



Cite this article: Parker A, Simpson MJ, Baker RE. 2018 The impact of experimental design choices on parameter inference for models of growing cell colonies. *R. Soc. open sci.* **5**: 180384. <http://dx.doi.org/10.1098/rsos.180384>

Received: 9 March 2018

Accepted: 18 July 2018

Subject Category:

Cellular and molecular biology

Subject Areas:

applied mathematics/cellular biology/
computational mathematics

Keywords:

approximate Bayesian computation,
cell spreading, experimental design,
cell migration, cell proliferation

Author for correspondence:

Ruth E. Baker

e-mail: ruth.baker@maths.ox.ac.uk

Electronic supplementary material is available online at <https://dx.doi.org/10.6084/m9.figshare.c.4200446>.

The impact of experimental design choices on parameter inference for models of growing cell colonies

Andrew Parker¹, Matthew J. Simpson²
and Ruth E. Baker¹

¹Mathematical Institute, University of Oxford, Oxford, UK

²School of Mathematical Sciences, Queensland University of Technology, Brisbane, Australia

 MJS, 0000-0001-6254-313X; REB, 0000-0002-6304-9333

To better understand development, repair and disease progression, it is useful to quantify the behaviour of proliferative and motile cell populations as they grow and expand to fill their local environment. Inferring parameters associated with mechanistic models of cell colony growth using quantitative data collected from carefully designed experiments provides a natural means to elucidate the relative contributions of various processes to the growth of the colony. In this work, we explore how experimental design impacts our ability to infer parameters for simple models of the growth of proliferative and motile cell populations. We adopt a Bayesian approach, which allows us to characterize the uncertainty associated with estimates of the model parameters. Our results suggest that experimental designs that incorporate initial spatial heterogeneities in cell positions facilitate parameter inference without the requirement of cell tracking, while designs that involve uniform initial placement of cells require cell tracking for accurate parameter inference. As cell tracking is an experimental bottleneck in many studies of this type, our recommendations for experimental design provide for significant potential time and cost savings in the analysis of cell colony growth.

1. Introduction

The study of how cell populations grow and spread is integral to understanding and predicting the invasion of cancer, the speed of wound repair and the robustness of embryonic development [1–3]. However, the extent to which cell populations grow and spread is governed by multiple processes, including motility,

proliferation, adhesion and cell death, making it difficult to elucidate the relative contributions of these processes to the growth and invasion of a cell colony [4]. As such, *in vitro* cell biology assays are routinely used to probe the mechanisms by which cells interact, and the key processes involved in the growth and expansion of cell colonies. These *in vitro* assays generally involve seeding a population of cells on a two-dimensional substrate, and observing the population as the individual cells move and proliferate and the density of the monolayer increases towards confluence. A useful approach to interpret the results of these assays involves using a mathematical model that incorporates mechanistic descriptions of processes such as cell motility and proliferation. By parametrizing and validating the models using quantitative data from *in vitro* assays, it is possible to provide quantitative insights into the mechanisms driving the growth and spreading of a cell population, and make experimentally testable predictions. However, it is not always clear how best to choose the experimental design, nor which summary statistics of the data to collect, in order to accurately and efficiently parametrize and validate models.

In this work, we use a two-dimensional lattice-based exclusion process model that incorporates both motility and proliferation mechanisms. Our goal is to assess how our ability to accurately infer model parameters is affected by changes in the experimental design. Parameter inference is performed in a Bayesian framework using approximate Bayesian computation (ABC), allowing us to quantify the uncertainty of our parameter estimates and bypass the need to compute a likelihood function for the mechanistic model. By quantifying the information gain using the different experimental protocols, we are able to provide guidelines for experimental design in terms of the selection of experimental geometry and the collection of relevant quantitative summary statistics from imaging data.

1.1. Experimental design

Typically, there are two main types of two-dimensional *in vitro* experiments that are considered at the level of the population. The first experiment, shown in figure 1*a*, is often referred to as a *growth-to-confluence assay*. Here, we observe a population of cells seeded, initially at low density, as the cells move and proliferate and the population increases in number to eventually occupy the whole domain under observation [7,8]. The second experiment we consider is shown in figure 1*b* and is often referred to as a *scratch assay*. It involves perturbing a population of cells at, or near, confluence by scraping away a region of the population and observing the resulting spread of the population into this, now empty, region [9,10]. In this work, we will explore the extent to which models of these two types of assays can be parametrized using various summary statistics of these experimental data.

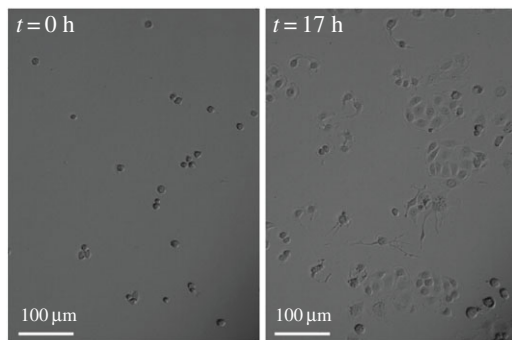
1.2. Approximate Bayesian computation and summary statistics

Parameter inference is approached generally in one of two ways, through either a frequentist approach or a Bayesian approach [11,12]. In frequentist inference, one generally seeks a point estimate of a parameter through maximum-likelihood estimation, and captures uncertainty in the estimate through the generation of confidence intervals. A Bayesian approach instead derives a predictive posterior distribution for the model parameters θ given observed data \mathcal{D}^{obs} [13]. The posterior, $\mathbb{P}(\theta | \mathcal{D}^{\text{obs}})$, satisfies $\mathbb{P}(\theta | \mathcal{D}^{\text{obs}}) \propto \mathcal{L}(\mathcal{D}^{\text{obs}} | \theta) \pi(\theta)$, where $\mathcal{L}(\mathcal{D}^{\text{obs}} | \theta)$ is the likelihood of obtaining the data from the given model and parameters, and the prior, $\pi(\theta)$, captures any previous knowledge of the parameters. For the mechanistic model we use in this work, the likelihood is intractable and so we use ABC to generate an approximate posterior distribution.

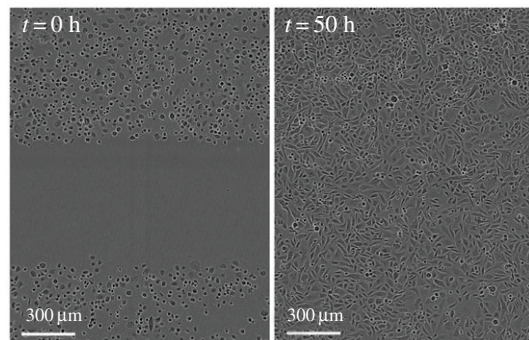
ABC has been used previously in the biological sciences, particularly to target complex problems in systems biology [14,15], population genetics [16] and ecology [17]. There is also growing popularity in the use of ABC specifically in the investigation of cell biology processes [18–20]. ABC relies on repeated simulation of the model using parameters from the prior distribution, and the acceptance of these parameter sets whenever the model output is sufficiently close to the experimental data. These accepted parameter sets are then used to estimate the posterior distribution. In using ABC in practice, there are several important user choices, most importantly the means of comparison between simulated and experimental data, typically through summary statistics and a suitable distance metric, and the threshold used for accepting or rejecting parameter sets [21,22]. In particular, the summary statistics represent lower dimensional descriptions of the data and their choice is vitally important as they impact the amount of information gained about model parameters through the use of ABC [23].

The major new insights provided in this work are a quantitative understanding of how both the choices of experimental geometry and summary statistics impact the quality of the posteriors

(a) growth-to-confluence assay



(b) scratch assay



(c) different experimental set ups considered in the model

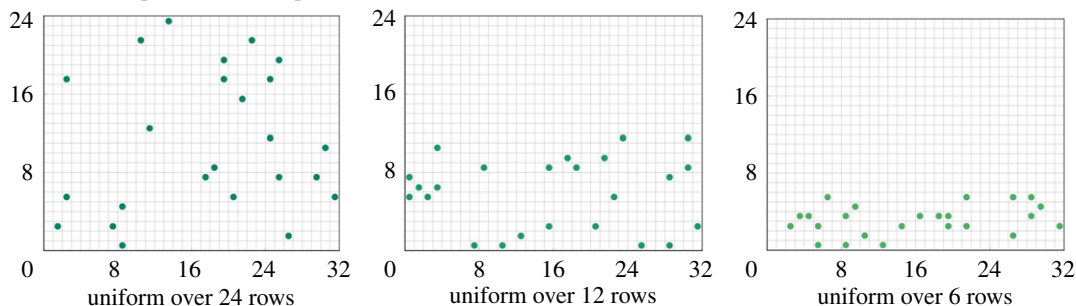


Figure 1. Two experimental designs considered in this work. (a) Images from a growth-to-confluence assay using MDA MB 231 breast cancer cells. See [5] for more information. (b) Images from a scratch assay with PC3 prostate cancer cells. See [6] for more information. (c) Schematics of the model. We use an on-lattice model, as described in the text, and vary the initial condition to replicate the experiments in (a,b): we seed cells uniformly at random over 24 rows (left), 12 rows (middle) or six rows (right). The images in (a) are reproduced from Simpson *et al.* [5] with kind permission, whereas the images in (b) are reproduced from Johnston *et al.* [6] with kind permission.

generated using ABC. The use of ABC rejection allows us to compare the quality of the posteriors resulting from a wide range of possible summary statistics in a computationally efficient way, and we quantify the information gain using the Kullback–Leibler divergence [24]. We also demonstrate how data-cloning ABC (ABC-DC) [25] can potentially be used to obtain maximum-likelihood estimates (MLEs) of model parameters more efficiently than ABC rejection. Instead of targeting the likelihood of the observed data, the data cloning approach targets the likelihood corresponding to a large number of copies (also known as clones) of the data, where each data clone is assumed independent of the others. Data cloning results in a posterior distribution that has the MLE as its mean, and the variance can be related to the asymptotic variance of the MLE [26]. ABC-DC uses ABC Markov chain Monte Carlo with data cloning to facilitate convergence towards a MLE [25].

