Bioinformatics, 36(9), 2020, 2813–2820
doi: 10.1093/bioinformatics/btaa021
Advance Access Publication Date: 23 January 2020
Original Paper

Systems biology

TRANSCOMPP: understanding phenotypic plasticity by estimating Markov transition rates for cell state transitions

N. Suhas Jagannathan [1,†], Mario O. Ihsan[2,3,†], Xiao Xuan Kin[2], Roy E. Welsch[4], Marie-Véronique Clément [2,*] and Lisa Tucker-Kellogg [1,*]

[1]Cancer and Stem Cell Biology Programme, Centre for Computational Biology, Duke-NUS Medical School, 169857 Singapore, [2]Department of Biochemistry, National University of Singapore, 117596 Singapore, [3]NUS Graduate School for Integrative Sciences and Engineering, National University of Singapore, 117456 Singapore and [4]Sloan School of Management and Center for Statistics and Data Science, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Jonathan Wren

Received on May 6, 2019; revised on December 10, 2019; editorial decision on January 7, 2020; accepted on January 17, 2020

Abstract

Motivation: Gradual population-level changes in tissues can be driven by stochastic plasticity, meaning rare stochastic transitions of single-cell phenotype. Quantifying the rates of these stochastic transitions requires time-intensive experiments, and analysis is generally confounded by simultaneous bidirectional transitions and asymmetric proliferation kinetics. To quantify cellular plasticity, we developed TRANSCOMPP (*T*ransition *R*ate *AN*alysis of *S*ingle *C*ells to *O*bserve and *M*easure *P*henotypic *P*lasticity), a Markov modeling algorithm that uses optimization and resampling to compute best-fit rates and statistical intervals for stochastic cell-state transitions.

Results: We applied TRANSCOMPP to time-series datasets in which purified subpopulations of stem-like or non-stem cancer cells were exposed to various cell culture environments, and allowed to re-equilibrate spontaneously over time. Results revealed that commonly used cell culture reagents hydrocortisone and cholera toxin shifted the cell population equilibrium toward stem-like or non-stem states, respectively, in the basal-like breast cancer cell line MCF10CA1a. In addition, applying TRANSCOMPP to patient-derived cells showed that transition rates computed from short-term experiments could predict long-term trajectories and equilibrium convergence of the cultured cell population.

Availability and implementation: Freely available for download at http://github.com/nsuhasj/Transcompp.

Contact: Lisa.Tucker-kellogg@duke-nus.edu.sg or bchmvc@nus.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Stochastic transitions allow cells of the same genotype to spontaneously switch between different phenotypic states in response to internal or external cues (Reyes and Lahav, 2018). The molecular mechanisms of stochasticity are often unknown but may include expression differences, epigenetic regulation or macromolecular changes. Stochastic transitions underlie temporal changes in many biological contexts including tissue regeneration, embryonic development, gene regulatory networks, epigenetic transformations (Armond *et al.*, 2014; Flöttmann *et al.*, 2012) and cancer plasticity (Dingli and Pacheco, 2011; Hoek and Goding, 2010).

In cancer, multiple studies have documented stochastic re-equilibration, a phenomenon where one phenotype purified from a heterogeneous population (such as stem-like or differentiated cells) can spontaneously give rise to a heterogeneous population resembling the original (Gupta *et al.*, 2011; Leong *et al.*, 2014). Such studies warn of the ability of tumors to recapitulate their cellular composition post-treatment and suggest a basis for the development of drug resistance in tumors (Chisholm *et al.*, 2016; Emmons *et al.*, 2016; Kemper *et al.*, 2014; Pisco and Huang, 2015). While transitions from stem cells to differentiated cells are unsurprising, such studies also showed that in many cancer types, some degree of de-differentiation (differentiated-to-stem transitions) may be occurring

spontaneously (Chaffer *et al.*, 2011), albeit infrequently. Even very infrequent transitions can have a significant impact on a system when they replenish an influential or proliferative subpopulation. Hence, to understand the inherent plasticity of any system, it is important to perform precise and sensitive quantification of stochastic state transition rates, even when the underlying transition rates are small.

Experimental quantification of stochastic state-transition rates is difficult for multiple reasons. Time-series experiments measure only snapshots of phenotypic abundance, and not transition rates. The analysis of such time-series experiments is further confounded by variables such as different proliferation rates for each phenotype (Supplementary Material S1). Also difficult is the deconvolution of observed bulk changes to obtain individual contributions of stochastic transitions in each direction. Computational modeling is ideally suited to handle such issues (Beerenwinkel *et al.*, 2015), and with sufficient experimental data, can estimate interdependent parameters simultaneously. The resulting transition rates then reveal the relative contributions of individual transitions to overall dynamics. The transition rates can also be used to simulate and predict long-term dynamics from short-term observations, with or without perturbations (e.g. phenotype depletion/enrichment, transition suppression/amplification). Since Markov state-transition models are particularly well suited to the analysis of stochastic transitions and the concept of re-equilibration (stationary distributions), we will use Markov modeling to estimate the transition rates in this study.

Previous work has successfully used Markov state-transition models (Fig. 1A) to compute transition rates between cell phenotypes in specific contexts (Buder *et al.*, 2017; Gupta *et al.*, 2011). However, their methods were specific to particular applications, without being generalizable to different contexts and experimental designs. For example, CellTrans (Buder *et al.*, 2017) has some practical limitations, such as inability to model variable proliferation rates, and requiring an invertible input matrix (meaning that it requires the number of replicates to be exactly equal to the number of phenotypes). Hence, there is an unmet need to develop a general-purpose tool to compute transition rate parameters from time-series measurements of single-cell phenotypes. The requirements of the proposed tool are as follows: to estimate transition rates that best agree with the collected time-series phenotypic snapshots (across any number of replicates), to measure uncertainty (noise) of the computed rates, to account for confounding factors, such as phenotype-specific differences in proliferation rates, and to analyze datasets with flexible configurations of the initial population compositions.

We have developed TRANSCOMPP (*T*ransition *R*ate *AN*alysis of *S*ingle *C*ells to *O*bserve and *M*easure *P*henotypic *P*lasticity), a novel computational tool to quantify transition rates and associated rate distributions, using data from time-series single-cell experiments. Experiments studying single-cell characteristics (flow cytometry, scRNA-seq, lineage tracing, etc.) can be used to obtain fractional compositions (fraction of total population) of different phenotypic states at multiple time-points. TRANSCOMPP uses these fractional compositions to estimate transition rates that best fit the observed data, with or without user-imposed constraints on transition rates. Proliferation rates of each phenotype can be known or partially bounded by experiments (for use as input), or completely unknown (simultaneously fit with transition rates). A resampling module (using the single-cell phenotype measurements) then quantifies the uncertainty interval associated with each computed rate of transition. Compared to previously published methods, such as CellTrans (Buder *et al.*, 2017), TRANSCOMPP fulfills the requirements above while providing robust performance and increased accuracy with increasing problem complexity (Supplementary Materials S2 and S3).

