# Stochastic inference of clonal dominance in gene therapy studies

L. Del Core[1*], M. A. Grzegorczyk[1*], E. C. Wit[2*]

**1** University of Groningen — Bernoulli Institute for Mathematics, Computer Science and Artificial Intelligence, Groningen, Netherlands
**2** Università della Svizzera italiana — Institute of Computing, Lugano, Switzerland

* l.del.core@rug.nl; m.a.grzegorczyk@rug.nl; ernst.jan.camiel.wit@usi.ch

## Abstract

Clonal dominance is a wake-up-call for adverse events in gene therapy applications. This phenomenon has mainly been observed as a consequence of a malignancy progression, and, in some rare cases, also during normal haematopoiesis. We propose here a random-effects stochastic model that allows for a quick detection of clonal expansions that possibly occur during a gene therapy treatment.

Starting from the Ito-type equation, the dynamics of cells duplication, death and differentiation at clonal level without clonal dominance can be described by a local linear approximation. The parameters of the base model, which are inferred using a maximum likelihood approach, are assumed to be shared across the clones. In order to incorporate the possibility of clonal dominance, we extend the base model by introducing random effects for the clonal parameters. This extended model is estimated using a tailor-made expectation maximization algorithm. The main idea of this paper is to compare the base and the extended models in high dimensional clonal tracking datasets by means of Akaike Information Criterion in order to detect the presence of clonal dominance. The method is evaluated using a simulation study, and is applied to investigating the dynamics of clonal expansion in a in-vivo model of rhesus macaque hematopoiesis.

## Author summary

Preventing or quickly detecting clonal dominance is an important aspect in gene therapy applications. Over the past decades, clonal tracking has proven to be a cutting-edge analysis capable to unveil population dynamics and hierarchical relationships in vivo. For this reason, clonal tracking studies are required for safety and long-term efficacy assessment in preclinical and clinical studies. In this work we propose a random-effects stochastic framework that allows to investigate events of clonal dominance using high-dimensional clonal tracking data. Our framework is based on the combination between stochastic reaction networks and mixed-effects generalized linear models. We have shown in a simulation study and in a real world application that our method is able to detect the presence of clonal expansions. Our tool can provide statistical support to biologists in gene therapy surveillance analyses.

# Introduction

The idea of gene therapy is that the correction of the defective gene(s) underlying the disease is, in principle, sufficient for inducing disease remission or even full recovery [1]. Since the blood system possesses a hierarchical structure with haematopoietic stem cells (HSC) at its root [2], correction of HSCs might be sufficient to eradicate a genetic disease in the blood system. Since transduction of stem cells has proven to be less efficient than transduction of more mature cells, it might be necessary to allow very high gene transfer rates, i.e., multiple vector copies, to ensure efficient genetic modification of HSCs [3,4]. But genetic modification of large numbers of cells is associated with the higher probability of unintentional vector insertion events near the growth-regulatory genes that may lead to insertional mutagenesis [5–7]. One particular drawback of insertional mutagenesis is the phenomenon of clonal dominance, which occurs if one or more clones dominate cell production [8]. The most extreme case of clonal dominance is monoclonality, where an entire tissue is dominated by the progeny of one particular cell. Although different gradients of clonal dominance (oligoclonality) exist, a precise threshold that defines dominance is hard to be specified in general, and thus a clear definition of what is meant by clonal dominance is required for any particular study.

Clonal dominance in malignant haematopoiesis has been previously identified as a consequence of a clonal competition that is corrupted by disease progression [9,10]. However, clonal dominance has also been observed in normal haematopoiesis, even in the case of truly neutral clonal markers [11–13]. Indeed, on the basis of various mathematical models, progression of monoclonality has been discussed also for normal (non-leukaemic) stem cell systems [14–18]. While there is strong evidence for clonal selection inducing monoclonal systems in the crypts of the small intestine [19–22], such a process has not been demonstrated for the haematopoietic system yet. To shed more light on those mechanisms, in this manuscript we extend the work of [23,24] and propose a random-effects cell differentiation network to model the dynamics of clonal expansions for high dimensional clonal tracking data.

More in detail, starting from the definition of the master equation [25], a set of Ito stochastic differential equations is derived to describe the first-two-order moments of the process. We estimate the parameters of the Ito system from its Euler-Maruyama local linear approximation (LLA) [26]. We propose a new inference procedure in the LLA formulation using a maximum likelihood approach, replacing the iterative weighted least square algorithm previously developed in [23,24]. Although the base LLA model formulation has been shown to be effective in modelling cell differentiation [24], it has some limitations as it only provides an average description of the dynamics across all the clones, and does not take into account possible extreme behaviour. Indeed in the base model all the dynamics parameters are shared across the clones, and thus is not possible to identify heterogeneous clonal patterns. Thus the base LLA formulation cannot be used to model clonal dominance. Therefore in this work we further increase the flexibility of the base LLA model to check if the process dynamics is mainly due to few clones and if those dominate a particular cell type. To this end we introduce random effects for the clones inside the LLA formulation, providing a mixed-effects LLA model. Then, if the mixed model outperforms the fixed one in terms of Akaike Information Criterion we use the former to infer the process parameters in order to identify which clones are mainly expanding and in which cell compartments. As every mixed-effects formulation, inference of parameters is performed by means of an expectation-maximization algorithm, for which we developed an efficient implementation. Effectively, our random-effects LLA formulation describes a stochastic process of clonal dominance on a network of cell lineages. We tested and validated our method using a simulation study. Finally, our model allowed to investigate the dynamics of clonal expansion in a in-vivo model of rhesus macaque hematopoiesis [27].

# Materials and Methods

This section contains background on clonal tracking data and a description of a Rhesus Macaques study. We also provide a concise description of the clonal dominance model and inference procedure. A more comprehensive, mathematical description can be found in Section 1 of S1 Text.

## Clonal tracking data

There are several high-throughput systems capable to quantitatively track cell types repopulation from an individual stem cell after a gene therapy treatment [28–30]. Tracking cells by random labeling is one of the most sensitive systems [31]. In gene therapy applications, haematopoietic stem cells (HSCs) are sorted from the bone marrow of the patient and uniquely labeled by the random insertion of a viral vector inside its genome. Each label, called clone, vector integration site (VIS), or barcode, is defined as the genomic coordinates where the viral vector integrates. After transplantation, all the progeny deriving through cell differentiation inherits the original labels. During follow-up, the labels are collected from tissues and blood samples using Next Generation Sequencing (NGS) [32–35]. Therefore NGS does allow identifying, quantifying and tracking clones arising from the same HSC ancestor. Over the past decades, clonal tracking has proven to be a cutting-edge analysis capable to unveil population dynamics and hierarchical relationships in vivo [36–39].

We consider single cell barcode data collected from an established hematopoietic stem cell gene therapy model previously used to investigate the hematopoietic reconstitution in Rhesus Macaques. [27] applied a lentiviral cellular barcoding technology to rhesus CD34+ HSPCs, thus allowing clonal tracking after myeloablative autologous transplantation. In particular, mobilized peripheral blood (MPB) CD34+ cells from three macaques were transduced with barcoded vectors, and 7.8-16.7 million autologous GFP+ cells were reinfused after an ablative total body irradiation. Following engraftment, myeloid Granulocytes (G), Monocytes (M), and lymphoid T, B, and Natural Killer (NK) cells were flow sorted (purity median 98.8%).

The authors showed with high confidence ( $> 95\%$) that a single barcode marked only one HSPC clone at these transplanted doses [40, 41]. Thus, only a minority of clones containing more than one barcode would skew calculations of the frequency of repopulating clones upward, but would not impact analysis of lineage contributions or kinetics. Barcode retrieval by PCR, Illumina sequencing, and custom data analysis was performed on purified hematopoietic lineage samples monthly for 9.5 months (ZH33), 6.5 months (ZH17), and 4.5 months (ZG66) [42]. They demonstrated high reproducibility of barcode retrieval and quantitation via sequencing several replicates on the same collected DNA samples. They also assayed independently processed replicate blood samples to identify a lower barcode read threshold that would result in 95% barcode retrieval between replicates. In particular they established a sampling error threshold of 1144 reads. Therefore we also considered the same reads threshold here, so as to be consistent with the previous studies. The total numbers of clones collected are 1165 (ZH33), 1280 (ZH17), and 1291(ZG66). To further remove bias, we only focused on the clones recaptured at least 5 times across lineages and time. This resulted in a subset of clones of size 481(ZH33), 139 (ZH17), and 202 (ZG66). Further details on transduction protocols and culture conditions can be found in the original study.

## A stochastic model for cell differentiation

We consider three event types, such as cell duplication, cell death and cell differentiation for a time counting process

$$Y_t = (Y_{1t}, \dots, Y_{Nt}) \tag{1}$$

of a single clone in $N$ distinct cell lineages. The time counting process $Y_t$ for a single clone in a time interval $(t, t + \Delta t)$ evolves according to a set of reactions $\{v_k\}_k$ and hazard functions $\{h_k\}_k$ defined as

$$v_k = \begin{cases} (0 \dots 1_i \dots 0)' \\ (0 \dots -1_i \dots 0)' \\ (0 \dots -1_i \dots 2_j \dots 0)' \end{cases} \qquad h_k(Y_t, \theta_i) = \begin{cases} Y_{it}\alpha_i & \text{for duplication} \\ Y_{it}^2\delta_i & \text{for death} \\ Y_{it}\lambda_{ij} & \text{for differentiation} \end{cases} \tag{2}$$

which contains a linear growth term with a duplication rate parameter $\alpha_i > 0$, a quadratic term for cell death with a death rate parameter $\delta_i > 0$, and a linear term to describe cell differentiation from lineage $i$ to lineage $j$ with differentiation rate $\lambda_{ij} > 0$ for each $i \neq j = 1, \dots, N$. Finally we use the LLA formulation of Section 1.3.1 from S1 Text with net-effect matrix and hazard vector defined as

$$V = \left[v_1 \cdots v_K\right]; \qquad h(Y_t; \theta) = \left[h_1(Y_t; \theta) \cdots h_K(Y_t; \theta)\right]' \tag{3}$$

In this formulation we implicitly assume that cells belonging to the same lineage obey to the same dynamics laws, that is all the clones share the same vector parameter $\theta$. In case we argue that clones behave differently in terms of dynamics we can use the random-effects LLA formulation of Eq. (6), where the random effects are defined on the vector parameter $\theta$ w.r.t. the clones. This is the case in our application study presented in next section, where we check wether there is heterogeneity in the clones for the duplication and death parameters, which we use as a proxy for a clonal expansion or contraction.

## LLA formulation of clonal dominance

Let $Y_t = (Y_{1t}, \dots, Y_{Nt})$ be a collection of "cells" of $N$ different types at time $t$ obeying to a network of stochastic biochemical reactions defined by a net-effect matrix $V \in \mathbb{Z}^{N \times K}$, a vector parameter $\theta$ and an hazard vector $h(Y, \theta) = (h_1(Y, \theta), \dots, h_K(Y, \theta))$ and let

$$\underbrace{\begin{bmatrix} \Delta Y_{t_0} \\ \vdots \\ \Delta Y_{t_{T-1}} \end{bmatrix}}_{\Delta Y} = \underbrace{\begin{bmatrix} M_{t_0} \\ \vdots \\ M_{t_{T-1}} \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} + \varepsilon; \quad \varepsilon \sim \mathcal{N}_{NT}\left(0, \overbrace{\underbrace{\begin{bmatrix} W_{t_0}(\theta) & & \\ & \ddots & \\ & & W_{t_{T-1}}(\theta) \end{bmatrix}}_{W(\theta)} + \sigma^2 I_{NT}}^{\Sigma(\theta, \sigma^2)}\right) \tag{4}$$

be the local linear approximation of an Ito-type equation written in generalized linear model formulation (see Section 1 of S1 Text for details) where

$$\Delta Y_t = V \overbrace{\begin{bmatrix} \prod_{i=1}^N \binom{Y_{it}}{r_{1i}} & & \\ & \ddots & \\ & & \prod_{i=1}^N \binom{Y_{it}}{r_{Ki}} \end{bmatrix}}^{M_t} \Delta t \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} + \left( \underbrace{V \begin{bmatrix} h_1(Y_t, \theta) & & \\ & \ddots & \\ & & h_1(Y_t, \theta) \end{bmatrix} V'}_{W_t(\theta)} + \sigma^2 I_N \right)^{1/2} dW(t) \tag{5}$$

$$dW(t) \sim \mathcal{N}_N(0, \Delta t I_N)$$

$\sigma^2$ is the noise variance, $M_t\theta$ the mean drift, $W_t(\theta)$ the diffusion matrix, and

