Article

Life histories of myeloproliferative neoplasms inferred from phylogenies

https://doi.org/10.1038/s41586-021-04312-6

Received: 16 October 2020

Accepted: 6 December 2021

Published online: 20 January 2022

Check for updates

Nicholas Williams[1], Joe Lee[1,2], Emily Mitchell[1,2,3,4], Luiza Moore[1], E. Joanna Baxter[3], James Hewinson[1], Kevin J. Dawson[1], Andrew Menzies[1], Anna L. Godfrey[4], Anthony R. Green[2,3,4,5], Peter J. Campbell[1,2,3,5] & Jyoti Nangalia[1,2,3,4,5] ✉

Mutations in cancer-associated genes drive tumour outgrowth, but our knowledge of the timing of driver mutations and subsequent clonal dynamics is limited[1–3]. Here, using whole-genome sequencing of 1,013 clonal haematopoietic colonies from 12 patients with myeloproliferative neoplasms, we identified 580,133 somatic mutations to reconstruct haematopoietic phylogenies and determine clonal histories. Driver mutations were estimated to occur early in life, including the in utero period. *JAK2*^{V617F} was estimated to have been acquired by 33 weeks of gestation to 10.8 years of age in 5 patients in whom *JAK2*^{V617F} was the first event. *DNMT3A* mutations were acquired by 8 weeks of gestation to 7.6 years of age in 4 patients, and a *PPM1D* mutation was acquired by 5.8 years of age. Additional genomic events occurred before or following *JAK2*^{V617F} acquisition and as independent clonal expansions. Sequential driver mutation acquisition was separated by decades across life, often outcompeting ancestral clones. The mean latency between *JAK2*^{V617F} acquisition and diagnosis was 30 years (range 11–54 years). Estimated historical rates of clonal expansion varied substantially (3% to 190% per year), increased with additional driver mutations, and predicted latency to diagnosis. Our study suggests that early driver mutation acquisition and life-long growth and evolution underlie adult myeloproliferative neoplasms, raising opportunities for earlier intervention and a new model for cancer development.

Human cancers harbour hundreds to many thousands of somatically acquired DNA mutations, a minority of which drive tumour initiation and progression[1]. These driver mutations occur in recurrently mutated cancer genes and stimulate the cell acquiring it to expand into a clone. With a large enough clonal expansion, typically abetted by further driver mutations, a cancer emerges. Little is known about the ages at which driver mutations occur, the timelines of clonal expansion over an individual's lifetime, or how these relate to clinical presentation with cancer. Some mutational processes accrue at a steady rate across life, representing a 'molecular clock'[4,5]. Knowing this tissue-specific rate of mutation accumulation has enabled broad estimates for the timing of mutations for some cancers[2,3].

In patients with blood cancers, the observation of normal blood counts months to years before diagnosis suggested that tumour development occurs quickly, and therefore driver mutations must occur late in life. Estimates from cancer incidences in people who survived the Hiroshima or Nagasaki atomic bombings who developed chronic myeloid leukaemia have suggested a mean latency time of only eight years between *BCR–ABL1* induction and clinical presentation[6]. However, the presence of driver mutations in normal tissues[7–12]—including blood from healthy individuals with clonal haematopoiesis[13–17], some of whom subsequently develop malignancies—supports a longer multi-hit trajectory of cancer. Understanding the absolute timelines of cancer evolution is critical for efforts aimed at early detection and intervention, especially if a given cancer takes decades to emerge. Myeloproliferative neoplasms (MPN) are blood cancers driven by somatic driver mutations in haematopoietic stem cells (HSCs) that result in increased mature myeloid cell production[18], and provide an opportunity to capture early tumourigenesis and disease evolution that are otherwise inaccessible in other malignancies. Most of these patients harbour *JAK2*^{V617F}—this can either be the sole driver mutation or occur in combination with additional mutations—whereas others lack known causative driver mutations with uncertainty over the underlying clonal origin[19]. Here we undertake whole-genome sequencing (WGS) of individual single-cell derived haematopoietic colonies and targeted resequencing of longitudinal blood samples in patients with MPN to infer timing of driver mutations and tumour evolutionary dynamics in vivo.

Lineage tracing in MPN using somatic mutations

The mutations present in a somatic cell's genome have accumulated throughout its ancestral lineage, passed from mother to daughter cells. We identified somatic mutations in individual haematopoietic cells,

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [2]Wellcome–MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, Cambridge, UK. [3]Department of Haematology, University of Cambridge, Cambridge, UK. [4]Department of Haematology, Cambridge University Hospitals NHS Foundation Trust, Cambridge, UK. [5]These authors jointly supervised this work: Anthony R. Green, Peter J. Campbell, Jyoti Nangalia. ✉e-mail: jn5@sanger.ac.uk

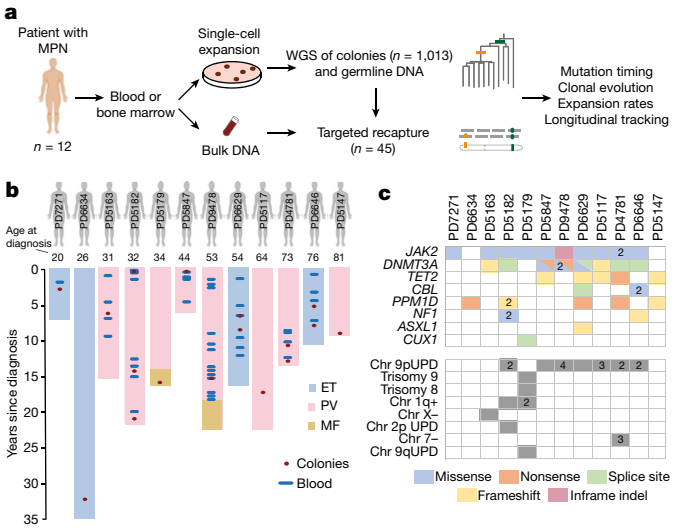


Fig. 1 | Patient cohort and experimental design. a, Experimental design. WGS, whole-genome sequencing. **b**, Patient cohort showing age at diagnosis, disease phase and duration, sample types and time points. ET, essential thrombocythemia; PV, polycythemia vera; MF, myelofibrosis. The length of the shaded bars represents the duration of disease, to last follow-up or to death. **c**, Driver mutations and copy number aberrations identified in at least one colony within each patient are shown. Shaded colours represent the type of mutation and the numbers within the squares represent the number of different events. Chr, chromosome; UPD, uniparental disomy.

using them to reconstruct lineage relationships among both malignant and normal blood cells in patients with MPN[20]. As somatic mutation burden does not differ significantly between HSCs and myeloid progenitors[20–23], we undertook WGS of in vitro expanded single-cell-derived haematopoietic colonies as faithful surrogates for the genomes of their parental HSCs. We 'recaptured' somatic mutations in bulk peripheral blood cells using targeted sequencing to longitudinally track clones (Fig. 1a). We sequenced 1,088 colonies to approximately 16.7× mean depth across 17 time points from 12 patients aged 20–81 years with different subtypes of MPN (Fig. 1b, Extended Data Fig. 1a). The majority of colonies were clonal (Extended Data Fig. 1b). Following filtering for low sequencing coverage and cross-colony contamination, 1,013 colonies were used in subsequent analyses (Supplementary Note 3). We identified 560,978 single nucleotide variants (SNV) and 19,155 small insertions and deletions (Supplementary Note 1). Ten out of 12 patients harboured mutated *JAK2*, and additional driver mutations were commonly observed in *DNMT3A* ($n = 12$), *PPM1D* ($n = 6$) and *TET2* ($n = 5$) (Fig. 1c).

We reconstructed phylogenetic trees using the presence or absence of SNV across colonies for individual patients (Fig. 2–4, Supplementary Note 2). A branch antecedent to more than one colony represents mutations shared by downstream colonies, and end branches comprise mutations in single colonies. A branch split in the tree ('coalescence') reflects an ancestral HSC symmetrically dividing into two daughter HSCs, the descendants of which have produced sampled blood cells. Each individual has a unique branching shape and phylogenetic tree structure, but many common themes emerge.

In patients with multiple driver mutations, additional driver mutations occurred both before and following *JAK2*^{V617F} as well as in independent HSCs, as previously reported[24–26]. All patients had mixtures of colonies with and without known driver mutations, suggesting that driver-mutation-bearing HSCs co-exist alongside normal blood production in patients with MPN. The colonies without driver mutations shared few mutations with one another—these were evident as long, isolated branches—demonstrating that residual non-malignant blood production remains highly polyclonal, as in healthy individuals[20]. Colonies with driver mutations typically shared tens to hundreds of mutations, including the driver mutation—this is evident as a 'clade', namely a set of lineages descending from a shared ancestral branch and confirms their single cell origin. Immediately beneath the shared branch containing the driver, we observe many short branches—this represents a 'clonal burst' in which the original mutated HSC expands to a population of cells. Coalescences are more frequent at the top of the tree and at the start of a clonal burst—their disappearance further down reflects the expanded size of the HSC population relative to the lineages sampled (Supplementary Note 4).

