Methods for Estimating Personal Disease Risk and Phylogenetic Diversity of Hematopoietic Stem Cells

Jack M. Craig\textsuperscript{1,2}, Glenn S. Gerhard\textsuperscript{3}, Sudip Sharma\textsuperscript{1,2}, Anastasia Yankovskiy\textsuperscript{1,2}, Sayaka Miura\textsuperscript{1,2*}, and Sudhir Kumar\textsuperscript{1,2*}

\textsuperscript{1}Institute for Genomics and Evolutionary Medicine, Temple University, Philadelphia, PA, USA
\textsuperscript{2}Department of Biology, Temple University, Philadelphia, PA, USA
\textsuperscript{3}Department of Medical Genetics and Molecular Biochemistry, Lewis Katz School of Medicine, Temple University, Philadelphia, PA, USA

*Corresponding authors. Email: s.kumar@temple.edu and sayaka.miura@temple.edu

\textbf{Keywords:} Biological age, hematopoietic stem cells, somatic mutations, cancer risk

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Abstract

An individual’s chronological age does not always correspond to the health of different tissues in their body, especially in cases of disease. Therefore, estimating and contrasting the physiological age of tissues with an individual’s chronological age may be a useful tool to diagnose disease and its progression. Here, we present novel metrics to quantify the loss of phylogenetic diversity in hematopoietic stem cells (HSC), which are precursors to most blood cell types and are associated with many blood-related diseases. These metrics showed an excellent correspondence with an age-related increase in blood cancer incidence, enabling a model to estimate the phylogeny-derived age (phyloAge) of HSCs present in an individual. HSC phyloAge was generally older than the chronological age of patients suffering from myeloproliferative neoplasms (MPN). We present a model that relates excess HSC aging with increased MPN risk. It predicted an over 200 times greater risk based on the HSC phylogenies of the youngest MPN patients analyzed. Our new metrics are designed to be robust to sampling biases and do not rely on prior knowledge of driver mutations or physiological assessments. Consequently, they complement conventional biomarker-based methods to estimate physiological age and disease risk.
Introduction

Somatic aging is characterized by an unrelenting accumulation of genetic variants that become a mutational burden with potentially significant health consequences, particularly in tissues with high cellular turnover (Fancello et al., 2019; Sha et al., 2020). HSCs bearing newly evolved variants can increase in numbers due to proliferative and/or survival advantage, leading to clonal expansions. Such clonally-expanded hematopoietic stem cells (HSCs) can result in clonal hematopoiesis of indeterminate potential (CHIP) in circulating blood cells. CHIP increases with age paralleling the increased risk of hematological neoplasms and cardiovascular disease (Bick et al., 2020; Nachun et al., 2021; Younes, Syler and Hamed, 2023). This pattern is often associated with driver mutations in some genetic loci, such as DNMT3A and TET2 (Buscarlet et al., 2017; Bailey et al., 2018).

Many studies identify CHIPs by the presence of such driver mutations, with their reported prevalence ranging from ~1% in young people (less than 40 years old) to >15% in those over 65 (Groarke and Young, 2019; Bick et al., 2020; Fabre et al., 2022). These drivers emerge from a background of somatic mosaicism that develops continuously over time (Groarke and Young, 2019; Bick et al., 2020; Fabre et al., 2022). However, few studies have focused on using putative passenger (non-driver) variation to assess disease risk. Advancements in single-cell sequencing of HSC genomes now provide base-resolution profiles of all genetic changes in somatic cells (Lee-Six and Kent, 2020; Van Egeren et al., 2021; Fabre et al., 2022; Mitchell et al., 2022), offering new avenues for developing quantitative models to complement the analysis of driver mutations.