We applied TRANSCOMPP to the basal-like breast cancer cell line MCF10CA1a, to assess the effect of different cell culture media supplements on the plasticity between stem-like and non-stem cancer cells (Santner *et al.*, 2001). Previous studies have identified cancer stem cells by virtue of the expression of the cell surface marker profile $CD44^{high}$ $CD24^{low}$ (Al-Hajj *et al.*, 2003; Bhat-Nakshatri *et al.*, 2010; Nakshatri *et al.*, 2009). Hence, we studied the dynamics of CD24 expression in the $CD44^{pos}$ MCF10CA1a cell line *in vitro*, by growing cells in either minimal basal medium (BM), or when supplemented with a cocktail of factors [insulin (INS), epidermal growth factor (EGF), hydrocortisone (HC), cholera toxin (CTX)] taken individually, or all together (complete medium, CM). Using the TRANSCOMPP-computed transition rates (and intervals), we identified novel regulators of plasticity (HC and CTX), whose effects were found to be greater than known modulators of plasticity, such as INS. HC (an analog of the stress hormone cortisol) enriched the $CD24^{neg}$ fraction and hence promoted pro-stem behavior. In contrast, CTX (an agonist of the PKA pathway) enriched the non-stem fraction ($CD24^{pos}$).

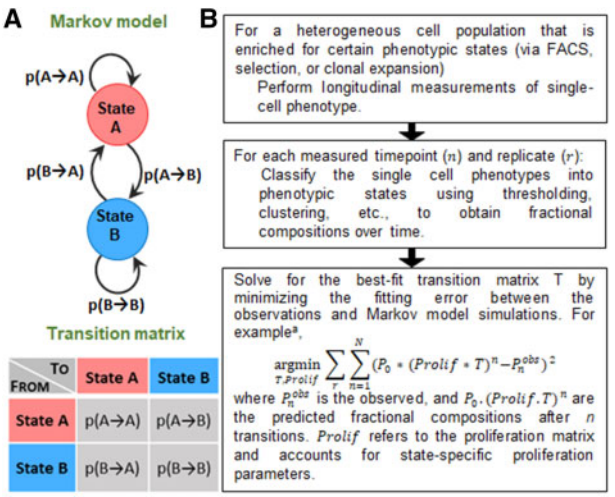

Fig. 1. Markov modeling framework for computing stochastic transition rates. (**A**) A sample Markov model showing two phenotypic states, A (red) and B (blue), with arrows showing transitions between the states. Arrow labels indicate the transition probabilities (rates) which we will compute using our method. Also shown is a representation of the model as a transition matrix, whose elements are the four transition rates. (**B**) The TRANSCOMPP flowchart for computing **T**, the best-fit transition matrix, using optimization based on multiple experimental replicates. r, replicates; *n*, experimental time-points monitored; P_0, initial populations; *Prolif*, proliferation matrix (see Section 2). [a]The choice of error metric affects the minimization. Shown is sum-of-squared residuals error metric. For other error metrics, see Supplementary Material S4

Because studies of gradual plasticity might require costly long-term observations, our final question was whether realistic experimental noise (from clinical samples) combined with TRANSCOMPP quantification would suffice to allow long-term plasticity to be studied using short-term experiments. Applying TRANSCOMPP to a previously published patient dataset (Leong *et al.*, 2014) revealed that the long-term population convergence and equilibrium between $ALDEFLUOR^{hi}$ (stem-like) and $ALDEFLUOR^{lo}$ (non-stem) tumor cell subpopulations could be accurately predicted using short-term observations.

2 Materials and methods

2.1 The TRANSCOMPP algorithm

Denote the set of K distinct phenotypic cell states by the set $\mathbf{S} = \{S_1, S_2 \ldots S_k\}$. The transition rates can be represented as a $K \times K$ transition matrix **T** such that the element **T**(a, b) of the transition matrix ($1 \leq a, b \leq K$) gives the probability that a cell in state S_a would transition to state S_b in unit time. Let the fractional compositions (relative abundance of each phenotype in the population) be given by the K-vector $\mathbf{F} = (f_1, f_2 \ldots f_K)$ such that $\sum_{i=1}^{K} f_i = 1$. Single-cell measurements of phenotypic states (such as flow cytometry) are performed at N time-points $t_0, t_1 \ldots t_{N-1}$, where time-point t_0 is the initial population. If the time-series experimental measures were generated

by an underlying Markov process with transition rates **T**, then in theory (in the absence of noise),

$$F_j = F_0 * \mathbf{T}^j \quad (1)$$

where, F_0 and F_j are the fractional compositions of the K states after 0 (initial) or j time intervals of arbitrary unit time ω, respectively. The time-step ω (hereafter called the iteration interval) for the transition rate units (e.g. day^{-1}) is chosen such that all experimental measurements are performed after integer numbers of ω. There is no requirement that the measurements be made at equal intervals of time. For any experimental time-point t_n, the number of elapsed iteration intervals until that measurement is then given by $(t_n - t_0)/\omega$.

A *simple trajectory* refers to a single replicate of a time-series study. At any given time-point t_n, the *population snapshot* P_n represents the measured fractional compositions of all states across each of R *simple trajectories*, and is given by $P_n = [\mathbf{F}^1(n), \mathbf{F}^2(n) \ldots \mathbf{F}^R(n)]'$. The *population trajectory* can then be defined as a series of *population snapshots* from all measured time-points $t_0, t_1 \ldots t_{N-1}$, given by $(P_0, P_1 \ldots P_{N-1})$.

For a given transition matrix **T** and initial population conditions, Markov modeling can be used to predict the estimated *population snapshot* at any time-point t_n, $(n > 0)$ using Equation (1) as,

$$P_n^{pred} = P_0 * \mathbf{T}^{(t_n - t_0)/\omega} \quad (2)$$

where P_0 is the *population snapshot* at time t_0. When the population dynamics are also affected by phenotype-specific proliferation rates, we convert the population fractions into population abundance and apply a discrete proliferation process at each time-step. The proliferation process is encoded by a size-K diagonal matrix called the *Prolif* matrix, whose elements are a function of the relative proliferation rate p_k of each phenotype. For each phenotypic state k, p_k represents the rate of division of cells in state k, relative to the rate of division in an arbitrarily chosen reference phenotype P1. For a phenotype that on average doubles twice as fast as the reference phenotype, $p_k = 2$. The elements of the *Prolif* matrix represent the factor of increase in the abundance of each phenotype, in one *relative iteration interval* (τ), and are given by $Prolif(k, k) = 2^{p_k \tau}$. τ is computed as the arbitrary time-unit ω (iteration interval in real time, e.g. 1 day) divided by the average doubling time of P1 (in real time). The reference phenotype, generally the first phenotype P1, has $p_1 = 1$, and hence, $Prolif(1,1) = 2^{\tau}$. The predicted *cellular abundance* (measure of number of cells of each phenotype) at any time-point t_n is, therefore,

$$A_n^{pred} = A_0 * [Prolif * \mathbf{T}]^{(t_n - t_0)/\omega} \quad (3)$$

where A_0 is the *cellular abundance* of each state initially and A_n^{pred} is the predicted *cellular abundance* of each state at t_n. The fractional compositions of the states at each time-point can then be obtained by normalizing the *cellular abundances* of all phenotypes such that they sum to 1.

