Inferring parameters for a lattice-free model of cell migration and proliferation using experimental data.

Alexander P. Browning[a], Scott W. McCue[a], Rachelle N. Binny[b,c,d], Michael J. Plank[b,d], Esha T. Shah[e], Matthew J. Simpson[a,*]

[a] *School of Mathematical Sciences, Queensland University of Technology (QUT), Brisbane, Australia.*
[b] *Landcare Research, Lincoln, Canterbury, New Zealand.*
[c] *Biomathematics Research Centre, University of Canterbury, Christchurch, New Zealand.*
[d] *Te Pūnaha Matatini, a New Zealand Centre of Research Excellence, New Zealand.*
[e] *Ghrelin Research Group, Translational Research Institute, QUT, 37 Kent St, Woolloongabba, Queensland, Australia.*

Abstract

Collective cell spreading takes place in spatially continuous environments, yet it is often modelled using discrete lattice-based approaches. Here, we use data from a series of cell proliferation assays, with a prostate cancer cell line, to calibrate a spatially continuous individual based model (IBM) of collective cell migration and proliferation. The IBM explicitly accounts for crowding effects by modifying the rate of movement, direction of movement, and the rate of proliferation by accounting for pair-wise interactions. Taking a Bayesian approach we estimate the free parameters in the IBM using rejection sampling on three separate, independent experimental data sets. Since the posterior parameter estimates from each experiment are similar, we combine the estimates. Performing simulations with parameters sampled from the combined distribution allows us to confirm the predictive power of the calibrated IBM by accurately forecasting the evolution of a fourth, experimental data set. Overall, we show how to calibrate a lattice-free IBM to experimental data, and our work highlights the importance of interactions between individuals. Despite great care taken to distribute cells as uniformly as possible experimentally, we find evidence of significant spatial clustering over short distances, suggesting that standard mean-field models could be inappropriate.

Keywords: individual based model, cell migration, model calibration, cell proliferation assay, approximate Bayesian computation

[*] Corresponding author at: Mathematical Sciences, QUT, Brisbane, Australia. Tel.:+617 3138 5241; fax:+617 3138 2310

Email address: matthew.simpson@qut.edu.au (Matthew J. Simpson)

1. Introduction

One of the most common *in vitro* cell biology experiments is called a *cell proliferation assay* (Bosco et al., 2015; Bourseguin et al., 2016; Browning et al., 2017). These assays are conducted by placing a monolayer of cells, at low density, on a two-dimensional substrate. Individual cells undergo proliferation and movement events, and the assay is monitored over time as the density of cells in the monolayer increases (Tremel et al., 2009). One approach to interpret a cell proliferation assay is to use a mathematical model. Calibrating the solution of a mathematical model to data from a cell proliferation assay can provide quantitative insight into the underlying mechanisms, by, for example, estimating the cell proliferation rate (Tremel et al., 2009; Sengers et al., 2007). A standard approach to modelling a cell proliferation assay is to use a mean-field model, which is equivalent to assuming that individuals within the population interact in proportion to the average population density and that there is no spatial structure, such as clustering, present (Tremel et al., 2009; Sengers et al., 2007; Maini et al., 2004b; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990). More recently, increased computational power has meant that individual based models (IBMs) have been used to directly model the cell-level behaviour (Binny et al., 2016a; Frascoli et al., 2013; Johnston et al., 2014). IBMs are attractive for modelling biological phenomena because they can be used to represent properties of individual agents, such as cells, in the system of interest (Binny et al., 2016a,b; Frascoli et al., 2013; Peirce et al., 2004; Read et al., 2012; Treloar et al., 2013). Typical IBMs use a lattice, meaning that both the position of agents, and the direction of movement, are restricted (Codling et al., 2008). In contrast, lattice-free IBMs are more realistic because they enable agents to move in continuous space, in any direction. However, this extra freedom comes at the cost of higher computational requirements (Plank and Simpson, 2012).

In this work we consider a continuous-space, continuous-time IBM (Binny et al., 2016b). This IBM is well-suited to studying experimental data from a cell proliferation assay with PC-3 prostate cancer cells (Kaighn et al., 1979), as shown in Figure 1(a)-(d). The key mechanisms in the experiments include cell migration and cell proliferation, and we note that there is no cell death in the experiments on the time scales that we consider. Therefore, agents in the IBM are allowed to undergo both proliferation and movement events. Crowding effects that are often observed in two-dimensional cell biology experiments (Cai et al., 2007) are explicitly incorporated into the IBM as the rates of proliferation and movement in the model are inhibited in regions of high agent density. In this study we specifically choose to work with the PC-3 cell line because these cells are known to be highly migratory, mesenchymal cells (Kaighn et al., 1979). This means that cell-to-cell adhesion is minimal for this cell line, and cells tend to migrate as individuals. We prefer to work with a continuous-space, lattice-free IBM as this framework gives us the freedom to identically replicate the initial location of all cells in the experimental data when we specify the initial condition in the IBM. In addition, lattice-free IBMs do not restrict the direction of movement like a lattice-based approach.

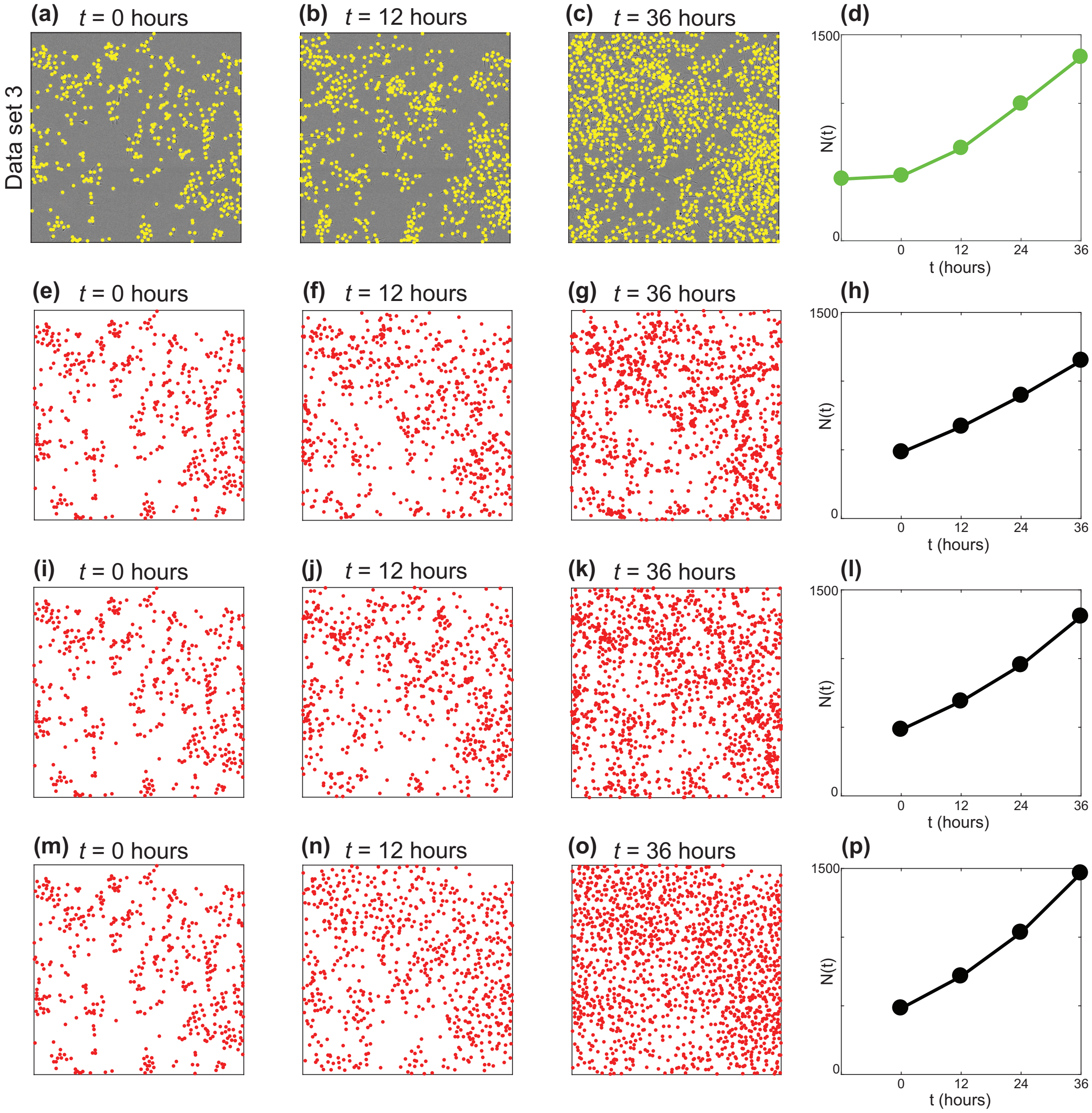

Fig. 1: (a)-(c) Experimental data set 3 at $t = 0$, 12 and 36 hours. The position of each cell is identified with a yellow marker. The field of view is a square of length 1440 μm. (d) Population size, $N(t)$ for experimental data set 3. (e)-(h) One realisation of the IBM with $\gamma_b = 0$ μm, leading to an overly clustered distribution of agents. (i)-(l) One realisation of the IBM with $\gamma_b = 4$ μm, leading to a distribution of agents with similar clustering to the experimental data. (m)-(p) One realisation of the IBM with $\gamma_b = 20$ μm, leading to an overly segregated distribution of agents. All IBM simulations are initiated using the same distribution of agents as in (a), with $m = 0.56$ /hour, $p = 0.041$ /hour, and $\sigma = 24$ μm.

