Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by
Jens Rößler, M.Sc. (Physics)
born in: Heidelberg

Oral-examination: ____________________

Referees:
Prof. Dr. Thomas Höfer
Prof. Dr. Ulrich Schwarz

Applications of *Polylox* barcoding to the hematopoietic system

Jens Rößler

January 25, 2019

Abstract: Understanding the development of tissues and organs at a single cell level remains a challenge. In this thesis I present a novel barcoding system *Polylox*, recently developed by Hans-Reimer Rodewald and colleagues. Based on the loxP-Cre recombination system, *Polylox* allows endogenous barcoding of single cells *in vivo*. Using a Markov chain model for the recombination process, I find that $n_B = 1,866,890$ individual barcodes can be generated. Due to the structure of *Polylox*, barcodes have different generation probabilities. The mathematical model presented in this thesis, calibrated against experimental *Polylox* data, allows the assignment of generation probabilities to each observed barcode and the selection of informative barcodes based on their generation probabilities for clonal analyses. Experimental collaborators induced barcodes in hematopoietic stem cells (HSC) and I analysed the clone size distributions, finding large clone sizes of up to 3.8% in young mice (< 1 year old) and 21.5% in old mice (2 years old) of HSC in the adult bone marrow. I show that the appearance of large HSC clones in older mice is explained by a neutral drift model.

Sampling from mature populations of the hematopoietic system revealed that a very large proportion of HSC contributes to adult hematopoiesis (85.7%). Additionally, many HSC realize multipotency *in vivo*, yet clustering analysis of barcode frequency distributions revealed a fundamental split between myelo- erythroid and common lymphoid lineage development. These findings support the long-held, but currently contested, view of a tree-like hematopoietic structure with few major branches.

The description of the potential of common myeloid progenitors (CMP) as myelo-erythroid restricted is largely dependent on transplantation and colony assays. Analysis of *Polylox* data places the CMP compartment downstream of the split inside the myelo-erythroid branch and shows the myelo-erythroid potential of CMP. By building a mathematical framework that allows the computing of the time evolution of the moments of barcode clone sizes, I show that *Polylox* data is consistent with previous work by Busch et al. further supporting the tree model of hematopoiesis.

In addition to the analyses of single barcodes, I use network analysis techniques on observed barcode sets. The connectivity of barcode sets reflects the proliferative state of the system during labelling. I found evidence for a strong proliferative burst of at least three divisions a day in HSC progenitor cells at the time point of fetal liver formation at embryonic day 9.5 in the mouse embryo.

Polylox proves to be a valuable technique for fate mapping that is not only applicable to hematopoiesis but a multitude of systems.

Zusammenfassung: Das Verständnis der Entwicklung von Gewebe und Organen auf Einzelzellebene bleibt eine Herausforderung. In dieser Dissertation stelle ich ein neuartiges Barcode System *Polylox* vor, das kürzlich von Hans-Reimer Rodewald und Kollegen entwickelt wurde. Basierend auf dem loxP-Cre Rekombinationssystem erlaubt *Polylox* das endogene Markieren von einzelnen Zellen *in vivo*. Mit Hilfe einer Markov-Kette, die den Rekombinationsprozess beschreibt, finde ich, dass $n_B = 1,866,890$ individuelle Barcodes erzeugt werden können. Aufgrund der Struktur von *Polylox* haben einzelne Barcodes unterschiedliche Erzeugungswahrscheinlichkeiten. Das mathematische Model, das in dieser Dissertation vorgestellt wird, erlaubt das Zuweisen von Erzeugungswahrscheinlichkeiten und somit die Auswahl von informativen Barcodes basierend auf ihrer Erzeugungswahrscheinlichkeiten zur klonalen Analyse. Experimentelle Kollaborateure erzeugten Barcodes in hämatopoetischen Stammzellen (HSC) und ich analysierte die Klongrößenverteilung und fand große Klongrößen von bis zu 3.8% (in jungen Mäusen; < 1 Jahr alt) und 21.5% (in alten Mäusen; 2 Jahre alt) aller HSC im adulten Knochenmark. Ich zeige, dass das Auftauchen von großen Klonen in älteren Mäusen mit einem neutralen Drift erklärbar ist.
Stichproben aus vollentwickelten Zellpopulationen des hämatopoetischen Systems zeigten, dass ein sehr großer Anteil der HSC zur adulten Hämatopoese beitragen (85.7%). Zusätzlich realisierten viele HSC Multipotenz in vivo, eine Clustering Analyse zeigte jedoch eine fundamentale Aufspaltung zwischen myelo-erythroider und gemeiner lymphoider Entwicklung. Diese Ergebnisse unterstützen die lange vertretene, aber kürzlich angefochtene Ansicht einer baumartigen Struktur der Hämatopoese mit einigen wenigen Ästen.
Die Beschreibung des Potential von gemeinen myeloiden Vorläufern (CMP) als myelo-erythroid beschränkt basiert weitestgehend auf Transplantationen und Kolonie Assays. Die Analyse von *Polylox* Daten platziert CMP der Aufspaltung nachgeschaltet im myelo-erthroiden Ast und zeigt das myelo-erythroide Potential von CMP.
Durch das Aufstellen eines mathematischen Frameworks, das die Berechnung der Zeitevolution der Momente der Barcode Verteilungen ermöglicht, konnte ich zeigen, dass *Polylox* Daten kompatibel mit vorhergehender Arbeiten von Busch et al. sind. Diese Übereinstimmung unterstützt das Baum Model der Hämatopoese weiter.
Zusätzlich zur Analyse von einzelnen Barcodes habe ich Netzwerk Analyse Techniken auf beobachtete Barcode Gruppen angewendet. Der Verbindungsgrad in Barcode Gruppen reflektiert den proliferativen Status des System zum Zeitpunkt der Markierung. Ich konnte Hinweise für einen starken proliferativen Schub von etwa drei Zellteilungen pro Tag in HSC Vorläufern zum Zeitpunkt der Formation der fötalen Leber am Tag 9.5 nach Befruchtung im Maus Embryo finden.
Polylox erweist sich als wertvolle Technik zur Kartierung von Zellgenealogien, die nicht nur am hämatopoietischen sonder auch an einer Vielzahl an anderen Systemen anwendbar ist.

Contents

1| Introduction

With a multitude of different cell types that are serving vastly varying functions, the hematopoietic system remains a topic of intense research. While traditional fate mapping techniques provided valuable insight, the deconvolution of the hematopoietic system at single cell resolution *in vivo* remains a challenge. In this thesis we present a novel fate mapping technique called *Polylox* , which is based on the Cre-Lox recombination system. This barcoding technique offers not only the tissue-specific induction of genetic barcodes, but also control over the time point of induction. The experimental basis of *Polylox* barcoding has been established in the laboratory of Hans-Reimer Rodewald at the DKFZ. Here, we develop the mathematical analysis of *Polylox* barcodes.

1.1| The hematopoietic system

Hematopoiesis, from Greek *αίμα* "blood" and *ποιέω* "bringing something into being" is the creation of the cellular contents of blood. At the top of this hematopoietic system are the hematopoietic stem cells (HSC), which have self-renewing capabilities. HSC have been discovered in the early 1960s and have since then sparked the interest of many researchers [1]. Residing in the bone marrow in adult mammals, HSC are characterized by their multipotency, as they give rise to the multitude of different cell types within the hematopoietic system [2].
In the last decades vigorous research has identified numerous multipotent and lineage restricted cell populations downstream of HSC. These mainly *in vitro* and transplantation experiments are the basis of the hierarchical hematopoietic differentiation tree. It is characterized by progressively specialized progenitor populations, that over the course of differentiation lose self-renewal and multipotency [3, 4].
Depending on their function, we distinguish three major lineages, thought to correspond to branches in the hematopoietic tree.
Red blood cells: Or erythrocytes, which transport oxygen. As they are matured they lose their nucleus. Since the proposed fate mapping technique uses a genetic barcode, fully matured erythrocytes also lose their barcode tag. To still be able to study their place in the tree, we sampled erythrocyte progenitors.
Lymphocytes: These are cells in the adaptive immune system that specifically rec-

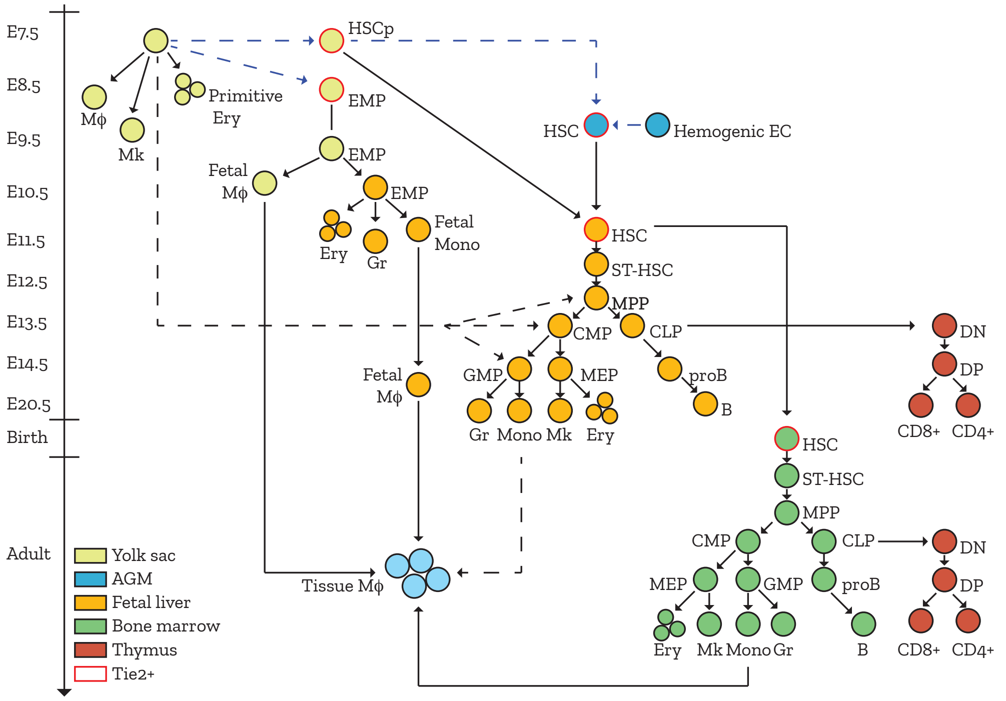


Figure 1.1: Hematopoietic system at different stages of development. Arrows denote differentiation steps, dashed arrows are hypothesized. Tie2 expressing cells are shown with a red outline. Location of development is color coded. This figure is adapted from [3]

ognize antigens. They are derived from common lymphoid progenitor cells (CLP), which are a lineage restricted population in the classical tree model. In this thesis we are studying B, CD4+ and CD8+ T cells.

Myeloid lineage: This lineage includes granulocytes, monocytes and macrophages. In this thesis we are mainly interested in the first two populations, since tissue macrophages can have a different origin than HSC [5]. These cells serve mainly in immune defense and tissue homeostasis by removing dead or damaged cells.

In figure 1.1 the tree model is pictured from embryonic stages into adulthood [3]. At the early phases of embryonic development, a preliminary hematopoietic system originates from the yolk sac (during embryonic days E7.5-E9.5). With the formation of the fetal liver at E9.5 the HSC production kicks off and moves from the yolk sac over the aorta-gonad-mesonephros (AGM) into the fetal liver [6, 7].

The structure of this model relies heavily on the identification of common lymphoid as well as common myeloid progenitor populations [8, 9]. Following the identification of CMP, multiple studies have shown that CMP have the potential to give rise to myeloid and erythroid fates [10, 11]. However, with the development of novel experimental techniques this long held view has become subject to controversy [12--14]. To find the backbone of the system and to answer these controversies many groups are now using single cell transcriptome data, with varying results. While some find evidence for fate restricted cell clusters within the CMP compartment [15], others argue that there are no lineage restricted progenitor populations that retain bipotency or oligopotency [14]. Perié et al. find, in single cell transplantation experiments, evidence for CMP as a step in the differentiation pathway towards myeloid and erythroid fates [16]. This controversy underscores the fact, that even though the hematopoietic system is well understood, some key elements remain unsolved.

Functions of transplanted multi- or bipotent progenitor populations might vary from unperturbed physiology, and transience of gene expression might lead to falsely classified connections [17]. We therefore need a method, that allows the analysis of progenitor-offspring relationships under an unperturbed physiology. *Polylox* barcoding enables the tagging of individual stem cells in a non-invasive manner and the tracing of their progeny.

1.2| Cre-Lox recombination system

Polylox makes use of a site-specific recombination system introduced into eukaryotic cells by Sauer et al. in the late 1980s [18, 19], the Cre-Lox recombination. It consists of the Cre recombinase and its respective *LoxP* target sequences, that are derived from the bacteriophage P1. With proper alignment of the target sequences, Cre-Lox enables the deletions and inversions of enclosed DNA sequences. This property is widely used to create knockout animals for *in vivo* studies of gene

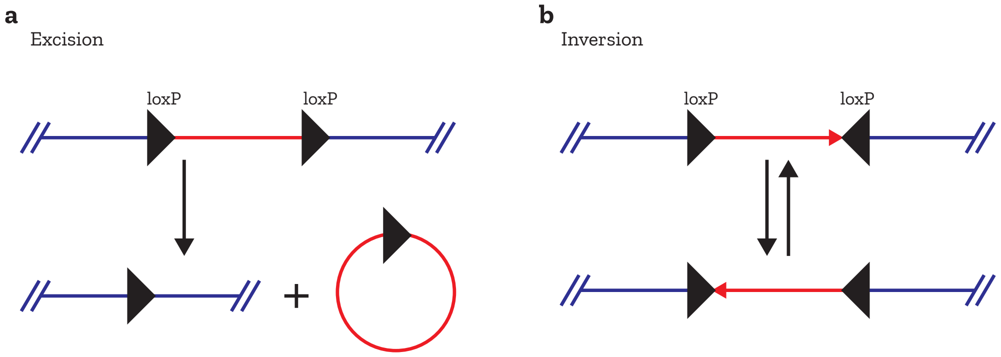


Figure 1.2: Cre-Lox recombination system: **a**, Cre-mediated excisions occur between two *loxP* -sites with the same orientation. The excised segment (red) is lost. **b**, Cre-mediated inversions occur between two *loxP* -sites with opposing orientation [21]. The blue and red line indicate DNA, the black triangles indicate *loxP* site with their respective orientation.

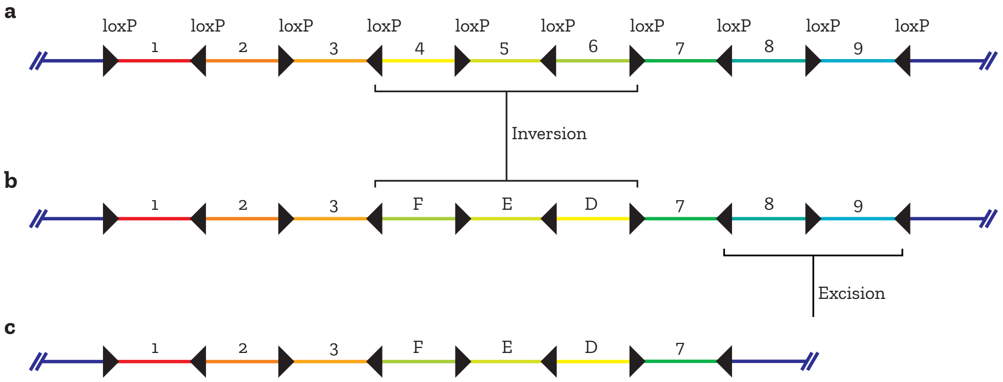


Figure 1.3: *Polylox* and nomenclature of barcodes: **a**, Unrecombined *Polylox*. Unique DNA-Elements are encoded by numbers 1-9, their inversions by letters A-I, *loxP* sites are shown as black triangles. **b**, *Polylox* after an inversion of elements 4,5 and 6 yielding the barcode '123FED789'. **c**, *Polylox* after another recombination event where elements 8 and 9 are cut out yields the final barcode '123FED7'

functions [20].
Specifically the binding of Cre to two *loxP* sites with the same orientation leads to an excision of the enclosed DNA sequence (fig. 1.2 a). Contrary, binding of Cre to two *loxP* site with different orientation leads to an inversion (fig. 1.2 b) [21].

1.2.1| *Polylox*

As the name suggest, *Polylox* consist of 10 *loxP* sites with alternating orientation [22]. Between each neighboring *loxP* pair is a unique DNA sequence of 178 base pairs (bp), which are based on the AT2G21770 gene from *Arabidopsis thaliana*. These 9 unique DNA blocks are labelled 1-9, their inversions are denoted by the letters A-I respectively (fig. 1.3 a, b).
The basic idea behind this DNA cassette is that a random recombination based on the Cre-Lox rules described in the previous section yields a high diversity of possible barcodes (fig. 1.3 c). As the Cre recombinase is inducible in certain tissues, recombination of *Polylox* is controllable in a temporal and tissue-specific manner. After recombination took place the genetic barcode is stable and heritable, allowing the tracing of output of stem cells. The key premise is that during a limited time window, Cre will not interact with all the *loxP* sites but only with a random selection of them.
To be able to study the hematopoietic systems under unperturbed conditions the locus in which *Polylox* is inserted needs to be chosen carefully. Here, a widely used locus with no known impact on physiology and function has been picked, *Rosa26*.

1.3| Overview

Since there are multiple ways of creating the same barcode, as well as endpoints of recombination such as '1' or '9', we need a way to identify rare i.e. barcodes that have been created only once. In the first chapter we will discuss a probability model for barcode creation based upon the established rules of Cre-Lox recombination. It will allow us to accurately describe *Polylox* barcode creation and will lay the foundation for the analysis presented in this thesis.

Next, we will quickly go over a computational framework that has been implemented and is easy to use for any group wanting to employ *Polylox* barcoding. The framework enables the assignment of probabilities for every found barcode and purges 'impossible' barcodes.

In the following chapter the experimental setup, that was used to create the data in this thesis is discussed.

Single cell and bulk data will be analyzed in the following chapter. Here we will take at a look at clone size distributions of HSC and their respective output into the adult hematopoiesis. We will also try to answer the question of progenitor-offspring relationships in the CMP and GMP populations, by employing hierarchical clustering strategies.

Further we will try to shed more light into barcode propagation and the expected correlation pattern between populations. For this, we build several toy models to understand the influence of kinetics and topological motifs on the measured barcode distributions.

Previous fate mapping studies have been able to deduce kinetics and topology in the hematopoietic system [23]. Using the findings of the previous chapter, we want to see whether the findings of *Polylox* barcoding is explained by the long standing tree model.

In the last chapter we will take a look at the information we get from whole barcode sets retrieved experimentally. For this we will discuss network analysis techniques and implement them for *Polylox* barcoding. This allows us to gain information in the state of the system at the time point of barcode creation.

2| Barcode probability model

In this chapter we will take a look at the theoretical barcode complexity of the *Polylox* cassette, with the aim to calculate a probability of generation P_{gen} for each possible barcode. Since some barcodes will be generated multiple times *in vivo*, we use P_{gen} to filter out highly abundant barcodes in order to gain single cell resolution of barcoding.

First, we will deduce a formula which describes the barcode complexity based on the simple rules discussed in chapter 1.2.1. To make this formula applicable in practice a library of all possible barcodes is created *in silico*. This library is then linked by possible recombination steps in an adjacency matrix, which is then used to form a transition matrix.

This simple Markov model allows us to accurately describe the process of barcode recombination and to assign to each barcode found in any given experiment the corresponding P_{gen} value. The model described in this chapter and the corresponding Matlab scripts have been made publicly available [22].

2.1| Barcode complexity

To label single cells with unique barcodes a high enough complexity in barcode creation is needed. We therefore ask how many barcodes that can be created with m unique DNA-segments. For that we create a formula for barcode complexity. The way *Polylox* is constructed, combined with the rules of the Cre-Lox recombination system, leads to the following observations:

- To avoid complete excision, *Polylox* must consist of an odd number of segments
- Suppose *Polylox* has a length of m segments then
  1. $\frac{(m+1)}{2}$ are at an odd location $(1/A, 3/C, ...)$
  2. $\frac{(m-1)}{2}$ are at an even location $(2/B, 4/D, ...)$
- Inversions have an odd length, excisions an even length

  $\Rightarrow$ As a consequence after any number of recombinations, odd segments end up at odd numbered locations, even segments end up at even numbered locations.

This gives $\frac{(m+1)}{2}!$ possibilities for the odd numbered segments and $\frac{(m-1)}{2}!$ for the even numbered segments. At every location there are again two possibilities: inverted and non-inverted. The number of barcodes for an uncut *Polylox* of length m is then given by:

$$N_{\text{inv}}(m) = \left(\frac{(m+1)}{2}\right)! \left(\frac{(m-1)}{2}\right)! \; 2^m \tag{2.1}$$

Now suppose our *Polylox* substrate of length m is originally cut down from a substrate of length l. There are now:

- $\frac{(l+1)}{2}$ different odd segments and $\frac{(m+1)}{2}$ odd numbered locations
- $\frac{(l-1)}{2}$ different even segments for $\frac{(m-1)}{2}$ even numbered locations.

To account for these permutations, we note that each one has the number of possible inversions given by eq. 2.1. The number of all possible *Polylox* barcodes of length m cut down from length l is then given by:

$$N(l,m) = \binom{\frac{(l+1)}{2}}{\frac{(m+1)}{2}} \binom{\frac{(l-1)}{2}}{\frac{(m-1)}{2}} \left(\frac{(m+1)}{2}\right)! \left(\frac{(m-1)}{2}\right)! \; 2^m \tag{2.2}$$

Summation over all possible lengths for a substrate of length l gives the total number of unique barcodes.

$$N(l) = \sum_{m=1,3,\dots}^{\frac{(l+1)}{2}} \binom{\frac{(l+1)}{2}}{\frac{(m+1)}{2}} \binom{\frac{(l-1)}{2}}{\frac{(m-1)}{2}} \left(\frac{(m+1)}{2}\right)! \left(\frac{(m-1)}{2}\right)! \; 2^m \tag{2.3}$$

For $l = 9$ this yields $n_b = 1,866,890$ unique barcodes (fig. 2.1 a). For barcoding the HSC compartment that consist of roughly 18,000 cells [23], the theoretical complexity of a length 9 *Polylox* cassette is clearly sufficient.

2.2| Library creation

To create a working probability model for barcode creation and to identify real barcodes from sequencing or PCR errors, we generated the complete library of all $n_b = 1,866,890$ *Polylox* barcodes. This is done *in silico* by applying all possible inversions and excisions exactly once to the unrecombined barcode. From that we get a list of all barcodes that are reachable within one recombination step. This step is then iterated on the so obtained barcodes until all $n_b = 1,866,890$ barcodes are created. After 10 recombination steps the list is complete.

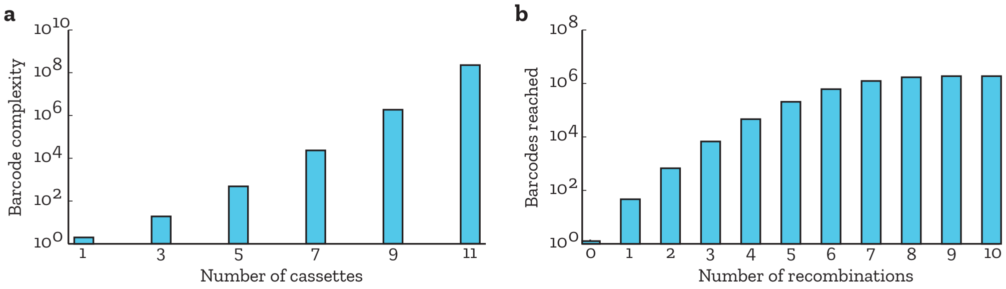


Figure 2.1: Combinatorial diversity: **a**, Barcode complexity rises faster than exponential with increasing cassette number. **b** Cumulative amount of barcodes of a *Polylox* substrate with 9 cassettes reached after a certain number of recombination steps. After 10 recombinations all $n_b = 1,866,890$ barcodes can be reached. In general, $l + 1$ recombination steps are needed to reach every barcode, where l is the number of cassettes.

2.3| Adjacency matrix

We now link the barcodes by their possible excisions and inversions in an adjacency matrix A. The elements of A correspond to connections, in this case possible recombinations, between barcodes. A is therefore a square matrix of dimension $n_b = 1,866,890$. The element $A_{ij} = 0$, if there is no possible recombination from barcode i to j, and $A_{ij} = 1$, if there is a possible recombination. Since inversions are reversible, all i, j for which $A_{ij} = A_{ji} = 1$ are encoding inversions between barcode i to barcode j. Similarly all i, j for which $A_{ij} = 1$ and $A_{ji} = 0$ are then excisions from barcode i to barcode j. The matrix product A^n at i, j gives the number of walks of length n between vertex i and j. For the smallest integer n for which $(A^n)_{ij} > 0$ holds true, n gives the minimal recombination distance between barcode i and barcode j. For $i = 1$ (unrecombined barcode) this gives the minimal recombination distance between the unrecombined *Polylox* cassette to the recombined product. This property is then used in chapter 2.6 to estimate Cre activity.

2.4| Length dependence of chromatin looping

Due to stiffness in the chromatin fiber, looping, which is needed for successful recombination events, depends on the distance between the two *loxP* sites involved in the recombination [24]. We therefore calculate the distance between two *loxP* sites from the $5'$ end of the $5'$ *loxP* site to the $5'$ end of the $3'$ *loxP* site in base pairs. Ringrose et al. (1999) showed that the looping probability P scales with the distance as follows [24]:

$$P(L) \propto \frac{125,000}{l^3} \left[\frac{4l}{10^4 L} \right]^{3/2} \exp\left(- \frac{510 l^2}{6.25 L + 50 l^2} \right) \tag{2.4}$$

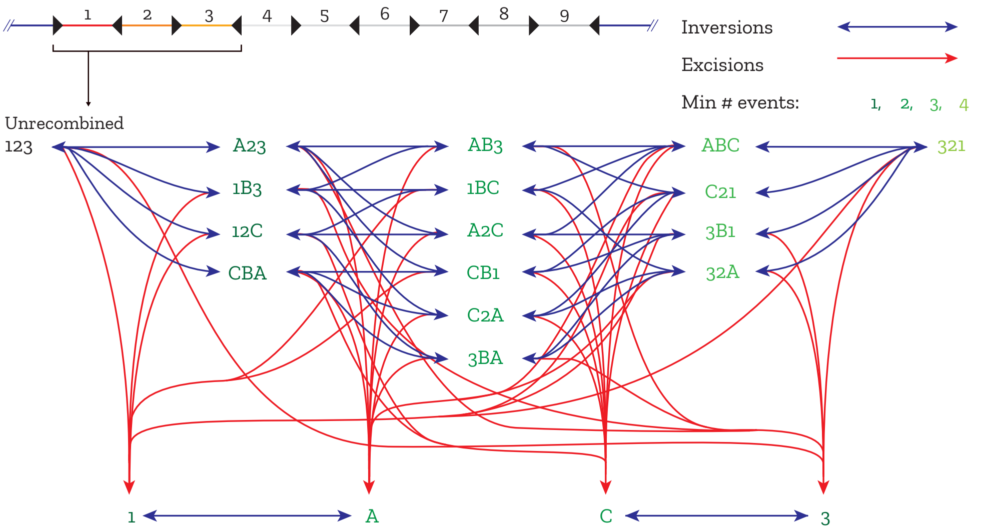


Figure 2.2: Example of combinatorics: Scheme for the combinatorics of a reduced *Polylox* of length 3. Red arrows denote excisions, blue arrows inversion. Shades of green encode the minimal number of recombinations needed to create a given barcode.

where L describes the genomic distance between to *loxP* sites and $l = 27$ nm is an *in vivo* value of the persistence length of chromatin. Since our length 9 *Polylox* cassette has a total of 10 *loxP* sites, there are 9 possible lengths of chromatin looping. When an excision occurs the number of possible recombination lengths scales with the number of segments remaining. In the next chapters the looping probability is denoted by $P_k(N)$, where N denotes the total remaining length of *Polylox* in segments, and k the distances associated with that length.

2.5| Transition matrix

In the next step of the model creation we weigh the entries of the adjacency matrix A by probabilities for each possible recombination to obtain a transition matrix T. By using the following relationship between P_{gen} and T we calculate the probability of generation given the number of recombination events t.

$$P_{\text{gen},t} = T^t P_{\text{gen},0} \tag{2.5}$$

where

$$P_{\text{gen},0} = (1, 0, 0, ...)^\top \tag{2.6}$$

encodes the unrecombined *Polylox* cassette. In chapter 2.3 we described a distinction between inversions and excisions, which we are going to use to differ between inversions and excisions in our model. To satisfy normalization, the inversion (P_{inv}) and excision probability (P_{exc}) must sum to one:

$$P_{\text{inv}} + P_{\text{exc}} = 1 \tag{2.7}$$

Entries encoding inversions are then given by:

$$T_{ij} = \frac{P_{\text{inv}}}{\sum_{j'} A_{ij'}} \tag{2.8}$$

In the next step chromatin looping probabilities $P_{k_j}(N_i)$ are added to the formula. Again, N_i denotes the total length of barcode i, k_j the genomic distance to from barcode i to barcode j. The final entries of T of an inversion are then given by:

$$T_{ij} = \frac{P_{\text{inv}} P_{k_j}(N_i)}{\sum_{j'} P_{k'_j}(N_i) A_{ij'}} \tag{2.9}$$

We treat excisions accordingly to inversions and get:

$$T_{ij} = \frac{P_{\text{exc}}}{\sum_{j'} A_{ij'}} \quad \text{and} \quad T_{ij} = \frac{P_{\text{exc}} P_{k_j}(N_i)}{\sum_{j'} P_{k'_j}(N_i) A_{ij'}} \tag{2.10}$$

for the entries of T encoding excisions without and with length dependence, respectively,

2.6| Estimation of Cre activity

Equation 2.5 allows the calculation of P_{gen} of all barcodes given an exact number of recombinations t. However, in an experiment different cells may experience different numbers of recombination steps. To estimate P_{gen} the distribution $\omega(t)$ of the number of recombination events is needed to weigh each $P_{\text{gen},t}$ (see eq. 2.11 below). In the experiment however the number of recombinations is dependent on the activity of Cre. Therefore an *a priori* estimate of $\omega(t)$ is not feasible. Due to the reversibility of inversions it also impossible to calculate the number of recombination steps for each barcode directly.
To solve this problem the minimum number of recombination steps for each given experimentally retrieved barcode is used as an estimate for the real number. Since, generally speaking, the probability of generation declines with rising t for the relevant barcodes of length 3 or greater, this estimate is appropriate.
The minimal number of recombinations needed to generate a certain barcode i is given by the minimal t, t_{min}, for which $P_{\text{gen},t_{min}}(i) > 0$. For an experimentally retrieved set of barcodes, we compute the distribution of recombination events $\omega(t)$ from these minimal numbers for each barcode. With these assumptions, we obtained narrow distributions of recombination events for all experimental data, with mean numbers between 2 and 3 and maximal numbers of 7 (fig. 2.3 c).

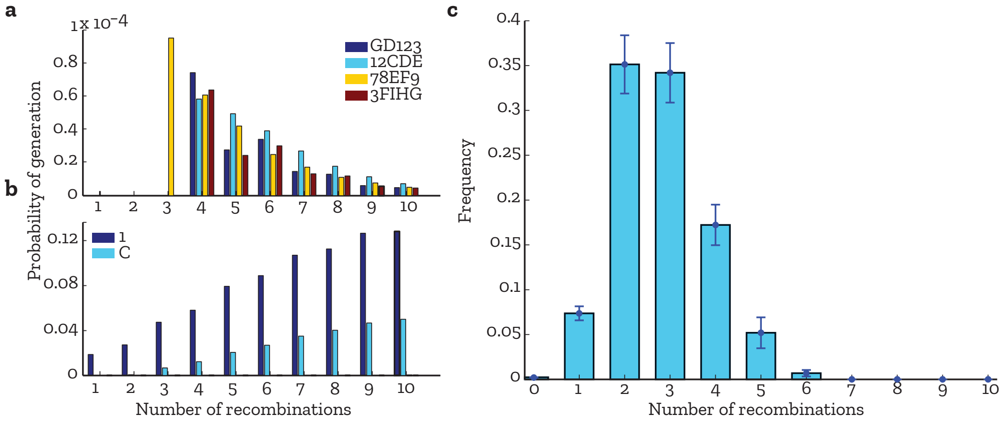


Figure 2.3: Distribution of minimal recombinations: **a**,**b**, P_{gen} for a certain number of recombinations for exemplary barcodes: **a**, Relevant, highly recombined barcodes. The minimal number needed to create the barcode shows the highest probability. **b**, Terminal barcodes show increased P_{gen} with more recombinations. **c**, Distribution of minimal recombination numbers. Dots denote the mean value, error bars show standard deviation ($n = 7$ mice).

2.7| Calculation of probability of generation

By using all of the above we then find the probability of generation for all barcodes, given an experimentally determined distribution of recombinations $\omega(t)$, as

$$P_{\text{gen}} = \sum_{t=1}^{t_{\text{max}}} \omega(t) P_{\text{gen,t}} \tag{2.11}$$

2.7.1| Adjusting the model parameters

The final estimation of P_{gen} is dependent on further parameters, mainly:

- the ratio of $P_{\text{inv}}/P_{\text{exc}}$
- the length dependence of chromatin looping

Length dependence of chromatin looping plays minor role

We therefore checked the difference between calculated P_{gen} with and without length dependence according to eq. 2.4. The P_{gen} distributions are highly correlated indicating very little influence of chromatin looping on the probability of generation for each barcode (fig. 2.4 a,b). For correctness throughout this thesis, length dependence is always taken into account.

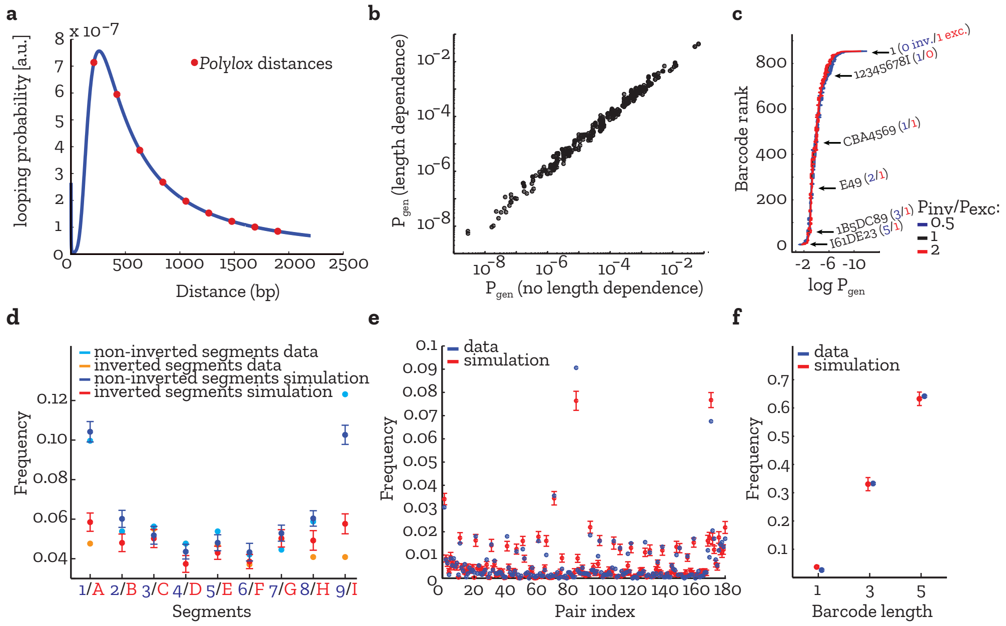


Figure 2.4: Testing the model parameters:a, Looping probability of chromatin from eq. 2.4 in a.u. [24]. Red dots denote possible loxP sites distances in *Polylox*. **b**, Difference of P_{gen} of the same barcodes with and without consideration of the looping probability of **a**. **c**, Testing of influence of the ratio of P_{inv} to P_{exc} on the calculated P_{gen}. The black line denotes $P_{\text{inv}}/P_{\text{exc}} = 1$, red and blue denote a ratio of 2 and 0.5. **d-f**, Best fit with $P_{\text{inv}}/P_{\text{exc}} = 1$ **d**, Frequency of segments compared between simulated and experimentally retrieved barcodes. $P_{\text{inv}}/P_{\text{exc}} = 1$ **e**, Comparison of the frequency of segment pairs (e.g 1-2,1-B,...) simulation in red and data in blue. **f**, Abundance of barcodes of different lengths 1,3 and 5 in simulation vs actual data. Due to known sequencing bias length 7 and 9 are omitted. **b**, Data from mouse #1, **c-f**, Data from mouse #6.

Inversion and excision have similar probability

To probe for the ratio of $P_{\text{inv}}/P_{\text{exc}}$ we calculated P_{gen} of all barcodes for three different ratios: 0.5, 1 and 2. Again the influence on P_{gen} is minimal, indicating a very robust model against those model parameters (fig. 2.4 c). To further test this, we fitted simulated barcodes according to the $\omega(t)$ of experimentally retrieved set of barcodes to that set of barcodes. To increase the pressure on the model we compared segment frequency (e.g 1/A, 2/B, ...), segment pair frequency (e.g (12,1B,...) and the length distribution.

In the fit, the parameters were not identifiable in the range between P_{inv} = 0.35 and P_{inv} = 0.6. The fact that P_{inv} does not strongly deviate from 0.5 indicates that inversion and excision are similarly likely.

Because of simplicity we do not put a bias towards either P_{inv} or P_{exc} into our model and set the ratio of $P_{\text{inv}}/P_{\text{exc}}$ to 1 (fig. 2.4 d-f).

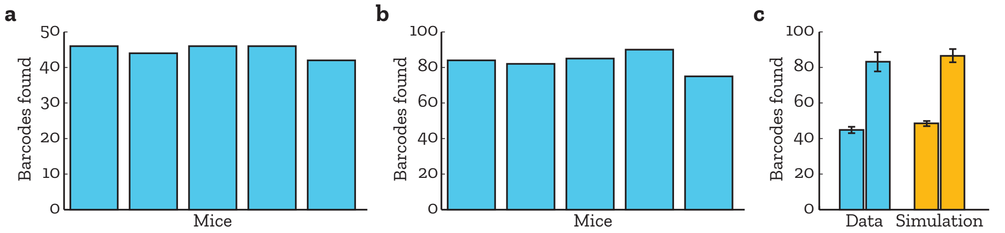


Figure 2.5: Barcodes found according to P_{gen}: **a**, Number of barcodes with the predicted 50 highest P_{gen} values found within different mice. **b**, Same as **a** but with 100 highest P_{gen} values. **c**, Summary of **a,b** in blue and simulated sets of barcodes according to the predicted P_{gen} in orange. Mean and standard deviation are shown.

