Bayesian Statistical Models of Cell-Cycle Progression at Single-Cell and Population Levels

by

Michael Benjamin Mayhew

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Supervisor

_____
Sayan Mukherjee

_____
Alexander J. Hartemink

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2014

Abstract

Bayesian Statistical Models of Cell-Cycle Progression at Single-Cell and Population Levels

by

Michael Benjamin Mayhew

Department of Statistical Science
Duke University

Date: _____

Approved:

_____
Edwin S. Iversen, Supervisor

_____
Sayan Mukherjee

_____
Alexander J. Hartemink

An abstract of a thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in the Department of Statistical Science in the Graduate School of Duke University
2014

Abstract

Cell division is a biological process fundamental to all life. One aspect of the process that is still under investigation is whether or not cells in a lineage are correlated in their cell-cycle progression. Data on cell-cycle progression is typically acquired either in lineages of single cells or in synchronized cell populations, and each source of data offers complementary information on cell division. To formally assess dependence in cell-cycle progression, I develop a hierarchical statistical model of single-cell measurements and extend a previously proposed model of population cell division in the budding yeast, *Saccharomyces cerevisiae*. Both models capture correlation and cell-to-cell heterogeneity in cell-cycle progression, and parameter inference is carried out in a fully Bayesian manner. The single-cell model is fit to three published time-lapse microscopy datasets and the population-based model is fit to simulated data for which the true model is known. Based on posterior inferences and formal model comparisons, the single-cell analysis demonstrates that budding yeast mother and daughter cells do not appear to correlate in their cell-cycle progression in two of the three experimental settings. In contrast, mother cells grown in a less preferred sugar source, glycerol/ethanol, did correlate in their rate of cell division in two successive cell cycles. Population model fitting to simulated data suggested that, under typical synchrony experimental conditions, population-based measurements of the cell-cycle were not informative for correlation in cell-cycle progression or heterogeneity in daughter-specific G1 phase progression.

I am dedicating this work to my grandfather, Francis Ignatius Stark. He taught me the value of honesty, both with myself and with others, and the importance of hard work which has helped me strive to understand the world and better myself every day.

Contents

List of Tables

List of Figures

List of Abbreviations and Symbols

Symbols

Φ	Standard normal cumulative distribution function.

Abbreviations

MCMC	Markov chain Monte Carlo
JAGS	Just Another Gibbs Sampler
BAR	Bifurcating Autoregressive
GFP	Green fluorescent protein
AR	Autoregressive
BIC	Bayesian Information Criterion
ML	Maximum likelihood
HPD	Highest posterior density
CLOCCS	Characterizing Loss of Cell-Cycle Synchrony

Acknowledgements

I'd like to thank my master's advisor, Professor Ed Iversen, for his unflagging support and guidance. Thanks also go to my doctoral advisor, Professor Alex Hartemink, and Professor Sayan Mukherjee for helping me develop this project and for supporting me in my pursuit of this master's degree in statistics.

I want to thank my wife, Diana Fusco, for her love and for standing by me through the exam preparation, the coursework, and the writing of this thesis. She has been and always will be the love of my life. In addition, I want to thank my family and friends for their love and support over my years of scholarly pursuit.

I'm thankful to the students of the Duke Department of Statistical Science, particularly Monika Hu, Jacopo Soriano, and Brian St. Thomas for their help in preparation for the first-year qualifying examination. I'm also grateful for the dedicated faculty of the Department of Statistical Science, especially Mike West and Merlise Clyde, who have both challenged me to push myself in my statistical training.

Lastly, I'd like to acknowledge the generous financial support I've received for my graduate training and participation in research conferences, particularly from the Defense Advanced Research Project Agency (USA), the National Institutes of Health (USA), the National Science Foundation (USA), the International Society of Computational Biology, the Duke University Program in Computational Biology & Bioinformatics, and the Duke University Graduate School.

1

Introduction

This work advances statistical methods needed to understand the vital process of cell division. In particular, this thesis centers on the question of whether cells (specifically of the budding yeast, *Saccharomyces cerevisiae*) correlate in their cell-cycle progression, and the statistical methods here proposed leverage complementary information from different types of cell-cycle data–those gathered from single cells and from populations of cells–to address this question. Study of cell-to-cell dependence in cell-cycle progression parameters has a long history in both experimental and theoretical research and is vital for understanding coordination of division in multicellular systems (e.g. organs) as well as other biological processes involving cell division.

1.1 Cell Division in the Budding Yeast, *Saccharomyces cerevisiae*

Cell division is a biological process fundamental to the life and proliferation of every living thing. The process underlies other vital and poorly understood processes such as tissue growth and organization, development and differentiation (Neufeld et al. (1998); Hawkins et al. (2009); Pauklin and Vallier (2013)). In addition, greater understanding of cell division will bridge gaps in mechanistic understanding of diseases

such as cancer which originate, at least in part, from dysregulation in cell division. In budding yeast cell division, a cell undergoes a series of regulated events that finally culminate in the partitioning of the cell's replicated genome and other cellular contents into a new daughter cell (Figure 1.1). The process is divided into four phases: an initial period of growth (G1 phase); a period in which the cell's DNA is replicated (S phase); a second period of growth (G2 phase); and finally a period of nuclear and cell division (mitotic or M phase). As the G2 and M phases largely overlap, the two phases are merged and treated as a single period called G2/M phase (Figure 1.1).

Transitions in the process are marked by genetic and morphological changes that facilitate monitoring and analysis of the cell cycle. For example, late in G1, a contractile ring composed of myosin as well as other proteins takes shape at the periphery of the mother cell. This structure, termed the myosin ring, marks the point of emergence of the nascent daughter cell or bud (Bi et al. (1998)). Shortly thereafter, having passed through the point of cell-cycle commitment called START (Hartwell et al. (1974)), the mother cell puts out a bud. Around the same time, replication of the mother's DNA commences. The bud continues to grow during this period of DNA synthesis. Following the completion of DNA replication, one of each of the two copies of the genome are partitioned into the mother and daughter cell during mitosis (Figure 1.1). The myosin ring then disappears at cytokinesis when the cytoplasms of the mother cell and newborn daughter cell pinch off from one another. It is at this point that the daughter can be considered a new cell, distinct from the mother. These different features indicative of cell-cycle state are known as markers of cell-cycle progression.

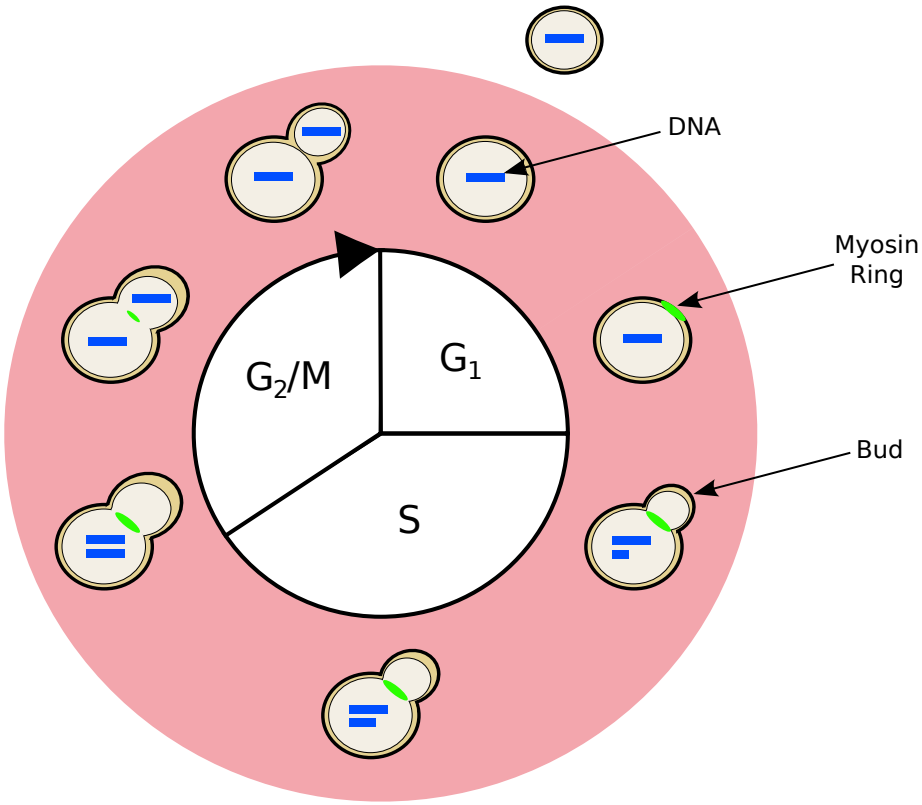

FIGURE 1.1: The process of cell division begins at the top of the diagram and proceeds clockwise. Depicted along the outer ring of the diagram are different morphological and genetic stages of division as reflected by the changing status of different markers of cell-cycle progression (myosin ring, bud, and DNA). The green feature at the neck joining the mother and daughter cell is the myosin ring. The myosin ring appears late in G1 phase, marking the point of emergence of the bud, and disappears with cytokinesis, indicating the separation of the mother and daughter cytoplasms. The bud is the nascent daughter cell that appears at the G1/S boundary and continues to grow over the course of S phase. The DNA (represented by the blue bars) is replicated during S phase and partitioned between the mother and daughter cell at M phase. Subsequent to cell wall separation, the mothers (shown inside the outer ring) and daughters (shown outside the outer ring) are free to undergo more rounds of division (top). In budding yeast, division is asymmetric and daughters are born smaller than their mothers (top). The cells shown are haploid in that they begin their lives with one copy of genomic DNA (single blue bar).

1.2 Population vs. Single-Cell Methods of Cell-Cycle Data Acquisition

1.2.1 Synchrony Experiments for Population-Based Analysis of Cell-Cycle Progression

Observations of markers of cell-cycle progression are acquired at one of two different scales: at the level of a cell population and at the level of single cells. Both methods have their relative strengths and drawbacks. In the former setting, a population of cells growing in liquid media is synchronized at a particular point in the cell cycle. Methods to achieve this synchronization can be mechanical, chemical, or genetic and can position the population at different points in the process. The cell population is subsequently released from synchrony, and the experimenter withdraws samples of the dividing population at intervals from the time of release. The samples are quickly mixed with a fixative to halt cell division. The experimenter then prepares a portion of each time point sample for marker quantitation: the evaluation of the proportion of cells with a particular cell-cycle marker state in the sample. For example, if the chosen marker is the bud, an experimenter would mount a portion of the time point sample on a microscope slide, count some predetermined number of cells, and record how many of the counted cells had buds.

Synchrony experiments are often straightforward to set up and run in terms of the equipment and reagents required to induce synchrony and culture the cell population over time. Also, the number of observations at each time point generally numbers in the several hundreds. However, synchrony experiments have their drawbacks. First and foremost, the observations acquired are independent of one another in time. Thus, estimating correlations in cell-cycle progression between two cells becomes challenging as information about the genealogical relationships between cells in any two time point samples is lost. Second, the observations at each time point are actually sample averages. As such, accuracy in estimation of the true underlying

proportion of budded cells in the entire time point sample depends on the total number of cells counted by the experimenter. Third, experimental constraints restrict the time point sampling frequency to no less than 4-5 minutes. Finally, progressively increasing asynchrony in the population (see below) limits the number of informative observations derived from the experiment.

Achieving perfect synchrony in synchrony experiments is practically impossible, and different sources of asynchrony have been identified that complicate characterization of cell-cycle progression (Orlando et al. (2007)). First, the synchronization procedures used to prepare a population of cells for cell-cycle analysis are not perfect. At best, cells in the population are concentrated near the same approximate point in the cell cycle, but they remain somewhat distributed around this point. Second, individual cells proceed at different rates through the process of cell division. Third, and specific to budding yeast, cells divide asymmetrically, and as a result, daughter cells tend to require more time to complete the G1 phase than mother cells. These different sources of asynchrony imply that observations at each time point are based on mixtures of cells at different stages of the cell cycle rather than a homogeneous population of cells moving in lock-step with one another through the process.

1.2.2 Single-Cell Analysis of Cell Division by Time-Lapse Microscopy

Another method for acquisition of cell-cycle data is time-lapse microscopy. Recent advances in time-lapse microscopy and fluorescence imaging allow for tracking of cells as well as features within cells to visualize and analyze biological processes over time (Tsien (1998); Muzzey and van Oudenaarden (2009) and references therein). In this paradigm, single cells are suspended in some culture medium under a microscope equipped with a time-lapse camera. The camera records images of the same field of view–and thereby the same cells–over time, usually at shorter intervals than in synchrony experiments ($\sim$1-3 minutes). In a cell-cycle setting, the expression of

fluorescent reporter proteins appended to cell-cycle-regulated proteins of interest signifies different cell-cycle events. Since repeated measurements are taken over the lives of single cells and since genealogical relationships are directly observed, time-lapse microscopy permits direct estimation of correlations in the behaviors and phenotypes of related cells.

However, time-lapse microscopy is not without its drawbacks. Equipment and time investments are more considerable compared with those required for synchrony experiments. Also, as cell density in a field of view increases, the ability to track individual cells and cell lineages diminishes. Consequently, the number of cells tracked in a single time-lapse movie is much smaller ($\sim$15-30) than the number of cells monitored in a typical synchrony experiment (several hundreds at each time point). As a result, an experimenter must acquire multiple time-lapse movies to achieve sufficient sample sizes for analysis. Increasing cell density over the course of a time-lapse movie doesn't just limit sample size. The number of cell divisions observed in any given experiment is usually no more than three or four. Thus, it is difficult to observe long-term patterns in cell-cycle progression of individual cells and their progeny. In addition, efforts to characterize the effects of increasing cell density on nutrient availability–say at the interior of the growing microcolony–and thereby on individual rates of growth and division have been limited. Nevertheless, both single-cell and population-level data are rich and complementary sources of information that can shed light on cell-to-cell dependence in cell-cycle progression.

1.3 Experimental Characterization of Cell-Cycle Progression

Studies dating as far back as the late 1950's and early 1960's have revealed heterogeneity in cell-cycle progression among presumably homogeneous cells (Powell (1955); Sisken and Morasca (1965); Kubitschek (1966); Lord and Wheals (1981)). Some of this early experimental work also revealed a dependence in cell-cycle progression

across cells in a lineage (Powell (1955)). These studies used time-lapse microscopy to monitor lineages of dividing bacterial cells and observed both positive, near zero, and negative correlations in the cell-cycle progression of mother and daughter cells. Positive correlations in the cell-cycle progression of sister cells (cells born of the same mother) were also observed. Recent experimental studies have also suggested inter-cell dependence in cell-cycle progression of lymphocytes (Hawkins et al. (2009)) and mouse embryos (Balbach et al. (2012)). In budding yeast cells, classical studies suggested the coordination of successive cell cycles as cells arrested at nuclear division did not emit new buds or initiate new rounds of DNA synthesis in subsequent cycles (Hartwell et al. (1974); Strathern et al. (1981)). However, comparatively little work has been done in budding yeast to determine if and the extent to which measures of cell-cycle progression are correlated between cells in a lineage.

1.4 Model-Based Characterization of Cell-Cycle Progression

Stochastic models that take into account correlations in the properties and cell-cycle dynamics of individual cells have existed since the late 1960's (Bell and Anderson (1967); Smith and Martin (1973); Lebowitz and Rubinow (1974); Cooper (1982); Rigney (1987); Webb (1987); Hejblum et al. (1988)). Two favored early models of symmetric cell division were based on complementary descriptions of cell-cycle progression and differed mainly in their specification of the G1 phase duration: these were Cooper's continuum model (Cooper (1982)) and the transition probability model of Smith and Martin (Smith and Martin (1973)). In the continuum model, G1 variability was the result of variability in the rate of production of an initiator, an unspecified molecular component responsible for the initiation of DNA synthesis. In this model, four major parameters determined cell-cycle duration: the rate of initiator synthesis; a threshold on the rate of synthesis needed for initiation of DNA replication; the duration of a replication-segregation phase (effectively S, G2,

and M phases combined); and the proportion of a mother's initiator allocated to the daughter cells. The model is deterministic though simulations were carried out with Gaussian noise added to each parameter to represent heterogeneity. In simulation studies, the continuum model predicted positive sister-sister correlation. In the transition probability model, some time after division a newly born cell entered a state (A) of indeterminate length with a probability of exit that remained constant over the course of the period. This A state was effectively the G1 phase. Once out of the A state, a cell entered a deterministic phase (B) dedicated to DNA replication and mitotic division (representing the combined durations of S, G2, and M phases). As such, this model assumed that all variability in cell-cycle duration was due to variability in G1 phase duration. In simulations with both the transition probability model and the continuum model, mother-daughter correlations were predicted to be either negative or close to 0.

Dependence in cell-cycle progression was explicitly modeled in dynamical representations of cell populations built on differential equations (Lebowitz and Rubinow (1974); Rigney (1987); Webb (1987); Hejblum et al. (1988)). Like experimental observations of the day, model-based estimates of correlations between sister cells and mothers and daughters varied with organism and experimental system (Hejblum et al. (1988) and references therein). Over time, these models grew in complexity to account for variability in cell-cycle progression due to cell age and growth. Lebowitz and Rubinow considered the effects of correlation between mother and daughter cells on the asymptotic behavior and age distribution of a population of proliferating bacterial cells (Lebowitz and Rubinow (1974)). Rigney's ground-breaking work proposed a model capturing heterogeneity in cell division times of mammalian cells and relating correlation in cell division times to a cell growth-based mechanism. Fitting of these models to data generally involved qualitative comparisons of model predictions with observed dynamical behaviors of the cell populations. While these structured

population dynamics models also found their way into studies of *Saccharomyces cerevisiae* (Vanoni et al. (1983); Tyson and Hannsgen (1986)), correlation in cell-cycle progression was not the focal point of the analysis.

Statistical models have also been developed to characterize cell-cycle progression from single-cell measurements on lineages. One such family of models, the bifurcating autoregressive or BAR models, was initially developed to characterize dependence in general cellular properties–including cell-cycle progression–observed from lineages of dividing cells (Cowan and Staudte (1986)). The model was proposed as an extension of the autoregressive (AR) models often found in time series analysis. The original BAR model is based on four main parameters: a mean (μ) and variance (σ^2) for cell-cycle duration as well as parameters for conditional sister-sister (ϕ, given the common mother) and marginal mother-daughter correlation (θ). Cell-cycle durations are assumed to be multivariate normal-distributed and sister cells are assumed to be marginally correlated with parameter $\rho = \theta^2 + (1 - \theta^2)\phi$. The authors used classical estimation techniques (e.g. maximum likelihood (ML)) for inference and the structure of the model lends itself to borrowing of information across replicate lineages. The BAR model has been successfully applied to lineages of symmetrically dividing bacterial and Chinese hamster cells (Staudte et al. (1996); Huggins and Basawa (1999)). Extensions of the BAR model have been proposed allowing for heterogeneity in the mean cell-cycle duration of each lineage tree, batch effects among different subsets of lineage trees, and long-range (multi-generational) cell-cycle dependencies (Staudte et al. (1996); Huggins and Basawa (1999); da Saporta et al. (2011)). As yet, neither BAR models nor similar statistical models for cell lineages have been applied to budding yeast single-cell lineages to analyze correlation in cell-cycle progression.

