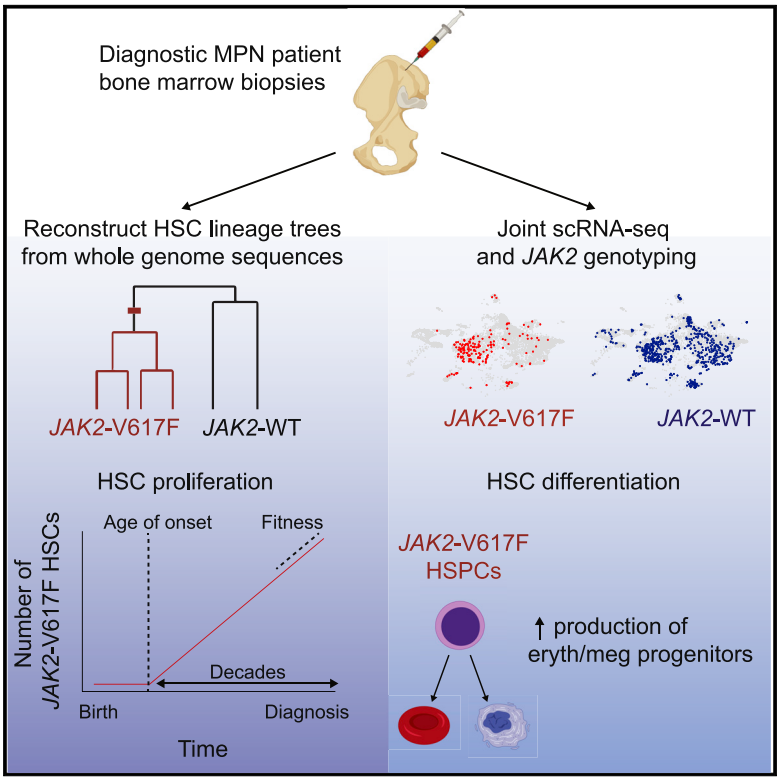
Cell Stem Cell

Short Article

Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms

Graphical Abstract



Highlights

- Single-cell transcriptome and whole-genome sequencing of HSPCs from individuals with MPNs
- The *JAK2*-V617F mutation occurs in a single HSC decades before diagnosis
- *JAK2*-V617F HSCs have increased fitness in native human hematopoiesis
- *JAK2* mutant fraction varies in myeloid progenitor compartments in the same individuals

Authors

Debra Van Egeren, Javier Escabi, Maximilian Nguyen, ..., Ann Mullally, Isidro Cortes-Ciriano, Sahand Hormoz

Correspondence

ann_mullally@dfci.harvard.edu (A.M.), icortes@ebi.ac.uk (I.C.-C.), sahand_hormoz@hms.harvard.edu (S.H.)

In Brief

Van Egeren et al. investigated the effect of the *JAK2*-V617F mutation in individuals with myeloproliferative neoplasms (MPNs) using single-cell profiling and found that the mutation occurs decades before MPN diagnosis and increases the fitness of HSCs. *JAK2*-V617F induces a megakaryocyte-erythroid differentiation bias. The *JAK2*-mutant fraction varies in myeloid compartments in the same individuals.

Van Egeren et al., 2021, Cell Stem Cell *28*, 514–523
March 4, 2021
https://doi.org/10.1016/j.stem.2021.02.001

CellPress

CellPress
OPEN ACCESS

Cell Stem Cell

Short Article

Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms

Debra Van Egeren,[1,2,3,14] Javier Escabi,[1,2,4,14] Maximilian Nguyen,[1,2,14] Shichen Liu,[2] Christopher R. Reilly,[5] Sachin Patel,[3] Baransel Kamaz,[6] Maria Kalyva,[7] Daniel J. DeAngelo,[5] Ilene Galinsky,[5] Martha Wadleigh,[5] Eric S. Winer,[5] Marlise R. Luskin,[5] Richard M. Stone,[5] Jacqueline S. Garcia,[5] Gabriela S. Hobbs,[8] Fernando D. Camargo,[3,9] Franziska Michor,[2,9,10,11,12,13] Ann Mullally,[5,6,10,*] Isidro Cortes-Ciriano,[7,*] and Sahand Hormoz[1,2,10,15,*]

[1]Department of Systems Biology, Harvard Medical School, Boston, MA 02115, USA
[2]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[3]Stem Cell Program, Boston Children's Hospital, Boston, MA 02115, USA
[4]Research Scholar Initiative, Harvard Graduate School of Arts and Sciences, Cambridge, MA 02138, USA
[5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[6]Division of Hematology, Brigham and Women's Hospital, Boston, MA 02115, USA
[7]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
[8]Leukemia Center, Massachusetts General Hospital, Boston, MA 02114, USA
[9]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, MA 02138, USA
[10]Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
[11]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[12]The Center for Cancer Evolution, Dana-Farber Cancer Institute, Boston, MA 02115, USA
[13]The Ludwig Center at Harvard, Boston, MA 02115, USA
[14]These authors contributed equally
[15]Lead Contact
*Correspondence: ann_mullally@dfci.harvard.edu (A.M.), icortes@ebi.ac.uk (I.C.-C.), sahand_hormoz@hms.harvard.edu (S.H.)
https://doi.org/10.1016/j.stem.2021.02.001

SUMMARY

Some cancers originate from a single mutation event in a single cell. Blood cancers known as myeloproliferative neoplasms (MPNs) are thought to originate when a driver mutation is acquired by a hematopoietic stem cell (HSC). However, when the mutation first occurs in individuals and how it affects the behavior of HSCs in their native context is not known. Here we quantified the effect of the *JAK2*-V617F mutation on the self-renewal and differentiation dynamics of HSCs in treatment-naive individuals with MPNs and reconstructed lineage histories of individual HSCs using somatic mutation patterns. We found that *JAK2*-V617F mutations occurred in a single HSC several decades before MPN diagnosis—at age 9 ± 2 years in a 34-year-old individual and at age 19 ± 3 years in a 63-year-old individual—and found that mutant HSCs have a selective advantage in both individuals. These results highlight the potential of harnessing somatic mutations to reconstruct cancer lineages.

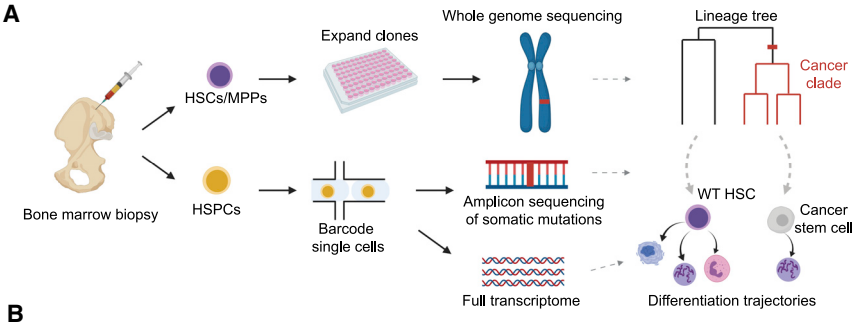
INTRODUCTION

In seminal studies, the *JAK2*-V617F mutation was identified to underlie the molecular pathogenesis of the majority of Philadelphia chromosome-negative myeloproliferative neoplasms (MPNs) (Baxter et al., 2005; James et al., 2005; Kralovics et al., 2005; Levine et al., 2005), a chronic blood cancer. More recently, *JAK2* was identified as one of the most commonly mutated genes in clonal hematopoiesis (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014). Notably, *JAK2*-V617F clonal hematopoiesis is associated with an increased risk of cardiovascular disease (Jaiswal et al., 2017) and venous thrombosis (Cordua et al., 2019; Wolach et al., 2018), in addition to sharing the same germline variants that predispose to the development of *JAK2* mutant MPN (Hinds et al., 2016).

The *JAK2*-V617F mutation results in activated JAK2 signaling, leading to increased production of mature blood cells of the myeloid lineage, ultimately resulting in MPNs. Some individuals with MPN present primarily with increased numbers of red blood cells (polycythemia vera [PV]), others with increased numbers of platelets (essential thrombocythemia [ET]), and, more rarely, some with scarring ("fibrosis") of the bone marrow (primary myelofibrosis [PMF]). Although in MPN it has been shown that the *JAK2*-V617F mutation is detectable in hematopoietic stem cells (HSCs) (Jamieson et al., 2006) and in all mature cell lineages (Delhommeau et al., 2007; Ishii et al., 2006), it is unclear how the mutation affects HSC differentiation

CellPress
OPEN ACCESS

A



B

	ET 1	ET 2	ET 3	PV 1	PV 2 (TET2)	PV 3 (TET2)	ET V617L
Age and sex	34 M	63 F	41 M	56 M	56 M	69 M	49 F
Allelic burden PB	V617F (16.6%)	V617F (25.3%)	V617F (10.6%)	V617F (62.7%)	V617F (68.0%)	V617F (40.9%)	V617L (45.0%)
Secondary mutations	-	-	-	*EZH2* E745* (4.2%)	*TET2* S268* (32.0%)	*TET2* frameshift (31.0%) *TET2* C1289S (31.1%)	-
# cells	7868	7349	7091	6732	7472	7911	7704
# *JAK2* mutant	174	217	68	308	469	541	107
# *JAK2* WT	646	525	567	216	320	687	285

Figure 1. Experimental Design

(A) Individual hematopoietic stem and progenitor cells (HSPCs) from bone marrow aspirates of individuals with MPNs were analyzed in two ways. First, hematopoietic stem cells (HSCs) and multipotent progenitors (MPPs) were expanded *in vitro* and characterized using WGS. Second, we simultaneously read out the transcriptional profiles and somatic mutations in single HSPCs.

(B) Information about the individuals with MPNs sampled in this study. "Allelic burden peripheral blood (PB)" and "secondary mutations" refer to VAFs of *JAK2* mutations and other hematopoiesis-associated mutations in PB, respectively. The numbers of *JAK2* WT and mutant cells identified in the HSPCs using scRNA-seq are given in the last two rows.

See also Figure S1.

and proliferation of human HSCs in their native bone marrow microenvironment to engender clonal hematopoiesis and MPN. Therefore, to understand the behavior of *JAK2*-V617F mutant HSCs and development of MPN, we studied unrelated individuals with newly diagnosed *JAK2*-V617F+ ET and PV and ascertained when the *JAK2*-V617F mutation first occurred in each individual, how the number of *JAK2* mutant cells expanded over time, and the extent to which the differentiation trajectories of the *JAK2* mutant cells deviated from those of cells without the mutation.

Although the effect of the *JAK2*-V617F mutation on HSCs *in vivo* has been modeled previously using *Jak2*-V617F transgenic mice and patient-derived xenograft MPN models, these experimental systems do not accurately recreate the native microenvironment or the behavior of *JAK2* mutant hematopoietic stem and progenitor cells (HSPCs) in humans. The discovery that *JAK2*-V617F causes clonal hematopoiesis (Genovese et al., 2014; Jaiswal et al., 2014; Xie et al., 2014) and the observation that *JAK2*-V617F is often the sole somatic mutation detected in individuals with MPNs (Grinfeld et al., 2018; Lundberg et al., 2014) suggest that the *JAK2* mutation promotes HSC self-renewal and confers a selective advantage. However, this has never been measured directly. Measurement of the self-renewal and differentiation ability of *JAK2* mutant HSCs in individuals with MPNs is not feasible because direct observation of dynamic cell behaviors is not possible in human bone marrow. However, static single-cell genomic and transcriptomic measurements can be used to reconstruct the self-renewal history and differentiation behavior in unperturbed cell populations (Lee-Six et al., 2018; Tusi et al., 2018).

Therefore, to directly assess the consequences of the *JAK2*-V617F mutation on HSC differentiation and self-renewal in their native microenvironment, we reconstructed lineage trees of *JAK2* mutant and wild-type HSCs obtained from individuals with MPN and inferred the history of MPN development in 2 people with ET. In addition, to determine how the *JAK2* mutation affects the differentiation trajectories of the progenies of HSCs, we profiled the transcriptomes of individual cells obtained from bone marrow aspirates of 7 individuals with MPN.

RESULTS

To investigate the effect of *JAK2* mutations in individuals with ET and PV, we performed single-cell transcriptomic profiling of HSPCs from 7 newly diagnosed, untreated individuals with PV (n = 3) and ET (n = 4) as well as healthy controls (n = 2) (Figure 1). The *JAK2*-V617F mutation was detected in 6 individuals, whereas the remaining individuals with ET had a *JAK2* variant previously unreported in humans (*JAK2*-V617L), with fibroblast germline testing confirming a somatic origin of the mutation. Of note, *JAK2*-V617L has been shown previously to induce cytokine independence and constitutive downstream signaling in Ba/F3 cells (Dusa et al., 2008). The individuals with ET did not harbor additional myeloid malignancy-associated mutations, as measured by a clinical next-generation sequencing (NGS) assay performed on whole white blood cells from peripheral blood (i.e., rapid heme panel; Kluk et al., 2016), whereas somatic truncating mutations in *TET2* (2 individuals) and *EZH2* (1 individual) were identified in people with PV (Figure 1B). From each individual with MPNs and healthy donor, we collected a bone marrow aspirate, isolated mononuclear cells, and then enriched for CD34 expression to isolate HSPCs (STAR Methods).

JAK2-Mutant HSPCs Exhibit Fate Bias

To determine how *JAK2* mutations affect HSPC differentiation dynamics in individuals with MPN, we simultaneously measured the full transcriptome and genotyped the *JAK2* mutation in individual CD34+ cells obtained from each bone marrow aspirate (Figure 1A). To do so, we developed a protocol for amplifying specific transcripts from single-cell RNA sequencing (RNA-seq) libraries. Briefly, we used the 10X platform to generate barcoded single-cell cDNA libraries. Before fragmenting the libraries

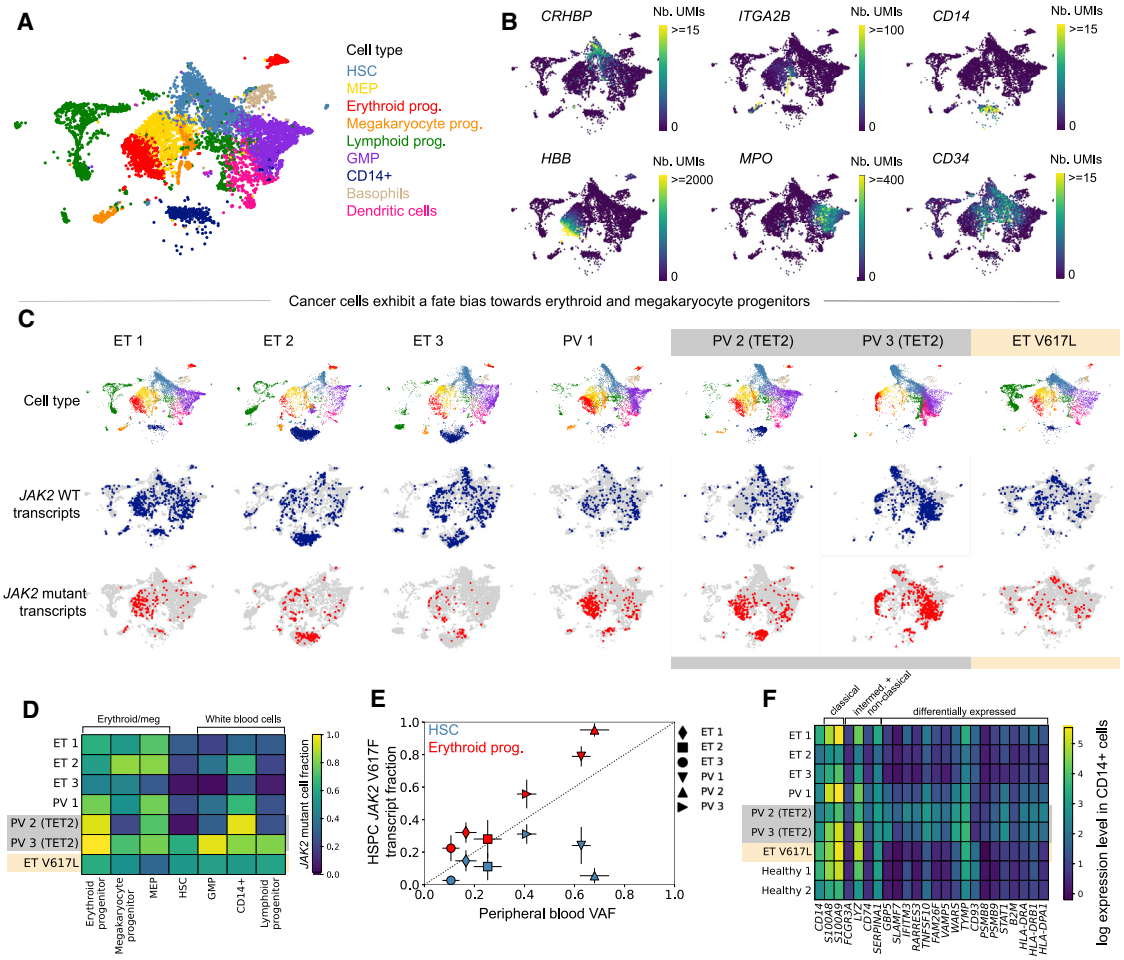


Figure 2. Erythroid and Megakaryocyte Progenitors from Individuals with MPNs Are More Likely to Have the *JAK2*-V617F Mutation than Other CD34+ Bone Marrow HSPCs

(A) UMAP of CD34-enriched bone marrow scRNA-seq data from ET 1, colored by cell type.

(B) Marker gene expression in ET 1 CD34-enriched bone marrow.

(C) Cell type classifications and *JAK2* WT/mutant transcript calls in scRNA-seq data from individuals with MPNs (columns).

(D) Fraction of *JAK2* mutant cells (colors) in different bone marrow cell types from individuals with MPNs.

(E) Relationship between the PB VAF and *JAK2* V617F mutant transcript fraction in bone marrow HSCs (blue) and erythroid progenitors (red). Error bars indicate 95% confidence intervals.

(F) Mean expression of selected marker genes in CD14+ cells that are upregulated in monocyte subsets or were differentially expressed in CD14+ cells between individuals with ET and PV or between individuals with MPNs and healthy controls.

See also Figure S2.

for sequencing, we generated amplicon libraries of the target loci for the somatic mutations of interest by performing three rounds of nested PCR with locus-specific reverse primers and generic forward primers (Figure S1; STAR Methods). The somatic mutations were mapped to the transcriptional profiles using the shared single-cell barcodes across the two libraries (Figure S1; STAR Methods).

Using this approach, we detected the *JAK2* mutation site (locus encoding codon 617) in at least one transcript in 5%–15% of cells (mean 9.5%) in the 7 libraries, improving over the existing methods for detecting somatic mutations in single-cell libraries (Nam et al., 2019; Psaila et al., 2020). We designated cells in which at least one mutated *JAK2* transcript was detected as *JAK2* mutant cells. Importantly, cells with a heterozygous *JAK2* mutation express wild-type (WT) *JAK2* and mutant *JAK2*. We accounted for this when computing the fraction of mutated cells in specific subpopulations.

Using the expression levels of marker genes, we identified all major hematopoietic lineage progenitors in all samples (Figures 2A and 2B) and found that individuals with *JAK2*-mutant MPNs showed a similar hematopoietic differentiation hierarchy as healthy controls (Figure S2). *JAK2*-V617F cells had similar gene expression profiles compared with WT cells from the same individuals and were found to be generally intermixed with the WT cells in UMAP visualizations (Figure 2C). A significant fraction of the HSC subpopulation was mutated in all individuals (ranging from 5%–62%). Interestingly, we found that the *JAK2*-V617F allele fraction varied in different myeloid compartments

in the same individual. The *JAK2*-V617F mutation frequency was higher in megakaryocyte/erythroid progenitors and lower in lymphoid progenitors and granulocyte-macrophage progenitors (GMPs) (combined $p < 10^{-10}$ for erythroid versus lymphoid and erythroid progenitors versus GMPs for all individuals with V617F mutation, Fisher's exact test with Fisher's method) (Figures 2C and 2D). In contrast, the *JAK2*-V617L mutation showed no significant megakaryocyte-erythroid lineage bias (Figures 2C and 2D). *TET2* mutations were amplified and identified similarly in the single-cell RNA-seq (scRNA-seq) libraries from individuals PV 2 and PV 3 and were present in *JAK2*-mutant cells in both individuals, suggesting that the *TET2* and *JAK2* mutations occurred in the same clone (Figure S2). Both individuals had a higher *JAK2* allele fraction than *TET2* allele fraction (Figure 1B; $p < 10^{-10}$ for PV 2, $p = 0.003$ for PV 3, Fisher's exact test).

In clinical practice, MPN clone size can be approximated by the peripheral blood *JAK2*-V617F variant allele frequency (VAF), which reflects the fraction of nucleated blood cells that harbor the mutation but does not measure the contribution from anucleated mature red blood cells and platelets (Steensma, 2006). Our single-cell analysis demonstrates that the peripheral blood VAF consistently underestimates the degree to which *JAK2* mutant cells contribute to steady-state erythropoiesis (Figure 2E). Indeed, the large fraction of *JAK2*-V617F mutant erythroid progenitor cells in individuals with PV (79% to more than 95%) suggests that nearly all erythropoiesis arises from the *JAK2*-V617F clone. Additionally, most of the erythroid progenitors in ET 1 and ET 2 are *JAK2* mutant, suggesting that the *JAK2*-V617F mutation induces an erythroid fate bias even in individuals diagnosed with ET. Furthermore, the peripheral blood VAF does not accurately reflect the extent of disease in HSCs (Figure 2E).

We identified a CD14+ population in our scRNA-seq data that showed enhanced expression of type I-II interferon-regulated genes in individuals with MPNs relative to healthy controls (Figure 2F; Figure S2). These cells did not express CD34 mRNA and therefore likely represent a contaminating CD34− monocyte-like bone marrow population. Interestingly, this CD14+ bone marrow population showed increased expression of *SLAMF7* compared with healthy controls (particularly in individuals with PV). *SLAMF7* is a cell surface protein reported recently to be highly expressed on monocytes from individuals with established *JAK2* mutant myelofibrosis (Maekawa et al., 2019).

Our observations show that *JAK2*-V617F HSPCs have gene expression profiles similar to those of WT HSPCs but show a clear bias toward the megakaryocyte-erythroid fate. In addition, a significant fraction of HSCs was mutated in each individual. To understand how this population of mutated stem cells emerged, we set out to determine when the *JAK2*-V617F mutation first occurred in each individual and how the population of mutated stem cells subsequently expanded.

Lineage Trees of Individual Mutated and WT Stem Cells

To infer the disease history prior to clinical presentation with MPN, we reconstructed the lineage trees of the *JAK2* mutant HSCs of two individuals with ET, ET 1 and ET 2, using the pattern of somatic mutations accrued by individual cells. Somatic mutations occur at random and are passed to a cell's descendants and can be used to establish lineage relationships. To read out the somatic mutations in each cell, we isolated individual HSCs and multipotent progenitor (MPP) cells using established cell surface markers (STAR Methods) from the CD34+-enriched bone marrow cells of ET 1 and ET 2. We then expanded each HSC or MPP cell *ex vivo* by culturing them for ~8 weeks and performed whole-genome sequencing (WGS) on the single-cell colonies (STAR Methods). We selected colonies to balance the number of *JAK2* mutant and WT cells sequenced: 22 *JAK2* mutant colonies and 20 WT colonies for ET 1 and 13 *JAK2* mutant colonies and 21 WT colonies for ET 2. We observed that *JAK2* mutant HSCs likely had a proliferative advantage under our culture conditions because the fraction of *JAK2* mutant HSCs and MPPs after culturing was higher than the fraction of *JAK2* mutant HSCs found in the scRNA-seq data (10%/2% of cultured MPPs and 73%/32% of cultured HSCs versus 29%/8% of HSCs identified by scRNA-seq for ET 1/ET 2, respectively; see STAR Methods for additional information).

We found that the younger individual (ET 1, 34 years old) had, on average, 713 ± 45 somatic point mutations in individual HSCs/MPPs, whereas the older individual (ET 2, 63 years old) had 1,185 ± 75 mutations in each cell. Using the number of point mutations found in each cell and the age of each individual, we estimated a constant somatic point mutation rate of 19 ± 1 per year, consistent with previous observations in healthy donors (Lee-Six et al., 2018; Osorio et al., 2018; Figure S3). In both individuals, the number of somatic mutations in *JAK2* mutant (ET 1, 732 ± 26; ET 2, 1,209 ± 35) and *JAK2* WT cells (ET 1, 690 ± 52; ET 2, 1,170 ± 89) was comparable, and after accounting for the shared ancestry of cancer cells, the difference in mutation rate was not significant (ET 1, $p = 0.06$; ET 2, 0.21; Figure S3; STAR Methods). Although the number of HSPCs analyzed in this study is limited, our results suggest that the somatic mutation rate was not altered by the *JAK2*-V617F mutation. However, the fraction of C > T mutations at NpCpG trinucleotides was increased significantly in *JAK2* mutant HSPCs in ET 2 ($p < 0.01$; Wilcoxon rank sum test), and the average telomere length was shorter in *JAK2* mutant HSPCs in both individuals ($p < 0.01$, Wilcoxon rank-sum test; Figure S3; STAR Methods), suggesting that *JAK2* mutant cells might have undergone more cell divisions than *JAK2* WT cells (Alexandrov et al., 2015). The number of somatic mutations was similar in HSCs and MPPs in both individuals ($p = 0.07$ for ET 1, $p = 0.21$ for ET 2; STAR Methods). No somatic structural variants or copy number aberrations were detected, except for loss of one copy of chromosome X in one colony from individual ET 2.