$\Delta Y_t = Y_{t+\Delta t} - Y_t$ is a finite-time increment. In system (4) all the cell counts $Y_1, \dots, Y_N$ share the same parameter vector $\theta$. To infer the parameters of (4)-(5) we developed a maximum likelihood algorithm which is fully described in Section 1.4 of S1 Text. In some cases it may happen that the cells being analysed are drawn from a hierarchy of $J$ different populations that possibly behave differently in terms of dynamics. In this case it might be of interest to quantify the population-average $\theta$ and the subject-specific effects $u$ around the average $\theta$ for the description of the subject-specific dynamics. Therefore we introduce here a novel stochastic framework which is more flexible then the base LLA model thus allowing for the quantification of clonal contribution to the process. In particular, to quantify the contribution of each subject $j = 1, \dots, J$ on the process dynamics we extend the LLA (4) with a mixed-effects model [43] introducing random effects $u$ for the $J$ distinct subjects on the parameter vector $\theta$, leading to a random-effects stochastic reaction network (RestoreNet). The extended random-effects formulation becomes

$$\Delta Y = \underbrace{\begin{bmatrix} M_1 & & 0 \\ & \ddots & \\ 0 & & M_J \end{bmatrix}}_{\mathbb{M} \in R^{N \times Jp}} u + \varepsilon \qquad u \sim \mathcal{N}_{Jp}\left( \underbrace{1_J \otimes \theta}_{\theta_u}, I_J \otimes \underbrace{\begin{bmatrix} \tau_1^2 & & 0 \\ & \ddots & \\ 0 & & \tau_p^2 \end{bmatrix}}_{\Delta_u} \right) \tag{6a}$$

$$\varepsilon \sim \mathcal{N}_{Jp}(0, \Sigma(\theta, \sigma^2)) \tag{6b}$$

where $\mathbb{M}$ is the block-diagonal design matrix for the random effects $u$ centered in $\theta$, where each block $M_j$ is subject-specific. As in the case of the null model (4), to explain additional noise of the data, which has the additional advantage of avoiding singularity of the covariance matrix $W(\theta)$, we add to its diagonal a small quantity $\sigma^2$ which we infer from the data. Under this framework it can be shown that

$$u|\Delta Y \sim \mathcal{N}_{Jp}(E_{u|\Delta Y;\psi}[u], V_{u|\Delta Y;\psi}(u)) \tag{7}$$

where $\psi = (\theta, \sigma^2, \tau_1^2, \dots, \tau_p^2)$ is the set of all the unknown parameters. Once the parameters are estimated (see next section for inference details), the conditional expectations $E_{u|\Delta Y;\psi}[u]$ can then be used as a proxy for the clone-specific rate parameters. This method allows to infer the clone-specific dynamic by extremely reducing the problem dimensionality from $J \cdot p$ to $2 \cdot p + 1$ ($J \gg 2$).

## Inference procedure

In order to infer the Maximum Likelihood estimator $\hat{\psi}$ for $\psi = (\theta, \sigma^2, \tau_1^2, \dots, \tau_p^2)$ we develop an efficient tailor-made Expectation-Maximization algorithm where the collected cell increments $\Delta Y$ and the random effects $u$ take the roles of the observed and latent states respectively. The full analytical expression of $E_{u|\Delta Y;\psi}[u]$, $V_{u|\Delta Y;\psi}(u)$, the E-step function $Q(\psi|\psi^*) = E_{u|\Delta Y;\psi^*}[\ell(\Delta Y, u; \psi)]$ and its partial derivatives $\frac{\partial}{\partial \psi_j} Q(\psi|\psi^*)$ are available (see Section 1.4 of S1 Text). In the EM-algorithm we iteratively update the E-function $Q(\psi|\psi^*)$ using the current estimate $\psi^*$ of $\psi$ and then we minimize the $-Q(\psi|\psi^*)$ w.r.t. $\psi$. As the E-step function $Q(\psi|\psi^*)$ is non-linear, we used the L-BFGS-B algorithm from the optim() base R function for optimization, to which we provided the objective function, along with its gradient $\nabla_\psi Q(\psi|\psi^*)$, as input. Given the high-dimensionality of the clones being analysed, and due to the sparsity of the clonal tracking datasets, the E-step function $Q(\psi|\psi^*)$ and its gradient $\nabla_\psi Q(\psi|\psi^*)$ are written in a sparse block-diagonal matrix fashion, so as to reduce computational complexity and memory usage. The EM algorithm is run until a convergence criterion is met, that is when the relative errors of the E-step function $Q(\psi|\psi^*)$ and the parameters $\psi^*$ are lower than a predefined tolerance.

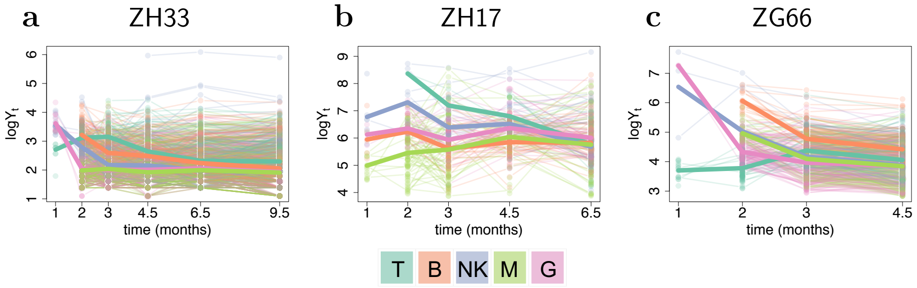


**Fig 1.** Logarithmic rescaled clonal abundance ($y$-axis) over time ($x$-axis) in each lineage (colors) of each treated animal (a-c). The thin lines are clone-specific, while the thick lines are the average across the clones. Data is rescaled according to equation (19) from S1 Text.

Once we get the EM estimate $\hat{\psi}$ for the parameters we evaluate the goodness-of-fit of the mixed-model according to the conditional Akaike Information Criterion [44]. As every EM algorithm, the choice of the starting point $\psi_s$ is very important from a computational point of view. We chose $\psi_s = (\theta_s, \sigma_s^2, \tau_1^2 = 0, \dots, \tau_p^2 = 0)$ as a starting point where $(\theta_s, \sigma_s^2)$ is the optimum found in the fixed-effects LLA formulation (4). This is a reasonable choice since we want to quantify how the dynamics $E_{u|\Delta Y;\hat{\psi}}[u]_j$ of each subject (clone) $j$ departs from the average dynamics $\theta_s$. With the help of simulation studies (see Section 2 of S1 Text) we empirically proved that this choice always led to a conditional expectation $E_{u|\Delta Y;\psi}[u]$ consistent with the true clone-specific dynamic parameters $\theta$. Computational details can be found in Section 1.4 of S1 Text. The pseudocode of the EM algorithm is provided in Algorithm 3 of S1 Text.

## Computational implementation

The maximimul likelihood inference for the basal model and the expectation maximization algorithm for the random-effects model are implemented in the R package RestoreNet. Few minimal working examples showing the usage of the package are provided in Section 5 of S1 Text.

# Results

A first comparative evaluation study on synthetic data, whose results are provided in Section 2 of S1 Text, shows how the proposed random-effects formulation is able to identify clonal dominance. We found that the random-effects model reached a significantly lower AIC than the null model, thus detecting the simulated dominance of a single clone into a cell type.

Next, we compared the base and random-effects models on the clonal tracking data of the rhesus macaque study fully described in Section “Clonal tracking data”. Although the sample DNA amount was maintained constant during the whole experiment (200 ng for ZH33 and ZG66 or 500 ng for ZH17), the sample collected resulted in different magnitudes of total number of reads (see Table 2 from S1 Text). This discrepancy makes all the samples not comparable across time and cell types. Therefore we rescaled the barcode counts according to Eq. (19) from S1 Text, and we report the rescaled cell counts, at clonal level, in Fig. 1. Since the CD34+ cells were not collected, we only estimated the duplication parameters $\alpha_T$, $\alpha_B$, $\alpha_{NK}$, $\alpha_M$, $\alpha_G$ and the death parameters $\delta_T$, $\delta_B$, $\delta_{NK}$, $\delta_M$, $\delta_G$ of the lymphoid (T, B, NK) and myeloid (M, G) cells. Therefore the differentiation parameters are not present in our model, and the net-effect matrix and the hazard vector are obtained from Eq. (2) - (3) accordingly. The

| | | $p$ | AIC | $KL_{div}$ | $KL_{div}/n$ |
|---|---|---|---|---|---|
| ZH33 | $M_0$ | 11.00 | 81377.27 | | |
| | $M_1$ | 434.16 | 38160.15 | 21062.95 | 1.87 |
| ZH17 | $M_0$ | 11.00 | 336752.11 | | |
| | $M_1$ | 478.43 | 29478.05 | 291854802.44 | 114228.89 |
| ZG66 | $M_0$ | 11.00 | 31194.60 | | |
| | $M_1$ | 410.92 | 21384.85 | 232030.37 | 83.77 |

**Table 1.** Comparison between fixed and mixed effects model: Number of parameters ($p$), Akaike Information Criterion (AIC), Kullback-Leibler divergence ($KL_{div}$) and rescaled $KL_{div}$ ($KL_{div}/n$) for the fixed ($M_0$) and the mixed ($M_1$) models in each treated individual.

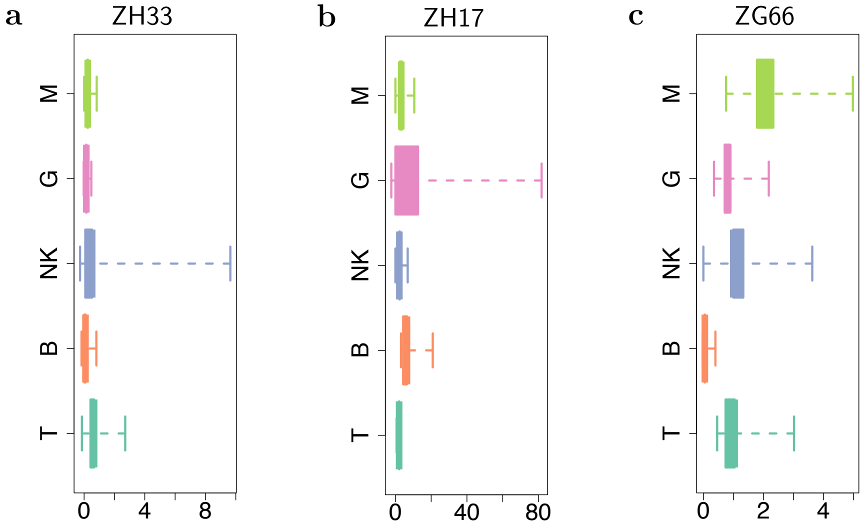


**Fig 2.** For each animal analyzed (a-c), the boxplots of the conditional expectations $E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_l}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_l}]$ computed from the estimated parameters $\hat{\psi}$ for the clone-specific net-duplication $\alpha_l - \delta_l$ in each cell lineage $l$ (different colors). The whiskers extend to the data extremes.

corresponding model becomes effectively a birth/death model. We fitted both the fixed model (4) and the mixed-effects model (6) separately to the data of each animal, where $J$ is equal to 481 (ZH33), 139 (ZH17), and 202 (ZG66) respectively. The size of the dynamic vector parameter $\theta$ is equal to 10, that is one scalar value for each combination of the five cell types with the duplication and death reactions. Also, $N$ equals 11275 (ZH33), 2555 (ZH17), and 2770 (ZG66), while the number of time-points $T$ is equal to 6 (ZH33), 5 (ZH17), and 4 (ZG66).

We report the results on model selection in Table 1 and the estimated parameters $\hat{\psi}$ in Table 3 of Section 4 from S1 Text. Then, from the estimated parameters $\hat{\psi}$ following Eq. (18) from S1 Text we computed the conditional expectations $E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_l}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_l}]$, which we use as a proxy for the $k$-th clone-specific net-duplication $\alpha_l - \delta_l$ in each cell lineage $l$. The resulting values are reported in Fig. 2 in a box-plot fashion. To visualize our findings at clonal level, in Fig. 3 we propose to use a weighted pie chart. Each pie corresponds to a particular clone and is weighted by the corresponding conditional expectations $E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_l}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_l}]$. The biological interpretation of this figure is that the larger the diameter, the more the corresponding clone is dominating cell production into the lineage associated to the largest slice.

