Master of Philosophy Thesis

Statistical Science Discipline

School of Mathematical Sciences Queensland University of Technology

Estimating parameters of a stochastic cell invasion model with fluorescent cell cycle labelling using Approximate Bayesian Computation

Michael John Carr

Bachelor of Mathematics, 2019

Queensland University of Technology

Submitted in fulfilment of the requirement for the degree of Masters of Philosophy

2021

Keywords: Sequential Monte Carlo; SMC-ABC; Cell proliferation; Cell motility; Random walk model

Statement of Original Authorship

The work contained in this thesis has not been previously submitted to meet the requirements for an award at this or any other higher education institution. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

Signature: _____ Date: _____

Acknowledgements

I would like to thank my supervisors Chris Drovandi and Matthew Simpson who without their mentorship this thesis would not be possible. I am grateful for your patience, encouragement and knowledge which have made me into a more knowledgeable and well rounded person. Additionally, I would also like to thank you both for the financial support that you have provided me.

To Lachlan and Ben who I have been lucky enough to be friends with since starting university and had the pleasure of sharing our journey of graduating our bachelor in mathematics and now Masters of Philosophy together. I would like to thank you both for being there as a friend and as someone I could talk to about challenging problems which without would have made university half as pleasant. I would also like to thank my friends Caleb and Henry who have also been there throughout the last year and a half to offer an escape from the stresses of the Masters of Philosophy to talk and play games together.

Lastly, I would like to thank my mum, who has always been there to support, encourage and guide me throughout my life. I could not have asked more from you in helping making the achievements I have made possible and making me into the person I am today.

Abstract

We develop a parameter estimation method based on Approximate Bayesian Computation (ABC) for a stochastic cell invasion model using fluorescent cell cycle labelling with proliferation, migration, and crowding effects. Previously, inference has been performed on a deterministic version of the model using cell density data, and not all the parameters could be identified. Working with a stochastic model allows us to harness more features of experimental data, including cell trajectories and cell count data, which overcomes the parameter identifiability problem. We demonstrate that, whilst difficult to collect, cell trajectory data can provide more information about the parameters of the cell invasion model. To handle the intractability of the likelihood function of the stochastic model, we use an efficient ABC algorithm based on sequential Monte Carlo.

Contents

Chapter 1

Introduction

1.1 Motivation and Context

Cancer cells can invade into neighbouring cell tissue in a process known as metastasis. In this thesis, we focus on the dynamics of cell invasion into the surrounding tissue similar to Maini et al. (2004) and Simpson et al. (2018) where malignant cells undergo combined proliferation and migration; although some other studies (see El-Hachem et al., 2021; Painter and Sherratt, 2003) choose to focus solely on cell migration. Insights into cell invasion dynamics can be obtained through collective cell spreading experiments. These experiments usually involve growing monocultures of cells on plastic tissue culture plates and observing how the population moves and evolves under a variety of different experimental conditions (Liang et al., 2007). Cells proliferate by progressing through a four-stage sequence of phases consisting of gap 1 (G1), synthesis (S), gap 2 (G2), and mitosis (M) where the cell divides into two daughter cells, each of which return to the G1 phase (Haass et al., 2014). Of particular interest in cancer research is the effect of applied drugs on the behaviour of cancerous cells (Desoize et al., 1998; Smalley et al., 2006). In addition, understanding the effects of these applied drugs with respect to the different stages of the cell cycle is becoming increasingly important as many drug treatment methods target different phases of the cell cycle (Haass & Gabrielli, 2017).

Fluorescent Ubiquitination-based Cell Cycle Indicator (FUCCI) technology (Sakaue-Sawano et al., 2008) allows us to visualise phases of the cell cycle in real time through the use of two fluorescent probes. When cells are in the G1 phase the probes emit a red fluorescence and when in the S/G2/M phases the probes emit a green fluorescence. Additionally, during the transition between G1 and S phase, both probes are active (giving the impression that the cell fluoresces yellow), allowing the visualisation of the early S phase, which we refer to as eS. This allows the visualisation of three unique phases in the cell cycle. Still images of a scratch assay experiment using WM983C

FUCCI-transduced melanoma cells are presented in Figure 1.1. In Figure 1.1 (c)-(f), cells appear to gradually migrate into the scratched region and proliferate with the abundance of space and nutrients as the experiment progresses. We note that the fluorescent intensity of the cells in Figure 1.1 (c)-(f) appears to fluctuate between cell phases. In this thesis, we choose to classify the cells into three phases previously mentioned using a classification rule based on the RGB decimal codes outlined in Section 3.5.2; however, this information may be useful to consider in future studies.

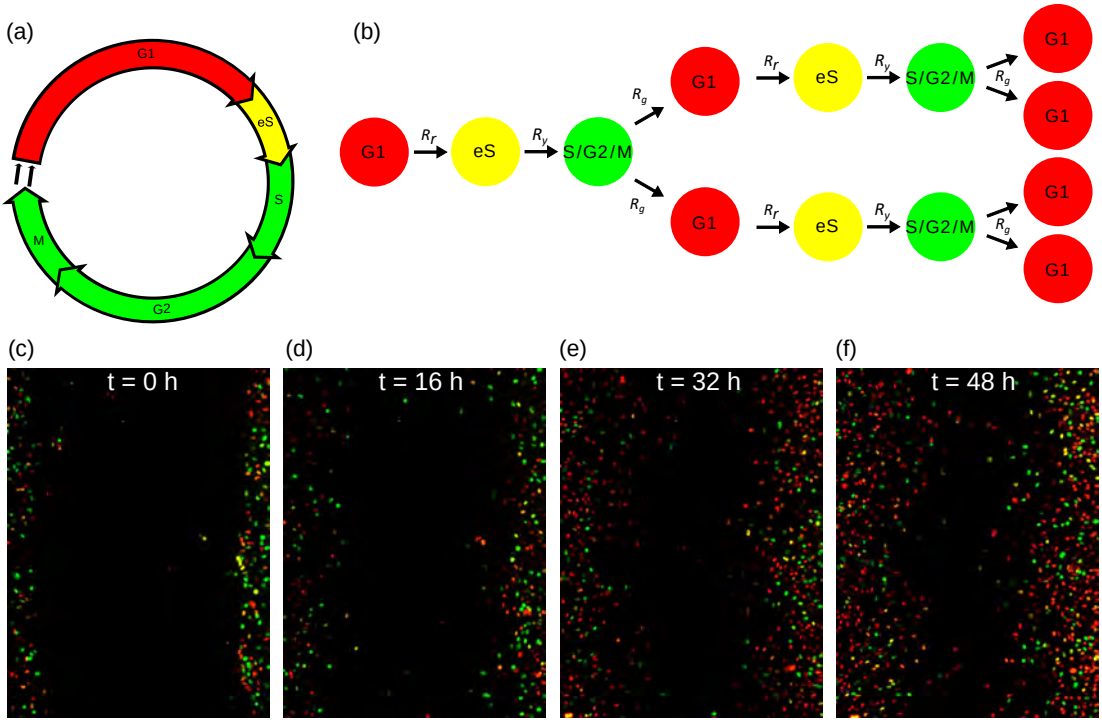

Figure 1.1: (a)-(b) Schematic of cell cycle. (c)-(f) Experimental images, all 1309.09 x 1745.35 μm, of WM983C FUCCI-transduced melanoma cells at 0, 16, 32 and 48 hours, respectively. Images reproduced with permission from Vittadello et al. (2018)

The process of reproducing these experiments with varying experimental conditions is often expensive and time consuming. An alternative to studying cell dynamics *in vitro* is the construction of mathematical models which are capable of reproducing experiment conditions virtually. There are a multitude of benefits to using these simulation models instead of *in vitro* experiments besides the obvious reduced time and expenses. Firstly, these models can be extended or altered in a variety of ways to aid in the development of theoretical mechanisms. Secondly, such mechanisms incorporated into the model are usually controlled by a collection of parameters, which can be quantified through statistical inference methods. Lastly, hypotheses can be tested by comparing model predictions with experiment data.

Collective cell spreading and invasion models usually use a deterministic modelling framework where the spatial location of cells in the experiment are defined by partial differential equations (PDE). These PDE's are generally extensions of the Fisher-

Kolmogorov-Petrovsky-Piscounov (Fisher-KPP) model (Fisher, 1937; Kolmogorov, 1937) which is a reaction-diffusion model. This modelling approach includes a diffusion source term which describes the cells movement and a logistic source term which describes proliferation with respect to the carrying capacity of the environment (Edelstein-Keshet, 2005; Murray, 2007). However, these modeling approaches are usually unable to accommodate multiple data types and can lead to parameter identifiability issues if the informativeness of the data that the model can produce is insufficient (see for example Simpson et al., 2020). Alternatively, a stochastic modelling approach can be used to consider individual cell behaviour (Codling et al., 2008). These models are constructed from Markov processes and often allow for a wider range of data types to be considered (e.g. cell trajectories).

In this thesis, we use a stochastic modelling approach proposed by Simpson et al. (2018) to mimic the the proliferation and movement of cells within a scratch assay experiment. This modelling approach involves describing a discrete exclusion based (meaning no two agents can occupy the same site) random walk on a two-dimensional (2D) hexagonal lattice. This model has yet to be calibrated to experimental data and have the unknown cell cycle transition and motility rate parameters estimated. We extend on the work of Simpson et al. (2018) by estimating these parameters through a Bayesian framework. Bayesian methods achieve this through estimating the posterior distribution which is a function of the likelihood of the data and the prior information. However, stochastic models of collective cell invasion and migration are often so complex that standard parameter estimation procedures are not feasible due to the intractability of the likelihood function. We overcome this limitation by applying Approximate Bayesian Computation (ABC methods). These methods bypass evaluating the intractable likelihood function by identifying parameter configurations which produce simulated data that closely resembles the observed data; where often the simulated and observed data are reduced to a set of low dimensional summary statistics. With respect to previous collective cell spreading modelling approaches (see Cai et al., 2007; Maini et al., 2004; Savla et al., 2004; Simpson et al., 2020; Swanson, 2008; Vo et al., 2015), our study is the first of its kind to successfully estimate the parameters of a stochastic cell invasion model with multiple phases of the cell cycle.

1.2 Aims and Objectives

The overall aim of this project is to develop a parameter estimation method for estimating parameters of a stochastic cell invasion model which considers proliferation, migration and crowding effects. By considering the stochastic model we can harness more features of experimental data, including cell trajectories and cell count data, which cannot be done with deterministic modelling. We achieve this aim through the

following objectives:

1. Assess the suitability of the existing ABC methods and consider their suitability to the stochastic cell invasion model
2. Analyse and compare the informativeness of various summary statistics with several biologically plausible synthetic data sets
3. Estimate the model parameters using a suitable set of summary statistics and likelihood-free inference method.
4. Assess the increase in information that can be obtained about the parameters by harnessing the additional cell trajectory and count data, which is not considered in previous deterministic modelling approaches.

1.3 Thesis Structure

In Chapter 2, we provide a background into Bayesian statistics, sampling methods and their likelihood-free adaptations to give the reader a better appreciation and understanding for the algorithms we later use to estimate model parameters in Chapter 3.

In Chapter 3, we describe how to estimate parameters of a stochastic collective cell spreading model by using likelihood-free inference methods. This involves a review of previous modelling methods, a review of the likelihood-free inference methods (objective 1), the procedure on how to calibrate the model to experimental data and data extraction, comparison of the informativeness of multiple data types (objective 2), estimation of the model parameters and validation (objective 3), and discussion on results (objective 4). The contents of this chapter have been published in *Journal of the Royal Society Interface* (see Carr et al., 2021).

Finally, in Chapter 4 we summarise our the main findings from the thesis and provide insight into limitations and possible future directions.

Chapter 2

Statistical Background

2.1 Bayesian Statistics

2.1.1 Bayes' Theorem

In Bayesian Statistics, the uncertainty in the unknown model parameters $\theta = (\theta_1, ..., \theta_p)^\top$ (where $\theta \in \Theta \subseteq \mathbb{R}^p$ and p is the number of parameters) conditional on the data $y = (y_1, ..., y_m)^\top$ (where $y \in Y \subseteq \mathbb{R}^m$ and m is the dimension of the data) can be quantified by the posterior distribution, given by Bayes' theorem:

$$\pi(\theta|y) = \frac{\pi(y|\theta)\pi(\theta)}{\int_\Theta \pi(y|\theta)\pi(\theta)\mathrm{d}\theta}$$

where $\pi(y|\theta)$ is the likelihood function, $\pi(\theta)$ is the prior and $\int_\Theta \pi(y|\theta)\pi(\theta)\mathrm{d}\theta$ is the normalising constant. Generally the normalising constant can be computationally intractable but is seldom required for applications where the focus is sampling from the posterior.

Their are numerous advantages to using Bayesian statistics (for a comprehensive review see Carlin and Louis, 2000; Gelman et al., 2013). To name a few, the inclusion of previous knowledge or expert opinion about the model parameters can be incorporated into the inference process through the prior distribution. Additionally, the uncertainty in the model parameters and predictions can be modeled with probabilistic distributions rather than frequentist point estimates and confidence intervals. These probabilistic distributions are generally more intuitive and contain much more information about the uncertainty in parameters. For these reasons, we adopt a Bayesian approach to inference in this thesis.

2.1.2 Monte Carlo

In Bayesian statistics the main objective is to reveal information about the unknown quantity θ. This can be achieved by computing the expectation with respect to the posterior:

$$\mathbb{E}[h(\theta)] = \int_{\Theta} h(\theta)\pi(\theta|y)\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N} h(\theta_i),$$

where $h(\theta)$ is the unknown quantity and it is assumed θ_i can be independently sampled from $\pi(\theta|y)$. This is known as Monte Carlo integration (Robert & Casella, 2013). Commonly $h(\theta)$ is set as θ or $(\theta - \mathbb{E}[\theta])^2$ to estimate the posterior mean and variance, respectively.

2.1.3 Markov Chain Monte Carlo

In practice, using Monte Carlo methods to the quantify the unknown parameter θ is often not possible due to the difficulty in independently sampling from the posterior distribution. However, we can construct an ergodic Markov chain with the posterior distribution as its stationary distribution to generate T samples from the posterior; although, these samples are still not independent. Nevertheless, we ensure that

$$\lim_{t\to\infty} \frac{1}{T}\sum_{t=1}^{T} h(\theta_t) \to \int_{\Theta} h(\theta)\pi(\theta|y)\mathrm{d}\theta,$$

by the Markov chain central limit theorem (Tierney, 1994) since we are using an ergodic Markov chain.

One of the most well know Markov Chain Monte Carlo (MCMC) sampling algorithms is the Metropolis-Hastings algorithm (Hastings, 1970; Metropolis et al., 1953) (see Algorithm 1). This method proposes θ^i, for $i = 1, ..., N$, from a proposal distribution, $q(\theta^i|\theta^{i-1})$, conditional on the current value θ^{i-1}. The proposed value, θ^i, is accepted as the next value in the chain with probability

$$p_{\text{acc}} = \min\left(1, \frac{\pi(y|\theta^i)\pi(\theta^i)q(\theta^{i-1}|\theta^i)}{\pi(y|\theta^{i-1})\pi(\theta^{i-1})q(\theta^i|\theta^{i-1})}\right),$$

otherwise the current value, θ^{i-1}, is accepted as the next value in the Markov chain.

Implementation of the Metropolis-Hastings algorithm requires an initial value θ^0, a proposal distribution $q(\theta^i|\theta^{i-1})$ and the total number of iterations T to be specified. The choice of θ^0 is somewhat arbitrary as after a large enough number of iterations the chain will converge to the stationary distribution; however this requires samples recovered prior to convergence to be discarded (referred to as burn-in) to avoid sample bias. Otherwise, θ^0 could be chosen such that it is already in the stationary distribution (see for example Flegal and Herbei, 2012) to improve efficiency. The choice of the

proposal distribution $q(\theta^i|\theta^{i-1})$ should be chosen such that the entire sample space of θ is explored. A simple choice is a multivariate normal distribution, $\mathcal{N}(\theta^i; \theta^{i-1}, \Sigma)$, with Σ as a tuning parameter. Although, more efficient proposal distributions which consider gradient information could be used instead to improve performance in higher dimensions (for example see Hoffman and Gelman, 2014; Roberts and Stramer, 2002). To tune the proposal distribution, a good rule of thumb to use when the proposal distribution is of the same shape as the target distribution is to target an acceptance rate between 23.4-44% (depending on the dimension of the parameter space) (Gelman et al., 1996). The choice of T is problem specific and will largely depend on the autocorrelation in samples and the desired level of precision in estimates. The Metropolis-Hastings algorithm is presented in Algorithm 1.