We frequently observed similar genetic changes occurring in different HSCs in the same patient—that is, 'parallel evolution'. Within individual patients, we observed (1) multiple acquisitions of chromosome (chr) 9p loss-of-heterozygosity (Figs. 2, 3, Extended Data Fig. 2), each with unique breakpoints as reported previously[27]; (2) independent acquisitions of chr1q^{+}, chr9q^{-} (PD5179) and *JAK2*^{V617F} (PD4781), each affecting different parental chromosomes (Extended Data Fig. 2); and (3) multiple different mutations affecting the same oncogene (Figs. 3, 4). This suggests that patient-specific factors shape selection on driver mutations and evolutionary trajectories in MPNs. In two patients lacking causative driver mutations in *JAK2*, *CALR* or *MPL*, we observed dominant clonal expansions driven by *PPM1D* and *TET2*, genes that are typically mutated in clonal haematopoiesis (Fig. 4a, b). *PPM1D* mutations have been reported in *JAK2*^{V617F}-mutated MPN[19] and following chemotherapy[28], but there was no known chemotherapy exposure prior to MPN therapy in either patient. We also observed clonal expansions without driver mutations in elderly patients (PD6646; Fig. 3), in keeping with our recent observations of haematopoietic oligoclonality in elderly individuals[23].

Mutation burden and telomere length

Point mutations accumulated linearly with age, with some heterogeneity both within and across patients (Fig. 4c). Wild-type cells were estimated to accrue 17.0 substitutions per year (95% confidence interval 14.8–19.2, mixed effects modelling; Supplementary Note 5), consistent with other studies[20–22], suggesting that mutation accumulation in wild-type blood cells in patients with MPN is similar to that in healthy individuals.

For some patients, mutation burden in mutant clades appeared to be increased compared with other colonies (Fig. 3). Using a Bayesian approach to model rates of mutation acquisition on a per patient and per clade basis, the mutation burden in mutant colonies appeared modestly higher than wild-type counterparts for several patients, but differences were less significant for individual patients when using non-parametric statistical methods (Extended Data Fig. 3a, b, Supplementary Note 5). However, at a cohort level, there was a significant increase in C>T transitions at CpG dinucleotides in *JAK2*-mutated clades compared with wild-type colonies (Extended Data Fig. 4a, b, Supplementary Note 6) raising the possibility that modest differences in mutation burden reflected differences in cell division history between mutant and wild-type cells[5]. In keeping with this, telomere lengths were significantly shorter in *JAK2*-mutated clades ($P = 0.002$; Fig. 4d), a relationship that was retained after correcting for their greater shared ancestry (−829 bp, confidence interval 662–968 bp, $P < 0.001$, Extended Data Fig. 4c). Telomere lengths followed the phylogenetic tree, that is, more closely related colonies had more similar telomere lengths (Extended Data Fig. 4d). The degree of telomere loss correlated with the size of *JAK2*-mutant clades and the inferred number of additional HSC cell divisions for clonal expansion, with telomeres estimated to shorten by approximately 57 bp per division (Extended Data Fig. 4e, Supplementary Note 7), similar to a previous report[29], although it is possible that telomere shortening also occurs during *JAK2*-mutant progenitor cell expansion[30]. Overall, our data suggest that telomere

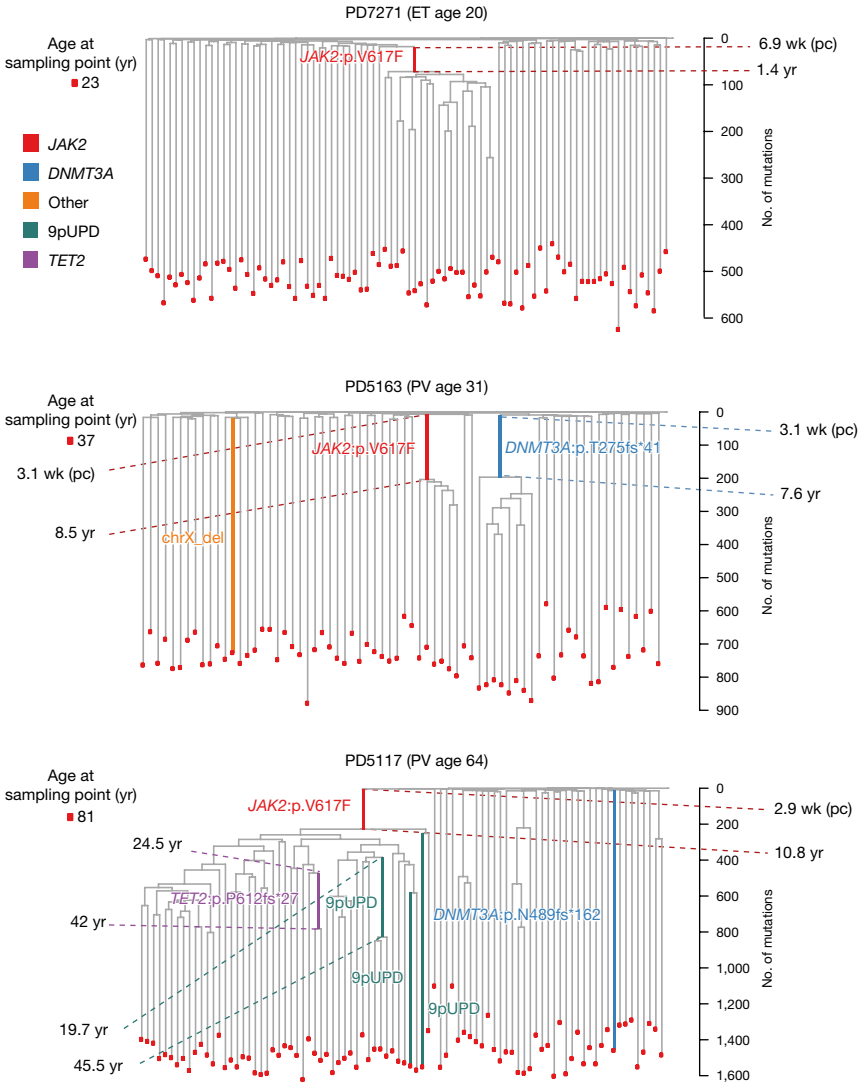


Fig. 2 | Phylogenetic histories of three patients with MPN driven by *JAK2*^{V617F}. The phylogenetic trees for three patients with stable *JAK2*^{V617F}-mutated MPN. PD7271 (ET) presented at age 20 with asymptomatic isolated thrombocytosis and received aspirin. PD5163 (PV) presented at age 31 with splanchnic vein thrombosis, a raised red cell mass and received interferon-α. PD5117 (PV) presented at age 64 with elevated blood counts and red cell mass and received hydroxycarbamide. The tips of the branches represent individual colonies (red dots). Shared branches represent mutations present in all downstream descendant colonies, and end branches represent mutations unique to single colonies. Branch lengths are proportional to mutation counts shown on the vertical axes. Branches containing different driver mutations and chromosomal aberrations are highlighted by colour. The inferred patient ages (yr, years post birth; wk (pc), weeks post conception) at the start and end of the shared branches harbouring driver mutations are depicted.

lengths, much more than mutation burden, reflect cell division differences between wild-type and mutant clades in patients with MPN, and are in keeping with the notion that the number of mutations that accumulate during haematopoietic cell division is very low, with mutation burden more determined by time[20,22,31,32].

Timing of driver mutation acquisition

Given the linear accumulation of somatic mutations with age, we inferred the time point in life when driver mutations in phylogenetic trees had occurred. Branches at the top of a tree comprise mutations acquired at a young age, with branches lower down representing mutations arising later in life. The traditional view of MPN might have predicted that *JAK2*^{V617F} occurs within a wild-type cell a few years before disease presentation, resulting in rapid clonal expansion. On a phylogenetic tree, this would manifest as a long vertical branch harbouring a single driver mutation (for example, *JAK2*) followed by a clonal burst with short emergent branches appearing low down in the tree. We did not observe any such instances in any of the patients. Instead, key observations can be made from qualitative assessment of the trees. First, despite the different ages of the patients, the earliest driver mutations are acquired within 200 mutations of the start of life, corresponding to the first decade of childhood. There are many such examples involving *JAK2* (patient IDs PD7271, PD5163, PD5117, PD5182 and PD9478), *DNMT3A* (PD5163, PD5182, PD5847 and PD6629) and *PPM1D* (PD6634) (Figs. 2–4). Driver mutations are observed on short branches near the very top of the trees, such as *JAK2*^{V617F} in PD5182 and *DNMT3A* mutations in PD5182 and PD5847 (Fig. 3). These earliest branches represent the first few dozen mutations of life, and correspond to the in utero period, given that blood cells consistently harbour around 50–55 somatic mutations at birth[21,23]. Secondly, whereas some patients have not acquired additional driver mutations in expanded clades (for example, PD7271, PD5163 and PD6634), the majority of patients accrue further driver mutations, each emerging several hundred mutations after the previous event (Fig. 3), suggesting that clonal evolution and sequential driver mutation acquisition is separated across decades over a lifetime.