As a result, according to the AIC values, in each treated individual the mixed model ($M_1$) outperformed the fixed one ($M_0$). This means that the clones did not follow the same average dynamics for the birth/death process. Instead, the dynamic of some

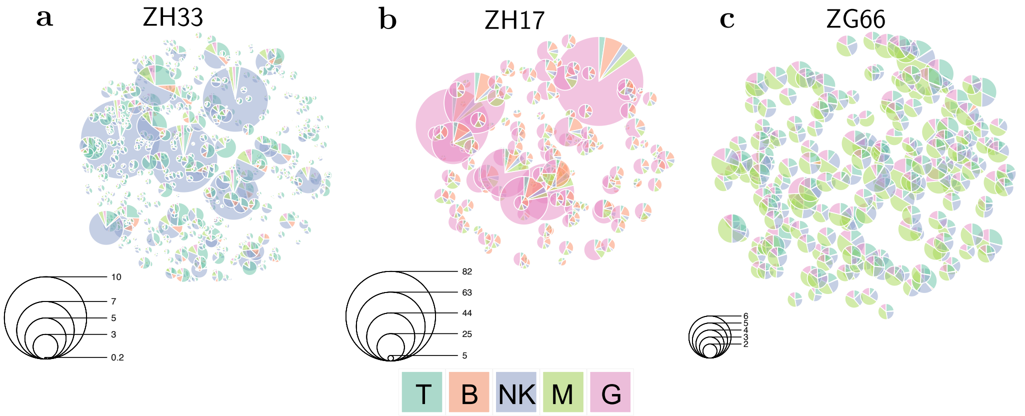


**Fig 3.** Graphical representation of the results obtained with the proposed the mixed effects model: Random effects for the clones to the process parameters of the rhesus macaques ZH33 (a), ZH17 (b) and ZG66 (c). Each $k$-th clone is identified with a pie whose slices are lineage-specific and weighted with $w_k$ defined as the difference between the corresponding duplication and death parameters, that is $w_k = E_{u|y}[u^k_{\alpha_{lin}}] - E_{u|y}[u^k_{\delta_{lin}}]$. The diameter of the $k$-th pie is proportional to the euclidean 2-norm of $w_k$. The legend scales are different across the three plot panels.

clones departed from the average dynamics with a significant (random) effect. In particular, the conditional net-duplication rates $E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_l}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_l}]$ of Fig. 2 - 3 suggest that there is clonal dominance in specific cell lineages. As an example, for the animals ZH33 and ZG66 we observed clonal expansions into NK cells with high conditional rates. Whereas, for the animal ZH17 we observed clonal expansions into G and B cell lineages with high conditional rates. Finally, for the animal ZG66 we also observed events of clonal dominance into M and T cell lineages. Furthermore, the weighted pie charts shown in Fig. 3 reveal different gradients of clonal dominance between the three rhesus macaques. As an example, looking at the size of the pies it is possible to observe an higher clonal dominance of NK cells in ZH33 and of G cells in ZH17 compared to the expansions of M, NK and T cells detected in ZG66, where the diameters of the clone-specific pies are rather similar. Not only does the proposed mixed effect model detect clonal dominance of certain cell types, it is also able to detect which clones are responsible.

## Discussion and conclusion

In this work we proposed a random-effects cell differentiation network which takes into account heterogeneity in the dynamics across the clones. Our framework extends the clone neutral local linear approximation of a stochastic quasi-reaction network, written in the Ito formulation, by introducing random-effects for the clones to allow for clonal dominance. To infer the parameter of the base (fixed-effects only) model we used a maximum likelihood approach. Whereas, to infer the parameters of the random-effects model, we have developed an expectation-maximization (EM) algorithm. We tested our framework with a $\tau$-leaping simulation study (see Section 2 from S1 Text for details), showing accurate performance of the method in the identification of a clonal expansion and in the inference of the true parameters. Subsequently, the application of the method on a rhesus macaque clonal tracking study revealed significant clonal dominance for specific cell types. Particularly interesting is that the NK clonal expansions detected by our model were already observed by former studies [27, 45, 46], and therefore our findings are consistent with those previously obtained. Indeed [45] described the oligoclonal expansions of NK cells and the long-term persistence of HSPCs and immature NK cells.

The main approximation in both the basal and random-effects formulations is the piece-wise linearity of the process. That is, in both cases we consider first a local linear approximation of the Ito equation, which then we use to infer the process parameters either with or without random-effects. Although the linearity assumption makes all the computations easier, this approximation becomes poor as the time lag increments (the $\Delta t$s) of the collected data increase. This can be addressed by introducing in the likelihood higher-order approximation terms than the ones considered by the Euler-Maruyama method. The Milstein approximation is a possible choice. Another, completely different, approach is to employ extended Kalman filtering (EKF) which is suitable for non-linear state space formulations. Also, our framework cannot consider false-negative errors or missing values of clonal tracking data. Also for this issue, an EKF formulation could be a possible extension.

Our tool can be considered as complementary to the classical Shannon entropy index [47] in detecting fast and uncontrolled growing of clones after a gene therapy treatment. Indeed, while the Shannon entropy measures the diversity of a population of clones as a whole, RestoreNet provides a clone-specific quantification of dominance in terms of conditional mean and variance of the expansion rates. In conclusion, our proposed stochastic framework allows to detect deviant clonal behaviour relative to the average dynamics of hematopoiesis. This is an important aspect for gene therapy applications where is crucial to quickly detect clonal dominance to prevent any adverse event that may be related to malignant scenarios. Therefore our tool can provide statistical support to biologists in gene therapy surveillance analyeses. With slight modifications our framework can be applied to every study of population dynamics that can be described with an Ito-type formulation, even when the whole population needs to be drawn from an hierarchical structure having subject-specific dynamics.

**S1 Text Stochastic inference of clonal dominance in gene therapy studies** (PDF)

## Author Contributions

- All authors analysed the data and wrote the paper.
- L.D.C. designed and implemented the stochastic framework.

## Data availability

The data that supports the findings of this study is openly available at [27].

## Code availability

- The code that supports the findings of this study is openly available at `https://github.com/delcore-luca/ClonalDominance`
- The stochastic framework is implemented in the R package RestoreNet available at `https://github.com/delcore-luca/RestoreNet`

## Acknowledgements

- This publication is based on work from COST Action CA15109 (COSTNET), supported by COST (European Cooperation in Science and Technology).
- E.C.W. acknowledges support from the Fondazione Leonardo (514.7.010.098-4) and funding from the Swiss National Science Foundation (SNSF 188534)
- The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

## References

1. Friedmann T, Roblin R. Gene Therapy for Human Genetic Disease? Science. 1972;175(4025):949–955. doi:10.1126/science.175.4025.949.
2. Bryder D, Rossi DJ, Weissman IL. Hematopoietic stem cells: the paradigmatic tissue-specific stem cell. The American journal of pathology. 2006;169(2):338–346.
3. Kustikova OS, Wahlers A, Kühlcke K, Stähle B, Zander AR, Baum C, et al. Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. Blood. 2003;102(12):3934–3937.
4. Fehse B, Kustikova O, Bubenheim M, Baum C. Pois (s) on–it's a question of dose.... Gene therapy. 2004;11(11):879–881.
5. Baum C, Düllmann J, Li Z, Fehse B, Meyer J, Williams DA, et al. Side effects of retroviral gene transfer into hematopoietic stem cells. Blood, The Journal of the American Society of Hematology. 2003;101(6):2099–2113.
6. Modlich U, Kustikova OS, Schmidt M, Rudolph C, Meyer J, Li Z, et al. Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. Blood. 2005;105(11):4235–4246.
7. Baum C, Kustikova O, Modlich U, Li Z, Fehse B. Mutagenesis and oncogenesis by chromosomal insertion of gene transfer vectors. Human gene therapy. 2006;17(3):253–263.
8. Fehse B, Roeder I. Insertional mutagenesis and clonal dominance: biological and statistical considerations. Gene therapy. 2008;15(2):143–153.
9. Catlin SN, Guttorp P, Abkowitz JL. The kinetics of clonal dominance in myeloproliferative disorders. Blood. 2005;106(8):2688–2692.
10. Roeder I, Horn M, Glauche I, Hochhaus A, Mueller MC, Loeffler M. Dynamic modeling of imatinib-treated chronic myeloid leukemia: functional insights and clinical implications. Nature medicine. 2006;12(10):1181–1184.
11. Müller-Sieburg CE, Cho RH, Thoman M, Adkins B, Sieburg HB. Deterministic regulation of hematopoietic stem cell self-renewal and differentiation. Blood, The Journal of the American Society of Hematology. 2002;100(4):1302–1309.
12. Roeder I, Kamminga LM, Braesel K, Dontje B, de Haan G, Loeffler M. Competitive clonal hematopoiesis in mouse chimeras explained by a stochastic model of stem cell organization. Blood. 2005;105(2):609–616.

13. Sieburg HB, Cho RH, Dykstra B, Uchida N, Eaves CJ, Muller-Sieburg CE. The hematopoietic stem compartment consists of a limited number of discrete stem cell subsets. Blood. 2006;107(6):2311–2316.

14. Loeffler M, Birke A, Winton D, Potten C. Somatic mutation, monoclonality and stochastic models of stem cell organization in the intestinal crypt. Journal of theoretical biology. 1993;160(4):471–491.

15. Loeffler M, Bratke T, Paulus U, Li Y, Potten C. Clonality and life cycles of intestinal crypts explained by a state dependent stochastic model of epithelial stem cell organization. Journal of Theoretical Biology. 1997;186(1):41–54.

16. Loeffler M, Roeder I. Tissue stem cells: definition, plasticity, heterogeneity, self-organization and models–a conceptual approach. Cells Tissues Organs. 2002;171(1):8–26.

17. Meineke FA, Potten CS, Loeffler M. Cell migration and organization in the intestinal crypt using a lattice-free model. Cell proliferation. 2001;34(4):253–266.

18. Roeder I, Braesel K, Lorenz R, Loeffler M. Stem cell fate analysis revisited: interpretation of individual clone dynamics in the light of a new paradigm of stem cell organization. Journal of biomedicine and biotechnology. 2007;2007.

19. Winton D, Blount M, Ponder B. A clonal marker induced by mutation in mouse intestinal epithelium. Nature. 1988;333(6172):463–466.

20. Park HS, Goodlad RA, Wright NA. Crypt fission in the small intestine and colon. A mechanism for the emergence of G6PD locus-mutated crypts after treatment with mutagens. The American journal of pathology. 1995;147(5):1416.

21. Bjerknes M, Cheng H. Modulation of specific intestinal epithelial progenitors by enteric neurons. Proceedings of the National Academy of Sciences. 2001;98(22):12497–12502.

22. Potten CS, Booth C, Pritchard DM. The intestinal epithelial stem cell: the mucosal governor. International journal of experimental pathology. 1997;78(4):219–243.

23. Pellin D. Stochastic modelling of dynamical systems in biology [PhD thesis]. University of Groningen; 2017.

24. Pellin D, Biasco L, Aiuti A, Di Serio MC, Wit EC. Penalized inference of the hematopoietic cell differentiation network via high-dimensional clonal tracking. Applied Network Science. 2019;4(1):1–26.

25. Bailey NTJ. The Elements of Stochastic Processes with Applications to the Natural Sciences. Wiley Classics Library. Wiley; 1990. Available from: https://books.google.it/books?id=yHPnwl4QOfIC.

26. Kloeden PE, Platen E. Numerical Solution of Stochastic Differential Equations. Stochastic Modelling and Applied Probability. Springer Berlin Heidelberg; 2011. Available from: https://books.google.it/books?id=BCvtssom1CMC.

27. Wu C, Li B, Lu R, Koelle SJ, Yang Y, Jares A, et al. Clonal tracking of rhesus macaque hematopoiesis highlights a distinct lineage origin for natural killer cells. Cell Stem Cell. 2014;14(4):486–499.

28. Lu R, Neff N, Quake S, Weissman I. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature Biotechnology. 2011;29:928–33. doi:10.1038/nbt.1977.

29. Nakamura T, Omasa T. Optimization of cell line development in the GS-CHO expression system using a high-throughput, single cell-based clone selection system. Journal of Bioscience and Bioengineering. 2015;120(3):323 – 329. doi:https://doi.org/10.1016/j.jbiosc.2015.01.002.

30. Gerrits A, Dykstra B, Kalmykowa OJ, Klauke K, Verovskaya E, Broekhuis MJC, et al. Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood. 2010;115(13):2610–2618. doi:10.1182/blood-2009-06-229757.

31. Harkey MA, Kaul R, Jacobs MA, Kurre P, Bovee D, Levy R, et al. Multiarm High-Throughput Integration Site Detection: Limitations of LAM-PCR Technology and Optimization for Clonal Analysis. Stem Cells and Development. 2007;16(3):381–392. doi:10.1089/scd.2007.0015.

32. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods. 2008;5(1):16–18. doi:10.1038/nmeth1156.

33. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005;437(7057):376–380. doi:10.1038/nature03959.

34. Demkow U, Ploski R. Clinical Applications for Next-Generation Sequencing. Elsevier Science; 2015. Available from: https://books.google.it/books?id=3f_IBAAAQBAJ.

35. Shendure J, Mitra RD, Varma C, Church GM. Advanced sequencing technologies: methods and goals. Nature Reviews Genetics. 2004;5(5):335–344. doi:10.1038/nrg1325.

36. Biasco L, Pellin D, Scala S, Dionisio F, Basso-Ricci L, Leonardelli L, et al. In¬†Vivo Tracking of Human Hematopoiesis Reveals Patterns of Clonal Dynamics during Early and Steady-State Reconstitution Phases. Cell Stem Cell. 2016;19(1):107 – 119. doi:https://doi.org/10.1016/j.stem.2016.04.016.

37. Wu C, Li B, Lu R, Koelle S, Yang Y, Jares A, et al. Clonal Tracking of Rhesus Macaque Hematopoiesis Highlights a Distinct Lineage Origin for Natural Killer Cells. Cell Stem Cell. 2014;14(4):486 – 499. doi:https://doi.org/10.1016/j.stem.2014.01.020.

38. Mazurier F, Gan OI, McKenzie JL, Doedens M, Dick JE. Lentivector-mediated clonal tracking reveals intrinsic heterogeneity in the human hematopoietic stem cell compartment and culture-induced stem cell impairment. Blood. 2004;103(2):545–552. doi:10.1182/blood-2003-05-1558.

39. Biasco L, Rothe M, Schott JW, Schambach A. Integrating Vectors for Gene Therapy and Clonal Tracking of Engineered Hematopoiesis. Hematology/Oncology Clinics. 2017;31(5):737–752. doi:10.1016/j.hoc.2017.06.009.

40. Kim HJ, Tisdale JF, Wu T, Takatoku M, Sellers SE, Zickler P, et al. Many multipotential gene-marked progenitor or stem cell clones contribute to hematopoiesis in nonhuman primates. Blood, The Journal of the American Society of Hematology. 2000;96(1):1–8.

41. Shepherd BE, Kiem HP, Lansdorp PM, Dunbar CE, Aubert G, LaRochelle A, et al. Hematopoietic stem-cell behavior in nonhuman primates. Blood, The Journal of the American Society of Hematology. 2007;110(6):1806–1813.

42. Lu R, Neff NF, Quake SR, Weissman IL. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nature biotechnology. 2011;29(10):928–933.

43. Dobson AJ, Barnett AG. An Introduction to Generalized Linear Models. Chapman & Hall/CRC Texts in Statistical Science. CRC Press; 2018. Available from: https://books.google.it/books?id=kIhnDwAAQBAJ.

44. Vaida F, Blanchard S. Conditional Akaike Information for Mixed-Effects Models. Biometrika. 2005;92(2):351–370.

45. Wu C, Espinoza DA, Koelle SJ, Yang D, Truitt L, Schlums H, et al. Clonal expansion and compartmentalized maintenance of rhesus macaque NK cell subsets. Science immunology. 2018;3(29):eaat9781.

46. Wu C, Mortlock RD, Shin T, Cordes S, Fan X, Brenchley J, et al. Tissue-Resident Clonal Expansions of Rhesus Macaque NK Cells. Blood. 2021;138:998.

47. Del Core L, Cesana D, Gallina P, Secanechia YNS, Rudilosso L, Montini E, et al. Normalization of clonal diversity in gene therapy studies using shape constrained splines. Scientific Reports. 2022;12(1):3836.

# S1 Text: Stochastic inference of clonal dominance in gene therapy studies

L. Del Core et al.
l.del.core@rug.nl

# 1 Mathematical details

## 1.1 Stochastic quasi-reaction networks

Stochastic quasi-reaction networks (S-QRN) allow to implement a particular class of stochastic differential equations that can be used to model biochemical reactions. More formally, let

$$Y_t = (Y_{1t}, \dots, Y_{Nt})$$

be a collection of molecules of $N$ different types observed at time $t$, and consider $K$ distinct (and competing) reactions

$$r_{j1}Y_1 + \cdots + r_{jN}Y_N \xrightarrow{\theta_j} p_{j1}Y_1 + \cdots + p_{jN}Y_N \qquad j = 1, \dots, K \tag{1}$$

each occurring with its own rate $\theta_j$. The coefficients $r_{ji}$'s defining the left-side of the reaction are called reagents and represent the minimum amount of molecules of type $i$ needed for the $j$-th reaction to occur. Similarly, the coefficients $p_{ji}$ defining the right-side of the reaction are called products and represents the amount of produced molecules of type $i$ after the $j$-th reaction is triggered. We assume that, if we observe $Y_0 = (r_{j1}, \dots, r_{jN})$ molecules at time $t = 0$, the $j$-th reaction will occur after

$$T_j \sim exp(\theta_j), \qquad j = 1, \dots, K$$

Namely, if exactly $r_{ij}$ molecules of each type $i$ would be present, then the $j$-th reaction can only take place in one way, with the exponential hazard rate $\theta_j$. The interpretation is that, after a waiting time $T_j$, $r_{ji}$ molecules of type $i$ collide with each other and produce $p_{ji}$ molecules of type $i$ ($\forall i = 1, \dots, N$), while the molecules move randomly in a hosting "cellular" environment. However, in general at time $t = 0$ we might observe $Y_{i0} \geq r_{ji}$ molecules of each type $i$ and, therefore, the $j$-th reaction can take place in a combinatorial number of ways leading to the following waiting time formulation

$$T_j \sim exp\underbrace{\left(\theta_j \prod_{i=1}^{N} \binom{Y_{i0}}{r_{ji}}\right)}_{h(Y_0,\theta)}, \qquad \texttt{where } \binom{x}{y} = 0 \texttt{ for } x < y \tag{2}$$

In this case, the effect will be that at time $t + T_j$ we have the following expression for the number of molecules of substrate $i$,

$$Y_{i,t+T_j} = Y_{it} + p_{ji} - r_{ji} = Y_{it} + v_{ji} \tag{3}$$

where $v_{ji} = p_{ji} - r_{ji}$ is the $j$-th net effect. More compactly, for a set of $K$ reactions and $N$ species, the molecular transfer from reagent to product species is a net change of

$$V = P - R$$

**Algorithm 1:** $\tau$-leaping algorithm

```
Input: $S$ (no. simulations), $Y_0$(initial state), $\tau$ (time lag),
       $\theta(t)$ (reaction rates)
Output: $\{Y_t\}_t$
$t \leftarrow 0$;
$Y_t \leftarrow Y_0$;
for $s = 1 : S$ do
    for $r = 1 : K$ do
        $\Theta_t^r = \int_t^{t+\Delta t} \theta_r(s) ds$ ;
    end
    $Y_{t+\Delta t} \leftarrow Y_t + V \begin{bmatrix} \Theta_t^1 \prod_{i=1}^p \binom{Y_{it}}{r_{1i}} \\ \vdots \\ \Theta_t^K \prod_{i=1}^p \binom{Y_{it}}{r_{Ki}} \end{bmatrix}$;
    $t \leftarrow t + \tau$;
end
```

where $P = [p_{ji}]'$ denotes the $N \times r$ dimensional matrix of products, $R = [r_{ji}]'$ is the $N \times r$ dimensional matrix of reactants, and $V = [v_{ji}]'$ is an $N \times r$ dimensional matrix called net-effect matrix. Therefore, a S-QRN of $K$-distinct reactions is fully identified by a net-effect matrix $V$ and by the hazard vector

$$h(Y, \theta) = \begin{bmatrix} h_1(Y, \theta) & \cdots & h_K(Y, \theta) \end{bmatrix}'$$

## 1.2 Simulating a trajectory of molecules

A $\tau$-leaping algorithm is an alternative method to a Gillespie algorithm for simulating triggering-chain events. Instead of simulating a waiting time for the first reaction to occur and selecting the corresponding winner reaction, a $\tau$-leaping algorithm simulates the number of occurrences of each possible event after a time-lag equal to $\tau$ elapsed. Formally, let $\{N_r(t)\}_{t \geq 0}$ be an inhomogeneous Poisson point process representing the number of reactions of type $r$ that took place up to (and including) time $t$. Therefore

$$N_r(t) \sim Poisson\left(\int_0^t \theta(s) ds\right) \quad \texttt{and} \tag{4}$$

$$E[N_r(t + \Delta t) - N_r(t)] = \int_t^{t+\Delta t} \theta(s) ds \mathrel{\widehat{=}} \Theta_t^r \tag{5}$$

The last equation gives an estimate of the expected number of reactions of type $r$ that took place within the time interval $[t, t + \Delta_t[$. Furthermore, the expected number of molecules $Y_{t+\Delta t}$ at time $t + \Delta t$ given the current number of molecules $Y_t$ is given by

$$Y_{t+\Delta t} = Y_t + V \begin{bmatrix} \Theta_t^1 \prod_{i=1}^p \binom{Y_{it}}{r_{1i}} \\ \vdots \\ \Theta_t^K \prod_{i=1}^p \binom{Y_{it}}{r_{Ki}} \end{bmatrix} \tag{6}$$

The pseudocode of the $\tau$-leaping algorithm ($\tau$-LA) is reported in Algorithm 1.

## 1.3 Inference of the rates

### 1.3.1 Local linear approximation

In order to estimate the rates $\theta = \begin{bmatrix}\theta_1 & \cdots & \theta_K\end{bmatrix}'$, we focus on the first two-order moments of the process, that is we consider the Ito equation

$$dY_t = \mu(dY_t;\theta)dt + \beta^{1/2}(dY_t;\theta)dW(t) \qquad dW(t) \sim N(0, dtI) \tag{7}$$

where $dY_t = Y_{t+dt} - Y_t$ is an infinitesimal time drift and $\mu(dY_t;\theta)$ and $\beta(dY_t;\theta)$ are called mean-drift and diffusion respectively. Given a $\sigma$-algebra $(\Omega, \mathscr{F}, \mathbb{P})$, the solution

$$Y : [0, +\infty) \times \Omega \to \mathbb{R}^N$$

of (7) is called Ito diffusion. Instead of finding the Ito diffusion itself, we focus on the first two-order moments $\mu(dY_t;\theta)$ and $\beta(dY_t;\theta)$ of the infinitesimal time drift $dY_t$ which can be approximated with the following Lemma and Proposition.

**Lemma 1.** *Given the hazard function as a limit of a conditional probability*

$$h(t) = \lim_{dt\to 0} \frac{\mathsf{P}(T < t + dt|T > t)}{dt}$$

*for small dt the following approximation holds*

$$h(t)dt \approx \mathsf{P}(T < t + dt|T > t)$$

*Furthermore, the event* $\{Y_{t+dt} - Y_t = V_{\cdot j}\} \equiv \{Y_{t+dt} = Y_t + V_{\cdot j}\}$ *occurs with probability* $\mathsf{P}(T_j < t + dt|T_j > t)$.