Analysis of the single-base substitution (SBS) mutation signatures revealed that spontaneous aging-associated clock mutations (COSMIC signatures 1 and 5) predominated in WT and *JAK2*-V617F HSCs/MPPs (Alexandrov et al., 2020; Figures 3A and 3B), consistent with previous analyses of somatic mutations in healthy HSPCs (Lee-Six et al., 2018; Machado et al., 2019; Osorio et al., 2018). Other than *JAK2*-V617F, no deleterious somatic mutations were detected in *JAK2* mutant cells. The other mutations we identified that occurred in genes that could potentially affect stem cell function, such as *ASH1L* and *FAT1*, were not predicted to affect protein function nor have they been reported previously to be pathogenic (STAR Methods). Therefore,

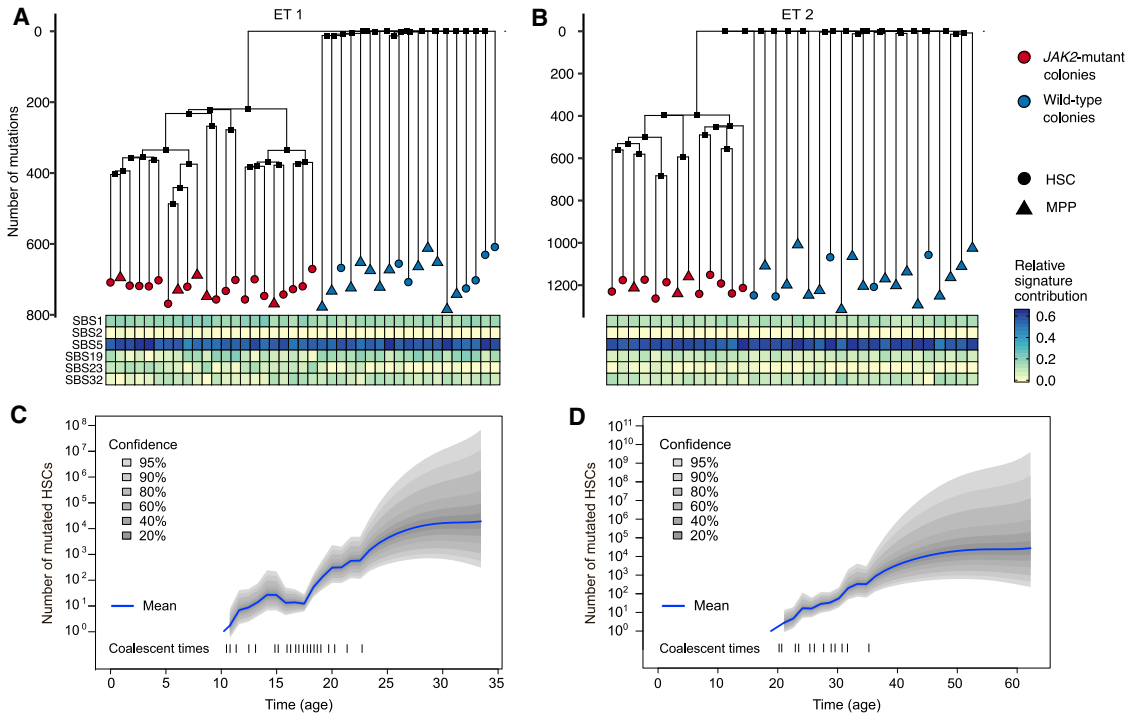


Figure 3. Somatic Mutations Can Be Used to Reconstruct the Lineage Trees of WT and Mutant HSCs

(A and B) Lineage trees constructed using somatic SNVs for ET 1 (A) and ET 2 (B). The heatmap below the lineage trees shows the relative contribution of the SBS mutational signatures SBS1, SBS2, SBS5, SBS19, SBS23, and SBS32 (STAR Methods) to the mutational spectrum defined by the private mutations detected in each HSC-derived colony.

(C) Number of mutant stem cells as a function of time inferred from the ET 1 lineage tree, assuming one generation per year. The dashed lines on the bottom show the times of the coalescent events in the tree.

(D) Same as (C) but for individual ET 2

See also Figure S3 and Table S2.

the *JAK2*-V617F mutation is likely the disease-initiating MPN driver mutation in these two individuals.

Next we used Wagner parsimony to reconstruct the phylogenies of the stem cells from the pattern of somatic mutations (Figures 3A and 3B; Table S2; STAR Methods). Two distinct clades were found in each individual that were defined by the presence or absence of the heterozygous *JAK2*-V617F mutation. These phylogenies suggest that, in both individuals, a single *JAK2* mutation event initiated the disease, followed by expansion of the mutated stem cells. No mutations were shared across all *JAK2* WT stem cells in ET 1 or in ET 2, suggesting that the common ancestor of *JAK2* WT stem cells dates back to embryonic development, before most somatic mutations occurred. However, there were many mutations shared across *JAK2* mutant stem cells (220 in ET 1 and 398 in ET 2), indicating that all mutated cells descended from a single common ancestor in which the *JAK2* mutation occurred. Using the inferred somatic mutation rate, we estimated that the disease-initiating mutation occurred ~25 years prior to sampling in ET 1 (the 34-year-old individual) and ~40 years prior to sampling in ET 2 (the 63-year-old individual).

Reconstructing the History of Disease Progression in Individuals

To reconstruct the history of disease development, we inferred the number of mutated stem cells in each individual from the time of the initial *JAK2* mutation event to the time of sampling by applying a phylogenetic dynamics inference algorithm (Karcher et al., 2017; Lee-Six et al., 2018) to the reconstructed lineage trees. This algorithm assumes that all mutated HSCs in the population are equivalent and that the population size over time can be modeled as a Gaussian process. Additionally, we assumed that HSCs divide symmetrically once per year (Abkowitz et al., 1996; Catlin et al., 2011; Lee-Six et al., 2018; Osorio et al., 2018); this assumption is required to infer the absolute population size but not to infer the rate of expansion of the population of mutated cells. No additional assumptions are made; e.g., whether the population of mutated cells expands or shrinks or at what rate. We found, in both individuals, that fewer than 100 mutated stem cells were present in the first decade after the *JAK2* mutation occurred. The number of mutated stem cells in both individuals grew exponentially for decades (Figures 3C and 3D).

To quantitatively estimate the difference in growth rates between *JAK2* WT and *JAK2* mutant HSCs *in vivo*, we constructed a mathematical model of stem cell self-renewal based on the Wright-Fisher model (Figure 4A; STAR Methods). Importantly, the Wright-Fisher model can be simulated efficiently to infer its parameter from the observed lineage trees (STAR Methods). Our model contains three parameters: the maximum number of mutated stem cells, the age at which the disease-initiating

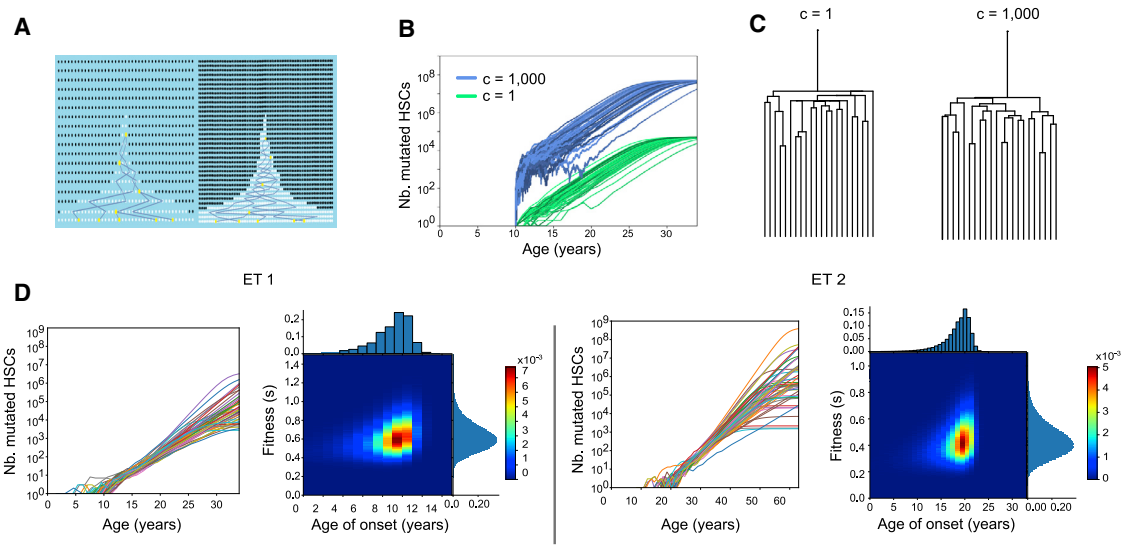


Figure 4. The History of *JAK2*-Mutant HSC Expansion Is Reconstructed from the Lineage Trees

(A) Schematics showing the effect of scaling the number of generations by a factor of 2 while keeping the onset of the disease and fitness the same. As a result, the number of mutant cells doubles because early on the number of mutant cells (shown in white) fluctuates to increase by a factor of 2 to escape stochastic extinction. Increasing the number of generations increases the coalescent rate. Increasing the number of mutant cells decreases the coalescent rate. These effects cancel each other, and the trees are indistinguishable.

(B) Green curves (c = 1) represent 50 simulated mutant HSC trajectories that survived extinction with fitness $s = 0.8$ and a maximum population size of 50,000. Blue curves are similar, except the number of generations and maximum population size are scaled by a factor of 1,000 (c = 1,000). This scaling results in 1,000 times as many mutant HSCs through time (blue) because a larger initial population is needed to escape stochastic extinction.

(C) Trees corresponding to the blue and green trajectories in (B) are statistically indistinguishable (STAR Methods).

(D) Inference on data from ET 1 and ET 2. For both individuals, we show 50 inferred trajectories of the number of mutant stem cells as a function of time. Heatmaps show the inferred joint distribution of the fitness of the cancer cells and the age when the disease initiating mutation occurred. The marginal distributions are shown as histograms.

See also Figure S4.

JAK2 mutation occurred, and the fitness *s* of the mutant stem cells, corresponding to the proliferative advantage of the mutated stem cells over WT stem cells. This model assumes that all mutated stem cells have the same constant fitness value. If the mutant stem cell population survives stochastic extinction early, when its size is small, it will grow exponentially as $(1+s)^t$, where t is time in years, until the mutant cells take over the majority of the population (Methods S1). Critically, unlike population size (Kimura, 1983; Lee-Six et al., 2018), fitness *s* and the time of occurrence of disease can be inferred without any knowledge of the HSC division rate (Figures 4A–4C; Methods S1). This is because changing the division rate scales the inferred population size and the minimum population size required to evade stochastic extinction in the same way. These two effects cancel each other, and *s* can be inferred directly from the observed lineage trees without knowledge of the division rate.

The model parameters were inferred from the reconstructed single-cell WGS lineage trees using approximate Bayesian computation (ABC). Briefly, we chose the parameter values from a prior distribution, simulated the model, randomly sampled a subset of cells, and obtained their lineage tree (Methods S1). We then compared the simulated tree with the observed tree (as measured by the lineage-through-time metric). If the simulated and observed trees were sufficiently similar, then we retained the parameter values. Otherwise, they were discarded, obtaining the posterior distribution for the parameter values (STAR Methods). We validated the inference procedure using simulated data, particularly in cases in which the observed tree was generated using slightly different dynamics than those used for ABC (Methods S1). In all cases, we were still able to accurately infer the effective fitness and age of onset. Therefore, our inference framework is robust to expected deviations in the actual dynamics of stem cell proliferation from the simplified Wright-Fisher model.

We then applied the inference procedure to the observed lineage trees of the *JAK2* mutant stem cells from ET 1 and ET 2 (Figure 4D; Figure S4). We inferred that in ET 1, the *JAK2*-V617F mutation first occurred at age 9 ± 2 and had a fitness effect of 63% ± 15%. Similarly, in ET 2, the *JAK2*-V617F mutation first occurred at age 19 ± 3 and had a fitness effect of 44% ± 13%. Our analyses show that *JAK2* mutant HSCs have a selective advantage over WT HSCs and increase in number over the decades before MPN diagnosis.

DISCUSSION

To determine the precise effect of the *JAK2*-V617F mutation on the behavior of human HSCs in their native bone marrow microenvironment, we performed WGS and single-cell profiling on HSPCs isolated from the bone marrow of newly diagnosed individuals with MPNs. Although it has long been known that MPN are clonal disorders (Adamson et al., 1976; Gilliland et al., 1991)

and it has been shown previously that the *JAK2*-V617F mutation is detectable in HSPCs in MPNs (Jamieson et al., 2006) and that *JAK2*-V617F cells are clonal in MPNs (Beer et al., 2009), in this study, we trace acquisition of the *JAK2*-V617F mutation to a single HSC decades before MPN diagnosis. Subsequently, the population of *JAK2* mutant stem cells grows exponentially but may exhibit large fluctuations and even stochastic extinction when its size is small in the first few years after occurrence of the mutation.

Using the WGS data, we estimated the time interval between *JAK2*-V617F acquisition and MPN development in individuals. Our findings that the *JAK2*-V617F mutation occurred in the first decade of life (9 ± 2 years) in a man who developed ET at age 34 and in the second decade of life (19 ± 3 years) in a woman who developed ET at age 63 are striking in terms of the young age at the time of *JAK2*-V617F acquisition and the decades-long interval to MPN development in both cases. While our paper was under review, similar findings were reported in an independent study (Williams et al., 2020).

We found that, at the time of MPN diagnosis, a significant fraction of HSCs (5% or more) are descendants of the original *JAK2* mutant HSC. In addition, we inferred the fitness advantage of *JAK2* mutant HSCs in 2 individuals with MPNs during the pre-diagnosis period to be approximately 63% ± 15% and 44% ± 13% in individuals ET 1 and ET 2, respectively. We emphasize that this fitness is inferred from the coalescent structure of the HSC lineage trees. Because most of the observed coalescent events occur close to the root of the tree (when the number of mutated stem cells is low), the inferred fitness value reflects the growth rate in the first decade after occurrence of the mutation. We did not sample enough HSCs to determine whether or when the growth in the population of mutated stem cells saturates. Our inferred fitness advantage is larger than that found in a population-level study of clonal hematopoiesis of indeterminate potential (CHIP), which analyzed peripheral blood variant allele fractions in large cohorts of healthy individuals (Watson et al., 2020). This discrepancy suggests that development of full-blown MPNs may require a faster-growing *JAK2* mutant clone than that observed in clonal hematopoiesis.

Our study focused on newly diagnosed, treatment-naive individuals with *JAK2* mutant ET and PV. We found that, in addition to modifying HSC proliferation dynamics, the *JAK2*-V617F mutation also affects the differentiation dynamics of their progenies. In contrast to a recent study of myelofibrosis that found marked aberrant megakaryopoiesis (Psaila et al., 2020), our study indicates that the hematopoietic differentiation hierarchy is largely preserved in individuals with ET and PV at diagnosis. However, *JAK2* mutant HSPCs showed a lineage bias toward the megakaryocyte-erythroid fate, and, strikingly, we found that the fraction of *JAK2* mutant cells varied significantly across different progenitor cell populations in the same patient. This suggests that peripheral blood monitoring may not accurately assess *JAK2* mutant allele burden, particularly in megakaryocyte-erythroid lineage cells and HSCs. Finally, the *JAK2*-V617F mutation has been shown to have cell-intrinsic effects not only in leukocytes (Rampal et al., 2014; Wolach et al., 2018) but also in erythroid cells (Chen et al., 2010; De Grandis et al., 2013) and platelets (Gangaraju et al., 2020; Guo et al., 2020). Our observation of high *JAK2*-V617F allele fractions in megakaryocyte-erythroid lineage cells in individuals with low peripheral blood *JAK2*-V617F mutational burden may help explain the development of thrombosis in these people.

Many cancers start when a genetic alteration arises in a single cell and confers a fitness advantage over other cells. By the time the disease manifests clinically, this cell has expanded to millions of cells or more. Naturally occurring somatic mutations provide a glimpse into the history of cancer in each individual, revealing when the driver mutations first occurred, how the population of cancer cells expanded, and how their proliferation and differentiation dynamics differ from healthy cells. The framework we developed to harness somatic mutations as a clock to reconstruct the lineage tree of cancer cells and follow the differentiation trajectories of their progenies is broadly applicable in oncology.

Limitations of Study

Although our study traced the acquisition of the *JAK2*-V617F mutation in newly diagnosed individuals with MPNs, it has some clear limitations. First, we studied a total of seven individuals with ET/PV and performed WGS to reconstruct lineage trees for two of these individuals with ET. It would be informative to expand the study to a larger cohort of individuals with MPNs and study sequential samples from the same person (including during *JAK2* mutant clonal hematopoiesis before development of MPNs). Second, we sequenced the whole genome of a limited number of HSCs from each individual. Although a sufficient number of coalescent events were observed to infer the behavior of mutated cells early in their history, extending the number of cells analyzed will provide further insights into the behavior of these cells closer to the time of diagnosis. More broadly, we inferred the history of disease expansion from final time point measurements. Any inference framework is only valid up to its assumptions. With larger lineage trees, we may be able to relax some of our assumptions and identify variations in fitness across the mutated stem cells over time.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead Contact
  - Materials Availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Human bone marrow biopsies
  - Cell lines
- METHOD DETAILS
  - Mononuclear cell isolation
  - CD34+ enrichment
  - Single-cell cDNA libraries
  - Locus-specific single-cell amplicon libraries
  - Stem cell genotyping and preparation for WGS
  - Phylodynamic inference
  - Inference of JAK2 mutant HSC fitness
- QUANTIFICATION AND STATISTICAL ANALYSIS

- scRNA-seq preprocessing and cell type identification
- Differential gene expression analysis between CD14+ cells from different patient groups
- Identification of JAK2 mutant cells in the scRNA-seq data
- Whole-genome sequencing data analysis
- Detection of somatic single-nucleotide variants and INDELs
- Detection of microsatellite mutations
- Detection of somatic structural variants
- Somatic copy number calling
- Mutational signature analysis
- Telomere length estimation
- Comparing the mutation rate between JAK2 mutant and JAK2-WT colonies
- Inference and validation of phylogenetic trees

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.stem.2021.02.001.

ACKNOWLEDGMENTS

We thank the individuals who participated in our study. We thank Drs. David Weinstock and Julie-Aurore Losman for valuable feedback on the manuscript. Portions of this research were conducted on the O2 High Performance Compute Cluster, supported by the Research Computing Group, at Harvard Medical School (https://it.hms.harvard.edu/our-services/research-computing/). S.H. acknowledges funding from NIH NIGMS R00GM118910 and NIH NHLBI R01HL158269, the DFCI BCB Fund Award, the Jayne Koskinas Ted Giovanis Foundation, The William F. Milton Fund at Harvard University, an AACR-MPM Oncology Charitable Foundation Transformative Cancer Research grant, and Gabrielle's Angel Foundation for Cancer Research. S.H. and A.M. acknowledge funding from the Claudia Adams Barr Program in Cancer Research. A.M. acknowledges funding from NIH NHLBI (R01HL131835) and the MPN Research Foundation. A.M. is a Scholar of The Leukemia & Lymphoma Society. C.R.R acknowledges funding from NIH NHLBI T32HL116324. D.V.E. acknowledges funding from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard, award number 1764269, and the Harvard Quantitative Biology Initiative. D.V.E and F.M. acknowledge support from the Ludwig Center at Harvard and the Dana-Farber Cancer Institute Physical Science-Oncology Center (NIH U54CA193461 to F.M.). I.C.-C. and M.K. acknowledge funding from EMBL. J.E. acknowledges funding from NIH NIGMS R25GM109436. F.D.C. is supported by a grant from the Edward Evans MDS Foundation and NIH NHLBI P01HL131477. We thank Dr. Francesc Mutas Remolar for fruitful discussions.

AUTHOR CONTRIBUTIONS

R.M.S., A.M., and S.H. conceived the project. M.N., S.L., and S.H. designed the experiments. M.N. devised and optimized the single-cell amplicon sequencing protocol with help from S.L. and supervision from S.H. M.N. and S.L. processed the samples and generated all sequencing libraries with help from B.K. C.R.R., G.S.H., and A.M. devised the individual selection criteria. D.J.D., I.G., M.W., E.S.W., M.R.L., R.M.S., J.S.G., G.S.H., and A.M. helped obtain samples, coordinated by C.R.R. D.V.E. analyzed all the single-cell data with help from C.R.R. and B.K., supervised by F.M. and S.H. M.K. and I.C.-C. analyzed the whole-genome sequencing data and reconstructed the lineage trees. S.P. isolated and cultured individual HSCs, supervised by F.D.C. J.E. devised and implemented the algorithms for inference of growth dynamics from lineage trees, supervised by S.H. D.V.E., J.E., M.N., S.L., C.R.R., F.M., A.M., I.C.-C., and S.H. wrote the manuscript with input from all authors. A.M., I.C.-C., and S.H. supervised the project.

DECLARATION OF INTERESTS

A.M. has consulted for Janssen, PharmaEssentia, Constellation, and Relay Therapeutics and receives research support from Janssen and Actuate Therapeutics. E.S.W. reports personal fees from Jazz Pharmaceuticals, Takeda Pharmaceutical Company, Novartis, and Pfizer. F.M. is the co-founder of an oncology company. J.S.G. has consulted for AbbVie, Takeda, and Astellas and receives research support from AbbVie, Genentech, Prelude, AstraZeneca, and Eli Lilly. D.J.D. receives research support from Glycomimetics, Novartis, AbbVie, and Blueprint Medicines and has consulted for Incyte, Jazz, Novartis, Pfizer, Shire, Takeda, Amgen, Forty-Seven, Agios, Autolos, and Blueprint Medicines. G.S.H. has received research support from Bayer, Merck, Incyte, and Constellation and has received honoraria from Constellation, Jazz, Novartis, and Celgene/BMS. R.M.S. has advisory board, DSMB, and/or steering committee membership at Syntrix/ACI Clinical, Takeda, Elevate Bio, Syndax Pharma, AbbVie, Syros, Gemoab, BerGenBio, Foghorn Thera, GSK, Aprea, Innate, Actinium, and OncoNova.

Received: October 19, 2020
Revised: December 1, 2020
Accepted: January 28, 2021
Published: February 22, 2021

SUPPORTING CITATIONS

The following references appear in the supplemental information: Fisher (1923); Kimura (1962); Kingman (1982a); Kingman (1982b); Kingman (1982c); Wright, (1931).

REFERENCES

Abkowitz, J.L., Catlin, S.N., and Guttorp, P. (1996). Evidence that hematopoiesis may be a stochastic process in vivo. Nat. Med. *2*, 190–197.

Adamson, J.W., Fialkow, P.J., Murphy, S., Prchal, J.F., and Steinmann, L. (1976). Polycythemia vera: stem-cell and probable clonal origin of the disease. N. Engl. J. Med. *295*, 913–916.

Alexandrov, L.B., Jones, P.H., Wedge, D.C., Sale, J.E., Campbell, P.J., Nik-Zainal, S., and Stratton, M.R. (2015). Clock-like mutational processes in human somatic cells. Nat. Genet. *47*, 1402–1407.

Alexandrov, L.B., Kim, J., Haradhvala, N.J., Huang, M.N., Tian Ng, A.W., Wu, Y., Boot, A., Covington, K.R., Gordenin, D.A., Bergstrom, E.N., et al.; PCAWG Mutational Signatures Working Group; PCAWG Consortium (2020). The repertoire of mutational signatures in human cancer. Nature *578*, 94–101.

Baxter, E.J., Scott, L.M., Campbell, P.J., East, C., Fourouclas, N., Swanton, S., Vassiliou, G.S., Bench, A.J., Boyd, E.M., Curtin, N., et al.; Cancer Genome Project (2005). Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. Lancet *365*, 1054–1061.

Beer, P.A., Jones, A.V., Bench, A.J., Goday-Fernandez, A., Boyd, E.M., Vaghela, K.J., Erber, W.N., Odeh, B., Wright, C., McMullin, M.F., et al. (2009). Clonal diversity in the myeloproliferative neoplasms: independent origins of genetically distinct clones. Br. J. Haematol. *144*, 904–908.

Blokzijl, F., Janssen, R., van Boxtel, R., and Cuppen, E. (2018). MutationalPatterns: comprehensive genome-wide analysis of mutational processes. Genome Med. *10*, 33.

Campbell, P.J., Getz, G., Korbel, J.O., Stuart, J.M., Jennings, J.L., Stein, L.D., Perry, M.D., Nahal-Bose, H.K., Ouellette, B.F.F., Li, C.H., et al.; ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium (2020). Pan-cancer analysis of whole genomes. Nature *578*, 82–93.

Catlin, S.N., Busque, L., Gale, R.E., Guttorp, P., and Abkowitz, J.L. (2011). The replication rate of human hematopoietic stem cells in vivo. Blood *117*, 4460–4466.

Chen, E., Beer, P.A., Godfrey, A.L., Ortmann, C.A., Li, J., Costa-Pereira, A.P., Ingle, C.E., Dermitzakis, E.T., Campbell, P.J., and Green, A.R. (2010). Distinct clinical phenotypes associated with JAK2V617F reflect differential STAT1 signaling. Cancer Cell *18*, 524–535.

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Källberg, M., Cox, A.J., Kruglyak, S., and Saunders, C.T. (2016). Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. Bioinformatics *32*, 1220–1222.

Cordua, S., Kjaer, L., Skov, V., Pallisgaard, N., Hasselbalch, H.C., and Ellervik, C. (2019). Prevalence and phenotypes of *JAK2* V617F and *calreticulin* mutations in a Danish general population. Blood *134*, 469–479.

De Grandis, M., Cambot, M., Wautier, M.-P., Cassinat, B., Chomienne, C., Colin, Y., Wautier, J.-L., Le Van Kim, C., and El Nemer, W. (2013). JAK2V617F activates Lu/BCAM-mediated red cell adhesion in polycythemia vera through an EpoR-independent Rap1/Akt pathway. Blood *121*, 658–665.

Delhommeau, F., Dupont, S., Tonetti, C., Massé, A., Godin, I., Le Couedic, J.P., Debili, N., Saulnier, P., Casadevall, N., Vainchenker, W., and Giraudier, S. (2007). Evidence that the JAK2 G1849T (V617F) mutation occurs in a lymphomyeloid progenitor in polycythemia vera and idiopathic myelofibrosis. Blood *109*, 71–77.

DePristo, M.A., Banks, E., Poplin, R., Garimella, K.V., Maguire, J.R., Hartl, C., Philippakis, A.A., del Angel, G., Rivas, M.A., Hanna, M., et al. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat. Genet. *43*, 491–498.

Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet. *24*, 2125–2137.

Dusa, A., Staerk, J., Elliott, J., Pecquet, C., Poirel, H.A., Johnston, J.A., and Constantinescu, S.N. (2008). Substitution of pseudokinase domain residue Val-617 by large non-polar amino acids causes activation of JAK2. J. Biol. Chem. *283*, 12941–12948.

Farmery, J.H.R., Smith, M.L., and Lynch, A.G.; NIHR BioResource - Rare Diseases (2018). Telomerecat: A ploidy-agnostic method for estimating telomere length from whole genome sequencing data. Sci. Rep. *8*, 1300.

Fisher, R.A. (1923). XXI.—On the Dominance Ratio. Proc. R. Soc. Edinb. *42*, 321–341.

Gangaraju, R., Song, J., Kim, S.J., Tashi, T., Reeves, B.N., Sundar, K.M., Thiagarajan, P., and Prchal, J.T. (2020). Thrombotic, inflammatory, and HIF-regulated genes and thrombosis risk in polycythemia vera and essential thrombocythemia. Blood Adv. *4*, 1115–1130.

