Modeling Stem Cell Fates Using Non-Markov Processes

Patrick S. Stumpf[1], Fumio Arai[2], Ben D. MacArthur[3,4,5,6,*]

[1] Joint Research Center for Computational Biomedicine, RWTH Aachen University, Aachen, 52074, Germany
[2] Department of Stem Cell Biology and Medicine, Graduate School of Medical Sciences, Kyushu University, 812-8582, Fukuoka, Japan
[3] Centre for Human Development, Stem Cells and Regeneration, University of Southampton, SO17 1BJ, United Kingdom
[4] Mathematical Sciences, University of Southampton, SO17 1BJ, United Kingdom
[5] Institute for Life Sciences, University of Southampton, SO17 1BJ, United Kingdom
[6] The Alan Turing Institute, London, NW1 2DB, United Kingdom
[*] Correspondence to bdm@soton.ac.uk

Epigenetic memories play an important part in regulating stem cell identities. Tools from the theory of non-Markov processes may help us understand these memories better and develop a more integrated view of stem cell fate and function.

It is becoming increasingly clear that epigenetic "memories" have a central role in regulating individual stem cell fates, and variations in individual cell histories can generate functional heterogeneity within apparently pure stem cell populations (Graf and Stadtfeld, 2008; Yu et al., 2016). These results suggest that there are deep connections between cell histories and cell fates, yet our understanding of these connections is currently incomplete.

In parallel to the experimental advances that are allowing us to probe these connections more deeply (Zhu et al., 2020), there has been progress in the mathematical and physical sciences on the theory of non-Markov stochastic processes. These developments are helping to provide a formal framework to understand how memories can be encoded and propagated in complex dynamical systems that have direct implications for our understanding of the relationship between stem cell histories and fates.

Stochastic processes are often used to model the dynamics of systems for which outcomes are not entirely predictable. A stochastic process for which the present state of the system provides all the information needed to determine the likelihood of future events is said to be Markov (Van Kampen, 1992). Markov processes are used to model dynamics in which history is not important and they are often described as being "memoryless" (see **Fig. 1A**). Markov processes are widely used to model stochastic processes in cell biology, and underpin some important modern data analysis techniques, such as trajectory inference methods for single cell data (Weinreb et al., 2018).

However, Markov processes are not appropriate in all circumstances. In some situations, the future of a system may depend on both its present state and on the particular path taken to get to the present state. Stochastic processes that account for such history are said to be non-Markov.

Non-Markov processes typically arise for one of two reasons.

First, there are features of the system being studied that are unobserved yet have an important effect on the observable dynamics. For example, there may be a key gene or molecular mechanism that directs a stem cell lineage choice that is not experimentally measured, yet has an important effect on the expression of the genes or mechanisms that are observed. In this case, the assessment of the "present state" of the cell does not include all relevant information needed to infer its future behaviour and the dynamics may be best modelled as a non-Markov process. This issue is particularly pertinent when considering regulation of cell fates by cis epigenetic signals (i.e., stable alterations to the DNA, such as methylation, acetylation, etc.). For instance, complex patterns of epigenetic regulators encode cell histories and play an important part in cell fate regulation (Yu et al., 2016), yet these patterns may not be easy to determine at the single cell level in their entirety (Zhu et al., 2020) and so may give rise to apparently unpredictable dynamics at the single cell level and unexplained dynamic heterogeneity at the population level.

Second, there is some persistence in the system. Unless our definition of the present state takes account of this persistence, then the dynamics may again be non-Markov (see **Fig. 1B**). Stem cell differentiation is a relevant example. To illustrate this, consider the expression of a hypothetical gene that acts as a marker of a transition between two cell fates A and B (see **Fig. 1C**). Suppose that this gene is highly expressed when the cell is in state A and lowly expressed when the cell is in state B. If the cell is in transition from state A to state B, then the expression of this gene will generally decrease over time. If, by contrast, the cell is in transition from state B to state A, then the expression of this gene will generally increase over time. In this case, knowledge of the cell's instantaneous internal molecular state (i.e., the expression of the gene), is not enough to infer its likely future because it does not contain information of whether the cell is moving from state A to state B or vice versa. More generally, stem cell differentiation involves movement from one functional state to another along a directed developmental trajectory in a high dimensional expression space, characterized by coordinated changes in gene expression patterns that occur in a particular order. Differentiation dynamics are therefore persistent and may be best modelled as a non-Markov process. This issue is particularly pertinent when considering trans epigenetic signals (i.e., molecular alterations that are stabilized by the dynamics of intracellular regulatory networks), which may alter on environmental stimulus and thereby confer a preferred direction to differentiation that is not directly encoded in the cell's instantaneous state.

