Developmental hematopoietic stem cell variation explains clonal hematopoiesis later in life

Jesse Kreger[a], Jazlyn A. Mooney[a], Darryl Shibata[b], Adam L. MacLean[a,*]

[a]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, CA, USA

[b]Department of Pathology, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA

[*]Correspondence: macleana@usc.edu

Abstract

Clonal hematopoiesis becomes increasingly common with age, but its cause is enigmatic because driver mutations are often absent. Serial observations infer weak selection indicating variants are acquired much earlier in life with unexplained initial growth spurts. Here we use fluctuating CpG methylation as a lineage marker to track stem cell clonal dynamics of hematopoiesis. We show, via the shared prenatal circulation of monozygotic twins, that weak selection conferred by stem cell variation created before birth can reliably yield clonal hematopoiesis later in life. Theory indicates weak selection will lead to dominance given enough time and large enough population sizes. Human hematopoiesis satisfies both these conditions. Stochastic loss of weakly selected variants is naturally prevented by the expansion of stem cell lineages during development. The dominance of stem cell clones created before birth is supported by blood fluctuating CpG methylation patterns that exhibit low correlation between unrelated individuals but are highly correlated between many elderly monozygotic twins. Therefore, clonal hematopoiesis driven by weak selection in later life appears to reflect variation created before birth.

Introduction

Hematopoietic stem cells (HSCs) maintain the blood system throughout life[1]. As humans age, clonal expansions of HSCs are frequently observed[2–4]. Such clonal hematopoiesis (CH) is associated with increased risks of hematopoietic neoplasia and other diseases[5–9]. Whereas in some cases driver mutations can be found[10–12], often CH is lacking identifiable drivers that can explain their expansions. Consistent with the frequent lack of strong driver mutations, serial observations of CH are often compatible with weak selection because clone sizes are stable or expanding slowly over many years. Such weak selection is problematic to explain CH because most somatic cells that acquire such a weak selective advantage would be randomly lost and therefore never attain detectable frequencies.

Weak selection and slow expansion suggest that HSC subclones can arise early in life, potentially even before birth, with explained initial growth spurts[1,10,13–15]. Such early acquisition of a selective advantage eventually leading to CH would explain two open questions about CH without the need for identifiable driver mutations. First, following classical evolutionary dynamics theory and its extensions[16–24], variants with weak selective advantages will become dominant given relatively large populations sizes, population turnover, and enough time. These conditions are present for human hematopoiesis, which exhibits a large HSC population size[25], competition between stem cells, and many decades of life. HSC variants acquired very early in life maximize the time needed for their subclones to reach dominance. Second, variants with weak selective advantages are uniquely protected from random loss and extinction during development due to the natural expansion of all HSC subclones at this time of growth. The interval before birth maximizes the possible time and HSC expansion. Hence, HSCs with weak selective advantages that arise before birth are naturally protected from stochastic loss by growth and have enough time to eventually reach detectable sizes later in life.

Studies of hematopoiesis in twins offer a unique opportunity to test whether stem cell variation arising before birth can explain CH. Approximately two-thirds of monozygotic (MZ) twins have monochorionic placentas and share blood circulation in utero[26,27], and therefore share HSC variation at birth. If HSC clonal variation is acquired after birth, blood clonal compositions will differ between aged MZ twins. If, on the other hand, aged MZ twins share the same clonal compositions, i.e. the same clones have grown towards fixation over decades[2,14,28], then the variation most likely arose before birth. Selection for variation before birth is supported by MZ twin studies of X-chromosome inactivation, which found a blood ratio skewing with aging, with concordance for either maternal or paternal X-chromosome dominance[29,30]. However, CH driver mutations are usually not concordant between MZ twins, indicating that driver mutations arise after birth and that MZ twins do not share a predisposition for specific mutations.

We hypothesize that HSC variation inevitably arises during development leading to subtle selective differences that eventually lead to CH later in life. This hypothesis is testable with MZ twins and lineage markers that become polymorphic before birth. Somatic mutations could be used to trace prenatal HSC subclones, but relatively few mutations occur in the brief time before birth and these prenatal mutations would be difficult to distinguish from postnatal mutations. An alternative to somatic mutations is DNA methylation that rapidly fluctuates between methylated and unmethylated states at certain CpG sites[31]. The higher rates of methylation fluctuation allow HSCs to become polymorphic before birth, and yet allow lineage tracing through life. Tracking fluctuating methylation sites is a convenient lineage marker due to its much higher error rate and the broad availability of suitable blood methylation datasets.

Here we show that fluctuating methylation patterns are often correlated in the blood between elderly MZ twins, relative to dizygotic twins or unrelated individuals. We develop a single-cell model of HSC clonal dynamics to study the origins of HSC clonal diversity, and show that variants arising before birth conferring weak selective survival advantages commonly become dominant later in life. Hence, the variation between many blood subclones that are weakly selected to grow to large sizes later in life was likely created before birth.

Results

Analysis of twins via fluctuating CpG methylation reveals patterns of HSC clonal dynamics over lifetime

To study HSC clonal dynamics over a human lifetime (Fig. 1A-B), we studied publicly available DNA methylation datasets (Table 1). Methylation profiles were extracted for the population of HSCs sampled for each individual in a dataset. As previously described[31], we selected CpG sites that have a high degree of intra-individual heterogeneity (indicating that they flip-flop) and which are unlikely to be under active regulation[31]. Full details of data and data processing steps to construct methylation profiles per time point per individual are given in Methods and Supplementary Information Section S1.

We leveraged the shared prenatal circulation of monozygotic (MZ) twins to investigate the impact of stem cell variation arising during development on lifetime clonal dynamics of HSCs. We extracted methylation profiles for three groups of individuals: MZ twins, DZ twins, and unrelated individuals. The data range from time points near birth to 86 years of age. To compare average methylation profiles between individuals, we used the Pearson correlation coefficient (R) to assess similarity between pairs of individuals. Values of R ranged from close to +1 (perfect positive correlation) to close to 0 (uncorrelated). For each set of pairs (MZ, DZ, unrelated), we calculated a line of best fit over time, as well as mean correlation values for each dataset (Fig. 1C-F).

For MZ twins (Fig. 1C, F), considerable variability was present in the correlation coefficients at birth, with some twin pairs showing much higher values than others. Variability exists in MZ twin development, whereby cleavage before formation of the blastocyst leads to dichorionic/diamniotic twins but cleavage after blastocyst formation can lead to either monochorionic or monochorionic/monoamniotic twins[32]. An estimated $\frac{2}{3}$ of MZ twins share circulation during development[33]: MZ twins with higher initial Pearson coefficients thus likely represent twins who shared circulation during development and the emergence of HSCs; MZ twins with lower initial Pearson coefficients (closer to DZ/unrelated pairs) likely characterize twins that split earlier and were thus offered less/no opportunity for developmental hematopoiesis. The mean Pearson coefficients at birth were approximately 0.87 for MZ twins (Fig. 1C), 0.72 for DZ twins (Fig. 1D), and 0.61 for unrelated individuals (Fig. 1E).

Throughout the human lifespan, similar trends were observed for each group of individuals, whereby the clonal relatedness declines slowly with age. The slopes of decline differ between groups, with the sharpest decline observed for DZ twins (slope of -0.0033), and more modest declines observed for unrelated individuals (slope of -0.0025) and MZ twins (slope of -0.0020). The variability at each time point was largest for MZ twins. This resulted in wide range of HSC clonal relatedness in MZ twins in later life, with Pearson coefficients ranging from low (~ 0.3) to very high (> 0.9). We also noted a change in the dynamics in later life: the Pearson coefficients decrease more sharply after 65 years of age. This is consistent with previous studies that show a significant loss of clonal diversity in HSCs at this age[10].

A single-cell resolved model of clonal hematopoiesis describes the somatic evolution of stem cells throughout life

To study the origins of clonal hematopoiesis and follow stem cell dynamics throughout life, we developed a mathematical model of methylation dynamics in HSCs, at the resolution of individual fluctuating CpG methylation (fCpG) sites in single cells (Fig. 2A; Methods). The clonal evolution of the HSC population as quantified by fluctuating methylation clocks (FMCs) was modeled from embryonic development through birth and up to a 100-year human lifespan.