A key contribution of this study is to demonstrate how the IBM can be calibrated to experimental data. In particular, we use approximate Bayesian computation (ABC) to infer the parameters in the IBM. Four sets of experimental images (Supplementary material 1), each corresponding to an identically-prepared proliferation assay, are considered. The experiments were conducted over a duration of 48 hours, which is unusual because proliferation assays are typically conducted for no more than 24 hours (Browning et al.,2017). Data from the first three sets of experiments (Figure 2) are used to calibrate the IBM and data from the fourth set of images is used to examine the predictive capability of the calibrated IBM. The IBM that we work with was presented very recently (Binny et al., 2016b). The description of the IBM by Binny et al. (2016b) involves a discussion of the mechanisms in the model and the derivation of a spatial moment continuum description (Binny et al.,2016b). IBMs are rarely calibrated to experimental data, and our current work is the first time experimental data has been used to provide parameter estimates for the new IBM.

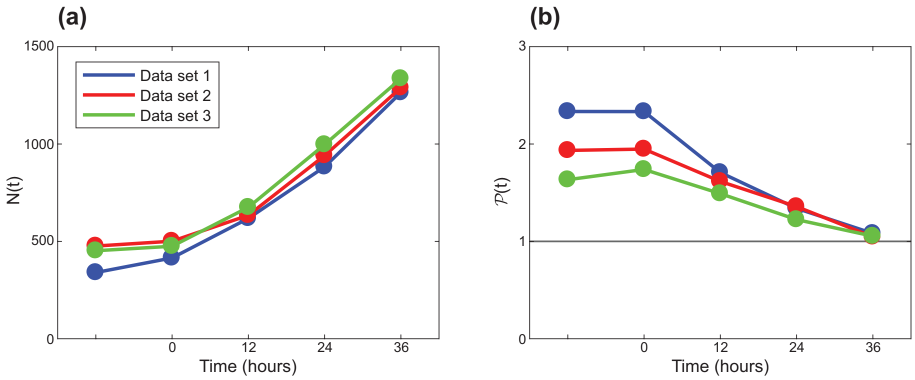

Fig. 2: Summary statistics for experimental data sets 1, 2 and 3, shown in blue, red and green, respectively. (a) Population size, $N(t)$. (b) Pair correlation, $\mathcal{P}(t)$. Unprocessed experimental data are given in Supplementary material documents 1 and 2.

Taking a Bayesian approach, we assume that cell proliferation assays are stochastic processes, and model parameters are random variables, allowing us to update information about the model parameters using ABC (Collis et al., 2017; Tanaka et al., 2006). For this purpose we perform a large number of IBM simulations using parameters sampled from a prior distribution. Previous work, based on mean-field models, suggests that the proliferation rate and cell diffusivity for PC-3 cells is $\lambda \approx 0.05$ /hour and $D \approx 175$ μm^2/hour, respectively (Johnston et al., 2015). The prior distribution for the IBM parameters are taken to be uniform and to encompass these previous estimates. We generate 10^6 realisations of the IBM using parameters sampled from the prior distribution, and accept 1% of simulations that provide the best match to the experimental data. Our approach to connect the experimental data and the IBM is novel, we are unaware of any previous work that has used ABC to parameterise a lattice-free IBM of a cell proliferation assay.

Applying the ABC algorithm to data from three sets of identically prepared experiments leads to three similar posterior distributions. This result provides confidence that the IBM is a realistic representation of the cell proliferation assays and leads us to produce a combined posterior distribution from which we use the mode to give point estimates of the model parameters. To provide further validation of the IBM, we use the combined posterior distribution and the IBM to make a prediction of the fourth experimental data set. Simulating the IBM with parameters sampled from the combined posterior distribution allows us to predict both the time evolution of the population size, $N(t)$, and the pair correlation within a small neighbourhood of radius 50 μm, $\mathcal{P}(t)$, which provides a measure of spatial structure. These results indicate that the *in silico* predictions are consistent with the experimental observations.

This manuscript is organised as follows. Sections 2.1-2.2 describe the experiments and the IBM, respectively. In Section 2.3 we explain how to apply the ABC algorithm to estimate the IBM parameters. In Section 3 we present the marginal posterior distributions of the IBM parameters using data from the first three sets of experiments. The predictive power of the calibrated IBM is demonstrated by using the combined marginal posterior distributions to predict the fourth experimental data set. The predictive power of the calibrated IBM is compared with the standard mean-field logistic equation (Murray, 2002). While both models can accurately predict $N(t)$, the logistic equation provides no information about the spatial structure in the experimental data. Finally, in Section 4, we conclude and summarise opportunities for further research.

2. Material and methods

2.1. Experimental methods

We perform a series of proliferation assays using the IncuCyte ZOOM$^{\text{TM}}$ live cell imaging system (Essen BioScience, MI USA) (Jin et al., 2017). All experiments are performed using the PC-3 prostate cancer cell line (Kaighn et al., 1979). These cells, originally purchased from American Type Culture Collection (Manassas, VA, USA), are a gift from Lisa Chopin (April, 2016). The cell line is used according to the National Health and Medical Research Council (NHMRC) National statement on ethical conduct in human research with ethics approval for the QUT Human Research Ethics Committee (QUT HREC 59644, Chopin). Cells are propagated in RPMI 1640 medium (Life Technologies, Australia) with 10% foetal calf serum (Sigma-Aldrich, Australia), 100 U/mL penicillin, and 100 μg/mL streptomycin (Life Technologies), in plastic tissue culture flasks (Corning Life Sciences, Asia Pacific). Cells are cultured in 5% CO_2 and 95% air in a Panasonic incubator (VWR International) at 37 °C. Cells are regularly screened for *Mycoplasma*.

Approximately 8,000 cells are distributed in the wells of the tissue culture plate as uniformly as possible. After seeding, cells are grown overnight to allow for attachment and some subsequent growth. The plate is placed into the IncuCyte ZOOM$^{\text{TM}}$ apparatus, and images showing a field of view of size 1440 × 1440 μm are recorded every 12 hours for a total duration of 48 hours. An example of a set of experimental images is shown in Figure 1(a)-(c), while images from the other three data sets are provided in Supplementary material 1.

Experimental images are recorded at five time points, at intervals of 12 hours, giving $t' = 0, 12, 24, 36$ and 48 hours. Comparing the evolution of $N(t')$ in Figure 2(a) shows the number of cells in some experiments do not increase appreciably during the first 12 hours. This suggests that the cells may experience a settling phase, so some time is required for the cells to commence normal proliferation (Tremel et al., 2009; Jin et al., 2017). Therefore, we treat the image at $t' = 12$ hours as the first image after the settling phase, and shift time, $t = t' - 12$ hours. Therefore, excluding the first experimental image at $t' = 0$ hours, we have images recorded at four time points after the settling time, $t = 0, 12, 24$ and 36 hours.