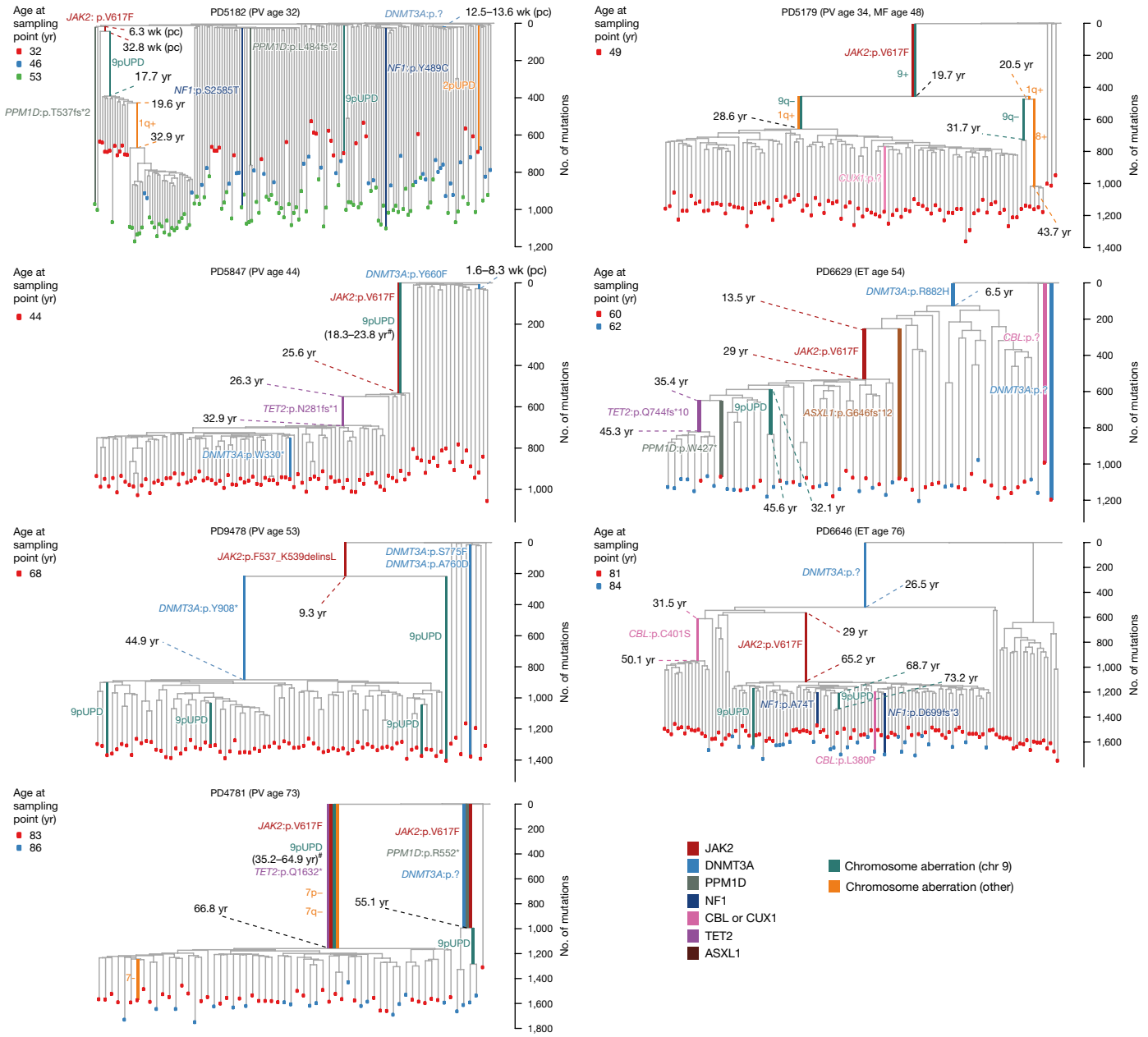


Fig. 3 | Phylogenetic histories of 7 patients with *JAK2*^{V617F}-mutated MPN and clonal evolution. The phylogenetic trees of seven patients with MPN who have evidence of multiple driver mutation-led expansions. The vertical axis shows mutation counts. The tips of the branches represent individual colonies and are coloured by the time point of sampling. The timing of 9pUPD events in PD4781 and PD5847 are calculated using the proportion of heterozygous versus homozygous mutations on the UPD regions indicated by the hash symbol (#). The two colonies in PD5182 that harbour 9pUPD and 2pUPD did not harbour detectable driver mutations in the copy number aberrant regions.

We scaled the phylogenetic trees to chronological time using patient-specific and clade-specific mutation rates, and took into account the accelerated rate of mutation acquisition reported in early life due to more rapid growth[21,22,31]. This provided an age estimate for the start and end of each branch containing a driver mutation together with confidence intervals, and thus a time interval during which the genetic event plausibly occurred, with the age at the end of the branch providing an upper bound on acquisition timing (Extended Data Fig. 5a, b). Of note, no direct measurements for driver mutations were made experimentally at the inferred times of acquisition.

Time-based phylogenetic trees delineated qualitative observations. In PD5182 and PD7271, *JAK2*^{V617F} acquisition was estimated to have occurred by 32.8 weeks post conception (pc) (95% confidence interval for upper age estimate 11 weeks pc–1.1 year) and 1.4 years (0.6–2.3 years) of age, respectively. Three patients acquired mutated-*JAK2* during childhood by 8.5 years (7.2–10.0 years, PD5163), 9.3 years (7.9–10.9 years, PD9478) and 10.8 years (9.2–12.5 years, PD5117) of age (Extended Data Fig. 5c). In these 5 patients in whom mutated *JAK2* was the first driver event within the MPN clone, the mean latency to diagnosis was 34 years (range 19–54 years). *JAK2*^{V617F} occurred as the second driver event within a mutated-*DNMT3A* clade in two patients, with diagnosis occurring 11 years (PD6646) and 25 years (PD6629) after *JAK2*^{V617F} acquisition. In the remaining three patients with mutated *JAK2* (PD4781, PD5847 and PD5179), we were unable to precisely time *JAK2*^{V617F} owing to the presence of additional driver events (for example, 9pUPD) on the same branch. However, timing estimates of 9pUPD acquisition often fell to decades before disease presentation, implying that *JAK2*^{V617F} acquisition occurred even earlier than this. Mutations in *DNMT3A*, the gene most commonly

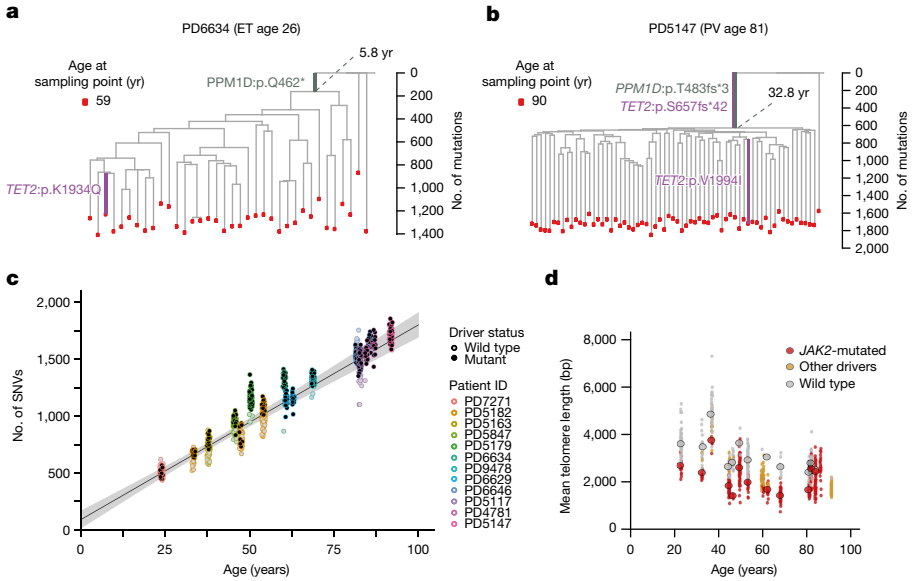


Fig. 4 | Phylogenetic trees, mutation burdens and telomere lengths. **a**, **b**, Phylogenetic trees for PD6634 (**a**) and PD5147 (**b**), patients with MPN lacking phenotypic driver mutations in *JAK2*, *CALR* or *MPL*. **c**, Total SNV and relationship with age. Dots represent single colonies analysed by WGS and coloured outlines represent individual patients. Total SNVs represent somatic SNV burden adjusted for depth of sequencing and clonality of the sample. Black filled dots represent mutant colonies and grey filled dots represent wild type. The black line shows the regression line and grey shading shows the 95% confidence interval. **d**, Relationship between mean telomere length and age for wild type (grey dots), *JAK2*-mutated colonies (red dots) and other mutant colonies (yellow dots).

detected later in life in clonal haematopoiesis[13–15], were also estimated to have occurred by the in utero period or childhood in several patients. *DNMT3A*$^{ess.splice}$ in PD5182 localized to the 20th or 21st point mutation of life in that lineage, corresponding to acquisition between 12 and 14 weeks pc. In PD5847, *DNMT3A*^{Y660F} occurred by the 22nd mutation in the lineage, corresponding to acquisition between 2 and 9 weeks pc. The canonical mutation *DNMT3A*^{R882H} was estimated to have occurred by 6.5 years (5.2–8.1 years) of age in PD6629, by 7.6 years (6.3–9.2 years) of age for *DNMT3A*^{T275fs*41} in PD5163, and mutated *PPM1D* was acquired by 5.8 years (4.6–7.2 years) of age in PD6634 (Extended Data Fig. 5c).