Genovese, G., Kähler, A.K., Handsaker, R.E., Lindberg, J., Rose, S.A., Bakhoum, S.F., Chambert, K., Mick, E., Neale, B.M., Fromer, M., et al. (2014). Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. N. Engl. J. Med. *371*, 2477–2487.

Gilliland, D.G., Blanchard, K.L., Levy, J., Perrin, S., and Bunn, H.F. (1991). Clonality in myeloproliferative disorders: analysis by means of the polymerase chain reaction. Proc. Natl. Acad. Sci. USA *88*, 6848–6852.

Grinfeld, J., Nangalia, J., Baxter, E.J., Wedge, D.C., Angelopoulos, N., Cantrill, R., Godfrey, A.L., Papaemmanuil, E., Gundem, G., MacLean, C., et al. (2018). Classification and Personalized Prognosis in Myeloproliferative Neoplasms. N. Engl. J. Med. *379*, 1416–1430.

Guo, B.B., Linden, M.D., Fuller, K.A., Phillips, M., Mirzai, B., Wilson, L., Chuah, H., Liang, J., Howman, R., Grove, C.S., et al. (2020). Platelets in myeloproliferative neoplasms have a distinct transcript signature in the presence of marrow fibrosis. Br. J. Haematol. *188*, 272–282.

Hinds, D.A., Barnholt, K.E., Mesa, R.A., Kiefer, A.K., Do, C.B., Eriksson, N., Mountain, J.L., Francke, U., Tung, J.Y., Nguyen, H.M., et al. (2016). Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. Blood *128*, 1121–1128.

Ishii, T., Bruno, E., Hoffman, R., and Xu, M. (2006). Involvement of various hematopoietic-cell lineages by the JAK2V617F mutation in polycythemia vera. Blood *108*, 3128–3134.

Jaiswal, S., Fontanillas, P., Flannick, J., Manning, A., Grauman, P.V., Mar, B.G., Lindsley, R.C., Mermel, C.H., Burtt, N., Chavez, A., et al. (2014). Age-related clonal hematopoiesis associated with adverse outcomes. N. Engl. J. Med. *371*, 2488–2498.

Jaiswal, S., Natarajan, P., Silver, A.J., Gibson, C.J., Bick, A.G., Shvartz, E., McConkey, M., Gupta, N., Gabriel, S., Ardissino, D., et al. (2017). Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. N. Engl. J. Med. *377*, 111–121.

James, C., Ugo, V., Le Couédic, J.-P., Staerk, J., Delhommeau, F., Lacout, C., Garçon, L., Raslova, H., Berger, R., Bennaceur-Griscelli, A., et al. (2005). A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature *434*, 1144–1148.

Jamieson, C.H.M., Gotlib, J., Durocher, J.A., Chao, M.P., Mariappan, M.R., Lay, M., Jones, C., Zehnder, J.L., Lilleberg, S.L., and Weissman, I.L. (2006). The JAK2 V617F mutation occurs in hematopoietic stem cells in polycythemia vera and predisposes toward erythroid differentiation. Proc. Natl. Acad. Sci. USA *103*, 6224–6229.

Karcher, M.D., Palacios, J.A., Bedford, T., Suchard, M.A., and Minin, V.N. (2016). Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference. PLoS Comput. Biol. *12*, e1004789.

Karcher, M.D., Palacios, J.A., Lan, S., and Minin, V.N. (2017). phylodyn: an R package for phylodynamic simulation and inference. Mol. Ecol. Resour. *17*, 96–100.

Kimura, M. (1962). On the probability of fixation of mutant genes in a population. Genetics *47*, 713–719.

Kimura, M. (1983). The Neutral Theory of Molecular Evolution (Cambridge University Press).

Kingman, J.F.C. (1982a). Exchangeability and the evolution of large populations. In Proceedings of the International Conference on Exchangeability in Probability and Statistics, B.D. Finetti, G. Koch, and F. Spizzichino, eds. (North-Holland Publishing Company), pp. 97–112.

Kingman, J.F.C. (1982b). On the genealogy of large populations. In Essays in Statistical Science, A.P. Moran, J.M. Gani, and E.J. Hannan, eds. (Applied Probability Trust), pp. 27–43.

Kingman, J.F.C. (1982c). The coalescent. Stochastic Process. Appl. *13*, 235–248.

Kluk, M.J., Lindsley, R.C., Aster, J.C., Lindeman, N.I., Szeto, D., Hall, D., and Kuo, F.C. (2016). Validation and Implementation of a Custom Next-Generation Sequencing Clinical Assay for Hematologic Malignancies. J. Mol. Diagn. *18*, 507–515.

Kralovics, R., Passamonti, F., Buser, A.S., Teo, S.-S., Tiedt, R., Passweg, J.R., Tichelli, A., Cazzola, M., and Skoda, R.C. (2005). A gain-of-function mutation of JAK2 in myeloproliferative disorders. N. Engl. J. Med. *352*, 1779–1790.

Layer, R.M., Chiang, C., Quinlan, A.R., and Hall, I.M. (2014). LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. *15*, R84.

Lee-Six, H., Øbro, N.F., Shepherd, M.S., Grossmann, S., Dawson, K., Belmonte, M., Osborne, R.J., Huntly, B.J.P., Martincorena, I., Anderson, E., et al. (2018). Population dynamics of normal human blood inferred from somatic mutations. Nature *561*, 473–478.

Levine, R.L., Wadleigh, M., Cools, J., Ebert, B.L., Wernig, G., Huntly, B.J.P., Boggon, T.J., Wlodarska, I., Clark, J.J., Moore, S., et al. (2005). Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell *7*, 387–397.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv, arXiv:1303.3997 https://arxiv.org/abs/1303.3997.

Lundberg, P., Karow, A., Nienhold, R., Looser, R., Hao-Shen, H., Nissen, I., Girsberger, S., Lehmann, T., Passweg, J., Stern, M., et al. (2014). Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. Blood *123*, 2220–2228.

Machado, H.E., Øbro, N.F., Mitchell, E., Davies, M., Green, A.R., Saeb-Parsy, K., Hodson, D.J., Kent, D., and Campbell, P.J. (2019). Life History of Normal Human Lymphocytes Revealed By Somatic Mutations. Blood *134*, 1045–1045.

Maekawa, T., Kato, S., Kawamura, T., Takada, K., Sone, T., Ogata, H., Saito, K., Izumi, T., Nagao, S., Takano, K., et al. (2019). Increased SLAMF7^{high} monocytes in myelofibrosis patients harboring *JAK2*V617F provide a therapeutic target of elotuzumab. Blood *134*, 814–825.

Nam, A.S., Kim, K.-T., Chaligne, R., Izzo, F., Ang, C., Taylor, J., Myers, R.M., Abu-Zeinah, G., Brand, R., Omans, N.D., et al. (2019). Somatic mutations

and cell identity linked by Genotyping of Transcriptomes. Nature *571*, 355–360.

Osorio, F.G., Rosendahl Huber, A., Oka, R., Verheul, M., Patel, S.H., Hasaart, K., de la Fonteijne, L., Varela, I., Camargo, F.D., and van Boxtel, R. (2018). Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. Cell Rep. *25*, 2308–2316.e4.

Psaila, B., Wang, G., Rodriguez-Meira, A., Li, R., Heuston, E.F., Murphy, L., Yee, D., Hitchcock, I.S., Sousos, N., O'Sullivan, J., et al.; NIH Intramural Sequencing Center (2020). Single-Cell Analyses Reveal Megakaryocyte-Biased Hematopoiesis in Myelofibrosis and Identify Mutant Clone-Specific Targets. Mol. Cell *78*, 477–492.e8.

Raine, K.M., Van Loo, P., Wedge, D.C., Jones, D., Menzies, A., Butler, A.P., Teague, J.W., Tarpey, P., Nik-Zainal, S., and Campbell, P.J. (2016). ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. Curr. Protoc. Bioinformatics *56*, 15.9.1–15.9.17.

Rampal, R., Al-Shahrour, F., Abdel-Wahab, O., Patel, J.P., Brunel, J.-P., Mermel, C.H., Bass, A.J., Pretz, J., Ahn, J., Hricik, T., et al. (2014). Integrated genomic analysis illustrates the central role of JAK-STAT pathway activation in myeloproliferative neoplasm pathogenesis. Blood *123*, e123–e133.

Rausch, T., Zichner, T., Schlattl, A., Stütz, A.M., Benes, V., and Korbel, J.O. (2012). DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics *28*, i333–i339.

Steensma, D.P. (2006). JAK2 V617F in myeloid disorders: molecular diagnostic techniques and their clinical utility: a paper from the 2005 William Beaumont Hospital Symposium on Molecular Pathology. J. Mol. Diagn. *8*, 397–411, quiz 526.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888–1902.e21.

Tusi, B.K., Wolock, S.L., Weinreb, C., Hwang, Y., Hidalgo, D., Zilionis, R., Waisman, A., Huh, J.R., Klein, A.M., and Socolovsky, M. (2018). Population snapshots predict early haematopoietic and erythroid hierarchies. Nature *555*, 54–60.

Wala, J.A., Bandopadhayay, P., Greenwald, N.F., O'Rourke, R., Sharpe, T., Stewart, C., Schumacher, S., Li, Y., Weischenfeldt, J., Yao, X., et al. (2018). SvABA: genome-wide detection of structural variants and indels by local assembly. Genome Res. *28*, 581–591.

Watson, C.J., Papula, A.L., Poon, G.Y.P., Wong, W.H., Young, A.L., Druley, T.E., Fisher, D.S., and Blundell, J.R. (2020). The evolutionary dynamics and fitness landscape of clonal hematopoiesis. Science *367*, 1449–1454.

Willems, T., Zielinski, D., Yuan, J., Gordon, A., Gymrek, M., and Erlich, Y. (2017). Genome-wide profiling of heritable and de novo STR variations. Nat. Methods *14*, 590–592.

Williams, N., Lee, J., Moore, L., Baxter, E.J., Hewinson, J., Dawson, K.J., Menzies, A., Godfrey, A.L., Green, A.R., Campbell, P.J., et al. (2020). Phylogenetic reconstruction of myeloproliferative neoplasm reveals very early origins and lifelong evolution. bioRxiv. https://doi.org/10.1101/2020.11.09.374710.

Wolach, O., Sellar, R.S., Martinod, K., Cherpokova, D., McConkey, M., Chappell, R.J., Silver, A.J., Adams, D., Castellano, C.A., Schneider, R.K., et al. (2018). Increased neutrophil extracellular trap formation promotes thrombosis in myeloproliferative neoplasms. Sci. Transl. Med. *10*, eaan8292.

Wolf, F.A., Angerer, P., and Theis, F.J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. *19*, 15.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.

Xie, M., Lu, C., Wang, J., McLellan, M.D., Johnson, K.J., Wendl, M.C., McMichael, J.F., Schmidt, H.K., Yellapantula, V., Miller, C.A., et al. (2014). Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nat. Med. *20*, 1472–1478.

Yu, G. (2020). Using ggtree to Visualize Data on Tree-Like Structures. Curr. Protoc. Bioinformatics *69*, e96.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
CD34-PB	BioLegend	Cat# 343512; RRID:AB_1877197
CD38-PE-Cy7	Thermo Fisher Scientific	Cat# 25-0388-42; RRID:AB_2573346
CD45RA-FITC	BioLegend	Cat# 304106; RRID:AB_314410
CD49f-APC-Cy7	BioLegend	Cat# 313628; RRID:AB_2616784
Thy1/CD90-PE	BioLegend	Cat# 328110; RRID:AB_893433
EasySep Human CD34 Positive Selection Kit II	STEMCELL Technologies	Cat# 17856
Biological Samples		
Whole bone marrow samples	Massachusetts General Hospital; Dana-Farber Cancer Institute	N/A
Peripheral blood samples	Dana-Farber Cancer Institute	N/A
Chemicals, Peptides, and Recombinant Proteins		
Recombinant Human SCF	PeproTech	Cat#300-07
Recombinant Human TPO	PeproTech	Cat#300-18
Recombinant Human FLT3-L	PeproTech	Cat#300-19
Recombinant Human IL-6	PeproTech	Cat#200-06
Recombinant Human IL-3	PeproTech	Cat#160-01
Chromium Single Cell 3′ GEM, Library & Gel Bead Kit v3, 16 rxns	10x Genomics	Cat#1000075
Chromium Chip B Single Cell Kit, 48 rxns	10x Genomics	Cat#1000073
Chromium i7 Multiplex Kit, 96 rxns	10x Genomics	Cat#120262
Buffer EB	QIAGEN	Cat#19086
High Sensitivity D5000 ScreenTape	Agilent	Cat#5067-5592
High Sensitivity D5000 Reagents	Agilent	Cat#5067-5593
MiSeq Reagent Kit v2 (500 cycles)	Illumina	Cat#MS-102-2003
NovaSeq 6000 SP Reagent Kit (200 cycles)	Illumina	Cat#20040326
Lymphoprep	STEMCELL Technologies	Cat#07801
EasySep Buffer	STEMCELL Technologies	Cat#20144
SPRIselect	Beckman Coulter	Cat#B23318
QIAmp UCP DNA Micro Kit	QIAGEN	Cat#56204
Qubit dsDNA HS Assay Kit	Invitrogen	Cat#32854
Monarch DNA Gel Extraction Kit	New England Biolabs	Cat#T1020L
Dulbecco's phosphate-buffered saline	Thermo Scientific	Cat#14040133
SFEM medium	STEMCELL Technologies	Cat#09650
DMEM-F12 medium	GIBCO	Cat#11320082
Iscove Modified Dulbecco Medium (IMDM)	GIBCO	Cat#31980030
Roswell Park Memorial Institute (RPMI)	GIBCO	Cat#12633012
Horse Serum	Thermo Scientific	Cat#16050130
Hydrocortisone Solution	Sigma-Aldrich	Cat#H6909-10ML
Fetal Bovine Serum	VWR	Cat#89510-186
Deposited Data		
scRNA-seq and WGS data	dbGAP	phs002308.v1.p1
Experimental Models: Cell Lines		
UKE-1	Ann Mullally	N/A
Molt4	Sahand Hormoz	N/A

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Oligonucleotides		
Primers	This paper	Table S1
Software and Algorithms		
StemCellSim	This paper	https://gitlab.com/hormozlab/stemcellsim
Scripts used for scRNA-seq analysis	This paper	https://gitlab.com/hormozlab/mpn-scrnaseq-analysis

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Sahand Hormoz (sahand_hormoz@hms.harvard.edu).

Materials Availability

All unique/stable reagents generated in this study are available from the Lead Contact with a completed Materials Transfer Agreement.

Data and code availability

We developed a C++ object called StemCellSim for simulating clonal expansions and inferring the parameters of our model. StemCellSim has the ability to generate simulated data under various models, and to infer model parameters from either simulated or real data with ABC. The StemCellSim code and the Python scripts used to analyze and plot the scRNA-seq data can be found on GitLab (https://gitlab.com/hormozlab). Raw scRNA-seq and whole-genome sequencing data have been deposited in dbGAP:phs002308.v1.p1.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Human bone marrow biopsies

Prospective MPN patients were identified through the outpatient clinic. Patients were required to have a confirmed diagnosis of PV or ET according to WHO criteria and next-generation sequencing (RapidHEME panel or SnapShot) documenting the presence of *JAK2*-V617F. Use of anti-platelet agents was permitted but disease-modifying treatments (e.g., hydroxyurea, interferon-alpha, ruxolitinib) were an exclusion criterion for the study. Consequently, our cohort consisted of newly diagnosed, treatment-naive *JAK2* mutant MPN patients. Bone marrow aspirate samples were uniformly collected at the time of diagnostic bone marrow biopsy under tissue banking protocols at the participating centers. The study was approved by and conducted in accordance with the Declaration of Helsinki protocol (Dana-Farber Cancer Institute IRB protocol no. 01-206 and Massachusetts General Hospital protocol 13-583). All patients provided informed consent. The healthy donor samples were purchased as de-identified samples from the Boston Children's Hospital. Healthy donor 1 was a 22-year-old female and healthy donor 2 was a 29-year-old female. The age and sex of all other study subjects can be found in Figure 1B. Bone marrow biopsies were performed on all donors. RapidHEME panel screen on peripheral blood from the same patients was performed as a part of clinical diagnostic testing. Approximately 10-20mL of bone marrow aspirate from each donor was collected in EDTA-coated tubes. The syringes and tubes used were sterile with no preservative-free heparin coating. The bone marrow aspirates (BMA) were kept at room temperature until use.

Cell lines

The UKE-1 and MOLT4 cell lines used in the control experiments were collected from a 59-year-old female and a 19-year-old male. UKE-1 cells were maintained in Iscove modified Dulbecco medium (IMDM) supplemented with 10% fetal bovine serum, 10% horse serum and 1 μM hydrocortisone. MOLT4 cells were maintained in Roswell Park Memorial Institute (RPMI) medium supplemented with 10% fetal bovine serum. All cultures were maintained in standard tissue culture conditions of 37°C and 5% CO_2.

METHOD DETAILS

Mononuclear cell isolation

Mononuclear cells (MNCs) were isolated from the BMA via a density gradient centrifugation protocol using StemCell Technologies, Inc.'s SepMate system. BMA, phosphate-buffered saline (PBS) with 2% fetal bovine serum (PBS + 2% FBS; StemCell Technologies, Inc. #07905), Lymphoprep (StemCell Technologies, Inc. #07801), and the centrifuge (Eppendorf #5810R) were all acclimated at room temperature. Approximately 20-22mL Lymphoprep was added to the 50mL SepMate tube (StemCell Technologies, Inc. #15450) by carefully pipetting it through the central hole of the SepMate insert, ensuring that as few air bubbles as possible were present. BMA

was diluted with an equal volume of PBS + 2% FBS and mixed gently with wide-bore pipette. Keeping the SepMate tube vertical, the diluted sample was added by slowly pipetting it down the side of the tube. The diluted BMW was centrifuged at 1200 x g for 20 minutes at room temperature, with the brake off. For BMA rich in platelets/plasma, the top layer of platelets/plasma was pipetted off. The remaining volume down to the SepMate central hole (containing all the enriched MNCs) was poured into a new 50mL tube. The MNCs were washed by topping up until 45mL with PBS + 2% FBS and mixing well with a wide-bore pipette. The MNCs were then centrifuged at 300 x g for 12 minutes at room temperature, with the brake low, and the supernatant was removed. The MNC pellet was topped up again with PBS + 2% FBS, and the volume was mixed well with a wide-bore pipette. Centrifuge at 120 x g for 12 minutes at room temperature, with the brake off. The supernatant was then removed. For BMA rich in platelets, an additional wash and centrifugation at 12 minutes at room temperature, with the break off was performed. The MNC pellet was resuspended in 1mL of EasySep buffer (StemCell Technologies, Inc. #20144) and mixed with a wide-bore pipette. Cell concentration was counted with a hemocytometer (Reichert) and a Tali Image Cytometer system (ThermoFisher #T10796) using Tali Image Analysis Slides (ThermoFisher #T10794). Cells were placed on ice until further use.

CD34+ enrichment

CD34+ MNCs were isolated using the protocol for EasySep Human CD34 Positive Selection Kit II (StemCell Technologies, Inc. #17856). MNCs (at concentration of > 10^8 cells/mL EasySep buffer) were added to 5mL (12 × 75 mm) polystyrene round-bottom tube (StemCell Technologies, Inc. #38007). EasySep Human CD34 Positive Selection Cocktail (StemCell Technologies, Inc. #17856C) was added at a concentration of 100 μL per 1mL of sample. The sample was then mixed and incubated at room temperature for 10 minutes. EasySep Dextran RapidSpheres (StemCell Technologies, Inc. #50100) were vortexed for 30 s. RapidSpheres were added at a concentration of 75 μL per 1 mL of sample. The sample was mixed and incubated at room temperature for 5 minutes. The tube was topped up to 2.5mL with EasySep buffer and gently mixed. The tube was placed in EasySep magnet (StemCell Technologies, Inc. #18000) and incubated at room temperature for 3 minutes. The supernatant was discarded by inverting the magnet with the tube inside. This process was repeated 4 more times for a total of 5 rounds of enrichment. Cells were resuspended in PBS+2% FBS after the last round of enrichment. Cell concentration was counted with a hemocytometer (Reichert #1492) and Tali Image Cytometer. Cells were placed on ice until further use. CD34+ MNCs were used to create single-cell cDNA, single-cell RNA-Seq (scRNA-seq) and *JAK2* amplicon libraries as described below.

Single-cell cDNA libraries

The isolated CD34+ MNC suspensions were used to generate single-cell gel bead emulsions (GEMs) using a 10x Genomics Chromium controller (10x Genomics #120223). Following steps 1 and 2 of the protocol "Chromium Single Cell 3′ Reagent Kit v3" (10x Genomics, CG000183 Rev A), single-cell cDNA libraries were constructed using the Chromium Single Cell 3′ Library & Gel Bead Kit v3 (10x Genomics #1000075). The protocol yields 40 μL of cDNA per sample after step 2.4. scRNA-Seq libraries were generated from 10 μL of single-cell cDNA libraries using Step 3 of the Chromium Single Cell 3′ Reagent Kit v3 user guide (10x Genomics, CG000183 Rev A).

Locus-specific single-cell amplicon libraries

We developed the following protocol to preferentially amplify transcripts containing loci-of-interest from single-cell cDNA libraries (ex. *JAK2*-V617F), thereby generating locus-specific single-cell amplicon libraries (Figures S1A–S1C).

A triple-nested PCR approach was used to amplify the transcripts carrying the loci-of-interest from single-cell cDNA libraries with high sensitivity and specificity. The approach used locus-specific reverse primers that flank the mutation site combined with generic forward primers that preserves the single-cell barcoding structure. In total, there were 5 total PCR steps (3 nested steps that increasingly filtered for a specific transcript, 1 step that added a Read2 sequence, and 1 step that added an Illumina P7 adaptor sequence). In Step 1, a PCR was conducted using a forward primer containing both the Illumina P5 sequence and part of the Read 1 sequence and a reverse primer containing a locus-specific sequence approximately ~300bp upstream of the mutation site. In Step 2, a PCR was conducted using a shortened version of the forward primer in Step 1 and a reverse primer containing a locus-specific sequence approximately ~150bp upstream of the mutation site. In Step 3, a PCR was conducted using the forward primer from Step 2 and a reverse primer containing a locus-specific sequence approximately ~50bp upstream of the mutation site. In Step 4, a PCR was conducted using the forward primer from Step 2 and a reverse primer containing part of the locus-specific sequence used in Step 3 combined with a Read2 sequence. In Step 5, a PCR was conducted using the forward primer used in Step 1 and a reverse primer containing the Read2 sequence combined with the Illumina P7 sequence.

The following was the specific protocol used for constructing the *JAK2*-V617F amplicon libraries. All PCR's were conducted in TempAssure PCR tubes (USA Scientific #14024700) on a Bio-Rad C1000 Touch Thermal Cycler.

In Step 1, a 25uL PCR mixture was made containing 12.5uL of Amplification Master Mix (10x Genomics #220125), 1.25uL of cDNA additive (10x Genomics #220067), 1.25uL of forward primer (P5-Partial Read 1, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC) at 20uM, 1.25ul of reverse primer (Reverse Ext 1, ACCAACCTCACCAACATTACAGAGGCCT) at 10uM, 3ng of cDNA library material, and remaining volume with nuclease-free water. With the thermal cycler lid set to 105°C, the following thermal cycling protocol is used: initial denaturation at 98°C for 45 s, 10 cycles of denaturing at 98°C for 20 s, annealing at 67°C for 30 s, extension at 72°C for 180 s, and a final extension at 72°C for 60 s.

The PCR reaction mixture was purified using SPRIselect (Beckman Coulter #B23318) as follows. 20uL (i.e., 0.8x) of SPRIselect reagent was added to the reaction mixture, pipette mixed, and incubated at room temperature for 5 minutes. The PCR tube was placed on the 10x Magnetic Separator (10x Genomics #230003) on High until solution clears (usually about a minute). Supernatant was removed and discarded. 200uL of 80% ethanol in nuclease-free water was added to the pellet and allowed to sit for 30 s. The ethanol wash was removed and repeated once more. The PCR tube was briefly centrifuged and put back into the magnet at Low setting. Any remaining ethanol wash was removed, and the pellet was allowed to air dry for 1 minute. The DNA was then eluted by removing the PCR tube from the magnet, pipetting 20uL Buffer EB (QIAGEN #19086) onto the pellet, pipette mixing, allowing the mixture to equilibrate for 2 minutes, placing the tube on the magnet on Low, and then eluting the liquid into a new tube.

In Step 2, a 25uL PCR mixture was made containing 12.5uL of Amplification Master Mix (10x Genomics #220125), 1.25uL of cDNA additive (10x Genomics #220067), 1.25uL of forward primer (Partial P5, AATGATACGGCGACCACCGAGATCT) at 20uM, 1.25ul of reverse primer (Reverse Ext 2, AGGAGACTACGGTCAACTGCATGAAACAGA) at 10uM, 5uL of the DNA product from Step 1, and 3.75uL of nuclease-free water. With the thermal cycler lid set to 105°C, the following thermal cycling protocol is used: initial denaturation at 98°C for 45 s, 10 cycles of denaturing at 98°C for 20 s, annealing at 67°C for 30 s, extension at 72°C for 180 s, and a final extension at 72°C for 60 s. The PCR reaction mixture was purified using SPRIselect (0.8x) as described previously.

In Step 3, a 25uL PCR mixture was made containing 12.5uL of Hot Start Taq 2x Master Mix (NEB #M0496S), 1.25uL of cDNA additive (10x Genomics #220067), 1.25uL of forward primer (Partial P5, AATGATACGGCGACCACCGAGATCT) at 20uM, 1.25ul of reverse primer (Reverse Ext 3, GCAGCAAGTATGATGAGCAAGCTTTCTCACA) at 10uM, 5uL of the DNA product from Step 2, and 3.75uL of nuclease-free water. With the thermal cycler lid set to 105°C, the following thermal cycling protocol is used: initial denaturation at 95°C for 30 s, 10 cycles of denaturing at 95°C for 30 s, annealing at 60°C for 60 s, extension at 68°C for 210 s, and a final extension at 68°C for 300 s. The PCR reaction mixture was purified using SPRIselect (0.8x) as described previously.

In Step 4, a 25uL PCR mixture was made containing 12.5uL of Amplification Master Mix (10x Genomics #220125), 1.25uL of cDNA additive (10x Genomics #220067), 1.25uL of forward primer (Partial P5, AATGATACGGCGACCACCGAGATCT) at 20uM, 1.25ul of reverse primer (Partial Reverse Ext 3-Read 2, GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCAGCAAGTATGATGAGCA) at 10uM, 2uL of the DNA product from Step 3, and 6.75uL of nuclease-free water. With the thermal cycler lid set to 105°C, the following thermal cycling protocol is used: initial denaturation at 98°C for 45 s, 10 cycles of denaturing at 98°C for 20 s, annealing at 67°C for 30 s, extension at 72°C for 195 s, and a final extension at 72°C for 60 s. The PCR reaction mixture was purified using SPRIselect (0.8x) as described previously.