2.2. Mathematical model

2.2.1. Individual based model

We consider an IBM describing the proliferation and movement of individual cells (Binny et al., 2016a,b). Since cell death is not observed in the experiments, the IBM does not include agent death. The IBM allows the net proliferation rate and the net movement rate of agents to depend on the spatial arrangement of other agents. To be consistent with previous experimental observations, the IBM incorporates a biased movement mechanism so that agents tend to move away from nearby crowded regions (Cai et al., 2007). We use the IBM to describe the dynamics of a population of agents on a square domain of length $L = 1440$ μm to match the field-of-view of the experimental data (Figure 1(a)-(c)). Agents in the model are treated as a series of points which we may interpret as a population of uniformly-sized discs with diameter $\sigma = 24$ μm (Supplementary material 1). Each agent has location $\mathbf{x}_n = (x_1, x_2)$, for $n = 1, ..., N(t)$. Since the field-of-view of each image is much smaller than the size of the well in the tissue culture plate, we apply periodic boundary conditions [16].

Proliferation and movement events occur according to a Poisson process over time (Binny et al., 2016b). The nth agent is associated with neighbourhood-dependent rates, $P_n \geq 0$ and $M_n \geq 0$, of proliferation and movement, respectively. These rates consist of intrinsic components, $p > 0$ and $m > 0$, respectively. Crowding effects are introduced by reducing the intrinsic rates by a contribution from other neighbouring agents. These crowding effects are calculated using a kernel, $w^{(\cdot)}(r)$, that depends on the separation distance, $r \geq 0$, so that

$$P_n = \max\left(0, p - \sum_{i \neq n}^{N(t)} w^{(p)}(r)\right), \tag{1}$$

$$M_n = \max\left(0, m - \sum_{i \neq n}^{N(t)} w^{(m)}(r)\right). \tag{2}$$

Following Binny et al.,(2016), we specify the kernels to be Gaussian with width corresponding to the cell diameter, σ, giving

$$w^{(p)}(r) = \gamma_p \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{3}$$

$$w^{(m)}(r) = \gamma_m \exp\left(-\frac{r^2}{2\sigma^2}\right). \tag{4}$$

Here, γ_p is the value of $w^{(p)}(0)$ and γ_m is the value of $w^{(m)}(0)$. These parameters provide a measure of the strength of crowding effects on agent proliferation and movement, respectively. The kernels, $w^{(p)}(r)$ and $w^{(m)}(r)$, ensure that the interactions between pairs of agents separated by more than roughly 2-3 cell diameters lead to a negligible contribution. For computational efficiency, we truncate the Gaussian kernels so that $w^{(p)}(r) = w^{(m)}(r) = 0$, for $r \geq 3\sigma$ (Law et al., 2003).

To reduce the number of unknown parameters in the IBM, we specify γ_p and γ_m by invoking an assumption about the maximum packing density of the population. Here we suppose that the net proliferation and net movement rates reduce to zero when the agents are packed at the maximum possible density, which is a hexagonal packing (Figure 3(a)). For interactions felt between the nearest neighbours only (Figure 3(b)), we obtain

$$\gamma_p = \frac{p}{6}\exp\left(\frac{1}{2}\right), \tag{5}$$

$$\gamma_m = \frac{m}{6}\exp\left(\frac{1}{2}\right), \tag{6}$$

which effectively specifies a relationship between γ_p and p, and between γ_m and m. Note that this assumption does not preclude a formation of agents in which some pairs have a separation of less than σ and densities greater than hexagonal packing, which can occur by chance.

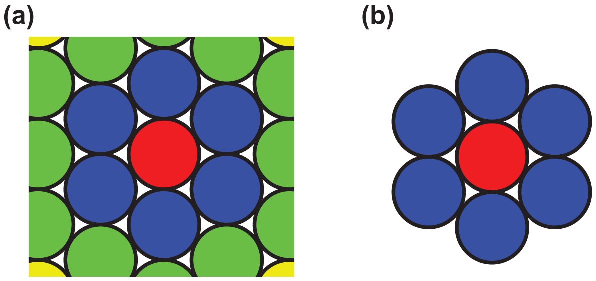

Fig. 3: (a) Hexagonal packing of uniformly sized discs. The focal agent (red) is surrounding by six nearest neighbouring agents (blue), and twelve next nearest neighbouring agents (green). (b) Hexagonal packing around a focal agent (red) showing the six nearest neighbours only.

When an agent at $\mathbf{x}_n$ proliferates, the location of the daughter agent is selected by sampling from a bivariate normal distribution with mean $\mathbf{x}_n$ and variance σ^2 (Binny et al., 2016b). Since mesenchymal cells in two-dimensional cell culture are known to move with a directional movement bias away from regions of high density (Cai et al., 2007), we allow the model to incorporate a bias so that the preferred direction of movement is in the direction of decreasing agent density. For simplicity, the distance that each agent steps is taken to be a constant, equal to the cell diameter, σ (Plank and Simpson, 2012).

To choose the movement direction, we use a crowding surface, $B(\mathbf{x})$, to measure the local crowdedness at location $\mathbf{x}$, given by

$$B(\mathbf{x}) = \sum_{i=1}^{N(t)} w^{(b)}(\|\mathbf{x} - \mathbf{x}_i\|). \tag{7}$$

The crowding surface is the sum of contributions from every agent, given by a bias kernel, $w^{(b)}(r)$. The contributions depend on the distance between $\mathbf{x}$ and the location of the ith agent, $\mathbf{x}_i$, given by $r = \|\mathbf{x} - \mathbf{x}_i\|$. Again, we choose $w^{(b)}$ to be Gaussian, with width equal to the cell diameter, and repulsive strength, $\gamma_b \geq 0$, so that

$$w^{(b)}(r) = \gamma_b \exp\left(-\frac{r^2}{2\sigma^2}\right), \tag{8}$$

where γ_b is value of $w^{(b)}(0)$, and has dimensions of length. Note that $B(\mathbf{x})$ is an increasing function of local density, and approaches zero as the local density decreases. A typical crowding surface is shown in Figure 4(b) for the arrangement of agents in Figure 4(a).

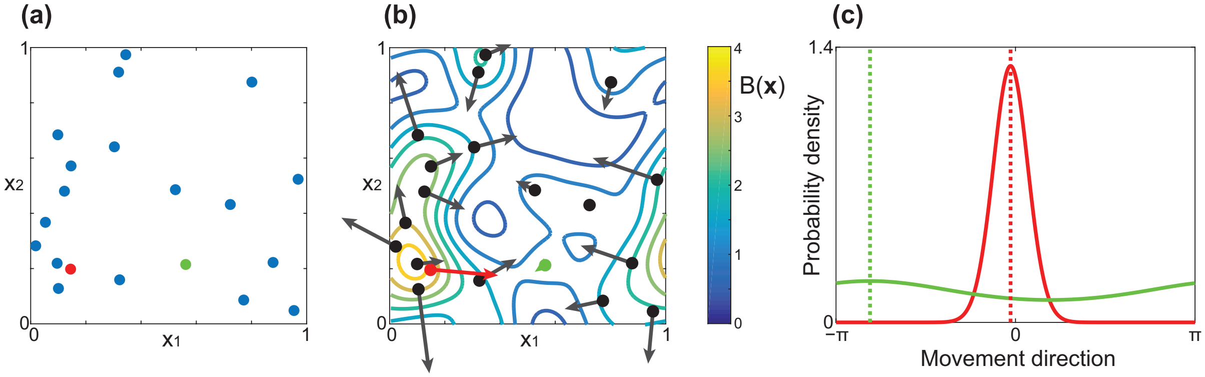

Fig. 4: (a) Example distribution of agents on a 1 × 1 periodic domain. (b) Level curves of the corresponding crowding surface, $B(\mathbf{x})$, for this arrangement of agents. The arrows show the preferred direction of movement, $\mathbf{B}_n$. To illustrate how the direction of movement is chosen, (c) shows the probability density of the von Mises distribution for the red and green agents highlighted in (a) and (b). The preferred direction, $\arg(\mathbf{B}_n)$, is shown as dotted vertical lines for both agents. The red agent is in a crowded region so $\|\mathbf{B}_n\|$ is large, meaning that the agent is likely to move in the preferred direction $\arg(\mathbf{B}_n)$. The green agent is in a low density region and $\|\mathbf{B}_n\|$ is small, meaning that the bias is very weak and the agent's direction of movement is almost uniformly distributed. To illustrate the effects of the crowding surface as clearly as possible, we set $\gamma_b = 1$, $\sigma = 0.1$, $L = 1$ in this schematic figure to draw attention to the gradient of the crowding surface.