**Proposition 1.** *An approximation of the mean drift* $\mu(dY_t;\theta)$ *and the diffusion* $\beta(dY_t;\theta)$ *for a small time increment dt is given by*

$$\mu(dY_t;\theta) \underset{\texttt{small}\ dt}{\approx} Vh(Y_t,\theta) \tag{8}$$

$$\beta(dY_t;\theta) \underset{\texttt{small}\ dt}{\approx} V \underbrace{\begin{bmatrix} h_1(Y_t;\theta) & & \\ & \ddots & \\ & & h_K(Y_t;\theta)\end{bmatrix}}_{diag(h(Y_t,\theta))} V' \tag{9}$$

*Proof.*

$$\begin{aligned}\mu(dY_t;\theta) &\,\widehat{=}\, \lim_{dt\to 0} \frac{E[dY_t|Y_t,\theta]}{dt} \\ &= \lim_{dt\to 0} \frac{\sum_{j=1}^K \mathsf{P}(T_j < t + dt|T_j > t)V_{\cdot j}}{dt} \\ &\underset{\texttt{small}\ dt}{\approx} \lim_{dt\to 0} \frac{\sum_{j=1}^K h_j(Y_t,\theta)V_{\cdot j}}{dt} = Vh(Y_t,\theta)\end{aligned}$$

$$\begin{aligned}\beta(dY_t;\theta) &\,\widehat{=}\, \lim_{dt\to 0} \frac{Cov(dY_t|Y_t,\theta)}{dt} \\ &= \lim_{dt\to 0} \frac{E[dY_t dY_t'|Y_t;\theta] - E[dY_t|Y_t;\theta]E[dY_t|Y_t;\theta]'}{dt} \\ &\underset{\texttt{small}\ dt}{\approx} \frac{\sum_{j=1}^K V_{\cdot j}V_{\cdot j}' h_j(Y_t;\theta)dt - Vh(Y_t;\theta)h'(Y_t;\theta)V'dt^2}{dt} \\ &= \sum_{j=1}^K V_{\cdot j}V_{\cdot j}' h_j(Y_t;\theta) = V \begin{bmatrix} h_1(Y_t;\theta) & & \\ & \ddots & \\ & & h_K(Y_t;\theta)\end{bmatrix} V'\end{aligned}$$

□

Using previous results and some linear algebra, the approximated Ito equation (7) can be further approximated as

$$\Delta Y_t = V \overbrace{\begin{bmatrix} \prod_{i=1}^{N} \binom{Y_{it}}{r_{1i}} & & \\ & \ddots & \\ & & \prod_{i=1}^{N} \binom{Y_{it}}{r_{Ki}} \end{bmatrix} \Delta t}^{M_t} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} + \left( \underbrace{V \begin{bmatrix} h_1(Y_t,\theta) & & \\ & \ddots & \\ & & h_1(Y_t,\theta) \end{bmatrix} V'}_{W_t(\theta)} + \sigma^2 I_N \right)^{1/2} \Delta W(t)$$
$$\Delta W(t) \sim \mathscr{N}_N(0, \Delta t I_N) \tag{10}$$

or more compactly

$$\Delta Y_t = M_t \theta + \varepsilon \qquad \Delta W(t) \sim \mathscr{N}_N(0, W_t(\theta) + \sigma^2 I_N) \tag{11}$$

where we included the term $\sigma^2 I_N$ so as to prevent singularity of the diffusion term, and to additionally explain noise variance. In practice, since we collect only discrete-time increments $\Delta Y_t = Y_{t+\Delta t} - Y_t$, we consider an Euler-Maruyama local linear approximation (LLA) of the approximated Ito equation. Indeed we also replaced the infinitesimal increments $dt$ and $dY_t$ with the discrete increments $\Delta t$ and $\Delta Y_t$. Then, all the time-specific blocks can be stacked together obtaining the full generalized linear model (GLM) formulation

$$\underbrace{\begin{bmatrix} \Delta Y_{t_0} \\ \vdots \\ \Delta Y_{t_{T-1}} \end{bmatrix}}_{\Delta Y} = \underbrace{\begin{bmatrix} M_{t_0} \\ \vdots \\ M_{t_{T-1}} \end{bmatrix}}_{M} \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} + \varepsilon; \quad \varepsilon \sim \mathscr{N}_{NT} \left( 0, \overbrace{\underbrace{\begin{bmatrix} W_{t_0}(\theta) & & \\ & \ddots & \\ & & W_{t_{T-1}}(\theta) \end{bmatrix}}_{W(\theta)} + \sigma^2 I_{NT}}^{\Sigma(\theta,\sigma^2)} \right) \tag{12}$$

which is convenient for parameters inference.

### 1.3.2 Maximum Likelihood (ML)

We infer the parameters $(\theta, \sigma^2)$ with a maximum likelihood approach, that is we solve the following constrained optimization problem

$$\hat{\theta}^p_{ML} \leftarrow \underset{\theta \geq 0; \sigma^2 \geq 0}{\texttt{argmin}} \, f(\theta, \sigma^2) \tag{13}$$

where the objective function is

$$f(\theta, \sigma^2) = log(|W_*|) + (dX - M\theta)' W_*^{-1} (dX - M\theta) \tag{14}$$

and we compactly write the diffusion matrix $W_* = W(\theta, \sigma^2)$ as a function of the free parameters. Using the rules of matrix calculus [1], the partial derivatives of $f$ w.r.t. $\theta$ and $\sigma^2$ can be written as

$$\begin{aligned} \nabla_\theta f(\theta, \sigma^2) = {} & \nabla_\theta log(|W_*|) + dX' \nabla_\theta W_*^{-1} dX + 2\theta' M' W_*^{-1} M + \\ & - 2(M' W_*^{-1} + \theta' M' \nabla_\theta W_*^{-1}) dX + \theta' M' \nabla_\theta W_*^{-1} M \theta \end{aligned} \tag{15}$$

$$\begin{aligned} \nabla_{\sigma^2} f(\theta, \sigma^2) = {} & \nabla_{\sigma^2} log(|W_*|) + dX' \nabla_{\sigma^2} W_*^{-1} dX + \\ & - 2\theta' M' \nabla_{\sigma^2} W_*^{-1} dX + \theta' M' \nabla_{\sigma^2} W_*^{-1} M \theta \\ & tr\left(W_*^{-1}\right) - (dX - M\theta)' W_*^{-1} W_*^{-1} (dX - M\theta) \end{aligned} \tag{16}$$

where

$$\frac{\partial}{\partial \theta_j} W_*^{-1} = -W_*^{-1} \frac{\partial}{\partial \theta_j} W_* W_*^{-1}; \quad \frac{\partial}{\partial \theta_j} W_* = W((\dots, \underset{j}{1}, \dots), 0)$$

```
Algorithm 2: Maximum Likelihood inference for the base (null) model.
  Input: $M$, $dX$, $\alpha$, $\Gamma$
  Output: $\hat{\theta}^p_{ML}$
  $\hat{\theta}^p_{ML} \underset{\theta_k}{\leftarrow} \underset{\theta \geq 0; \sigma^2 \geq 0}{\texttt{argmin}} \left\{ log(|W_*|) + (dX - M\theta)' W_*^{-1} (dX - M\theta) \right\}$
```

$$\frac{\partial}{\partial \sigma^2} W_*^{-1} = -W_*^{-1} W_*^{-1}; \quad \frac{\partial}{\partial \theta_j} log|W_*| = tr\left(W_*^{-1} \frac{\partial}{\partial \theta_j} W_*\right); \quad \frac{\partial}{\partial \sigma^2} log|W_*| = tr\left(W_*^{-1}\right)$$

Then, we solve (26) by using the objective function (14) and its gradients (15)-(16) inside the `L-BFGS-B` optimization algorithm from the `optim()` function of the `stats` R package. The inference procedure is summarised in Algorithm 2.

## 1.4 A mixed effects LLA model

In the system (24) all the molecules $Y_1, \ldots, Y_N$ share the same parameter vector $\theta$. In some cases it may happen that the molecules being analysed are drawn from a hierarchy of $J$ different populations having different properties. In this case it might be of interest to quantify the population-average $\theta$ and the subject-specific effects $u$ around the average $\theta$ for the description of the subject-specific dynamics. Therefore, to quantify the contribution of each subject $j = 1, \ldots, J$ on the process's dynamics we extended the LLA (24) by introducing random effects $u$ for the $J$ distinct subjects on the parameter vector $\theta$, leading to the following mixed-effects [2] formulation

$$\Delta Y = \underbrace{\begin{bmatrix} M_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & M_J \end{bmatrix}}_{\mathbb{M} \in R^{N \times Jp}} u + \varepsilon \qquad \varepsilon \sim N(0, \Sigma(\theta, \sigma^2)) \tag{17a}$$

$$u \sim N_{Jp}\left( \underbrace{\mathbf{1}_J \otimes \theta}_{\theta_u}, I_J \otimes \underbrace{\begin{bmatrix} \tau_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \tau_p^2 \end{bmatrix}}_{\Delta_u} \right) \tag{17b}$$

where $\mathbb{M}$ is the block-diagonal design matrix for the random effects $\mathbf{u}$ centered in $\theta$, where each block $M_j$ is subject-specific. As in the case of the null model (24), to explain additional noise of the data and to avoid singularity of the stochastic covariance matrix $W(\theta)$ we added to its diagonal a small unknown quantity $\sigma^2$ which we infer from the data. In order to infer the maximum likelihood estimator $\hat{\psi}$ for $\psi = (\theta, \sigma^2, \tau_1^2, \ldots, \tau_p^2)$ we developed an efficient tailor-made expectation-maximization algorithm where $\Delta Y$ and $u$ take the roles of the observed and latent states respectively. Under this framework

$$p(u|\Delta Y) \propto_u p(\Delta Y|u) p(u)$$
$$\propto_u exp\left( -\frac{1}{2} u' (\mathbb{M}' \Sigma^{-1}(\theta, \sigma^2) \mathbb{M} + \Delta_u^{-1}) u + u' (\mathbb{M}' \Sigma^{-1}(\theta, \sigma^2) \Delta Y + \Delta_u^{-1} \theta_u) \right)$$

and therefore

$$u|\Delta Y \sim \mathscr{N}_{Jp}(E_{u|\Delta Y;\psi}[u], V_{u|\Delta Y;\psi}(u))$$

where

$$\begin{aligned} E_{u|\Delta Y;\psi}[u] &= V_{u|\Delta Y;\psi}(u)(\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \Delta_u^{-1}\theta_u) \\ V_{u|\Delta Y;\psi}(u) &= (\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\mathbb{M} + \Delta_u^{-1})^{-1} \end{aligned} \tag{18}$$

Also, the joint log-likelihood of $\Delta Y$ and $u$ is given by

$$\begin{aligned} l(\Delta Y, u;\psi) &\propto_\psi l(\Delta Y|u;\psi) + l(u;\psi) \\ &\propto_\psi -\frac{1}{2}log|\Sigma(\theta,\sigma^2)| - \frac{1}{2}(\Delta Y - \mathbb{M}u)'\Sigma^{-1}(\theta,\sigma^2)(\Delta Y - \mathbb{M}u) + \\ &- \frac{1}{2}log|\Delta_u| - \frac{1}{2}(u-\theta_u)'\Delta_u^{-1}(u-\theta_u) \end{aligned}$$

which only depends on $u$ linearly via its first two-order conditional moments (18). Therefore, it follows for the `E`-step function that

$$\begin{aligned} Q(\psi|\psi^*) &= E_{u|\Delta Y;\psi^*}[l(\Delta Y, u;\psi)] \\ = -\frac{1}{2}log|\Sigma(\theta,\sigma^2)| &- \frac{1}{2}\{\Delta Y'\Sigma^{-1}(\theta,\sigma^2)\Delta Y - 2E_{u|\Delta Y;\psi^*}[u]'\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \\ &+ \texttt{tr}\left(\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\mathbb{M}[V_{u|\Delta Y;\psi^*}(u) + E_{u|\Delta Y;\psi^*}[u]E_{u|\Delta Y;\psi^*}[u']]\right)\} + \\ &- \frac{1}{2}log|\Delta_u| - \frac{1}{2}\texttt{tr}(\Delta_u^{-1}[V_{u|\Delta Y;\psi^*}(u) + E_{u|\Delta Y;\psi^*}[u]E_{u|\Delta Y;\psi^*}[u]']) + \\ &+ E_{u|\Delta Y;\psi^*}[u]'\Delta_u^{-1}\theta_u - \frac{1}{2}\theta_u'\Delta_u^{-1}\theta_u \end{aligned}$$