In Step 5, a 25uL PCR mixture was made containing 12.5uL of Amplification Master Mix (10x Genomics #220125), 0.5uL of SI-PCR primer (10x Genomics #220111), 2.5ul of from one well (recorded for future reference) from the Chromium i7 Sample Index Plate (10x Genomics #220103), 2uL of the DNA product from Step 4, and 7.5uL of nuclease-free water. With the thermal cycler lid set to 105°C, the following thermal cycling protocol is used: initial denaturation at 98°C for 45 s, 10 cycles of denaturing at 98°C for 20 s, annealing at 54°C for 30 s, extension at 72°C for 195 s, and a final extension at 72°C for 60 s. The PCR reaction mixture was purified using SPRIselect (0.8x) as described previously.

Quality control of each step was verified using High Sensitivity D5000 ScreenTape (Agilent #5067-5592) and High Sensitivity D5000 ScreenTape (Agilent #5067-5593) on an Agilent 2200 Tapestation system (Figure S1E). Amplicon libraries display multiple peaks in the TapeStation trace (Figure S1E), which is believed to be due to promiscuous binding of primers to poly-A like regions. Successful enrichment with correct DNA product were typically associated with "sawtooth"-like traces. After using different indexing primers in Step 5 for different libraries, several libraries were pooled and subsequently sequenced together on an Illumina Novaseq 6000 sequencing machine. The sequencing cycle settings were as follows: 28 cycles for Read 1, 8 cycles for the i7 index, and 91 cycles for Read 2.

cDNA material generated in the single-cell cDNA library construction step described in the previous section was generally adequate for making the amplicon libraries described here. If additional cDNA was needed, the single-cell cDNA library was amplified by repeating Step 2.2 of the Chromium Single Cell 3′ Reagent Kit v3 user guide (10x Genomics, CG000183 Rev A).

Depending on the mutation being targeted, optimization of the location of the locus-specific primers and their biochemical properties were needed to minimize non-specific primer binding. For the *JAK2*-V617F mutation, the three locus-specific primers used were ~30bp and have melting temperatures of around 68°C. Since the PCR cycle number is dependent on the expression level of the gene harboring the targeted mutations, the number of cycles was optimized experimentally for other mutation targets (ex. *TET2*).

Stem cell genotyping and preparation for WGS

Several types of cells were genotyped and prepared for WGS: (1) HSCs, MPPs, and stromal cells from bone marrow biopsies, and (2) fibroblasts from a skin biopsy.

1. HSCs, MPPs, and stromal cells

10mL of bone marrow aspirates were used to isolate HSCs, MPPs, and stromal cells. Erythrocytes were removed from 9mL of bone marrow aspirate samples using red blood cell lysis buffer. CD34-enrichment was performed using magnetic-assisted cell sorting with anti-CD34 magnetic beads (Miltenyi Biotech #130-046-703). Different cell populations were purified through using FACSAria (Becton Dickinson). The following combinations of cell surface markers were used to define cell populations. HSC: CD34+CD38-CD45RA-CD90+CD49f+; MPP: CD34+CD38-CD45RA-CD90-CD49f-.

Cells were first sorted into a collection tube, and a second index sorting step was performed to seed single-cells into round bottom 384-well plates, with ~90 μL of growth medium. Colonies were grown in StemSpan SFEM medium (StemCell Technologies, Inc. #09650) supplemented with: SCF (100 ng/ml), Flt3-L (100 ng/mL), TPO (50 ng/mL), IL3 (10 ng/mL), Tpo (50 ng/mL), Epo (10 ng/mL), and GM-CSF (10 ng/mL). Colonies were grown at 37°C in 5% CO_2 for 4-6 weeks before collection, with partial media exchange and fresh cytokine repletion every 2 weeks.

Polyclonal mesenchymal stem cells (MSCs) cultures were established from 1mL of whole bone marrow aspirate samples after red blood cell lysis, cells were plated in tissue culture treated dishes in DMEM-F12 medium (GIBCO #11320082), supplemented with 10% fetal bovine serum (VWR # 89510-186). MSCs were kept in culture for a week and medium was replaced each day to remove non-adherent cells. Stromal cells were ready for collection after reaching 80%–100% confluency.

2. Skin biopsy

One skin biopsy sample was obtained from the *JAK2* V617L mutant patient. The tissue was dissociated using collagenase I (StemCell Technologies, Inc. #07415) and genomic DNA was extracted as described below.

For both samples, genomic DNA was extracted from cells using QIAmp UCP DNA Micro Kits (QIAGEN #56204) and eluted into a final volume of 30 μL. The DNA concentration was quantified using a Qubit fluorometer (Invitrogen #Q32866) and Qubit dsDNA HS Assay Kit (Invitrogen #32854). Some of the genomic DNA (1 ng) was amplified using *JAK2*-V617F specific primers and screened for the mutation using Sanger sequencing.

To generate PCR products for the *JAK2* target loci for Sanger sequencing, we performed three rounds of nested PCR with locus-specific reverse primers and generic forward primers. First, a 611bp fragment of *JAK2* was amplified to obtain a sufficient amount of DNA containing the mutation site. PCR components included 1ng of gDNA template, Phusion Hot Start Flex 2x Master Mix (M0536L), forward and reverse primers ACTCTTGCTCTCTCTCACTTTG and ACCTGCCATAATCTCTTTTGCT (DNA oligos synthesized by IDT), respectively, and nuclease-free water. The amplification protocol was as follows: (1) initial denaturation at 98°C for 30 s; (2) 40 cycles of denaturation at 98°C for 10 s; (3) annealing at 63°C for 30 s; (4) extension at 72°C for 30 s; and (5) final extension at 72°C for 10 minutes. PCR reaction mixtures were slowly run on a 1.5% agarose gel in EDTA. PCR products were extracted from the gel using a Monarch DNA Gel Extraction Kit (NEB T1020L).

Second, the PCR products were Sanger sequenced in four separate reactions (each with one of four primers) through the Psomagen's gDNA sequencing service. The four primers used were flanking both directions of the mutation site (DNA oligos synthesized by IDT):

1. TGGCAGAGAGAATTTTCTGAAC (147bp upstream of the mutation)
2. ACTCTTGCTCTCTCTCACTTTG (304bp upstream of the mutation)
3. GTCCTACAGTGTTTTCAGTTTCA (166bp downstream of the mutation)
4. ACCTGCCATAATCTCTTTTGCT (306bp downstream of the mutation)

Electropherograms from Sanger sequencing were annotated. A sample of cells was marked as likely having the mutation if all of the corresponding electropherograms for that sample contained the mutation. Samples were then selected and their gDNA submitted for WGS sequencing (Broad Genomics). The WGS sequencing results were always consistent with the Sanger sequencing, that is the *JAK2* mutation was detected in the whole-genome sequencing data for the colonies that were designated as mutated using Sanger sequencing.

Of the ~600 MPPs and ~600 HSCs cultured from ET 1, we recovered sufficient genomic DNA from 62 MPP colonies and 22 HSC colonies after expansion. Of these, 6 MPP colonies and 16 HSC colonies had the *JAK2*-V617F mutation, indicating that mutated HSCs proliferated more in our culture conditions compared with WT HSCs. A similar bias was also observed in cells cultured from the patient ET 2. 384 HSC and 384 MPP colonies were cultured from the ET 2 patient. Of these, we extracted sufficient genomic DNA from 44 HSC colonies and 84 MPP colonies. Sanger genotyping revealed that 14 out of the 44 HSC colonies and 2 out of the 84 MPP colonies had a heterozygous *JAK2*-V617F mutation. The total sequencing depth was 1.6 billion reads for ET 1 and 480 million reads for ET 2.

Phylodynamic inference

To infer the clonal expansion of mutant HSCs, we first used BNPR (Karcher et al., 2016), an algorithm that infers population size multiplied by a constant factor from lineage trees, where the constant factor is the time between generations. BNPR assumes a Gaussian process prior on the clonal expansion and infers the marginal posterior distributions of the population size (multiplied by a constant factor) at different time points from the coalescent times of a tree. To infer population dynamics with BNPR, we used the Phylodyn package (Karcher et al., 2017).

For the 34-year-old patient tree, we used an averaging algorithm (the averaging algorithm is described in the ABC sections) to make the length from any leaf to the root of the tree the same. We then converted the branch lengths from mutations to generations by assuming 1 cell division per year (Abkowitz et al., 1996; Catlin et al., 2011), so that we could infer population size without the constant factor. The coalescent times of the tree were given to BNPR as input. The only parameter we set was lengthout = 28, which determines the number of time slices at which the population size is estimated, and the remaining were default parameters. The BNPR inference on the 63-year-old patient tree was done in an identical manner.

The BNPR inference was not sensitive to the priors we chose. In particular, changing the covariance associated with the Gaussian process did not change the interpretation of the results. We also tested BNPR on simulated clonal expansions under various scenarios, including simple exponential growth and population bottlenecks, and reliably inferred the population size over time.

Inference of JAK2 mutant HSC fitness

To more precisely infer the clonal expansion of mutated HSCs, we carried out ABC (Approximate Bayesian Computation) using the Wright-Fisher model with selection (see Methods S1 for a detailed description of ABC, the modeling, and for simulations that show that our inference is robust to model assumptions).

On each patient tree, we ran our ABC algorithm and inferred the model parameters.

For the patient with age 34, we used the following specifications for ABC:

1. s was drawn from a uniform distribution on (0, 2).
2. N was drawn from 10^X, where X is uniformly distributed on (1, 9).
3. L was drawn from round(Y), where Y is a Gaussian with mean 35 and std 5. If $L < 2$, we redrew L until $L >= 2$ since at least 2 generations are necessary to produce a tree.
4. g was drawn uniformly on 2, …, L.
5. $k = 22$ cells were sampled
6. the mutation rate was (total length of patient tree in mutations)/(L - 1). In this case, the total length of the patient tree in mutations was 723.
7. An epsilon threshold of 0.0225 was used.

For the patient with age 63, we used the following specifications for ABC:

1. s was drawn from a uniform distribution on (0, 2).
2. N was drawn from 10^X, where X is uniformly distributed on (1, 9).
3. L was drawn from round(Y), where Y is a Gaussian with mean 65 and std 10. If $L < 2$, we redrew L until $L >= 2$ since at least 2 generations are necessary to produce a tree.
4. g was drawn uniformly on 2, …, L.
5. $k = 13$ cells were sampled
6. The mutation rate was (total length of patient tree in mutations)/(L - 1). In this case, the total length of the patient tree in mutations was 1205.
7. An epsilon threshold of 0.0125 was used.

Note that we drew s from (0, 2) instead of from (0, 5) as done on simulated data in Methods S1. This was done to speed up the simulations and is justified because preliminary runs showed the distribution of s converging to a much smaller value.

For each patient we ran 400 parallel simulations. For the 34-year-old we collected 1,038,712 data points from the posterior, and for the 63-year-old we collected 8,816,199 data points from the posterior. The posterior joint distributions are plotted in Figure 4D and Figure S4.

As indicated by our analysis (see Methods S1), fitness s could be inferred from the patient trees. Our analysis also suggests that if we assume a division rate of one per year (Lee-Six et al., 2018), n can be inferred for the 34-year-old patient as 4.74 ± 0.68 of the posterior distribution. n, however, cannot be inferred for the 63-year-old patient, and the inferred distribution of n is just the prior information. This is due to the fact that the coalescent events occur in the very early history of the disease, and the information about the population size is lost. We can, however, put bounds on the point of saturation if we assume one division per year. As seen in simulation results presented in Methods S1, when the number of mutant cells approaches N, the growth of the mutant population slows down and starts to exhibit neutral dynamics. If N is sufficiently small, it changes the coalescent structure, and ABC then assumes the saturation point is n = N. Trajectories generated by ABC if N is below a certain threshold value produced coalescent structures that did not match that of the patient data. Any value of N larger than this threshold had no effect on the coalescent structure and therefore was retained as a possible inferred value. Therefore, N could not be precisely determined.

QUANTIFICATION AND STATISTICAL ANALYSIS

scRNA-seq preprocessing and cell type identification

scRNaseq libraries and the amplicon libraries were sequenced on the NovaSeq platform. The resulting bcl files were run through the Cell Ranger 4.0.0 pipeline to generate the fastq files and the count matrices. The fastq files for the amplicon libraries were analyzed as described below. Count matrices from each patient were loaded into Scanpy (Wolf et al., 2018). Genes expressed in < 3 cells and cells with < 2,000 total UMIs or > 20% mitochondrial transcripts were excluded from further analysis. Total count normalization was performed so that each cell had 100,000 total transcripts. Log-transformed expression values were used for UMAP visualization and clustering after regressing out % of mitochondrial transcripts and total counts. UMAP coordinates were calculated using Scanpy default parameter values. To assign an HSPC cell type to each cell, the scRNA-seq data from all patients were merged and batch corrected using Seurat's data integration workflow with the default parameter values (Stuart et al., 2019). Louvain clustering was

performed on the merged and batch-corrected dataset in Scanpy and each cluster was assigned an HSPC cell type identity by manually reviewing the expression levels of marker genes in that cluster. Identification of monocyte subsets was performed similarly on CD14+ cells from all donors.

Differential gene expression analysis between CD14+ cells from different patient groups

Scanpy's implementation of the Student's t test was used to compare gene expression between ET, PV, and healthy CD14+ cells using total count normalized gene expression values without batch correction. To limit the impact of batch effects, we performed all pairwise comparisons between each patient in both groups (e.g., for the ET versus PV comparison, we separately compared ET 1 and PV 1, ET 1 and PV 2, etc) and identified genes that were differentially expressed in all comparisons for gene set enrichment analysis using GSEApy (https://pypi.org/project/gseapy/).

Identification of JAK2 mutant cells in the scRNA-seq data

To identify individual cells in the scRNaseq library as either WT or *JAK2* mutant, we separately analyzed the fastq files of the amplicon libraries derived from the same cells. First, the reads in the fastq files were discarded if the average Illumina base quality value was less than 30. Next, only reads were retained whose single-cell barcode uniquely matched (up to at most 2 bps differences) a barcode from the list of single-cell barcodes in the scRNaseq library of the same cells. A threshold for the number of reads was determined by inspecting the plot of the number of reads from each molecule (unique cell-barcode and UMI) after rank ordering the molecules by their number of reads, corresponding to the knee in the plot, usually around 100 to 1,000 reads depending on the sequencing depth (Figure S1F). Molecules that had fewer reads than the threshold were discarded. To correct for sequencing errors in the UMIs, those molecules that shared the same barcode but had the same UMI sequence up to 2 mismatches were merged. Next, the mutation site was inspected in the remaining reads. Only reads were retained that had the expected WT nucleotide, or the expected mutated version of the nucleotide, and the where the 10 bps upstream and downstream of the mutation site matched the reference genome. A molecule was designated as "mutated" if more than half of its reads carried the mutated nucleotide, and WT otherwise. The above analysis pipeline was implemented in MATLAB R2018. *JAK2* genotyping results for all cells sequenced can be found in Table S3.

Cells with at least one WT amplicon call were marked as "WT." It is important to note that in cells with a heterozygous *JAK2* mutation, the presence of a WT transcript does not guarantee that the cell is homozygous WT. However, cells with at least one *JAK2*-mutant transcript were definitively classified as *JAK2*-mutant cells. Therefore, to correct for *JAK2* mutation heterozygosity in patients with < 50% peripheral blood *JAK2* mutation VAF, the fraction of *JAK2* mutant cells in a cell population was estimated as the fraction of *JAK2* mutant molecules in the cell population multiplied by 2. This correction factor comes from the observation that, since most cells only have one *JAK2* transcript call, cells with a heterozygous *JAK2* mutation have approximately a 50% chance of having a *JAK2* mutant transcript sequenced so approximately half of true *JAK2* mutant cells have a mutant *JAK2* transcript sequenced.

Whole-genome sequencing data analysis

Raw sequencing reads were mapped to the GRCh38 build of the human reference genome using BWA-MEM (Li, 2013) version 0.7.17-r1188. Aligned reads in BAM format were processed following the Genome Analysis Toolkit (GATK, version 4.1.2.0) Best Practices workflow to remove duplicates and recalibrate base quality scores (DePristo et al., 2011).

Detection of somatic single-nucleotide variants and INDELs

The germline short variant discovery workflow from GATK version 4.1.2.0 was used to detect somatic single-nucleotide variants (SNVs) and small insertions and deletions (INDELs) in the single-cell-derived WGS data. In brief, intermediate GVCF files were generated for each colony and chromosome using HaplotypeCaller in GVCF mode. Default parameter values were used except for the output-mode argument, which was set to "EMIT_ALL_SITES." Next, GVCF files for all colonies from each patient were consolidated into a single GVCF file using the GATK functionality CombineGVCFs using default options. Finally, colonies were jointly genotyped across all sites using GenotypeGVCFs with the "–include-non-variant-site" parameter set to true.

In order to identify somatically acquired point mutations and indels in the colonies the following steps were followed.

All sites with a genotype quality of at least 50 and showing variation in at least one colony were selected.
Variants mapping less than 10bp upstream or downstream of a simple repeat reported in the RepeatMasker track from the UCSC Genome Browser were discarded.
Variants mapping less than 100bp apart from each other were removed, as in our experience these are likely artifacts.
Variants that could not be genotyped in 10 or more colonies in each patient were discarded.
To remove subclonal mutations acquired during *in vitro* culture the mean variant allele frequency (VAF) value across all mutated colonies was required to be between 0.3 and 0.7 for patient ET 2 (female) and for the autosomes in the case of patient ET 1 (male). Additionally, we required a minimum coverage of at least 6 sequencing reads. Variants mapping to chromosomes X and Y in the case of patient ET 1 and chromosome X in colony MPP-73 from patient ET 2, which harbors only one copy of this chromosome, were required to show a VAF value of at least 0.9 and the coverage threshold was set to 3 sequencing reads.
Sites supporting more than 4 genotypes across all colonies were removed, as after manual inspection of a number of such cases we concluded that these were likely artifacts.

We required the genotype quality in the bulk sequencing data from stromal cells to be at least 80 in order to remove variants in low-quality mapping regions.
Only variants with a homozygous reference genotype in the bulk sample were kept in order to filter out germline heterozygous polymorphisms.
Given that all cancer cells share 220 and 398 mutations in patients ET 1 and ET 2, respectively, we reasoned that any mutations occurring early in development and giving rise to both the cancer and wild-type cells should be present in all cancer colonies and in a subset of the normal colonies, but not in normal colonies and just a subset of the cancer colonies. Therefore, all mutations detected in just a subset of the cancer cells and one or more wild-type colonies were discarded, as these are likely germline polymorphisms or artifacts. We did not find any mutation present in all cancer cells and one or more normal cells. In this analysis we only focused on the discovery of heterozygous variants given that all colonies, with the exception of the loss of chromosome X in colony MPP-73 from ET 2, show diploid karyotypes with no copy number alterations.

All variants remaining after applying the filters described above were visually inspected using BAMsnap (https://github.com/parklab/bamsnap), and those deemed to be false positives were removed. The remaining variants were deemed to be somatic and were considered for further analysis. Annovar (version 2018Apr16) was used to annotate variants. Missense variants predicted to be deleterious by MetaLR and MetaSVM were considered pathogenic (Dong et al., 2015).

Detection of microsatellite mutations

Somatic mutations at microsatellite loci were detected using HipSTR version 0.6.2 (Willems et al., 2017) using *de novo* stutter estimation and allele generation, and the reference set of microsatellite loci provided by the authors. Subsequently, microsatellite calls were filtered and only calls satisfying the following criteria were considered for further analysis: (1) Posterior probability for the genotype higher than 0.95; (2) the fraction of indels in the reads mapping to the flanking regions of the microsatellite under consideration smaller than 0.15; (3) the fraction of reads estimated to contain a stutter artifact smaller than 0.15; (4) at least 3 sequencing reads spanning each of the supported alleles for the microsatellite under consideration; (5) $\log_{10}$ *P value* for the allele bias test implemented in HipSTR higher than 2; (6) $\log_{10}$ *P value* for the Fisher strand bias test higher than 2; (7) the ratio of the number of reads supporting each allele higher than 0.7. This filter served to remove low-VAF mutations likely arising during *in vitro* culture or PCR noise; and (8) a depth of at least 10 sequencing reads. Finally, only microsatellite loci with a reliable call in at least 30 samples and with at least 2 different genotypes across all colonies were considered for further analysis. All mutations satisfying the criteria listed above were further validated through visual inspection of raw sequencing reads.

Detection of somatic structural variants

Structural variants were called in each colony using Manta (version 1.6.0), LUMPY (version 0.2.13), SvABA (version 1.1.3), and Delly (version 0.8.3) (Chen et al., 2016; Layer et al., 2014; Rausch et al., 2012; Wala et al., 2018). Each algorithm was run independently on each colony using the bulk sequencing data for stromal cells from the corresponding patient as control, and in a second run using a randomly selected *JAK2*-WT colony as control. The calls generated by each algorithm were merged using the Python library mergevcf. (https://github.com/ljdursi/mergevcf) and only calls generated by at least two algorithms were kept for further analysis.

Somatic copy number calling

The software package ascatNGS (Raine et al., 2016) was used to detect somatic copy number alterations in each colony and to estimate their purity and ploidy. The bulk sequencing data from bone marrow stromal cells from the same patient was used as the normal sample in all cases.

Mutational signature analysis

Mutational signature analysis was performed using the R package *MutationalPatterns* (Blokzijl et al., 2018). To quantify the contribution of mutational processes known to be operative in MPNs (Alexandrov et al., 2020) (namely SBS1, SBS2, SBS5, SBS19, SBS23, and SBS32) to the observed spectrum of somatic point mutations in each colony, we used the function *fit_to_signatures* using default parameter options. The goodness of fit was determined by computing the cosine similarity between the observed mutational pattern and the reconstructed one using the estimated signature contributions. In all cases we obtained cosine similarity values > 0.95, suggesting that our analysis explained most of the variance related to the contribution of different mutational processes to the observed mutational spectra.

Telomere length estimation

The length of telomeres was estimated for all colonies from the same patient jointly using Telomerecat version 3.4.0 (Farmery et al., 2018) and the default options except for batch correction. The average telomere length across 100 runs was considered for further analysis.

Comparing the mutation rate between JAK2 mutant and JAK2-WT colonies

To assess whether the mutation rate in *JAK2* mutant and *JAK2*-WT colonies is statistically significant, we had to account for the fact that *JAK2* mutant colonies are clonally related, as they share hundreds of mutations in both ET 1 and ET 2. To account for this shared

ancestry, we computed the difference between the mean number of mutations in *JAK2* mutant and *JAK2*-WT colonies. Next, we computed the expected variance by accounting for the clonal relatedness of *JAK2* mutant colonies. Specifically, we scaled the variance of the number of mutations in *JAK2*-WT colonies by the number of years at which the clonal expansion started (that is, 9/34 and 19/63 in the case of ET 1 and ET 2, respectively), and computed the square root. We scaled the number of mutations by the variance rather than by the standard deviation given that we assume that the accumulation of mutations in HSPCs can be modeled as a Poisson process. If we then consider the distribution of mean differences to be Gaussian with mean zero, we can compute a *z* score by computing the mean difference divided by the estimated standard deviation, and then estimate the corresponding one-sided *P value*.

Inference and validation of phylogenetic trees

The somatic mutations detected across all colonies in a given patient were used to reconstruct phylogenetic trees using the software package PHYLIP version 3.695 (https://evolution.genetics.washington.edu/phylip.html). For each patient and mutation type, namely, SNVs, INDELs, and microsatellite mutations, as well as for these three combined, we constructed a binary matrix with rows indexed by somatic mutations and columns by colonies such that the *i*,*j* entry in each matrix was set to one if mutation *i* is present in colony *j*, and to zero otherwise. Only mutations detected in at least two colonies were considered to build lineage trees, as private mutations are uninformative to establish the phylogeny of the colonies. We detected a total of 21,699 SNVs (935 present in at least two colonies), 1,396 (60) indels, and 482 (31) microsatellite mutations across the single-cell-colonies derived from patient ET 1 (Table S2). In the case of ET 2, we detected a total of 33,994 SNVs (1,245), 2,464 (94) indels, and 891 (70) microsatellite mutations (Table S2).

For each input mutation matrix, we generated 100 bootstrap replicates by sampling with replacement using the *Seqboot* method. Lineage trees were then estimated for each resample using the Wagner parsimony algorithm as implemented in the *Mix* method using the bulk data from stromal cells as the outgroup. The consensus tree across all bootstrap samples was generated using the extended majority rule method as implemented in the program *Consense*. Once the consensus tree was determined, we assigned to each branch those mutations that were present in all the descendant colonies of that branch and in none of the other colonies. Lineage tree representations were generated using the R package *ggtree* (Yu, 2020).

As expected given the high number of mutations shared across cancer colonies, the clonal architecture of cancer cells was largely consistent across bootstrap resamples irrespective of the type of mutations considered for lineage tree inference. In fact, the clonal architecture for the cancer colonies was the same across all resamples when using somatic SNVs as input. More variability was observed when the lineage trees were constructed using indels or microsatellite mutations as input, although the majority of splits in the tree were consistent across more than 90% of resamples. This is expected given that variant callers generally show lower sensitivity and specificity for the detection of small insertions and deletions as compared to point mutations (Campbell et al., 2020). This is also consistent with the fact that the highest rates of private INDELs and microsatellite are detected for those colonies with the lowest sequencing quality in our cohort (e.g., HSC-49 from ET 1). The clonal architecture of WT colonies varied across resamples, as indicated by the low bootstrap values we obtained for nodes splitting clades of WT colonies. The low concordance observed for node splits across resamples is likely due to the low number of somatic mutations detected in more than one WT colony, consistent with previous lineage tree analyses of human HSPCs using somatic mutations (Lee-Six et al., 2018). Overall, the reliability of the consensus trees we have generated is supported by the following: (1) the clonal architecture, in particular for cancer colonies, observed across lineage trees inferred using different types of somatic mutations is overall consistent, (2) the nodes in the trees are largely concordant across bootstrapping resamples for cancer colonies, and (3) 96% and 99% of the SNVs detected in at least 2 colonies from ET 1 and ET 2, respectively, could be unambiguously assigned to the consensus lineage tree generated using SNVs.