Model predicts barcode outcome

Finally, we compared the number of barcodes with the predicted highest P_{gen} to actually found barcodes. From the 50 barcodes with the highest P_{gen} we found 44.8 ± 1.8 (mean and standard deviation of 5 mice), from the top 100 we found 83.2 ± 5.5 in experimental data. The values of individual mice are found in fig. 2.5 a,b. In chapter 4.3 we will estimate the number of labelled cells in our experiments to be in the order of 1000 cells. Under the assumption that our model is indeed correct, we have therefore sampled sets of 1000 barcodes according to their P_{gen} and compared the emergence of barcodes with our experimental data. We found a strong concordance between model prediction and data, indicating that the model reliable predicts the generation of barcodes (fig. 2.5 c).

3| Computational framework

Due to the complexity of *Polylox* with $n_b = 1,866,890$ possible barcodes, calculating the power of the transition matrix T^t is computationally quite heavy. Therefore we calculated and stored each T^t for $t = 0, 1, 2, \dots, 10$ together with the minimal recombination steps for each barcode. The pipeline then executes the following steps:

- **Purging of impossible or incomplete barcodes:** due to incomplete PCR or errors during sequencing, it is possible to obtain incomplete and impossible barcodes. While barcodes with the missing 5' or 3' ends are discarded beforehand, there may still be non identifiable segments left. The pipeline checks every experimentally retrieved barcode against the library of possible barcodes and purges those that are not found.

- **Finding minimal recombinations:** For the remaining barcodes the minimal number of recombination steps is looked up in a pre-calculated list. From the result a frequency distribution $\omega(t)$ of numbers of recombination events for the given experiment is calculated.

- **Calculation of P_{gen}:** The frequency distribution $\omega(t)$ is then used to calculate the probability of generation for each barcode according to eq. 2.11

- **Calculation of frequencies:** In the last step, read counts are used to calculate barcode frequencies to make different cell populations more comparable.

The pipeline generates the following outputs for further analysis:

- **Purged barcodes:** A list of all possible and identifiable barcodes that are found

- **Purged reads:** The read counts for each respective barcode for every population

- **Purged frequency:** The barcode frequency for each barcode

- **Minimal recombinations:** A list of the minimal number of recombination events to generate each barcode in the purged barcodes list

- **Frequency distribution:** The frequency distribution $\omega(t)$ of minimal numbers of recombination
- $P_{\mathbf{gen}}$: The probability of generation for every barcode in the purged barcodes list
- **Annotation:** The names of the analyzed population.

The computational pipeline is available and has been published in [25]. It is being used by other laboratories that are currently establishing *Polylox* barcoding (e.g., Shosei Yosheida, personal communication)

4| Experimental setup

The group of Hans-Reimer Rodewald developed an experimental protocol to induce *Polylox* barcodes in HSC. This protocol, along with several crucial control experiments, is discussed here briefly. A more thorough description and a step by step guide is found in Nature Protocol [26]. The experimental setup described in this chapter is used in the whole thesis with the exception of the single cell sequencing data discussed in chapter 5.1. All experiments have been performed by Weike Pei, Thorsten Feyerabend, Kay Klapproth, Daniel Postrach and Thorsten Benz (Hans-Reimer Rodewald Lab, DKFZ). Sequence alignment has been performed by Xi Wang (Thomas Höfer Lab, DKFZ), and single molecule real time (SMRT) bulk sequencing was provided by Claudia Quedenau (MDC Berlin).

4.1| Barcode induction *in vivo*

Crossing of *Rosa26*^{Polylox}

For barcode induction *in vivo* the *Rosa26*^{Polylox} allele was crossed into two mice with different inducible Cre variants:
First, mice with CreERT2, an ubiquitously expressed but tamoxifen-dependend Cre to produce *Rosa26*$^{\mathrm{Polylox\,/CreERT2}}$ mice. Tamoxifen application generates barcodes in all cells, and hence this mouse model has been mainly used in control experiments.
Second, mice with MerCreMer in the *Tie2* locus to create *Rosa26*^{Polylox} *Tie2*^{MCM} mice. *Tie2*^{MerCreMer} has been shown to have selective high expression in fetal and adult HSC and is therefore used in HSC fate mapping experiments [3].
In both mouse models, the treatment with tamoxifen causes the drug to act on the inducible Cre, making it active (fig. 4.1 a). Cre then recombines the *Polylox* construct randomly. Since the inducible Cre is not active itself and needs activation, the absence of tamoxifen due to degradation or stops *Polylox* recombination.

Experimental design

During this thesis two major experimental designs are used: embryonic and adult labelling. In the case of embryonic labelling a single dose of 2.5 mg tamoxifen is

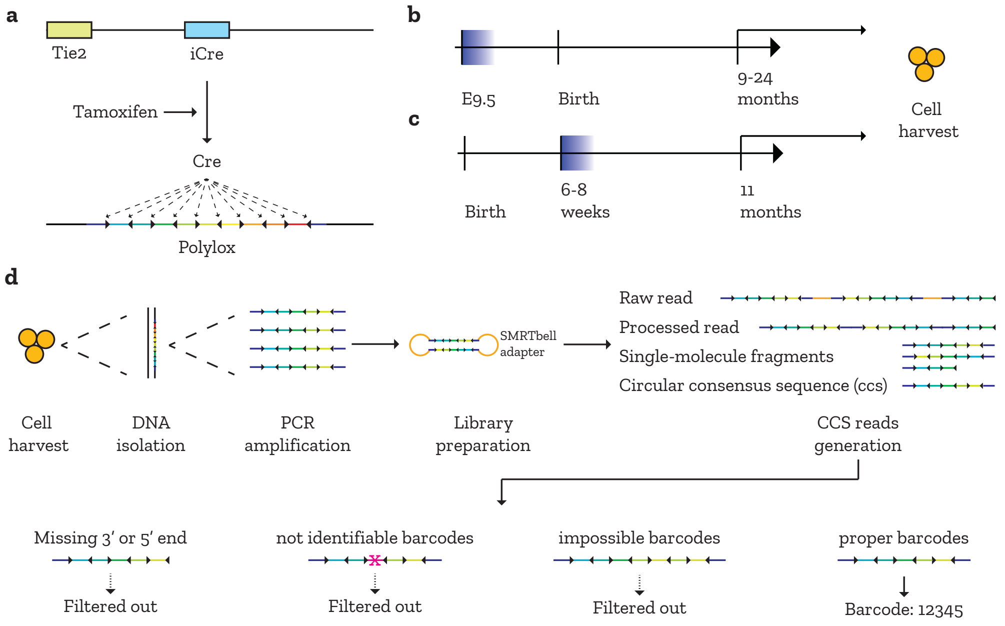


Figure 4.1: Experimental setup: **a**, Inducible Cre (iCre) in the *Tie2* locus. iCre is expressed in fetal and adult HSC. Tamoxifen translocates iCre to the cell nucleus producing active Cre that recombines the *Polylox* cassette. **b,c**, Scheme of the experimental design: Tamoxifen (blue area) is given to the mice at various stages; **b**, embryonic labelling, **c**, adult labelling. After some waiting time the cells are harvested. **d**, After cell harvesting, DNA is isolated and *Polylox* is amplified. Next, the *Polylox* cassettes are being sequenced by SMRT sequencing and CCS are generated. In the last step, invalid barcodes are filtered out.

administered by oral gavage to the mother during various time points in midgestation (E7.5 - E10.5) (fig. 4.1 b). The mothers were also treated with 1.25 mg progesterone to sustain the pregnancy.

For the experiments with adult labelling, mice were injected with 1 mg tamoxifen once per day on five consecutive days (fig. 4.1 c) 6-8 weeks after birth. Next, 9-24 months after birth, the mice in both designs were sacrificed and the cells of interest harvested.

For the experiments with *Rosa26*^{Polylox}/$^{\text{CreERT2}}$ a single injection of 1 mg tamoxifen was given intraperitoneally. In this case, only B cells were harvested a few days after induction.

Barcode retrieval

After the cells were harvested, they were sorted by FACS into different cell types (FACS gating is found in the Appendix). Then the DNA is isolated and the *Polylox* locus is amplified by PCR. For all bulk experiments, SMR by Pacific Biosciences was used. This methods allows sequencing of up to 30 kb length, which is sufficient to sequence the complete *Polylox* locus. First the amplified barcodes are linked with SMRTbell adapters to prepare the library. Next, multiple passes of sequencing are done. From the raw reads, the SMRTbell adapters are removed to obtain processed reads. Those are cut into single-molecule fragments. The fragments are then aligned, and a circular consensus sequence (CCS) is created. The usage of CCS and multiple passes of sequencing allows a high accuracy when sequencing a single molecule. In the last step, the individual segments are aligned with the CCS and labelled according to the nomenclature discussed in chapter 1.2.1. Due to the large difference in sequence between each DNA segment of the *Polylox* cassette, we map 99% of all intact reads to barcodes [22].

However, we still need to apply a number of filtering criteria to make the so obtained barcodes usable. Barcodes with missing 3' or 5' end are discarded, as only complete barcodes are properly identified and processed further. Sometimes single segments are not identifiable due to sequencing errors. Since we are not able to reconstruct the original barcode, those are also filtered out. Finally, we observe impossible barcodes. Those barcodes, for example, contain the same segment multiple times, or be longer than the original *Polylox* cassette. We also filter those out (fig. 4.1 d). This filtering process leaves us with only proper barcodes, that are used for analysis (see also chapter 3).

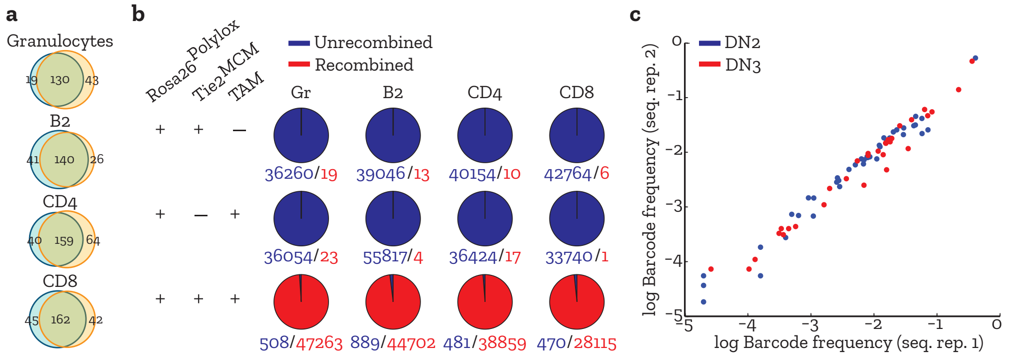


Figure 4.2: Control experiments: a, Venn diagrams of barcode overlap of sample replicates of different mature populations. The high overlap indicates a good sample depth. **b**, Recombined vs. unrecombined barcodes (read counts) in three different setups. a *Rosa26*^{Polylox} *Tie2*^{MCM} mouse without tamoxifen treatment (top row), a *Rosa26*^{Polylox} mouse with tamoxifen (middle row), and a *Rosa26*^{Polylox} *Tie2*^{MCM} mouse with tamoxifen treatment (bottom row). Only the last experiment showed recombination on a large scale. **c,** Log barcode frequency of two sequencing repeats of DN2 and DN3 cells from mouse #3. Each dot represents a single unique barcode

4.2| Control experiments

A number of control experiments have been conducted.

Incomplete recombination is possible

First, *Polylox* was recombined by Cre *in vitro* to see whether incomplete recombination is possible. Indeed, in the gel electrophoresis five bands were clearly visible, corresponding to the five possible fragment lengths expected from largely incomplete and random recombination.

Recombined products are stable

Another crucial step was to check how stable the recombined *Polylox* cassettes are. Here *Polylox* was targeted in the *Rosa26* locus of embryonic stem (ES) cells and transfected with MerCreMer. Treatment with 4-hydroxy-tamoxifen (4-OH-TAM), the active form of tamoxifen, yielded again the five fragments. In those pulse-chase experiments the distribution of recombined fragments remained constant. This indicates that once Cre is inactive, the recombination stops, yielding stable barcodes.

No recombination in the absence of Cre and TAM

In the next step, we studied the leakiness of the system. If we find recombined barcodes in mice without *Tie2*^{MCM} or tamoxifen, this would mean that already recombined barcodes could slowly change over time, making fate mapping difficult. We conducted three experiments for this: first, mice with the complete construct (*Rosa26*^{Polylox} *Tie2*^{MCM}) and without tamoxifen treatment; second, mice without the inducible Cre but with tamoxifen treatment; last, the complete construct with tamoxifen treatment. Only in the last experiment we found recombination of *Polylox* on a large scale, indicating that the *Rosa26*^{Polylox} *Tie2*^{MCM}) mouse model has no discernible rate of background recombination of *Polylox* (fig. 4.2 b).

Sample replicates show high overlap

Another key aspect is sampling depth. As the absence of barcodes in one population could have implication on fate restriction and the topology of the system, one needs to make sure to sample enough cells to be able to exclude sampling issues as explanation for the absence. Sample replicates were taken for four populations (granulocytes, spleenic B2, CD4+ T cells and CD8+ T cells) with 30,000 cells each. In general the overlap of found barcodes was between 60%-67%, indicating a good sampling depth of barcodes (fig. 4.2 a). Nevertheless, the possibility that a barcode is not found due to incomplete sampling must always be considered.

Sequencing repeats are concordant

In the last control experiment, a single sample of DN2 and DN3 cells were sequenced twice and compared. We found a high concordance between both sets of sequencing repeats (fig 4.2 c). Both sets had a correlation coefficient of 0.96-0.98. This indicates that the barcode frequency obtained from read counts is robust and is therefore used as a measure of barcode abundance in the analysis.

4.3| Threshold setting for P_{gen}

Since some barcodes have a very high probability of generation, we use the estimated P_{gen} (chapter 2) to filter out those barcodes that are generated more than once. However the threshold for P_{gen} depends on the number of cells that are initially barcoded. It might be hard to estimate this number accurately (e.g. HSC expand during fetal labelling).

We therefore used data from two mice to set the threshold for P_{gen} specifically for the experiments that were conducted. We induced barcodes as described above. After 9-11 months, to allow for equilibration, the barcodes were then sequenced and binned according to their probability (fig. 4.3). Since highly likely barcodes were generated more than once, their frequency is expected to be higher than the frequency of unlikely barcodes. At one point, the P_{gen} threshold, the correlation between frequency and P_{gen} vanishes. This indicates the point below which every barcode was created in the smallest possible unit, a single cell. We performed a two sample Kolmogorov-Smirnoff test to see whether the binned frequency distribution for different P_{gen} ranges differ from each other in embryonic labelling

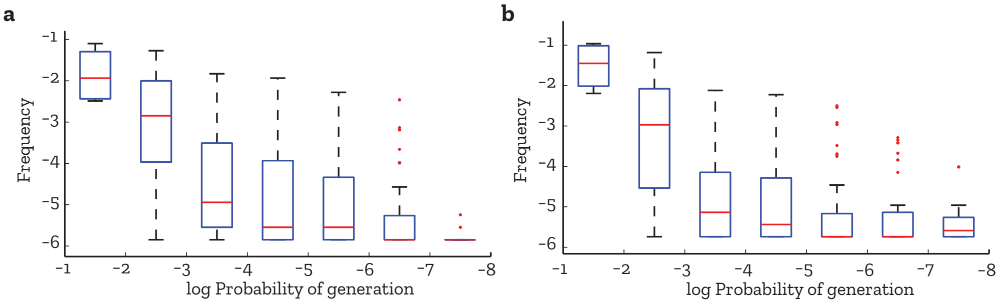


Figure 4.3: Barcode frequencies versus $P_{\mathbf{gen}}$ for threshold setting. The red bar indicates the median, the box ends the 25th and 75th percentile. Black bars indicate the most extreme value not considered outlier, whereas outliers are marked with red dots. As expected the frequency drops with decreasing probability of barcode generation.
After $P_{\text{gen}} = 10^{-4}$ the correlation vanishes, indicating that below that, the smallest possible unit, a single cell has been labelled.

experiments. Up to $P_{\text{gen}} = 10^{-4}$ we found significant differences between the binned frequency distributions, below $P_{\text{gen}} = 10^{-4}$ no differences could be found. We therefore set the threshold for all experiments to $P_{\text{gen}} = 10^{-4}$.
This indicates that the original cell number labelled has to be in the order of 1,000 cells, as barcodes with $P_{\text{gen}} = 10^{-4}$ are most likely generated only once in such a pool of cells.
In adult barcoding experiments, sampling played a major role in barcode detection. Even though the number of cells that are barcoded is much higher, we also chose $P_{\text{gen}} = 10^{-4}$ as a threshold. This value is a trade off between high confidence of single cells barcodes and finding enough barcodes for analysis.

5| Data analysis and clustering

In this chapter we analyse data obtained by inducing *Polylox* barcodes in HSC in either embryonic mice or adult mice.
From these experiments we will get information about the clone size distribution and examine aging effects in the HSC compartment in the adult bone marrow. We also analyse the output of HSC into progenitor and mature cell populations, gaining inside into the topology of the hematopoietic system. By using hierarchical clustering of barcode frequencies we find a strong dichotomy between a myelo-erthyroid and a common lymphoid differentiation branch.
Several results discussed in this chapter are published in [22].

5.1| Embryonic labelling

In the mouse embryo, the HSC that drive definitive hematopoiesis begin to form at around embryonic day (E)9.5 in the fetal liver and migrate into the bone marrow at birth (fig. 1.1). Because this process is not fully understood, we labelled embryonic mice in the time range of E7.5 - E12.5, where the earliest labelling time-point coincides with the appearance of HSC progenitors (HSCp) in the yolk sac [3]. Labeling was done by administering a single dose of tamoxifen orally to the mother. After 9-24 months the mice were sacrificed and analysed. For five mice (#1, #2, #10, #12, #14) in addition to bulk sequencing, single cell sorting and sequencing of HSC was performed. First we will look at the barcode clone distributions of labelled HSC and possible aging effects in older mice. Then, we will study the overall output of labelled HSC and investigate evidence of fate/lineage restriction. Finally, we will take a look at the relationships of mature populations in the context of shared barcodes and their respective frequencies.

5.1.1| Clone size distributions of HSC

Large fetal HSC clones play a major role in adult bone marrow

To determine HSC clone size distributions in the adult bone marrow, barcodes of approximately 500 sorted single HSC were analysed 9-11 months after induction. Since in both mice (#1, #2) at least 95% of all cells had recombined barcodes, the

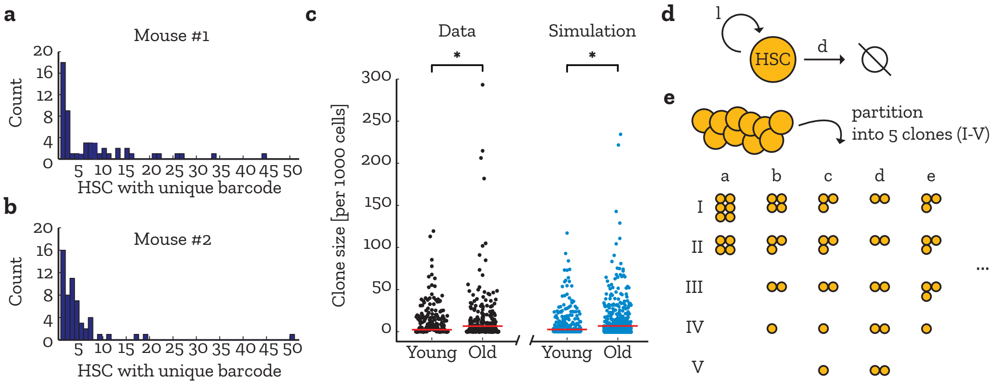


Figure 5.1: Clone size distributions of HSC in adult bone marrow: a,b, measured clone size distribution of mouse #1 and #2 at 9 and 11 months of age. Some barcodes show large clone sizes, while the majority of barcodes was found in 10 or less HSC (a, n=54; b, n=56 barcodes) [22]. **c**, measured clone sizes extrapolated to 1000 cells of mice younger than 1 year (young, n=2) and 2 years of age (old, n=2). Simulation (blue dots) of neutral drift explains the data (black dots) qualitatively. The red bar indicates the mean. **d**, scheme of the model to explain the drift to larger clone sizes. **e**, exemplary scheme of the partitioning of cells with the same barcode into different clones. Here 10 cells are divided into 5 clones (I-V). Different possibilities are taken into account (a-e) to correct the measured clone size distribution.

Tie2-driven induction labels almost all cells during midgestation that later form the adult HSC compartment. Both mice show a wide variety of barcode clone sizes (fig. 5.1 a,b), where the largest clones made up almost 10% of all HSC. However, since this also includes highly abundant barcodes, this does not properly reflect real clone size distribution in the adult bone marrow. We therefore focused on rare barcodes with $P_{\text{gen}} < 10^{-4}$ and found 14 barcodes that were created most likely only once. Here the clone sizes ranged between 0.2 and 3.8%. We found relatively large clones (> 1.5%) in both mice, corresponding to at least 150 adult HSC. Large clones therefore seem to play a major role in the clonal makeup of HSC in adult bone marrow [22].

Neutral drift explains clone size spread

In order to study the effects of aging we also analysed the barcode clone size distribution of mice of 2 years of age (#10, #12) in the same manner as described above. In both older mice we found enlarged HSC clones in comparison with the younger mice. The largest clone labelled with a rare barcode made up 21.5% of the adult bone marrow. In all experiments a different number of cells were analysed, making it difficult to directly compare measured barcode clone sizes. The clone sizes were therefore extrapolated to number per 1000 cells.

We asked whether the appearance of large HSC clones in older mice could by explained by neutral drift. To this end, we build a model (fig. 5.1 d) where HSC proliferate with rate l and leave the HSC compartment by either differentiating or cell death with combined rate d. To initialize the model, we sampled from the distribution of the young mice and simulated the individual clones with Gillespie's algorithm [27]. The rate parameters for the model where taken from [23] and set to $l = d = \frac{1}{110}\frac{1}{\text{day}}$. With this simple model the spread of barcode clone sizes is qualitatively described (fig. 5.1). This result is remarkable in the face of the slow HSC turnover.

Corrected clone size distribution gives inside into aging

Due to highly abundant barcodes the measured clone size distribution does not properly reflect the true make up of HSC clones in the adult bone marrow. While focusing on rare barcodes with $P_{\text{gen}} < 10^{-4}$ corrects this, the sample size becomes very small by this filtering (< 10 rare barcodes per single cell sequencing experiment). We reassured that a good estimate of the clone size distribution is obtained by considering all barcodes and factoring in their generation probabilities. First, we estimated the number of expected clones for a given barcode with P_{gen} by using the following formula:

$$E\left[k\right] = \sum_{k=0}^{\infty} k \binom{n}{k} P_{\text{gen},\, i}^{k} (1 - P_{\text{gen},\, i})^{n-k} \tag{5.1}$$

$$= nP_{\text{gen},\, i} \tag{5.2}$$

Here $E\left[k\right]$ is the expected number of times barcode i is generated in a sample of n cells and hence the number of clones that share barcode i. The assumption is independent barcode creation, so that the number of expected times barcode i is created is therefore described by a binomial distribution. We then partition the uncorrected, measured clone size into $E\left[k\right]$ parts. Here we also allow a clone to have a size of zero to take the death of a clone into account (fig. 5.1 e). Since the number of possibilities rises very quickly with rising cell numbers, we sample 10,000 times from all possibilities of a single barcode clone. Doing this for all the barcode clones allows us to compute the summary statistics of the corrected clone size distributions to get more insights into the aging process seen in fig. 5.1 c.

Due to HSC expansion during midgestation clone size distributions at birth are not known. To see whether the simple neutral drift model (fig. 5.1 d) explains the changing of the moments of the clone size distributions we therefore use the corrected clone size distributions of the youngest mouse as initial values. One has to keep in mind that the chosen initial values might not reflect the expected distribution perfectly.

Mean, variance, skewness and kurtosis are explained well by the simple neutral drift model, where we have a rising mean number of cells per clone and variance,

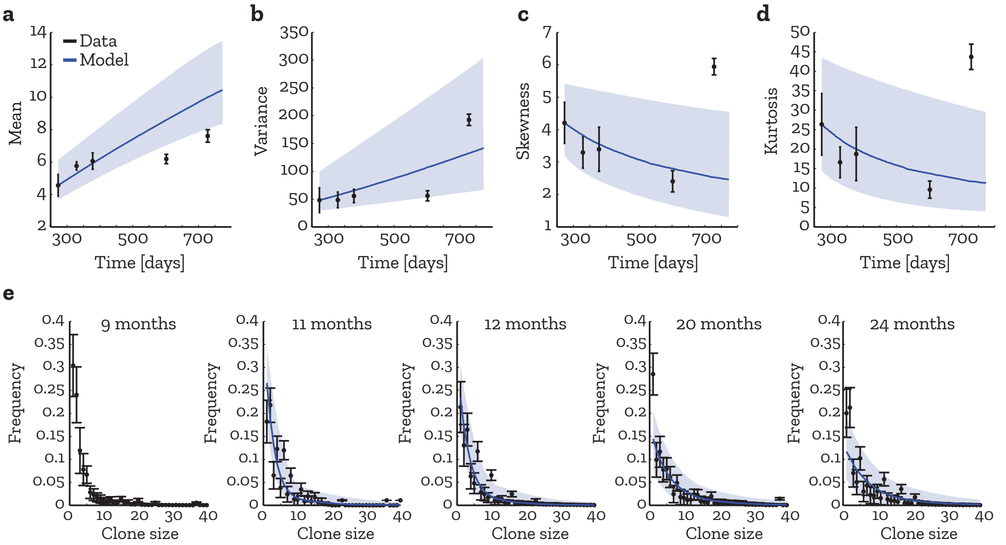


Figure 5.2: Neutral drift model on corrected distributions a-d, Corrected mean, variance, skewness and kurtosis values with standard deviations (black dots), and the neutral drift model (blue line). The blue shaded area indicates the 95% confidence interval. **e**, Time snapshots of the simulated clone size distribution (blue line) over corrected distributions (black dots). The model is in good agreement with the data.

but declining skewness and kurtosis (fig. 5.2 a-d). In the oldest mouse we found a very large rare clone that is overrepresented. Due to the still relatively small number of clones this leads to a huge deviation in the skewness and kurtosis, explaining the outliers there (fig. 5.2 c-d).

In addition to the moments, the simulated and measured, corrected clone size distributions match well (fig. 5.2 e), indicating that clone size distributions age by neutral drift. During aging this leads to a smaller number of clones that make up more of the HSC compartment [28--33].

However, the small sample size (few hundreds cells from 5 mice) makes it difficult to get quantitative information. Especially earlier time points could shed more light on the dynamics of clonal aging in the HSC compartment.

5.1.2| Output of HSC into mature compartments

In addition to single cell sequencing of sorted HSC, peripheral population were also analysed in bulk. Here we focused on the output of HSC and possible lineage restriction that might occur.

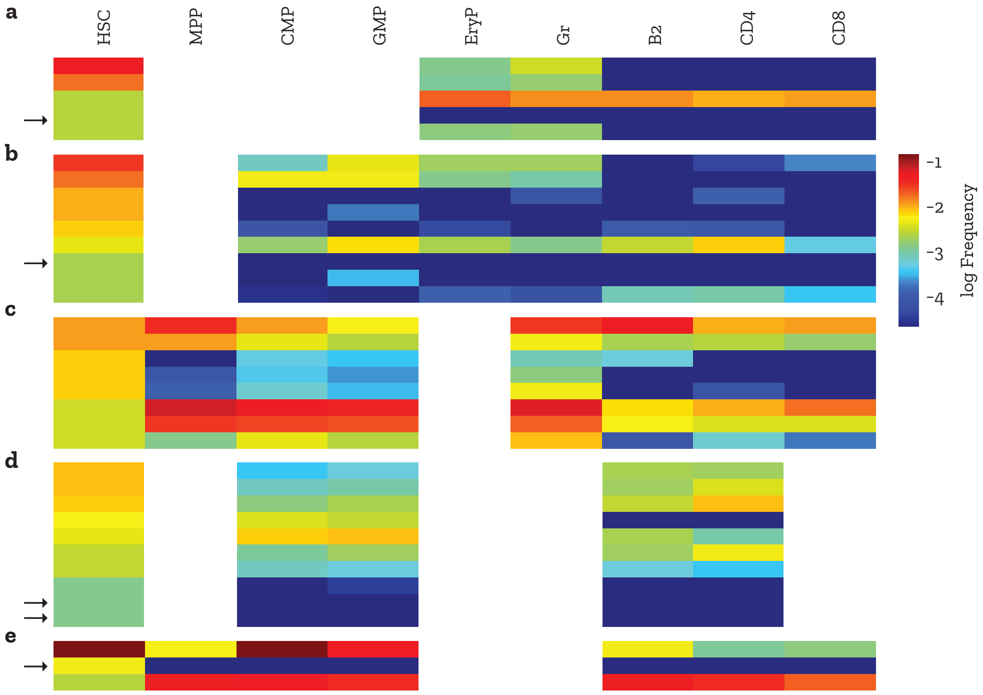


Figure 5.3: Output of single cell HSC: a-e Shows the output of rare barcodes found in HSC by single cell sequencing into different compartments (#1,#2,#10,#12, and #14 respectively), white areas depict populations that were not sampled. Black arrows indicate HSC with no output into mature populations.

A large proportion of HSC is active

First, we analysed the output of HSC labelled with barcodes with $P_{\text{gen}} < 10^{-4}$ (fig. 5.3 a-e). Out of the 35 rare barcodes we found in the HSC compartment of 5 mice of varying age (9-24 months) 85.7% were found in the peripheral populations. However those HSC that show no output into mature populations tend to have a relatively small clone size (fig. 5.3 a-e, black arrows). This leads to a small expected output, which could be due to undersampling. In addition to sampling, barcodes also leave the system via cell death, which is much more probable for small clone sizes. In line with BrdU studies, which showed that in a span of 6 months 99% of HSC divided at least once [33, 34], this indicates that a very large proportion of HSC contribute to adult hematopoiesis over the time span of these experiments.

Indications of multilineage potential

To analyse possible lineage restrictions of HSC, we focused again on rare barcodes with $P_{\text{gen}} < 10^{-4}$, to make sure that barcode clones represent HSC clones. 63.6% of the analysed barcodes were found in both major lineage branches (myelo-

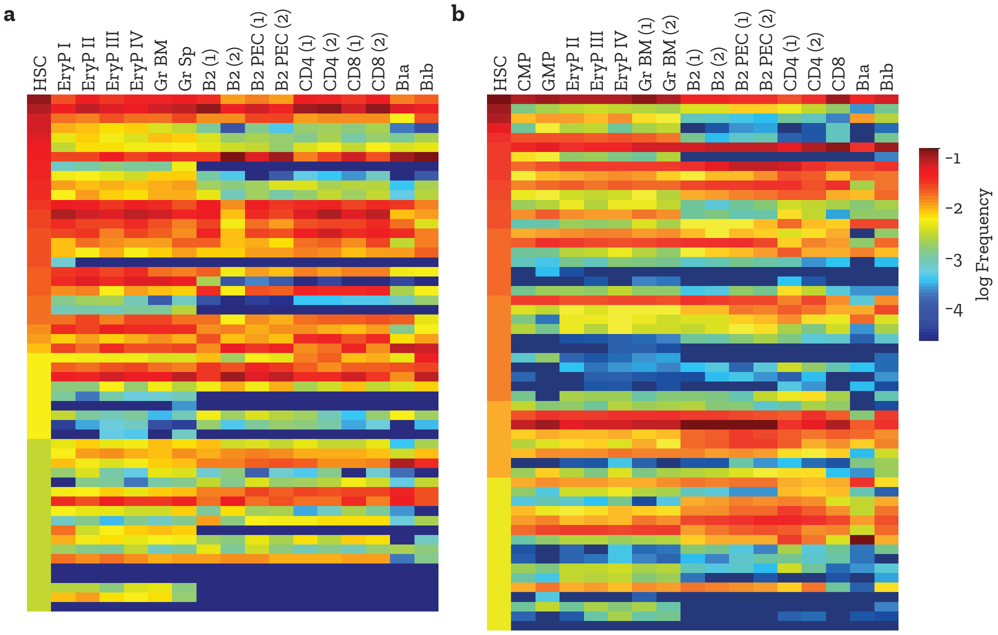


Figure 5.4: Output of HSC: a,b, All Barcodes found by performing single cell sequencing of HSC and their respective output into mature cell populations in two experiments

erythroid and lymphoid). Still, since we can only track clones we cannot say whether individual cells also have multilineage potential or not. In addition to that, we also found lineage restricted clones (26.4% of analysed barcodes). This fate restriction indicates that at least some proportion of HSC clones show a coherent fate potential. Importantly, one should keep in mind that due to sampling, some barcodes might be missed leading to an underestimation of the observed fates.

Barcode usage differs across different fates

Releasing the P_{gen} filter, one notices the difference in barcode usage across different fates. While generally speaking, a large barcode clone in the HSC compartment leads to a large output into the peripheral populations, there are some exceptions. There are less abundant barcodes in HSC that are found at high frequency in mature cells and vice versa (fig. 5.4). This lack of strict correlation is explained, at least in part, by the slow differentiation of HSC into the downstream populations. Downstream of HSC massive expansion takes place. This causes stochasticity that decorrelates barcode usage between HSC and peripheral populations (more on this in chapter 6). However, there is a coherent barcode usage in both major branches visible (fig. 5.4: myelo-erythroid: CMP, GMP, EryP, Gr; lymphoid: B2, CD4, CD8), while B1 cells and especially B1a cells differ from both. In the next section we will analyse this different barcode usage in more depth.

5.1.3| Cluster Analysis

Besides information regarding potential fate restrictions, barcode usage may also provide inside into the topology of differentiation pathways. The general assumption here is that if two populations share a high number of barcodes as well as their respective frequencies, the two lineages belong likely to a common developmental pathway. Conversely, if they share fewer barcodes at dissimilar frequencies the two examined lineages emerged more independently from each other.

Barcodes sharing provides insight into lineage topology

We analysed the shared barcodes and frequencies of mature populations. We focused on rare barcodes again ($P_{\text{gen}} < 10^{-4}$), and in addition we also filtered for reliably sampled barcodes. Barcodes are considered reliably sampled if they are found in both sample repeats of at least one population. This additional filter guards against considering undersampled barcodes.
When comparing lymphoid populations (B2 cells, CD4+ T and CD8+ T cells) with myeloid and erythroid populations (EryP, Gr), the percentage of shared barcodes is around 40-50%. Further, the frequencies of the shared barcodes appear to be uncorrelated (fig. 5.5 a-c). Taken together, these findings point to substantially distinct lymphoid and myelo-erythroid pathways. By contrast the percentage of shared barcodes between EryP and Gr, and B2, CD4+ T cells and CD8+, respectively, varies between 73-82%. Here we also see a strong correlation between observed barcode frequencies (fig. 5.5 d-e), indicating a strong developmental link between B and T cells as well as between EryP and Gr. This general pattern was observed in all examined mice.

Cluster analysis reveals dichotomy

We further examined the relationships between different populations by hierarchical clustering. Since the correlation of barcode frequencies also contains information on the overlap, we use correlation as measure of barcode concordance. As measured barcode frequencies are not necessarily quantifiable by a linear relationship model and vary over orders of magnitude, we use Spearman's rank correlation ρ. Spearman's ρ allows us to asses monotonic and non-linear relationships and is more robust against outliers [35].
To assess relationships between populations we calculated $\rho_{i,j}$ for every combination of populations i,j. Next we used as a distance measure $d_{i,j} = 1 - \rho_{i,j}$. Thus the distance between populations gets higher the lower the correlation is. The distance $d_{i,j}$ quantifies the observations made in the previous section (fig. 5.5). We used $d_{i,j}$ to cluster all measured populations hierarchically. Again, we focused on rare and reliably sampled barcodes.

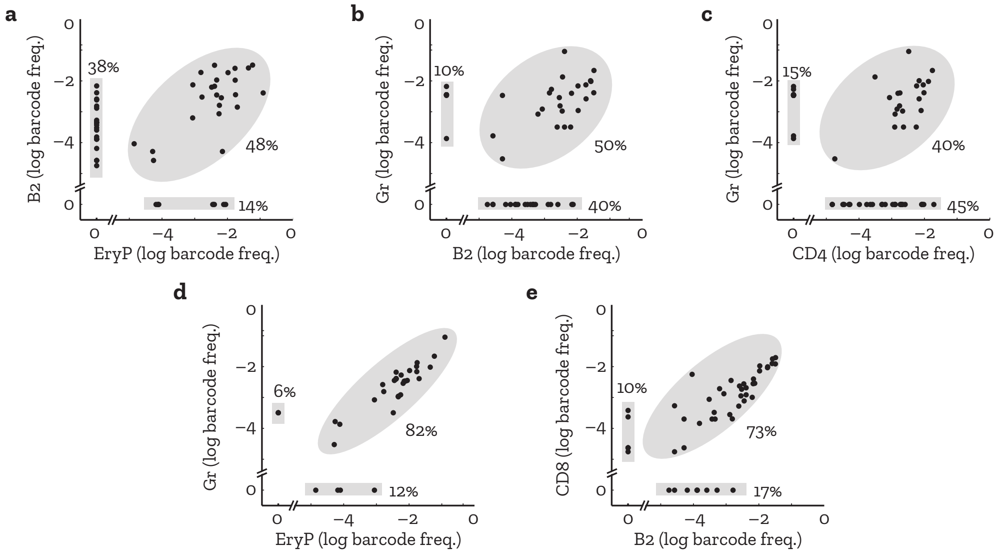


Figure 5.5: Scatterplots of retrieved barcodes in mature populations: Barcode frequencies of **a**, Erythrocyte progenitors (EryP) versus B2 lymphocytes **b**, B2 lymphocytes versus granulocytes **c**, CD4+ T cells versus granulocytes **d**, EryP versus granulocytes and **e** B2 lymphocytes versus CD8+ T cells. Data retrieved from mouse #1, each dot represents a single barcode

We observed a major split between myelo-erythroid and common lymphoid lineages (fig. 5.6 a). Populations inside one of the branches showed very high correlations (0.61 - 0.93; min.-max. values), correlations between the two branches tend to be much lower (0.18-0.53). This pattern was found in all measured mice.

Localization plays no role in barcode usage

Both from granulocytes and B2 lymphocytes samples where taken from different localizations (bone marrow and spleen versus spleen and peritoneal cavity). However there is no apparent difference in barcode usage between the different localizations, indicating that the origin of cells from the same population in different localizations is the same (fig. 5.6). Most likely mature cells in different locations are created from common pools of progenitors and then migrate to the different tissues.

