Cell Reports
Report

# Determining Lineage Pathways from Cellular Barcoding Experiments

Leïla Perié,[1,2,*] Philip D. Hodgkin,[3,4] Shalin H. Naik,[3,4] Ton N. Schumacher,[1] Rob J. de Boer,[2] and Ken R. Duffy[5]
[1]Division of Immunology, Netherlands Cancer Institute, 1066 CX Amsterdam, the Netherlands
[2]Theoretical Biology and Bioinformatics, Utrecht University, 3584 CH Utrecht, the Netherlands
[3]Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, VIC 3052, Australia
[4]Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia
[5]Hamilton Institute, National University of Ireland, Maynooth, Ireland
*Correspondence: l.perie@nki.nl

http://dx.doi.org/10.1016/j.celrep.2014.01.016

## SUMMARY

**Cellular barcoding and other single-cell lineage-tracing strategies form experimental methodologies for analysis of in vivo cell fate that have been instrumental in several significant recent discoveries. Due to the highly nonlinear nature of proliferation and differentiation, interrogation of the resulting data for evaluation of potential lineage pathways requires a new quantitative framework complete with appropriate statistical tests. Here, we develop such a framework, illustrating its utility by analyzing data from barcoded multipotent cells of the blood system. This application demonstrates that the data require additional paths beyond those found in the classical model, which leads us to propose that hematopoietic differentiation follows a loss of potential mechanism and to suggest further experiments to test this deduction. Our quantitative framework can evaluate the compatibility of lineage trees with barcoded data from any proliferating and differentiating cell system.**

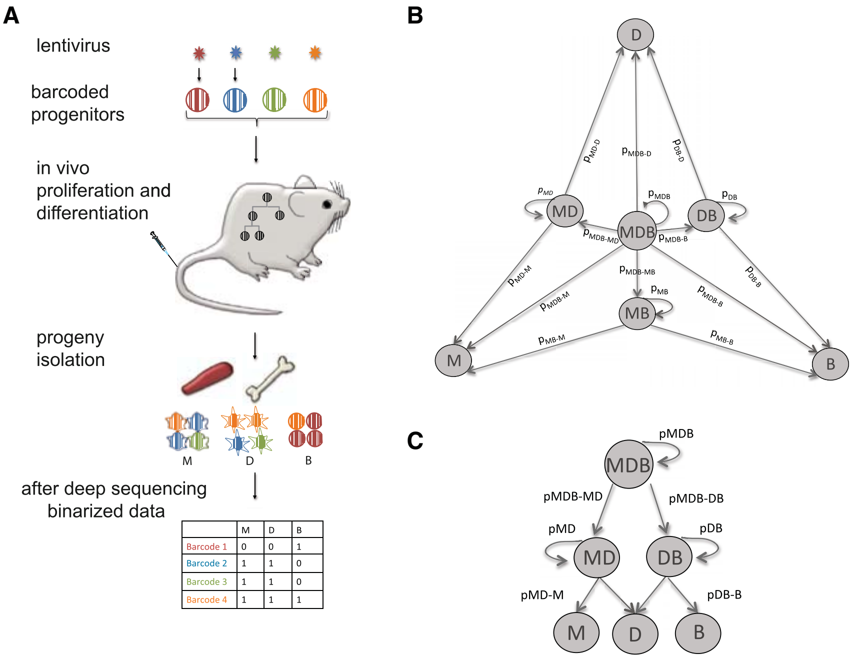
## INTRODUCTION

Cell lineage-tracing techniques are powerful tools to study development, tissue maintenance, and repair. They aim to decipher lineage pathways and to understand cell-fate decisions. Cellular barcoding is an in vivo lineage tracing technique for simultaneously determining the fate of the progeny of multiple initial cells. Extending pioneering approaches that rely on retroviral tagging (Lemischka et al., 1986; Lemischka, 1992), it labels cells with a unique and heritable genetic barcode enabling the identification of familial relationships in progeny (Figure 1A). Fluorescent markers and cell surface markers have also been utilized to follow the output of individual cells (Kretzschmar and Watt, 2012; Livet et al., 2007; Snippert et al., 2010; Buchholz et al., 2013). Cellular barcoding has recently led to significant discoveries in fields such as immunology, hematopoiesis, and cancer (Schepers et al., 2008; van Heijst et al., 2009; Gerrits et al., 2010; Lu et al., 2011; Kreso et al., 2013; Naik et al., 2013). Data from continuously self-renewing stem cells typically reveal dominance of a small number of stem cells, providing little information on the downstream structure of the lineage pathway (Lu et al., 2011; Naik et al., 2013; Grosselin et al., 2013; Verovskaya et al., 2013). In contrast, cellular barcoding data from multipotent progenitors (MPPs) have revealed substantial heterogeneity in the proliferation and differentiation outcomes of apparently identical cells (Naik et al., 2013) and have greater utility for lineage pathway inference.

Although substantial development of the experimental methodology has taken place, a framework for mathematical modeling and statistical testing is required to draw quantitative inferences about the underlying proliferation and differentiation processes. In particular, identification of lineage pathways from cellular barcoding data is fraught with difficulty because the nonlinear dynamics of proliferation and differentiation mislead intuitive deductions. We developed such a quantitative framework for multipotent cells based on transient multitype branching processes, which enabled us to capture the heterogeneity of individual cell fates. The framework identifies the best-fit parameters for any given lineage pathway structure and queries whether the resulting model is statistically consistent with the data. Pathways that are inconsistent with the data are identified and rejected.

We applied this framework to our previously published data from the hematopoietic system (Naik et al., 2013), which has emerged as a model system to study stem cell development (Orkin and Zon, 2008). Hematopoiesis describes the continuous formation of blood cells, which are grouped into two broad lineages: the lymphoid and myeloid lineage. Results from hematopoietic research have been applied in stem cell transplantation to cure blood cell deficiencies caused by irradiation, chemotherapy, and genetic defects (Weissman, 2000). Despite its clinical successes, the hematopoietic pathway of cell differentiation remains poorly understood.