1.3. Aims and outline

The aim of this work is to demonstrate the use of Bayesian inference methods to characterize the motile and proliferative behaviour of individual cells within growing cell colonies. In particular, we aim to estimate parameters of a lattice-based volume exclusion model for cell colony growth in a variety of experimental geometries. In addition, we demonstrate the use of ABC-DC for more efficient maximum-likelihood estimation. In §2, we introduce our model, and the various ABC algorithms and summary statistics we employ in this work. In §3, we assess how experimental design choices impact the quality of estimated posterior distributions, using both ABC rejection and ABC-DC, and we conclude with a discussion of our results in §4. To confirm the accuracy of our inference process, and the relevance of the mechanistic model to the experimental data, we conclude by sampling from the posterior distributions and exploring whether key summary statistics predicted by the parametrized model, such as the size of the population and the displacement of individual cells, reflect our experimental observations within some reasonable confidence interval.

2. Methods

2.1. Mechanistic model

We employ a simple two-dimensional lattice-based exclusion process model akin to that of Simpson *et al.* [27], whereby $N(t)$ cells occupy a square lattice with R rows and C columns at time t . During each time step of duration τ , we choose $N(t)$ cells at random with replacement to attempt a movement and proliferation event into orthogonally adjacent lattice sites with probabilities P_m and P_p , respectively. For each cell, we draw a uniform random number, $r \sim \mathcal{U}(0, 1)$. If $r \leq P_m$, the cell attempts a movement event into one of the four orthogonally adjacent lattice sites with equal probability, and if $P_m < r \leq P_m + P_p$ the cell attempts a proliferation event, whereby a daughter cell is placed into one of these lattice sites, each with equal probability. If $P_m + P_p < r \leq 1$, no movement or proliferation event is attempted. If a cell attempts to move or to place a daughter cell into an occupied lattice site, or outside of the domain, the attempted movement or proliferation event is aborted. These parameters in the discrete model are related to the classical diffusion coefficient, D , and cell proliferation rate, λ , by $D = \lim_{\Delta \rightarrow 0, \tau \rightarrow 0} P_m \Delta^2 / (4\tau)$ and $\lambda = \lim_{\Delta \rightarrow 0, \tau \rightarrow 0} P_p / \tau$ [28].

To replicate experimental images, we take $R = 24$, $C = 32$, where lattice sites have length $\Delta = 18.75 \mu\text{m}$ (corresponding to the approximate cell diameter of the cells considered in typical experiments). Simulations are initialized with cell positions randomly distributed in the first \hat{R} rows of the domain, where \hat{R} is chosen to mimic potential experimental conditions. To interpolate between the growth-to-confluence and scratch assay designs, we choose three initial conditions (figure 1c) initializing cells uniformly at random across either 24, 12 or six rows of the domain. We note that, in reality, the size of the experimental domain is much larger than that captured in the experimental images [6]. To confirm that our results are insensitive to this choice of domain size, we repeated our investigations with simulations carried out on a larger domains and data collected on a subset of this domain that corresponds to the experimental image size. Results are shown in electronic supplementary material, figure S6.

2.2. *In silico* data

As our aim in this work is to better understand how experimental design impacts our ability to infer model parameters, we use our mechanistic model to generate *in silico* (observed) data that closely replicate that available from experiments (figure 1a,b) and attempt to infer the parameters used to generate these *in silico* data. We use a time step of $\tau = 1/24$ h, and model parameters $P_m = 0.25$ and $P_p = 0.0025$ (motivated by estimates for the diffusivity and proliferation rates from similar experiments [6]). We record the final positions of cells in our *in silico* experiments after $T = 12$ h, equating to 288 time steps. We also record trajectory data for five randomly chosen cells by recording their positions every eight time steps (corresponding to every 20 min [5]). As experiments are typically repeated several times to ensure reproducibility of results [5], we repeat simulations $M = 10$ times and average the resulting statistics. To confirm the consistency of our results for larger values of the final simulation time, T , we repeat our methodology for longer periods of time, either $T = 24$ h or $T = 36$ h. Results are shown in the electronic supplementary material, figure S5.

2.3. Inference

We use ABC to estimate posterior distributions for the model parameters, $\theta = (P_m, P_p)$. ABC rejection is performed by repeatedly sampling from the prior distribution, $\pi(\theta)$, simulating the model, and accepting parameters that result in simulation output sufficiently close to the observed data. The accepted parameters are used to compute the approximate posterior distribution, $\mathbb{P}(\theta | \mathcal{D}^{\text{obs}})$. In order to assess how close the simulated and observed data are, we consider a range of summary statistics (detailed in §2.3.2) and an appropriate distance function (described in §2.3.3).

In order to quantify the performance of different summary statistics in inferring model parameters, we choose a uniform (uninformative) prior,

$$\pi(\theta) = \frac{1}{0.99 \times 0.01}, \quad \theta \in (0, 0.99) \times (0, 0.01), \quad (2.1)$$

to ensure $P_m + P_p \leq 1$.

2.3.1. Approximate Bayesian computation rejection

Algorithm 1 describes the ABC rejection algorithm we use in this work. In algorithm 1, we avoid directly specifying an acceptance threshold, ϵ , by accepting the 1st percentile of the samples (ranked in terms of the distance between the simulated and observed data) and taking the number of samples, K , sufficiently large to obtain an accurate approximation to the true posterior. The accepted samples are used to estimate a posterior distribution using bivariate kernel density estimation with a Gaussian kernel [29].

Algorithm 1. ABC rejection sampler.

```

1: for  $k = 1$  to  $K$  do
2:   sample  $\theta_k$  from  $\pi(\theta)$ 
3:   simulate  $\mathcal{D}^k$  from the model with parameter  $\theta_k$ 
4:   compute summary statistics,  $s_j^k(l)$ , for  $j = 1, \dots, 13$  and  $l = 1, \dots, L_j$  (see §2.3.3)
5:   calculate distance  $d^k$  (equations (2.15) and (2.16))
6: end for
7: calculate the acceptance threshold,  $\epsilon$  (see §2.3.1).
8: for  $k = 1$  to  $K$  do
9:   if  $d^k < \epsilon$  then
10:    accept  $\theta_k$ 
11:   else
12:    reject  $\theta_k$ 
13:   end if
14: end for

```

2.3.2. Summary statistics

In order to compare observed and simulated data, we reduce the dimension of the data using summary statistics. How best to choose summary statistics for parameter inference is an ongoing research question, and some automated procedures have been developed for this choice, typically through either minimizing a loss of information function [30], or maximizing the gain of information [23] through a measure such as the Kullback–Leibler divergence [24].

Kullback–Leibler divergence. The Kullback–Leibler divergence is a measure of the relative difference in two continuous distributions, F and G , and defined as

$$I_{\text{KL}}(F \| G) = \int_{\theta} F(\theta) \ln \left(\frac{F(\theta)}{G(\theta)} \right) d\theta. \quad (2.2)$$

In this work, we generate bivariate posteriors and priors which are each functions of $\theta = (P_m, P_p)$. We discretize these distributions onto a fine mesh, 512×512 in size, for plotting. We compute the Kullback–Leibler divergence, I_{KL} , by numerically integrating over both dimensions.

The observed data consist of the positions of all cells at the terminal time of the assay together with the tracks of five randomly chosen cells, $i = 1, \dots, 5$. We let $N(t_n)$ be the number of cells at time $t_n = \tau \times n$, i.e. at the n th iteration of a simulation, so that $n = 288$ corresponds to the final simulation time, $T = 12$ h. We also let $X_i(t_n) = (X_i(t_n), Y_i(t_n))$ be the (x, y) lattice coordinates of cell i at time t_n , so that $\|\delta X_{ij}(t_n)\| = \|X_i(t_n) - X_j(t_n)\|$ is the horizontal distance between cell i and cell j at time t_n , and similarly for δY_{ij} . We summarize these data using statistics motivated by random walk models [5,20,23], considering thirteen statistics in total, labelled as in table 1.

- The first summary statistic we consider is the final cell number, $N(T)$.
- In order to assess whether cells are forming clusters, the second and third summary statistics we consider relate to the size of the largest cluster of cells, computed using the Matlab function `bwconncomp`.¹ We consider either 4-connected clusters (cells orthogonally adjacent are part of a single cluster), $\kappa_4(T)$, or 8-connected clusters (cells diagonally adjacent are also part of a single cluster), $\kappa_8(T)$.

¹See <https://uk.mathworks.com/help/images/ref/bwconncomp.html> (accessed March 2018).