The gradient of $Q(\psi|\psi^*)$ is defined by the following partial derivatives

$$\begin{aligned} \frac{\partial}{\partial\theta_j}Q(\psi|\psi^*) &= -\frac{1}{2}\texttt{tr}\left(\Sigma^{-1}(\theta,\sigma^2)\frac{\partial}{\partial\theta_j}\Sigma(\theta,\sigma^2)\right) + \\ &- \frac{1}{2}\{-\Delta Y'\Sigma^{-1}(\theta,\sigma^2)\frac{\partial}{\partial\theta_j}\Sigma(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \\ &+ 2E_{u|\Delta Y;\psi^*}[u]'\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\frac{\partial}{\partial\theta_j}\Sigma(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \\ &+ \texttt{tr}(-\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\frac{\partial}{\partial\theta_j}\Sigma(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2))\mathbb{M}[V_{u|\Delta Y;\psi^*}(u) + E_{u|\Delta Y;\psi^*}[u]E_{u|\Delta Y;\psi^*}[u]']\} + \\ &+ E_{u|\Delta Y;\psi^*}[u]'\Delta_u^{-1}\frac{\partial}{\partial\theta_j}\theta_u - \theta_u'\Delta_u^{-1}\frac{\partial}{\partial\theta_j}\theta_u \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\tau_j}Q(\psi|\psi^*) &= -\frac{1}{2}\texttt{tr}\left(\Delta_u^{-1}\frac{\partial}{\partial\tau_j}\Delta_u^{-1}\right) + \\ &- \frac{1}{2}\texttt{tr}\left(\Delta_u^{-1}\frac{\partial}{\partial\tau_j}\Delta_u^{-1}\Delta_u^{-1}\left[V_{u|\Delta Y;\psi^*}(u) + E_{u|\Delta Y;\psi^*}[u]E_{u|\Delta Y;\psi^*}[u]'\right]\right) + \\ &- E_{u|\Delta Y;\psi^*}[u]'\Delta_u^{-1}\frac{\partial}{\partial\tau_j}\Delta_u^{-1}\Delta_u^{-1}\theta_u + \frac{1}{2}\theta_u'\Delta_u^{-1}\frac{\partial}{\partial\tau_j}\Delta_u^{-1}\Delta_u^{-1}\theta_u \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\sigma^2}Q(\psi|\psi^*) &= -\frac{1}{2}\texttt{tr}(\Sigma^{-1}(\theta,\sigma^2)) - \frac{1}{2}\{-\Delta Y'\Sigma^{-1}(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \\ &+ 2E_{u|\Delta Y;\psi^*}[u]'\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2)\Delta Y + \\ &+ \texttt{tr}(-\mathbb{M}'\Sigma^{-1}(\theta,\sigma^2)\Sigma^{-1}(\theta,\sigma^2)\mathbb{M}[V_{u|\Delta Y;\psi^*}(u) + E_{u|\Delta Y;\psi^*}[u]E_{u|\Delta Y;\psi^*}[u]'])\} \end{aligned}$$

```
Algorithm 3: EM algorithm for the mixed-effects LLA model
Input: $(\theta_s, \sigma_s^2) = (\hat{\theta}_{fixed}, \hat{\sigma}^2_{fixed})$ starting point
Output: $\hat{\psi}_{EM}$
chose a small tolerance tol and set $\epsilon_k = +\infty$ ;
while $\epsilon_k > tol$ do
    update $E_{u|\Delta Y;\psi}[u]$ and $V_{u|\Delta Y;\psi}(u)$ as defined in (18) ;
    set to zero the negative elements of $E_{u|\Delta Y;\psi}[u]$ ;
    update $Q(\psi|\psi^*)$ and $\nabla_\psi Q(\psi|\psi^*)$ according to (1.4) - (1.4) ;
    set $\psi_{old} \leftarrow \psi^*$ ;
    update $\psi^* \leftarrow \underset{\psi \geq 0}{\text{ARGMAX}} Q(\psi|\psi^*)$ ;
    update $\epsilon_k = |Q(\psi_{old}|\psi_{old}) - Q(\psi^*|\psi^*)|$ ;
$\hat{\psi}_{EM} = \psi^*$
```

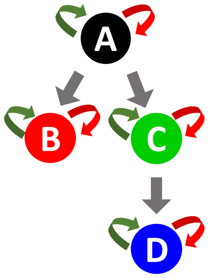


**Fig 1.** Differentiation structure of four synthetic cell types A, B, C, D. Duplication, death and differentiation moves are indicated with green, red and grey arrows respectively.

In the EM-algorithm we iteratively update the E-function $Q(\psi|\psi^*)$ using the current estimate $\psi^*$ of $\psi$ and then we minimize the $-Q(\psi|\psi^*)$ w.r.t. $\psi$. The EM algorithm is run until a convergence criterion is met, that is when the relative errors of both the E-step function $Q(\psi|\psi^*)$ and the vector parameter $\psi$ are lower than a predefined tolerance. Once we get the EM estimate $\hat{\psi}$ for the parameters we evaluate the goodness-of-fit of the mixed-model according to the conditional Akaike Information Criterion [3]. As every EM algorithm, the choice of the starting point $\psi_s$ is very important from a computational point of view. We chose as a starting point $\psi_s = (\theta_s, \sigma_s^2, \tau_1^2 = 0, \ldots, \tau_p^2 = 0)$ where $(\theta_s, \sigma_s^2)$ is the optimum found in the fixed-effects LLA formulation (24). This is a reasonable choice since we want to quantify how the dynamics $E_{u|\Delta Y;\hat{\psi}}[u]_j$ of each subject $j$ departs from the average dynamics $\theta_s$. The EM pseudocode is given in Algorithm 3.

## 2 Simulation studies

Here we mimic the dynamics of $J = 3$ distinct clones in four synthetic cell types A, B, C, D following the differentiation network structure of Figure 4. The net-effect matrix $V$ and hazard vector $h(Y, \theta)$ can be derived from equations (6)-(7) of the main paper. To simulate the clonal tracking data we used the $\tau$-leaping Algorithm 1 with a time lag $\tau = 1$, that has been run independently for each clone. We designed each simulation so that the first clone dominates lineage D and the third clone dominates lineage C. We first run a single simulation under different magnitudes for the noise variance $\sigma^2$. Then we fitted both the base (24) and random-effects (27) models to the simulated data using Algorithms 2 and 3. We reported in Figure 2 the simulated trajectories and a scatterplot of the estimated conditional expectation $E_{u|\Delta Y;\hat{\psi}}[u]$ for the random-effects

| | $\sigma^2 = 1$ | | | $\sigma^2 = 1$ | | | $\sigma^2 = 10$ | | | $\sigma^2 = 100$ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ | $c_1$ | $c_2$ | $c_3$ |
| $\alpha_A$ | 0.183 | 0.191 | 0.198 | 0.183 | 0.191 | 0.198 | 0.151 | 0.139 | 0.127 | 0.000 | 0.105 | 0.050 |
| $\alpha_B$ | 0.146 | 0.148 | 0.145 | 0.146 | 0.148 | 0.145 | 0.163 | 0.148 | 0.137 | 0.442 | 0.337 | 0.358 |
| $\alpha_C$ | 0.163 | 0.168 | 0.518 | 0.163 | 0.168 | 0.518 | 0.166 | 0.175 | 0.649 | 0.300 | 0.214 | 0.958 |
| $\alpha_D$ | 0.450 | 0.100 | 0.121 | 0.450 | 0.100 | 0.121 | 0.479 | 0.199 | 0.319 | 0.499 | 0.691 | 0.778 |
| $\delta A$ | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.000 | 0.001 | 0.000 | 0.000 | 0.000 |
| $\delta B$ | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.008 | 0.007 | 0.007 | 0.016 | 0.014 | 0.015 |
| $\delta C$ | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.004 | 0.005 | 0.005 | 0.008 | 0.006 | 0.008 |
| $\delta D$ | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.002 | 0.003 | 0.004 | 0.002 | 0.010 | 0.007 |
| $\delta_{A \to B}$ | 0.129 | 0.130 | 0.133 | 0.129 | 0.130 | 0.133 | 0.127 | 0.126 | 0.110 | 0.119 | 0.177 | 0.173 |
| $\delta_{A \to C}$ | 0.148 | 0.149 | 0.151 | 0.148 | 0.149 | 0.151 | 0.154 | 0.155 | 0.153 | 0.197 | 0.197 | 0.133 |
| $\delta_{B \to C}$ | 0.079 | 0.080 | 0.078 | 0.079 | 0.080 | 0.078 | 0.082 | 0.079 | 0.058 | 0.081 | 0.084 | 0.059 |

**Table 1.** Conditional expectations $E_{u|\Delta Y;\hat{\psi}}[u]$ of the random-effects obtained from the estimated parameters $\hat{\psi}$ under different magnitudes of the noice variance $\sigma^2$ (outer columns) for each clone (inner columns).

model against the true clone-specific parameters. In the same figure we also show a piechart where each clone $k$ is identified with a pie whose slices are lineage-specific and weighted with $w_k$, defined as the difference between the conditional expectations of the duplication and death parameters, that is $w_k = E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_{lin}}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_{lin}}]$, where `lin` is a cell lineage. The diameter of the $k$-th pie is proportional to the euclidean 2-norm of $w_k$. Therefore, the larger the diameter, the more the corresponding clone is expanding into the lineage associated to the largest slice. The values of the estimated conditional expectations are reported in Table 1. The scatterplot of Figure 2 clearly indicate strong correlation between the true parameters and the conditional expectations $E_{u|\Delta Y;\hat{\psi}}[u]$. In particular, as expected, as the noise variance $\sigma^2$ increases, the parameter estimates gradually move away from the diagonal, so that the correlation decreases. Also, our model correctly detected the dominance of clones 1 and 3 in lineages D and C respectively even for large values of $\sigma^2$, as suggested by the pie-charts of Figure 2 and by the values of Table 1.

Subsequently, to check parameter uncertainty we run $n_{sim} = 100$ independent simulations separately for each noise variance setting. After fitting both the base (24) and the random-effects (27) models, the latter always reached a significantly lower AIC compared to the null model as suggested by the boxplots of Figure 3. In Figure 3 we also report the boxplots of the estimated conditional expectation $E_{u|\Delta Y;\hat{\psi}}[u]$, obtained from the independent simulations, divided by the true parameters $\theta_{true}$. Not surprisingly, as the noise variance $\sigma^2$ increases, the parameter estimates get poor, but they still fluctuate around the true values, even under extreme magnitudes of $\sigma^2$. These results clearly show accurate performance of the method in the identification of a simulated clonal dominance and in the inference of the true parameters, regardless of the noise level of the data.

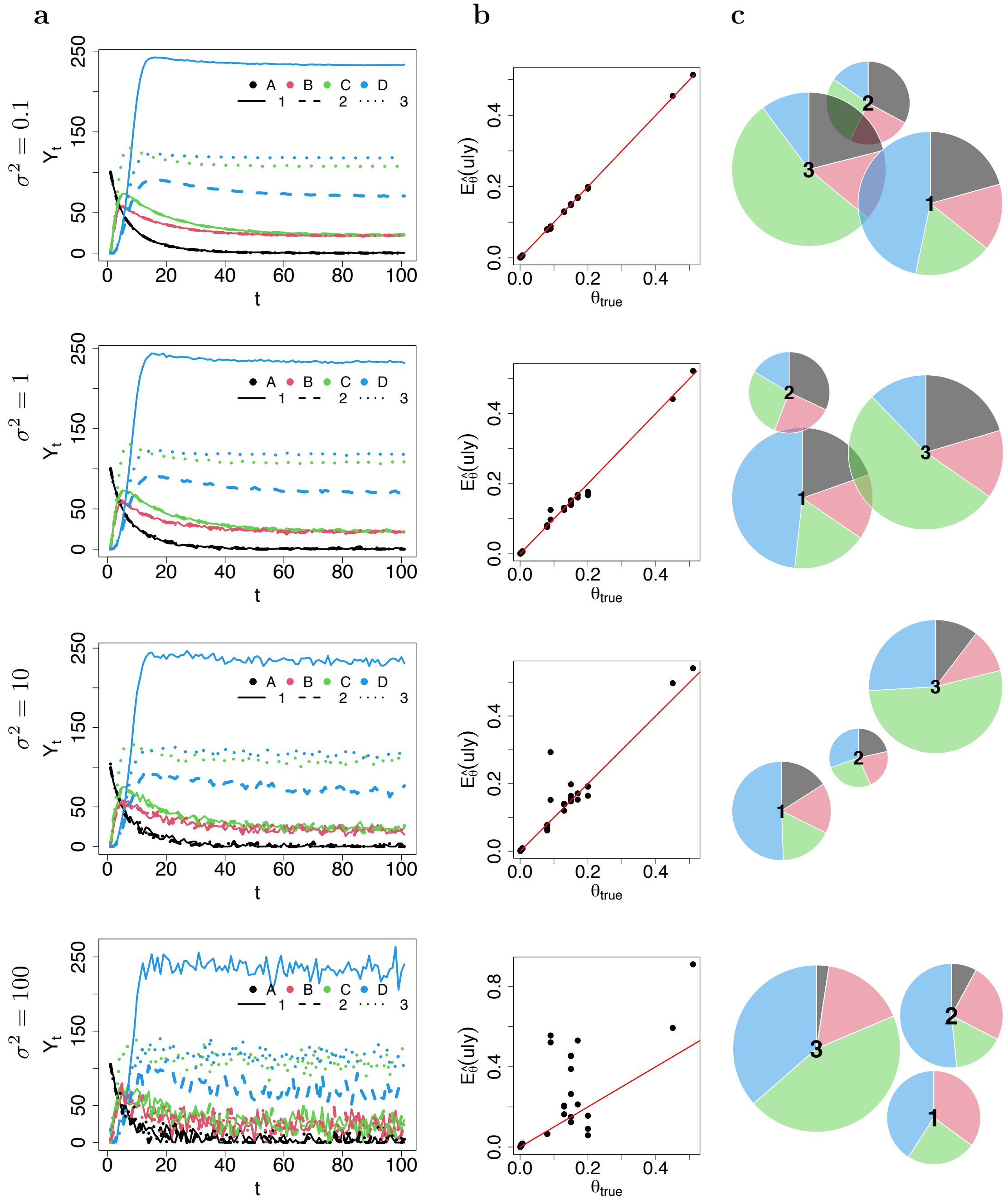


**Fig 2.** (a): Simulated trajectories. (b): Scatterplot between the clone-specific true parameters $\theta_{true}$ and the conditional expectation $E_{u|\Delta Y;\hat{\psi}}[u]$ of the random effects obtained from the estimated parameters $\hat{\psi}$ of the random-effects model. (c): Clonal pie-charts where each clone $k$ is identified with a pie whose slices are lineage-specific and weighted with $w_k = E_{u|y}[u^k_{\alpha_{lin}}] - E_{u|y}[u^k_{\delta_{lin}}]$. The diameter of the $k$-th pie is proportional to the euclidean 2-norm of $w_k$. Each row refers to specific values of the noise variance $\sigma^2$ used for simulations.