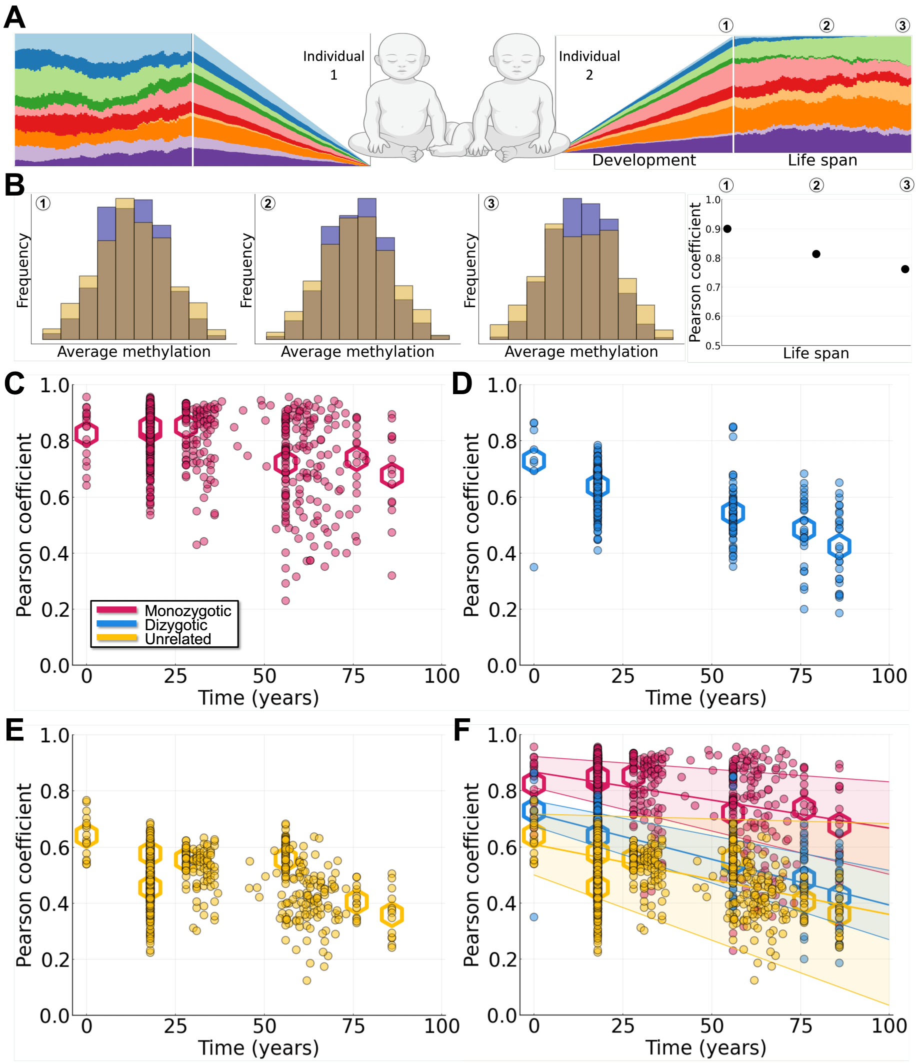

Figure 1: **Clonal dynamics of HSCs are characterized by methylation profiles over lifetime.** **A**. HSC population dynamics over time, from embryo development through human lifespan. **B**. β distributions of two individuals over time, from embryo development through human lifespan. Pearson correlation coefficients are also included (right panel). **C**. Pearson correlation coefficients for average methylation profiles between MZ twins. Dots represent individual comparisons and hexagons represent means of datasets. **D**. Pearson correlation coefficients between DZ twins. **E**. Pearson correlation coefficients between unrelated individuals. **F**. Pearson correlation coefficients combined for the three sets of comparisons. Lines of best fit with corresponding 90% confidence intervals are also shown.

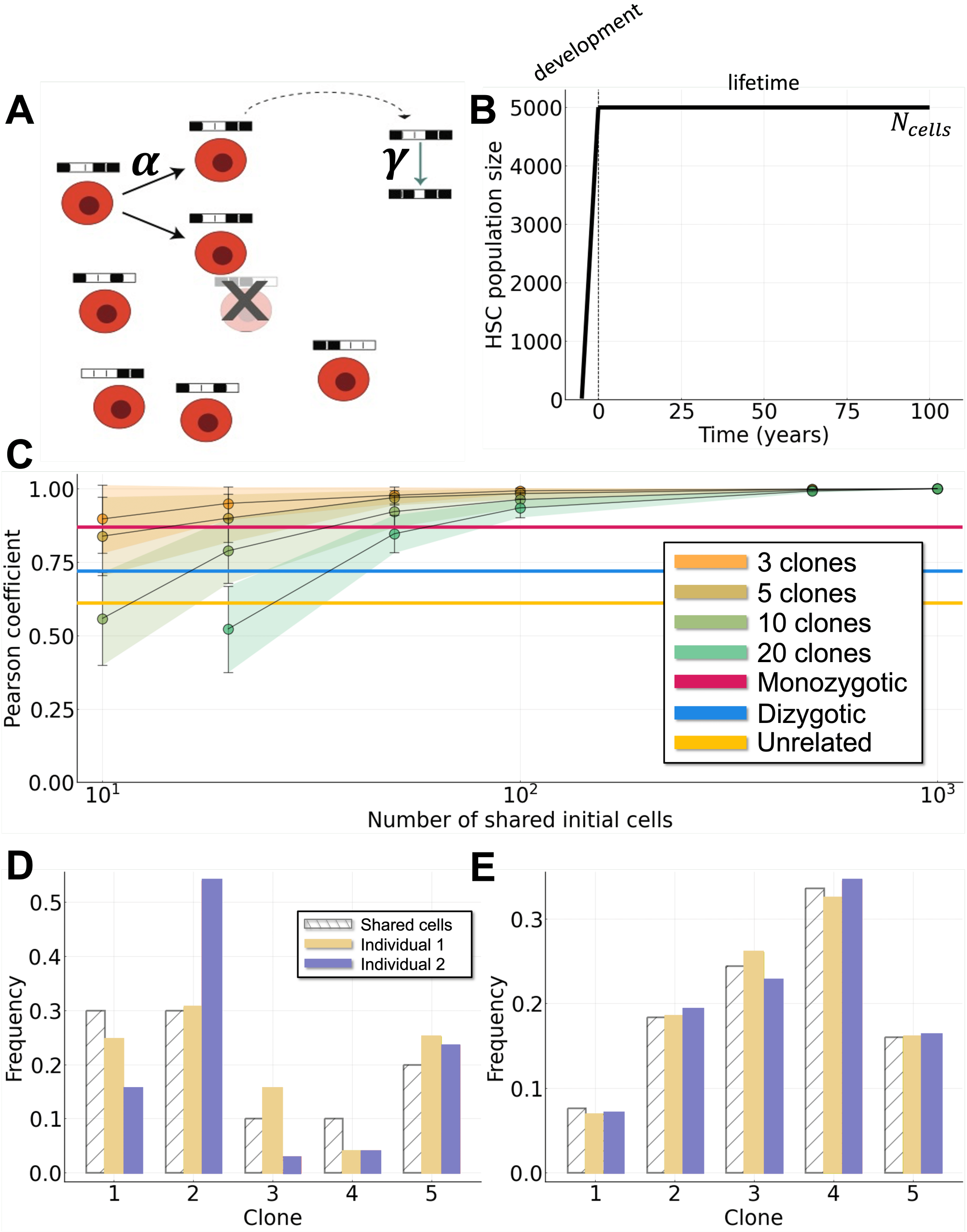

Figure 2: **Variation during development is needed in order for model to match initial Pearson coefficients at birth.** **A**. Schematic diagram of the mathematical model, with probability of cell replacement α and probability of change in methylation γ. **B**. HSC population dynamics over a lifetime. **C**. Parameter estimation using initial Pearson coefficients of twins and unrelated individuals. Given a choice for the initial number of clones, the number of cells at which the embryo splits (N_{split}) can be determined. Data points (dots) represent the mean Pearson coefficient for 10^3 simulations with a given number of clones (N_{clones}) and number of shared cells (N_{split}). Shaded areas denote one standard deviation from the mean. The horizontal lines represent the data: initial Pearson coefficients at birth for MZ twins (red), DZ twins (blue) and unrelated individuals (yellow). **D-E**: Effect of varying N_{split}. White striped bars represent the clonal distribution in the shared embryo when it reaches N_{split} cells and splits into two embryos. Yellow/blue bars represent individuals 1 or 2 at the point at which the embryo has reached N_{cells} cells and finished growing. D: $N_{\text{split}} = 10$. E: $N_{\text{split}} = 500$. All other parameter values can be found in Table 2.

At birth, a population of N_{cell} stem cells represents the entire HSC pool, and we assume a constant population size throughout life (Fig. 2B). Variation in the population size during life does occur, e.g. the HSC expands during aging, however these changes are small relative to both the growth phase during development and the total population size after birth. Each cell is modeled by N_{site} fCpG sites (Fig. 2A), where at each site, given two alleles, there are three possible methylation states: 0 (unmethylated), 0.5 (hemimethylated) and 1 (fully methylated). A population of cells is defined by $\mathbf{x}$, $\mathbf{x} = \left[x_i^j\right]$, where x_i^j represents the methylation status of the i^{th} cell at the j^{th} fCpG site, i.e. we have $x_i^j \in \{0, 1, 2\}$, for $i \in \{1, 2, \ldots N_{\text{cells}}\}$ and $j \in \{1, 2, \ldots N_{\text{sites}}\}$. We model stem cell methylation dynamics throughout life by a Markov process. Dynamics occur by cell replacement (with probability α) and changes in fCpG methylation. During a cell replacement event, each fCpG site can flip with probability γ, and a methylation step change of ±1. See Table 2 for full description of model parameters along with values used for simulation.