However, statistical methods have been developed and applied with great success to population measurements of cell division in budding yeast. One such model of population division called CLOCCS (Characterizing Loss of Cell Cycle Synchrony)

was developed to characterize budding yeast cell-cycle progression while accounting for the different sources of asynchrony inherent in synchrony experiments (Orlando et al. (2007); Orlando et al. (2009); Mayhew et al. (2011)). More specifically, the model captures initial heterogeneity in cell-cycle position, cell-to-cell heterogeneity in cell-cycle velocity, and differences in expected cell-cycle duration between mother and daughter cells. The model is built on a branching process representation of cell division in which the branch lengths represent the expected cell-cycle durations of budding yeast cells. The model has been extended to fit a variety of different data types including image-derived binary marker observations and flow cytometric measurements of DNA content (Orlando et al. (2009); Mayhew et al. (2011)).

1.5 Organization of Thesis

The central aim of this work is to advance our understanding of the dynamics of budding yeast cell-cycle division by analyzing dependence in cell-cycle progression. I address this aim by developing complementary approaches to characterize the process using both single-cell and population-based measurements. In Chapter 2, I introduce a hierarchical model of cell division. The model captures cell-to-cell heterogeneity as well as correlations in cell-cycle progression and is applied to previously published lineage data. The hierarchical formulation of the model allows for borrowing of information across replicate lineages, and fully Bayesian parameter inference is carried out. I discuss results of model fitting to three different lineage datasets comprising two different genetic backgrounds and two different nutrient conditions. In Chapter 3, I draw on results from the single-cell analysis of Chapter 2 to extend the CLOCCS model of population division. The model extensions allow for correlations in cell-cycle durations as well as variation in daughter-specific G1 cell-cycle progression. I fit the model by Markov chain Monte Carlo (MCMC), and test the estimation properties of the extended model on simulated data for which the true model is known. Finally,

in Chapter 4, I summarize the major findings from Chapters 2 and 3 and discuss potential future directions for this work in the larger context of computational and statistical systems biology.

2

A Hierarchical Model of Budding Yeast Cell Division at Single-Cell Level

In this chapter, I develop a hierarchical statistical model of budding yeast cell division at the single-cell level to characterize dependence in cell-cycle progression. Encoded within the model hierarchy is an asymmetric autoregressive process describing the generation of a lineage tree. The model reflects biological characteristics of the budding yeast: the model formally describes the asymmetric division typical of budding yeast. This asymmetry is captured with parameters representing differential correlation structure in cell-cycle progression between parent cells and their progeny and extended G1 phases for daughter cells. Parameter inference is fully Bayesian allowing for the incorporation of prior information about budding yeast cell division. The model is fit to three recently published time microscopy datasets consisting of cell division measurements from multiple lineage trees and comprising different genetic and nutrient environment conditions. Parameter inferences are discussed under the three different experimental settings and different models of cell-cycle progression are formally compared in a Bayes factor analysis.

2.1 Single-Cell Measurements of *Saccharomyces cerevisiae* Division

Single-cell haploid yeast data was acquired for 26 wild-type lineages (213 cells) grown in glucose, 19 6xCLN3 lineages (99 cells) grown in the preferred sugar source glucose, and 21 lineages of 157 wild-type cells grown in the less preferred sugar glycerol/ethanol all provided by the authors of Di Talia et al. (2007). The data were derived from time-lapse microscopy experiments in which images of cells grown in glucose were taken every 3 minutes while images of cells growing in glycerol/ethanol were taken every 6 minutes. A single yeast cell (the founder cell) growing on an agar plate was identified at the outset of the time-lapse experiment. The times of occurrence of two landmark cell-cycle events were recorded for each yeast cell on the plate: the appearance and disappearance of the myosin ring (see Figures 1.1 and 2.1). Myosin ring appearance and disappearance times are hereafter referred to as budding and cycle times, respectively. The myosin ring is a contractile structure that appears late in G1 phase just prior to the appearance of the bud (Bi et al. (1998)), and was visualized by tagging Myo1p with green fluorescent protein (GFP) (Figure 1.1). The disappearance of the myosin ring marks the end of cytokinesis and hence the separation of the shared cytoplasm of the mother cell into mother and daughter cytoplasms (Figure 1.1). These cells and their progeny were subsequently monitored for additional budding and division times until cell density prevented accurate measurements (Figure 2.1).

In each lineage, the time at which the founder cell began its cell cycle was not known, and so the cell-cycle progression of the founder cell could not be determined. Thus, each lineage was divided into two sub-lineages. The first sub-lineage had as an initial cell the founder cell in its second observed cell cycle. The second sub-lineage took the founder cell's first daughter as its initial cell. Analysis was carried out on these sub-lineages for each of the three datasets. In contrast to the bifurcating

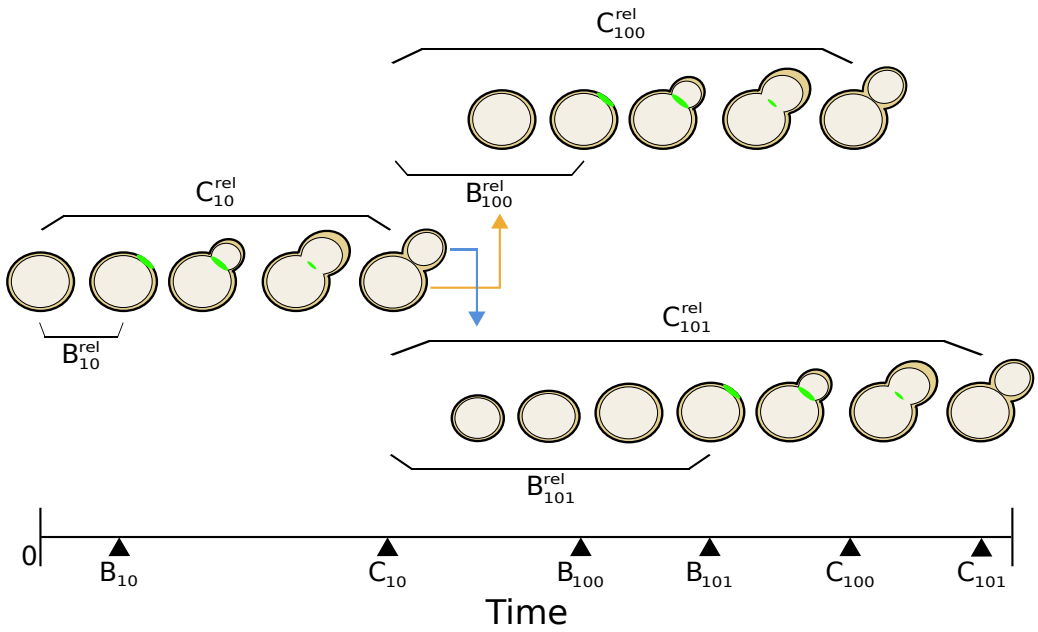

FIGURE 2.1: Here, a mother cell (cell 10) of the lineage proceeds through the cell cycle, undergoing budding and division. Once divided from her daughter (cell 101; follow blue arrow), the mother (now cell 100; follow orange arrow) undergoes another round of division. As budding yeast divide asymmetrically, the daughter cell (101) is born smaller than the mother (100). This small birth size contributes to a longer G1 phase. The budding (B_{10}) and division (D_{10}) events for each cell were originally observed in absolute time (shown on timeline at base of figure). In our analysis, we transformed the absolute measurements to relative measurements or durations of budding (B^{rel}_{10}) and division (D^{rel}_{10}).

autoregressive (BAR) model (Cowan and Staudte (1986)), correlation between the initial cells of each sub-lineage (sister cells) is not modeled explicitly in the analysis that follows.

I adopted the binary indexing scheme of Di Talia et al. (2007) to refer to cell division characteristics specific to each individual cell cycle. The founder cell in each pedigree was considered cell 1. After undergoing its first division, the founder cell became mother origin cell 10. The newly divided daughter cell was labeled cell 11. In this way, a 0 was appended to the label of a cell for each division subsequent to its first cell cycle as a daughter cell. So, aside from the lineage founder cell, only newborn daughter cells took labels ending in 1 (Figure 2.1). In adopting this numbering

scheme, cells were divided into four different categories of cells: mother origins, daughter origins, mother offspring, and daughter offspring. Daughter offspring cells are cells in their first daughter cell cycles (excluding daughter origin cells). Mother offspring cells are cells undergoing at least their second cell cycle.

2.2 Modeling Budding Yeast Cell Division at the Single-Cell Level

Time-lapse measurements of cell-cycle progression (as in Figure 2.1) can be viewed as noisy observations of an underlying branching process in which the branches (cell-cycle times) have a certain dependence on one another. Characterizing cell-cycle progression in a given experimental condition requires inferring the branch lengths (cell-cycle times) of the lineage tree as well as the correlations between them. In the following sections, I develop a hierarchical model of cell division. To perform Bayesian inference on model parameters, I appeal to Bayes' theorem:

$$\mathsf{Pr}(\Theta_{pop}, \tilde{\lambda}, \beta_m, \beta_d, \tau^2 \mid \tilde{B}_i^{rel}, \tilde{C}_i^{rel}) \propto \tag{2.1}$$

$$\prod_{i=1}^{\mathcal{L}} \mathsf{Pr}(\tilde{B}_i^{rel}, \tilde{C}_i^{rel} \mid \tilde{\lambda}_i, \beta_m, \beta_d, \tau^2, \Theta_{pop}) \mathsf{Pr}(\tilde{\lambda}_i \mid \Theta_{pop}) \mathsf{Pr}(\Theta_{pop}, \beta_m, \beta_d, \tau^2) \tag{2.2}$$

That is, the posterior distribution of model parameters (first line; parameters to be described in the following sections) is proportional to the product of the likelihood of the budding and division observations and the prior distribution on model parameters. Here, $\mathcal{L}$ is the number of lineages in a particular experimental setting. Development of the model begins with specification of a likelihood and error model for the budding and division measurements of a particular lineage:

$$\mathsf{Pr}(\tilde{B}_i^{rel}, \tilde{C}_i^{rel} \mid \tilde{\lambda}_i, \beta_m, \beta_d, \tau^2, \Theta_{pop}). \tag{2.3}$$

Next, I specify a branching process–and corresponding probability distribution–underlying the mean structure for the observations:

$$\mathsf{Pr}(\tilde{\lambda}_i \mid \Theta_{pop}). \tag{2.4}$$

Finally, I specify the population-level parameters in the model hierarchy that facilitate borrowing of information across replicate lineages and the prior distributions on those parameters:

$$\mathsf{Pr}(\Theta_{pop}, \beta_m, \beta_d, \tau^2). \tag{2.5}$$

The parameters of the hierarchical model are: $\tilde{\lambda}_i$, a vector of cell-specific parameters to specify the expected values of the budding and division observations; Θ_{pop}, the subset of population-level parameters $\{\Lambda, \Delta, \sigma^2_\lambda, \sigma^2_\delta, \psi, \rho, \phi\}$ (see Table 2.2.1 for a description); β_m and β_d, the proportions of the mother and daughter cell's λs during which the cell is unbudded. $\tilde{\lambda}$ from Equation 2.1 is a vector of the cell-specific parameters from all lineages.

2.2.1 Likelihood and Error Model for Budding and Cycle Observations

For each cell j in lineage i, an absolute budding ($B_{i,j}$) and cell division or cycle time ($C_{i,j}$) is observed. These measurements are absolute in that each event is recorded at the time elapsed since the beginning of the movie for that lineage (Figure 2.1). In constructing an error model for the budding and cycle times, I assume:

$$B_{i,j} \sim \mathsf{Norm}(\mu_{B_{i,j}}, \tau^2) \tag{2.6}$$

and

$$C_{i,j} \sim \mathsf{Norm}(\mu_{C_{i,j}}, \tau^2) \tag{2.7}$$

Combining these absolute budding and division observations into a multivariate vector:

$$\begin{pmatrix} \tilde{B}_i \\ \tilde{C}_i \end{pmatrix} \sim \mathsf{MVNorm}(\tilde{\mu}_i, \tau^2 I) \tag{2.8}$$

where $\tilde{B}_i$ and $\tilde{C}_i$ are vectors of absolute budding and division times for lineage i and $\tilde{\mu}_i$ is a vector of corresponding means $\mu_{B_{i,j}}$ and $\mu_{C_{i,j}}$. The means of the observations will be described in more detail in subsequent sections. From the joint specification for the absolute measurements (above), it is assumed that the budding and division times within a lineage are independent of one another. As cell-cycle durations of each cell are of interest, I transform the absolute measurements into relative measurements $B^{rel}_{i,j}$ and $C^{rel}_{i,j}$ by subtracting the total time spent to reach the division that produced cell j (Figure 2.1). More formally:

$$B^{rel}_{i,j} = B_{i,j} - C_{i,Ant(j)} \tag{2.9}$$

$$C^{rel}_{i,j} = C_{i,j} - C_{i,Ant(j)} \tag{2.10}$$

where $C_{i,Ant(j)}$ is the division time of the cell in lineage i that is cell j's direct antecedent (e.g. Ant(110) = 11).

After this linear transformation, the likelihood for relative budding and division measurements for lineage i is:

$$\begin{pmatrix} \tilde{B}^{rel}_i \\ \tilde{C}^{rel}_i \end{pmatrix} | \tilde{\lambda}_i, \Theta_{pop}, \beta_m, \beta_d, \tau^2 \sim \mathsf{MVNorm}(A\tilde{\mu}_i, \tau^2 AA')$$

A is the linear transformation matrix and $\tilde{\mu}_i$ is the vector of the mean absolute budding and division times. It is assumed that the lineage observations are independent from one another (hence the product in 2.2). The cell-specific branch lengths are used to specify the expected value of budding and division times. Given the branch lengths that arise from the branching process, the population parameters Θ_{pop}, and the cell type of each cell j, the expected value of the budding and cycle times ($\mu_{B_{i,j}}$ and $\mu_{C_{i,j}}$) will be fully specified:

$$E[B^{rel}_{i,j} | \tilde{\lambda}_i, \Theta_{pop}, \beta_m, \beta_d] = \begin{cases} \beta_m \lambda_{i,j} & \text{if } j \in \mathcal{M}_i \\ \delta_{i,j} + \beta_d \lambda_{i,j} & \text{if } j \in \mathcal{D}_i \end{cases} \tag{2.11}$$

Table 2.1: Description of Population-Level Parameters in Hierarchical Model

Parameter	**Description**
Λ	population average mother cell-cycle duration
Δ	population average daughter cell G1 phase extension
σ^2_λ	variance in cell-specific λ branch lengths
σ^2_δ	variance in cell-specific δ branch lengths
ψ	correlation in λs from two successive mother cycles
ρ	correlation between mother λ and daughter λ
ϕ	correlation between mother λ and daughter δ
β_m	unbudded proportion of mother cell cycle
β_d	unbudded proportion of daughter λ branch
τ^2	measurement error variance

$$E[C^{rel}_{i,j}|\tilde{\lambda}_i, \Theta_{pop}, \beta_m, \beta_d] = \begin{cases} \lambda_{i,j} & \text{if } j \in \mathcal{M}_i \\ \delta_{i,j} + \lambda_{i,j} & \text{if } j \in \mathcal{D}_i \end{cases} \tag{2.12}$$

Here, $\mathcal{M}_i$ and $\mathcal{D}_i$ are the sets of indices in lineage i of mother and daughter cells, respectively. The parameters β_m and β_d are not cell-specific and are shared across all mother and daughter cells in each lineage of a given experimental setting.

2.2.2 An Asymmetric Autoregressive Process Describing Dependence in Cell-Cycle Progression, $Pr(\tilde{\lambda} \mid \Theta_{pop})$

To capture correlation in cell-cycle duration and describe mean structure in budding and cycle times, I developed an autoregressive branching process. The process formalizes the construction of a lineage tree (Figure 2.2). The cell-specific branch lengths that make up the lineage tree ($\tilde{\lambda}$ from above) are based on two sets of parameters: $\lambda_{i,j}$'s and $\delta_{i,j}$'s. $\lambda_{i,j}$ represents the baseline cell-cycle duration for cell j of lineage i while $\delta_{i,j}$ represents a daughter-specific extension to G1 phase duration. As budding yeast cells divide asymmetrically, daughter cells are born smaller than mothers. It is believed that daughter cells must spend more time in G1 to compensate for their smaller birth sizes and reach a critical size required for cell-cycle entry (Johnston et al. (1977); Di Talia et al. (2007)). To capture the correlation

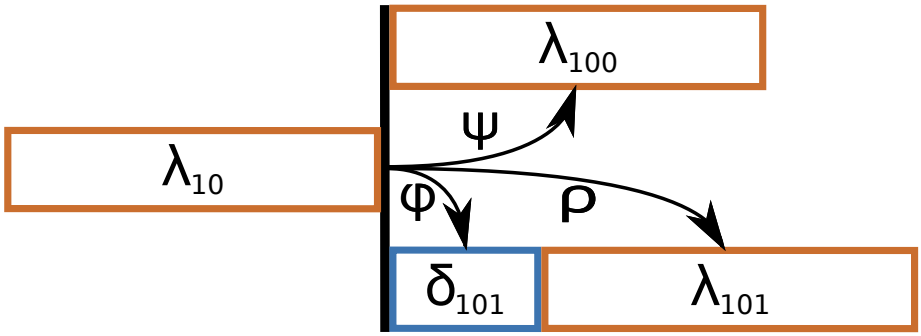

FIGURE 2.2: The diagram is drawn to indicate the branch lengths underlying the single-cell budding and division measurements in Figure 2.1. For this lineage, λ_{10} is the expected cell-cycle duration of mother cell 10. The expected cell-cycle duration of her subsequent cycle is λ_{100} which depends on her first cell cycle through the correlation parameter ψ. For the daughter branch, two parameters specify expected cell-cycle duration: δ_{101} and λ_{101}. In general, the λ branch represents a baseline cell-cycle duration for the daughter cell to which the δ branch is added to account for the longer G1 phases observed in daughters. These branch lengths depend on the mother's cell-cycle duration through the correlation parameters ρ and ϕ respectively.

between these branch lengths, I introduce three correlation parameters: ψ, ρ, and ϕ (see Table 2.2.1 for description). The branching process is asymmetric both in terms of the branch lengths that make up the lineage and in terms of the correlation structure: correlation between mother and daughter branches is different from that between successive mother branches (Figure 2.2).