To determine the direction of movement we use the shape of $B(\mathbf{x})$ to specify the bias, or preferred direction, of agent n, $\mathbf{B}_n$, given by

$$\mathbf{B}_n = -\nabla B(\mathbf{x}_n), \tag{9}$$

which gives the magnitude and direction of steepest descent. Results in Figure 4(b) show $\mathbf{B}_n$ for the arrangement of agents in Figure 4(a). To determine the direction of movement, we consider the magnitude and direction of $\mathbf{B}_n$, and sample the actual movement direction from a von Mises distribution, von Mises(arg($\mathbf{B}_n$), $\|\mathbf{B}_n\|$) (Binny et al., 2016b; Forbes et al., 2011). Therefore, agents are always most likely to move in the direction of $\mathbf{B}_n$, however as $\|\mathbf{B}_n\| \to 0$, the preferred direction becomes uniformly distributed.

To illustrate how the direction of movement is chosen, we show, in Figure 4(b), the bias vector for each agent, $\mathbf{B}_n$. Note that $\mathbf{B}_n$ does not specify the movement step length, and the direction of $\mathbf{B}_n$ does not necessarily specify the actual direction. Rather, arg($\mathbf{B}_n$) specifies the preferred direction. To illustrate this property, we highlight two agents in Figure 4(a). The red agent is located on a relatively steep part of the crowding surface, so $\|\mathbf{B}_n\|$ is large. The green agent is located on a relatively flat part of the crowding surface, so $\|\mathbf{B}_n\|$ is close to zero. Figure 4(c) shows the von Mises distributions for the red and green agent. Comparing these movement distributions confirms that the crowded red agent is more likely to move in the direction of $\mathbf{B}_n$. The bias is weak for the green agent, so the direction of movement is almost uniformly distributed since $\|\mathbf{B}_n\|$ is smaller.

IBM simulations are performed using the Gillespie algorithm (Gillespie, 1977). To initialise each simulation we specify the initial number and initial location of agents to match to the experimental images at $t = 0$ hours (Supplementary material 1) for experimental data sets 1, 2, 3 and 4. In all simulations we set $\sigma = 24$ μm and $L = 1440$ μm. The remaining three parameters, m, p and γ_b, are varied with the aim of producing posterior distributions using a Bayesian framework.

If $\gamma_m = \gamma_b = 0$, and the variance of the dispersal distribution is large, the IBM corresponds to logistic growth (Binny et al.,2016b, Browning at al. 2017). Under these simplified conditions, a uniformly distributed initial population of agents will grow, at rate p, to eventually reach a uniformly distributed maximum average density of $p/(2\pi\gamma_p\sigma_p^2)$. We do not consider this case here as our initial distribution of cells in the experiments is clustered, and so the logistic growth model is, strictly speaking, not valid (Binny et al.,2016b).

2.2.2. Summary statistics

To match the IBM simulations with the experimental data we use properties that are related to the first two spatial moments (Law et al., 2003). The first spatial moment, the average density, is characterised by the number of agents in the population, $N(t)$. The second spatial moment characterises how agents are spatially distributed, and is often reported in terms of a pair correlation function (Binny et al., 2016a,b; Law et al., 2003). In this work we consider the pair correlation within a distance of δr, given by

$$\mathcal{P}(t) = \frac{L^2 \sum\limits_{i=1}^{N(t)} \sum\limits_{\substack{j=1 \\ j\neq i}}^{N(t)} \mathbb{I}_{\|\mathbf{x}_i - \mathbf{x}_j\| \leq \delta r}}{N(t)^2 \pi \delta r^2}, \tag{10}$$

where $\mathbb{I}$ is an indicator function so that the double sum in Equation (10) gives twice the number of distinct pairs within a distance δr, which we set to be 50 μm. Therefore, $\mathcal{P}(t)$ is the ratio of the number of pairs of agents, separated by a distance of less than 50 μm, to the expected number of pairs of agents separated by a distance of less than 50 μm, if the agents were randomly distributed. This means that, $\mathcal{P}(t) = 1$ corresponds to randomly placed agents; $\mathcal{P}(t) > 1$ corresponds to a locally clustered distribution; and, $\mathcal{P}(t) < 1$ corresponds to a locally segregated distribution.

2.3. Approximate Bayesian computation

We consider m, p and γ_b as random variables, and the uncertainty in these parameters is updated using observed data (Collis et al., 2017; Tanaka et al., 2006). To keep the description of the inference algorithm succinct, we refer to the unknown parameters as $\mathbf{\Theta} = \langle m, p, \gamma_b \rangle$.

In the absence of any experimental observations, information about $\mathbf{\Theta}$ is characterised by specified prior distributions. The prior distributions are chosen to be uniform on an interval that is wide enough to encompass previous estimates of m and p (Johnston et al., 2015). To characterise the prior for γ_b, we note that this parameter is related to a length scale over which bias interactions are felt. Preliminary results (not shown) use a prior in the interval $0 \leq \gamma_b \leq 20$ μm and suggest that a narrow prior in the interval $0 \leq \gamma_b \leq 10$ μm is appropriate. In summary, our prior distributions are uniform and independent, given by

$$\pi(m) = \mathrm{U}(0, 10)\ /\mathrm{hour}, \tag{11}$$

$$\pi(p) = \mathrm{U}(0, 0.1)\ /\mathrm{hour}, \tag{12}$$

$$\pi(\gamma_b) = \mathrm{U}(0, 10)\ \mu\mathrm{m}. \tag{13}$$

We always summarise data, $\mathbf{X}$, with a lower-dimensional summary statistic, S. Data and summary statistics from the experimental images are denoted $\mathbf{X}_{\mathrm{obs}}$ and S_{obs}, respectively. Similarly, data and summary statistics from IBM simulations are denoted $\mathbf{X}_{\mathrm{sim}}$ and S_{sim}, respectively. Information from the prior is updated by the likelihood of the observations, $\pi(S_{\mathrm{obs}}|\mathbf{\Theta})$, to produce posterior distributions, $\pi(\mathbf{\Theta}|S_{\mathrm{obs}})$. We employ the most fundamental ABC algorithm, known as ABC rejection (Liepe et al., 2014; Tanaka et al., 2006), to sample from the approximate posterior distribution. The approximate posterior distributions are denoted $\pi_u(\mathbf{\Theta}|S_{\mathrm{obs}})$.

In this work we use a summary statistic that is a combination of $N(t)$ and $\mathcal{P}(t)$ at equally spaced intervals of duration 12 hours. A discrepancy measure, $\rho(S_{\mathrm{obs}}, S_{\mathrm{sim}})$, is used to assess the closeness of S_{obs} and S_{sim},

$$\rho(S_{\mathrm{obs}}, S_{\mathrm{sim}}) = \sum_{j=1}^{3} \left(\frac{[N_{\mathrm{sim}}(12j) - N_{\mathrm{obs}}(12j)]^2}{N_{\mathrm{obs}}(12j)^2} + \frac{[\mathcal{P}_{\mathrm{sim}}(12j) - \mathcal{P}_{\mathrm{obs}}(12j)]^2}{\mathcal{P}_{\mathrm{obs}}(12j)^2} \right). \tag{14}$$

Algorithm 1 is used to obtain 10^6u samples, $\{\mathbf{\Theta}_i\}_{i=1}^{10^6u}$, from the approximate joint posterior distribution, $\pi_u(\mathbf{\Theta}|S_{\mathrm{obs}})$, for each data set. Here, $u \ll 1$ is the accepted proportion of samples.