Collectively, these considerations should cause us to pause and consider what we can, and cannot, determine from an experiment – and how we can better design experiments. For example, remarkable advances in modern single cell profiling techniques that allow us to explore cellular identities in exquisite detail do not, alone, resolve these issues when considering dynamics, such as differentiation, that are out of equilibrium. Computational trajectory inference methods, which aim to infer local velocities via pseudotemporal ordering of multiple single cell expression profiles are also needed. Similarly, methods that seek to infer mRNA expression velocities for individual cells from snap-shot data, for instance by comparing the abundance of spliced and unspliced mRNA or through metabolic labelling of nascent RNA, are important.

Yet, powerful as many of these computational methods are, they are subject to some limitations. Typically, they are presently developed for one modality – usually single cell RNA sequencing data since this is the dominant current methodology – and make an implicit assumption that effects of unmeasured "hidden variables", such as chromatin state, and expression of proteomic, metabolic or epigenetic factors, etc., are negligible and dynamics along an inferred trajectory are therefore Markov (Weinreb et al., 2018). By doing so they are potentially able to detect persistence in expression dynamics due to trans mechanisms but are less well-equipped to address the effects of hidden cis mechanisms on cell fate dynamics. Trajectory inference methods that merge data from multiple modalities are therefore needed, and some work in this area is now emerging (Chen et al., 2019). A strong theoretical basis in the theory of non-Markov processes could help develop the next generation of such methods.

Integrating such theory with experiment is hard and better methods are needed to do this. Nevertheless, recent years have seen some progress (Armond et al., 2014; Stumpf et al., 2017; Zhang and Zhou, 2019) and some simple heuristics may help. For example, so-called wait-times between events are of central importance in stochastic analysis (Van Kampen, 1992). The extent to which an observed wait-time distribution deviates from the exponential distribution expected for Markov processes is a straightforward, experimentally obtainable, indication that important information is being missed. As a simple example, cell cycle time distributions show strong deviation from exponential, and analysis of cycle time distributions can accordingly be used to infer the presence of unobserved intermediate stages. Similarly, analysis of the distribution of exit times from pluripotency has been used to reveal the presence of hidden metastable states in embryonic stem cell differentiation trajectories (Stumpf et al., 2017). These results prompt two general observations: (1) theoretical considerations can help guide the design of experiments and maximize the information obtained from complex data sets; (2) to make better use of theory we need to move toward collecting and analysing properties of distributions (e.g. of cell cycle/exit times) and comparing experimentally observed distributions with those expected from theory, rather than anchoring our analysis on statistical comparison of distributional moments, such as the mean and variance. To do so will require collection of fine-grained data on the dynamics of large numbers of individual cells, for example using advances in live cell labelling and continuous imaging strategies or observation of cellular genealogies via genetic lineage tracing.

Such notions may be particularly useful in understanding the origins and functional consequences of "heterogeneity" in stem cell populations.

It has been widely observed that apparently pure stem cell populations can, in fact, be highly heterogeneous in their molecular expression patterns (Graf and Stadtfeld, 2008). This variability is thought to play an important part in regulating stem cell population function and has numerous genetic and epigenetic origins that we do not yet fully understand. One intriguing possibility is that mitotic histories may have a central role (Bernitz et al., 2016).

Consider a population of stem cells proliferating under homeostatic conditions in vivo. If individual stem cells in the population divide in a temporally stochastic, uncoordinated (i.e., unsynchronized) way, then proliferation will naturally generate a heterogeneous, age-structured population (illustrated schematically in **Fig. 2A-B**). In principle this age-structure need not be of functional importance. However, if divisional history confers bias to individual stem cells, then the potency of the population as a whole will depend on the collective dynamics of an inherently heterogeneous mix of cells, each with different innate regenerative abilities. There is developing evidence that this is indeed the case. For example, the accumulation of cell divisions is directly associated with loss of hematopoietic stem cell potency both during native haematopoiesis and under conditions of stress (Bernitz et al., 2016).

