficient of variation (CV), and is measured using the results from five independent PSM fits of the cell cycle (see Chapter 4), in combination with bootstrapping (100 resamplings). We see that the precision is worse for the extreme size groups. This can be attributed to the fact that in every stage of the cell cycle there are fewer cells with sizes near the extremes. This error can be reduced by up-scaling the protocol in order to analyse a larger sample of the population.

To look at the effects of cell size and cell cycle separately, we plot the transcription rate as a function of cell cycle for each size group separately, see Figure 5.9. Here, we see that indeed the cell size has a much smaller effect on transcription rate than the cell cycle has. The effect of cell size appears to be strongest towards the final stages of the cell cycle, though it is worth noting that this is the same region for which the cell cycle resolution is poorest. As a result, the effect of cell size cannot be distinguished from the effect of the cell cycle during this phase, and thus the stronger effect of cell size during the end of the cell cycle could be due to the difference in size between cells at different stages of the end of the cell cycle that cannot be resolved. We further plot the reciprocal view of transcription rate as a function of cell size for each cell cycle phase 5.10, where the same observation can be made.

To investigate the relationship between cell size and transcription rate further, we normalise the measured 5EU incorporation and cell size with respect to their lowest values in order to compare the respective relative changes. As we can see in Figure 5.11, there is a large span in relative cell sizes at any given phase, which is not accompanied by a similar span in transcription rates. Instead, transcription rate correlates more strongly with cell cycle than cell size.

Although this is contrary to what is described by Padovan-Merhar et al. (2015),

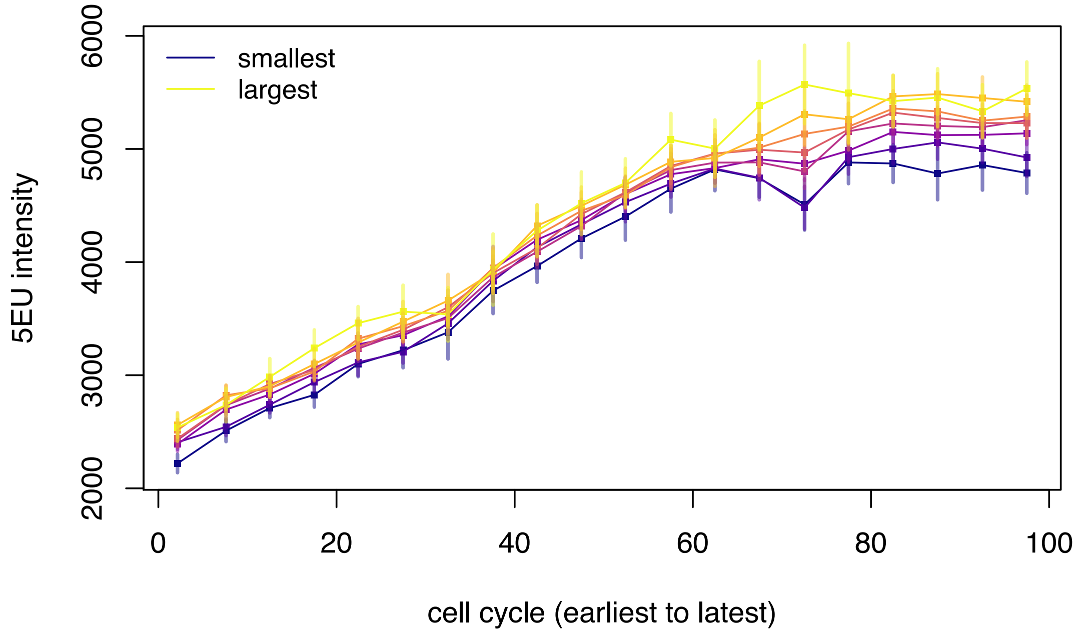

Figure 5.9: **Transcription rate as a function of cell cycle.** Colours reflect different size groups.

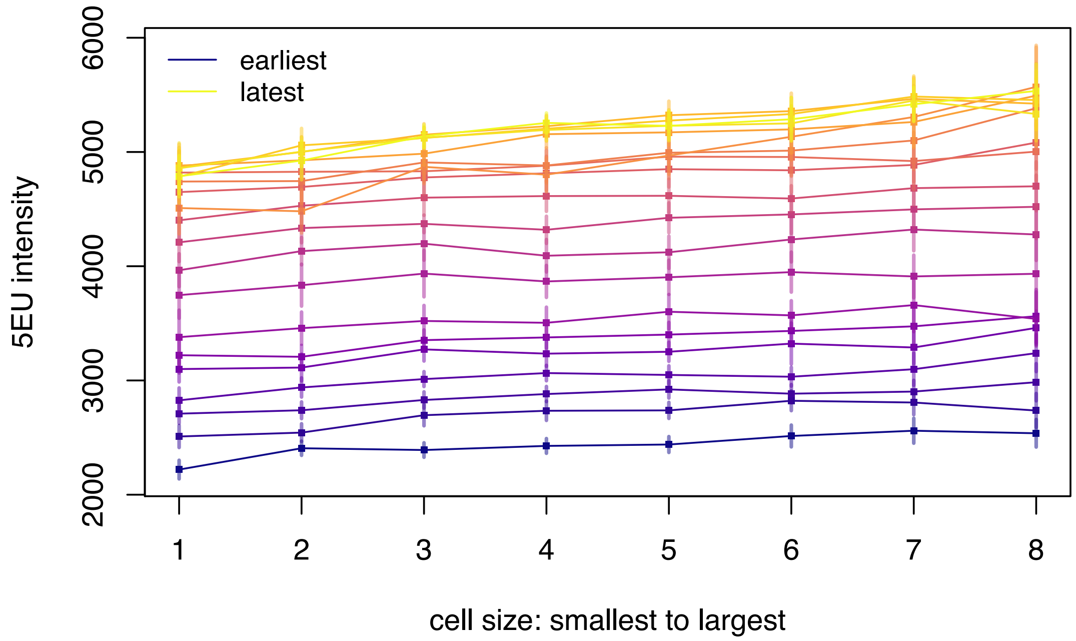

Figure 5.10: **Transcription rate as a function of cell size.** Colours reflect cell cycle phases.'

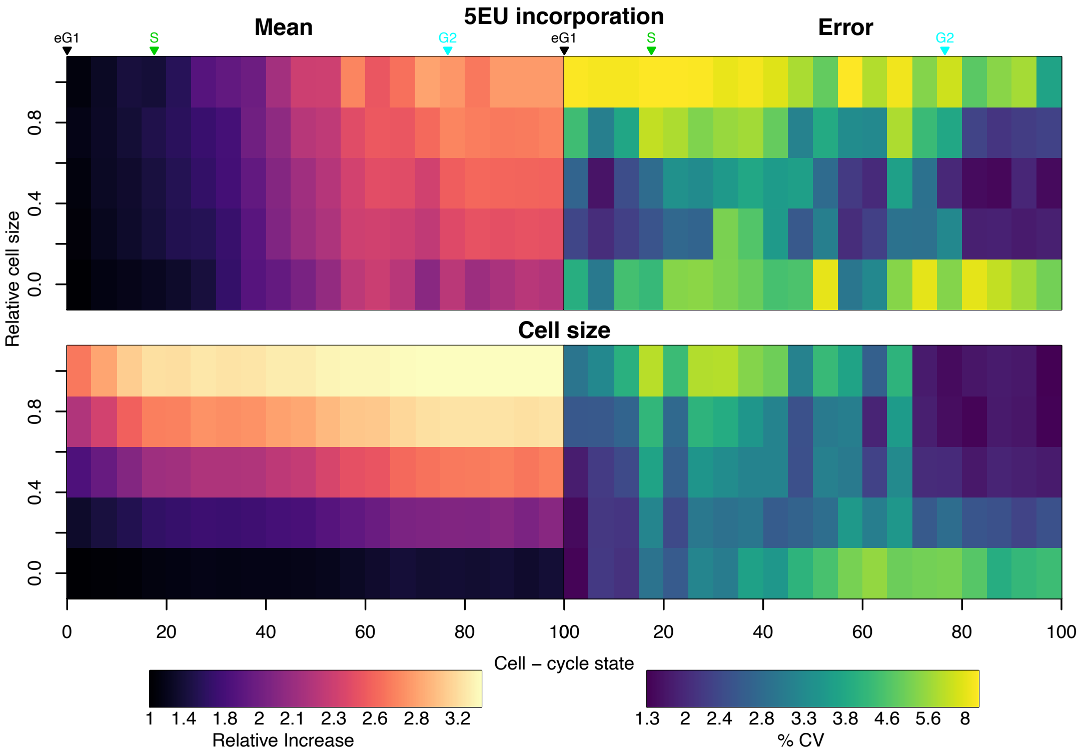

Figure 5.11: **Comparisson of relative changes in transcription rate and cell volume.** See caption in Figure 5.11 for details. Means are normalised with respect to the lowest measurements respectively.

upon closer inspection of the manuscript it appears that the authors did not control for cell cycle phase in their metabolic labeling experiment. This would naturally lead to cells from later phases in the cell cycle being over-represented in the larger group compared to cells from earlier stages in the cell cycle, and vice versa for smaller cells, thus allowing for changes in transcription rate due to cell cycle effects to skew the volume - transcription rate relationship.

On the other hand, Padovan-Merhar et al. (2015) do account for the cell cycle in their determination of RNA steady state using smFISH on a selection of 25 genes, where the correlation with size appears to be strong. In light of Figure 5.11, it appears that transcription rate alone does not sufficiently explain the size related changes in RNA abundance shown in (Padovan-Merhar et al., 2015). Although an adjustment of the decay rate was ruled out by the same authors, this conclusion was derived by chemical inhibition of transcription followed by time-measurements of RNA disappearance.

It has since been demonstrated in yeast (Das et al., 2017) that transcription and degradation of RNA are coupled by feedback mechanisms, with evidence suggesting similar mechanisms existing in mammals (Timmers and Tora, 2018). Therefore, chemical inhibition of transcription may have led to an underestimation of the role of RNA decay in preserving homeostasis. Furthermore, as the degradation of RNA is known to be carried out actively by specialised nuclease enzymes, which are themselves affected by dilution effects as cells grow, it is possible that a passive mechanism of transcript homeostasis may exist, regulated by the concentration of such enzymes. Another possibility is that a higher number of ribosomes protect the mRNA from being degraded (Chan et al., 2018).

The implications of a slower RNA turnover would include a slower progression through the cell cycle, as the relevant expression profiles for each phase would shift more slowly. Such an observation has been made previously, though has been attributed to alternative mechanisms, such as the reduced mitochondrial function seen in larger cells (Miettinen and Björklund, 2016), or DNA concentration becoming limiting (Neurohr et al., 2019), with cells exceeding a certain size becoming senescent. Furthermore, decay modulation has been implicated in RNA homeostasis before in yeast (García-Martínez et al., 2016). In light of the above, the possibility of changes in RNA stability being responsible for adjusting the abundance of RNA in cells of different sizes should be further investigated in the future.

In order to better understand the scale of this effect, we quantify the relative transcription rate per unit of cell size. In Figure 5.12 5.13, we see that smaller cells produce far more RNA per unit of cell volume than larger cells do (roughly 2 - 3 fold). This is consistent with previous results from measurements of steady state mRNA levels. Padovan-Merhar et al. (2015) noted that although transcript abundance scaled strongly with cell size in the genes investigated, there was a 1.2 to 3 fold higher concentration of RNA seen in smaller cells. This effect, which was highlighted more recently by Neurohr et al. (2019), raises two questions. First, why do larger cells have a lower concentration of RNA, and second, how do larger cells cope with this effect in terms of cell growth.

In Figure 5.12, we observe that the number of RNA molecules produced per unit of volume is up to 3 times lower in larger cells than in smaller cells, suggesting that RNA transcription is sufficient to explain this discrepancy. There are two hypotheses that can readily describe this effect. Either larger cells do not require a higher transcription rate to keep growing, or smaller cells are already transcribing at a rate near the biological limit, therefore rendering larger cells unable to increase their transcription rate, in spite of higher transcriptional demands.

As each RNA molecule can be translated hundreds of times by ribosomes (Pérez-Ortín et al., 2019b), it is possible that increased translation, conferred in part by increased RNA stability, could potentially be a mechanism for compensating for the lower concentration of RNA in larger cells. Such a regime would lead to a noisier expression of proteins, as stochastic fluctuations in RNA numbers would be amplified by the increased translation rate (Hausser et al.,

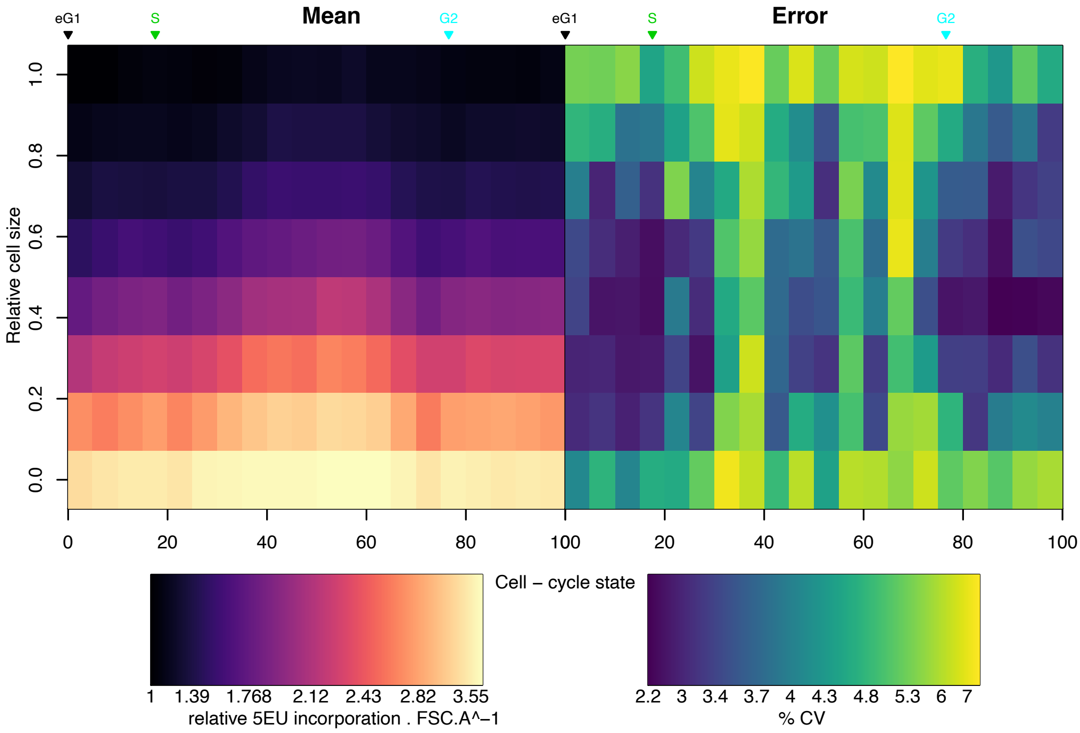

Figure 5.12: **Transcription rate relative to cell size.** Transcription rate estimated by 5EU metabolic labelling is normalised by cell volume, obtained by flow cytometry FSC.A measurements. Cell cycle analysed by PSM. Error measurements obtained using 5 independent fittings of the cell cycle parameters and bootstrapping (100 resamplings).

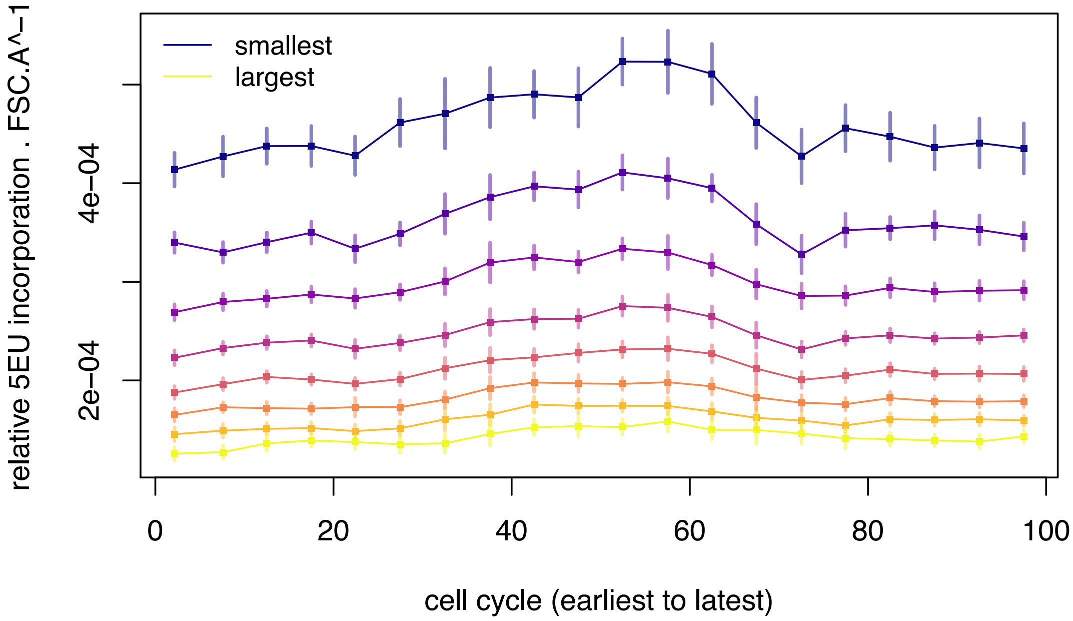

Figure 5.13: **Transcription rate relative to cell size (Lines).** Transcription rate estimated by 5EU metabolic labelling is normalised by cell volume, obtained by flow cytometry FSC.A measurements. Cell cycle analysed by PSM. Error measurements obtained using 5 independent fittings of the cell cycle parameters and bootstrapping (100 resamplings).

2019). Therefore it is unlikely that such a mechanism would be preferentially selected and more likely that it occurs by necessity, which suggests that transcription being limiting is more probable. In a way, this is reasonable to expect, as the amount of DNA template is constant between cells of different sizes, and therefore much more likely to pose a bottleneck during gene expression than RNA, which is amenable to amplification.