B1 cells form a distinct lineage

In the next step we also included B1a and B1b cells into the clustering. B1 cells are a distinct subset of B lymphocytes that, unlike the more conventional B2 cells, do not develop into memory cells during an immune response. In both experiments B1a and B1b sublineages cluster together and form a distinct B1 lineage inside the

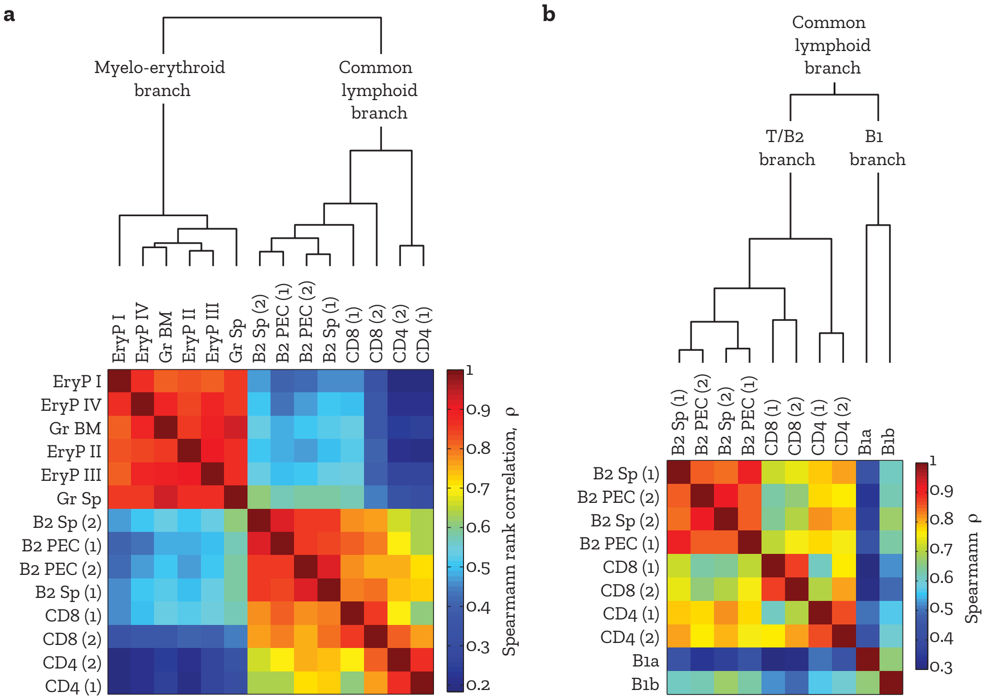


Figure 5.6: Hierarchical clustering of mature populations: applied with distance measure $d_{i,j} = 1 - \rho_{i,j}$: **a**, Various erythrocyte stages (Ery I-IV), granulocytes, B2 lymphocytes and T cells (CD4,CD8). Only rare and reliable sampled barcodes are taken into account. A major split between myelo-erythroid and lymphoid populations is visible. **b**, Analysis as in a, including all lymphoid populations, B1a and B1b cells.

common lymphoid branch (fig. 5.6 b). This indicates that B1 cells arise from a different developmental path compared to B2 lymphocytes.

We found a slightly lower percentage of recombined reads in B1 (~85%) in comparison to both common lymphoid (~96%) and myelo-erythroid (~96%) branches. In steady state one would expect the percentage of labelled cells to adjust to the labelling frequency of their origin, namely HSC. Since in embryonic labelling steady state is reached after 2 weeks [23], a deviation from the HSC labelling frequency may indicate the existence of another origin . During development of the embryo several waves of B1 generation have been observed, where the first wave emerges before E9.0 [36]. While later waves are HSC derived, the first wave seems to have restricted progenitors [36] or occur before labelling takes place. Since B1 populations show almost self-renewing capabilities [37], this first wave of B1 cells is still visible in the adult hematopoiesis and explains the lower percentage in recombined reads. Of all barcodes retrieved, we found an overlap of 40% in both the common lymphoid and B1 branch. In summary, the number of shared barcodes and the overall low correlation suggest that most B1 cells arise from HSC but split early from the common lymphoid branch. This is in line with transplantation studies showing an early split in development of B1 cells [38--41].

5.1.4| Robustness against filtering criteria

In the previous section only rare ($P_{\text{gen}} < 10^{-4}$) and reliably sampled barcodes were taken into account. This ensures that only barcodes that were generated exactly once are considered and guards against undersampling. To analyse the robustness of the barcode usage patterns we found, we applied modified filtering criteria. Cluster analysis was performed with three alternative criteria on two different experiments (mouse #1 and #2):

1. all retrieved peripheral barcodes (no filter);
2. rare barcodes that are found in at least three different populations, hence restricting the analysis to multilineage barcodes;
3. codes found in adult HSC as retrieved by single cell sequencing.

All three cases showed the same correlation pattern (fig. 5.7) as the clustering of rare barcodes, with a strong dichotomy between common lymphoid and myeloerythroid development. In particular multilineage barcodes showing the same pattern indicate that the correlation is a stronger indicator than the overlap itself, as almost all barcodes were shared between all the examined populations. The mechanistic implications of this will be explored in chapter 6.

5.1.5| Progenitor data

In the next step of the data analysis we focused on stem and progenitor cell populations. Here we also analysed: long-term HSC (abbreviated LT or LT-HSC in the figures), short-term HSC (ST-HSC, ST), multipotent progenitors (MPP), common myeloid progenitors (CMP), granulocyte-macrophage progenitors (GMP), pro B (pooled from Fr.B/C and Fr.D cells) and pre T cells (double negative thymocytes, pooled from DN2 and DN3).

Slow differentiation kinetics lead to lower correlation

Due to low labelling efficiency (~50%) in respective experiment, we included all barcodes in the analysis. As shown in chapter 5.1.4 this increases the overall correlation, but has only minor effects on the clustering pattern.
While long term (LT) stem cells are known to directly feed into the ST compartment, the overall correlation we found was rather low (0.65) (fig. 5.8 a,k). This is explained by the very slow differentiation of the LT compartment into the ST-HSC [23]. A cell rarely differentiates into the ST compartment, where proliferation is much faster. Therefore the barcode clones in both compartments develop

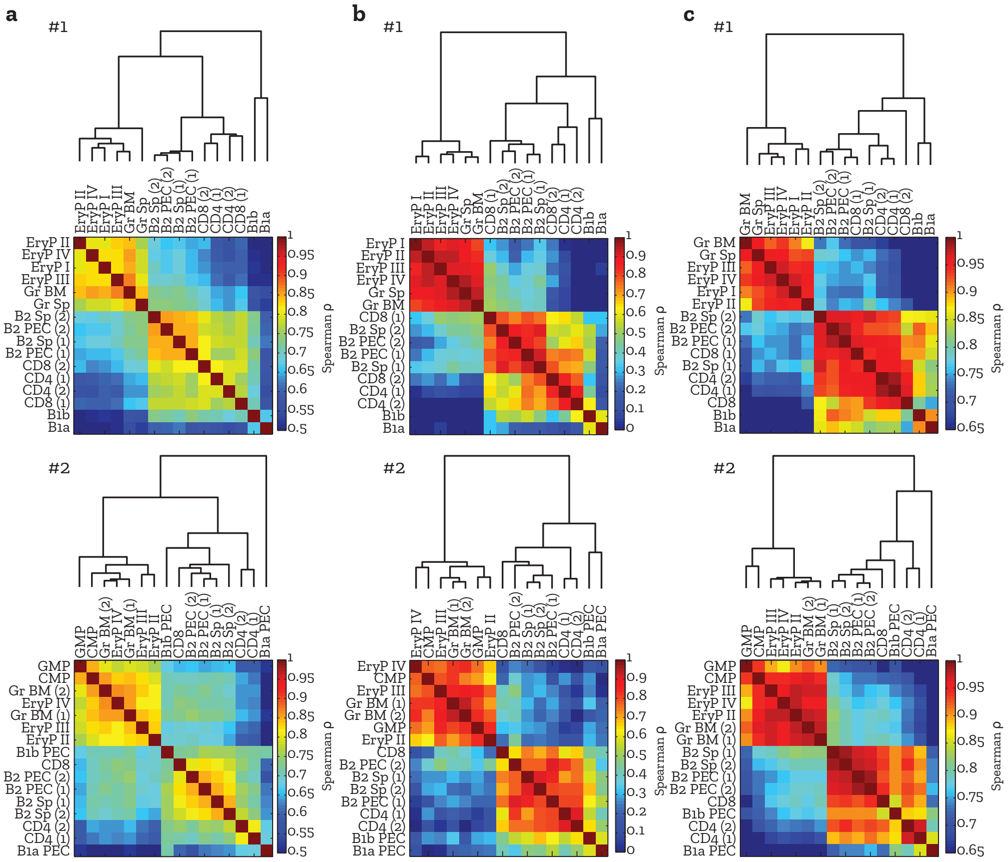


Figure 5.7: Robustness of correlation patterns: Hierarchical clustering with a distance measure of $d = 1 - \rho_{ij}$ applied to **a**, All peripheral codes; **b**, Rare barcodes found in at least three different populations; **c** All HSC barcodes found in the periphery. The major split between myeloid-erythroid and lymphoid branch is robust and reproducible.

rather independently of each other. For the differentiation from ST to MPP we already have larger clone sizes, reducing stochasticity and increasing correlation (0.73) (fig. 5.8 b,k). This indicates that in addition to the topology of the system, the kinetics play a substantial role in the observed correlation patterns.

MPP feed into CMP and CLP

MPP are thought to be the last multipotent lineage before the split in the topology into myelo-erythroid and lymphoid branch happens, although they may already contain lineage-restricted progenitors (as seen in in transplantation [16, 42]). To investigate this, we also analysed MPP the downstream compartments CMP and CLP. Indeed MPP showed a high correlation with both populations (0.79 vs CMP,

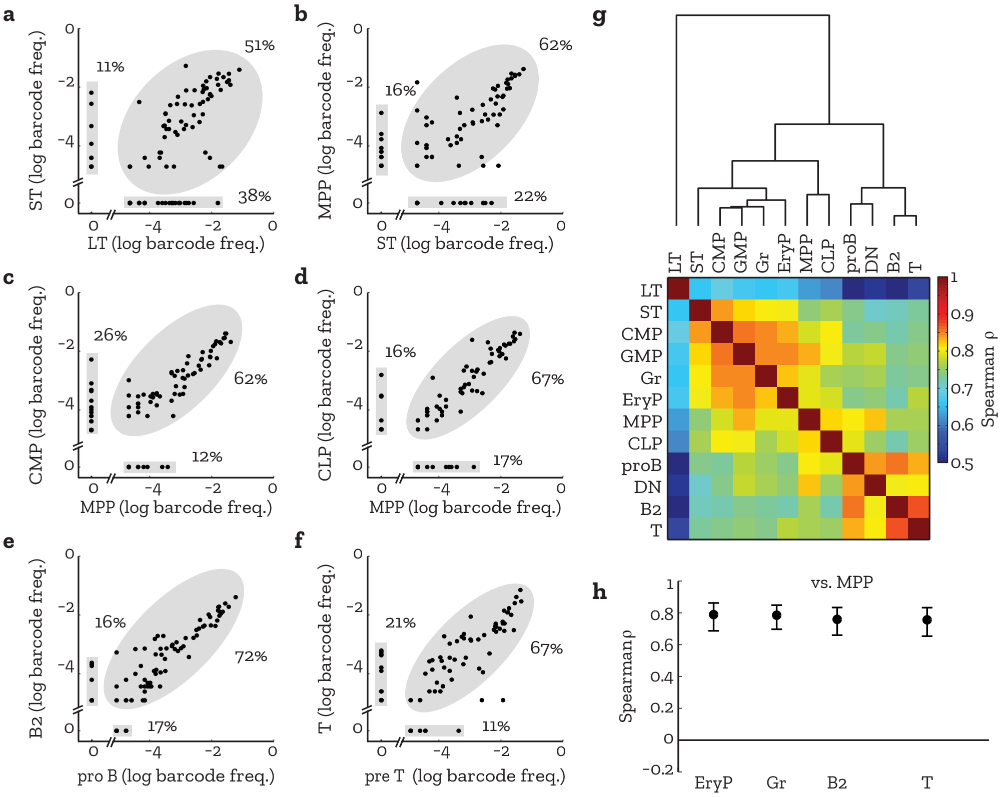


Figure 5.8: Hierarchical clustering of progenitor populations: a-f, Scatter plots of barcode usage of indicated populations. Each dot represents a unique barcode. Percentages show proportion of shared/individual barcodes. **g**, Hierarchical clustering of indicated populations with 1-ρ_{ij} as distance measure. **h** Spearman's rank correlation of different mature populations vs MPP. Errors are obtained by non-parametric bootstrapping and indicate 95% confidence interval

0.82 vs CLP), indicating that MPP are closely related to those lineages (fig. 5.8 c-d). Further, while there is a clear split between myelo-erythroid and lymphoid lineages, there was no apparent difference in barcode usage between MPP and both branches (fig. 5.8 h). Correlations found were high with all analysed mature populations independent of their branch, indicating that MPP do harbor multilineage potential and are placed before the split.

Progenitor are closely related with offspring

In our assertion that a close lineage relationship shows up in the correlation holds true, we would expect barcode usage in pro B and pre T cells to be highly correlated with B2 lymphocytes and T cells respectively. Indeed we found a high concordance in barcode usage between progenitor and offspring (fig. 5.8 e,f). In

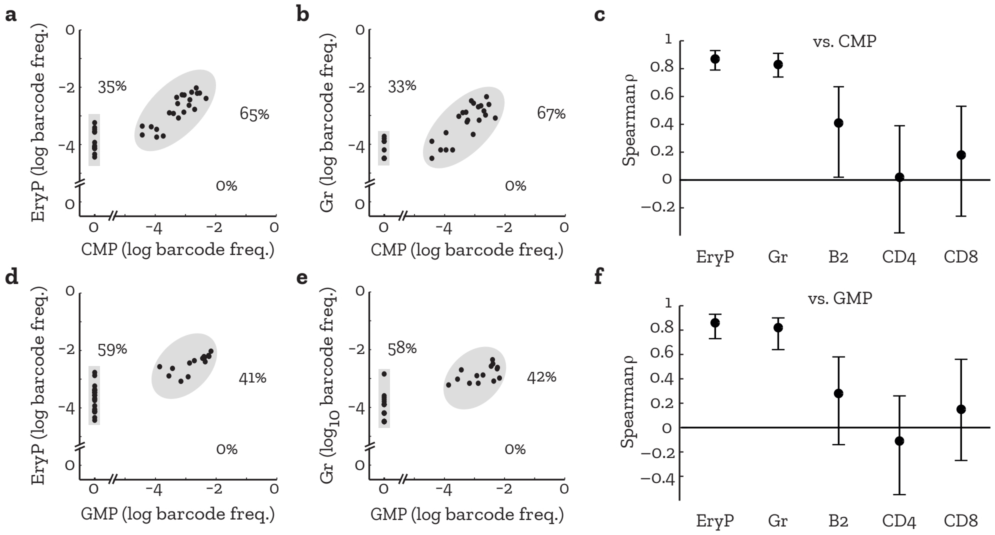


Figure 5.9: CMP and GMP correlations: a-e, Scatter plots of barcode usage of indicated populations. Each dot represents a single rare barcode. Percentages show proportion of shared/individual barcodes. Spearman's rank correlation ρ of **c**, CMP and **f**, GMP versus the indicated populations. Barcodes are extracted from mouse #2. Mean and 95 % confidence interval calculated by non-parametric bootstrapping are shown.

the cluster analysis pro B and pre T cells clustered clearly in the lymphoid branch (fig. 5.8 g). While they did not form subcluster with their respective offspring, pro B had the highest correlation with B2 lymphocytes (0.86). DN showed a lower correlation with T cells of 0.8, which could be explained by selection processes occurring during T cell development [43]. Both DN and pro B cells are strongly correlated with CLP (0.76 versus DN cells, 0.79 versus pro B cells), indicating that CLP may feed both populations in unperturbed hematopoiesis.

CMP and GMP act as myeloid progenitors

In the experiment described above, sampling proved to be an issue. We therefore included all barcodes found in the analysis. However, to study CMP and GMP in more depth, another experiment with fewer cell populations was done. Here we again found a very high efficiency of recombination (98% of reads are recombined). A better efficiency leads to a higher number of barcodes created. This allows us to use the P_{gen} $=10^{-4}$ filter as well as focusing on reliably sampled barcodes to gain single cell resolution.

While the potential of CMP as myelo-erythroid lineage restricted has been described largely by transplantation and colony assays [16], *Polylox* allows us to follow the fate of this lineage directly *in vivo*. Since both populations are expected to give rise to granulocytes a high correlation would indicate this.

Indeed we found a high concordance in barcode usage and frequency with granulocytes and erythrocytes (fig. 5.9 a-e), suggesting that CMP are *in vivo* progenitors of both lineages. The high correlation of EryP and GMP implies a close developmental relationship between the two, placing GMP into the myelo-erythroid branch. In this case, all barcodes that were retrieved from the progenitor populations have also been found in the mature populations (fig. 5.9 a-e). However in different experiments we also found CMP and GMP barcodes that were not present in neither EryP nor Gr, which is easily explained by barcode propagation and is explored in the next chapter.

As expected, the correlation between CMP/GMP and populations from the common lymphoid branch are very low (fig. 5.9 c,f). This places CMP and GMP downstream of the split of the two branches into the myelo-erythroid branch.

Another interesting feature of the data is that about 1/3 of Gr and EryP barcodes were not retrieved in CMP, although the CMP compartment was more comprehensively sampled. This finding could indicate the existence of differentiation pathways to these mature populations that bypass phenotypic CMP.

5.2| Adult labelling

In the final part of this chapter we explore output of adult HSC into mature populations. Three mice (#3, #4, #5) were treated by injecting tamoxifen intraperitoneally at around 8 weeks of age. Barcodes were retrieved 11-13 months after induction. Because Tie2 is expressed in ST-HSC, MPP and CMP in the adult mouse, barcodes were also induced in these populations and was not only in HSC. However recombination in ST-HSC, MPP and CMP was much lower than in HSC [22], so that the vast majority of barcodes were generated in HSC.

5.2.1| Cluster Analysis and progenitor data

We repeated the analysis steps as described above for the data retrieved from embryonic labelled mice.

Smaller HSC barcode clone sizes lead to less sample overlap

In embryonic labelled mice, we obtain large barcoded HSC clones in the adult. This is due to the emergence of HSC at the time of barcode induction and the following rapid expansion of the compartment [7]. In the case of adult labelling, the starting pool of HSC is much larger. In combination with a slow differentiation and proliferation this creates smaller barcode clones in adult labelling.

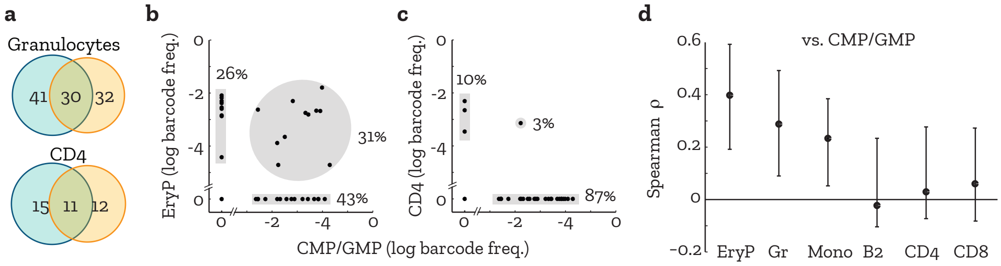


Figure 5.10: Adult labelled mice: a, Low overlap in sample repeats hint at lower barcode clone sizes. **b,c**, barcode usage of pooled CMP and GMP cells in comparison with EryP **b** and CD4+ T-cells **c**. Each dot represents a single rare barcode. **d**, spearman's rank correlation ρ of pooled CMP/GMP rare barcodes versus mature populations.

This effect should be visible when comparing sample repeats. We therefore looked at the overlap of rare barcodes in sample repeats, analyzing two samples of granulocytes and CD4+ T-cells. As expected, the measured overlap is much lower than in embryonic labelling (fig. 5.10). Consequently, measured correlations are expected to be lower and observed clustering patterns not as clear-cut as in embryonic labelled mice.

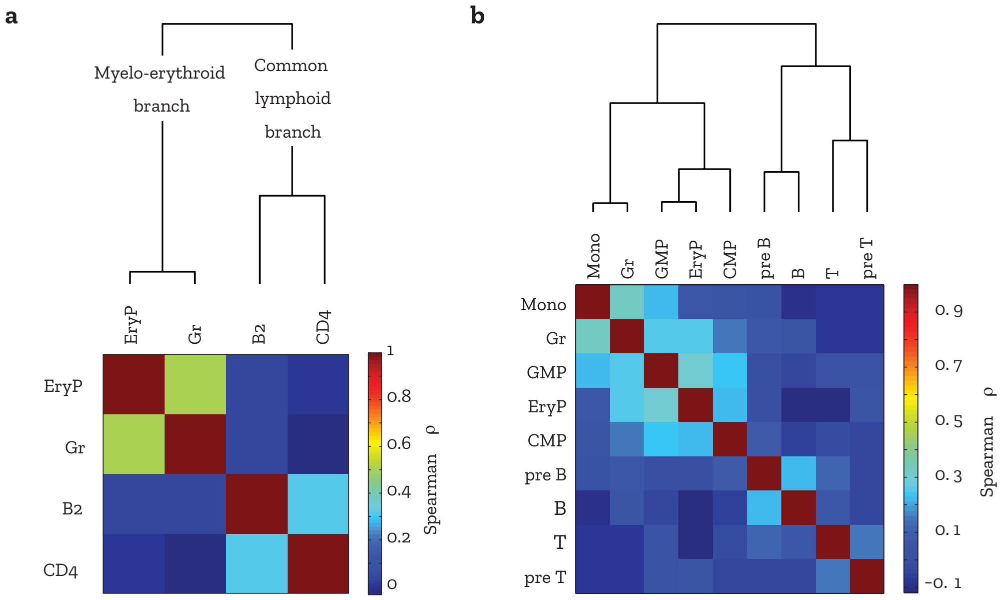


Figure 5.11: Hierarchical clustering of barcodes from adult labelled mice: clustering applied to **a**, Codes from mouse #5 with increased sample size and **b**, Mouse #3. A clear dichotomy between myelo-erythroid and common lymphoid branch is visible for both experiments. Progenitor barcodes cluster with their respective mature population in **b**.

CMP/GMP are related to the myelo-erythroid branch

To see whether the observations in chapter 5.1.5 regarding the potential of CMP and GMP as myeloid progenitors hold true in adult labelled , we analyzed the data in the same manner as described before. Due to smaller clone size and the attached sampling issues CMP and GMP barcode reads were pooled.
As before the number of shared barcodes was higher in the case of EryP than with CD4+ T-cells. Here only 3% of the retrieved rare barcodes were shared in both populations, where 31% of barcodes were shared with erythrocyte progenitors (fig. 5.10 b,c). The almost complete absence of shared barcodes between CMP/GMP and CD4+ T-cells indicates a high distance in the developmental path.
We then compared correlations and the 95% confidence interval obtained by non-parametric bootstrapping of CMP/GMP versus mature populations. In addition to the populations in chapter 5.1.5 we also sampled monocytes. While the correlations with populations from the common lymphoid branch (B2, CD4, CD8) was practically zero, populations from the myelo-erythroid branch (EryP, Gr, Mono) showed much higher correlation. Again, this places CMP and GMP populations downstream of the split between myelo-erythroid and common lymphoid branches and as progenitors of the myelo-erythroid populations.

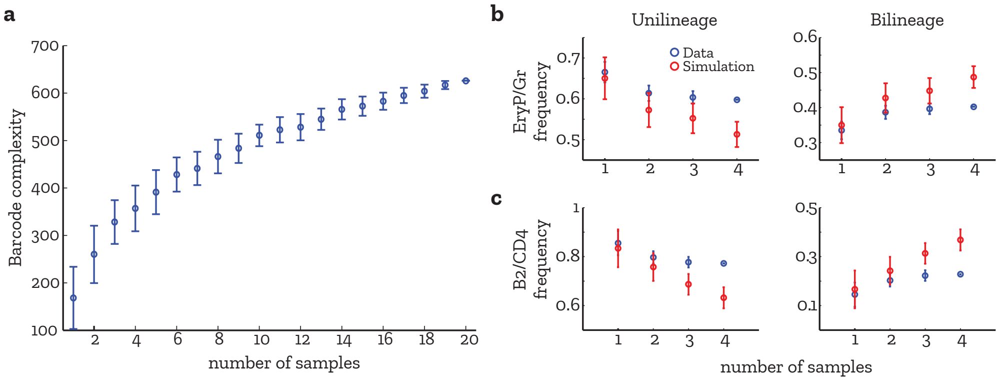


Figure 5.12: Barcode enrichment analysis: a, Unique barcodes found over number of drawn samples. Mean and standard deviation are shown. The curve flattens out in the end, suggesting that the majority of barcodes have been found. **b,c**, Proportion of uni- and bilineage barcodes. Data in blue, assumption of ubiquitous barcodes in red. Mean and standard deviation.

Cluster analysis reveals same correlation patterns as in embryonic labelled mice

Cluster analysis as described in chapter 5.1.3 was performed. Additionally we included pro B cells and pre T cells as progenitor populations in the experiment. To address the known undersampling issue caused by smaller barcode clone sizes, we increased the sampling depth 3-4 times in one experiment.
The overall correlation pattern was again strongly visible, splitting the populations into the two branches (fig. 5.11). Next, we compared all mature populations to their respective progenitors. Despite undersampling, the dichotomy between myelo-erythroid and common lymphoid pathways appears. To minimize the influence of undersampling, CD4+ and CD8+ T cells are pooled, as well as DN and DP cells as T cell progenitors and Fr. A, Fr. B and Fr. C B cell progenitors as pro B cells. Further, pro B and B2 lymphocytes and pre T and T-cells are clustering together in their respective lineage. This suggest that they are indeed downstream of the split. Again GMP and CMP clustered inside the myelo-erythroid branch, showing that those populations are more closely related to myelo-erythroid than lymphoid lineages.

5.2.2| Multilineage fates

In the case of adult labelling the pool of cells in which barcodes are generated is much larger than in the case of embryonic labelling. We therefore expect more but smaller barcode clones in the adult system. To circumvent the sampling issues that arise with smaller barcode clone sizes, we analysed multiple sample repeats. In addition to a better sampling depth used in the previous section for hierarchical clustering, this allows for a barcode enrichment analysis to check whether unilineage barcodes are truly unilineage or due to undersampling.

Sampling reveals unilineage clones

To check whether undersampling plays a role in assigning a barcode as unilineage, or if there are true unilineage fates we used a "null" model. We assumed barcodes are found in all lineages with the mean frequency obtained in the experiment (EryP/Gr and B2/CD4 respectively). Next we sampled 30,000 times from this frequency distribution to obtain a total of 4 samples (each experimental sample contains 30,000 cells). We then analysed the proportion of unilineage to bilineage barcodes as more samples are added. The data deviate from the null model substantially. While in the model the proportion of bilineage barcodes tends to 1, it stagnates in the data at around 0.4 (EryP/Gr) and 0.2 (B2/CD4) fig. 5.12 b,c). This indicates that indeed unilineage clones exist even inside a major branch. The occurrence of unilineage clones does not necessarily mean fate restriction as it might also be possibly explainable by stochasticity in the system, which we will study in the next chapter.

6| Theoretical study of barcode propagation

Basic correlation analysis of barcode distributions in different populations revealed a major split between the myeloid-/erythroid and the common lymphoid branch. To gain a deeper inside how these observed correlation patterns arise it is crucial to understand the underlying mechanics. With *Polylox* we are able to label single stems cells. Therefore, stochastic effects need to be taken into account when investigating the dynamics of their proliferation and differentiation. In this chapter, we introduce the theoretical framework and study in general terms the implications of topology, kinetics and clone sizes on observable correlation patterns.

6.1| Moment equations for proliferation and differentiation

We are interested in finding the time evolution of the moments of barcode clone sizes. To this end, we formulate a master equation of a standard Markov model [44]. We will then translate the master equation (as infinite-dimensional system of ordinary differential equations, ODE) into a partial differential equation (PDE) for the probability generating function F (PGF). F is given, in general, by

$$F(z_1, \dots, z_j, t) = \sum_{n_1, \dots, n_j} z_1^{n_1} \cdot \dots \cdot z_j^{n_j} P(n_1, \dots, n_j, t) \tag{6.1}$$

for a system with j state variables $n_1, \dots, n_j$. By sequentially differentiating F with respect to z_i and setting $z_i = 1$ we obtain a set of ODE for the moments of P. The numerical solution of ODE is usually faster than stochastic simulations.
In this chapter several toy models are studied. Because of their linearity, we treat differentiation and proliferation separately and rebuild the probability generating function from these building blocks:

Differentiation:

The master equation of a simple differentiation process from A to B with n_A and n_B cells respectively and rate d is obtained by balancing influx and efflux of a given state:

$$\dot{P}(n_A, n_B) = d[(n_A + 1)P(n_A + 1, n_B - 1) - n_A P(n_A, n_B)] \tag{6.2}$$

Using the definition of the PFG from eq. 6.1 we rewrite the master equation as a PDE for F. By multiplying $z_A^{n_A} z_B^{n_B}$ and summing over n_A and n_B, we express the master equation in means of F, term by term:

$$\sum_{n_A=0}^{\infty} \sum_{n_B=0}^{\infty} z_A^{n_A} z_B^{n_B} \dot{P}(n_A, n_B) = \dot{F}(z_A, z_B) \tag{6.3}$$

$$\begin{aligned}
\sum_{n_A=0}^{\infty} \sum_{n_B=0}^{\infty} z_A^{n_A} z_B^{n_B} (n_A+1) P(n_A, n_B) &= \frac{z_B}{z_A} \sum_{n'_A=0}^{\infty} \sum_{n'_B=0}^{\infty} z_A^{n'_A} z_B^{n'_B} n'_A P(n'_A, n'_B) \\
&= \frac{z_B}{z_A} z_A \sum_{n'_A=0}^{\infty} \sum_{n'_B=0}^{\infty} z_A^{n'_A - 1} z_B^{n'_B} n'_A P \\
&= \frac{z_B}{z_A} z_A \partial_{z_A} F(z_A) \\
&= z_B \partial_{z_A} F(z_A)
\end{aligned} \tag{6.4}$$

and finally

$$\begin{aligned}
\sum_{n_A=0}^{\infty} \sum_{n_B=0}^{\infty} z_A^{n_A} z_B^{n_B} (n_A) P(n_A, n_B) &= z_A \sum_{n_A=0}^{\infty} \sum_{n_B=0}^{\infty} z_A^{n_A - 1} z_B^{n_B} n_A P(n_A, n_B) \\
&= z_A \partial_{z_A} F(z_A)
\end{aligned} \tag{6.5}$$

Combining eq. 6.3 - 6.5 we the following PDE for cell differentiation:

$$\partial_t F(z_A, z_B) = d(z_B - z_A) \partial_{z_A} F(z_A, Z_B) \tag{6.6}$$

Proliferation:
Using the same approach we find

$$\dot{F}(z_A) = l(z_A^2 - z_A) \partial_{z_A} F(z_A) \tag{6.7}$$

for symmetric cell division with rate l.
In a similar manner, we write down equations governing F for asymmetric cell division and symmetric differentiating division. For simplicity, we assume that symmetric self-renewing division and differentiation independent of division are the dominant processes [45]. The moments of P be expressed as the partial derivatives of $F|_{z_i=1}$. Up to the first two partial derivatives we get:

$$\begin{aligned}
\partial_{z_i} F|_1 &= \langle n_i \rangle \\
\partial_{z_i}^2 F|_1 &= \langle n_i (n_i - 1) \rangle \\
\partial_{z_i} \partial_{z_j} F|_1 &= \langle n_i n_j \rangle
\end{aligned} \tag{6.8}$$

For the mean values μ_i, the variances $var(n_i)$, the coefficient of variation CV_i and the pairwise correlations r_{ij} we find:

$$\begin{aligned} \mu_i &= \partial_{z_i} F|_1 \\ var(n_i) &= \partial^2_{z_i} F|_1 + \partial_{z_i} F|_1 - (\partial_{z_i} F|_1)^2 \\ CV_i &= \frac{\sqrt{var(n_i)}}{\mu_i} \\ r_{ij} &= \frac{\partial_{z_i}\partial_{z_j} F|_1 - \partial_{z_i} F|_1 \partial_{z_j} F|_1}{\sqrt{var(n_i)}\sqrt{var(n_j)}} \quad \text{for} \quad i \neq j \end{aligned} \tag{6.9}$$

This framework allows us to compute observable quantities such as mean clone sizes, CVs and correlation coefficients by solving a set of ordinary differential equations (ODE). Other properties of interest, such as barcode concordance, Spearman's ρ and the influence of sampling cannot be calculated this way. To study these, we will resort to stochastic simulations.

6.2| Topological motifs

To study the effect of topology on the measured correlation pattern we use simplified "toy" models. We will focus on how different topological motifs affect barcode correlations.

6.2.1| Branch points

Consider the three toy models shown in fig. 6.1. Within the framework of the previous section, we compute the ODE system that allows us to calculate the moments of P.

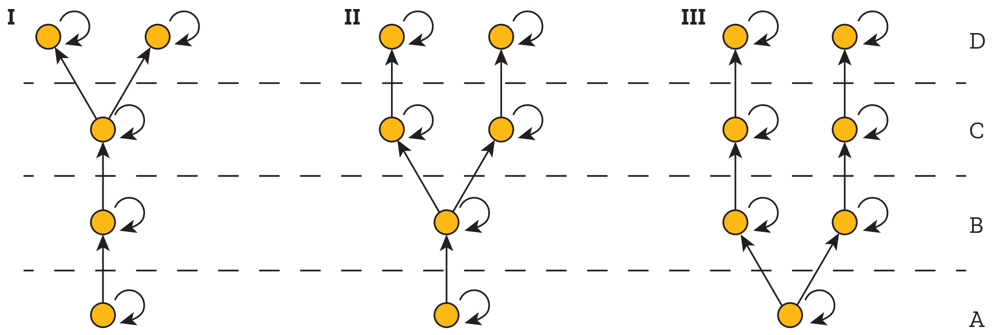


Figure 6.1: Toy models to study branching points: The effects of branching points are studied by three toy models: every population A-D has a net proliferation rate l (indicated by self-pointing arrows) and a differentiation with rate d (straight arrows). One may think of population A as the stem cell, populations B and C as progenitor cell and populations D_1 and D_2 as two different types mature cell compartments.

For the mean values we get for toy model I, where l denotes net proliferation and d differentiation.

$$\begin{aligned}
\langle \dot{n}_A \rangle &= (l_A - d_A)\langle n_A \rangle \\
\langle \dot{n}_B \rangle &= (l_B - d_B)\langle n_B \rangle + d_A \langle n_A \rangle \\
\langle \dot{n}_C \rangle &= (l_C - d_C - d_{D_1})\langle n_C \rangle + d_B \langle n_B \rangle \\
\langle \dot{n}_{D_1} \rangle &= l_{D_1} \langle n_{D_1} \rangle + d_C \langle n_C \rangle \\
\langle \dot{n}_{D_2} \rangle &= l_{D_2} \langle n_{D_2} \rangle + d_{D_1} \langle n_C \rangle
\end{aligned} \tag{6.10}$$

The second derivatives at $z_i = 1$ yield the following set of ODEs:

$$\dot{F}_{A,A} = 2l_A \langle n_A \rangle - 2(d_A - l_A)F_{A,A}$$

$$\begin{aligned}
\dot{F}_{A,B} &= (l_A + l_B - d_A - d_B)F_{A,B} + d_A F_{A,A} \\
\dot{F}_{A,C} &= (l_A + l_C - d_A - d_C - d_{D_1})F_{A,C} + F_{A,B} \\
\dot{F}_{A,D_1} &= (l_A + l_{D_1} - dA)F_{A,D_1} + d_C F_{A,C} \\
\dot{F}_{A,D_2} &= (l_A + l_{D_2} - dA)F_{A,D_2} + d_{D_1} F_{A,C} \\
\dot{F}_{B,B} &= 2(d_A F_{A,B} - d_B F_{B,B} + l_B(\langle n_B \rangle + F_{B,B})) \\
\dot{F}_{B,C} &= (l_B + l_C - d_B - d_C - d_{D_1})F_{B,C} + d_B F_{B,B} + d_A F_{A,C} \\
\dot{F}_{B,D_1} &= (l_B + lD_1 - dB)F_{B,D_1} + d_C F_{B,C} + d_A F_{A,D_1} \\
\dot{F}_{B,D_2} &= (l_B + lD_2 - dB)F_{B,D_2} + d_{D_1} F_{B,C} + d_A F_{A,D_2} \\
\dot{F}_{C,C} &= 2(d_B F_{B,C} - (d_C + d_{D_1})F_{C,C} + l_C(\langle n_c \rangle + F_{C,C})) \\
\dot{F}_{C,D_1} &= (l_C + l_{D_1} - d_C - d_{D_1})F_{C,D_1} + d_C F_{C,C} + d_B F_{B,D_1} \\
\dot{F}_{C,D_2} &= (l_C + l_{D_2} - d_C - d_{D_1})F_{C,D_2} + d_{D_1} F_{C,C} + d_B F_{B,D_2} \\
\dot{F}_{D_1,D_1} &= 2(l_{D_1} \langle n_{D_A} \rangle + l_{D_1} F_{D_1,D_1} + d_C F_{C,D_1}) \\
\dot{F}_{D_1,D_2} &= (l_{D_1} + l_{D_2})F_{D_1,D_2} + d_C F_{C,D_2} + d_{D_1} F_{C,D_1} \\
\dot{F}_{D_2,D_2} &= 2(l_{D_2} \langle n_{D_B} \rangle + l_{D_2} F_{D_2,D_2} + d_{D_1} F_{C,D_2})
\end{aligned} \tag{6.11}$$

The equivalent ODS systems for toy models II and III are found in the appendix. To minimize the influence of kinetics, parameters are restricted as follows:

- the differentiation rates at a branch point (e.g., A → B1/B2) are half of that of unbranched differentiation (e.g A, → B).
- proliferation rate is the same in all compartments

These parameter settings ensure, that the mean values are the same in all three models:

$$\begin{aligned}
&\langle n_{A,\mathrm{I}} \rangle &&= \langle n_{A,\mathrm{II}} \rangle &&= \langle n_{A,\mathrm{III}} \rangle \\
&\langle n_{B,\mathrm{I}} \rangle &&= \langle n_{B,\mathrm{II}} \rangle &&= \langle n_{B1,\mathrm{III}} \rangle + \langle n_{B2,\mathrm{III}} \rangle \\
&\langle n_{C,\mathrm{I}} \rangle &&= \langle n_{C1,\mathrm{II}} \rangle + \langle n_{C2,\mathrm{II}} \rangle &&= \langle n_{C1,\mathrm{III}} \rangle + \langle n_{C2,\mathrm{III}} \rangle \\
&\langle n_{D1,\mathrm{I}} \rangle + \langle n_{D2,\mathrm{I}} \rangle &&= \langle n_{D1,\mathrm{II}} \rangle + \langle n_{D2,\mathrm{II}} \rangle &&= \langle n_{D1,\mathrm{III}} \rangle + \langle n_{D2,\mathrm{III}} \rangle
\end{aligned}$$

To determine the effect of parameter values on the behaviour of the system we performed latin hypercube sampling (100 times to obtain 100 parameter sets in the interval of [0,1]). Latin hypercube sampling allows to generate near-random samples from a multidimensional distribution as follows: Instead of randomly drawing parameter sets from [0,1], the interval is divided into N equal portions. Suppose you have N samples in the interval [0,1] in k dimensions. First, for every dimension, we draw a random variable in the interval [0,1/N]. In the next step we draw a

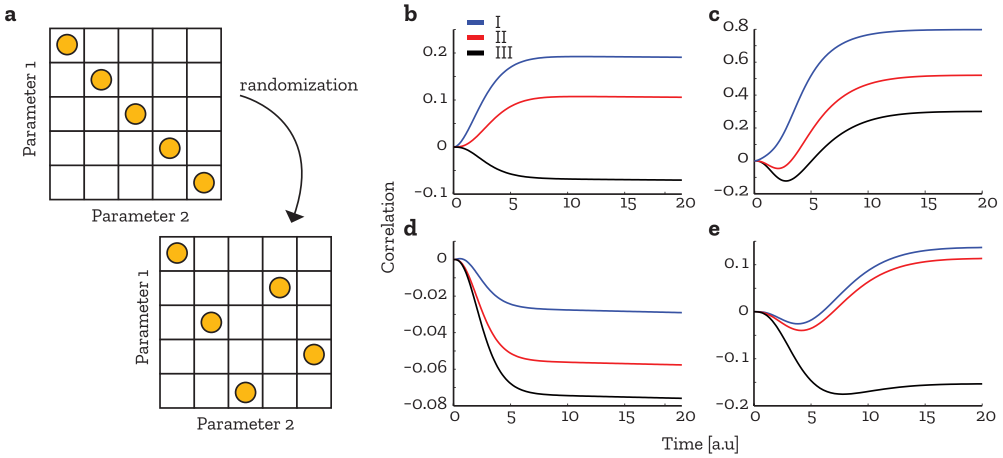


Figure 6.2: Latin hypercube sampling and effects of branching points: **a**, Scheme for parameter sampling in 2 dimensions. First each parameter is divided into equally likely parts. One value is drawn from every interval at random. In the last step the parameter pairs are randomized to reduce correlation. **b-e**, Exemplary time evolution of the correlation for the three different toy models. Blue denotes the latest split, magenta the earliest. The red line shows the model with the split in between. In the whole parameter space we found $\mathrm{r}_{\mathrm{I}}(D_1, D_2) > \mathrm{r}_{\mathrm{II}}(D_1, D_2) > \mathrm{r}_{\mathrm{III}}(D_1, D_2)$. Parameters for the shown examples can be found in the appendix

random variable from the next interval, namely [1/N,2/N]. This process is repeated until all N samples have been drawn. We then randomly reassign the drawn variables to length k vectors. This ensures sampling from the complete space with relatively small samples [46].