This reasoning suggests that population heterogeneity and individual cell histories may be intertwined. This is an interesting hypothesis, yet because it implies that the regenerative potential of the population depends on the entire mitotic history of each of its constituent cells, it is hard to experimentally explore. However, some theoretical notions may again help. In particular, if mitotic history (or indeed any other functionally important aspect of cellular history) is not experimentally observed, then an important part of the stem cell identity will remain hidden, and proliferation may be best modelled as a non-Markov process. In situations such as this, hidden Markov models – which are widely used in the physical sciences to model processes for which the "true" dynamics cannot be directly or fully monitored – provide a powerful way to infer the presence of regulatory mechanisms that cannot be observed from the dynamics that can be observed, and so can be used to interpret sparse data (see **Fig. 2C**). Higher-order Markov models – which assume that the future of a system depends on its present state and immediate, but not distant, past, and therefore allow for "short-term memory" – may be similarly used to formally encode hypothesised memory mechanisms in explanatory models. Moreover, because both hidden and higher-order Markov models allow putative candidate mechanisms to be compared with each other, they can be used to weigh the empirical evidence for competing hypotheses and so provide biological clarity – for example, by ruling out some candidate mechanisms. Substantial benefit can therefore be gained by taking advantage of theory to interpret experiment and make better use of data that can be collected. Indeed, we propose that these considerations highlight a general principle of widespread importance: experimental advances are not enough; we also need better methods and models to extract biological information from the data we collect.

Collectively, these considerations indicate that relationships between cellular histories, memories and fates are intrinsically complex, but they are not impenetrable. They also raise numerous fundamental questions that may help guide future work in this area. For instance: Do different epigenetic regulatory mechanisms leave different characteristic “signatures” in observable dynamics that can be dissected by appropriate non-Markov, hidden Markov or higher-order Markov models? How can mathematical models be developed to account for complex cis regulatory mechanisms that may occur stochastically and independently at numerous different loci and persist over time? How can advances in multimodal live cell tracking and cell linage tracing be best combined with mathematical models to decipher the formation and propagation of memories at the single cell level? To approach these most challenging and interesting problems will require that we foster new ways of working in which theoretical and experimental methods are developed concurrently and guide each other. Doing so may help us develop a more integrated perspective of epigenetic memories and their effects on stem cell identities.

Figure Captions

Figure 1. Markov and non-Markov processes. (**A)** The unbiased random walk is a simple example of a Markov process. In the unbiased random walk, a walker moves up and down on a one-dimensional domain. At each time-step the walker moves up with probability 0.5 and down with probability 0.5. Three simulations are shown. (**B)** The persistent random walk is a simple example of a non-Markov process. In the persistent random walk, a walker moves up and down on a one-dimensional domain. At each time-step the walker persists in the direction it is currently going with probability p and changes direction with probability 1-p. In this case, transition probabilities depend explicitly on both the current and previous position of the walker, and the dynamics are accordingly non-Markov (the persistent random walk is a second-order Markov process). Three simulations are shown. (**C)** Transitions between cell fates A and B, assessed by the expression of a hypothetical marker gene. A transition from fate A to B is in red; a transition from fate B to A is in blue. The current state of the cell (i.e., expression of the hypothetical gene in the cell) is not enough to determine its future, since it depends on which way the fate transition is occurring. The dotted pink line shows a set expression level encountered in both transitions. If a transition from fate A to B is occurring then it is likely that the expression level will decrease, while if a transition from fate B to A is occurring, it is likely that it will increase. This persistence, which may in turn relate to the expression dynamics of a myriad of unobserved genes, means that the dynamics may be best modelled as a non-Markov process.

Figure 2: Stem cell proliferation as a non-Markov process. (**A**) Proliferation in a homogeneous stem cell population. If all stem cells are equivalent in their self-renewal ability, then proliferation may be modelled as a Markov process. (**B**) Proliferation in a heterogeneous population. If stem cells are distinguished in their self-renewal ability by their mitotic history and cell divisions occur asynchronously then the stem cell pool will become inherently heterogeneous. If mitotic history is not experimentally observed, then proliferation may be best modelled as a non-Markov process. (**C**) Hypotheses concerning unobserved “hidden” mechanisms can be tested against experimentally observed dynamics using hidden Markov models. In this hypothetical example, stem cell proliferation in vivo is monitored over time (left panel), but the divisional history of each cell is not observed. As an illustration, it is hypothesised that stem cells divide four times before entering a dormant state (second panel; (Bernitz et al., 2016)). This hypothesis can be encoded in a mathematical model and compared with experimentally observed cell numbers (third panel; model is in blue). Based on model fit to

observed dynamics, the number of cells that have divided 0, 1, 2, 3 and 4 times can be inferred (right panel). Although apparently successful, this may not be the only model that explains the experimentally observed data. If alternative models of proliferation also explain the data, then model selection tools can be used to weigh the empirical evidence for competing hypotheses, suggest those that are compatible for further experimental investigation, and exclude those that lack experimental support.