In this study, we measured the accumulation of single nucleotide alterations (SNAs) in somatic genomes of HSCs and assessed temporal changes in HSC phylogenetic diversity with age. We present novel approaches to measure the decay of phylogenetic diversity of HSC genomes generated from single-cell sequencing data. We explored the relationships between the number of SNAs and the decay of phylogenetic diversity with the age-related incidence of cancer in populations (Fig. 1). Based on the observed patterns, we developed a model to estimate HSC phylogeny-derived age (phyloAge) to predict the increased risk of disease. We applied these measures to HSC phylogenies of individuals suffering from myeloproliferative neoplasms (MPN) to estimate increased cancer risk independent of commonly used panels of driver mutations or other biomarkers.
Results and Discussion

Accumulation of genetic variation in HSCs in the healthy

We utilized the most comprehensive single-cell sequencing data publicly available for HSCs (Mitchell et al., 2022). This consists of whole genome sequences of DNA obtained from 3,579 colonies of cells derived from single immunophenotypic HSC/MPPs (Lin–CD34+CD38–CD45RA–) that were sorted using flow cytometry and then cultured to produce single-cell-derived hematopoietic colonies from 10 healthy individuals spanning the human lifespan.

We calculated the number of genetic differences (GDs) between HSC sequences for each individual. We divided these by two to generate per-lineage estimates for each pair of HSCs. The distribution of these GDs was unimodal in infants, consistent with the rapid initial HSC expansion during embryogenesis, which results in limited HSC lineage divergence and thus similar GDs (Fig. 2a). With increasing age, the GDs increase due to the steady accumulation of new variants, resulting in longer GDs in 38-year-olds and 63-year-olds (Fig. 2b and c). These adult distributions are characterized by long tails of shorter GDs, representing the gradual accumulation of more recently diverged subclonal HSCs that are more closely related to one another than to any of the founder HSCs. In an 81-year-old individual, these developed into a secondary peak reflecting the increasing prevalence of subclonal HSCs (Fig. 2d), consistent with CHIPs that arise after the establishment of the initial population of HSC clones during embryogenesis and become more prominent with age (Groarke and Young, 2019; Ayachi, Buscarlet and Busque, 2020; Fabre et al., 2022).

The tempo of sequence variation accumulation in HSCs

Figure 3a shows the relationship between the central values of primary peaks in GD distributions and the ages of seven healthy adults. The relationship is nearly linear, but the null hypothesis of a linear fit was rejected at $P < 0.01$ when compared to a second-order polynomial fit. The curvilinear relationship was also found for single nucleotide alterations (SNAs), which were detected directly by comparing HSC genomes with the germline genome ($P < 0.01$; Fig. 3b). These findings differ from previous reports of a constant molecular clock for HSC evolution (Osorio et al., 2018; Brown et al., 2019; Dietlein et al., 2020; Mitchell et al., 2022; Williams et al., 2022). Importantly, the curvilinear relationship between age and SNA counts and GDs does not explain the exponential increase in the age-related incidence of blood cancers (Fig. 3c).
Decay of phylogenetic diversity with age

To seek an explanation (or at least a better correlation) for the exponential age-related increase in cancer risk, we explored phylogenies reconstructed using SNAs in HSC genomes as an alternative method for characterizing patterns of age-related change in HSC. An individual's total number of HSC lineages has been shown to be established during embryogenesis and remains relatively stable throughout life (Lee-Six et al., 2018; Jaiswal and Ebert, 2019; Ayachi, Buscarlet and Busque, 2020; Mitchell et al., 2022). As expected then, the HSC phylogeny of a 38-year individual exhibits extensive polyclonality, visible as many early-branching lineages, with only rare instances of evolutionary bifurcations (Fig. 4a). In contrast, the HSC phylogeny of an elderly individual (81 years old) has many expanding lineages (CHIPs) with many descendant subclones (Fig. 4b). In this case, subclonal HSCs comprise over one-third of all HSCs, which means that subclonal HSCs which originated after birth appear to replace ancestral HSC lineages. Since the ancestral HSCs diversified early and evolved independently of others, they represent a greater phylogenetic diversity than the more recently diverged subclonal lineages that share a genetic history with closely related subclonal HSCs. Therefore, the displacement of ancestral HSC lineages by subclonal HSCs represents a decay in phylogenetic diversity with age (Mitchell, Gopakumar and Jaiswal, 2021). The lineages-through-time (LTT) plots reveal these phenomenological trends in biodiversity loss based on the temporal distribution of branch points in the phylogenetic trees. In the LTT plot of the 38-year-old (Fig. 4c), the total number of HSC lineages is established early and remains similar over time. In contrast, the LTT plot of the elderly individual shows an additional phase of HSC diversification and the loss of phylogenetic diversity (Fig. 4d).