Topology has major influence on barcode correlations

In the complete parameter space we found that the correlations between D_1 and D_2 obey $\mathrm{r}_{\mathrm{I}}(D_1, D_2) > \mathrm{r}_{\mathrm{II}}(D_1, D_2) > \mathrm{r}_{\mathrm{III}}(D_1, D_2)$. This finding implies that the point of branching has a big impact on observed correlation patterns. The correlation distinguishes between early or late divergence in the development. Additionally the correlation also encodes information on fate restriction. For example, if a lineage that is considered bipotent actually consists of two distinct populations that are restricted to one respective fate, one would expect a lower correlation between both offspring populations than if there were no fate restriction.
Additionally, the time evolution of the correlation depends on the pathway topology. This finding implies that multiple measured time points contain meaningful information on the topology and could be used in constraining models of differentiation pathways.

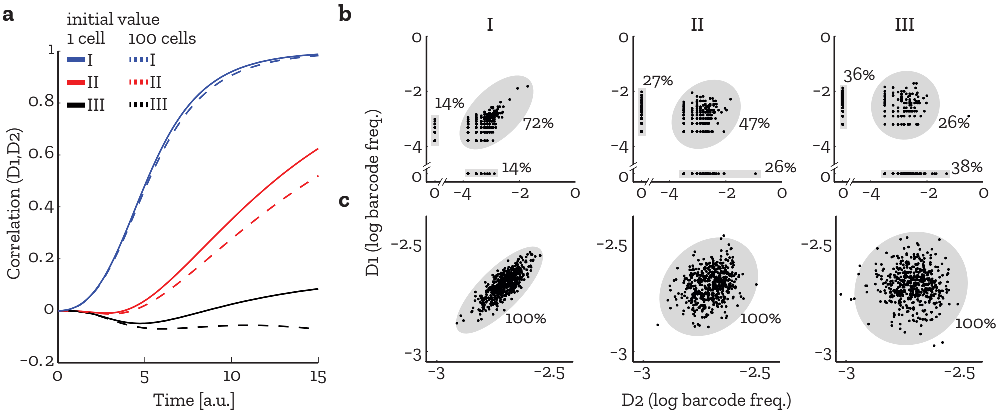


Figure 6.3: Effects of barcode clone size on the measured correlation: **a**, Time evolution of correlation between D_1 and D_2 for the three toy models. The full lines show an initial value of 1 cell, the dotted lines an initial value of 100 cells. **b,c**, Simulated clone size concordance for models I, II and III. Each dot corresponds to a single clone. Percentage indicates overlap. Initial vaules: **b**, 1 cell, **c**, 100 cells.

Clone size affects correlation patterns

To study the effects of clone sizes, we used clones with size 1 and with size 100 as initial values in all three models. While the change in clone size had only a small effect, it overall slightly reduced barcode correlations (fig. 6.3 a) between the mature population D1 and D2. This is in line with the observations made in chapter 5.1.4, where an increase in clone size decreased the correlation inside a major branch. In addition to the moment based approach, we also performed a stochastic simulation with the same parameters to see the effects of clone size on barcode concordance. For smaller clone sizes, the overlap seems to be a good indication of close developmental relationship. Even though increasing the number of starting cells leads to 100 percent shared codes, this has little impact on the correlations, with model III still leading to negative correlations (fig. 6.3 b,c). This indicates that the correlation is a better measure of developmental proximity than the overlap itself. Nevertheless, the expected percentage of shared barcodes might provide some inside into clone sizes. The high overlap for larger clone sizes is also in line with the observation that large clones in our data set tend to show multilineage potential.

6.2.2| Influence of kinetics on correlations

From our observations of correlations in chapter 5, we assumed that high correlations are not necessarily only due to close relationships of lineages but are also influenced by the kinetics of cell proliferation and differentiation. We are particularly interested in the interplay of kinetics and topology. In this section we study

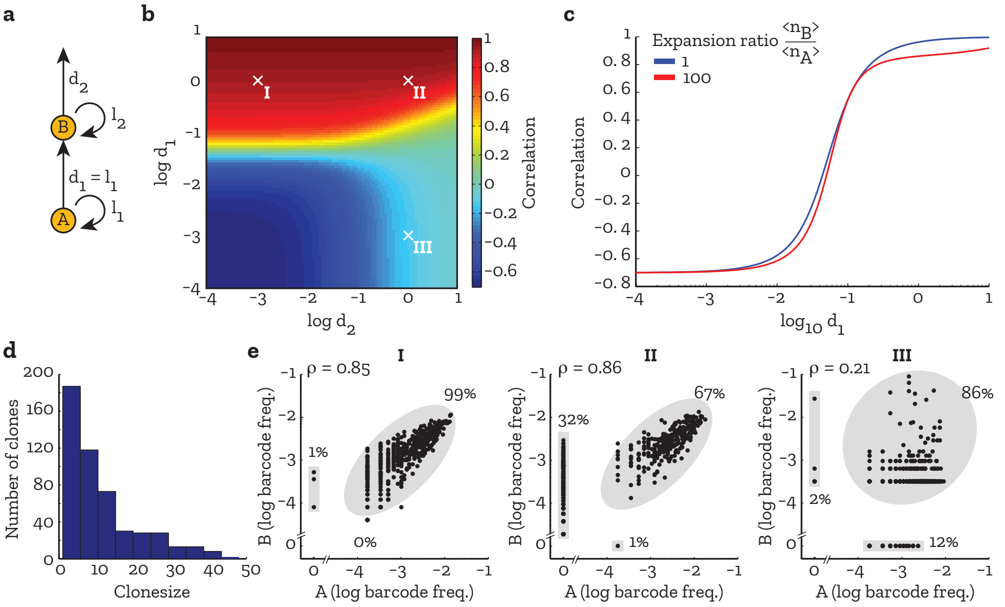


Figure 6.4: Implications of kinetics in a simple motif: **a**, Topological motif with two populations and indicated parameters. **b**, Correlation heatmap for different parameter combinations of d_1 and d_2. **c**, Line profile of **b** at $d_2 = 10^{-3}$ for different expansion ratios. **d**, Clonesize distribution used as initial values for stochastic simulation of the indicated model. **e**, Simulated barcode frequencies at steady state with different parameters as indicated in **b** (white markings). Each dot indicates a single barcode clone.

two distinct topological motifs found in the full tree: a simple linear differentiation and split into two offspring populations.

Slow kinetics cause low correlations

First we take a look at a simple differentiation process depicted in fig. 6.4 a. Here population A proliferates with rate l_1 and differentiate into B with rate d_1. B again proliferates and differentiates "out of the system" with rates l_2 and d_2 respectively. To reduce the number of parameters we set $d_1 = \mathrm{l}_1$ and therefore A is self-renewing. Since LT-HSC have been observed to be self-renewing [47], the assumption is appropriate. We then set l_2 in such a way that the expansion ratio $\frac{\langle n_\mathrm{B} \rangle}{\langle n_\mathrm{A} \rangle}$ is fixed. From the steady state solution we get:

$$\frac{\langle n_\mathrm{B} \rangle}{\langle n_\mathrm{A} \rangle} = \frac{\mathrm{d}_1}{\mathrm{d}_2 - \mathrm{l}_2} \tag{6.12}$$

We then systematically calculated the steady state correlation for all parameter combinations of d_1 and l_2 for different expansion ratios. The differentiation rate from A to B has the largest impact on the correlation (fig. 6.4 b). In the range of

10^{-4} to 10^{-2} for d_1 there is little impact on the correlation. This is followed by a steep increase as d_1 is raised further. A fast proliferation of B causes the correlation between A and B to approach zero. In particular, to keep the expansion ratio constant a rise in the d_2 also means a rise in l_2. Due to low residence times of cells in population B caused by high d_2 this then leads to a correlation of zero. If however the differentiation from A to B is also fast enough, the individual clone sizes reflect the distribution in A more closely, leading to an increase in correlation again (fig. 6.4 b, right side).
The expansion ratio itself has only a minor influence in the correlation between A and B (fig. 6.4 c). In general, a high correlation is strongly dependent on fast differentiation. Indeed, in the hematopoiesis data, we saw low barcode correlations between LT-HSC and ST-HSC, where LT-HSC differentiate very slowly, but a higher correlation between ST-HSC and MPP, where ST-HSC differentiate more rapidly ([23], see section 5.1.5).
For three parameter pairs (fig. 6.4,b; white markings) a stochastic simulation was run. To resemble the experiment more closely, the initial clone size distribution for A was drawn from an exponential distribution and is depicted in fig. 6.4 d. The simulation ran until steady state was reached. As in the ODE model, a slow differentiation from A to B had a major impact on the observed rank correlations, leading to uncorrelated barcode distributions in A and B. In contrast a fast differentiation from A to B lead to a high rank correlation largely independent on the differentiation rate d_2.

Effect of differentiation rate on downstream compartments

We now add another population C emerging from B to study the effects of different combinations of differentiation rates. Again we set $d_1 = \mathrm{l}_1$ and chose the parameters l_2 and l_3 such, that a steady state solution is possible. To better control the system we introduced again expansion ratios for the different populations instead of systematically calculating the correlations in 4 dimensions. The expansion ratios directly constrain the net proliferation rates of B and C. Since we found little to no influence of the expansion ratios on the correlation patterns we set them to 1. As in the previous section, the main driver of the correlation between A and B is the differentiation rate d_1. The same holds true for the correlation between A and C, where d_1 has a stronger impact than d_2. Generally speaking $\mathrm{r}_{\mathrm{A,C}}$ is slightly lower than $\mathrm{r}_{\mathrm{A,B}}$ (fig. 6.5 c). The correlation between B and C then again depends on both differentiation rates d_1 and d_2. Faster differentiation leads to stronger correlations.
In the last step we further increased the number of populations to 4. Here each subsequent population has a lower correlation with A than the previous populations. For slower rates the correlation tends to zero (fig. 6.5 e-h).

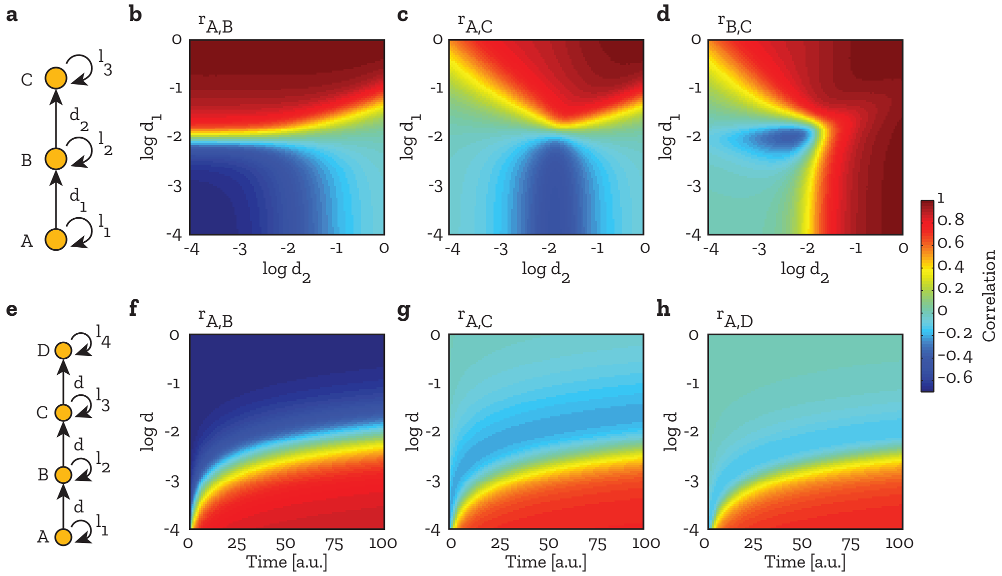


Figure 6.5: Linear progression model: a, Scheme of a linear progression model. **b-d**, Effects on the correlation for different parameter combinations. **e-h**, Same plot as above with an expanded linear model

Proliferation and differentiation kinetics influence progenitor-offspring correlations

In the next step we study the influence of kinetics on the relationship between offspring and progenitor populations, focusing on toy model I. We systematically changed the proportion of the differentiation rates to the two different types of offspring D_1 and D_2, $\frac{d_{D1}}{d_{D2}}$, thus allowing a kinetic bias in the two branches.
First, parameter sets discussed in the previous section were studied. In line with the observations made in the previous section fast differentiation drives high correlations. A heavy bias towards one of the branches lowers the expected correlation between both offspring populations drastically (fig. 6.6 b,c red line). Indeed, the highest possible offspring correlation occurs when there is no bias and differentiation is equally fast. Moreover a bias decreases the correlation between offspring and progenitor in the slower branch (fig. 6.6 b,c black and blue lines). This is to be expected, since it slows down the differentiation in one branch, which we could see in the previous section has a major impact on the correlation. Interestingly, there exist parameter combinations where offspring pairs are higher correlated than progenitor/offspring pairs and vice versa. This has also been observed in our experiments, where EryP showed a slightly higher correlation with GMP than Gr (fig. 5.9 f). In the conventional tree like model EryP and GMP would be offspring pairs (both arising from CMP), where GMP and Gr are in a progenitor-offspring relationship.

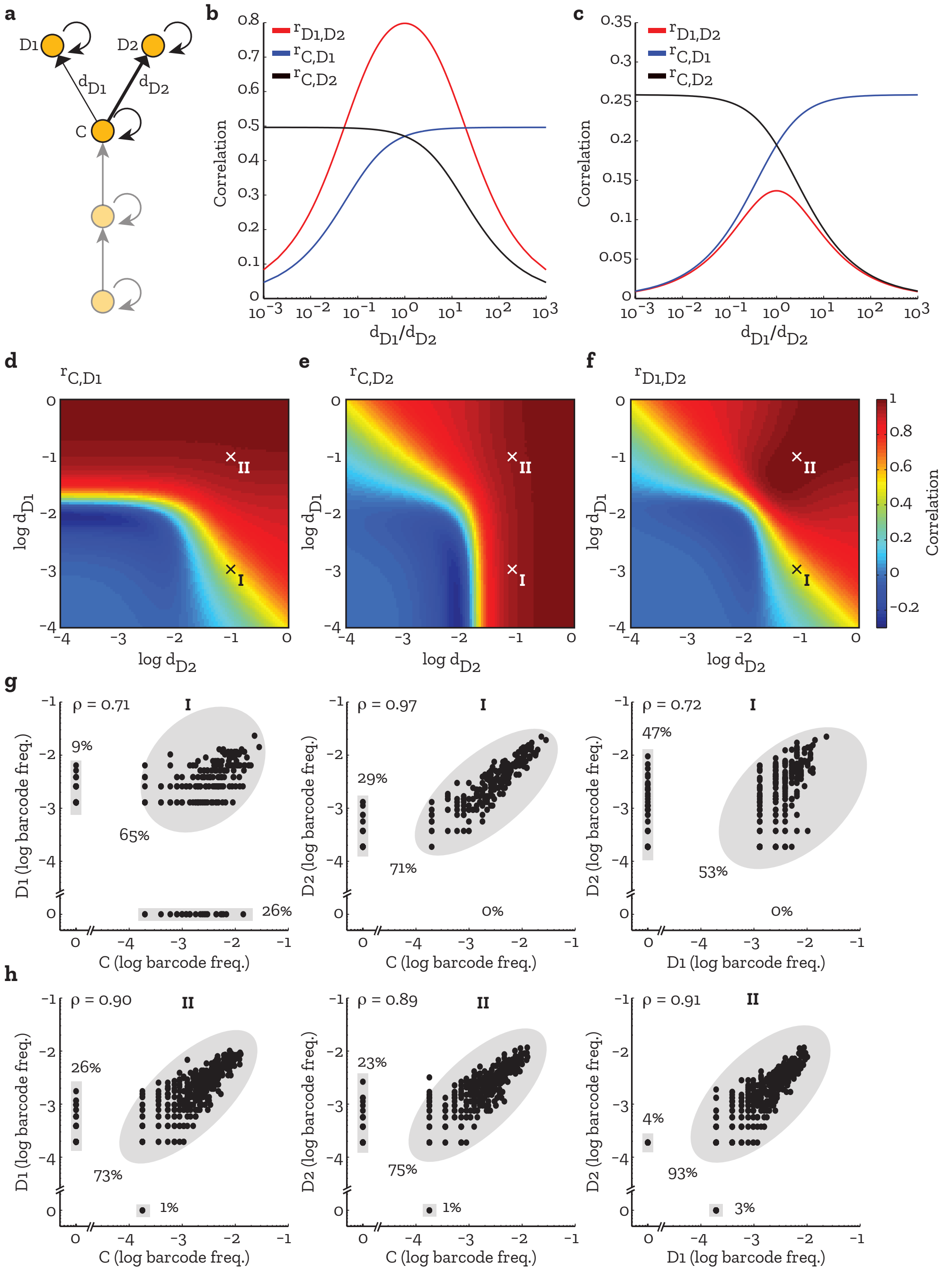


Figure 6.6: Offspring correlations at a branch point: **a**, Toy model I with emphasis on the branching point. Highlighted populations are of interest. **b**, effects on the correlation for a bias in the branch points. Parameters are the same as in fig. 6.2 c. **c**, Same plot for a different parameter set (fig. 6.2 e). **d-f**, A more general look at the influence on kinetics at a branch point. Only the highlighted populations of **a** have been taken into account. **g,h**, Simulated barcode frequencies at marked parameter values at steady state. Each dot indicates a single barcode clone.

To investigate more general parameter choices , we focused on populations C, D_1 and D_2 (highlighted part of the model shown in fig. 6.6 a). To achieve a steady state solution some constraints on the parameters are applied.

- Proliferation rate $l_C = d_{D1} + d_{D2}$, preventing C on average from dying out;
- D_1 and D_2 leave the system with rates d_{D1} and d_{D2} respectively (since we found little influence of an expansion ratio we set it to 1 here)

This leaves us with two major parameters d_{D1} and d_{D2}. As in the previous section we calculated the pairwise correlations between the three populations for every parameter combination. As expected, an increase in the differentiation rate also increased the correlation between progenitor and offspring (fig. 6.6 d,e). Due to the steady state restriction increasing d_{D2} leads also to an increase in l_C, which in turn also increases the correlation between C and D_1 (fig. 6.6 d lower right quadrant). The correlation between D_1 and D_2 appears to be a convolution of the correlation between C and D_1 and C and D_2. Keeping both parameters the same, increasing the rates also increases the correlation (fig. 6.6 f). Decreasing one of the rates while increasing the other leads to a decrease in correlations.
Further, we chose two parameter pairs for d_{D1} and d_{D2} (fig. 6.6 d-f; markings) and run a stochastic simulation using the clone size distribution from fig. 6.4 d as the initial distribution for C. At steady-state a kinetic bias in differentiation rates towards D_2 leads to a high rank correlation between C and D_2 but a lower rank correlation between C and D_1 as well as between D_1 and D_2 (fig. 6.6 g). If both differentiation steps are equally fast, rank correlations between all three populations are equal (fig. 6.6 h).

6.2.3| Barcode correlations reflect developmental relationship

In chapter 5 we analysed the data under the assumption that high correlation in barcode usage could be interpreted as a close developmental relationship. The theoretical exploration of correlation patterns in various scenarios presented in this chapter, the general assumption, that high correlations reflect a close developmental relationship, holds true. A high correlation between two populations not only indicates a close relationship in a topological but also in a kinetic sense. Therefore, measured correlations patterns may provide insight into both pathway topology and differentiation kinetics. Indeed topology is a "limiting case" of kinetics characterized by some differentiation rates being zeros. For rigorous inference of differentiation rates and hence lineage topology, correlations alone will not be sufficient [48].

7| Modeling differentiation dynamics in the hematopoietic system

The fate mapping of HSC populations previously allowed the inference of key parameters of hematopoietic stem and progenitor cells during physiological hematopoiesis [23]. In this chapter, we extend this modeling framework to include not only mean values of the cell population numbers but also quantities derived from the second moments (coefficient of variation, CV, and correlation coefficients), following the framework outlined in the previous chapter. We ask whether data obtained with *Polylox* barcoding are consistent with previous work [23].

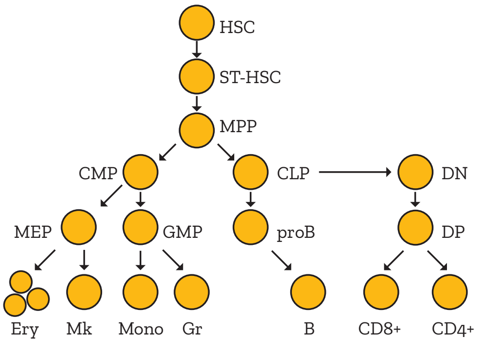


Figure 7.1: The hematopoietic system: Classical model of adult hematopoiesis. Black arrows denote differentiation steps.

7.1| Parameters of steady state hematopoiesis

As in Busch et al. [23], we focus on adult steady state hematopoiesis and use the topology described in fig. 7.1. Balancing influx and efflux leads to the following steady state equations for the mean barcode clone sizes $\langle n_i \rangle$:

$$\begin{aligned}
\langle n_{\mathrm{ST}} \rangle &= \frac{d_{\mathrm{LT,ST}}}{d_{\mathrm{ST,MPP}} - l_{\mathrm{ST}}} \langle n_{\mathrm{LT}} \rangle \\
\langle n_{\mathrm{MPP}} \rangle &= \frac{d_{\mathrm{ST,MPP}}}{d_{\mathrm{MPP,CMP}} + d_{\mathrm{MPP,CLP}} - l_{\mathrm{MPP}}} \langle n_{\mathrm{ST}} \rangle
\end{aligned} \tag{7.1}$$

$$\begin{aligned}
\langle n_{\mathrm{CMP}} \rangle &= \frac{d_{\mathrm{MPP,CMP}}}{d_{\mathrm{CMP,MEP}} + d_{\mathrm{CMP,GMP}} - l_{\mathrm{CMP}}} \langle n_{\mathrm{MPP}} \rangle \\
\langle n_{\mathrm{GMP}} \rangle &= \frac{d_{\mathrm{CMP,GMP}}}{d_{\mathrm{GMP,Gr}} - l_{d\mathrm{GMP}}} \langle n_{\mathrm{CMP}} \rangle \\
\langle n_{\mathrm{Gr}} \rangle &= -\frac{d_{\mathrm{GMP,Gr}}}{l_{\mathrm{Gr}}} \langle n_{\mathrm{GMP}} \rangle
\end{aligned} \tag{7.2}$$

$$\begin{aligned}
\langle n_{\mathrm{CLP}} \rangle &= \frac{d_{\mathrm{MPP,CLP}}}{d_{\mathrm{CLP,proB}} + d_{\mathrm{CLP,DN}} - l_{\mathrm{CLP}}} \langle n_{\mathrm{MPP}} \rangle \\
\langle n_{\mathrm{proB}} \rangle &= \frac{d_{\mathrm{CLP,proB}}}{d_{\mathrm{proB,B2}} - l_{\mathrm{proB}}} \langle n_{\mathrm{CLP}} \rangle \\
\langle n_{\mathrm{B2}} \rangle &= -\frac{d_{\mathrm{proB,B2}}}{l_{\mathrm{B2}}} \langle n_{\mathrm{proB}} \rangle \\
\langle n_{\mathrm{DN}} \rangle &= \frac{d_{\mathrm{CLP,\ DN}}}{d_{\mathrm{DN,DP}} - l_{\mathrm{DN}}} \langle n_{\mathrm{CLP}} \rangle \\
\langle n_{\mathrm{DP}} \rangle &= \frac{d_{\mathrm{DN,DP}}}{d_{\mathrm{DP,\ CD4}} + d_{\mathrm{DP,CD8}} - l_{\mathrm{DP}}} \langle n_{\mathrm{DN}} \rangle \\
\langle n_{\mathrm{CD4}} \rangle &= -\frac{d_{\mathrm{DP,\ CD4}}}{l_{\mathrm{CD4}}} \langle n_{\mathrm{DP}} \rangle \\
\langle n_{\mathrm{CD8}} \rangle &= -\frac{d_{\mathrm{DP,CD8}}}{l_{\mathrm{CD8}}} \langle n_{\mathrm{DP}} \rangle,
\end{aligned} \tag{7.3}$$

where l_i denote the net proliferation rates of population i (proliferation minus death rates), $d_{i,j}$ is the differentiation rate from population i to j. Note that the net proliferation of mature populations that do not divide (e.g. granulocytes) is less than zero.

Estimating barcode clone sizes

Due to potentially skewed PCR amplification from a single genomic *Polylox* locus per cell, read counts are not easily converted into clone size distributions. However, due to sufficiently large sample sizes (e.g. 10,000-120,000 cells) the central limit theorem allows us to estimate the population mean frequency by the sample mean frequency. The size of the HSC compartment to ~16,000 cells [23]. In addition to net proliferation and differentiation rates, Busch et al. also estimated the population sizes of the adult hematopoietic system [23].

LT	ST	MPP				
1.6	4.7	14.7				
CMP	MEP	GMP	Gr			
62.7	160	86.9	261			
CLP	proB	B2	DN	DP	CD4	CD8
206	2450	1900	2780	8360	759	245

Table 7.1: Population sizes adapted from [23]: Estimated population sizes for different cell types [x10^4]

The number of cells with barcode j in population i is then given by:

$$s_{i,j} = f_{i,j}S(i)$$

where $f_{i,j}$ the frequency of barcode j in population i and $S(i)$ the population size of population i.
The standard error of the mean (SEM) was estimated by non-parametric bootstrapping of the mean and calculating the standard deviation of the mean distribution.

Mean values are explained by refined parameters

We used the parameters estimated by Busch et al. [23] and a standard χ^2-minimization for the objective function to refine parameters for the given *Polylox* data:

$$\chi^2 = \sum_{i=1}^{k} \left(\frac{\langle n_i \rangle - \bar{n}_i}{\sigma_{\bar{n}_i}} \right) \tag{7.4}$$

The resulting best fit is depicted in fig. 7.2, showing that the tree model reproduces the observed barcode mean clone sizes. The prediction band was calculated by bootstrapping the measured mean values 10,000 times and fitting the model to each value. We then calculated the 2.5% and 97.5% quantiles of the parameters, which is found in table 7.2.
Due to the linear dependency of differentiation and net proliferation rates, rates inferred from steady state are not identifiable. Since we aim to show consistency between *Polylox* data and previous work, we used the rates estimated by Busch et al. as initial values [23] and looked for a local minimum of χ^2 eq. 7.4.
To further test the model we calculated the moments for the tree model as described in chapter 6, namely the coefficient of variation (CV) and the correlation.

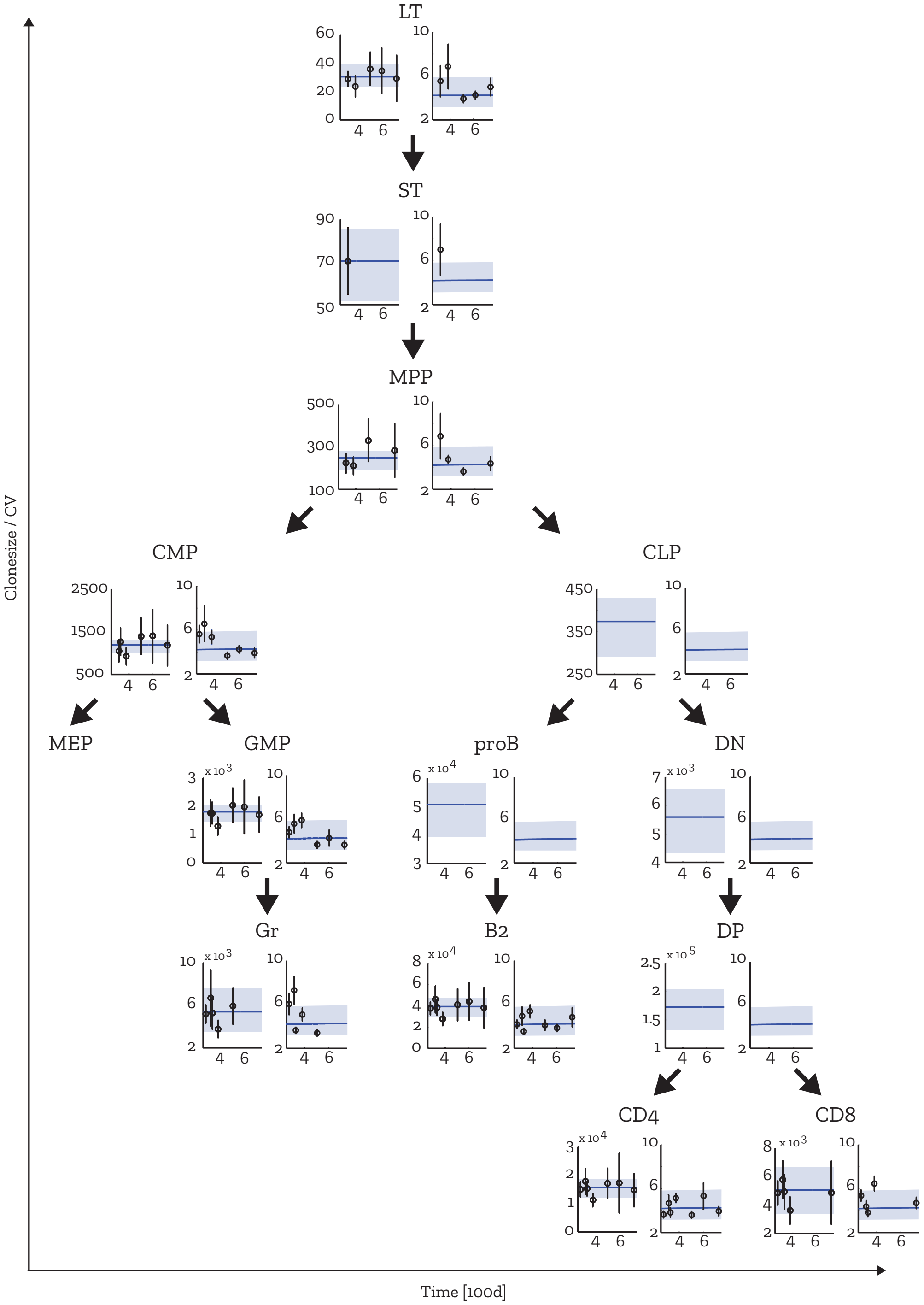


Figure 7.2: Steady-state solution: The calculated mean barcode clonesizes and the CV explains the data from embryonic *Polylox* mice. Arrows show differentiation steps along the tree. Blue line indicates the mean, light blue area indicates 95% confidence interval calculated by bootstrapping experimental values and calculating the 95% intervals of the respective best-fit parameters

Parameter: net proliferation l_i $[d^{-1}]$	Best fit	95% CI
LT	0.0088	[0.0083;0.0090]
ST	0.045	[0.044;0.045]
MPP	4	[3.99;4]
CMP	3.97	[3.84;3.99]
GMP	1.34	[1.1;1.5]
Gr	-1.33	[-1.59;-1.07]
CLP	2.64	[2.63;2.64]
proB	0.02	[0.019;0.021]
B2	-0.045	[-0.046;-0.044]
DN	3.95	[3.95;3.95]
DP	0.0024	[0.0011;0.0036]
CD4	-1.43	[-1.44;-1.43]
CD8	-0.073	[-0.073;0.072]
Parameter: differentiation $d_i[d^{-1}]$	Best fit	95% CI
LT → ST	0.0088	[0.0083;0.0090]
ST → MPP	0.048	[0.048,0.049]
MPP → CMP	3.99	[3.99,3.99]
MPP → CLP	0.022	[0.02,0.024]
CMP → MEP	0.81	[0.74,0.88]
CMP → GMP	4	[3.89,4]
GMP → Gr	3.98	[3.74,4]
CLP → proB	2.02	[2.02,2.02]
CLP → DN	0.64	[0.63,0.64]
proB → B2	0.035	[0.034,0.036]
DN → DP	4	[4,4]
DP → CD4	0.13	[0.012,0.013]
DP → CD8	0.0021	[0.0019,0.0024]
Parameter: clone size [cells]	Best fit	95% CI
LT clone size	30.35	[23.53;39.41]
LT clone size (rare)	1.28	[0.87;1.81]

Table 7.2: Best fit parameters for differentiation and LT clone sizes. Parameters where obtained by fitting the steady state solution to the barcode clone sizes. 95% intervals are calculated by fitting to bootstrapped mean values and calculating the 2.5% and 97.5% quantiles of the resulting distributions.

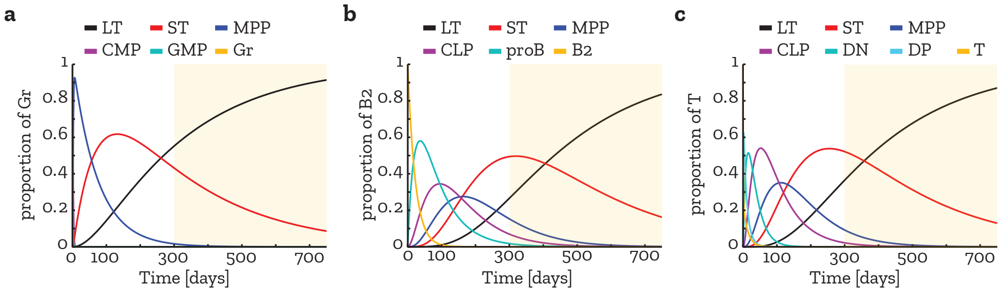


Figure 7.3: Proportion of lineages derived from progenitor populations at different time points at steady state. **a**, Gr, **b**, B2 cells, **c**, T cells. Yellow shaded area indicates experimental time frame. Barcodes retrieved experimentally are mainly LT-HSC derived and to a lesser extent also stem from ST-HSC or MPP.

7.2| Moment based modeling

We now expand the steady state model as described in chapter 6 (eq. 6.9) to calculate the coefficient of variation and the pairwise correlation. If the parameters estimated by Busch et. al and refined in the previous section explains the measured CV and pairwise correlation, this would provide further support to the differentiation model from fig. 7.1.

For the sake of brevity and readability the full ODE model of the moment based approach is cut from the main text and is found in the appendix.

Solving the initial conditions

A full covariance matrix is required as part of the initial conditions to properly calculate the steady state solutions of CVs and correlations. Since we do not have data on every population some initial values are not available.

In the first two weeks after birth the hematopoietic system rapidly equilibrates to approximately steady state [23]. Due to labelling occurring during midgestation, barcodes in mature populations in adult mice may have been subject to this fast equilibration and therefor not reflect the physiological differentiation and net proliferation rates. Given the parameters estimated previously, most of these peripheral "early" barcodes have already left the system at the time of experimental measurement. Barcodes obtained in the periphery are mainly derived from LT-HSC, ST-HSC and MPP at steady state kinetics (fig. 7.3). We therefore calculated the covariance matrix only from these populations (LT-HSC, ST-HSC and MPP) from the data. The ODE-model was then initialized for a non-steady state at birth. At the experiments time point, the system is in steady state.

The tree model is not only able to reproduce the mean but observed CV values as shown in fig. 7.2.