It is assumed that the budding and cycle times of specific cells in a given experimental setting are marginally drawn from some overarching population distributions of $\lambda_{i,j}$'s and $\delta_{i,j}$'s:

$$\lambda_{i,j} \sim \mathsf{Norm}(\Lambda, \sigma_\lambda^2) \tag{2.13}$$

and

$$\delta_{i,j} \sim \mathsf{Norm}(\Delta, \sigma_\delta^2) \tag{2.14}$$

σ_λ^2 and σ_δ^2 represent variation in the $\lambda_{i,j}$'s and $\delta_{i,j}$'s, respectively, denoting cell-to-cell variability in cell-cycle progression. Λ is the average population-level "base" cell-cycle duration: the expected cell-cycle duration of a mother cell. The expected

cell-cycle duration of a daughter cell can be longer. Δ represents the population mean daughter-specific G1 extension that explains this difference. In describing the dependence between branch lengths, I assume that the branch lengths of each new cell j are conditionally independent of the branch lengths of any previous cell in the lineage given the cell-cycle progression of the new cell's parent cell, Ant(j) (a first-order autoregressive or AR(1) assumption). I also further assume that the branch lengths of cell j take a conditional Normal distribution given the λ of the cell's mother (or $\lambda_{i,Ant(j)}$).

$$\delta_{i,j}|\lambda_{i,Ant(j)}, \Theta_{pop} \sim \mathsf{Norm}\Big(\Delta + \phi\big(\frac{\sigma_\delta}{\sigma_\lambda}\big)(\lambda_{i,Ant(j)} - \Lambda), (1-\phi^2)\sigma_\delta^2\Big) \text{ for } j \in \mathcal{D}_i \quad (2.15)$$

$$\lambda_{i,j}|\lambda_{i,Ant(j)}, \Theta_{pop} \sim \mathsf{Norm}\Big((1-\rho)\Lambda + \rho\lambda_{i,Ant(j)}, (1-\rho^2)\sigma_\lambda^2\Big) \text{ for } j \in \mathcal{D}_i \quad (2.16)$$

$$\lambda_{i,j}|\lambda_{i,Ant(j)}, \Theta_{pop} \sim \mathsf{Norm}\Big((1-\psi)\Lambda + \psi\lambda_{i,Ant(j)}, (1-\psi^2)\sigma_\lambda^2\Big) \text{ for } j \in \mathcal{M}_i \quad (2.17)$$

The above conditional Normal assumptions lead to a joint specification for the branch lengths of lineage i conditional on the population parameters, Θ_{pop}:

$$\tilde{\lambda}_i|\Theta_{pop} \sim \mathsf{MVNorm}(\mu_{\tilde{\lambda}_i}, \Sigma_{\tilde{\lambda}_i}) \quad (2.18)$$

where, due to the AR(1) assumption,

$$\Sigma^{-1}_{\tilde{\lambda}_i,j,k} = 0 \text{ if } j \neq k, j \neq Ant(k) \text{ and } k \neq Ant(j) \quad (2.19)$$

In other words, branch lengths of cells that are not part of the same "triad" (first mother cycle, second mother cycle, and daughter cycle; Figure 2.2) are independent of one another given all other branch lengths in the lineage tree. Also, in contrast to the BAR model, the daughter and second mother cycles within a triad are conditionally independent of one another given the first mother cycle.

2.2.3 Prior Distributions on Model Parameters

In adopting a fully Bayesian approach to parameter inference, I incorporated empirical knowledge of budding yeast cell division in prior distributions on different model parameters. Specifically:

$$\beta_m, \beta_d \sim \mathsf{Beta}(2.4, 17.6) \tag{2.20}$$

$$\Lambda \sim \mathsf{Norm}(78.2, 18.2^2) \tag{2.21}$$

$$\Delta \sim \mathsf{Norm}(55.0, 22.5^2) \tag{2.22}$$

$$\psi, \rho, \phi \sim \mathsf{Unif}(-1.0, 1.0) \tag{2.23}$$

$$\sigma_\delta^{-2}, \sigma_\lambda^{-2} \sim \mathsf{Gamma}(0.95, 4.48) \tag{2.24}$$

$$\tau^{-2} \sim \mathsf{Gamma}(4.08, 10.18) \tag{2.25}$$

These priors were biologically motivated and many were used in previous work with cell populations (Orlando et al. (2009); Mayhew et al. (2011)). The $\mathsf{Beta}(2.4, 17.6)$ prior on β_m and β_d indicates an *a priori* expectation that cells spend 12% of their λ branch durations in the unbudded state. The $\mathsf{Unif}(-1.0, 1.0)$ priors on the correlation parameters reflect the lack of strong *a priori* information. The $\mathsf{Gamma}(0.95, 4.48)$ priors on the precision rather than the variance in G1 phase delays (σ_δ^2) and cell-cycle durations (σ_λ^2) reflect prior beliefs that the standard deviations of the cell-specific branch lengths are less than 15 and greater than 1 with high probability ($\sim$0.965). The $\mathsf{Gamma}(4.08, 10.18)$ prior on precision in measurements of budding and division

($\frac{1}{\tau^2}$) is based on the belief that measurements are approximately within 6 minutes of their true values.

2.2.4 Markov Chain Monte Carlo Sampling of Posterior Distribution

I fit the model to the single cell data using JAGS or Just Another Gibbs Sampler (Plummer (2003); mcmc-jags.sourceforge.net). The program builds a Markov chain Monte Carlo (MCMC) sampler based on the dependencies between and distributional assumptions on variables in a graphical model. The first 10k iterations were counted as burn-in to allow the sampler to find modes of high posterior probability. The subsequent 250k iterations were retained for each parameter. The MCMC chain was initialized for each dataset with sample estimates roughly corresponding to each parameter. As I did not directly observe Λ or Δ, I used sample estimates of the budded and unbudded durations of each cells cell cycle, respectively. Likewise, I initialized σ_λ^2 and σ_δ^2 with the sample variances of the budded and unbudded cell-cycle durations. The initial points of the correlation parameter chains were the sample correlations between successive mother budded durations (ψ), mother and daughter budded durations (ρ), and mother budded and daughter unbudded durations (ϕ). For τ^2, I used the prior mean as the MCMC starting point. The number of samples for each parameter was sufficient to estimate the 2.5th quantile of each parameter's marginal posterior distribution with a 0.01 margin error with 90% probability by the Raftery-Lewis convergence diagnostic (Raftery and Lewis (1992); cran.r-project.org/package=coda).

2.3 Model Fitting to Budding and Division Times from Di Talia *et al.*

2.3.1 Budding Yeast Cell-Cycle Progression Varies Across Experimental Settings

Inferring the parameters of the hierarchical model for the three different datasets (wild-type in glucose, 6xCLN3 in glucose, and wild-type in glycerol/ethanol), revealed distinct patterns of cell-cycle progression. Posterior inferences (modes and 95% highest posterior density (HPD) intervals) are shown in Table 2.3.1. Population average mother cell-cycle duration (Λ) was approximately 88 minutes for wild-type cells in glucose. In contrast, mother cells divided nearly twice as slowly in glycerol/ethanol ($\sim$148 minutes). As Cln3 is a rate-limiting factor for cell-cycle entry (Cross and Blake (1993); Di Talia et al. (2007)), daughters with 6 copies of CLN3 show comparatively short G1 extensions relative to wild-type daughters grown in glucose (Table 2.3.1). Likewise, the estimated variation in 6xCLN3 daughter G1 extensions (σ_δ^2) was much smaller compared to the corresponding estimates for wild-type cells, reflecting the greater availability of the regulator Cln3 (Table 2.3.1). In contrast, wild-type daughters in glycerol/ethanol take nearly 90 minutes more on average to complete G1 than their mothers. Consistent with experimental evidence, cell-cycle progression in glycerol/ethanol is in general slower than in the more preferred sugar source, glucose (Broach (2012)). The estimates of the measurement error variance (τ^2) for wild-type and 6xCLN3 cells are comparable to one another though distinct from the estimates for wild-type cells in glycerol/ethanol. This difference between measurement error variance in the two sugar sources reflects differences in temporal resolution of the time-lapse experiments (imaging every 3 vs. 6 minutes).

Classical studies have suggested that the combined S/G2/M duration is a roughly constant interval shared by mother and daughter cells and that the primary source of variability in cell-cycle duration is due to G1 phase variability (Hartwell and Unger

Table 2.2: Posterior Inferences (Modes and 95% Highest Posterior Density Intervals in parentheses) for Single-Cell Hierarchical Model

	Wild-Type Glucose	**6xCLN3 Glucose**	**Wild-Type GlyEth**
Parameter	**Estimate**	**Estimate**	**Estimate**
Λ	87.57 (84.50,90.95)	95.32 (91.52,100.06)	147.81 (140.21,155.87)
Δ	25.74 (19.93,31.27)	18.94 (14.20,23.30)	90.13 (76.91,101.15)
β_m	0.17 (0.16,0.19)	0.15 (0.13,0.16)	0.27 (0.24,0.29)
β_d	0.13 (0.08,0.18)	0.05 (0.01,0.09)	0.24 (0.19,0.29)
ψ	0.03 (-0.28,0.38)	0.25 (-0.28,0.68)	0.88 (0.68,0.98)
ρ	-0.19 (-0.46,0.11)	-0.38 (-0.72,0.19)	0.24 (-0.19,0.62)
ϕ	-0.11 (-0.39,0.19)	-0.77 (-1.00,0.16)	0.00 (-0.29,0.27)
σ_δ	17.60 (14.69,20.54)	5.03 (1.43,7.37)	37.40 (30.23,45.98)
σ_λ	16.70 (14.75,18.84)	17.17 (14.57,20.63)	25.45 (21.04,30.96)
τ	5.46 (4.82,6.23)	4.41 (3.72,5.32)	12.52 (11.27,14.19)

(1977)). The inferences for β_m and β_d can be used to address this hypothesis. The mother budded period (S/G2/M; $1-\beta_m$), is mildly (4% of total λ on average) shorter than the budded proportion of daughter λ in wild-type cells growing in glucose (Table 2.3.1), with the posterior probability of $\beta_m > \beta_d$ being 0.953. In 6xCLN3 cells, this difference in S/G2/M increases with mothers showing a nearly 10% (of total λ) longer average S/G2/M duration and the posterior probability of $\beta_m > \beta_d =$ 0.999. The pattern appears reversed with wild-type cells grown in glycerol/ethanol. Although the difference in unbudded proportion of the cell cycle between mothers and daughters is still present ($\beta_m > \beta_d = 0.808$), it is less pronounced compared with the differences in glucose ($\sim$2% of total λ longer S/G2/M on average). Taken together, the estimates suggest that the combined S/G2/M duration not only varies between mother and daughter cells but that daughters tend to have longer S/G2/M durations than mothers. The extent of these differences appears to depend on experimental conditions.

2.3.2 *Analysis of Dependence in Cell-Cycle Progression Across Experimental Conditions*

Inferences on the correlation parameters ρ, ψ, and ϕ also revealed differences in dependence of cell-cycle progression in the three different experimental settings. Mothers growing in glycerol/ethanol appear to retain a very similar rate of cell division throughout their lives as indicated by strongly positive (posterior mode of 0.88; Table 2.3.1) correlations between durations of consecutive mother cell cycles (ψ). In contrast, wild-type cells growing in glucose seem to show little to no correlation in cell-cycle progression (Table 2.3.1; HPD intervals include 0). Thus, wild-type mothers and daughters appear to divide in glucose at a rate largely independent of the rates at which either they divided previously (for mothers) or their mothers divided (for daughters). This pattern is also true of 6xCLN3 cells, as 6xCLN3 mothers and daughters appear to divide at rates independent of the rates of immediately preceding cells (Table 2.3.1). Despite the correlation observed between mother cycles, daughters growing in glycerol/ethanol appear to divide independently of their mothers' cell-cycle progression.

2.4 Comparison of Different Cell-Cycle Models with Approximate Bayes Factors

2.4.1 *Computing Approximate Bayes Factors by Mixture Importance Sampling*

To more formally investigate the dependence between successive mother cell cycles in glycerol/ethanol, I computed Bayes factors (Kass and Raftery (1995)). A Bayes factor is a tool for model comparison that evaluates the weight of evidence in data for two competing models. More formally, it is a ratio of the marginal likelihoods of data under the two models.

$$BF_{1,2} = \frac{\int Pr(D|\Theta_{M_1})Pr(\Theta_{M_1})d\Theta_{M_1}}{\int Pr(D|\Theta_{M_2})Pr(\Theta_{M_2})d\Theta_{M_2}} \tag{2.26}$$

$$BF_{1,2} = \frac{Pr(D|M_1)}{Pr(D|M_2)} \tag{2.27}$$

Here M_1 and M_2 represent two competing models, Θ_{M_1} and Θ_{M_2} represent the corresponding parameters of each model and D is the data. The models may potentially have non-nested parameter spaces. A log_{10} Bayes factor close to 0 indicates a lack of evidence in the data to distinguish the two competing models while large positive or negative log_{10} Bayes factors indicate a preference in the data for the model in the numerator or denominator, respectively. To assess evidence in the data supporting a model with nonzero correlation in cell-cycle progression, I compared different models where the correlation parameters ψ, ρ, and ϕ were set to 0. With three correlation parameters, the total number of models to compare is eight. For this analysis, the larger model is always in the numerator of the Bayes factor.

To compute the marginal likelihood under a model first required integrating out the cell-specific branch lengths ($\tilde{\lambda}$). Since the cell division observations are multivariate normal distributed and $\tilde{\lambda}$ is multivariate normal distributed:

$$\tilde{B}_i^{rel}, \tilde{C}_i^{rel} | \tilde{\lambda}_i, \beta_m, \beta_d, \tau^2, \Theta_{pop} \sim \mathsf{MVNorm}(A\tilde{\mu}_i, \tau^2 AA') \tag{2.28}$$

$$\tilde{\lambda}_i | \Theta_{pop} \sim \mathsf{MVNorm}(\tilde{\mu}_{\tilde{\lambda}_i}, \Sigma_{\tilde{\lambda}_i}) \tag{2.29}$$

$$\tilde{B}_i^{rel}, \tilde{C}_i^{rel} | \beta_m, \beta_d, \tau^2, \Theta_{pop} \sim \mathsf{MVNorm}(A\tilde{\mu}_{\tilde{\lambda}_i}, \tau^2 AA' + \Sigma_{\tilde{\lambda}_i}) \tag{2.30}$$

As I did not have a closed form for the unnormalized posterior distribution of each model, I used importance sampling (IS) to compute an approximate Bayes factor. In IS, one approximates an integral by performing Monte Carlo integration, sampling from a density other than the density of interest in the integral and correcting for the difference between the densities. For this analysis, the target density is the unnormalized posterior distribution under a given model M. Specifically:

$$\prod_{i=1}^{\mathcal{L}} \mathsf{Pr}(\tilde{B}_i^{rel}, \tilde{C}_i^{rel} | \beta_m, \beta_d, \tau^2, \Theta_{pop,M}) \mathsf{Pr}(\Theta_{pop,M}, \beta_m, \beta_d, \tau^2) \tag{2.31}$$

where $\Theta_{pop,M}$ is the set of the population-level parameters corresponding to a particular model M. For general importance sampling:

$$Pr(D|M_1) = \int Pr(D|\Theta_{M_1})Pr(\Theta_{M_1})\frac{g(\Theta_{M_1})}{g(\Theta_{M_1})}d\Theta_{M_1} \tag{2.32}$$

$$Pr(D|M_1) \approx \frac{1}{S}\sum_{i=1}^{S}\frac{Pr(D|\Theta_{M_1}^{(i)})Pr(\Theta_{M_1}^{(i)})}{g(\Theta_{M_1}^{(i)})} \tag{2.33}$$

The second of the two above equations shows the IS approximation where the summation is over S samples ($\Theta_M^{(i)}$) from the IS density g. The ratio of the unnormalized posterior density to the IS density of the sampled parameters ($\Theta_{M_1}^{(i)}$) is the importance sampling weight. A central task in performing importance sampling is identifying an appropriate IS density, g. To get the importance sampling distribution of each sub-model, I fit the sub-model to budding and division observations with JAGS, again generating 10k burn-in iterations and retaining 250k posterior samples. Due to potential asymmetry in the posterior densities estimated from some of the samples, I fit a mixture of normal distributions by expectation-maximization to the MCMC samples for each sub-model (Fraley et al. (2012)). The number of components in the mixture (one to twelve possible components) was automatically determined for each sample by Bayesian information criterion (BIC). The fitted component means and covariance matrices were then used as the corresponding locations and scale matrices, respectively, in a mixture of multivariate t densities (25 degrees of freedom). This multivariate t distribution was used as the IS density, g. The marginal likelihoods under each sub-model were computed from 25k importance samples generated by each fitted sub-model mixture of t distributions. The 25k importance samples were sufficient to achieve an effective sample size of at least 7-10k for all model comparisons (Liu (2008)), indicating that the IS weights generally had low variance. Thus, the posterior distribution of parameters under each sub-model was being explored

Table 2.3: $\log_{10}$ Bayes Factors for Wild-Type Cells Grown in Glucose

	$\psi = 0$	$\rho = 0$	$\phi = 0$	$\psi, \rho = 0$	$\psi, \phi = 0$	$\rho, \phi = 0$	$\psi, \rho, \phi = 0$
Full	-0.649	-0.412	-0.601	-1.066	-1.251	-0.961	-1.606
$\psi = 0$	-	-	-	-0.416	-0.603	-	-0.957
$\rho = 0$	-	-	-	-0.656	-	-0.546	-1.193
$\phi = 0$	-	-	-	-	-0.648	-0.356	-1.005
$\psi, \rho = 0$	-	-	-	-	-	-	-0.541
$\psi, \phi = 0$	-	-	-	-	-	-	-0.357
$\rho, \phi = 0$	-	-	-	-	-	-	-0.645

Table 2.4: $\log_{10}$ Bayes Factors for 6xCLN3 Cells Grown in Glucose

	$\psi = 0$	$\rho = 0$	$\phi = 0$	$\psi, \rho = 0$	$\psi, \phi = 0$	$\rho, \phi = 0$	$\psi, \rho, \phi = 0$
Full	-0.311	-0.152	0.276	-0.472	0.027	0.164	-0.078
$\psi = 0$	-	-	-	-0.162	0.344	-	0.232
$\rho = 0$	-	-	-	-0.320	-	0.313	0.069
$\phi = 0$	-	-	-	-	-0.243	-0.118	-0.357
$\psi, \rho = 0$	-	-	-	-	-	-	0.388
$\psi, \phi = 0$	-	-	-	-	-	-	-0.110
$\rho, \phi = 0$	-	-	-	-	-	-	-0.243

Table 2.5: $\log_{10}$ Bayes Factors for Wild-Type Cells Grown in Glycerol/Ethanol

	$\psi = 0$	$\rho = 0$	$\phi = 0$	$\psi, \rho = 0$	$\psi, \phi = 0$	$\rho, \phi = 0$	$\psi, \rho, \phi = 0$
Full	6.021	-0.313	-0.737	5.627	5.327	-1.061	4.925
$\psi = 0$	-	-	-	-0.395	-0.700	-	-1.100
$\rho = 0$	-	-	-	5.941	-	-0.746	5.242
$\phi = 0$	-	-	-	-	6.060	-0.322	5.661
$\psi, \rho = 0$	-	-	-	-	-	-	-0.703
$\psi, \phi = 0$	-	-	-	-	-	-	-0.397
$\rho, \phi = 0$	-	-	-	-	-	-	5.985

with reasonable efficiency by the importance samplers.