In the most commonly held model of hematopoiesis, which we call the classical model, hematopoietic stem cells (HSCs) self-renew and generate progenitors that differentiate and produce



|  | M | D | B |
|---|---|---|---|
| Barcode 1 | 0 | 0 | 1 |
| Barcode 2 | 1 | 1 | 0 |
| Barcode 3 | 1 | 1 | 0 |
| Barcode 4 | 1 | 1 | 1 |

**Figure 1. Binarizing Cellular Barcoding Data and Possible Lineage Pathways**

(A) In cellular barcoding experiments, progenitors are transduced so that each receives a unique, heritable DNA barcode. After proliferation and differentiation, progeny of the barcoded progenitors are isolated from the bone marrow and the spleen and analyzed for their barcode repertoire by deep sequencing. The myeloid cell group consisted of neutrophils and monocytes measured in the bone marrow and the spleen, the dendritic cell group consisted of CD8$^+$ dendritic cells and plasmacytoid dendritic cells measured in the spleen, and the B cells were measured in the bone marrow and the spleen. The data obtained from deep sequencing are then binarized to identify the cell types in which the barcode was recovered.

(B) In the general tree, multipotent progenitors can divide or differentiate by losing the potential to generate one or two cell types, as shown by the arrow associated with the name of the probability of the event. The letters of their name encode their cell type potentials; for example, an MDB cell has the potential to generate myeloid cells, dendritic cells, and B cells.

(C) The classical tree is a quantitative interpretation of the classical model and is a restriction of the general tree with the probability to lose the dendritic cell potential at the MDB stage and the probability of losing two potentials simultaneously set to zero. MD and DB represent respectively the myeloid and lymphoid branches.

all blood cells. Immediate progenitors of a HSC lose their self-renewal capacities but remain multipotent (Adolfsson et al., 2001; Morrison et al., 1997). These MPPs commit to two separate branches, becoming either common lymphoid progenitors (CLPs) or common myeloid progenitors (CMPs) (Akashi et al., 2000; Kondo et al., 1997; Reya et al., 2001). CLPs give rise to further committed progenitors that produce lymphoid cells, such as T and B lymphocytes and natural killer (NK) cells, whereas CMPs give rise to progenitors that produce granulocytes and monocytes (GMPs), among others, and progenitors that only produce megakaryocytes and erythrocytes (MEPs). Dendritic cells, another group of blood cells, derive from both CLPs and GMPs (Manz et al., 2001; Traver et al., 2000). Not all available data, however, appear consistent with the classical model (Graf, 2008; Kawamoto and Katsura, 2009). For example, in contrast to the presumed myeloid and lymphoid origin of dendritic cells, our recent cellular barcoding data (Naik et al., 2013) from lymphoid-primed multipotent progenitors (LMPPs) (Adolfsson et al., 2001, 2005) established that many LMPPs produced dendritic cells without generating detectable lymphoid and myeloid output, leading us to propose that dendritic cells should be considered a separate lineage of hematopoiesis.

As an illustration of the power of our framework, we used it to test potential hematopoietic lineage pathways for consistency with barcode-labeled LMPPs (Naik et al., 2013). We found that the distribution of cell types generated by LMPPs was statistically consistent across mice, suggesting that the hematopoietic pathway is mouse independent. Furthermore, we found that the data are incompatible with a quantitative interpretation of the classical model. Rather, this analysis provides evidence for additional differentiation paths beyond those found in the classical model, leading us to propose a revised model of the hematopoietic pathway. In this model, hematopoietic differentiation follows a loss of potential mechanism that is proportionally equal at every step of differentiation.

More generally, the quantitative framework is suitable for the analysis of transient multipotent, proliferating, and differentiating cells utilizing any single-cell lineage tracing methodology.

## RESULTS

### The Quantitative Framework

The framework includes binarization of the cellular barcoding data as well as the development and analysis of a stochastic model for drawing inference on lineage pathways. For each lineage pathway, best-fit model results are compared to data with a statistical test. Lineage pathways that are consistent with the data are simulated to garner information on the dynamic properties of proliferation and differentiation and to test for confidence in parameter fits.

The quantitative framework for drawing inferences on lineage pathways begins with recording the presence or absence of each barcode in each cell type, a process that we call binarization (Figure 1A). The reasons for this binarization are 3-fold: (1) due to nonlinearities in PCR amplification for the low-abundant barcodes (see Figure S2 of Naik et al., 2013), cell counts are at best semiquantitative; (2) not all cells in the animal are investigated for their barcodes, and the binarized data are more robust to this sampling; and (3) binarization has the advantage that the proliferation of the final cell types need not be modeled.

Because the output from apparently identical cells consistently exhibits heterogeneity (Snippert et al., 2010; Kaech and Wherry, 2007; Gerlach et al., 2013; Buchholz et al., 2013; Duffy et al., 2012; Gomes et al., 2011; Rieger and Schroeder, 2008; Hasbold et al., 2004; Shortman and Naik, 2007), a stochastic approach was chosen to encapsulate this diversity. Randomness acts as a proxy for the uncertainty in outcome regardless of whether the source is the execution of truly stochastic proliferation and differentiation programs or due to an unidentified heterogeneity in the initial barcoded cohort. The stochastic model is based on a multitype, transient branching process that captures potency restrictions due to differentiation (see Experimental Procedures).

The model begins with the definition of a lineage pathway structure that encodes the differentiation outcomes that are possible. When cells differentiate, they lose the potential to make one or more cell types. Within the model, each cell is assumed to proliferate or differentiate independently of all other cells with a probability that depends upon its current potentialities. This process of division and differentiation from a barcoded progenitor ends when each of its progeny are left with the potential to produce only one cell type. With regards to both intuition and mathematical analysis, it is this proliferation that is the primary confounding factor.

The model quantifies a given lineage pathway by a set of probabilistic parameters. From this we determined explicit mathematical expressions for the probability that a barcoded progenitor produces a given binarized output, which facilitates the fitting of the model to the data. The maximum likelihood best-fit model to data is identified by computer-optimization and compared to the cellular barcoding data via a statistically consistent likelihood-ratio multinomial test that determines if the pathway is rejected by the data. If a lineage pathway with its best-fit parameters is statistically inconsistent with the data, then it is rejected.