To test whether translation is in fact increased in larger cells, where transcription may be limiting, we could use metabolic labelling of translation. Specifically, a chemically labelled amino acid, which can subsequently be detected, could be supplied in the growth media in order to measure the relative translation rate, followed by the same analysis used here for the transcription rate. In order to test whether transcription is indeed a limiting factor for cell growth in larger cells,

Furthermore, the global rate of translation itself depends directly on the concentration of tRNAs, rRNAs, and ribosome encoding mRNAs. Therefore, the upregulation of translation in larger cells is contingent on an increase in the transcription rate of these RNA species, which in turn is limited by the available DNA template. Thus, increased translation can only compensate in a limited way for the continuously increasing demand of transcription during cell growth. Taken together, these results suggest that DNA becoming limiting in larger cells, as seen recently in (Neurohr et al., 2019), could explain the reduced relative amount of transcription in larger cells. In Section 5.4, we look at how this effect impacts cell growth.

5.4 Ergodic Rate Analysis

So far we have seen that even though the global transcription rate scales strongly with progression in the cell cycle, there is only a modest relationship with cell size within any given cell phase. Furthermore, we hypothesised that this may be due to DNA becoming limiting in larger cells, as has been suggested recently by (Neurohr et al., 2019). Here, we ask what downstream effects this has on cell growth. To do so, we employ ergodic rate analysis (ERA) (Kafri et al., 2013), a method which enables us to compare the growth rate between cells of different sizes using a single snapshot of a growing population (see Equation (2.1)).

We first use ERA to look at how the growth rate varies during the cell cycle for cells of average size. We do this by solving Equation (2.1) along the cell cycle timeline obtained by PSM in Chapter 4. In Figure 5.14, we see that the growth rate of the average cell drops during the G1/S phase, as shown in (Kafri et al., 2013). Interestingly, we further identify that the highest rate in the cell cycle occurs during early G1, which was not seen in (Kafri et al., 2013). This discrepancy can be attributed to the increased resolution provided by the additional cell cycle reporter, Cdt1, which enables us to detect changes within the G1 phase with greater detail. It is important to note that the resolution is too low during the G2 phase to accurately measure the growth rate, so the decreased growth rate observed in both replicates during that phase cannot be unambiguously attributed to a biological effect.

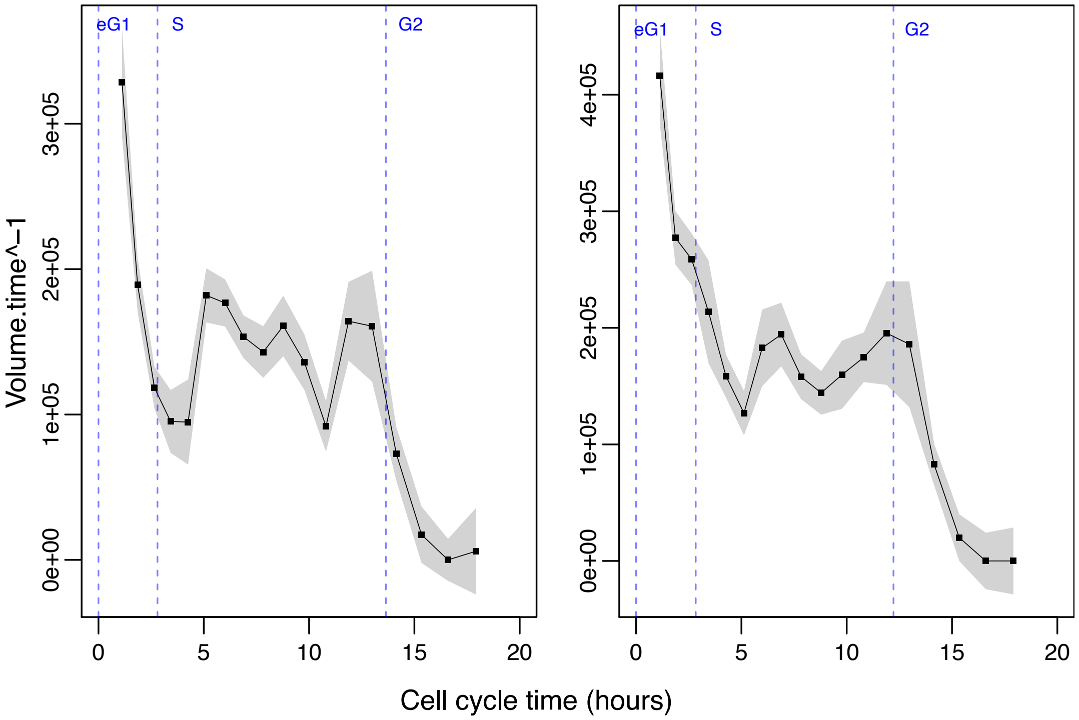

Figure 5.14: **Mean growth rate calculated by ERA.** Growth rate is calculated using ERA for the mean cell size along the cell cycle predicted by PSM. The two figures represent two biological replicates independently analysed.

Following from the discussion in Section 5.3, here we use ERA to look at how the growth rate varies between cells of different sizes at each stage in the cell cycle. Figures 5.15 and 5.16 show that in terms of absolute growth rate (volume added per unit time), there are both cell cycle and cell size related effects that can readily be observed. The expected decrease of growth rate at the start of the S phase (Kafri et al., 2013) can be seen in all size groups. Cells of all sizes appear to grow fastest during the start of the cell cycle, with the unexpected exception of large cells, which show a growth burst maximum at the late S phase. Furthermore, growth appears to slow down for all cells prior to the G2 phase, consistent with the presence of an S/G2 growth checkpoint, recently observed in Hela cells by real-time volume tracking (Cadart et al., 2018).

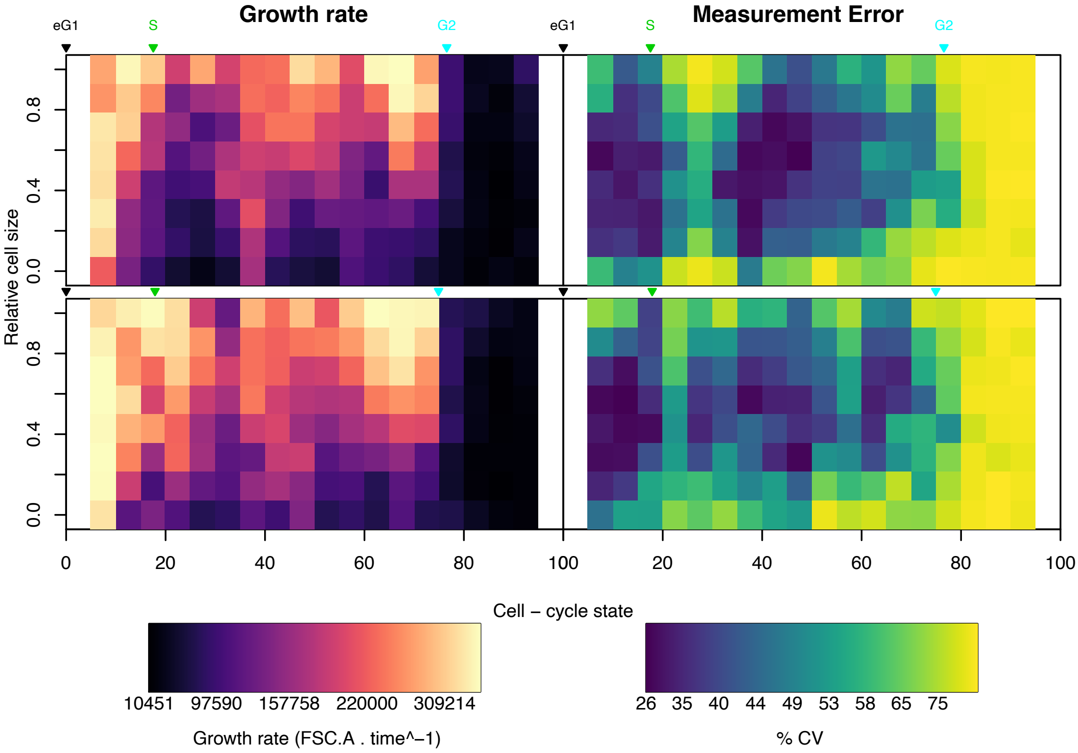

Figure 5.15: **Growth rate for cells of different sizes.** The two rows represents independent biological replicate. Growth rates calculated for each size group using ERA, as described in the main text. The cell cycle analysed using PSM. The left column corresponds to growth rate, the right column to associated error, obtained using bootstrapping (100 resamplings).

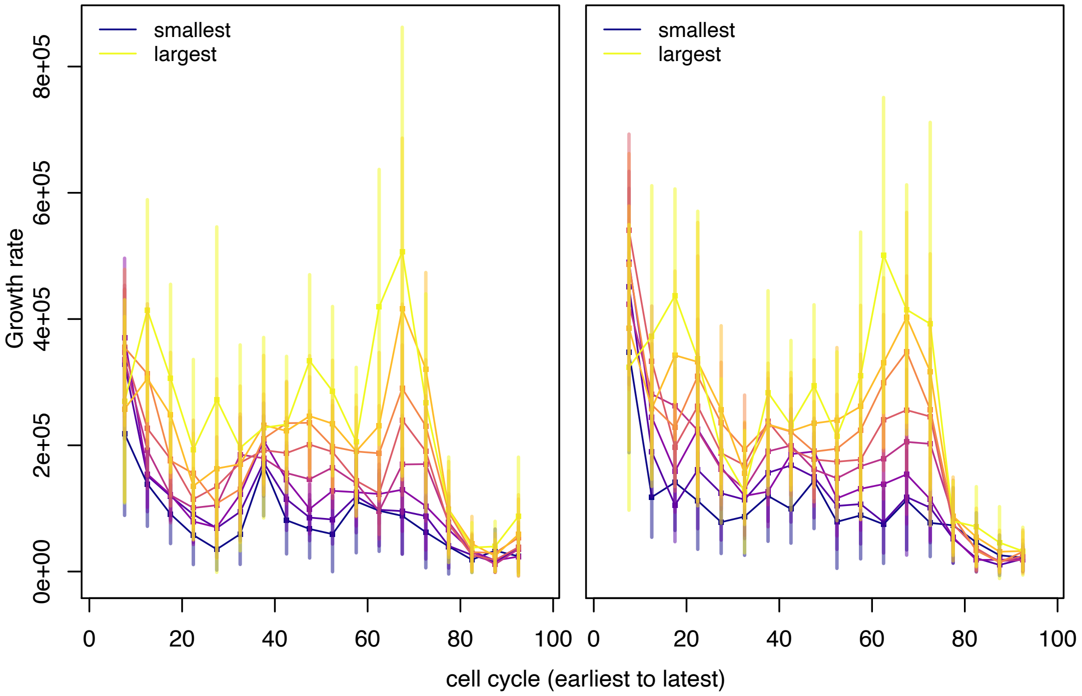

Figure 5.16: **Growth rate for cells of different sizes (Lines).** The two panels represents independent biological replicate. Growth rates calculated for each size group using ERA, as described in the main text. The cell cycle analysed using PSM. The left column corresponds to growth rate, the right column to associated error, obtained using bootstrapping (100 resamplings).

Although comparing absolute growth rate between samples of different sizes is illuminating, a more biologically relevant measure is the relative growth rate, in other words the percentage by which cells increase in size during a given phase of the cell cycle, as this reveals the presence of homeostatic mechanisms controlling the growth rate better than the absolute growth rate. In Figure 5.17, we investigate these effects.

Specifically, we see that all sizes of cells have the highest percentage increase during the early to middle G1 phase, though we can now identify a size dependent component, with smaller cells growing relatively faster than larger cells during that phase. We further see that smaller cells show an additional burst in growth rate during the S phase (approximately 40^{th} percentile), with a roughly 50% faster relative growth rate compared to cells in neighbouring cell phases or larger sizes. This is consistent with the existence of a previously described homeostatic mechanism at the G1/S transition, which ensures that cells are over a certain size threshold when progressing into the S phase by modulating the length of the G1 phase (Liu et al., 2018).

Strikingly, even when controlling for cell size, we find a population of large cells growing roughly 30% faster than other cells during the late S phase. We must highlight at this point that the above analysis cannot differentiate between changes in growth rate and changes in cell phase duration, as it does not incorporate distinct cell cycle length measurements for the different size groups. Therefore, we cannot tell whether cells are growing faster during a given phase, or whether instead these cells are spending a larger amount of time in that phase, as seen in (Liu et al., 2018) for smaller cells during G1. In this sense, the cell cycle trajectories of all cell size groups have been *in silico* synchronised to the average trajectory, which, although useful for comparison purposes, should be interpreted accordingly.

To disambiguate between the two interpretations, the growth rate could be measured more directly using metabolic labelling of translation, using a labelled

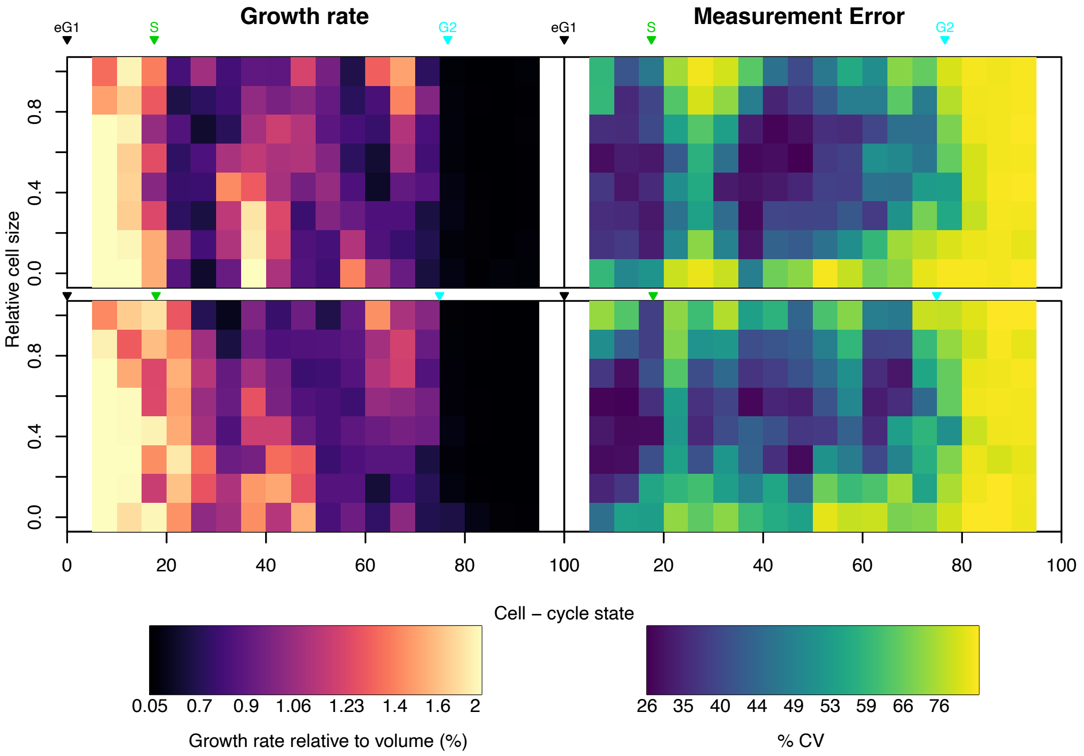

Figure 5.17: **Relative growth rate for cells of different sizes.** The two rows represent biological replicates. Relative growth rate is obtained by normalising the growth rate, obtained by ERA, by the mean cell size within each bin. The growth rate is thus expressed as the fraction of growth with respect to cell size. The error was estimate using bootstrapping (100 resamplings).

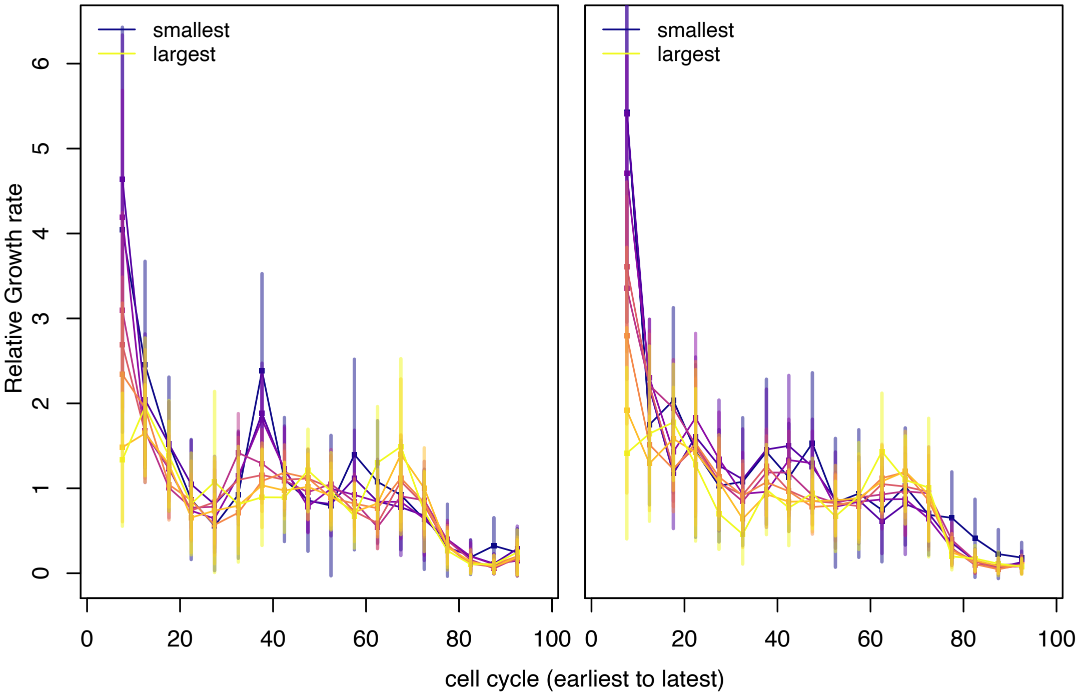

Figure 5.18: **Relative growth rate for cells of different sizes (Lines).** The two panels represent biological replicates. Relative growth rate is obtained by normalising the growth rate, obtained by ERA, by the mean cell size within each bin. The growth rate is thus expressed as the fraction of growth with respect to cell size. The error was estimate using bootstrapping (100 resamplings).

amino acide, as mentioned in Section 5.3. Thus, if larger cells are indeed growing faster during the late S phase, we would detect an increased incorporation of labelled amino acids, compared to smaller cells.