In development, hematopoiesis was modeled starting from a small initial number of HSCs. This population of initial stem cells (of size N_{clone}) each seeds a clone that grows during development until the population reaches its final size of N_{cell} cells (Fig. 2B). Each fCpG site in each clone is initialized randomly in state either 0 (unmethylated) or 1 (fully methylated). To model selection in development and throughout life, we define that each of the N_{clone} clones has a fitness coefficient, $1 + s_i$, for $i \in \{1, 2, \ldots, N_{\text{clone}}\}$, with s_i chosen from a Gamma distribution (see Supplementary Table **??**).

The model (Fig. 2A) is developed to follow single-site single-cell methylation dynamics. At the population level, it is similar to previous FMCs models that allow for analysis of average methylation profiles in a population of cells[10,13,31]. In particular, for increasing number of clones and/or increasing rate of flipping, we see the average methylation profile switch from a multi-modal distribution to a unimodal distribution centered around 0.5[31]. For more details on population level similarities to previous models see Supplementary Information Section S2.

We characterized the effects of each model event: cell replacement (with rate α; Fig. S1) and (de)methylation (with rate γ; Fig. S2) by exploring the parameter space of these parameters (Section S3 in the Supplementary Information). We also analyzed the impact of population-level model parameters on the lifetime stem cell dynamics, namely by varying the threshold number of cells at which embryos split, N_{split} (Fig. 2C-E), or by varying the total number of cell clones, N_{clones} (see Section S4 in the Supplementary Information). As the total number of cell clones is reduced, so is the total observed variability in the clonal distributions, resulting in overall higher Pearson coefficients between individuals (Fig. S3).

Stem cell variation arising during development explains the blood clonal composition at birth

During development, a nascent population of HSCs must grow large enough to entirely support hematopoiesis by birth. We model a population of N_{clone} cells that grows to a size of N_{cell} cells, by either uniform or frequency-dependent growth[15,34] (Fig 1A). The variation observed between pairs of twins/unrelated individuals at birth (Fig. 1F) amounts to Pearson coefficients observed ranging from 0.5 to 1.0. In the case of MZ twins, we initialize a shared set of clones and uniform growth of these clones was not able to explain the data (Fig. 3A). The methylation distributions observed under the uniform growth model were highly similar with Pearson coefficients ≥ 0.95 (Fig. 3A). Frequency-dependent growth, in contrast, permitted greater variation in methylation profiles between individuals at birth, due to more divergent HSC clonal dynamics in development (Fig. 3B). While the frequency-dependent growth model for developmental hematopoiesis is simplistic with regards to the evolutionary dynamics, it provides parsimonious means with which to generate diverse clonal distributions as observed in the data for pairs of individuals at birth.

Under the frequency-dependent growth model, we model differences in development between twins and unrelated individuals as follows. In the case of twins, we model the embryo growing from

N_{clone} cells to N_{split} cells with identical clonal dynamics, at which point the embryos split and two independent embryos are simulated where, in each, the HSC population grows from N_{split} cells to its full size of N_{cells} cells. The assumptions required for this model are that: 1. MZ twins that develop from a single embryo can directly share early hematopoietic cells for a varying length of time depending on their developmental state, ranging from dichorionic/diamniotic to monochorionic/monoamniotic, with a corresponding variation in the extent to which early hematopoietic cells may be shared; 2. due to genetic similarity, DZ twins share some clonal growth characteristics during early development; 3. unrelated individuals that share no developmental history are seeded by HSC populations that each grow independently. The data characterizing relatedness of HSCs at birth between individuals (time point 0 in Fig. 1F) are consistent with this model.

To determine values for N_{split} for DZ and MZ twins, we fit the model parameters (N_{clone}, N_{split}) to FMC data characterizing twins and unrelated individuals (Fig. 2C). We fit curves to characterize the observed Pearson coefficients as we vary the number of shared cells, and find that, assuming $N_{\text{clone}} = 10$, the following values of embryo splitting are consistent with the data: for MZ twins $N_{\text{split}} \approx 36$; for DZ twins $N_{\text{split}} \approx 15$; and for unrelated individuals $N_{\text{split}} = 10$ (i.e. no shared development). For further details, see Supplementary Information Section S5 and Figs. S4 and S5.

HSC clonal dynamics in monozygotic twins are constrained by weak selection

In order to explore the effects of selection on clonal dynamics of HSCs, we simulated FMC dynamics in HSCs over a human lifetime under different models for selection. We considered three models: no selection, weak selection, and strong selection. In Fig. 4A-C, simulations of the lifetime HSC dynamics for MZ twins are shown, under each of three different selection models. In the case of no selection, where all clones have the same fitness post-birth (clonal dynamics during development are subject to frequency-dependent selection), Pearson coefficients later in life are too low, as clonal distributions diverge between individuals (Fig. 4A for a representative simulation; and Fig. 4D, for trajectories of 100 pairs of individuals). In this regime characterized by drift, many twin pair methylation profiles become under-correlated over time relative to the data.

For non-neutral models of lifetime HSC dynamics, stem cell fitness values were drawn from a Gamma distribution parameterized by shape a and size θ (see Methods for details). For weak selection $a = 0.05$ and $\theta = 0.01$ (see Table **??**). Variants with weak fitness advantages arising in development can experience clonal expansions due to the developmental growth phase (Fig. 4B and E). Even if the variant frequency is low at birth, the long time range of a human lifespan enables the effects of weak selection to become evident in later life. In simulations, we observed that many MZ twin pairs remain correlated under weak selection throughout life, but without individual clones fixating. We observe that weak selection results in good agreement between our model simulations and the MZ twin data. In particular, through directly comparing model simulations to the data (see the last section in Methods and Table S2 in the Supplementary Information) we find that weak selection results in a better fit to the data compared to neutral selection, as measured by the mean distance from each model simulation to simulated data trajectories.

For strong selection, clone fitness values were sampled from a Gamma distribution with $a = 0.05$ and $\theta = 0.05$ (see Table **??**). In this case we observed frequent, dramatic reductions in clonal diversity whereby one clone with relatively high fitness would rapidly expand and fixate in the HSC population (Fig. 4C and 4F). This scenario mimics the possible endpoint following from trajectories of clonal hematopoiesis of indeterminate potential (CHIP)/hematopoietic malignancies in the sense that all clonal diversity has been lost. When the same clone nears/reaches fixation in two people due to its fitness, the Pearson coefficient between individuals approaches 1. In this scenario, the methylation distribution in both individuals becomes bimodal with most density near 0 and 1. The bimodal distribution (rather than trimodal or "W-shaped" as in[31]) observed in each twin when a clone fixates (Fig. 4C) results from the large population size of HSCs (10^4 cells) and the relatively low CpG flipping rate.

We also observed infrequently scenarios where two or more clones have relatively high fitness,