TRANSCOMPP optimizes the transition matrix **T** and, optionally, the proliferation rate matrix *Prolif*, in order to minimize the cumulative error-of-fit between observed and predicted *population snapshots* across all measured time-points and replicates (Fig. 1B). The error-of-fit is dependent on the type of error measure. Shown in Equation (4) is the optimization performed using SSR (sum of squared residuals) as the measure. See Supplementary Material S4 for two other measures: least trimmed squares and L1 norm.

$$\underset{\mathbf{T},\ Prolif}{\operatorname{argmin}} \sum_{n=1}^{N-1} \left(P_0 * [Prolif * \mathbf{T}]^{(t_n - t_0)/\omega} - P_n^{obs} \right)^2 \quad (4)$$

Note that by definition, *population snapshots* include data from R simple trajectories (replicates) and hence the overall minimization is performed over all replicates and time-points. TRANSCOMPP implements the minimization using the optimization toolbox of MATLAB (Mathworks). The above minimization is repeated from a random initial seed until convergence, for a user-specified number of times (default = 50). The optimization can include upper- and lower-bound constraints to ensure that the transition rates favor specific directions (e.g. self-transitions) and that the proliferation rates for phenotypes are within a range of fold-changes from each other.

2.2 Stochastic resampling

To compute uncertainty measures for the best-fit transition rates, we obtain a probability distribution for each transition rate, by repeatedly generating *pseudo population trajectories* (*pPT*) from the original data, and computing the best-fit transition rates for each pseudo-trajectory. Each *pPT* is equivalent to a single hypothetical *simple trajectory* (replicate) composed of *pseudo population snapshots* (*pPS*) $(P_0, P_1 \ldots P_{N-1})$ derived from the original data. When the input data includes single-cell measures that allow phenotype classification (flow cytometry, scRNA-seq, etc.), the *pPS* for each time-point t_n is calculated as follows:

Pick a predetermined number μ (default = 100) of single-cell measurements across all R replicates, with each replicate assigned an equal probability of being chosen. Using these measurements, bin the chosen μ cells into one of the K phenotypic states through user-defined thresholds or clustering algorithms. Then compute F_n^{pseudo}, the fraction of μ cells in each phenotypic state.

When the input data do not contain single-cell measures of state but only fractional compositions, the *pPS* is computed using techniques explained in Supplementary Material S5. The choice of μ (the sampling breadth) affects how the sampled *pPS* will be distributed around the mean experimental population snapshots. Higher values of μ result in greater convergence toward the mean experimental distribution (Supplementary Material S6).

To compute the best fit transition matrix **T** for each *pPT*, we perform the minimization from Equation (4). If the experiments were originally performed starting with k enriched or purified states, then the *pPT* will also be constructed with enrichment of the same states, and contain k rows of F_n^{pseudo}. Each row of this *pPT* corresponds to a *simple trajectory* of hypothetical enrichment of a different phenotypic state. This ensures that the computed **T** for any one *pPT* is not biased by unbalanced initial configurations. We repeat the generation of *pPT* (and solving for **T**) B times (default $B = 1000$), to obtain a frequency distribution for each transition rate.

2.3 Analytical flow cytometry

MCF10CA1a cells at 70–80% confluency were harvested with StemPro® Accutase® Cell Dissociation Reagent (Life Technologies). Collected cells were re-suspended in wash buffer (serum-free DMEM/F12 w/o phenol red, 0.5% Bovine Serum Albumin, 2 mM Ethylenediaminetetraacetic acid) and stained with fluorophore-conjugated antibodies for CD24 (BD Biosciences, San Jose, CA, USA) for 30 min at 4°C in the dark. For each sample, fluorescence intensity (>10 000 cells) was detected with a BD LSRFortressaTM cell analyzer (BD Biosciences). Intensity thresholds for $CD24^{neg/pos}$ were established at the 99th percentile of unstained cells.

2.4 Further experimental and statistical methods

See Supplementary Material S7 for further experimental and statistical methods.

3 Results

3.1 Development of the TRANSCOMPP algorithm

We developed TRANSCOMPP for computing best-fit transition rates for Markov models. The TRANSCOMPP algorithm uses a discrete-time Markov model to capture state-change dynamics of cell phenotypes. A Markov model consists of a set of discrete states (phenotypic states) each of which can transition to other states at a fixed probability per time-step (transition rate) (Fig. 1A). Input to TRANSCOMPP is a time-series dataset of single-cell measures of cell phenotype (e.g. biomarker expression), or bulk statistics showing the relative abundance of each phenotype in the cell population (fractional compositions). TRANSCOMPP minimizes an objective function (a measure of

distance between modeled and observed populations) to solve for transition rates that are the minimum-error fit across all replicates and time-points (Fig. 1B). In addition to computing transition rates, TRANSCOMPP can also account for phenotype-specific proliferation parameters (when they are known or solve for them when they are unknown) and compute uncertainty estimates for the computed transition rates.

3.2 Benchmarking the performance of TRANSCOMPP

To assess the accuracy of using a simple proliferation matrix at discrete time-steps to approximate continuous proliferation dynamics under phenotype-specific rates, we developed an agent-based simulation of stochastic single-cell proliferation and phenotypic transition with fine-grained temporal events (Supplementary Material S2). Figure 2A shows the agreement (error-of-fit) between the fractional composition trajectories of a two-state system, generated by the agent-based simulation versus that generated by using a discrete proliferation matrix (Equation 3). The proliferation rate of the phenotype P1 is kept constant at 1.0 (reference phenotype), while the relative proliferation rate of the phenotype P2 is varied in the range 0.1 to 10 (10× slower to 10× faster). The *relative iteration interval* was varied from 0.1 to 2, indicating that pseudo-measurements would be made periodically at time-points ranging from one-tenth to twice the doubling time of P1. From 2A, it can be seen that using the discrete proliferation rate provides a close approximation of a continuous proliferation process, in the range of variability expected to be observed physiologically in proliferation rates (0.5× – 2× of any phenotype).

Using similar agent-based simulations, we generated pseudo-experimental data for different conditions (systems with 2–12 phenotypes, random transition matrices, similar/random variable proliferation rates, purified/unsorted initial conditions, 1–16 time-points of measurement and 1–20 replicates). TRANSCOMPP was then applied to these datasets and benchmarked as follows.

To benchmark runtimes, we applied TRANSCOMPP to datasets of unsorted initial populations of phenotypes with variable proliferation rates (solving for both transition and proliferation rates). Figure 2B shows the runtime of TRANSCOMPP when run in parallel (one dataset per core) on an Intel(R) Xeon(R) CPU E5-2650 machine as a function of the number of phenotypes in the system (top) and the number of replicates in the input data (bottom). Note that each run of TRANSCOMPP to compute a transition matrix used 50 restarts of optimization, to avoid local minima. To benchmark performance accuracy, we checked how well the best-fit transition rates (output by TRANSCOMPP) would recapitulate the nominal transition rates used originally to generate the pseudo-data by the agent-based simulations. Figure 2C shows the agreement between the nominal transition rates and predicted transition rates when TRANSCOMPP was applied to the same datasets as in Figure 2B. On the left are plots of systems with 2, 5, 8 or 11 phenotypes when approximate estimates (± 25%) for the relative proliferation rates of all phenotypes (except P1) are provided. On the right are similar plots corresponding to TRANSCOMPP runs where no information is provided about proliferation rates, and hence TRANSCOMPP simultaneously solves for best-fit transition rates and proliferation rates. It can be seen that even in the latter case, the TRANSCOMPP-estimated rates show good agreement with the nominal rates (lowest correlation = 0.951, for Phe = 11).