Nevertheless, the burst of growth rate during the late S phase (or the reciprocal extension of said phase's duration) seen in large cells has not been described before, and is surprising in the context of cell size homeostasis. In order to establish whether this effect is not an artifact of the 5EU labelling protocol, we repeat these measurements on a live cell population. Figures 5.19 and 5.20 show that although the cell cycle distribution is quite different when compared to Figure 5.17, which is likely due to variation in the culturing routine, the observations made before are still apparent. Again, we find that growth is highest in the early G1 phase for cells of all sizes, with smaller cells growing relatively faster. We see the same increased growth in the early S phase in small cells though the effect here is more modest (~ 20% increase instead of 50%). Notably, larger cells show roughly 30% more growth in the late S phase than all other groups, similar to Figure 5.17, suggesting that the effect is real, constituting an 'anti-checkpoint' in cell growth.

5.5 Volume added per transcript synthesised

In order to better understand the relationship between growth rate and transcription rate, in Figures 5.21 and 5.22 we look at how much volume is added per RNA synthesised in cells of different sizes, as a function of cell cycle time. On first inspection, the amount of volume added per RNA synthesised varies greatly across different cell sizes and throughout the cell cycle. Specifically, larger cells produce much more volume per synthesised RNA molecule than smaller cells throughout the whole cell cycle (between 2 and 4 fold), which suggests that larger cells have a higher translation capacity, potentially due to a combination of greater RNA stability and a higher number or activity of ribosomes.

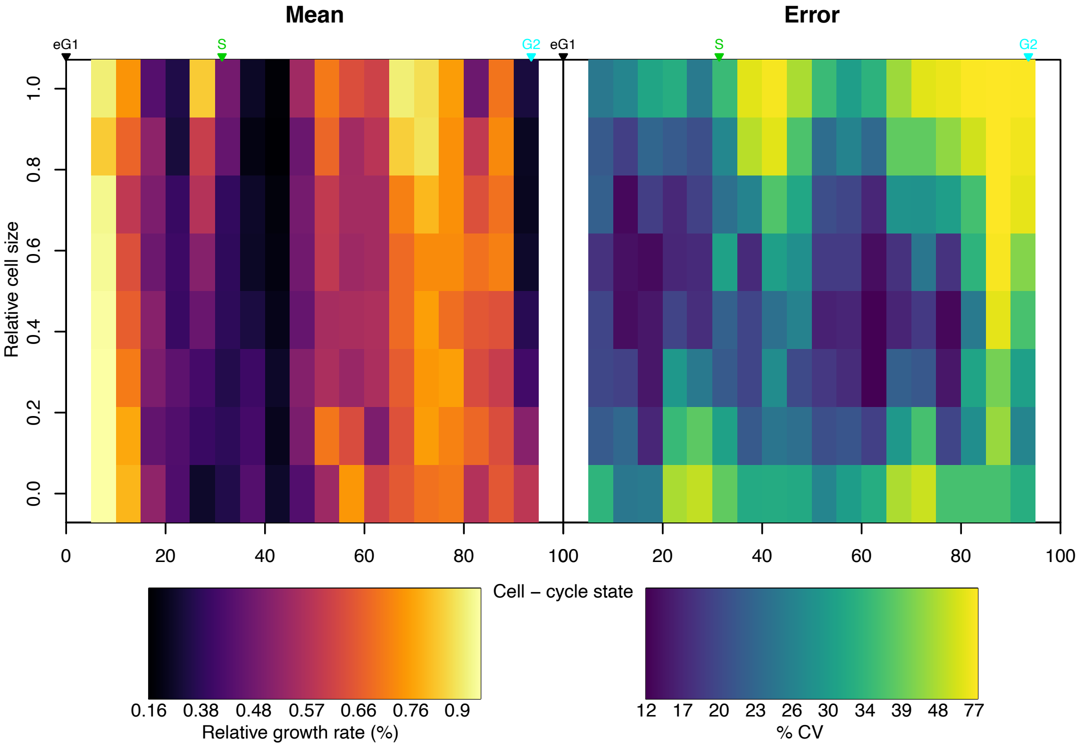

Figure 5.19: **Relative growth rate in live cells.** 30,000 live cells were analysed at high resolution by flow cytometry. The cell cycle was analysed using PSM. The growth rate obtained by ERA. The error was estimated by bootstrapping (100 resamplings).

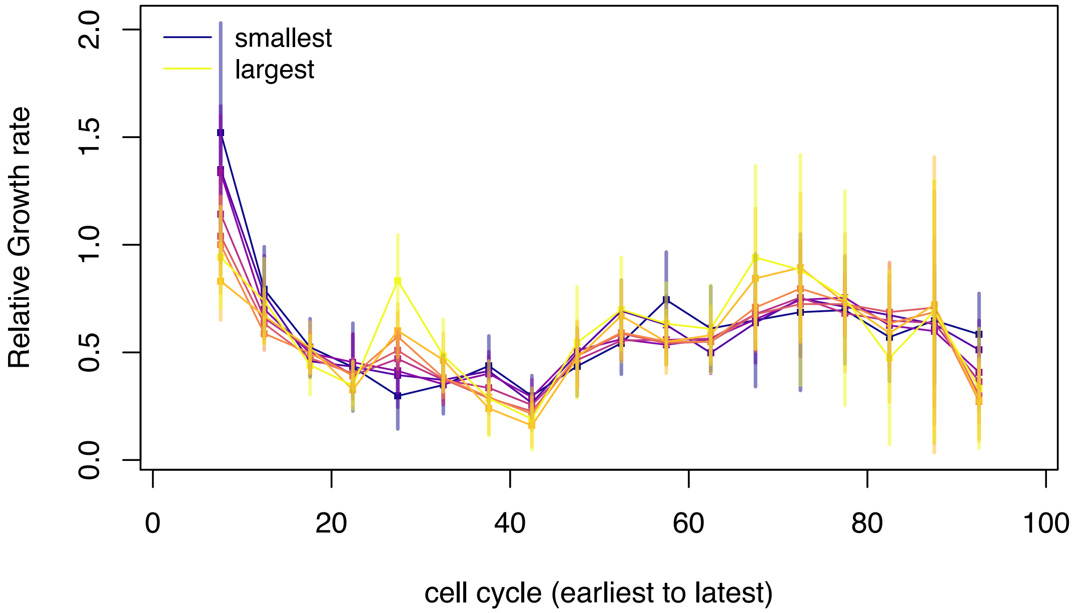

Figure 5.20: **Relative growth rate in live cells (Lines).** 30,000 live cells were analysed at high resolution by flow cytometry. The cell cycle was analysed using PSM. The growth rate obtained by ERA. The error was estimated by bootstrapping (100 resamplings).

Furthermore, all size groups produce the most cell volume per RNA molecule during G1. This effect diminishes as cells progress towards the S phase, but in a size dependent manner, with larger cells continuing to produce more volume per synthesised RNA for longer into the S phase. During the S phase, the translation capacity of all cell sizes drops briefly at the G1/S transition, and then rapidly rises in all cells, again in a size dependent manner. Interestingly, we find that the late S phase growth burst in large cells observed in Figures 5.19 and 5.17 is not explained by a corresponding rise in global transcription. It must be noted, however, that transcription and cell growth are not necessarily strictly synchronised. Specfically, cell growth in the sense of accumulation of cell mass is primarily determined by translation (Pérez-Ortín et al., 2019b), which is temporaly distinct from transcription, especially in nucleated cells such as eukaryotes. For this reason a more in depth analysis is required to better understand the relation between changes in transcription and growth rate.

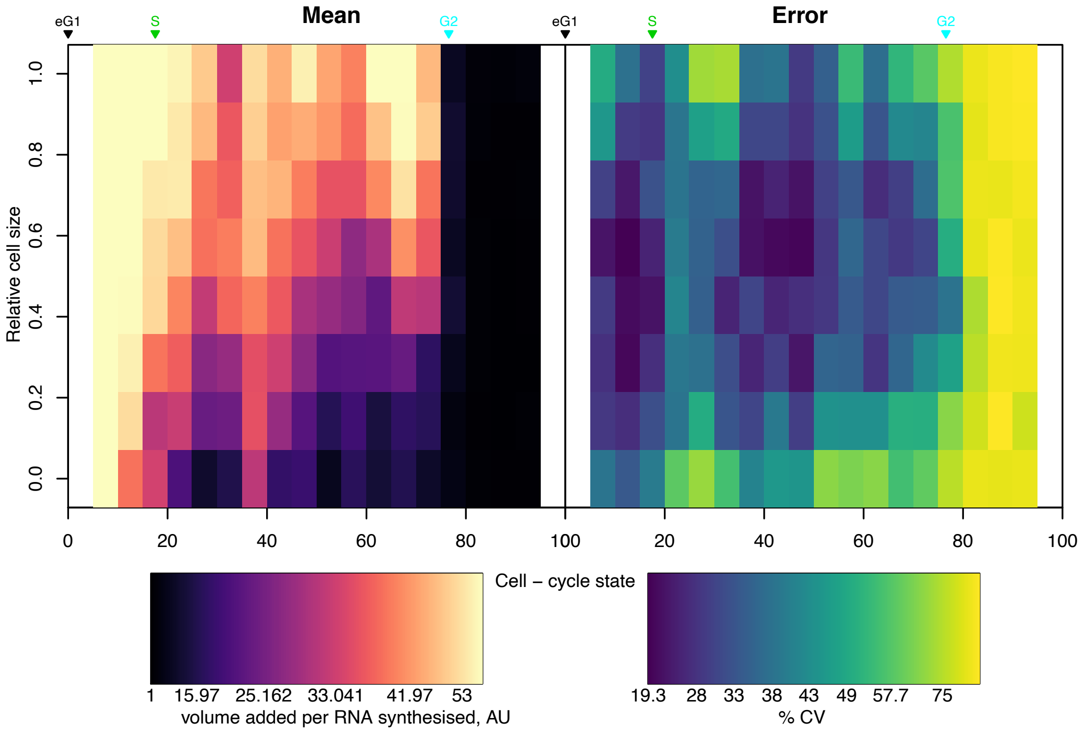

Figure 5.21: **Volume added per transcript synthesised.** Volume per transcript obtained by dividing the growth rate, obtained by ERA, by the rate of 5EU incorporation, measured by flow cytometry. Error obtained by bootsrtapping (100 resamplings).

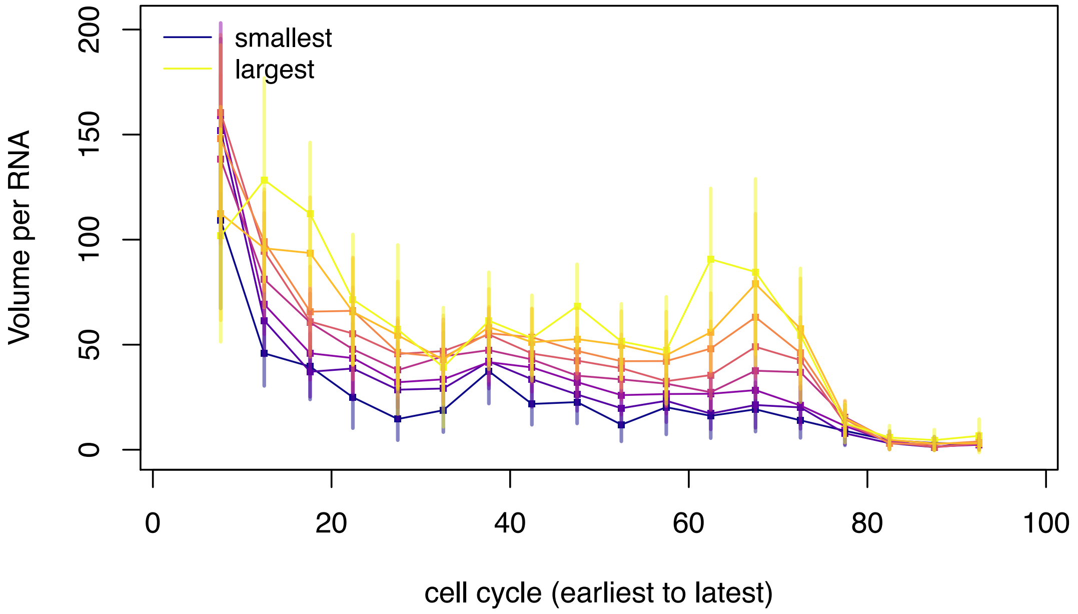

Figure 5.22: **Volume added per transcript synthesised.** Volume per transcript obtained by dividing the growth rate, obtained by ERA, by the rate of 5EU incorporation, measured by flow cytometry. Error obtained by bootsrtapping (100 resamplings).

5.6 Discussion

We saw that the global transcription rate in Hela cells does not correlate strongly with cell size, contrary to what has been suggested previously in mouse fibroblasts (Padovan-Merhar et al., 2015). Upon closer inspection of the results in (Padovan-Merhar et al., 2015), we found that cell cycle had not been controlled for in their metabolic labelling experiment, which could explain this discrepancy. Recently, it has been shown that DNA can become limiting as cells exceed a certain size, with profound effects on cell growth rate (Neurohr et al., 2019). Here, I suggest that DNA becoming limiting in larger cells can explain the fact that RNA transcription does not correlate strongly with cell size. This raises the question of whether larger mammalian cells maintain RNA concentration homeostasis by regulating the stability of transcripts, as seen previously in yeast (García-Martínez et al., 2016). Although this suggestion has previously been ruled out by Padovan-Merhar et al. (2015), their use of a transcriptional inhibitor for measuring the decay rate is likely to have confounded their results due to the existence of transcription-decay coupling mechanisms recently suggested to exist in mammals (Timmers and Tora, 2018). Careful kinetic experiments need to be performed to shed light in this direction (Chan et al., 2018).

To investigate the effect limiting transcription has on the growth of cells, we revisited a study on the growth rate of cells of different sizes, which identified a size checkpoint at the G1/S boundary (Kafri et al., 2013). More recently, the p38 MAPK pathway has been linked to regulating the transition of smaller cells from G1 to S, by extending the duration of G1, thus allowing for further growth (Liu et al., 2018). Interestingly, no mechanism has so far been proposed to explain the decreased growth rate seen in larger cells during this transition.

In light of the above results, I propose that decreased RNA expression due to DNA becoming limiting can explain this phenomenon. Specifically, if we assume that DNA can become limiting as cells accumulate volume during their progression in G1, we would expect the effect to be stronger in larger cells, as the transcriptional requirements for maintaining a constant concentration of RNA would be higher in these cells. This could explain the size dependent curbing of cell growth rate during the G1/S transition. As the S phase progresses, the DNA template of relevant genes becomes doubled, thus enabling larger cells to resume a higher growth rate. DNA limiting transcription during the cell cycle has been proposed in the past by Pfeiffer and Tolmach (1968), who noted that inhibiting DNA replication at different stages of the S phase led to a proportional decrease in transcription rate.

Upon repeating the growth rate analysis method developed by Kafri et al. (2013), we identify the characteristic G1/S checkpoint described before, as well as a surprising ‘anti-checkpoint’, whereby large cells experience a burst of growth during the late S phase. This could be another sign of DNA becoming limiting in larger cells, the effect of which we would expect would become diminished once certain late-replicating genes were duplicated. In support of this suggestion, it is worth noting that the rate of DNA replication is not constant, with the bulk of DNA becoming replicated in the later S phase (Li et al., 2014).

The fact that this phenomenon was not observed by Kafri et al. (2013) could be attributed to their cell size measurements reflecting total protein mass, while ours reflects cell volume. Repeating our experiment by measuring protein mass rather than volume would clarify this. Furthermore, our analysis benefits from a higher resolution of the cell cycle than that by Kafri et al. (2013), conferred by an additional reporter (Cdt1), as well as a more advanced cell cycle analysis algorithm (PSM), which could also explain this discrepancy.

It is also worth noting that although the use of cell lines such as the HeLa-based *fucci* cells makes the study of cell cycle effects much more feasible, many aspects of the physiology of these cells do not reflect the native state of mammalian cells. For example, most mammalian cells exist in tissues and have thus evolved to respond to multiple types of extracellular cues relevant to their neighbouring microenvironment, such as contact with the surfaces of other cells, mechanical stresses and hormonal signals (Glazier, 2018). It is reasonable to assume that, although basic insights can be gained using HeLa cells, the relations underlying the mechanisms described in this chapter are more complex. These results should therefore not be directly related to the multicellular setting without further experiments in primary cells.

Another limitation in the chosen approach is that, although light scatter measured by flow cytometry has been extensively used for approximating cell size (see introduction in (Tzur et al., 2011) for a brief review), it is not a direct measurement of size. Specifically, the acquired measurement is the result of multiple factors such as the difference in refraction index between the suspension fluid and the cells, the angle of beam-to-cell incidence, as well as internal and external surface irregularities, among others. For this reason, the obtained size estimates are quite noisy, and further verifications are required to make sure that meaningful distinctions can be made between the different size groups in Figures 5.8 to 5.21.

Moreover, as stated at the start of the chapter, the use of RNA metabolic labelling and quantification by 5EU results in measurements of the global transcription rate. As the vast majority of RNA in a cell consists of ribosomal RNA (rRNA), our measurements of global transcription rate are likely dominated by that of rRNA. To specifically study the effects of growth rate, cell size and cell cycle phase on the kinetics all other RNA species, an alternative approach combining the powers of metabolic labelling and RNA sequencing can be emploeyd (Herzog et al., 2017). This avenue is explored in the next chapter.