Branches harbouring more than one driver mutation were often observed, particularly in older patients (PD4781 and PD5147) and for *JAK2*^{V617F}/chr9 aberrations (PD5847 and PD5179). Such branches could reflect the out-competition of the ancestral mutant clade(s) harbouring the initial driver mutation(s) by subsequent driver events, or that the ancestral clade was not sampled. We assessed the allele fractions in bulk blood for somatic mutations present on such branches for the presence of more than one clone but could not find evidence that ancestral clades were present at a substantial fraction in blood, suggesting extinction of earlier mutant clones.

Clonal expansion dynamics in patients

The pattern of branching in mutant clades reflect their historical rates of clonal expansion. A clonal burst comprising long emergent branches downstream of a branch harbouring a driver mutation implies slow expansion. Examples include *DNMT3A*- and *PPM1D*-mutated clades in PD6629 and PD6634, respectively (Figs. 3, 4), and the *JAK2*^{V617F} clade in PD5117, a patient who presented with asymptomatic MPN more than 50 years after *JAK2*^{V617F} acquisition. By contrast, a clonal burst comprising multiple short emergent branches suggests more rapid historical expansion—this is seen for PD7271, a young patient with MPN, in whom branching downstream of *JAK2*^{V617F} occurs over approximately 200 mutations, and is especially evident for clones harbouring multiple driver mutations (for example, the *JAK2*/9pUPD/*TET2*-mutated clade in PD5847) (Figs. 3, 4). Large and rapid clonal expansions are often observed to outcompete ancestral mutant clones (for example, multiply mutated clades in PD4781 and PD5147). Such inter-clonal dynamics are highlighted by phylogenetic trees comprising multiple colony time points. For example, in PD5182, colonies were grown at age 32 (diagnosis), 46 and 53 years. By diagnosis (Fig. 3, PD5182 red dots), there is evidence of a clonal expansion harbouring both *JAK2*^{V617F} and 9pUPD that has outcompeted the ancestral heterozygous *JAK2*^{V617F} clone acquired very early in life. By the next time point, the patient was receiving interferon therapy, and a smaller fraction of mutant colonies are captured, which now harbour additional Chr1q+ (blue dots). By age 53, at which time the patient is refractory to interferon, the mutant clonal fraction is larger, and almost entirely comprised of the clonally evolved *JAK2*^{V617F} homozygous/1q+ clade (green dots), demonstrating how sequential driver mutation acquisition decades apart can lead to successive clonal sweeps during MPN disease course. Similarly, in PD6646, sampling a few years later captures a larger fraction of the *DNMT3A*/*JAK2*-mutated clade, raising the possibility that it may be in the process of outgrowing other clades in this individual.

Modelling the fitness of mutant clones

We simulated a forward time continuous birth–death process, in which the introduction of a driver mutation increases the symmetrical HSC division rate by a selection coefficient (s) with corresponding growth rate per year (S) (Methods) and used approximate Bayesian computation (ABC) to estimate S for clades with five or more downstream lineages (Extended Data Fig. 6a). Our approach provided orthogonal estimates for absolute driver mutation timing, confirming estimates from branch length intervals (Extended Data Fig. 6b). Estimates of S correlated well with an alternative method that we developed ('Phylofit' versus ABC correlation coefficient $r = 0.96$; Extended Data Fig. 6c). Our modelling provided an average historical growth rate for clades and showed that single-mutant *JAK2*^{V617F} clades expanded at different rates in different patients (Fig. 5a)—73% yr^{-1} (confidence interval 43–108% yr^{-1}) in PD7271, the youngest patient with MPN in our cohort, and 18% yr^{-1} in PD5117, who was diagnosed 54 years after *JAK2*^{V617F} acquisition. Average

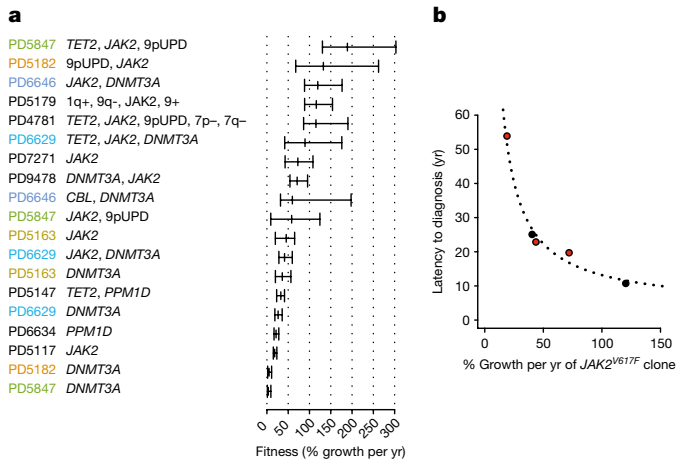


Fig. 5 | Clonal fitness and latency to diagnosis. a, The inferred fitness of clones, as measured by *S*, the growth rate per year, with 95% confidence intervals (CI). *S* is highest for multiply mutated clades (up to 189% yr^{-1}), and lowest for driver mutations common in clonal haematopoiesis (as low as 3% yr^{-1}). Coloured patient IDs are those harbouring several different clones. **b**, The latency to diagnosis in relation to *S* following acquisition of mutated *JAK2* is shown for five individuals (PD7271, PD5163, PD5117, PD6646 and PD6629). Red dots represent patients with mutated *JAK2* as the only driver mutation. Black dots represent *JAK2* mutation acquisition after mutated *DNMT3A*. The latency to diagnosis fits the model latency = $a/(\log(1+S)^b)$, where *a* is the log of the number of mutant cells at diagnosis and $b = 1$.

historical rates of clonal expansion following the acquisition of $JAK2^{V617F}$ fitted well with the latency between mutation acquisition and disease diagnosis (Fig. 5b). In some patients, successive increases in expansion rates occurred with sequential driver mutation acquisition (PD6629 $DNMT3A^{R882H}$, 26% yr^{-1} (confidence interval 19–36% yr^{-1}); $DNMT3A^{R882H}$/$JAK2^{V617F}$ 41% yr^{-1} (confidence interval 28–60% yr^{-1}); $DNMT3A^{R882H}$/$JAK2^{V617F}$/$TET2^{Q744fs*10}$ 90% yr^{-1} (confidence interval 42–176% yr^{-1})). The most rapidly growing clades in the cohort all harboured multiple driver mutations, with growth rates up to 189% yr^{-1} (confidence interval 130–303% yr^{-1}, PD5847), translating to the clone doubling in size every 8 months. Lowest selection estimates were observed for two in utero-acquired *DNMT3A* mutated clones (PD5847 and PD5182, 3–5% yr^{-1}). Given that one would predict a higher probability of stochastic extinction of HSCs harbouring driver mutations with low selection coefficients (Supplementary Note 8), it is plausible that such clones associated with weak fitness advantages survived to expand only by hitch-hiking on the rapid population growth occurring in utero. Indeed, with such low growth rates, clones would take several decades to even reach low detectable clonal fractions in blood.

Somatic mutations from the phylogenetic trees were deep sequenced using targeted recapture in bulk mature blood cells from the same patients. We observed that the clonal fractions of clades from phylogenetic trees were broadly concordant with bulk population fractions and inferred clonal trajectories from ABC (Extended Data Fig. 7a). We did not find evidence of lineages missing from the trees, suggesting that any in vitro culture selection bias was not significant (Extended Data Fig. 7b). Our model of HSC population dynamics closely recapitulated the pattern of branching in individual clades, however, there are some assumptions. First, the modelling assumes that the HSC population size remains constant over life and following MPN development. Secondly, we assume that selection is constant during the period of time captured by each clonal burst. We searched for clades in which there was a rapid clonal burst early in the tree (as suggested by short emergent branches) but yet only a small final clonal fraction, suggesting that *S* may have been higher initially and decreased later in life. We found three such clades (Extended Data Fig. 7c), one associated with treatment with interferon (PD5163 $JAK2^{V617F}$ clade), and two *DNMT3A*-mutated in utero clades, suggesting more rapid growth earlier in life. Given the multiple selective pressures, including treatment, bone marrow environment and inter-clonal competition, our estimates of fitness should be viewed as average historical expansion rates of clones under a simple population model and more complex dynamics may exist in vivo.