Algorithm 1 Metropolis-Hastings Algorithm

1: Initialise θ^0
2: **for** $i = 1$ to T **do**
3: Propose $\theta^i \sim q(\cdot|\theta^{i-1})$
4: Compute $p_{\text{acc}} = \min\left(1, \frac{\pi(y|\theta^i)\pi(\theta^i)q(\theta^{i-1}|\theta^i)}{\pi(y|\theta^{i-1})\pi(\theta^{i-1})q(\theta^i|\theta^{i-1})}\right)$
5: Accept proposal θ^i with probability p_{acc} otherwise set $\theta^i = \theta^{i-1}$ (reject proposal)
6: **end for**

2.1.4 Importance Sampling

Importance sampling offers refinement to Monte Carlo methods in situations when samples are difficult or unable to be drawn from the posterior distribution. The basic principal is to instead sample from an importance distribution $g(\theta)$ which is easier to sample from. To transform these samples from the importance distribution to be samples from the posterior we first calculate the weights $w_i = \pi(\theta_i|y)/g(\theta_i)$, then the normalised weights $\tilde{w_i} = w_i / \sum_j w_j$ and then weight the samples with respect to the normalised weights. In effect, importance sampling computes the expectation:

$$\mathbb{E}[h(\theta)] = \int_\Theta h(\theta)\pi(\theta|y)\mathrm{d}\theta = \int_G h(\theta)\frac{\pi(\theta|y)}{g(\theta)}g(\theta)\mathrm{d}\theta \approx \frac{1}{N}\sum_{i=1}^{N}\tilde{w_i}h(\theta_i),$$

where $h(\theta)$ is the unknown quantity and θ_i is independently sampled from the importance distribution $g(\theta)$.

A requirement of importance sampling to work well is that the importance distribution should be close to the posterior distribution but with heavier tails. However, this is often difficult to find in practice as information about the surface of the posterior is often unknown. An extension to importance sampling, known as adaptive importance sampling, uses a family of importance distributions $g_{\phi_t}(\theta)$ with hyperparameter ϕ_t that is sequentially tuned from the weighted sample $\{\tilde{w}_i^{t-1}, \theta_i^{t-1}\}_{i=1}^N$ to bring it closer to the posterior. We can compare the performance of each importance distribution

$g_{\phi_t}(\theta)$ by computing the effective sample size ESS $= 1/\sum_i \tilde{w}_i^2$ between iterations. We continually iterate importance distributions bringing the importance distribution closer to the posterior until the change in ESS is below a pre-specified tolerance. We present the adaptive importance sampling algorithm in Algorithm 2.

Algorithm 2 Adaptive Importance Sampling Algorithm

1: Set t = 0 and initialise ϕ_0
2: Sample $\{\theta_i\}_{i=1}^N$ from g_{ϕ_t}
3: Compute sample weights $w_i \propto \pi(\theta_i|y)/g_{\phi_t}(\theta_i)$, for $i = 1, ..., N$
4: Compute normalised weights $\tilde{w}_i = w_i/\sum_j w_j$, for $i = 1, ..., N$
5: Compute ϕ_t from $\{\tilde{w}_i, \theta_i\}_{i=1}^N$
6: Compute ESS $= 1/\sum_i \tilde{w}_i^2$
7: Set $t = t + 1$ and repeat from 2 until there is little improvement in ESS

2.1.5 Sequential Monte Carlo

One of the shortcomings of importance sampling is the difficulty in specifying an appropriate family of importance distribution which closely resembles the posterior. Sequential Monte Carlo (SMC) offers an alternative method which forms a sequence of distributions based on likelihood annealing, given by:

$$\pi_t(\theta|y) \propto \pi(y|\theta)^{\gamma_t}\pi(\theta), \text{ for } t = 1, ..., T$$

where when $\gamma_1 = 0$ the prior distribution is used as the importance distribution and when $\gamma_t = 1$ the posterior is used. In this way, the importance distribution is gradually brought closer to the target posterior by increasing the intermediate temperatures $\gamma_2 < ... < \gamma_{T-1}$ in a smooth fashion.

One of the challenges of SMC and importance sampling is the increasing variability of the particle weights in the collection of particles $\{\tilde{w}_i^t, \theta_i^t\}_{i=1}^N$ which lowers the ESS as t increases. Chopin (2002) proposes an algorithm to address the issues of sample degeneracy by implementing a resampling and move step once the ESS falls below an undesirable threshold E. The intuition behind this procedure is that the resampling step will discard particles with insignificant weights and duplicate particles with high weights by resampling particles with probabilities given by their weights. The move step then diversifies the population with an MCMC kernel using a proposal distribution, q_t, with invariant distribution $\pi_t(\theta|y)$. Moreover, the tuning parameters for the proposal distribution can be adaptively tuned from the population of particles $\{\tilde{w}_i^t, \theta_i^t\}_{i=1}^N$. We present the SMC algorithm of Chopin (2002) in Algorithm 3

Algorithm 3 SMC Algorithm

Draw $\theta_i^0 \sim \pi(\cdot)$ and set $w_i^0 = 1/N$ for $i = 1, ..., N$
for t = 1 to T **do**
    Compute sample weights $w_i^t \propto \pi_t(\theta_i^{t-1}|y)/\pi_{t-1}(\theta_i^{t-1}|y)$, for $i = 1, ..., N$
    Set $\theta_i^t = \theta_i^{t-1}$, for $i = 1, ..., N$
    Compute normalised weights $\tilde{w}_i^t = w_i^t / \sum_j w_j^t$, for $i = 1, ..., N$
    Compute ESS $= 1/\sum_i (\tilde{w}_i^t)^2$
    if ESS $<$ E **then**
        Resample particles producing $\{\theta_i^t\}_{i=1}^N$ and reset weights to be equal to $1/N$
        Compute tuning parameters of MCMC kernel q_t
        Move θ_i^t with MCMC kernel R_t times, for $i = 1, ..., N$
    end if
end for

2.2 Likelihood-free Inference

While Bayesian statistics can be beneficial to use, a key requirement is that the likelihood function is tractable. However, this may not be the case for sufficiently complex models if the likelihood function is either too computationally expensive to compute or is not analytically tractable. For example, the stochastic model used in this thesis has an intractable likelihood function due to the computational cost of computing the matrix exponential on the high dimensional generator matrix (a matrix of rate parameters which describe the rate of transitioning between states in a Markov process). Rather than reverting to simpler models with tractable likelihoods, these types of problems can be instead analysed using likelihood-free methods that avoid evaluating the likelihood function. Likelihood-free inference is the set of methods which quantify the uncertainty in the unknown model parameters θ by simulating data from the model, $x \sim \pi(\cdot|\theta)$, and identifying which parameter values yield simulated data x that closely resembles observed data y. It can be often impractical to compare the full data sets of x and y, so likelihood-free inference methods usually reduce the full data sets down to a set of low dimensional summary statistics by some summarising function $S(\cdot)$; where the summary statistics for x and y are denoted $S_x = S(x)$ and $S_y = S(y)$, respectively. Provided that these summary statistics are highly informative about the model parameters, then $\pi(\theta|y) \approx \pi(\theta|S_y)$ is a good approximation (Blum et al., 2013). By this process, likelihood-free methods avoid having to evaluate the intractable likelihood function by instead simulating data from a model. However, these methods tend to be computationally intensive due to the large number of data sets required to be simulated. Therefore, the strengths and weaknesses of the likelihood-free inference method relative to the application should be considered carefully. In this study, the likelihood-free method we adopt is ABC, which we now explore in further detail.

2.2.1 Approximate Bayesian Computation

Introduced by Pritchard et al. (1999) and later popularised by Beaumont et al. (2002), ABC samples from the approximate posterior:

$$\pi_\epsilon(\theta|S_y) \propto \pi(\theta) \int_x K_\epsilon(\rho(S_y, S_x))\pi(x|\theta)\ \mathrm{d}x, \tag{2.1}$$

where $\rho(S_y, S_x)$ is the discrepancy function which measures the difference between the two summary statistics and $K_\epsilon(\cdot)$ is the kernel weighting function which weighs $\rho(S_y, S_x)$ conditional on the tolerance ϵ. A common choice for the discrepancy function is the Euclidean distance, $\rho(S_y, S_x) = \|S_y - S_x\|_2$, and for the kernel weighting function is the indicator function, $\mathbb{1}(\cdot)$, which is equal to one if $\rho(S_y, S_x) \leq \epsilon$ and is zero otherwise. The approximate posterior in Equation 2.1 converges to the posterior conditional on the observed summary (often referred to as the partial posterior) in the limit as $\epsilon \to 0$ (Beaumont et al., 2002).

Generally the limitations of ABC are related to the dimension of the parameter space or the number of summary statistics used being too high. Fearnhead and Prangle (2012) show that the average acceptance probability of a proposed parameter configuration is $\mathcal{O}(\epsilon^d)$, where d is the dimension of the summary statistic. Therefore, for problems which have high dimensional parameter spaces or require high dimensional summary statistics, ABC can perform poorly. However, strategies such as regression adjustment (Beaumont et al., 2002; Blum & François, 2010) and semi-automatic ABC (Fearnhead & Prangle, 2012; Harrison & Baker, 2020) can be used to improve ABC's performance in higher dimensions. Another limitation of ABC is the requirement for the tolerance, ϵ, to be small for $\pi_\epsilon(\theta|S_y)$ to be a good approximation of $\pi(\theta|S_y)$. However, as ϵ decreases the acceptance rate also decreases. This requires a greater number of model simulations to be performed which can make ABC computationally expensive or intractable when the simulation model is computationally intensive.

2.2.2 ABC sampling algorithms

The most rudimentary ABC sampling algorithm is ABC rejection (Pritchard et al., 1999). This method generates N independently and identically distributed (iid) samples from the approximate posterior by accepting parameter configurations where the discrepancy measure $\rho(S_y, S_x)$ is less than the desired target tolerance ϵ_T. Under this method, proposed parameter values are generated from the prior distribution $\pi(\theta)$. The ABC rejection algorithm is presented in Algorithm 4. ABC rejection is desirable to use because it is relatively simple to implement and acquires iid samples (Marjoram et al., 2003). However, if the prior is relative diffuse compared to the posterior, then the simulated data sets will mostly not resemble the observed data because parameter values will be predominantly sampled in regions of low posterior density (Sisson et al.,

2007). This results in very few samples being accepted and can make ABC rejection computationally prohibitive, especially when the model is computationally expensive to simulate.

Algorithm 4 ABC rejection

1: **for** $i = 1$ to N **do**
2: Propose $\theta^i \sim \pi(\cdot)$
3: Simulate $x \sim \pi(\cdot|\theta^i)$
4: Compute $S_x = S(x)$
5: Accept θ^i if $\rho(S_y, S_x) \leq \epsilon_T$
6: **end for**

Marjoram et al. (2003) developed a likelihood-free Markov chain Monte Carlo algorithm (MCMC-ABC) in an effort to increase the acceptance rates relative to ABC rejection by constructing a Markov chain with a stationary distribution identical to the ABC posterior. This algorithm proposes parameters θ^i, for $i = 1, ..., N$, from a carefully tuned proposal distribution $\theta^i \sim q(\cdot|\theta^{i-1})$, conditional on the current parameter θ^{i-1}. Proposals are to be accepted with probability p_{acc} based on the Metropolis-Hastings ratio (Hastings, 1970; Metropolis et al., 1953):

$$p_{\text{acc}} = \min\left(1, \frac{\pi(\theta^i)q(\theta^{i-1}|\theta^i)K_\epsilon(\rho(S_y, S_x^i))}{\pi(\theta^{i-1})q(\theta^i|\theta^{i-1})K_\epsilon(\rho(S_y, S_x^{i-1}))}\right). \tag{2.2}$$

The choice of the initial value, θ^0, can be chosen arbitrarily if a pilot run is conducted until the chain has converged to the stationary distribution and all samples collected prior to convergence are discarded (referred to as burn-in). While MCMC-ABC can be more efficient than ABC rejection (Marjoram et al., 2003), due to searching around locally for regions of high posterior probability, it is possible for the Markov chain to spend many iterations in areas of low posterior probability (Sisson et al., 2007). Additionally, tuning the proposal distribution can take considerable effort. The MCMC-ABC algorithm is presented in Algorithm 5.

In recognition of the disadvantages of MCMC-ABC, Sisson et al. (2007) develop a likelihood-free Sequential Monte Carlo algorithm (SMC-ABC). This method sequentially approximates the posterior with T non-increasing tolerance levels $\epsilon_1 \geq \cdots \geq \epsilon_T$:

$$\pi_{\epsilon_t}(\theta|S_y) \propto \pi(\theta) \int_x K_{\epsilon_t}(\rho(S_y, S_x))\pi(x|\theta)\, \mathrm{d}x, \text{ for } t = 1, ..., T. \tag{2.3}$$

Initially θ is independently sampled from the prior distribution and ϵ_1 is set as the maximum sample tolerance. Thereafter, to propagate samples between target distributions, θ is resampled from the previous population proportional to their weights, $\theta_t \sim \{\theta_{t-1}^i, W_{t-1}^i\}_{i=1}^N$, and perturbed according to a Markov transition kernel $\theta^i \sim M_t(\cdot|\theta^i i - 1)$. Corrections on the weighting function from Beaumont et al. (2009) sug-

Algorithm 5 MCMC-ABC

After burn-in initialise $\theta^0, S_x, \psi^0 = K_\epsilon(\rho(S_y, S_x))$
for $i = 1$ to T **do**
    Propose $\theta^i \sim q(\cdot|\theta^{i-1})$
    Simulate $x \sim \pi(\cdot|\theta^i)$
    Compute $S_x = S(x)$
    Compute $\psi^i = K_\epsilon(\rho(S_y, S_x))$
    Compute $p_{\text{acc}} = \frac{\pi(\theta^i)q(\theta^{i-1}|\theta^i)\psi^i}{\pi(\theta^{i-1})q(\theta^i|\theta^{i-1})\psi^{i-1}}$
    if $\mathcal{U}(0,1) < p_{\text{acc}}$ **then**
        Accept θ^i
    else
        Set $\theta^i = \theta^{i-1}$, $\psi^i = \psi^{i-1}$
    end if
end for

gest using $W_t \propto \pi(\theta_t)/\sum_{j=1}^{N} W_{t-1}^j M_t(\theta_t|\theta_{t-1}^j)$. However, this approach still requires considerable effort and the intermediate intermediate target distances $\epsilon_2, ..., \epsilon_T$ to be specified.

An alternative approach from Drovandi and Pettitt (2011) suggest using a weighting function $W_t \propto W_{t-1}\mathbb{1}(\rho(S_y, S_{x_{t-1}}) \leq \epsilon_t)/\mathbb{1}(\rho(S_y, S_{x_{t-1}}) \leq \epsilon_{t-1}))$, where $\mathbb{1}(\cdot)$ is the indicator function, and an MCMC kernel to perturb samples. This method results in parameter configurations only being kept if they satisfy the target tolerance ϵ_t; however, this may cause sample degeneracy (too many duplicated particles). To overcome this, θ is resampled from the best $100(1-\alpha)\%$ (with respect to minimising tolerance) until N samples are attained. However, this approach will cause particle duplication. Hence, Drovandi and Pettitt (2011) recommend iterating the MCMC kernel R_t times to guarantee sample diversity, where $R_t = \lceil \log(c)/\log(1 - \overline{p_{\text{acc}}}) \rceil$ with $\overline{p_{\text{acc}}}$ as the overall MCMC acceptance rate which can be computed from $R_{t-1}/2$ pilot iterations and the ceiling function $\lceil \cdot \rceil$ is used to be conservative. R_t is dependent on the sequence level t because as ϵ_t decreases, the number of iterations required to accept a proposal (with probability $1 - c$) tends to increase since the criterion for accepting a proposal, $\rho(S_y, S_x) < \epsilon_t$, becomes more stringent. Appropriate values for the tuning parameters are $\alpha = 0.5$ and $c = 0.01$ (Drovandi & Pettitt, 2011). The algorithm is completed once the sample maximum tolerance is less than or equal to the target tolerance ϵ_T or when the overall MCMC acceptance rate reaches an unacceptable level t_{acc}.

The main innovation of this approach is that the intermediate target distances $\epsilon_2, ..., \epsilon_T$ are adaptively set (along with the number of targets $T-2$) and are replaced by a single tuning parameter α. Furthermore, the tuning parameters for the proposal distribution, $q_t(\cdot|\cdot)$, can be dynamically computed from the $\{\theta\}_{i=1}^{N-N_a}$ samples because they are already distributed according to the next target distribution. Consequently, the main drawback of SMC-ABC is that is it relatively more complex to implement than ABC

rejection and MCMC-ABC. The SMC-ABC replenishment algorithm of Drovandi and Pettitt (2011) is presented in Algorithm 6.