Table 1. A list of the summary statistics and the corresponding notation adopted. Note, apart from cell trajectory statistics, all summary statistics are evaluated at the final time, T .

summary statistic	notation
number of cells	N
size of largest cluster, k -connected	κ_k , $k = 4, 8$
binning variance, bin size = k	Q_k , $k = 2, 4, 8$
Manhattan displacement of cell i	$\ x\ $
Manhattan tortuosity of cell i	Γ
smallest gyration tensor eigenvalue	λ
correlation function	$C_{XY}(l)$
normalized two-dimensional correlation function	$\hat{C}_{XY}(l)$
one-dimensional correlation function	$C_Y(l)$
normalized one-dimensional correlation function	$\hat{C}_Y(l)$

- Summary statistics four to six correspond to the binning variance [5], which quantifies the deviation from the average number of cells expected in quadrats of the domain, and is computed as

$$Q_k(T) = \sum_{b=1}^{B_k} \left(n(b, T) - \frac{N(T)}{B_k} \right)^2 \quad \text{for } B_k = \frac{RC}{k^2} \quad \text{and } k = 2, 4, 8, \quad (2.3)$$

where $n(b, T)$ is the number of cells in bin b at time T , and B_k the number of bins when the bin width is k .

- As summary statistics seven and eight, we consider trajectory statistics from the five tracked cells (labelled $i = 1, \dots, 5$): either the total Manhattan displacement (the sum of horizontal and vertical distances moved) of the cells, $\|x\|$, where

$$\|x\| = \frac{1}{5} \sum_{i=1}^5 \|x_i\| \quad \text{and} \quad \|x_i\| = \sum_{n=1}^{288/8} \|X_i(t_{8n}) - X_i(t_{8(n-1)})\|_1, \quad (2.4)$$

or the tortuosity of the trajectory, Γ , where

$$\Gamma = \frac{1}{5} \sum_{i=1}^5 \frac{\|x_i\|}{\|X_i(T) - X_i(0)\|_1}. \quad (2.5)$$

- Summary statistic nine is the smallest eigenvalue, λ , of the gyration tensor [31],

$$G = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix}, \quad (2.6)$$

where

$$\left. \begin{aligned} G_{11} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta X_{ij}(T), \\ G_{12} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta Y_{ij}(T), \\ G_{21} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta X_{ij}(T) \delta Y_{ij}(T), \\ G_{22} &= \frac{1}{N(T)^2} \sum_{i=1}^{N(T)} \sum_{j=i+1}^{N(T)} \delta Y_{ij}(T) \delta Y_{ij}(T), \end{aligned} \right\} \quad (2.7)$$

and

and the smallest eigenvalue quantifies the spread of cells along the lesser of the two principal axes.

- Summary statistics ten and eleven consider the distribution of pairs of cells, using the pairwise correlation functions, where the argument l indicates the number of pairs of cells separated by distance l . We consider pairwise correlations that measure only the vertical separation between

cells, C_Y (due to the heterogeneity of initial condition in the y direction), or the total separation of cells, C_{XY} , where

$$C_Y(l) = \frac{1}{2} \sum_{i=1}^{N(T)} \sum_{j=1}^{N(T)} 1_{\{l\}} [\|\delta Y_{ij}(T)\|] \quad (2.8)$$

and

$$C_{XY}(l) = \frac{1}{2} \sum_{i=1}^{N(T)} \sum_{j=1}^{N(T)} 1_{\{l\}} [\|\delta X_{ij}(T)\| + \|\delta Y_{ij}(T)\|], \quad (2.9)$$

and $1_{\{l\}}$ is the indicator function which, in this case, counts pairs of cells separated by lattice distance l .

- Finally, we consider the correlation functions normalized by the expected number of pairs at each distance $l = 1, \dots, L$, which accounts for the density of cells in the domain,

$$\hat{C}_Y(l) = \frac{C_Y(l)}{q_Y(l)}, \quad \hat{C}_{XY}(l) = \frac{C_{XY}(l)}{q_{XY}(l)}, \quad (2.10)$$

where

$$q_Y(l) = C^2(R - l)\rho\bar{\rho}, \quad (2.11)$$

$$q_{XY}(l) = \left[C(R - l) + 2 \sum_{j=1}^{l-1} (C - j)(R - (l - j)) + (C - l)R \right] \rho\bar{\rho} \quad (2.12)$$

and

$$\rho = \frac{N}{CR}, \quad \bar{\rho} = \frac{N - 1}{CR - 1}, \quad (2.13)$$

and C and R are the numbers of columns and rows in the lattice, respectively.

2.3.3. Computing distances

Experiments are typically performed multiple times in similar conditions, so we simulate multiple datasets, and call each set a replicate. The observed data, \mathcal{D}^{obs} , consist of M replicates, $\mathcal{D}^{\text{obs}} = \{\mathcal{D}_m^{\text{obs}} : m = 1, \dots, M\}$. The simulated data, \mathcal{D} , consist of M replicates and K samples, $\mathcal{D} = \{\mathcal{D}_m^k : m = 1, \dots, M; k = 1, \dots, K\}$. From the simulated data, we compute each summary statistic, $s_j(l)$, where $j = 1, \dots, 13$ denotes the 13 different summary statistics, and $l = 1, \dots, L_j$ is the l th element of statistic ($L_j = 24$ for the pairwise correlation function statistics because there are 24 rows in the lattice, and $L_j = 1$ otherwise).

First we average over the replicates,

$$\bar{s}_j^k(l) = \frac{1}{M} \sum_{m=1}^M s_{j,m}^k(l) \quad \text{and} \quad \bar{s}_j^{\text{obs}}(l) = \frac{1}{M} \sum_{m=1}^M s_{j,m}^{\text{obs}}(l), \quad (2.14)$$

then we compute the median absolute deviation (MAD) for each statistic, $\sigma_j(l)$. This is defined for a univariate dataset, $X = \{X_i\}$, as $\text{MAD} = \text{median}(|X_i - \text{median}(X)|)$. We then compute the distances for each statistic, $j = 1, \dots, 13$, as

$$d_j^k = \sqrt{\frac{1}{L_j} \sum_{l=1}^{L_j} \left(\frac{\bar{s}_j^k(l) - \bar{s}_j^{\text{obs}}(l)}{\sigma_j(l)} \right)^2}. \quad (2.15)$$

Note that this choice of distance function weights vectors of summary statistics of different lengths equally. This is of particular importance when performing parameter inference using combinations of summary statistics, as in this work. We consider combinations of $A = 1, 2$ or 3 summary statistics from the 13 statistics listed in table 1 to give

$$d^k = \sqrt{\sum_{a=1}^A (d_a^k)^2}, \quad (2.16)$$

where the subscript a indexes the specific summary statistics used.

2.3.4. Regression adjustment

After performing ABC rejection (and during ABC-DC), we perform regression adjustment to derive a more accurate estimate of the posterior distribution [16,25,32]. We assume that θ satisfies the following regression model:

$$\theta_k = \alpha + (s^k - s^{\text{obs}})^T \beta + \xi_k, \quad k = I(1), \dots, I(B), \quad (2.17)$$

where $I = \{k : d^k < \delta\}$, $B = |I|$, s denotes the vector of summary statistics considered in the distance function (2.15) and the ξ_k are uncorrelated Gaussian random variables with zero mean and common variance σ^2 . The regression model is solved to find the least-squares estimates, $(\hat{\alpha}, \hat{\beta})$, and the accepted parameter sets are adjusted according to

$$\theta_k^* = \theta_k - (s^k - s^{\text{obs}})^T \hat{\beta}, \quad k = I(1), \dots, I(B). \quad (2.18)$$

2.3.5. Data-cloning approximate Bayesian computation

Data-cloning ABC involves considering a dataset, $\mathcal{D}_K^{\text{obs}}$, containing K clones of the experimental data, \mathcal{D}^{obs} , that is, $\mathcal{D}_K^{\text{obs}} = (\mathcal{D}^{\text{obs}}, \mathcal{D}^{\text{obs}}, \dots, \mathcal{D}^{\text{obs}})$, where each clone is assumed independent of the others. The likelihood of the cloned data is then [25,26]

$$\mathcal{L}(\mathcal{D}_K^{\text{obs}} | \theta) = (\mathcal{L}(\mathcal{D}^{\text{obs}} | \theta))^K. \quad (2.19)$$

Hence, the posterior distribution resulting from cloned data satisfies

$$\mathbb{P}(\theta | \mathcal{D}_K^{\text{obs}}) \propto \mathcal{L}(\mathcal{D}_K^{\text{obs}} | \theta) \pi(\theta) = (\mathcal{L}(\mathcal{D}^{\text{obs}} | \theta))^K \pi(\theta), \quad (2.20)$$

and the MLE of θ is equivalent to the mean of $\mathbb{P}(\theta | \mathcal{D}_K^{\text{obs}})$ as $K \rightarrow \infty$ [26]. Intuitively, we can see this by reasoning that, if all of the independent model-generated datasets are close to the experimental data, we are K times as likely to have selected a sensible candidate parameter.

The ABC-DC algorithm was proposed by Picchini *et al.* [25], and it uses ABC Markov chain Monte Carlo [33]. The algorithm works in two stages first the acceptance threshold, ϵ , is decreased according to a tolerance scheme $\{\epsilon_1, \epsilon_2, \dots, \epsilon_P\}$, and then the number of clones, K , is increased according to a clone scheme $\{K_{P+1}, K_{P+2}, \dots, K_{P+Q}\}$, where subscripts index the population number. The MLE is approximated by averaging the final parameter population, and the 90% credible interval can be used for comparison to ABC rejection.