Discussion

MPN are a common chronic blood cancer prevalent in up to 1 in every 1,000–2,000 individuals, with an increasing incidence with age[33,34]. Our results suggest that MPN driver mutations, such as $JAK2^{V617F}$, as well as *DNMT3A* and *PPM1D* mutations commonly associated with clonal haematopoiesis[13–15], often occur during childhood, including in utero. Driver mutations in all single-driver mediated clonal expansions were robustly estimated to have occurred in the first half of life in the patients, making this probabilistically likely to be a general feature of MPN. Our data are consistent with $JAK2^{V617F}$ detection and mutation timing estimation before MPN diagnosis[35,36] and in cord blood[37], and with the detection of low-burden clonal haematopoiesis in younger adults[16,38]. Inferred rates of clonal expansion following $JAK2^{V617F}$ were variable across patients, raising the possibility that additional factors influence the consequences of $JAK2^{V617F}$, such as cytokine homeostasis, bone marrow microenvironment, inflammation or germline influences[39–43], which may dynamically contribute over the lifetime of an individual. Factors influential during embryogenesis may also shape the future trajectories of nascent clones. Overall, growth rate estimates for $JAK2^{V617F}$ in patients with MPN were greater than that reported for $JAK2^{V617F}$-mediated clonal haematopoiesis—which may account for the relative lack of clinical manifestation in the latter group. Growth estimates for *DNMT3A*-mutated clones were consistent with that reported in clonal haematopoiesis[44].

Clinically, MPN diagnosis is defined phenotypically by blood count parameters and bone marrow histomorphology[45]. Our results indicate that current clinical diagnosis is made at a late time point in the expansion and evolution of phenotype-driving clones. This latency potentially encompasses a disease phase during which no phenotype has occurred, as low to moderately fit clones with driver mutations take decades to reach detectable levels that may be required for clinical manifestations. However, earlier detection may also be opportune to prevent life threatening thromboses that often trigger current diagnosis, and to better manage individuals in the general population harbouring low levels of $JAK2^{V617F}$ who have increased risk of thrombosis[46]. The cornerstone of MPN management is aimed at reducing the risk of vascular events; such treatments are mostly safe and well-tolerated and could be offered to individuals with high-risk molecular profiles if clinical trials supported their use. Given the lifelong trajectories to MPN, our data also provide a strong rationale for the evaluation of measures targeting the mutant clone[47,48] in order to curb clonal expansion and evolution. This model for blood cancer development may also be relevant to other cancers and organs, given the abundance of mutations under selection in histologically normal tissues[7,9–11,13–16].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-04312-6.

1. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
2. Gerstung, M. et al. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
3. Mitchell, T. J. et al. Timing the landmark events in the evolution of clear cell renal cell cancer: TRACERx Renal. *Cell* **173**, 611–623.e17 (2018).

4. Welch, J. S. et al. The origin and evolution of mutations in acute myeloid leukemia. *Cell* **150**, 264–278 (2012).
5. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet.* **47**, 1402–1407 (2015).
6. Radivoyevitch, T., Hlatky, L., Landaw, J. & Sachs, R. K. Quantitative modeling of chronic myeloid leukemia: insights from radiobiology. *Blood* **119**, 4363–4371 (2012).
7. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
8. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
9. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
10. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
11. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
12. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
13. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
14. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
15. Xie, M. et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat. Med.* **20**, 1472–1478 (2014).
16. Young, A. L., Challen, G. A., Birmann, B. M. & Druley, T. E. Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat. Commun.* **7**, 12484 (2016).
17. Abelson, S. et al. Prediction of acute myeloid leukaemia risk in healthy individuals. *Nature* **559**, 400–404 (2018).
18. Vainchenker, W. & Kralovics, R. Genetic basis and molecular pathophysiology of classical myeloproliferative neoplasms. *Blood* **129**, 667–679 (2017).
19. Grinfeld, J. et al. Classification and personalized prognosis in myeloproliferative neoplasms. *N. Engl. J. Med.* **379**, 1416–1430 (2018).
20. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
21. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316.e4 (2018).
22. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).
23. Mitchell, E. et al. Clonal dynamics of haematopoiesis across the human lifespan. Preprint at https://doi.org/10.1101/2021.08.16.456475 (2021).
24. Nangalia, J. et al. DNMT3A mutations occur early or late in patients with myeloproliferative neoplasms and mutation order influences phenotype. *Haematologica* **100**, 438–442 (2015).
25. Ortmann, C. A. et al. Effect of mutation order on myeloproliferative neoplasms. *N. Engl. J. Med.* **372**, 601–612 (2015).
26. Lundberg, P. et al. Clonal evolution and clinical correlates of somatic mutations in myeloproliferative neoplasms. *Blood* **123**, 2220–2228 (2014).
27. Godfrey, A. L. et al. JAK2V617F homozygosity arises commonly and recurrently in PV and ET, but PV is characterized by expansion of a dominant homozygous subclone. *Blood* **120**, 2704–2707 (2012).
28. Kahn, J. D. et al. PPM1D-truncating mutations confer resistance to chemotherapy and sensitivity to PPM1D inhibition in hematopoietic cells. *Blood* **132**, 1095–1105 (2018).
29. Vaziri, H. et al. Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. *Proc. Natl Acad. Sci. USA* **91**, 9857–9860 (1994).
30. Anand, S. et al. Effects of the JAK2 mutation on the hematopoietic stem and progenitor compartment in human myeloproliferative neoplasms. *Blood* **118**, 177–181 (2011).
31. Chapman, M. S. et al. Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
32. de Kanter, J. K. et al. Antiviral treatment causes a unique mutational signature in cancers of transplantation recipients. *Cell Stem Cell* **28**, 1726–1739 (2021).
33. Titmarsh, G. J. et al. How common are myeloproliferative neoplasms? A systematic review and meta-analysis. *Am. J. Hematol.* **89**, 581–587 (2014).
34. Mehta, J., Wang, H., Iqbal, S. U. & Mesa, R. Epidemiology of myeloproliferative neoplasms in the United States. *Leuk. Lymphoma* **55**, 595–600 (2014).
35. Van Egeren, D. et al. Reconstructing the lineage histories and differentiation trajectories of individual cancer cells in myeloproliferative neoplasms. *Cell Stem Cell* **28**, 514–523.e9 (2021).
36. McKerrell, T. et al. JAK2 V617F hematopoietic clones are present several years prior to MPN diagnosis and follow different expansion kinetics. *Blood Adv.* **1**, 968–971 (2017).
37. Hirsch, P. et al. Clonal history of a cord blood donor cell leukemia with prenatal somatic JAK2 V617F mutation. *Leukemia* **30**, 1756–1759 (2016).
38. Wong, W. H. et al. Engraftment of rare, pathogenic donor hematopoietic mutations in unrelated hematopoietic stem cell transplantation. *Sci. Transl. Med.* **12**, eaax6249 (2020).
39. Olcaydu, D. et al. A common JAK2 haplotype confers susceptibility to myeloproliferative neoplasms. *Nat. Genet.* **41**, 450–454 (2009).
40. Hinds, D. A. et al. Germ line variants predispose to both JAK2 V617F clonal hematopoiesis and myeloproliferative neoplasms. *Blood* **128**, 1121–1128 (2016).
41. Fleischman, A. G. Inflammation as a driver of clonal evolution in myeloproliferative neoplasm. *Mediators Inflamm.* **2015**, 606819 (2015).
42. Bick, A. G. et al. Inherited causes of clonal haematopoiesis in 97,691 whole genomes. *Nature* **586**, 763–768 (2020).
43. Bao, E. L. et al. Inherited myeloproliferative neoplasm risk affects haematopoietic stem cells. *Nature* **586**, 769–775 (2020).
44. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
45. Arber, D. A. et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405 (2016).
46. Nielsen, C., Birgens, H. S., Nordestgaard, B. G. & Bojesen, S. E. Diagnostic value of *JAK2* V617F somatic mutation for myeloproliferative cancer in 49 488 individuals from the general population. *Br. J. Haematol.* **160**, 70–79 (2013).
47. Kiladjian, J. J. et al. Pegylated interferon-α-2a induces complete hematologic and molecular responses with low toxicity in polycythemia vera. *Blood* **112**, 3065–3072 (2008).
48. Pieri, L. et al. *JAK2*V617F complete molecular remission in polycythemia vera/essential thrombocythemia patients treated with ruxolitinib. *Blood* **125**, 3352–3353 (2015).

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Patients and samples

Peripheral blood and bone marrow samples were obtained from patients with myeloproliferative neoplasms attending Cambridge Universities NHS Trust following written informed consent and ethics committee approval. Patients were selected from a cohort that had undergone previous whole exome sequencing of their blood[49]. Apart from ensuring a wide representation of ages, patients were chosen at random and captured a broad range of MPN. Recapture samples obtained in 10 of 12 patients were predominantly peripheral blood derived granulocytes. Constitutional samples were obtained from either buccal epithelium or T cells.