2.4.2 Evidence of Dependence in Mother Cell-Cycle Progression in Glycerol/Ethanol

The results of this analysis corroborated the results of the hierarchical model fitting of the three datasets. Namely, the data are consistent with a model in which a cell growing in glycerol/ethanol, whether slow- or fast-dividing, retains her rate of cell-cycle progression throughout her life. The $\log_{10}$ Bayes factors comparing larger

models with sub-models in which ψ equals 0 are approximately 5.0 or larger, measures of decisive evidence for the models with nonzero ψ (Kass and Raftery (1995); Table 2.4.1). In contrast, the Bayes factors for correlation between mother and daughter cell-cycle progression in glycerol/ethanol suggest that the data is not consistent with either model. In fact, all three datasets show little evidence against zero correlation between mother and daughter cell-cycle progression (Tables 2.4.1, 2.4.1, and 2.4.1). In particular, the 6xCLN3 data is not consistent with nonzero correlation in cell-cycle progression between successive mother cell cycles and between mothers and daughters. Moreover, the data on wild-type cells growing in glucose seem to support a model in which the correlation parameters are 0 ($\log_{10}$ Bayes factor with full model of -1.606; Table 2.4.1). A picture emerges of contrasting patterns of division across sugar sources. For the sake of explanation, the mechanism of assigning a cell a rate of division for one cycle is analogous to sampling the rate from some distribution at its birth. In rich media (glucose), the rate at which a cell divides is drawn independent of the rate of division of any cells preceding it. In poor media (glycerol/ethanol), daughters also draw their rates of cell-cycle progression independently of the rates of their mothers. However, once drawn, the cell's rate of division, whether slow or fast, appears to be retained over the cell's lifespan.

2.5 Discussion of Single-Cell Model Fitting Results

This analysis has revealed evidence of dependence in budding yeast cell-cycle progression between successive mother cycles. However, this dependence was only detected among mother cells growing in glycerol/ethanol. Indeed, data from cells growing in glucose (both wild-type and 6xCLN3 cells) seems to support a model in which cell-cycle progression is not correlated. It is worth noting that the lineage trees in each dataset do not span more than 3-4 generations. Thus, it is difficult to claim without more data that the observed correlations are maintained throughout a cell's

lifespan. Moreover, the lack of correlations observed for 6xCLN3 cells might be at least partially attributable to the smaller sample size. With more observations, we can determine whether the lack of dependence in cell-cycle progression we observe is due to low power from smaller samples sizes. Nevertheless, this analysis suggests that when drawing conclusions about cell-cycle progression, the genetic and environmental context of the cells must be taken into account: patterns of cell-cycle progression were quite different across the three experimental settings. Elucidating the biological mechanism by which this dependence arises in glycerol/ethanol is beyond the scope of this work, but future analysis building upon the proposed hierarchical model offer different potential directions for further investigation of cell division (discussed in Chapter 4).

3

Using Single Cell Insights to Enhance the CLOCCS Model of Budding Yeast Population Division

The hierarchical model of cell division at the single-cell level revealed correlation between parameters of cell-cycle progression as well as considerable cell-to-cell variation in cell-cycle duration and daughter-cell-specific G1 phase delays. However, time-lapse microscopy and other single-cell methods are but one class of a variety of ways of monitoring and analyzing cell-cycle progression. Indeed, the method of choice among cell-cycle biologists has been analysis of cell-cycle progression in populations of cells. Population-based methods require synchronization of a cell culture. Synchronization of haploid budding yeast cells can be carried out by means of treatment with mating pheromones, size-based separation of cells by centrifugation, or use of genetic mutants who, upon changes in sugar source or temperature, fail to express functional forms of required cell-cycle regulatory proteins (Hartwell et al. (1974)). After releasing the synchronized population into the cell cycle, one collects independent samples of the population over time. The samples are immediately placed in fixative to prevent any further cell-cycle progression, prepared for microscopy, and

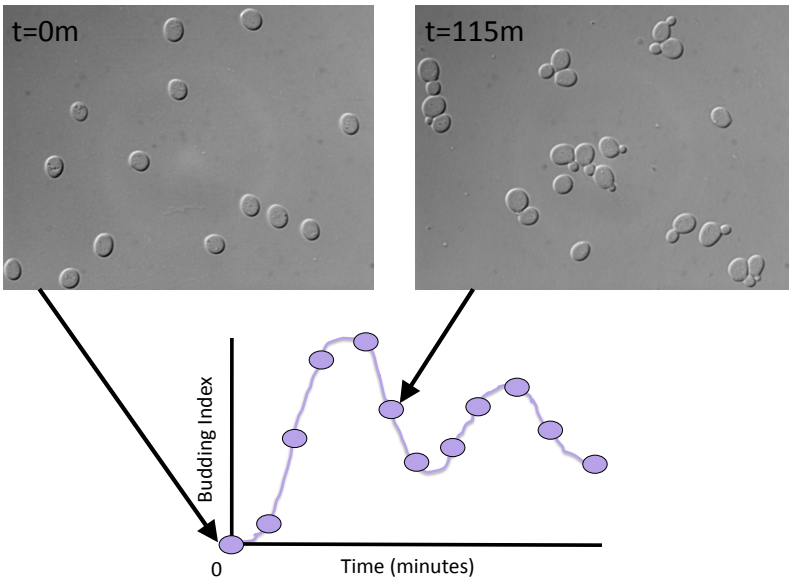

FIGURE 3.1: In a synchrony experiment, an experimenter records the proportion of cells with a particular marker of cell-cycle progression (here, the bud) at each time point. As the population advances through the cell cycle, the proportion of budded cells changes. Shown in the upper left and upper right panels of the diagram are differential interference contrast microscopy images of two fields of view of budding yeast cells at different points in the synchrony experiment. The times indicated are in minutes. Records of the proportion of budded cells at each time point are compiled to make a budding index curve (bottom panel).

scored for the presence of markers of interest. This scoring procedure results in a dynamic index of the presence of a marker of interest in the dividing cell population (Figure 3.1). The presence of a binary-valued cell-cycle marker for each cell is a periodic event (e.g. the bud). Thus, the marker index for a perfectly synchronized population of cells with no cell-to-cell variation in marker appearance would be a square wave function.

However, owing to caveats of synchronization and different sources of biological variation in a budding yeast population, perfect population synchrony is nearly im-

possible to achieve in practice. Synchronization procedures tend to induce delays in cell-cycle entry (hereafter termed recovery), and cells in the population can be heterogeneous in their starting position relative to the point of cell-cycle entry. In addition, daughter cells tend to spend more time on average in G1 phase than mother cells due to the asymmetric nature of budding yeast cell division (Hartwell and Unger (1977)) which produces daughters smaller than mothers at birth. Also, cells in the population do not progress through the cell cycle at the same rate. These three sources of variation contribute to asynchrony in the population thereby complicating estimation of parameters of cell-cycle progression.

3.1 Introduction to CLOCCS Model

The CLOCCS (Characterizing Loss Of Cell Cycle Synchrony) model was developed to infer parameters of cell-cycle progression from population data while accounting for these different sources of asynchrony (Orlando et al. (2007); Orlando et al. (2009)). The basic CLOCCS model specifies the distribution of the position at time t (P_t) of a randomly sampled cell and assumes that cells take an initial position at the onset of the time course that is normal distributed. The model also assumes that such a cell–as well as its progeny–proceeds linearly through each cell cycle with a normal distributed rate. Put more formally:

$$P_t = P_0 + Vt \tag{3.1}$$

Here P_0 is a normal random variable representing the cell's initial position and V is a normal random variable representing the distribution of cell-cycle progression rates. This is a linear combination of normal distributions, and hence the distribution of P_t is normal.

In CLOCCS, the process of cell division is represented by a branching process (see Figure 3.2) where the branch lengths are distances cells traverse. At each branch, a

cell both begins a new cell cycle and contributes to the population a new daughter cell who traverses her own branch in the lineage tree. The set of model parameters, Θ, consists of components of the branching process like a mother cell's expected cell-cycle duration (Λ), a daughter's additional time spent in G1 (Δ), and the expected period of recovery from synchronization by the population (μ_0). Also included in Θ are parameters that describe the different sources of asynchrony in the population such as heterogeneity in the initial population's recovery from synchronization (σ_0^2) and heterogeneity in rates of cell-cycle progression (σ_v^2). σ_0^2 and σ_v^2 are the variances of the normal distributions for P_0 and V, respectively. It is assumed that the mean cell-cycle velocity μ_v is constant over time, and equal to one cell-cycle unit/minute, and that the cell-cycle velocities of mother and daughter cells are drawn from the same distribution. A cell-cycle unit is $\frac{1}{\Lambda}$.

3.1.1 Representing Sub-Populations of Cells with Cohorts

Due to loss of synchrony, a time point sample represents a mixture of cells of different genealogical ages in different phases of the cell cycle (Figure 3.2, part B). To model the mixture of effects contributed by these sub-populations of cells, we consider cells as belonging to different cohorts. Each cohort or subpopulation of cells is represented by a normal density (different colored densities in Figure 3.2, part A). A cohort is indexed by its generation (g) and its reproductive instance (r) where g is the number of daughter cell-specific delays undergone by the cohort, and r is the index of the cell division that produced the current cohort. The probability of randomly sampling a cell from a given cohort {g,r} is:

$$\mathsf{Pr}(g, r \mid \Theta, t) = \frac{M_\Theta(g, r, t)}{Q_\Theta(t)} \tag{3.2}$$

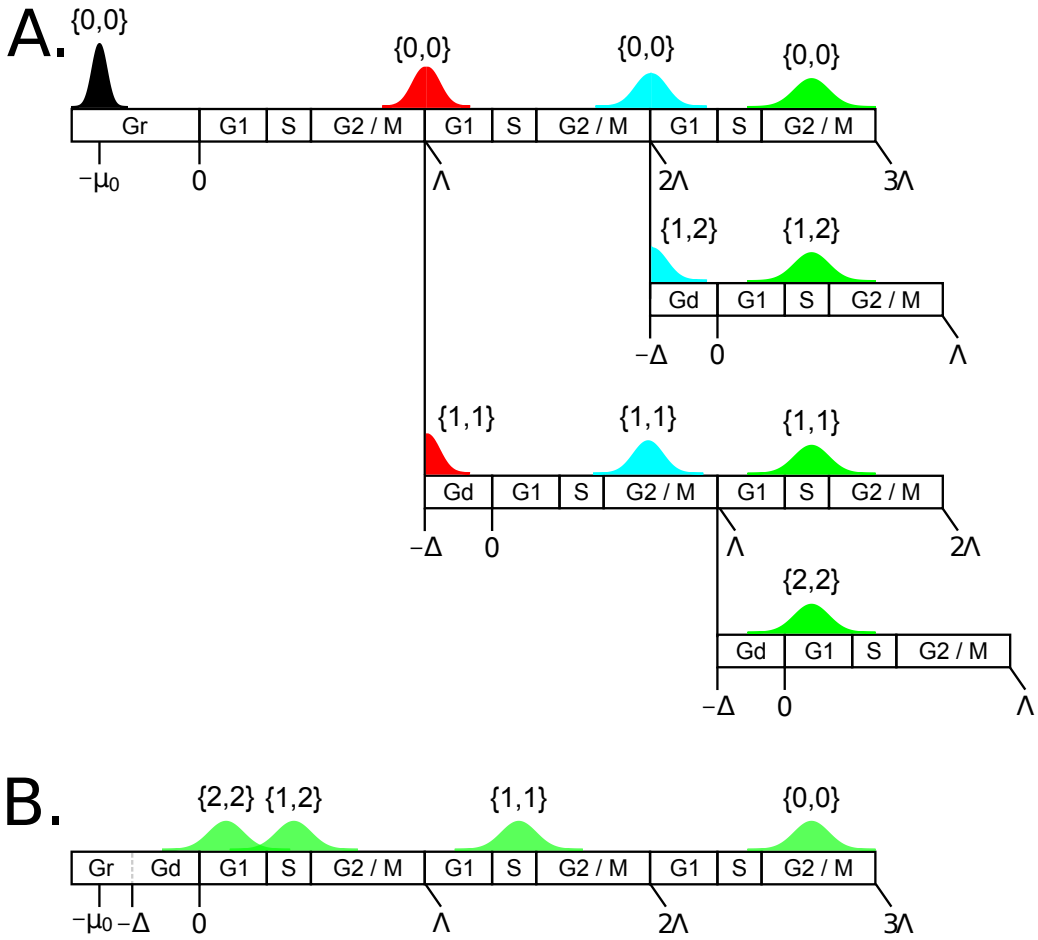

FIGURE 3.2: Shown in (A) is an example branching diagram depicting the CLOCCS model process of division. Cohorts of the same color make up the population at the correspond to time points in the synchrony experiment. The experiment begins with cohort {0,0} distributed about the expected recovery time from synchronization, μ_0. Over the course of the synchrony experiment, the cohorts move along the branching diagram at a linear rate. The cohorts contribute different amounts of probability mass to different positions in the cell-cycle as they move along the branching diagram. Each cell cycle is Λ minutes long and daughter cells spend an additional Δ minutes in their first cycle. At division, a cohort contributes a new cohort to the population that proceeds through the cell cycle on its branch (e.g. {0,0} contributes cohort {1,1}). At a given time point in the synchrony experiment, the probability that a cell is at a particular position in cell division depends on the cohort to which it belongs, and the cell-cycle position distribution at a time point is a mixture of the cohort position distributions at that time point (B).

where

$$M_\Theta(g,r,t) = \begin{cases} 1 & g = r = 0 \\ \left(1 - \Phi\left(\frac{-\Delta-\mu_{grt}}{\sigma_t}\right)\right) \cdot \binom{r-1}{g-1} & 1 \leqslant g \leqslant r \\ 0 & \text{otherwise} \end{cases} \tag{3.3}$$

and

$$Q_\Theta(t) = \sum_{\mathcal{C}} M_\Theta(g,r,t) \tag{3.4}$$

Here, $\mathcal{C}$ represents all possible cohorts, $\Theta = \{\mu_0, \sigma_0^2, \sigma_v^2, \Lambda, \Delta\}$ is the set of all CLOCCS branching process parameters, $\mu_{grt} = -\mu_0 + t - r \cdot \Lambda - g \cdot \Delta$, $\sigma_t = \sqrt{\sigma_0^2 + t^2 \cdot \sigma_v^2}$, and Φ is the standard normal cumulative density function.

3.1.2 CLOCCS Specification of Cell-Cycle Position Distribution

Over time, the cohorts move along the lineage tree, contributing different amounts of probability mass to different parts of the cell-cycle time line (Figure 3.2, parts A and B). To determine the probability of a randomly sampled cell from the population being at cell-cycle position P_t requires marginalizing over the cohorts:

$$\mathsf{Pr}(P_t \mid \Theta, t) = \sum_{\mathcal{C}} \mathsf{Pr}(P_t \mid \Theta, g, r, t)\mathsf{Pr}(g, r \mid \Theta, t) \quad \text{where} \tag{3.5}$$

$$\mathsf{Pr}(P_t|\Theta, g, r, t) = \phi\left(\frac{P_t - (-\mu_0 + t)}{\sqrt{\sigma_0^2 + t^2\sigma_v^2}}\right) \tag{3.6}$$

when g=0 and r=0 or

$$\mathsf{Pr}(P_t|\Theta, g, r, t) = \frac{\phi\left(\frac{P_t-(-\mu_0+t-r\Lambda-g\Delta)}{\sqrt{\sigma_0^2+t^2\sigma_v^2}}\right)}{\sqrt{\sigma_0^2 + t^2\sigma_v^2}\left(1 - \Phi\left(\frac{-\Delta-(-\mu_0+t-r\cdot\Lambda-g\cdot\Delta)}{\sqrt{\sigma_0^2+t^2\sigma_v^2}}\right)\right)} \tag{3.7}$$

when $P_t \geqslant -\Delta$, $g \neq 0$ and $r \neq 0$. Here ϕ is the standard normal density function. $\mathcal{C}$ is the set of all cohorts. A cell from cohort {g,r} has to first undergo the recovery period (μ_0) as well as the g Δ durations and r Λ durations to reach the expected

beginning of its cell cycle. Cells from cohort {0,0} are always part of the population and have infinite support on the negative real line while cells from other cohorts only appear in the population starting from $-\Delta$ hence the cohort-specific truncation and normalization evident in equation 3.7 (Figure 3.2).