Algorithm 1 ABC rejection sampling algorithm to obtain 10^6u samples from the approximate posterior distribution, $\pi_u(\mathbf{\Theta}|S_{\mathrm{obs}})$.

```
1: Set $\sigma = 24$ $\mu$m, $L = 1440$ $\mu$m, and set $\mathbf{x}_n$ to match experimental data $\mathbf{X}_{\mathrm{obs}}$ at $t = 0$.
2: Draw parameter samples from the prior $\mathbf{\Theta}_i \sim \pi(\mathbf{\Theta})$.
3: Simulate cell proliferation assay with $\mathbf{\Theta}_i$ and $t \leq 36$ hours.
4: Record summary statistic $S_{\mathrm{sim}_i} = \{N_{\mathrm{sim}}(12j), \mathcal{P}(12j)\}_{j=1}^{3}$, where $j$ is an index that denotes the
   three observation time points, $t = 12, 24$ and $36$ hours.
5: Compute the discrepancy measure $\epsilon_i = \rho(S_{\mathrm{obs}}, S_{\mathrm{sim}_i})$, given in Equation 14.
6: Repeat steps 2-5 until $10^6$ samples $\{\mathbf{\Theta}_i, \epsilon_i\}_{i=1}^{10^6}$ are simulated.
7: Order $\{\mathbf{\Theta}_i, \epsilon_i\}_{i=1}^{10^6}$ by $\epsilon_i$ such that $\epsilon_1 < \epsilon_2 < \ldots$.
8: Retain the first 1% ($u = 0.01$) of prior samples $\mathbf{\Theta}_i$, as posterior samples, $\{\mathbf{\Theta}_i\}_{i=1}^{10^6u}$.
```

To present and perform calculations with posterior samples, we use a kernel density estimate to form approximate marginal posterior distributions, for each parameter, and each data set using the `ksdensity` function in MATLAB (Math- works, 2017). This is done by treating the components of the joint posterior samples as samples from each marginal distribution. The `ksdensity` function gives a discrete distribution for each marginal posterior, with grid spacing $\Delta m = 0.01$, $\Delta p = 0.0001$ and $\Delta \gamma_b = 0.01$, for m, p and γ_b, respectively. This discretisation ensures that the marginal posterior densities are approximated using 1000 equally spaced values across the prior support.

2.3.1. Generating and sampling from the combined posterior distribution

The marginal posterior distributions for each parameter are similar for each independent experimental data set. Therefore, we combine the marginal posterior distributions for each independent experimental data set to obtain a combined posterior distribution. If the approximate marginal posterior distribution for m is $\pi_u(m|S_{\text{obs}}^{(k)})$, where $S_{\text{obs}}^{(k)}$ is the summary statistic from the kth experimental data set, then the combined marginal posterior distribution for m is

$$\pi_u(m|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3}) \propto \prod_{k=1}^{3} \pi_u(m|S_{\text{obs}}^{(k)}). \tag{15}$$

Combined marginal posterior distributions for p and γ_b are calculated similarly.

To test the predictive power of the calibrated IBM, we sample parameters from the combined joint posterior distribution by sampling each parameter separately from the corresponding combined marginal posterior distributions. This approach amounts to assuming that m, p and γ_b are independent random variables, and we will make a comment on the validity of this assumption later. For m, we generate a discrete combined posterior distribution, $\pi_u(m|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$, using the kernel-density estimate for each data set and Equation (15). This gives a discrete distribution with bin width $\Delta m = 0.01$, where each bin is denoted by an index, $l = 0, 1, ...$, and has probability density $\pi_u(l\Delta m|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$. If m is uniformly distributed within each bin, we apply Algorithm 2 to obtain 10^4 samples. Repeating this process in a similar way to gives 10^4 samples for both p and γ_b.

Algorithm 2 Rejection sampling algorithm for sampling from the combined approximate posterior distribution, $\pi_u(m|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$.

```
1: Set $\Delta m = 0.01$, $m_{\max} = 10$, which is the upper limit of the prior support.
2: Set maximum density $\nu = \max \pi_u(m|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$.
3: Sample proposal bin index $l_*$ from $\{0, ..., m_{\max}/\Delta m - 1\}$.
4: Sample $r_1 \sim \mathrm{U}(0, \nu)$.
5: If $r_1 < \pi_u(l_* \Delta m; \{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$, accept $l_*$, else repeat steps 3-5.
6: Sample the location within the chosen bin, $m_i \sim \mathrm{U}(l_* \Delta m, (l_* + 1)\Delta m)$.
7: Repeat steps 3-6 until $10^4$ samples, $\{m_i\}_{i=1}^{10^4}$, are obtained.
```

2.3.2. Predicting experimental data set 4 using the combined posterior distribution

We sample 10^4 parameter sets, $\{\mathbf{\Theta}_i\}_{i=1}^{10^4}$, from the combined posterior distribution, $\pi_u(\mathbf{\Theta}|\{S_{\text{obs}}^{(k)}\}_{k=1}^{3})$. Using these samples, we simulate the IBM initialised with the actual initial arrangement of cells in data set 4 at $t = 0$. For each parameter combination S_{sim} is recorded at 12 hour intervals, and used to construct distributions of $N(t)$ and $\mathcal{P}(t)$. These distributions are represented as box plots and compared with summary statistics from experimental data set 4.

3. Results and discussion

To qualitatively illustrate the importance of spatial structure we show, in rows 2-4 of Figure 1, snapshots from the IBM with different choices of parameters. In each case the IBM simulations evolve from the initial condition specified in Figure 1(a). Results in the right-most column of Figure 1 compare the evolution of $N(t)$ and we see that the parameter combination in the second row underestimates $N(t)$, the parameter combination in the fourth row overestimates $N(t)$, and the parameter combination in the third row produces a reasonable match to the experimental data. A visual comparison of the spatial arrangement of agents in rows 2-4 of Figure 1 suggests that these different parameter combinations may lead to different spatial structures. This illustration of how the IBM results vary with the choice of parameters motivates us to use ABC rejection to estimate the joint distribution of the parameters. To do this we will use summary statistics from three identically prepared, independent sets of experiments. The summary statistics for these experiments, $N(t)$ and $\mathcal{P}(t)$, are summarised in Figure 2, and tabulated in Supplementary material 1.

The approximate marginal posterior distributions for m, p and γ_b are shown in Figure 5(a)-(c), respectively, for experimental data sets 1, 2 and 3. There are several points of interest to note. In each case, the posterior support is well within the interior of the prior support, suggesting that our choice of priors is appropriate. An interesting feature of the marginal posterior distributions for all parameters is that there is significant overlap for each independent experimental data set. There is some variation in the mode between experimental data sets, for each parameter, which is expected under the assumption that cell proliferation assays are stochastic processes.

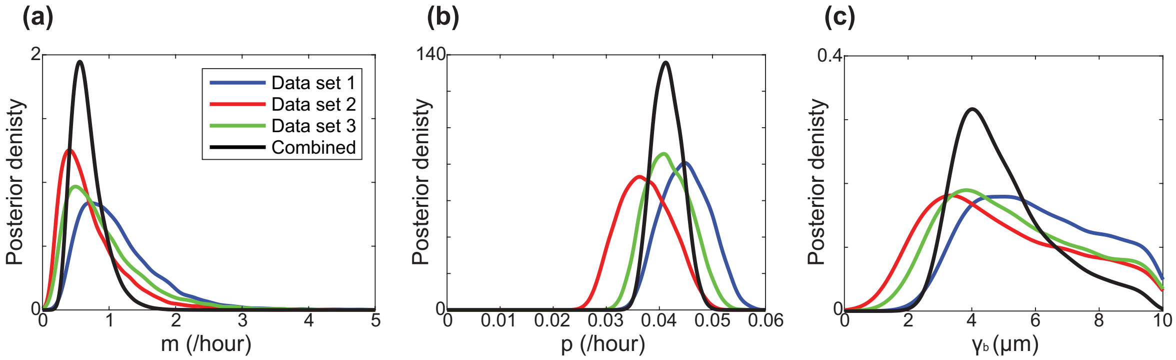

Fig. 5: (a)-(c) Kernel-density estimates of the approximate marginal posterior distributions for each data set, for parameters m, p and γ_b, respectively, with $u = 0.01$. The combined posterior distribution (black), given by Equation (15), is superimposed. The modes of the combined marginal posterior distributions are $m = 0.56$ /hour, $p = 0.041$ /hour and $\gamma_b = 4.0$ μm. All distributions are scaled so that the area under the curve is unity.

Since the marginal posterior distributions for each experimental data set overlap, we produce a combined marginal posterior distribution for each parameter using Equation (15). The combined marginal posterior distributions are superimposed, and the mode is given by 0.56 /hour, 0.041 /hour and 4.0 μm for m, p and γ_b, respectively. These estimates of p and m give a cell doubling time of $\ln(2)/p \approx 17$ hours, and a cell diffusivity of approximately 320 μm^2/hour, which are typical values for PC-3 cells at low density [18, 15]. All results in the main document correspond to retaining the top 1% of samples ($u = 0.01$) and additional results (Supplementary material 1) confirm that the results are relatively insensitive to this choice.