In vitro colony culture for whole genome sequencing and targeted recapture of bulk DNA samples

Peripheral blood mononuclear cells were isolated and cultured for 14 days in MethoCult 4034 (Stemcell) and single erythroid haematopoietic colonies (burst forming unit-erythroid, BFU-E) were plucked and lysed in individual 50 µl aliquots of RLT lysis buffer (Qiagen). Ten to twenty microlitres of this lysed sample was used for library preparation for whole genome sequencing. Due to the low DNA input, the NEBNext Ultra II low input kit (NEB) with enzymatic fragmentation and 8 cycles of PCR was used[50]. Sequencing was 150 bp paired end on either Illumina HiSeqX or Novaseq machines. Reads were aligned to the human reference genome (NCBI build37) using BWA-MEM. For targeted re-sequencing of bulk peripheral blood (recapture) samples, we used Agilent SureDesign to design a baitset that captured all unmasked shared mutations across colonies. For private branches, we assayed up to 4 randomly selected unmasked variants per year of approximate branch length, capping the total number per private branch variants at 80 mutations. Variants that passed the SureDesign 'most stringent' filter were preferentially selected. Sequencing on Illumina Novaseq was undertaken to a depth of 300–400× for recapture samples.

Somatic mutation identification and filtering

Single nucleotide variants (SNV) were identified using CaVEMan[51] for each colony by comparison to an unmatched normal colorectal crypt sample (PD26636b) that had previously undergone whole genome sequencing. CaVEMan was run with the 'normal contamination of tumour' parameter set to zero, and the tumour/normal copy numbers set to 5/2. In addition to standard filters[3], reads supporting an SNV had to have a median BWA-MEM alignment Score ≥ 140 and less than half of the reads clipped. Filtering designed for quality control following processing through the low-input sequencing pipeline, as recently described[10,50], were also applied. The use of the unmatched normal meant that this process called both somatic and germline SNVs. The removal of germline SNVs and artefacts of sequencing required further filtering, and we used pooled information across per-patient colonies and read counts from a matched germline WGS sample, either buccal or T cell DNA, to ensure that genuine somatic variants that may have been present in the germline sample, either as embryonic variants or due to tumour-in-normal contamination were also identified. (Supplementary Note 1). Short Insertions and deletions were called using cgpPindel[52] with the standard cgpPindel VCF filters applied, except the F018 Pindel filter was disabled as it excludes loci of depth <10. Copy-number aberrations (CNA) were identified using ASCAT[53] with comparison to a matched normal sample. Given the clonal nature of samples identification of sub-clonal copy number changes was not required. The union of colony SNVs and insertion–deletions (indels) was then taken and reads counted across all samples belonging to the patient (colonies, recapture samples, buccal and T cells) using VAFCorrect. VAFCorrect avoids double counting of overlapping reads and minimises reference bias in the variant allele fractions (VAFs) of indels by performing a pileup mapping to both the reference genome and a locally altered reference genome that incorporates the alternative allele in the reference sequence.

Creating a genotype matrix

The genotype at each locus within each sample was either 1 (present), 0 (absent) or NA (unknown). We inferred the genotype in a depth sensitive manner. We assumed the observed mutant read count for a colony at a given site was MTR ~ Binomial(n = Depth, P = VAF), if the site was mutant, and MTR ~ Binomial(n =depth, P = 0.01), if the site was wild type. The genotype was set to the most likely of the two possible states provided one of the states was at least 20 times more likely than the other. Otherwise the genotype is set to missing (NA). The VAF was usually 0.5 for autosomal sites, but for chromosomes X, Y and CNA sites, it was conservatively set to 1/ploidy. For loss-of-heterozygosity (LOH) sites, the genotype was overridden and set to missing if it was originally 0.

Phylogenetic tree topology

We constructed phylogenetic tree topologies using maximum parsimony with MPBoot[54]. The inputs for MPBoot were the binary genotype matrices with missing values per patient. This method also enables the rapid generation of bootstrap trees that correspond to the maximum parsimony trees that would be obtained by resampling the per site genotypes with replacement. Only SNVs were used to infer the topology, but both SNVs and indels were subsequently assigned to the branches of phylogenetic trees. Variants with zero genotype in genomic regions identified as carrying a chromosomal LOH or deletion were set to missing for MPBoot.

Assignment of mutations to branches and branch length estimation

We developed an expectation maximisation method (R package treemut) to soft assign mutations to trees and to estimate branch length (https://github.com/NickWilliamsSanger/treemut). The starting point for the method is Bayes rule: P(mutation has genotype) ∝ P(read count | genotype). P(genotype). Each variant has a corresponding true genotype vector where the elements are 1 for colonies that descend from a particular branch and 0 for colonies that do not. The prior probability of a mutation having a genotype, P(genotype), is therefore proportional to branch length. The read counts are modelled using a binomial distribution with per sample specific error rates and an assumed VAF = 1/ploidy for variant sites and VAF = 0 for wild type sites. The above process is iterated where each iteration updates the branch lengths (that is, P(genotype)). Having estimated the maximum likelihood probability that each mutation belongs to each branch we then hard assigned mutations. Simulations indicated that this approach did not exhibit obvious biases in branch length estimation vs true branch lengths and that using the edge length implied by the hard assignment had a minimal effect on the deviation from the true edge length (Supplementary Note 2, Supplementary Fig. 2). The sensitivity for detecting somatic variants would be expected to increase both with depth of sequencing as well as with the degree of clonality of a colony. Given that heterogeneous sensitivity of somatic variant identification across the colonies would inadvertently affect branch lengths, particularly private branches, all branch lengths were adjusted for the sensitivity of identified SNVs. We used an approach where fully clonal SNV detection sensitivity was directly estimated from the per colony sensitivity for identification of germline SNVs, in conjunction with a correction for the clonality of the colonies (Supplementary Note 2, Supplementary Figs. 3, 4). In addition to SNVs and indels, the patient colonies exhibited a variety of LOH and CNA events. The events were curated as being present or absent in each of the colonies giving an event genotype vector similar to that obtained

for SNVs and indels. Once the tree topology was inferred using the SNV genotypes, the branches that exactly matched the event genotype were identified and the event assigned to the corresponding branch. In the case where the event did not map to the tree topology then the event was further reviewed, and disentangled into repeated acquisitions of the same events and validated by distinctness of CNA breakpoints and haplotypes (Extended Data Fig. 2).

Phylogenetic tree quality assessment

Colonies were initially removed if their CaVEMan somatic mutation detection sensitivity was below 60%. In addition, colonies were tested for cross-contamination with other colonies from the same patient following phylogenetic tree construction. If a colony was a mixture of more than one colony then how this will manifest depended on whether the mixed colonies were from the same mutant clade or not. If the mixed colonies belonged to the same clade with a long shared branch, then the overall sample VAF would appear quite clonal, however the private branches would exhibit a lower than expected VAF. To check that a colony was sufficiently clonal we required that the mean VAF of variants from a colony that mapped to ancestral branches was not significantly less than 0.35 (Bonferroni-adjusted one-sided binomial test). Additionally, we expected the VAF to be zero in non-ancestral branches and so excluded colonies that had significantly more than 5% VAF on non-ancestral branches (Bonferroni-adjusted one-sided binomial test). The thresholds in the above test reflected the requirement that total contamination be less than 10%. For highly shared early branches, some true somatic variants may inadvertently be filtered out due to their presence as tumour-in-normal contamination in the matched germline DNA sample. We screened all phylogenetic trees to rescue any such variants (Supplementary Fig. 5). Detailed quality assessment of phylogenetic trees and colony filtering is provided in Supplementary Note 3 and a summary of the proportion of colonies that failed quality thresholds is shown in Supplementary Fig. 6 and Supplementary Table 1.

Mutation rate estimation and timing of branches harbouring driver mutations on phylogenetic trees

We created time-based ('ultrametric') trees, wherein the *y*-axis of the trees is converted from mutations to time. We jointly fitted wild type rates, mutant rates and absolute time branch lengths using a Bayesian per patient tree-based model under the assumption that the observed branch lengths are Poisson distributed with mean = duration × sensitivity × mutation rate. We also evaluated modelling observed branch lengths using a negative binomial distribution and results were similar. We consider a rooted tree where each edge i consists of an observed mutation count m_i and a true duration t_i. We refer to a given edge and its child node interchangeably by the same label. $D(i)$ is the set of terminal nodes (tips) that descend from node i and $A(i)$ is its corresponding set of ancestral nodes excluding the root. We assume that each tip of the tree k has a known corresponding time T_k (for example, the post conception age in years of the patient at sampling of the cell) and we therefore have the following constraint:

$$T_k = t_k + \sum^{i \in A(k)} t_i$$

and $T_k > t_i > 0$.

