Manuscript Category

Systems biology

Methods for analysing lineage tracing datasets

Vasiliki Kostiou[1], Huairen Zhang[1], Michael WJ Hall[1,2], Philip H Jones[1,2,*] and Benjamin A Hall[1,*]

[1]MRC Cancer Unit, University of Cambridge, Hutchison-MRC Research Centre, Cambridge Biomedical Campus, Cambridge, UK, [2]Wellcome Trust Sanger Institute, Hinxton CB10 1SA, UK

*To whom correspondence should be addressed.

Abstract

Motivation: A single population of stem cells maintains many epithelial tissues. A mathematical model of this tissue maintenance requires three parameters to describe the growth dynamics observed in transgenic mouse cell tracking- a division rate, a stratification rate, and the probability of dividing symmetrically. Deriving these parameters is time intensive and the results highly sensitive to biological variation.

Results: We compare the alternative strategies for analyzing experimental data, identifying an approximate Bayesian computation- based approach as the best in terms of efficiency and appropriate error estimation. All tested solutions are made available to allow new datasets to be analysed following our workflows. Based on our findings we make recommendations for future experimental design.

Availability: All code is made available as python notebooks under the MIT license.

Contact: bh418@mrc-cu.cam.ac.uk, pj3@sanger.ac.uk

1 Introduction

The stem cell dynamics of squamous epithelial cells is a major subject of study in biomedicine. Squamous epithelial tissues cover the external surface of the body, the mouth, and the esophagus, and the most common human cancers develop from these tissues. Epithelial tissues consist of layers of keratinocytes, and in mice, skin and esophageal tissues are maintained by a single layer of cells at the base of the tissue. Stem cells in this basal layer stochastically differentiate, cease cell division, and then stratify into the upper layers of the tissue, migrating to the surface before eventually being shed. Human squamous tissues have a more complex organization, but share several features and are believed to be maintained in a similar manner.

The 'single progenitor' model has been shown over several years to accurately describe the observed stem dynamics in transgenic lineage tracing experiments (Clayton *et al.*, 2007; Doupé *et al.*, 2010, 2012; Lim *et al.*, 2013). Using three parameters, the division rate, the stratification rate, and the probability of symmetric division, this model predicts average clone size, clone size distributions, tissue homeostasis, and cell survival probabilities. Whilst simulation based techniques have been used to fit experimental data, an analytical solution has also been described (Antal and Krapivsky, 2010) allowing for maximum likelihood calculations. Model parameterization and accurate representation of uncertainty however remains problematic, particularly in light of short-term data from in vivo histone dilution assays and live imaging that show cell cycle distribution times do not follow the exponential distributions assumed in the analytical model (Murai *et al*, 2018).

Here we report the results of exploring alternative approaches applied to the analysis of both published and synthetic datasets. We find that simulation-based maximum likelihood methods require extensive sampling to find a distribution of parameters making them intractable for many analyses. We further find that use of the published analytical solution allows identification of a single parameter with narrow confidence intervals. However, the analysis of synthetic datasets with realistic cell-cycle distribution times and biological variation between samples suggest that these intervals are too narrow. We conclude that an approximate Bayesian computation (ABC) based approach using a non-Markovian simulator gives appropriate error bars at an acceptable computational cost, and present our methods in a python notebook, enabling the easy analysis of newly collected datasets.

2 Methods

The single progenitor model has been described in detail elsewhere (Clayton *et al.*, 2007; Doupé *et al.*, 2010, 2012). Briefly, the progenitor cell compartment is modelled as containing a mixture of stem cells, which go on to divide, and differentiated cells, which go on to stratify into upper layers of the tissue. Cell fate is determined on cell division; a stem cell divides to either give one differentiated and one stem cell daughters (asymmetric division), or either a pair of differentiated cells or a pair of stem cells (symmetric division). The probability of either symmetric division is r, and of asymmetric division is 1-2r. The model has two further parameters, the division rate (λ) and the stratification rate (Γ). The stratification rate can be expressed in terms of the division rate and the fraction of cells in the basal layer that go on to divide (ρ). Therefore, parameterising the model requires specific values to be found for r, λ, and ρ.