Although the mathematical model provides explicit formulae for any finite number, N, of final cell types, the ability to reject a network depends on the quantity and nature of the data. If cellular barcoding experiments were performed for N cell types, then each barcode can be recovered in one of $2^N - 1$ combinations. Due to the nonlinear dynamics of the model, the statistical power is, however, a function not only of N but also of the structure of the empirical multinomial obtained from the data. As a result, we advocate that for a given pathway structure of interest, random sets of binarized outcomes should be independently simulated from the data and the model refit to each set to determine the consistency of parameter estimates.

A typical limitation of cellular barcoding experiments is that animals must be sacrificed, so there is no time-course data. For the best-fit model parameters, a simulation of the lineage pathway can identify timescales within the model in terms of generations (rounds of proliferation). This quantifies the dynamics of events within the model, which can be compared with what is physiologically known.

## Cellular Barcoded LMPPs

As application of the framework is most readily understood by example, we consider its use for data from the hematopoietic system. From the data that we published in (Naik et al., 2013) for barcoded LMPPs, we categorized the final cell types as being of the myeloid (M), dendritic (D), or B cell (B) families and binarized the output of individual barcodes (Figure 1A). The general tree (Figure 1B) encapsulates the lineage pathway of a barcoded LMPP, which has the potential to make M, D, and B cells, with all possible links. Within this framework, our quantitative interpretation of the classical model of hematopoiesis, which we call the classical tree, is a restriction of the general tree such that the initial progenitors are incapable of losing dendritic cell potential (Figure 1C). This restriction is consistent with the early separation between lymphoid and myeloid branches, and the myeloid and lymphoid origin of dendritic cells in the classical model (Akashi et al., 2000; Kondo et al., 1997; Traver et al., 2000; Manz et al., 2001). The framework assumes that individual cell fates are chosen independently so that no feedback mechanism exists between a cell's state and any other's. In addition, we assume that the cellular barcoding data are the result of final differentiation events. This hypothesis seems sound because the heterogeneity of LMPPs has been reported not to dramatically change when assessed on different days (Naik et al., 2013).

Progeny of 978 distinct barcoded LMPPs were identified in four wild-type mice. As reported previously, individual LMPPs were not all multioutcome, but instead LMPPs produced heterogeneous patterns of limited types of blood cells (Figure 2A). The offspring of a single barcoded LMPP cell can be found in one of seven possible combinations of cell categories (M, D, B, M+D, M+B, D+B, or M+D+B), depending on the cell types it produced. The proportion of barcodes recovered in each combination of cell types did not show statistically significant differences between mice (Figure 2B and multinomial test in Table 1). This consistency across mice suggests that the lineage pathway is not a mouse-specific property and justifies pooling the data when considering model fits. Thus, it was appropriate to use our quantitative framework to identify hematopoietic lineage pathways that can explain this data.

## Inferring LMPP Lineage Pathways

We first assessed whether the classical tree (Figure 1C) was consistent with the data. When the six best-fit probabilities of division and differentiation of the classical tree were determined, we obtained visual similarity between the cellular barcoding data and the fit (Figure 2C). The multinomial test, however, showed significant differences (Table 1), strongly rejecting the hypothesis that the classical tree can account for the heterogeneous outcomes observed. The best-fit model to the pooled data was checked against the barcodes taken from each individual mouse and was also rejected as being inconsistent (Table 1). The classical tree does not describe the cellular barcoding data because the prevalence of progenitors producing both myeloid and lymphoid cells without producing dendritic cells (M+B) cannot be made consistent with the number that generate only one of the cell types B or M without D (Table S1). Note that in the classical tree, few progenitors can give rise to myeloid and lymphoid cells without producing dendritic cells. If a progenitor divides at least once before engaging in both the lymphoid (DB) and myeloid (MD) branches, it will most likely produce dendritic cells due to the large numbers of progenitors producing only dendritic

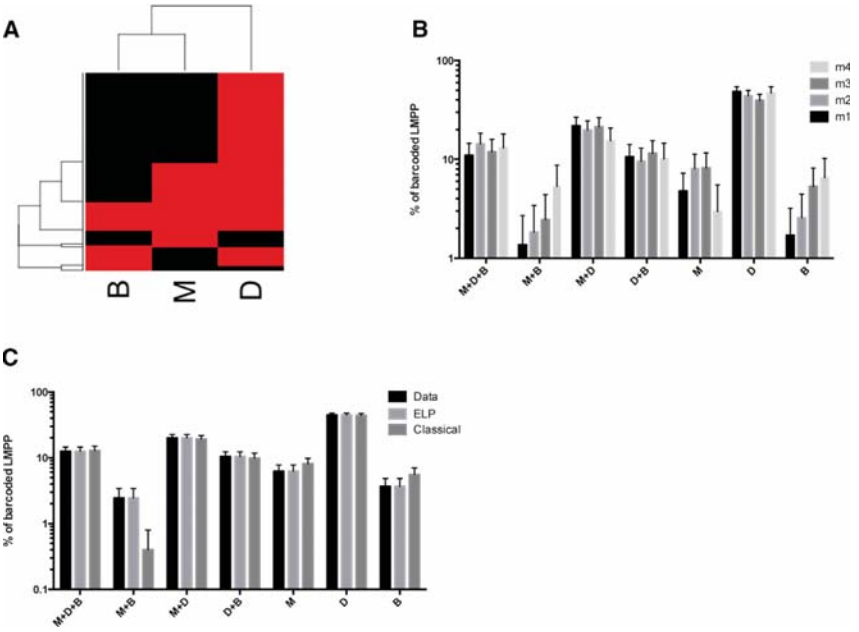


**Figure 2. Application of the Framework to Barcoded Multipotent Progenitors Data**

(A) For one representative mouse, the heatmap shows the binarized output of individual barcoded multipotent progenitors (LMPP in the rows) to the three different cell types (in the columns). The clustering is done using complete linkage and Manhattan distance. Red indicates the presence of the barcode in the cell type and black its absence.

(B) Proportions of progenitors in each of the seven possible combinations of cell types for each of the four mice (with, respectively, 292, 273, 244, and 169 individual barcoded LMPPs).