Attenuation of correlation

Since sampling introduces errors in the measured barcode frequency distributions, the expected correlation from the model overestimates the observed correlation [35, 49]. In this section we will introduce a way to correct for an attenuation of correlation due to sampling and sequencing.
Let f' and g' be sampled barcode frequency distributions from f and g. ϵ_f and ϵ_g are then the associated measurement errors due to sampling.

$$f' = f + \epsilon_f \qquad \text{and} \qquad g' = g + \epsilon_g \tag{7.5}$$

The measured correlation of the sample is then transformed into the population correlation by the following steps:

$$\begin{aligned} \text{corr}(f', g') &= \frac{\text{cov}(f', g')}{\sqrt{\text{var}(f')\text{var}(g')}} \\ &= \frac{\text{cov}(f + \epsilon_f, g + \epsilon_g)}{\sqrt{\text{var}(f + \epsilon_f)\text{var}(g + \epsilon_g)}} \\ &= \frac{\text{cov}(f, g)}{\sqrt{\text{var}(f)\text{var}(g)}} \frac{\text{var}(f)\text{var}(g)}{\sqrt{\text{var}(f + \epsilon_f)\text{var}(g + \epsilon_g)}} \\ &= \text{corr}(f, g)\sqrt{R_f R_g}\,, \end{aligned} \tag{7.6}$$

where R_f and R_g are called reliability coefficients under the assumption of uncorrelated ϵ and is calculated from eq. 7.6 as follows:

$$R_f = \frac{\text{var}(f)}{\text{var}(f) + \text{var}(\epsilon_f)} \qquad \text{and} \qquad R_g = \frac{\text{var}(g)}{\text{var}(g) + \text{var}(\epsilon_g)} \tag{7.7}$$

Since only f' and g' is known experimentally, it is not possible to calculate the reliability coefficients directly. Instead, to obtain population frequency distributions, a stochastic simulation is used [27]. By using the parameters from table 7.2 we obtain population barcode distributions. By sampling from those distributions the reliability coefficients are estimated *in silico*, given a mean sample depth for each population. Where the sample depth of population i is given by:

$$s_i = \frac{\#\text{sampled cells}_i}{\text{population size}_i}\,, \tag{7.8}$$

where the number of sampled cells is between 10,000 and 50,000 cells.
While we already showed in fig. 4.2 that sequencing repeats are robust, sequencing introduces noise contributing to an attenuation of correlation. Repeating the steps from eq. 7.6 for two distributions k_1 and k_2 obtained through repeated sequencing of the same sample and using $\text{corr}(k_1, k_2) = 1$ we find:

$$\begin{aligned} \text{corr}(k_1', k_2') &= \text{corr}(k_1, k_2)\sqrt{R_s R_s} \\ &= R_s \end{aligned} \tag{7.9}$$

Estimated reliability coefficients from sequencing repeats and simulations are found in table 7.3.

reliability coefficient	mean	95% CI	mean (rare)	95% CI (rare)
R_s	0.918	[0.758;0.982]		
R_{LT}	0.991	[0.954;1]	0.931	[0.686;1]
R_{ST}	0.966	[0.951;0.978]	0.781	[0.709;0.835]
R_{MPP}	0.971	[0.950;0.995]	0.939	[0.875;0.974]
R_{CMP}	0.982	[0.957;0.996]	0.962	[0.898;0.991]
R_{GMP}	0.983	[0.958;0.995]	0.961	[0.876;0.991]
R_{Gr}	0.993	[0.987;0.997]	0.985	[0.971;0.994]
R_{CLP}			N.D.	
R_{proB}			N.D.	
R_{B_2}	0.983	[0.972;0.993]	0.965	[0.926;0.9862]
R_{DN}	0.886	[0.834;0.924]	0.799	[0.665;0.868]
R_{DP}			N.D.	
R_{CD4}	0.976	[0.952;0.987]	0.950	[0.898;0.979]
R_{CD8}	0.982	[0.968;0.991]	0.962	[0.985;0.985]

Table 7.3: Reliability coefficients estimated from sequencing repeats with eq. 7.9 (R_s), or by sampling from a simulated data-sets using eq. 7.6

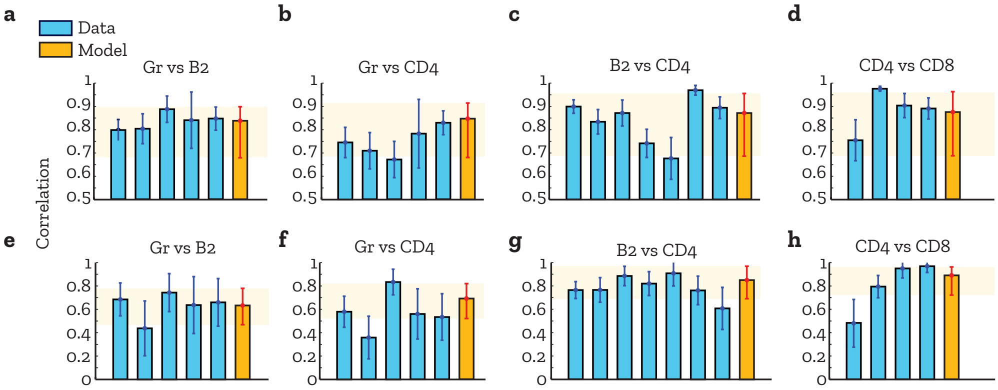


Figure 7.4: Pairwise correlation of mature populations: Data of different mice compared with model prediction of the shown populations. Data is indicated in light blue bars, model prediction in yellow. Model prediction bands are shown as yellow shaded area. Mean values and standard deviation obtained by non-parametric bootstrapping are shown for data; mean and prediction bands for the model. Top row shows all barcodes, bottom row rare barcodes with $P_{\text{gen}} < 10^{-4}$.

7.2.1| Emerging correlation patterns

The moment based approach allows us to calculate the pairwise correlation. Using the reliability coefficients estimated in the previous section we correct for attenuation effects introduced by the experimental setup.

Model predicts correlations of mature populations

We focused on the mature populations (Gr, B2, CD4 and CD8) first. The estimated parameters (table 7.2) allow a good prediction of the expected pairwise correlation values (fig. 7.4). With a few exceptions all experimental values are in the prediction bands of the model, for both all and rare barcodes with $P_{\text{gen}} < 10^{-4}$, further supporting the tree model.

CMP/GMP correlations are predicted correctly

As one of the conclusions of chapter 5.1.5 was the potential of CMP/GMP as progenitors of myelo-erythroid branch, we also compared measured correlations with model predictions. In the model CMP/GMP are fate restricted into the myelo-erythroid branch. We compared data from CMP and GMP with the mature populations in the previous section. We considered all barcodes and rare barcodes with $P_{\text{gen}} < 10^{-4}$. The model accurately predicts the lower correlation between CMP/GMP and the lymphoid lineages, as well as the correlation values itself (fig. 7.5). The accurate model prediction of the pairwise correlation of CMP/GMP with the mature populations is a strong indicator that CMP/GMP indeed possess myelo-

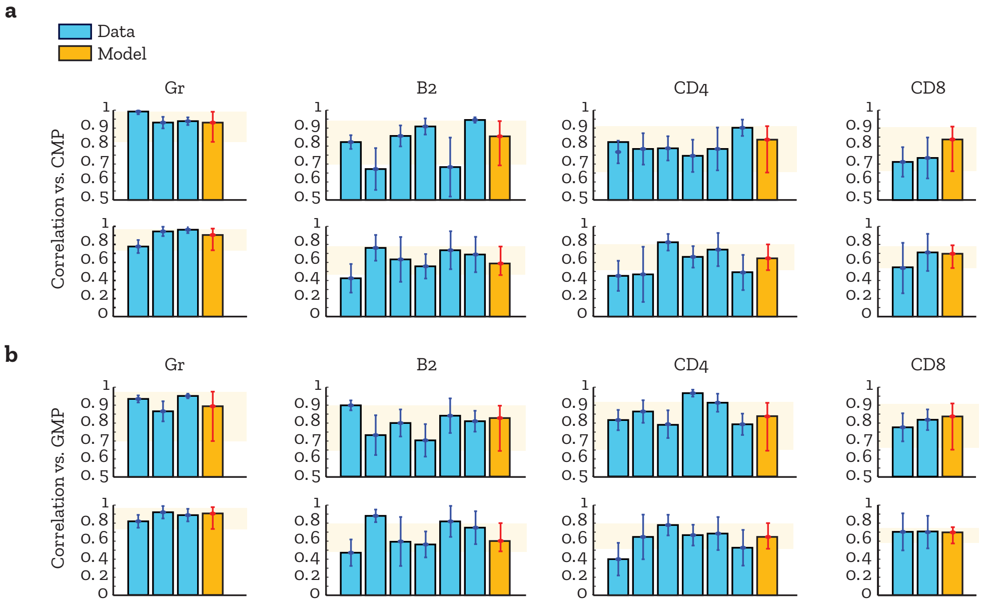


Figure 7.5: Pairwise correlation of progenitor populations: Data of different mice compared with model prediction of the shown populations. Data is indicated in light blue bars, model prediction in yellow. Model prediction bands are shown as yellow shaded area. Mean values and standard deviation obtained by non-parametric bootstrapping are shown for data; mean and prediction bands for the model. Top row shows all barcodes, bottom row rare barcodes with $P_{\text{gen}} < 10^{-4}$.

erythroid potential in vivo and act as direct progenitor for granulocytes.

7.3| Stochastic simulations

In the last section of this chapter, we will use stochastic simulations to calculate Spearman's rank correlation ρ and compare model predictions with observed correlations. While the model predicts pairwise correlations, Spearman's ρ is a more robust measure of correlation. In this section we used Gillespie's algorithm to simulate barcode propagation through the hematopoietic system. The refined parameters from the previous section were used (table 7.2).

Rank correlations are explained by model

We focused on the mature populations first and rare barcodes with $P_{\text{gen}} < 10^{-4}$. Here, we ask whether the model reproduces the observed Spearman's rank correlation. The simulation has been initialized by drawing from the rare barcode clone sizes from the data for LT, ST and MPP populations and ran for 400 days in

order to allow for equilibration. We then sampled from the resulting distributions according to sampling depths used in the experiment and compared barcode frequencies as done in chapter 5.1.3.
We found that the barcode usage in unrelated populations was lower and less concordant, reflecting in a lower ρ, than in closely related populations. The simulation of the tree like system with the refined parameters accurately describes the data from various mice (fig. 7.6). In the simulation of the model as well as in the data we find a strong barcode concordance between lymphoid lineages ($\rho_{\textbf{model}}$ = 0.89-0.97, $\rho_{\textbf{data}}$ = 0.78-0.88, min/max values). Granulocytes showed a significantly lower correlation with the common lymphoid lineages ($\rho_{\textbf{model}}$ = 0.51-0.61, $\rho_{\textbf{data}}$ = 0.39-0.56). We found that the model is in good agreement with the experimental data. In total, the correlations predicted by the model were systematically slightly higher than the correlations observed in the data. Since we do not take sequencing errors into account in the simulations, this is to be expected.

CMP/GMP progenitor explained by model

From the same simulation we also extracted CMP and GMP data and compared barcode usage with different mature populations (fig. 7.7). We found a strong con-

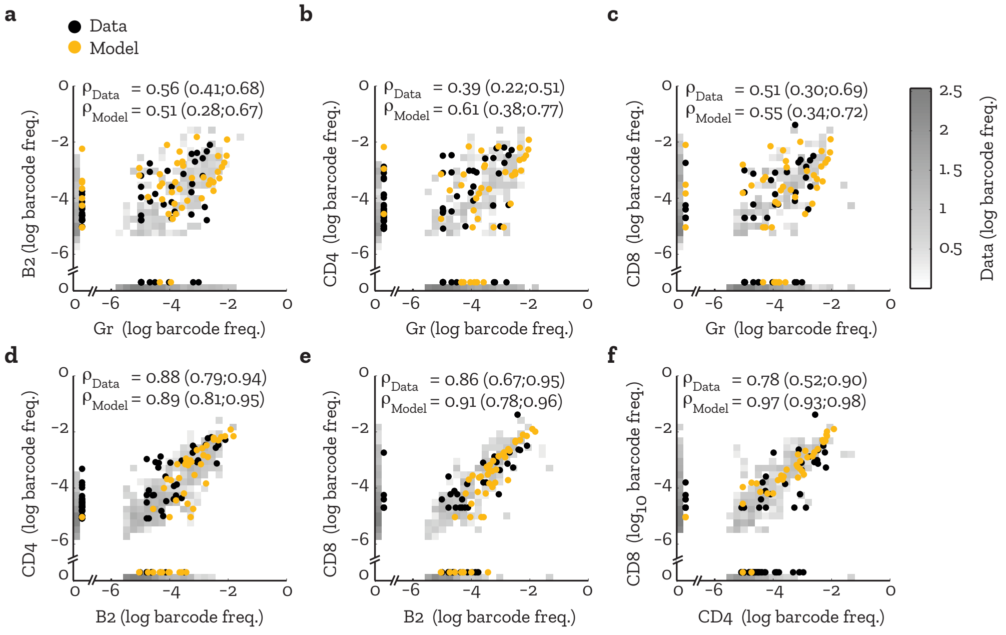


Figure 7.6: Comparison of barcode usage: Every dot indicates a single rare barcode. Yellow dots indicate unique barcodes of a single simulation run, black dots indicate data. Underlying heatmap is a summary of all experiments. ρ of model and data on top of each scatter plot, 95% confidence interval in brackets.

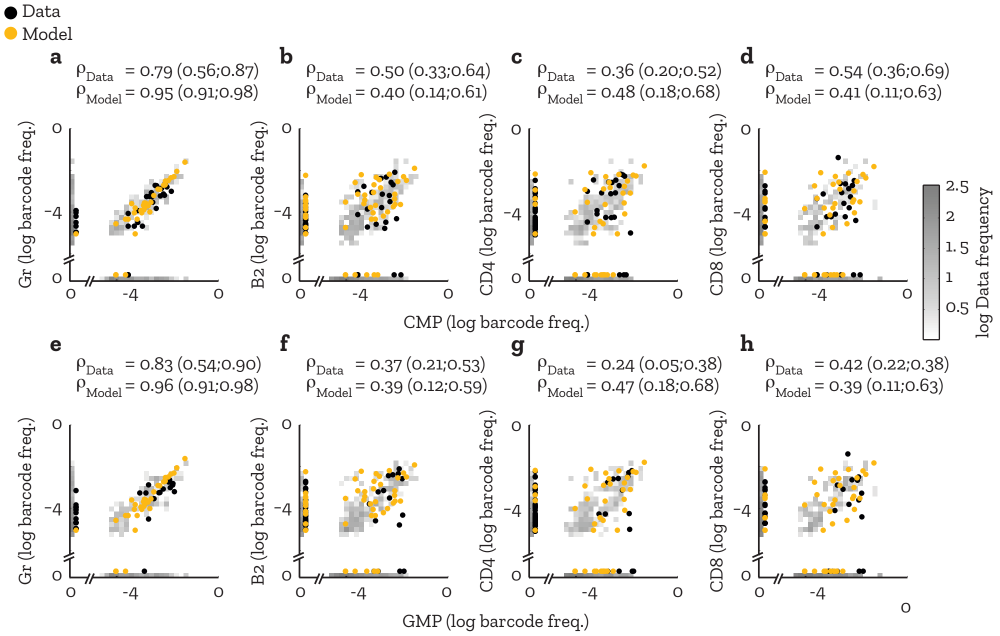


Figure 7.7: Comparison of barcode usage in progenitor populations: Every dot indicates a unique rare barcode. Yellow dots indicate a single simulation run, black dots indicate data. Underlying heatmap is a summary of all experiments. ρ of model and data on top of each scatter plot, 95% confidence interval in brackets. Top row: indicated populations vs. CMP, bottom row vs. GMP

cordance in barcode usage in CMP and GMP with Gr ($\rho_{\textbf{model}}$ = 0.95/0.96 vs CMP and GMP respectively, $\rho_{\textbf{data}}$ = 0.79/0.83). CMP and GMP shared a significantly lower number of barcodes with lymphoid lineages (B2, CD4, CD8) in both the simulation and the data ($\rho_{\textbf{model}}$ = 0.39-0.48, $\rho_{\textbf{data}}$ = 0.24-0.54). As in the previous section, the tree model of hematopoiesis explains emerging correlation patterns, as well as frequency distributions of barcode clones in the measured lineages. This concordance of model and data provides further support for the classical tree model of hematopoiesis (fig. 7.1).

Hierarchical clustering compares to data

In the next step we performed the same hierarchical clustering analysis as described in section 5.1.3 on the simulated barcode distributions. We found a correlation pattern that closely resembles the data (fig. 7.8). Two distinct clusters are observed: one with myeloid progenitors as well as granulocytes resembling the myelo-erythroid branch and one cluster containing all the lymphoid populations, the common lymphoid branch. Due to the slow differentiation from LT to ST, LT are far removed from the other populations. This has also been observed in our

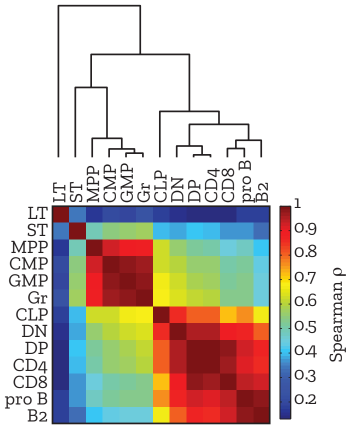


Figure 7.8: Cluster of simulation data: a, Hierarchical clustering of simulated rare barcodes of indicated populations. Distance measure used is $d_{i,j} = 1 - \rho_{i,j}$. **b**, A single barcode simulation over time of indicated populations in steady state.

data (see fig. 5.8). Since a strong bias of MPP towards CMP in terms of differentiation has been observed [23], MPP cluster inside the myelo-erythroid branch. We see the diminishing correlations downstream from MPP (e.g. $\rho_{\text{MPP;CMP}} > \rho_{\text{MPP;GMP}} > \rho_{\text{MPP;Gr}}$) as explored in chapter 6.

8| Barcode network analysis

In the previous chapters we focused on single barcodes. While the identity of each barcode allows us to assign a generation probability P_{gen} to it, the correlated generation of closely related barcodes has been ignored. In this chapter we want to analyse the total barcode composition of different experiments and the resulting implications on barcode creation. Since *Polylox* barcodes are created in distinct steps, the presence of intermediate products reveal information on the creation process itself.

8.1| Connected Barcodes

In this section we explore what relationships between barcodes tell us about the creation process and the state of the system at time of creation. We call two barcodes related if one excision or inversion event transform one into the other.

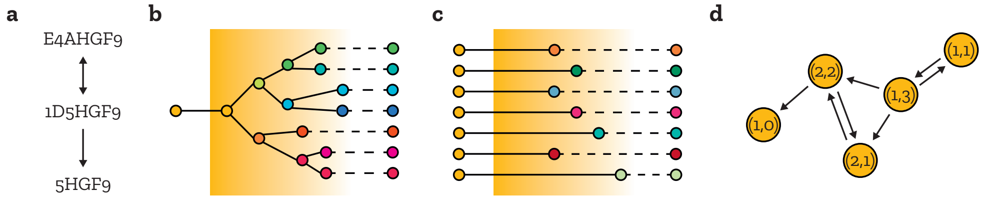


Figure 8.1: Connection of barcodes: **a**, Barcodes relation can be one-sided by excision and two-sided by inversion events. **b**, Scheme of barcode creation with strong proliferation. Recombination timeframe is indicated as shaded yellow area. Different barcodes are coded by different color. **c**, same as **b** but without proliferation. **d**, Exemplary network of barcodes. In brackets are in-degree and out-degree of each barcode respectively.

This relationship can be one-sided in case of an excision event (e.g. 1D5HGF9 →5HGF9) or two-sided for an inversion event (e.g. 1D5HGF9 ↔ E4AHGF9) (fig. 8.1). Now, suppose the system consists of a single unlabelled cell. During tamoxifen treatment and therefore recombination, the cell proliferates while *Polylox* is recombined (fig. 8.1 b). The barcode composition will be different than that of a system where no proliferation occurs in the time frame of recombination (fig. 8.1 c). The two systems will not only differ in clone sizes for each barcode, but the

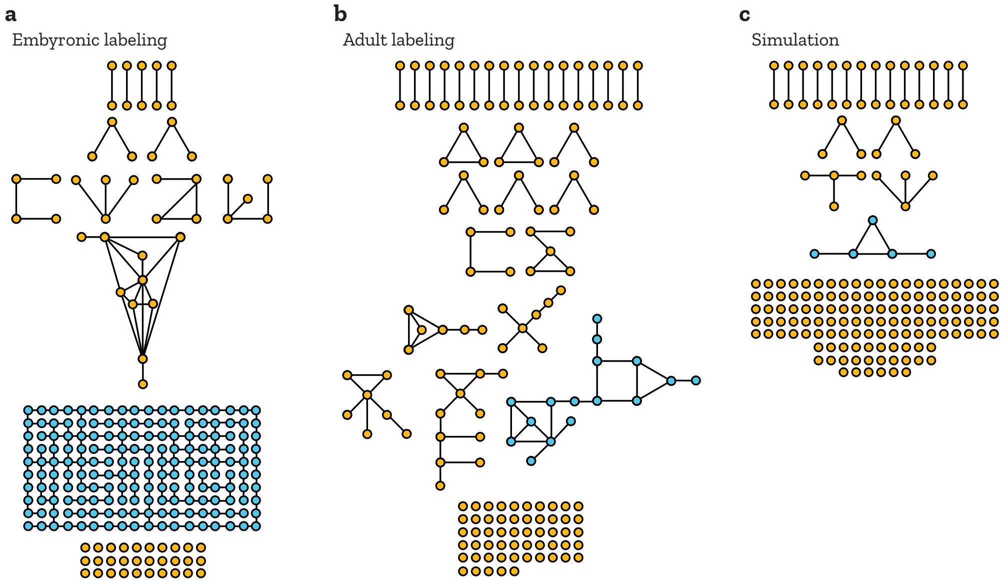


Figure 8.2: Comparison of subcluster sizes of barcodes sets: Each dot represents a unique rare barcode, black line denotes relationship. Light blue indicates the largest subcluster. For readability there is no distinction between two-sided and one-sided relationships. Sets of rare barcodes retrieved from a, An embryonic labelled mouse, **b**, Adult labelled mouse. **c**, Random barcodes according to P_{gen}

barcodes in fig. 8.1 b will generally be more related than in fig. 8.1 c.

Proliferation is visible in barcode composition

To see whether concurrent proliferation and recombination has an influence on the barcode composition we analysed barcodes from three different experimental setups. First, mice labelled at the embryonic stage: here we expect a strong proliferation as the HSC compartment massively expands during midgestation [6, 7]. Second, mice labelled at the adult stage: from the steady state solution from the previous chapter we expect the proliferation to be very slow in comparison to recombination. Third, drawing random barcodes according to P_{gen} estimated from experiments: here no proliferation is taken into account.

Since all setups generate the unrecombined barcode as well as barcodes of length 1 (e.g. fully recombined barcodes), we focused on rare barcodes only. The barcodes in the embryonic labelling experiment where highly related, where most barcodes formed one big cluster of more than ~70% of all rare barcodes (fig. 8.2 a). In the case of adult labelling, we found much smaller and fewer subclusters of barcodes. Here the biggest cluster contain only about 8% of all barcodes (fig. 8.2 b). For the simulation without proliferation we found that the vast majority of barcodes had

no relationships at all (fig. 8.2 c).
This shows that the connectivity of the subset of retrieved barcodes reflects the proliferative state of the system during labelling.

8.2| Degree distributions

To properly quantify such properties as relationships we introduce two measures: in-degree and out-degree. The in-degree gives the number of total barcodes that are recombined in one step to the barcode in question, while the out-degree gives the number of total barcodes reachable within one step by recombination from the given barcode (fig. 8.1 c). From the adjacency matrix A from section 2.3 we calculate the degrees, by summing over the subset of found barcodes.

$$d_i^+ = \sum_{j \,\in\, \text{sample}} A_{ij} \quad \text{and} \quad d_i^- = \sum_{j \,\in\, \text{sample}} A_{ji} \tag{8.1}$$

where d_i^+ is the out-degree of barcode i, d_i^- the in-degree.

Proliferation is visible in degree distributions

In order to assess the influence of proliferation on the observed degree distributions we compare data from embryonic labelled mice with a random sampling of 1000 barcode sets. Each set contains similar barcode numbers than retrieved from the experiment. To make the distributions more comparable we sample barcodes according to their P_{gen} we estimated from the embryonic labelled mice (see. chapter 2). In fig.8.3 degree distributions for embryonic labelled mice are shown together with a random sampling of barcodes. In the top row the degree distributions for all sampled barcodes are shown, while in the bottom row we filtered for rare barcodes. There is a significant difference in both, the in- and out-degree distribution, indicating that concurrent proliferation and recombination is detectable in the experiment.

Minimal recombination distributions reveal recombination rate

In the next step, we want to see whether one can estimate the proliferation rate at the time point of labelling with the help of those distributions. Before we estimate the proliferation rate from the degree distribution we also need to estimate the rate and timeframe of *Polylox* recombination. As discussed in chapter 2, due to inversions being reversible, it is not possible to directly estimate the number of recombination events every experimentally retrieved barcode has undergone. Under the assumption that individual recombination events occur independently of each other, the probability for m events is described by a Poisson distribution

$$p(m \text{ events} \mid \lambda) = e^{-\lambda} \frac{\lambda^m}{m!} \tag{8.2}$$

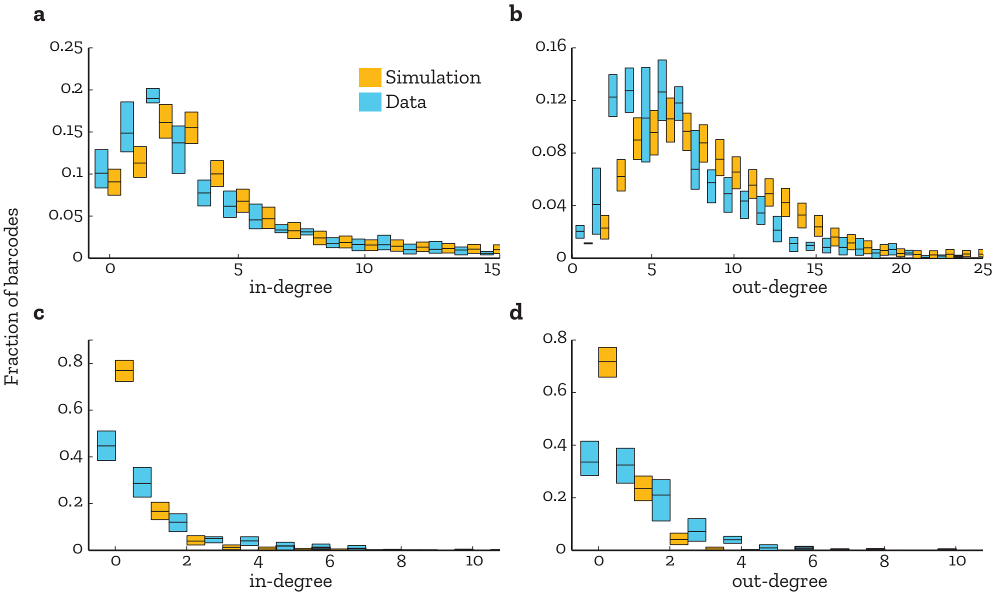


Figure 8.3: Degree distribution of barcodes induced in the embryo: Comparison of degree distributions of all barcodes (top row) and barcodes with $P_{\text{gen}} < 10^{-4}$ (bottom row) of different experiments and simulation. In the simulation no proliferation was taken into account when creating the barcode sets analyzed. The data comes from embryonic labelled mice, where a strong proliferative behaviour is expected during recombination. Simulated and experimentally derived distribution diverge significantly.

with rate parameter λ, which is directly connected to the recombination rate r by $\lambda = rt$. λ is the average number of recombination events per labelling. We use this probability distribution as weights $\omega(m)$ to calculate P_{gen} for different parameters λ. P_{gen} is then given by:

$$P_{\text{gen},\lambda} = \sum_{m=1}^{m_{\max}} p(\lambda, m) T^m P_{\text{gen},0} \quad \text{and} \quad P_{\text{gen},0} = (1, 0, 0, ...)^\top, \tag{8.3}$$

where T is the transition matrix from eq. 2.5 ff. For different values of λ we draw a set of barcodes according to $P_{\text{gen},\lambda}$. We then calculate the distribution of minimal numbers of recombinations for every barcode, as described in section 2.6. This allows us to compare the simulated frequency distributions with the experimentally retrieved sets of barcodes.

In fig. 8.4 a, the frequency of minimal number of recombinations of experimentally retrieved barcodes is shown as a bar plot. Different rates λ are overlaid. A recombination rate parameter of $\lambda = 3$ (fig. 8.4 a, yellow line) produces a fitting distribution. As expected, a low recombination parameter results in a distribution that is shifted to the left, while increasing the recombination rate parameter broadens the distribution and shifts to the right.

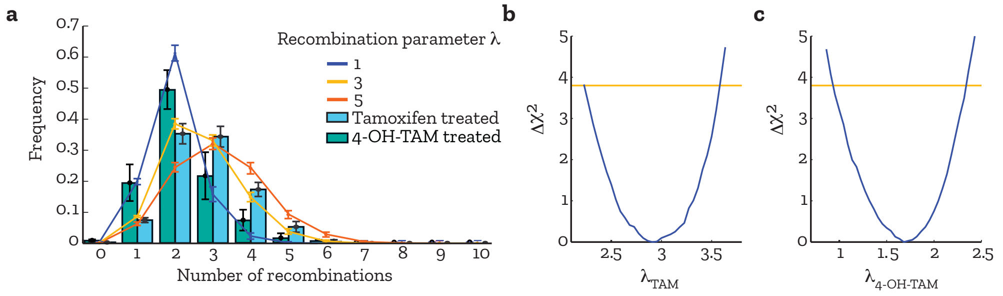


Figure 8.4: Recombination distribution of barcodes induced in the embryo: a, Frequency distribution of minimal numbers of recombination for different experimental setups. Teal bars show tamoxifen treated mice, green bars show 4-OH-Tam treated mice. Colored lines show simulation with different recombination parameters λ. Mean and standard deviations of 100 simulated barcode sets are shown, as well as mean and standard deviation of the data. **b,c**, Profile likelihood curves for different λ parameters and both experimental setups. The yellow line indicates $\chi^2_{95\%}$ = 3.8 and therefore the 95% confidence interval on the parameter.

In addition to mice treated with tamoxifen, we repeated the analysis on data from mice directly treated with the active component of tamoxifen, 4-OH-Tam. In contrast to tamoxifen which has a depot effect [50], 4-OH-Tam is believed to degrade much faster. Since in both experiments 4-OH-Tam is the active component, it is reasonable to assume the same recombination rate r. However, since the 4-OH-Tam degrades more quickly, we expect a lower λ, as $\lambda = rt$, where t is the length of the time interval with present 4-OH-Tam.

Data from mice treated only with 4-OH-Tam has been plotted as bars in green in fig. 8.4 a. Indeed we see that the distribution is much narrower and shifted to the left, as expected for a lower λ. Here, among the used values of λ, $\lambda = 1$ reproduces the experimentally observed distributions the best.

To further quantify λ we scanned the relevant λ-range for both experimental setups and calculated χ^2 for every $P_{gen,\lambda}$.

$$\chi^2 = \sum \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2 \tag{8.4}$$

where x_i is the mean fraction of barcodes that are reachable within i recombination events of the simulation, μ_i denotes the mean of the data, σ_i the standard deviation of the data. In order to reduce stochastic effects introduced by simulating, for every value of λ, 1000 sets of barcodes were simulated. Parameter estimates and their uncertainties were calculated by means of calculating the profile likelihood [51]. Here the best fit parameter is λ for which χ^2 is minimal, denoted as χ^2_{min}. The 95% confidence region is then defined as containing every λ for which

$$\chi^2(\lambda) - \chi^2_{\text{min}} \leq \chi^2_{95\%} \tag{8.5}$$

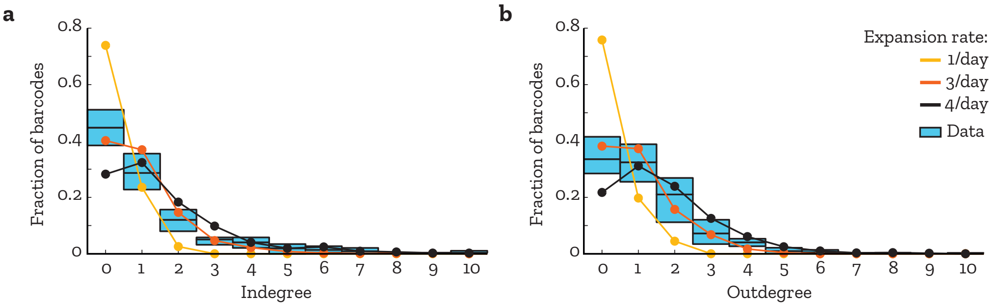


Figure 8.5: Degree distribution with proliferative behaviour: In- (**a**) and out-degree (**b**) of experimental data (teal, middle bar shows mean, top and bottom indicates 95% confidence interval) compared to different expansion rates during the time interval of recombination. Only a high expansion rate explains the experimental data. The proliferation rate is kept constant over the time frame of recombination.

holds true. Fig. 8.4 b and c show the $\Delta\chi^2(\lambda)$ of tamoxifen and 4-OH-TAM treated mice respectively. With $\chi^2_{95\%}$ = 3.8, we obtain λ_{TAM} = 2.91 (2.23;3.63) and $\lambda_{\text{4-OH-TAM}}$ = 1.68 (0.94;2.32). This indicates that tamoxifen is active for about 1.7 times as long as 4-OH-TAM.

Pharmacokinetic experiments have shown that 4-OH-Tam degrades rapidly within 24 hours in mice at a dose of 0.04mg [52]. A higher dose in the case of the experiments described in this thesis (2.5mg) will also mean a longer residence time of 4-OH-TAM [53]. This scenario is in line with the observation made here. At E8.25 macrophage progenitors arise (see fig. 1.1), which are Tie2 positive, and are therefore labelled. Later, at E9.5 HSC progenitor arise in the AGM (Aorta-gonad-mesonephros). In the case of labelling mice at E7.5 with 4-OH-Tam we found recombined barcodes in macrophage progenitor-derived populations, but not in HSC. In the case of regular tamoxifen treatment, recombination could be seen in both populations. This result puts a limit on the active time frame of 4-OH-Tam and we estimate the time interval in which 4-OH-Tam is active to be between 24 and 48 hours. Tamoxifen is therefore active for approximately 48-96 hours after treatment.

With $\lambda_{\text{TAM}} = rt_{\text{TAM}}$ and $\lambda_{\text{4-OH-TAM}} = rt_{\text{4-OH-TAM}}$ we estimate the recombination rate during the time interval of treatment to be about $r = 1.00\ (0.89;1.11)\frac{1}{\text{day}}$

Fast proliferation in embryo

The estimate of the recombination rate r then allows us to simulate the generation of barcodes with different proliferation rates using a stochastic simulation approach. We then calculated the in- and out-degree distributions of the set of rare barcodes and compared it to the distributions obtained from experimental data. In this simulations we found that a higher proliferation rate leads to a highly

connected network of barcodes, where a low proliferation rate produces more independent barcodes.
These simulations allow us to estimate the proliferation rate of Tie2+ HSCp in the embryo to be around $p = 3\frac{1}{\text{day}}$ (fig. 8.5). The indirect nature of this measurement makes it difficult to obtain a more exact number on the proliferation rate. Rapidly proliferating Tie2+ HSCp during labelling are needed to explain the experimentally observed barcode sets.

8.3| Qualitative description of embryonic development

Taking into account everything discussed in this chapter so far, our aim in this section is to formulate a qualitative description of embryonic development. Using the experimentally observed barcode sets we will estimate the proliferation rate during the time interval of recombination. Further, the barcode complexity will allow us to estimate a lower bound of cell numbers at the time point when recombination stopped.

Estimating barcode complexity

The absolute lowest cell number needed to explain a given barcode complexity i.e. the total number of unique barcodes, is the barcode complexity itself. However this does not apply to highly abundant barcodes, which are most probably generated more than once (e.g. fully recombined barcodes), but counted only once. Therefore for abundant barcodes the minimal required cell number is likely higher then than the barcode complexity. For a better estimate of cell numbers, we therefore sample barcodes according to their generation probability P_{gen}. This allows us find the number of total barcodes needed to explain a given number of unique barcodes.
Due to possible undersampling and barcode clones leaving the system via cell

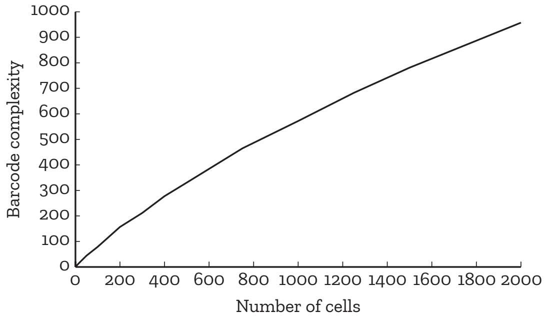


Figure 8.6: Barcode complexity given cell number: Random sampling of a given number of barcodes and the retrieved number of unique barcodes (barcode complexity).

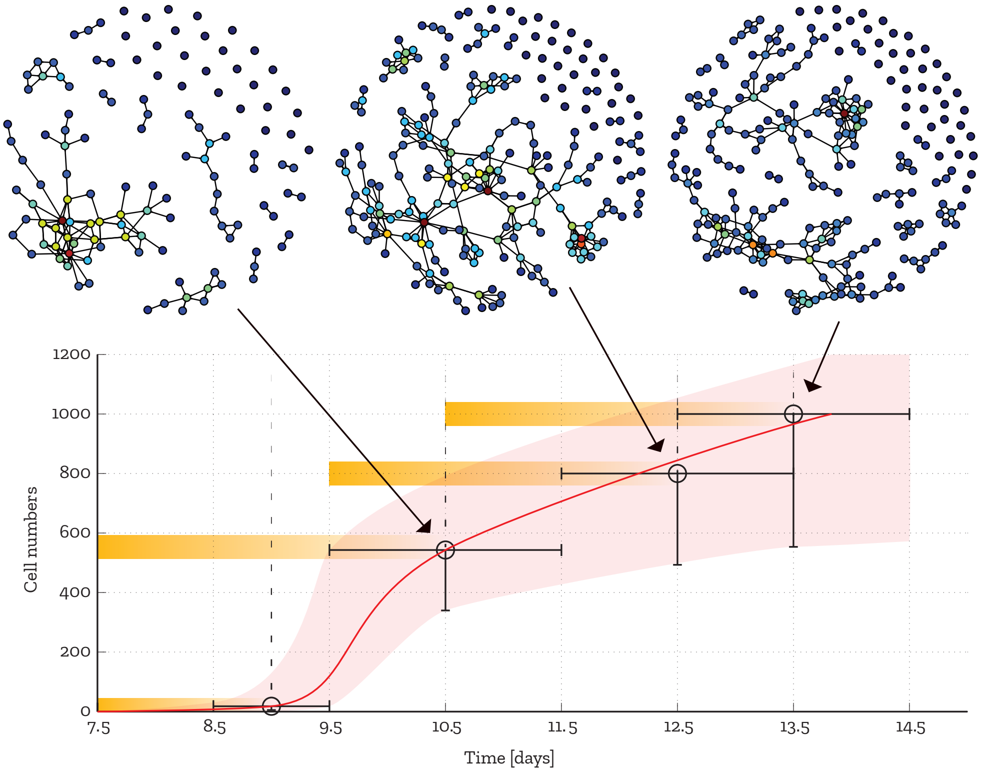


Figure 8.7: Qualitative description of embryonic development: Orange bands indicate Tamoxifen treatment. Lower bound is obtained from the barcode complexity. Data points are extrapolated cell numbers based on barcode complexity. On top are barcode networks from indicated time points showing increased connectivity. Each node is a unique rare barcode with $P_{\text{gen}} < 10^{-4}$

death, this estimated cell number still only provides a lower bound.