To assess the predictive power of the calibrated IBM, we attempt to predict the time evolution of a separate, independently collected data set, experimental data set 4, as shown in Figure 6(a)-(d). We use the mode of the combined posterior distribution and the initial arrangement of agents in experimental data set 4 to produce a typical prediction in Figure 6(e)-(h). Visual comparison of the experimental data and the IBM prediction suggests that the IBM predicts a similar number of agents, and a similar spatial structure, with some clustering present. To quantify our results, we compare the evolution of $N(t)$ in Figure 6(i) which reveals an excellent match. Furthermore, we predict the evolution of $\mathcal{P}(t)$ in Figure 6(j) confirming similar trends. The quality of match between the predicted distribution of $N(t)$ and $\mathcal{P}(t)$ supports our assumption that m, p and γ_b can be treated as independent random variables as posited in Section 2.3. Although the predicted decay in $\mathcal{P}(t)$ is not as rapid as in the experimental data. There are many potential explanations for this, including the choice of summary statistics, and assumption relating p and γ_p, and m and γ_m.

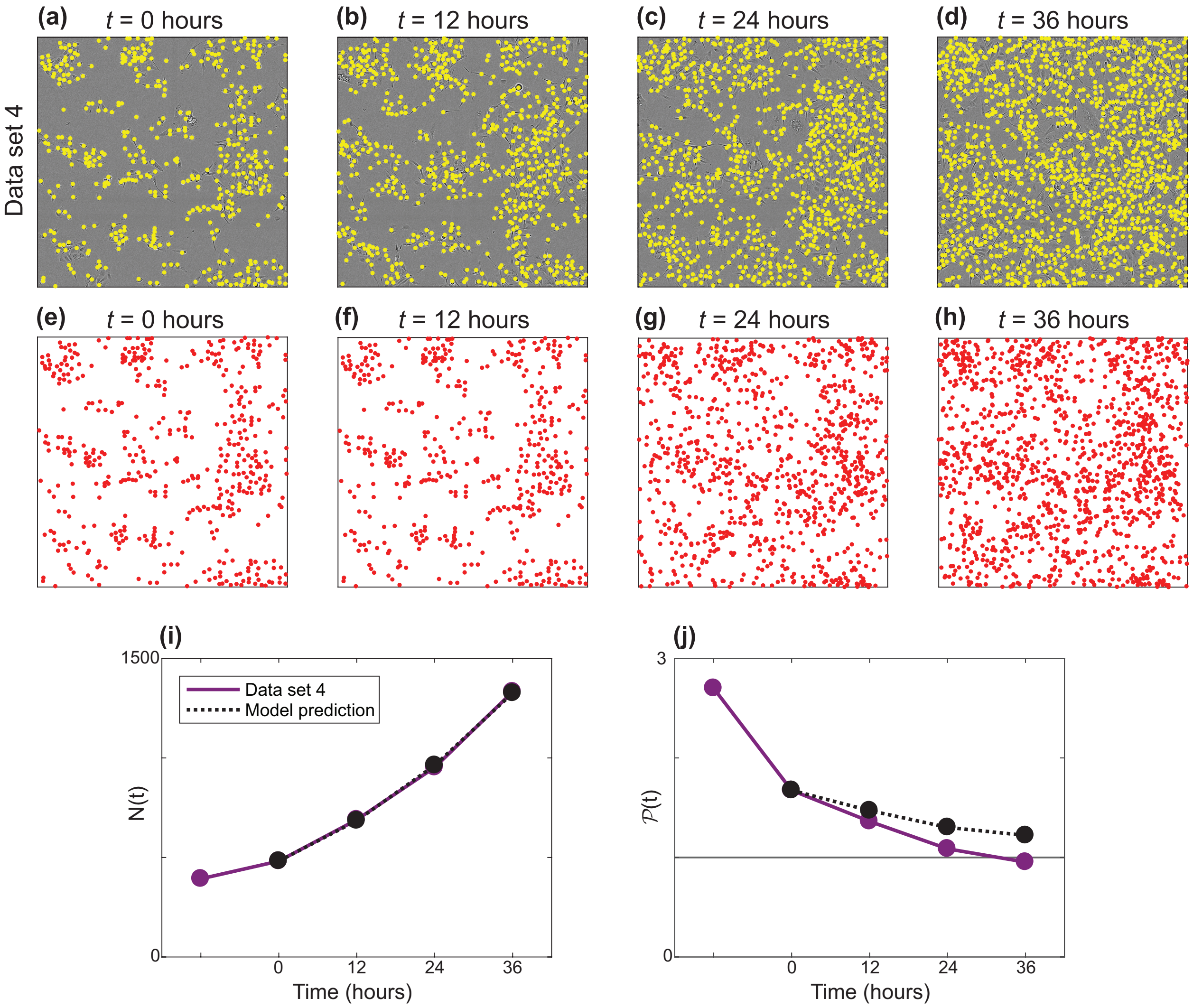

Fig. 6: (a)-(d) Experimental images for data set 4. The position of each cell is identified with a yellow marker. The field of view is a square of length 1440 μm. (e)-(h) One realisation of the IBM with parameters corresponding to the posterior mode: $m = 0.56$ /hour, $p = 0.041$ /hour and $\gamma_b = 4.0$ μm, with the same initial arrangement of agents as in (a). (i) $N(t)$ for the experimental data (purple) and the IBM prediction (dashed black). (j) $\mathcal{P}(t)$ for the experimental data (purple) and the IBM prediction (dashed black).

In addition to examining a single, typical realisation of the calibrated model, we now examine a suite of realisations of the calibrated IBM, and compare results with experimental data set 4. The suite of IBM realisations is obtained by sampling from the joint posterior distribution. Results in Figure 7(a) compare $N(t)$ from experimental data set 4 with distributions of $N(t)$ from the suite of IBM simulations, showing an excellent match. The spread of the distributions of $N(t)$ increases with time, which is expected. Results in Figure 7(b) compare the evolution of $\mathcal{P}(t)$ from experimental data set 4 with distributions of $\mathcal{P}(t)$ from the suite of IBM simulations, showing the predicted distributions of $\mathcal{P}(t)$ overlap with the experimental data. Overall, the quality of the match between the prediction and the experimental data is high, as the prediction captures both qualitative and quantitative features of the data.

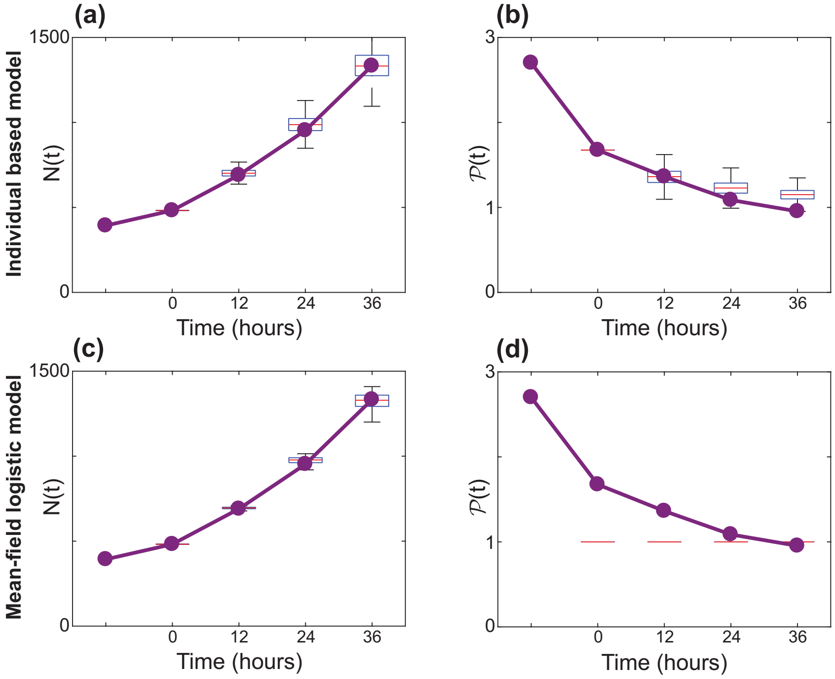

Fig. 7: (a)-(b) Predictive distributions for $N(t)$ and $\mathcal{P}(t)$, respectively, generated using the IBM. 10^4 parameter samples were taken from the combined posterior distribution, and a model realisation produced for each sample, initiated as in Figure 6(a). Box plots show the distribution of $N(t)$ and $\mathcal{P}(t)$ across these realisations in (a) and (b), respectively. (c)-(d) Show the equivalent predictive distributions as box plots, using the same procedure for the mean-field logistic growth model. The procedure and kernel-density estimates of the marginal distributions for the mean-field logistic model are outlined in Supplementary material 1.