To quantify the impact of the amount of input data on transition rate estimation, we applied TRANSCOMPP to the same datasets as in Figure 2B (solving for both transition and proliferation rates) while increasing the number of replicates in the input data. Figure 2D shows the distribution of the root-mean-squared deviations (RMSDs) of transition rates (approximate measure of error in each TRANSCOMPP-predicted rate of the transition matrix), as a function of the number of replicates in the input data. Results showed that the RMSD usually approaches its minimum value within 4–5 replicates. Further benchmarking (scatter plots of all cases from Phe = 2 to Phe = 12, and RMSD values of transition rates accuracy of TRANSCOMPP applied to other types of problems) can be found in Supplementary Material S2.

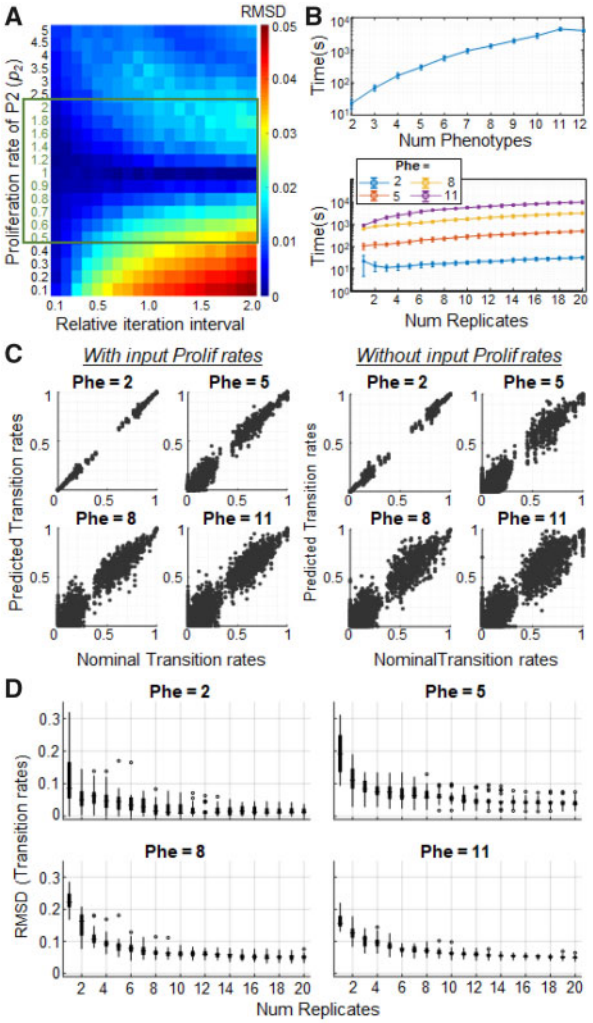

Fig. 2. Benchmarking TRANSCOMPP runtime and accuracy. A spectrum of test problems was created using an agent-based simulation with variable numbers of phenotypic states, variable transition rates and proliferation rates, and variable amounts of data provided as input to TRANSCOMPP. See Supplementary Material S2. TRANSCOMPP was run on each test problem and the accuracy of the model trajectory, the accuracy of the transition rates, the dependence on data availability, and the runtime were assessed. (**A**) To assess the intrinsic error of simplifying a continuous proliferation process using a discrete-time proliferation model, we computed the error-of-fit between the phenotype trajectories generated by a two-state simulation of continuous proliferation, versus the trajectories obtained using a discrete-time Markov model with the same transition rates and the same average proliferation rate. The heat map shows different combinations of relative iteration interval and proliferation rate (parameterized as p_2, the proliferation rate of the second phenotype, relative to 1.0 rate for the first phenotype), and the green box delineates a range of p_2 that is most physiologically relevant ($0.5 \leq p_2 \leq 2$). (**B**) Runtimes of TRANSCOMPP as a function of the number of phenotypes in the system (top) and the number of replicates in the input data (bottom). *Y* axis in log-scale. (**C**) Scatter plots showing the agreement between the nominal transition rates used to generate pseudo-experimental data in the agent-based model, versus the best-fit transition rates computed by TRANSCOMPP, for systems with 2, 5, 8 or 11 different phenotypic states. The predicted rates are highly correlated with the original rates when an approximate proliferation rate is known for each phenotype (left). When proliferation rates are variable and not known (right), then the TRANSCOMPP-estimated rates show slightly greater errors. (**D**) The RMSD (error-of-fit) between predicted and nominal transition rates, as a function of the amount of data (number of replicates) supplied as input to TRANSCOMPP. Even in a system with many states (Phe = 11), the RMSD is close to minimum with 4–5 replicates

3.3 Re-equilibration between stem-like CD24^{neg} and non-stem CD24^{pos} cells in the basal-like breast cancer cell line MCF10CA1a

CD24^{neg} (stem-like) and CD24^{pos} (non-stem) cells from the basal-like breast cancer cell line MCF10CA1a were obtained by

fluorescence-assisted cell sorting (FACS) (Fig. 3A–C) and cultured for 32 days ($n = 4$ replicates) in BM.

To verify the existence of a dynamic equilibrium between the CD24^{neg} and CD24^{pos} phenotypic states, single-cell measurements of CD24 intensity were performed for all replicates of each sorted population, on days 4, 8, 12, 19, 25 and 32 post-sorting. Data showed that, regardless whether the population had been originally enriched for CD24^{neg} or CD24^{pos} cells, the BM-treated populations converged within 12 days to the same fractional composition of ~20% CD24^{neg} and ~80% CD24^{pos} cells, and this ratio remained stable over the 32-day monitored period (Fig. 3D).

3.4 TRANSCOMPP-computed transition rates reveal that external environment cues affect the stemness equilibrium in MCF10CA1a

A supplemented form of BM, called CM is also widely used to maintain the MCF10A series of cell lines *in vitro*. We noticed that MCF10CA1a cells grown in CM exhibited morphological differences in comparison to those grown in BM (Fig. 4A), after three passages (around 2 weeks). Cells grown in BM formed compact epithelial colonies, whereas cells grown in CM appeared as looser colonies with scattered cells. Interestingly, the observed morphologies were found to be reversible on changing the medium from CM to BM or vice versa.

This reversibility demonstrates that interconversion between the epithelial-like and scattered morphologies could be triggered by BM and CM as environmental cues. Hence, to assess if the same environmental cues could affect equilibrium dynamics of MCF10CA1a, we cultured sorted-CD24^{neg} and CD24^{pos} cells in either BM or CM for 12 days (flow cytometry every 4 days) (Fig. 4B). TRANSCOMPP was used to compute best-fit transition rates. Computed transition rates indicated that compared to BM, CM induced a 2-fold increase in the rate of transition of the non-stem CD24^{pos} cells to the stem-like CD24^{neg} cells, with minimal effect on the CD24^{neg} to CD24^{pos} cell transition rate (Fig. 4C and D).