As is often the case in biology, the mechanisms underlying our observations are unlikely to be clear cut. In Figure 5.21, it appears that larger cells are able to grow more with relatively less RNA synthesis than smaller cells. On the other hand, it looks like the growth of larger cells may be limited by DNA, as cells reach a critical size towards the end of the G1 phase. It is possible that RNA decay rates, translational activity, cell cycle duration, or all the above, can be modulated to allow growth in conditions where DNA becomes limiting. A systems approach will enable us to interpret the presented and future measurements by integrating into a mathematical model, which in turn will let us form new, testable hypotheses about the role of these different mechanisms and their contributions to gene expression noise.

To look more closely at the underlying mechanisms, we decided to measure the transcriptional kinetics of the whole genome, at different stages in the cell cycle and for varrying cell sizes. In Chapter 6, we look at how careful consideration of the experimental constraints allows us to optimally design such an ambitious experiment.

Chapter 6

Transcriptomic kinetic analysis

In order to measure the contribution of cell-growth to the observed gene expression variability in an asynchronously growing cell population, we can use metabolic labelling to measure changes in the kinetics of RNA at different stages in the cell cycle and at different cell sizes within each phase. Specifically, once cells have been administered the labelled nucleotide for a specified amount of time (called a ‘pulse’), they can be sorted according to cell size and cell cycle, prior to analysis. A similar approach has been used to discern the rates associated with gene expression at different stages in the cell-cycle using synchronised yeast cells (Eser et al., 2014), whereby the labelled RNA is purified and quantified using microarray analysis.

Here, we use an asynchronously growing population in order to avoid artefacts from the disruption that chemical synchronisation methods cause, or clouding from incomplete synchronisation. Furthermore, we use a more modern approach for measuring the amount of metabolically labelled nucleotides based on sequencing (Baptista and Dölken, 2018, Herzog et al. (2017)). This bypasses the need for biochemical purification, making it more straightforward and thus less prone to technical biases. Specifically, the aim of the experiment is to metabolically label RNA in *fucci* cells, followed by cell sorting based on two variables: cell cycle and cell size. Once the cells have been sorted, they will be analysed by sequencing, in order to quantify the amount of incorporated labelled nucleotide and thus infer the change in RNA kinetics between the different cell-states.

6.1 Revisiting published data

Herzog et al. (2017) the first to use metabolic sequencing to measure the kinetic rates of thousands of mRNA species. For planning our own experiment, it was important to first understand the limitations of this method. To do so, we started by re-analysing the results from (Herzog et al., 2017). Here, we use the reported mRNA turnover rates to predict how many T to C conversions we would expect to see for each gene, and compare these results with the experimentally derived ones. To get the predicted T to C conversions from the RNA turnover rate, we use the below formulation, based on the analysis by Herzog et al. (2017).

Assuming that the rate of T to C mutations observed in reads, $Rate_{TC} = \frac{N_{TC}}{N_T}$, has two sources (background mutations and 4sU incorporation events), we can express it as

$$Rate_{T>C} = \theta_n(p_n + p_o) + (1 - \theta_n)p_o,$$

or

$$Rate_{T>C} = \theta_n p_n + p_o,$$

were p_n is the combined rate of incorporation and chemical conversion of 4sU, p_o is the background rate, θ_n is the fraction of new mRNA, and $Rate_{T>C}$ is the ratio of T to C conversions detected over T's covered, as defined by Herzog et al. (2017).

By rearranging we get a solution for the new mRNA fraction,

$$\theta_n = \frac{Rate_{T>C} - p_o}{p_n}, \tag{6.1}$$

When comparing the number of expected T to C conversions to the observed (see Figure 6.1), we find that the published inferred rates correlate very poorly for the earliest time point (see 45 minute pulse panel in Figure 6.1). Specifically, there is an under-representation in detected T to C conversions compared to those predicted by the published rate of mRNA turnover.

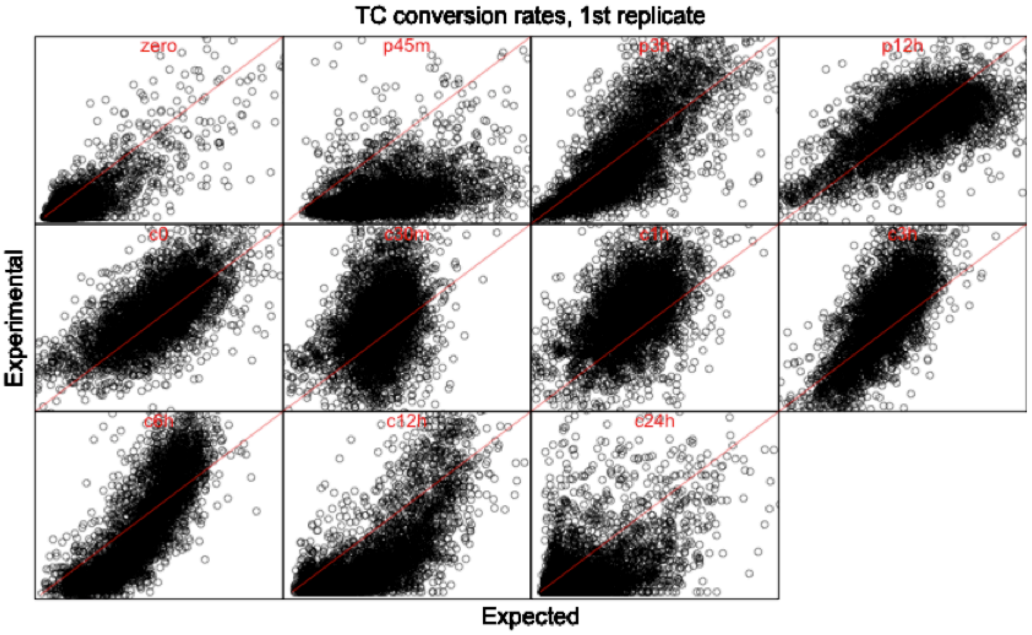

Figure 6.1: **Experimental data vs expected TC conversions based on the fitted degradation rates.**

This can be shown to be consistent between the different replicates, and is likely to be due to the loss of pre-RNA, which may consist of a substantial proportion of the nascent pool of intragenic RNA at early timepoints. Specifically, the library preparation method used in (Herzog et al., 2017) employs poly-A capture, a method for enriching mature mRNA. This warrants further investigation, as in order to minimise the contamination of cells between cell-cycle phases, the duration of the metabolic labelling step needs to be minimised (see Section 6.4).

To that end, we analyse the rates for introns and exons separately next, and compare them. To do so, we use data from another study where pre-RNA is not lost (Baptista and Dölken, 2018). To get these rates, we plug an expression given by Jürges et al. (2018),

$$\lambda = -t\frac{log(2)}{log(1-\theta_n)},$$

which relates the half life λ of a given gene's RNA to the θ_n fraction of new RNA at a single pulse timepoint, t. We substitute θ_n with Equation (6.1) to get

$$\lambda = -t\frac{log(2)}{log(1-\frac{Rate_{T>C}-p_o}{p_n})}.$$

which can be used to obtain an estimate of the half lives directly from the (background subtracted) T to C conversion rate, which is measured experimentally. We use this relation to obtain the half lives of introns and exons separately from the raw data provided by Baptista and Dölken (2018) using their stated p_n and p_o parameters. As expected, introns indeed have a much higher turnover than exons (Figure 6.2). This can also be seen in Figure 6.3, which shows that according to the calculate rates, more than half of the introns (~56%, for genes in chromosome 1) have a labelled fraction of 20% or larger within 20 minutes of labelling, in contrast to less than a quarter of exons (~24%, for genes in chromosome 1). This is especially relevant as it can be shown by simulation that, for our given experimental platform, the nascent RNA fraction and by extension the half-life for the majority of genes is most accurately determined when at least 20% of the counted transcripts are labelled (see Section 6.2).

This suggests that alternative methods of library preparation which retain the pre-RNA, such as ribodepletion instead of polyA-capture may be preferable in our case, as the pre-RNA labelled fraction expands much more rapidly providing a more reliable measurement at shorter pulses.

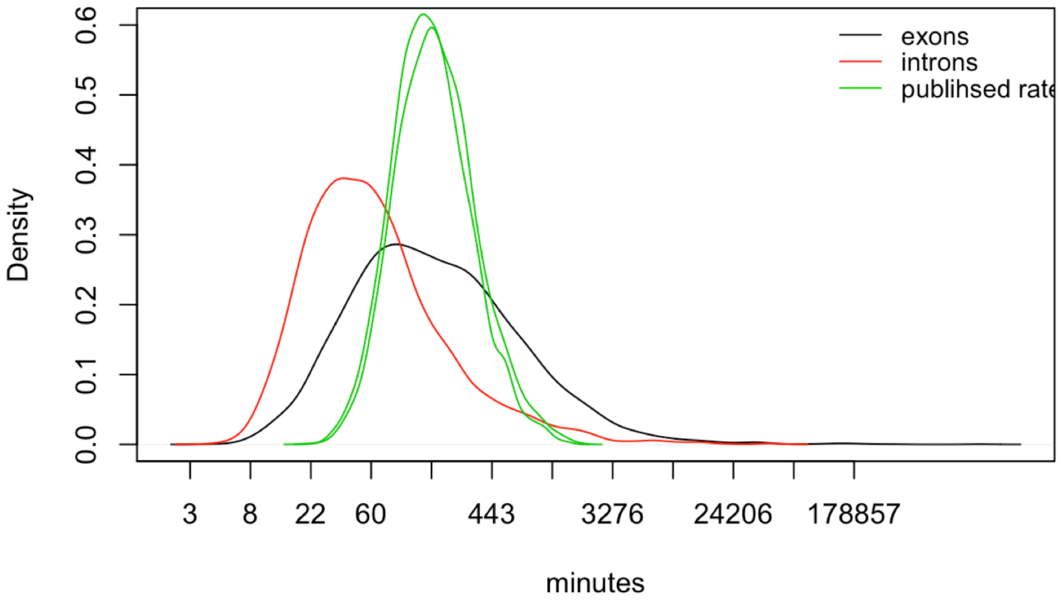

Figure 6.2: **Halflives of introns and exons.** Raw data SLAM-seq data from @Baptista2018 analysed for introns and exons separately. Data for genes on chromosome 1.

6.2 Simulation of 4sU-RNA-seq experiment

As menioned in Chapter 2, a simulation can be used to test how well we can infer the true RNA kinetics from a given SLAM-seq experiment. This simulation can be briefly outlined as

1) Generate n_{reads} number of new reads of length l_r nucleotides.

2) Assign n_u number of uracil moieties by sampling from a binomial with l_r trials and rate U_{bias}, equal to the fraction of U nucleotides in the specific gene.

3) Decide whether the read derives from a new transcript by sampling from a Bernoulli distribution with probability equal to the *true* fraction of new transcripts, which is a function of the specific RNA turnover rate associated with each gene and the duration of the labelling pulse - chase steps.

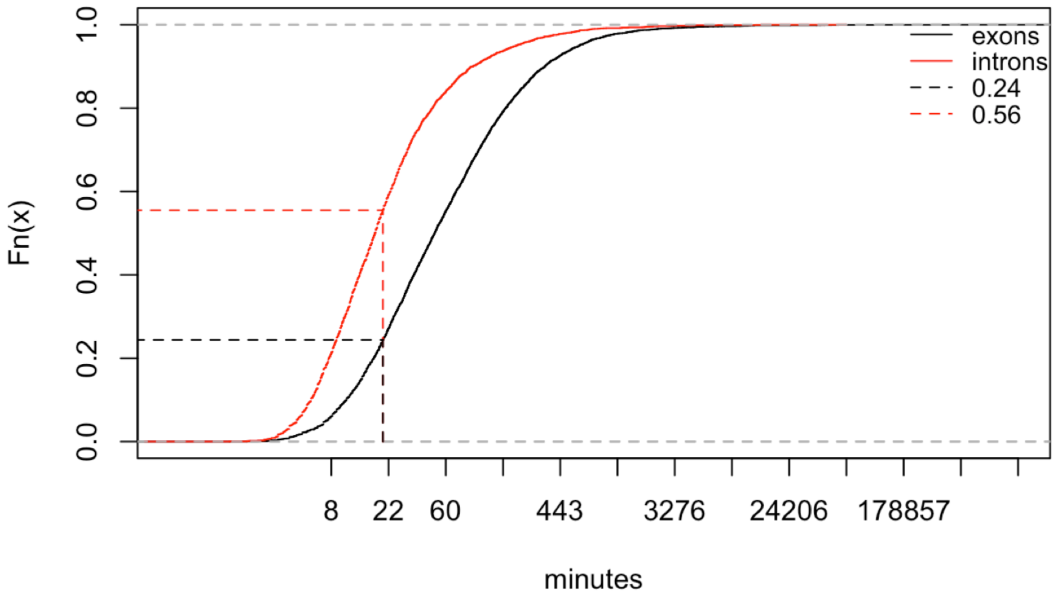

Figure 6.3: **Cumulative Distribution Function of the time it takes for different species of RNA to reach a labelled fraction of 0.2.**

4) According to the result of step 3), assign the number of T-to-C mutations. For new transcripts, use a Binomial with n_u trials and rate equal to the incorporation rate p_n times the chemical efficiency of the conversion step ($y_{chem} = 94\%$ for slam-seq) plus the background mutation rate p_o specific to the sequencing method used, where p_n is the probability of 4sU being incorporated in a given U position. For old transcripts, use a Binomial with n_u trials and probability equal to that of the background mutation p_o.

Using a simple Poisson mixture model, see eq. (6.2), it is possible to infer the fraction of new reads, θ_n (Schofield et al., 2018).

$$f(y_i|\lambda_o, \lambda_n, \theta_n) = \theta_n \text{Poisson}(y_i; \lambda_n) + (1 - \theta_n)\text{Poisson}(y_i; \lambda_o), \tag{6.2}$$

where y_i is the number of T to C mutations detected in read i, θ_n is the fraction of reads deriving from new mRNA (synthesised during the pulse). λ_n and λ_o are rates which correspond to the number of T to C conversions per read, deriving from new and old transcripts respectively. Relating these back to the parameters used for the simulations, $\lambda_n = n_u(p_n y_{chem} + p_o) = l_r U_{bias}(p_n y_{chem} + p_o)$ and $\lambda_o = n_u p_o = l_r U_{bias} p_o$.

Using the simulation described above and the mixture model in Equation (6.2), we can investigate how effectively the fraction of new mRNA, θ_n, can be obtained for different sets of experimental parameters. Specifically, using the published turnover rates of transcripts in conjunction with their relative abundances, we can investigate how different combinations of read depth and pulse times affect the number of genes we can reliably detect. Here, we use the specific transcript abundances measured in the SLAM-seq study by Herzog et al. (2017) and their respective turnover rates to project where the bulk of the genes will lie on a reliability heat map for a given pulse duration and sequencing depth, see Figure 6.4.

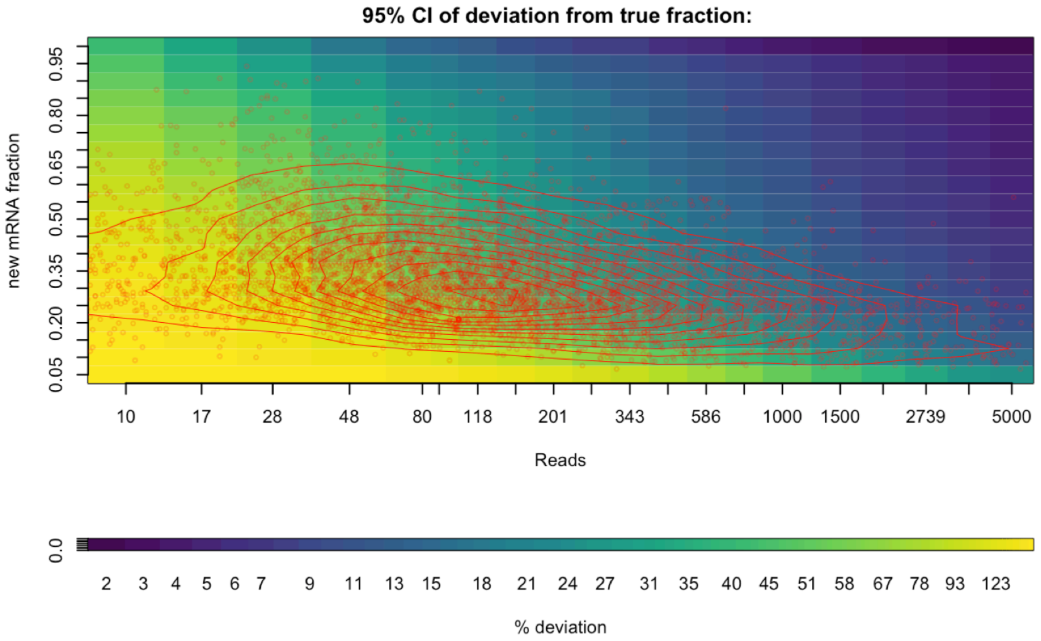

Figure 6.4: **Effect of sequencing depth and metabolic labelling duration on measurement error.** Error defined as a percentage of the 95% confidence interval of the measured fraction, obtained by simulation, over the true fraction. Simulation was run 1000 times for each cell of the heatmap, in order to obtain the relevant distributions. Red circles correspond to individual genes, the locations of which are obtained using published half lives and relative abundances. Red contour lines correspond to the density of genes.

As expected, Figure 6.4 shows that the more reads sequenced per transcript, the higher the accuracy with which we can identify the true fraction of newly transcribed RNA, and by extension the turnover rate of that fraction. Similarly, the larger the fraction of labelled RNA, the more accurately it can be estimated. In this way we can find the region in which we can reliably infer the kinetics.

Using either published results or results from a pilot study, we can use the heat map in Figure 6.4 to predict the proportion of genes for which the newly transcribed fraction can be accurately estimated. Figure 6.4 thus constitutes a guide on how deeply to sequence, as well as how long to administer the labelled nucleotide for, when planning such a kinetic experiment. Furthermore, the effect of other parameters such as read-length and background error can be considered, by changing the parameters of the simulation accordingly.