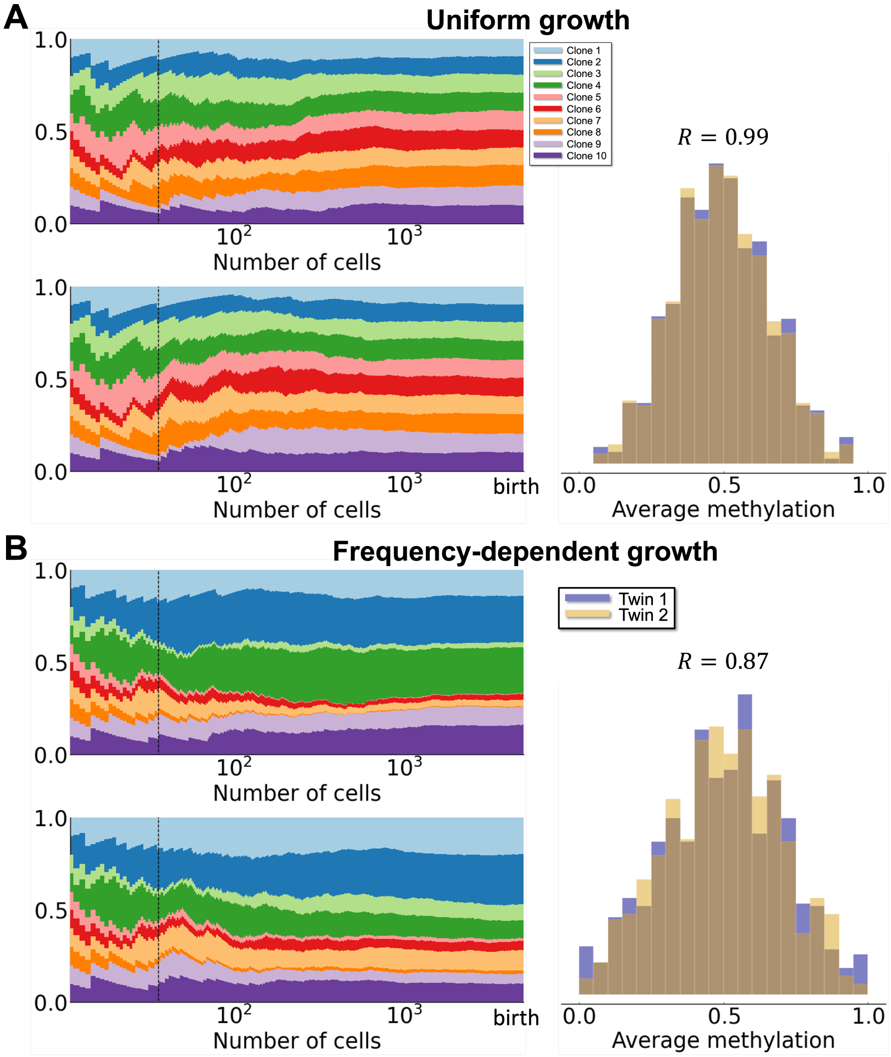

Figure 3: **Frequency-dependent growth produces clonal variation during development. A**. Uniform growth during development with $N_{\text{split}} = 36$ and no selection ($s_i = 0$) for all clones. Left panel: Clonal dynamics during development, colors denote different clones. Right panel: β distributions at the end of development, i.e. birth. The initial Pearson correlation coefficient at birth is 0.99. **B**. Same as A for frequency-dependent growth model during development. The initial Pearson correlation coefficient at birth is 0.87. All other parameter values can be found in Table 2.

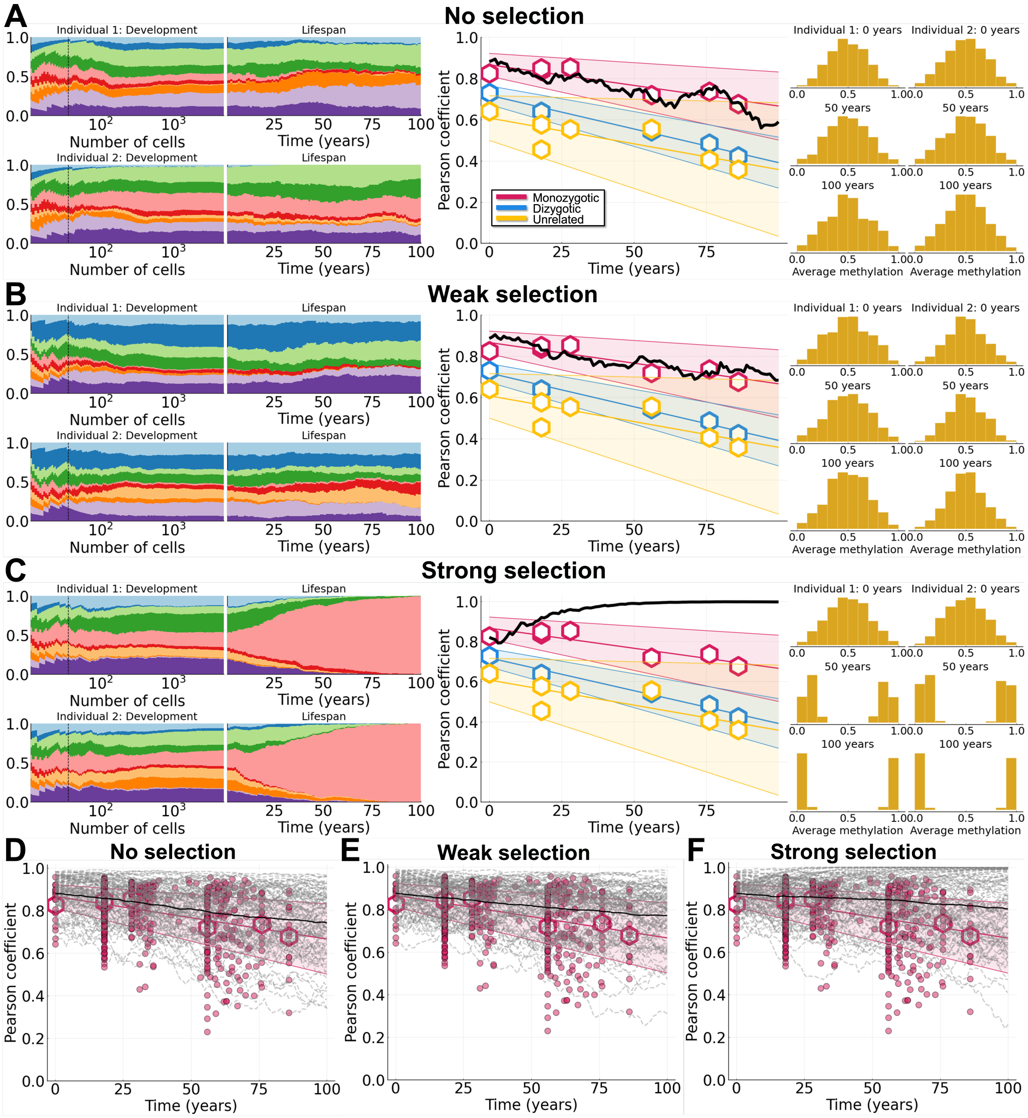

Figure 4: **Variants with weak selection arise during development and explain FMC dynamics for monozygotic twins.** Simulations shown for MZ twins ($N_{\text{split}} = 36$) with variation (frequency-dependent growth) during development. For plots showing Pearson correlation data (middle column and bottom row), dots represent individual comparisons, hexagons represent means of datasets, and shaded bands represent 90% confidence intervals. **A-C**: Clone growth frequency plots for both individuals during development and life (dashed vertical line represents N_{split}), Pearson correlation coefficient, and β distributions at 0, 50, and 100 years of life. A: No selection. B: Weak selection ($a = 0.05$ and $\theta = 0.01$). C: Strong selection ($a = 0.05$ and $\theta = 0.05$). **D-f**: Results from 10^2 simulations are shown (dashed lines are individual simulations and solid lines are mean trajectories). D: No selection. E: Weak selection ($a = 0.05$ and $\theta = 0.01$). F: Strong selection ($a = 0.05$ and $\theta = 0.05$). All other parameter values can be found in Table 2.

and then due to differences arising from frequency-dependent growth during development different clones will tend towards fixation in the two twins. This results in a Pearson coefficient near or trending toward 0 (Fig. 4F). Moreover, we observed in the case of strong selection that the clonal distributions at birth had little effect on the lifetime clonal dynamics, as long as the fittest clones were not lost due to random cell loss. This is due to the relative fitness advantages of these clones and the length of the possible human lifetime: even if the fittest clone starts with few initial cells at birth relative others, over many decades it will grow to dominate, even in the large size of the HSC pool (the classical probability of fixation in a large population for a single cell with fitness advantage s is given by $2s$[16]). The timing of when variants arise under strong selection does not greatly affect HSC clonal dynamics in later life since a loss of clonal diversity is effectively ensured. Under this model, the timing of when variants arise would affect only the time at which the clone takes over (earlier generation/faster initial growth resulting in earlier fixation).

Different lifetime clonal dynamics of twins vs unrelated individuals can be explained with no tunable parameters other than n_{split}

In the previous section we studied the clonal dynamics of HSCs over lifetime in MZ twins, under different models of selection. We analyzed simulations of DZ twins and unrelated individuals in a similar manner and found evidence that in both cases strong selection is not consistent with the data (see Supplementary Information Section S6 and Figs. S6-S8). In the case of DZ twins/unrelated individuals, lifetime FMC dynamics may be consistent with either neutral dynamics or weak selection (Table S2), although as we have seen neutral dynamics are not entirely consistent with clonal dynamics in MZ twins.