Tie2+ HSCp arise around E9.5

For different experiments we find different barcode complexities. In experiments where the mouse is labelled at E7.5 with 4-OH-Tam, only 0-5 barcodes with low read counts are found. In this case, 99.98% of read counts are from the unrecombined barcode. This finding suggests that at the end point of labelling at around ~E9 only 0-5 Tie2+ HSCp/HSC are found, and the majority of HSCp arise at a later time point.
By contrast labelling at E7.5 with tamoxifen, which is active for approximately 3 days, already produces ~350 unique barcodes, which corresponds to about 550 cells (fig. 8.6). Here, the unrecombined barcode only accounts for less than 5% of total read counts. Importantly, this strongly suggests that all Tie2+ HSCp/HSC

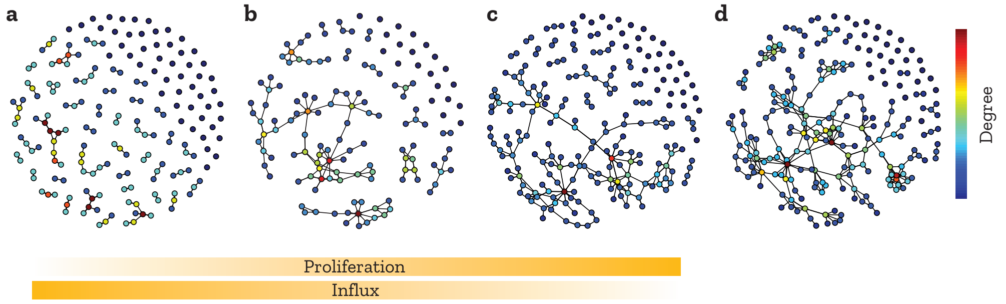


Figure 8.8: Effect of concurrent influx, proliferation and recombination: **a**, Barcode networks obtained by simulation of barcode creation with proliferation rate $p = 0.5\frac{1}{\text{day}}$, influx rate $i = 100\frac{1}{\text{day}}$. **b**, $p = 2.5\frac{1}{\text{day}}$; $i = 50\frac{1}{\text{day}}$. **c**, $p = 4.5\frac{1}{\text{day}}$; $i = 0\frac{1}{\text{day}}$.**d**, Experimental data from E9.5 labelling with Tamoxifen. Each node is a unique rare barcode with $P_{\text{gen}} < 10^{-4}$

arise between E9 and E10.5.

The delayed emergence of Tie2+ HSCp/HSC in the case of labelling at E7.5 is also reflected in the recombination distribution. Mice labelled at E7.5 with Tamoxifen show a mean minimal recombination number of $\mu_{\text{E7.5;TAM}} = 2.64 \pm 0.1$, where as mice labelled at E9.5 have a higher mean minimal recombination number of $\mu_{\text{E9.5;TAM}} = 2.85 \pm 0.01$. This suggest a shorter time interval of recombination for E7.5 labelling.

In line with this observation mice labelled at E9.5 and E10.5 show rising barcode complexity. At E12.5, the end point of E9.5 treatment, the barcode complexity reaches around 500 unique barcodes, indicating that around 800 cells have been labelled (fig. 8.6). One day later we find 550 unique barcodes, which puts the expected cell number for E13.5 at around 1000 cells (fig. 8.6). In both experimental setups the unrecombined barcode only accounts for less than 5% of total read counts, further supporting the notion that all Tie2+ HSCp/HSC arise early. The findings here are visualized in fig. 8.7.

Proliferative burst at E9.5

Combining the barcode complexity and the cell number estimation with the observed barcode networks, we also infer some information about the proliferative state of Tie2+ HSCp/HSC at various time points. At all three labelling time points (E7.5, E9.5, E10.5) a highly connected barcode set was retrieved. Since we found only a small number of barcodes at E9 and already 350 unique barcodes at E10.5, there has to be either a strong proliferative burst, a high influx of progenitor cells, or both, starting around E9.5.

In order to assess whether a high influx of progenitor cells, a high proliferation or a combination of both can explain the observed connectivity, we build a stochastic model. The model allows unlabelled progenitor cells to become Tie2+ with a constant influx rate i. Recombination occurs only in Tie2+ cells, which also proliferate with rate p. The results of the simulation with different proliferation and influx rates are shown in fig. 8.8 a-c. Since a high influx rate cannot explain the observed connectivity of barcodes in the experimental data (fig. 8.8 d), we exclude influx as a potential explanation.

As the proliferative burst occurs at around E9.5, labelling at E7.5 catches the burst only at the end of the recombination time interval. This leads to slightly less combined barcode sets (largest cluster: ~27% of rare barcodes) . Labelling at E9.5 coincides strongly with the burst, and we observe that almost all rare barcodes share a common cluster (largest cluster: ~70% of rare barcodes). After this initial burst, the proliferation slows down, as cell numbers rise slower at the later time points. This is also reflected again in smaller cluster sizes at E10.5 labelling (cluster shown in fig 8.7 top row).

The strong proliferative burst observed here is compatible with different studies reporting rapid expansion of HSCp in the mouse embryo during mid-gestation at around E9.5 [6, 7, 54, 55]. This time point coincides with the population of the fetal liver. Future *Polylox* experiments could help pin down the exact timing of proliferation during midgestation. Additional labelling time points with 4-OH-Tam will narrow down developmental steps much more closely, than what was possible during the writing of this thesis, as the main focus of this project was the understanding of the adult hematopoietic system.

8.4| Barcode clusters

In the previous section we have established that related barcodes are generated through simultaneous proliferation and recombination. Since the absence of proliferation leads to very few random connections, most observed barcode relationships are a direct measurement of intermediate recombination products. We are interested if related barcodes share a similar fate in adult hematopoiesis. If one could find such fate restriction, this would imply that fate and specialization are determined at a very early stage.

8.4.1| Fate of barcode clusters

Here we focus on finding barcode clusters using the adjacency matrix of barcode subsets via a Dulmage-Mendelsohn decomposition. We then check for output of these clusters generated in adult hematopoiesis in both the myelo-erythroid and common lymphoid branch.

Finding barcode clusters

To find barcode cluster in the measured subset M of barcodes, the adjacency matrix A' is calculated from the complete matrix A containing all codes. This is done by finding all rows and columns in A corresponding to barcodes in subset M (fig. 8.9 a and b).

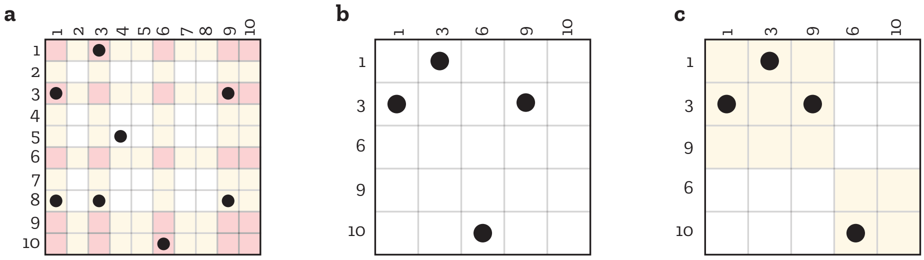


Figure 8.9: Barcode cluster using Dulmage-Mendelsohn decomposition: **a**, From the complete adjacency matrix A, a subset is taken (red squares) corresponding to all rows and columns in M, **b**, The reduced adjacency matrix A' is formed. **c**, A reordering of the entries is performed such that block entries are created. Those blocks correspond to the subgraphs.

In the next step a Dulmage-Mendelsohn decomposition is carried out [56]. This reorders the entries of A' in such a way that block matrices are formed. These unique block entries then correspond to the subgraphs of A' (fig. 8.9 b and c) and therefore barcode clusters.

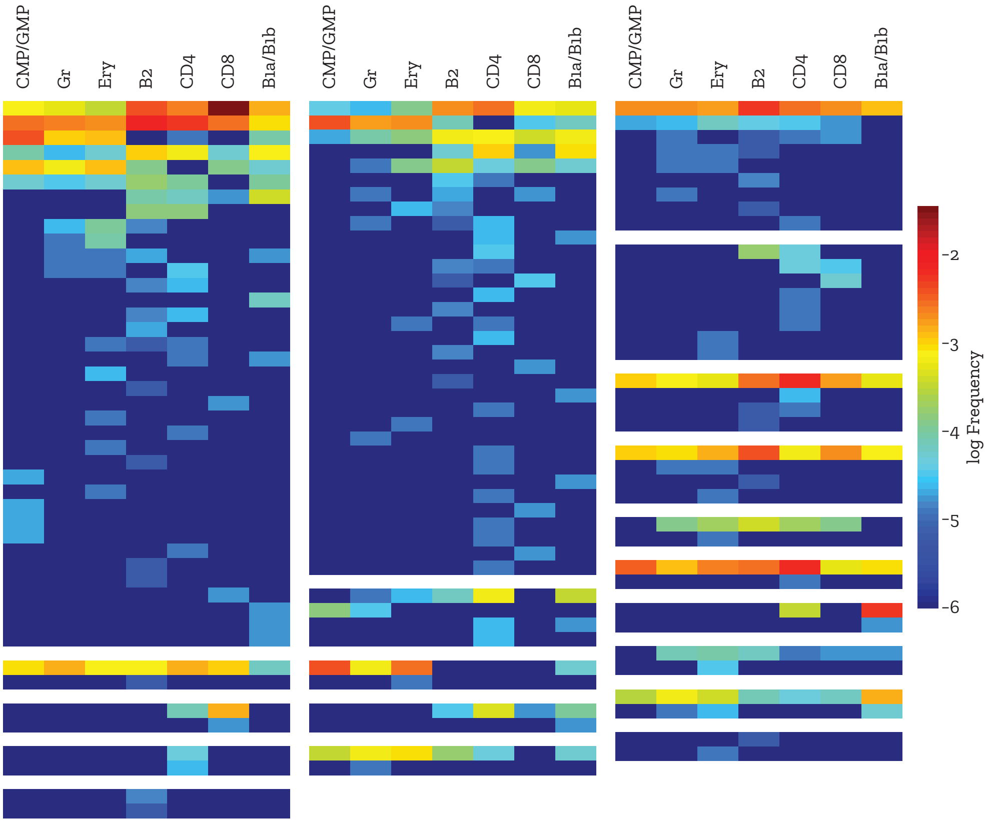


Figure 8.10: Output of barcode cluster: Blocks correspond to a single barcode cluster. Row inside the blocks show the output of a unique barcode. Columns correspond to the denoted populations on top. Log frequency of read counts is color coded.

No fate restriction evident

After finding the related barcode clusters, we now take a look at the output of these clusters in the adult hematopoiesis. If there is fate restriction already at the time point of labelling, barcodes in the same cluster should share a similar fate. Here we find only a few small clusters that only contribute to a particular branch or single lineage, which could be due to sampling. The vast majority of clusters however show no sign of similar lineage output (fig. 8.10). While one cannot conclude that fate restriction occurs in some cells, the majority of barcode clones in the same cluster do not share the same fate. It is therefore highly unlikely, that fate restriction occurs during the timeframe of labelling, if at all. In the next section we further elaborate this observation by reconstructing possible recombination pathways of rare barcodes with $P_{\text{gen}} < 10^{-4}$.

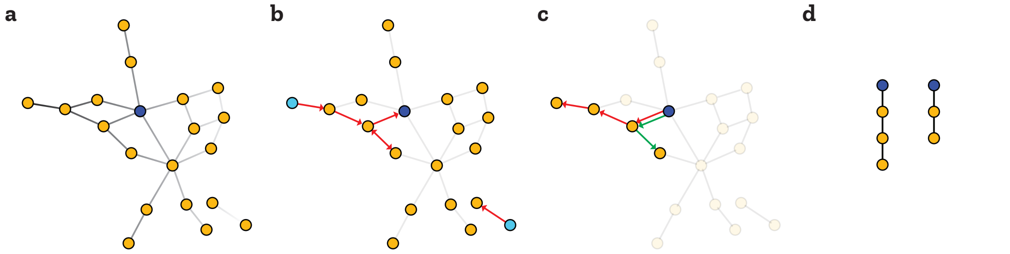


Figure 8.11: Finding possible recombination pathways: **a**, Connecting all experimentally found barcodes based on their relation. **b**, Starting from a single rare barcode (teal) a breadth-first search is done until the unrecombined barcode (dark blue) is found. If there is no available path to the unrecombined barcode, the starting barcode is discarded. **c**, Using only the connected barcodes from **b**, the search is repeated starting from the unrecombined barcode. To avoid looping due to inversions, only connections to barcodes with increasing number of minimal recombinations needed are taken into account. **d**, Multiple linear pathways are created and connected in trees. **b-d** Is repeated for every rare barcode. Orange nodes are unique barcodes, the blue node symbolizes the unrecombined barcode.

8.4.2| Fate evolution in recombination pathways

In this next step, we will find possible recombination pathways of rare barcodes that are still traceable in the data set and examine their fate output evolution during labelling.

Finding possible recombination pathways

In the previous sections of this chapter we have already established that the high proliferation during labelling allows intermediate barcodes to still be found after recombination stopped. Here we will use this property to build an algorithm to trace back recombination pathways. In order to find these pathways, a list of all rare barcodes is created. As in the previous chapters the cutoff has been set to $P_{\text{gen}} < 10^{-4}$. From each of those rare barcodes a backwards breadth-first search is done until the unrecombined barcode '123456789' is reached (fig. 8.11 a). In the breadth-first search, a list of all barcodes reachable within one step of recombination is created. This step is repeated recursively on each barcode in this, until a given depth has been reached. Since we know the minimal number of recombinations needed for each barcode, we use this as a depth cutoff for the search algorithm. Since we required to reconstruct the complete pathway, rare barcodes that cannot be connected to the unrecombined barcode are discarded, as at least one of their required intermediate barcodes is missing from the data (fig. 8.11 b). These first steps allow us not only to filter for barcodes with a complete pathway, but also discard barcodes that are not part of any given pathway. All barcodes in all possible pathways from barcode i to the unrecombined barcode form a subnetwork N_i (fig. 8.11 c).

Building trees from unrecombined barcodes

In the next step, only barcodes within a path to a given rare barcode i are taken into account (fig. 8.11 c). Again, the search is repeated on the subnetwork N_i, but with an additional restriction: the unrecombined barcode is the starting node for the search, the end node is the rare barcode i. Also connections with decreasing numbers of minimal recombinations are cut from the network. This allows us to avoid looping of the algorithm due to reversible inversions.

This search then leaves us with multiple linear pathways to barcode i. This is repeated for every subnetwork N_i found in the previous section. From these linear pathways trees are created by building the adjacency matrix and visualizing it. For mouse #2 we found 24 such trees which are shown in figure 8.12.

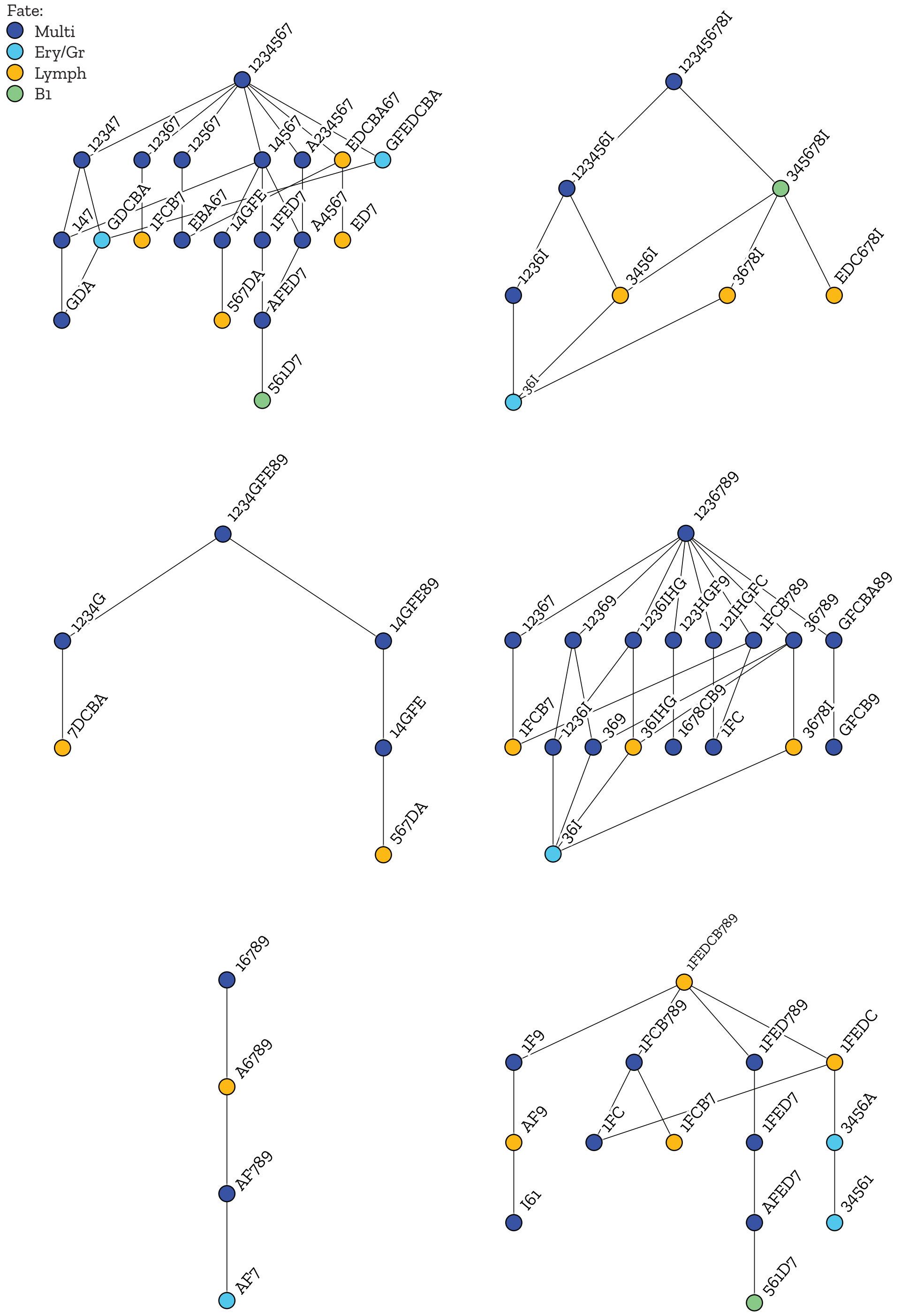
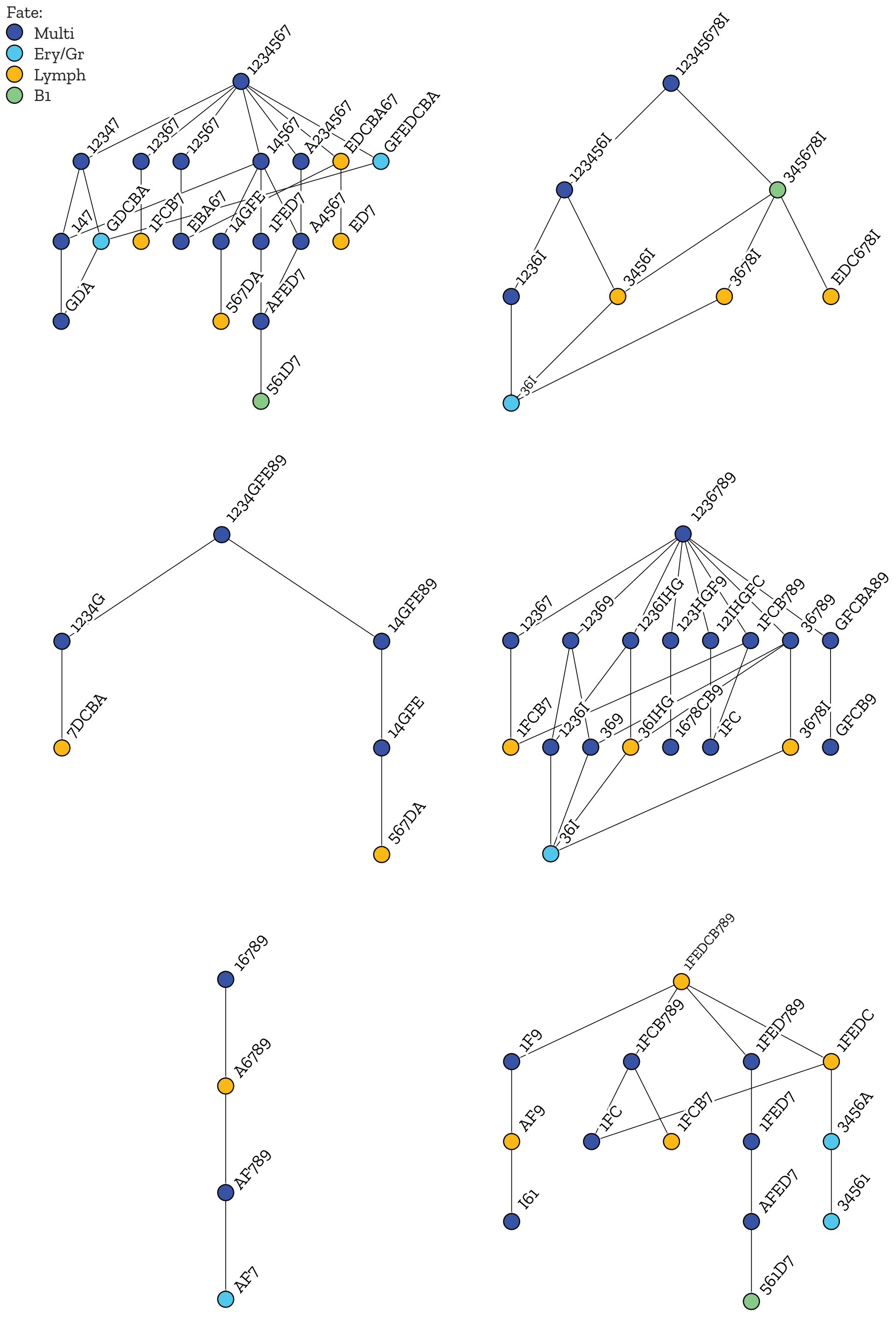
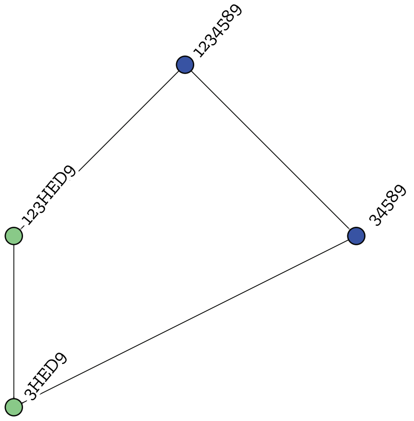
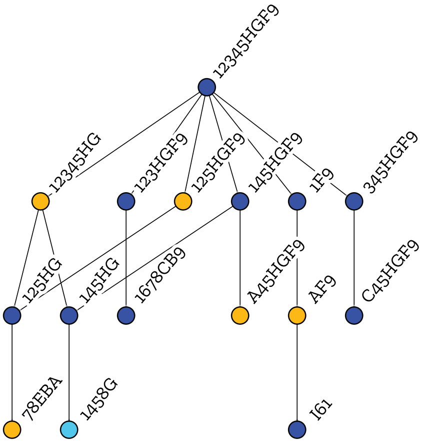
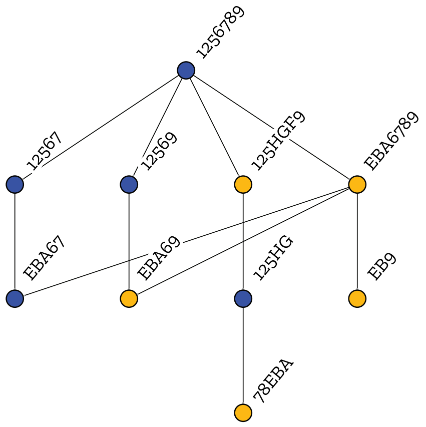


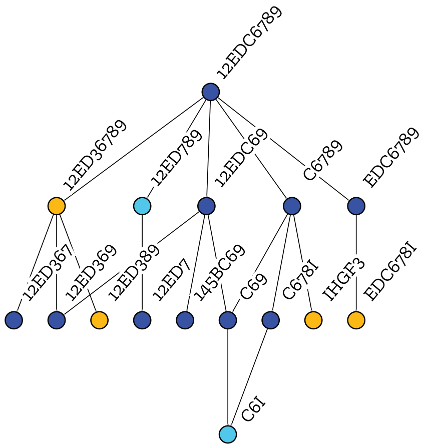
Figure 8.12: Recombination pathway trees: Each nodes denotes a barcode, color coded is the output into adult hematopoiesis. Top nodes of each tree are barcodes reachable within one recombination step, each consequent layer is a new recombination step.

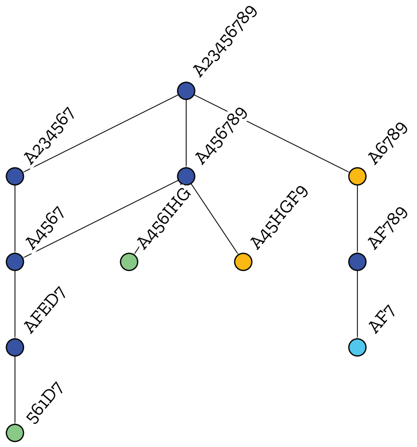

Fate:
● Multi
● Ery/Gr
● Lymph
● B1
1234567
12347
12367
12567
14567
A234567
EDCBA67
GFEDCBA
147
GDCBA
1FCB7
EBA67
14GFE
1FED7
A4567
ED7
GDA
567DA
AFED7
561D7
123456781
1234561
345678I
1236I
3456I
3678I
EDC678I
36I
1234GFE89
1234G
14GFE89
7DCBA
14GFE
567DA
1236789
12367
12369
1236IHG
123HGF9
12IHGFC
1FCB789
36789
GFCBA89
1FCB7
1236I
369
36IHG
1678CB9
1FC
3678I
GFCB9
36I
16789
A6789
AF789
AF7
1FEDCB789
1F9
1FCB789
1FED789
1FEDC
AF9
1FC
1FCB7
1FED7
3456A
161
AFED7
34561
561D7

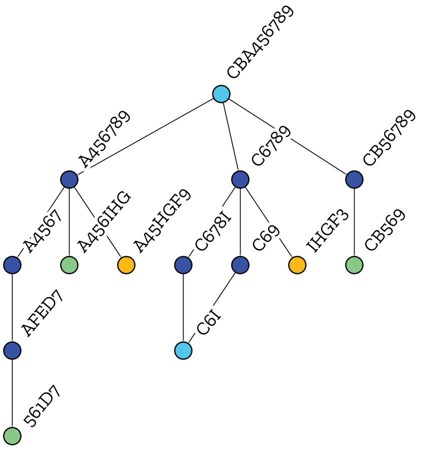

1234589
123HED9
34589
3HED9

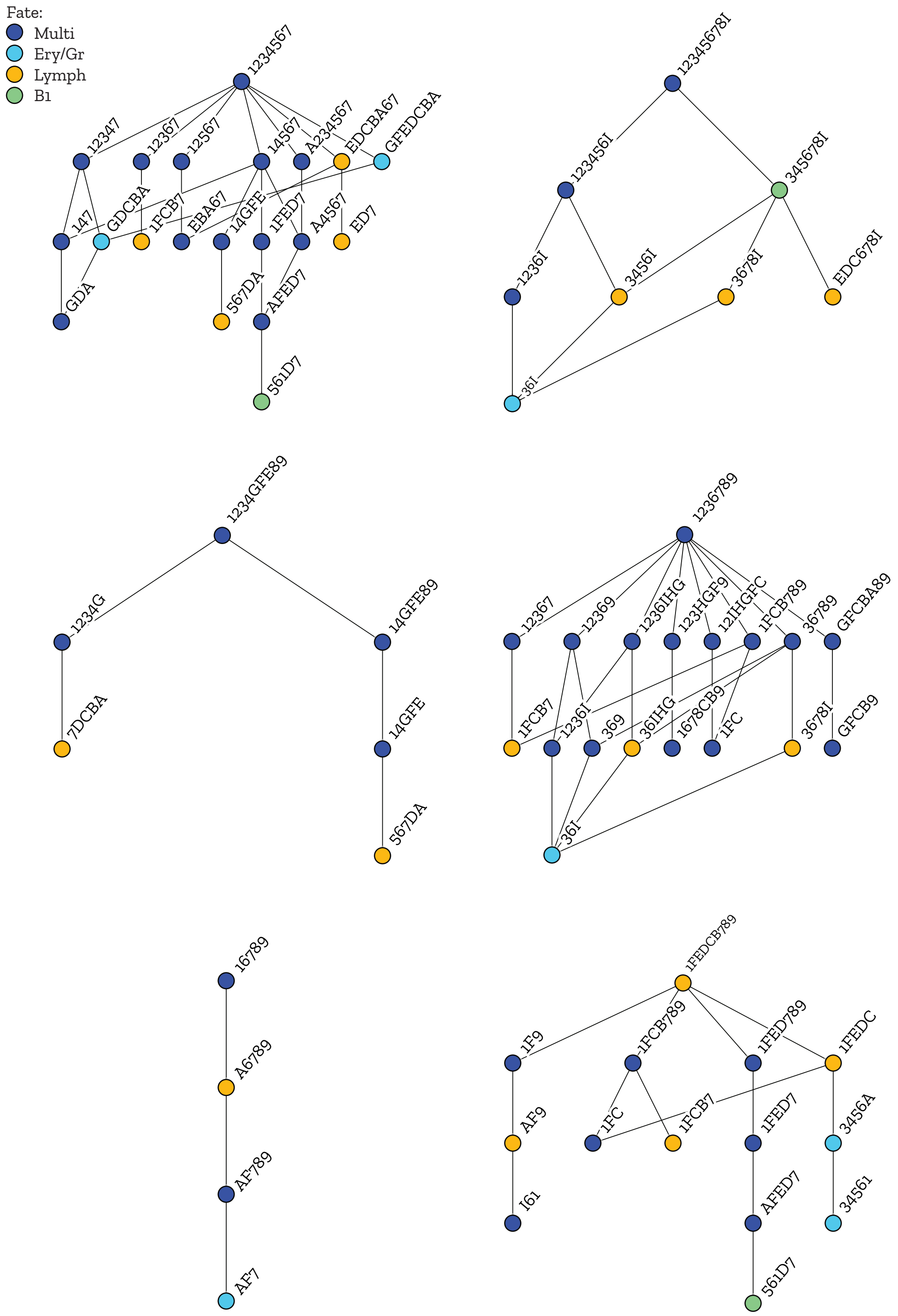

12345HGF9
12345HG
123HGF9
125HGF9
145HGF9
1F9
345HGF9
125HG
145HG
1678CB9
A45HGF9
AF9
C45HGF9
78EBA
1458G
161


1256789
12567
12569
125HGF9
EBA6789
EBA67
EBA69
125HG
EB9
78EBA


12EDC6789
12ED36789
12ED789
12EDC69
C6789
EDC6789
12ED367
12ED369
12ED389
12ED7
145BC69
C69
C678I
IHGF3
EDC678I
C6I


A23456789
A234567
A456789
A6789
A4567
A456IHG
A45HGF9
AF789
AFED7
AF7
561D7


CBA456789
A456789
C6789
CB56789
A4567
A456IHG
A45HGF9
C678I
C69
IHGF3
CB569
AFED7
C6I
561D7


Fate:
Multi
Ery/Gr
Lymph
B1
1234567
12347
12367
12567
14567
A234567
EDCBA67
GFEDCBA
147
GDCBA
1FCB7
EBA67
14GFE
1FED7
A4567
ED7
GDA
567DA
AFED7
561D7
123456781
1234561
345678I
1236I
3456I
3678I
EDC678I
36I
1234GFE89
1234G
14GFE89
7DCBA
14GFE
567DA
1236789
12367
12369
1236IHG
123HGF9
12IHGFC
1FCB789
36789
GFCBA89
1FCB7
1236I
369
36IHG
1678CB9
1FC
3678I
GFCB9
36I
16789
A6789
AF789
AF7
1FEDCB789
1F9
1FCB789
1FED789
1FEDC
AF9
1FC
1FCB7
1FED7
3456A
I61
AFED7
34561
561D7

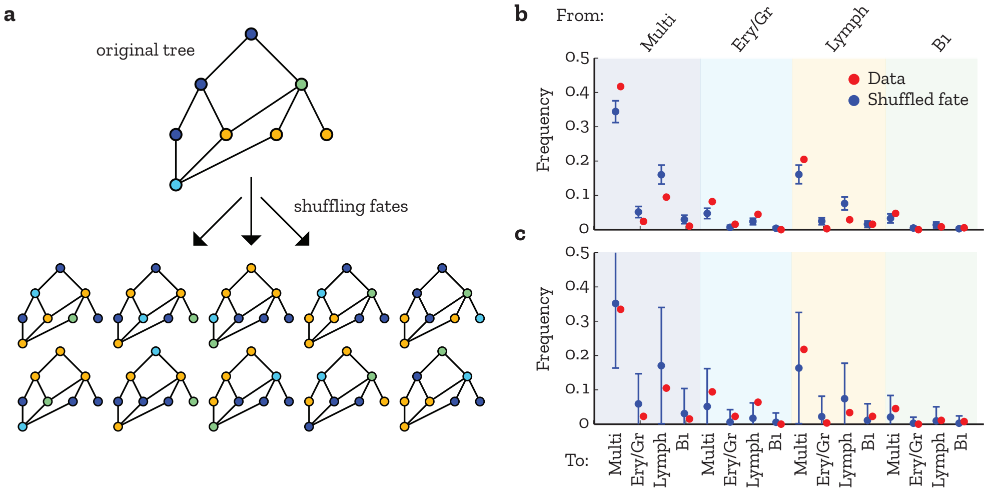


Figure 8.13: Fate restriction analysis: **a**, From the original tree, several iterations with shuffled fates are created. **b**, Fate transitions are counted and compared. Shuffled data is shown in blue, actual data in red. **c**, Same as above only taking rare barcodes into account.

No development of streamlined HSC

Since a connection of two individual barcodes hints at a possible relatedness of the cells, we are interested in the output in the adult hematopoiesis. If HSC subtypes with pre-determined fates are present at the embryonic state, more precisely at the time point of labelling, we should see a clear bias for cells with similar output to also have related barcodes (i.e. two connected barcodes are likely to come from the same progenitor cell, and should therefore have the same fate, if there are fate restricted subtypes of HSC). We therefore analyzed the fates of barcode i and barcode j, if i and j are connected (fig. 8.13 b). Next we repeated the same analysis, but with fates of each barcode shuffled randomly (fig. 8.13 a). Indeed there are significant differences between random fates and the measured data (fig. 8.13 b). However, the general trend of the data can be described.

As the top layers of each tree are most likely generated more than once, the nodes in those layers tend to show a multipotent output. This is expected to happen, with or without fate restriction. In addition, multiple possible pathways for each barcode can be found, as a result of the general connectedness of all barcodes. Both points create a bias in the observed fate changes. This bias leads to an underestimation of fate transitions towards multilineage barcodes and an overestimation of fate transitions coming from multilineage barcodes in the case of random shuffling.

To circumvent both issues, we restricted the shuffling procedure to rare barcodes and repeated the analysis. In this case, there is no evident deviation between randomly shuffled fate and the data (fig. 8.13 c). This finding suggests that there is no apparent hardwired fate in early HSC during midgestation. However, one cannot conclude that there is no fate restriction at all, as information of later time points of embryonic development or even adulthood cannot be conveyed in this kind of analysis.

9| Discussion

In this thesis, I have discussed a novel technique to mark single progenitor cells with a unique genetic barcode. *Polylox* barcoding allows us to achieve single cell resolution, while retaining advantages of bulk analysis like speed and high throughput. Single cell resolution makes it possible to gain valuable insight into the developmental relationships of cell types within the hematopoietic system. The main finding was that a dichotomy seems to exist between two major branches. In addition to the application of *Polylox* for fate mapping, I have studied different angles of this barcoding system, which provided new insights into clonal dynamics in the hematopoietic system of the embryo and adult mouse.

9.1| Results

Probability Model

One major advantage of the *Polylox* system is, that while it offers a high diversity with $n_b = 1,866,890$ possible barcodes, its simple rules of recombination make the creation process of barcodes predictable. This allowed us to create a complete library of possible barcodes, such that we can easily discriminate sequencing errors from real barcodes. Calculation of recombination events enabled the creation of an adjacency matrix containing the connections of all barcodes. Using this matrix, we can weigh connections between any two barcodes and build a simple but accurate Markov model of barcode creation. The calculation of the generation probability P_{gen} from this model is crucial in order to identify highly abundant versus rare barcodes, which is the only way to achieve single cell resolution. The model proposed in this thesis is very robust against variation of model parameters and is able to explain the barcode statistics found in several experiments. This probability model is the groundwork for all analyses that followed.

Theoretical Implications

In order to be able to interpret the measured barcode distributions, an understanding of barcode propagation is needed. In this thesis a moment based modeling approach was used to gain insight into barcode usage correlations. Here we found that the interpretation that first comes to mind is indeed correct. We suspected that populations that share not only the same barcodes but also show similar barcode frequencies are developmentally close. However, not only the topology of the system but also the dynamics play a major role in the observed correlation values. Directly related populations show a higher correlation if the differentiation from one population to the other is sufficiently fast. Early divergence of two populations result in a generally lower correlation.

Dichotomy in hematopoiesis

With the theoretical implications of barcode propagation in mind we analyzed the measured cell populations and calculated the pairwise correlations. The resulting correlation matrix was then used to cluster the populations hierarchically.

We found a major dichotomy splitting lineages into a myeloid-erythroid and a common lymphoid branch, while not ruling out additional routes. This finding is in line with the idea of a tree model proposed in multiple studies [23, 64]. Of note, B1 cell types did not share many barcodes with the other cell types, indicating a separate origin of these cells, as has been reported before in several experimental studies [36, 38--40].

Another major finding is the relationship of common myeloid progenitor cells (CMP) and fully matured cells. Where CMP shared a high number of barcodes with myeloid and erythroid cell types, the proportion of shared barcodes is much lower with lymphoid populations. Further, the measured correlations between myelo-erythroid populations and CMP where significantly higher than between CMP and lymphoid cells. Former correlations scattered around zero, indicating a distant relationship between CMP and the common lymphoid branch. The high correlation between CMP and populations from the myelo-erythroid branch places CMP as direct progenitors of granulocytes and erythrocytes. Refining progenitor-offspring relationships will require different Cre-drivers, active in progenitor stages such as CMP.

In the last years the proposed existence of such a dichotomy sparked much controversy in the field [12--14]. However, *Polylox* allows us to study progenitor-offspring connections very closely in great detail in an unperturbed system *in vivo*. These are important characteristic traits other experiments cannot yet provide.

The data presented here is supported by different studies, who agree on a major dichotomy in hematopoiesis [4, 23, 64, 65].

Fate restriction

One major question we addressed with *Polylox* next to developmental relationships is fate restriction. Namely we asked whether subpopulations within the HSC compartment exist, which produce output only in certain branches or populations.

While undersampling and different time scales of propagation for different cell populations make it difficult to answers this question definitely, we observed several single HSC with rare barcodes that showed a multilineage potential.

In addition to multilineage barcodes, several barcodes only had output in one of the two major branches. While we cannot rule out stochastic effects in barcode propagation as the cause of this observation, finding lineage restricted barcodes suggests that at least some fraction of HSC is limited in fate potential. The observed multipotency in *Polylox* data is supported, among others, by a study of Rodriguez-Fraticelli et al., who found evidence of multipotent as well as fate-restricted HSC with a different barcoding strategy [65, 66].