6.3 Simulation of multi-step experiment

As shown in Section 6.2, we can use simulations in order to measure how accurately we can obtain the kinetics of transcription and decay for different genes in a single step 4sU labelling RNA-seq experiment. It is common in metabolic labelling experiments to obtain multiple measurements in the form of a time series, in order to get a better estimate of the rates, as well as to see what type of model best describes the results. As for the single labelling time point, each of these time points will have an associated sequencing depth and labelling window, which need to be picked accordingly. In order to identify the best distribution of time points and sequencing depths, we formulate a likelihood equation which describes these steps, and test how well the parameters of data simulated as in Section 6.2 can be inferred.

6.3.1 Experiment likelihood equation

As shown by Jürges et al. (2018), the fraction of labelled RNA depends solely on the decay rate, and is given by

$$\theta_n = 1 - e^{-t\mu},$$

where μ is the specific decay rate and t the duration of the labelling pulse. A similar equation can be derived for the chase step. Taking this logic further, we can include both pulse and chase, yielding

$$\theta_n = e^{-\mu t_p} - e^{-\mu(t_p+t_c)},$$

where t_p and t_c are the times for pulse and chase respectively. This way we can calculate the fraction of new transcripts following both pulse and chase steps. Substituting this in (6.2), we get a new time dependent distribution of T to C conversions,

$$\begin{aligned} \mathrm{P}(y_i|\mu) = f(y_i|\mu, t_p, t_c, \lambda_o, \lambda_n) = \\ &= (e^{-\mu t_p} - e^{-\mu(t_p+t_c)})\mathrm{Poisson}(y_i;\lambda_n) \\ &+ (1 - (e^{-\mu t_p} - e^{-\mu(t_p+t_c)}))\mathrm{Poisson}(y_i;\lambda_o). \end{aligned} \tag{6.3}$$

Using the simulation from Section 6.2, we can fit the above distribution to synthetic data produced using different combinations of parameter values, in order to assess how well we can retrieve the true decay rate in each case. This is shown for a set of parameters in Figure 6.5, where we assess the effectiveness of the inference by measuring the fraction of runs for which the inferred decay rate can be obtained within a given precision threshold. This number can be compared between different sets of experimental parameters to help us decide which to choose. Furthermore, a global optimisation algorithm can be used in conjunction with the above construction, such as the Bayesian Optimisation package (Bischl et al., 2017) used in Section 4, in order to automatically find the best set of parameters for our experiment.

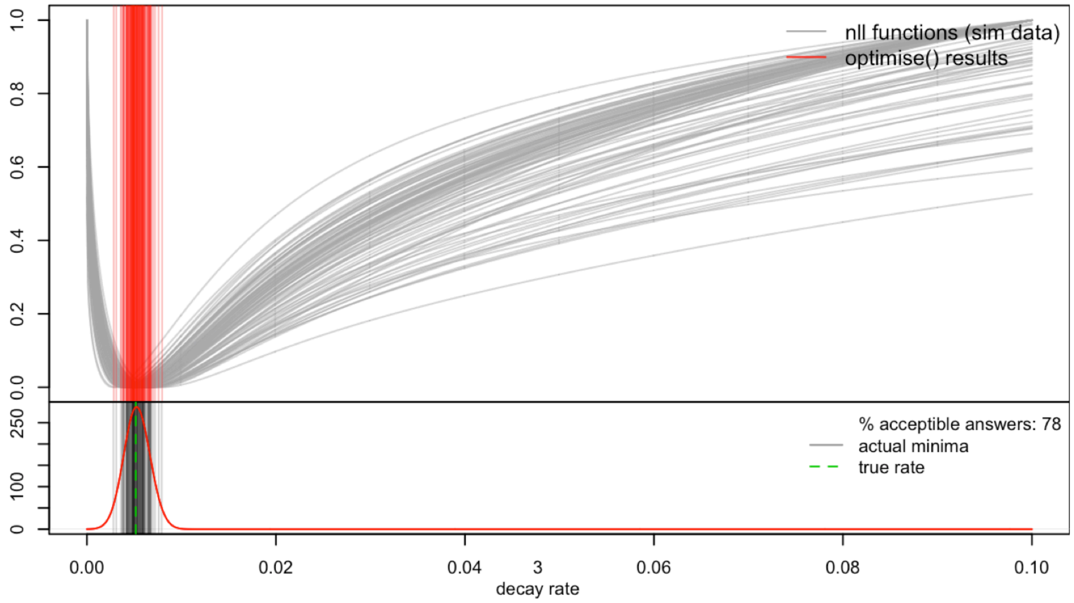

Figure 6.5: **Negative Log Likelihood profile for pulse experiment.** Each line (grey) represents the likelihood equation for a single gene simulation (N = 100), with read depth = 100 and decay rate equal to that of the average intron (0.8 per hour). Minima found using the golden section method. Solutions accepted on the basis of their relative distance (< 10%) from the true decay rate.

6.3.2 Accounting for experimental limitations

As well as testing the overall inference power of the whole experiment and thus enabling us to choose the most suitable experimental parameter sets, we can use this method to test the extent to which different steps in the protocol can affect the reliability of our estimates. For example, an important concern in kinetic studies in general is the bias caused by the time between the end of each pulse-chase step and the actual measurement of the labelled fraction. Although cells can be kept on ice for part of this duration, which is known to reduce the rates of RNA metabolism (Scholtissek, 1967), there are steps such as the removal of the labelling solution and re-suspension of the cells, which together can take up to

20 minutes. This amount of time is comparable to the half-lives of many of the faster - turnover RNA species, and can therefore lead to a significant systematic overestimation of the decay rates.

For this reason, it is useful to know whether we can take this into account in our model. Figure 6.6 shows the resulting likelihood profiles of several repetitions of the simulation, using a single set of parameters which include a sample processing time of 10 minutes. The sample processing time is easily implemented into the model as an additional chase - step at the end of each pulse or chase step. The effect of cooling cells on ice is implicitly accounted for by assuming that the rates of degradation and transcription are zero during the duration of this step. The validity of this assumption can be tested using a separate experiment, whereby cells are loaded with the labelled nucleotide prior to incubation on ice, and the deterioration of the signal while on ice measured. If the degradation rate on ice is found to be non-negligible, as found in certain cases by Sensky and Rees (1976), the above formulation can be used to account for it in the same way as for the wash steps, using an additional set of rates.

Interestingly, simulating the time it takes to perform the required wash-steps leads to a bi-modal likelihood profile, thus complicating the fitting (compare with results with no processing time, Figure 6.6).

Although in this example the global minimum is still correct, this is not true in all cases (not shown). Specifically, shorter timepoints are more prone to leading to erroneus estimations of the decay rates, as the fraction of time spent processing the sample after the incubation step increases. In our case, shorter labelling steps are preferred in order to minimise the clouding effects caused by cells travelling between phases, as discussed in Section 6.4. For that reason, the option of reversably fixing cells was investigated.

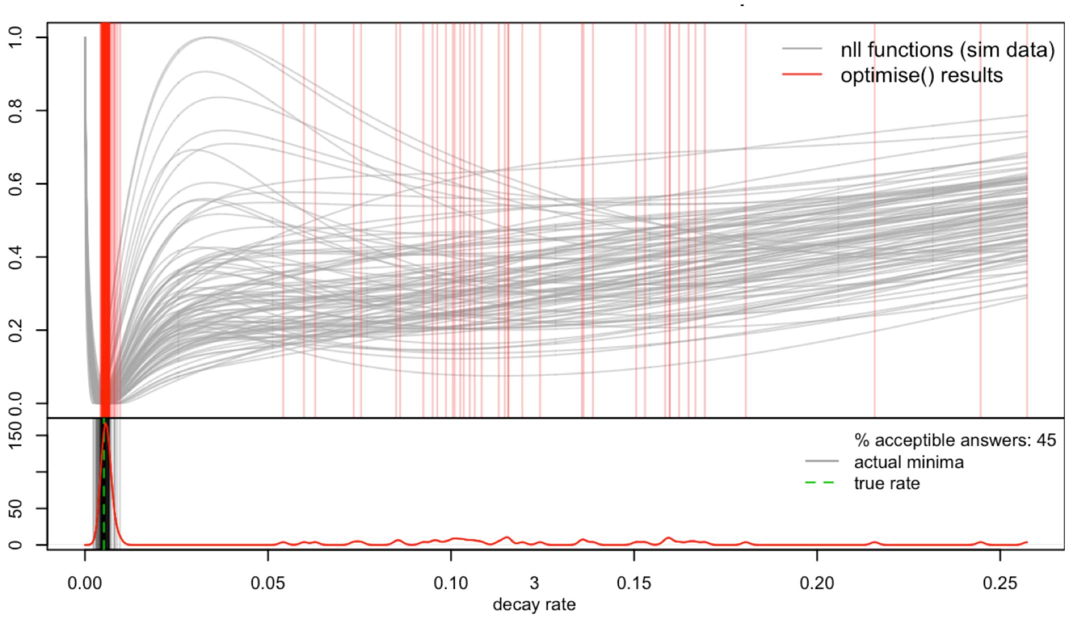

Figure 6.6: **Negative Log Likelihood profile for pulse experiment with wash step**. Each line (grey) represents the likelihood equation for a simulated intron (N = 100), with read depth = 100 and decay rate equal to that of the average intron (0.8 per hour). Minima found using the golden section method. Solutions accepted on the basis of their relative distance (10 percent) from the true decay rate.

6.4 Integration of phases during metabolic labelling

Expression fold change can vary by up to 10 fold during the cell cycle in many genes (Kuang et al., 2012), which suggests that the transcription and degradation rates need to be adjusted accordingly. In order to measure these rates accurately between phases, it is important that we account for the cell cycle time that passes during the labelling step. Specifically, during the pulse of the chemically labelled nucleotide, cells will travel from one phase to another, as shown in Figure 6.7. The contribution of cells from the previous phase affects the average amount of labelled RNA within each phase to varying extents, depending on the ratio of $\frac{T_{pulse}}{T_{phase}}$, where T_{pulse} and T_{phase} are the duration of the present cell cycle phase and labeling pulse, respectively.

In order to account for this effect we need to model the incorporation of labelled nucleotide occurring over two adjacent phases with two sets of rates, assuming that no cells cross to a third phase during the labeling pulse, and integrate over the 'effective duration' of the first phase (time spent in first phase before entering second phase), in order to get the average of all the possible such 'effective durations'.

From mass action kinetics, we have

$$\frac{dx}{dt} = \lambda - \mu x(t),$$

where x is the number of mRNAs (as a unit of concentration). We wish to model the progression of a cell from one phase to the next. We first solve for $x(t)$ for the first phase, with $x(0) = 0$.

We get the time varying equation

$$x(t) = \frac{\lambda_1 e^{\mu_1(-t)} \left(e^{\mu_1 t} - 1\right)}{\mu_1}$$

,

where λ_1 and μ_1 are the mRNA synthesis and degradation rates during the first phase, respectively.

Now we use this solution to get the initial conditions for the next phase. Specifically,

$$x(t_1) = \frac{\lambda_1 e^{\mu_1(-t_1)} \left(e^{\mu_1 t_1} - 1\right)}{\mu_1}$$

,

which we use to solve $\frac{dx}{dt} = \lambda_2 - \mu_2 x(t)$, where λ_2 and μ_2 are the mRNA synthesis and degradation rates during the second phase, respectively, and t_1 is the duration of the time that a cell spends in the previous phase.

This results in the below equation:

$$x(t) = \frac{e^{\mu_1(-t_1)-\mu_2 t} \left(\lambda_2 \mu_1 e^{\mu_1 t_1+\mu_2 t} - \lambda_2 \mu_1 e^{\mu_1 t_1+\mu_2 t_1} - \lambda_1 \mu_2 e^{\mu_2 t_1} + \lambda_1 \mu_2 e^{\mu_1 t_1+\mu_2 t_1}\right)}{\mu_1 \mu_2}, \tag{6.4}$$

where t > t1.

As mentioned above, we need to average over all the possible durations of time a cell can spend in the previous phase, which can be done by integrating t_1 from zero to the duration of the pulse (t), and re-scaling by dividing by the duration of the pulse. This yields the below function

$$x(t) = \frac{\frac{\lambda_1 \mu_2 \left(\mu_1 + \mu_1 \left(-e^{\mu_2(-t)}\right) + \mu_2 \left(e^{\mu_1(-t)} - 1\right)\right)}{\mu_1(\mu_1 - \mu_2)} + \lambda_2 \left(\mu_2 t + e^{\mu_2(-t)} - 1\right)}{\mu_2^2 t}, \tag{6.5}$$

where t is the duration of the pulse.

As mentioned earlier, the contribution of cells from the previous phase depends on the ratio of $\frac{T_{pulse}}{T_{phase}}$. This is due to the fact that there are two populations of cells contributing to the resulting average. These can be described as 1) cells that started in the previous phase when the staining started and ended up in the next phase, and 2) cells which started and ended within the second phase, corresponding to the red and green populations seen in figure 6.7, respectively. The relative contributions of these two population to the overall average will depend on the relative duration of the staining pulse and the second phase. Specifically,

$$\frac{pop1}{pop2} = \frac{T_{pulse}}{T_{phase} - T_{pulse}}$$

The first population will obey the integrated two-phase model, while the second population will obey the single-phase model. For fitting, we thus need to take the weighted average of the contributing populations. It is the focus of future research to determine the extent to which this construction can be used to take into account the progression of cells from phase to phase during the labelling step, by simultaneously fitting the data from all cell-phases.

6.5 Experimental Design

The results from this chapter lead us to suggest three important alterations to the basic metabolic sequencing protocol suggested by Herzog et al. (2017). Firstly, in Section 6.1 we saw that shorter labelling timepoints suffered from an understimation of the RNA turnover, which we hypothesised to be due to the choice of a library preparation method which enriches for mature RNA (polyA-capture) and thus lead to a loss of much of the newly labelled preRNA. This prompted me to suggest an alternative method of library preparation based on ribodepletion, which conserves the nascent RNA fraction.

Secondly, using the simulation by Baptista and Dölken (2018) combined with

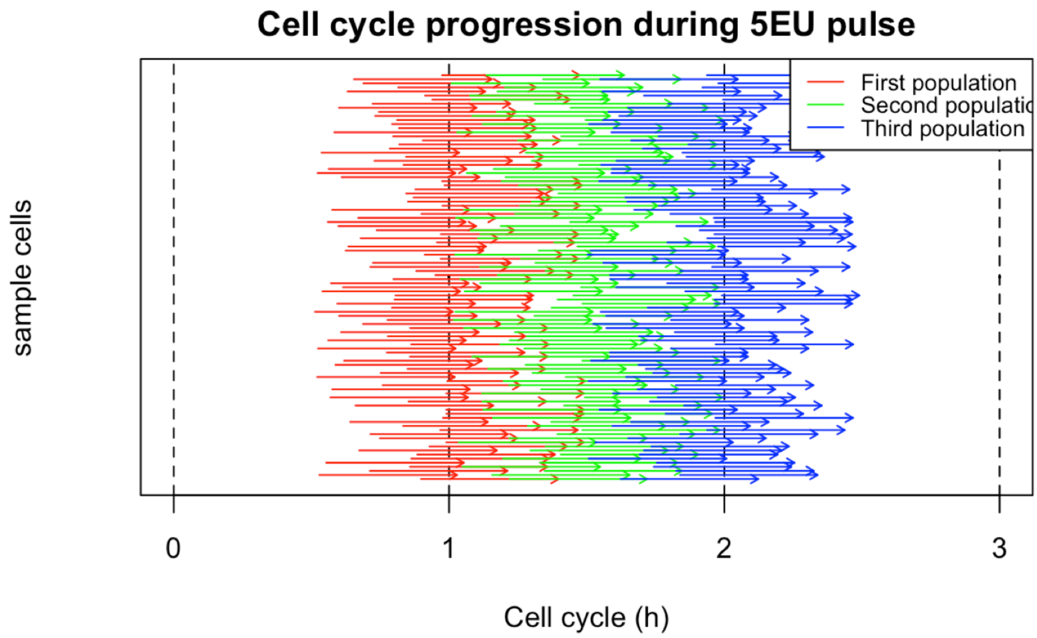

Figure 6.7: **Cells progressing through adjacent stages of the cell cycle during the metabolic labelling pulse.**

mathematical modelling, in Section 6.2 we saw for which experimental conditions (labelling time and sequencing depth) the RNA turnover rates could be accurately inferred for the majority of genes. We found that shorter labelling times (smaller new RNA fraction) were associated with a higher measurement error, which could be mitigated to an extent by sequencing more deeply. An alternative approach could be to enrich for the nascent transcripts, by means of nuclear fractionation of the cells prior to libary preparation. As we are interested in shorter timepoints in order to avoid the cell-cycle clouding effects discussed in Section 6.4, I suggested we employ nuclear fractionation in order to enrich for the labelled RNA fraction.

Thirdly, the analysis in Section 6.3.2 showed us that sample preparation time can lead to biased estimations of RNA kinetics. Our experiment relies on extensive sample processing following the labelling step, such as cell sorting into different cell cycle phases and distinct cell sizes. In order to minimise the error caused by the elapse time during these processing steps, we looked into reversible fixation protocols that will allow us to preserve the state of the cells immediately following the labelling step. I thus identified reversible fixation using dithio-bis(succinimidyl propionate), (DSP), also known as Lomant's Reagent, which has been shown to be useful in preserving the state of cells prior to sequencing (Attar et al., 2018).

Using the above suggestions, a pilot experiment was performed, the results of which are shown in Figures 6.8 and 6.9. In pilot 1, we pulsed cells for 10 min, 20 min, and 1h. In Figure 6.8 the average TC/AG counts can be seen to increase with labeling time, while the background mutation rate of other baseflips remains constant and at the same level as the negative control. In each of the cases the intron signal (red) is higher, consistent with the higher turnover rate of pre-RNA. Also, the DSP fixation does not seem to affect the signal much, though more repeats are required to confirm that.