3.1.3 Sampling Distribution of Budding Observations

CLOCCS is able to infer parameters of cell-cycle progression by fitting measurements of binary-valued cell-cycle markers (e.g. presence of a bud) as well as continuous-valued flow cytometric measurements of DNA content (Orlando et al. (2007); Orlando et al. (2009); Mayhew et al. (2011)). To fit the CLOCCS model to population measurements (e.g. budding) requires a likelihood. The parameter β represents the expected proportion of the cell cycle during which a cell is unbudded. As such, β takes values in the range 0 to 1. The probability of observing a budded cell derives from the cell-cycle position distribution described above: the probability of observing a budded cell under the CLOCCS model is the probability of observing a cell positioned in the budded proportion of the cell cycle (so, in the interval $(1-\beta)\Lambda$). More formally:

$$\mathsf{Pr}(b_{j,t} = 1|\beta, \Theta, t) = \sum_{\mathcal{C}} \mathsf{Pr}(b_{j,t} = 1|\beta, \Theta, g, r, t)\mathsf{Pr}(g, r|\Theta, t) \tag{3.8}$$

where

$$\mathsf{Pr}(b_{j,t} = 1|\beta, \Theta, g, r, t) = \sum_{c=0}^{C} \left[\Phi\Big(\frac{(c+1)\Lambda - (-\mu_0 + t)}{\sqrt{\sigma_0^2 + t^2\sigma_v^2}}\Big) - \Phi\Big(\frac{(c+\beta)\Lambda - (-\mu_0 + t)}{\sqrt{\sigma_0^2 + t^2\sigma_v^2}}\Big) \right] \tag{3.9}$$

when $g = 0$ and $r = 0$ and

$$\mathsf{Pr}(b_{j,t} = 1|\beta, \Theta, g, r, t) = \sum_{c=0}^{C} \frac{\left[\Phi\left(\frac{(c+1)\Lambda-(-\mu_0+t-r\Lambda-g\Delta)}{\sqrt{\sigma_0^2+t^2\sigma_v^2}}\right) - \Phi\left(\frac{(c+\beta)\Lambda-(-\mu_0+t-r\Lambda-g\Delta)}{\sqrt{\sigma_0^2+t^2\sigma_v^2}}\right)\right]}{1 - \Phi\left(\frac{-\Delta-(-\mu_0+t-r\Lambda-g\Delta)}{\sqrt{\sigma_0^2+t^2\sigma_v^2}}\right)} \tag{3.10}$$

for given $g > 0$ and $r \geqslant g$. Here, C is the number of cell cycles in the lineage tree (e.g. in Figure 3.2 $C = 2$). The normalization for daughter cohorts (i.e. other than the initial cohort {0,0}) is due to the fact that cells in those cohorts do not have positive support for being positioned in a cell cycle until they enter the population. After marginalizing over the cohorts, the probability $p_{j,t} = \mathsf{Pr}(b_{j,t} = 1|\beta, \Theta, t)$ is taken as a binomial success probability of finding a budded cell at time t. We consider time point samples in the synchrony experiment as conditionally independent given the CLOCCS model parameters (Θ) and the unbudded proportion parameter (β). The likelihood of a set of budding observations is:

$$\mathcal{L}(\Theta, \beta) = \prod_{t=1}^{T} \binom{N_t}{n_t} p_{j,t}^{n_t}(1 - p_{j,t})^{N_t - n_t} \tag{3.11}$$

where N_t and n_t are the total number of cells and the number of budded cells, respectively, counted at time t.

3.2 Allowing Correlated Branch Lengths and Variability in Daughter G1 Progression in CLOCCS Model

The analysis of single cell data in the previous chapter revealed structure in cell-cycle progression as well as marked variation in G1 phase durations of daughter cells under different experimental conditions. In CLOCCS, it was previously assumed that the branch lengths (Λ) of the branching process were constant or fixed effects and that rates of cell-cycle progression were perfectly correlated from one mother branch to

the next and from mother branches to daughter branches. Put in terms of single cells, a hypothetical mother cell in the initial cohort would draw a cell-cycle velocity from the normal distribution $N(1.0, \sigma_v^2)$. The mother cell would not only retain this cell-cycle velocity over each of her subsequent cell cycles but would also bestow the same cell-cycle velocity to every one of her daughters.

In this section, I explore the possibility of extending the model to be more comparable to the single-cell model from the previous chapter, allowing for different cell-cycle branch lengths in the population lineage tree as well as arbitrary (not perfect) correlations between them. The CLOCCS model construction also assumed that daughter-specific G1 extensions (Δ) were fixed effects that did not vary between daughter branches. I test these two assumptions in the following analysis, extending the CLOCCS population model to allow for heterogeneity and dependence in cell-cycle and daughter G1 branch lengths. I evaluate the information in population data to recover correlation parameters and variances in cell-cycle and daughter G1 branch lengths by simulating budding data from models with known parameter settings.

3.2.1 Parameterizing Model in Terms of Branch Lengths Rather Than Velocities

Incorporating branch-specific cell-cycle velocities into the CLOCCS model is complicated by their non-linear contribution to position. Thus, we re-parameterized the model in terms of branch-specific durations (see Figure 3.3). That is, the likelihood of a randomly sampled cell from the population at time t falling at cell-cycle position P_t given the model parameters Θ is:

$$\mathsf{Pr}(P_t|\Theta, t) = \int_{\tilde{\lambda}} \sum_{B} \mathsf{Pr}(P_t|\Theta, B, t, \tilde{\lambda})\mathsf{Pr}(B|t, \Theta, \tilde{\lambda})\mathsf{Pr}(\tilde{\lambda}|\Theta) \tag{3.12}$$

Here $\tilde{\lambda}$ is a vector of the branch lengths of the population lineage tree (both λs and δs) and B is the set of branch-specific indices (e.g. 1, 10, 11, etc.). The binary indexing scheme for each branch is needed to trace the antecedence of a cell belonging to a

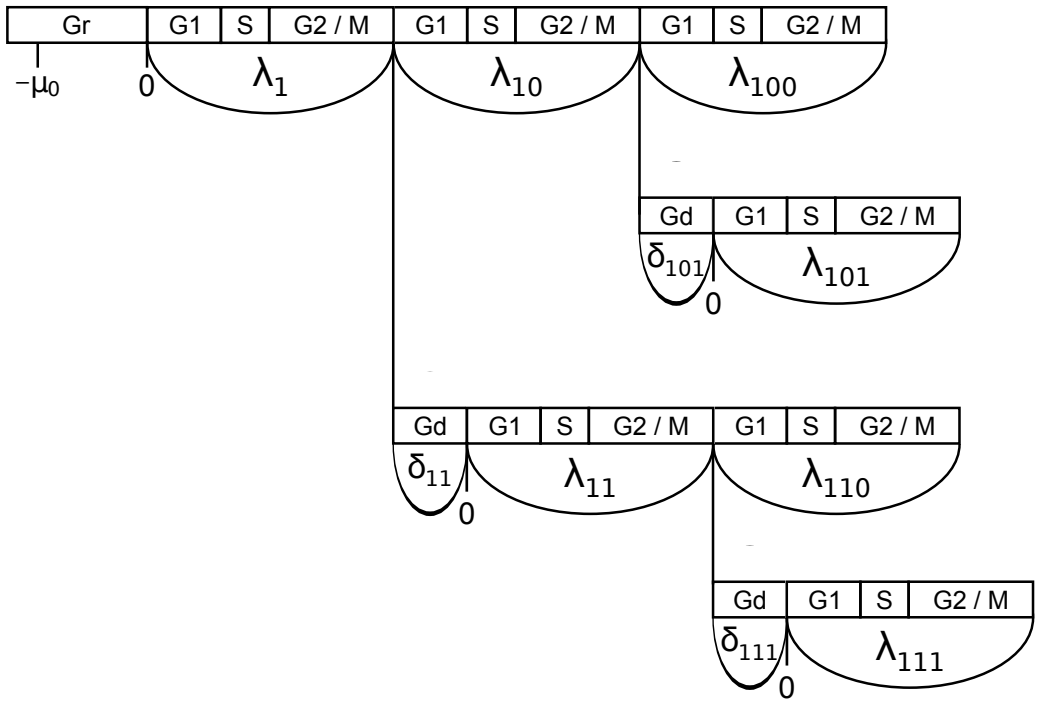

FIGURE 3.3: Shown is an adapted branching diagram from Figure 3.2. In the extended CLOCCS model, branch lengths (Λ and Δ) are not constant and shared throughout the diagram but rather are means of population-level distributions from which branch-specific durations (λs and δs arise). The extended model focuses on capturing variation in and correlation between these branch-specific durations.

particular branch. Rather than fixing the cell-cycle branch lengths and daughter-specific G1 extensions to the same durations (Λ and Δ respectively) as in the previous CLOCCS formulation, this extended model allows for branch lengths to vary and depend on one another. Allowing for dependence in these branch-specific durations requires specifying a multivariate probability distribution on the durations ($\tilde{\lambda}$).

3.2.2 Probability Distribution on Branch Lengths ($\tilde{\lambda}$) of the Population Lineage Tree

To allow for dependence and heterogeneity in branch lengths, I augment the set of CLOCCS model parameters, Θ, to include the five new parameters: the branch correlation parameters ψ, ρ, and ϕ; and the variances in branch lengths σ^2_λ and σ^2_δ.

ψ represents the correlation between "mother" cell-cycle branches (e.g. between λ_1 and λ_{10} in Figure 3.3) in the population lineage tree. ρ represents the correlation between "mother" and "daughter cell-cycle branches (e.g. between λ_1 and λ_{11} in Figure 3.3). ϕ corresponds to the correlation between a mother cell-cycle branch and a daughter G1 extension branch (e.g. between λ_1 and δ_{11} in Figure 3.3). As in the single cell model of the previous chapter, it is assumed that the branch lengths ($\tilde{\lambda}$) are distributed multivariate normal:

$$\tilde{\lambda} \sim \mathsf{MVNorm}(\mu_{\tilde{\lambda}}, \Sigma_{\tilde{\lambda}}) \tag{3.13}$$

Marginally speaking:

$$\lambda_i \sim \mathsf{Norm}(\Lambda, \sigma_\lambda^2) \tag{3.14}$$

and

$$\delta_i \sim \mathsf{Norm}(\Delta, \sigma_\delta^2) \tag{3.15}$$

Thus, as σ_λ^2 and σ_δ^2 approach 0, the expected cell-cycle and daughter G1 branch lengths become Λ and Δ, respectively, as in the original CLOCCS model. Here the mean vector, $\mu_{\tilde{\lambda}}$ of the multivariate normal distribution for $\tilde{\lambda}$ consists of a number of λs equal to the the total number of links in the population lineage tree followed by a number of δs equal to the number of daughter links in the population lineage tree. As indicated, the branch lengths of the population lineage tree ($\tilde{\lambda}$) are conditionally independent of time t given the model parameters Θ.

To specify the covariance matrix, $\Sigma_{\tilde{\lambda}}$, I use an autoregressive structure following (Ripley (1981)):

$$\tilde{\lambda} = \mu_{\tilde{\lambda}} + A(\tilde{\lambda} - \mu_{\tilde{\lambda}}) + \epsilon_{\tilde{\lambda}} \tag{3.16}$$

$$\tilde{\lambda} - \mu_{\tilde{\lambda}} = (I - A)^{-1}\epsilon_{\tilde{\lambda}} \tag{3.17}$$

with

$$\epsilon_{\tilde{\lambda}} \sim \mathsf{MVNorm}(0, \Sigma_\epsilon) \tag{3.18}$$

By multivariate normal theory:

$$\Sigma_{\tilde{\lambda}} = Var(\tilde{\lambda} - \mu_{\tilde{\lambda}}) = (I - A)^{-1}\Sigma_{\epsilon}((I - A)^{-1})' \tag{3.19}$$

The matrix A has zero-valued diagonal and upper-diagonal elements. For the lower diagonal elements of A:

$$A_{i,j} = \begin{cases} \psi & \text{if } i \in \Lambda_{\mathcal{M}} \text{ and } j = Ant(i) \\ \rho & \text{if } i \in \Lambda_{\mathcal{D}} \text{ and } j = Ant(i) \\ \phi\frac{\sigma_\delta}{\sigma_\lambda} & \text{if } i \in \Delta_{\mathcal{D}} \text{ and } j = Ant(i) \end{cases} \tag{3.20}$$

Here, $Ant(i)$ is the index of the branch that immediately precedes branch i in the population lineage tree (e.g. Ant(110) = 11) . Λ_M and Λ_D are sets of the indices of mother and daughter cell-cycle branch durations, respectively. $\Delta_{\mathcal{D}}$ is the set of indices of daughter G1 extension branches.

The covariance matrix, Σ_ϵ, has zero-valued off-diagonal elements. In particular, $\sigma^2_{i,i} = Var(\lambda_i|\lambda_{Ant(i)})$ in the case of cell-cycle branches and $\sigma^2_{i,i} = Var(\delta_i|\lambda_{Ant(i)})$ in the case of daughter branch G1 extensions. If $i = 1$ then $\sigma^2_{i,i} = \sigma^2_\lambda$. If $i > 1$ is associated with a "mother" cell-cycle branch (e.g. branch λ_{10}) then $\sigma^2_{i,i} = (1-\psi^2)\sigma^2_\lambda$. If i corresponds to a "daughter" cell-cycle branch (e.g. branch λ_{11}) then $\sigma^2_{i,i} = (1 - \rho^2)\sigma^2_\lambda$. Likewise, if $i > 1$ corresponds to a daughter G1 extension branch such as δ_{11} then $\sigma^2_{i,i} = (1 - \phi^2)\sigma^2_\delta$. The diagonal elements in the covariance matrix (with the exception of $i = 1$) are the conditional variances of the corresponding branch length given $\lambda_{Ant(i)}$, its antecedent branch. This construction guarantees that the marginal variances of the λ_i and δ_i branches are σ^2_λ and σ^2_δ respectively. So, as in the single cell model of the previous chapter, the λ_i and δ_i branch lengths are jointly distributed multivariate normal and thereby conditionally normal distributed given their antecedent λ branch and the CLOCCS model parameters Θ. Also as previously developed, branch lengths are assumed to be conditionally independent of all other

preceding branch lengths given the immediately preceding antecedent branch length (e.g. $\lambda_{110} \perp\!\!\!\perp \lambda_1 | \lambda_{11}$; first order autoregressive assumption). Specifically:

$$\delta_i | \lambda_{Ant(i)}, \Theta_{pop} \sim \mathsf{Norm}\Big(\Delta + \phi\big(\frac{\sigma_\delta}{\sigma_\lambda}\big)(\lambda_{Ant(i)} - \Lambda), (1-\phi^2)\sigma_\delta^2\Big) \text{ for } i \in \mathcal{D} \tag{3.21}$$

$$\lambda_i | \lambda_{Ant(i)}, \Theta_{pop} \sim \mathsf{Norm}\Big((1-\rho)\Lambda + \rho\lambda_{Ant(i)}, (1-\rho^2)\sigma_\lambda^2\Big) \text{ for } i \in \mathcal{D} \tag{3.22}$$

$$\lambda_i | \lambda_{Ant(i)}, \Theta_{pop} \sim \mathsf{Norm}\Big((1-\psi)\Lambda + \psi\lambda_{Ant(i)}, (1-\psi^2)\sigma_\lambda^2\Big) \text{ for } i \in \mathcal{M} \tag{3.23}$$

where $\mathcal{M}$ and $\mathcal{D}$ are the sets of indices corresponding to mother and daughter branches, respectively, in the population lineage tree.

3.2.3 Determining the Probability that a Cell is Positioned on a Given Branch of the Population Lineage Tree

In contrast to the original CLOCCS formulation, cells are categorized not as belonging to cohorts but rather as being positioned on different branches. The probability of a cell being on a particular branch, B, in the population lineage tree is the probability that the current time, t, is between the beginning and end points of that branch. Probability masses associated with each branch will vary with time owing to the cell-cycle progression of the population.

$$\mathsf{Pr}(B|\Theta, \tilde{\lambda}, t) = \frac{M(B|\Theta, \tilde{\lambda}, t)}{\sum_{B^*} M(B^*|\Theta, \tilde{\lambda}, t)} \tag{3.24}$$

where

$$M(B|\Theta, \tilde{\lambda}, t) = \Phi\left(\frac{t - (\mu_0 + I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2 + I_B^{+\prime}\Sigma_{\tilde{\lambda}} I_B^+}}\right) \tag{3.25}$$

if $B = 1$ and

$$M(B|\Theta, \tilde{\lambda}, t) = \Phi\left(\frac{t - (\mu_0 + I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2 + I_B^{+\prime}\Sigma_{\tilde{\lambda}} I_B^+}}\right) - \Phi\left(\frac{t - (\mu_0 + I_B^{-\prime}\tilde{\lambda}}{\sqrt{\sigma_0^2 + I_B^{-\prime}\Sigma_{\tilde{\lambda}} I_B^-}}\right) \tag{3.26}$$

otherwise. Φ is the standard normal CDF, and both I_B^- and I_B^+ are vectors that select different elements of $\tilde{\lambda}$. I_B^- selects all branch lengths (λ's and δ's) leading up to but not including the durations associated with branch B. For example, I_{100}^- would contain 1s at the positions indexed by λ_1 and λ_{10} since those branch lengths precede branch 100. No positions in I_{100}^- indexed by δs would take a value of 1, and all other elements in the vector would be zero-valued. Likewise, I_B^+ selects all branch lengths leading up to and including those associated with branch B. As an example, I_{101}^+ would be zero-valued except for 1's at positions corresponding to λ_1 and λ_{10} (the preceding branch lengths in branch 101's lineage). The vector would also contain 1's at positions corresponding to λ_{101} and δ_{101} since those branch lengths are part of branch 101.

The branch indices (B), like the cohorts in the original CLOCCS model, are a device to evaluate the probability of different sub-populations of the cell population. In addition, the branch indices are necessary to keep track of the genealogical relationships between the branches and thereby to specify the trajectory followed by a particular cell that could fall on a particular branch. In this way, the above distribution specifies the probability that a cell at time t has reached the cell cycle associated with a given branch. Unlike the previous CLOCCS formulation in which variation in a cell's cell-cycle position was a quadratic function of time ($\sigma_0^2 + t^2\sigma_v^2$), the current formulation involves stepwise increases in cell-cycle position variance due to the transition of a cell from one branch to another or even within a branch (e.g. from the G1 extension to the λ of a daughter branch).

3.2.4 Sampling Distribution for Budding Observations Under Extended Model

As previously, the probability of a randomly sampled cell j being budded at time t corresponds to the probability of the cell being in the budded proportion of the cell cycle. In contrast to the original CLOCCS model, the probability of a particular cell

being budded at time t depends on the branch (B) on which it is positioned rather than the cohort to which it belongs:

$$\mathsf{Pr}(b_{j,t} = 1|\Theta, t) = \sum_B \int_{\tilde{\lambda}} \mathsf{Pr}(b_{j,t} = 1|\Theta, B, t, \tilde{\lambda})\mathsf{Pr}(B|\Theta, \tilde{\lambda}, t)\mathsf{Pr}(\tilde{\lambda}|\Theta)d\tilde{\lambda} \tag{3.27}$$

$$= \sum_B \mathsf{Pr}(b_{j,t} = 1|\Theta, B, t)\mathsf{Pr}(B|\Theta, t) \tag{3.28}$$

The distribution over the different branch lengths in the lineage tree ($\tilde{\lambda}$) as well as the probability of sampling a cell from a particular branch ($\mathsf{Pr}(B|\Theta, \tilde{\lambda}, t)$) have been specified in previous sections. It is important to note that this sampling model is applicable to all binary-valued markers of cell-cycle progression, though we focus on the bud in our analysis.