Cell Stem Cell, Volume *28*

Supplemental Information

Reconstructing the Lineage Histories and Differentiation Trajectories of Individual Cancer Cells in Myeloproliferative Neoplasms

Debra Van Egeren, Javier Escabi, Maximilian Nguyen, Shichen Liu, Christopher R. Reilly, Sachin Patel, Baransel Kamaz, Maria Kalyva, Daniel J. DeAngelo, Ilene Galinsky, Martha Wadleigh, Eric S. Winer, Marlise R. Luskin, Richard M. Stone, Jacqueline S. Garcia, Gabriela S. Hobbs, Fernando D. Camargo, Franziska Michor, Ann Mullally, Isidro Cortes-Ciriano, and Sahand Hormoz

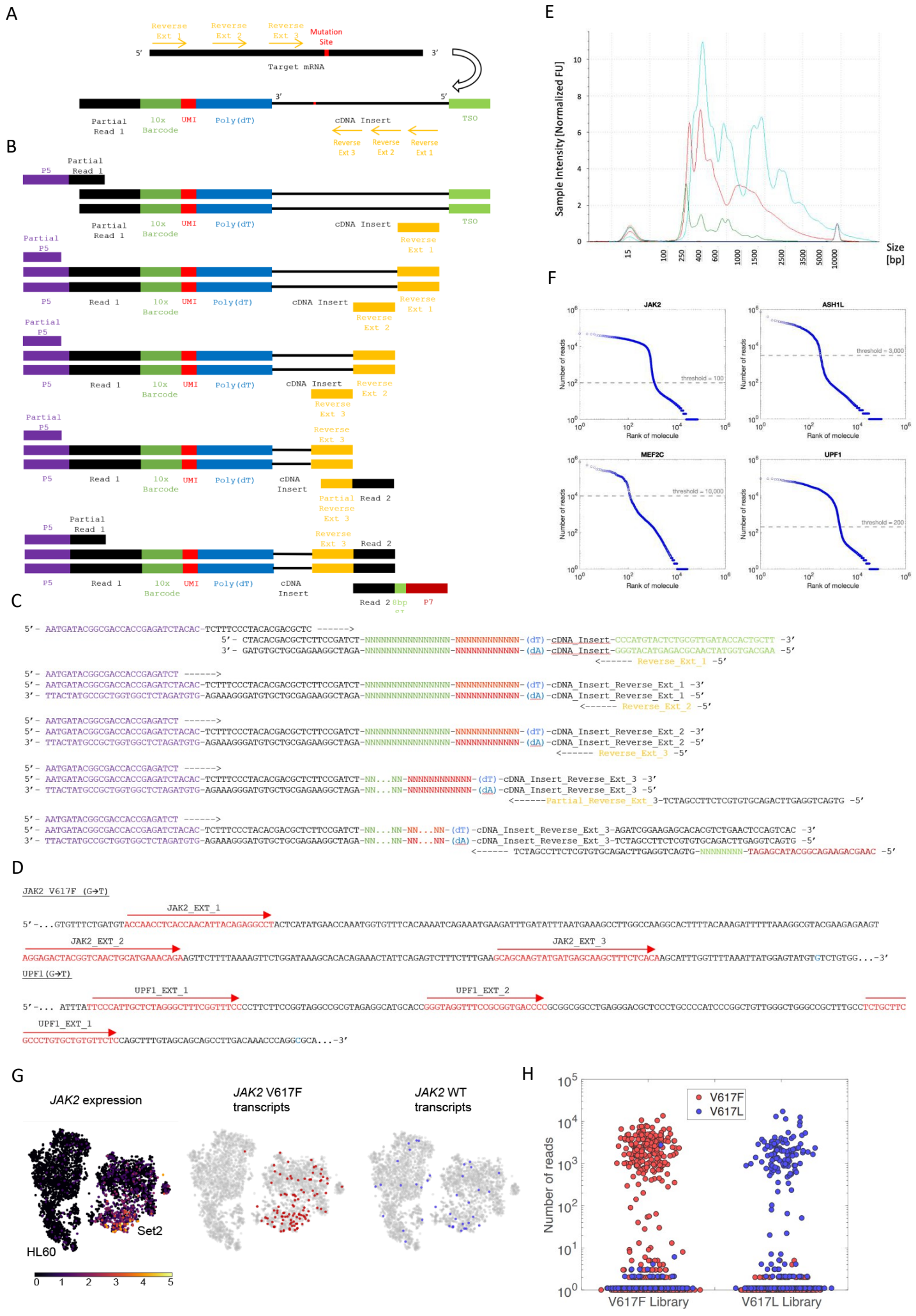

A
Reverse Ext 1
Reverse Ext 2
Reverse Ext 3
Mutation Site
5′
3′
Target mRNA
Partial Read 1
10x Barcode
UMI
Poly(dT)
cDNA Insert
TSO
B
P5
Read 1
Read 2
Read 2 8bp
P7
C
5′- AATGATACGGCGACCACCGAGATCTACAC-TCTTTCCCTACACGACGCTC ------>
5′- CTACACGACGCTCTTCCGATCT-NNNNNNNNNNNNNNNN-NNNNNNNNNNNN-(dT)-cDNA_Insert-CCCATGTACTCTGCGTTGATACCACTGCTT -3′
3′- GATGTGCTGCGAGAAGGCTAGA-NNNNNNNNNNNNNNNN-NNNNNNNNNNNN-(dA)-cDNA_Insert-GGGTACATGAGACGCAACTATGGTGACGAA -5′
<------ Reverse_Ext_1 -5′
<------ Reverse_Ext_2 -5′
<------ Reverse_Ext_3 -5′
<------Partial_Reverse_Ext_3-TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG -5′
<------ TCTAGCCTTCTCGTGTGCAGACTTGAGGTCAGTG-NNNNNNNN-TAGAGCATACGGCAGAAGACGAAC -5′
D
JAK2 V617F (G→T)
JAK2_EXT_1
JAK2_EXT_2
JAK2_EXT_3
UPF1 (G→T)
UPF1_EXT_1
UPF1_EXT_2
E
Sample Intensity [Normalized FU]
Size [bp]
F
JAK2
ASH1L
MEF2C
UPF1
Number of reads
Rank of molecule
threshold = 100
threshold = 3,000
threshold = 10,000
threshold = 200
G
JAK2 expression
JAK2 V617F transcripts
JAK2 WT transcripts
HL60
Set2
H
V617F
V617L
Number of reads
V617F Library
V617L Library

Figure S1. Primer designs, directions, locations, sequences of target mutation amplification and troubleshooting; Accurate identification of the mutated cells from the amplicon libraries. Related to Figure 1. A. Schematic illustration of primer direction against target mRNA and the change of primer directionality during amplification of sc-cDNA. **B.** Schematic diagrams of the nested PCR from step 1 to step 5, respectively. **C.** Oligonucleotide sequences and localization of common primers and adaptors. **D.** Example primer positions and sequences of targeted mutation. **E.** Example TapeStation trace for QC and optimization from step 1 to 5. **F.** Number of reads vs rank of molecule and threshold of cell calling. **G.** In a control experiment HL60 (WT cells) were mixed with SET2 cells (heterozygous *JAK2*-V617F mutation) and ran through the experimental and analysis pipeline. The two cell populations could be distinguished based on their transcriptional profiles: two distinct clusters were seen when transcriptomes of the cells were visualized using tSNE. Marker genes were used to identify the clusters as either HL60 or SET2 cells. Cells in which a mutated *JAK2* transcript (middle panel) or a WT *JAK2* transcript (right panel) were detected in the amplicon libraries are shown as colored points. All other cells are shown in gray. *JAK2* mutation site was detected in ~4% of cells. The detection limit is primarily determined by the JAK2 expression levels (shown in the left panel). The false positive rate of detecting a mutated transcript in a WT HL60 cell is less than 1%. **H.** In another control experiment we combined the single-cell libraries of a *JAK2*-V617F patient (ET 1) and that of the *JAK2*-V617L patient before the libraries were fragmented and indexed. We then ran the combined library through the experimental and analysis pipeline. The *JAK2* amplicon sequences could be mapped back to the library from which they originated based on their single-cell barcode. The plot shows the number of reads of each *JAK2* transcript detected for the V617F library (left) and V617L library (right). The colors of the points denote whether the transcript sequence contained a V617F or V617L mutation. Blue dots on the left side and red dots on the right side correspond to incorrect mapping of a mutation to a single-cell barcode, most likely due to PCR crossover events during amplification. Above the threshold of 100 reads, the false positive rate is negligible.

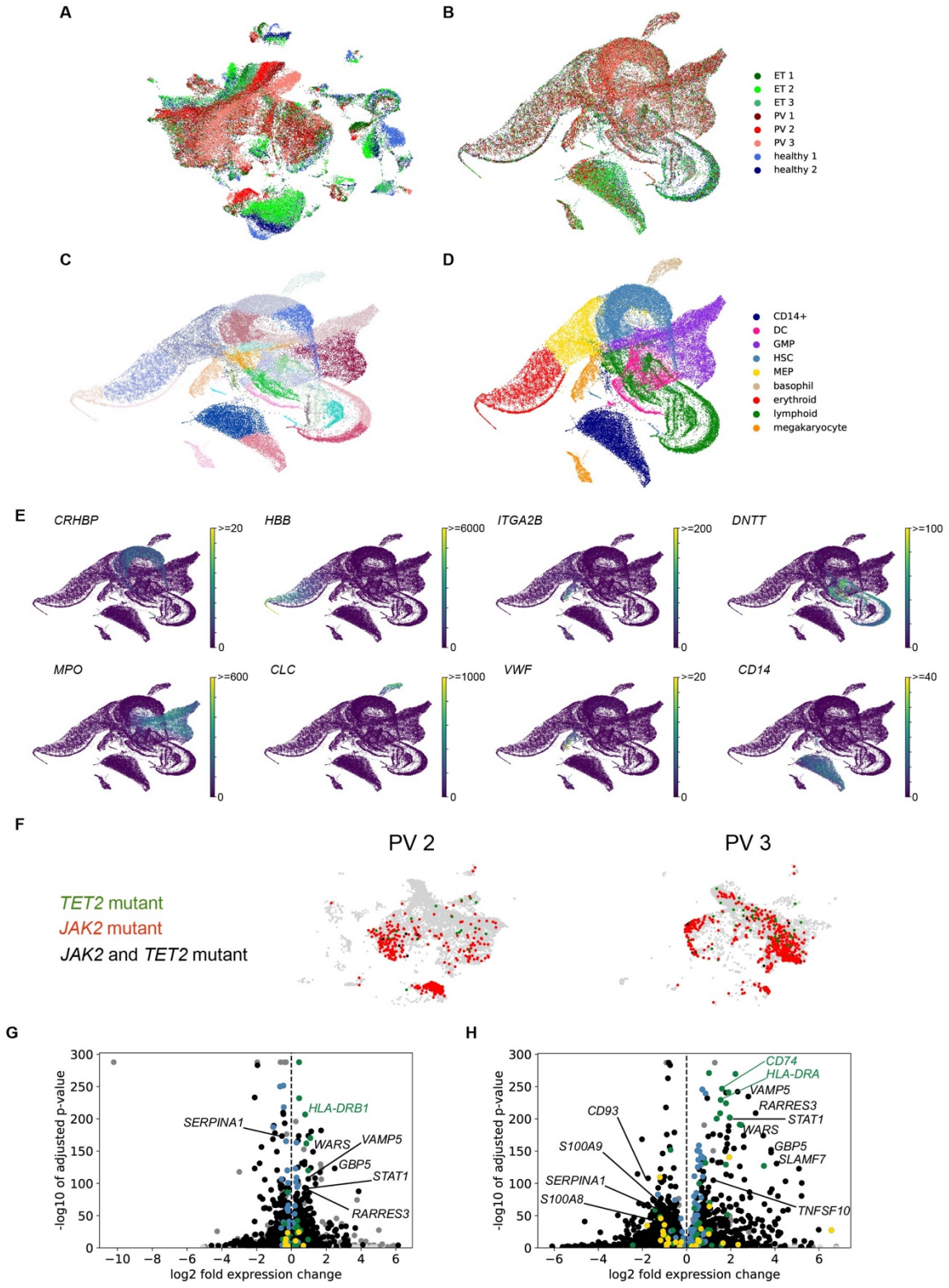


Figure S2. scRNA-seq analysis and genotyping of *JAK2*-mutant MPNs. Related to Figure 2. A. UMAP of CD34-enriched bone marrow scRNA-seq data from all patients before batch correction, colored by donor as in **B**. **B-D.** UMAP of CD34-enriched bone marrow scRNA-seq data from all patients after Seurat batch correction, colored by donor (**B**), Louvain cluster (**C**), and final cell type identification (**D**). **F.** UMAPs for patient PV 2 (left) and PV 3 (right) with cells with *JAK2* and/or *TET2* mutant transcripts highlighted. G-H. Volcano plots of differential expression for all CD14+ cells between ET patients and healthy controls (**G**) or PV patients and

healthy controls (**H**). Genes found to be differentially expressed in all pairwise comparisons between different patient subsets are highlighted and colored by KEGG 2019 biological process group (gold: hematopoiesis-related, green: antigen presentation, blue: ribosomal, black: other).

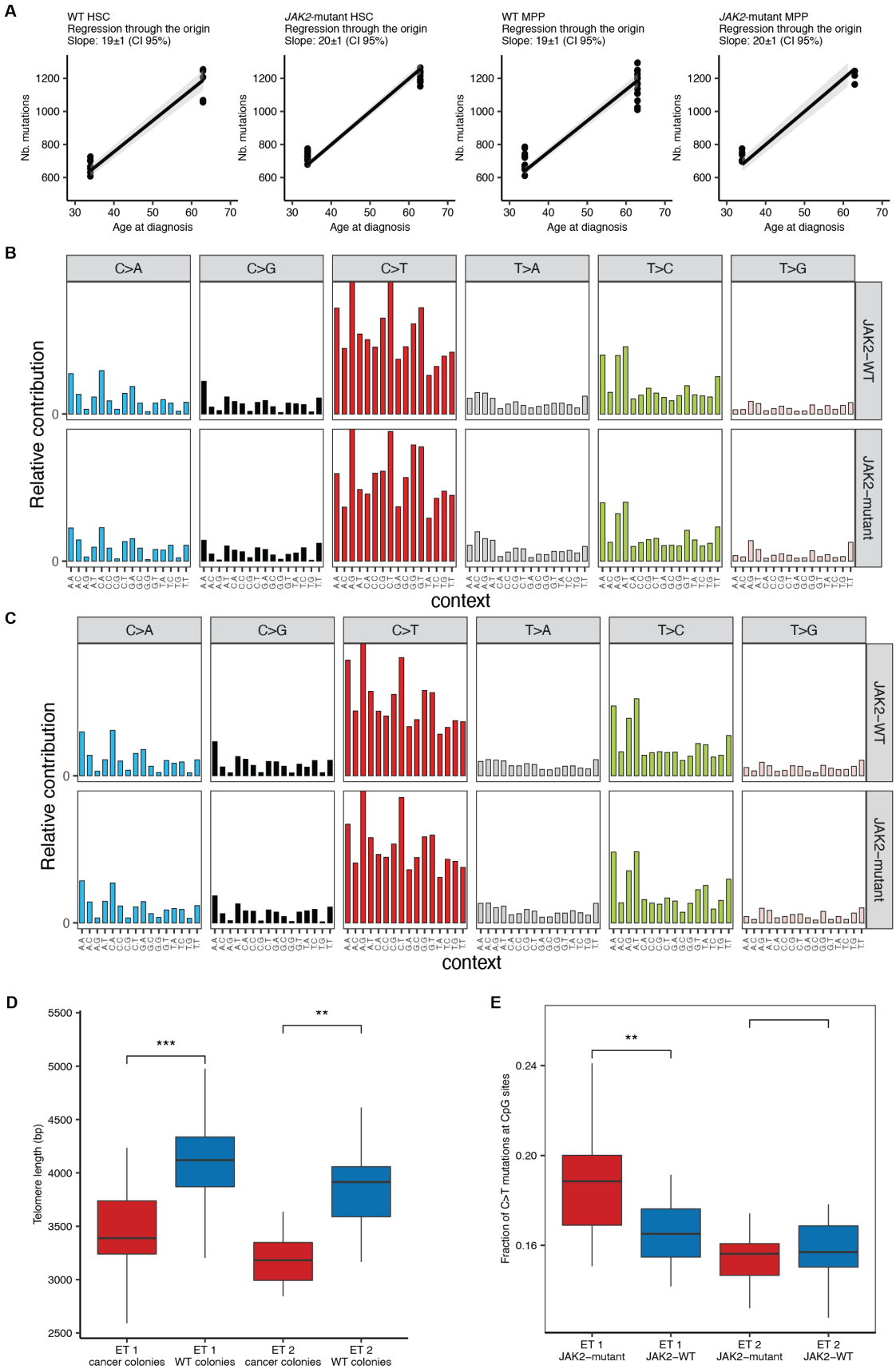

A
WT HSC
Regression through the origin
Slope: 19±1 (CI 95%)
JAK2-mutant HSC
Regression through the origin
Slope: 20±1 (CI 95%)
WT MPP
Regression through the origin
Slope: 19±1 (CI 95%)
JAK2-mutant MPP
Regression through the origin
Slope: 20±1 (CI 95%)
Nb. mutations
Age at diagnosis
B
C>A
C>G
C>T
T>A
T>C
T>G
JAK2−WT
JAK2−mutant
Relative contribution
context
C
D
Telomere length (bp)
***
**
ET 1 cancer colonies
ET 1 WT colonies
ET 2 cancer colonies
ET 2 WT colonies
E
Fraction of C>T mutations at CpG sites
ET 1 JAK2−mutant
ET 1 JAK2−WT
ET 2 JAK2−mutant
ET 2 JAK2−WT

Figure S3. Mutational analysis of individual HSCs and MPPs from MPN patients. Related to Figure 3. A. Correlation between age at diagnosis (years; x-axis) and the number of somatic SNVs (y-axis) detected in WT HSCs, *JAK2*-mutant HSCs, WT MPPs, and *JAK2*-mutant MPPs. Each dot corresponds to a single-derived colony, and the lines represent the regression through the origin. The estimated values for the slope and the 95% confidence intervals (CI) are shown. **B-C.** Patterns of somatic mutations for *JAK2*-mutant or *JAK2*-WT colonies from patients ET 1 (**B**) and ET 2 (**C**). The relative fraction of each mutation type in the catalogue of point mutations detected in each colony is reported. Base substitutions are further stratified into categories based on the trinucleotide context in which the mutation occurs. **D.** Distribution of telomere lengths estimated using the whole-genome sequencing data for *JAK2*-mutant and *JAK2*-WT HSPCs from patients ET 1 and ET 2. **E.** Distribution of the number of C>T mutations at CpG dinucleotides in *JAK2*-mutant and *JAK2*-WT HSPCs from patients ET 1 and ET 2. The box plots in **D-E** show the median, first and third quartiles (boxes), and the whiskers encompass observations within a distance of 1.5× the interquartile range from the first and third quartiles. The box plots in **D-E** show the median, first and third quartiles (boxes), and the whiskers encompass observations within a distance of 1.5× the interquartile range from the first and third quartiles. The single, double, and triple asterisks indicate statistical significance at $P < 0.05$, $P < 0.01$, and $P < 0.001$, respectively (Wilcoxon rank sum test).

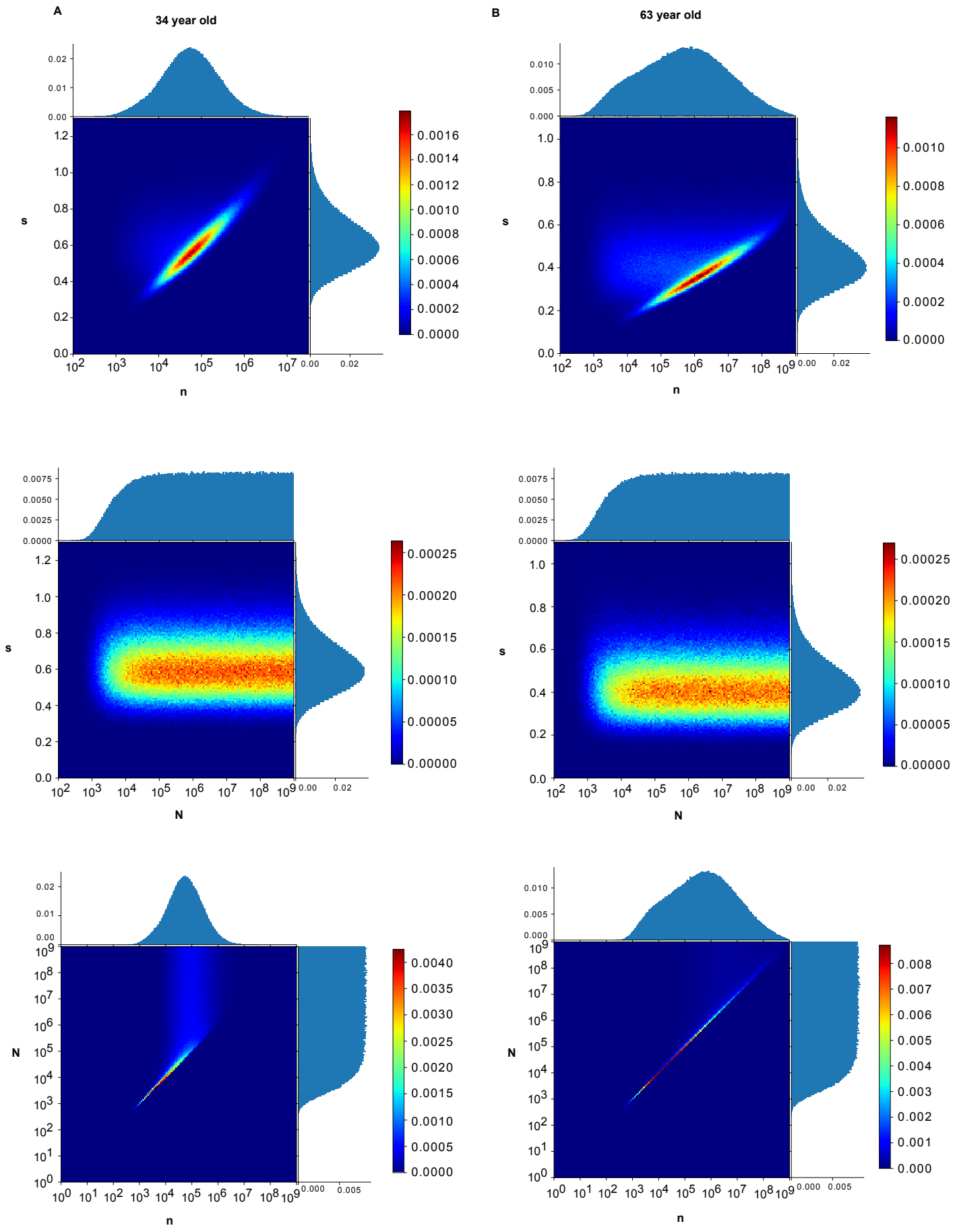

A
34 year old
B
63 year old
s
n
N

Figure S4. Inference on patient data. Related to Figure 4. ABC was run on the patient data, and the model parameters were inferred. Joint distributions of the parameter values were plotted along with their marginals. *s* is in growth per year, and age of onset of the disease is in years. **A.** Distributions for inference on 34-year-old patient data. **B.** Distributions for inference on 63-year- old patient data. As observed, inference of *n* can indicate saturation when the number of cancer cells approaches *N*. In this case, the cancer expansion slows down and begins to exhibit neutral dynamics. This changes the coalescent structure, allowing ABC to detect the saturation and infer a saturation parameter value of $n = N$. When *N* is too large to affect the exponential growth dynamics of the cancer cells, ABC can only put a bound on *N*, namely that *N* must be larger than the number of cancer cells at the final time-point.

Table S1. Primers and sequences for mutation-specific single-cell amplicon libraries (5′→3′). Related to STAR Methods.

INTERNAL_FORWARD	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
SHORT_INT_FOR	AATGATACGGCGACCACCGAGATCT
JAK2_EXT1	ACCAACCTCACCAACATTACAGAGGCCT
JAK2_EXT2	AGGAGACTACGGTCAACTGCATGAAACAGA
JAK2_EXT3	GCAGCAAGTATGATGAGCAAGCTTTCTCACA
JAK2_EXT4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGCAGCAAGTATGATGAGCAA
ASH1L_Ext1	GCATCTCACTCCTATCTGAAAAGTTGACAAGC
ASH1L_Ext2	TGGCCACAAAGAAAAACCTAGACCATGTCA
ASH1L_Ext3	GGAAATGTCCCCTTCAGGCTGTCGTATCAA
ASH1L_Ext4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGGAAATGTCCCCTTCAGGCT
HSPA9_EXT1	ACCTGACAAGAGTCTTAAGCAACCAAAGCA
HSPA9_EXT2	GTGGGTCATGCCTGTAATCCCAACACTTG
HSPA9_EXT3	GTGTGGGAGTTGAAGATCACCCTAGGCAA
HSPA9_EXT4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGTGTGGGAGTTGAAGATCAC
NRROS_EXT1	GAATCCATCTGTCTCCTTTCCTCAGCTTTGCCT
NRROS_EXT2	AGTCCCGGAGCTGGTGGCAAAGA
NRROS_EXT3	TCTCACGGGCCCAGCCTTACTCA
NRROS_EXT4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTCACGGGCCCAGCCTTAC
UPF1_EXT1	TTCCCATTGCTCTAGGGCTTTCGGTTTCC
UPF1_EXT2	GGGTAGGTTTCCGCGGTGACCCC
UPF1_EXT3	TCTGCTTCGCCCTGTGCTGTGTTCTC
UPF1_EXT4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTTCTGCTTCGCCCTGTGCTGT
TET2B_EXT1	CACATAACTGCAGTGGGCCTGAAAATCCAG
TET2B_EXT2	TAATGGTGCTACAGTTTCTGCCTCTTCCGT
TET2B_EXT3	ACATCTCACATAAATGCCATTAACAGTCAGGC
TET2B_EXT4	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTACATCTCACATAAATGCCAT

Methods S1, related to STAR Methods.

Outline. In Section 1, we first define our mathematical model for mutated HSC growth dynamics, followed by the model with feedback where the fitness of the mutant cells decreases over time, and then describe an efficient way to simulate the mutant growth dynamics for inference. We then go on to derive analytical mathematical expressions for the average growth rate of mutant cells in our model, and use these to show that the fitness in terms of growth per year can be inferred from lineage trees without knowledge of the mutation rate. In Section 2, we include a detailed description of the simulations, analyses, and figures to verify our theoretical calculations and to demonstrate the robustness of our inference with respect to the model assumptions. Section 2 can be read before Section 1.