We used the high proliferation rate of embryonic HSC at around E9.5 to create barcode lineage trees, which allowed us to study the process of barcode creation in more detail. As concurrent fast proliferation and recombination leaves intermediate barcodes behind, we can reconstruct recombination pathways for barcodes. Comparing fates along recombination pathways under the assumption that all barcodes in these pathways are likely to belong to the same HSC clone, we found no evidence of fate restriction. Noteworthy, the deduction from pathway analysis is restricted to the timeframe of active recombination.

This observation does not contradict the existence of lineage-restricted fates, as this clone-intrinsic fate restriction may arise at a later time point than the time of labelling E9.5. A study from Yu et al. suggests that realized fates are controlled by DNA methylation, as they have observed color-coded HSC clones realizing stereotypic fate behaviour *in situ*, after transplantation and tissue injury [67].

Markov based modeling of hematopoiesis

With data from fate mapping experiments using a Tie2+YFP mouse model, Busch et al. inferred net proliferation and differentiation kinetics in the hematopoietic tree model [23]. Since in this thesis, barcode distributions are measured at steady state, it is not possible to infer the same kind of information from the data presented here, or to discriminate between different topological models. Instead, we asked whether the net proliferation and differentiation rates inferred by Busch et al. can explain the correlation structures we observed. To this end, we built a Markov model that allowed us to compute the first two moments of barcode distributions directly using a probability generating function. A significant advantage of this moment-based approach is that it can reduce the computational load significantly, compared to stochastic simulations of this system, in particular when cell number become very large. Furthermore, it allows us to refine the inferred parameters of Busch et al. inside their computed confidence intervals.

After refining the parameters for net proliferation and differentiation rates using the mean barcode clone size only, we found that the tree model (fig. 7.1) and parameters indeed predict the found correlation measures very well. This finding further supports the tree model of hematopoiesis [22, 23, 65].

The moment-based ODE model approach allows the calculation of Pearson's correlation coefficient as well as the moments of the distribution. To assess the actual barcode distributions we resorted to stochastic simulations instead. Again, using the refined parameters we found a good agreement of data and model.

HSC dynamics

Single cell analysis of *Polylox* directly allows us to study HSC clone size distributions. Due to the high workload for such experiments and the fact that only barcodes created once yield a direct insight into HSC clone size distributions, the data gained per experiment is limited (3-10 rare barcodes per experiment). By using information on the expected number of occurrence of all barcodes, we dissected common barcode clones into cell numbers. We analyzed data from several mice at different ages; we found that the clone size distribution broadens with age. Assuming net proliferation and differentiation rates estimated by Busch et al., a stochastic simulation of these distributions indicates that a stochasticity-induced neutral drift of clone sizes is sufficient to explain this effect [23]. The drift leads to a reduced number of clones making up more of the compartment [28--33].
While HSC dynamics in the adult murine hematopoiesis is characterized by slow proliferation and differentiation kinetics, the role of HSC during midgestation is still not fully understood. We argued in this thesis that the proliferative behaviour of labelled cells during recombination is preserved in the identity of the measured set of barcodes. By analyzing the connectivity of barcode sets and estimating cell numbers at different time points, we found evidence of a reported proliferative burst at around E9.5-E10.5, where HSCp/HSC proliferate roughly 3 times a day [6, 7, 54, 55].
From this burst we can estimate the cell numbers populating the fetal liver at E9.5 to around 100. Literature puts this number between 1 and 1000, which elucidates the lack of precise quantitative data in this area of research [7, 55, 68].
Polylox barcoding can help shed more light on the development of embryonic hematopoiesis. As of writing this thesis more experiments with 4-OH-TAM are planned and underway. These experiments offer a much higher temporal resolution which is needed to pin down time windows of proliferation and to enable mathematical modeling with increased precision.

9.2| *Polylox* barcoding and other fate mapping tools

With a plethora of fate mapping tools widely available, the question is raised why *Polylox* deserves a spot in the fate mapping toolkit. *Polylox* offers a number of advantages but also some disadvantages over traditional and novel techniques. These will be discussed in this section.

Transplantation experiments

In transplantation experiments, marked cells are transplanted in recipient organisms. The differentiated output of these donor cells can then be traced and analyzed. These experiments require some kind of perturbation of the studied system in order to be successful [2]. One major advantage of *Polylox* compared to transplantation is that it allows a vast number of possible genetic barcodes to be used *in vivo* in an unperturbed system.

Fluorescent markers

Fate mapping experiments using inducible fluorescent markers such as Brainbow or Tie2+YFP [23, 57, 58] enable the study of dynamics of a system even at steady state. However, they lack the high diversity of labels necessary for single cell resolution. While *Polylox* can in theory be used to study dynamic properties of a given system, the high complexity of barcodes makes it very difficult to compare barcode distributions at different timepoints. In addition to that, the experimental workload for *Polylox* is also much higher, making it generally unappealing for dynamical measurements.

Transposon integration sites

Another proposed barcoding technique is the method of hyperactive Sleeping Beauty (HSB) mediated transposon integration sites (TIS) [59]. In this technique HSB expression is controlled by Doxycycline and causes the transposon to be integrated at a random point in the genome. These integration sites offer a practically infinite number of markers, however the barcodes are induced ubiquitous and are not specific to a certain tissue or cell type, making it difficult to study progenitor-offspring relationships. In addition, studies have shown that depending on the affected region in the genome the cellular functions can be impaired [60, 61]. The transposon insertion may target regulatory genes potentially leading to cellular selection. Additionally, transpositions occurs preferentially in G1 phase and is known to slow down the cell cycle [60,61]. In contrast, the locus of *Polylox* is known, widely used and there is no known impact on cell physiology.

An additional advantage of *Polylox* compared to TIS is the absence of background noise (induced barcodes without active Cre or Doxycycline). In the case of *Polylox*

barcoding, the absence of noise is cell type independent (< 0.1% barcode induction in non targeted cells; fig. 4.2). For TIS the background noise can reach up to 4% and is depending on the cell type [59].
A comparison between TIS as a sample for other barcoding techniques and *Polylox* is shown in table 9.1 a.

	TIS [59]	*Polylox*
Number of markers	practically infinite	1,866,890
Specificity	ubiquitous	Cre-driven targeting
Marker neutrality	TIS might be neutral, but can cause gains or losses of cellular function depending on affected region [60, 61]	*Rosa26* reporter locus is widely used, with no known impact on physiology
Possible applications	avaibility of Doxycycline regulated systems limits application	different tissue specific Cre-driver are available
Background noise	up to 4% depending on cell type	< 0.1% cell type independent
Frequency of labelled cells	~30%	up to 99% of HSC progenitors
Measured diversity per mouse	~300	~550 of which roughly 230 are rare

Table 9.1: Comparison between transposon integration site [59] and *Polylox* This table has been adapted from [22]

CRISPR/Cas9 barcoding

New genome editing techniques like CRIPSR/Cas9 also bring new barcoding possibilities for developmental studies [62, 63]. One of these approaches is genome editing of synthetic target arrays for lineage tracing (GESTALT) [62]. Here a genetic barcode is mutated with random deletions and insertions over the development of the organism. Mutations shared in a large fraction of cells indicate an early event and a common progenitor. This enables the construction of huge lineage trees.
While this slow mutation over several proliferation and differentiation events offers time resolved data even with a single measured time point, this technique has some disadvantages over *Polylox* barcoding. First, there is the chance of already mutated sequences to be mutated again in a subsequent step, which results in the loss of progenitor information [62]. In contrast, the high stability of *Polylox* barcodes allows to trace back the output of single cells reliably.
The simple rules of *Polylox* barcoding makes the whole system calculable, and we

can distinguish common from rare barcodes. In GESTALT however, there have been reports of mutations biases towards certain locations in the genetic barcode [62]. The randomness of these events makes it hard to predict these biases. Common mutations may strongly influence the obtained lineage tree.

On the other hand, building lineage trees (as show in fig. 8.12) from *Polylox* is possible. However since the induction is limited to a short timeframe and the recombination events are not entirely random but have to satisfy certain rules, drawing clear conclusions from these lineages proves difficult. In contrast to the one time labelling and subsequent propagation of *Polylox* barcodes, GESTALT offers time resolved data in a single experiment as mutations are accumulated over the whole timeframe of the experiment.

General remarks

While the high complexity of *Polylox* barcodes affords single cell resolution, clone size information can be lost due to PCR and sequencing bias towards certain lengths or identities of barcodes. The lost information can be regained however with single cell sequencing, trading in the advantage of high throughput.

One strong advantage of *Polylox* is that it allows fate restriction as well as progenitor-offspring connections to be studied. Here the single cell resolution enables tracking of the output of a single HSC into mature cell populations. One limitation is that undersampling and different time scales of propagation through the system can hinder a clear statement regarding fate restriction, while multipotency can be assessed easily.

For all techniques that can achieve single cell resolution by barcoding, undersampling can be a limitation, because it restricts assessment of fate restriction and progenitor-offspring connections.

Nonetheless, *Polylox* provides a highly complex, yet calculable barcoding system. Its major strength is single cell resolution in an unperturbed system *in vivo* that is not only restricted to hematopoiesis, but can be used in a multitude of systems. The main limitation is

9.3| Outlook

The *Polylox* system is being used in an increasing number of groups world-wide. In the following we highlight two ongoing research projects that illustrate the great potential of *Polylox* barcoding. In addition to this two highlighted projects there are currently efforts of understanding the early stages of hematopoiesis using 4-OH-Tam. The lower time window of recombination in comparison to regular tamoxifen treatment allows a higher temporal resolution. Yoshida et al. are, as of writing of this thesis, employing *Polylox* as a fate mapping technique to understand spermatogenesis [69].

Macrophage origin

While most hematopoeitic populations are dependent on influx from upstream compartments and ultimately HSC, macrophages can self-maintain without HSC [5]. The origin and possible contributions of HSC in these macrophage cell populations remains unclear. Previous studies have shown that tissue macrophages in the brain (microglia/MG), liver (Kupffer Cells/KC) as well as in the lung (alveolar macrophages/AM) are created by erythro-myeloid progenitor cells (EMP) [5] during midgestation.

By labelling Tie2+ EMP at different time points, preliminary *Polylox* experiments suggest that differentiation occurs over a few days between E7.5 and E10.5 in a wave-like manner. *Polylox* barcoding allowed the tracing of individual clones, making it possible to observe this wave-like differentiation into the different compartments for the first time.

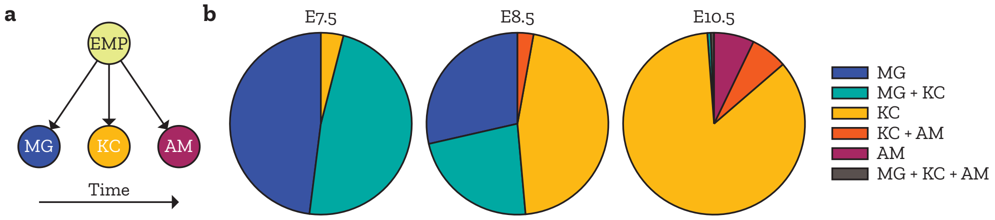


Figure 9.1: Preliminary macrophage data: **a**, Proposed model of differentiation. As the different organs are developed, differentiation from EMP into the respective tissues is favored. **b**, Time resolved output of Tie2+ EMP into tissue-resident macrophages using *Polylox* barcodes. Shown is the fraction of unique rare barcodes with indicated output.

PolyExpress

Another research topic revolving around *Polylox* barcoding is *PolyExpress*. Here the goal is to complement *Polylox* barcode information with single cell expression data. The aim of this project is to express the *Polylox* Barcode ubiquitously as RNA barcode. Subsequent single cell RNA sequencing then allows to combine barcode fate mapping with single cell expression. While there are some challenges to overcome, this project has the potential to resolve the long standing question if and at which developmental time point fate restriction occurs, and what role expression and certain regulatory genes play in this process.

References

[1] J. E. Till and E. A. McCulloch. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiation Research*, 14(2):213--222, 1961.

[2] K. Busch and H.-R. Rodewald. Unperturbed vs. post-transplantation hematopoiesis: both in vivo but different. *Current opinion in hematology*, 23(4):295--303, 07 2016.

[3] T. Höfer, K. Busch, K. Klapproth, and Rodewald H.-R. Fate Mapping and Quantitation of Hematopoiesis In Vivo. *Annual Review of Immunology*, 34(1):449--478, 2016.

[4] T. Höfer and H.-R. Rodewald. Differentiation-based model of hematopoietic stem cell functions and lineage pathways. *Blood*, 01 2018.

[5] Elisa Gomez Perdiguero, Kay Klapproth, Christian Schulz, Katrin Busch, Emanuele Azzoni, Lucile Crozet, Hannah Garner, Celine Trouillet, Marella F. de Bruijn, Frederic Geissmann, and Hans-Reimer Rodewald. Tissue-resident macrophages originate from yolk-sac-derived erythro-myeloid progenitors. *Nature*, 518:547 EP --, 12 2014.

[6] S. Taoudi, C. Gonneau, K. Moore, J. M. Sheridan, C. C. Blackburn, E. Taylor, and A. Medvinsky. Extensive hematopoietic stem cell generation in the agm region via maturation of ve-cadherin+cd45+ pre-definitive hscs. *Cell Stem Cell*, 3(1):99 -- 108, 2008.

[7] A. Batsivari, S. Rybtsov, C. Souilhol, A. Binagui-Casas, D. Hills, S. Zhao, P. Travers, and A. Medvinsky. Understanding hematopoietic stem cell development through functional correlation of their proliferative status with the intra-aortic cluster architecture. *Stem Cell Reports*, 8(6):1549 -- 1562, 2017.

[8] Motonari Kondo, Irving L. Weissman, and Koichi Akashi. Identification of clonogenic common lymphoid progenitors in mouse bone marrow. *Cell*, 91(5):661--672, 1997.

[9] Koichi Akashi, David Traver, Toshihiro Miyamoto, and Irving L. Weissman. A clonogenic common myeloid progenitor that gives rise to all myeloid lineages. *Nature*, 404:193 EP --, 03 2000.

[10] G. Terszowski, C. Waskow, P. Conradt, D. Lenze, J. Koenigsmann, D. Carstanjen, I. Horak, and H.-R. Rodewald. Prospective isolation and global gene expression analysis of the erythrocyte colony-forming unit (cfu-e). *Blood*, 105(5):1937, 03 2005.

[11] C. J. H. Pronk, D. J. Rossi, R. Månsson, J. L. Attema, G. L. Norddahl, C. K. F. Chan, M. Sigvardsson, I. L. Weissman, and D. Bryder. Elucidation of the phenotypic, functional, and molecular topography of a myeloerythroid progenitor cell hierarchy. *Cell Stem Cell*, 1(4):428--442, 2007.

[12] F. Notta, S. Zandi, N. Takayama, S. Dobson, Olga I. Gan, G. Wilson, K. B. Kaufmann, J. McLeod, E. Laurenti, C. F. Dunant, J. D. McPherson, L. D. Stein, Y. Dror, and J. E. Dick. Distinct routes of lineage development reshape the human blood hierarchy across ontogeny. *Science*, 351(6269), 01 2016.

[13] H. Kawamoto, T. Ikawa, K Masuda, H Wada, and Y. Katsura. A map for lineage restriction of progenitors during hematopoiesis: the essence of the myeloid-based model. *Immunological Reviews*, 238(1):23--36, 2018/11/08 2010.

[14] L. Velten, S. F. Haas, S. Raffel, S. Blaszkiewicz, S. Islam, B. P. Hennig, C. Hirche, C. Lutz, E. C. Buss, D. Nowak, T. Boch, W.-K. Hofmann, A. D. Ho, W. Huber, A. Trumpp, M. A. G. Essers, and L. M. Steinmetz. Human haematopoietic stem cell lineage commitment is a continuous process. *Nature Cell Biology*, 19:271–281, 03 2017.

[15] F. Paul, Y. Arkin, A. Giladi, D. Ad. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, E. David, N. Cohen, F. K. B. Lauridsen, S. Haas, A. Schlitzer, A. Mildner, F. Ginhoux, S. Jung, A. Trumpp, B. T. Porse, A. Tanay, and I. Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663--1677, 2015.

[16] L. Perié, K. R. Duffy, L. Kok, R. J. de Boer, and T. N. Schumacher. The branching point in erythro-myeloid differentiation. *Cell*, 163(7):1655--1662, 2015.

[17] S. Huang, Y.-P. Guo, G. May, and T. Enver. Bifurcation dynamics in lineage-commitment in bipotent progenitor cells. *Developmental Biology*, 305(2):695--713, 2007.

[18] B. Sauer. Functional expression of the cre-lox site-specific recombination system in the yeast saccharomyces cerevisiae. *Molecular and Cellular Biology*, 7(6):2087--2096, 1987.

[19] B. Sauer and N. Henderson. Site-specific dna recombination in mammalian cells by the cre recombinase of bacteriophage p1. *Proceedings of the National Academy of Sciences of the United States of America*, 85(14):5166--5170, 07 1988.

[20] K Rajewsky, H Gu, R Kühn, U A Betz, W Müller, J Roes, and F Schwenk. Conditional gene targeting. *The Journal of Clinical Investigation*, 98(3):600--603, 8 1996.

[21] N. Sternberg and D. Hamilton. Bacteriophage p1 site-specific recombination: I. recombination between loxp sites. *Journal of Molecular Biology*, 150(4):467 -- 486, 1981.

[22] W. Pei, T. B. Feyerabend, J. Rößler, X. Wang, D. Postrach, K. Busch, I. Rode, K. Klapproth, N. Dietlein, C. Quedenau, W. Chen, S. Sauer, S. Wolf, T. Höfer, and H.-R. Rodewald. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. *Nature*, 548(7668):456--460, Aug 2017.

[23] K. Busch, K. Klapproth, M. Barile, M. Flossdorf, T. Holland-Letz, S. M. Schlenner, M. Reth, T. Höfer, and H.-R. Rodewald. Fundamental properties of unperturbed haematopoiesis from stem cells in vivo. *Nature*, 518(7540):542--546, Feb 2015.

[24] L. Ringrose, S. Chabanis, P. O. Angrand, C. Woodroofe, and A. F. Stewart. Quantitative comparison of DNA looping in vitro and in vivo: chromatin increases effective DNA flexibility at short distances. *EMBO J.*, 18(23):6630--6641, Dec 1999.

[25] W. Pei, T. B. Feyerabend, J. Rößler, X. Wang, D. Postrach, T. Höfer, and H.-R. Rodewald. Protocol for the use of polylox -endogenous barcoding for high resolution in vivo lineage tracing. *Protocol Exchange*, 09 2017.

[26] W. Pei, X. Wang, J. Rössler, T. B. Feyerabend, T Höfer, and Rodewald. H.-R. Cre-recombinase-driven polylox barcoding in mice - under revision. *Nature Protocols*, 2019.

[27] D. T. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[28] A. Rundberg Nilsson, S. Soneji, S. Adolfsson, D. Bryder, and C. J. Pronk. Human and murine hematopoietic stem cell aging is associated with functional impairments and intrinsic megakaryocytic/erythroid bias. *PLoS ONE*, 11(7):e0158369, 2016.

[29] M. Kim, H.‐B. Moon, and G. J. SPANGRUDE. Major age‐related changes of mouse hematopoietic stem/progenitor cells. *Annals of the New York Academy of Sciences*, 996(1):195--208.

[30] S. M. Chambers, C. A. Shaw, C. Gatza, C. J. Fisk, L. A. Donehower, and M. A. Goodell. Aging hematopoietic stem cells decline in function and exhibit epigenetic dysregulation. *PLoS Biology*, 5(8):e201, 08 2007.

[31] D. J Rossi, D. Bryder, J. M. Zahn, H. Ahlenius, R. Sonu, A. J. Wagers, and I. L. Weissman. Cell intrinsic alterations underlie hematopoietic stem cell aging. *Proceedings of the National Academy of Sciences of the United States of America*, 102(26):9194--9199, 06 2005.

[32] S. J. Morrison, A. M. Wandycz, K. Akashi, A. Globerson, and I. L. Weissman. The aging of hematopoietic stem cells. *Nature Medicine*, 2:1011 -- 1016, 09 1996.

[33] A. Nakamura-Ishizu, H. Takizawa, and T. Suda. The analysis, roles and regulation of quiescence in hematopoietic stem cells. *Development*, 141(24):4656--4666, 2014.

[34] S. H Cheshier, S. J. Morrison, X. Liao, and I. L. Weissman. In vivo proliferation and cell cycle kinetics of long-term self-renewing hematopoietic stem cells. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):3120--3125, 03 1999.

[35] C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1):72--101, Jan 1904.

[36] E. Montecino-Rodriguez and K. Dorshkind. B-1 b cell development in the fetus and adult. *Immunity*, 36(1):13--21, 01 2012.

[37] A. B. Kantor and L. A. Herzenberg. Origin of murine b cell lineages. *Annual Review of Immunology*, 11(1):501--538, 1993. PMID: 8476571.

[38] A. M. Stall, S. Adams, L. A. A. Herzenberg, and A. B. Kantor. Characteristics and development of the murine b-lb (ly-1 b sister) cell population. *Annals of the New York Academy of Sciences*, 651(1):33--43, 1992.

[39] K Hayakawa, R. R. Hardy, and L. A. Herzenberg. Progenitors for ly-1 b cells are distinct from progenitors for other b cells. *Journal of Experimental Medicine*, 161(6):1554--1568, 1985.

[40] B. de Andrés, P. Gonzalo, S. Minguet, J. A. Martınez-Marın, P. G. Soro, M. A. R. Marcos, and M. L. Gaspar. The first 3 days of b-cell development in the mouse embryo. *Blood*, 100(12):4074--4081, 2002.

[41] R. R. Hardy and K. Hayakawa. A developmental switch in b lymphopoiesis. *Proceedings of the National Academy of Sciences*, 88(24):11550--11554, 1991.

[42] H. Oguro, L. Ding, and S. J. Morrison. Slam family markers resolve functionally distinct subpopulations of hematopoietic stem cells and multipotent progenitors. *Cell Stem Cell*, 13(1):102 -- 116, 2013.

[43] T. K. Starr, S. C. Jameson, and K. A. Hogquist. Positive and negative selection of t cells. *Annual Review of Immunology*, 21(1):139--176, 2003. PMID: 12414722.

[44] C. Gardiner. Stochastic Methods: A Handbook for the Natural and Social Sciences (Springer Series in Synergetics). *Springer*, 4th edition, 2009.

[45] K. R. Duffy, C. J. Wellard, J. F. Markham, J. H. S. Zhou, R. Holmberg, E. D. Hawkins, J. Hasbold, M. R. Dowling, and P. D. Hodgkin. Activation-induced b cell fates are selected by intracellular stochastic competition. *Science*, 335(6066):338--341, 2012.

[46] R.L. Iman, J.M. Davenport, and D.K. Zeigler. Latin hypercube sampling (program user's guide). 1980.

[47] J. Seita and I. L. Weissman. Hematopoietic stem cell: Self-renewal versus differentiation. *Wiley interdisciplinary reviews. Systems biology and medicine*, 2(6):640--653, Nov-Dec 2010.

[48] V. R. Buchholz, M. Flossdorf, I. Hensel, L. Kretschmer, B. Weissbrich, P. Gräf, A. Verschoor, M. Schiemann, T. Höfer, and D. H. Busch. Disparate individual fates compose robust cd8+ t cell immunity. *Science*, 340(6132):630--635, 2013.

[49] C. Frost and S. G. Thompson. Correcting for Regression Dilution Bias: Comparison of Methods for a Single Predictor Variable. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 163(2):173--189, 2000.

[50] J. I. Macgregor and V. C. Jordan. Basic guide to the mechanisms of antiestrogen action. *Pharmacological Reviews*, 50(2):151--196, 1998.

[51] D. J. Venzon and S. H. Moolgavar. A method for computing profile-likelihood-based confidence intervals. *Applied Statistics*, 37(1):225--234, 1988.

[52] Q. Zhong, C. Zhang, Q. Zhang, L. Miele, S. Zheng, and G. Wang. Boronic prodrug of 4-hydroxytamoxifen is more efficacious than tamoxifen with enhanced bioavailability independent of cyp2d6 status. *BMC Cancer*, 15(1):625, Sep 2015.

[53] M. Valny, P. Honsa, D. Kirdajova, Z. Kamenik, and M. Anderova. Tamoxifen in the mouse brain: Implications for fate-mapping studies using the tamoxifen-inducible cre-loxp system. *Frontiers in Cellular Neuroscience*, 10:243, 2016.

[54] H. Ema and H. Nakauchi. Expansion of hematopoietic stem cells in the developing liver of a mouse embryo. *Blood*, 95(7):2284--2288, 2000.

[55] Stanislav Rybtsov, Andrejs Ivanovs, Suling Zhao, and Alexander Medvinsky. Concealed expansion of immature precursors underpins acute burst of adult hsc activity in foetal liver. *Development*, 143(8):1284--1289, 2016.

[56] A. L. Dulmage and N. S. Mendelsohn. Coverings of bipartite graphs . *Canad. J. Math.*, (10):517--534, 1958.

[57] J. Livet, T. A. Weissman, H. Kang, R. W. Draft, J. Lu, R. A. Bennis, J. R. Sanes, and J. W. Lichtman. Transgenic strategies for combinatorial expression of fluorescent proteins in the nervous system. *Nature*, 450:56–62, 11 2007.

[58] T. A Weissman and Y. A. Pan. Brainbow: New resources and emerging biological applications for multicolor genetic labeling and analysis. *Genetics*, 199(2):293--306, 02 2015.

[59] J. Sun, A. Ramos, B. Chapman, J. B. Johnnidis, . Le, Y.-J. Ho, A. Klein, O. Hofmann, and F. D. Camargo. Clonal dynamics of native haematopoiesis. *Nature*, 514:322–327, 10 2014.

[60] O. Walisko, Z. Izsvák, K. Szabó, Christopher D. Kaufman, S. Herold, and Z. Ivics. Sleeping beauty transposase modulates cell-cycle progression through interaction with miz-1. *Proceedings of the National Academy of Sciences of the United States of America*, 103(11):4062, 03 2006.

[61] Neal G. Copeland and Nancy A. Jenkins. Harnessing transposons for cancer gene discovery. *Nature Reviews Cancer*, 10:696 EP --, 09 2010.

[62] A. McKenna, G. M. Findlay, J. A. Gagnon, M. S. Horwitz, A. F. Schier, and J. Shendure. Whole organism lineage tracing by combinatorial and cumulative genome editing. *Science*, 2016.

[63] R. Kalhor, K. Kalhor, L. Mejia, K. Leeper, A. Graveline, P. Mali, and G. M. Church. Developmental barcoding of whole mouse via homing crispr. *Science*, 08 2018.

[64] S. M. Schlenner, V. Madan, K. Busch, A. Tietz, C. Läufle, C. Costa, C. Blum, H. J. Fehling, and H.-R. Rodewald. Fate mapping reveals separate origins of t cells and myeloid lineages in the thymus. *Immunity*, 32(3):426--436, 2010.

[65] A. E. Rodriguez-Fraticelli, Samuel L. Wolock, C. S. Weinreb, R. Panero, S. H. Patel, M. Jankovic, J. Sun, R. A. Calogero, A. M. Klein, and F. D. Camargo. Clonal analysis of lineage fate in native haematopoiesis. *Nature*, 553:212–216, 01 2018.

[66] C. M. Sawai, S. Babovic, S. Upadhaya, D. J. H. F. Knapp, Y. Lavin, C. M. Lau, A. Goloborodko, J. Feng, J. Fujisaki, L. Ding, L. A. Mirny, M. Merad, C. J. Eaves, and B. Reizis. Hematopoietic stem cells are the major source of multilineage hematopoiesis in adult animals. *Immunity*, 45(3):597--609, 2016.

[67] V. W. C. Yu, R. Z. Yusuf, T. Oki, J. Wu, B. Saez, X. Wang, C. Cook, N. Baryawno, M. J. Ziller, E. Lee, H. Gu, A. Meissner, C. P. Lin, P. V. Kharchenko, and D. T. Scadden. Epigenetic memory underlies cell-autonomous heterogeneous behavior of hematopoietic stem cells. *Cell*, 168(5):944--945, 2017.

[68] M. Ganuza, T. Hall, D. Finkelstein, A. Chabot, G. Kang, and S. McKinney-Freeman. Lifelong haematopoiesis is established by hundreds of precursors throughout mammalian ontogeny. *Nature Cell Biology*, 19:1153 EP --, 09 2017.

[69] S. Yoshida. Elucidating the identity and behavior of spermatogenic stem cells in the mouse testis. *Reproduction*, 144(3):293--302, Sep 2012.

[70] D. Karamitros, B. Stoilova, Z. Aboukhalil, F. Hamey, A. Reinisch, M. Samitsch, L. Quek, G. Otto, . Repapi, J. Doondeea, B. Usukhbayar, J. Calvo, S. Taylor, N. Goardon, E. Six, F. Pflumio, C. Porcher, R. Majeti, B. Göttgens, and P. Vyas. Single-cell analysis reveals the continuum of human lympho-myeloid progenitor cells. *Nature Immunology*, 19(1):85--97, 2018.

[71] A.-L. Barabási. Network Science. *Northeastern University, Boston*, 2016.

[72] S. P. Robinson, S. M. Langan-Fahey, D. A. Johnson, and C. Jordan. Metabolites, Pharmacodynamics, and Pharmacokinetics of Tamoxifen in Rats and Mice Compared to the Breast Cancer Patients. *Drug Metabolism and Disposition*, (19):36–43, 1991.

[73] E. R. Kisanga, J. Gjerde, G. Mellgren, and E. A. Lien. Tamoxifen administration and metabolism in nude mice and nude rats. *The Journal of Steroid Biochemistry and Molecular Biology*, (84):361–367, 2003.

[74] M.L. Kauts, C. S. Vink, and E. Dzierzak. Hematopoietic (stem) cell development - how divergent are the roads taken? *Febs Letters*, 590(22):3975--3986, 2016.

[75] A. Cumano and I. Godin. Ontogeny of the hematopoietic system. *Annual Review of Immunology*, 25(1):745--785, 2007.

[76] R. Lanza, H. Blau, J. Gearhart, B. Hogan, D. Melton, M. Moore, R. Pedersen, E.D. Thomas, J.A. Thomson, C. Verfaillie, I. Weissman, and M.D. West. Handbook of stem cells. *Elsevier Inc.*, 2004.

[77] P. Kumaravelu, L. Hook, A. M. Morrison, J. Ure, S. Zhao, S. Zuyev, J. Ansell, and A. Medvinsky. Quantitative developmental anatomy of definitive haematopoietic stem cells/long-term repopulating units (hsc/rus): role of the aorta-gonad-mesonephros (agm) region and the yolk sac in colonisation of the mouse embryonic liver. *Development*, 129(21):4891--4899, 2002.

[78] H Huang and R Auerbach. Identification and characterization of hematopoietic stem cells from the yolk sac of the early mouse embryo. *PNAS*, 90(21):10110--10114, 1993.

[79] M.J. Sánchez, A. Holmes, C. Miles, and E. Dzierzak. Characterization of the first definitive hematopoietic stem cells in the agm and liver of the mouse embryo. *Immunity*, 5(6):513--525, 1996.

[80] I. Godin, J.. Garcia-Porrero, F. Dieterlen-Lièvre, and A. Cumano. Stem cell emergence and hemopoietic activity are incompatible in mouse intraembryonic sites. *Journal of Experimental Medicine*, 190(1):43--52, 1999.

[81] R. Lu, N. F. Neff, S. R. Quake, and I. L. Weissman. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. *Nature biotechnology*, 29:928--933, 10 2011.

[82] S. P. Robinson, S. M. Langan-Fahey, D. A. Johnson, and V. C. Jordan. Metabolites, pharmacodynamics, and pharmacokinetics of tamoxifen in rats and mice compared to the breast cancer patient. *Drug Metabolism and Disposition*, 19(1):36--43, 1991.

[83] H. K. A. Mikkola and S. H. Orkin. The journey of developing hematopoietic stem cells. *Development*, 133(19):3733--3744, 2006.

[84] S. Feil, N. Valtcheva, Feil R., W. Wurst, and Kühn R. Inducible cre mice. in: Gene knockout protocols. methods in molecular biology (methods and protocols). *Humana Press*, 530, 2009.

[85] T. S. Weber, M. Dukes, D. C. Miles, S. P. Glaser, S. H. Naik, and K. R. Duffy. Site-specific recombinatorics: in situ cellular barcoding with the cre lox system. *BMC Systems Biology*, 10:43, 2016.

[86] I. D. Peikon, D. I. Gizatullina, and A. M. Zador. In vivo generation of dna sequence diversity for cellular barcoding. *Nucleic Acids Research*, 42(16):e127--e127, 09 2014.

[87] J. Carrelha, Y. Meng, L. M. Kettyle, T. C. Luis, R. Norfo, V. Alcolea, H. Boukarabila, F. Grasso, A. Gambardella, A. Grover, K. Högstrand, Allegra M. Lord, A. Sanjuan-Pla, P. S. Woll, C. Nerlov, and S. E. W. Jacobsen. Hierarchically related lineage-restricted fates of multipotent haematopoietic stem cells. *Nature*, 554:106–111, 01 2018.

[88] C. M. Arends, J. Galan-Sousa, K. Hoyer, W. Chan, M. Jäger, K. Yoshida, R. Seemann, D. Noerenberg, N. Waldhueter, H. Fleischer-Notter, F. Christen, C. A. Schmitt, B. Dörken, U. Pelzer, M. Sinn, T. Zemojtel, S. Ogawa, S. Märdian, A. Schreiber, A. Kunitz, U. Krüger, L. Bullinger, E. Mylonas, M. Frick, and F. Damm. Hematopoietic lineage distribution and evolutionary dynamics of clonal hematopoiesis. *Leukemia*, 2018.

[89] O. Akinduro, T. S. Weber, H. Ang, M. L. R. Haltalli, N. Ruivo, D. Duarte, N. M. Rashidi, E. D. Hawkins, K. R. Duffy, and C. Lo Celso. Proliferation dynamics of acute myeloid leukaemia and haematopoietic progenitors competing for bone marrow space. *Nature Communications*, 9(1):519, 2018.

[90] P. R. Dharampuriya, G. Scapin, C. Wong, K.J. Wagner, J.L. Cillis, and D.I. Shah. Tracking the origin, development, and differentiation of hematopoietic stem cells. *Current Opinion in Cell Biology*, 49:108 -- 115, 2017. Cell Differentiation and Development.

[91] N. Roquet, A. P. Soleimany, A. C. Ferris, S. Aaronson, and T. K. Lu. Synthetic recombinase-based state machines in living cells. *Science*, 353(6297), 2016.

[92] B. Spanjaard and J.P. Junker. Methods for lineage tracing on the organism-wide level. *Current Opinion in Cell Biology*, 49:16 -- 21, 2017.

[93] B. Efron and R.J. Tibshirani. An introduction to the bootstrap. *Chapman and Hall*, 1993.

[94] R. E. Korf. Depth-first iterative-deepening: An optimal admissible tree search. *Artif. Intell.*, 27(1):97--109, September 1985.

[95] E.F. Moore. The shortest path through a maze. 1959.

[96] J. Palis. Hematopoietic stem cell-independent hematopoiesis: emergence of erythroid, megakaryocyte, and myeloid potential in the mammalian embryo. *FEBS Letters*, 590(22):3965--3974, 2018/11/06 2016.

[97] L. Perié and K. R. Duffy. Retracing the in vivo haematopoietic tree using single-cell methods. *FEBS Letters*, 590(22):4068--4083, 2018/11/06 2016.

A| Appendix

A.1| Mouse List

Mouse	Strain	Treatment	Analysis	internal
			(time after treatment)	mouse ID
#1	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E9.5)	39 weeks	#20
#2	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E9.5)	47 weeks	#18
#3	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	5 x TAM i.p.	49 weeks	#35577
#4	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	5 x TAM i.p.	51 weeks	#35587
#5	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	5 x TAM i.p.	18 hours	#36384
#6	$Rosa26^{\text{Polylox/CreERT2}}$	1 x TAM i.p.	18 hours	#36409
#7	$Rosa26^{\text{Polylox/CreERT2}}$	1 x TAM i.p.	18 hours	#36410
#8	$Rosa26^{\text{Polylox/CreERT2}}$	1 x TAM i.p.	18 hours	#36411
#9	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E10.5)	45 weeks	#8
#10	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E7.5)	104 weeks	#21
#11	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E7.5)	72 weeks	#46
#12	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E7.5)	54 weeks	#47
#13	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E10.5)	43 weeks	#62
#14	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E10.5)	86 weeks	#49
#15	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x 4-OHT (E7.5)	21 days	#K3
#16	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x 4-OHT (E7.5)	21 days	#K5
#17	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x 4-OHT (E7.5)	11 weeks	#K9
#18	$Rosa26^{\text{Polylox}}Tie2^{\text{MCM}}$	1 x TAM (E7.5)	87 weeks	#34

Table A.1: "In-utero labelling for embryonic treatment (E9.5, E10.5) was done by oral gavage to the mother. Labeling of adult mice was done by intraperitoneal tamoxifen injection (i.p.). Time after treatment is given as the time span between first treatment and the day of analysis [22]."