For each sample, a proposal $\theta^\#$ is accepted or rejected according to the acceptance probability [25]

$$J_j^\epsilon(\mathcal{D}^\#, \mathcal{D}^{\text{obs}}) = \exp \left\{ -\frac{1}{2\epsilon^2} \left(\frac{1}{L_j} \sum_{l=1}^{L_j} \left(\frac{\bar{s}_j^\#(l) - \bar{s}_j^{\text{obs}}(l)}{\sigma_j(l)} \right)^2 \right) \right\} = \exp \left\{ -\frac{(d_j^\#)^2}{2\epsilon^2} \right\}, \quad (2.21)$$

which approximates a Gaussian centred on the experimental observation with variance ϵ , weighted according to the MAD, $\sigma_j(l)$, of the summary statistic j under consideration, where the MAD is estimated using ABC rejection.

The ABC-DC method is described in full in algorithm 2. The number of samples, r , for the decreasing tolerance stage is denoted by $\{r_1, \dots, r_P\}$, and $\{r_{P+1}, \dots, r_{P+Q}\}$ denotes the number of samples for the increasing clones stage. The scheme we use has $P = 5$ and $Q = 4$, and other variables are as follows:

- number of samples, $r = (1000, 1000, 1000, 1000, 1000, 1000, 1000, 1000, 2000)$,
- number of clones, $K = (1, 1, 1, 1, 1, 1, 2, 3, 4)$,
- value of tolerance, $\epsilon = (5, 2, 1, 0.5, 0.25, 0.25, 0.25, 0.25, 0.25)$.

Algorithm 2. ABC-DC algorithm.

```

1: initialize  $\theta^* = (0.5, 0.005)$ 
2: for  $p = 1$  to  $P$  do
3:   set  $\theta_1 = \theta^*$ , and fix  $\varepsilon = \varepsilon_p$ ,  $K = 1$ 
4:   simulate  $\mathcal{D}^*$  from the model with parameter  $\theta^*$ , and compute  $q^* = J_j^\varepsilon(\mathcal{D}^*, \mathcal{D}^{\text{obs}})$ 
5:   for  $k = 1$  to  $r_p - 1$  do
6:     sample a new  $\theta^\# \sim N(\theta^*, (2.38)^2 \Sigma_p / 2)$ , where  $\Sigma_p$  is the covariance of the previous  $\theta$ 
       population (or the prior if  $p = 1$ )a.
7:     simulate  $\mathcal{D}^\#$  from the model with parameter  $\theta^\#$ , and compute  $q^\# = J_j^\varepsilon(\mathcal{D}^\#, \mathcal{D}^{\text{obs}})$ 
8:     compute  $\alpha = \min \left[ 1, \frac{q^\#}{q^*} \right]$  and generate  $\omega \sim U(0, 1)$ 
9:     if  $\omega > \alpha$  then
10:       set  $\theta_{k+1} := \theta_k$ 
11:     else
12:       set  $\theta_{k+1} := \theta^\#$ ,  $\theta^* = \theta^\#$  and  $q^* := q^\#$ .
13:     end if
14:     if  $p = P$  and  $\max \pi_p < q^\# \pi(\theta^\#)$  then
15:       update current maximum-likelihood estimate: set  $\max \pi \varepsilon_p = q^\# \pi(\theta^\#)$  and  $\theta_{\text{MLE}} = \theta^\#$ .
16:     end if
17:   end for
18: end for
19: perform regression adjustment on the final  $\theta$  population, set  $\Sigma_p$  as the covariance of the adjusted  $\theta$ .
20: for  $q = 1$  to  $Q$  do
21:   set  $\theta_1 = \theta^*$ , and fix  $\varepsilon = \varepsilon_p$ ,  $K = K_{p+q}$ 
22:   simulate  $K$  independent data sets from the model,  $\mathcal{D}^{*(1)}, \dots, \mathcal{D}^{*(K)}$ , with parameter  $\theta^*$ , and
       compute  $q^* = \prod_{k=1}^K J_j^\varepsilon(\mathcal{D}^{*(k)}, \mathcal{D}^{\text{obs}})$ 
23:   for  $k = 1$  to  $r_{p+q} - 1$  do
24:     sample a new  $\theta^\# \sim N(\theta_{\text{MLE}}, \Sigma_{p+q-1})$ 
25:     simulate  $K$  independent data sets from the model,  $\mathcal{D}^{\#(1)}, \dots, \mathcal{D}^{\#(K)}$ , with parameter  $\theta^*$ ,
       and compute  $q^\# = \prod_{k=1}^K J_j^\varepsilon(\mathcal{D}^{\#(k)}, \mathcal{D}^{\text{obs}})$ 
26:     compute  $\alpha = \min \left[ 1, \frac{q^\#}{q^*} \right]$  and generate  $\omega \sim U(0, 1)$ 
27:     if  $\omega > \alpha$  then
28:       set  $\theta_{k+1} := \theta_k$ 
29:     else
30:       set  $\theta_{k+1} := \theta^\#$ ,  $\theta^* = \theta^\#$  and  $q^* := q^\#$ .
31:     end if
32:   end for
33:   update  $\Sigma_{p+q}$  as the covariance of the current  $\theta$  population
34: end for

```

^aThe factor $2.38^2/d$ is optimal for Markov chain Monte Carlo exploration [11].

3. Results

Our aim in this work is to provide insights into how both the choices of experimental geometry and summary statistics impact the quality of the posteriors generated using ABC for experiments that are typically used to characterize growing and spreading cell populations. We generate *in silico* (observed) data using the model outlined in §§2.1 and 2.2 for the range of experimental designs illustrated in figure 1c, and quantify the quality of the posteriors resulting from the use of ABC rejection with various summary statistics using the Kullback–Leibler divergence between the prior and posterior distributions (details of the methods used can be found in §2.3). In §§3.1 and 3.2, we ask whether it is possible to quantify, using only a single summary statistic of the data, the relative contributions of proliferation and motility in a number of cell spreading experiments. In §3.3, we explore the extent to which parameter estimates can be improved using multiple summary statistics. Finally, in §3.4 we demonstrate the use of ABC-DC to efficiently find MLEs, and compare these estimates to those generated using ABC rejection.

3.1. Can we infer both model parameters using a single summary statistic?

We first ask whether we can jointly infer both model parameters, P_m and P_p , using a single summary statistic in the distance function (2.15). To do this, we perform ABC rejection on *in silico* (observed) data gathered from simulations with $P_m = 0.25$ and $P_p = 0.0025$, where we replicate different experimental designs by varying the model initial conditions: we initialize the model with 24 cells in all 24 rows, or in 12 rows, or in six rows, of the domain, as shown in figure 1c. Note that this means we change the average density within the region into which cells are initialized, but that the overall cell number, and hence the average total density, is constant. Figure 2a–f shows the posteriors generated for selected summary statistics of the data. The average information gain, and its deviation, for each of the three experimental designs and all summary statistics, is summarized in figure 2g.

Figure 2a,b shows that for a growth-to-confluence assay design (where cells are initialized across all 24 rows of the domain), an accurate estimate of P_p can be obtained using the number of cells, N , as a summary statistic, and an accurate estimate of P_m can be obtained using the Manhattan displacement, $\|x\|$, as a summary statistic. However, no single summary statistic can be used to infer both P_p and P_m with any degree of confidence. Figure 2c,d, however, demonstrates that, in contrast to a growth-to-confluence assay design, using a scratch assay design (where cells are initialized only in six rows of the domain) enables both P_m and P_p to be accurately estimated using a single summary statistic, for example the one-dimensional correlation function summary statistic, C_Y . Finally, figure 2e,f demonstrates how the predicted posterior can vary for a single summary statistic as the experimental design is changed. We use as an example the binning variance statistic, Q_8 , which is the deviation in cell numbers from the mean when the domain is divided into 8×8 bins. Increasing the rate of movement, P_m , means clusters are broken up more rapidly, while increasing the rate of proliferation, P_p , makes clusters larger. As a result, *in silico* (observed) data generated using small P_m and P_p result in the same binning variance as data generated with large P_m and P_p . The effect is more pronounced for a scratch assay design due to the initial confinement of cells.

Figure 2g summarizes the quality of the predicted posteriors resulting from the use of each of the summary statistics listed in table 1 for each experimental design considered in figure 1c. Our results demonstrate that statistics based on cell numbers provide the same or less information as cells are increasingly confined initially, whereas the opposite is true for the summary statistics that relate to correlations in cell positions. Both model parameters, P_m and P_p , can be inferred using a single summary statistic, without the need for cell trajectory information, but only when cells are initially confined to a scratch-assay-like geometry. For growth-to-confluence assays, cell trajectory data are necessary to accurately infer both model parameters.

3.2. What is the impact of the initial number of cells on the posteriors?

Next, we consider how changing the initial number of cells in the assay (24, 48 or 72 cells) affects the quality of the predicted posterior distribution. We consider either a growth-to-confluence assay (cells initialized uniformly at random across the domain, figure 3a–c) or a scratch assay design (cells initialized uniformly into six rows of the domain, figure 3d–f). Figure 3a–c demonstrates that, for the growth-to-confluence assay, increasing the initial cell number does not facilitate accurate inference of both model parameters using a single summary statistic. In line with the results presented in figure 2, figure 3d–f indicates that both model parameters can be accurately estimated using a single summary statistic when the scratch assay is used; however, increasing the initial cell number does not significantly increase the quality of the posterior. We quantify the quality of the estimated posterior for each summary statistic and each experimental design in the bar plots in figure 3c,f. For almost every summary statistic, the quality of the posterior increases as the initial cell number increases, likely due to a reduction in the variance of each summary statistic.