We simulated 100 pairs of individuals from each group (MZ twins, DZ twins, and unrelated individuals) over a human lifetime under the assumptions of 1) frequency-dependent growth during development and 2) weak selection (Fig. 5). All parameters of the model were held constant over all pairs of individuals except for N_{split}, which is determined from twin status: $N_{\text{split}} = 36$ for MZ twins (Fig. 5A); $N_{\text{split}} = 15$ for DZ twins (Fig. 5B); and $N_{\text{split}} = 10$ for unrelated individuals (Fig. 5C). We observed that the balance between clonal diversity arising from frequency-dependent growth and weak selection throughout life gave rise to more similar (MZ) or divergent (unrelated) methylation distributions in later life (Fig. 5A-C, right hand column). Various methylation distributions can be observed in cases with reduced clonal diversity, leading to multimodal distributions with three models if two clones dominate (Fig. 5A; individual 2 at 100 years), or four modes if three clones dominate (Fig. 5B; individual 2 at 100 years). For all of the total of 100 simulations for each twin status (Fig. 5D-F, dashed lines), we observed a good fit between model and data in all of the three twin cases modeled.

In summary, when variants subject to weak selection arise during development, they oftentimes experience clonal expansion because of growth, which mitigates against stochastic loss. Then, post-birth, over the relatively long possible human lifespan, the effects of the weak fitness advantages of particular clones can become evident. By studying the lifetime dynamics of HSCs in twins and unrelated individuals, we show how variation is most likely generated before birth yet only becomes evident in later life.

Discussion

Clonal hematopoiesis becomes increasingly detectable with aging. In many cases, driver mutations are absent and inferred weak selection indicate variants arise much earlier in life with unexplained initial growth spurts[10]. Here, we have shown that a simple model of HSC clonal dynamics with weak fitness advantages can explain lifetime hematopoiesis when variants arise before birth.

Publicly available twin methylation data offer a unique opportunity to study when hematopoietic variants arise. Through theory and simulation, we have shown that weakly selective variants

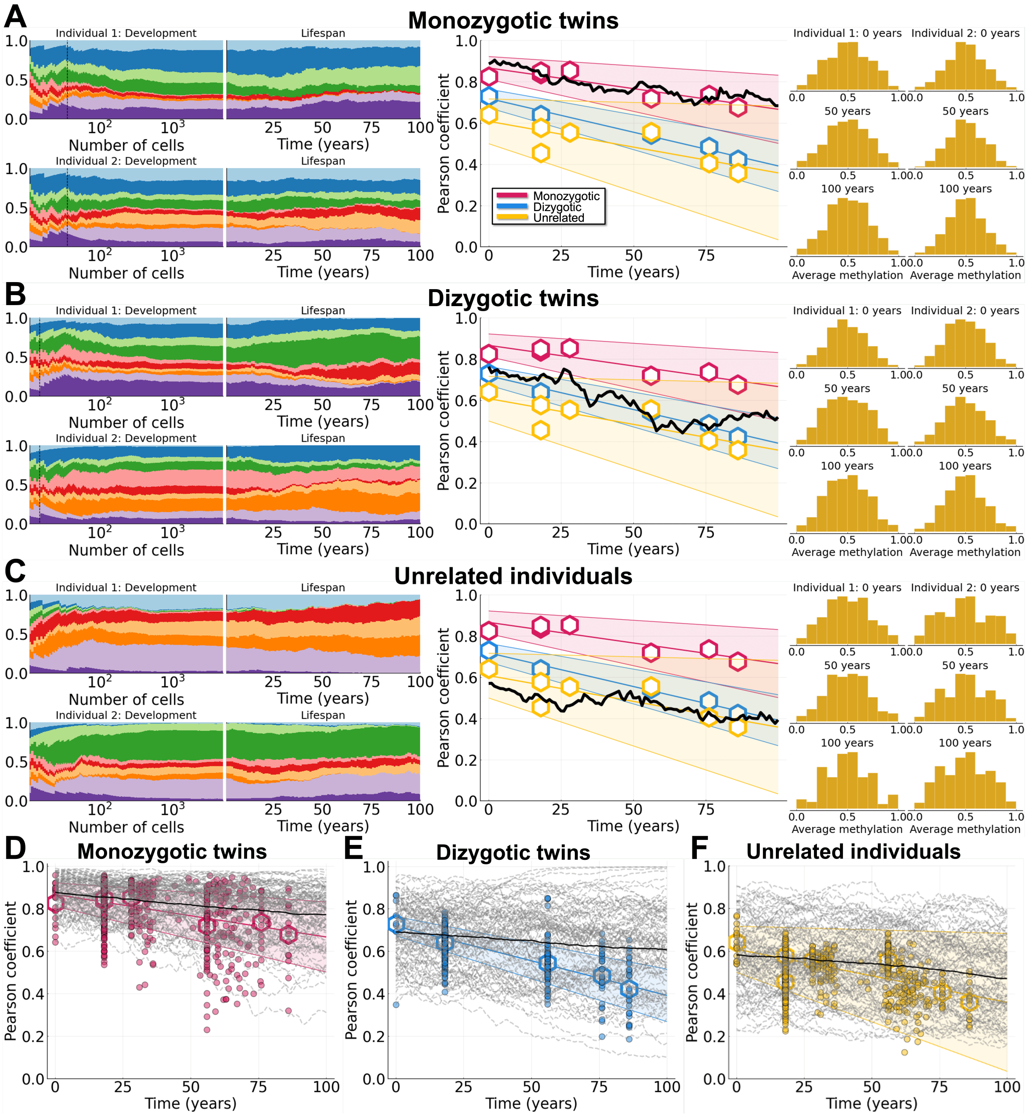

Figure 5: **Weak selection with variants arising in development explains lifetime dynamics of twins and unrelated individuals.** Simulations shown for weak selection ($a = 0.05$ and $\theta = 0.01$) with variation (frequency-dependent growth) during development. For plots showing Pearson correlation data (middle column and bottom row), dots represent individual comparisons, hexagons represent means of datasets, and shaded bands represent 90% confidence intervals. **A-C**: Clone growth frequency plots for both individuals during development and life (dashed vertical line represents N_{split}), Pearson correlation coefficient, and β distributions at 0, 50, and 100 years of life. A: MZ twins, $N_{\text{split}} = 36$. B: DZ twins, $N_{\text{split}} = 15$ C: Unrelated individuals, $N_{\text{split}} = 10$ **D-F**: Results from 10^2 simulations are shown (dashed lines are individual simulations and solid lines are mean trajectories). D: MZ twins. E: DZ twins. F: Unrelated individuals. All other parameter values can be found in Table 2.

present in embryos can undergo expansion as a byproduct of developmental growth, which mitigates against random loss, and can permit clones to tend towards fixation much later in life after a period of latency. Analysis of the clonal dynamics of twins vs unrelated individuals showed that some HSC variation is required to occur before birth. Given the assumption that some HSC clonal variation is present by birth, we are able to explain the different lifetime trajectories of pairs of unrelated vs twinned individuals using a single parameter that characterizes the developmental growth phase of hematopoiesis. With no further fitting, this model is consistent with the increased frequency of clonal expansions observed in individuals over the age of 65, without the need for any driver mutations to arise later in life.

The fate of genetic variants within a population has a rich theoretical foundation[16–18,35], although in cases of time-varying population sizes, these inferences are more nuanced[22,23,36]. A growing body of literature has revealed in depth the clonal diversity and clonal dynamics of HSCs over life[4,10,11]. Evidence from these studies suggests that HSC variation is likely to arise early, perhaps even before birth. However, sampling hematopoiesis before birth for longitudinal study is impractical. By overcoming this challenge using twins data from which we can estimate variation occurring before vs after birth, we are able to show that indeed the developmental phase of hematopoiesis offers a window of opportunity for genetic variants to persist even when they have only very slight fitness advantages. In contrast, previous models assumed higher selection coefficients to explain clonal distributions in individuals ≥ 65 years old[10], which were necessary if these variants arose in mid life, and not earlier. In our model, low frequency non-driver mutations are "passengers" that hitchhike with the selected prenatal HSC subclones.

Genetic heterogeneity of HSCs is but one factor influencing hematopoiesis as we age, since HSCs are also a product of their niche, and the stem cell niche/microenvironment plays a crucial role in defining stem cell function[37]. Additional non-genetic factors influencing hematopoiesis include stem cell heterogeneity (mediated by transcriptional or epigenetic variation)[38–41], feedbacks from and contributions of progenitor cells to hematopoiesis[42] and other environmental signals, e.g. resulting from diet or the immune system[43,44]. Future models building on this current work ought to consider some of these factors and their influence on healthy hematopoiesis in development and throughout life.

Other limitations of the current model include scaled population sizes (HSC pool approx. 1/6 of the estimated total, to ease computational cost), the assumption of constant selection throughout life, and a simplified model of embryonic hematopoietic development. Future work could include building further biological details into the framework of the mathematical model in order to analyze their effects and increase the predictive power of the model.