To determine the respective contributions of the individual supplements to the CM-induced plasticity, we sorted CD24^{neg} and CD24^{pos} cells and cultured them for 12 days in BM, supplemented individually with each CM component (HC/INS/EGF/CTX). For each treatment, we performed flow cytometry every 4 days (Supplementary Material S8) and computed transition rates using TRANSCOMPP (Fig. 5). The estimated transition rates were also used to simulate population dynamics, starting from purified states at time t_0 as per Equation (2) (see Materials and Methods). The fit of the simulated populations to the experimental observations, and the predicted equilibrium compositions for each treatment can be seen in Supplementary Material S9.

Interestingly, the computed transition rates suggested that exposure to HC caused the most pro-CD24^{neg} dynamic as shown by the lowest CD24^{neg} to CD24^{pos} transition rates compared to BM, CM, EGF, INS and CTX, while CTX treatment was found to show the most pro-CD24^{pos} dynamic with the lowest CD24^{pos} to CD24^{neg} transition rate among all culture conditions. However, before concluding that the underlying transition rates were truly different across treatments, we had to assess if the observed differences between treatments were greater than the difference that might be seen within replicates of the same treatment, and thus establish statistical significance.

3.5 Resampling method and its application

We developed a resampling method to quantify uncertainty intervals for the computed transition rates, and perform statistical comparisons between experiments. The resampling method works by randomly subsampling the experimental measurements (input single-cell or fractional composition data) to obtain a *pseudo-data trajectory* (Fig. 6).

We used the resampling method to compute CD24 transition rate distributions for each of the six media treatments used (Table 1).

Transition rate distributions were computed using SSR error during resampling.

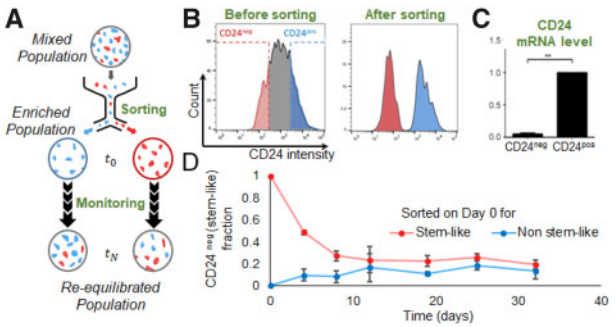

Fig. 3. Re-equilibration between CD24^{neg} (stem-like) and CD24^{pos} (non-stem) phenotypes in the basal-like breast cancer cell line MCF10CA1a. (**A**) Schematic showing enrichment of phenotypic states (red or blue) from a heterogeneous population of cells, by sorting/enrichment through FACS, dilution, etc. For each enriched state, single-cell measurements of phenotypic state (e.g. flow cytometry) are performed periodically to monitor changes over time. (**B**) Flow cytometry of the untreated MCF10CA1a population shows heterogeneity in the CD24 stemness marker (left). Cells were sorted into stem-like CD24^{neg} and non-stem CD24^{pos} cells by FACS. After sorting, the CD24 populations were validated by flow cytometry (right). (**C**) RT-PCR of CD24 mRNA for CD24^{neg}-sorted and CD24^{pos}-sorted cells. (**D**) The fraction of stem-like CD24^{neg} cells over time, in populations that were originally sorted into stem-like CD24^{neg} (red) and non-stem CD24^{pos} (blue). Flow cytometry was performed on days 4, 8, 12, 19, 25 and 32 post-sorting. Regardless of initial enrichment, the cellular population re-converged toward ~15–20% CD24^{neg} stem-like fraction

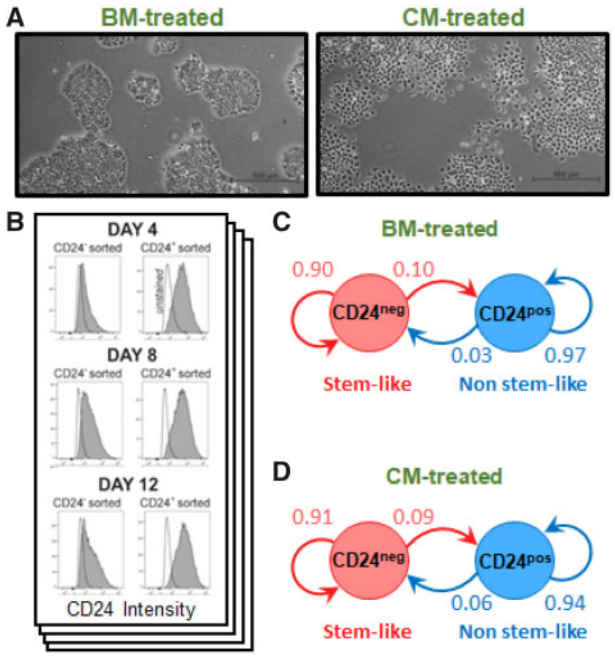

Fig. 4. Phenotype and re-equilibration dynamics of the MCF10CA1a cell line is affected by choice of growth medium. (**A**) Compared to MCF10CA1a cells grown in BM, cells grown in complete medium (CM = BM + 20 ng/ml epidermal growth factor, 10 μg/ml human insulin, 100 ng/ml cholera toxin and 0.5 μg/ml hydrocortisone) showed scattered morphology after 90 days (same magnification. Scale bar = 500 μm). (**B**) Sorted CD24^{neg} and CD24^{pos} populations were cultured in either control BM or CM and for 12 days ($n = 4$ replicates). Flow cytometry was performed on days 4, 8 and 12 post-sorting. Stacked rectangles denote the four replicates used for each treatment (BM or CM). Within each rectangle, each square plot is a flow cytometry dataset, with the shaded region indicating the histogram of CD24 intensities, and the unshaded indicating the unstained control. (**C**) The computed best-fit transition rates for 12-day BM cultures, computed using TRANSCOMPP. (**D**) The computed best-fit transition rates for 12-day CM cultures

Statistical comparisons across all conditions (Fig. 7A) indicated that the transition rates induced by most supplements were significantly different from the rates found in BM (Supplementary Material S10). EGF and INS caused a significant increase in the rate of $CD24^{pos}$ to $CD24^{neg}$ transition, without much effect on the $CD24^{neg}$ to $CD24^{pos}$ transition. In contrast, HC acted bi-directionally and exerted a stronger effect, by suppressing the transition from $CD24^{neg}$ to $CD24^{pos}$ while increasing the transition from $CD24^{pos}$ to $CD24^{neg}$. Finally, CTX caused opposite changes in transition rates, bi-directionally - increasing the rate of $CD24^{neg}$ to $CD24^{pos}$ while decreasing the rate of $CD24^{pos}$ to $CD24^{neg}$.