New phylogenetic measures of biodiversity decay

We explored the use of two biodiversity metrics to quantify the decline in phylogenetic biodiversity with age in single-cell HSC phylogenies: HSC phylogeny imbalance ($\pi$) and the number of HSC lineages ($n$). Phylogeny imbalance (Colless, 1982) is the sum of absolute differences in the sizes of the descendant clades for every internal node in the tree. Computationally, the absolute size difference between the two descendant clades of a node is calculated for every node and these values are summed for all the internal nodes in the phylogeny to obtain the imbalance metric. Due to the CHIP events, the imbalance grows in the HSC phylogeny with age. However, $\pi$ and $n$ depend on the number of HSCs sampled, as...
shown in **Fig. 5a** and **c**, respectively. In these analyses, we randomly removed subsets of HSC lineages from the HSC phylogeny of a 38-year-old healthy individual (KX002) at 10% intervals, resulting in a set of ten phylogenies ranging from 100% to 10% of the total lineages sampled. The probability of a given lineage appearing in a phylogeny is called “sampling fraction,” which may be treated as a measure of data richness in the HSC phylogeny. We then calculated phylogenetic imbalance and lineage counts for each. The imbalance of HSC phylogenies ($\pi$) scaled linearly with data richness ($R^2 = 0.98$; **Fig. 5a**), as did the number of tips (n) present in the HSC phylogeny ($R^2 = 1$; **Fig. 5c**).

So, we developed two novel metrics (α and β) to quantify the decline in phylogenetic biodiversity with age in single-cell HSC phylogenies. These metrics are designed to be robust to the number of HSCs sampled. The first metric (α) measures the standardized change in $\pi$ over an individual’s lifetime after birth as the natural logarithm of the ratio of contemporary and ancestral values of $\pi$:

$$\alpha = \log_2 \left( \frac{\pi}{\pi_a} \right) = \log_2 (\pi) - \log_2 (\pi_a).$$  \hspace{1cm} \text{(Eq. 1)}$$

The contemporary value of $\pi$ is estimated using the whole HSC phylogeny. The ancestral imbalance ($\pi_a$) is obtained by cropping the phylogeny at a time point equivalent to the inflection point in its LTT plot, where the initial HSC lineage diversification ends at or near birth (see Material and Methods). The α metric has a minimum value of 0 at birth, because $\pi$ and $\pi_a$ will be equal. A base-2 logarithm is used because phylogenetic branching is typically a doubling process. HSC phylogenies’ α estimates were not significantly correlated with the number of HSCs sampled ($R^2 = 0.16; P > 0.05$), unlike the estimate of $\pi$ (compare **Fig. 5a** and **b**).

The second metric (β), based on the number of HSC clonal lineages (n), is also intrinsically normalized. It is the logarithm of the ratio of observed lineages (n) and inferred ancestral lineages that retained ($n_a$):

$$\beta = \log_2 \left( \frac{n}{n_a} \right) = \log_2 (n) - \log_2 (n_a)$$  \hspace{1cm} \text{(Eq. 2)}$$

The β metric is 0 for young, healthy individuals who have not lost any ancestral HSC lineages. However, it increases over time as expanding subclones displace the ancestral HSCs, reducing overall diversity. We find no significant relationship between β and data richness ($R^2 = 0.16; P > 0.05$; $R^2 = 0.16; P > 0.05$; **Fig. 5c** and **d**).
Relationship of $\alpha$ and $\beta$ with age-related increase in blood cancer

We estimated $\alpha$ and $\beta$ metrics using HSC phylogenies of eight healthy adults. Both exhibit an exponential relationship with age (Fig. 6). The diversity decay increases exponentially, remaining small until about 65 years and increasing rapidly thereafter. These patterns were highly concordant with that observed for the cancer risk, with the correlation coefficients between the MPN risk and biodiversity metrics of $R^2 = 0.89$ ($P < 0.01$) for $\alpha$ and $R^2 = 0.87$ for $\beta$ ($P < 0.01$).