In pilot 2 we only used 1 timepoint (1h), and tested the effect of sorting, UV exposure and nuclear enrichment after DSP fixation. As in pilot 1, the DSP fixation does not affect the signal intensity (compare lane 1 and 3), and it causes no increase in the background mutation rate (compare lane 2 and 4). Finally, nuclear enrichment leads to a big increase in the signal (see lane 5, Nuc), which is maintained after sorting, even with the UV laser on (lanes 6 and 7). The below table can be used to identify the different samples.

Sample / Steps	4sU pulse	DSP	FACS	UV	nuclear ext
4sU-10min	0.5mM 10min	-	-	-	-
4sU-20min	0.5mM 20min	-	-	-	-
4sU-60min	0.5mM 60min	-	-	-	-
control	-	-	-	-	-
DSP	0.5mM 60min		-	-	-
4sU	1mM 60min	-	-	-	-
ve-minus	-				
DSP-plus	1mM 60min		-	-	-
DSP-minus	-		-	-	-
Nuc	1mM 60min		-	-	

Sample / Steps	4sU pulse	DSP	FACS	UV	nuclear ext
UV-Plus	1mM 60min				
UV-Minus	1mM 60min			-	

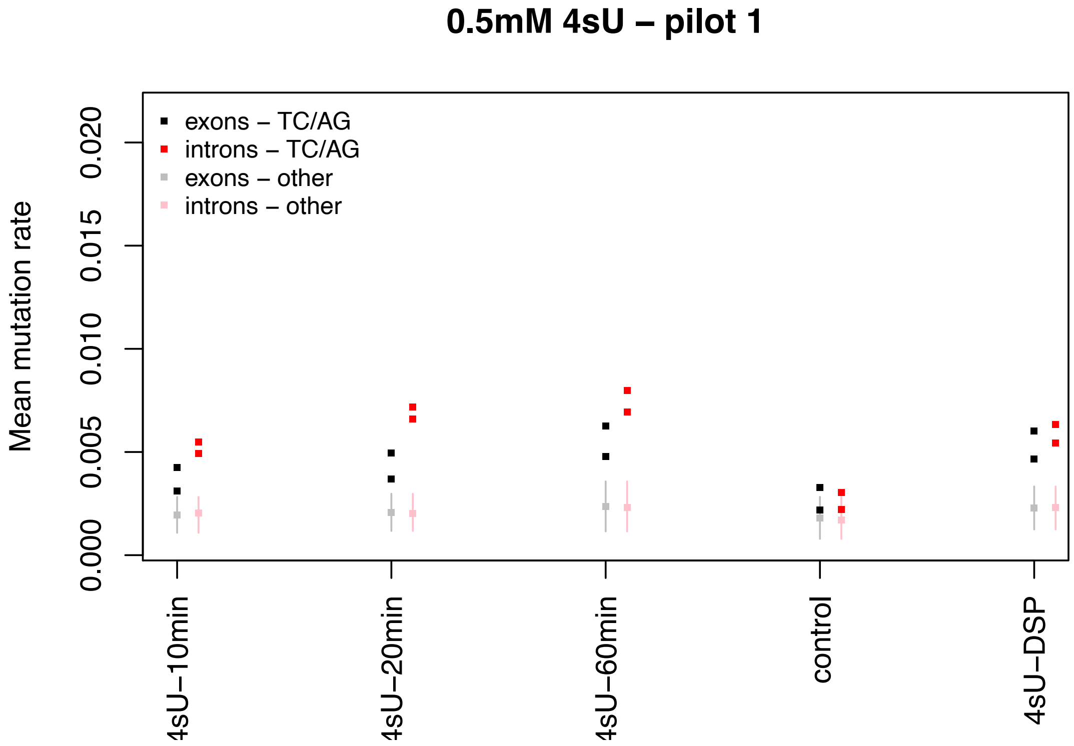

Figure 6.8: **SLAM-seq pilot experiment 1.** Cells were labelled with 500 micromolar 4sU for either 10, 20 or 60 minutes prior to reversible crosslinking with DSP. Library preparation and sequence processing performed by Dr Mark Walsh following the SMART-seq method (TAKARA BioSciences). Mutation rate obtained using the analysis in Section 6.1.

6.6 Discussion

As is often the case with high throughput experiments, careful planning is required to get the most value out of this method. Using computer simulations and parameters from published data we can understand what the required settings such as labelling time, processing time and sequencing depth are, in order to optimise the experiment. Such considerations are especially important when

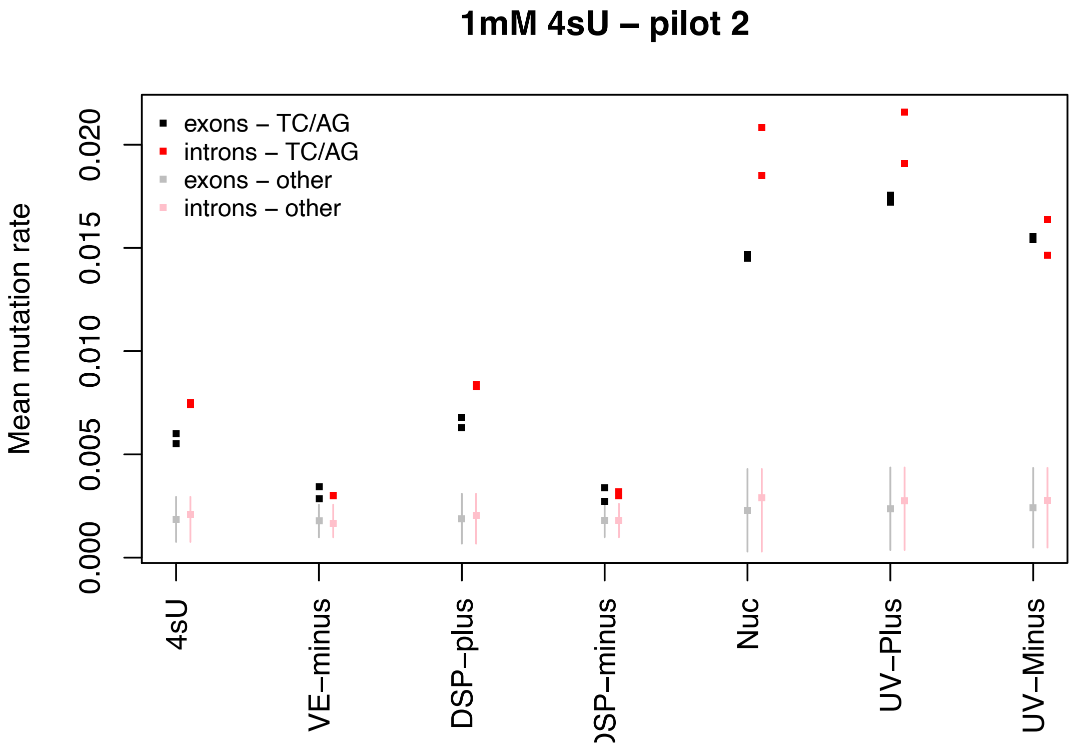

Figure 6.9: **SLAM-seq pilot experiment 2.** Cells were labelled with 1000 micromolar 4sU for 60 minutes prior to reversible crosslinking with DSP. Library preparation and sequence processing performed by Dr Mark Walsh following the SMART-seq method (TAKARA BioSciences). Mutation rate obtained using the analysis in Section 6.1.

planning a large experiment on a budget.

This type of work falls under the category of optimal design of experiments. Recently, Uvarovskii et al. (2019) demonstrated a way of optimising the above experiment for a single timepoint by deriving an analytical solution to the problem, which uses Fisher Information maximisation as the optimisation criterion. An alternative approach to optimise any type of process is via simulation (see (Hong et al., 2015) for a review of optimisation via simulation methods). We have chosen the latter type of approach as it allows for greater flexibility in the structure of the experiment designed.

Using simulations and modelling, we saw that the processing time of the samples following a pulse step can be potentially accounted for (Figure 6.6), though a more robust algorithm is required for finding the global minimum instead of the bisection method currently employed. Once such an algorithm has been implemented, Bayesian Optimisation (Bischl et al., 2017) or some other equivalent optimiser can be used to find the best experimental parameters. Similarly, we saw that while cells transitioning between adjacent phases in the cell cycle during the metabolic labelling step can cloud our estimation of rates specific to each phase, this can potentially be accounted for by fitting the data from different phases simultaneously.

Using the results from the analysis in this chapter we designed an experiment which best suits the requirements of this study. Using two pilot experiments we saw that the suggested changes can be succesfully employed to suit our needs. Specifically, the pilot experiments show that we can fix the samples after short 4sU pulses which will allow us to sort into different sizes and cell cycle stages without worrying about time passing. Furthermore, nuclear enrichment gives a strong boost (~3 fold) in the signal, which suggests we can perform multiple shorter timepoints in order to fit the rates accurately, without worrying about the averaging between cell-cycle phases that longer 4sU pulses would involve.

Chapter 7

Conclusion

Much of modern biology has been concerned with identifying the causes and effects underpinning biological functions. This has led to a vast wealth of knowledge, and the understanding of thousands of molecular mechanisms underlying health and disease, development and aging. This understanding has in turn led to the birth of synthetic biology, which promises to tackle many of the worlds greatest challenges such as food security, sustainable energy, currently incurable diseases, and more.

However, relationships between cause and effect are seldom linear in biology. The complex interplay between gene expression homeostasis and cell growth, which relies on feedback mechanisms that are only now starting to become understood, is a good example. In such cases, it is crucial to consider the system as a whole, since individual components in isolation often do not capture the observed phenomena. Molecular systems biology allows us to express such complex, dynamic interactions, in a way that enables us to test hypotheses which encompass the whole system at once.

The utility of systems biology comes with its own set of challenges. Attempting to understand how a system behaves requires a good level of understanding of the individual components, for each of which there are often entire fields devoted. It is thus necessary to be able to distill the available information in a way that enables implementation into a model, while preserving all key mechanistic aspects. Furthermore, systems biology is a truly interdisciplinary field. For example, systems biology publications can be found across physical, biological, mathematical and computational journals. Therefore an appreciation of all these distinct scientific domains is required in order to fully utilise the sum of accomplished work. Such an effort has been made here, where we looked at methods for studying gene expression noise and cell growth.

In the Chapter 3, we saw how a simple set of models encompassing aspects of cell growth and gene expression can be derived. This will be useful in interpreting future experimental results, and can be used as a tool for testing hypotheses. Although more complex models of this kind exist, here we wanted to see whether a simpler model without stochastic promoter switching can adequately explain the observed behaviours, as suggested in recent experimental results (Ietswaart et al., 2017; Zopf et al., 2013; Battich et al., 2015; Klein et al., 2015). A similar modelling approach was followed by Soltani et al. (2016). Specifically, we used assumptions of stationarity and spatial homogeneity to see how RNA dynamics based on mass action kinetics govern the observed distribution of RNA molecules numbers in a population of growing cells. We modeled the cell cycle as having three phases corresponding to G1, S and G2/M phases of the eukaryotic cell cycle, which can be readily resolved experimentally by DNA staining, and thus directly compared to the model.

In Chapter 4 we saw how PSM, a method developed for delineating paths of immune cell differentiation, can be re-purposed for resolving the cell cycle. In the process, we identified certain technical limitations of PSM, for which suggestions were made. Specifically, we found that fitting one measurement at a time as suggested by Bagwell et al. (2015b) can ultimately lead to suboptimal solutions, as the optimum found using the first measurement usually does not correspond to the true solution, thus constraining the discovery of the global optimum when considering further measurements. We saw how this can be overcome by fitting all available measurements simultaneously. Although this leads to a combinatorial explosion in the explored parameter state-space, this was addressed using a Bayesian Optimisation approach, thus allowing for intelligent exploration of the parameters.

Another observed limitation of PSM was the inability to capture the apparent population heterogeneity in certain cases. Specifically, we saw that, within a population of growing cells, Cdt1 can take two levels of mean expression at the same time during the cell cycle, as has been noted by Grant et al. (2018). Although this was interesting in its own right and should be the topic of future work, it meant that the continuous piecewise models used in PSM could not adequately describe the cell cycle profile of Cdt1, proving an obstacle in resolving the cell cycle. Here, we saw how to get around this by selectively excluding uninformative regions of a given marker's cell cycle path, during which we can rely on more informative markers. In this way, we constructed a descriptive cell cycle model based on the *fucci* reporters geminin and Cdt1 (Sakaue-Sawano et al., 2008), in combination with measurements of DNA quantity. This model can be used to either study processes of the cell cycle, or to specifically control for the effects of the cell cycle when studying unrelated process. This is particularly important when studying gene expression noise, for which extrinsic contributors such as the cell cycle can play an important role.

In Chapter 5 we specifically looked at how cell growth contributes to gene expression noise by affecting the kinetics of RNA molecules. We used the proposed *fucci* probability state models from the previous chapter to look at how the rate of RNA synthesis varies with respect to both cell size and cell cycle. Interestingly, we saw that, contrary to previous findings (Schmidt and Schibler, 1995; Padovan-Merhar et al., 2015), the rate of RNA synthesis alone cannot explain the variation seen in RNA numbers in cells of different sizes, suggesting that modulation of RNA stability may play a more important role than previously thought. Although this is contrary to previous findings, the discrepancy could be explained by the strong relation between transcription rate and cell cycle progression, which had not been previously controlled for. Furthermore, recent findings in yeast and mammalian cells (Neurohr et al., 2019) suggest that the rate of transcription can become limitted by the decreased concentration of DNA in larger cells, although this idea has been also been challenged (Sun et al., 2019b). Surprisingly, our high resolution cell cycle analysis, in combination with ergodic rate analysis, uncovered a burst in cell growth rate seen in larger cells during the late S phase. Further experimental repeats alongside orthogonal measurements of cell size and growth rate are required to verify these findings.

In Chapter 6 we set the ground for performing a high throughout metabolic sequencing experiment (Herzog et al., 2017), for measuring the kinetics of thousands of RNA species at different stages in the cell cycle and for different cell sizes. Such an experiment is expected to yield a comprehensive, detailed view on cell size and cell cycle transcriptional regulation, as well as an understanding of the genetic determinants underlying the observed variation in RNA molecule numbers in growing populations of cells. Using simulations based on the work from (Baptista and Dölken, 2018), and revisiting previous data, we saw that high turnover RNA species can be more easily detected by metabolic sequencing, and thus decided to use preRNA synthesis as a proxy for measuring transcriptional activity. This highlighted the need for preserving the state of the cell at the time of sampling throught the processing steps of the expriment. This was achieved using reversible fixation of the cells (Attar et al., 2018), which was optimised for our experimental conditions. We further discuss how the simulations and mathematical models used in this chapter can be emploeyd to optimise the parameters of this experiment, as recenlty explored by Uvarovskii et al. (2019). The methods used and developed here will hopefully prove useful for further exploration of the mechanisms underlying gene expression noise and cell growth.

7.1 Future work

In the future, the number of phases modelled in Chapter 3 can be extended to an arbitrary number, in order to enable comparison with more sophisticated cell cycle resolution methods such as the one described herein. Although we focused on the effect of cell growth on gene expression noise, feedback relationships between cell size, cell growth rate and gene expression should be considered in the future, which could be explored using coarse grained modelling (Shahrezaei and Marguerat, 2015). Finally, once available, it will be informative to see how well the models developed in this chapter compare to cell cycle-resolved smFISH data.

In order to verify the metabolic labelling results in Chapter 5, more experiments would be required. Specifically, the cell size measurements which are based on cytometry light scatter need to be repeated using more precise methods such as using an impedence based Coulter counter. Such measurements could be used to test the resolution and calibrate the light scatter measurements of the flow cytometer, in order to correlate with other parameters, such as cell cycle and transcription rate.

More direct measurements of growth rate could be achieved by using alternative metabolic labelling approaches, such as that of translation or lipid biosynthesis. Once reliable measurements have been obtained, a model which combines the interactions described in this chapter would be helpful.

PSM should be extended to cope with discontinuous functions such as the one determined for Cdt1 in Chapter 4. Alternative cell cycle reporters based on established markers such as cyclins should also be tried (Gookin et al., 2017), in order to confirm the results. Finally, PSM should be compared to other existing pseudotime detection algorithms.

Parameter constraints should be incorporated into the mlrMBO package so that parameter space can be explored more efficiently for the PSM models. Optimisation of the metabolic sequencing experiment should be performed using mlrMBO or equivalent, in order to find the best timepoints to sample the level of metabolic labelling incorporation at, so that the number of genes that can be reliably measured is maximised. Finally, the model proposed by Baptista and Dölken (2018) should be used to extract the turnover rates from our pilot experiments, while simulations should be used to calculate the associated confidence of each rate. Together, the above should be used to plan a larger scale experiment, with measurements for each cell cycle phase and cell size group.

In the future, it would be interesting to investigate whether there are commonalities in the sequence of genes with similar kinetics or similar patterns during the cell cycle. This could be achieved once the data becomes available using deep learning (Agarwal and Shendure, 2018; Washburn et al., 2019).

Bibliography

Acar, M., Becskei, A., and van Oudenaarden, A. (2005). Enhancement of cellular memory by reducing stochastic transitions. *NATURE*, 435(7039):228–232.

Adan, A., Alizada, G., Kiraz, Y., Baran, Y., and Nalbant, A. (2017). Flow cytometry: basic principles and applications. *Critical Reviews in Biotechnology*, 37(2):163–176.

Agarwal, V. and Shendure, J. (2018). Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *bioRxiv*, page 416685.

Angelidis, I., Simon, L. M., Fernandez, I. E., Strunz, M., Mayr, C. H., Greiffo, F. R., Tsitsiridis, G., Ansari, M., Graf, E., Strom, T.-M., Nagendran, M., Desai, T., Eickelberg, O., Mann, M., Theis, F. J., and Schiller, H. B. (2019). An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics. *NATURE COMMUNICATIONS*, 10.

Antolovic, V., Miermont, A., Corrigan, A. M., and Chubb, J. R. (2017). Generation of Single-Cell Transcript Variability by Repression. *CURRENT BIOLOGY*, 27(12):1811+.