1. Mathematical Description of Growth Dynamics of Mutated Cells

WF model with selection

We chose to model mutated HSC growth (or the growth of a population of mutant cells within a population of wild-type cells) as a variation of the Wright-Fisher stochastic process (Fisher, 1923; Wright, 1931) with selection included. We begin with an initial population of N stem cells at $t = 0$. The stochastic process arises by iterating the following rule on the initial population: the cells at generation t give birth to N stem cells which correspond to generation $t + 1$. Each cell in generation $t + 1$ then selects a cell at random from generation t as its parent, and the cells at generation t die off. At generation t', a mutation arises in one of the stem cells that gives it a selective advantage. From then on, the iterative rule changes in that instead of the cells from the parent generation being selected at random, each wild-type cell has probability p of being selected as a parent, and each mutant cell has probability $(1 + s)p$ of being selected as a parent. The cells always inherit the phenotypic state of their parents (mutant or wild-type). After L iterations, we produce an evolving population of stem cells for $t = 0,1, \dots, t', t' + 1, \dots, L$ along with a set of genealogical relationships.

Note that here there are a total of $L + 1$ generations of cells, since the first generation corresponds to $t = 0$. In the computational section (Section 2), we instead define L to be the total number of generations. The way we have defined L here is more convenient for the mathematical derivations, but it should be noted that any expression we have derived here in the mathematical section (Section 1) that uses L will be replaced with $L - 1$ in the computational section (Section 2).

Computing the number of mutant cells as a function of time

Here, we derive the distributions for the number of mutant cells as a function of time. Given the current generation of mutant cells, the number of mutant cells in the generation that follows is binomially distributed with parameter values that depend on the number of mutant cells in the current generation. This fact is used in the subsequent subsections to compute the mean growth of mutant cells and for efficient simulation of clonal expansions.

Suppose that there are n mutant cells at generation t (Note that in Section 2, we define n to be the number of mutant cells at the final time-point. This n is not the same). Since each wild-type cells has probability p of being selected as a parent, and each mutant cell has probability $(1+s)p$ of being selected as a parent, then the probability that a cell at t + 1 choses a wild-type cell is $(N-n)p$, and the probability it chooses a mutant cell is (1 + s)np. Since probabilities must sum to one, p can be derived from the condition that $(N-n)p + (1+s)np = 1$, and so we obtain $p = \frac{1}{N+ns}$. After substituting, we obtain the probability a cell selects a mutant cell as its parent as:

$$\frac{(1+s)n}{N+ns}$$

Since each of the N cells at t + 1 either chooses a mutant cell or it does not, and since the choices are independent, then the number of mutant cells at generation t + 1 is binomially distributed with parameters N and $\frac{(1+s)n}{N+ns}$. It then follows that we can compute the number of mutant cells as a function of time by beginning with an initial condition of $n = 1$ mutant cell, and then carrying out a series of binomial draws where we update n before each draw to equal the current number of mutant cells.

Wright-Fisher model with feedback

The Wright-Fisher model is an idealized model that ignores a wide range of biologically plausible scenarios. For example, as the number of mutant cells increases, it is conceivable that there are underlying biological mechanisms that slow the growth of the mutant cells. To simulate such scenarios, we decided to incorporate feedback into the Wright-Fisher model. This is accomplished by letting the value of s change depending on the current number of mutant cells. In particular, if the number of mutant cells at generation t is n, then the selection parameter at t is $s(1-\frac{n-1}{N-1})^x$. Notice that when $x = 0$, the selection parameter remains constant and we recover the usual Wright-Fisher model with selection.

To simulate clonal expansions for mutant cells with feedback, we simply draw a series of numbers from binomial distributions as described in the previous section, except that instead of just updating n before each draw, we first update n and then $s(1-\frac{n-1}{N-1})^x$. Clonal expansions with feedback in Section 2 are simulated in this way.

Dynamics of average population size

Define n(t) as the number of mutant cells as a function of time. We now compute the expectation of $n(t)$, which we will call the mean trajectory. We consider the expected value of $n(t)$, conditioned on the mutant clone consisting of n cells at time $t-1$. Upon conditioning, $n(t)$ reduces to a binomially distributed random variable as shown before, and so its mean is given by:

$$E_N[n(t) \mid n(t-1) = n] \;=\; N\frac{(1+s)\,n}{N+ns}$$

Subscript N is used to emphasize the dependence on population size. We then assume $N >> ns$, and drop subscript $N$to obtain:

$$E[\,n(t) \mid n(t-1) = n\,] \;=\; (1+s)n$$

Rewriting the expected value as a conditional expectation gives us

$$E[\,n(t) \mid n(t-1)\,] \;=\; (1+s)\,n(\,t-1\,)$$

We then take the expectation of both sides to generate the following recursion

$$E[n(t)] \;=\; (1+s)\,E[n(t-1)]$$

initial condition: $n(0) = 1$

where without loss of generality we have let the time at which the mutation arrives be $t = 0$.

The recursion is then easily solved to obtain

$$E[n(t)] \;=\; (1+s)^t$$

Dynamics of average population size conditioned on survival.

When the number of mutant cells is small, they are susceptible to stochastic fluctuations and extinction. After growing to a sufficiently large size, their growth dynamics become deterministic and fluctuations can be safely ignored.

When using ABC to infer our model's parameters, we only consider trees where mutant cells have not gone extinct. We are thus interested in the growth dynamics conditioned on no stochastic extinction.

In the previous section, we computed the expectation value of the number of mutant cells as a function time across all trajectories. Here, we will constrain the expectation value to trajectories that do not go extinct. As would be expected, the average population size is larger when extinction is not allowed.

We begin by defining the conditioning event for our trajectories as $F = \{fixation\ will\ occur\}$. Then we use Bayes' theorem to compute

$$E_N[\, n(t) \mid F \,] \;=\; \sum_{n\,=\,1}^{N} n\, P(\, n(t) = n \mid F \,) \;=\; \sum_{n\,=\,0}^{N} \frac{n\, P(\, n(t) = n \,)\, P(\, F \mid n(t) = n \,)}{P(\, F \,)}$$

The probability of fixation of a clone of size n within a sufficiently large population of size N and with fitness $1 + s$ is given by Kimura's diffusion approximation (Kimura, 1962)

$$\frac{1 - e^{-2sn}}{1 - e^{-2sN}}$$

We therefore put

$$P(\, F \mid n(t) = n \,) = \frac{1 - e^{-2sn}}{1 - e^{-2sN}}$$

$$P(\, F \,) \;=\; \frac{1 - e^{-2s}}{1 - e^{-2sN}}$$

The probability of fixation $P(\, F \,)$ independent of the clone size is simply the probability that a clone of size $n = 1$ will eventually fix. Substituting both probabilities back into the sum and cancelling terms gives us

$$E_N[\, n(t) \mid F \,] \;=\; \frac{1}{1 - e^{-2s}} \sum_{n\,=\,0}^{N} \quad n\, P(\, n(t) = n \,)\, (1 - e^{-2sn})$$

Next, we make a key biologically motivated assumption: we assume that t is sufficiently large so that the population of mutant cells is:

1) either large enough to exhibit deterministic dynamics, or
2) has gone extinct.

Therefore, $P(\, n(t) = n \,)$ vanishes except for when n is large, or $n = 0$. Since the only nonzero terms in the expectation are those for large n, and since $1 - e^{-2sn}$~1 when n is large, then we may replace $1 - e^{-2sn}$ with 1 in the expectation as an approximation. After replacing $1 - e^{-2sn}$ with 1 and observing that the sum is now the unconditional expectation, we obtain

$$E_N[\, n(t) \mid F \,] \;=\; \frac{1}{1 - e^{-2s}}\; E_N[\, n(t) \,]$$

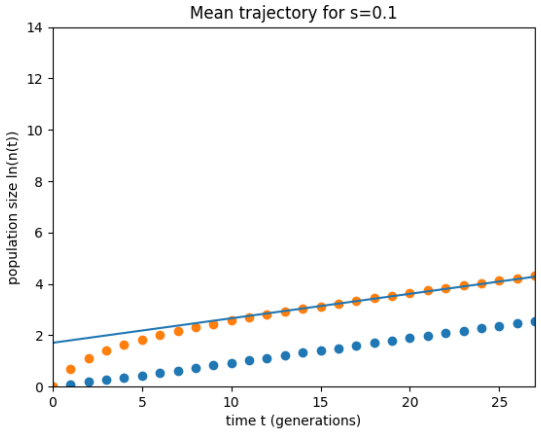
If we let $N >> ns$ and drop the subscripts, the expectation on the right-hand side becomes the mean trajectory $E[n(t)] \;=\; (1+s)^t$ that we derived in the previous section. Substituting then gives us

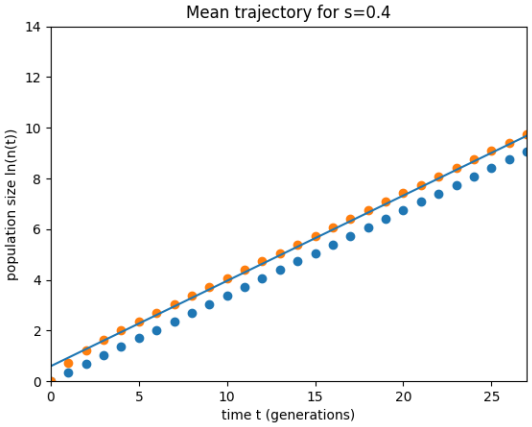
$$E[\, n(t) \mid F \,] \;=\; \frac{1}{1 - e^{-2s}}\; (1+s)^t$$

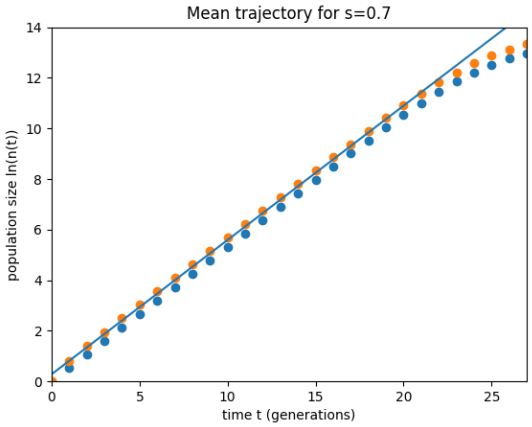
Note that the above approximation is not valid for small values of t, for example, evaluating at $t = 0$ does not give an average population size of 1, because we assumed that t must be sufficiently large.

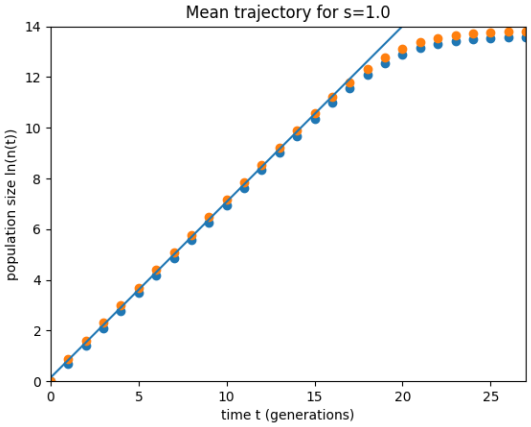
Importantly, the above approximation has an interesting biological interpretation. The growth dynamics excluding extinction events is functionally equivalent to the growth dynamics of unconditional trajectories that begin with a clone size $\frac{1}{1-e^{-2s}}$. Later, we will use this observation to show that s could be inferred without prior knowledge of mutation rate.

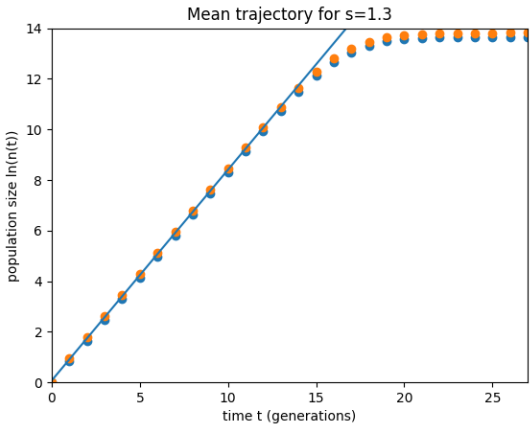
To validate the approximation derived above, we simulated the average growth dynamics of the mutant cells. Figures below show the simulated mean trajectory (blue dots), the simulated mean trajectories conditioned on no stochastic extinction (orange dots), and then our approximation of the mean trajectories conditioned on no stochastic extinction (blue curve) for different values of s. The simulated mean trajectories were collected by taking the average number of mutant cells at each time slice over a large number of simulated clonal expansions. The simulated mean trajectories conditioned on no stochastic extinction were generated similarly by first letting the simulated clones expand until they either went extinct or fixed, discarding the clones that went extinct, and then taking the average number of mutant cells at each time slice over the remaining clonal expansions. All simulated expansions were run with $N = 10^6$ and are shown for $g = 28$ generations.











Decoupling of fitness and mutation rate for weakly expanding clones

Under the neutral Wright-Fisher model, it is not possible to separately infer population size without knowledge of the mutation rate per generation. If we underestimate the mutation rate, we will

overestimate the number of generations between coalescent events in the tree, and thereby overestimate the population size. Conversely, if we overestimate the mutation rate, we will underestimate the number of generations between coalescent events in the tree, and thereby underestimate the population size. It is therefore necessary to have a priori information about one parameter, for example mutation rate, to extract any information about the other, for example population size, from the tree.

Fitness, or growth rate per year s_y, can be inferred from the shape of the reconstructed lineage tree of a small number of cells randomly sampled from the population at the final time point. If s_y is large, the population of mutated stem cells grows rapidly, therefore the coalescent events on the lineage tree will be confined to the first few generations, when the population size was small. Conversely, if s_y is small, coalescent events are more likely to occur in the last few generations. Critically, unlike population size, s_y can be inferred without any knowledge of the mutation rate, or equivalently, the total number of generations along the lineage tree. To intuitively understand this, note that rescaling the number of generations by a given factor scales the inferred population size by the same factor. Because at the onset of disease there is only one mutant cell, it might be expected that the growth rate, or s, must also be changed to achieve the scaled population size at the final time point. However, rescaling the number of generations also scales the minimum population size required before the mutated cells can escape stochastic extinction and grow exponentially. This is because more generations implies that the fitness advantage per generation is smaller and therefore the population is more susceptible to going extinct from random birth and death events. Taken together, these two competing effects precisely cancel and thus s_y can be inferred directly from the observed lineage trees without knowledge of the mutation rate or the number of generations.

We will make the above intuition precise by deriving the analytical expression for coalescent statistics as a function of s_y and showing that when expressed as growth per unit time, s_y does not vary with the mutation rate.

Mathematical analysis of mutation rate per generation and rate of population growth per unit time

Here, we will show that it is possible to infer the population growth rate per unit time without knowledge of the mutation rate per generation.

First, we will define the population growth rate per unit time. Then, we will derive an expression for the expected coalescent times of a random sample of mutant cells, and use it to estimate the impact of mutation rate per generation on the inferred population growth rate.

Define L to be the total number of divisions that an HSC would have undergone averaged across all HSCs, or equivalently the total number of generations in our trees. Note that knowing L is equivalent to knowing the mutation rate per generation, since we know the number of mutations accrued throughout the patient's life.

Definition of population growth rate per unit time

Note that $1 + s$, where s is the selection parameter, is the average growth per generation. We can also define a related quantity s_y as the average percent growth per unit time (for example percent growth rate per year). If we let a be the age of the patient, the number of generations per unit time is $\frac{L}{a}$. Hence, per unit time, the mutant clone is expected to grow by a factor of $(1 + s)^{\frac{L}{a}}$, and so we arrive at the expression

$$s_y(s, L) = (1 + s)^{\frac{L}{a}} - 1$$

Estimating the coalescent times

We now derive an expression for the expected time for coalescence of k randomly sampled mutant cells given that the clone has expanded for g generations.

Let t denote time in number of generations measured from the leaves of the tree towards the root. If there are $n(t)$ mutants at generation t, then the amount of coalescence time that passes from generation t to $t + 1$is $\frac{1}{n(t)}$, whereby coalescence time refers to the timescale in the Kingman Coalescent model (Kingman, 1982a, 1982b, 1982c). To understand what we are doing intuitively, note that for the standard Wright-Fisher model without selection, where the population size N is constant over time, the average time for coalescence of k randomly sampled lineages is $\frac{N}{\binom{k}{2}}$generations. This is generally computed by scaling time so that N generations correspond to 1 unit of time, and then letting N become large. In doing so, the times of coalescence of k randomly

sampled lineages converge to the Kingman Coalescent where the coalescence times are known to be $\frac{1}{\binom{k}{2}}$. The coalescence times in generations can then be recovered through an inverse time-scale transformation. Note that this is equivalent to scaling time so that the time between two neighboring generations is $\frac{1}{N}$. To account for a variable population size, we let the time between two neighboring generations t and $t+1$ be the inverse of the population size at t and assume the population size is always large. In doing so, the times until coalescence also converge to the Kingman Coalescent model. The statistics of coalescence times are then recovered by transforming back to time in generations.

Since the expected coalescence time of k lineages is $\frac{1}{\binom{k}{2}}$ in units of coalescence, and since the population size in our simulations grows approximately as $n(t) = \frac{1}{1-e^{-2s}}(1+s)^t$, the expected coalescence time of k lineages in units of generations is the t satisfying:

$$\frac{1}{\binom{k}{2}} = \sum_{k=0}^{t-1} \frac{1}{\frac{1}{1-e^{-2s}}(1+s)^{g-k}}$$

We then notice the R.H.S. is a geometric sum and rewrite to obtain:

$$1 + \frac{s}{1-e^{-2s}} \frac{(1+s)^g}{\binom{k}{2}} = (1+s)^t$$

Solving for t we obtain:

$$t = \frac{log\left(1 + \frac{s}{1-e^{-2s}} \frac{(1+s)^g}{\binom{k}{2}}\right)}{log\,(1+s)}$$

We use $\frac{L}{a}$ to convert 3) from generational time to real time (such as years), thereby obtaining

$$t_k(s,g,L) = \frac{a}{L}\frac{log\left(1+\frac{s}{1-e^{-2s}}\frac{(1+s)^g}{\left(\frac{k}{2}\right)}\right)}{log\,(1+s)}$$

Invariance theorem for weakly expanding clones

We now show that for a weakly expanding clone, we always infer the correct percent growth rate per unit time (for example per year) independent of our assumption of L.

We begin by assuming that $s_p, g_p,$ and L_p are the parameter values associated with our patient's tree, and that $|s_p| \ll 1$ so that selection is weak.

We then assume that we have incorrectly estimated our g and L parameters (i.e., our mutation rate) so that we erroneously believe they are $g_c = cg_p$ and $L_c = cL_p$ respectively. Note that we have kept the ratio $\frac{g_p}{L_p} = \frac{g_c}{L_c}$, and so we have treated the arrival time of the first mutated cell in real time as known.

Then we show that if we incorrectly assume our parameter values to be g_c and L_c, we then infer $s_c = (1+s_p)^{\frac{1}{c}} - 1$ for our s parameter, where

$$s_y(s_c, L_c) = s_y(s_p, L_p)$$

That is, we always infer the same percent growth per year. The way we show our inferred s is s_c is by plugging in g_c, L_c and s_c into the expected coalescent time expression we derived, and then showing that the coalescent times are identical to having plugged in $s_p, g_p,$ and L_p. In other words, we show that when we erroneously assume g_c and L_c are our parameter values, the s value that generates trees that match our patient's is s_c, and that s_c we infer combined with the L_c we've

assumed give us the same inference for yearly percent growth as the correct parameter values. We first show that s_c is our inferred s:

Begin by recalling that our coalescent time expression is given by

$$t_k(s,g,L) = \frac{a}{L}\frac{log\left(1+\frac{s}{1-e^{-2s}}\frac{(1+s)^g}{\left(\frac{k}{2}\right)}\right)}{log\,(1+s)}$$

Since $|s| \ll 1$, a Taylor expansion lets us make the following approximation:

$1-e^{-2s}$ ~$2s$

where we have let the 2nd order terms vanish. Substituting above and cancelling gives us

$$t_k(s,g,L) = \frac{a}{L}\frac{log\left(1+\frac{1}{2}\frac{(1+s)^g}{\left(\frac{k}{2}\right)}\right)}{log\,(1+s)}$$

But then using $s_c = \left(1+s_p\right)^{\frac{1}{c}} - 1$, $g_c = cg_p$ and $L_c = cL_p$ we can show:

$$t_k(s_c,g_c,L_c) = \frac{a}{L_c}\frac{log\left(1+\frac{1}{2}\frac{(1+s_c)^{g_c}}{\left(\frac{k}{2}\right)}\right)}{log\,(1+s_c)} = \frac{a}{cL_p}\frac{log\left(1+\frac{1}{2}\frac{\left(1+\left[(1+\,s_p)^{\frac{1}{c}}-1\right]\right)^{cg_p}}{\left(\frac{k}{2}\right)}\right)}{log\left(1+\,[(1+\,s_p)^{\frac{1}{c}}-1]\right)}$$

$$= \frac{a}{L_p} \frac{log\left(1 + \frac{1}{2}\frac{(1 + s_p)^{g_p}}{\left(\frac{k}{2}\right)}\right)}{log\,(1 + s_p)} = t_k(s_p, g_p, L_p)$$

so that s_c is our inferred s.

We then show our inference of percent growth is identical using $s_c = (1 + s_p)^{\frac{1}{c}} - 1$ and $L_c = cL_p$:

$$s_y(s_c, L_c) = (1 + s_c)^{\frac{L_c}{a}} - 1 = \left(1 + \left[\left[(1 + s_p)^{\frac{1}{c}} - 1\right]\right)^{\frac{cL_p}{a}} - 1$$

$$= (1 + s_p)^{\frac{L_p}{a}} - 1 = s_y(s_p, L_p)$$

2. Description of Simulations for Inference and Validation

Description of ABC

Approximate Bayesian Computation, or ABC, is an algorithm used to infer the parameter values of a stochastic model. ABC works by simulating data with the model using parameter values drawn from a prior distribution, and then computing a metric distance between the simulated data and the observed data. If their distance is smaller than a predetermined threshold, the parameter values are retained, otherwise, they are discarded. This procedure is iterated until a sufficient number of parameter values are retained to construct the posterior distribution.

ABC implementation

To perform ABC, it is first necessary to define a model. We briefly describe our model (see Section 1 for more details), and then give a detailed description of our ABC inference algorithm.

The model we used to infer the population dynamics of mutant cells is a variation of the Wright-Fisher model with selection (Fisher, 1923; Wright, 1931). Briefly, we consider a population of N stem cells that exists in discrete generations. There are L generations in total. At each generation, each cell chooses a parent cell at random from the previous generation. After t' generations, a cancerous mutation is acquired by one of the cells. Critically, the mutant cells are $1 + s$ times as likely to be chosen as a parent than the wild type cells. As a result, the number of mutant cells grows as $\sim (1 + s)^i$, for $i = 1, \ldots , g$, where $g = L - t$' corresponds to the disease duration. For convenience we use g as opposed to t' as a parameter in the following sections. However, provided L is given, if we know the value of g, we also know the value of t' and *vice versa*, and so the two parameters are equivalent. To summarize, the parameters of our model are:

N = saturation parameter (the total number of stem cells)

L = total number of generations

g = disease duration ($g <= L$)

s = selection parameter

We also define n as the number of mutant cells at the final time-point.

We now outline the steps of the ABC algorithm, and then elaborate on the details.

1. Draw s from its prior distribution.
2. Draw N, L, and g from their prior distributions.
3. Simulate a clonal expansion with our model for g generations.

4. If the final number of mutant cells is $n < k$, where k is the number of mutant cells we sample from the final population, back to 2). Else, move on to 5).
5. Sample k mutant cells from the final population and simulate their lineage history.
6. Simulate the number of mutations along the branches of the tree with the given mutation rate.
7. Perform the averaging algorithm on the tree so that the number of mutations from any leaf to the root of the tree is the same.
8. Convert the tree to an LTT plot.
9. If the area between the LTT plot of our simulated tree and the LTT plot of our patient tree is smaller than epsilon, retain the parameter values, otherwise discard them.
10. If a sufficient number of parameters to construct a distribution has been retained, finish. Else, back to 1).

Note that to perform the ABC, we must first specify the prior distributions on s, N, L and g (to test the robustness of our simulation *in silico*, we will sometimes fix parameter values rather than drawing from a distribution), the number of cells we will sample from the final population k, the mutation rate, and the epsilon threshold for retaining or discarding parameter values.
We now elaborate on the details of each step. We begin with 3) since 1) and 2) simply involve assigning a distribution, which will be specified when the simulations are described below.

After drawing the parameter values in 1) and 2), we simulate a clonal expansion for g generations. By a clonal expansion, we mean the number of mutant cells as a function of time. This can be attained in linear time complexity through a series of binomial draws. We begin with an initial condition of one mutant cell, since the number of mutant cells is always one when the mutation first arises. Then, assuming there are $n(i)$ cells in the ith generation, the number of mutant cells in the $(i + 1)$th generation is drawn from a binomial distribution with parameters N and $p = n(i)*(1+s)/(N + n(i)*s)$ (see Section 1 for the derivation). After iterating the binomial draw g times beginning with the initial condition, we recover the number of mutant cells as a function of time (see figure below that shows the ABC schematic). If the mutant clone does not grow to at least k cells, we redraw the parameters in 2) and re-simulate 3), iterating until we have acquired an expansion that does.
It is important to note that the clonal expansion is conditioned on $n >= k$, which is equivalent to conditioning on the mutant cells escaping stochastic extinction. In general, if the clonal expansion is simulated for a sufficient number of generations, that is for a sufficiently large g, then the mutant

clone will either go extinct or grow to a large size and exhibit deterministic dynamics. Therefore, when g is sufficiently large, if the mutant clone has grown to more than *k* cells, its size will be much larger than *k* and will have escaped stochastic extinction. Conditioning on escaping stochastic extinction has the important consequence of allowing us to infer the fitness in percent growth per year from the lineage trees (see Section 1 for details).

After obtaining a clonal expansion, we randomly sample *k* cells from the population of mutant cells at the final timepoint and simulate the lineage history of only the random sample, while ignoring the lineage history of all other cells. The lineage history is constructed by letting each sampled cell choose a mutant cell at random from the previous generation to be its parent. Each mutant cell chosen from the previous generation then chooses its parent at random from the generation of mutants before, etc. This is repeated until all lineages have coalesced (figure below).