Algorithm 6 SMC-ABC (Drovandi & Pettitt, 2011)

Set t_{acc}, ϵ_T and $N_\alpha = \lfloor \alpha N \rfloor$
Set S_t the initial number of pilot MCMC iterations
for $i = 1$ to N **do**
    Draw $\theta^i \sim \pi(\cdot)$
    Simulate $x^i \sim \pi(\cdot|\theta^i)$
    Compute $S_x^i = S(x^i)$
    Compute $\rho^i = \rho(S_y, S_x^i)$
end for
Sort θ by ρ such that $\rho^1 \leq \rho^2 \leq ... \leq \rho^N$
Set $\epsilon_t = \rho^{N-N_\alpha}$
while $\overline{p_{\text{acc}}} > t_{\text{acc}}$ OR $\rho^N > \epsilon_T$ **do**
    Compute tuning parameters of MCMC kernel $q_t(\cdot|\cdot)$ using $\{\theta^i\}_{i=1}^{N-N_\alpha}$
    for $j = N - N_\alpha + 1$ to N **do**
        Resample θ^j from $\{\theta^i\}_{i=1}^{N-N_\alpha}$
        for $k = 1$ to S_t **do**
            Propose $\theta^* \sim q_t(\cdot|\theta^j)$
            Simulate $x \sim \pi(\cdot|\theta^*)$
            Compute $S_x = S(x)$
            Compute MH ratio $p_{\text{acc}} = \dfrac{\pi(\theta^*)q(\theta^j|\theta^*)}{\pi(\theta^j)q(\theta^*|\theta^j)} \mathbb{1}(\rho(S_y, S_x) < \epsilon_t)$
            if $\mathcal{U}(0,1) < p_{\text{acc}}$ **then**
                Set $\theta^j = \theta^*$, $\rho^j = \rho(S_y, S_x)$ and $S_x^j = S_x$
            end if
        end for
    end for
    Calculate $\overline{p_{\text{acc}}}$ based on the overall acceptance rate from the pilot MCMC runs
    Set $R_t = \lceil \dfrac{\log(c)}{\log(1 - \overline{p_{\text{acc}}})} \rceil$
    Repeat steps 13-24 with $S_t = \max(0, R_t - S_t)$
    Set $\epsilon_t = \rho^{N-N_\alpha}$ and $S_t = \lceil R_t/2 \rceil$
end while

Chapter 3

Estimating parameters of a stochastic cell invasion model with fluorescent cell cycle labelling using Approximate Bayesian Computation

3.1 Statement of Authorship

This chapter has been written as a journal article and published in the *Journal of the Royal Society Interface*. The authors listed below have certified that:

1. They meet the criteria for authorship as they have participated in the conception, execution or interpretation of at least the part of the publication in their field of expertise;
2. They take public responsibility for their part of the publication, except for the responsible author who accepts overall responsibility for the publication;
3. There are no other authors of the publication according to these criteria;
4. Potential conflicts of interest have been disclosed to granting bodies, the editor or publisher of the journals of other publications and the head of the responsible academic unit; and
5. They agree to the use of the publication in the student's thesis and its publication on the Australian Digital Thesis database consistent with any limitations set by publisher requirements.

The reference for the publication associated with this chapter is (Carr et al., 2021), Carr, M. J., Simpson, M. J. Drovandi, C. (2021). Estimating parameters of a stochastic

cell invasion model with fluorescent cell cycle labelling using approximate Bayesian computation. Journal of the Royal Society Interface, 18(182).

Contributor	Statement of Contribution
Michael Carr	Implemented the methodology, performed all data analysis, produced all figures, supplementary material and wrote the paper. In addition, they wrote and packaged the code used within this study onto Github.
Signature and Date:	_____
Matthew Simpson	Oversaw and directed the research, provided technical assistance, aided in interpreting results and critically reviewed the paper.
Christopher Drovandi	Oversaw and directed the research, provided technical assistance, aided in interpreting results and critically reviewed the paper.

Principal Supervisor Confirmation: I have sighted email or other correspondence for all co-authors confirming their authorship.

Name: _____ Signature: _____ Date: _____

3.2 Introduction

Australia and New Zealand have the highest incidence rates of melanoma in the world, followed by northern America and northern Europe (Parkin et al., 2005). In Australia, melanoma is the third most common diagnosed form of cancer (Australian Institute of Health and Welfare, 2018). Since the 1960s, Australia's primary strategy to reduce overall mortality rates has been targeted at early prevention and detection (Giblin & Thomas, 2007). However, a better understanding of the mechanisms which control cell invasion is necessary in order to improve or establish new treatment measures.

The underlying mechanisms of cell invasion we consider are combined cell proliferation and cell migration. Cell proliferation is a four-stage sequence consisting of gap 1 (G1), synthesis (S), gap 2 (G2), and mitosis (M) where the cell divides into two daughter cells, each of which return to the G1 phase (Haass et al., 2014). Improvements in technology have enabled us to visualise different phases of the cell cycle in real time using Fluorescent Ubiquitination-based Cell Cycle Indicator (FUCCI) technology (Sakaue-Sawano et al., 2008). FUCCI technology involves two fluorescent probes which emit

red fluorescence when the cells are in G1 phase and green fluorescence when in S/G2/M phases. During the transition between G1 and S phase, both probes are active (giving the impression that the cell fluoresces yellow), allowing the visualisation of the early S phase, which we refer to as eS. Experiments using FUCCI-transduced melonama cells are becoming increasingly important in cancer research because many drug treatments target different phases of the cell cycle (Haass & Gabrielli, 2017).

The development of simulation models offer us a quick and inexpensive alternative to *in vitro* experiments. Although, existing mathematical models have had a long history without incorporating cell cycle information until more recently (for example see Perez-Carrasco et al., 2020; Simpson et al., 2018). In this study, we adopt the cell invasion model of scratch assay experiments developed by Simpson et al. (2018). This model describes a discrete exclusion based random walk on a two-dimensional (2D) hexagonal lattice. Furthermore, this model involves treating the entire population of agents as three subpopulations that correspond to the red, yellow and green phases of the cell cycle as identified by FUCCI. Agents transition through the cell cycle, while simultaneously undergoing a nearest neighbour random walk, with exclusion, to model cell migration. This model is discussed in more detail in Section 3.4.1. This previous study did not perform any parameter inference or calibrate the model to experimental data. The primary focus of this present work is to apply Bayesian methods to recover parameter estimates for the model and the associated distribution of uncertainly around them. However, standard Bayesian approaches rely on the computation of the likelihood function which is often intractable in complex stochastic models. We overcome this limitation by applying Approximate Bayesian Computation (ABC) methods, which is discussed later in Section 3.4.2.

Simpson et al. (2020) investigate practical parameter identifiability in a deterministic partial differential equation of FUCCI scratch assay experiments. Practical parameter identifiability is a term that describes whether it is possible to produce precise estimates with finite regions of confidence levels (Raue et al., 2009). We adopt this terminology here since it is consistent with Simpson et al. (2020). Nevertheless, by using a simpler model, their study was able to adopt standard Bayesian approaches to parameter estimation since the likelihood function is tractable. Using a Markov Chain Monte Carlo (MCMC) framework and cell density data, their study found cell diffusivities were practically non-identifiable when they considered the case where the cell migration rate depends on the cell cycle phase. Although, their study does not consider other types of data which may be more informative of the underlying mechanisms. Here, we address the limitations Simpson et al. (2020) identify by modelling individual cell behaviour with a stochastic model which allows the generation of numerous data types. Indeed, we take full advantage of the flexibility of the stochastic model in this study and combine multiple data types (the number of cells in each phase and cell trajectory

data accounting for different phases) to improve parameter identifiability. However, working with cell trajectory data can be challenging, and these challenges include time consuming effort to manually track cells and the need for the cell density to be low to make cell tracking easier. Models which can avoid using cell trajectory data is an active area of research (Hywood et al., 2021), but we find using the Simpson et al. (2018) model, which incorporates cell trajectory data, leads to a good outcome.

Many other studies have explored modelling and/or parameter estimation in cell invasion models (Cai et al., 2007; Maini et al., 2004; Savla et al., 2004; Swanson, 2008; Takamizawa et al., 1997; Vo et al., 2015). Notably, Vo et al. (2015) estimate the parameters of a stochastic cell spreading model of an expanding population of fibroblast cells in a 2D circular barrier assay without cell cycle labelling. While ABC methods have previously been considered in stochastic cell spreading models, such as the Vo et al. (2015) study, they have never before been considered with FUCCI models and/or data. Prior to Vo et al. (2015), cell invasion models were usually defined by deterministic partial differential equation and when performing parameter inference, they usually used trial and error based approaches (Takamizawa et al., 1997) or non-linear least squares estimation (Cai et al., 2007; Maini et al., 2004; Savla et al., 2004; Swanson, 2008). However, these approaches to parameter estimation are unable to quantify the uncertainty around the point estimates. In this study, we show that using a discrete stochastic model is necessary to identify the transition and motility parameters when multiple phases of the cell cycle are considered. This difference is due to the wider range of data types that are available since individual cells are modelled rather than working with a simple cell density profile. This allows data types that are more informative about the model parameters, which have previously been unavailable to deterministic modelling approaches, to be considered. That is, we find cell count and cell trajectory data to be the most informative data types as they can produce practically identifiable parameters for the transition and motility parameters, respectively. Rcpp and MATLAB implementations of the simulation model and ABC algorithm used in this study are available at https://github.com/michaelcarr-stats/FUCCI.

The paper is structured as follows. In Section 3.2, we introduce the experimental data and the process by which it is collected. Section 3.3 describes the simulation model, the parameter inference method used, and our prior knowledge on the model parameters. In Section 3.4 we explain the image analysis process and present the inference results when using synthetic and experimental data sets. Discussion of results, future work and concluding remarks are presented in Section 3.5.

3.3 Data

2D scratch assay experiments are a good screening tool for more complex experimental models, as they are low cost, allow for easy data interpretation and readily allow control of oxygen, nutrients and drug supply (Beaumont et al., 2014; Santiago-Walker et al., 2009). We adopt data from a study conducted by Vittadello et al. (2018) where a scratch assay is used to examine melanoma cell proliferation and migration in real time with FUCCI technology. The experiment is initialised by placing a small population of cells and a growth medium in a culture dish (Figure 3.1 (a)) to create a uniform 2D monolayer of cells. Next, a sharp-tipped instrument is used to make a scratch in the monolayer of cells (Figure 3.1 (b)). Finally, the cells are observed at regular intervals as they proliferate and migrate into the newly created gap over the following 48 hours. For this study, we adopt the data from the experiments with WM983C FUCCI-transduced melonoma cells and present still images captured at 0 and 48 hours in Figure 3.1 (c)-(d), respectively. A major advantage of 2D scratch assay experiments is the multitude of different data types which can be easily recovered. The data types which we explore later include the number of cells in each population, position of cell populations, and cell trajectory data (Figure 3.1 (e)). It is important to consider the size of the imaged region compared to the culture plate (Figure 3.1 (b)) because the boundaries of the imaged region are not physical boundaries. Since the cell density outside of the scratched region is approximately uniform, with no macroscopic density gradients away from the leading edge, the net flux of cells across the boundary will be zero (Simpson et al., 2018). Therefore, the appropriate mathematical boundary conditions along the vertical boundaries will be zero net flux.

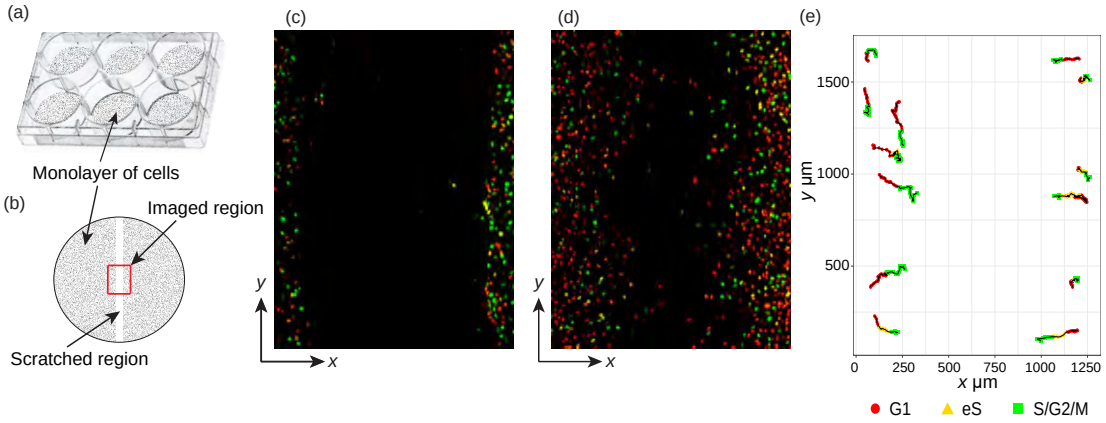

Figure 3.1: Experimental procedure and data. (a)-(b) Explains the experimental procedure and boundary conditions for simulation models. (a) Photograph of 6 culture plates commonly used with a uniform monolayer of cells. (b) Schematic showing the uniform cell monolayer (shaded), scratched region (white), and imaged region (outlined in red) in a 35 mm culture plate. (c)-(d) Experimental images, both 1309.09 x 1745.35 μm, of WM983C FUCCI-transduced melanoma cells at 0 and 48 hours, respectively. Images reproduced with permission from Vittadello et al. (2018). (e) Cell trajectory data of a select few cells recorded through red to green phases travelling inward to fill scratched region.

3.4 Methods

3.4.1 Simulation model

We adopt the discrete random walk model developed by Simpson et al. (2018) on a 2D hexagonal lattice. Each lattice site has diameter $\Delta = 20$ μm, which is the average cell diameter (Treloar et al., 2013), and is associated with a set of unique Cartesian coordinates,

$$(x_i, y_j) = \begin{cases} \big((j-1/2)\Delta\sqrt{3}/2, i\Delta\big) & \text{if } i \text{ is even,} \\ \big((j-1)\Delta\sqrt{3}/2, i\Delta\big) & \text{if } i \text{ is odd,} \end{cases} \tag{3.1}$$

where i and j are the respective row and column indices. To mimic scratch assay experiments, cells in G1 phase are represented by red agents, cells in eS phase are represented by yellow agents, and cells in S/G2/M phase are represented by green agents. Agents are permitted to transition through phases of the cell cycle and undergo a nearest neighbour random walk by simulating from a Markov process using the Gillespie algorithm (Gillespie, 1977) where the time between events is exponentially distributed. The algorithm is presented in Section 3.12.1 of the Supplementary Material.

To simulate cell migration, agents undergo a nearest neighbour random walk at rates M_r, M_y, M_g per hour for red, yellow and green agents, respectively (Figure 3.2 (a)-(f)). Potential movement events involve randomly selecting the target site from the set of six nearest-neighbouring lattice sites, with the movement event being successful only if the target site is vacant. In this way crowding effects are simply accommodated. To

simulate transitions through the cell cycle, red agents are allowed to transition into yellow agents at rate R_r per hour (Figure 3.2 (h)-(i)), yellow agents to green agents at rate R_y per hour (Figure 3.2 (i)-(j)) and green agents into two red daughter agents at rate R_g per hour (Figure 3.2 (j)-(k)). While we assume that the red-to-yellow and yellow-to-green transitions are unaffected by crowding, we model crowding effects for the green-to-red transition by aborting transitions where the additional red daughter agent would be placed onto an occupied lattice site. By prohibiting multiple agents from occupying the same lattice site, we are able to realistically incorporate crowding effects (Ermentrout & Edelstein-Keshet, 1993; Johnston et al., 2016).

The Simpson et al. (2018) model is dependent on the initial geometry, boundary conditions, the lattice spacing Δ, and the cell cycle transition and motility rates. Since we have reasonable estimates for Δ (Treloar et al., 2013) and we calibrate the initial geometry and boundary conditions to the experimental data, our study is concerned with estimating the unknown cell cycle transition and motility parameters. In a Bayesian setting, the unknown model parameters, $\theta = (R_r, R_y, R_g, M_r, M_y, M_g)$, and the uncertainty around them can be quantified by the posterior distribution, which is dependent on the likelihood and the prior distribution. However, while the Markov process model can capture the stochastic nature of cell proliferation and migration, when the dimension of the generator matrix (a matrix of rate parameters which describe the rate of transitioning between states) is too high the likelihood function consequently becomes intractable due to the computational cost of computing the matrix exponential (see Ho et al., 2018; Moler and Van Loan, 2003; Schnoerr et al., 2017; Sidje, 1998). Since conventional Bayesian approaches to parameter estimation are no longer feasible, we are motivated to use likelihood-free methods.