We incorporate this constraint by performing the optimisation over the interior branches of the tree with reparameterised branch durations x_i transformed to be in the range $0 > x_i > 1$, where x_i can be thought of as stick breaking fractions. If j is an edge whose parent node is the root then: $t_j = x_j \min(T_k : k \in D(j))$.

For other interior edges, i, we have

$$t_i = \left(\min\{T_k : k \in D(i)\} - \sum_{j \in A(i)} t_j\right) x_i$$

The duration of the terminal edges is fixed by the duration of the ancestral interior edges and the overall duration constraint:

$$t_i = T_i - \sum^{j \in A(i)} t_j$$

We assume that there are $p-1$ change points in the tree corresponding to the acquisition of driver mutations. This results in p mutation rates λ_j applying throughout the tree where we allow at most one change point per branch and the initial ancestral (or wild-type) rate is λ_0 and additional rate change points occur a fraction α_j along branch j and descendent branches have the rate λ_j unless there are additional change points in descendant branches. The effective rate on branches with a change point going from λ_l to λ_j is just the weighted average $\alpha_j\lambda_l + (1-\alpha_j)\lambda_j$ where we use a uniform unit interval prior for the α values.

We assume the underlying mutation process follows a Poisson distribution with the above piecewise constant driver specific mutation rates, the number of observed mutations accrued on branch i in time t_i measured in years: $m_i \sim \text{Poisson}(\lambda t_i S_i)$ where $S_i \sim \text{Beta}\left(\alpha = c, \beta = c\frac{1-s_i}{s_i}\right)$ where we have chosen the parameter $c = 100$. This reflects only modest uncertainty in our estimates in sensitivity and also allows the model to mitigate larger than expected variability in the branch lengths. In addition, $\lambda \sim \mathcal{N}(\widehat{\lambda}, 0.25\widehat{\lambda})$ where $\widehat{\lambda}$ is the naive estimation of a single rate λ as the per patient median of the ratio of the root to tip mutation count and the tip sampling age, and finally we use the weakly informative prior for the stick breaking fractions:

$$x_i \sim \text{Beta}\left(\alpha = 1 - \frac{1}{1 - \sum^{j \in A(i)} p_j}, \beta = 1\right)$$

where the p_i is an initial approximation of the duration of the branch length expressed as a fraction of the sampling time:

$$p_i = \min_{j \in D(i)}\left\{\frac{m_j + 1}{\sum^{k \in A(j} m_k + 1}\right\}$$

Copy number regions were masked for the above analysis and the final rates rescaled by the reciprocal of assayed proportion of the autosomal genome (as determined by the background local mutation rate model – see below). The above model will overestimate the duration of early branches because it does not take into account that mutation burden during development is higher[31]. Therefore, we used information from two studies[21,31] to add a time-dependent mutation excess rate into the model and fitted a logistic function to the mutation burden at 3 different ages given across the two studies:

$$\lambda(t) = \frac{149.2}{1 + e^{50.0(t - 0.225)}}$$

This initially maintains an excess instantaneous mutation rate of approximately 150 yr^{-1} before rapidly dropping to zero between 2 and 4 months and integrating to an excess of 33.5 mutations by 6 months – when added to the steady state rate of 18 mutations per year this implies a mutation burden at birth of around 50 mutations. The above models were coded in R and Rstan and inferred using the Rstan implementation of Stan's No-U-Turn sampler variant of Hamiltonian Monte Carlo method[55]. Models were fitted across four chains each with 20,000 iterations including 10,000 burn-in iterations. The code is available as an R package 'Rtreefit' at https://github.com/NickWilliamsSanger/rtreefit. Mutation clades for which mutation rates were fitted are shown in Extended Data Fig. 5. The trees are guaranteed to have a root to tip distance that matches the sampling age of the colony.

Estimating the timing of chromosomal aberration events along a branch

We estimated the timing of LOH and CNA events along branches by first estimating the number of expected detectable mutations, L, in LOH/CNA regions for the duration of the branches. We estimated a local relative somatic mutation rate for mutations detectable by CaVEMan in autosomal regions. This background local mutation rate was measured by counting distinct mutations across a panel of samples consisting of those colonies in the 12 patients that do not exhibit copy number aberrations (457,884 mutations across 734 colonies). The genome was divided into 100 kb bins and the number of passed somatic mutations was counted across all samples in the panel, to give a count c_i for bin b_i. The probability that a given mutation occurred in bin i was estimated by $p(\text{mut} \in b_i) = \frac{c_i}{\sum_j c_j}$ and with standard error in that estimate of $p_{se} = \frac{p(\text{mut} \in b_i)}{\sqrt{c_j}}$. For a given copy number region C then $p(\text{mut} \in C) = \sum_{b_i \in C} p(\text{mut} \in b_i)$. For a branch of duration t and with global mutation rate λ then $E(L) = \lambda t p(\text{mut} \in C)$ and $\text{Var}(L) = \lambda^2 t^2 \sum_{b_i \in C} p_{se}(\text{mut} \in b_i)^2$ where it should be noted that errors in t and λ were not included in the variance estimation. To time copy number neutral LOH events, all somatic mutations that occur prior to the LOH event, occurring a fraction x along the branch, will be homozygous with detection sensitivity s_{HOM} and those after will be heterozygous with detection sensitivity s_{HET}. We modelled the mutations as arriving at a constant rate along the branch and fit the following model for:

$$N_{HET} \sim \text{Poisson}((1-x)Ls_{HET})$$

$$N_{HOM} \sim \text{Poisson}(xLs_{HOM})$$

with priors $x \sim \text{Uniform}(0,1)$ and $L \sim N(E(L), \text{Var}(L))$ and where $s_{HOM} = 0.5$ (assuming perfect detection of homozygous mutant variants) and s_{HET} is estimated from germline SNPs as previously discussed.

To time chromosomal duplications, somatic mutations that occur prior to the CNA event, occurring a fraction x along the branch, have an equal chance of exhibiting VAF = 1/3 or VAF = 2/3, whereas those occurring after the event will always have VAF = 1/3 and are expected to occur at a 50% greater rate.

$$N_{2/3} \sim \text{Poisson}\left(\frac{xLs_{2/3}}{2}\right)$$

$$N_{1/3} \sim \text{Poisson}\left(s_{1/3}\left(\frac{xL}{2} + (1-x)\frac{3L}{2}\right)\right)$$

where the priors are as in the LOH model above. The detection sensitivities $s_{1/3}$ and $s_{2/3}$ are similar to s_{het} because of the additional sequencing depth afforded by the duplication. Unlike the LOH case, the value of x is relatively unaffected by L because of the similarity of $s_{1/3}$ and $s_{2/3}$.

Estimation of the rate of mutation in wild-type and mutant colonies

We used two orthogonal methods to estimate the rate of mutation acquisition in wild-type cells over age. Mixed effects modelling was used to calculate the relationship between age and mutation acquisition. To account for inter-patient rate heterogeneity as well as intra-patient variance we fitted a linear mixed model for the wild type adjusted mutation burden, with the patient as a random effect using the nlme package in R. For each patient we included the time point with the most wild-type colonies and we excluded time points with <3 wild-type colonies. The model was specified as: lme(nsnv_count-age_at_sample_pcy, random =~1 + age_at_sample_pcy|patient (Supplementary Note 5). The second method utilised the time-based 'ultrametric' trees created for individual patients (Extended Data Fig. 5). The resulting posterior mean and standard error of each patient's wild type rate were combined using a random effects meta-analysis. The model incorporated an excess mutation rate in early life that added an average of 33.5 mutations during the first 6 months post conception, this corresponds to fixing the intercept in a linear model to a mean of 33.5 mutations at conception, or around 50 at birth. The mean branch timings were directly sampled from the MCMC posterior distribution and by construction the resulting trees are guaranteed to have a root to tip distance that matches the sampling age of the colony. The resulting posterior mean and standard error of each patient's wild type rate were combined in a random effects meta-analysis using the R package metafor. The cohort wide wild type rate from this approach is depicted on Extended Data Fig. 3a. The within patient difference in burden between wild type clades and mutant clades were also tested using a non-parametric method limma's rankSumTestWithCorrelation where the non-independence of the samples in the mutant clade is corrected using a single estimate of the average correlation between samples. The pairwise correlation, $c_{i,j}$, of colonies i and j in the mutant clade were estimated as

$$c_{i,j} = \frac{b_{i,j}}{\sqrt{b_i b_j}}$$

where b_i is the total adjusted burden of colony i and $b_{i,j}$ is the total adjusted length of branches shared by i and j. The underlying assumption behind this correlation estimation is that the expected variance in burden is directly proportional to the burden. Note that rankSumTestWithCorrelation requires one of the comparison sets to be independently sampled. We therefore only carried out comparisons for cases where the mean correlation between wild type colonies is less than 0.01.