(C) Same as (B) but for the pooled data from the four mice, for the maximum-likelihood fit values of the classical tree and the equal loss of potential (ELP) model. Using a multinomial test, the best-fit classical tree is rejected ($p = 10^{-10}$), but not the ELP model ($p = 1$). The 95% confidence intervals, based on a normal approximation, are shown as error bars.

cells. This property of the classical tree can be established mathematically for any paramerization of its links (Supplemental Experimental Procedures).

To test the robustness of our conclusions to potential contamination or detection issues, we applied a range of barcode abundance thresholds for the binarization. As a consequence, a barcode that had a read-count below the threshold in a given cell type is not recorded as present in that cell type (Supplemental Experimental Procedures). The percentage of barcodes in each category is not dramatically changed upon application of these different thresholds (Figure S1). When refitted to the new thresholded data, the classical model is only no longer rejected after the least-abundant 37% barcodes are eliminated (Figure S2), adding confidence that our conclusions are robust to potential contamination or detection issues.

Another possible caveat to our rejection of the quantitative interpretation of classical model is that we cannot exclude that the high proportion of progenitors that produce only dendritic cells was due to the effect of irradiation and may not represent the physiological situation. This, however, doesn't change our deduction, because the classical model was itself derived from experiments in irradiated recipient mice (e.g., for the identification of the lymphoid and myeloid progenitor [Akashi et al., 2000; Kondo et al., 1997]).

Removing the restrictions of the classical tree (Figure 1B) and repeating the fitting procedure for the general tree produces a fit that is consistent with the cellular barcoding data (Table 1), indicating that additional differentiation paths are necessary to explain the data. Importantly, no other lineage pathway structure than the general tree is consistent with the cellular barcoding data, because deleting any of the links of the general tree and refitting the corresponding pathway resulted in significant statistical differences (Table 2). Nevertheless, the general tree, which has 12 parameters (4 proliferation and 8 differentiation probabilities), is overparameterized, with several equally good combinations found during fitting. Consequently, we looked for a biologically plausible alternative that includes all the links but reduces the number of parameters.

## Equal Loss of Potential

Several aspects of hematopoietic differentiation have already been modeled using branching processes, such as self-renewal of stem cells (Till et al., 1964; Nakahata et al., 1982; Macken and Perelson, 1988) and proliferation and differentiation of MPPs (Tsuji and Nakahata, 1989; Kurnit et al., 1985), although not for the purpose of lineage pathway identification. While studying hematopiesis with in vitro assays, Ogawa and coworkers (Ogawa et al., 1983; Tsuji and Nakahata, 1989) proposed a model where each cell loses each potential with a fixed probability, irrespective of its current cell type. Based on that intuitive biological principle, we developed a parameterization of the general tree, which we call the equal loss of potential (ELP), that is suitable for the additional difficulties of binarized in vivo cellular barcoding data where proliferation is not directly observed and proliferation rates of different cell types are allowed to be distinct.

In the ELP model, the lineage pathway retains all of the transitions of the general model (Figure 1B) but strictly couples the transition probabilities by insisting that the probability of losing a potential to produce a certain cell type (M, B, or D) remains proportionally equal at every step of differentiation (Figure 3A; Experimental Procedures). The number of parameters is thus reduced to seven: four division probabilities and three loss-of-potentiality rates. The fit obtained from this model didn't show significant differences with the data (Figure 2C; Table 1; Table S1) and is not rejected with a multinomial test. Interestingly, the by-far most probable paths obtained from the ELP fit, starting from MDB cells, are to the lymphoid (DB) and myeloid (MD) branches, drawing a lineage pathway whose main initial routes are akin to those of the classical tree

**Table 1. Multinomial Test Result for the Data and the Different Model Tested**

| Mouse | Multinomial | Multinomial with 5% CIs Holm-Bonferroni Correction | Classical Tree | General Tree | ELP Model |
|---|---|---|---|---|---|
| Pooled mice | NA | | $7.5 \times 10^{-11}$ | 1 | 1 |
| Mouse 1 | 0.19 | 0.02 | 0.0008 | 0.19 | 0.19 |
| Mouse 2 | 0.71 | 0.05 | 0.0399 | 0.71 | 0.70 |
| Mouse 3 | 0.53 | 0.03 | 0.0280 | 0.53 | 0.53 |
| Mouse 4 | 0.04 | 0.01 | $6 \times 10^{-7}$ | 0.04 | 0.04 |

NA, not applicable.

**Table 2. Multinomial Test Result for the Pathway with Each Single Link Removed**

| Link | p Value | Likelihood Ratio |
|---|---|---|
| pMDB | 0 | infinity |
| pMDB→M | $7.5 \times 10^{-11}$ | 58.89 |
| pMDB→D | $7.5 \times 10^{-11}$ | 58.89 |
| pMDB→B | $7.5 \times 10^{-11}$ | 58.89 |
| pMDB→MD | 0 | 118.1 |
| pMDB→MB | $7.5 \times 10^{-11}$ | 58.89 |
| pMDB→DB | 0 | 91.36 |
| pMD | 0 | 118.1 |
| pMB | $7.5 \times 10^{-11}$ | 58.89 |
| pDB | 0 | 91.36 |
| pMD→M | 0 | 117.9 |
| pMD→D | 0 | 117.9 |
| pMB→M | $7.5 \times 10^{-11}$ | 58.89 |
| pMB→B | $7.5 \times 10^{-11}$ | 58.89 |
| pDB→D | 0 | 89.56 |
| pDB→B | 0 | 90.81 |

(Figure 3B). The primary quantitative differences between the best-fit classical tree and the best ELP model are the addition of a rare MB progenitor and the infrequent loss of two potentialities simultaneously. The small contribution of the additional transitions is a possible explanation for their rareness in the data that led to the classical model.

### Simulation of Cellular Barcoding Data

As is typically the case with cellular barcoding data, we have no time-course data and so no estimates of cell lifetimes. Monte Carlo simulation of the best-fit lineage pathway showed that the number of rounds of proliferation until barcoded LMPPs produce their committed cell types is of the order of 20, a value that would allow the process to be completed before the mice were sacrificed after 2 weeks. An example of the generational time course of one such simulated experiment is shown in (Figure 3C). The stochastic nature of the framework leads to significant heterogeneity in repeated simulations with a single LMPP, but population level consistency through the law of large numbers.