To illustrate the importance of considering spatial structure in the IBM, we also calibrate the solution of the classical mean-field logistic equation (Murray, 2002) to experimental data sets 1, 2 and 3. The logistic equation is given by

$$\frac{\mathrm{d}N(t)}{\mathrm{d}t} = \lambda N(t)\left(1 - \frac{N(t)}{N_{\mathrm{max}}}\right), \qquad (16)$$

where λ is the cell proliferation rate and N_{max} is the maximum number of agents (Murray, 2002; Jin et al., 2017). Following a similar procedure (Supplementary material 1), we use ABC rejection to form combined posterior distributions of λ and N_{max}. The modes of the combined posterior distributions are $\lambda = 0.036$ /hour and $N_{\mathrm{max}} = 4017$. This estimate leads to a doubling time of approximately 19 hours, which is slightly longer than the doubling time predicted using the calibrated IBM. We then examine a suite of solutions to Equation (16), where we sample from the joint posterior distribution for λ and N_{max}. The predicted distribution of $N(t)$ is compared with experimental data set 4 in Figure 7(c), revealing an excellent match. However, implicit in the logistic equation is the mean-field assumption, which amounts to ignoring spatial structure. Therefore, the logistic equation effectively predicts $\mathcal{P}(t) = 1$ for all $t > 0$, which clearly is unable to match the spatial structure inherent in the experiments, as demonstrated in Figure 7(d). Overall, both calibrated models are able to predict the evolution of $N(t)$ over 36 hours. However, the logistic model is unable to describe, or predict, any information relating to spatial structure in the arrangement of cells. The differences in the way that the logistic model and the IBM treat interactions between individuals could explain why the calibration process leads to different estimates of the low density cell proliferation rates, λ and p. These differences affirm that the interactions between individuals at different spatial scales appear to be important for our experimental data.

4. Conclusions

In this work we explore how to connect a spatially continuous IBM of cell migration and cell proliferation to novel data from a cell proliferation assay. Previous work parameterising IBM models of cell migration and cell proliferation to experimental data using ABC have been restricted to lattice-based IBMs (Johnston et al., 2014). This is partly because ABC methods require large numbers of IBM simulations, and lattice-based IBMs are far less computationally expensive than lattice-free IBMs (Plank and Simpson, 2012). We find it is preferable to work with a lattice-free IBM when dealing with experimental data as a lattice-based IBM requires approximations when mapping the distribution of cells from experimental images to a lattice (Johnston et al., 2014; Johnston et al., 2016). This mapping can be problematic. For example, if multiple cells in an experimental image are equally close to one lattice site, ad hoc assumptions have to be introduced about how to arrange those cells on the lattice without any overlap. These issues are circumvented using a lattice-free method.

To help overcome the computational cost of using ABC with a lattice-free IBM, we introduce several realistic, simplifying assumptions. The IBM originally presented by Binny et al. (2016b) involves 12 free parameters, which is a relatively large number for standard inference techniques. The model is simplified by noting that our experiments do not involve cell death, and specifying the width of the interaction kernels to be constant, given by the cell diameter. Another simplification is given by assuming that crowding effects reduce the proliferation and movement rates to zero when the agents are packed at the maximum hexagonal packing density. This leads to a simplified model with three free parameters: m, p and γ_b. Using ABC rejection, we arrive at posterior distributions for these parameters for three independent experimental data sets. The marginal posterior distributions for the three parameters are similar, leading us to combine the marginal posterior distributions. The mode of the combined posterior distributions for m and p are consistent with previous parameter estimates (Johnston et al., 2015) and the mode for γ_b is consistent with previous observations that mesenchymal cells in this kind of two-dimensional experiment tend to move away from regions of high cell density (Cai et al., 2007).

In the field of mathematical biology, questions about how much detail to include in a mathematical model, and what kind of mathematical model is preferable for understanding a particular biological process are often settled in an *ad hoc* manner, as discussed by Maclaren et al. (2015). Our approach in this work is to use a mathematical model that incorporates just the key mechanisms, with an appropriate number of unknown parameters. Other approaches are possible, such as using much more complicated mathematical models that describe additional mechanisms such as: (i) detailed information about the cell cycle in individual cells (Fletcher et al., 2012); (ii) concepts of leader and follower cells (Kabla, 2012); (iii) explicitly coupling cell migration and cell proliferation to the availability of nutrients and growth factors (Tang at al., 2014); or (iv) including mechanical forces between cells (Stichel at al., 2017). However, we do not include these kinds of detailed mechanisms because our experimental data does not suggest that these mechanisms are relevant to our situation. Furthermore, it is not always clear that using a more complicated mathematical model, with additional mechanisms and additional unknown parameters, necessarily leads to improved biological insight. In fact, simply incorporating additional mechanisms and parameters into the mathematical model often leads to a situation where multiple parameter combinations lead to equivalent predictions which limits the usefulness of the mathematical model (Simpson et al., 2006). In this study, our approach is to be guided by experimental data and our ability to infer the parameters in a mathematical model based on realistic amounts of experimental data (Maclaren et al. 2015). In particular we use three experimental data sets to calibrate the IBM, and an additional data set to separately examine the predictive capability of the calibrated IBM. We find that the process of calibrating the IBM leads to well defined posterior distributions of the model parameters, and that the calibrated IBM produces a reasonable match to the experimental data. The process of calibrating the IBM, and then separately testing the predictive capability of the calibrated IBM, provides some confidence that the level of model complexity is appropriate for our purposes.

An interesting feature of our approach is that the ABC marginal posterior distributions for each parameter overlap for each independent experimental data set. This is reassuring as it suggests that the same IBM mechanism matches the three independent experimental data sets using similar parameters. Another approach would be to use ABC to parameterise the IBM by matching all the experimental data sets simultaneously. Although this alternative approach is valid, it does not allow us to examine whether the parameter estimates are consistent across the three independent experiments. Additional confidence in the calibrated IBM is provided by predicting the evolution of a fourth independent experimental data set by performing IBM simulations with parameters sampled from the combined marginal posterior distributions.

An interesting feature of all experimental data at early time, when the cell density is relatively low, is that the pair correlation measure suggests that the cells are clustered at short intervals, and that this clustering becomes less pronounced with time. This observation is very different to the way that previous theoretical studies have viewed the role of spatial structure. For example, previous simulation-based studies assume that some initial random spatial arrangement of cells can lead to clustering at later times (Baker and Simpson, 2010). In contrast, our experimental data suggests it could be more realistic to consider that the spatial structure is imposed by the initial arrangement of cells. Moreover, since all of our experimental data involves some degree of spatial clustering, our work highlights the importance of using appropriate models to provide a realistic representation of key phenomena. Almost all continuum models of collective behaviour in cell populations take the form of ordinary differential equations and partial differential equations that implicitly invoke a mean-field assumption (Tremel et al., 2009; Sengers et al., 2007; Maini et al., 2004b; Sarapata and de Pillis, 2014; Sherratt and Murray, 1990). Such assumptions ignore the role of spatial structure. While pair-wise models that avoid mean-field assumptions are routine in some fields, such as disease spreading (Sharkey et al., 2006; Sharkey, 2008) and ecology (Law et al., 2003), models that explicitly account for spatial structure are far less common for collective cell behaviour.

Using our parameter estimates, the continuum spatial moment description could be used to interpret experimental data sets with larger numbers of cells (Binny et al., 2016b), such as experimental images showing a wider field-of-view, or experiments initiated with a higher density of cells. Our approach to estimate the parameters in the model is to work with the IBM since this allows us more flexibility in connecting with the experimental data, such as choosing the initial locations of the agents in the IBM to precisely match the initial locations of cells in the experimental images.

There are many ways that our study could be extended. For example, here we choose a summary statistic encoding information about the first two spatial moments. However, other summary statistics may provide different insight, and it could be of interest to explore the effect of this choice. For example, here we describe the spatial structure over a relatively short spatial interval, approximately 2σ. It could be of interest to repeat our analysis with a wider interval, however this would incur additional computational costs. Another approach to extend our work would be to repeat the inference procedure without making any assumptions relating p and γ_p, and m and γ_m. Such an approach would be more computationally expensive and probably require additional experimental data. Therefore, we leave these topics for future consideration.