Attar, M., Sharma, E., Li, S., Bryer, C., Cubitt, L., Broxholme, J., Lockstone, H., Kinchen, J., Simmons, A., Piazza, P., Buck, D., Livak, K. J., and Bowden, R. (2018). A practical solution for preserving single cells for RNA sequencing. *Scientific Reports*, 8(1):1–10.

Avva, J., Weis, M. C., Sramkoski, R. M., Sreenath, S. N., and Jacobberger, J. W. (2012). Dynamic Expression Profiles from Static Cytometry Data: Component Fitting and Conversion to Relative, “Same Scale” Values. *PLoS ONE*, 7(7):e38275.

Bagwell, C. B. and Adams, G. E. (1993). Fluorescence Spectral Overlap Compensation for Any Number of Flow Cytometry Parameters. *Annals of the New York Academy of Sciences*, 677(1 Clinical Flow):167–184.

Bagwell, C. B., Hill, B. L., Wood, B. L., Wallace, P. K., Alrazzak, M., Kelliher, A. S., and Preffer, F. I. (2015a). Human B-cell and progenitor stages as determined by probability state modeling of multidimensional cytometry data. *Cytometry Part B - Clinical Cytometry*, 88(4):214–226.

Bagwell, C. B., Hunsberger, B. C., Herbert, D. J., Munson, M. E., Hill, B. L., Bray, C. M., and Preffer, F. I. (2015b). Probability state modeling theory. *Cytometry Part A*, 87(7):646–660.

Bahar Halpern, K., Caspi, I., Lemze, D., Levy, M., Landen, S., Elinav, E., Ulitsky, I., and Itzkovitz, S. (2015). Nuclear Retention of mRNA in Mammalian Tissues. *Cell Reports*, 13(12):2653–2662.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., and Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *SCIENCE*, 305(5690):1622–1625.

Baptista, M. A. and Dölken, L. (2018). RNA dynamics revealed by metabolic RNA labeling and biochemical nucleoside conversions. *Nature Methods*, 15(3):171–172.

Bar-Even, A., Paulsson, J., Maheshri, N., Carmi, M., O’Shea, E., Pilpel, Y., and Barkai, N. (2006). Noise in protein expression scales with natural protein abundance. *NATURE GENETICS*, 38(6):636–643.

Barkai, N. and Leibler, S. (2000). Biological rhythms - Circadian clocks limited by noise. *NATURE*, 403(6767):267–268.

Battich, N., Stoeger, T., and Pelkmans, L. (2015). Control of Transcript Variability in Single Mammalian Cells. *Cell*, 163(7):1596–1610.

Bendall, S. C., Davis, K. L., Amir, E. A. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'Er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725.

Bertaux, F., Marguerat, S., and Shahrezaei, V. (2018). Division rate , cell size and proteome allocation : impact on gene expression noise and implications for the dynamics of genetic circuits. *bioRxiv*.

Bischl, B., Richter, J., Bossek, J., Horn, D., Thomas, J., and Lang, M. (2017). mlrMBO: A Modular Framework for Model-Based Optimization of Expensive Black-Box Functions. *arXiv*.

Blake, W. J., Kaern, M., Cantor, C. R., and Collins, J. J. (2003). Noise in eukaryotic gene expression. *NATURE*, 422(6932):633–637.

Brooks, R. F., Bennett, D. C., and Smith, J. A. (1980). Mammalian cell cycles need two random transitions. *Cell*, 19(2):493–504.

Cadart, C., Monnier, S., Grilli, J., Sáez, P. J., Srivastava, N., Attia, R., Terriac, E., Baum, B., Cosentino-Lagomarsino, M., and Piel, M. (2018). Size control in mammalian cells involves modulation of both growth rate and cell cycle duration. *Nature Communications*, 9(1).

Chan, L. Y., Mugler, C. F., Heinrich, S., Vallotton, P., and Weis, K. (2018). Non-invasive measurement of mRNA decay reveals translation initiation as the major determinant of mRNA stability. *eLife*, 7:1–32.

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., and Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, 453(7194):544–547.

Charvin, G., Cross, F. R., and Siggia, E. D. (2008). A Microfluidic Device for

Temporally Controlled Gene Expression and Long-Term Fluorescent Imaging in Unperturbed Dividing Yeast Cells. *PLOS ONE*, 3(1):1–12.

Chen, X. and Zhang, J. (2016). The Genomic Landscape of Position Effects on Protein Expression Level and Noise in Yeast. *Cell Systems*, 2(5):347–354.

Cheung, P., Vallania, F., Warsinske, H. C., Donato, M., Schaffert, S., Chang, S. E., Dvorak, M., Dekker, C. L., Davis, M. M., Utz, P. J., Khatri, P., and Kuo, A. J. (2018). Single-Cell Chromatin Modification Profiling Reveals Increased Epigenetic Variations with Aging. *CELL*, 173(6):1385+.

Chiorino, G., METZ, J. A. J., TOMASONI, D., and UBEZIO, P. (2001). Desynchronization Rate in Cell Populations: Mathematical Modeling and Experimental Data. *Journal of Theoretical Biology*, 208(2):185–199.

Colman-Lerner, A., Gordon, A., Serra, E., Chin, T., Resnekov, O., Endy, D., Pesce, C. G., and Brent, R. (2005). Regulated cell-to-cell variation in a cell-fate decision system. *NATURE*, 437(7059):699–706.

Cozy, L. M. and Kearns, D. B. (2010). Gene position in a long operon governs motility development in Bacillus subtilis. *MOLECULAR MICROBIOLOGY*, 76(2):273–285.

Cramer, P. (2019). Organization and regulation of gene transcription. *Nature*.

Das, S., Sarkar, D., and Das, B. (2017). The interplay between transcription and mRNA degradation in Saccharomyces cerevisiae. *Microbial Cell*, 4(7):212–228.

Dietrich, J.-E. and Hiiragi, T. (2007). Stochastic patterning in the mouse pre-implantation embryo. *DEVELOPMENT*, 134(23):4219–4231.

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2012). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21.

Dowling, M. R., Kan, A., Heinzel, S., Zhou, J. H., Marchingo, J. M., Wellard, C. J., Markham, J. F., and Hodgkin, P. D. (2014). Stretched cell cycle model for proliferating lymphocytes. *Proceedings of the National Academy of Sciences of the United States of America*, 111(17):6377–6382.

Duursma, A. and Agami, R. (2005). p53-Dependent Regulation of Cdc6 Protein Stability Controls Cellular Proliferation. *Molecular and Cellular Biology*, 25(16):6937–6947.

Eldar, A. and Elowitz, M. B. (2010). Functional roles for noise in genetic circuits. *Nature*, 467(7312):167–173.

Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002). Stochastic gene expression in a single cell. *Science Signaling*, 297(5584):1183.

Enge, M., Arda, E., Mignardi, M., Beausang, J., Bottino, R., Kim, S. K., and Quake, S. R. (2017). Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *CELL*, 171(2):321+.

Eser, P., Demel, C., Maier, K. C., Schwalb, B., Pirkl, N., Martin, D. E., Cramer, P., and Tresch, A. (2014). Periodic mRNA synthesis and degradation co-operate during cell cycle gene expression. *Molecular systems biology*, 10 VN - r:717.

Franceschini, N., Kirschfeld, K., and Minke, B. (1981). FLUORESCENCE OF PHOTORECEPTOR CELLS OBSERVED INVIVO. *SCIENCE*, 213(4513):1264–1267.

Fraser, H. B., Hirsh, A. E., Giaever, G., Kumm, J., and Eisen, M. B. (2004). Noise minimization in eukaryotic gene expression. *PLOS BIOLOGY*, 2(6):834–838.

Fuhrmann, F., Lischke, T., Gross, F., Scheel, T., Bauer, L., Kalim, K. W., Radbruch, A., Herzel, H., Hutloff, A., and Baumgrass, R. (2016). Adequate immune response ensured by binary IL-2 and graded CD25 expression in a murine transfer model. *ELIFE*, 5.

García-Martínez, J., Delgado-Ramos, L., Ayala, G., Pelechano, V., Medina, D. A., Carrasco, F., González, R., Andrés-León, E., Steinmetz, L., Warringer, J., Chávez, S., and Pérez-Ortín, J. E. (2016). The cellular growth rate controls overall mRNA turnover, and modulates either transcription or degradation rates of particular gene regulons. *Nucleic Acids Research*, 44(8):3643–3658.

Ge, H., Wu, P., Qian, H., and Xie, X. S. (2018). Relatively slow stochastic gene-state switching in the presence of positive feedback significantly broadens the region of bimodality through stabilizing the uninduced phenotypic state. *PLOS COMPUTATIONAL BIOLOGY*, 14(3).

Gilchrist, W. (2000). *Statistical Modelling with Quantile Functions*. Chapman and Hall/CRC.

Glazier, D. S. (2018). Rediscovering and Reviving Old Observations and Explanations of Metabolic Scaling in Living Systems. *Systems*, 6(1).

Golding, I. and Cox, E. C. (2004). RNA dynamics in live Escherichia coli cells. *Proceedings of the National Academy of Sciences*, 101(31):11310–11315.

Golding, I., Paulsson, J., Zawilski, S. M., and Cox, E. C. (2005). Real-Time Kinetics of Gene Activity in Individual Bacteria. *Cell*, 123(6):1025–1036.

Golkaram, M., Hellander, S., Drawert, B., and Petzold, L. R. (2016). Macromolecular Crowding Regulates the Gene Expression Profile by Limiting Diffusion. *PLOS Computational Biology*, 12(11):e1005122.

Gookin, S., Min, M., Phadke, H., Chung, M., Moser, J., Miller, I., Carter, D., and Spencer, S. L. (2017). A map of protein dynamics during cell-cycle progression and cell-cycle exit. *PLoS Biology*, 15(9):1–25.

Grant, G. D., Kedziora, K. M., Limas, J. C., Cook, J. G., and Purvis, J. E. (2018). Accurate delineation of cell cycle phase transitions in living cells with PIP-FUCCI. *Cell Cycle*, 17(21-22):2496–2516.

Groisman, A., Lobo, C., Cho, H. J., Campbell, J. K., Dufour, Y. S., Stevens, A. M., and Levchenko, A. (2005). A microfluidic chemostat for experiments with bacterial and yeast cells. *Nature Methods*, 2(9):685–689.

Gut, G., Tadmor, M. D., Pe'Er, D., Pelkmans, L., and Liberali, P. (2015). Trajectories of cell-cycle progression from fixed cell populations. *Nature Methods*, 12(10):951–954.

Hashimoto, M., Nozoe, T., Nakaoka, H., Okura, R., Akiyoshi, S., Kaneko, K., Kussell, E., and Wakamoto, Y. (2016). Noise-driven growth rate gain in clonal cellular populations. *Proceedings of the National Academy of Sciences*, 113(12):3251–3256.

Hausnerová, V. V. and Lanctôt, C. (2017). Transcriptional Output Transiently Spikes Upon Mitotic Exit. *Scientific Reports*, 7(1):12607.

Hausser, J., Mayo, A., Keren, L., and Alon, U. (2019). Central dogma rates and the trade-off between precision and economy in gene expression. *Nature Communications*, 10(1).

Hawley, T. S. and Hawley, R. G., editors (2018). *Flow Cytometry Protocols*, volume 1678 of *Methods in Molecular Biology*. Springer New York, New York, NY.

Hazen, A. L., Bushnell, T., and Haviland, D. L. (2018). The importance of area scaling with FACS DIVA software. *Methods*, 134-135:130–135.

Hebenstreit, D. (2013). Are gene loops the cause of transcriptional noise? *Trends in Genetics*, 29(6):333–338.

Herzog, V. A., Reichholf, B., Neumann, T., Rescheneder, P., Bhat, P., Burkard, T. R., Wlotzka, W., von Haeseler, A., Zuber, J., and Ameres, S. L. (2017). Thiol-linked alkylation of RNA to assess expression dynamics. *Nature Methods*, 14(12):1198–1204.

Hong, L. J., Nelson, B. L., and Xu, J. (2015). *Handbook of Simulation Optimization*, volume 216 of *International Series in Operations Research & Management Science*. Springer New York, New York, NY.

Hood, S. and Amir, S. (2017). The aging clock: circadian rhythms and later life. *The Journal of clinical investigation*, 127(2):437–446.

Hughes, S. M. and Salinas, P. C. (1999). Control of muscle fibre and motoneuron diversification. *CURRENT OPINION IN NEUROBIOLOGY*, 9(1):54–64.

Huh, D. and Paulsson, J. (2011a). Non-genetic heterogeneity from stochastic partitioning at cell division. *Nature Genetics*, 43(2):95–100.

Huh, D. and Paulsson, J. (2011b). Random partitioning of molecules at cell division. *Proceedings of the National Academy of Sciences*, 108(36):15004–15009.

Ietswaart, R., Rosa, S., Wu, Z., Dean, C., and Howard, M. (2017). Cell-Size-Dependent Transcription of FLC and Its Antisense Long Non-coding RNA COOLAIR Explain Cell-to-Cell Expression Variation. *Cell Systems*, 4(6):622–635.e9.

Jacobberger, J. W., Avva, J., Sreenath, S. N., Weis, M. C., and Stefan, T. (2012). Dynamic Epitope Expression from Static Cytometry Data: Principles and Reproducibility. *PLoS ONE*, 7(2):e30870.

Jao, C. Y. and Salic, A. (2008). Exploring RNA transcription and turnover in vivo by using click chemistry. *Proceedings of the National Academy of Sciences*, 105(41):15779–15784.

Ji, N., Middelkoop, T. C., Mentink, R. A., Betist, M. C., Tonegawa, S., Mooijman, D., Korswagen, H. C., and van Oudenaarden, A. (2013). Feedback Control of Gene Expression Variability in the Caenorhabditis elegans Wnt Pathway. *CELL*, 155(4):869–880.

Jia, D., Jolly, M. K., Kulkarni, P., and Levine, H. (2017). Phenotypic Plasticity and Cell Fate Decisions in Cancer: Insights from Dynamical Systems Theory. *Cancers*, 9(7).

Johnston Jr., R. J. and Desplan, C. (2010). Stochastic Mechanisms of Cell Fate

Specification that Yield Random or Robust Outcomes. In Schekman, R and Goldstein, L and Lehmann, R., editor, *ANNUAL REVIEW OF CELL AND DEVELOPMENTAL BIOLOGY, VOL 26*, volume 26 of *Annual Review of Cell and Developmental Biology*, pages 689–719. ANNUAL REVIEWS, 4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. *Journal of Global Optimization*, 13(4):455–492.

Jürges, C., Dölken, L., and Erhard, F. (2018). Dissecting newly transcribed and old RNA using GRAND-SLAM. *Bioinformatics*, 34(13):i218–i226.

Kafri, R., Levy, J., Ginzberg, M. B., Oh, S., Lahav, G., and Kirschner, M. W. (2013). Dynamics extracted from fixed cells reveal feedback linking cell growth to cell cycle. *Nature*, 494(7438):480–483.

Kempe, H., Schwabe, A., Cremazy, F., Verschure, P. J., and Bruggeman, F. J. (2015). The volumes and transcript counts of single cells reveal concentration homeostasis and capture biological noise. *Molecular Biology of the Cell*, 26(4):797–804.

Kiviet, D. J., Nghe, P., Walker, N., Boulineau, S., Sunderlikova, V., and Tans, S. J. (2014a). Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514(7522):376–379.

Kiviet, D. J., Nghe, P., Walker, N., Boulineau, S., Sunderlikova, V., and Tans, S. J. (2014b). Stochasticity of metabolism and growth at the single-cell level. *Nature*, 514:376.

Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A., and Kirschner, M. W. (2015). Droplet Barcoding for Single-Cell Transcriptomics Applied to Embryonic Stem Cells. *Cell*, 161(5):1187–1201.

Kollmann, M., Lovdok, L., Bartholome, K., Timmer, J., and Sourjik, V. (2005). Design principles of a bacterial signalling network. *NATURE*, 438(7067):504–507.

Kuang, C.-y., Yu, Y., Wang, K., Qian, D.-h., Den, M.-y., and Huang, L. (2012). Knockdown of Transient Receptor Potential Canonical-1 Reduces the Proliferation and Migration of Endothelial Progenitor Cells. *Stem Cells and Development*, 21(3):487–496.

Larsen, J. K., Jensen, P. Ø., and Larsen, J. (2001). Flow cytometric analysis of RNA synthesis by detection of bromouridine incorporation. *Current Protocols in Cytometry*, Chapter 7:Unit 7.12.

Lehner, B. (2008). Selection to minimise noise in living systems and its implications for the evolution of gene expression. *MOLECULAR SYSTEMS BIOLOGY*, 4.

Lenstra, T. L., Rodriguez, J., Chen, H., and Larson, D. R. (2016). Transcription Dynamics in Living Cells. *Annual Review of Biophysics*, 45(1):25–47.

Li, B., Zhao, H., Rybak, P., Dobrucki, J. W., Darzynkiewicz, Z., and Kimmel, M. (2014). Different rates of DNA replication at early versus late S-phase sections: Multiscale modeling of stochastic events related to DNA content/EdU (5-ethynyl-2 deoxyuridine) incorporation distributions. *Cytometry Part A*, 85(9):785–797.

Liu, S., Ginzberg, M. B., Patel, N., Hild, M., Leung, B., Li, Z., Chen, Y.-c., Chang, N., Wang, Y., Tan, C., Diena, S., Trimble, W., Wasserman, L., Jenkins, J. L., Kirschner, M. W., and Kafri, R. (2018). Size uniformity of animal cells is actively maintained by a p38 MAPK-dependent regulation of G1-length. *eLife*, 7:1–27.

Losick, R. and Desplan, C. (2008). Stochasticity and cell fate. *SCIENCE*, 320(5872):65–68.

Lu, Y., Biancotto, A., Cheung, F., Remmers, E., Shah, N., McCoy, J. P., and Tsang, J. S. (2016). Systematic Analysis of Cell-to-Cell Expression Variation of T Lymphocytes in a Human Cohort Identifies Aging and Genetic Associations. *IMMUNITY*, 45(5):1162–1175.