References

Armond, J.W., Saha, K., Rana, A.A., Oates, C.J., Jaenisch, R., Nicodemi, M., and Mukherjee, S. (2014). A stochastic model dissects cell states in biological transition processes. Sci. Rep. *4*, 1–9.

Bernitz, J.M., Kim, H.S., MacArthur, B., Sieburg, H., and Moore, K. (2016). Hematopoietic Stem Cells Count and Remember Self-Renewal Divisions. Cell *167*, 1296-1309.e10.

Chen, S., Lake, B.B., and Zhang, K. (2019). High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. Nat. Biotechnol. *37*, 1452–1457.

Graf, T., and Stadtfeld, M. (2008). Heterogeneity of Embryonic and Adult Stem Cells. Cell Stem Cell *3*, 480–483.

Van Kampen, N.G. (1992). Stochastic Processes in Physics and Chemistry (Elsevier Science).

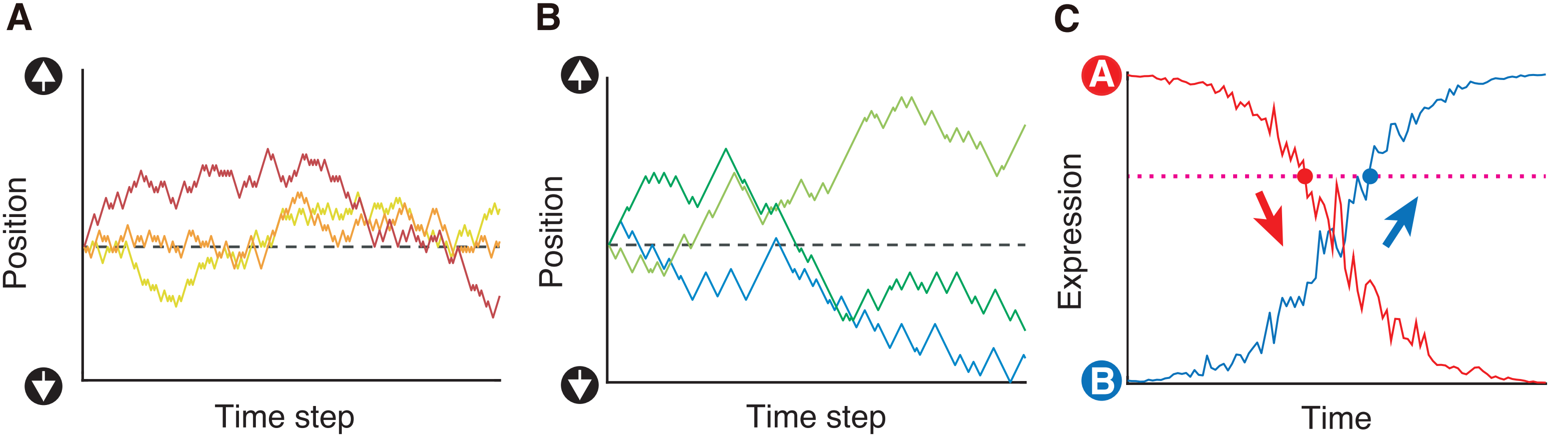
Stumpf, P.S., Smith, R.C.G., Lenz, M., Schuppert, A., Müller, F.-J., Babtie, A., Chan, T.E., Stumpf, M.P.H., Please, C.P., Howison, S.D., et al. (2017). Stem Cell Differentiation as a Non-Markov Stochastic Process. Cell Syst. *5*, 268-282.e7.