The probability mass associated with sampling a budded cell from a particular branch B is the probability that the current time t falls between the two endpoints of the budded period of that branch.

$$\mathsf{Pr}(b_{j,t} = 1|\Theta, B, t, \tilde{\lambda}) = \mathsf{Pr}(I_B^{bud\prime}\tilde{\lambda} \leqslant t < I_B^{+\prime}\tilde{\lambda}|\Theta, B, t, \tilde{\lambda}) \tag{3.29}$$

More formally:

$$\mathsf{Pr}(b_{j,t} = 1|\Theta, B, t, \tilde{\lambda}) = \frac{\Phi\left(\frac{t-(\mu_0+I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{+\prime}\Sigma_{\tilde{\lambda}}I_B^{+}}}\right) - \Phi\left(\frac{t-(\mu_0+I_B^{bud\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{bud\prime}\Sigma_{\tilde{\lambda}}I_B^{bud}}}\right)}{\Phi\left(\frac{t-(\mu_0+I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{+\prime}\Sigma_{\tilde{\lambda}}I_B^{+}}}\right)} \text{ if } B = 1 \tag{3.30}$$

$$\mathsf{Pr}(b_{j,t} = 1|\Theta, B, t, \tilde{\lambda}) = \frac{\Phi\left(\frac{t-(\mu_0+I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{+\prime}\Sigma_{\tilde{\lambda}}I_B^{+}}}\right) - \Phi\left(\frac{t-(\mu_0+I_B^{bud\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{bud\prime}\Sigma_{\tilde{\lambda}}I_B^{bud}}}\right)}{\Phi\left(\frac{t-(\mu_0+I_B^{+\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{+\prime}\Sigma_{\tilde{\lambda}}I_B^{+}}}\right) - \Phi\left(\frac{t-(\mu_0+I_B^{-\prime}\tilde{\lambda})}{\sqrt{\sigma_0^2+I_B^{-\prime}\Sigma_{\tilde{\lambda}}I_B^{-}}}\right)} \text{ otherwise} \tag{3.31}$$

The vector I_B^{bud} is a selector vector similar to I_B^+ except that instead of a 1 at the position in the vector corresponding to cell-cycle branch length λ_B, there is β. As the time course necessarily begins with cells on the "1" link of the lineage tree, the

normalizing constant of the budding probability $Pr(b_{j,t} = 1|\Theta, B = 1, t, \tilde{\lambda})$ includes all mass up to and including the end of the "1" link. Similarly to the original CLOCCS model, our assumptions of normality mean that cells on later generation branches (e.g. branch 111) in the tree can exist and be budded early in the time course. Likewise, cells on early generation branches in the tree have zero probability of being alive and budded throughout the synchrony experiment. However, in both of these cases, the probability of such events is vanishingly small. Here, the branch lengths $\tilde{\lambda}$ as well as the branch indices (B) of the population lineage tree are nuisance parameters to be integrated out.

As in previous developments, the probability of sampling a budded cell at a time t (that is, $\mathsf{Pr}(b_{j,t} = 1|\Theta, t)$) is taken as a success probability in a binomial likelihood. Certain combinations of parameter values can produce budding probabilities greater than 1. For example, large cell-to-cell variation in daughter cell-specific G1 phase delay (σ^2_δ) permits negative branch-specific δ's. Parameter combinations that resulted in biologically impossible phenomena (such as daughter cells being budded before they were born) were assigned a likelihood of 0.

3.2.5 Prior Distributions and Model Fitting

Prior Distributions

Taking a Bayesian approach to inference dictates the specification of prior distributions on model parameters. Based on previous observations of wild-type haploid cell-cycle durations (Orlando et al. (2009); Mayhew et al. (2011)), a $\mathsf{Norm}(78.2, 18.2)$ prior was placed on Λ and a $\mathsf{Norm}(41.4, 11.25)$ prior on Δ. To constrain the inferences for the variances on the cell-cycle and daughter G1 extension branch lengths, a $\mathsf{Gamma}(9.12, 35.15)$ prior distribution was placed on $\frac{\sigma_\delta}{\Delta}$ and $\frac{\sigma_\lambda}{\Lambda}$. These distributions concentrated prior mass on those combinations in which standard deviation was less than the mean. The mean of these distributions was set to ~ 0.25, indicat-

ing a prior expectation that individual daughter G1 extensions or cell-cycle branch lengths would fall outside a range of 0.5Δ (or 0.5Λ) with approximately 5% probability. The correlation parameters (ρ, ψ, and ϕ) were given a $\mathsf{Unif}(0,1)$ distribution. This distribution is equivalent to a $\mathsf{Beta}(1,1)$ distribution. A linear transformation of the correlation parameters allowed them to take values in the range (-1,1). Without strong prior beliefs that correlations between branch lengths would be present, the transformed prior distribution incorporated weak information (a re-scaled Beta distribution with $\alpha = 2$ and $\beta = 2$). The resulting prior distribution for the correlation parameters was symmetric and slightly peaked with a mean and mode at 0.0. For the recovery period parameters (μ_0 and σ_0), some recovery time was expected. It was also expected that recovery time would not exceed the time required for a single cell division. So, an exponential prior was placed on μ_0 ($\mathsf{Exp}(\lambda = 1/78.2)$). An $\mathsf{Inv\text{-}Gamma}(2, 78.2/3)$ prior was placed on σ_0, reflecting strong *a priori* beliefs that recovery time was both positive and less than two cell divisions in length. Finally, a $\mathsf{Beta}(2.4, 17.6)$ prior distribution was placed on the budding proportion parameter β (Orlando et al. (2009)).

Model Fitting by Random Walk Markov chain Monte Carlo Sampling

To fit our population model to each simulated dataset, we used a random walk Markov chain Monte Carlo (MCMC) sampler with Metropolis updates (Metropolis et al. (1953)). The sampler was run for a burn-in period of 25k iterations. During this burn-in period we tuned the random walk algorithm based on parameter acceptance rates. We then retained the following 500k iterations for further analysis. To facilitate sampling along the real line, we performed the logit transformation for β, $\frac{1+\rho}{2}$, $\frac{1+\psi}{2}$, and $\frac{1+\phi}{2}$. We also performed the log transformation for σ_0. The log transform was also performed and the corresponding prior distribution was derived for $\frac{\sigma_\delta}{\Delta}$ and $\frac{\sigma_\lambda}{\Lambda}$. Initial parameter settings were $\mu_0 = -90.0$, $\mathsf{log}(\sigma_0) = 2.5$, $\Lambda = 85.0$, $\Delta = 41.4$,

$\mathsf{log}(\sigma_\delta) = 2.5$, $\mathsf{log}(\sigma_\lambda) = 2.5$, $\mathsf{logit}(\frac{1+\rho}{2}) = 0.0$, $\mathsf{logit}(\frac{1+\psi}{2}) = 0.0$, $\mathsf{logit}(\frac{1+\phi}{2}) = 0.0$, and $\mathsf{logit}(\beta) = -1.386294$.

The Raftery diagnostic was computed to assess MCMC convergence (Raftery and Lewis (1992)). By this diagnostic, the 500k retained iterations from the fitting of each simulated dataset (described in more detail in the next section) were sufficient to estimate the 2.5th quantile within a 0.005 margin of error with 95% probability for all but one parameter: Δ. In fact, sampling of the posterior for Δ seemed very inefficient with significant autocorrelation between samples more than 100 iterations apart. In addition, posterior samples of Λ and β also showed significant long-range (more than 50 iterations) autocorrelation despite passing the Raftery diagnostic. As random walk MCMC is known to be a less efficient method for posterior sampling, exploring other sampling approaches might address the observed inefficiencies.

3.3 Simulation Study with Extended CLOCCS Model

3.3.1 Simulation of Dividing Populations of Budded Yeast Cells

To evaluate the quality of fits of the extended model, I implemented simulator software to both construct random lineage trees and represent changes over time in the position of cells on the trees based on our modeling assumptions. The simulator was built to mimic a population budding time course experiment, accounting for the different effects of asynchrony. The simulator took the following as input: values of the different parameters of the population model; the number of cells in the population at the onset of the experiment; the time points in minutes at which budding counts were collected; the number of cells counted at each time point; and the number of divisions the population of cells underwent. The software returned a vector consisting of the number of budded cells at each specified time point.

In evaluating the extended model's fit to simulated data with known parameters, five different simulated datasets were generated. For all five datasets, $\mu_0 = 95.97$

minutes, $\log(\sigma_0) = 2.7$, $\Lambda = 80$ minutes, $\Delta = 35$ minutes, $\sigma_\lambda = e^{2.5} = 12.18$ minutes, and $\beta = 0.14$. The true size of the initial population was set to 10k, the number of cells to be collected at each time point to 200, and the number of cycles undergone by the cell population to 5. In the first three simulated datasets, there was either no ($\rho = 0.0$, $\psi = 0.0$, $\phi = 0.0$), mild ($\rho = -0.3$, $\psi = 0.3$, $\phi = -0.3$), or strong ($\rho = -0.9$, $\psi = 0.9$, $\phi = -0.9$) correlation in cell-cycle duration, respectively. For these three datasets, $\sigma_\delta = e^2 = 7.39$ minutes. In the fourth and fifth datasets, variances in daughter G1 phase extension branches varied. For these two datasets, the correlation parameters (ρ, ψ, and ϕ) were set to 0.0 and σ_δ was set to $e = 2.72$ minutes and $e^3 = 20.09$ minutes, respectively. These five different models correspond to five different functions for the probability of budding over time in the true underlying cell population.

Figure 3.4 depicts the expected fraction in the cell population as a function of time. Under the five different models, one can see differences in the amplitude of peaks in budding as well as the rate of damping in the curve (Figure 3.4). In particular, models with lower variation in daughter G1 phase extension (smaller σ_δ) lose synchrony at a lower rate and, consequently, higher amplitudes in peaks in the budding curve after the first cell cycle. Models with higher absolute magnitude correlations between branch lengths also show higher peak amplitude and longer maintenance of budding oscillations. Importantly, varying either G1 phase extension variability or correlation between branch lengths seems to have no effect on the behavior of the budding curve through the first peak. This result is not surprising in that G1 phase extensions are unique to daughter cells which generally do not arise in the population until after the first peak in budding. Also, correlations in branch lengths cannot be estimated until at least two cycles have been observed.

In synchrony experiments, at least three variables are under the direct control of the biologist: the number of cell divisions observed; the number of cells counted

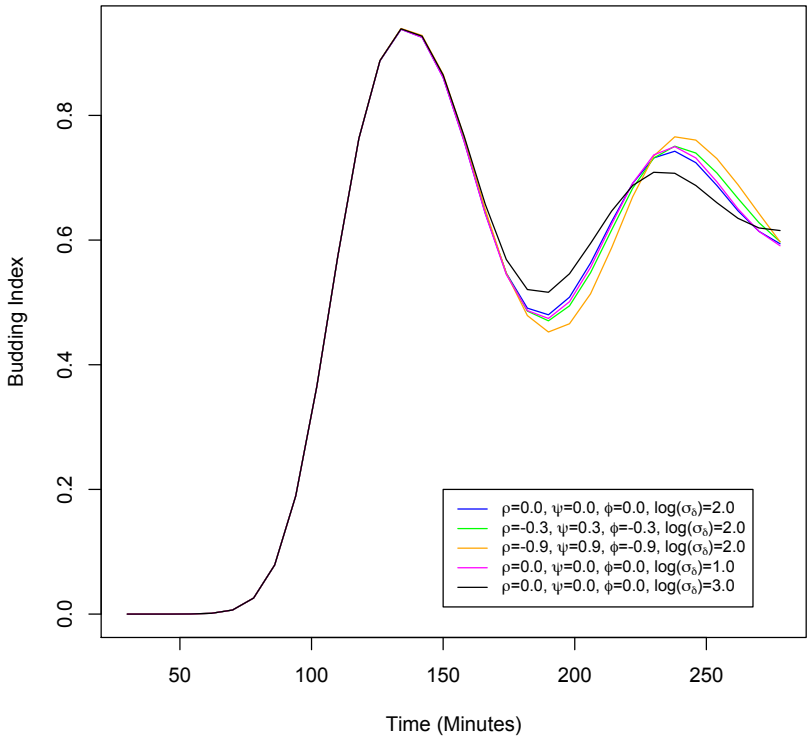

FIGURE 3.4: Shown are the probabilities of randomly sampling a budded cell over a period of observation of two cycles. Note that the budding probabilities are those of the "true" population and do not include sampling variability inherent to counting a subset of cells from the population (Figure 3.9). The simulated budding curve under a model of no correlation is difficult to distinguish from the curve under either a model of weak correlation (compare blue curve with green curve) or a model of relatively reduced variation in G1 phase delay (compare blue curve with magenta curve). Though, as the absolute magnitude of correlation increases, so too does the amplitude of peaks in the simulated curve (compare blue curve with green and orange curves). Peak amplitude seems inversely related to variation in G1 phase delay (compare magenta curve with blue and black curves).

at each time point; and the sampling rate or frequency of time points in the time course. Changing one or more of these variables will undeniably have an effect on the information content in the data and, thereby, on our ability to estimate model parameters. Knowing the times at which and the extent to which these different experimental variables resolve the true population budding curve and thereby distinguish one model from another (e.g. high vs. low inter-branch correlation) would provide valuable and actionable information for population-based synchrony experiment design. Thus, in a separate analysis, an additional dataset was simulated in which the number of cell divisions observed was varied. Specifically, simulated budding counts were generated over three–rather than two–cycles (Figure 3.5). The three cycle budding curves in Figure 3.5 seem to suggest that extending the number of cycles observed in a synchrony experiment can aid in the distinction of different cell-cycle models. For each combination of model and experimental setting, 10 simulated datasets were generated. The population model was then fitted to each dataset to verify consistency in sampling-based inferences.

3.3.2 Parameter Inferences from Simulation Study

After model fitting, we evaluated the accuracy of the population model in capturing the true values of each parameter, particularly those of ρ, ψ, ϕ and σ_δ. Encouragingly, 95% posterior credible intervals for nearly all of the model parameters included the true parameter values (Figures 3.6, 3.7, and 3.8). However, inferences for the correlation parameters were quite wide and included 0 (Figure 3.7). In addition, inferences for σ_δ appeared removed from the true values (Figure 3.8). Parameter estimates are inherently biased by the prior distribution in Bayesian inference, and influence from the prior did appear to be a factor as inferences for σ_δ were closer to the prior means than the true values (dashed vs. solid black lines in Figure 3.6). Furthermore, estimates were consistent across each of the 10 datasets simulated under

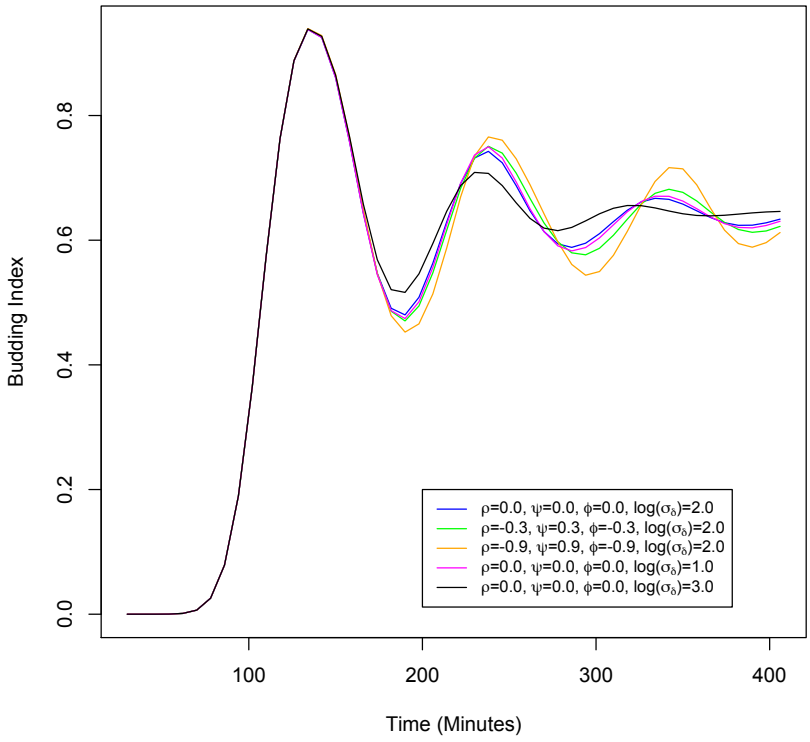

FIGURE 3.5: Shown are the probabilities of randomly sampling a budded cell over a period of observation of three cycles. We note that the budding probabilities are those of the "true" population and do not include sampling variability inherent to counting a subset of cells from the population. The differences between budding curves under the five models become more exaggerated with the observation of the additional cycle (compare with Figure 3.4).

each model suggesting that this effect was not the result of sampling variability in the simulated data. In addition, in those cases where model inferences did not agree with the true parameter values, estimates for other parameters also appeared to deviate from their true values. This effect can be seen for model 5 (black lines in Figures 3.6 and 3.8) in which the inferences for μ_0, Λ, Δ, and β are systematically shifted from the true values. This pattern can also be seen, to a lesser extent, with model 3 (orange lines in Figures 3.6 and 3.8) in which the correlation parameters are

set to values of high magnitude.

To determine whether this bias was the result of a high degree of correlation between the different model parameters, I used the retained MCMC samples to compute the sample correlation matrix, C. With the sample correlation matrix, C, I performed an eigen-decomposition to compute the condition number, $\kappa(C)$.

$$\kappa(C) = \left| \frac{\lambda_{max}}{\lambda_{min}} \right| \tag{3.32}$$

The closer the correlation matrix's condition number is to infinity, the more ill-conditioned is the matrix. A condition number close to infinity indicates the existence of a near singularity or strong linear dependence between model parameters while a condition number closer to 1 suggests that the parameters are, for the most part, linearly independent of one another. I computed condition numbers for the sample correlation matrices from one two-cycle data fitting of each of the five models. I found that the condition numbers were consistent across models (model 1 - 118.891, model 2 - 134.363, model 3 - 141.360, model 4 - 143.261, model 5 - 118.189) and did not exceed 144. Further analysis of the eigenvalues is required to determine the extent of linear dependence between parameters in the model posterior.