3.3. Are trajectory statistics necessary for maximal information gain?

It is natural to ask whether multiple summary statistics can be combined to increase the quality of the estimated posteriors and, further, whether combining different summary statistics can avoid the need to collect cell trajectory data to accurately infer parameters of a growth-to-confluence assay. To answer this question, we considered all combinations of two or three summary statistics from the 13 under consideration, computing the resulting posteriors and information gain for each combination. Figure 4 demonstrates that, for a growth-to-confluence assay, cell trajectory data are necessary to generate a

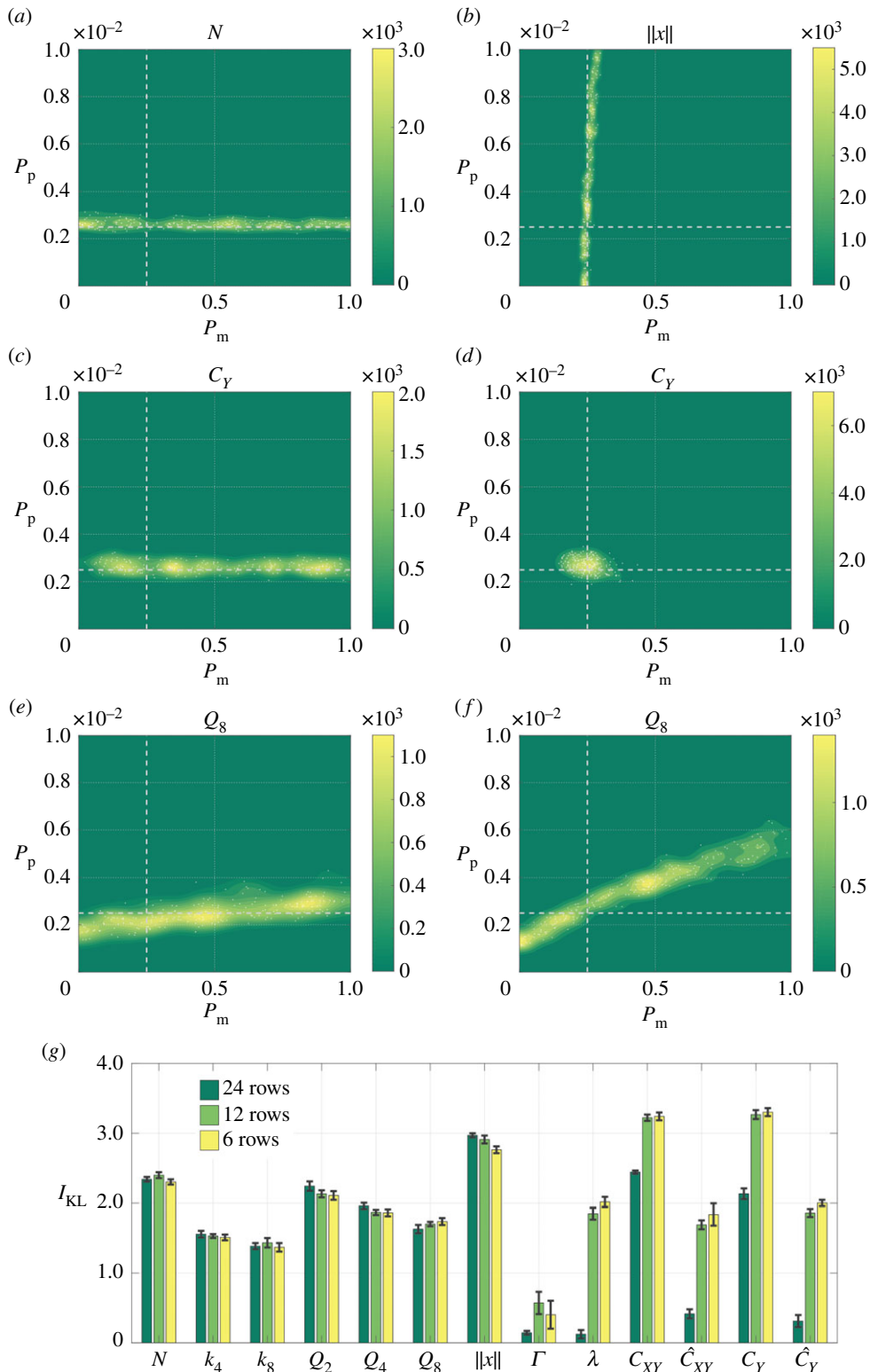


Figure 2. Posteriors generated using a single summary statistic as the experimental design is varied. The parameters used to generate the *in silico* (observed) data are indicated using dashed white lines. (a,b) The posterior distributions corresponding to the most informative summary statistic for inferring P_p ((a), N) or P_m ((b), $\|x\|$) for the growth-to-confluence assay (cells initialized uniformly at random over 24 rows). (c,d) The posterior distributions generated using the summary statistic C_Y for the growth-to-confluence assay (c) and the scratch assay (d) (cells initialized uniformly at random over six rows). (e,f) The posterior distributions for summary statistic Q_8 for the growth-to-confluence assay (e) and the scratch assay (f). (g) The information gain in moving from the prior to the posterior for each summary statistic and for each experimental design, where the error bars denote the standard deviation.

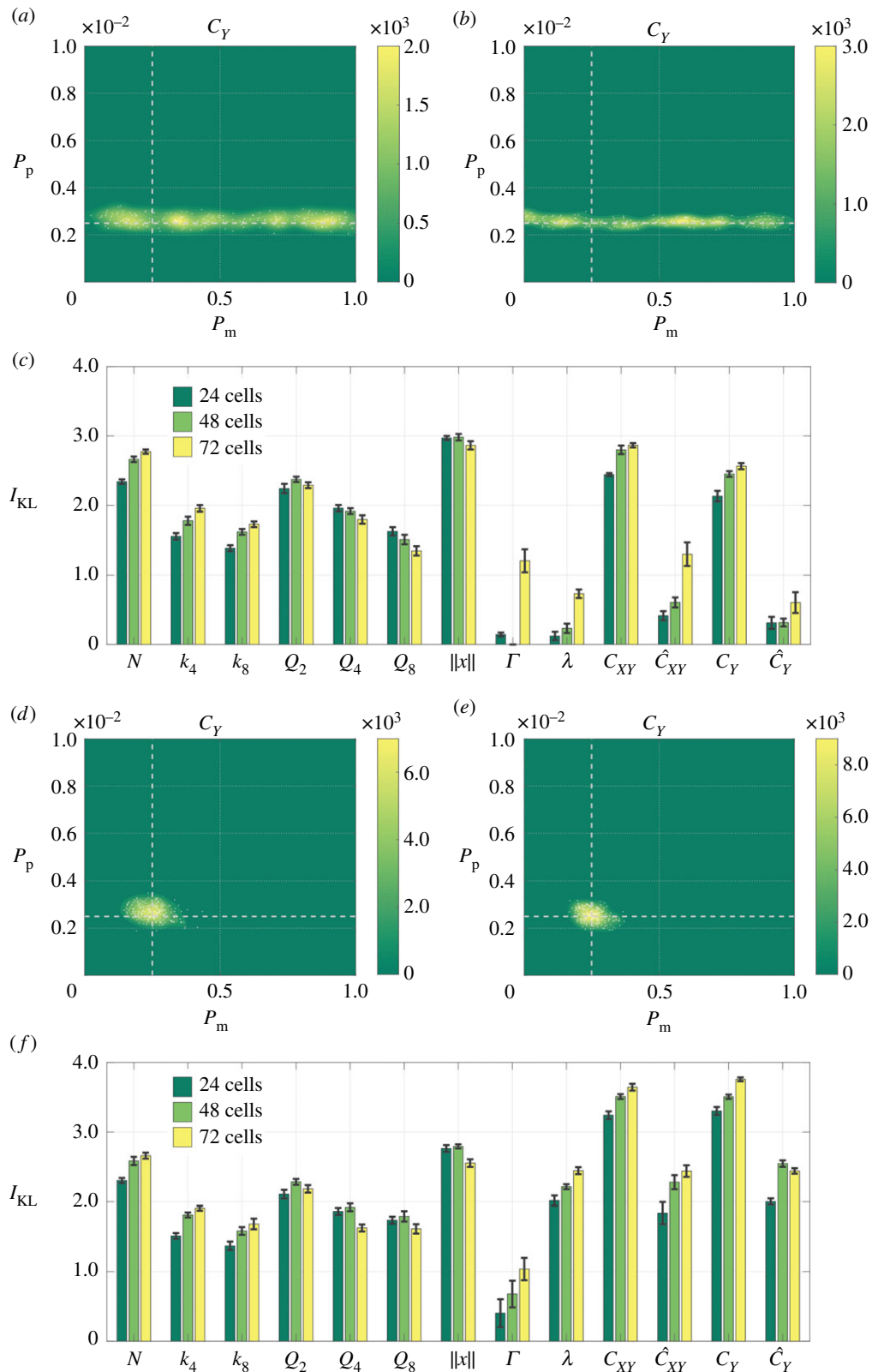


Figure 3. Posteriors generated using a single summary statistic as the initial cell number is varied. The parameters used to generate the *in silico* (observed) data are indicated using dashed white lines. (a–c) Results for a growth-to-confluence assay generated using 24 cells (a) and 72 cells (b). (d–f) Results for a scratch assay generated using 24 cells (d) and 72 cells (e). The posterior distributions in (a,b) and (d,e) are generated using the pairwise correlations summary statistic C_Y . (c,f) The information gained in moving from the prior to the posterior for each summary statistic for each experimental design, where the error bars denote the standard deviation.