To test if the 978 barcodes are sufficient for the population-level consistency, we generated simulated sets of four mice using the best-fit equal loss of potential model, mimicking the experimental data. These simulated data are statistically indistinguishable from their experimental counterpart (Figure 3D). The ELP model was fitted to these simulated data and the best-fit parameters were consistent across fits, giving an indication of their robustness (Figure S3). We repeated this procedure with different numbers of recovered barcodes (Figure S3). The results indicate that for the seven categories of barcoded LMPPs outcome, recovery of approximately 500 barcodes is sufficient to build a clear statistical image in the quantitative framework.

## DISCUSSION

The application of our quantitative framework to data from the hematopoietic development reveals surprising features that result in experimentally testable predictions. The statistical consistency that was found across mice, despite per-progenitor heterogeneity, suggests that the lineage pathway is a robust feature (Figure 2B). Repeating these experiments in other mouse strains would aid in the identification of genetic sources of variability in hematopoietic development. If the distributions between the categories of progenitors are statistically indistinguishable in distinct strains, then it would suggest that much of the lineage pathway is determined from conserved or intrinsic properties of the progenitors and not heavily influenced by the environment. If, instead, the distribution is inconsistent across mouse strains but identical within them, then this would aid in the identification of significant fate factors. These perturbing factors would need to be described and characterized and would inform further experimentation.

Lineage pathway inference establishes that additional links beyond those in the classical model are necessary to explain the data from Naik et al. (2013). Due to the prevalence of single-outcome dendritic cell progenitors in Naik et al. (2013), we previously deduced that dendritic cells form a separate lineage. Results from the quantitative framework suggest that many of the single-outcome dendritic cell progenitors arise through intermediate MPPs (MD or DB) that only give rise to dendritic cells.

Motivated by a biologically meaningful process, the equal loss of potential at every step, we propose a hematopoietic lineage pathway that includes all possible links. Note that the best-fit pathway subject to that constraint possesses the strong myeloid-lymphoid split of the classical model but with essential, albeit less frequently observed, links that could have proved difficult to observe in experiments on populations of cells.

Even though the ELP model gives results in accordance with the cellular barcoding data, one cannot deduce that proliferation and differentiation are genuine stochastic processes, because the results from the ELP model can be interpreted in at least two ways. If the model reflects truly stochastic proliferation and differentiation, then it predicts that all the intermediate cell states of differentiation described in the general tree should be identifiable downstream of LMPPs. The identification of single-outcome progenitors for the lymphoid lineage in the population of MPPs (Medina et al., 2001; Lai et al., 2005) is consistent with this prediction. Other types of progenitors that derive

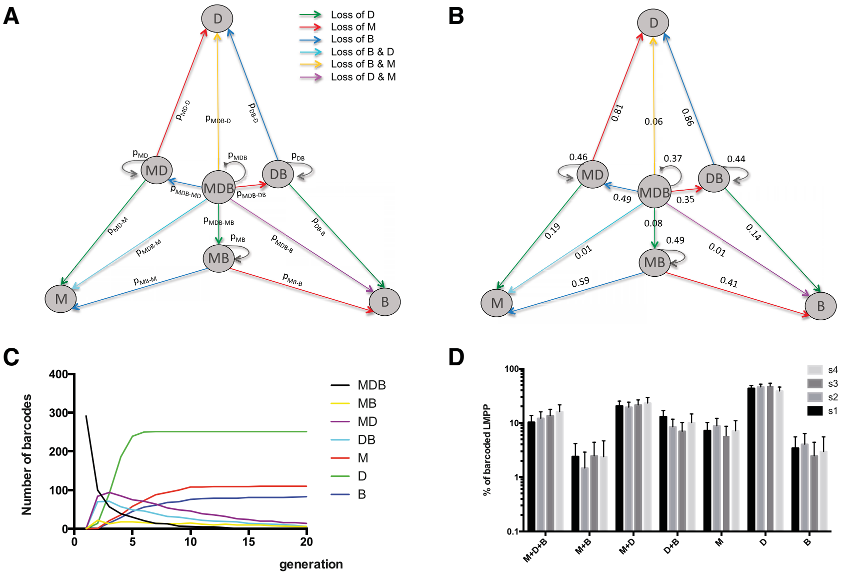


**Figure 3. The Equal Loss of Potential Model**

(A) The equal loss of potential (ELP) model has all of the links of the general model, but its transition probabilities are constrained: the probability of losing a potential to produce a certain cell type (M, red; B, blue; or D, green) remains proportionally equal at every step of differentiation, which is illustrated by use of the same color. The probability to lose two potentials at the same time is the product of the probability of losing each of the potentials.

(B) The best-fit value of the probabilities computed from the seven parameters of the ELP model.

(C) Using the best-fit values from (B), a mouse with 300 barcoded MDB cells was simulated. The number of barcodes at each stage of the pathway at each round of division (generation) is shown.

(D) Similar to Figure 2B, the proportions of progenitors in each of the seven possible combinations of cell type are shown for four simulated mice with the same number of individual barcoded progenitors detected (292, 273, 244 and 169 respectively). Standard deviations are shown as error bars. These simulated mice are statistically consistent with the experimental mice, indicating the sufficiency of the data.

directly from MPPs, such as single-outcome dendritic cells and bioutcome MB progenitors, could then also be identified. If instead the ELP were to encapsulate lineage priming in the initial barcoded cohort that is already present before major clonal expansion, then the model predicts the frequency of occurrence of each unidentified primed state. Another way to test the ELP model further would be to increase the number of cell types recovered, for example by adding erythrocytes. One could then reuse the framework to determine whether the extended data are still consistent with the ELP model.

Although the application of our framework to barcoded hematopoietic MPPs demonstrates the method's utility and power, it is suitable for the study of any single-cell lineage tracing methodology, from multipotent, transient, proliferating, and differentiating cells. Our framework may prove instrumental for lineage pathway inference in other systems of tissue development, such as those found in the development of the breast and intestine in homeostasis or cancer (Barker et al., 2008; Visvader, 2009).