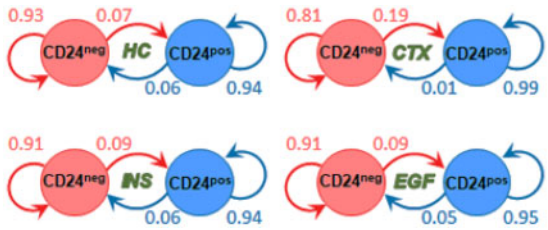

Fig. 5. Computing transition rates for MCF10CA1a cells exposed to four media supplements. TRANSCOMPP was used to compute best-fit transition rates for cells grown in BM with the addition of individual components of CM: HC, CTX, INS and EGF

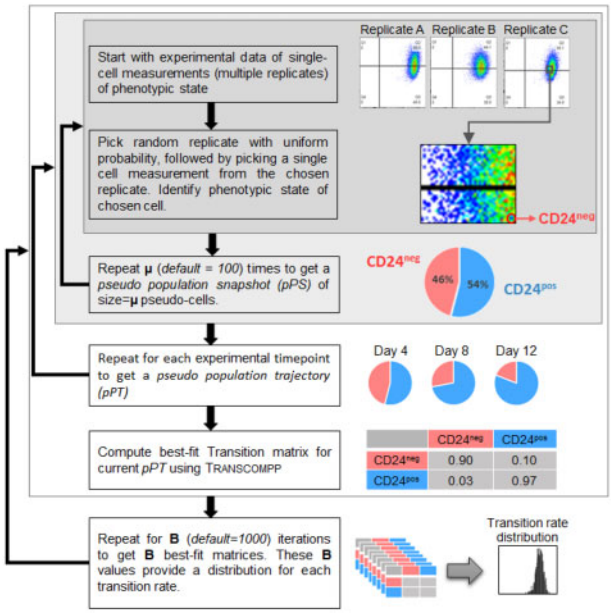

Fig. 6. Schematic of the resampling method, to compute transition rate distributions. The resampling method is a technique to estimate the variability in the best-fit transition rates. It involves iteratively creating B pseudo datasets from the experimental data. Each pseudo dataset is created by picking μ cell intensities at random from all replicates (such that each replicate is given equal weightage), and then computing the fraction of cells in each state. After repeating for each time-point, we obtain a *pseudo population trajectory*, from which we compute the best-fit transition matrix. Repeating this for the B *pseudo population trajectories* allows us to obtain a distribution for each transition rate. Shown on the left are the individual steps involved in resampling, and on the right are illustrative examples of the data types utilized at each stage of the algorithm

In summary, of all the tested supplements, HC and CTX caused the greatest change in transition rates, compared with BM control conditions. The computed transition rate distributions were also used to simulate the *expected range* of $CD24^{neg}$ fractions (Fig. 7B shaded regions) over 12 days for each treatment. In all cases, this predicted range was found to align well with the observed $CD24^{neg}$ dynamics. We also observed that trajectories simulated using the combined effects (transition rates) of the four supplements (plus BM) did not recapitulate either the experimental 12-day CM data, or the simulated dynamics using the CM transition rates. This indicates that the overall effect of the CM cocktail is different from the sum of its parts (Supplementary Material S11).

3.6 TRANSCOMPP predicts long-term dynamics for head and neck cancer patients using transition rates computed from early time-points

In a clinical setting, re-equilibration was earlier demonstrated in patient-derived cells from head and neck cancer (HNCC) (Leong *et al.*, 2014), where the authors noted that populations would revert to an equilibrium with ~5% stem-like (ALD^{Hi}) and ~95% non-stem (ALD^{Lo}) cells, starting from enriched populations of either phenotype. Treatment with growth factors (Insulin + EGF), shifted the equilibrium toward the stem-like (ALD^{Hi}) phenotype, but no transition rates were computed. TRANSCOMPP analysis (Fig. 8A) computed transition rates confirming that non-stem cells had very low plasticity (0.2% rate of state-change per passage) and stem-like cells had much greater plasticity (6% rate of state-change per passage) in two different patient samples. Model fit for the other HNCC datasets can be found in Supplementary Material S12.

To assess the predictive power of Markov modeling in this case, we asked whether the rates computed using Markov assumptions on short-term data could be extrapolated to predict future behavior of patient-derived cells. Given that the HNCC experiments were performed over 36 passages (>100 days), we retrospectively asked how much monitoring would have been sufficient for TRANSCOMPP to predict long-term population dynamics and steady-state convergence in the experimental data. We truncated the experimental dataset to the first 3, 4, 5…12 time-points, and applied TRANSCOMPP to each dataset. Figure 8C shows fits obtained from using only the early 3, 6 or 9 time-points.

Calculations revealed that for both patients included in the study, four early time-points were sufficient to extrapolate a predicted trajectory of convergence that had an RMSD of 2.37% and 3.17% (ALD^{Hi} fraction) with the observed trajectory, for the two patients, respectively. In comparison, the RMSD using rates computed with the full 12 time-point dataset was found to be 2.11% and 2.97% respectively (a sub-10% difference in RMSD) showing how good a fit could be obtained from using four early time-points only (Supplementary Material S13).

4 Discussion

We developed TRANSCOMPP, a tool to compute rates (and distributions) of stochastic transitions between phenotypic states.

Table 1. Transition rate distributions for MCF10CA1a cells in different environments (reported as mean ± SD)

Media	$CD24^{neg} \rightarrow CD24^{neg}$	$CD24^{neg} \rightarrow CD24^{pos}$	$CD24^{pos} \rightarrow CD24^{neg}$	$CD24^{pos} \rightarrow CD24^{pos}$
BM	0.90 ± 0.010	0.10 ± 0.010	0.03 ± 0.005	0.97 ± 0.005
CM	0.91 ± 0.010	0.09 ± 0.010	0.06 ± 0.010	0.94 ± 0.010
HC	0.93 ± 0.010	0.07 ± 0.010	0.06 ± 0.009	0.94 ± 0.009
CTX	0.81 ± 0.015	0.19 ± 0.015	0.01 ± 0.004	0.99 ± 0.004
INS	0.92 ± 0.010	0.08 ± 0.010	0.06 ± 0.008	0.94 ± 0.008
EGF	0.91 ± 0.011	0.09 ± 0.011	0.05 ± 0.007	0.95 ± 0.007

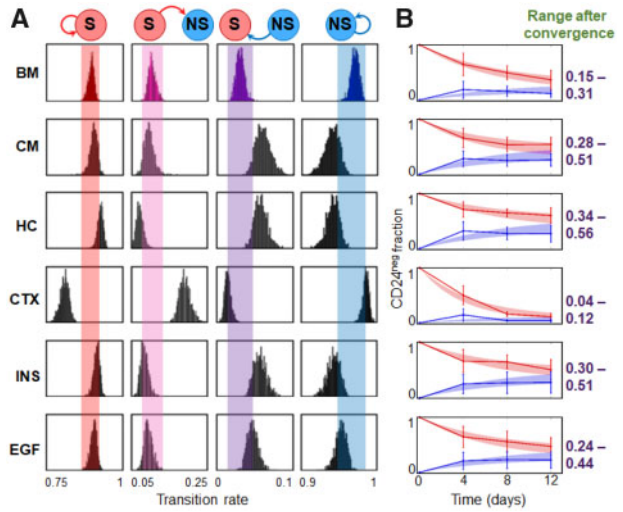

Fig. 7. Transition rate distributions computed for each treatment show statistically-significant differences, and result in different long-term equilibria. (**A**) Using the resampling method, we obtained statistical distributions for each transition rate for each environmental treatment. Shown here are the distributions for each of the four transition rates. HC strongly promotes stem-like $CD24^{neg}$ population while CTX promotes the non-stem $CD24^{pos}$ population (statistical tests show significance compared with control and other environments). (**B**) The predicted trajectories and range of steady-state $CD24^{neg}$ fractions, for each media condition, based on the 95% bootstrap uncertainty interval of transition rates from (A)