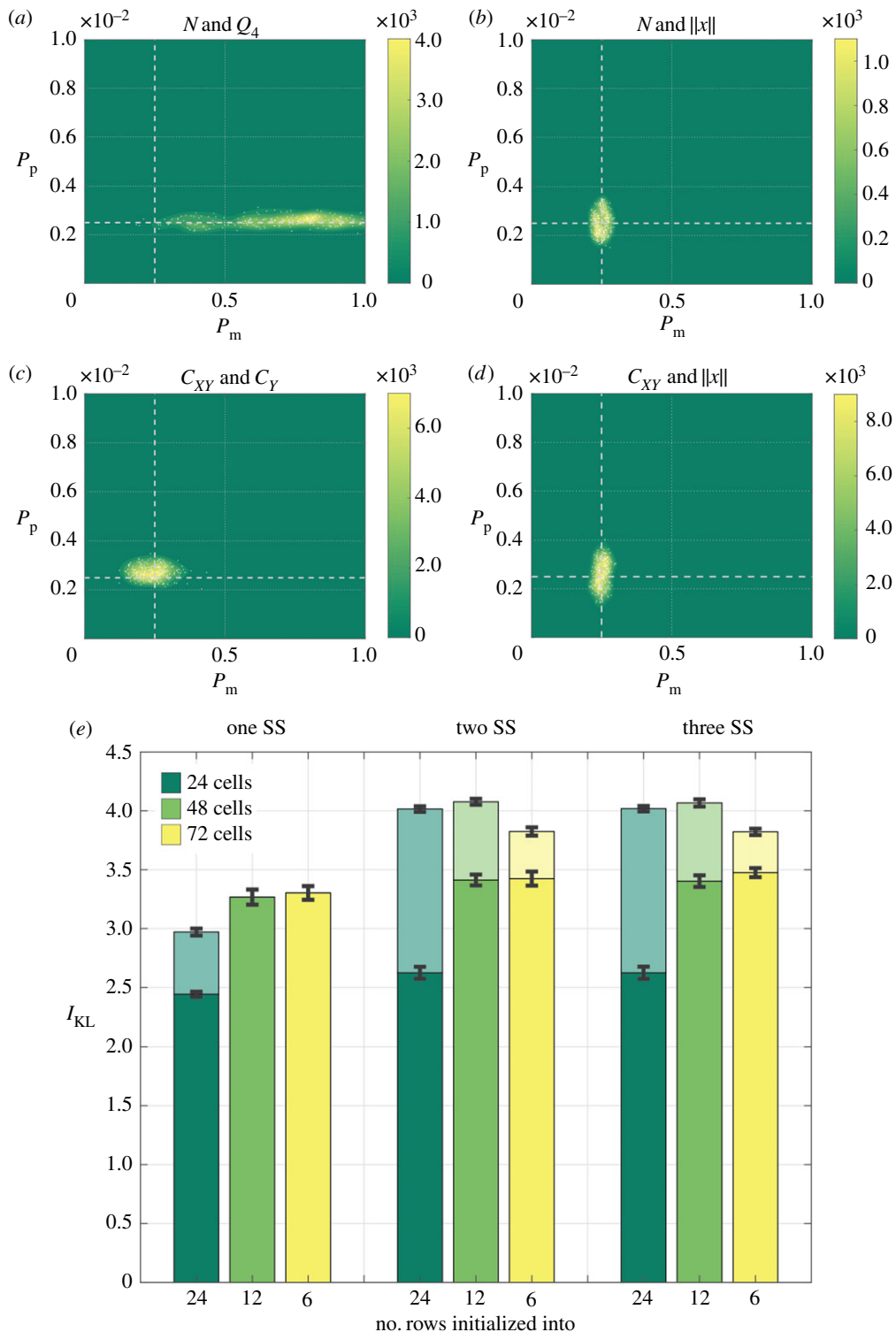


Figure 4. Posterior distributions and information gain when considering combinations of summary statistics (SS). (a–d) Posteriors resulting from the most informative two-summary-statistic combination when considering either a growth-to-confluence assay (a,b), or a scratch assay (c,d), excluding trajectory information (a,c), or including trajectory information (b,d). (e) The information gain in moving from the prior to the posterior for each of the best-performing summary statistic combinations, varying the experimental design, and excluding trajectory information (dark shading) or including trajectory information (light shading), where the error bars denote the standard deviation.

posterior that allows for accurate estimation of both parameters P_m and P_p (figure 4a,b). In particular, trajectory information is necessary to estimate the motility parameter, P_m . It is interesting to note that, in both cases (with and without trajectory information), the number of cells summary statistic, N , is

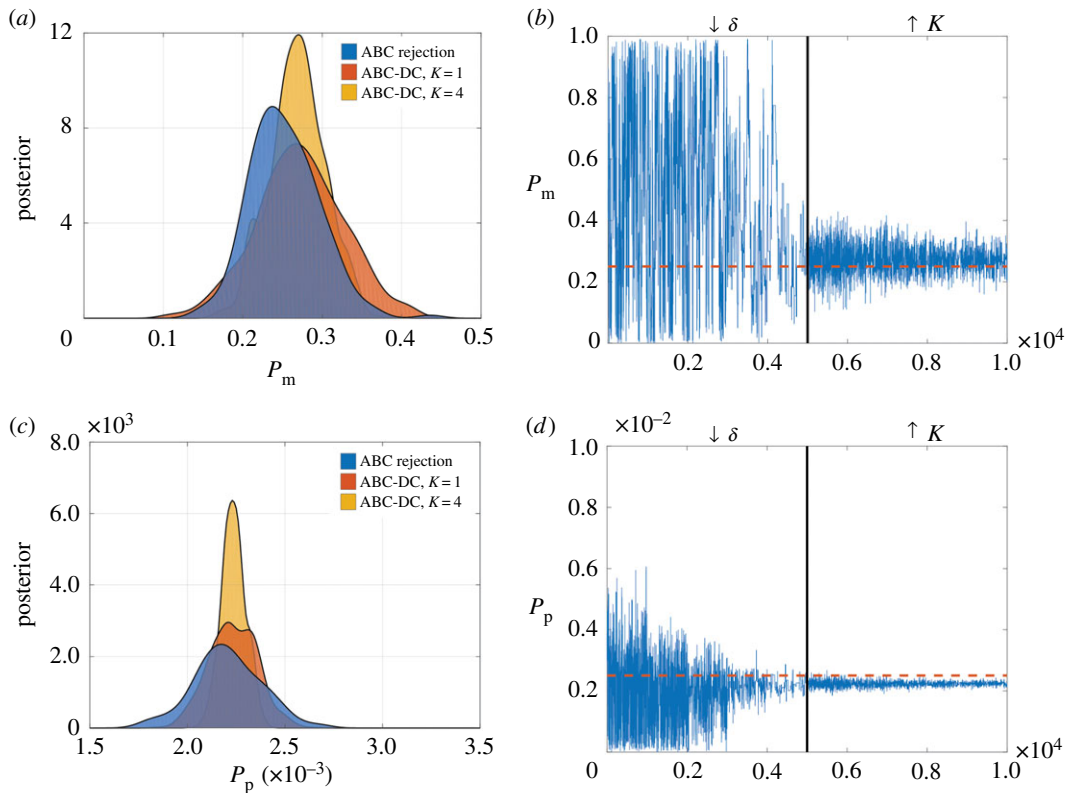


Figure 5. Example results from the ABC-DC algorithm for a scratch assay using the two-dimensional pairwise correlations statistic, C_{XY} . (a,c) Estimates of the posteriors for P_m and P_p , respectively, generated using kernel density estimation. We compare the posteriors for P_m and P_p generated using ABC-DC (yellow), ABC-DC (before the clone numbers are increased) (red) and ABC rejection (blue). (b,d) The trajectories of parameters accepted during the Markov chain Monte Carlo step of the ABC-DC algorithm.

included in the best-performing two-summary-statistic combinations, likely because it facilitates accurate estimation of the proliferation parameter, P_p . For a scratch assay, cell trajectory information is not necessary to accurately infer both model parameters; however, as expected, including cell trajectory data enables more accurate inference of the cell motility parameter, P_m . Figure 4e quantifies the quality of the estimated posteriors for a range of experimental designs, initial cell densities and for combinations of one, two or three summary statistics. Results demonstrate that little extra information is gained by considering more than two summary statistics and that, as expected, accurate parameter inference for growth-to-confluence assays requires cell trajectory data.