## EXPERIMENTAL PROCEDURES

### Introduction

The framework is suitable for initially multipotent cells that have any finite number of potential fates, but for clarity we present it for three final fates. This setting reveals all of the complexity of the approach while being sufficient for the analysis of the experimental data described in the main text.

Consider a cohort of initially multipotent cells that can produce cells of type M, D, and B (shorthand for myeloid, dendritic, and B cells). Each initial cell is of type MDB, indicating it has all three potentials, whereas subsequent cells can be of type MDB, MD, MB, DB, M, D, or B. For example, a cell of type MD could produce cells of types M and D but has lost the B potential. A barcode initially placed in an MDB cell can ultimately be found in one of seven combinations: cells of only one type (M, B, D), cells of two types (M+B, M+D, D+ B), or cells of all three types (M+B+D). For the application to hematopoiesis, because the mice have been irradiated and are being reconstituted, we assume that cells proliferate and differentiate but do not die before becoming final cell types. For alternate applications, one can readily include an additional state corresponding to dead cells.

### Overview of the Stochastic Model

As depicted in Figure 1B, the general tree allows the loss of any number or combination of potentials at any stage. For example, a cell of type MD can become a cell of type M or D. Starting with an initial cell of type MDB, letting $N_M$, $N_D$, and $N_B$ denote the number of offspring of each type produced by that cell, we are interested in knowing properties of the joint probability mass function $P(N_M = m, N_D = d, N_B = b)$ for all m, d, b in {0,1,2,...}. For any parameterization of the general tree, Equation S2 in Supplemental Experimental Procedures identifies an explicit expression for the Laplace transform of this probability mass function, its probability-generating function (PGF):

$$\varrho(s_M,\ s_D,\ s_B) = \sum_{m,d,b} s_M^m\ s_D^d\ s_B^b\ P(N_M = m,\ N_D = d,\ N_B = b),$$

(Equation 1)

where $s_M$, $s_D$, and $s_B$ are elements of the real line. From this, we deduce explicit expressions for the likelihood that a given barcode appears in cells of type M, D, B, M+D, M+B, D+B, or M+D+B, which is crucial in enabling us to find the parameters that provide the best fit to data.

### Maximum-Likelihood Fitting Procedure

Denote the vector of 12 transition probabilities of the model by

$$\mathbf{p} = (p_{MDB},\ p_{MD},\ p_{MB},\ p_{DB},\ p_{MDB\to MD},\ p_{MDB\to MB},\ p_{MDB\to DB}, \\ p_{MDB\to M},\ p_{MDB\to D},\ p_{MDB\to B},\ p_{MD\to M},\ p_{MB\to B},\ p_{DB\to D}).$$

Using the expressions for the probabilities of barcoded outcomes in Table S2 and the function fmincon from the Optimization Toolbox of Matlab R2011a, we determine the maximum likelihood parameters of the model by solving the following optimization problem

$$\begin{aligned}\mathbf{p}_{max} = \arg\max_{\mathbf{p}}\big(&n_{M+D+B}\log \pi_{M+D+B}(\mathbf{p}) + n_{M+D}\log \pi_{M+D}(\mathbf{p}) \\ &+ n_{M+B}\log \pi_{M+B}(\mathbf{p}) + n_{D+B}\log \pi_{D+B}(\mathbf{p}) \\ &+ n_M\log \pi_M(\mathbf{p}) + n_D\log \pi_D(\mathbf{p}) + n_B\log \pi_B(\mathbf{p})\big),\end{aligned}$$

(Equation 2)

where $n_{M+D+B}$ is the number of barcodes observed to produce M+D+B and $\pi_{M+D+B}(\mathbf{p})$ is the probability that, given the lineage pathway is parameterized with **p**, a barcode results in cells of all three types, and so forth for the other six potential outcomes (see Table S2). In this fitting procedure, we can restrict to any pathway by insisting that the probability of disallowed transitions are set to zero.

### Evaluating Fits via a Multinomial Test

Having identified the best-fit parameters for a given pathway, $\mathbf{p}_{max}$, in Equation 2, we use a consistent multinomial test based on a log likelihood ratio to determine whether the pathway is inconsistent with the data. Namely, letting

$$n = n_{M+D+B} + n_{M+D} + n_{M+B} + n_{D+B} + n_M + n_D + n_B$$

be the total number of barcodes recovered, we evaluate twice the log likelihood ratio,

$$\begin{aligned}
&2(n_{M+D+B}/n \log(n_{M+D+B}/(n\, \pi_{M+D+B}(\mathbf{p}_{max})))\\
&\quad + n_{M+D}/n \log(n_{M+D}/(n\, \pi_{M+D}(\mathbf{p}_{max})))\\
&\quad + n_{M+B}/n \log(n_{M+B}/(n\, \pi_{M+B}(\mathbf{p}_{max})))\\
&\quad + n_{D+B}/n \log(n_{D+B}/(n\, \pi_{D+B}(\mathbf{p}_{max})))\\
&\quad + n_M/n \log(n_M/(n\, \pi_M(\mathbf{p}_{max})))\\
&\quad + n_D/n \log(n_D/(n\, \pi_D(\mathbf{p}_{max}))) + n_B/n \log(n_B/(n\, \pi_B(\mathbf{p}_{max})))),
\end{aligned}$$

where the $\pi$ are the probabilities in Table S2 for the best-fit parameters determined from Equation 2. For a sufficiently large number of recovered barcodes, n, because the model has seven mutually exclusive potential outcomes, we use a $\chi^2$ test with six degrees of freedom. Note that the maximum likelihood fit to the data minimizes this $\chi^2$ value, so that the best-fit parameter for a given pathway structure is the one that minimizes the likelihood we will reject the pathway given the data.