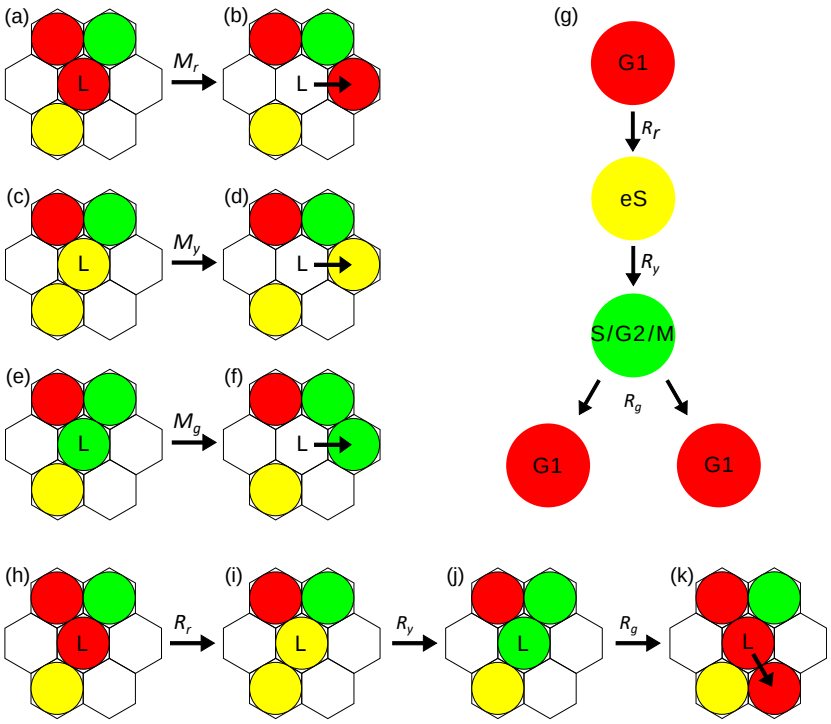

Figure 3.2: Cell migration and proliferation. (a-f) An agent at lattice site L will attempt to migrate to the six neighbouring lattice sites, successfully migrating if the selected site is vacant. (g) Schematic showing the progression through the G1 phase (red), early S phase (yellow) and S/G2/M phase (green) for FUCCI. (h-k) Agent transition through the cell cycle and proliferation. (k) A green agent (S/G2/M phase) at lattice site L will successfully divide and transition if the randomly selected neighbouring site is vacant.

3.4.2 Approximate Bayesian Computation

Using a Bayesian framework, the uncertainty about the unknown parameter θ with respect to the data y can be quantified by sampling from the posterior distribution $\pi(\theta|y) \propto \pi(y|\theta)\pi(\theta)$, where $\pi(y|\theta)$ is the likelihood function and $\pi(\theta)$ is the prior. However, the likelihood function for sufficiently complex models becomes intractable (see examples in biology Johnston et al., 2016; Kursawe et al., 2018; Vo et al., 2015, in ecology Guillemaud et al., 2010; Toni et al., 2009 and in cosmology Weyant et al., 2013). Rather than reverting to simpler models with tractable likelihoods, these types of problems can be instead analysed using likelihood-free methods that avoid evaluating the likelihood function.

One popular likelihood-free approach is ABC (Sisson et al., 2018). ABC involves simulating data from the model $x \sim \pi(\cdot|\theta)$ instead of evaluating the intractable likelihood; accepting configurations of θ which produce simulated data x that is close to the observed data y. It can be impractical to compare the full data sets of x and y, so ABC often relies on reducing the full data sets to summary statistics by some summarising function $S(\cdot)$, where the summary statistics for x and y are denoted $S_x = S(x)$ and $S_y = S(y)$, respectively. Provided the summary statistics are highly informative about

the model parameters, then $\pi(\theta|y) \approx \pi(\theta|S_y)$ is a good approximation or exact if sufficient statistics are used (Blum et al., 2013). However, the latter are usually difficult to attain in practice and so this study focuses on the use of summary statistics. In effect, ABC samples from the approximate posterior:

$$\pi_\epsilon(\theta|S_y) \propto \pi(\theta) \int_x K_\epsilon(\rho(S_y, S_x))\pi(x|\theta)\ \mathrm{d}x, \tag{3.2}$$

where $\rho(S_y, S_x)$ is the discrepancy function which measures the difference between the two data sets and $K_\epsilon(\cdot)$ is the kernel weighting function which weighs $\rho(S_y, S_x)$ conditional on the tolerance ϵ. A common choice for the discrepancy function is the Euclidean distance, $\rho(S_y, S_x) = \|S_y - S_x\|_2$, and for the kernel weighting function is the indicator function, $1(\cdot)$, which is equal to one if $\rho(S_y, S_x) \leq \epsilon$ and is zero otherwise. The approximate posterior in Equation 3.2 converges to the posterior conditional on the observed summary (often referred to as the partial posterior) in the limit as $\epsilon \to 0$ (Beaumont et al., 2002).

To sample from the approximate posterior, commonly ABC-rejection (Pritchard et al., 1999; Tavaré et al., 1997), Markov Chain Monte Carlo ABC (MCMC-ABC) (Marjoram et al., 2003), or Sequential Monte Carlo ABC (SMC-ABC) (for examples see Drovandi and Pettitt, 2011; Harrison and Baker, 2020; Sisson et al., 2007) algorithms are used. ABC-rejection samples particles from the prior distribution and accepts particles with a discrepancy measure $\rho(S_y, S_x)$ less than the desired tolerance ϵ. In cases when the prior distribution is relatively diffuse compared to the posterior density (such as our application), lower acceptance rates are common because particles are predominantly sampled in regions of low posterior density (Sisson et al., 2007). To increase efficiency, one could instead use MCMC-ABC which constructs a Markov chain with a stationary distribution identical to the approximate posterior by proposing particles from a carefully-tuned proposal distribution, $\theta^i \sim q(\cdot|\theta^{i-1})$, and accepting those with probability

$$p_{\text{acc}} = \min\left(1, \frac{\pi(\theta^i)q(\theta^{i-1}|\theta^i)K_\epsilon(\rho(S_y, S_x^i))}{\pi(\theta^{i-1})q(\theta^i|\theta^{i-1})K_\epsilon(\rho(S_y, S_x^{i-1}))}\right), \tag{3.3}$$

which is based on the Metropolis-Hastings ratio (Hastings, 1970; Metropolis et al., 1953). While MCMC-ABC tends to be more computationally efficient compared to ABC-rejection (Marjoram et al., 2003), it is possible for the Markov chain to spend many iterations in areas of low posterior probability. In our application we found MCMC-ABC to take a considerable effort to tune the proposal distribution while still being computationally cumbersome. However, SMC-ABC or more specifically the SMC-ABC replenishment algorithm (Drovandi & Pettitt, 2011) requires very little tuning comparatively and allows for simulations to be performed in parallel to increase computational efficiency.

The SMC-ABC replenishment algorithm traverses a set of distributions defined by T non-increasing tolerance levels $\epsilon_1 \geq \cdots \geq \epsilon_T$ to sample from the approximate posterior:

$$\pi_{\epsilon_t}(\theta|S_y) \propto \pi(\theta) \int_x 1(\|S_y - S_x\|_2 \leq \epsilon_t)\pi(x|\theta)\ \mathrm{d}x, \text{ for } t = 1, ..., T$$

where the first target distribution is constructed by sampling from the prior distribution to attain a collection of parameter values (called particles) and their discrepancies, $\{\theta^i, \rho^i\}_{i=1}^N$. The first tolerance threshold, ϵ_1, is set as the maximum of the set of discrepancies. Thereafter, to propagate particles through the sequence of target distributions, particles are first sorted in ascending order by their discrepancy and the new tolerance is set as $\epsilon_t = \rho^{N-N_\alpha}$ where $N_\alpha = \lfloor \alpha N \rfloor$, α is the proportion of particles discarded and $\lfloor \cdot \rfloor$ is the floor function. Particles, $\{\theta^i\}_{i=N-N_\alpha+1}^N$, which do not satisfy the new tolerance are discarded and resampled, with replacement, from the remaining particles to replenish the population. To prevent sample degeneracy (too many duplicated particles), resampled particles are then perturbed according to an MCMC kernel R_t times with an invariant distribution given by the current approximate posterior $\pi_{\epsilon_t}(\theta|S_y)$. In each of the R_t iterations, the proposed particles are drawn from an automatically tuned proposal distribution $\theta^i \sim q(\cdot|\theta^{i-1})$ and accepted with probability $p_{\text{acc}}^{i,j}$ (Equation 3.3) where i denotes the ith particle and j the jth MCMC iteration. To ensure sample diversity, R_t can be dynamically set based on the overall MCMC acceptance rate, $\overline{p_{\text{acc}}} = \frac{1}{N_\alpha \times R_t} \sum_{i=1}^{N_\alpha} \sum_{j=1}^{R_t} p_{\text{acc}}^{i,j}$, such that there is a $1 - c$ chance that all particles are moved at least once and is given by

$$R_t = \left\lceil \frac{\log(c)}{\log(1 - \overline{p_{\text{acc}}})} \right\rceil,$$

where the ceiling function $\lceil \cdot \rceil$ is used to be conservative and an estimate for $\overline{p_{\text{acc}}}$ is calculated from $R_{t-1}/2$ pilot MCMC iterations. A popular choice for the proposal distribution is the multivariate normal distribution, $q(\theta^i|\theta^{i-1}) = \mathcal{N}(\theta^i; \theta^{i-1}, \Sigma)$, where the covariance matrix Σ is the tuning parameter. To create a more efficient proposal distribution, we can adaptively tune Σ by computing the empirical covariance matrix of the $\{\theta\}_{i=1}^{N-N_a}$ particles which are already distributed according to the current target distribution. In this application we do not worry about scaling the covariance matrix as there are only six parameters. The algorithm finally stops once the overall MCMC acceptance rate, $\overline{p_{\text{acc}}}$, is unreasonably low ($\leq 1\%$) or the desired tolerance threshold is reached. For the two tuning parameters, Drovandi and Pettitt (2011) suggest setting $\alpha = 0.5$ and $c = 0.01$. The SMC-ABC replenishment algorithm is presented in Algorithm 7 and is hereafter referred to as SMC-ABC.

A crucial limitation of ABC methods is the curse of dimensionality, where despite the addition of more data, the approximation to the posterior can become distorted as a

result of the discrepancy between observed data and simulated data $\rho(S_y, S_x)$ naturally increasing with the dimension (Beaumont et al., 2002). In applications where increasing the dimension of the summary statistics cannot be avoided, the discrepancy between observed and summary statistics can be accounted for, at least approximately, with regression adjustment (Beaumont et al., 2002; Blum et al., 2013). Regression adjustment involves explicitly modelling the parameters against the discrepancy between observed and simulated data. Assume for the moment that θ is a scalar parameter. Consider the following regression model

$$\theta^i = \beta_0 + (S_x^i - S_y)^\top \beta + \varepsilon^i,$$

where $i = 1, \ldots, N$ is the parameter sample index, β is the regression coefficients, β_0 is the intercept and ε^i is the error term. Estimates for β_0 and β can be computed by minimising the weighted least squares criterion $\sum_{i=1}^N w^i(\theta^i - \beta_0 - (S_x^i - S_y)^\top \beta)^2$. Here we choose to use the popular Epanechnikov weighting function (Epanechnikov, 1969), defined as $w^i = 0.75(1 - (\rho^i / \max(\{\rho^i\}_{i=1}^N))^2)$, but other weighting functions could also be used. Using the estimated regression coefficients $\hat{\beta}$, we then make the adjustment

$$\theta^{i^*} = \theta^i - (S_x^i - S_y)^\top \hat{\beta} \quad \text{for } i = 1, \ldots, N.$$

The adjusted sample $\{\theta^{i^*}\}_{i=1}^N$ can often give a more accurate approximation of the posterior. To ensure that the adjusted parameters remain within the support of the prior distribution (if bounded), Hamilton et al. (2005) suggest transforming parameter values before applying the regression adjustment. We use a logit transformation, $\tilde{\theta} = \log((\theta - a)/(b - \theta))$, where a and b are the respective lower and upper bounds of the prior. Given that we have a vector of parameters, we apply a regression adjustment to each component of the parameter vector separately.

Algorithm 7 SMC-ABC (Drovandi & Pettitt, 2011)

1: Set f_{acc}, ϵ_T and $N_\alpha = \lfloor \alpha N \rfloor$
2: Set S_t the initial number of pilot MCMC iterations
3: **for** $i = 1$ to N **do**
4: Draw $\theta^i \sim \pi(\cdot)$
5: Simulate $x^i \sim \pi(\cdot|\theta^i)$
6: Compute $S_x^i = S(x^i)$
7: Compute $\rho^i = \rho(S_y, S_x^i)$
8: **end for**
9: Sort θ by ρ such that $\rho^1 \leq \rho^2 \leq ... \leq \rho^N$
10: Set $\epsilon_t = \rho^{N-N_\alpha}$
11: **while** $\overline{p_{\text{acc}}} > f_{\text{acc}}$ OR $\rho^N > \epsilon_T$ **do**
12: Compute tuning parameters of MCMC kernel $q_t(\cdot|\cdot)$ using $\{\theta^i\}_{i=1}^{N-N_\alpha}$
13: **for** $j = N - N_\alpha + 1$ to N **do**
14: Resample θ^j from $\{\theta^i\}_{i=1}^{N-N_\alpha}$
15: **for** $k = 1$ to S_t **do**
16: Propose $\theta^* \sim q_t(\cdot|\theta^j)$
17: Simulate $x \sim \pi(\cdot|\theta^*)$
18: Compute $S_x = S(x)$
19: Compute MH ratio $r = \dfrac{\pi(\theta^*)q(\theta^j|\theta^*)}{\pi(\theta^j)q(\theta^*|\theta^j)} 1(\rho(S_y, S_x) < \epsilon_t)$
20: **if** $\mathcal{U}(0,1) < r$ **then**
21: Set $\theta^j = \theta^*$, $\rho^j = \rho(S_y, S_x)$ and $S_x^j = S_x$
22: **end if**
23: **end for**
24: **end for**
25: Calculate $\overline{p_{\text{acc}}}$ based on the overall acceptance rate from the pilot MCMC runs
26: Set $R_t = \lceil \dfrac{\log(c)}{\log(1 - \overline{p_{\text{acc}}})} \rceil$
27: Repeat steps 13-24 with $S_t = \max(0, R_t - S_t)$
28: Set $\epsilon_t = \rho^{N-N_\alpha}$ and $S_t = \lceil R_t/2 \rceil$
29: **end while**
30: Implement regression adjustment (optional)

3.4.3 Prior Knowledge

The cell cycle time, which is related to the doubling time, can be thought of as the summation of the time spent in each phase of the cell cycle. Estimates of the cell doubling time for melanoma cells range from 16-47 h (Haass et al., 2014; Simpson et al., 2020). Furthermore, Simpson et al. (2020) estimate the average time 1205Lu FUCCI-transduced melanoma cells spend in the G1 to be between 8-30 h. This means that the

transition from G1 to S/G2/M phase is approximately $1/30 - 1/8$ /h. Additionally, the duration spent in S/G2/M phases were reported as 8-17 h. This means that the transition from S/G2/M to G1 phase is approximately $1/17 - 1/8$ /h. Therefore, we propose that our prior information of the transition rates is uniform over the range $0 - 1$ /h to be conservative.

Cell diffusivity, D, the measurement of motility rate for particles undergoing random diffusive migration, can be used to quantify the cell motility rate $M \in \{M_r, M_y, M_g\}$, by $D = M\Delta^2/4$ (Codling et al., 2008), where Δ is the cell diameter. Empirical evidence finds estimates for cell diffusivity to range from $0 - 3304$ μm^2/h (Cai et al., 2007; Maini et al., 2004; Treloar et al., 2013). Furthermore, Simpson et al. (2018) suggest the cell diffusivity to be approximately 400 μm^2/h, so that the rates are approximately 4 /h. We propose that the prior information of the motility rates to be uniform over the range $0 - 10$ /h; attributing the larger interval to the greater variation of cell diffusivity estimates in existing literature.

3.5 Results

For SMC-ABC we generate samples from the approximate posteriors using $N = 1000$ particles. From preliminary trials, we found it more useful to use the overall MCMC acceptance rate as the stopping rule for the SMC-ABC algorithm and adopt the sensible choice for the final acceptance rate as $f_{\text{acc}} = 1\%$ and $\epsilon_T = 0$ for the target tolerance.

3.5.1 Developing summary statistics and validation with synthetic data

The accuracy and precision of ABC methods in approximating the posterior distribution is sensitive to the quality of the summary statistics used (Beaumont et al., 2002). We first trial and validate different summary statistics with multiple synthetic data sets such that the true parameter values are known. In this way, we are able to compare the performance of different summary statistics and determine which are the most effective. While trying to replicate the environment of the experimental data as close as possible, such as domain size, boundary conditions and initial number of cells, we do not calibrate the initial location of cells but rather randomly distribute the cells within a 200 μm by 1745.35 μm region on either side of the scratch. We attain the initial cell counts of red, yellow, and green cells by using the procedure outlined in section 3.5.2 (steps 1-4) and report them here to be 119, 35 and 121, respectively.