Estimating phyloAge of HSC phylogenies

The relationships of $\alpha$ and $\beta$ estimates with the age of healthy adults and their cancer risk prompted us to develop a model to estimate what we call HSC phylogeny-derived age (phyloAge). In this case, both $\alpha$ and $\beta$ were used as phylogenetic markers analogous to epigenetic markers of age (Hannum et al., 2013; Horvath, 2013; Simpson and Chandra, 2021). We optimized the values of parameters using a numerical Gauss-Newton algorithm, aiming to minimize the sum of squared differences between the observed and predicted age (see Methods).

In leave-one-out (LOO) analyses, the phyloAge model predictions achieved a correlation coefficient of 0.82 with the chronological age. The absolute difference between phyloAges and chronological ages, referred to as the residualAge, was 6.6 years on average (Fig. 7). The predictions for younger individuals, where $\alpha$ and $\beta$ do not change much over time, had higher residualAge (3.8 to 15.0 years). The residualAge was less than 3.7 years for individuals 65 years and older. We also explored using Shannon's biodiversity index (Whittaker, 1972; Mitchell et al., 2022), but found its inclusion in the phyloAge model not to improve the accuracy of predictions.

The performance of the phyloAge model was comparable to that reported for epigenetic clocks in some recent studies. For example, GrimAge2 was trained using thousands of methylation profiles to predict chronological age and showed a correlation of 0.78–0.95 between true and predicted ages (Lu et al., 2019; Lu et al., 2022), whereas the correlation was 0.94 for phyloAges. DeepMAge, a deep learning method trained on 4,930 methylation profiles (Galkin et al., 2021), was reported to predict age with a median error of 2.8 years, which is better than phyloAges for young people but similar for older people. These results suggest that phyloAge
could yield results comparable to existing methods, with the possibility of improvement when single-cell HSC sequencing data becomes available for additional healthy adults.

**Estimates of phyloAges for HSC phylogenies of MPN patients**

Next, we applied our metrics and methods to the HSC phylogenies of 12 MPN patients aged 23–83 years (Williams *et al.*, 2022). These data were acquired similarly to those of healthy individuals (Mitchell *et al.*, 2022). The MPN phyloAge estimates were consistently older than the chronological ages (Fig. 8a). For example, a 49-year-old MPN patient received a phyloAge of 94.4 years, indicating that their HSC biodiversity had decayed by an additional 45.4 years beyond their chronological age. This is evident upon comparing their HSC phylogeny (Fig. 8b) with that of the oldest healthy person (81 years old; Fig. 4b). In fact, the MPN patient has lost all but three ancestral HSC lineages, which represents a greater loss in genetic diversity and a greater change in phylogeny shape than that of a healthy person almost four times their age (Fig. 4b). In general, young MPN patients’ LTTs, and even cellular phylogenies, resemble those of much older healthy individuals. That is, MPN is associated with HSC diversity loss earlier in life due to the dominance of subclonal HSC lineages. Overall, residualAges discriminated between the healthy adults and MPN patients until the age of 80, with the residualAges > 33 for MPN patients less than 80 years old (Fig. 8a).

We sought to compare the performance of residual age based on the phyloAge model for MPN patients with those reported for biological ages derived from methylation biomarkers, which have been widely used since their introduction (Horvath, 2013). One study which used DNA methylation markers to estimate physiological ages in MPN patients specifically found an average residualAge of just 0.5 years. However, the residualAge estimates based on DNAm vary extensively among cancer types. Weidner *et al.* (2014) reported small differences for aplastic anemia (average 11.7 years) and dyskeratosis congenita (average 16.5 years), whereas (Zhu *et al.*, 2019) reported much larger differences for many cancers (up to 50 years). DeepMAge predicted an average residualAge of 1.7 years for ovarian cancer. Therefore, residualAge differences can be large for some types of cancers, including MPN examined here.

Unlike residualAges of MPN phylogenies, we did not see a significant difference between the counts of cancer-associated driver variants between the healthy individuals and those with MPN, as the average number of drivers in healthy individuals and MPN patients were quite similar (4.5 and 3.9 per individual, respectively) and not significantly different ($P > 0.75$). Therefore, while drivers are frequently implicated in causing CHIPs (Brown *et al.*, 2019; Dietlein
et al., 2020), their counts do not discriminate between healthy and diseased individuals.