### The ELP Model

The ELP model retains all links of the general tree but reduces the number of parameters from 12 to seven by strictly coupling transition probabilities through a biologically motivated means. We assume that irrespective of current cell type there is a rate of loss of each of the three potentialities $\boldsymbol{\alpha} = (\alpha_M, \alpha_D, \alpha_B)$, which are all nonnegative real numbers. The transition probabilities of the models are written in terms of these so that, for example, the probability that an MDB cell loses its B potential is

$$p_{MDB \to MD}(\boldsymbol{\alpha}) = \alpha_B / (\alpha_M + \alpha_D + \alpha_B + \alpha_M\alpha_D + \alpha_M\alpha_B + \alpha_D\alpha_B),$$

whereas the probability that an MDB cell loses both its D and B potentials is

$$p_{MDB \to M}(\boldsymbol{\alpha}) = (\alpha_B + \alpha_D) / (\alpha_M + \alpha_D + \alpha_B + \alpha_M\alpha_D + \alpha_M\alpha_B + \alpha_D\alpha_B)$$

and the probability that an MD cell loses both its D potential is

$$p_{MD \to M}(\boldsymbol{\alpha}) = \alpha_D / (\alpha_M + \alpha_D).$$

These rules are consistently applied across all cell types and correspond to a notion of approximately autonomous processes underway in each cell leading to the loss of each potential.

### Simulation

Given the best-fit ELP model, we implemented a stochastic simulation in MATLAB that follows precisely the rules of the model. Although the mathematical fits are determined by the final outcome of the proliferation and differentiation processes, the simulation enables us to investigate the per-generation (i.e., per-division) transient process, giving an indication as to the consistency of the model time frame with the data time frame. One typical sample of the cell types produced by 300 barcodes as a function of generation is presented in Figure 3C, illustrating that the process completes within a biologically reasonable number of generations.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and two tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2014.01.016.

## AUTHOR CONTRIBUTIONS

L.P., P.H., S.N., T.S., R.D.B., and K.D. conceived the study, discussed the framework, and contributed to the deductions; L.P. and R.D.B. proposed the ELP model for the binarized data; S.N., L.P., and T.S. provided the experimental data; and L.P. and K.D. developed the framework, performed the analysis, and wrote the paper incorporating feedback from all coauthors.

## ACKNOWLEDGMENTS

We thank T. Krohn for helpful feedback on the manuscript and L. Riboli-Sasco for useful discussions. L.P. was supported by a Marie Curie Intra European FP7 fellowship and a Bettencourt-Schueller foundation award. The work of P.H., S.N., T.S., and K.D. was supported by Human Frontier Science Program grant RGP0060/2012. R.D.B. and T.S. were supported by a Nederlandse Organisatie voor Wetenschappelijk Onderzoek TOP GO grant (TOPGO.L.10.042). K.D. was also supported by Science Foundation Ireland grant 12 IP 1263. This research was supported in part by the National Science Foundation under grant NSF PHY11-25915 from the Kavli Institute for Theoretical Physics.

Received: October 14, 2013
Revised: December 9, 2013
Accepted: January 14, 2014
Published: February 6, 2014

## REFERENCES

Adolfsson, J., Borge, O.J., Bryder, D., Theilgaard-Mönch, K., Astrand-Grundström, I., Sitnicka, E., Sasaki, Y., and Jacobsen, S.E. (2001). Upregulation of Flt3 expression within the bone marrow Lin(-)Sca1(+)c-kit(+) stem cell compartment is accompanied by loss of self-renewal capacity. Immunity *15*, 659–669.

Adolfsson, J., Månsson, R., Buza-Vidas, N., Hultquist, A., Liuba, K., Jensen, C.T., Bryder, D., Yang, L., Borge, O.J., Thoren, L.A., et al. (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. Cell *121*, 295–306.

Akashi, K., Traver, D., Miyamoto, T., and Weissman, I.L. (2000). A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. Nature *404*, 193–197.

Barker, N., van de Wetering, M., and Clevers, H. (2008). The intestinal stem cell. Genes Dev. *22*, 1856–1864.

Buchholz, V.R., Flossdorf, M., Hensel, I., Kretschmer, L., Weissbrich, B., Gräf, P., Verschoor, A., Schiemann, M., Höfer, T., and Busch, D.H. (2013). Disparate individual fates compose robust CD8+ T cell immunity. Science *340*, 630–635.

Duffy, K.R., Wellard, C.J., Markham, J.F., Zhou, J.H., Holmberg, R., Hawkins, E.D., Hasbold, J., Dowling, M.R., and Hodgkin, P.D. (2012). Activation-induced B cell fates are selected by intracellular stochastic competition. Science *335*, 338–341.

Gerlach, C., Rohr, J.C., Perié, L., van Rooij, N., van Heijst, J.W., Velds, A., Urbanus, J., Naik, S.H., Jacobs, H., Beltman, J.B., et al. (2013). Heterogeneous differentiation patterns of individual CD8+ T cells. Science *340*, 635–639.

Gerrits, A., Dykstra, B., Kalmykowa, O.J., Klauke, K., Verovskaya, E., Broekhuis, M.J., de Haan, G., and Bystrykh, L.V. (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. Blood *115*, 2610–2618.

Gomes, F.L., Zhang, G., Carbonell, F., Correa, J.A., Harris, W.A., Simons, B.D., and Cayouette, M. (2011). Reconstruction of rat retinal progenitor cell lineages in vitro reveals a surprising degree of stochasticity in cell fate decisions. Development *138*, 227–235.

Graf, T. (2008). Immunology: blood lines redrawn. Nature *452*, 702–703.

Grosselin, J., Sii-Felice, K., Payen, E., Chretien, S., Roux, D.T., and Leboulch, P. (2013). Arrayed lentiviral barcoding for quantification analysis of hematopoietic dynamics. Stem Cells *31*, 2162–2171.

Hasbold, J., Corcoran, L.M., Tarlinton, D.M., Tangye, S.G., and Hodgkin, P.D. (2004). Evidence from the generation of immunoglobulin G-secreting cells that stochastic mechanisms regulate lymphocyte differentiation. Nat. Immunol. *5*, 55–63.

Kaech, S.M., and Wherry, E.J. (2007). Heterogeneity and cell-fate decisions in effector and memory CD8+ T cell differentiation during viral infection. Immunity *27*, 393–405.

Kawamoto, H., and Katsura, Y. (2009). A new paradigm for hematopoietic cell lineages: revision of the classical concept of the myeloid-lymphoid dichotomy. Trends Immunol. *30*, 193–200.