In summary, we have shown that HSC variants created before birth can determine HSC clonality much later in life. Fluctuating CpG methylation provides a lineage marker sufficient to reveal clonal variation in development and early life. By modeling this variation, we have seen that weak selection conferred by the stem cell variation that arose during development yields clonal hematopoiesis decades later. Although CH is associated with increased disease risks, HSC variants created before birth combined with decades of selection within a large HSC pool can help explain why CH is so common with normal aging. Further more detailed studies of MZ twin blood populations, especially elderly MZ twins with documented monochorionic or dichorionic placentas, can help further explore the roles of early HSC variation on normal aging and disease predisposition.

Methods

Data and fCpG site filtering

Methylation datasets are publicly available on the Gene Expression Omnibus (GEO)[45]. For this study, we use data listed in Table 1, which represent a collection of DNA methylation profiles from both monozygotic (MZ) and dizygotic (DZ) twins. We extract average methylation profiles from

GEO ID	Reference	Participant age (years)	Num. monozygotic twin pairs	Num. dizygotic twin pairs
GSE36642	[47]	Prenatal (gestational week 32-38)	17	8
GSE154566	[48]	18	116	-
GSE105018	[49]	18	426	306
GSE43975	[50]	28 (average)	39	-
GSE61496	[51]	30-74	142	-
GSE89093	[52]	38-79	46	-
GSE100227	[53]	56 (average)	65	66
GSE73115	[54]	76 (average)	28	52
GSE73115	[54]	86 (average)	28	52

Table 1: **Description of datasets.** The first column is the dataset GEO ID, the second column is the reference, the third column is the age of the individuals, the fourth column is the number of MZ twin pairs, and the fifth column is the number of DZ twin pairs.

groups of MZ and DZ twins at different points throughout the human lifespan (Fig. 1). Unrelated individuals comparisons are made between twins that are not related. Similar to[31], we study the dynamics of fCpG sites that are "neutral", i.e. CpG loci that are not actively regulated and that show a high degree of intraindividaul heterogeneity (see Supplementary Information Section S1 for further details). To compare the average methylation profiles between individuals we use the Pearson correlation coefficient. For each category (MZ twins/DZ twins/unrelated individuals) we calculate the line of best fit for the data as well as the 90% confidence interval using the LsqFit.jl Julia curve fitting package[46].

All raw data (DNA methylation profiles) used in this study are publicly available on the Gene Expression Omnibus (GEO) (see above for GSE numbers and citations). All processed data (Pearson correlation coefficients between twin/unrelated individuals at different ages) is publicly available here: github.com/maclean-lab/scFMC-model. Data used in this study can also be seen in Fig. 1C-F. A full list of "neutral" fCpG sites used in this study can also be found at our publicly available github repository: github.com/maclean-lab/scFMC-model, see also Supplementary Information Section S1.

Correlations between individuals

In order to compare the correlation between average methylation profiles at different ages we use the Pearson correlation coefficient[55], which is given by

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}, \tag{1}$$

where x_i (y_i) is the average methylation value of individual 1 (2) at the i^{th} fCpG site and $\bar{x}$ ($\bar{y}$) is the mean value for individual 1 (2) over all fCpG sites. The Pearson coefficient measures the similarity between the two average methylation profiles, where a Pearson coefficient of 1 represents a positive correlation, a Pearson coefficient of -1 represents a negative correlation, and a Pearson coefficient of 0 represents no relationship.

Model description: modeling HSC dynamics postnatally

We develop a theoretical model of methylation dynamics in a population of cells, similar to[31]. We model dynamics at the resolution of individual fCpG sites in individual hematopoietic stem cells (HSCs). We assume there are N_{cells} cells, each with N_{sites} fCpG sites. Since there are two alleles at each site, there are three possible methylation states at each site

- 0 (which represents 0% methlyation),
- 1 (which represents 50% methlyation),
- 2 (which represents 100% methlyation).

This can be seen in Fig. 2A (for a similar representation see Fig. 1 in[31]).

We represent $\mathbf{x}$ as our population of cells, where

$$\mathbf{x} = \begin{bmatrix} x_1^1 & x_2^1 & \dots & x_i^1 & \dots & x_{N_{\text{cells}}}^1 \\ x_1^2 & x_2^2 & \dots & x_i^2 & \dots & x_{N_{\text{cells}}}^2 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_1^j & x_2^j & \dots & x_i^j & \dots & x_{N_{\text{cells}}}^j \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_1^{N_{\text{sites}}} & x_2^{N_{\text{sites}}} & \dots & x_i^{N_{\text{sites}}} & \dots & x_{N_{\text{cells}}}^{N_{\text{sites}}} \end{bmatrix},$$

and each $x_i^j \in \{0, 1, 2\}$ represents the methylation status of the i^{th} cell at the j^{th} fCpG site for $i \in \{1, 2, \dots N_{\text{cells}}\}$ and $j \in \{1, 2, \dots N_{\text{CpG}}\}$. We develop a discrete Markovian model where each discrete step represents one day, and we simulate the model up to 100 years (assuming 365 days a year). For post-birth dynamics (embryo development described in the following paragraph), we assume a constant population of cells. Dynamics occur with cell replacement (a birth/death event), which happens for each cell with probability α each discrete step. We assume the cell chosen to be replaced (die) is chosen uniformly from the population of cells, and the cell chosen to reproduce is chosen based on fitness (described below). In this way, a constant population size of HSCs is maintained from birth until 100 years of age. Furthermore, during each replacement event, each fCpG site in the new cell has a γ chance to flip, with maximum methylation step ±1 (Fig. 2A).

Model description: modeling HSC dynamics during development

To model development (before birth), we assume there are N_{clones} cell clones and we start initially with one cell of each clone. Each fCpG site for each clone is initially randomly either 0 (homozygously unmethylated) or 2 (homozygously methylated). This is because during the early stages of embryogenesis, the inherited methylation patterns from parental gametes are largely erased before a large wave of de novo methylation remodels the entire genome, resulting in a bimodal methylation distribution[31,56]. To model selection, we assume that each of the N_{clones} clones has a fitness coefficient, $1 + s_i$, for $i \in \{1, 2, \dots, N_{\text{clones}}\}$, where each s_i is chosen from a Gamma distribution[57] with shape parameter a and scale parameter θ[58–60]. For MZ/DZ twins the clonal fitness coefficients are the same for both individuals across individual comparisons whereas for unrelated individuals the clonal fitness coefficients are different (but chosen from the same distribution) for both individuals across individual comparisons. (De)methylation at fCpG sites will change the methylation profile of cells within a clonal lineage, but we assume that this has a negligible effect on fitness. We note that MZ twin dynamics over lifetime (as characterized by methylation profiles) match the case of no selection as the fitness differences go to zero, i.e. as clones have fitness coefficients $s_i = 0$ $\forall i = 1, 2, \dots, N_{\text{clones}}$).

$$\lim_{a \to 0 \text{ or } \theta \to 0} \text{clonal distributions} = \text{clonal distributions}_{s_i = c, \text{ for all } i}. \tag{2}$$

We allow the N_{clones} cells to then grow to N_{cells} cells (see Fig. 2B) in either of two different growth scenarios-

- **No/little variation during development**: uniform growth, where each clone has an equal chance throughout developmental growth to reproduce. Here, each embryo will have frequencies of approximately $\frac{1}{N_{\text{clones}}}$ for each of the N_{clones} clones.

Notation	Description	Value	Units	Reference
N_{sites}	number of fCpG sites per cell	10^3	-	-
N_{cells}	number of cells	5×10^3	cells	-
N_{clones}	number of cell clones	10, varies	cells	-
N_{split}	cell threshold at which the embryos split	$10,\ 15,\ 36$	cells	-
α	cell replacement (birth/death) rate	$\frac{1}{365}$, varies	days^{-1}	10 24
γ	(de)methylation rate	10^{-3}, varies	per fCpG site per replacement event	31 63 64
$1 + s_i$	fitness coefficient for i^{th} clone	Table **??**	-	13
a	gamma distribution shape parameter	0.05	-	58 59 60
θ	gamma distribution scale parameter	$0.01,\ 0.05$	-	58 59 60

Table 2: **Description of model parameters and values.** Estimated from the literature, see in particular[10,31]. The first column is the parameter notation, the second column is the parameter description, the third column is the parameter estimated value, the fourth column is the parameter units (if applicable), and the fifth column is the citation of the reference for the parameter estimate.

- **Variation during development**: frequency-dependent growth, where clones that grow to larger numbers during early development are more likely to reach higher frequencies in the embryo. Here, we also assume growth which is weighted by each clone's fitness coefficient $(1 + s_i)$. This increases the variation in the clonal frequencies during development and at birth.

In the case of MZ and DZ twins, which share a single embryo early during development (and blood circulation in the case of MZ twins), we allow a single embryo to grow from N_{clones} cells to N_{split} cells before splitting into two embryos. These two embryos then grow independently to the full N_{cells} cells. The value of N_{split} is determined based on simulating embryo development and finding the best fit to Pearson coefficients at birth from twins from methylation data (Fig. 2C). Larger values of N_{split} result in less variability between individuals (Fig. 2D-E). For unrelated individuals, we assume that $N_{\text{split}} = N_{\text{clones}}$ (i.e. there is no shared embryo growth).