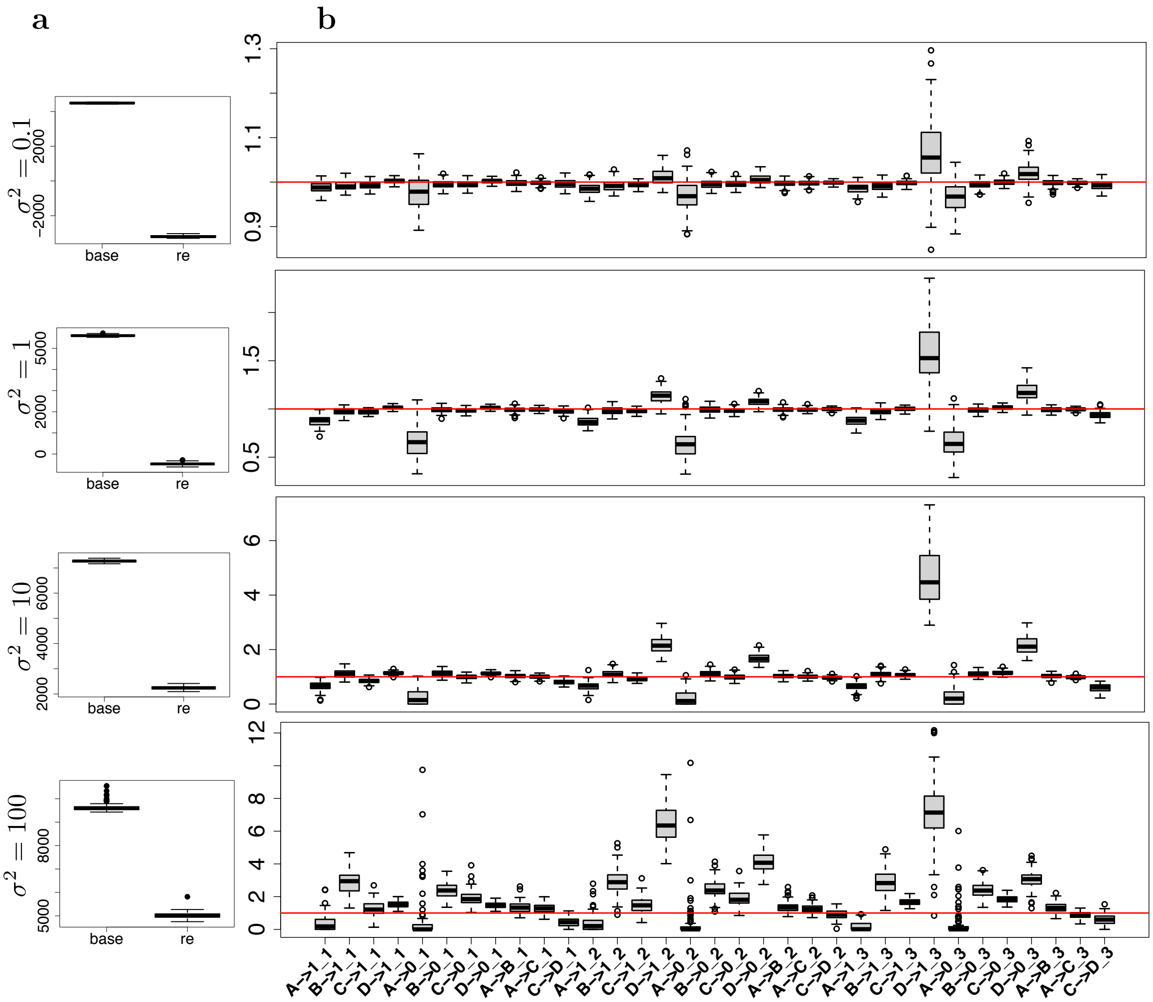


**Fig 3.** Boxplot of the AICs of the base and random-effect models (a) and boxplots of the estimated conditional expectation $E_{u|\Delta Y;\hat{\psi}}[u]$ divided by the corresponding true parameters $\theta_{true}$ obtained under 100 independent simulations (b). Each row refers to a specific value of noise variance $\sigma^2$ used for simulation.

## 3 Rhesus macaque data rescaling

Although the sample DNA amount was maintained constant during the whole experiment (200 ng for ZH33 and ZG66 or 500 ng for ZH17), the sample collected resulted in different magnitudes of total number of reads. Table 2 shows the total number of reads collected in each sample of the rhesus macaque clonal tracking dataset. This discrepancy makes all the samples not comparable across time and cell types. Therefore we rescaled the barcode counts according to

$$Y_{ijk} \leftarrow Y_{ijk} \cdot \frac{\min_{ij} \sum_c Y_{ijc}}{\sum_c Y_{ijc}} \tag{19}$$

where $Y_{ijk}$ is the $ijk$-entry of the barcode matrix with dimensions $(i, j, k)$ mapping respectively time, cell type and clone.

| | | T | B | NK | M | G |
|---|---|---|---|---|---|---|
| ZH33 | 1 | 1465289 | 74735 | 135092 | 119331 | 2831 |
| | 2 | 225797 | 216844 | 335789 | 1035270 | 908685 |
| | 3 | 243986 | 413757 | 663184 | 886682 | 816990 |
| | 4.5 | 485542 | 479493 | 834064 | 985821 | 987171 |
| | 6.5 | 645005 | 676413 | 926089 | 895309 | 911637 |
| | 9.5 | 829073 | 962325 | 1057398 | 1229233 | 1220506 |
| ZH17 | 1 | 51802 | 1347050 | 1288718 | 1351450 | 707382 |
| | 2 | 826190 | 1342700 | 1350703 | 1354355 | 1213749 |
| | 3 | 1303922 | 1347692 | 1338024 | 1347177 | 1283250 |
| | 4.5 | 190591 | 1206361 | 489098 | 572877 | 1195585 |
| | 6.5 | 887851 | 610999 | 1344488 | 381552 | 1339299 |
| ZG66 | 1 | 752127 | 0 | 211350 | 13382 | 0 |
| | 2 | 692133 | 58890 | 308800 | 363310 | 145252 |
| | 3 | 339292 | 209137 | 424458 | 808404 | 704331 |
| | 4.5 | 617281 | 338977 | 718472 | 887183 | 897672 |

**Table 2.** Total number of reads (sum across the different clones) collected in each treated animal at each time point and for all the cell types.

## 4 Estimated parameters

| | ZH33 | | ZH17 | | ZG66 | |
|---|---|---|---|---|---|---|
| | $\theta$ | $\tau^2$ | $\theta$ | $\tau^2$ | $\theta$ | $\tau^2$ |
| $\alpha_T$ | 0.813 | 1.176 | 2.246 | 1.051 | 1.081 | 2.702 |
| $\alpha_B$ | 0.193 | 0.597 | 6.503 | 4.648 | 0.055 | 0.876 |
| $\alpha_{NK}$ | 0.758 | 2.253 | 2.435 | 2.364 | 1.095 | 1.943 |
| $\alpha_G$ | 0.197 | 0.403 | 10.931 | 53.216 | 0.847 | 1.318 |
| $\alpha_M$ | 0.360 | 0.547 | 3.298 | 4.256 | 2.198 | 1.800 |
| $\delta_T$ | 0.155 | 0.074 | 0.172 | 0.741 | 0.039 | 0.059 |
| $\delta_B$ | 0.102 | 0.059 | 2.159 | 36.268 | 0.006 | 0.051 |
| $\delta_{NK}$ | 0.228 | 0.089 | 0.223 | 0.406 | 0.098 | 0.100 |
| $\delta_G$ | 0.039 | 0.029 | 13.211 | 70.756 | 0.018 | 0.017 |
| $\delta_M$ | 0.100 | 0.059 | 0.012 | 0.018 | 0.035 | 0.019 |

**Table 3.** Parameter estimated for proposed mixed effects model: Fixed effects ($\theta$) and variance ($\tau^2$) of the random effects for both the duplication $\alpha$ and death $\delta$ parameters for each cell lineage and each treated animal.

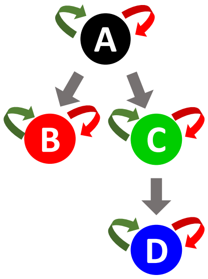


**Fig 4.** Cell differentiation structure of four synthetic cell types A, B, C and D. Duplication, death and differentiation moves are indicated with green, red and grey arrows respectively.

# 5 RestoreNet package: minimal working examples

This section reviews some key functionalities of RestoreNet package. Section 5.1 shows how to simulate a clonal tracking dataset from a stochastic quasi-reaction network. In particular, we show how to simulate clone-specific trajectories, following a given set of biochemical reactions. Sections 5.2 and 5.3 show how to fit the null (base) model and the random-effects model to a simulated clonal tracking dataset. Finally in Section 5.4 we show how to visualize the results at clonal level.

## 5.1 Simulating clonal tracking datasets

A clonal tracking dataset compatible with RestoreNet's functions must be formatted as a 3-dimensional array $Y$ whose $ijk$-entry $Y_{ijk}$ is the number of cells of clone $k$ for cell type $j$ collected at time $i$. The function `get.sim.tl()` can be used to simulate a trajectory of a single clone given an initial conditions $Y_0$ for the cell counts, and obeying to a particular cell differentiation network defined by a list `rct.lst` of biochemical reactions. In particular, our package considers only three cellular events, such as cell duplication ($Y_{it} \to 1$), cell death ($Y_{it} \to \emptyset$) and cell differentiation ($Y_{it} \to Y_{jt}$) for a clone-specific time counting process

$$Y_t = (Y_{1t}, \dots, Y_{Nt}) \tag{20}$$

observed in $N$ distinct cell lineages. The time counting process $Y_t$ for a single clone in a time interval $(t, t + \Delta t)$ evolves according to a set of biochemical reactions defined as

$$v_k = \begin{cases} (0 \dots 1_i \dots 0)' & \text{dup. of the } i\text{-th cell type} \\ (0 \dots -1_i \dots 0)' & \text{death of the } i\text{-th cell type} \\ (0 \dots -1_i \dots 1_j \dots 0)' & \text{diff. of the } i\text{-th type into the } j\text{-th type} \end{cases} \tag{21}$$

with the $k$-th hazard function given by

$$h_k(Y_t, \theta_i) = \begin{cases} Y_{it}\alpha_i & \text{for duplication} \\ Y_{it}^2\delta_i & \text{for death} \\ Y_{it}\lambda_{ij} & \text{for differentiation} \end{cases} \tag{22}$$

Finally, the net-effect matrix and hazard vector are defined as

$$V = \left[v_1 \cdots v_K\right]; \qquad h(Y_t; \theta) = \left[h_1(Y_t; \theta) \cdots h_K(Y_t; \theta)\right]' \tag{23}$$