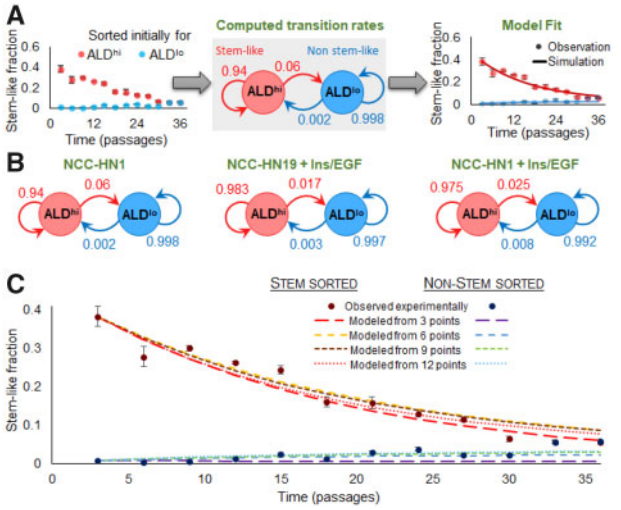

Fig. 8. Application of TRANSCOMPP to patient-derived cells. (**A**) At left, time-series measurements of stem-like fractions (shown as mean and SD of three replicates), using patient NCC-HN19 from Leong *et al.* (2014). Central panel shows the best-fit transition rates computed by TRANSCOMPP. At right is the simulated trajectory using the Markov model with the computed transition rates, plotted to show fit with experimental observations. (**B**) Computed transition rates for the other datasets in Leong *et al.* (2014), including cells from patient NCC-HN1, and cells treated with insulin + EGF to favor stemness. (**C**) Points are experimental observations of cells enriched for stem-like and non-stem cells in patient NCC-HN1. Dashed lines show Markov model predictions computed from using 3, 6, 9 or 12 time-points (reflecting 9, 18, 27 or 36 passages, respectively)

Quantifying the rate of interconversion between the identified phenotypic states will help us understand which key transitions dominate dynamics and help us bias equilibria toward less aggressive states.

Although previous work has dealt with similar issues, the following features of TRANSCOMPP make it the first general-purpose method developed toward this goal. First and foremost, our method integrates the ability to handle states with different proliferation rates, and hence can be used to run simulations where change in bulk population is a result of both stochastic transitions and proliferation, which can strongly affect the results of rate-fitting (Supplementary Material S1). Second, our method employs optimization to minimize the fitting error between the modeled trajectory and experimental observations, computing the transition matrix that best explains the variation *across all replicates and time-points* (Supplementary Material S2). Third, our method computes transition rate distributions and uncertainty intervals around the computed transition matrix, through repeated resampling. This allows us to perform statistical tests comparing transition rates from different treatments/environments. Fourth, our method does not impose constraints on the number of replicates (sorted/unsorted) that can be used as input, and hence is compatible with a wider variety of experimental designs. Lastly, our method can be customized to work with different error metrics (for model fit). Benchmarking Transcompp performance suggests that Transcompp is scalable, can be applied to a wider variety of datasets, and provides more robust performance than existing tools like CellTrans (Supplementary Material S2 and S3).

Applying TRANSCOMPP to the MCF10CA1a cell line model of basal-like breast cancer, we showed that the equilibrium between $CD24^{neg}$ stem-like and $CD24^{pos}$ non-stem cells could be influenced by the composition of the cell culture medium (Figs 3–5). We also used the resampling module of TRANSCOMPP (Fig. 6) to compute the impact of experimental noise on the Markov transition rates, and this suggested that HC or CTX supplementation exerted the strongest influence on the transition rates (Fig. 7), although in opposite directions. CTX (an agonist of the cAMP signaling pathway) promoted accumulation of the non-stem-like $CD24^{pos}$ fraction. This is in agreement with a recent report that PKA activation by CTX and Forskolin causes mesenchymal–epithelial transitions in multiple mammary cell lines (Pattabiraman *et al.*, 2016). In contrast, HC was found to favor an equilibrium with a much greater proportion of stem-like $CD24^{neg}$ cells (~46%), compared to the equilibria seen with BM (~23%), or other supplements such as INS (~40%) and EGF (~36%) (Supplementary Material S9), which are known inducers of plasticity (Abhold *et al.*, 2012; Castaño *et al.*, 2013; Ma *et al.*, 2013; Malaguarnera and Belfiore, 2014; Tominaga *et al.*, 2017; Xu *et al.*, 2017). Since HC is commonly prescribed for nausea in cancer patients, our study raises questions about the safety of HC for certain subtypes of breast cancer. Likewise, since HC is an analog of the stress hormone cortisol, our finding suggests a novel avenue for why psychosocial stress is detrimental for certain subtypes of breast cancer.

Applying TRANSCOMPP to published data from patient-derived cells (Leong *et al.*, 2014) showed that retrospective TRANSCOMPP predictions made from short-term experiments matched long-term population behavior and steady-state population compositions (Fig. 8). Short-term assays of long-term plasticity could become a powerful tool for precision medicine because future anti-cancer therapies will require not just targeting tumor-specific mutations but also anticipating likely future trajectories [evolutionary-enlightened design (Brown *et al.*, 2017; Jonsson *et al.*, 2017; Yeang and Beckman, 2016)], in order to decrease the probability of chemo-induced epithelial-mesenchymal transition (Kim *et al.*, 2015; Sun *et al.*, 2012).

One caveat of TRANSCOMPP is that the computed rates are appropriate and relevant only when the user definitions of 'state' represent all the distinct, physiologically relevant phenotypes in the sample. Ongoing research to discover unbiased definitions of phenotypic clusters from single-cell sequencing (Kanter *et al.*, 2019; Patel *et al.*, 2014; Poirion *et al.*, 2018; Zemmour *et al.*, 2018) might resolve this in the future. Additionally, TRANSCOMPP assumes that observable population dynamics result only from Markovian stochastic transitions and phenotype-specific proliferation, a commonly held assumption. However, if the underlying biology in particular cases is non-Markovian, one of two things might happen. Failure of the best-fit Markov model to recapitulate experimentally observed dynamics might be suggestive of non-Markovian biology. In contrast, if the biology is truly non-Markovian but still can be fit by a

Markov model, our approach would not detect any problem. In addition, perceived non-Markovian behavior could also be a result of switching between different Markov equilibria (which might occur due to binary activation/inactivation of key pathways). Non-Markovian cases are a focus of our ongoing and future work.

In the present work, we use single-cell measurements of antibody markers to define phenotype, but with the growth of droplet technology and single-cell measurements (e.g. scRNA-seq, single-cell Westerns), we could in the future classify phenotypic categories at a much higher resolution, creating a phenotypic landscape with fine-grained differences in 'state'. Since TRANSCOMPP is compatible with all forms of single-cell data, it can be used to describe the 'flows' of cells across this landscape.

In conclusion, we provide a novel tool to compute rates (and distributions) of stochastic transitions between phenotypic states. We have employed TRANSCOMPP in analyzing immortalized and patient-derived cancer cells *in vitro* and show the underappreciated pervasiveness of plasticity. While applied to cancer in this study, the technique of transition rate quantification has the potential to capture novel and interesting cell population dynamics throughout the life sciences.