All code was implemented in python 3.6, in Jupyter notebooks, using numpy. Sequential Monte Carlo ABC was used through the PyABC library. Code is available in supplementary information.

3 Results

To explore the limitations of different methods for parameterising stem cell clone dynamics, we initially applied maximum likelihood based on simulations and a published analytical solution to the analysis of both experimental (Doupé *et al.*, 2012) and synthetic datasets with exponentially distributed cell cycle time distributions. Both approaches broadly agreed on the estimated parameters for the experimental data. The analytical solution was able to infer the expected parameter values efficiently, producing narrow confidence intervals and a smooth likelihood distribution (Fig 1a). In contrast, maximum likelihood based simulations required a sample size larger than 100,000 simulations per parameter set to produce an appropriate distribution, increasing computational effort substantially (Fig 1b). Strikingly, both experimental and synthetic datasets gave extremely narrow 95 % confidence intervals, suggesting extremely little biological variation, in contrast to estimates of division times where confidence intervals were ~10 % of the mean.

A known issue however in the analytical solution is the assumption of an exponential distribution of stem cell cycle times, which is known to be incorrect based on histone labelling and live imaging experiments (Piedrafita *et al.*, 2019; Rompolas *et al.*, 2016). Furthermore, the original synthetic datasets do not take into account biological variation between mice, and so do not allow us to test whether error estimates are accurate. To test these properties explicitly both features were included in synthetic datasets and rerun.

Application of the analytical engine to the variable synthetic datasets revealed that the estimated likelihoods were both inaccurate, and overly precise. Whilst the introduction of more non-Markovian cell cycle times undermines the maximum likelihood parameters, the underestimation of errors by maximum likelihood methods has been reported as a result of poor sampling (Jain and Wang, 2008; Sugasawa and Noma, 2019), (Fig 1c). Using a simulation based maximum likelihood approach improves the parameter estimation but error estimation, and computational efficiency remain issues.

To address this, a Sequential Monte Carlo ABC (SMC-ABC) approach was applied. This method efficiently and accurately identified the input parameters, producing a smooth distribution with more realistic confidence intervals (Fig 1d). Together, these findings demonstrate that the SMC-ABC inference technique can be considered as the most appropriate method for analyzing lineage tracing datasets.

The analysis of synthetic datasets with introduced noise in their parameters further allowed us to explore the effect of inter-mice variability on the analysis. The likelihood distributions calculated from individual time points vary across time and are insufficient to estimate parameters on their own. When compared with likelihood distributions calculated from experimental measurements, whilst we find that the distributions are generally highly similar the small number of biological replicates at each timepoint (two or three) leaves them prone to distortion by chance parameter combinations. Given that the appearance of individual timepoints has been used to justify the consideration of more complex models of stem cell hierarchy (Mascré *et al.*, 2012), this presents a problem for experimental design. Based on our findings here we suggest that increasing sampling at individual timepoints by three-fold whilst reducing the total number of timepoints by a third would minimize this effect without impacting parameter estimation. This tool enables the analysis of epithelial tissues, but also other systems where cohesive clones are observed.