3.4. Can data-cloning approximate Bayesian computation provide an efficient means to estimate model parameters?

Finally, we investigate whether the ABC-DC algorithm proposed by Picchini *et al.* [25] can provide an efficient means to estimate model parameters. ABC-DC essentially involves using a number of copies of the observed data (called ‘clones’) that are assumed to be independently generated datasets. ABC-DC uses ABC Markov chain Monte Carlo [33] over two stages: first the acceptance threshold, ϵ , is decreased; and then the number of clones, K , is increased. This has the impact of concentrating the posterior around the MLE, with the posterior mean an approximation to the MLE, and the 90% credible interval available for quantification of the uncertainty in parameter predictions. Figure 5 shows the results of using ABC-DC with the two-dimensional pairwise correlation function, C_{XY} , to estimate posterior distributions for both P_m and P_p . Figure 5a,c shows how the posterior changes as the number of clones is increased, relative to ABC rejection, while figure 5b,d shows the Markov chains of both P_m and P_p . Our results indicate that, as expected, the ABC-DC algorithm results in a tighter posterior focused approximately about the MLEs found using ABC-DC with a single instance of the data, $K=1$. To understand how the MLEs generated using ABC-DC differ from those generated using ABC rejection, we look at the posteriors generated using each algorithm together with three summary statistics: the number of cells, N ; the two-dimensional pairwise correlation

Table 2. ABC-DC for a scratch assay experiment. The MLEs and the 5%–95% empirical percentiles are shown for parameters accepted using ABC rejection and ABC-DC, for different summary statistics. The speed up of the ABC-DC method over ABC rejection is also given.

statistic: parameter	ABC rejection MLE (90% CI)	ABC-DC MLE (90% CI)	speed up
$\ x\ : P_m$	0.250 (0.231, 0.300)	0.247 (0.239, 0.255)	1.65
$C_{XY}: P_m$	0.239 (0.174, 0.310)	0.236 (0.186, 0.280)	4.50
$C_{XY}: P_p (\times 10^{-2})$	0.241 (0.215, 0.270)	0.228 (0.219, 0.236)	4.50
$N: P_p (\times 10^{-2})$	0.243 (0.220, 0.282)	0.246 (0.235, 0.257)	1.72

statistic, C_{XY} ; and the Manhattan displacement, $\|x\|$. We calculate MLEs and the 5% and 95% empirical percentiles of the accepted parameters. Results in table 2 indicate that the MLEs are similar for both ABC rejection and ABC-DC, though the credible intervals for ABC-DC are much smaller, and not a good representation of the true uncertainty in each parameter. However, ABC-DC can provide significant computational savings in estimation of the MLE, with maximal efficiency displayed with summary statistics that are relatively computationally demanding to compute, such as C_{XY} .

3.5. Predictions for experimental data

We can use our model to infer parameter distributions for different cell types in a growth-to-confluence assay, and subsequently make predictions about long-time behaviours, such as the time taken to reach confluence. In figure 6, we use data from the growth-to-confluence experiments detailed in [5] to infer motility and proliferation rates for two different cell types: 3T3 fibroblast cells and MDA MB 231 breast cancer cells. In particular, for each cell type, we use 10 experimental replicates to estimate the bivariate posterior distribution for P_m and P_p , with data collected 12 h into the experiment. Figure 6*a,b* shows the posteriors generated using ABC rejection with the cell number, N , and Manhattan displacement, $\|x\|$, as summary statistics, and the priors

$$\pi_{3T3}(\theta) = \pi_{231}(\theta) = \frac{1}{0.1 \times 0.01}, \quad \theta \in (0, 0.1) \times (0, 0.01).$$

For the 3T3 fibroblast cell line, we obtain $P_m = 0.0465 \pm 0.00341$ and $P_p = 0.00163 \pm 0.000403$, while for the MDA MB 231 breast cancer cells we obtain $P_m = 0.0227 \pm 0.00254$ and $P_p = 0.000931 \pm 0.000427$. In each case, we give the posterior mean and the standard error. In real terms, for the 3T3 cells this gives a diffusivity of $D = 97.986 \pm 7.196 \mu\text{m}^2 \text{h}^{-1}$ and a proliferation rate of $\lambda = 0.0392 \pm 0.00968 \text{h}^{-1}$, while for the 231 cells we have $D = 47.850 \pm 5.347 \mu\text{m}^2 \text{h}^{-1}$ and $\lambda = 0.0223 \pm 0.0102 \text{h}^{-1}$.

To check the validity of our model and inferred posterior distributions, we collected data from the same experimental replicates at 24 h and used these to perform a posterior predictive check. In figure 6*c–f*, we present the results of performing the posterior predictive check using both the cell number, N , and the Manhattan displacement, $\|x\|$, as summary statistics. We produce results using, for each dataset, 1000 samples from the estimated posteriors, and visually compare the statistics for each experimental replicate with the average of the predicted statistics and their 95% confidence intervals. These plots show strong support for the model for each cell type. We see that for the number of cells statistic, all of the experimental statistics fall within the 95% confidence intervals. For the Manhattan displacement statistic, two of the 10 replicates for the 231 cells have an experimental statistic that falls outside the 95% confidence interval of the simulated statistic. In this case, we suspect translation of the positions of slow-moving cells (as in the case of the 231 cells) onto a lattice has introduced artefacts, causing discrepancy between simulation displacement and experimental displacement.

4. Discussion

Mechanistic models together with experimental data from cell biology assays provide an excellent opportunity to characterize the extent to which cell proliferation and cell motility can drive the growth and expansion of a cell colony. To be able to draw quantitative distinctions between the relative contributions of cell proliferation and motility for different cell types and biological scenarios requires the accurate inference of model parameters using quantitative summaries of experimental data.

In this work, we have applied ABC to explore the extent to which different experimental designs and quantitative summary statistics impact our ability to accurately infer parameters of a simple mechanistic

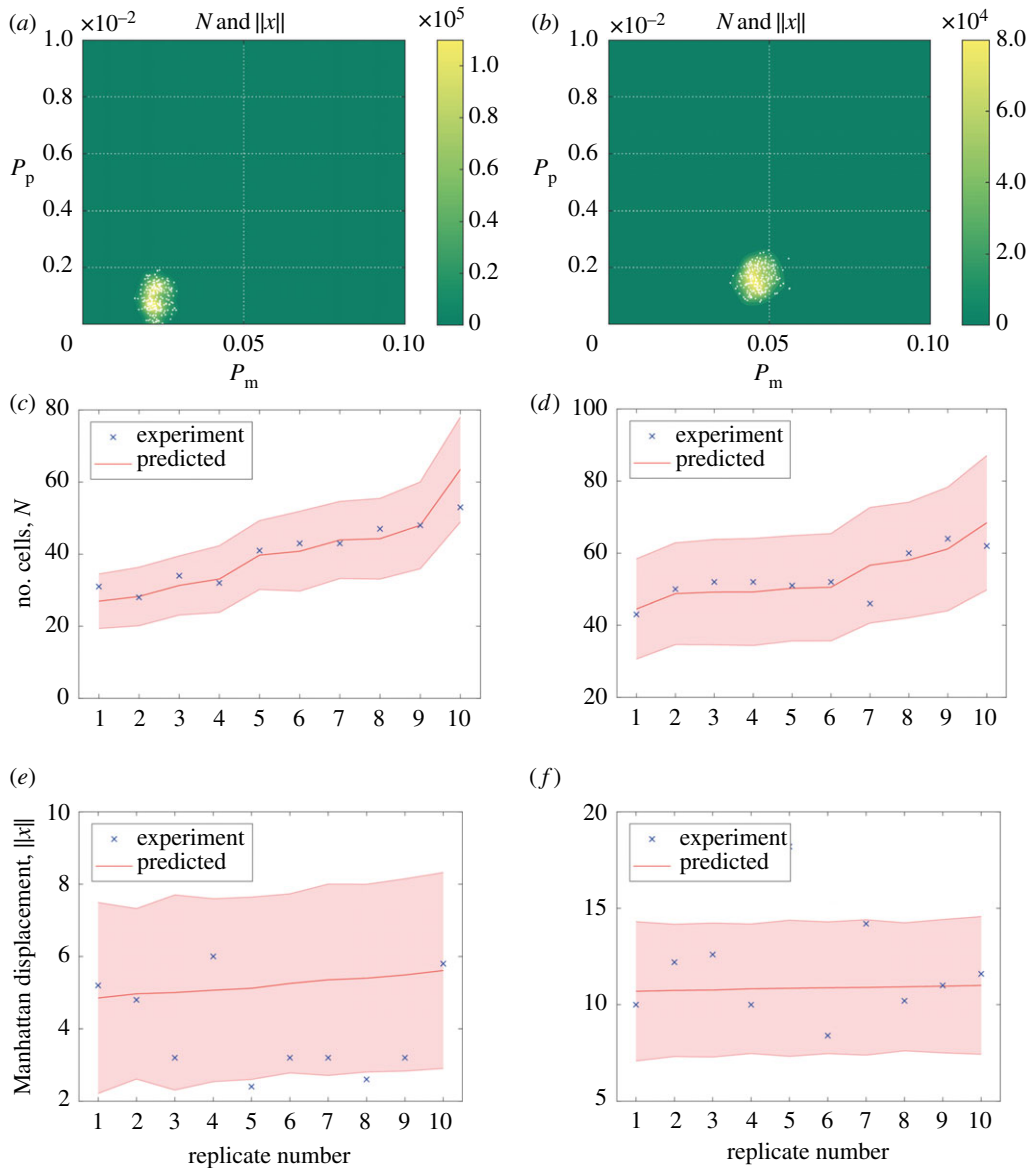


Figure 6. Posterior distributions for the results of posterior predictive checks. (a,b) Posterior distributions for 231 breast cancer and 3T3 fibroblast cell lines, respectively. (c,d) Posterior predictive check using data collected 24 h into the experiment, for the 231 breast cancer and 3T3 fibroblast cell lines, respectively, and the cell number, N , as a summary statistic. The ten experimental replicates are ordered according to increasing cell number at 24 h. For each replicate, we plot the experimental data (blue cross), the mean cell number (red line) and the 95% confidence interval (shaded region). (e,f) Posterior predictive check using data collected 24 h into the experiment, for the 231 breast cancer and 3T3 fibroblast cell lines, respectively, and the Manhattan displacement, $\|x\|$, as a summary statistic. Details are as for (c,d).

model of cell growth and spreading. The model we present is the simplest possible individual-based model of a two-dimensional cell biology experiment in which we can simulate cell motility and cell proliferation with a simple mechanism for contact inhibition. Our rationale for working with this fundamental model is that it is crucial to explore whether it is possible to parametrize this simple model using routinely collected experimental data, before attempting inference for a more biologically realistic model with a greater number of unknown parameters.