5. Acknowledgements

This work is supported by the Australian Research Council (DP140100249, DP170100474) and the Royal Society of New Zealand Marsden Fund (11-UOC-005). Computational resources provided by the High Performance Computing and Research Support Group at QUT are appreciated. We also thank David Warne for technical advice.

6. References

[1] Baker RE, Simpson MJ. 2010. Correcting mean-field approximations for birth-death-movement processes. *Phys Rev E* **82**, 041905.

[2] Binny RN, Haridas P, James A, Law R, Simpson MJ, Plank MJ. 2016a. Spatial structure arising from neighbour-dependent bias in collective cell movement. *PeerJ* **4**, e1689.

[3] Binny RN, James A, Plank MJ. 2016b. Collective cell behaviour with neighbour-dependent proliferation, death and directional bias. *Bull Math Biol* **78**, 2277–2301.

[4] Bosco DB, Kenworthy R, Zorio DAR, Sang QXA. 2015. Human mesenchymal stem cells are resistant to paclitaxel by adopting a non-proliferative fibroblastic state. *PLoS One* **10**, e0128511.

[5] Bourseguin J, Bonet C, Renaud E, Pandiani C, Boncompagni M, Giuliano S, Pawlikowska P, Karmous-Benailly H, Ballotti R, Rosselli F, Bertolotto C. 2016. FANCD2 functions as a critical factor downstream of MiTF to maintain the proliferation and survival of melanoma cells. *Sci Rep* **6**, 36539.

[6] Browning AP, McCue SW, Simpson MJ. 2017. A Bayesian computational approach to explore the optimal duration of a cell proliferation assay. *Bull Math Biol* **10** 1888–1906.

[7] Cai AQ, Landman KA, Hughes BD. 2007. Multi-scale modeling of a wound-healing cell migration assay. *J Theor Biol* **245**, 576–594.

[8] Codling EA, Plank MJ, Benhamou S. 2008. Random walk models in biology. *J R Soc Interface* **5**, 813–834.

[9] Collis J, Connor AJ, Paczkowski M, Kannan P, Pitt-Francis J, Byrne HM, Hubbard ME. 2017. Bayesian calibration, validation and uncertainty quantification for predictive modelling of tumour growth: a tutorial. *Bull Math Biol* **79**, 939–974.

[10] Fletcher AG, Breward CJW, Chapman SJ. 2012. Mathematical modelling of monoclonal conversion in the colonic crypt. *J Theor Biol* **300**, 118–133.

[11] Forbes C, Evans M, Hastings N, Peacock B. 2011. *Statistical distributions*. 4th ed. John Wiley & Sons, New Jersey.

[12] Frascoli F, Hughes BD, Zaman MH, Landman KA. 2013. A computational model for collective cellular cotion in three dimensions: general framework and case study for cell pair dynamics. *PLoS ONE* **8**, e59249.

[13] Gillespie DT. 1977. Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* **81**, 2340–2361.

[14] Kabla AJ. 2012. Collective cell migration: leadership, invasion and segregation. *J R Soc Interface* **9** 20120448.

[15] Jin W, Shah ET, Penington CJ and McCue SW, Chopin LK, Simpson MJ. 2016. Reproducibility of scratch assays is affected by the initial degree of confluence: Experiments, modelling and model selection. *J Theor Biol* **390**, 136–145.

[16] Jin W, Shah ET, Penington CJ, McCue SW, Maini PK, Simpson MJ. 2017. Logistic proliferation of cells in scratch assays is delayed. *Bull Math Biol* **79**, 1028–1050.

[17] Johnston ST, Simpson MJ, McElwain DLS, Binder BJ, Ross JV. 2014. Interpreting scratch assays using pair density dynamics and approximate Bayesian computation. *Open Biol* **4**, 140097.

[18] Johnston ST, Shah ET, Chopin LK, McElwain DLS, Simpson MJ. 2015. Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte ZOOM^{TM} assay data using the Fisher-Kolmogorov model. *BMC Syst Biol* **9**, 38.

[19] Johnston ST, Ross JV, Binder BJ, McElwain DLS, Haridas P, Simpson MJ. 2016. Quantifying the effect of experimental design choices for in vitro scratch assays. *J Theor Biol* **400**, 19–31.

[20] Kaighn ME, Narayan KS, Ohnuki Y, Lechner JF, Jones LW. 1979. Establishment and characterization of a human prostatic carcinoma cell line (PC-3). *Invest Urol* **17**, 16–23.

[21] Law R, Murrell DJ, Dieckmann U. 2003. Population growth in space and time: Spatial logistic equations. *Ecology* **84**, 252–262.

[22] Liepe J, Kirk P, Filippi S, Toni T, Barnes CP, Stumpf MPH. 2014. A framework for parameter estimation and model selection from experimental data in systems biology using approximate Bayesian computation. *Nat Protoc* **9**, 439–456.

[23] Maclaren OJ, Byrne HM, Fletcher AG, Maini PK. 2015. Models, measurement and inference in epithelial tissue dynamics. *Math Biosci Eng* **12**, 1321 – 1340.

[24] Maini PK, McElwain DLS, Leavesley DI. 2004. Traveling wave model to interpret a wound-healing cell migration assay for human peritoneal mesothelial cells. *Tissue Eng* **10**, 475–482.

[25] Mathworks. 2017. Kernel smoothing function estimate for univariate and bivariate data. http://www.mathworks.com/help/stats/ksdensity.html. Accessed: June 2017.

[26] Murray JD. 2002. *Mathematical Biology.* Springer, Berlin.

[27] Peirce SM, Van Gieson EJ, Skalak TC. 2004. Multicellular simulation predicts microvascular patterning and in silico tissue assembly. *FASEB J* **18**, 731–733.

[28] Plank MJ, Simpson MJ. 2012. Models of collective cell behaviour with crowding effects: comparing lattice-based and lattice-free approaches. *J R Soc Interface* **9**, 2983-2996.

[29] Read M, Andrews PS, Timmis J, Kumar V. 2012. Techniques for grounding agent-based simulations in the real domain: a case study in experimental autoimmune encephalomyelitis. *Math Comp Model Dyn* **18**, 67–86.

[30] Sarapata EA, de Pillis LG. 2014. A comparison and catalog of intrinsic tumor growth models. *Bull Math Biol* **76**, 2010–2024.

[31] Sengers BG, Please CP, Oreffo ROC. 2007. Experimental characterization and computational modelling of two-dimensional cell spreading for skeletal regeneration. *J R Soc Interface* **4**, 1107.

[32] Sharkey KJ, Fernandez C, Morgan KL, Peeler E, Thrush M, Turnbull JF, Bowers RG. 2006. Pair-level approximations to the spatio-temporal dynamics of epidemics on asymmetric contact networks. *J Math Biol* **53**, 61–85.

[33] Sharkey KJ. 2008. Deterministic epidemiological models at the individual level. *J Math Biol* **57**, 311–331.

[34] Sherratt JA, Murray JD. 1990. Models of epidermal wound healing. *P Roy Soc Lond B* **241**, 29.

[35] Simpson MJ, Landman KA, Hughes BD, Newgreen DF. 2006. Looking inside an invasion wave of cells using continuum models: Proliferation is the key. *J Theor Biol* **243**, 343–360.

[36] Stichel D, Middleton AM, Müller BF, Depner S, Klingmüller U, Breuhahn K, Matthäus F. 2017. An individual-based model for collective cancer cell migration explains speed dynamics and phenotype variability in response to growth factors. *NPJ Syst Biol Appl* **3**, 5.

[37] Tanaka MM, Francis AR, Luciani F, Sisson SA. 2006. Using approximate Bayesian com- putation to estimate tuberculosis transmission parameters from genotype data. *Genetics* **173**, 1511–1520.

[38] Tang L, van de Ven AL, Guo D, Andasari V, Cristini V, Li KC, Zhou X. 2014. Computational modeling of 3D tumor growth and angiogenesis for chemotherapy evaluation. *PLoS One* **9**, e83962.

[39] Treloar KK, Simpson MJ, Haridas P, Manton KJ, Leavesley DI, McElwain DLS, Baker RE. 2013. Multiple types of data are required to identify the mechanisms influencing the spatial expansion of melanoma cell colonies. *BMC Syst Biol* **7**, 137.

[40] Tremel A, Cai A, Tirtaatmadja N, Hughes BD, Stevens GW, Landman KA, O'Connor AJ. 2009. Cell migration and proliferation during monolayer formation and wound healing. *Chem Eng Sci* **64**, 247–253.