A.2| FACS-Gating

Population	FACS-Gates
HSC	Lin– Sca+ Kit+ CD48– CD150+
ST-HSC	Lin– Sca+ Kit+ CD48– CD150–
MPP	Lin– Sca+ Kit+ CD48+ CD150–
CMP	Lin– Sca– Kit+ CD16/32lo CD34lo
GMP	Lin– Sca– Kit+ CD16/32+ CD34+
EryPI	Lin– Ter119low CD71+ CD44hi FSChi
EryPII	Lin– Ter119+ CD71+ CD44+ FSC+
EryPIII	Lin– Ter119+ CD71+ CD44med FSCmed
EryPIV	Lin– Ter119+ CD71+ CD44low FSClow
Gr	CD4– CD8– CD19– CD11b+ Gr-1+
Mono	CD4– CD8– CD19– Ter119– CD45+ Ly6G– CD11b+ CD115+ MHCII–
CLP	Lin– Kitlo CD127+ CD135+ B220lo
pre B cells	Lin– CD43+ B220+ CD24+ BP1- and BP1+
B cells	CD4– CD8– CD19+ CD11b– Gr-1–
B2 PEC	CD5– CD11b– CD19+ CD21– CD23+ CD93– IgM+
pre T cells	CD11b-, CD19-, NK1.1-, Gr1-, Ter119-, CD3-, CD4-, CD8
DN2	CD44+, CD25+
DN3	CD44-, CD25+
CD4+ T cells	CD4+ CD8– CD19– CD11b– Gr-1–
CD8+ T cells	CD4– CD8+ CD19– CD11b– Gr-1–
B1a PEC	CD5+ CD11b+ CD19+ CD21– CD23– CD93– IgM+
B1b PEC	CD5– CD11b+ CD19+ CD21– CD23– CD93– IgM+

Table A.2: FACS-Gating for populations [22]

A.3| ODE-Systems

A.3.1| Examplary parameters for toy models

Parameter	fig. 6.2 b	fig. 6.2 c	fig. 6.2 d	fig. 6.2 e
l_A	0.41	0.79	0.30	0.27
l_C	0.63	0.49	0.19	0.34
l_B	0.41	0.68	0.19	0.05
l_D	0.93	0.41	0.86	0.32
d_{AB}	0.19	0.94	0.59	0.58
d_{BC}	0.49	0.77	0.71	0.45
d_{CD}	0.76	0.84	0.29	0.38

Table A.3: Exemplary parameters of toy models used in chapter 6

A.3.2| Toy Model II

$$
\begin{aligned}
dn =&(l_1 - d_1)\langle n_1\rangle\\
\langle \dot{n_2}\rangle =&(l_2 - d_2 - d_3)\langle n_2\rangle + d_1\langle n_1\rangle\\
\langle \dot{n_3}\rangle =&(l_3 - d_4)\langle n_3\rangle + d_2\langle n_2\rangle\\
\langle \dot{n_4}\rangle =&(l_3 - d_5)\langle n_3\rangle + d_3\langle n_2\rangle\\
\langle \dot{n_5}\rangle =&l_5\langle n_5\rangle + d_4\langle n_3\rangle\\
\langle \dot{n_6}\rangle =&l_6\langle n_6\rangle + d_5\langle n_4\rangle
\end{aligned}
$$

$$
\begin{aligned}
\dot{F}_{1,1} =&2l_1\langle n_1\rangle + 2(l_1 - d_1)F_{1,1}\\
\dot{F}_{1,2} =&-(d_1 + d_2 + d_3 - l_1 - l_2)F_{1,2} + d_1F_{1,1}\\
\dot{F}_{1,3} =&(-d_1 - d_4 + l_1 + l_3)F_{1,3} + d_2F_{1,2}\\
\dot{F}_{1,4} =&(-d_1 - d_5 + l_1 + l_4)F_{1,4} + d_3F_{1,2}\\
\dot{F}_{1,5} =&(-d_1 + l_1 + l_5)F_{1,5} + d_4F_{1,3}\\
\dot{F}_{1,6} =&(-d_1 + l_1 + l_6)F_{1,6} + d_5F_{1,4}\\
\dot{F}_{2,2} =&2(-(d_2 + d_3)F_{2,2} + l_2(\langle n_2\rangle + F_{2,2}) + d_1F_{1,2})\\
\dot{F}_{2,3} =&-(d_2 + d_3 + d_4 - l_2 - l_3)F_{2,3} + d_2F_{2,2} + d_1F_{1,3}\\
\dot{F}_{2,4} =&-(d_2 + d_3 + d_5 - l_2 - l_4)F_{2,4} + d_3F_{2,2} + d_1F_{1,4}\\
\dot{F}_{2,5} =&(-d_2 - d_3 + l_2 + l_5)F_{2,5} + d_4F_{2,3} + d_1F_{1,5}\\
\dot{F}_{2,6} =&(-d_2 - d_3 + l_2 + l_6)F_{2,6} + d_5F_{2,4} + d_1F_{1,6}\\
\dot{F}_{3,3} =&2(-d_4F_{3,3} + l_3(\langle n_3\rangle + F_{3,3}) + d_2F_{2,3})\\
\dot{F}_{3,4} =&(-d_4 - d_5 + l_3 + l_4)F_{3,4} + d_2F_{2,4} + d_3F_{2,3}\\
\dot{F}_{3,5} =&(-d_4 + l_3 + l_5)F_{3,5} + d_4F_{3,3} + d_2F_{2,5}\\
\dot{F}_{3,6} =&(-d_4 + l_3 + l_6)F_{3,6} + d_5F_{3,4} + d_2F_{2,6}\\
\dot{F}_{4,4} =&2(-d_5F_{4,4} + l_4(\langle n_4\rangle + F_{4,4}) + d_3F_{2,4})
\end{aligned}
$$

$$\begin{aligned}
\dot{F}_{4,5} =&(-d_5 + l_4 + l_5)F_{4,5} + d_4 F_{3,4} + d_3 F_{2,5} \\
\dot{F}_{4,6} =&(-d_5 + l_4 + l_6)F_{4,6} + d_5 F_{4,4} + d_3 F_{2,6} \\
\dot{F}_{5,5} =&2(l_5(\langle n_5\rangle + F_{5,5}) + d_4 F_{3,5}) \\
\dot{F}_{5,6} =&(l_5 + l_6)F_{5,6} + d_5 F_{4,5} + d_4 F_{3,6} \\
\dot{F}_{6,6} =&2(l_6(\langle n_6\rangle + F_{6,6}) + d_5 F_{4,6})
\end{aligned} \tag{A.1}$$

A.3.3 Toy Model III

$$\begin{aligned}
\langle\dot{n}_1\rangle =& -(d_1 + d_2 - l_1)\langle n1\rangle \\
\langle\dot{n}_2\rangle =&(-d_3 + l_2)\langle n_2\rangle + d_1\langle n_1\rangle \\
\langle\dot{n}_3\rangle =&(-d_4 + l_3)\langle n_3\rangle + d_2\langle n_1\rangle \\
\langle\dot{n}_4\rangle =&(-d_5 + l_4)\langle n_4\rangle + d_3\langle n_2\rangle \\
\langle\dot{n}_5\rangle =&(-d_5 + l_5)\langle n_5\rangle + d_4\langle n_3\rangle \\
\langle\dot{n}_6\rangle =&l_6\langle n_6\rangle + d_5\langle n_4\rangle \\
\langle\dot{n}_7\rangle =&l_7\langle n_7\rangle + d_6\langle n_5\rangle
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{1,1} =&2l_1\langle n_1\rangle - 2(d_1 + d_2 - l_1)F_{11} \\
\dot{F}_{1,2} =& -(d_1 + d_2 + d_3 - l_1 - l_2)F_{1,2} + d_1 F_{1,1} \\
\dot{F}_{1,3} =& -(d_1 + d_2 + d_4 - l_1 - l_3)F_{1,3} + d_2 F_{1,1} \\
\dot{F}_{1,4} =& -(d_1 + d_2 + d_5 - l_1 - l_4)F_{1,4} + d_3 F_{1,2} \\
\dot{F}_{1,5} =& -(d_1 + d_2 + d_6 - l_1 - l_5)F_{1,5} + d_4 F_{1,3} \\
\dot{F}_{1,6} =&(-d_1 - d_2 + l_1 + l_6)F_{1,6} + d_5 F_{1,4} \\
\dot{F}_{1,7} =&(-d_1 - d_2 + l_1 + l_7)F_{1,7} + d_6 F_{1,5} \\
\dot{F}_{2,2} =&2(-d_3 F_{2,2} + l_2(\langle n_2\rangle + F_{2,2}) + d_1 F_{1,2}) \\
\dot{F}_{2,3} =&(-d_3 - d_4 + l_2 + l_3)F_{2,3} + d_1 F_{1,3} + d_2 F_{1,2} \\
\dot{F}_{2,4} =&(-d_3 - d_5 + l_2 + l_4)F_{2,4} + d_3 F_{2,2} + d_1 F_{1,4} \\
\dot{F}_{2,5} =&(-d_3 - d_6 + l_2 + l_5)F_{2,5} + d_4 F_{2,3} + d_1 F_{1,5} \\
\dot{F}_{2,6} =&(-d_3 + l_2 + l_6)F_{2,6} + d_5 F_{2,4} + d_1 F_{1,6} \\
\dot{F}_{2,7} =&(-d_3 + l_2 + l_7)F_{2,7} + d_6 F_{2,5} + d_1 F_{1,7} \\
\dot{F}_{3,3} =&2(-d_4 F_{3,3} + l_3(\langle n_3\rangle + F_{3,3}) + d_2 F_{1,3}) \\
\dot{F}_{3,4} =&(-d_4 - d_5 + l_3 + l_4)F_{3,4} + d_3 F_{2,3} + d_2 F_{1,4} \\
\dot{F}_{3,5} =&(-d_4 - d_6 + l_3 + l_5)F_{3,5} + d_4 F_{3,3} + d_2 F_{1,5} \\
\dot{F}_{3,6} =&(-d_4 + l_3 + l_6)F_{3,6} + d_5 F_{3,4} + d_2 F_{1,6} \\
\dot{F}_{3,7} =&(-d_4 + l_3 + l_7)F_{3,7} + d_6 F_{3,5} + d_2 F_{1,7} \\
\dot{F}_{4,4} =&2(-d_5 F_{4,4} + l_4(\langle n_4\rangle + F_{4,4}) + d_3 F_{2,4}) \\
\dot{F}_{4,5} =&(-d_5 - d_6 + l_4 + l_5)F_{4,5} + d_4 F_{3,4} + d_3 F_{2,5} \\
\dot{F}_{4,6} =&(-d_5 + l_4 + l_6)F_{4,6} + d_5 F_{4,4} + d_3 F_{2,6}
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{4,7} =&(-d_5 + l_4 + l_7)F_{4,7} + d_6F_{4,5} + d_3F_{2,7}\\
\dot{F}_{5,5} =&2(-d_6F_{5,5} + l_5(\langle n_5\rangle + F_{5,5}) + d_4F_{3,5})\\
\dot{F}_{5,6} =&(-d_6 + l_5 + l_6)F_{5,6} + d_5F_{4,5} + d_4F_{3,6}\\
\dot{F}_{5,7} =&(-d_6 + l_5 + l_7)F_{5,7} + d_6F_{5,5} + d_4F_{3,7}\\
\dot{F}_{6,6} =&2(l_6(\langle n_6\rangle + F_{6,6}) + d_5F_{4,6})\\
\dot{F}_{6,7} =&(l_6 + l_7)F_{6,7} + d_6F_{5,6} + d_5F_{4,7}\\
\dot{F}_{7,7} =&2(l_7(\langle n_7\rangle + F_{7,7}) + d_6F_{5,7})
\end{aligned} \tag{A.2}$$

A.3.4| Full Tree Model

$$\begin{aligned}
\langle \dot{n}_1\rangle =&-(d_1 - l_1)\langle n_1\rangle\\
\langle \dot{n}_2\rangle =&-(d_2 - l_2)\langle n_2\rangle + d_1\langle n_1\rangle\\
\langle \dot{n}_3\rangle =&-(d_{3,4} + d_{3,5} - l_3)\langle n_3\rangle + d_2\langle n_2\rangle\\
\langle \dot{n}_4\rangle =&-(d_{4,6} + d_{4,8} - l_4)\langle n_4\rangle + d_{3,4}\langle n_3\rangle\\
\langle \dot{n}_5\rangle =&-(d_{5,10} + d_{5,12} - l_5)\langle n_5\rangle + d_{3,5}\langle n_3\rangle\\
\langle \dot{n}_6\rangle =&-(d_{6,7} - l_6)\langle n_6\rangle + d_{4,6}\langle n_4\rangle\\
\langle \dot{n}_7\rangle =&l_7\langle n_7\rangle + d_{6,7}\langle n_6\rangle\\
\langle \dot{n}_8\rangle =&-(d_{8,9} - l_8)\langle n_8\rangle + d_{4,8}\langle n_4\rangle\\
\langle \dot{n}_9\rangle =&l_9\langle n_9\rangle + d_{8,9}\langle n_8\rangle\\
\langle \dot{n}_{10}\rangle =&-(d_{10,11} - l_{10})\langle n_{10}\rangle + d_{5,10}\langle n_5\rangle\\
\langle \dot{n}_{11}\rangle =&l_{11}\langle n_{11}\rangle + d_{10,11}\langle n_{10}\rangle\\
\langle \dot{n}_{12}\rangle =&-(d_{12,13} - l_{12})\langle n_{12}\rangle + d_{5,12}\langle n_5\rangle\\
\langle \dot{n}_{13}\rangle =&-(d_{13,14} + d_{13,15} - l_{13})\langle n_{13}\rangle + d_{12,13}\langle n_{12}\rangle\\
\langle \dot{n}_{14}\rangle =&l_{14}\langle n_{14}\rangle + d_{13,14}\langle n_{13}\rangle\\
\langle \dot{n}_{15}\rangle =&l_{15}\langle n_{15}\rangle + d_{13,15}\langle n_{13}\rangle
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{1,1} =&_2l_1\langle n_1\rangle -_2 (d_1 - l_1)F_{1,1}\\
\dot{F}_{1,2} =&-(d_1 + d_2 - l_1 - l_2)F_{1,2} + d_1F_{1,1}\\
\dot{F}_{1,3} =&-(d_1 + d_{3,4} + d_{3,5} - l_1 - l_3)F_{1,3} + d_2F_{1,2}\\
\dot{F}_{1,4} =&-(d_1 + d_{4,6} + d_{4,8} - l_1 - l_4)F_{1,4} + d_{3,4}F_{1,3}\\
\dot{F}_{1,5} =&-(d_1 + d_{5,10} + d_{5,12} - l_1 - l_5)F_{1,5} + d_{3,5}F_{1,3}\\
\dot{F}_{1,6} =&-(d_1 + d_{6,7} - l_1 - l_6)F_{1,6} + d_{4,6}F_{1,4}\\
\dot{F}_{1,7} =&-(d_1 - l_1 - l_7)F_{1,7} + d_{6,7}F_{1,6}\\
\dot{F}_{1,8} =&-(d_1 + d_{8,9} - l_1 - l_8)F_{1,8} + d_{4,8}F_{1,4}\\
\dot{F}_{1,9} =&-(d_1 - l_1 - l_9)F_{1,9} + d_{8,9}F_{1,8}\\
\dot{F}_{1,10} =&-(d_1 + d_{10,11} - l_1 - l_{10})F_{1,10} + d_{5,10}F_{1,5}\\
\dot{F}_{1,11} =&-(d_1 - l_1 - l_{11})F_{1,11} + d_{10,11}F_{1,10}
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{1,12} =& -(d_1 + d_{12,13} - l_1 - l_{12})F_{1,12} + d_{5,12}F_{1,5}\\
\dot{F}_{1,13} =& -(d_1 + d_{13,14} + d_{13,15} - l_1 - l_{13})F_{1,13} + d_{12,13}F_{1,12}\\
\dot{F}_{1,14} =& -(d_1 - l_1 - l_{14})F_{1,14} + d_{13,14}F_{1,13}\\
\dot{F}_{1,15} =& -(d_1 - l_1 - l_{15})F_{1,15} + d_{13,15}F_{1,13}
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{2,2} =& 2(-(d_2)F_{2,2} + l_2(\langle n_2\rangle + F_{2,2}) + d_1F_{1,2})\\
\dot{F}_{2,3} =& -(d_2 + d_{3,4} + d_{3,5} - l_2 - l_3)F_{2,3} + d_2F_{2,2} + d_1F_{1,3}\\
\dot{F}_{2,4} =& -(d_2 + d_{4,6} + d_{4,8} - l_2 - l_4)F_{2,4} + d_{3,4}F_{2,3} + d_1F_{1,4}\\
\dot{F}_{2,5} =& -(d_2 + d_{5,10} + d_{5,12} - l_2 - l_5)F_{2,5} + d_{3,5}F_{2,3} + d_1F_{1,5}\\
\dot{F}_{2,6} =& -(d_2 + d_{6,7} - l_2 - l_6)F_{2,6} + d_{4,6}F_{2,4} + d_1F_{1,6}\\
\dot{F}_{2,7} =& -(d_2 - l_2 - l_7)F_{2,7} + d_{6,7}F_{2,6} + d_1F_{1,7}\\
\dot{F}_{2,8} =& -(d_2 + d_{8,9} - l_2 - l_8)F_{2,8} + d_{4,8}F_{2,4} + d_1F_{1,8}\\
\dot{F}_{2,9} =& -(d_2 - l_2 - l_9)F_{2,9} + d_{8,9}F_{2,8} + d_1F_{1,9}\\
\dot{F}_{2,10} =& -(d_{10,11} + d_2 - l_{10} - l_2)F_{2,10} + d_{5,10}F_{2,5} + d_1F_{1,10}\\
\dot{F}_{2,11} =& -(d_2 - l_{11} - l_2)F_{2,11} + d_{10,11}F_{2,10} + d_1F_{1,11}\\
\dot{F}_{2,12} =& -(d_{12,13} + d_2 - l_{12} - l_2)F_{2,12} + d_{5,12}F_{2,5} + d_1F_{1,12}\\
\dot{F}_{2,13} =& -(d_{13,14} + d_{13,15} + d_2 - l_{13} - l_2)F_{2,13} + d_{12,13}F_{2,12} + d_1F_{1,13}\\
\dot{F}_{2,14} =& -(d_2 - l_{14} - l_2)F_{2,14} + d_{13,14}F_{2,13} + d_1F_{1,14}\\
\dot{F}_{2,15} =& -(d_2 - l_{15} - l_2)F_{2,15} + d_{13,15}F_{2,13} + d_1F_{1,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{3,3} =& 2(-(d_{3,4} + d_{3,5})F_{3,3} + l_3(\langle n_3\rangle + F_{3,3}) + d_2F_{2,3})\\
\dot{F}_{3,4} =& -(d_{3,4} + d_{3,5} + d_{4,6} + d_{4,8} - l_3 - l_4)F_{3,4} + d_{3,4}F_{3,3} + d_2F_{2,4}\\
\dot{F}_{3,5} =& -(d_{3,4} + d_{3,5} + d_{5,10} + d_{5,12} - l_3 - l_5)F_{3,5} + d_{3,5}F_{3,3} + d_2F_{2,5}\\
\dot{F}_{3,6} =& -(d_{3,4} + d_{3,5} + d_{6,7} - l_3 - l_6)F_{3,6} + d_{4,6}F_{3,4} + d_2F_{2,6}\\
\dot{F}_{3,7} =& -(d_{3,4} + d_{3,5} - l_3 - l_7)F_{3,7} + d_{6,7}F_{3,6} + d_2F_{2,7}\\
\dot{F}_{3,8} =& -(d_{3,4} + d_{3,5} + d_{8,9} - l_3 - l_8)F_{3,8} + d_{4,8}F_{3,4} + d_2F_{2,8}\\
\dot{F}_{3,9} =& -(d_{3,4} + d_{3,5} - l_3 - l_9)F_{3,9} + d_{8,9}F_{3,8} + d_2F_{2,9}\\
\dot{F}_{3,10} =& -(d_{10,11} + d_{3,4} + d_{3,5} - l_{10} - l_3)F_{3,10} + d_{5,10}F_{3,5} + d_2F_{2,10}\\
\dot{F}_{3,11} =& -(d_{3,4} + d_{3,5} - l_{11} - l_3)F_{3,11} + d_{10,11}F_{3,10} + d_2F_{2,11}\\
\dot{F}_{3,12} =& -(d_{12,13} + d_{3,4} + d_{3,5} - l_{12} - l_3)F_{3,12} + d_{5,12}F_{3,5} + d_2F_{2,12}\\
\dot{F}_{3,13} =& -(d_{13,14} + d_{13,15} + d_{3,4} + d_{3,5} - l_{13} - l_3)F_{3,13} + d_{12,13}F_{3,12} + d_2F_{2,13}\\
\dot{F}_{3,14} =& -(d_{3,4} + d_{3,5} - l_{14} - l_3)F_{3,14} + d_{13,14}F_{3,13} + d_2F_{2,14}\\
\dot{F}_{3,15} =& -(d_{3,4} + d_{3,5} - l_{15} - l_3)F_{3,15} + d_{13,15}F_{3,13} + d_2F_{2,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F}_{4,4} =& 2(-(d_{4,6} + d_{4,8})F_{4,4} + l_4(\langle n_4\rangle + F_{4,4}) + d_{3,4}F_{3,4})\\
\dot{F}_{4,5} =& -(d_{4,6} + d_{4,8} + d_{5,10} + d_{5,12} - l_4 - l_5)F_{4,5} + d_{3,4}F_{3,5} + d_{3,5}F_{3,4}\\
\dot{F}_{4,6} =& -(d_{4,6} + d_{4,8} + d_{6,7} - l_4 - l_6)F_{4,6} + d_{4,6}F_{4,4} + d_{3,4}F_{3,6}\\
\dot{F}_{4,7} =& -(d_{4,6} + d_{4,8} - l_4 - l_7)F_{4,7} + d_{6,7}F_{4,6} + d_{3,4}F_{3,7}
\end{aligned}$$

$$
\begin{aligned}
\dot{F_{4,8}} =& -(d_{4,6}+d_{4,8}+d_{8,9}-l_4-l_8)F_{4,8}+d_{4,8}F_{4,4}+d_{3,4}F_{3,8}\\
\dot{F_{4,9}} =& -(d_{4,6}+d_{4,8}-l_4-l_9)F_{4,9}+d_{8,9}F_{4,8}+d_{3,4}F_{3,9}\\
\dot{F_{4,10}} =& -(d_{10,11}+d_{4,6}+d_{4,8}-l_{10}-l_4)F_{4,10}+d_{5,10}F_{4,5}+d_{3,4}F_{3,10}\\
\dot{F_{4,11}} =& -(d_{4,6}+d_{4,8}-l_{11}-l_4)F_{4,11}+d_{10,11}F_{4,10}+d_{3,4}F_{3,11}\\
\dot{F_{4,12}} =& -(d_{12,13}+d_{4,6}+d_{4,8}-l_{12}-l_4)F_{4,12}+d_{5,12}F_{4,5}+d_{3,4}F_{3,12}\\
\dot{F_{4,13}} =& -(d_{13,14}+d_{13,15}+d_{4,6}+d_{4,8}-l_{13}-l_4)F_{4,13}+d_{12,13}F_{4,12}+d_{3,4}F_{3,13}\\
\dot{F_{4,14}} =& -(d_{4,6}+d_{4,8}-l_{14}-l_4)F_{4,14}+d_{13,14}F_{4,13}+d_{3,4}F_{3,14}\\
\dot{F_{4,15}} =& -(d_{4,6}+d_{4,8}-l_{15}-l_4)F_{4,15}+d_{13,15}F_{4,13}+d_{3,4}F_{3,15}
\end{aligned}
$$

$$
\begin{aligned}
\dot{F_{5,5}} =& 2(-(d_{5,10}+d_{5,12})F_{5,5}+l_5(\langle n_5\rangle+F_{5,5})+d_{3,5}F_{3,5})\\
\dot{F_{5,6}} =& -(d_{5,10}+d_{5,12}+d_{6,7}-l_5-l_6)F_{5,6}+d_{4,6}F_{4,5}+d_{3,5}F_{3,6}\\
\dot{F_{5,7}} =& -(d_{5,10}+d_{5,12}-l_5-l_7)F_{5,7}+d_{6,7}F_{5,6}+d_{3,5}F_{3,7}\\
\dot{F_{5,8}} =& -(d_{5,10}+d_{5,12}+d_{8,9}-l_5-l_8)F_{5,8}+d_{4,8}F_{4,5}+d_{3,5}F_{3,8}\\
\dot{F_{5,9}} =& -(d_{5,10}+d_{5,12}-l_5-l_9)F_{5,9}+d_{8,9}F_{5,8}+d_{3,5}F_{3,9}\\
\dot{F_{5,10}} =& -(d_{10,11}+d_{5,10}+d_{5,12}-l_{10}-l_5)F_{5,10}+d_{5,10}F_{5,5}+d_{3,5}F_{3,10}\\
\dot{F_{5,11}} =& -(d_{5,10}+d_{5,12}-l_{11}-l_5)F_{5,11}+d_{10,11}F_{5,10}+d_{3,5}F_{3,11}\\
\dot{F_{5,12}} =& -(d_{12,13}+d_{5,10}+d_{5,12}-l_{12}-l_5)F_{5,12}+d_{5,12}F_{5,5}+d_{3,5}F_{3,12}\\
\dot{F_{5,13}} =& -(d_{13,14}+d_{13,15}+d_{5,10}+d_{5,12}-l_{13}-l_5)F_{5,13}+d_{12,13}F_{5,12}+d_{3,5}F_{3,13}\\
\dot{F_{5,14}} =& -(d_{5,10}+d_{5,12}-l_{14}-l_5)F_{5,14}+d_{13,14}F_{5,13}+d_{3,5}F_{3,14}\\
\dot{F_{5,15}} =& -(d_{5,10}+d_{5,12}-l_{15}-l_5)F_{5,15}+d_{13,15}F_{5,13}+d_{3,5}F_{3,15}\\
\dot{F_{6,6}} =& 2(-(d_{6,7})F_{6,6}+l_6(\langle n_6\rangle+F_{6,6})+d_{4,6}F_{4,6})
\end{aligned}
$$

$$
\begin{aligned}
\dot{F_{6,7}} =& -(d_{6,7}-l_6-l_7)F_{6,7}+d_{6,7}F_{6,6}+d_{4,6}F_{4,7}\\
\dot{F_{6,8}} =& -(d_{6,7}+d_{8,9}-l_6-l_8)F_{6,8}+d_{4,6}F_{4,8}+d_{4,8}F_{4,6}\\
\dot{F_{6,9}} =& -(d_{6,7}-l_6-l_9)F_{6,9}+d_{8,9}F_{6,8}+d_{4,6}F_{4,9}\\
\dot{F_{6,10}} =& -(d_{10,11}+d_{6,7}-l_{10}-l_6)F_{6,10}+d_{5,10}F_{5,6}+d_{4,6}F_{4,10}\\
\dot{F_{6,11}} =& -(d_{6,7}-l_{11}-l_6)F_{6,11}+d_{10,11}F_{6,10}+d_{4,6}F_{4,11}\\
\dot{F_{6,12}} =& -(d_{12,13}+d_{6,7}-l_{12}-l_6)F_{6,12}+d_{5,12}F_{5,6}+d_{4,6}F_{4,12}\\
\dot{F_{6,13}} =& -(d_{13,14}+d_{13,15}+d_{6,7}-l_{13}-l_6)F_{6,13}+d_{12,13}F_{6,12}+d_{4,6}F_{4,13}\\
\dot{F_{6,14}} =& -(d_{6,7}-l_{14}-l_6)F_{6,14}+d_{13,14}F_{6,13}+d_{4,6}F_{4,14}\\
\dot{F_{6,15}} =& -(d_{6,7}-l_{15}-l_6)F_{6,15}+d_{13,15}F_{6,13}+d_{4,6}F_{4,15}\\
\dot{F_{7,7}} =& 2(l_7\langle n_7\rangle+l_7F_{7,7}+d_{6,7}F_{6,7})
\end{aligned}
$$

$$
\begin{aligned}
\dot{F_{7,8}} =& -(d_{8,9}-l_7-l_8)F_{7,8}+d_{6,7}F_{6,8}+d_{4,8}F_{4,7}\\
\dot{F_{7,9}} =& (l_7+l_9)F_{7,9}+d_{8,9}F_{7,8}+d_{6,7}F_{6,9}\\
\dot{F_{7,10}} =& -(d_{10,11}-l_{10}-l_7)F_{7,10}+d_{6,7}F_{6,10}+d_{5,10}F_{5,7}\\
\dot{F_{7,11}} =& (l_{11}+l_7)F_{7,11}+d_{10,11}F_{7,10}+d_{6,7}F_{6,11}\\
\dot{F_{7,12}} =& -(d_{12,13}-l_{12}-l_7)F_{7,12}+d_{6,7}F_{6,12}+d_{5,12}F_{5,7}
\end{aligned}
$$

$$\begin{aligned}
\dot{F_{7,13}} =& -(d_{13,14} + d_{13,15} - l_{13} - l_7)F_{7,13} + d_{12,13}F_{7,12} + d_{6,7}F_{6,13} \\
\dot{F_{7,14}} =&(l_{14} + l_7)F_{7,14} + d_{13,14}F_{7,13} + d_{6,7}F_{6,14} \\
\dot{F_{7,15}} =&(l_{15} + l_7)F_{7,15} + d_{13,15}F_{7,13} + d_{6,7}F_{6,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F_{8,8}} =&2(-(d_{8,9})F_{8,8} + l_8(\langle n_8\rangle + F_{8,8}) + d_{4,8}F_{4,8}) \\
\dot{F_{8,9}} =& -(d_{8,9} - l_8 - l_9)F_{8,9} + d_{8,9}F_{8,8} + d_{4,8}F_{4,9} \\
\dot{F_{8,10}} =& -(d_{10,11} + d_{8,9} - l_{10} - l_8)F_{8,10} + d_{5,10}F_{5,8} + d_{4,8}F_{4,10} \\
\dot{F_{8,11}} =& -(d_{8,9} - l_{11} - l_8)F_{8,11} + d_{10,11}F_{8,10} + d_{4,8}F_{4,11} \\
\dot{F_{8,12}} =& -(d_{12,13} + d_{8,9} - l_{12} - l_8)F_{8,12} + d_{5,12}F_{5,8} + d_{4,8}F_{4,12} \\
\dot{F_{8,13}} =& -(d_{13,14} + d_{13,15} + d_{8,9} - l_{13} - l_8)F_{8,13} + d_{12,13}F_{8,12} + d_{4,8}F_{4,13} \\
\dot{F_{8,14}} =& -(d_{8,9} - l_{14} - l_8)F_{8,14} + d_{13,14}F_{8,13} + d_{4,8}F_{4,14} \\
\dot{F_{8,15}} =& -(d_{8,9} - l_{15} - l_8)F_{8,15} + d_{13,15}F_{8,13} + d_{4,8}F_{4,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F_{9,9}} =&2(l_9\langle n_9\rangle + (l_9)F_{9,9} + d_{8,9}F_{8,9}) \\
\dot{F_{9,10}} =& -(d_{10,11} - l_{10} - l_9)F_{9,10} + d_{8,9}F_{8,10} + d_{5,10}F_{5,9} \\
\dot{F_{9,11}} =&(l_{11} + l_9)F_{9,11} + d_{10,11}F_{9,10} + d_{8,9}F_{8,11} \\
\dot{F_{9,12}} =& -(d_{12,13} - l_{12} - l_9)F_{9,12} + d_{8,9}F_{8,12} + d_{5,12}F_{5,9} \\
\dot{F_{9,13}} =& -(d_{13,14} + d_{13,15} - l_{13} - l_9)F_{9,13} + d_{12,13}F_{9,12} + d_{8,9}F_{8,13} \\
\dot{F_{9,14}} =&(l_{14} + l_9)F_{9,14} + d_{13,14}F_{9,13} + d_{8,9}F_{8,14} \\
\dot{F_{9,15}} =&(l_{15} + l_9)F_{9,15} + d_{13,15}F_{9,13} + d_{8,9}F_{8,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F_{10,10}} =&2(-(d_{10,11})F_{10,10} + l_{10}(\langle n_{10}\rangle + F_{10,10}) + d_{5,10}F_{5,10}) \\
\dot{F_{10,11}} =& -(d_{10,11} - l_{10} - l_{11})F_{10,11} + d_{10,11}F_{10,10} + d_{5,10}F_{5,11} \\
\dot{F_{10,12}} =& -(d_{10,11} + d_{12,13} - l_{10} - l_{12})F_{10,12} + d_{5,10}F_{5,12} + d_{5,12}F_{5,10} \\
\dot{F_{10,13}} =& -(d_{10,11} + d_{13,14} + d_{13,15} - l_{10} - l_{13})F_{10,13} + d_{12,13}F_{10,12} + d_{5,10}F_{5,13} \\
\dot{F_{10,14}} =& -(d_{10,11} - l_{10} - l_{14})F_{10,14} + d_{13,14}F_{10,13} + d_{5,10}F_{5,14} \\
\dot{F_{10,15}} =& -(d_{10,11} - l_{10} - l_{15})F_{10,15} + d_{13,15}F_{10,13} + d_{5,10}F_{5,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F_{11,11}} =&2(l_{11}\langle n_{11}\rangle + (l_{11})F_{11,11} + d_{10,11}F_{10,11}) \\
\dot{F_{11,12}} =& -(d_{12,13} - l_{11} - l_{12})F_{11,12} + d_{10,11}F_{10,12} + d_{5,12}F_{5,11} \\
\dot{F_{11,13}} =& -(d_{13,14} + d_{13,15} - l_{11} - l_{13})F_{11,13} + d_{12,13}F_{11,12} + d_{10,11}F_{10,13} \\
\dot{F_{11,14}} =&(l_{11} + l_{14})F_{11,14} + d_{13,14}F_{11,13} + d_{10,11}F_{10,14} \\
\dot{F_{11,15}} =&(l_{11} + l_{15})F_{11,15} + d_{13,15}F_{11,13} + d_{10,11}F_{10,15}
\end{aligned}$$

$$\begin{aligned}
\dot{F_{12,12}} =&2(-(d_{12,13})F_{12,12} + l_{12}(\langle n_{12}\rangle + F_{12,12}) + d_{5,12}F_{5,12}) \\
\dot{F_{12,13}} =& -(d_{12,13} + d_{13,14} + d_{13,15} - l_{12} - l_{13})F_{12,13} + d_{12,13}F_{12,12} + d_{5,12}F_{5,13} \\
\dot{F_{12,14}} =& -(d_{12,13} - l_{12} - l_{14})F_{12,14} + d_{13,14}F_{12,13} + d_{5,12}F_{5,14} \\
\dot{F_{12,15}} =& -(d_{12,13} - l_{12} - l_{15})F_{12,15} + d_{13,15}F_{12,13} + d_{5,12}F_{5,15}
\end{aligned}$$

$$
\begin{aligned}
\dot{F}_{13,13} =& 2(-(d_{13,14} + d_{13,15})F_{13,13} + l_{13}(\langle n_{13}\rangle + F_{13,13}) + d_{12,13}F_{12,13}) \\
\dot{F}_{13,14} =& -(d_{13,14} + d_{13,15} - l_{13} - l_{14})F_{13,14} + d_{13,14}F_{13,13} + d_{12,13}F_{12,14} \\
\dot{F}_{13,15} =& -(d_{13,14} + d_{13,15} - l_{13} - l_{15})F_{13,15} + d_{13,15}F_{13,13} + d_{12,13}F_{12,15} \\
\\
\dot{F}_{14,14} =& 2(l_{14}\langle n_{14}\rangle + (l_{14})F_{14,14} + d_{13,14}F_{13,14}) \\
\dot{F}_{14,15} =& (l_{14} + l_{15})F_{14,15} + d_{13,14}F_{13,15} + d_{13,15}F_{13,14} \\
\\
\dot{F}_{15,15} =& 2(l_{15}\langle n_{15}\rangle + (l_{15})F_{15,15} + d_{13,15}F_{13,15})
\end{aligned}
\tag{A.3}
$$

A.4| Barcode Cluster

Barcodes forming cluster as described in chapter 8, exemplary barcodes from mouse #2 as seen in fig. 8.10.

- '1FC' 'GF3' '561' '1234987' '12ED7' 'EDCBA6987' 'G45' 'ED7' 'AFED7' '1458G69' '987' '145HG' 'IHGF3' '561D7' '12GF3' '34987' '12ED367' '16987' '1458G' 'GF349' 'GF345' 'CBA6987' 'C6987' 'AF7' '56987' '56149' '5612G89' '56129' '3456A' '34561' '1FCB7' '1F3' '14987' '145HGF3' '145FC' '1456987' '12C456987'
- '78965' '3HG' '5416789' '56789DC' '16547' '965' '541' '12965' '1278965' '789DC' '16789DC' '1478965' '145678C' 'CBA8965' '5678C' '3HGFE' '18965' '125678C' 'G8965' 'C65' '9DC' '54789' '547' '54169' '541678C' '54167' '5412IHG' '189DC' '16945' '1654CB9' '165' '12365' '123478965'
- '72189' '721' '723' '729' '725' '72369' '72569' '7214E' '3472189'
- '1HGDC' '7892A' 'G892A' '38IBA' '38A' '3892A' '3478A' '347892A'
- '1DE2369' '1DE29' '1DE23' '1DE2367'
- '7D9' '7D92A' '167D9' 'C67D9'
- 'C678I' '36I' 'EDC678I' 'C6I'
- '125HG' '78EBA'
- '9FE' '129FE'
- 'I61' 'I6147'
- '123FG' '3FG'
- '1458329' '14583'
- 'GHA69' 'GHA'
- '129D5' '9D5'
- '56GBA89' '56GBA'
- 'IDG' '38IDG'
- '145BC69' '145B9'
- '34IFE' '34IFGHE'
- '1278369' '127834569'
- 'EB9' 'EBA69'
- '12I4567' '12I45'

A.5| Abbreviations

HSC	hematopoietic stem cell
LT	Long-term hematopoietic stem cell
ST	Short-term hematopoietic stem cell
MPP	Multipotent progenitor
CMP	Common myeloid progenitor
CLP	Common lymphoid progenitor
GMP	Granulocyte/macrophage progenitors
MEP	Megakaryocyte/erythrocyte progenitors
Mono	Monocytes
Gr	Granulocytes
EryP	Erythrocyte progenitors
eq.	equation
P_{gen}	Probability of generation
FACS	Fluorescence-activated cell sorting
pacBio	Pacific Biosciences
fig.	figure

A.6| List of figures

A.7| List of tables

Barcode statistics

While this section does not provide any scientific insight, some barcode statistics are fun and interesting. Those will be highlighted here:
The total number of unique possible barcodes are 1.866.890, however due to the treatment methods used only a subset of 931.427 barcodes can practically be reached. This is due to the fact, that the highly complex barcodes need more recombination events than the number occuring in our experiments. Taking all mice into account, that have been studied in this thesis, we found a total number of 2146 unique barcodes. This is merely 0.23% of all reachable barcodes. This raises the question how many mice are needed to find all reachable barcodes with a 95% chance. And the answer to that is an astounding number of 2.6 x 10^{12} mice, where as the number of total mice in the world are estimated to around ~10^9. It is therefor safe to say, that we will never run out of unique barcodes to find using *Polylox* barcodes.
Hurray!

Total number of possible barcodes	1.866.890
Possible barcodes given treatment	931.427
Of which we have found	2.146 (0.23%)
Mice needed to find all possible barcodes (95% chance, given treatment)	2.6 x 10^{12}
Number of mice in the world	~10^9

Acknowledgments

I want to thank first and foremost my supervisor Thomas Höfer for granting me the opportunity for this PhD thesis and the trust he placed in me. He created an atmosphere of support not only in scientific matters, but also allowed enough independence to pursue my own research interest.

I also want to thank Hans-Reimer Rodewald for the deep biological discussions and a highly engaging collaboration. As a physicist, without them I would partly be lost in FACS gates and phenotypes.

Gratitude is also owed to my office mates Lisa and Nils for all the discussions and good times we had. I would also like to thank the whole group of Thomas Höfer. In this group support was always around the corner.

I would also like to thank Weike Pei, Thorsten Feyerabend, Kai Klapproth and all the other members of Hans-Reimer's lab which whom I had the pleasure to work with. This collaboration was coined by respect and mutual interest and made it a very enjoyable process.

Thanks also to Nils and Thomas for proofreading this thesis.

I want to thank Ulrich Schwarz and Frederik Graw for being part of my examination commision.

I also want to thank my former flatmates Kathrin and Sarah, for being part of my life for so long. Thanks for all the white wine and fun evenings. I will never forget our makeshift balcony.

Last but not least thanks to my family, and especially to my wife Katja, who had to endure all the wild rambles about my work, if everything went wrong, and were the first to celebrate when everything went right. Thanks for being there and picking me up.