Maamar, H., Raj, A., and Dubnau, D. (2007). Noise in gene expression determines cell fate in Bacillus subtilis. *SCIENCE*, 317(5837):526–529.

Maheshri, N. and O 'shea, E. K. (2007). Living with Noisy Genes: How Cells Function Reliably with Inherent Variability in Gene Expression. *Annu. Rev. Biophys. Biomol. Struct*, 36:413–34.

Mair, F., Hartmann, F. J., Mrdjen, D., Tosevski, V., Krieg, C., and Becher, B. (2016). The end of gating? An introduction to automated analysis of high dimensional cytometry data. *European Journal of Immunology*, 46(1):34–43.

Marciano, R., Leprivier, G., and Rotblat, B. (2018). Puromycin labeling does not allow protein synthesis to be measured in energy-starved cells. *Cell Death & Disease*, 9(2):39.

McHugh, M. L. (2013). The Chi-square test of independence Lessons in biostatistics. *Biochemia Medica*, 23(2):143–9.

Meng, X., Firczuk, H., Pietroni, P., Westbrook, R., Dacheux, E., Mendes, P., and McCarthy, J. E. (2017). Minimum-noise production of translation factor eIF4G maps to a mechanistically determined optimal rate control window for protein synthesis. *Nucleic Acids Research*, 45(2):1015–1025.

Mettetal, J. T., Muzzey, D., Pedraza, J. M., Ozbudak, E. M., and van Oudenaarden, A. (2006). Predicting stochastic gene expression dynamics in single cells. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA*, 103(19):7304–7309.

Miettinen, T. P. and Björklund, M. (2016). Cellular Allometry of Mitochondrial

Functionality Establishes the Optimal Cell Size. *Developmental Cell*, 39(3):370–382.

Miettinen, T. P., Kang, J. H., Yang, L. F., and Manalis, S. R. (2019). Mammalian cell growth dynamics in mitosis. *eLife*, 8:1–29.

Mugler, A., Kittisopikul, M., Hayden, L., Liu, J., Wiggins, C. H., Sueel, G. M., and Walczak, A. M. (2016). Noise Expands the Response Range of the Bacillus subtilis Competence Circuit. *PLOS COMPUTATIONAL BIOLOGY*, 12(3).

Nabbi, A. and Riabowol, K. (2015). Rapid isolation of nuclei from cells in vitro. *Cold Spring Harbor Protocols*, 2015(8):769–772.

Nachman, I., Regev, A., and Ramanathan, S. (2007). Dissecting timing variability in yeast meiosis. *CELL*, 131(3):544–556.

Nathans, J. (1999). The evolution and physiology of human color vision: Insights from molecular genetic studies of visual pigments. *NEURON*, 24(2):299–312.

Neurohr, G. E., Terry, R. L., Lengefeld, J., Bonney, M., Brittingham, G. P., Moretto, F., Miettinen, T. P., Vaites, L. P., Soares, L. M., Paulo, J. A., Harper, J. W., Buratowski, S., Manalis, S., van Werven, F. J., Holt, L. J., and Amon, A. (2019). Excessive Cell Growth Causes Cytoplasm Dilution And Contributes to Senescence. *Cell*, 176(5):1083–1097.e18.

Newman, J. R. S., Ghaemmaghami, S., Ihmels, J., Breslow, D. K., Noble, M., DeRisi, J. L., and Weissman, J. S. (2006). Single-cell proteomic analysis of S. cerevisiae reveals the architecture of biological noise TL - 441. *Nature*, 441 VN -(7095):840–846.

Nikolic, N., Schreiber, F., Dal Co, A., Kiviet, D. J., Bergmiller, T., Littmann, S., Kuypers, M. M. M., and Ackermann, M. (2017). Cell-to-cell variation and specialization in sugar metabolism in clonal bacterial populations. *PLOS GENETICS*, 13(12).

Ozbudak, E. M., Thattai, M., Lim, H. N., Shraiman, B. I., and van Oudenaarden,

A. (2004). Multistability in the lactose utilization network of Escherichia coli. *NATURE*, 427(6976):737–740.

Padovan-Merhar, O., Nair, G. P., Biaesch, A. G., Mayer, A., Scarfone, S., Foley, S. W., Wu, A. R., Churchman, L. S., Singh, A., and Raj, A. (2015). Single Mammalian Cells Compensate for Differences in Cellular Volume and DNA Copy Number through Independent Global Transcriptional Mechanisms. *Molecular Cell*, 58(2):339–352.

Paek, A. L., Liu, J. C., Loewer, A., Forrester, W. C., and Lahav, G. (2016). Cell-to-Cell Variation in p53 Dynamics Leads to Fractional Killing. *CELL*, 165(3):631–642.

Paldi, A. (2003). Stochastic gene expression during cell differentiation: order from disorder? *Cellular and Molecular Life Sciences CMLS*, 60(9):1775–1778.

Paulsson, J. (2005). Models of stochastic gene expression. *Physics of Life Reviews*, 2:157–175.

Peccoud, J. and Ycart, B. (1995). Markovian Modeling of Gene-Product Synthesis.

Pérez-Ortín, J. E., Tordera, V., and Chávez, S. (2019a). Homeostasis in the Central Dogma of Molecular Biology. *bioRxiv*, page 599050.

Pérez-Ortín, J. E., Tordera, V., and Chávez, S. (2019b). Homeostasis in the Central Dogma of molecular biology: the importance of mRNA instability. *RNA Biology*, 16(12):1659–1666.

Pfeiffer, S. E. and Tolmach, L. J. (1968). RNA synthesis in synchronously growing populations of HeLa S3 cells. I. Rate of total RNA synthesis and its relationship to DNA synthesis. *Journal of Cellular Physiology*, 71(1):77–93.

Pires, J. C. and Conant, G. C. (2016). Robust Yet Fragile: Expression Noise, Protein Misfolding, and Gene Dosage in the Evolution of Genomes. 50:113–131.

Rabani, M., Levin, J. Z., Fan, L., Adiconis, X., Raychowdhury, R., Garber, M., Gnirke, A., Nusbaum, C., Hacohen, N., Friedman, N., Amit, I., and Regev,

A. (2011). [SUPP] Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells. *Nature biotechnology*, 29(5):436–442.

Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y., and Tyagi, S. (2006). Stochastic mRNA synthesis in mammalian cells. *PLoS Biology*, 4(10):1707–1719.

Raj, A., Rifkin, S. A., Andersen, E., and van Oudenaarden, A. (2010). Variability in gene expression underlies incomplete penetrance. *NATURE*, 463(7283):913–U84.

Rao, C. V., Wolf, D. M., and Arkin, A. P. (2002). Control, exploitation and tolerance of intracellular noise. *Nature*, 420(6912):231–237.

Raser, J. M. and Shea, E. K. O. (2006). Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 304(5678):1811–1814.

Richard, A., Boullu, L., Herbach, U., Bonnafoux, A., Morin, V., Vallin, E., Guillemin, A., Gao, N. P., Gunawan, R., Cosette, J., Arnaud, O., Kupiec, J.-J., Espinasse, T., Gonin-Giraud, S., and Gandrillon, O. (2016). Single-Cell-Based Analysis Highlights a Surge in Cell-to-Cell Molecular Variability Preceding Irreversible Commitment in a Differentiation Process. *PLOS BIOLOGY*, 14(12).

Roorda, A. and Williams, D. R. (1999). The arrangement of the three cone classes in the living human eye. *NATURE*, 397(6719):520–522.

Russo, J., Heck, A. M., Wilusz, J., and Wilusz, C. J. (2017). Metabolic labeling and recovery of nascent RNA to accurately quantify mRNA stability. *Methods*, 120:39–48.

Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). *Computational flow cytometry: Helping to make sense of high-dimensional immunology data.* PhD thesis.

Saitou, T. and Imamura, T. (2016). Quantitative imaging with Fucci and mathematics to uncover temporal dynamics of cell cycle progression. *Development, Growth & Differentiation*, 58(1):6–15.

Sakaue-Sawano, A., Kurokawa, H., Morimura, T., Hanyu, A., Hama, H., Osawa, H., Kashiwagi, S., Fukami, K., Miyata, T., Miyoshi, H., Imamura, T., Ogawa, M., Masai, H., and Miyawaki, A. (2008). Visualizing Spatiotemporal Dynamics of Multicellular Cell-Cycle Progression. *Cell*, 132(3):487–498.

Schmidt, E. E. and Schibler, U. (1995). Cell size regulation, a mechanism that controls cellular RNA accumulation: Consequences on regulation of the ubiquitous transcription factors Oct1 and NF-Y, and the liver-enriched transcription factor DBP. *Journal of Cell Biology*, 128(4):467–483.

Schofield, J. A., Duffy, E. E., Kiefer, L., Sullivan, M. C., and Simon, M. D. (2018). TimeLapse-seq: Adding a temporal dimension to RNA sequencing through nucleoside recoding. *Nature Methods*, 15(3):221–225.

Scholtissek, C. (1967). NUCLEOTIDE METABOLISM IN TISSUE CULTURE CELLS AT LOW. 45:228–237.

Schrom, E. C. and Graham, A. L. (2017). Instructed subsets or agile swarms: how T-helper cells may adaptively counter uncertainty with variability and plasticity. *CURRENT OPINION IN GENETICS & DEVELOPMENT*, 47:75–82.

Schuh, L., Saint-Antoine, M., Sanford, E., Emert, B. L., Singh, A., Marr, C., Goyal, Y., and Raj, A. (2019). Gene networks with transcriptional bursting recapitulate rare transient coordinated expression states in cancer. *bioRxiv*.

Sensky, T. E. and Rees, K. R. (1976). Effects of low temperature on RNA metabolism in different cell lines. *Exprl Cell Res*, (Cvi):200–204.

Setty, M., Tadmor, M. D., Reich-Zeliger, S., Angel, O., Salame, T. M., Kathail, P., Choi, K., Bendall, S., Friedman, N., and Pe'Er, D. (2016). Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nature Biotechnology*, 34(6):637–645.

Shaffer, S. M., Dunagin, M. C., Torborg, S. R., Torre, E. A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P. A., Xiao, M., Ggan, E. E., Anastopoulos, I. N., Vargas-Garcia, C. A., Singh, A., Nathanson, K. L., Herlyn, M., and Raj, A. (2017). Rare cell variability and drug- induced reprogramming as a mode of cancer drug resistance. *NATURE*, 546(7658):431+.

Shah, N. M., Groves, A. K., and Anderson, D. J. (1996). Alternative neural crest cell fates are instructively promoted by TGF beta superfamily members. *CELL*, 85(3):331–343.

Shahrezaei, V. and Marguerat, S. (2015). Connecting growth with gene expression: Of noise and numbers. *Current Opinion in Microbiology*, 25:127–135.

Sherwood, S. W., Rush, D. F., Kung, A. L., and Schimke, R. T. (1994). Cyclin B1 expression in hela S3 cells studied by flow cytometry.

Simpson, M. L., Cox, C. D., Allen, M. S., Mccollum, J. M., Dar, R. D., Karig, D. K., and Cooke, J. F. (2009). Noise in biological circuits.

Skinner, S. O., Xu, H., Nagarkar-Jaiswal, S., Freire, P. R., Zwaka, T. P., and Golding, I. (2016). Bursting through the cell cycle. *eLife*, 5.

Snijder, B. and Pelkmans, L. (2011). Origins of regulated cell-to-cell variability. *Nature reviews. Molecular cell biology*, 12(2):119–125.

Soltani, M., Vargas-Garcia, C. A., Antunes, D., and Singh, A. (2016). Intercellular Variability in Protein Levels from Stochastic Expression and Noisy Cell Cycle Processes. *PLoS Computational Biology*, 12(8).

Son, S., Tzur, A., Weng, Y., Jorgensen, P., Kim, J., Kirschner, M. W., and Manalis, S. R. (2012). Direct observation of mammalian cell growth and size regulation. *Nature Methods*, 9(9):910–912.

Stapel, L. C., Zechner, C., and Vastenhouw, N. L. (2017). Uniform gene expression in embryos is achieved by temporal averaging of transcription noise. *GENES & DEVELOPMENT*, 31(16):1635–1640.

Sun, Q., Jiao, F., Lin, G., Yu, J., and Tang, M. (2019a). The nonlinear dynamics and fluctuations of mRNA levels in cell cycle coupled transcription. *PLOS Computational Biology*, 15(4):1–27.

Sun, X.-M., Bowman, A., Priestman, M., Bertaux, F., Martinez-Segura, A., Tang, W., Dormann, D., Shahrezaei, V., and Marguerat, S. B. (2019b). Size-dependent increase in RNA Polymerase II initiation rates mediates gene expression scaling with cell size. *bioRxiv Molecular Biology*.

Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002). Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(20):12795–12800.

Thattai, M. and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(15):8614–9.

Thattai, M. and van Oudenaarden, A. (2004). Stochastic gene expression in fluctuating environments. *GENETICS*, 167(1):523–530.

Timmers, H. T. M. and Tora, L. (2018). Transcript Buffering: A Balancing Act between mRNA Synthesis and mRNA Degradation. *Molecular Cell*, 72(1):10–17.

Tirosh, I., Weinberger, A., Carmi, M., and Barkai, N. (2006). A genetic signature of interspecies variations in gene expression. *NATURE GENETICS*, 38(7):830–834.

To, T.-L. and Maheshri, N. (2010). Noise Can Induce Bimodality in Positive Transcriptional Feedback Loops Without Bistability. *SCIENCE*, 327(5969):1142–1145.

Tonn, M. K., Thomas, P., Barahona, M., and Oyarzun, D. A.

(2019). Stochastic modelling reveals mechanisms of metabolic heterogeneity. *COMMUNICATIONS BIOLOGY*, 2.

Tsang, J., Zhu, J., and Oudenaarden], A. v. (2007). MicroRNA-Mediated Feedback and Feedforward Loops Are Recurrent Network Motifs in Mammals. *Molecular Cell*, 26(5):753–767.

Tsuboi, A., Yoshihara, S., Yamazaki, N., Kasai, H., Asai-Tsuboi, H., Komatsu, M., Serizawa, S., Ishii, T., Matsuda, Y., Nagawa, F., and Sakano, H. (1999). Olfactory neurons expressing closely linked and homologous odorant receptor genes tend to project their axons to neighboring glomeruli on the olfactory bulb. *JOURNAL OF NEUROSCIENCE*, 19(19):8409–8418.

Tzur, A., Kafri, R., LeBleu, V. S., Lahav, G., and Kirschner, M. W. (2009). Cell Growth and Size Homeostasis in Proliferating Animal Cells. *Science*, 325(5937):167–171.

Tzur, A., Moore, J. K., Jorgensen, P., Shapiro, H. M., and Kirschner, M. W. (2011). Optimizing optical flow cytometry for cell volume-based sorting and analysis. *PLoS ONE*, 6(1):1–9.

Uvarovskii, A., Naarmann-de Vries, I. S., and Dieterich, C. (2019). On the optimal design of metabolic rna labeling experiments. *PLOS Computational Biology*, 15(8):1–22.

van Dyken, J. D. (2017). Noise slows the rate of Michaelis-Menten reactions. *JOURNAL OF THEORETICAL BIOLOGY*, 430:21–31.

Voichek, Y., Bar-Ziv, R., and Barkai, N. (2016a). A role for Rtt109 in buffering gene-dosage imbalance during DNA replication. *Nucleus*, 7(4):00–00.

Voichek, Y., Bar-Ziv, R., and Barkai, N. (2016b). Expression homeostasis during DNA replication. *Science*, 351(6277):1087–1090.

Wagner, A. (2005). Energy constraints on the evolution of gene expression. *MOLECULAR BIOLOGY AND EVOLUTION*, 22(6):1365–1374.

Walter, D., Hoffmann, S., Komseli, E.-S., Rappsilber, J., Gorgoulis, V., and Sørensen, C. S. (2016). SCFCyclin F-dependent degradation of CDC6 suppresses DNA re-replication. *Nature Communications*, 7(1):10530.

Wang, P., Robert, L., Pelletier, J., Dang, W. L., Taddei, F., Wright, A., and Jun, S. (2010). Robust Growth of <em>Escherichia coli</em>. *Current Biology*, 20(12):1099–1103.

Wang, Z. and Zhang, J. (2011). Impact of gene expression noise on organismal fitness and the efficacy of natural selection. *Proceedings of the National Academy of Sciences*, 108(16):E67—-E76.

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., and Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, 116(12):5542–5549.

Weinberger, L. S., Dar, R. D., and Simpson, M. L. (2008). Transient-mediated fate determination in a transcriptional circuit of HIV. *NATURE GENETICS*, 40(4):466–470.

Wernet, M. F., Mazzoni, E. O., Celik, A., Duncan, D. M., Duncan, I., and Desplan, C. (2006). Stochastic spineless expression creates the retinal mosaic for colour vision. *NATURE*, 440(7081):174–180.

Williams, G. H. and Stoeber, K. (2012). The cell cycle and cancer. *The Journal of Pathology*, (October 2011):352–364.

Woods, H. A. (2014). Mosaic physiology from developmental noise: within-organism physiological diversity as an alternative to phenotypic plasticity and phenotypic flexibility. *JOURNAL OF EXPERIMENTAL BIOLOGY*, 217(1, SI):35–45.

Yan, X., Hoek, T. A., Vale, R. D., and Tanenbaum, M. E. (2016). Dynamics of Translation of Single mRNA Molecules in Vivo. *Cell*, 165(4):976–989.

Zhao, X., Luo, C., and Wang, H. (2019). Protein dynamic analysis of the budding yeast sporulation process at the single-cell level in an air-enriched microfluidic device. *INTEGRATIVE BIOLOGY*, 11(3):79–86.

Zopf, C. J., Quinn, K., Zeidman, J., and Maheshri, N. (2013). Cell-Cycle Dependence of Transcription Dominates Noise in Gene Expression. *PLoS Computational Biology*, 9(7).

Zuleta, I. A., Aranda-Díaz, A., Li, H., and El-Samad, H. (2014). Dynamic characterization of growth and gene expression using high-throughput automated flow cytometry. *Nature methods*, 11(4):443–448.