In our analysis of the simulation model, agents with relatively higher transition rates were found to correspond to lower population sizes, and vice versa. Therefore, we use the number of agents in each population (N_r, N_y, N_g) at the end of the experiment as summary statistics that may be informative about the transition rates R_r, R_y, R_g,

respectively. We test the suitability of this summary statistic on four synthetic data sets produced by varying the transition rates amongst biologically plausible values and keeping the motility rates known and constant. Estimates in existing literature of cell transition rates are similar (Haass et al., 2014; Simpson et al., 2020) and the efficiency of the simulation model is dependent on the number of agents in the system (higher transition rates increase the overall proliferation rate and the frequency of events). Therefore, we choose to keep the transition rates rather close to the estimates of Haass et al. (2014), instead of varying them over the extents of the prior domain. Thus, the four parameter configurations we choose to generate the synthetic data sets are $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.25, 0.15, 0.22, 4, 4, 4), (0.12, 0.07, 0.03, 4, 4, 4), (0.3, 0.36, 0.28, 4, 4, 4)\}$. In Section 3.12.2 of the Supplementary Material we present the marginal posterior distributions produced and confirm the suitability of this summary statistic.

For the motility parameters we explore and compare two sets of summary statistics, namely cell density and cell trajectory data. Of these two data sets, cell density data is desirable due to less manual effort needed to generate the data while cell trajectory data could offer more information but is more challenging to collect. For the cell density data, we first segment the imaged region at the end of the experiment ($t = 48$ h) directly down the centre of the image in the y direction and calculate the median position and interquartile ranges of the red, yellow, and green agent populations in the x direction for cells on the left and right sides. For the cell trajectory data, we average the distance of multiple cell trajectories through each cell phase until the cell returns to the initial phase or the simulation is terminated. We select cells to be tracked provided that the cell is initially in G1 (red) phase and the cell is located on the leading edge of the cell monolayer toward the gap in the scratch assay. The reasoning behind beginning tracking from the G1 phases is due to a short period of fluorescent negativity in between S/G2/M phases and G1 phase (Haass et al., 2014) which makes tracking between these phases difficult. Therefore, by starting from the G1 phase we avoid having to track between these two phases. Furthermore, cells were identified as being on the leading edge if their path toward the middle of the scratch was unhindered. The process of manually tracking cell trajectories can be time consuming. Thus, we are interested in finding the minimum number of cells to track such that sufficient information is acquired. In Section 3.12.2 of the Supplementary Material, we draw samples from the posterior distribution using 10, 20, 30, 40 and 50 cell trajectories with four synthetic data sets generated from $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.04, 0.17, 0.08, 2, 5, 8), (0.04, 0.17, 0.08, 8, 2, 5), (0.04, 0.17, 0.08, 5, 8, 2)\}$. Our analysis finds diminishing returns of parameter precision as the number of cell trajectories increases. We find that using 20 cell trajectories achieves a good balance between precision and the number of cell trajectories used. Using the same four synthetic data sets we also attempt to draw

samples from the posterior distribution using cell density data with the cell transition rates held constant in Section 3.12.2 of the Supplementary Material. However, under these settings, we found the motility parameters to be non-identifiable.

We now combine the summary statistics formulated to estimate the cell cycle transition and motility rates together with four synthetic data sets. Due to the similarity in estimates for cell cycle transition rates in existing literature (see Haass et al., 2014; Simpson et al., 2020), we adopt estimates for the cell cycle transition rates from Haass et al. (2014) for all four parameter configurations. Since estimates for motility rates have been reported to vary by two orders of magnitude (see Cai et al., 2007; Maini et al., 2004; Treloar et al., 2013), we choose to vary the motility rates over the range of the prior for the four parameter configurations. That is, we generate four synthetic data sets with $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.04, 0.17, 0.08, 2, 5, 8), (0.04, 0.17, 0.08, 8, 2, 5), (0.04, 0.17, 0.08, 5, 8, 2)\}$. In Figure 3.3 we present the marginal posterior distributions produced when using the number of cells in each subpopulation and cell density data as summary statistics. In Figure 3.4 we present the marginal posterior distributions when using cell trajectory data in place of cell density data. Again, we see that the motility parameters are practically non-identifiable when cell density data is used while both cell cycle transition and motility parameters are practically identifiable when cell trajectory data is included. Furthermore, it is clear from the concentration of the marginal posterior distributions around the true parameter values (dashed line) in Figure 3.4 that the cell count and cell trajectory data are highly informative about the transition and motility parameters, respectively. We note that the precision of these distributions is greater for the cell cycle transition parameters than the motility parameters. Importantly, these results show for the first time that practical parameter inference on both transition and motility parameters of a FUCCI scratch assay experiment using Bayesian inference techniques is possible. These results justify the choice of the Markov process model compared to simpler continuum models which do not give insight into cell trajectory data (see Simpson et al., 2020).

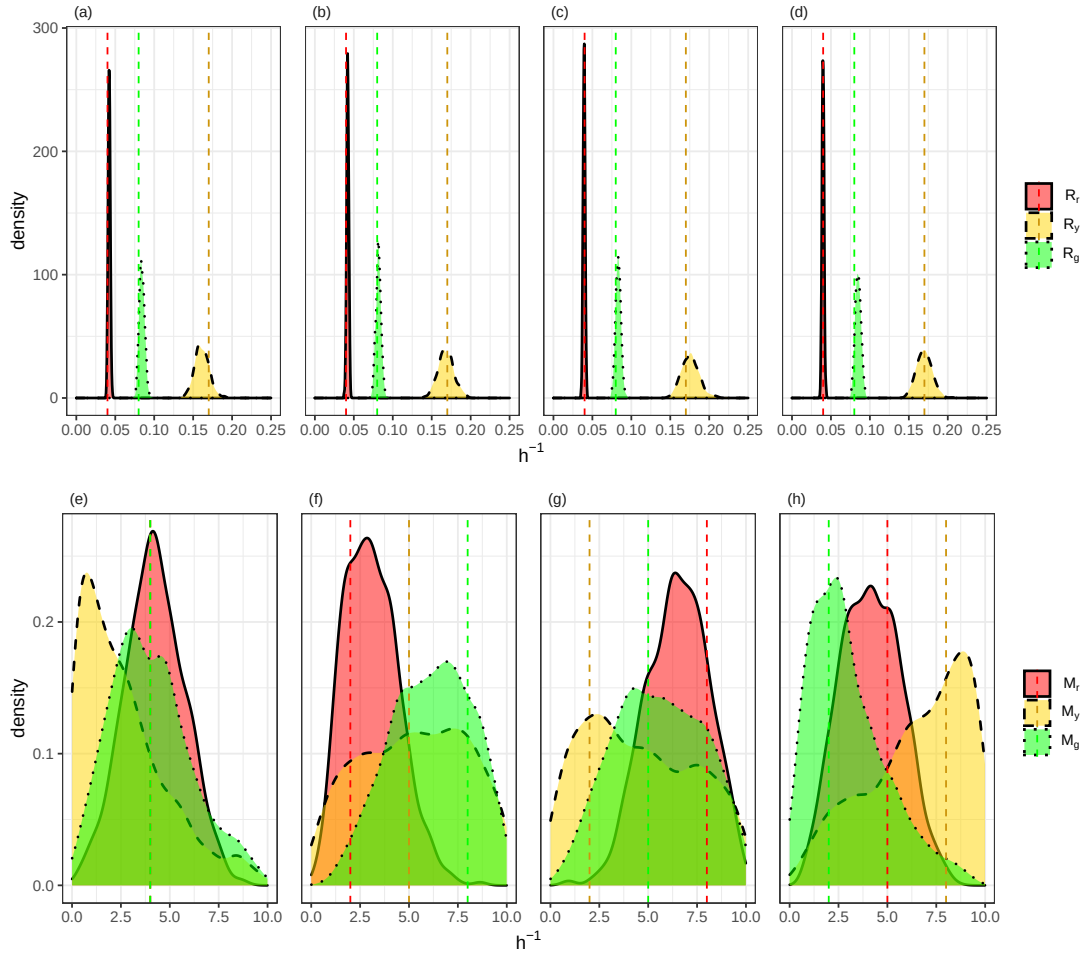

Figure 3.3: Estimating cell cycle transition and cell motility parameters, $\theta = (R_r, R_y, R_g, M_r, M_y, M_g)$, with the number of cells in each phase at $t = 48$ h and cell density data as summary statistics across several synthetic data sets. Synthetic data sets were produced from simulations with true parameter values indicated by dashed vertical lines (note that in (e) the lines overlap). (a,e) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 4, 4, 4)$. (b,f) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 2, 5, 8)$. (c,g) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 8, 2, 5)$. (d,h) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 5, 8, 2)$.

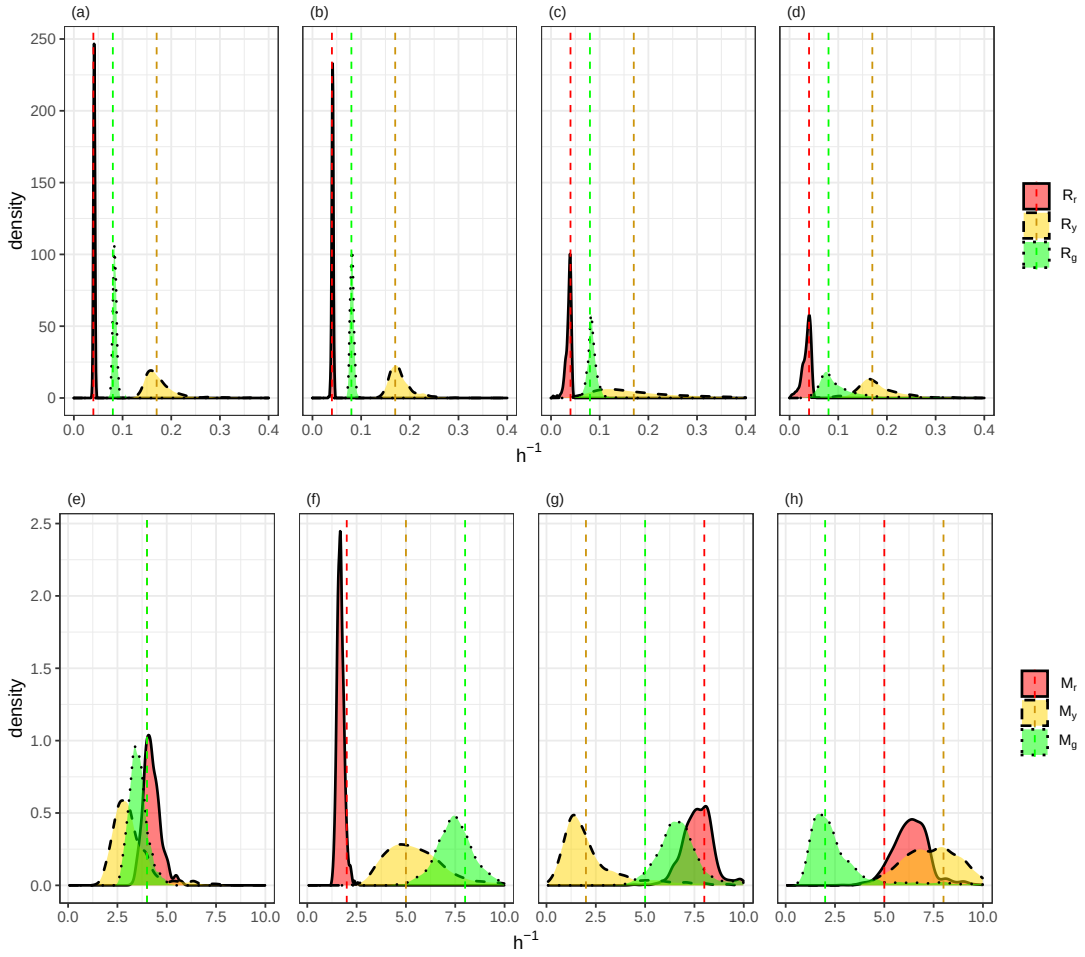

Figure 3.4: Estimating cell cycle transition and cell motility parameters, $\theta = (R_r, R_y, R_g, M_r, M_y, M_g)$, with the number of cells in each phase at $t = 48$ h and cell tracking data as summary statistics across several synthetic data sets. Synthetic data sets were produced from simulations with true parameter values indicated by dashed vertical lines (note that in (e) the lines overlap). (a,e) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 4, 4, 4)$. (b,f) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 2, 5, 8)$. (c,g) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 8, 2, 5)$. (d,h) Estimated marginal posteriors produced with $\theta = (0.04, 0.17, 0.08, 5, 8, 2)$.

3.5.2 Image analysis of experimental data

We analyse the experimental images using *ImageJ* (Rueden et al., 2017) to record the Cartesian coordinates of cells. Of primary interest is processing the initial frame such that we can replicate the experimental settings as accurately as possible in the simulation but we also repeat this procedure for the final frame to retrieve the final cell counts and cell density data, which we use as summary statistics. The process is as follows:

Step 1: *Read in image:* File > Open > *select image* (Figure 3.5 (a)).

Step 2: *Convert image to 8-bit:* Image > Type > 8-bit (Figure 3.5 (b)).

Step 3: *Identify cell edges:* Convert the image to black and white (Process > Binary > Convert to Mask) and then distinguish conjoined cells (Process > Binary > Watershed) (Figure 3.5 (c)).

Step 4: *Compute Cartesian coordinates:* Analyze > Analyze Particles... > OK.

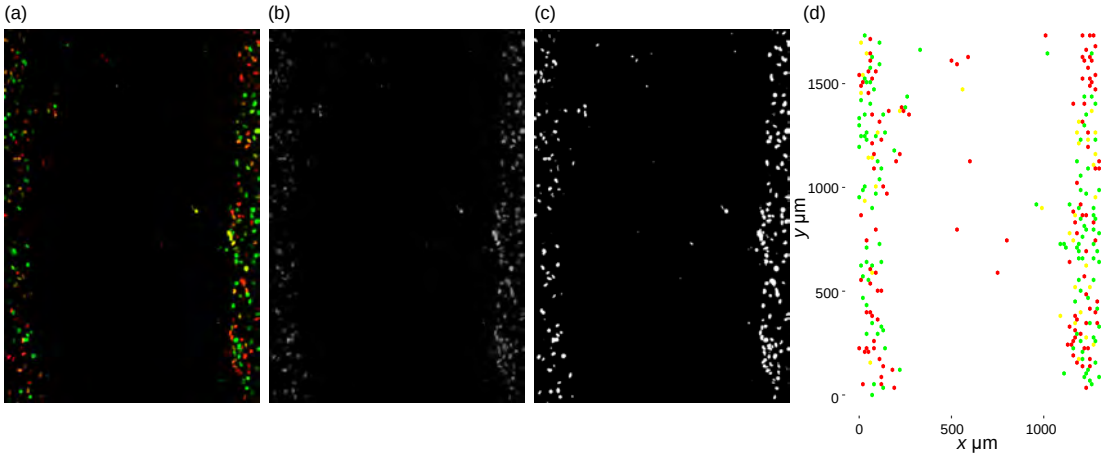

Figure 3.5: *ImageJ* procedure. (a) Original image loaded (WM983C FUCCI-transduced melanoma cells). (b) Image after compression to 8-bit. (c) Image after converting to black and white and watershedding. (d) Simulation initial geometry recovered from data processing of WM983C FUCCI-transduced melanoma cells in *ImageJ* and *R*

A limitation of using the watershed tool is that we must convert the image to black and white. In doing so, we lose the cell phase identity associated with the cell coordinates recovered from *ImageJ*. To overcome this, we use *R* (R Core Team, 2020) to retrieve the RGB decimal color code and Cartesian coordinates of pixels. Matching pixel coordinates recovered from *R* and the coordinates of the centroid of the cells recovered in *ImageJ*, we create a data set of cell coordinates and their associated RGB decimal codes. To classify the RGB coordinates into one of the three cell cycle phases we use the conditions outlined in Table 3.1.

To extract summary statistics from the experimental data, we repeat the image pro-

Table 3.1: Cell phase classification rule using RGB decimal codes.

State	RGB decimal code	
	Red	Green
G1	>100	≤100
eS	>100	>100
S/G2/M	≤100	>100

cessing procedure previously outlined above with the final frame ($t = 48$ h) and extract the final cell counts and cell density data. Additionally, we extract cell trajectory data by processing the entire sequence of still images in *ImageJ* with the "Multi-point" tool to manually track cell coordinates between frames. We use a similar process as before to identify cell phases in these summary statistics using R and present them in Table 3.2 and the cell trajectory data in Figure 3.6.