In particular, the cellular events of duplication, death and differentiation are respectively coded in the package with the character labels `"A->1"`, `"A->0"`, and `"A->B"`, where `A` and B are two distinct cell types. The following R code chunk shows how to simulate clone-specific trajectories of cells via a $\tau$-leaping simulation algorithm. In particular, as an illustrative example we focus on a simple cell differentiation network structure of four synthetic cell types A, B, C and D and only one clone, as illustrated in Figure 4.

```
> library(RestoreNet)
> rcts <- c("A->1", "B->1", "C->1", "D->1",
            "A->0", "B->0", "C->0", "D->0",
            "A->B", "A->C", "C->D") ## set of reactions
> S <- 100 ## trajectory length
> tau <- 1 ## for tau-leaping algorithm
> theta <- c(.2,.15,.17,.09*5,
             .001 , .007 , .004 , .002 ,
             .13, .15, .08) ## parameters
> names(theta) <- rcts
> Y0 <- c(100,0,0,0) ## initial state names(Y0) <- rownames(V)
> names(Y0) <- head(LETTERS ,4)
> s20 <- 1 ## noise variance
> Y <- get.sim.tl(Yt = Y0,
                  theta = theta,
                  S = S,
                  s2 = s20,
                  tau = tau,
                  rct.lst = rcts) ## simulation
> head(Y) ## look at the simulated data
          A           B          C          D
0 100.61983  0.06136727  0.7714631  0.3255576
1  82.64798 25.80389091 30.2276346  0.0000000
2  67.38059 44.75329724 52.8111779  4.9761676
3  59.22818 57.88492115 64.9075555 15.2798701
4  49.95502 57.19943051 73.4204752 32.5405621
5  43.79580 56.15629549 73.4675043 57.1191486
```

## 5.2 Fitting the base model

The base model is defined as

$$\underbrace{\begin{bmatrix} \Delta Y_{t_0} \\ \vdots \\ \Delta Y_{t_{T-1}} \end{bmatrix}}_{\Delta Y} = \underbrace{\begin{bmatrix} M_{t_0} \\ \vdots \\ M_{t_{T-1}} \end{bmatrix}}_{M} \theta + \varepsilon; \quad \varepsilon \sim \mathscr{N}_{NT}\left(0, \overbrace{\underbrace{\begin{bmatrix} W_{t_0}(\theta) & & \\ & \ddots & \\ & & W_{t_{T-1}}(\theta) \end{bmatrix}}_{W(\theta)} + \sigma^2 I_{NT}}^{\Sigma(\theta,\sigma^2)}\right) \tag{24}$$

where

$$\underbrace{Y_{t+\Delta t} - Y_t}_{\Delta Y_t} = \overbrace{V \begin{bmatrix} h_1(Y_t,\theta) & & \\ & \ddots & \\ & & h_K(Y_t,\theta) \end{bmatrix}}^{M_t} \Delta t \underbrace{\begin{bmatrix} \theta_1 \\ \vdots \\ \theta_K \end{bmatrix}}_{\theta} + \left( \underbrace{V \begin{bmatrix} h_1(Y_t,\theta) & & \\ & \ddots & \\ & & h_1(Y_t,\theta) \end{bmatrix} V'}_{W_t(\theta)} + \sigma^2 I_N \right)^{1/2} \Delta W(t) \tag{25}$$

$$\Delta W(t) \sim \mathscr{N}_N(0, \Delta t I_N)$$

Further details can be found in []. The package RestoreNet allows to infer the parameters $(\theta, \sigma^2)$ of (24) with a maximum likelihood approach, that is by solving the following constrained optimization problem

$$\hat{\theta}^p_{ML} \leftarrow \underset{\theta \geq 0; \sigma^2 \geq 0}{\texttt{argmin}}\, f(\theta, \sigma^2) \tag{26}$$

where the objective function $f$ is the negative log-likelihood of the multivariate normal distribution $\mathscr{N}_{NT}\left(M\theta, \Sigma(\theta, \sigma^2)\right)$. The following R code chunk shows how to accomplish this on a clonal tracking dataset simulated from the same differentiation network structure of previous section. In this case we simulate the trajectories of three independent clones following different dynamics of clonal dominance, that is we use clone-specific values for the vector parameter $\theta$.

```
> library(RestoreNet)
> rcts <- c("A->1", "B->1", "C->1", "D->1",
            "A->0", "B->0", "C->0", "D->0",
            "A->B", "A->C", "C->D") ## set of reactions
> ctps <- head(LETTERS ,4)
> nC <- 3 ## number of clones
> S <- 100 ## trajectory length
> tau <- 1 ## for tau-leaping algorithm
> u_1 <- c(.2, .15, .17, .09*5,
           .001, .007, .004, .002,
           .13, .15, .08)
> u_2 <- c(.2, .15, .17, .09,
           .001, .007, .004, .002,
           .13, .15, .08)
> u_3 <- c(.2, .15, .17*3, .09,
           .001, .007, .004, .002,
           .13, .15, .08)
> theta_allcls <- cbind(u_1, u_2, u_3) ## clone-specific parameters
> rownames(theta_allcls) <- rcts
> s20 <- 1 ## additional noise
> Y <- array(data = NA,
           dim = c(S + 1, length(ctps), nC),
           dimnames = list(seq(from = 0, to = S*tau, by = tau),
                           ctps,
                           1:nC)) ## empty array to store simulations
> Y0 <- c(100,0,0,0) ## initial state
> names(Y0) <- ctps
> for (cl in 1:nC) { ## loop over clones
>   Y[,,cl] <- get.sim.tl(Yt = Y0,
                        theta = theta_allcls[,cl],
                        S = S,
                        s2 = s20,
                        tau = tau,
                        rct.lst = rcts)
> }
> null.res <- fit.null(Y = Y, rct.lst = rcts) ## null model fitting
> null.res$ fit ## model fitting results
$par
 [1] 6.788801e-02 2.125983e-02 9.192739e-03 2.753155e-03
        1.000000e-07 2.102263e-03 8.510596e-05 7.137124e-05
 [9] 7.727499e-02 1.147283e-01 3.631258e-02 1.297511e+00

$value
[1] 3419.932

$counts
function gradient
     673      673

$convergence
[1] 0
```

```
$message
[1] "CONVERGENCE:␣REL_REDUCTION_OF_F␣<=␣FACTR*EPSMCH"
> null.res$stats ## model statistics
nPar    cll     mll     cAIC    mAIC    Chi2    p-value
12.000  -2812.692 -2812.692 5649.384 5651.691 337324.840 0.000

> head(null.res$design$M) ## design matrix
6 x 11 sparse Matrix of class "dgCMatrix"
1 100.61983 .   .       .       -10124.350      .       .
1       .       0.06136727      .       .       .       -0.003765942
1       .       .       0.7714631       .       .       .
1       .       .       .       0.3255576       .       .
1  82.64798     .       .       .       -6830.688       .

> null.res$design$V ## net-effect matrix
  A->1 B->1 C->1 D->1 A->0 B->0 C->0 D->0 A->B A->C C->D
A    1    0    0    0   -1    0    0    0   -1   -1    0
B    0    1    0    0    0   -1    0    0    2    0    0
C    0    0    1    0    0    0   -1    0    0    2   -1
D    0    0    0    1    0    0    0   -1    0    0    2
```

### 5.3 Fitting the random-effects model

The random-effects model is defined as

$$\Delta Y = \underbrace{\begin{bmatrix} M_1 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & M_J \end{bmatrix}}_{M \in R^{N \times Jp}} u + \varepsilon \qquad u \sim \mathcal{N}_{Jp}\left( \underbrace{\mathbf{1}_J \otimes \theta}_{\theta_u}, I_J \otimes \underbrace{\begin{bmatrix} \tau_1^2 & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \tau_p^2 \end{bmatrix}}_{\Delta_u} \right) \tag{27a}$$

$$\varepsilon \sim \mathcal{N}_{Jp}(0, \Sigma(\theta, \sigma^2)) \tag{27b}$$

where $Y_{t+\Delta t} - Y_t = \Delta Y$ is the vector of cellular increments that took place in the time interval $\Delta t$, M is the block-diagonal design matrix for the random effects $\mathbf{u}$ centered in $\theta$, $J$ is the number of clones, and each block $M_j$ is clone-specific. As in the case of the null model (24), to explain additional noise of the data, which has the additional advantage of avoiding singularity of the covariance matrix $W(\theta)$, we add to its diagonal a small quantity $\sigma^2$ which we infer from the data. Under this framework (see [] for details) the conditional distribution of the random effects $u$ given the data $\Delta Y$ has the following explicit formulation

$$u|\Delta Y \sim \mathcal{N}_{Jp}(E_{u|\Delta Y;\psi}[u], V_{u|\Delta Y;\psi}(u))$$

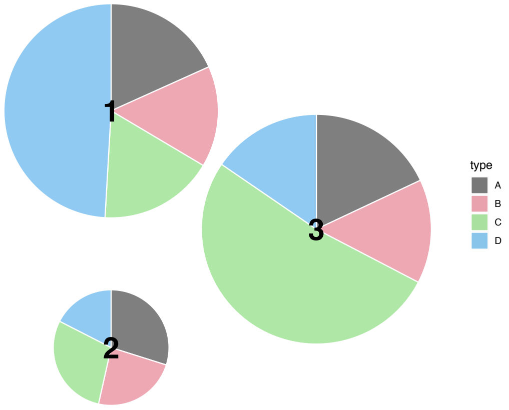
where $E_{u|\Delta Y;\psi}[u]$ and $V_{u|\Delta Y;\psi}(u)$ provide clone-specific mean and variance of the (random) reaction rates. The package RestoreNet allows to infer the vector parameter $\psi = (\theta, \sigma^2, \tau_1^2, \dots, \tau_p^2)$, and in turn to get the corresponding conditional first two-order moments $E_{u|\Delta Y;\psi}[u]$ and $V_{u|\Delta Y;\psi}(u)$, by the means of an efficient tailor-made Expectation-Maximization algorithm where $\Delta Y$ and $u$ take the roles of the observed and latent states respectively. The following R code chunk shows how to accomplish this on the simulated clonal tracking dataset of previous section. In this example we use the optimal parameter vector $\hat{\theta}_0$ estimated for the null model in the previous section, as initial guess for the corresponding parameters in the random-effects model.

```
> re.res <- fit.re(theta_0 = null.res$fit$par,
                   Y = Y,
                   rct.lst = rcts,
                   maxemit = 100) ## random-effects model fitting
> re.res$fit$par ## estimated parameters
 [1] 1.000000e-07 1.843245e-03 1.000000e-07 1.036969e-04
        5.255077e-04 1.000000e-07 1.000000e-07 1.000000e-07
 [9] 1.026921e-03 5.080835e-03 1.000000e-07 3.837475e-02
        2.862468e-02 7.111302e-02 6.109796e-02 1.000000e-07
[17] 4.675422e-05 1.550055e-05 4.952111e-06 1.416910e-02
        2.576975e-02 1.106758e-02 1.720079e+00

> re.res$fit$VEuy$euy ## conditional expected values of u|y
33 x 1 Matrix of class "dgeMatrix"
             [,1]
 [1,] 0.1693522400
 [2,] 0.1478834088
 [3,] 0.1643743275
 [4,] 0.4553735855
 [5,] 0.0006527738
.
.
.
> re.res$fit$VEuy$vuy ## conditional covariance matrix of u|y
33 x 33 sparse Matrix of class "dsCMatrix"

 [1,]  3.552098e-04  2.910616e-05  2.925707e-05 .       .       .
 [2,]  2.910616e-05  2.095979e-04 -3.311544e-07 .       .       .
 [3,]  2.925707e-05 -3.311544e-07  1.478458e-04 .       .       .
        .             .             .
        .             .             .
        .             .             .
```

### 5.4 Visualizing results

The main graphical output of RestoreNet is a clonal piechart. In this representation each clone $k$ is identified with a pie whose slices are lineage-specific and weighted with $w_k$, defined as the difference between the conditional expectations of the random-effects on duplication and death parameters, that is $w_k = E_{u|\Delta Y;\hat{\psi}}[u^k_{\alpha_l}] - E_{u|\Delta Y;\hat{\psi}}[u^k_{\delta_l}]$, where $l$ is a cell lineage. The diameter of the $k$-th pie is proportional to the euclidean 2-norm of $w_k$. Therefore, the larger the diameter, the more the corresponding clone is expanding into the lineage associated to the largest slice. The package RestoreNet includes the function `get.scatterpie()` which returns a clonal piechart given a fitted random-effects model previously obtained with the function `fit.re()`. The following R code chunk illustrates how to obtain a clonal piechart with few lines of R code.

```
> re.res <- fit.re(theta_0 = null.res$fit$par,
                   Y = Y,
                   rct.lst = rcts,
                   maxemit = 100) ## random-effects model fitting
> get.scatterpie(re.res, txt = TRUE) ## get the clonal piechart
```



## References

1. Petersen KB, Pedersen MS. The Matrix Cookbook; 2012. Available from: `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`.

2. Dobson AJ, Barnett AG. An Introduction to Generalized Linear Models. Chapman & Hall/CRC Texts in Statistical Science. CRC Press; 2018. Available from: `https://books.google.it/books?id=kIhnDwAAQBAJ`.

3. Vaida F, Blanchard S. Conditional Akaike Information for Mixed-Effects Models. Biometrika. 2005;92(2):351–370.