Our results suggest that, for growth-to-confluence experiments, cell trajectory information, in addition to summary statistics such as cell numbers and spatial correlations, are required for accurate parameter inference. On the other hand, for a scratch assay geometry, it is possible to accurately infer both motility and proliferation parameters without recourse to tracking individual cells over multiple frames of a microscopy video. As cell tracking provides the bottleneck in the analysis of many experimental studies, this renders experimental geometry an important aspect to consider in the design of experiments.

In addition, our results demonstrate that increases in the initial cell number in cell biology assays can increase the quality of estimated posterior distributions by decreasing noise in the summary statistics, a result also found for a similar model by Ross and co-workers [18]. We also demonstrated that posteriors of significantly increased quality can be generated using a combination of two summary statistics, but that further increases in the number of summary statistics provided little increase in the quality of the posteriors. In electronic supplementary material, figures S1–S4, we demonstrate that our results are consistent across a range of parameter values and a range of experimental durations. However, it should be noted that, because we collected data at only the final time point of the experiment, some summary statistics lose efficacy if experiments are run over a sufficiently long time period that the population grows to confluence. For example, in this case, the number of cells, N , at the final time point provides little information on the model parameters. Finally, we have shown that the ABC-DC approach provides an efficient means to approximate the MLEs of model parameters, with significant improvements in computational efficiency provided where the summary statistics are time-consuming to calculate.

5. Conclusion

In summary, our results suggest that experimental design choices that incorporate initial spatial heterogeneities in cell positions facilitate parameter inference without the requirement of cell tracking, while those that seed cells uniformly initially require cell tracking for accurate parameter inference. As cell tracking is often a major technical limitation of many experimental studies of this type, our recommendations for experimental design choice could lead to significant potential time and cost savings in the analysis of these kinds of commonly used experiments.

Data accessibility. Codes used to generate the data are available from <https://github.com/andrew-parker/Impact-of-experimental-design>.

Authors' contributions. A.P., M.J.S. and R.E.B. conceived the study, which was performed by A.P. A.P., M.J.S. and R.E.B. wrote the manuscript.

Competing interests. We declare we have no competing interests.

Funding. A.P. thanks the UK's Engineering and Physical Sciences Research Council (EPSRC, EP/G03706X/1) for funding through a studentship at the Systems Biology programme of the University of Oxford's Doctoral Training Centre. M.J.S. appreciates support from the Australian Research Council (DP170100474). R.E.B. is a Royal Society Wolfson Research Merit Award holder and thanks the Leverhulme Trust for a Research Fellowship and the UK's Biotechnology and Biological Sciences Research Council (BBSRC, BB/R000816/1).

Acknowledgements. We thank the anonymous referees for their helpful suggestions.

References

- Cheung KJ, Ewald AJ. 2014 Illuminating breast cancer invasion: diverse roles for cell-cell interactions. *Curr. Opin. Cell Biol.* **30**, 99–111. (doi:10.1016/j.cob.2014.07.003)
- Theveneau E, Mayor R. 2012 Neural crest delamination and migration: from epithelium-to-mesenchyme transition to collective cell migration. *Dev. Biol.* **366**, 34–54. (doi:10.1016/j.ydbio.2011.12.041)
- Schneider L *et al.* 2010 Directional cell migration and chemotaxis in wound healing response to PDGF-AA are coordinated by the primary cilium in fibroblasts. *Cell. Physiol. Biochem.* **25**, 279–292. (doi:10.1159/000276562)
- Vicente-Manzanares M, Horwitz AR. 2011 Cell migration: an overview. *Methods Mol. Biol.* **769**, 1–24. (doi:10.1007/978-1-61779-207-6_1)
- Simpson MJ, Binder BJ, Haridas P, Wood BK, Treloar KK, McElwain DLS, Baker RE. 2013 Experimental and modelling investigation of monolayer development with clustering. *Bull. Math. Biol.* **75**, 871–889. (doi:10.1007/s11538-013-9839-0)
- Johnston ST, Shah ET, Chopin LK, McElwain DLS, Simpson MJ. 2015 Estimating cell diffusivity and cell proliferation rate by interpreting IncuCyte ZOOM assay data using the Fisher-Kolmogorov model. *BMC Syst. Biol.* **9**, 38. (doi:10.1186/s12918-015-0182-y)
- Cai AQ, Landman KA, Hughes BD. 2007 Multi-scale modeling of a wound-healing cell migration assay. *J. Theor. Biol.* **245**, 576–594. (doi:10.1016/j.jtbi.2006.10.024)
- Binder BJ, Simpson MJ. 2016 Cell density and cell size dynamics during in vitro tissue growth experiments: implications for mathematical models of collective cell behaviour. *Appl. Math. Model.* **40**, 3438–3446. (doi:10.1016/j.apm.2015.01.016)
- Yue PYK, Leung EPY, Mak NK, Wong RNS. 2010 A simplified method for quantifying cell migration/wound healing in 96-well plates. *J. Biomol. Screen.* **15**, 427–433. (doi:10.1177/1087057110361772)
- Martínez-Poveda B, Quesada AR, Medina MA. 2005 Hypericin in the dark inhibits key steps of angiogenesis in vitro. *Eur. J. Pharmacol.* **516**, 97–103. (doi:10.1016/j.ejphar.2005.03.047)
- Bolker B. 2008 *Ecological models and data in R*. Princeton, NJ: Princeton University Press.
- Aguilar O, Allmaras M, Bangerth W, Tenorio L. 2015 Statistics of parameter estimates: a concrete example. *SIAM Rev.* **57**, 131–149. (doi:10.1137/130929230)
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. 2014. *Bayesian data analysis*, vol. 3. London, UK: Chapman & Hall.
- Kirk P, Thorne T, Stumpf MPH. 2013 Model selection in systems and synthetic biology. *Curr. Opin. Biotechnol.* **24**, 767–774. (doi:10.1016/j.copbio.2013.03.012)
- 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. (doi:10.1038/nprot.2014.025)

16. Beaumont M, Zhang W, Balding D. 2002 Approximate Bayesian computation in population genetics. *Genetics* **162**, 2025–2035.
17. Toni T, Stumpf MPH. 2010 Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics* **26**, 104–110. (doi:10.1093/bioinformatics/btp619)
18. Ross RJH, Baker RE, Parker A, Ford MJ, Mort RL, Yates CA. 2017 Using approximate Bayesian computation to quantify cell-cell adhesion parameters in a cell migratory process. *npj Syst. Biol. Appl.* **3**, 9. (doi:10.1038/s41540-017-0010-7)
19. Vo BN, Drovandi CC, Pettitt AN, Simpson MJ. 2015 Quantifying uncertainty in parameter estimates for stochastic models of collective cell spreading using approximate Bayesian computation. *Math. Biosci.* **263**, 133–142. (doi:10.1016/j.mbs.2015.02.010)
20. 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. (doi:10.1098/rsob.140097)
21. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. 2013 Approximate Bayesian computation. *PLoS Comput. Biol.* **9**, e1002803. (doi:10.1371/journal.pcbi.1002803)
22. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MPH. 2009 Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**, 187–202. (doi:10.1098/rsif.2008.0172)
23. Barnes C, Filippi S, Stumpf M, Thorne T. 2012 Considerate approaches to achieving sufficiency for ABC model selection. *Stat. Comput.* **22**, 1181–1197. (doi:10.1007/s11222-012-9335-7)
24. Kullback S, Leibler RA. 1951 On information and sufficiency. *Ann. Math. Stat.* **22**, 79–86. (doi:10.1214/aoms/1177729694)
25. Picchini U, Anderson R. 2017 Approximate maximum likelihood estimation using data-cloning ABC. *Comput. Stat. Data Anal.* **105**, 166–183. (doi:10.1016/j.csda.2016.08.006)
26. Lele SR, Dennis B, Lutstcher F. 2007 Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. *Ecol. Lett.* **10**, 551–563. (doi:10.1111/j.1461-0248.2007.01047.x)
27. Simpson MJ, Merrifield A, Landman KA, Hughes BD. 2007 Simulating invasion with cellular automata: connecting cell-scale and population-scale properties. *Phys. Rev. E* **76**, 21918. (doi:10.1103/physreve.76.021918)
28. Simpson MJ, Landman KA, Hughes BD. 2010 Cell invasion with proliferation mechanisms motivated by time-lapse data. *Physica A* **389**, 3779–3790. (doi:10.1016/j.physa.2010.05.020)
29. Botev Z, Grotowski J, Kroese D. 2010 Kernel density estimation via diffusion. *Ann. Stat.* **38**, 2916–2957. (doi:10.1214/10-AOS799)
30. Aeschbacher S, Beaumont M, Futschik A. 2012 A novel approach for choosing summary statistics in approximate Bayesian computation. *Genetics* **192**, 1027–1047. (doi:10.1534/genetics.112.143164)
31. Rudnick J, Gaspari G. 2004 *Elements of the random walk: an introduction for advanced students and researchers*. Cambridge, UK: Cambridge University Press.
32. Nunes MA, Balding DJ. 2010 On optimal selection of summary statistics for approximate Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* **9**, 34. (doi:10.2202/1544-6115.1576)
33. Marjoram P, Molitor J, Plagnol V, Tavaré S. 2003 Markov chain Monte Carlo without likelihoods. *Proc. Natl Acad. Sci. USA* **100**, 15 324–15 328. (doi:10.1073/pnas.0306899100)