Finally, we calibrate the hexagonal lattice used in the simulation model with the data set of Cartesian coordinates recovered previously by rearranging Equation 3.1 to find their associated lattice row and column indices denoted

$$(i, j) = \left(\lfloor \frac{2x}{\sqrt{3}\Delta} + 1 \rceil, \lfloor \frac{y}{\Delta} \rceil \right),$$

where $\lfloor \cdot \rceil$ rounds to the nearest integer. We treat the rare instances (<1%) where multiple coordinates are mapped to the same lattice space as duplicated values and omit them rather than place them on the next closest lattice site. The result from this translation of data is presented in Figure 3.5 (d). We repeat this process for the initial frame of the cell trajectory data to identify the starting position. However, due to manually tracking cell trajectories, often the coordinate retrieved was not centred on the cell which in some cases caused the starting position to be mapped to an unoccupied lattice site. We intervene prior to transforming the starting position and adjust the coordinate values to the closest occupied lattice site which is chosen such that the radial distance between the coordinate and the lattice site is minimised.

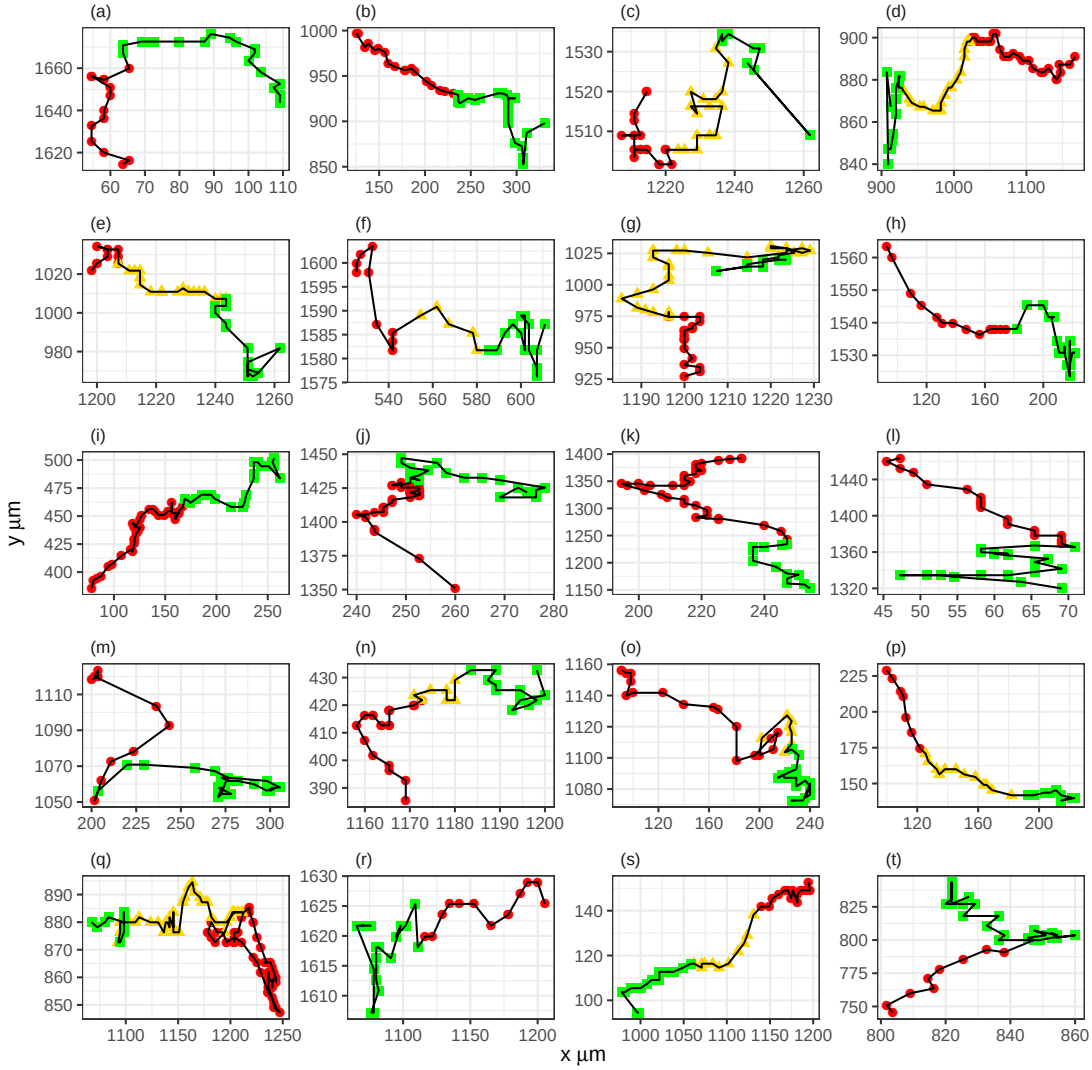

Figure 3.6: Trajectories of WM983C FUCCI-transduced melanoma cells where each box ((a)-(t)) corresponds to one of the twenty cell trajectories. Tracking begins in red phase (red circles) then progresses through the yellow phase (yellow triangles) and is terminated at end of green phase (green squares).

Table 3.2: Observed summary statistics of WM983C FUCCI-transduced melanoma cells

Summary Statistic	Description	Value
S_1	Number of red cells at 48 hours	566 cells
S_2	Number of yellow cells at 48 hours	111 cells
S_3	Number of green cells at 48 hours	166 cells
S_4	Average distance travelled through red phase by 20 cells	105 μm
S_5	Average distance travelled through yellow phase by 20 cells	40 μm
S_6	Average distance travelled through green phase by 20 cells	100 μm
S_7	Median position of red cells on the left and right side	(155, 1170) μm
S_8	Median position of yellow cells on the left and right side	(158, 1189) μm
S_9	Median position of green cells on the left and right side	(177, 1129) μm
S_{10}	Interquartile range of the red cells position on the left and right side	(196, 197) μm
S_{11}	Interquartile range of the yellow cells position on the left and right side	(164, 144) μm
S_{12}	Interquartile range of the green cells position on the left and right side	(213, 207) μm

3.5.3 Estimating Model Parameters with Experimental Data

After calibrating the simulation to the experimental data of WM983C FUCCI-transduced melanoma cells, we first attempt to sample from the posterior distribution using the number of cells in each subpopulation and cell density data (summary statistics S_1 to S_3 and S_7 to S_{12} in Table 3.2, respectively) and present the samples from the posterior distribution in Figure 3.7. Consistent with results found in Section 3.5.1 and Simpson et al. (2020), estimates for the motility rates are practically non-identifiable when cell density data is used.

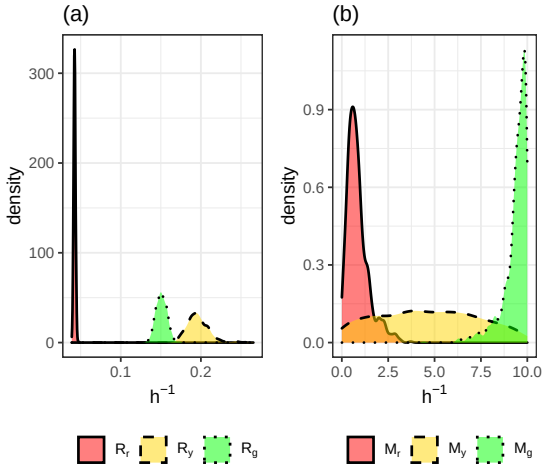

Figure 3.7: Marginal posterior distributions using number of cells in each subpopulation and cell density data. (a) Marginal posterior distributions for transition rates of WM983C FUCCI-transduced melanoma cells. (b) Marginal posterior distributions for motility rates of WM983C FUCCI-transduced melanoma cells.

Next, we attempt to sample from the posterior distribution using the number of cells in each subpopulation and cell trajectory data (summary statistics S_1 to S_3 and S_4 to S_6 in Table 3.2, respectively). We present the marginal posterior distributions produced in Figures 3.8 (a)-(b) along with the mean, standard deviation, (2.5%, 50%, 97.5%) quantiles, and the coefficient of variation (CV) in Table 3.3. The practical identifiability in the transition and motility parameters clearly shows the benefits of using cell tracking data as the distributions are unimodal and concentrated. We estimate the cell cycle transition rates to be between $0.0411 - 0.193$ /h which is consistent with estimates in existing literature (Haass et al., 2014; Simpson et al., 2020). Our estimates for cell motility were found to range between $0.316 - 1.12$ /h which corresponds to estimates of cell diffusivity between 31.6 and 112 μm^2/h which is reasonable considering the degree of uncertainty in existing estimates which can vary between 0 and 3304 μm^2/h (Cai et al., 2007; Maini et al., 2004; Treloar et al., 2013). The precision in parameter estimates can be quantified by the CV which is a standard measure for the dispersion of data around the mean. Using the CV, the dispersion in the transition rates range from $2.65 - 5.31\%$ and the motility rates range from $10.9 - 18.4\%$. To validate the para-

meter estimates recovered, we also present the posterior predictive distributions for the summary statistics retained from each parameter value in the posterior in Figures 3.8 (c)-(d). These distributions are formed by plotting the distribution of simulated summary statistics produced from the posterior samples and is compared to the observed summary statistics (dashed line). These results suggest that the Markov process model developed by Simpson et al. (2018) is promising as it is able to recover the observed summary statistics of the experimental data. However, further model validation should be considered in future research to determine if the simulated cell trajectories produce similar paths to those which were observed.

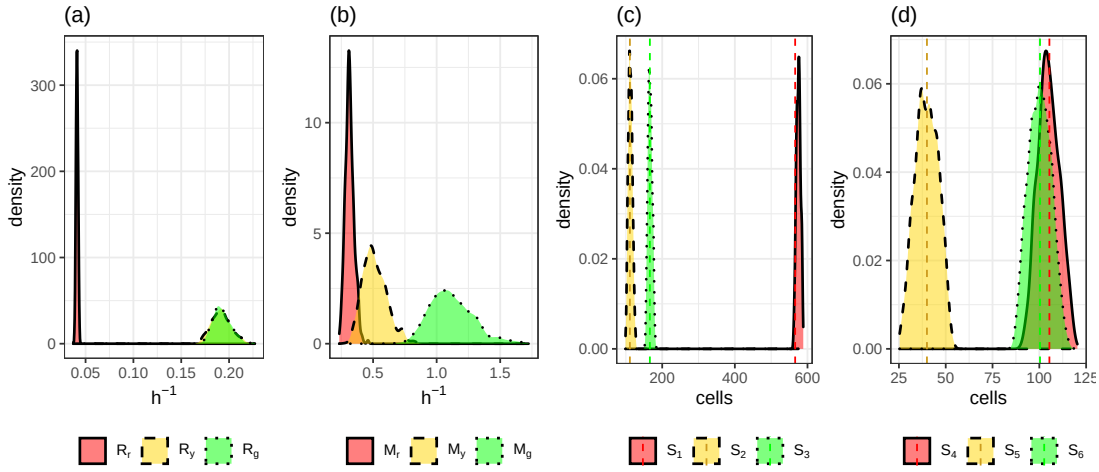

Figure 3.8: Marginal posterior distributions using number of cells in each subpopulation and cell trajectory data. (a) Marginal posterior distributions for transition rates of WM983C FUCCI-transduced melanoma cells. (b) Marginal posterior distributions for motility rates of WM983C FUCCI-transduced melanoma cells. (c) Distribution of simulated summary statistics (informative of transition rates) compared to observed summary statistics (dashed line). (d) Distribution of simulated summary statistics (informative of motility rates) compared to observed summary statistics (dashed line).

Table 3.3: Posterior summaries (3 significant figures): mean, standard deviation, (2.5%, 50%, 97.5%) quantiles, and the coefficient of variance (CV).

Parameter	Mean	Std. Dev.	(2.5%, 50%, 97.5%)	CV (%)
R_r	0.0411	0.00109	(0.039, 0.0411, 0.0432)	2.65
R_y	0.192	0.0102	(0.173, 0.191, 0.214)	5.31
R_g	0.193	0.00957	(0.177, 0.192, 0.213)	4.96
M_r	0.316	0.0343	(0.256, 0.313, 0.385)	10.9
M_y	0.514	0.0945	(0.353, 0.502, 0.725)	18.4
M_g	1.12	0.169	(0.836, 1.10, 1.48)	15.1

3.6 Discussion

In this study, we calibrate the 2D hexagonal-lattice random walk model developed by Simpson et al. (2018) to scratch assay data where the cell cycle is revealed in real time

using FUCCI technology. While this model is well suited to describing the stochastic nature of cell proliferation and migration, the likelihood function consequently becomes intractable. This makes conventional Bayesian approaches to parameter inference infeasible. We resort to using the class of Bayesian methods known as ABC which bypass evaluating the likelihood function. After evaluating the appropriateness of different ABC algorithms in Section 3.4.2 we find the SMC-ABC replenishment algorithm developed by Drovandi and Pettitt (2011) to be suitable.

In this study we work with uniform prior distributions. This may be considered as a reasonably vague prior, but it can also be interpreted as providing more support to larger rate parameters when the prior ranges over several orders of magnitude. To discourage larger rate parameter values, other priors could be considered, such as Jeffry's prior over all positive reals. If a proper prior (i.e. integrates to unity) is desired, the Jeffrey's prior could be truncated or an exponential prior used instead. We leave such extensive prior sensitivity analysis and performance for future research.

In Section 3.4.1 we previously discussed the intractability of the likelihood function. Although, we did not discuss particle filtering methods to construct a continuous time likelihood function which can be considerably more tractable than those based off discrete data. However, given that the model is highly stochastic, very different cell trajectories can be produced with the same parameter values. This would make filtering approaches difficult to apply. Instead, it is more efficient to match summary statistics of the cell trajectories (here we consider the average time spent in each phase). Additionally, Pseudo-marginal MCMC (Andrieu & Roberts, 2009) could be used to sample from the exact posterior distribution if an unbiased likelihood estimator (based on the full dataset) with a small enough variance can be constructed. Unfortunately, due to the complexity of the model and the need to summarise the data, it does not seem feasible to construct such a likelihood estimator here. Therefore, the nature of the modelling approach and the use of summary statistics naturally lends itself to using ABC methods.

The accuracy of ABC methods in approximating the posterior distribution is sensitive to the quality of the summary statistics used (Beaumont et al., 2002). We trial various summary statistics with multiple synthetic data sets to determine which summary statistics are the most informative. We find using the number of cells in each cell cycle phase at the end of the experiment to be highly informative about the cell cycle transition rates. We trial and compare two sets of summary statistics for the motility parameters: the median position and interquartile range of the cells in the x direction on the left and right side of the scratch assay (which we refer to as cell density data); and the average distance travelled through each cell phase by 20 individual cells (which we refer to as cell trajectory data). Using these two sets of summary statistics

in conjunction with the cell count data, we attempt to draw samples from the posterior distribution using the SMC-ABC replenishment algorithm with multiple biologically plausible synthetic data sets. We find that when using cell trajectory data as summary statistics the parameters are practically identifiable; however this is not the case when cell density data is used. Importantly, this is the first time practical parameter identifiability for both cell cycle transition and motility has been successfully conducted with fluorescent cell cycle labelling scratch assay experiments.

In this study we summarise cell trajectory data by taking the average distance travelled in each phase across 20 cell trajectories. However, additional features of cell trajectories could also be considered (for example the variance). Although, the addition of more summary statistics may increase ABC error due to the increased dimensionality of the summary statistic despite our efforts to treat this with regression adjustment. An additional method which could be used is semi-automatic ABC (Fearnhead & Prangle, 2012) which constructs a set of summary statistics with the same dimension as the parameter space by modelling the importance of the initial set of summary statistics. However, exploration of these additional summary statistics and the the effects on the precision of the posterior distribution are left for future research.

We extend on the work of previous studies (Simpson et al., 2020; Simpson et al., 2018) by calibrating our model to real data and performing Bayesian inference. Using experimental data of WM983C FUCCI-transduced melanoma cells, we estimate the approximate posterior using the SMC-ABC algorithm with our cell cycle transition rate summary statistics and our two sets of motility summary statistics. Under the experimental setting, our results again find the estimates for the motility parameters to be practically non-identifiable when cell density data is used but practically identifiable when cell trajectory data is used. These results are consistent with Simpson et al. (2020) and justify the motivation to use a stochastic model capable of generating multiple data types. When using the number of cells in each subpopulation and cell trajectory data, we find estimates for the average cell cycle transition rates to range between $0.0411 - 0.193$ /h and estimates for average cell motility to range between $0.316 - 1.12$ /h. Interestingly, we find that the motility rates appear to depend upon the cell cycle phase and for this data the motility of cells in S/G2/M phase is higher than the motility rate in the in G1 or eS phase. We quantify the precision of these estimates through the CV which is a standard measure of dispersion about the mean. We find the CV to be suitably small for all parameters as it ranges from $2.65 - 5.31\%$ and $10.9 - 18.4\%$ for the transition and motility marginal posteriors, respectively. To validate our results we also draw samples from the posterior predictive distribution to determine whether the simulated data sets recovered accurately reflect the observed data sets. These results confirm that the model and summary statistics are recovering the underlying mechanisms present in the experiment.