All models were developed in the Julia programming language[61,62]. All code developed for this study is available at a public github repository located here: github.com/maclean-lab/scFMC-model.

Parameter estimates

Model parameters are chosen based on previous literature estimates and/or best fit to available data. For a full list of model parameters and their values, see Table 2. For the number of cells (N_{cells}) and number of fCpG sites per cell (N_{sites}), we choose values that are reasonable from both a biological and computational standpoint. Examples of fitness coefficients $(1 + s_i)$ for each clone in the case of strong, weak, and no selection are included in Table **??** in the Supplementary Information.

Comparing model simulations vs methylation data

To directly compare model simulations to the methylation data we generate 10^2 data trajectories for monzygotic twins, dizygotic twins, and unrelated individuals, each in the case of no selection, weak selection, and strong selection during life. Simulated data trajectory points are generated for each dataset (e.g. at $t = 0, 18, 28, 56, 76, 86$ for MZ twins) via a normal distribution with mean of the dataset and standard deviation of the dataset at that time point. These trajectories are then compared with the 10^2 mathematical model simulations by computing the function d, which measures the mean distance from each model simulation to each data trajectory at the given time points. Results can be seen in Table S2 in the Supplementary Information.

Author statements

Author Contributions

J. Kreger: Conceptualization, software, investigation, methodology, writing–original draft, and editing. **J. Mooney**: Investigation, methodology, writing–original draft, and editing. **D. Shibata**: Conceptualization, investigation, methodology, supervision, writing–original draft, and editing. **A.L. MacLean**: Conceptualization, software, investigation, methodology, supervision, writing–original draft, and editing.

Funding Statement

J. Kreger acknowledges support from the Momental Foundation Mistletoe Research Fellowship. **A.L. MacLean** acknowledges support from the NIH (R35GM143019).

Conflict of Interest Statement

The authors declare no competing interests.

Data Availability Statement

All code and data analysis associated with this study are available on GitHub at: github.com/maclean-lab/scFMC-model. All raw datasets used in this study are publicly available on the Gene Expression Omnibus (GEO): see Table 1.

References

[1] Haas, S., Trumpp, A. & Milsom, M. D. Causes and Consequences of Hematopoietic Stem Cell Heterogeneity. *Cell Stem Cell* **22**, 627–638 (2018). URL `https://www.cell.com/cell-stem-cell/abstract/S1934-5909(18)30165-6`. Publisher: Elsevier.

[2] Zink, F. *et al.* Clonal hematopoiesis, with and without candidate driver mutations, is common in the elderly. *Blood* **130**, 742–752 (2017). URL `https://doi.org/10.1182/blood-2017-02-769869`.

[3] Watson, C. J. *et al.* The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science (New York, N.Y.)* **367**, 1449–1454 (2020).

[4] Fabre, M. A. *et al.* The longitudinal dynamics and natural history of clonal haematopoiesis. *Nature* **606**, 335–342 (2022).

[5] Ahmad, H., Jahn, N. & Jaiswal, S. Clonal Hematopoiesis and Its Impact on Human Health. *Annual Review of Medicine* **74**, 249–260 (2023). URL `https://doi.org/10.1146/annurev-med-042921-112347`. _eprint: https://doi.org/10.1146/annurev-med-042921-112347.

[6] MacLean, A. L., Lo Celso, C. & Stumpf, M. P. H. Concise Review: Stem Cell Population Biology: Insights from Hematopoiesis. *Stem Cells (Dayton, Ohio)* **35**, 80–88 (2017).

[7] Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012). URL `https://www.nature.com/articles/nature10762`. Number: 7381 Publisher: Nature Publishing Group.

[8] Wu, W.-C. *et al.* Circulating hematopoietic stem and progenitor cells are myeloid-biased in cancer patients. *Proceedings of the National Academy of Sciences* **111**, 4221–4226 (2014). URL `https://www.pnas.org/doi/10.1073/pnas.1320753111`. Publisher: Proceedings of the National Academy of Sciences.

[9] Stadler, T., Pybus, O. G. & Stumpf, M. P. H. Phylodynamics for cell biologists. *Science* **371**, eaah6266 (2021). URL `https://www.science.org/doi/10.1126/science.aah6266`. Publisher: American Association for the Advancement of Science.

[10] Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022). URL `https://www.nature.com/articles/s41586-022-04786-y`. Number: 7913 Publisher: Nature Publishing Group.

[11] Stacey, S. N. *et al.* Genetics and epidemiology of mutational barcode-defined clonal hematopoiesis. *Nature Genetics* 1–11 (2023). URL `https://www.nature.com/articles/s41588-023-01555-z`. Publisher: Nature Publishing Group.

[12] Weinstock, J. S. *et al.* Aberrant activation of TCL1A promotes stem cell expansion in clonal haematopoiesis. *Nature* **616**, 755–763 (2023). URL `https://www.nature.com/articles/s41586-023-05806-1`. Number: 7958 Publisher: Nature Publishing Group.

[13] Johnson, B., Shuai, Y., Schweinsberg, J. & Curtius, K. clonerate: fast estimation of single-cell clonal dynamics using coalescent theory. *Bioinformatics* **39**, btad561 (2023). URL `https://doi.org/10.1093/bioinformatics/btad561`.

[14] Fabre, M. A. *et al.* Concordance for clonal hematopoiesis is limited in elderly twins. *Blood* **135**, 269–273 (2020). URL `https://doi.org/10.1182/blood.2019001807`.

[15] Ghersi, J. J. *et al.* Haematopoietic stem and progenitor cell heterogeneity is inherited from the embryonic endothelium. *Nature Cell Biology* **25**, 1135–1145 (2023). URL `https://www.nature.com/articles/s41556-023-01187-9`. Number: 8 Publisher: Nature Publishing Group.

[16] Haldane, J. B. S. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Mathematical Proceedings of the Cambridge Philosophical Society* **23**, 838–844 (1927). URL `https://www.cambridge.org/core/journals/mathematical-proceedings-of-the-cambridge-philosophical-society/article/abs/mathematical-theory-of-natural-and-artificial-selection-part-v-selection-and-mutation/9B6F4FE68136A70E06133E2E389EFA5B`. Publisher: Cambridge University Press.

[17] Wright, S. Evolution in Mendelian populations. *Genetics* **16**, 97 (1931). URL `http://www.genetics.org/content/16/2/97.abstract`.

[18] Fisher, R. A. The Evolution of Dominance in Certain Polymorphic Species. *The American Naturalist* **64**, 385–406 (1930). URL `https://www-journals-uchicago-edu.libproxy2.usc.edu/doi/abs/10.1086/280325`. Publisher: The University of Chicago Press.

[19] Moran, P. A. P. Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society* **54**, 60–71 (1958). Publisher: Cambridge University Press.

[20] Levin, S. A. Dispersion and Population Interactions. *The American Naturalist* **108**, 207–228 (1974). URL https://doi.org/10.1086/282900. Publisher: The University of Chicago Press.

[21] Levin, S. A. & Powell, T. M. *Patch Dynamics*, vol. 96 of *Lecture Notes in Biomathematics* (Springer-Verlag Berlin Heidelberg, 1993). URL 10.1007/978-3-642-50155-5.

[22] Otto, S. P. & Whitlock, M. C. The Probability of Fixation in Populations of Changing Size. *Genetics* **146**, 723–733 (1997). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208011/.

[23] Patwa, Z. & Wahl, L. The fixation probability of beneficial mutations. *Journal of The Royal Society Interface* **5**, 1279–1289 (2008). URL https://royalsocietypublishing.org/doi/10.1098/rsif.2008.0248. Publisher: Royal Society.

[24] Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909–918.e8 (2018).

[25] Cosgrove, J., Hustin, L. S. P., de Boer, R. J. & Perié, L. Hematopoiesis in numbers. *Trends in Immunology* **42**, 1100–1112 (2021).

[26] Marceau, K. *et al.* The Prenatal Environment in Twin Studies: A Review on Chorionicity. *Behavior Genetics* **46**, 286–303 (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858569/.

[27] van Beijsterveldt, C. E. M. *et al.* Chorionicity and Heritability Estimates from Twin Studies: The Prenatal Environment of Twins and Their Resemblance Across a Large Number of Traits. *Behavior Genetics* **46**, 304–314 (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4858554/.

[28] Hansen, J. W. *et al.* Clonal hematopoiesis in elderly twins: concordance, discordance, and mortality. *Blood* **135**, 261–268 (2020).