However, we found SBS9 mutational signatures in CHIP lineages of elderly MPN patients only (ages > 80 years), which showed limited discrimination when using phyloAges. SBS9 is a key cancer-associated mutational signature characterized by T>G mutations and is induced by somatic hypermutation, which is often reported in the lymphoid (50% and 46% of ICGC and GEL samples, respectively) and myeloid cancer cells (31% of GEL samples) (Alexandrov et al., 2020; Degasperi et al., 2022). SBS9 was not present in any of the healthy individuals.

Therefore, SBS9 may be useful as an additional biomarker for detecting the emergence of MPN in conjunction with phyloAge when the residualAge is small. But, more data is needed to test these suggestions.

**Assessing increased cancer risk using phyloAge**

We found the residualAge, the difference between the HSC phyloAge and chronological age, to be naturally related to the fold-increase in cancer risk due to excess aging, because the reported trends in MPN cancer incidence with age are described by an exponential function 

\( (0.04 \times e^{0.08 \times \text{age}}, \ R^2 = 0.99) \). In this case, the equation for the fold increase in the risk of MPN:

\[
\text{Fold-increased risk} = \frac{0.04 \times e^{0.08 \times \text{phyloAge}}}{0.04 \times e^{0.08 \times \text{chronological age}}}
\]

\( = e^{0.08 \times (\text{phyloAge} - \text{chronological age})} \)  

(Eq. 3)

This equation shows that the increased risk is a function of the difference between phyloAge and chronological age. It translates premature HSC aging into an estimate of increased cancer risk. Using this framework, it should be possible to develop similar equations for other types of cancers and diseases that are influenced by changes in the HSCs.

Applying the above equation to the HSC phylogeny in MPN patients predicts that a 23-year-old person, whose phyloAge is 64.76, will have a 34-fold increase in their probability of developing MPN. Similarly, individuals aged 40-65 exhibited a 27-fold increase in risk, while those aged 65-81 showed a more than two-fold increase in risk. Based on these results, phyloAge shows promise as a tool for forecasting MPN risk based solely on the phenomenological metrics of phylogenetic diversity decay.
Conclusions

HSCs give rise to myeloid cells that differentiate into red blood cells, platelets, neutrophils, basophils, monocytes, eosinophils, and lymphoid progenitors that include T and B lymphocytes and natural killer cells (Ogawa, 1993; Mikkola and Orkin, 2006). CHIP appears to arise from the acquisition of somatic mutations in HSC genomes that permit and/or drive clonal expansion over time to produce an age-dependent increase in blood and immune cell mosaicism (Nachun et al., 2021; Ahmad, Jahn and Jaiswal, 2023; Goldman, Spellman and Agarwal, 2023; Singh and Singh, 2023). Consequently, CHIP is associated with an increased risk of hematological malignancies and cardiovascular and pulmonary disease (Jaiswal et al., 2017; Wong et al., 2023).

We have presented two novel phylogeny shape metrics (α and β), based on traditional ecological biodiversity measures, to capture the decay of diversity of HSC phylogenies reconstructed using SNAs. These measures increase exponentially with age and are concordant with the age-related increases in cancer incidence, fulfilling the need for quantitative phenomenological descriptions of the time-dependent structure of HSC phylogenies. We presented a method to estimate HSC phyloAge using α and β that significantly exceeds the chronological age of HSCs in MPN patients (average 59%, up to 182% in extreme cases).