Acknowledgements

The authors are grateful to N. Gopalakrishna Iyer, Leong Hui Sun and Elysia C. Saputra for data sharing and collaborative discussions.

Funding

This work was supported by the Singapore Ministry of Education's Tier 2 Grant [MOE2019-T2-1-138 to L.T.K.] and Tier 1 Grant [T1-2016Apr-07 to M.V.C.]; the St. Baldrick's Foundation [Duke-NUS-SBF/2018/0006 to L.T.K.]; the Singapore Ministry of Health's National Medical Research Council (NMRC) under its Open Fund Large Collaborative Grant [NMRC/OFLCG/003/2018 to L.T.K.]; and a seed grant from the National University Health System [NUHSRO/2018/091/T1/Seed-Nov/01 to MVC].

Conflict of Interest: none declared.

References

Abhold,E.L. *et al.* (2012) EGFR kinase promotes acquisition of stem cell-like properties: a potential therapeutic target in head and neck squamous cell carcinoma stem cells. *PLoS One*, **7**, e32459.

Al-Hajj,M. *et al.* (2003) Prospective identification of tumorigenic breast cancer cells. *Proc. Natl. Acad. Sci. USA*, **100**, 3983–3988.

Armond,J.W. *et al.* (2014) A stochastic model dissects cell states in biological transition processes. *Sci. Rep.*, **4**, 3692.

Beerenwinkel,N. *et al.* (2015) Cancer evolution: mathematical models and computational inference. *Syst. Biol.*, **64**, e1–e25.

Bhat-Nakshatri,P. *et al.* (2010) SLUG/SNAI2 and tumor necrosis factor generate breast cells with CD44+/CD24- phenotype. *BMC Cancer*, **10**, 411.

Brown,J.S. *et al.* (2017) Aggregation effects and population-based dynamics as a source of therapy resistance in cancer. *IEEE Trans. Biomed. Eng.*, **64**, 512–518.

Buder,T. *et al.* (2017) CellTrans: an R package to quantify stochastic cell state transitions. *Bioinform. Biol. Insights*, **11**, 1177932217712241.

Castaño,Z. *et al.* (2013) Stromal EGF and IGF-I together modulate plasticity of disseminated triple-negative breast tumors. *Cancer Discov.*, **3**, 922–935.

Chaffer,C.L. *et al.* (2011) Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state. *Proc. Natl. Acad. Sci. USA*, **108**, 7950–7955.

Chisholm,R.H. *et al.* (2016) Cell population heterogeneity and evolution towards drug resistance in cancer: biological and mathematical assessment, theoretical treatment optimisation. *Biochim. Biophys. Acta*, **1860**, 2627–2645.

Dingli,D. and Pacheco,J.M. (2011) Stochastic dynamics and the evolution of mutations in stem cells. *BMC Biol.*, **9**, 41.

Emmons,M.F. *et al.* (2016) The role of phenotypic plasticity in the escape of cancer cells from targeted therapy. *Biochem. Pharmacol.*, **122**, 1–9.

Flöttmann,M. *et al.* (2012) A stochastic model of epigenetic dynamics in somatic cell reprogramming. *Front. Physiol.*, **3**, 216.

Gupta,P.B. *et al.* (2011) Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells. *Cell*, **146**, 633–644.

Hoek,K.S. and Goding,C.R. (2010) Cancer stem cells versus phenotype-switching in melanoma. *Pigment Cell Melanoma Res.*, **23**, 746–759.

Jonsson,V.D. *et al.* (2017) Novel computational method for predicting polytherapy switching strategies to overcome tumor heterogeneity and evolution. *Sci. Rep.*, **7**, 44206.

Kanter,I. *et al.* (2019) A cluster robustness score for identifying cell subpopulations in single cell gene expression datasets from heterogeneous tissues and tumors. *Bioinformatics*, **35**, 962–971.

Kemper,K. *et al.* (2014) Phenotype switching: tumor cell plasticity as a resistance mechanism and target for therapy. *Cancer Res.*, **74**, 5937–5941.

Kim,A.-Y. *et al.* (2015) Epithelial-mesenchymal transition is associated with acquired resistance to 5-fluorocuracil in HT-29 colon cancer cells. *Toxicol. Res.*, **31**, 151–156.

Leong,H.S. *et al.* (2014) Targeting cancer stem cell plasticity through modulation of epidermal growth factor and insulin-like growth factor receptor signaling in head and neck squamous cell cancer. *Stem Cells Transl. Med.*, **3**, 1055–1065.

Ma,L. *et al.* (2013) Cancer stem-like cell properties are regulated by EGFR/AKT/β-catenin signaling and preferentially inhibited by gefitinib in nasopharyngeal carcinoma. *FEBS J.*, **280**, 2027–2041.

Malaguarnera,R. and Belfiore,A. (2014) The emerging role of insulin and insulin-like growth factor signaling in cancer stem cells. *Front. Endocrinol.*, **5**, 10.

Nakshatri,H. *et al.* (2009) Breast cancer stem cells and intrinsic subtypes: controversies rage on. *Curr. Stem Cell Res. Ther.*, **4**, 50–60.

Patel,A.P. *et al.* (2014) Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, **344**, 1396–1401.

Pattabiraman,D.R. *et al.* (2016) Activation of PKA leads to mesenchymal-to-epithelial transition and loss of tumor-initiating ability. *Science*, **351**, aad3680.

Pisco,A.O. and Huang,S. (2015) Non-genetic cancer cell plasticity and therapy-induced stemness in tumour relapse: 'What does not kill me strengthens me'. *Br. J. Cancer*, **112**, 1725–1732.

Poirion,O. *et al.* (2018) Using single nucleotide variations in single-cell RNA-seq to identify subpopulations and genotype-phenotype linkage. *Nat. Commun.*, **9**, 4892.

Reyes,J. and Lahav,G. (2018) Leveraging and coping with uncertainty in the response of individual cells to therapy. *Curr. Opin. Biotechnol.*, **51**, 109–115.

Santner,S.J. *et al.* (2001) Malignant MCF10CA1 cell lines derived from premalignant human breast epithelial MCF10AT cells. *Breast Cancer Res. Treat.*, **65**, 101–110.

Sun,L. *et al.* (2012) MiR-200b and miR-15b regulate chemotherapy-induced epithelial-mesenchymal transition in human tongue cancer cells by targeting BMI1. *Oncogene*, **31**, 432–445.

Tominaga,K. *et al.* (2017) Addiction to the IGF2-ID1-IGF2 circuit for maintenance of the breast cancer stem-like cells. *Oncogene*, **36**, 1276–1286.

Xu,Q. *et al.* (2017) EGF induces epithelial-mesenchymal transition and cancer stem-like cell properties in human oral cancer cells via promoting Warburg effect. *Oncotarget*, **8**, 9557–9571.

Yeang,C.-H. and Beckman,R.A. (2016) Long range personalized cancer treatment strategies incorporating evolutionary dynamics. *Biol. Direct*, **11**, 56.

Zemmour,D. *et al.* (2018) Single-cell gene expression reveals a landscape of regulatory T cell phenotypes shaped by the TCR. *Nat. Immunol.*, **19**, 291–301.