[29] Trejo, V. *et al.* X Chromosome Inactivation Patterns Correlate with Fetal-Placental Anatomy in Monozygotic Twin Pairs: Implications for Immune Relatedness and Concordance for Autoimmunity. *Molecular Medicine* **1**, 62–70 (1994). URL https://molmed.biomedcentral.com/articles/10.1007/BF03403532. Number: 1 Publisher: BioMed Central.

[30] Kristiansen, M. *et al.* Twin study of genetic and aging effects on X chromosome inactivation. *European Journal of Human Genetics* **13**, 599–606 (2005). URL https://www.nature.com/articles/5201398. Number: 5 Publisher: Nature Publishing Group.

[31] Gabbutt, C. *et al.* Fluctuating methylation clocks for cell lineage tracing at high temporal resolution in human tissues. *Nature Biotechnology* **40**, 720–730 (2022). URL https://doi.org/10.1038/s41587-021-01109-w.

[32] Müllers, S. M., McAuliffe, F. & Malone, F. D. 44 - Multiple Pregnancy. In Pandya, P. P., Oepkes, D., Sebire, N. J. & Wapner, R. J. (eds.) *Fetal Medicine (Third Edition)*, 532–553.e6 (Elsevier, London, 2020). URL https://www.sciencedirect.com/science/article/pii/B9780702069567000440.

[33] Wenstrom, K. D., Tessen, J. A., Zlatnik, F. J. & Sipes, S. L. Frequency, Distribution, and Theoretical Mechanisms of Hematologic and Weight Discordance in Monochorionic Twins. *Obstetrics & Gynecology* **80**, 257 (1992). URL https://journals.lww.com/greenjournal/Abstract/1992/08000/Frequency,_Distribution,_and_Theoretical.20.aspx?casa_token=QG9NueefvJAAAAAA:

TahmJmedoHqtdinxWxbUDQoSX3Zbe9x-vC-iQA3QRMOpWQB5rbOs11WzjJCN8FaaVTI13cKYO_fCX8_JKTCheoo.

[34] Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021). URL https://www.nature.com/articles/s41586-021-03786-8. Number: 7876 Publisher: Nature Publishing Group.

[35] Kimura, M. On the probability of fixation of mutant genes in a population. *Genetics* **47**, 713–719 (1962). URL https://pubmed.ncbi.nlm.nih.gov/14456043.

[36] Waxman, D. A Unified Treatment of the Probability of Fixation when Population Size and the Strength of Selection Change Over Time. *Genetics* **188**, 907–913 (2011). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3176099/.

[37] Lo Celso, C. & Scadden, D. T. The haematopoietic stem cell niche at a glance. *Journal of Cell Science* **124**, 3529–3535 (2011). URL https://doi.org/10.1242/jcs.074112. https://journals.biologists.com/jcs/article-pdf/124/21/3529/1441798/3529.pdf.

[38] Paul, F. *et al.* Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors. *Cell* **163**, 1663–1677 (2015). URL https://www.cell.com/cell/abstract/S0092-8674(15)01493-2. Publisher: Elsevier.

[39] Rodriguez-Fraticelli, A. E. *et al.* Single-cell lineage tracing unveils a role for TCF15 in haematopoiesis. *Nature* **583**, 585–589 (2020). URL https://www.nature.com/articles/s41586-020-2503-6. Number: 7817 Publisher: Nature Publishing Group.

[40] Ranzoni, A. M. *et al.* Integrative Single-Cell RNA-Seq and ATAC-Seq Analysis of Human Developmental Hematopoiesis. *Cell Stem Cell* **28**, 472–487.e7 (2021). URL https://www.sciencedirect.com/science/article/pii/S1934590920305531.

[41] Pei, W. *et al.* Resolving Fates and Single-Cell Transcriptomes of Hematopoietic Stem Cell Clones by *PolyloxExpress* Barcoding. *Cell Stem Cell* **27**, 383–395.e8 (2020). URL https://www.sciencedirect.com/science/article/pii/S1934590920303568.

[42] Sun, J. *et al.* Clonal dynamics of native haematopoiesis. *Nature* **514**, 322–327 (2014). URL https://www.nature.com/articles/nature13824. Number: 7522 Publisher: Nature Publishing Group.

[43] Haltalli, M. L. R. *et al.* Manipulating niche composition limits damage to haematopoietic stem cells during Plasmodium infection. *Nature Cell Biology* **22**, 1399–1410 (2020). URL https://www.nature.com/articles/s41556-020-00601-w. Number: 12 Publisher: Nature Publishing Group.

[44] Tang, D. *et al.* Dietary restriction improves repopulation but impairs lymphoid differentiation capacity of hematopoietic stem cells in early aging. *The Journal of Experimental Medicine* **213**, 535–553 (2016).

[45] Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41**, D991–D995 (2013). URL https://doi.org/10.1093/nar/gks1193.

[46] K Mogensen, P. & N Riseth, A. Optim: A mathematical optimization package for Julia. *Journal of Open Source Software* **3**, 615 (2018). URL http://joss.theoj.org/papers/10.21105/joss.00615.

[47] Gordon, L. *et al.* Neonatal DNA methylation profile in human twins is specified by a complex interplay between intrauterine environmental and genetic factors, subject to tissue-specific influence. *Genome Research* **22**, 1395–1406 (2012).

[48] Kandaswamy, R. *et al.* DNA methylation signatures of adolescent victimization: analysis of a longitudinal monozygotic twin sample. *Epigenetics* **16**, 1169–1186 (2021).

[49] Hannon, E. *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS genetics* **14**, e1007544 (2018).

[50] Ollikainen, M. *et al.* Genome-wide blood DNA methylation alterations at regulatory elements and heterochromatic regions in monozygotic twins discordant for obesity and liver fat. *Clinical Epigenetics* **7**, 39 (2015).

[51] Tan, Q. *et al.* Epigenetic signature of birth weight discordance in adult twins. *BMC genomics* **15**, 1062 (2014).

[52] Roos, L. *et al.* Integrative DNA methylome analysis of pan-cancer biomarkers in cancer discordant monozygotic twin-pairs. *Clinical Epigenetics* **8**, 7 (2016).

[53] Li, S. *et al.* Genetic and Environmental Causes of Variation in the Difference Between Biological Age Based on DNA Methylation and Chronological Age for Middle-Aged Women. *Twin Research and Human Genetics: The Official Journal of the International Society for Twin Studies* **18**, 720–726 (2015).

[54] Tan, Q. *et al.* Epigenetic drift in the aging genome: a ten-year follow-up in an elderly twin cohort. *International Journal of Epidemiology* **45**, 1146–1158 (2016). URL `https://doi.org/10.1093/ije/dyw132`.

[55] Pearson, K. & Galton, F. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58**, 240–242 (1997). URL `https://royalsocietypublishing.org/doi/10.1098/rspl.1895.0041`. Publisher: Royal Society.

[56] Cedar, H. & Bergman, Y. Programming of DNA methylation patterns. *Annual Review of Biochemistry* **81**, 97–117 (2012).

[57] Besançon, M. *et al.* Distributions.jl: Definition and modeling of probability distributions in the juliastats ecosystem. *Journal of Statistical Software* **98**, 1–30 (2021). URL `https://www.jstatsoft.org/v098/i16`.

[58] Kousathanas, A. & Keightley, P. D. A Comparison of Models to Infer the Distribution of Fitness Effects of New Mutations. *Genetics* **193**, 1197–1208 (2013). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3606097/`.

[59] Kim, B. Y., Huber, C. D. & Lohmueller, K. E. Inference of the Distribution of Selection Coefficients for New Nonsynonymous Mutations Using Large Samples. *Genetics* **206**, 345–361 (2017). URL `https://doi.org/10.1534/genetics.116.197145`.

[60] Huber, C. D., Kim, B. Y., Marsden, C. D. & Lohmueller, K. E. Determining the factors driving selective effects of new nonsynonymous mutations. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 4465–4470 (2017). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5410820/`.

[61] Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: A Fresh Approach to Numerical Computing. *SIAM Review* **59**, 65–98 (2017). URL `https://epubs.siam.org/doi/10.1137/141000671`.

[62] Roesch, E. *et al.* Julia for biologists. *Nature Methods* **20**, 655–664 (2023). URL `https://www.nature.com/articles/s41592-023-01832-z`. Number: 5 Publisher: Nature Publishing Group.

[63] Ushijima, T. *et al.* Fidelity of the Methylation Pattern and Its Variation in the Genome. *Genome Research* **13**, 868–874 (2003). URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC430912/`.

[64] Crofts, S. J. C., Latorre-Crespo, E. & Chandra, T. DNA methylation rates scale with maximum lifespan across mammals (2023). URL `https://www.biorxiv.org/content/10.1101/2023.05.15.540689v1`. Pages: 2023.05.15.540689 Section: New Results.