This age difference was found to distinguish between healthy and diseased individuals more effectively than other quantitative descriptions utilizing DNA methylation (Hannum et al., 2013; Horvath, 2013; Weidner et al., 2014; Bell et al., 2019; Lu et al., 2019; Galkin et al., 2021; Ake T. Lu et al., 2022; A. T. Lu et al., 2022; Seale et al., 2022; Dabrowski et al., 2023). HSC phyloAge was more effective at predicting physiological ages and discerning between healthy people and those with blood cancer than the GD or total number of SNAs accumulated. Further, while few of the somatic SNAs an individual acquires in their lifetime are expected to be cancer-associated driver mutations (Bailey et al., 2018; Brown et al., 2019; Nussinov et al., 2019), even restricting our analysis to these drivers was not sufficient to discern between healthy people and those with cancer, as the counts of drivers in HSC genomes of individuals suffering from MPN were not sufficiently different from those in healthy individuals (Mitchell et al., 2022; Williams et al., 2022). Interestingly, the difference between phyloAge and chronological age forecasts a fold-increase in cancer risk. This could find application in clinical and research settings to track the temporal rate of deterioration of HSCs when personal single-cell sequencing at different time points becomes more routine.
As the field of biological aging research continues to evolve, we anticipate the development of new and more useful approaches, such as chronic sterile inflammation (Franceschi et al., 2018),
glycomics (Borelli, 2014; Krištić et al., 2014), and lipidomics (Beyene et al., 2020; Slade et al., 2021). These tools are already showing a promising ability to identify many serious diseases (Horvath and Ritz, 2015; Ding and Rexrode, 2020; Miyoshi, Kamada and Suzuki, 2020; Wang et al., 2020; Gaunitz, Tjernberg and Schedin-Weiss, 2021). There are even many new tools being developed to give consumers access to their blood health data (Csordas et al., 2022), though these have not yet been able to bridge the gap between physiological age and disease risk.

Regarding phyloAge, we anticipate that advancements in single-cell sequencing technology and the availability of richer cancer incidence data with age will enable more accurate HSC phylogeny-based evolutionary modeling for age and risk assessment. With an increase in single-cell sequencing data from younger healthy individuals, where changes in genomic diversity are more subtle, we expect to understand the natural trajectory of HSC evolution better. These analyses and predictive models will also characterize aging and blood cancers.

Methods

Data acquisition

HSC sequences and phylogenies for ten healthy people (age range infant–81) were retrieved from Mitchell et al. (2022) and for 15 individuals with MPN (age range 20–83) from Williams et al. (2022). An alignment was not available for one healthy individual (KX007), but the phylogeny was so it was used. HSC sequences were composed of genomic positions with SNAs. For MPN patients, the first sampling event for each individual was considered in order to eliminate bias stemming from cancer treatments. Two healthy infants were excluded from rate estimation and model fitting because they were still experiencing rapid HSC diversification among founder lineages, and thus unrepresentative of the stable adult phase of HSC growth which we aimed to model. The incidence of MPN by age was obtained from (Hultcrantz et al., 2020).

Genetic distance estimation

We produced counts of sequence differences using MEGA to calculate pairwise genetic distances (GDs) between HSC sequences (Tamura, Stecher and Kumar, 2021). In all these comparisons, positions with missing data were ignored (pairwise-deletion option). This number
was divided by two to generate per-sequence GD estimates. To determine the peak of the GD
distribution, we generated a histogram with 300 GD bins and then identified the bin containing
the highest point in the resulting GD distribution. We also counted the number of single
nucleotide alterations, SNAs, in an HSC sequence by comparing it to the respective germline
genomes.

**Constant versus relaxed HSC molecular clocks**

We used a likelihood ratio test (Wilks, 1938; Glover and Dixon, 2004) to compare fits between
linear and polynomial regression models for GDs and SNAs accumulated in the HSCs of 7
adults. The polynomial models fit the data significantly better ($P < 0.01$) with the following
models obtained for GDs and SNAs: $GD = 0.06 \times Age^2 + 11.90 \times Age$ and $SNAs = 0.06 \times Age^2 +
12.43 \times Age$.

**\( \alpha \) and \( \beta \) parameter estimation and normalization**

To calculate Colless' index of phylogeny imbalance (Colless, 1982), we used apTreeshape
(Bortolussi et al., 2006) in R (Core Development Team, 2020). We plotted a lineages-through-
time (LTT) plot for each phylogeny using the R package ape (Paradis and Schliep, 2019),
identified the point of inflection where lineage diversification ceased, corresponding to the end of
the embryonic phase of HSC diversification (Lee-Six et al., 2018; Mitchell, Spencer Chapman,
et al., 2022). Because our metrics assess the number and branching patterns of phylogenetic
tips, an ultrametric HSC phylogeny was only required to make the initial crop of LTT. No
estimates of an absolute time scale were necessary, which avoids biases associated with
molecular dating approaches that may not work well when the evolutionary rates converge
throughout the tree.