Similar results were obtained from fitting three cycles of budding data as opposed to two cycles (results not shown). Considering the sampling variability inherent of later cycles in a synchrony experiment, this result is not surprising (Figure 3.9). More specifically, due to asynchrony in the population, the probability of sampling a budded cell is closer to 0.5 the longer the synchrony experiment is conducted. At that point, binomial sampling variability is also at its maximum, complicating the distinction of different models from the second cycle onward. The differences between competing models are subtle before even factoring in this sampling variability. Perhaps expectedly, these results for the correlation parameters suggest that it might be difficult to extract information about single cells (at least in their dependence in

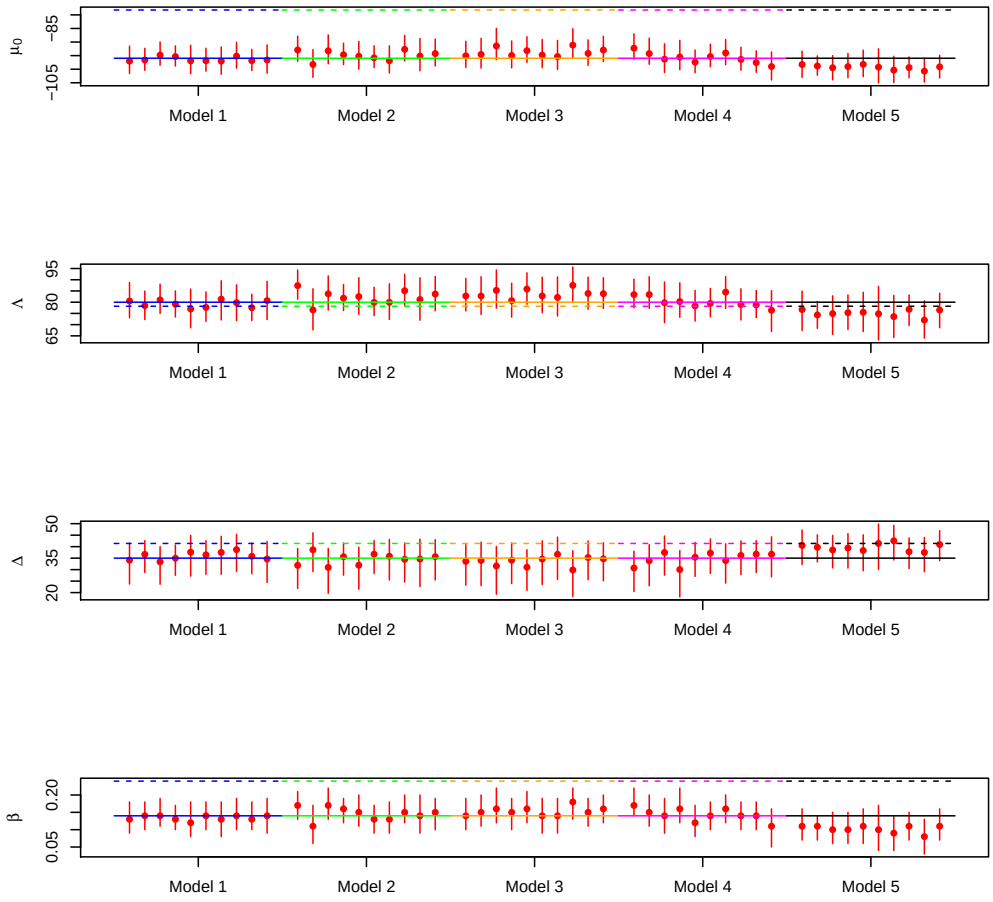

FIGURE 3.6: Shown are inferences for population model parameters μ_0, Λ, Δ, and β based on two cycles of simulated data with 200 cells sampled at time points 8 minutes apart. Each red dot and bar represents the posterior mean and 95% posterior credible interval for the plotted parameter from the fitting of one two-cycle dataset simulated under one of the models. Models are represented here by colored lines, and the colors correspond to the colors of the true population budding curves in Figure 3.4. Solid lines demarcate the true values of the plotted parameter under the particular model while dashed lines indicate the prior mean of the plotted parameter. Nearly all credible intervals overlap the true model parameter values suggesting that the population model fitting is performing adequately. Also, the inferences appear internally consistent as the posterior means and credible intervals within each model seem to fall on a line with one another. Though the credible intervals do for the most part overlap the true parameter values, the parameter inferences appear biased for some models (e.g. model 3 - orange lines, model 5 - black lines)

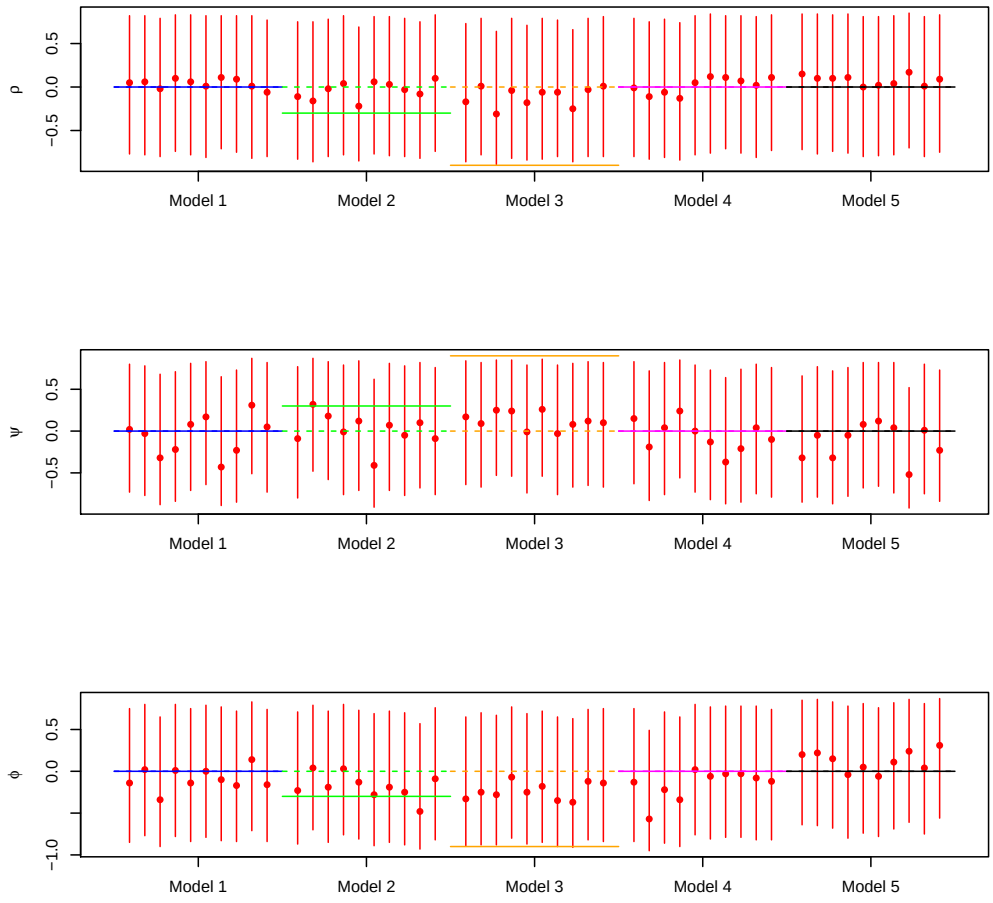

FIGURE 3.7: Shown are inferences for the branch length correlation parameters ρ, ψ, and ϕ based on two cycles of simulated data with 200 cells sampled at time points 8 minutes apart. As indicated by wide credible intervals for model parameters that include the value 0.0, the population model cannot recover the true values of the correlation parameters. Instead, the inferences seem to be closer to the prior mean suggesting limited information in the data. Line colors and type are as described in Figure 3.6.

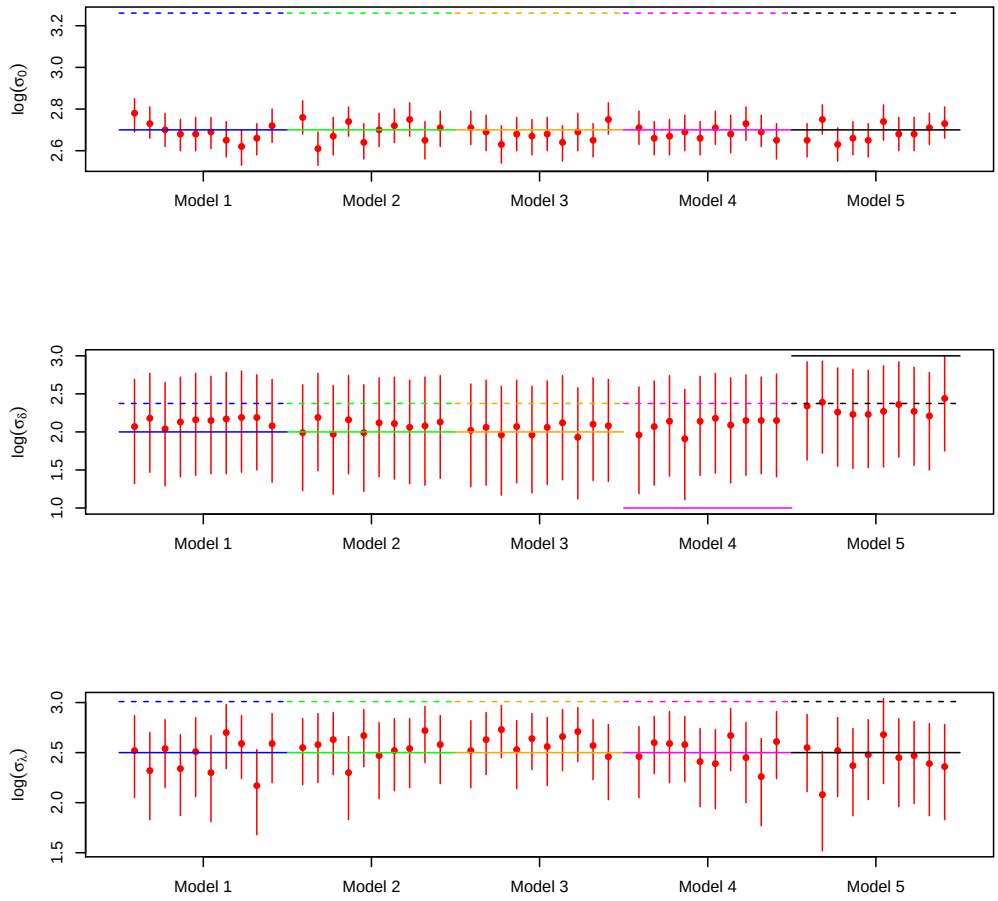

FIGURE 3.8: Inferences for the model standard deviation parameters, σ_0, σ_λ, and σ_δ based on two cycles of simulated data with 200 cells sampled at time points 8 minutes apart. While inferences for σ_0 and σ_λ accurately capture the true parameter value, the inferences for σ_δ (middle panel) appear to be influenced by the prior distribution and are removed from their true values under models 4 and 5 (magenta and black lines). Line colors and type are as described in Figure 3.6.

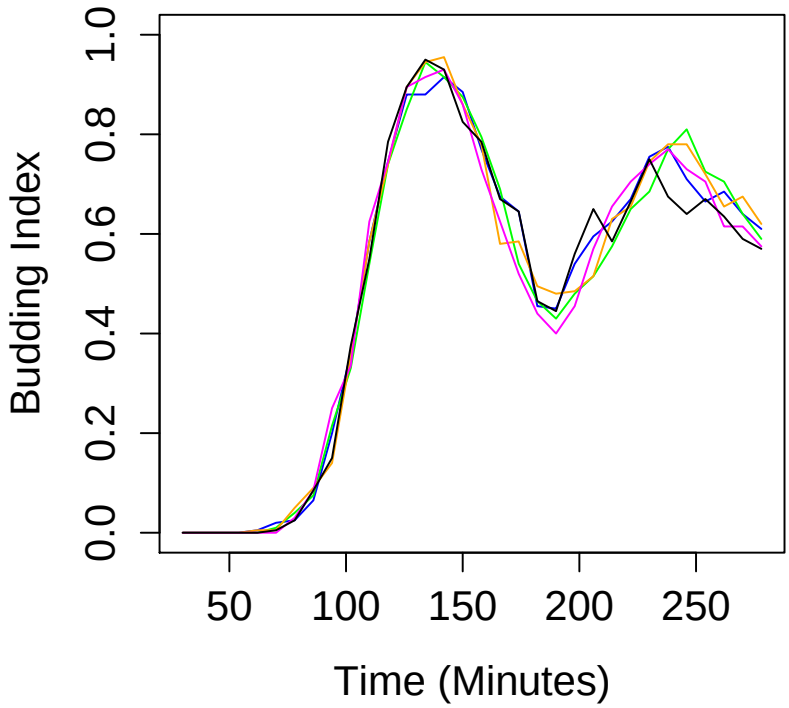

FIGURE 3.9: Shown are proportions of budded cells (out of 200 cells) sampled from the expected budding curves in Figure 3.4. The colors of the five different curves correspond to the five models for which simulated data was generated. As shown, the models are indistinguishable from one another during the first cycle (first peak). However, sampling variability complicates distinction of the models in the second cycle (second peak).

cell-cycle progression on one another) from population-level measurements.

3.3.3 Discussion of Results from Simulation Study

The results suggest, at least under the experimental settings in the simulation study, that information content in the data were not sufficient to distinguish the true underlying cell-cycle models from one another. Posterior inferences for the correlation parameters and σ_δ^2 appeared to be largely influenced by their prior distributions

and were removed from their true values. The observation of an additional cycle of data did not appear to affect these inferences. However, as previously mentioned, other experimental parameters can be tuned beyond the number of cycles observed: namely, the number of cells counted at each time point and the number and spacing of sampled time points. In this analysis, I have only just begun to explore the effect of sampling variability on model fitting. While the budding success probability curves under the different models appear (at least by eye) to be distinguishable from one another, sampling variability could potentially blur this distinction especially at points in the synchrony experiment at which the success probability is close to 0.5 and the total number of counted cells is relatively small (e.g. < 200). Using the simulator as a foundation for further analysis, one can explore the effects of these different design parameters on estimation and power to distinguish different cell-cycle models. Such results would have important implications for experimental design and could provide useful information to experimenters to optimize their collection of cell-cycle progression data.

Another area of investigation involves the apparent posterior dependencies between parameters under certain models. For example, when the true value of σ_δ^2 was large such as in simulated model 5 (black line in Figure 3.8), estimates for the variance were biased and consequently estimates for other model parameters were biased. While an analysis of the correlation between posterior samples did not reveal strong evidence for collinearity among the parameters, it is possible that σ_δ^2 shares a non-linear dependence with other model parameters such as Δ and β. Δ, β, and σ_δ^2 each describe some aspect of daughter G1 phase cell-cycle progression and so might be correlated with one another in some way.

4

Future Directions and Conclusions

This work is a first step towards more in-depth statistical characterization of the cell cycle. We've developed new models for cell-cycle dynamics at both the single cell and population levels that more flexibly fit a variety of types of cell-cycle observations and allow researchers to more effectively explore the connections between cell division and other biological processes.

4.1 Considerations for the Hierarchical Model of Single-Cell Division

4.1.1 Extending the Single-Cell Hierarchical Model to Fit Growth and Division Measurements

The single-cell analysis provides evidence that cell-cycle progression is correlated between related cells at least under certain experimental conditions. The hierarchical model of single-cell division was developed to fit budding and division observations. However, it's important to note that cell-specific growth measurements based on a fluorescent protein reporter construct are also available for the same cells (Di Talia et al. (2007)). The reporter construct facilitates analysis of a constitutively expressed protein thereby providing a proxy measurement for cell mass. Coordination between

cell growth and division has been established in the cell-cycle literature though the mechanism by which the processes are coupled is still under investigation (Johnston et al. (1977); Polymenis and Schmidt (1997); Di Talia et al. (2007); Ferrezuelo et al. (2012)). Moreover, the extent to which growth and division parameters depend on one another across cells in a lineage has not been determined. One potential future direction would be to extend the hierarchical model to fit growth measurements in addition to budding and division times.

Based on experimental studies, one model for cell growth is that single cells grow exponentially. More formally:

$$M_{Div} = M_0 e^{\alpha t_{Div}} \tag{4.1}$$

where M_0 and M_{Div} are the sizes (or masses) of the cell at birth and division respectively and α is the cell's growth rate. These parameters can be incorporated into the asymmetric branching process model as cell-specific parameters arising from some population distribution in a hierarchy. To investigate the dependence between growth and division, one could expand the multivariate normal distribution of the cell-specific branch lengths to also include the growth parameters, introducing parameters to capture correlation between growth rates, birth sizes, and cell-cycle durations within each lineage. Such an analysis would not only permit more in-depth analysis of the connections between growth and division but also generate experimentally testable hypotheses about the effects of nutrient conditions and molecular backgrounds on the coordination between the processes.

4.1.2 Application of the Model to General Cellular Characteristics

The branches in the asymmetric autoregressive branching process correspond to cell-cycle durations. However, the hierarchical model could also be used to describe other quantitative cellular characteristics (e.g. Hawkins et al. (2009)). Time-lapse

microscopy studies have become more common and have been conducted in a variety of symmetrically and asymmetrically dividing organisms. Thus, with some modifications, the hierarchical model could be used to fit other cellular measurements such as numbers of intracellular organelles or expression of different fluorescent protein reporters indicating biological process activity or DNA content. Such an analysis would require careful consideration and design of the fluorescent reporter constructs used in the time-lapse experiment. Also, some adjustments to the model would have to be made for non-normal observations (e.g. numbers of intracellular organelles). In such cases, generalized linear models could be adapted to help specify the likelihood of the cellular observations.

4.1.3 Accounting for Replicative Age of Cells in Hierarchical Model

In the hierarchical model, population average baseline cell-cycle duration (Λ) is considered constant over the course of the time-lapse experiment. However, experimental studies have suggested that average cell-cycle duration might drift with replicative cell age (Egilmez and Jazwinski (1989); Lee et al. (2012)). In other words, the more divisions a cell undergoes, the longer on average it takes to complete cell division. However, mean cell-cycle duration remains roughly constant for at least the first 10 cycles of a cell's life (Egilmez and Jazwinski (1989); Lee et al. (2012)). Furthermore, owing to the exponential growth of the population, younger cells dominate older cells after continued rounds of division. Thus, the probability of finding a cell of age 10 or greater becomes vanishingly small. Nevertheless, it would be straightforward to extend the hierarchical model to allow for drift in population average cell-cycle duration due to replicative aging (or other genealogical characteristics). The observations could be augmented to include cell ages and an additional layer in the hierarchy would be introduced in which age-specific cell-cycle duration means (Λ_a)

arise from a population distribution:

$$\Lambda_a \sim \mathsf{Norm}(\Lambda, \sigma_a^2) \tag{4.2}$$

Below this new level of the hierarchy would be the asymmetric autoregressive branching process of the initial specification with the modification that the conditional mean of each $\lambda_{i,j}$ and $\delta_{i,j}$ given its mother's λ would include these age-specific means (Λ_a) as opposed to the overall population mean (Λ).