Kondo, M., Weissman, I.L., and Akashi, K. (1997). Identification of clonogenic common lymphoid progenitors in mouse bone marrow. Cell *91*, 661–672.

Kreso, A., O'Brien, C.A., van Galen, P., Gan, O.I., Notta, F., Brown, A.M., Ng, K., Ma, J., Wienholds, E., Dunant, C., et al. (2013). Variable clonal repopulation dynamics influence chemotherapy response in colorectal cancer. Science *339*, 543–548.

Kretzschmar, K., and Watt, F.M. (2012). Lineage tracing. Cell *148*, 33–45.

Kurnit, D.M., Matthysse, S., Papayannopoulou, T., and Stamatoyannopoulos, G. (1985). Stochastic branching model for hemopoietic progenitor cell differentiation. J. Cell. Physiol. *123*, 55–63.

Lai, A.Y., Lin, S.M., and Kondo, M. (2005). Heterogeneity of Flt3-expressing multipotent progenitors in mouse bone marrow. J. Immunol. *175*, 5016–5023.

Lemischka, I.R. (1992). What we have learned from retroviral marking of hematopoietic stem cells. Curr. Top. Microbiol. Immunol. *177*, 59–71.

Lemischka, I.R., Raulet, D.H., and Mulligan, R.C. (1986). Developmental potential and dynamic behavior of hematopoietic stem cells. Cell *45*, 917–927.

Livet, J., Weissman, T.A., Kang, H., Draft, R.W., Lu, J., Bennis, R.A., Sanes, J.R., and Lichtman, J.W. (2007). Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. Nature *450*, 56–62.

Lu, R., Neff, N.F., Quake, S.R., and Weissman, I.L. (2011). Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat. Biotechnol. *29*, 928–933.

Macken, C.A., and Perelson, A.S. (1988). Stem Cell Proliferation and Differentation: A Multitype Branching Process Model (Berlin: Springer-Verlag).

Manz, M.G., Traver, D., Miyamoto, T., Weissman, I.L., and Akashi, K. (2001). Dendritic cell potentials of early lymphoid and myeloid progenitors. Blood *97*, 3333–3341.

Medina, K.L., Garrett, K.P., Thompson, L.F., Rossi, M.I., Payne, K.J., and Kincade, P.W. (2001). Identification of very early lymphoid precursors in bone marrow and their regulation by estrogen. Nat. Immunol. *2*, 718–724.

Morrison, S.J., Wandycz, A.M., Hemmati, H.D., Wright, D.E., and Weissman, I.L. (1997). Identification of a lineage of multipotent hematopoietic progenitors. Development *124*, 1929–1939.

Naik, S.H., Perié, L., Swart, E., Gerlach, C., van Rooij, N., de Boer, R.J., and Schumacher, T.N. (2013). Diverse and heritable lineage imprinting of early haematopoietic progenitors. Nature *496*, 229–232.

Nakahata, T., Gross, A.J., and Ogawa, M. (1982). A stochastic model of self-renewal and commitment to differentiation of the primitive hemopoietic stem cells in culture. J. Cell. Physiol. *113*, 455–458.

Ogawa, M., Porter, P.N., and Nakahata, T. (1983). Renewal and commitment to differentiation of hemopoietic stem cells (an interpretive review). Blood *61*, 823–829.

Orkin, S.H., and Zon, L.I. (2008). Hematopoiesis: an evolving paradigm for stem cell biology. Cell *132*, 631–644.

Reya, T., Morrison, S.J., Clarke, M.F., and Weissman, I.L. (2001). Stem cells, cancer, and cancer stem cells. Nature *414*, 105–111.

Rieger, M.A., and Schroeder, T. (2008). Exploring hematopoiesis at single cell resolution. Cells Tissues Organs (Print) *188*, 139–149.

Schepers, K., Swart, E., van Heijst, J.W., Gerlach, C., Castrucci, M., Sie, D., Heimerikx, M., Velds, A., Kerkhoven, R.M., Arens, R., and Schumacher, T.N. (2008). Dissecting T cell lineage relationships by cellular barcoding. J. Exp. Med. *205*, 2309–2318.

Shortman, K., and Naik, S.H. (2007). Steady-state and inflammatory dendritic-cell development. Nat. Rev. Immunol. *7*, 19–30.

Snippert, H.J., van der Flier, L.G., Sato, T., van Es, J.H., van den Born, M., Kroon-Veenboer, C., Barker, N., Klein, A.M., van Rheenen, J., Simons, B.D., and Clevers, H. (2010). Intestinal crypt homeostasis results from neutral competition between symmetrically dividing Lgr5 stem cells. Cell *143*, 134–144.

Till, J.E., McCulloch, E.A., and Siminovitch, L. (1964). A stochastic model of stem cell proliferation, based on the growth of spleen colony-forming cells. Proc. Natl. Acad. Sci. USA *51*, 29–36.

Traver, D., Akashi, K., Manz, M., Merad, M., Miyamoto, T., Engleman, E.G., and Weissman, I.L. (2000). Development of CD8alpha-positive dendritic cells from a common myeloid progenitor. Science *290*, 2152–2154.

Tsuji, K., and Nakahata, T. (1989). Stochastic model for multipotent hemopoietic progenitor differentiation. J. Cell. Physiol. *139*, 647–653.

van Heijst, J.W., Gerlach, C., Swart, E., Sie, D., Nunes-Alves, C., Kerkhoven, R.M., Arens, R., Correia-Neves, M., Schepers, K., and Schumacher, T.N. (2009). Recruitment of antigen-specific CD8+ T cells in response to infection is markedly efficient. Science *325*, 1265–1269.

Verovskaya, E., Broekhuis, M.J., Zwart, E., Ritsema, M., van Os, R., de Haan, G., and Bystrykh, L.V. (2013). Heterogeneity of young and aged murine hematopoietic stem cells revealed by quantitative clonal analysis using cellular barcoding. Blood *122*, 523–532.

Visvader, J.E. (2009). Keeping abreast of the mammary epithelial hierarchy and breast tumorigenesis. Genes Dev. *23*, 2563–2577.

Weissman, I.L. (2000). Translating stem and progenitor cell biology to the clinic: barriers and opportunities. Science *287*, 1442–1446.