Now that the recovery of precise parameter estimates from a fluorescent cell cycle labelling model has been demonstrated, further models can be built which are more biologically realistic. For instance, the Markov process model we used in this study describes a discrete exclusion based random walk on a 2D hexagonal lattice. However, a more biologically realistic and meaningful model would incorporate a three-dimensional (3D) environment (for example Jin et al., 2021). By constraining our model to a 2D hexagonal lattice, we ultimately omit realistically modelling: the spatial supply of oxygen, nutrients and drugs; the orientation in 3D space; and interactions with the extracellular matrix (Beaumont et al., 2014; Smalley et al., 2006). Although, increasing model complexity tends to require additional parameters in the model which in some applications may render ABC methods ill suited to inference due to their poorer performance in higher dimensions (Fearnhead & Prangle, 2012). Such modeling and inference implications would need to be considered in future work. Nevertheless, we demonstrate that the 2D stochastic model developed by Simpson et al. (2018) is able to recover key features of the experimental data set we examined and can be used to provide a quick and inexpensive alternative to *in vitro* experiments.

Finally, to bypass evaluating the likelihood function we resort to using ABC techniques. However, ABC requires many model simulations, which can be computationally expensive if the simulation model is relatively inefficient. In our application, the computation time for the model is largely dependent on the value of the transition and motility parameters, where larger values will require more computation. We compute the computational cost of 1000 simulations with parameter configurations drawn from the prior distribution and report the computational cost of the model as a 95% empirical confidence interval that ranges between 1.08-57.33 seconds per simulation. Using an Intel(R) Xeon(R) Gold 6140 CPU at 2.3GHz and paralysing over the 16 cores results in the total computation time of the SMC-ABC algorithm taking approximately 23 hours to run when using cell count and cell trajectory data and 16 hours when using cell count and cell density data. We find these computation times to be reasonable but future work may need to consider more computationally efficient modelling and/or statistical methods, particularly if more summary statistics are to be considered.

3.7 Acknowledgments

We would like to acknowledge the services of Queensland University of Technologies High Performance Computing (QUT HPC) for allowing the code used within this study to be ran on their servers. Additionally, MJC would like to thank funding provided by MJS and CD. We thank the four referees for their helpful suggestions.

3.8 Data accessibility

Rcpp and MATLAB implementations of the simulation model and SMC-ABC algorithm along with the data used in this study are available at https://github.com/michaelcarr-stats/FUCCI.

3.9 Authors' contributions

All authors designed the research. MJC performed the research, wrote the manuscript, produced the figures, and implemented the computational algorithms used in this study. All authors edited and approved this manuscript.

3.10 Competing interests

We declare we have no competing interests.

3.11 Funding

MJS is supported by the Australian Research Council (DP200100177) and CD is supported by the Australian Research Council (DP200102101).

3.12 Supplementary Material

3.12.1 Gillespie Algorithm

Algorithm S1 Simulation Model utilising Gillespie Algorithm (Gillespie, 1977) with input parameter $\theta = (R_r, R_y, R_g, M_r, M_y, M_g)$

Calculate number of red, yellow and green agents in system as N_r, N_y, N_g
while $t < t_{\max}$ **do**
    Set $a_r = M_r \times N_r$ and $t_r = R_r \times N_r$
    Set $a_y = M_y \times N_y$ and $t_y = R_y \times N_y$
    Set $a_g = M_g \times N_g$ and $t_g = R_g \times N_g$
    Set $a_0 = a_r + a_y + a_g + t_r + t_y + t_g$
    Draw $R \sim \mathcal{U}(0, 1)$
    if $R < a_r/a_0$ **then**
        Do red migration
    else if $R < (a_r + a_y)/a_0$ **then**
        Do yellow migration
    else if $R < (a_r + a_y + a_g)/a_0$ **then**
        Do green migration
    else if $R < (a_r + a_y + a_g + t_r)/a_0$ **then**
        Do red transition
        Set $N_r = N_r - 1$
        Set $N_y = N_y + 1$
    else if $R < (a_r + a_y + a_g + t_r + t_y)/a_0$ **then**
        Do yellow transition
        Set $N_y = N_y - 1$
        Set $N_g = N_g + 1$
    else if $R < (a_r + a_y + a_g + t_r + t_y + t_g)/a_0$ **then**
        Do green transition
        Set $N_g = N_g - 1$
        Set $N_r = N_r + 1$
    end if
    Draw time step $\tau \sim \exp(a0)$
    Increment time $t = t + \tau$
end while

3.12.2 Developing Summary Statistics

Cell Cycle Transition Rates

The summary statistic we explore for the cell cycle transition rates is the number of agents within each phase of the cell cycle (red, yellow, green) at time t = 48 hours. We

believe that the count of each cell type to be a good choice because the transition rates will only influence the number of cells. Therefore, we expect higher relative transition rates to correlate to lower cell counts and vice versa. For simplicity, we will assume the motility rates to be known and equal while we estimate the cell cycle transition parameters using multiple synthetic data sets generated from $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.25, 0.15, 0.22, 4, 4, 4), (0.12, 0.07, 0.03, 4, 4, 4), (0.3, 0.36, 0.28, 4, 4, 4)\}$. Using the SMC-ABC algorithm with the same summary statistics for the simulated data, we present the marginal posterior distribution of the cell cycle transition rates in Figure S1. Since the posterior distributions are all centred on the "true" value we confirm the summary statistics suitability at identifying the cell cycle transition parameters when the motility parameters are held constant and known.

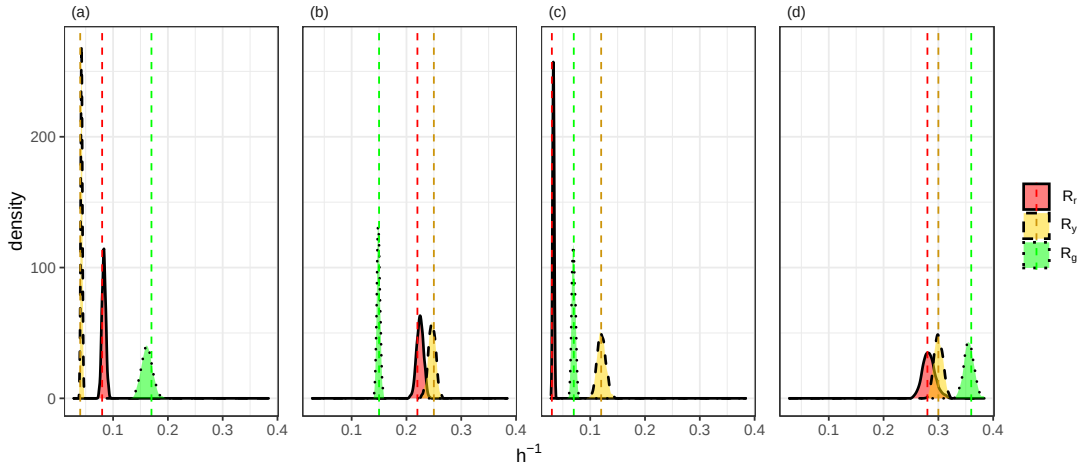

Figure S1: Using number of cells in each cell cycle phase as summary statistics for the transition rates. (a)-(d) Posterior distributions produced using synthetic data sets generated from $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.25, 0.15, 0.22, 4, 4, 4), (0.12, 0.07, 0.03, 4, 4, 4), (0.3, 0.36, 0.28, 4, 4, 4)\}$ (respectively) with true parameter values indicated by vertical dotted line.

Cell Motility Rates

We analyse the effectiveness of two summary statistics which are used to estimate the motility parameters with four synthetic data sets generated with $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.04, 0.17, 0.08, 2, 5, 8), (0.04, 0.17, 0.08, 8, 2, 5), (0.04, 0.17, 0.08, 5, 8, 2)\}$ where the transition rates are held constant. The first summary statistic we consider is the median position and interquartile range of each cell type on the left and right side of the scratched region; which we refer to as cell density data. The marginal posterior distributions are presented in Figure S2. We see that the estimates for cell motility are practically non-identifiable when cell density data is used, which is consistent with findings from Simpson et al. (2020). We believe that this may be due to interference from cells transitioning between phases and the associated difficulty in attributing the distance travelled in a phase with a single time point.

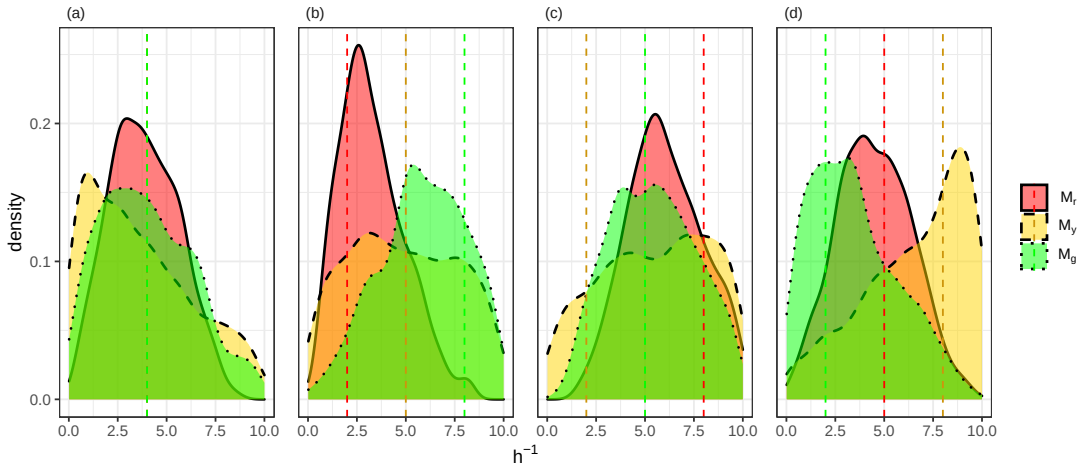

Figure S2: Using cell density data as summary statistics for the motility rates. (a)-(d) Posterior distributions produced using synthetic data sets generated from $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.04, 0.17, 0.08, 2, 5, 8), (0.04, 0.17, 0.08, 8, 2, 5), (0.04, 0.17, 0.08, 5, 8, 2)\}$ (respectively) with true parameter values indicated by vertical dotted line.

We next consider the average distance cells travels through each cell phase of the cell cycle until the cell returns to the G1 phase or the simulation is terminated. We refer to this summary statistic as cell trajectory data and test its effectiveness with the same four synthetic data sets which were used with the cell density data with 10, 20, 30, 40 and 50 individual cell trajectories. We present the marginal posterior distributions in Figure S3. We see from the concentration of the distributions around the true value (dashed line) that cell trajectory data is highly informative about the motility rates. Furthermore, we analyse the marginal benefit of increasing the number of cells to track by 10 and find that the benefit plateaus after 20 cells. Therefore, we chose the minimally suitable number of cell trajectories which produced well defined distributions to be 20 (corresponding to Figure S3 (b,g,l,q)).

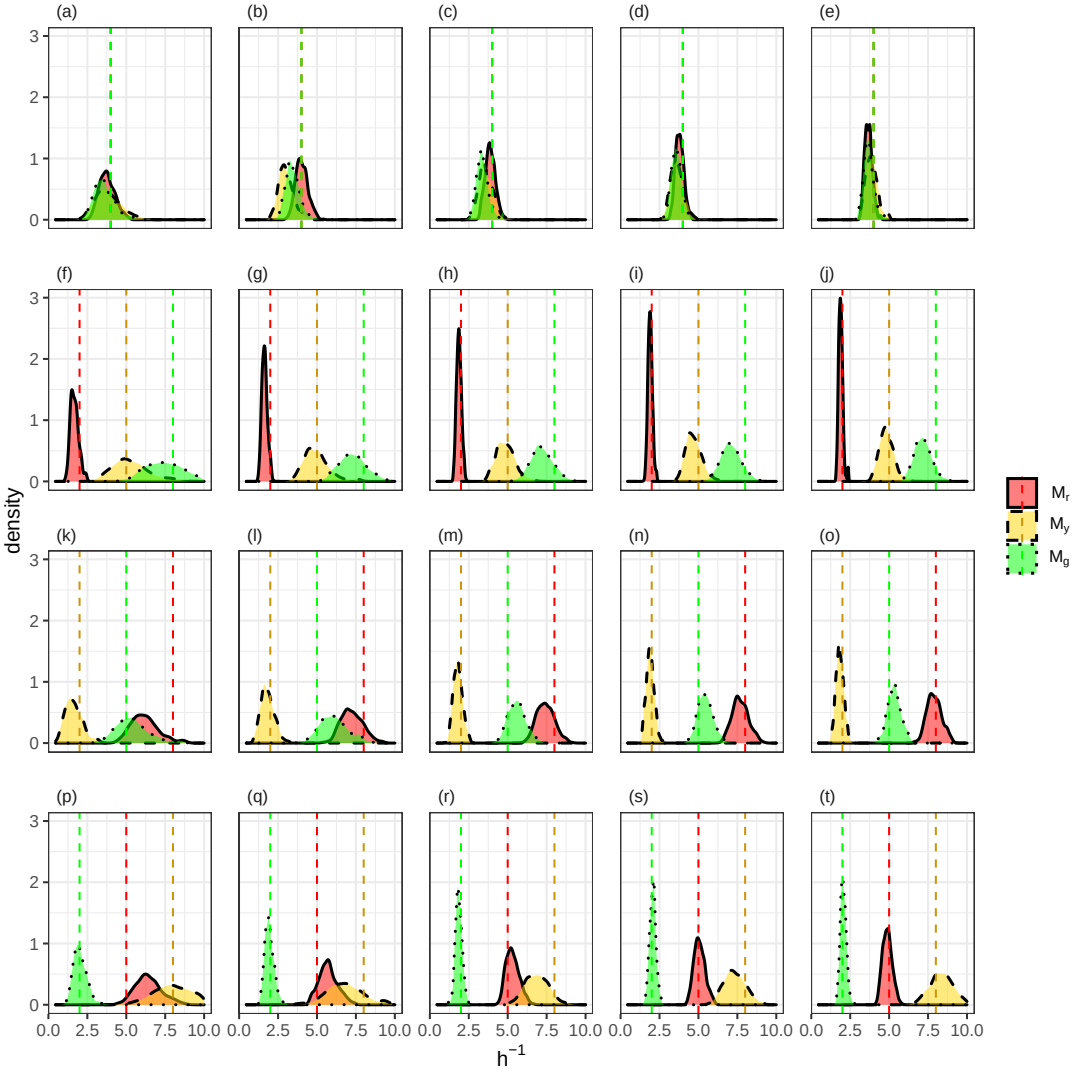

Figure S3: Using cell tracking data as summary statistics for the motility rates. Synthetic data sets generated from $\theta \in \{(0.04, 0.17, 0.08, 4, 4, 4), (0.04, 0.17, 0.08, 2, 5, 8), (0.04, 0.17, 0.08, 8, 2, 5), (0.04, 0.17, 0.08, 5, 8, 2)\}$ are varying down the rows and number of cells tracked increases by 10 across the columns.

Chapter 4

Conclusion

4.1 Summary

In this thesis, we develop a parameter estimation method for estimating parameters of a stochastic cell invasion model which considers proliferation, migration and crowding effects. To achieve this goal, in Chapter 3, we assess the strengths and weaknesses of multiple ABC algorithms relative to the applicaiton and find the SMC-ABC algorithm (Drovandi & Pettitt, 2011) to be suitable for this application. Additionally, we explore the informativeness of several summary statistics to estimate the model parameters with respect to multiple biologically plausible synthetic data sets. Our analysis finds using the number of cells in each phase of the cell cycle at the end of the experiment and the average distance travelled by 20 cells to be highly informative about the cell cycle transition and motility rates, respectively. Finally, using the experimental data, we successfully estimate the parameters of the stochastic cell spreading model using the aforementioned summary statistics. Furthermore, we demonstrate that cell density data (median position and interquartile range of cells in each cell phase on left and right side of scratched region) in place of cell trajectory data provides insufficient information about the motility parameters such that they are non-identifiable. These results are consistent with Simpson et al. (2020) who investigate a simpler deterministic model of collective cell spreading with multiple cell cycle phases. Importantly, this thesis demonstrates the first successful parameter calibration of a stochastic model with multiple phases of the cell cycle.

4.2 Discussion and Future Research

In this thesis, we adopt the stochastic cell invasion model developed by Simpson et al. (2018). Although, there are several limitations associated with this modelling approach. Firstly, we consider a discrete exclusion based random walk on a 2D hexagonal

lattice to mimic scratch assay experiments. However, a more biologically realistic and meaningful model would incorporate 3D environment (for example see Jin et al., 2021) which mimics tumour spheroids. Previous studies have indicated that results from experiments which have been conducted under a 2D study do not necessarily translate to a 3D environment (Desoize et al., 1998). This is due to the 2D model omitting rational features such as the spatial supply of oxygen, nutrients and drugs; the orientation in 3D space; and interactions with the extracellular matrix (Beaumont et al., 2014; Smalley et al., 2006). Nevertheless, the 2D model used in this study still provides a quick and inexpensive alternative to *in vitro* experiments. In future research, one possible extension of this study would be to construct a 3D model and estimate the parameters.