We generated the ancestral (embryonic) tree by cropping the HSC phylogeny at this inflection
point and estimated ancestral estimates of \( \pi \) and \( n \). For some HSC phylogenies, apTreeshape
failed to provide any estimates because the embryonic phylogeny contained fewer than four
ancestral HSC lineages. In this case, we assumed $\log(\pi_a) = 0$ such that imbalance was the
smallest. By the way, this approach will work even when the HSC tree contains only the
subclonal lineages, because they will be distinguishable due to a long stem branch connecting
them to the germline reference.

For individuals who were sampled multiple times in the progression of their cancer, we take only
the first sampling event as their adult phylogeny condition. This avoids the confounding effects
of the different treatments these individuals underwent to eliminate cancerous HSCs in their blood, necessarily impacting their HSC phylogenies.

**Modeling HSC phylogeny age**

Both $\alpha$ and $\beta$ increase exponentially with age (Fig. 6). Therefore, we fitted non-linear models for estimating age using $\alpha$ and $\beta$ separately. The model parameters were optimized using a numerical Gauss-Newton algorithm, aiming to minimize the sum of squared differences between the observed and predicted age of healthy individuals. Next, we combined these two models using a meta-regression framework (Viechtbauer, 2010) with a maximum likelihood approach (Hardy and Thompson, 1996). This process predicted age and the SE for each healthy individual by utilizing random-effects meta-analysis in the metafor package in R (Viechtbauer, 2010). We validated the combined model, predicted age, and estimated the prediction interval for each healthy individual using the leave-one-out (LOO) approach, in which individual ages were estimated using a model that excluded the individual of interest.

**Annotations of driver mutations**

Driver mutation counts per individual were obtained from the original studies, i.e., Mitchell et al. (2022) for healthy individuals and Williams et al. (2022) for individuals with MPN. Overall, we obtained information on the following genes that contained driver mutations: ASXL1, CBL, DMNT3A, JAK2, KRAS, PPM1D, SF3B1, TET2, and TP53 in healthy people, and CBL, DMNT3A, JAK2, PPM1D, TET2, 1q+, 7p-, 7q-, 9pUPD, 9q-, and 9+ in the MPN patients.

**Mutational signature analysis**

We annotated a clade of many cells separated from the other cells with a long branch in a phylogeny. We identified singletons in cell genomes from the selected clade and inferred mutational signatures using Signal (Degasperi et al., 2020).

**Data Availability**

Source code and sequence data are deposited in the Code Ocean online resource.
Author contributions

S.K. conceived and designed the study, developed diversity and risk metrics, designed statistical analyses, and co-wrote the manuscript. J.C. implemented and investigated all the metrics, conducted data analysis, and co-wrote the manuscript. S.M. contributed to the initial study design and led the preparation and some analyses of genome sequences. S.S. contributed to the statistical analysis and model fitting. G.G. provided key insights into aging and helped with the study design. A.Y. assisted in genome preparation and conducted mutation signature analysis. All other authors contributed to the manuscripts intellectually and assisted with the writing of the results.

Acknowledgments

We would like to thank Chenguir Duy, Jeff Townsend, and Li Liu for their insightful comments on an early draft of this manuscript. This study was funded by grants from the National Institutes of Health to S.K. (LM013385-04) and S.M. (LM014005-02).
References


Ayachi, S., Buscarlet, M. and Busque, L. (2020) ‘60 Years of clonal hematopoiesis research: From X-chromosome inactivation studies to the identification of driver mutations’, *Experimental hematology*, 83(Xm), pp. 2–11.


Dabrowski, J.K. *et al.* (2023) ‘Probabilistic inference of epigenetic age acceleration from cellular dynamics’, *bioRxiv*. Available at: https://doi.org/10.1101/2023.03.01.530570.


Available at: https://doi.org/10.1101/2021.01.18.426733.


Figure captions

Figure 1: Incidence of blood cancer by age per 100K individuals. The incidence of MPN increases exponentially by age according to the following function: Incidence (per 100K) = 0.04×e^{0.08×Age}. Rates of incidence were obtained from Hultcrantz et al. (2020). This pattern mirrors the known incidence rates of other blood cancers such as leukemia and myeloma (Cancer Research UK 2016-2018, ICD-10 C91-C95 and 2016-2018, ICD-10 C90, respectively).