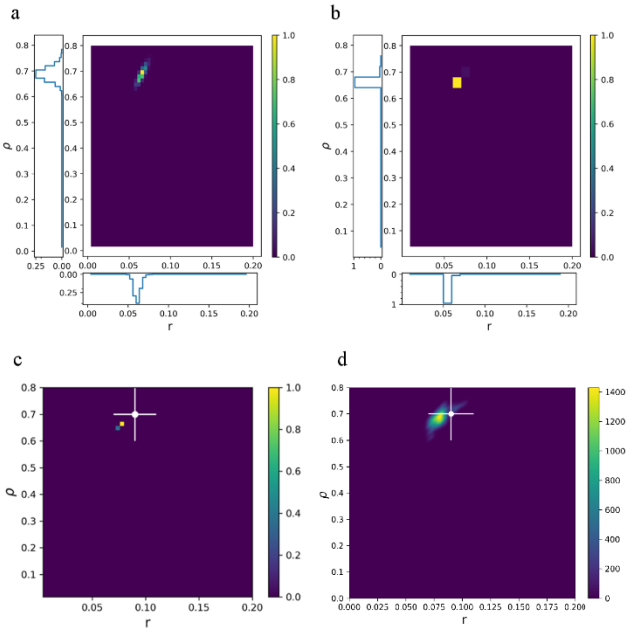

Fig. 1. ABC approaches efficiently and appropriately parameterize the single progenitor model from lineage tracing data. a) A maximum likelihood approach based on the analytical engine, applied to data from Doupe et al, gives a narrow distribution of likelihoods in each parameter. Heatmap shows the log-likelihood distribution, whilst likelihood estimates for each parameter are shown alongside. b) With 100,000 simulations per parameter set, inferred likelihoods from the same dataset are noisy and fail to give smooth distributions in r or ρ. Heatmap shows the log-likelihood distribution, whilst likelihood estimates for each parameter are shown alongside c) Synthetic datasets where realistic cell cycle distribution times and biological variation are modelled show that likelihoods calculated from the analytical engine are overly conservative and inaccurate. Input parameters and error estimates indicated by cross and error bars (2xSD), whilst the heatmap plots likelihood distribution calculated using the analytical solution. d) An ABC approach to inferring parameters gives a smooth distribution and reasonable confidence intervals. Input parameters and error estimates indicated by cross and error bars (2xSD), whilst the heatmap plots KDE of final population of parameter sets.

Acknowledgements

We thank Allon Klein, and the Hall and Jones groups for useful discussions.

Funding

This work has been supported by the Royal Society (URF to BAH grant no. UF130039, studentship to VK). BAH is supported by a Medical Research Council (MRC) Grant-in-Aid to the MRC Cancer unit. M.W.J.H. acknowledges support from the Harrison Watson Fund at Clare College, Cambridge. P.H.J. is supported by a Cancer Research UK Programme grant no. (C609/A17257) and core grants from the Wellcome Trust to the Wellcome Sanger Institute, 098051 and 206194.

Conflict of Interest: none declared.

References

Antal,T. and Krapivsky,P.L. (2010) Exact solution of a two-type branching process: Clone size distribution in cell division kinetics. *J. Stat. Mech. Theory Exp.*, P07028.

Clayton,E. *et al.* (2007) A single type of progenitor cell maintains normal epidermis. *Nature*, **446**, 185–189.

Doupé,D.P. *et al.* (2012a) A single progenitor population switches behavior to maintain and repair esophageal epithelium. *Science*, **337**, 1091–1093.

Doupé,D.P. *et al.* (2012b) A single progenitor population switches behavior to maintain and repair esophageal epithelium. *Science*, **337**, 1091–3.

Doupé,D.P. *et al.* (2010) The Ordered Architecture of Murine Ear Epidermis is Maintained by Progenitor Cells with Random Fate. *Dev. Cell*, **18**, 317–323.

Jain,R.B. and Wang,R.Y. (2008) Limitations of maximum likelihood estimation procedures when a majority of the observations are below the limit of detection. *Anal. Chem.*, **80**, 4767–4772.

Lim,X. *et al.* (2013) Interfollicular epidermal stem cell self-renew cia autocrine Wnt signaling. *Science (80-.).*, **342**, 1226–1230.

Mascré,G. *et al.* (2012) Distinct contribution of stem and progenitor cells to epidermal maintenance. *Nature*, **489**, 257–262.

Murai,K *et al* (2018) Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. *Cell Stem Cell*, **23(5),** 687-699.

Piedrafita,G. *et al.* (2019) The single-progenitor model as the unifying paradigm of squamous epithelial maintenance. *bioRxiv*.

Rompolas,P. *et al.* (2016) Spatiotemporal coordination of stem cell commitment during epidermal homeostasis. *Science*, **9**, 855–861.

Sugasawa,S. and Noma,H. (2019) A unified method for improved inference in random effects meta-analysis. *Biostatistics*, 1–29.