One of the key motivations for adopting a stochastic modelling approach was its flexibility to generate multiple data types. In our study, we found that using the number of cells in each cell phase and the average distance travelled by 20 cell trajectories to be informative summary statistics and lead to a well defined posterior being produced. However, working with cell trajectory data can be challenging due to the time consuming nature of manually tracking cells and the need for experiments to be performed in low density to make tracking easier. An alternative approach to cell tracking of interest uses cell positional data over multiple time points to measure cells aggregation toward an area (Hywood et al., 2021). This study measures cell aggregation by repeatedly simulating a 2D biased agent based random walk at each time step to estimate the drift and diffusion coefficients in the PDE model. The directional bias of the cells is governed by the idealised chemokine concentration with the angle of movement drawn from the von Misses distribution. They demonstrate the effectiveness of this method with a PDE model of cytotoxic T cells interacting with tumour spheroid cells and find that good estimates for the drift and diffusivity are produced. Using a similar approach, where the directional bias of melanoma cells in a scratch assay is modeled to estimate cell diffusivity and drift, could be a worthwhile future endeavour.

In our research we resolve the issue of the computationally intractable likelihood function by using ABC techniques. However, since many simulations are required ABC itself can be computationally intensive - especially when the model is expensive to simulate. Furthermore, development of more realistic but more complex models can be hindered by poor performance of ABC in higher dimensions if too many additional parameters are incorporated. Therefore, alternative approaches which are both more computationally efficient and scale better to higher dimensions are necessary to develop more realistic models. A more recent likelihood-free inference method which we did not explore in this study is Bayesian Synthetic Likelihood (BSL). This method assumes a Gaussian parametric form of the likelihood $\pi(S_y|\theta) \approx \mathcal{N}(S_y; \mu_n(\theta), \Sigma_n(\theta))$ where unbiased estimates for $\mu_n(\theta)$ and $\Sigma_n(\theta)$ can be estimated (provided that the summary

statistics are Gaussian (Andrieu & Roberts, 2009)) with Monte Carlo integration:

$$\mu_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(x_i), \text{and}$$
$$\Sigma_n(\theta) = \frac{1}{n-1}\sum_{i=1}^{n}(S(x_i) - \mu_n(\theta))(S(x_i) - \mu_n(\theta))^\top.$$

Importantly, the value of n can be chosen to maximise computational efficiency, where large n produces precise estimates but slow estimates and small n produces imprecise but fast estimates (Price et al., 2018). Frazier et al. (2019) demonstrate, under some assumptions, that BSL is more computationally efficient than ABC for any dimension of the summary statistic due to using a parametric approximation to the likelihood. Moreover, the difference in computational efficiency has empirically been shown to be more prominent as the dimension of the summary statistics increases (Price et al., 2018). However, BSL is ill suited when the distribution of summary statistics is non-Gaussian and can produce unreliable estimates when the model is misspecified. Although, estimates can still be reasonable when the summary statistics are non-Gaussian but the distribution remains regular (An et al., 2020; Price et al., 2018). Furthermore, Frazier and Drovandi (2021) addresses BSL's poor performance under model misspecification with a robust BSL algorithm. Hence, in future research, the use of BSL may be of interest to explore for computationally expensive stochastic cell models and/or those with high dimensional parameter spaces.

Bibliography

An, Z., Nott, D. J. & Drovandi, C. (2020). Robust Bayesian synthetic likelihood via a semi-parametric approach. *Statistics and Computing*, *30*(3), 543–557.

Andrieu, C. & Roberts, G. O. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, *37*(2), 697–725.

Australian Institute of Health and Welfare. (2018). Cancer in Australia: Actual incidence data from 1982 to 2013 and mortality data from 1982 to 2014 with projections to 2017. *Asia-Pacific Journal of Clinical Oncology*, *14*(1), 5–15.

Beaumont, K. A., Mohana-Kumaran, N. & Haass, N. K. (2014). Modeling melanoma in vitro and in vivo. *Healthcare*, *2*(1), 27–46.

Beaumont, M. A., Cornuet, J.-M., Marin, J.-M. & Robert, C. P. (2009). Adaptive approximate bayesian computation. *Biometrika*, *96*(4), 983–990.

Beaumont, M. A., Zhang, W. & Balding, D. J. (2002). Approximate Bayesian computation in population genetics. *Genetics*, *162*(4), 2025–2035.

Blum, M. G. & François, O. (2010). Non-linear regression models for approximate Bayesian computation. *Statistics and computing*, *20*(1), 63–73.

Blum, M. G., Nunes, M. A., Prangle, D. & Sisson, S. A. (2013). A comparative review of dimension reduction methods in approximate Bayesian computation. *Statistical Science*, *28*(2), 189–208.

Cai, A. Q., Landman, K. A. & Hughes, B. D. (2007). Multi-scale modeling of a wound-healing cell migration assay. *Journal of Theoretical Biology*, *245*(3), 576–594.

Carlin, B. P. & Louis, T. A. (2000). *Bayes and empirical Bayes methods for data analysis*. Chapman & Hall/CRC,

Carr, M. J., Simpson, M. J. & Drovandi, C. (2021). Estimating parameters of a stochastic cell invasion model with fluorescent cell cycle labelling using approximate Bayesian computation. *Journal of the Royal Society Interface*, *18*(182).

Chopin, N. (2002). A sequential particle filter method for static models. *Biometrika*, *89*(3), 539–552.

Codling, E. A., Plank, M. J. & Benhamou, S. (2008). Random walk models in biology. *Journal of the Royal Society Interface*, *5*(25), 813–834.

Desoize, B., Gimonet, D. & Jardiller, J. (1998). Cell culture as spheroids: An approach to multicellular resistance. *Anticancer Research*, *18*(6A), 4147.

Drovandi, C. C. & Pettitt, A. N. (2011). Estimation of parameters for macroparasite population evolution using approximate Bayesian computation. *Biometrics*, *67*(1), 225–233.

Edelstein-Keshet, L. (2005). *Mathematical models in biology*. SIAM.

El-Hachem, M., McCue, S. W. & Simpson, M. J. (2021). Travelling wave analysis of cellular invasion into surrounding tissues. *arXiv preprint arXiv:2105.04730*.

Epanechnikov, V. A. (1969). Non-parametric estimation of a multivariate probability density. *Theory of Probability & Its Applications*, *14*(1), 153–158.

Ermentrout, G. B. & Edelstein-Keshet, L. (1993). Cellular automata approaches to biological modeling. *Journal of Theoretical Biology*, *160*(1), 97–133.

Fearnhead, P. & Prangle, D. (2012). Constructing summary statistics for approximate Bayesian computation: Semi-automatic approximate Bayesian computation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *74*(3), 419–474.

Fisher, R. A. (1937). The wave of advance of advantageous genes. *Annals of eugenics*, *7*(4), 355–369.

Flegal, J. M. & Herbei, R. (2012). Exact sampling for intractable probability distributions via a bernoulli factory. *Electronic Journal of Statistics*, *6*, 10–37.

Frazier, D. T. & Drovandi, C. (2021). Robust approximate Bayesian inference with synthetic likelihood. *Journal of Computational and Graphical Statistics*, 1–19.

Frazier, D. T., Nott, D. J., Drovandi, C. & Kohn, R. (2019). Bayesian inference using synthetic likelihood: Asymptotics and adjustments. *arXiv preprint arXiv:1902.04827*.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013). *Bayesian data analysis*. CRC press.

Gelman, A., Roberts, G. O. & Gilks, W. R. (1996). Efficient metropolis jumping rules. *Bayesian statistics*, *5*(599-608), 42.

Giblin, A. V. & Thomas, J. M. (2007). Incidence, mortality and survival in cutaneous melanoma. *Journal of Plastic, Reconstructive & Aesthetic Surgery*, *60*(1), 32–40.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry*, *81*(25), 2340–2361.

Guillemaud, T., Beaumont, M. A., Ciosi, M., Cornuet, J.-M. & Estoup, A. (2010). Inferring introduction routes of invasive species using approximate Bayesian computation on microsatellite data. *Heredity*, *104*(1), 88–99.

Haass, N. K., Beaumont, K. A., Hill, D. S., Anfosso, A., Mrass, P., Munoz, M. A., Kinjyo, I. & Weninger, W. (2014). Real-time cell cycle imaging during melanoma growth, invasion, and drug response. *Pigment Cell & Melanoma Research*, *27*(5), 764–776.

Haass, N. K. & Gabrielli, B. (2017). Cell cycle-tailored targeting of metastatic melanoma: Challenges and opportunities. *Experimental Dermatology*, *26*(7), 649–655.

Hamilton, G., Stoneking, M. & Excoffier, L. (2005). Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proceedings of the National Academy of Sciences*, *102*(21), 7476–7480.

Harrison, J. U. & Baker, R. E. (2020). An automatic adaptive method to combine summary statistics in approximate Bayesian computation. *PloS one*, *15*(8), e0236954.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Ho, L. S. T., Xu, J., Crawford, F. W., Minin, V. N. & Suchard, M. A. (2018). Birth/birth-death processes and their computable transition probabilities with biological applications. *Journal of mathematical biology*, *76*(4), 911–944.

Hoffman, M. D. & Gelman, A. (2014). The no-u-turn sampler: Adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, *15*(1), 1593–1623.

Hywood, J. D., Rice, G., Pageon, S. V., Read, M. N. & Biro, M. (2021). Detection and characterization of chemotaxis without cell tracking. *Journal of the Royal Society Interface*, *18*(176), 20200879.

Jin, W., Spoerri, L., Haass, N. K. & Simpson, M. J. (2021). Mathematical model of tumour spheroid experiments with real-time cell cycle imaging. *Bulletin of Mathematical Biology*, *83*(5), 1–23.

Johnston, S. T., Ross, J. V., Binder, B. J., McElwain, D. S., Haridas, P. & Simpson, M. J. (2016). Quantifying the effect of experimental design choices for in vitro scratch assays. *Journal of Theoretical Biology*, *400*, 19–31.

Kolmogorov, A. N. (1937). Étude de l'équation de la diffusion avec croissance de la quantité de matière et son application à un problème biologique. *Bull. Univ. Moskow, Ser. Internat., Sec. A*, *1*, 1–25.

Kursawe, J., Baker, R. E. & Fletcher, A. G. (2018). Approximate bayesian computation reveals the importance of repeated measurements for parameterising cell-based models of growing tissues. *Journal of theoretical biology*, *443*, 66–81.

Liang, C.-C., Park, A. Y. & Guan, J.-L. (2007). In vitro scratch assay: A convenient and inexpensive method for analysis of cell migration in vitro. *Nature protocols*, *2*(2), 329.

Maini, P. K., McElwain, D. S. & Leavesley, D. I. (2004). Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Engineering*, *10*(3-4), 475–482.

Marjoram, P., Molitor, J., Plagnol, V. & Tavaré, S. (2003). Markov chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *100*(26), 15324–15328.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of chemical physics*, *21*(6), 1087–1092.

Moler, C. & Van Loan, C. (2003). Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, *45*(1), 3–49.

Murray, J. D. (2007). *Mathematical biology: I. An introduction* (Vol. 17). Springer Science & Business Media.

Painter, K. J. & Sherratt, J. A. (2003). Modelling the movement of interacting cell populations. *Journal of theoretical biology*, *225*(3), 327–339.

Parkin, D. M., Bray, F., Ferlay, J. & Pisani, P. (2005). Global cancer statistics, 2002. *CA: a cancer journal for clinicians*, *55*(2), 74–108.

Perez-Carrasco, R., Beentjes, C. & Grima, R. (2020). Effects of cell cycle variability on lineage and population measurements of messenger rna abundance. *Journal of the Royal Society Interface*, *17*(168), 20200360.

Price, L. F., Drovandi, C. C., Lee, A. & Nott, D. J. (2018). Bayesian synthetic likelihood. *Journal of Computational and Graphical Statistics*, *27*(1), 1–11.

Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. (1999). Population growth of human y chromosomes: A study of y chromosome microsatellites. *Molecular Biology and Evolution*, *16*(12), 1791–1798.

R Core Team. (2020). R: A language and environment for statistical computing. https://www.R-project.org/

Raue, A., Kreutz, C., Maiwald, T., Bachmann, J., Schilling, M., Klingmüller, U. & Timmer, J. (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics*, *25*(15), 1923–1929.

Robert, C. & Casella, G. (2013). *Monte carlo statistical methods*. Springer Science & Business Media.

Roberts, G. O. & Stramer, O. (2002). Langevin diffusions and metropolis-hastings algorithms. *Methodology and computing in applied probability*, *4*(4), 337–357.

Rueden, C. T., Schindelin, J., Hiner, M. C., DeZonia, B. E., Walter, A. E., Arena, E. T. & Eliceiri, K. W. (2017). Imagej2: Imagej for the next generation of scientific image data. *BMC bioinformatics*, *18*(1), 1–26.

Sakaue-Sawano, A., Kurokawa, H., Morimura, T., Hanyu, A., Hama, H., Osawa, H., Kashiwagi, S., Fukami, K., Miyata, T., Miyoshi, H., Imamura, T., Ogawa, M., Masai, H. & Miyawaki, A. (2008). Visualizing spatiotemporal dynamics of multicellular cell-cycle progression. *Cell*, *132*(3), 487–498.

Santiago-Walker, A., Li, L., Haass, N. & Herlyn, M. (2009). Melanocytes: From morphology to application. *Skin Pharmacology and Physiology*, *22*(2), 114–121.

Savla, U., Olson, L. E. & Waters, C. M. (2004). Mathematical modeling of airway epithelial wound closure during cyclic mechanical strain. *Journal of Applied Physiology*, *96*(2), 566–574.

Schnoerr, D., Sanguinetti, G. & Grima, R. (2017). Approximation and inference methods for stochastic biochemical kinetics—a tutorial review. *Journal of Physics A: Mathematical and Theoretical*, *50*(9), 093001.

Sidje, R. B. (1998). Expokit: A software package for computing matrix exponentials. *ACM Transactions on Mathematical Software (TOMS)*, *24*(1), 130–156.

Simpson, M. J., Baker, R. E., Vittadello, S. T. & Maclaren, O. J. (2020). Practical parameter identifiability for spatio-temporal models of cell invasion. *Journal of the Royal Society Interface*, *17*(164), 20200055.

Simpson, M. J., Jin, W., Vittadello, S. T., Tambyah, T. A., Ryan, J. M., Gunasingh, G., Haass, N. K. & McCue, S. W. (2018). Stochastic models of cell invasion with fluorescent cell cycle indicators. *Physica A: Statistical Mechanics and its Applications*, *510*, 375–386.

Sisson, S. A., Fan, Y. & Beaumont, M. (2018). *Handbook of approximate Bayesian computation*. CRC Press.

Sisson, S. A., Fan, Y. & Tanaka, M. M. (2007). Sequential Monte Carlo without likelihoods. *Proceedings of the National Academy of Sciences*, *104*(6), 1760–1765.

Smalley, K. S., Lioni, M. & Herlyn, M. (2006). Life isn't flat: Taking cancer biology to the next dimension. *In Vitro Cellular & Developmental Biology-Animal*, *42*(8-9), 242–247.

Swanson, K. R. (2008). Quantifying glioma cell growth and invasion in vitro. *Mathematical and Computer Modelling*, *47*(5-6), 638–648.

Takamizawa, K., Niu, S. & Matsuda, T. (1997). Mathematical simulation of unidirectional tissue formation: In vitro transanastomotic endothelialization model. *Journal of Biomaterials Science, Polymer Edition*, *8*(4), 323–334.

Tavaré, S., Balding, D. J., Griffiths, R. C. & Donnelly, P. (1997). Inferring coalescence times from dna sequence data. *Genetics*, *145*(2), 505–518.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *the Annals of Statistics*, 1701–1728.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A. & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, *6*(31), 187–202.

Treloar, K. K., Simpson, M. J., Haridas, P., Manton, K. J., Leavesley, D. I., McElwain, D. S. & Baker, R. E. (2013). Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. *BMC Systems Biology*, *7*(1), 137.

Vittadello, S. T., McCue, S. W., Gunasingh, G., Haass, N. K. & Simpson, M. J. (2018). Mathematical models for cell migration with real-time cell cycle dynamics. *Biophysical Journal*, *114*(5), 1241–1253.

Vo, B. N., Drovandi, C. C., Pettitt, A. N. & Simpson, M. J. (2015). Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Mathematical Biosciences*, *263*, 133–142.

Weyant, A., Schafer, C. & Wood-Vasey, W. M. (2013). Likelihood-free cosmological inference with type Ia supernovae: Approximate Bayesian computation for a complete treatment of uncertainty. *The Astrophysical Journal*, *764*(2), 116.