Figure 2: Distributions of genetic differences (GDs) between HSCs. Distributions are shown for (a) an infant, (b) a 38-year-old, (c) a 63-year-old, and (d) an 81-year-old. Subclones arising from later in life are marked by an arrow, which form a tail or a distribution to the left of the primary peak corresponding to the HSCs that arose during embryogenesis.

Figure 3: Accumulation of sequence variation over time in HSCs. Relationship of (a) GDs and (b) SNA counts with age for 7 healthy adult individuals. These relationships are curvilinear, as a second degree polynomial fits the data better than a linear regression in both cases (P < 0.01). (c) Tempos of SNA and GD increases do not explain the much more rapid increase in the age-related incidence of MPN.

Figure 4: The HSC phylogenies of two healthy individuals and changes in their phylogenetic biodiversity. (a) The HSC phylogeny of a 38-year-old healthy individual, which exhibits <10 HSC lineage divergences after the embryonic phase. (b) The HSC phylogeny of an 81-year-old healthy individual, which has many large clades containing subclonal HSCs. Only one-third of the tips are direct descendants of the embryonic HSCs. (c) The LTT plot from the 38-year-old exhibits an initial rapid diversification in HSCs, followed by a period of growth, with minimal increase in total lineages thereafter. (d) The LTT plot from an 81-year-old healthy individual exhibits an initial rapid diversification in HSCs, a period of growth, then a second period of increase later in life.

Figure 5: Normalized metrics account for lineage sampling. (a) The relationship of the imbalance of adult HSC phylogenies (πa) with the data richness. (R^2 = 0.98) and other info. (b) The standardized phylogeny imbalance (α) shows a low correlation with the data richness (R^2 = 0.16; P > 0.05). (c) The relationship of the number of tips present in the adult HSC phylogeny (n) is the same as the sampling fraction (R^2 = 1). (d) The standardized tip count (β) is also not highly correlated with data richness (R^2 = 0.16; P > 0.05).

Figure 6: The relationship of phylogenetic diversity decay metrics with age and age-
related cancer risk. (a) Normalized phylogeny imbalance ($\alpha$) in HSC phylogenies from healthy individuals increases following an exponential function: $\alpha = 0.002 \times e^{0.09 \times \text{age}}$ ($R^2 = 0.89$). (b) Normalized HSC lineage count ($\beta$) increases following an exponential function: $\beta = 0.003 \times e^{0.08 \times \text{age}}$ ($R^2 = 0.87$).

**Figure 7:** The relationship of estimated phyloAge with the chronological age of healthy adults. We estimated phyloAges using the composite model incorporating $\alpha$ and $\beta$ with 95% confidence intervals derived from the leave-one-out (LOO) analysis.

**Figure 8:** Application of the phyloAge model to phylogenies of MPN patients. (a) PhyloAges of individuals with MPN are substantially greater than their true chronological ages. A 1:1 relationship (dashed line) is shown for comparison. (b) The HSC phylogeny of a 49-year-old individual with MPN. Only three of 92 HSCs trace their direct origin back to the root, and one subclonal lineage gives rise to a vast majority of HSCs present at age 49 that displayed almost all the primary HSC lineages. (c) The logged LTT plot from a 49-year-old individual with MPN exhibits an initial rapid diversification in HSCs, a period of growth, then a second period of increase later in life, much like the older healthy person, but much sooner.
Figure 1

Incidence (per 100K)

\[ y = 0.04e^{0.08x} \]

\[ R^2 = 0.99 \]

Age

FIG. 1

165x146 mm (x DPI)
FIG. 2

Figure 2

165x187 mm (x DPI)
Figure 3

165x71 mm (x DPI)
Figure 4

165x147 mm (x DPI)
Figure 5
165x156 mm (x DPI)
Figure 6

Normalized phylogenetic imbalance (a) vs. Age.

Normalized HSC lineages (b) vs. Age.
Figure 7

Chronological age

PhyloAge

Fig. 7
FIG. 8

Figure 8
165x97 mm (x-DPI)