It is worth noting that simulating the number of mutant cells as a function of time and thus initially ignoring all genealogical relationships, and then simulating only the genealogical history of the random sample backwards in time is statistically equivalent to simulating the genealogical process forward in time, and then producing a tree by following the lineages of the random sample back to common ancestry. This equivalence follows from the fact that each mutant cell can be descended from any of the mutant cells in the previous generation with equal probability. Simulating the lineage history for only the subset of the k chosen cells significantly increases the speed of the simulations without the loss of any information.

Once we have simulated the lineage tree, we then simulate the mutational process. To each edge of the tree (edge refers to a single line connecting two nodes on the tree, see figure below), we assign the number of mutations drawn from a Poisson distribution, where the mean is equal to the mutation rate (in units of mutations per generation). The mutation rate is generally computed empirically from the patient tree by dividing the total length of the patient tree in mutations by the value of *L* - 1, where *L* was drawn in step 2). It is very important to note that in this case, when we refer to the length of the tree, we mean the total number of mutations from the very bottom of the tree, which corresponds to the present time, to the very top of the tree which corresponds to the birth of the patient (not to the common ancestor of the mutant cells).

Before computing the mutation rate empirically from the data tree, we need to rescale the branches of the tree so that the distance (in mutations) from any leaf to the root of the tree is the same. If we don't do this, the number of mutations from each leaf to the root of the tree would not be the same, resulting in a tree length and mutation rate that is not well-defined. We accomplished

this by applying an averaging algorithm described below. The same averaging algorithm is also applied to simulated trees immediately after they are constructed before computing the metric distance between the simulated tree and the data tree. Therefore, any information loss from the algorithm will be expressed as uncertainty in the error bars of our inference.

The averaging algorithm we designed is based on the principle that the best estimate of time to common ancestry between two lineages is the average number of mutations between the two. In particular, let's define a tree as well-averaged if the distance from any leaf to the root is the same. In pseudocode, the algorithm works by calling the following function on the parent of any two sisters:

```
Average( currentNode )
{
```

If the left subtree of currentNode is not well-averaged:

```
    Average(left child of currentNode)
```

If the right subtree of currentNode is not well-averaged:

Average(right child of currentNode)

If both the left and right subtrees are well-averaged:

Compute the average length of the left and right subtrees. Then, for both the left and right subtree, rescale the branches of the subtree proportionally so that the length of the subtree equals the average.

```
    if currentNode != root:
        Average( parent of currentNode)
    else:
        break
}
```

We begin at the parent of two sisters, where the subtrees are single branches connecting a parent node to two leaf nodes. Note that we may start at the parent of any two sisters (or even more generally, at any node) and produce the same averaged tree since averaging the two subtrees of any node produces a unique value. For implementation of this algorithm refer to our GitHub repository. See figure below for a schematic of the averaging algorithm.

After producing a well-averaged tree, we construct its LTT (Lineages Through Time) plot. The LTT plot of a tree shows the number of lineages as a function of time in mutations (figure below). The LTT plot of a tree loses all information about its topology (the way the branches are connected). However, since the mutant cells in each generation pick their parents at random from the mutant cells in the previous generation, any topology on the tree is equally likely, and thus the tree topology contains no information about the parameter values that gave rise to the tree. Therefore, LTT contains all possible information about the parameter values.

After converting the simulated tree to an LTT plot, we compute the distance between the LTT plot of the simulated tree to the LTT plot of the data tree, defined as the area between the two LTT curves. The LTT plot of the data tree is always constructed before the ABC begins by first applying the averaging algorithm we previously described so that the leaf nodes line up side by side, and then converting it to an LTT plot. If the area between the two plots is smaller than the epsilon threshold, we retain the parameter values s, N, L, and g drawn from the priors, as well as the cancer trajectories $n(t)$ and the LTT curve produced, and if the area is >= epsilon we discard them. This process is iterated until a sufficient number of parameter values (along with the trajectories and LTT plots) to construct a convergent posterior distribution is retained.

The LTT curves start at zero but may end at different values because of different tree lengths. The area between two LTT curves that do not end at the same point on the x axis is undefined. To address this, we extend the end points of LTT curves, which corresponds to a value of 1, to infinity.

Since the lengths of LTT curves tended to vary, we decided to divide the area by k^*(the length of the data tree) before checking if epsilon was smaller than the threshold. This allowed us to run ABC without having to choose a new epsilon for each tree, since a smaller epsilon would be required for a tree of smaller length, and a larger epsilon for a tree with a larger length. Intuitively, this is equivalent to taking the percent difference between the data tree and the simulated trees.

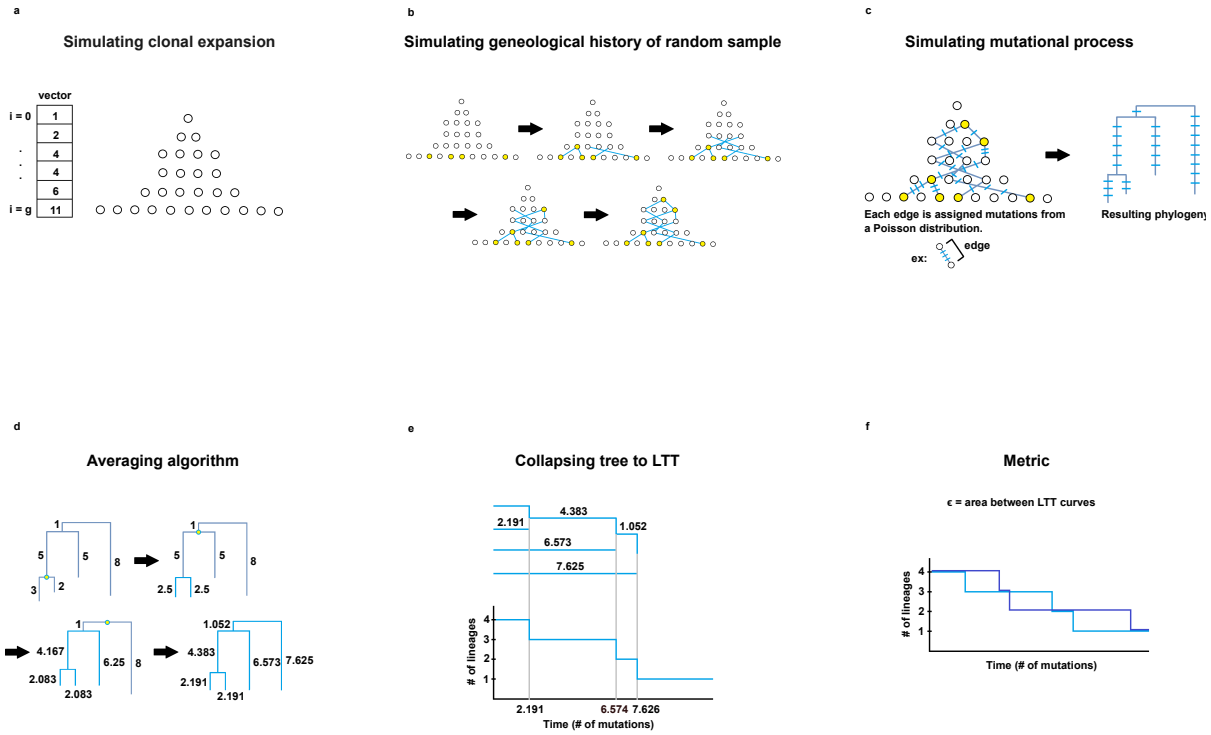


Figure. Schematic of the Approximate Bayesian Computation. First, the parameters that determine the growth dynamics (i.e. fitness, population size, total number of stem cells, and age of onset) are randomly drawn from a prior distribution. a. The clonal expansion of the cancer cells given the selected parameters is simulated. b. A specified number of cells is randomly chosen from the final population. The lineage tree is reconstructed for these cells. c. We assign a number of mutations accrued along each branch by drawing from a Poisson distribution at each edge with the mutation rate. Those mutations are shown pictorially as blue dashes on the tree. The resulting tree is a phylogeny with branch lengths in mutations as opposed to generations. d. Next, the branch lengths are scaled so that the total number of mutations from the root of the tree to each leaf node is the same. e. The rescaled trees are then plotted as LTT curves. f. A distance is computed between the LTT curve of the simulated tree and the observed tree. If the distance is below a threshold, the initial set of parameters is retained, otherwise they are discarded. This process is iterated.

Simulating data

To test the robustness of our ABC inference, we were interested in inferring the model parameters from simulated data where the ground truth is known. The way we simulated data as follows:

1. Draw *s* from its prior distribution.
2. Draw *N*, *L*, and *g* from their prior distributions.

3. Simulate a clonal expansion for g generations.
4. If the final number of mutant cells $n < k$, where k is the number of mutant cells, we sample from the final population, back to 2). Else, move on to 5).
5. Sample k mutant cells from the final population and simulate their lineage history.
6. Simulate the mutational process on the tree with the given mutation rate.
7. Perform the averaging algorithm on the tree.

The resulting tree is used as the data tree in ABC, and its parameters are inferred. Note that these steps are simply the first 7 steps of ABC.

In silico validation of the inference algorithm

To validate our inference, we decided to test the inference on simulated data over a wide range of parameter values. We began by simulating 30 trees as data with our model, where the underlying parameter values were known. Each tree was constructed using the following specifications:

1. s was drawn from a uniform distribution on (0, 1.2).
2. N was drawn from 10^X, where X is uniformly distributed on (1, 9).
3. L was drawn from round(Y), where Y is a Gaussian with mean 35 and std 5. If we drew $L < 2$, we redrew L until L >= 2 since at least 2 generations are necessary to produce a tree.
4. g was drawn uniformly on 2, … , L.
5. k = 22 mutant cells were randomly sampled. (22 is the number of mutant stem cells sampled for ET 1 patient data)
6. The mutation rate was 723/(L - 1). 723 was the number of mutations observed in ET 1 patient data.

We then inferred the parameters for each tree using ABC. For the ABC, we used the exact same specifications as the data to generate trees for comparison, except that we instead drew s from a uniform distribution on (0, 5) in Step 1, and the mutation rate was instead estimated empirically as (total length of tree in data in # of mutations)/(L - 1) in Step 6. Epsilon was set to 0.03, since this threshold was sufficient to obtain an inferred distribution that converges to the posterior distribution for most inferences, while also allowing a large number of points to be retained for the inferred distribution. The simulations were run until they accrued ~10,000 or more points for the posterior.

In the figure called “In silico validation of the inference algorithm without feedback,” we show a representative set of 10 inferences out of the 30 inferences we ran.

To quantify the accuracy of our ABC inference, we then simulated 200 trees in precisely the same way as above, except that the value of *s* for the tree data was drawn uniformly on (0, 2) instead to obtain data across a much wider range of fitness values. We then carried out ABC inferences on each tree with epsilon = 0.0225 until most of the posterior distributions had accrued ~400 or more points. Tree data where the inference accrued less than 30 points for the posterior were excluded.

We then applied the following filters to the data:

1. We excluded data trees where the ratio of the standard deviation to the mean of the posterior of *s* (in percent growth per year) was greater than 0.425
2. We excluded data trees where the std of *n* was larger than 1.15

We arrived at the first filtering criterion by noting that inferences for small s tended to have large error bars relative to their inferred means (or large coefficient of variation), and their inferred means were generally inaccurate and much larger than the true values. ABC inference cannot determine whether a small number of cells at the final time-point is due to a small growth rate *s* or small saturation limit (see figure “In silico validation of the inference algorithm with feedback,” row 2 column 2). In both scenarios the population size is small and coalescence events occur rapidly, producing similar trees. We reasoned that if expansions produced by small s produce ABC inferences characterized by large coefficient of variation, then by eliminating ABC inferences exhibiting this characteristic we could exclude inaccurate inferences without any knowledge of the ground truth.

Similarly, we arrived at the second filtering criterion because simulations that expanded to sufficiently large population sizes generated inferred log(*n*) distributions with large standard deviations and mean values distributed around 10^6. This suggested that ABC was extracting little information from the data and that the log(*n*) distributions were almost identical to the prior. We reasoned that by filtering out inferences with large log(*n*) standard deviations we could exclude inaccurate inferences without any knowledge of the ground truth value. We emphasize that this filtering procedure does not use the ground truth values in any way. The inference is deemed inaccurate if the posterior distribution width is too large regardless of the ground truth value. Therefore, the filtering procedure can also be applied to actual data where the ground truth is not known. Finally, when devising the filtering criteria, we were conservative with our choices. As such, the interpretation of these data was not sensitively dependent on the filters we chose.

The inferred vs true values of the inferences is plotted in the figure called “ABC accurately infers model parameters”.

Next, we quantified the accuracy of our ABC inference for a 63-year-old patient in a similar manner by simulating 200 trees, applying filters to the data, and then plotting the inferred vs true values (figure “ABC accurately infers model parameters”). The data were produced in a similar manner as for the 34-year-old patient using the following criteria:

We simulated 200 trees as data for a 63-year-old patient using the following specifications:

1. *s* was drawn from a uniform distribution on (0, 2).
2. *N* was drawn from 10^X, where *X* is uniformly distributed on (1, 9).
3. *L* was drawn from round(*Y*), where *Y* is a Gaussian with mean 64 and std 10. If we drew *L* < 2, we redrew *L* until L >= 2 since at least 2 generations are necessary to produce a tree.
4. *g* was drawn uniformly on 2, … , *L*.
5. we sampled *k* = 13 cancer cells, which is the number of sampled cells in the ET 2 patient data.
6. The mutation rate was 1205/*L*. 1205 was the number of mutations observed in ET 2 patient data.
7. No feedback was included.

We then carried out ABC inferences on each tree using the same specifications as the data, except that we drew s from (0, 5) uniformly in Step 1, and the mutation rate was estimated empirically as (total length of tree in data)/(*L*-1) for 6). Epsilon was set to 0.0125. The inferences were left running until about half of them (many ABC inferences accrued little to no points for the inferred distributions) had accrued ~100 or more points for the posterior distribution. Many simulations accrued little or no points, and so we excluded trees with less than 30 points.

We then applied the following filters to the data:

1. We excluded data trees where the mean of the posterior of s (in percent growth per generation) was larger than 1.5.
2. We excluded data trees where the ratio of the std to the mean of the posterior of s (in percent growth per year) was greater than 1.5.

Similar reasoning was applied to devise the above filtering criteria. Mainly, values outside of above criteria contain little information beyond the prior distributions.

ABC on simulated data with feedback

Our model of growth dynamics of the mutant cells only approximates the actual growth dynamics. In particular, the population of mutant cells seems to saturate when it has expanded to a certain fraction of the total population of stem cells. Therefore, it is conceivable that the mutants lose their fitness advantage as their population size increases. Here, we set out to test whether the inference of the parameters of the simple model of growth dynamics remains accurate if the actual dynamics is simulated using a different model. To do so, we constructed a model with feedback, whereby the fitness advantage of the mutant cells decreases compared to wild-type cells as their population size increases. The model with feedback is described in detail in Section 1.

We then repeated the simulations carried out for 34-year-old and 63-year-old patients, except that we used feedback with x = 30 when generating the simulated data (see Section 1 for definition of x parameter). The ABC did not incorporate feedback in the model, since we were interested in how well we could infer the parameter values if the ground truth incorporated feedback. For the 10 example trees shown in the figure called “In silico validation of the inference algorithm with feedback,” we ran ABC until most of the inferences had accrued ~5,000 or more points for the posterior. For the 200 trees for the 34-year-old (figure “ABC accurately infers model parameters”), we ran ABC until most of the inferences had accrued ~300 or more points for the posterior. For the 200 trees for the 63-year-old (figure “ABC accurately infers model parameters”), we ran ABC until about half of the inferences had accrued ~50 or more points for the posterior. Inferences that accrued less than 30 points were always excluded. The filters were applied in an identical fashion as for the inferences without feedback.

Taken together, the inferences suggest that the ABC inference can infer model parameters over a wide range of parameter values, regardless of whether or not feedback is incorporated in the underlying model. In particular, *s* in percent per year and the age of onset of the disease can be inferred from lineage trees, even if feedback is incorporated. However, it appears n can only be inferred for the 34-year-old patient. The inferred *n* vs true *n* plots for the 63-year-old patient indicate that the trees have no information about the number of mutant cells at the final time point. This is likely due to the fact that we have sampled only 13 lineages (as opposed to 22 lineages for the 34-year-old patient), and that the clone has expanded for much longer. Because of this, most coalescent events occur in the early history of the expansion, and information about the dynamics of the later history are lost.

Age = 34 and x = 0 (no feedback)

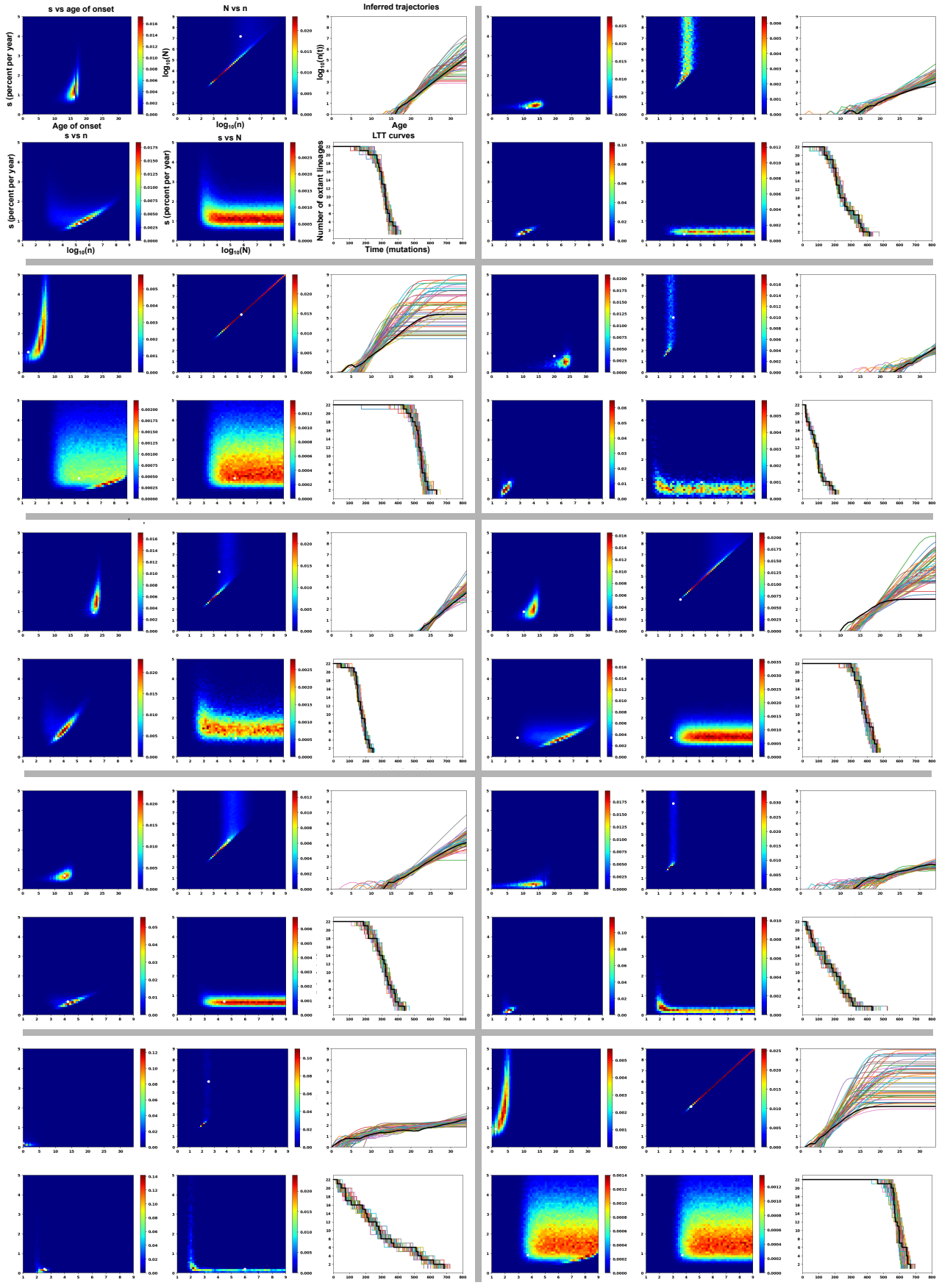


Figure. In silico validation of the inference algorithm without feedback. To validate the inference algorithm, we generated simulated ground truth growth dynamics and then inferred the parameters using ABC. In these simulations, the ground truth dynamics were simulated using the same model as in the ABC. In all the heatmaps, the ground truth parameters are shown as white dots. In the traces, the ground truth is shown in black. 10 illustrative examples are shown. For each example, the heatmaps of inference of s (fitness parameter), n (number of cancer cells), N (total populations size), and g (age of onset) are shown alongside the LTT curves that were retained and the inferred trajectories of population growth.

Age = 34 and x = 30 (with feedback)

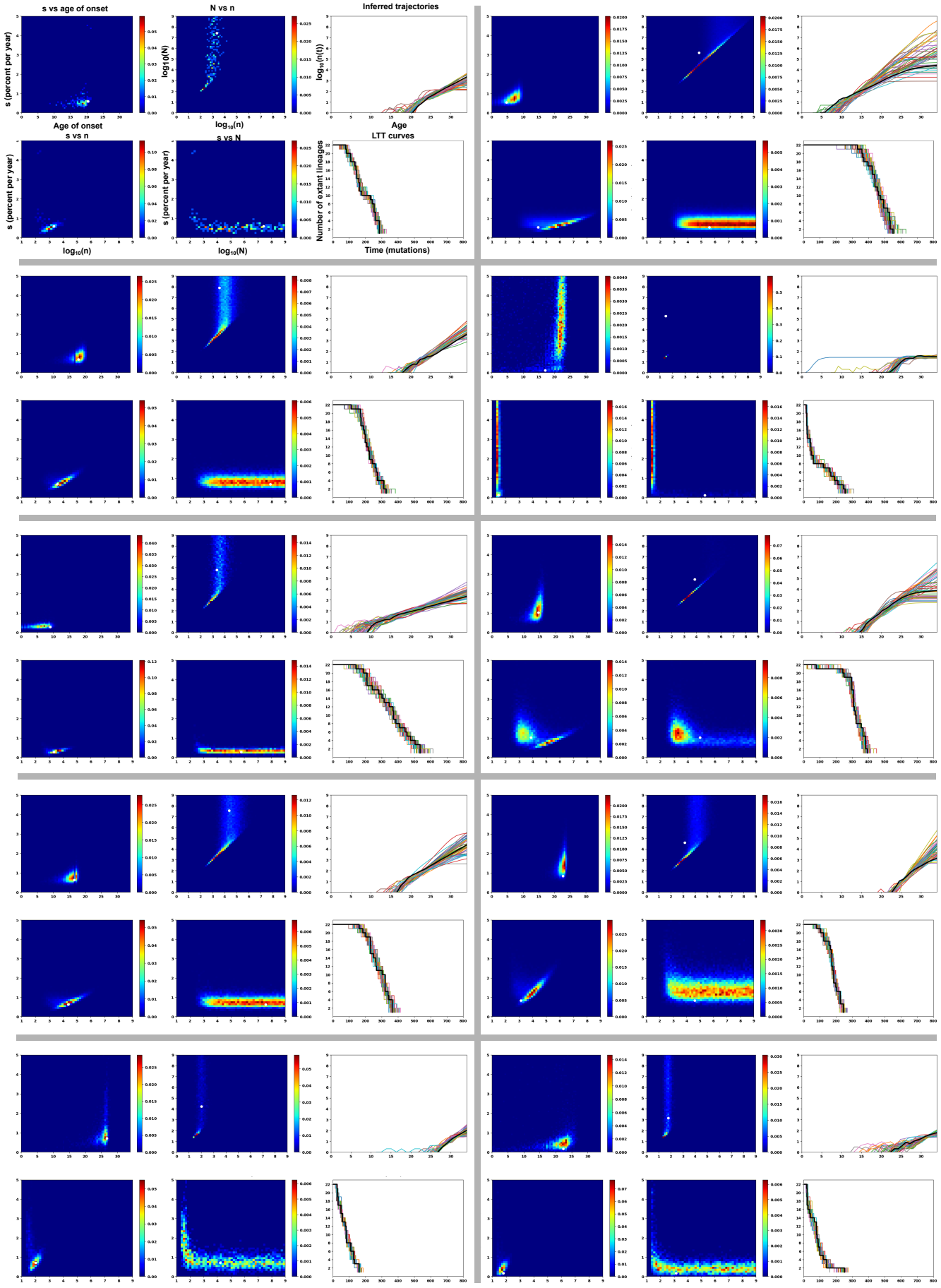


Figure. In silico validation of the inference algorithm with feedback. To validate the inference algorithm, we generated simulated ground truth growth dynamics and then inferred the parameters using ABC. In these simulations, the ground truth dynamics were simulated with a feedback where the fitness of cancer cells decreased as the population size increased. The growth dynamics for generating trajectories for ABC did not incorporate the feedback. In all the heatmaps, the ground truth parameters are shown as white dots. In the traces, the ground truth is shown in black. 10 illustrative examples are shown. For each example, the heatmaps of inference of s (fitness parameter), n (number of cancer cells), N (total populations size), and g (age of onset) are shown alongside the LTT curves that were retained and the inferred trajectories of population growth. Taken together, accurate inference is possible even if additional features, such as feedback, are not incorporated in the ABC dynamics.