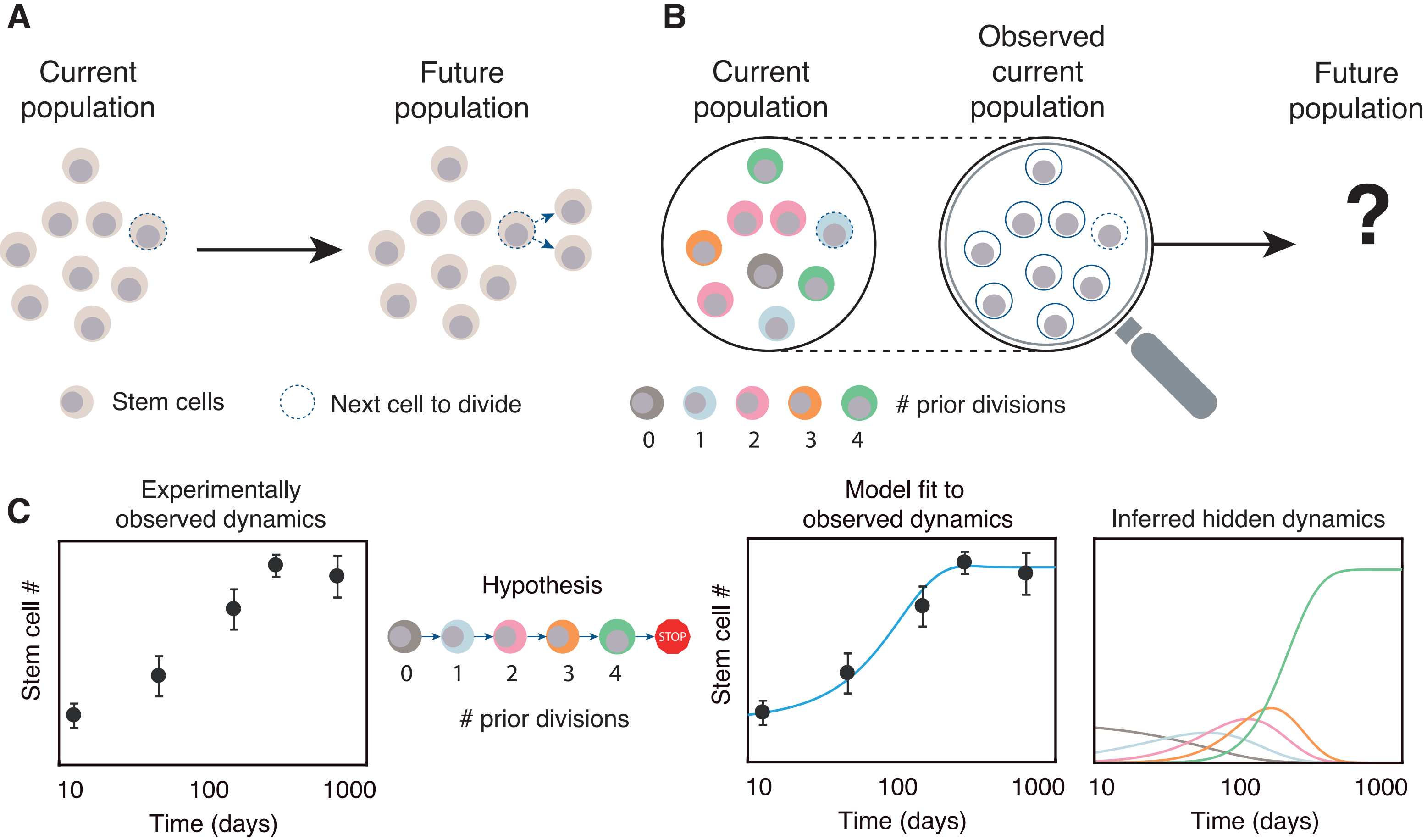
Weinreb, C., Wolock, S., Tusi, B.K., Socolovsky, M., and Klein, A.M. (2018). Fundamental limits on dynamic inference from single-cell snapshots. Proc. Natl. Acad. Sci. U. S. A. *115*, E2467–E2476.

Yu, V.W.C., Yusuf, R.Z., Oki, T., Wu, J., Saez, B., Wang, X., Cook, C., Baryawno, N., Ziller, M.J., Lee, E., et al. (2016). Epigenetic Memory Underlies Cell-Autonomous Heterogeneous Behavior of Hematopoietic Stem Cells. Cell *167*, 1310-1322.e17.

Zhang, J., and Zhou, T. (2019). Markovian approaches to modeling intracellular reaction processes with molecular memory. Proc. Natl. Acad. Sci. U. S. A. *116*, 23542–23550.

Zhu, C., Preissl, S., and Ren, B. (2020). Single-cell multimodal omics: the power of many. Nat. Methods *17*, 11–14.


A
Position
Time step
B
Position
Time step
C
A
Expression
B
Time


A
Current population
Future population
Stem cells
Next cell to divide
B
Current population
Observed current population
Future population
?
# prior divisions
0
1
2
3
4
C
Experimentally observed dynamics
Stem cell #
10
100
1000
Time (days)
Hypothesis
0
1
2
3
4
STOP
# prior divisions
Model fit to observed dynamics
Stem cell #
10
100
1000
Time (days)
Inferred hidden dynamics
10
100
1000
Time (days)