4.1.4 Comparison of Hierarchical Model with Bifurcating Autoregressive (BAR) Models

The hierarchical model of single-cell division was developed based on previous work in populations of cells (Orlando et al. (2007); Orlando et al. (2009); Mayhew et al. (2011)). During the course of this research, I discovered the BAR or bifurcating autoregressive family of models (Cowan and Staudte (1986); Huggins and Basawa (1999); da Saporta et al. (2011)). The BAR family is a natural extension of the AR or autoregressive family of models used to model time series data, and BAR models have been successfully applied to joint analysis of independent bacterial and mammalian cell lineages (Staudte et al. (1996); Huggins and Basawa (1999)). Our model has similar structure. In fact, when mother and daughter branch lengths are constrained to have the same distribution and upon integrating out the cell-specific branch lengths ($\tilde{\lambda}$) in the hierarchical model, one can recover the BAR model. The chief differences between our model and the BAR model are twofold. First, parameter estimation in the BAR setting up to this point has largely been classical, based on maximum likelihood, method of moments, and least squares estimation. We instead opted for Bayesian inference to allow for incorporation of prior knowledge about cell division and to regularize parameter estimates for datasets (such as the 6xCLN3 dataset) comprised of smaller numbers of lineage trees. Second, and most importantly, we make a stricter assumption of zero conditional correlation in sister

cell λ's given the mother's λ. Conversely, the BAR model allows for nonzero conditional correlation in sister cell characteristics, making the compelling argument that sisters might correlate in those characteristics due to their shared environment or some inheritance mechanism independent of the mother's cell-cycle duration. More formally, sister-sister correlation in λ's in the BAR setting (borrowing notation from our hierarchical model) is

$$\rho_{sis} = \rho\psi + (1 - \rho\psi)\gamma \tag{4.3}$$

where γ is the conditional sister-sister correlation coefficient. In our model, $\gamma = 0$ and so $\rho_{sis} = \rho\psi$.

We assumed zero conditional correlation between sisters because of the fact that budding yeast cells divide asymmetrically, producing inherent cellular and molecular differences between mothers and daughters. While the hierarchical model does allow for nonzero marginal dependence between sister cell-cycle durations, it is possible that under certain experimental conditions or in specific genetic backgrounds sisters may be correlated in their cell-cycle progression (at least in part) in a non-inheritable way. Hence, it would be useful to allow for this nonzero conditional dependence between sisters by introducing a new correlation parameter in the model (the γ parameter as shown above; equivalent to ϕ in the BAR literature; Cowan and Staudte (1986)). Introducing this correlation parameter will allow inference of other types of dependence in cell-cycle progression, and results will be more comparable to previous analyses with the BAR model.

4.2 Considerations for Population-based Models of Cell Division

4.2.1 Information in Population Data from Synchrony Experiments for Correlation Parameters and Variation in Daughter G1 Extensions

A shown by parameter estimates, the population data–even under very different true models–appeared weakly informative for the correlation parameters (ψ,ρ, and ϕ) and

σ^2_δ. For the correlation parameters, 95% HPD intervals were wide and included 0. For σ^2_δ, inferences were consistently closer to the prior mean than the true value. While curves of budding probability over time under different models could be distinguished by eye (Figure 3.4), sampling variability in actual budding observations might be making identification of these different models more difficult (Figure 3.9). Inferences for the correlation parameters did not change when three cycles of data were fit rather than just two cycles. However, the number of cell cycles observed is just one tunable parameter of the synchrony experiment. Other aspects of synchrony experiments like the number of cells counted at each time point and the number and spacing of time points were not varied in the simulation study. Increasing the frequency of sampling as well as the number of cells counted might increase power to distinguish different underlying models. Future work should expand the simulation study to look at the effects of varying these experimental factors on model fitting. This analysis will have important implications for design of synchrony experiments, identifying vital points in the time course at which it would be advisable to count more cells more frequently.

4.2.2 More Flexible Representations of the Initial Cell Population

One potential reason for poor fit to population measurements (experimental, not simulated) is the assumption that the initial population distribution is a symmetric normal distribution. In fact, different types of synchronization procedures can affect the shape of the initial population distribution. For example, synchronization with the mating pheromone, α-factor, tends to produce a sharply peaked initial population distribution around the point of cell-cycle entry, START (Hartwell et al. (1974)). However, it is unclear whether the initial population is symmetrically distributed around this point, and, more than likely, some cells might still be lagging in earlier stages of G1 resulting in a skew towards the beginning of the cell cycle (Figure 4.1). The extent of this skew often depends on the amount of time the cell

population spends in α-factor treatment before release into the cell cycle. In contrast, when synchronizing by centrifugal elutriation, cells are selected by size with the smallest cells forming the initial population. The assumption behind this approach to synchronization is that small cells will be concentrated in the G1 phase. While this is not an unreasonable assumption, small cells will undoubtedly show heterogeneity in their cell-cycle position within G1. Moreover, some "small" cells may already be budded and so will fall outside G1 phase.

Hence, one direction for follow-up with CLOCCS is extending the model to have a more flexible representation that can capture technical and biological heterogeneity in initial cell-cycle position. A straightforward approach would be to represent the initial population with a finite mixture of normal distributions (Escobar and West (1995)). In this way, the model would treat the population as being composed of different subpopulations with each subpopulation following its own branching process and corresponding to one component of the initial mixture. In fitting a more flexible model to the initial population, the immediate concern is a potential sacrifice of biological interpretability in exchange for modeling artifice that produces better fits to data. To address this issue, one can invoke highly structured priors on the component means and variances as well as cap the number of components in the mixture to not allow too much flexibility and to fit strongly unimodal (though not necessarily symmetric) initial population distributions (Roeder and Wasserman (1997)).

4.2.3 Other Extensions to CLOCCS

Finally, CLOCCS is currently used to independently fit measurements from experimental replicates in which the cells have been synchronized in G1 phase (e.g. α-factor treatment and centrifugal elutriation). As has been previously noted (Orlando et al. (2009)), parameters of the branching process for each replicate can be tied in a

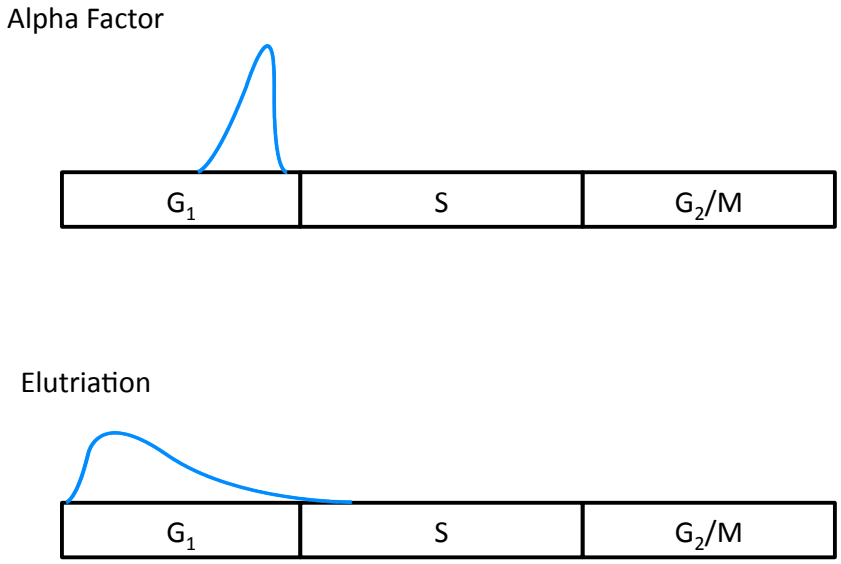

FIGURE 4.1: Shown are (idealized) initial population distributions resulting from two common methods of synchronization. Treatment with the mating pheromone, α-factor produces a sharply peaked and likely skewed initial population distribution (outlined in blue) late in G1 phase (top panel). On the other hand, centrifugal elutriation preferentially selects small cells. While the cell-cycle position distribution of smaller cells is likely concentrated in G1 phase, the initial position distribution of such a cell population might be more diffuse than in an α-factor treatment (bottom panel).

hierarchical model, facilitating sharing of information across replicated synchrony experiments and analysis of cell-cycle progression in a given experimental setting. Another potential direction worth pursuing is extending the CLOCCS model to accommodate M phase synchronizations which are commonplace in mammalian cells (e.g. Lane et al. (2013)). Such an extension would require modifications in the description of the initial cell-cycle position distribution since cells would be concentrated at the S/G2/M boundary of the first cycle.

4.3 Conclusions

Taken together, this work provides fertile ground for continued development of statistical model-based approaches to characterize cell-cycle progression. Coupled with increasingly available high-resolution time-lapse microscopy datasets, methods for integration of data across single-cell and population scales, and development of better microscopy-based approaches for observing gene expression dynamics, these models provide a powerful platform not just for cell-cycle analysis but for analysis of general cellular phenomena–and underlying molecular interactions–that comprise many dynamical biological processes. Indeed, as newer forms of molecular and cellular data continue to come online–giving complementary views of complex biological processes–statistical models have a vital role to play in facilitating mechanistic descriptions of biological systems and capturing uncertainty in biological model space.

Bibliography

Balbach, S. T., Esteves, T. C., Houghton, F. D., Siatkowski, M., Pfeiffer, M. J., Tsurumi, C., Kanzler, B., Fuellen, G., and Boiani, M. (2012), "Nuclear reprogramming: kinetics of cell cycle and metabolic progression as determinants of success." *PloS one*, 7, e35322.

Bell, G. and Anderson, E. (1967), "Cell Growth and Division I. A Mathematical Model with Applications to Cell Volume Distributions in Mammalian Suspension Cultures," *Biophys J.*, 7, 329–351.

Bi, E., Maddox, P., Lew, D., Salmon, E., McMillan, J., Yeh, E., and Pringle, J. (1998), "Involvement of an actomyosin contractile ring in Saccharomyces cerevisiae cytokinesis." *J Cell Biol.*, 142, 1301–1312.

Broach, J. R. (2012), "Nutritional control of growth and development in yeast." *Genetics*, 192, 73–105.

Cooper, S. (1982), "The continuum model: statistical implications." *Journal of theoretical biology*, 94, 783–800.

Cowan, R. and Staudte, R. (1986), "The bifurcating autoregression model in cell lineage studies." *Biometrics*, 42, 769–83.

Cross, F. R. and Blake, C. M. (1993), "The yeast Cln3 protein is an unstable activator of Cdc28." *Molecular and cellular biology*, 13, 3266–71.

da Saporta, B., Gegout-Petit, A., and Marsalle, L. (2011), "Parameters estimation for asymmetric bifurcating autoregressive processes with missing data." *Elec J Stats*, 5, 1313–1353.

Di Talia, S., Skotheim, J., Bean, J., Siggia, E., and Cross, F. (2007), "The effects of molecular noise and size control on variability in the budding yeast cell cycle." *Nature*, 448, 947–951.

Egilmez, N. and Jazwinski, S. (1989), "Evidence for the involvement of a cytoplasmic factor in the aging of the yeast *Saccharomyces cerevisiae*," *J Bacteriol*, 171, 37–42.

Escobar, M. D. and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–88.

Ferrezuelo, F., Colomina, N., Palmisano, A., Gari, E., Gallego, C., Csikasz-Nagy, A., and Aldea, M. (2012), "The critical size is set at a single-cell level by growth rate to attain homeostasis and adaptation." *Nature Commun*, 3, 1012.

Fraley, C., Raftery, A., Murphy, T., and Scrucca, L. (2012), "mclust Version 4 for R: Normal Mixture Modeling for Model-Based Clustering, Classification, and Density Estimation." Tech. Rep. 597, Department of Statistics, University of Washington, Seattle, Washington.

Hartwell, L. and Unger, M. (1977), "Unequal division in *Saccharomyces cerevisiae* and its implications for the control of cell division," *J Cell Biol*, 75, 422–435.

Hartwell, L., Culotti, J., Pringle, J., and Reid, B. (1974), "Genetic control of the cell division cycle in yeast," *Science*, 183, 46–51.

Hawkins, E. D., Markham, J. F., McGuinness, L. P., and Hodgkin, P. D. (2009), "A single-cell pedigree analysis of alternative stochastic lymphocyte fates." *Proceedings of the National Academy of Sciences of the United States of America*, 106, 13457–62.

Hejblum, G., Costagliola, D., Valleron, A.-J., and Mary, J.-Y. (1988), "Cell Cycle Models and Mother-Daughter Correlation," *J. Theor. Biol.*, 131, 255–262.

Huggins, R. M. and Basawa, I. V. (1999), "Extensions of the bifurcating autoregressive model for cell lineage studies," *Journal of Applied Probability*, 1233, 1225–1233.

Johnston, G., Pringle, J., and Hartwell, L. (1977), "Coordination of growth with cell division in the yeast *Saccharomyces cerevisiae*," *Exp Cell Res*, 105, 79–98.

Kass, R. and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Kubitschek, H. E. (1966), "Normal distribution of cell generation rates." *Nature*, 209, 1039–40.

Lane, K. R., Yu, Y., Lackey, P. E., Chen, X., Marzluff, W. F., and Cook, J. G. (2013), "Cell cycle-regulated protein abundance changes in synchronously proliferating HeLa cells include regulation of pre-mRNA splicing proteins." *PloS one*, 8, e58456.

Lebowitz, J. and Rubinow, S. (1974), "A Theory for the Age and Generation Time Distribution of a Microbial Population," *J Math Biol*, 1, 17–36.

Lee, S. S., Avalos Vizcarra, I., Huberts, D. H. E. W., Lee, L. P., and Heinemann, M. (2012), "Whole lifespan microscopic observation of budding yeast aging through a microfluidic dissection platform." *Proceedings of the National Academy of Sciences of the United States of America*, 109, 4916–20.

Liu, J. (ed.) (2008), *Monte Carlo Strategies in Scientific Computing*, Springer Press.

Lord, P. G. and Wheals, a. E. (1981), "Variability in individual cell cycles of Saccharomyces cerevisiae." *Journal of cell science*, 50, 361–76.

Mayhew, M., Robinson, J., Jung, B., Haase, S., and Hartemink, A. (2011), "A generalized model for multi-marker analysis of cell cycle progression in synchrony experiments." *Bioinformatics*, 27, i295–303.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953), "Equation of State Calculations by Fast Computing Machines," *The Journal of Chemical Physics*, 21, 1087.

Muzzey, D. and van Oudenaarden, A. (2009), "Quantitative time-lapse fluorescence microscopy in single cells." *Annual review of cell and developmental biology*, 25, 301–27.

Neufeld, T. P., de la Cruz, a. F., Johnston, L. a., and Edgar, B. a. (1998), "Coordination of growth and cell division in the Drosophila wing." *Cell*, 93, 1183–93.

Orlando, D., Lin, C., Bernard, A., Iversen, E., Hartemink, A., and Haase, S. (2007), "A probabilistic model for cell cycle distributions in synchrony experiments," *Cell Cycle*, 6, 478–488.

Orlando, D. A., Iversen Jr., E. S., Hartemink, A. J., and Haase, S. B. (2009), "A branching process model for flow cytometry and budding index measurements in cell synchrony experiments," *Annals of Applied Statistics*, 3, 1521–1541.

Pauklin, S. and Vallier, L. (2013), "The cell-cycle state of stem cells determines cell fate propensity." *Cell*, 155, 135–47.

Plummer, M. (2003), "JAGS: A Program for Analysis of Bayesian Graphical Models Using Gibbs Sampling." in *Proceedings of Workshop on Distributed Statistical Computing 2003*, eds. H. K., F. Leisch, and A. Zeileis.

Polymenis, M. and Schmidt, E. (1997), "Coupling of Cell Division to Cell Growth by Translational Control of the G1 Cyclin CLN3 in Yeast." *Genes Dev*, 11, 2522–2531.

Powell, E. O. (1955), "Some Features of the Generation Times of Individual Bacteria," *Biometrika*, 42, 16–44.

Raftery, A. E. and Lewis, S. M. (1992), "One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo," *Statistical Science*, 7, 493–497.

Rigney, D. (1987), "Inherited Rate Model of the Cell Cycle: Kinetics of Related Cells, Epi-Genetics of Ribosomal DNA Transcription and the Evaluation of Cancer-Therapy Fractionation Schedules and Doses," *Comput Math Applic*, 14, 699–739.

Ripley, B. (ed.) (1981), *Spatial Statistics*, John Wiley & Sons, Inc.

Roeder, K. and Wasserman, L. (1997), "Practical Bayesian Density Estimation Using Mixtures of Normals," *Journal of the American Statistical Association*, 92, 894–902.

Sisken, J. and Morasca, L. (1965), "Intrapopulation Kinetics of the Mitotic Cycle," *Journal of Cell Biology*, 25, 179–189.

Smith, J. a. and Martin, L. (1973), "Do Cells Cycle?" *Proceedings of the National Academy of Sciences*, 70, 1263–1267.

Staudte, R. G., Zhang, J., Huggins, R. M., and Cowan, R. (1996), "A reexamination of the cell-lineage data of E. O. Powell." *Biometrics*, 52, 1214–22.

Strathern, J., Jones, E., and Broach, J. (eds.) (1981), *The Molecular Biology of the Yeast* Saccharomyces*: Life Cycle and Inheritance*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA.

Tsien, R. Y. (1998), "The green fluorescent protein." *Annual review of biochemistry*, 67, 509–44.

Tyson, J. and Hannsgen, K. (1986), "Cell growth and division: a deterministic/probabilistic model of the cell cycle," *Journal of Mathematical Biology*, 23, 231–246.

Vanoni, M., Vai, M., Popolo, L., and Alberghina, L. (1983), "Structural heterogeneity in populations of the budding yeast Saccharomyces cerevisiae." *Journal of bacteriology*, 156, 1282–91.

Webb, G. F. (1987), "Random Transitions, Size Control, and Inheritance in Cell Population Dynamics," *Mathematical Biosciences*, 85, 71–91.