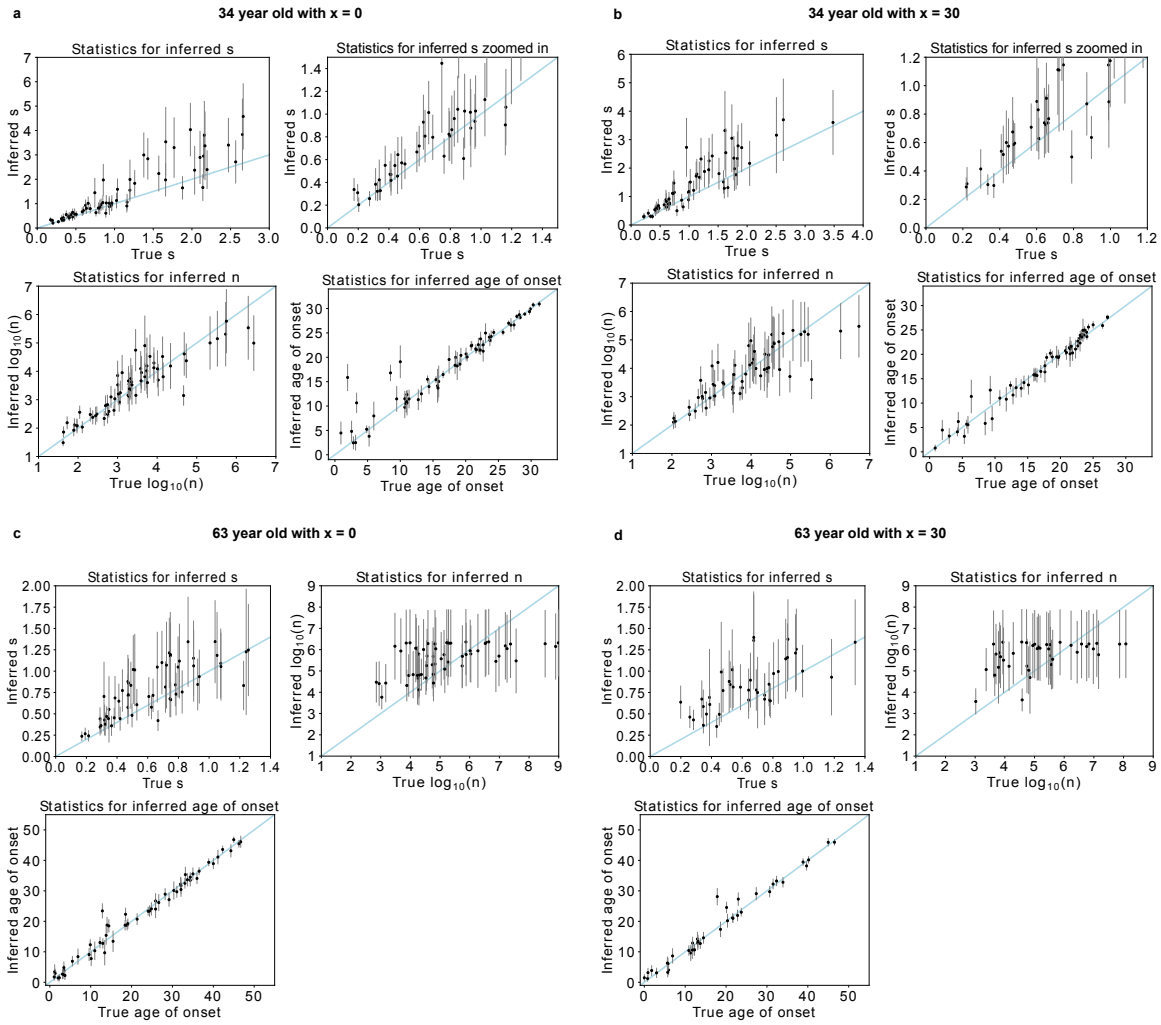


Figure. ABC accurately infers model parameters. To verify that the inference framework can accurately infer model parameters, we simulated growth dynamics across a range of scenarios (corresponding to different parameter values) and inferred the parameters using ABC. The inferred parameter values were plotted against their true values for s (percent per year), n, and age of onset of the disease (years), and error bars were included to denote 1std in the inference. Filters were applied to exclude inferences with large error margins, and were done completely agnostic of the ground truths. a-b Many iterations of ground truths were simulated for a 34-year-old patient. In a., the model used to simulate ground truths did not incorporate feedback, while in b. the model did. In both cases, the model used for the ABC inference did not incorporate feedback but was able to infer the parameters correctly within the statistical error. c. same as a. and d. same as b., except the ground truth simulations and ABC inferences were carried out for a 63-year-old patient. In both c. and d., the model used for the ABC inference also did not incorporate feedback, but was able to infer the parameter values of s and the age of onset of disease within the statistical error. However, for c. and d., n could not be inferred since there were not enough coalescent events in the later history to extract information about its dynamics after the initial expansion. For the inferred n vs true n plots in c. and d., we decided to show them with no filter so the reader could see that the inferences contain no information.

Fitness can be inferred without knowing the number of generations

So far, we have shown that our inference is robust to feedback but have not shown how well we can infer the parameter values if our assumption about the total number of generations is incorrect. Surprisingly, fitness, when converted to percent growth per year, can always be inferred without knowing the number of generations (Section 1). To validate this, we simulated ~10 data trees for a 34-year-old patient and carried out ABC on them. We then selected a data tree where the ABC inference precisely inferred the parameter values. The data tree had been simulated with the following specifications:

1. Parameter values were fixed at $s = 0.264911$, $N = 10^9$, $g = 50$, $L = 70$, arbitrary values for which the inference was accurate.
2. $k = 22$ cells were sampled
3. The mutation rate was 723/69 per generation

We then inferred *s*, *n*, and the age of onset of the disease having kept all other parameter values fixed, but assuming that the number of generations *L* was *c* times 70 (the ground truth *L*) in the

ABC model. More precisely, for each c = 0.5, 1, 2, 4, 8, 16, 32, 64 we ran an ABC inference on the data tree with the following specifications:

1. s was drawn uniformly on (0, 10/c)
2. We fixed $N = 10^9$ and $L = c*70$
3. g was drawn uniformly on 2, … , L
4. A mutation rate of 723/(70*c - 1) was used
5. An epsilon distance of 0.02 was used

The inferences were run until the ABC had accrued ~15,000 points for the posterior distribution. We then plotted inferred joint distributions for s in growth per generation, s in growth per year, age of onset of the disease, and n (figure below).

As expected, when increasing the number of generations assumed by the model for ABC inference, the inferred s in percent growth per generation decreased while the inferred percent growth per year remained invariant. In theory, the decrease in growth per generation will increase the rate of stochastic extinction, and so the number of mutant cells must fluctuate to a larger population size to escape stochastic extinction. As expected, the number of mutant cells increased at the final time point.

Taken together, our simulation results are consistent with our theoretical calculations (Section 1) in that fitness, in percent growth per year, and the age of onset in years can be inferred from lineage trees without prior knowledge of L, while prior knowledge of L is necessary to infer n.

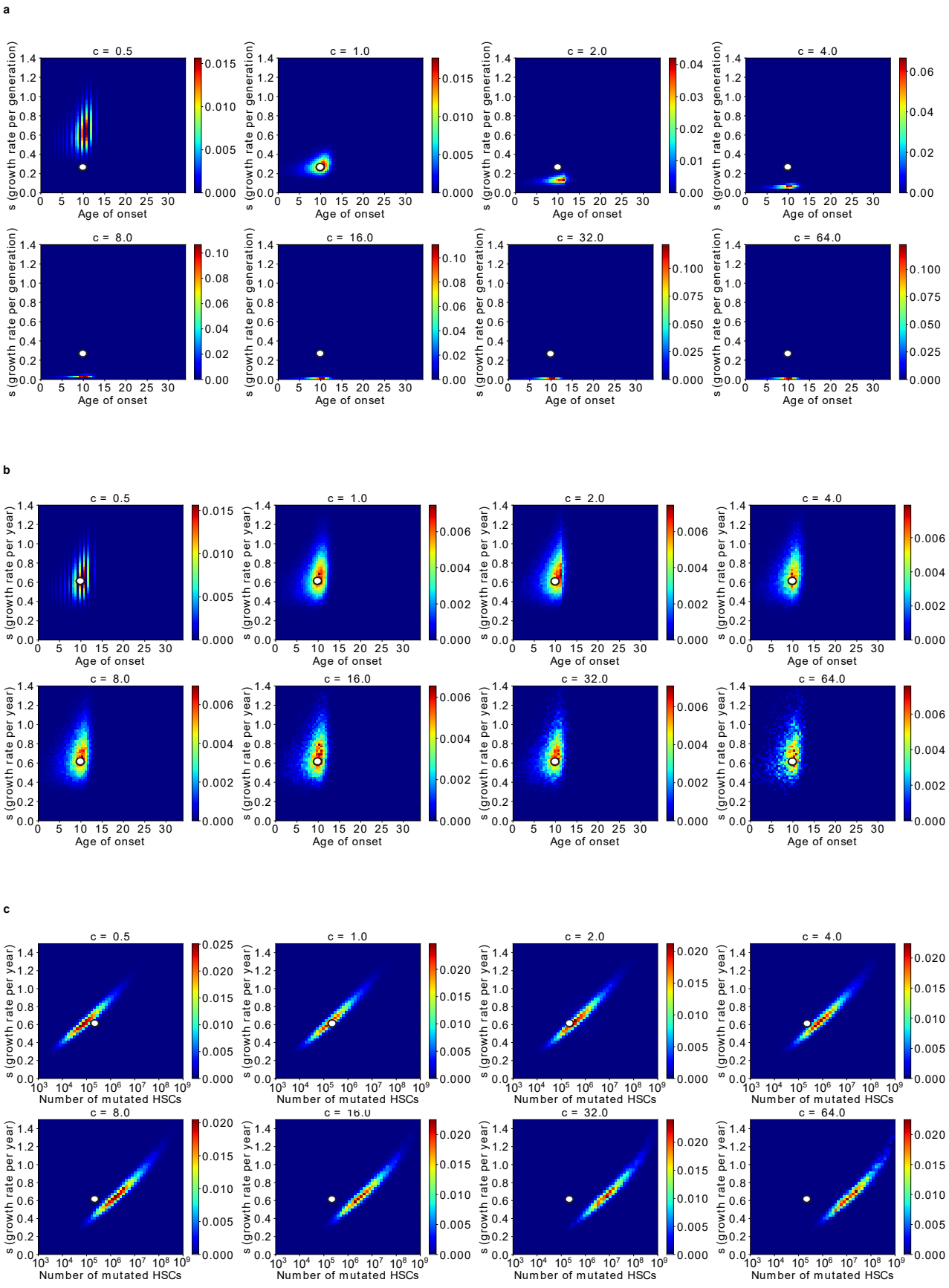

a
c = 0.5
c = 1.0
c = 2.0
c = 4.0
c = 8.0
c = 16.0
c = 32.0
c = 64.0
s (growth rate per generation)
Age of onset
b
c = 0.5
c = 1.0
c = 2.0
c = 4.0
c = 8.0
c = 16.0
c = 32.0
c = 64.0
s (growth rate per year)
Age of onset
c
c = 0.5
c = 1.0
c = 2.0
c = 4.0
c = 8.0
c = 16.0
c = 32.0
c = 64.0
s (growth rate per year)
Number of mutated HSCs

Figure. Fitness can be inferred without knowing the number of generations. Ground truth growth dynamics was simulated, and multiple ABC inferences were carried out under different c values. For each value of c, the number of generations used in ABC was multiplied by c for the ABC inference. a. inferred s (percent per generation) vs inferred age of onset of disease (years) are plotted for different values of c. As expected, increasing c, and hence the number of generations, decreases the growth per generation to keep the percent growth per year invariant. b. Inferred s was then converted to growth per year. As suggested by our derivation, the inferred s in growth per year remains invariant as the number of generations is scaled. c. The inferred s (percent per year) vs the inferred number of mutated HSCs at the final time-point is plotted for different values of c. As expected, increasing c, and hence the number of generations, decreases the growth per generation and thereby the rate of stochastic extinction. The number of mutated HSCs must then fluctuate to a larger population size early on to escape extinction. This results in a larger number of mutated stem cells at the final time-point.

The analytical calculation of coalescent times matches the simulations results

In Section 1, we provide an analytical calculation of the average coalescent times of our model, and show that the average coalescence times do not change if we scale the number of generations while keeping the percent growth the same (suggesting that fitness can be inferred without prior knowledge of the number of generations).

To validate our analytical calculation of coalescent times, we performed the following simulations. For each s = 0.1, 0.3, … 1.5 (growth per generation), we constructed thousands of data trees using the following specifications:

1. g = 25, L = 35, $N = 10^9$ (with the corresponding s)
2. k = 22 cells were randomly sampled
3. A mutation rate of 723/34 was used

We then converted the branches of each data tree to years, assuming the tree was for a 34-year-old patient, by multiplying the branch lengths by 34 / 723. Then, for each value of s separately, we constructed a distribution for each of the i = 1, …, 21 coalescence times using the corresponding trees. We computed the means and standard deviations of those distributions and plotted them (figure below).

The analytical derivation of coalescence times therefore matches the simulated coalescence times within a standard deviation. Our derivations also predict that the coalescence times of a tree should not change when scaling the number of generations while keeping the percent growth per year fixed. To verify this occurs in our simulated trees, we did the following:

For each s' = 0.1, 0.6, 1.1 and for each c = 0.5,1, 2, 10, 100, we simulated thousands of data trees using the following specifications:

1. $s = (1 + s')^{1/c} - 1$ (this s, in growth per generation, along with the L in specification 2), keep the percent growth per year invariant (See supplemental).
2. $L = c*35$, $g = c*25$, $N = 10^9$,
3. 22 cells were randomly sampled
4. A mutation rate of $723/(c*35 - 1)$ was used

We then converted the branches of each data tree to years, assuming the tree was for a 34-year-old patient, by multiplying the branch lengths by $(c*35 - 1)/723$. For each combination of s and c, we constructed distributions for each of the $i = 1, \ldots, 21$ coalescence times of the corresponding trees. We then computed the means and standard deviations of the distributions and plotted them (figure below).

Consistent with our theoretical predictions, scaling the number of generations while keeping the percent growth per year invariant appears to not significantly change the average times until coalescence. This implies that trees are indistinguishable when the percent growth per year is the same, even if the number of generations is different, showing that percent growth per year can be inferred from lineage trees without knowing L.

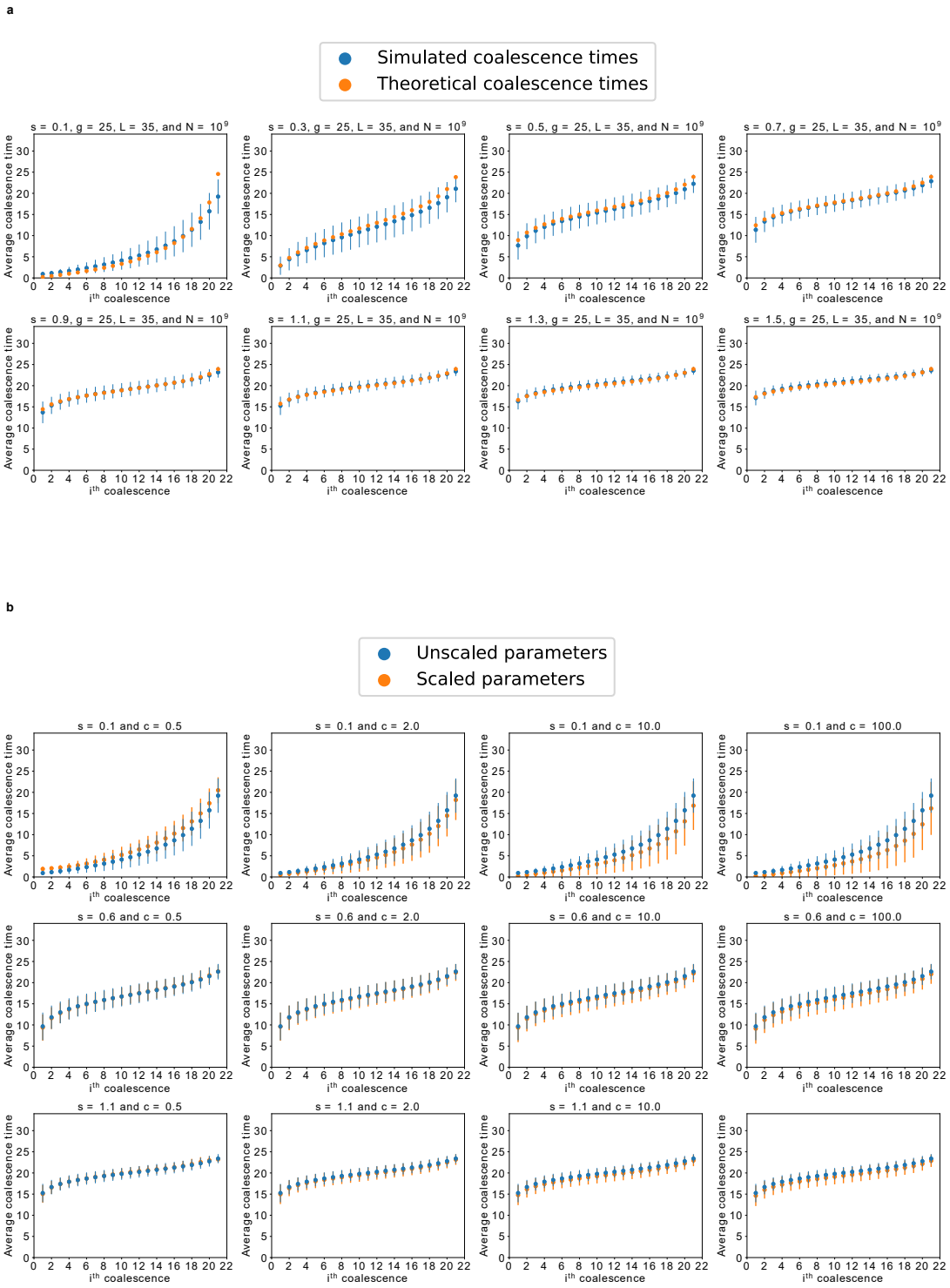


Figure. The analytical calculation of coalescent times matches the simulations results. a. We verify that our analytical calculation of the average coalescence times of our model matches the empirically observed coalescence times. Blue dots show the average coalescence times computed empirically for different s by simulating a large number of cancer expansions,

constructing trees, and then averaging over the times until coalescence. Error bars denote 1 std in the coalescence times. Orange dots show the average coalescence times derived analytically for comparison. b. Our mathematical derivation shows that scaling the number of generations by a factor of c while maintaining the percent growth per year, s, fixed does not change the average coalescence times of trees. We take s = 0.1, 0.6, and 1.1 in percent growth per year corresponding to rows 1, 2, and 3 respectively, scale the number of generations by different factors of c, and compute the average coalescent times empirically. Blue dots are the empirically computed coalescent times before scaling the number of generations by c, and the orange dots are the empirically computed coalescent times after scaling the number of generations by c. The average coalescent times using scaled parameter values always lie within 1 std of the average coalescent times using unscaled parameters. The small deviations are due to the fact that s in growth per year is not sufficiently small when the number of generations is not scaled. As c increases, increasing the number of generations, the growth per generation decreases to maintain constant percent growth per year. As the growth per generation decreases, the coalescent times converge according to our mathematical derivation.

ABC on simulated data with fitness heterogeneity

The Wright-Fisher model assumes that each cell at each generation has the same fitness value. However, it is conceivable that there is heterogeneity in the fitness of cells, or more precisely that the fitness of each cell comes from a distribution. To test the robustness of our inference to the assumption of homogeneous s, we simulated 20 data trees using the following specifications:

1. We fixed parameter values $N = 10^7$, $g = 25$, $L = 35$.
2. Clonal expansions were simulated for g generations, but with a slight modification to the Wright-Fisher process with selection that we use for ABC. In particular, each generation of cells creates the next generation of cells according to the following rule: Non-cancer cells have probability p of being chosen as a parent by a cell in the next generation, and the i^{th} cancer cell has probability $(1 + s_i)$*p of being chosen, where the s_i are i.i.d. Gaussians with mean 0.6 and std 0.2. That is, a fitness value is assigned to each cell in the current generation, and then each cell from the next generation picks a parent cell randomly according to the probabilities determined by the fitness values drawn from a Gaussian.
3. $k = 22$ mutant cells were randomly sampled
4. The mutation rate was $723/(L - 1)$

We then inferred the parameter values for each tree using the same specifications as the ABC model used for the 34-year-old patient data (see Inference on patient data at the end of the document). Results were plotted in the figure below and show that the mean growth per year can be accurately inferred, even when incorporating Gaussian noise.

To test how sensitive our results are to the assumed fitness distribution, we decided to simulate 20 data trees in an identical fashion as before, except that instead of drawing from a Gaussian we chose to draw from a uniform distribution on (0, 1.2). Results were plotted in the figure below and suggest that the mean growth per year can be accurately inferred, even when the fitness distribution is highly variable.

The Wright-Fisher model assumes a specific amount of genetic drift, an assumption that is often neglected. To test the impact of this assumption on our inference, we simulated 20 data trees using the following specifications:

1. We fixed parameter values $s = 0.6$, $N = 10^7$, $g = 25$, $L = 35$.
2. Clonal expansions were simulated for g generations. At each generation, 25% of the total population of stem cells were randomly “inactivated”, so that when the lineage trees were simulated, the cells could not select an “inactivated” cell. This forces a smaller number of stem cells to contribute to the exponential growth, so that genetic drift occurs more rapidly. If all the cells in a generation were randomly inactivated by chance, the clonal expansion was re-simulated since this represents an extinction event.
3. $k = 22$ mutant cells were randomly sampled
4. The mutation rate was $723/(L - 1)$

Similar to the previous simulations, we used the ABC model specifications that were applied to the data tree for the 34-year-old. The results are plotted in the figure below, and show that the inference is robust to a different amount of genetic drift.

We then decided to simulate a more realistic version of the previous scenario, where cells are continuously killed at a constant rate. 20 simulated trees were constructed in an identical fashion as above, except that instead of simulating the entire clonal expansion first and then inactivating

cells at each generation, the cells were killed off simultaneously with the exponential growth. In particular, before a new generation was produced, 25% of the total number of stem cells were randomly killed, and this new smaller pool of stem cells was used to create the next generation of cells. Note that in this case, the p-value of the binomial parameter would be computed using *N**0.75 instead of *N*, before *N* cells are chosen from current generation to create the next generation of cells. This is subtlety distinct from the drift scenario, where the parameter values for the binomial draws do not factor the temporary reduction in population size due to the random killing. The results are plotted in the figure below and show that the inference is also robust to randomly killing cells.

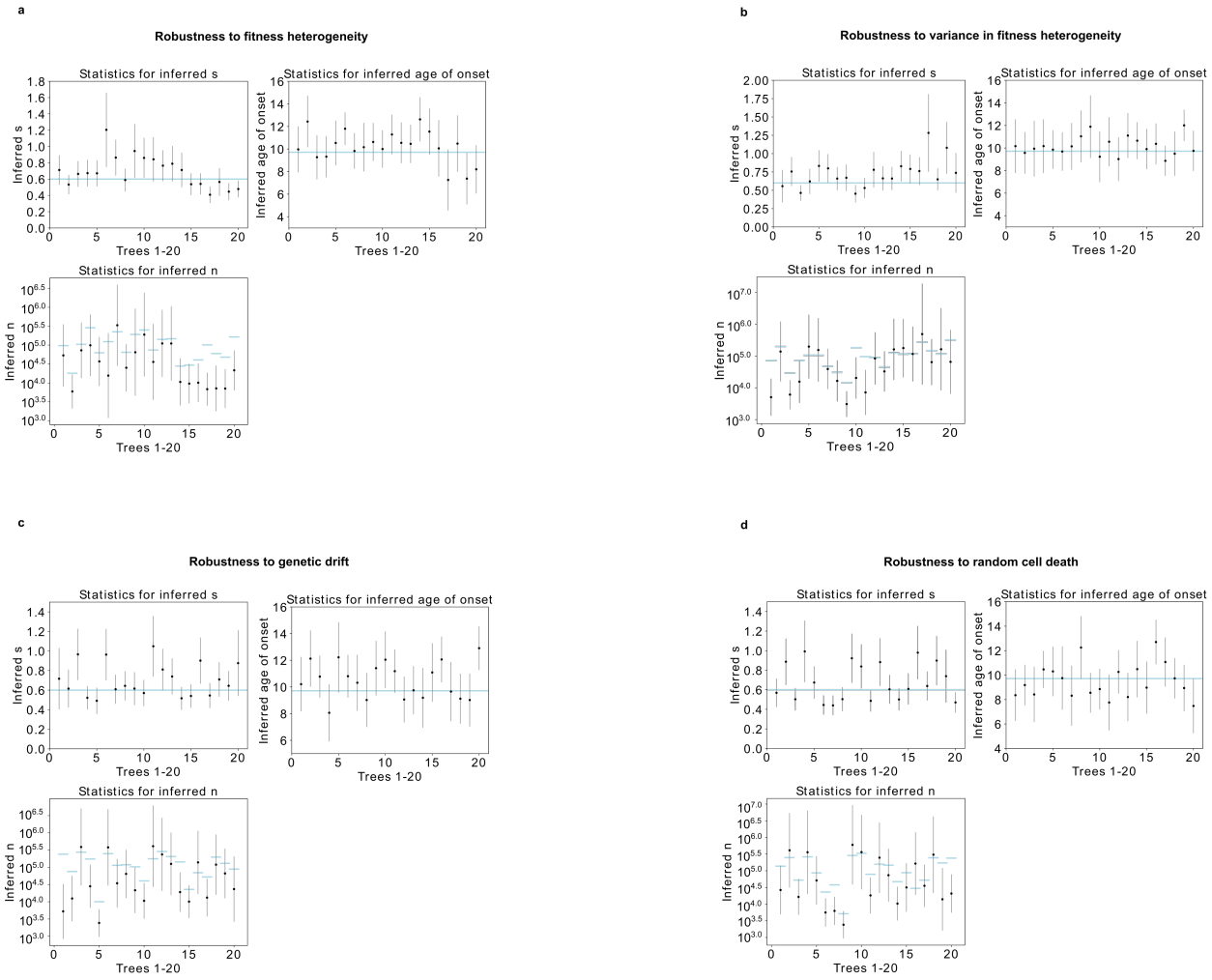


Figure. Robustness to cell heterogeneity. a. 20 data trees were simulated for a 34-year-old patient using parameter values g = 25, L = 35, and N = 10^7, and where the fitness of each cell was drawn from a Gaussin with mean 0.6 and std 0.2. ABC inferences were then carried out on each data tree using the usual model that was applied to data, which assumes s is the same for

each cell. The inferred mean of the posterior distributions for each tree were plotted with error bars to denote 1 std. The blue lines represent the ground truth parameter values. For the inferred vs true s figure, the blue line is just the mean growth per year of the clonal expansions. b. These figures were constructed in an identical fashion to figures 1b, except that s was instead drawn from a uniform distribution on (0, 1.2). c. 20 data trees were simulated for a 34-year-old patient by running clonal expansions with parameter values s = 0.6, g = 25, L = 35, and N = 10^7. After each clonal expansion, genetic drift was increased by inactivating 25% of the total number of stem cells at each generation so that ~75% of the cells were contributing to the exponential growth at each time step. ABC inferences were then carried out on each data tree using the usual model that was applied to data, which did not include the additional genetic drift. The inferred mean of the posterior distributions were plotted for each tree with error bars to denote 1 std. The blue lines represent the ground truth parameter values. d. 20 data trees were simulated for a 34-year-old patient by running clonal expansions with parameter values s = 0.6, g = 25, L = 35, and N = 10^7, except that 75% of the total population of stem cells was randomly killed at each generation before the next generation of cells was birthed. ABC inferences were then carried out on each data tree using the usual model that was applied to data, which does not include random killing. The inferred mean of the posterior distributions were plotted for each tree with error bars to denote 1 std. The blue lines represent the ground truth parameter values.