HSC simulator Rsimpop

A birth death process was used to model the evolution of a population of cells, tracking each cell as it symmetrically divides. Each cell had a rate of symmetric division and a rate of symmetric differentiation (or death). Asymmetric divisions did not affect the HSC genealogy and were therefore not explicitly included in the model. The wild type rate of symmetric division was, measured in divisions per day. We modelled selective advantage s as an increased rate of symmetric division $\alpha_{mut} = \alpha(1+s)$. We assumed during the growth phase that the cells population grows unrestrained by death. Once the specified equilibrium population size, N, was reached then the death rate β, which is the same for every cell, matched the average division rate:

$$\sum^{\text{cells}} \beta = \sum^{\text{wild-type cells}} \alpha + \sum^{\text{mutant cells}} (1+s)\alpha$$

Thus giving

$$\beta = \frac{N_m(1+s)\alpha + (N - N_m)\alpha}{N}$$

For large N_m the expected number of symmetric divisions in some small time period δt is:

$$E(N_m(t+\delta t)) = N_m(t) + N_m(t)(\alpha(1+s) - \beta)\delta t$$

Where here we assumed δt is sufficiently small for the relative change in N_m to be small, but sufficiently large so that $N_m(t)(\alpha(1+s) - \beta)\delta t$ is in the regime where the change in clone size is deterministic. Substituting in for β:

$$E(N_m(t+\delta t)) = N_m(t)\alpha s\left(1 - \frac{N_m(t)}{N}\right)\delta t$$

Gives the following differential equation:

$$\frac{dN_m}{dt} = N_m(t)\alpha s\left(1 - \frac{N_m(t)}{N}\right)$$

Which is solved by the logistic function:

$$N_m = N\frac{1}{1 + exp\left(-\alpha s\left(t - t_m\right)\right)}$$

For some constant t_m.

Following the acquisition of a driver mutation the mutant cell population grows stochastically until the population is sufficiently large, when the growth becomes essentially deterministic following a logistic growth function where in the early stages the exponential growth process exhibits an annual rate of growth S, given by:

$$S = \exp(\alpha s) - 1.$$

The above model is implemented using the Gillespie algorithm where the waiting time until the next event is exponentially distributed with a rate given by the total division rate + total death rate, this event is then: division with probability = total division rate/(total division rate + total death rate). If the event is division then the choice of which cell is given by a probability proportional to the cell's division rate whereas if the event is death then all cells are equally likely to be chosen.

Implementation was in C++ with an R based wrapper as an R package rsimpop (see https://github.com/NickWilliamsSanger/rsimpop). The simulator maintains a genealogy of the extant cells, together with a record of the number of symmetric divisions on each branch, the absolute timing of any acquired drivers and the absolute timings of branch start and end. Assessment of the expected behaviour of rsimpop is detailed in Supplementary Note 8. The package also provides mechanisms for sub-setting simulated genealogies whilst preserving the above per branch information.

Approximate Bayesian computation

ABC is a flexible inference procedure that enables the inference of parameters governing a well-defined process that is simple to simulate but for which it might be difficult to calculate the likelihood. This approach has been previously used to infer the HSC population and cell division rate parameters under the assumption of a neutrally evolving population with a stable population size[20]. The general approach was to recapitulate the experiment in-silico, to reproduce the timing and shape of clonal expansions in our phylogenetic trees, using our simulator to generate sampled trees. If the mutation rate, lambda, is the mean root to tip distance divided by the age of sampling, and the mutation counts at the start and end of the branch carrying the driver in the experimental tree are denoted M_{start} and M_{end}, respectively, then the procedure was as follows:

- Fix symmetric division rate at 1 division per year.
- Sample N from $\log_{10}(N) \sim \text{Uniform}(3,6.5)$
- Sample age of driver acquisition by resampling mutation counts from a Poisson distribution:
  - $T_{\text{driver}} \sim \text{Uniform}\left(\frac{\text{Poisson}(M_{\text{start}})}{\lambda}, \frac{\text{Poisson}(M_{\text{end}})}{\lambda}\right)$
  - For small M_{start} use a more refined version of the above†
- Sample S from $S \sim \text{Uniform}(0.05,4)$
- Simulate the population evolution:
  - Simulate the tree with initial division rate of 0.1 per day until population has grown to the equilibrium population size.
  - Simulate neutral evolution until time T_{driver}.
  - Save the state of the simulation (*)
  - Introduce the driver with the specified selection coefficient.
  - If the driver lineage dies out before the sampling age is reached then return to the saved state (*)

A tree with the observed number of mutant samples is subsampled from the population of extant cells.

For the case where M is small (<50) we need to account for the discrete character of M and also allow for the drivers being acquired during the growth phase of the population, as might occur early in life. We wished to find the probability distribution for the age at which a given observed mutation count M occurred. We used Baye's rule and assumed a uniform prior for λt between 0 and λT, and that the mutations are Poisson distributed with rate λt. This gives the following cumulative distribution function for λt:

$$Q(\lambda t|m, \lambda T) = \frac{1 - e^{-\lambda t}\sum_{i=0}^{m}(\lambda t)^i/i!}{1 - e^{-\lambda T}\sum_{i=0}^{m}(\lambda T)^i/i!}$$

This distribution is referred to as $\text{PoisInv}(m, \lambda T)$ and used for sampling λt in 'mutation time' space. To account for the elevated mutation rate that occurs during embryogenesis we used the same model as described above (section 'Estimation of the rate of mutation in wild-type and mutant colonies'). This implies a time varying rate with the following average accumulation of mutations between time t_0 and t_1:

$$M(t_0, t_1) = \lambda t + \frac{L}{k(t_1 - t_0)}(\log(1 + e^{k(t_1 - t_m)}) - \log(1 + e^{k(t_0 - t_m)}))$$

Where, as before, we set $L = 149.2$, $k = 50.0$ and $t_m = 0.225$. For the timing of the start of the branch we sample $a \sim \text{PoisInv}(M_{\text{start}}, \lambda T)$ and then solve the following for T_{start}:

$$a = M(0, T_{\text{start}})$$

For the timing of the end of the branch conditional on the timing of the start of the branch we sample $b \sim \text{PoisInv}(M_{\text{end}} - M_{\text{start}})$ and solve the following for T_{end}:

$$b = M(T_{\text{start}}, T_{\text{end}})$$

Finally, the driver acquisition time is sampled from:

$$T_{\text{driver}} \sim \text{Uniform}(T_{\text{start}}, T_{\text{end}})$$

For each simulation the following summary statistics were calculated (i) Total deviation in the simulation mutant clade's number of lineages through time (LTT) with respect to the patient clade of interest, and (ii) deviation of the simulation-based population clonal fraction with respect to the patient clade's aberrant cell fraction (ACF). In all cases but two, ACF was calculated as the proportion of sampled mutant clades. For PD5163_JAK2 and PD5182_JAK2 clades, the aberrant cell fraction at diagnoses were used prior to the commencement of interferon-alpha treatment which led to clone size reduction. The total distance score is then calculated as the sum of the above two scores where each score is expressed as a rank. For each clade approximately 1 million simulations were run (Extended Data Fig. 6b). In each case, the posterior distribution was approximated by the top 0.01% simulations.

Calculation of aberrant cell fraction in recapture samples

Recapture samples were used to validate inferences derived from phylogenetic trees. In recapture samples, a per branch ACF was calculated as twice the aggregate mutant read fraction where only autosomal variants that map to the branch and are outside of the copy number aberrant regions are included. For each patient, the trajectory of the ACF was retrieved from the top 0.01% of simulations. The simulator periodically takes snapshots at approximately daily intervals and measures ACF as the current fraction of cells that carry the driver mutation. Subsequently, for each simulation these daily snapshots are binned into a common sequence of 10-day periods over each of which the ACF is

averaged. We present the 2.5%, 50% and 97.5% quantiles for each binned period calculated across the simulations.

Stochastic Extinction

The probability of extinction in a homogenous birth death process is the ratio of death rate to birth rate[56] $\frac{\alpha}{\alpha+\log(1+S)}$ which in our case is $\frac{1}{1+s}$. In the ABC simulations we record the number of attempts required to introduce the drivers. We then verified that the simulator behaved as expected by gathering all simulations across the analysed mutant clades and then restricted to the 13,048,861 simulations where the driver was introduced after development (>1 year post conception). The simulations were binned into selection coefficient bins of width = 0.05 and $\log_{10}(N)$ bins of width = 0.1. The extinction probability was estimated in each bin using the maximum likelihood estimator $1-\frac{q}{\sum_{i=1}^{q} a_i}$ where q is the number of simulations in the bin and each simulation, i, fixes after a_i attempts.

Mutational signatures

De novo signature extraction was carried out using SigProfiler. The SNVs that were mapped to the 12 patient trees were divided into per patient driver/wild-type clade-based groupings, where the mutations mapped to the clades were treated as an individual sample for the purposes of mutational signature extraction as detailed in Supplementary Note 6.

Telomere analysis

The mean telomere lengths of the colonies were estimated using Telomerecat with batch correction using cohort wide information to correct the error in F2a counts as detailed. Given the slight discrepancies in length found on multiple readings, each telbam was analysed 10 times and the average length was used. To compare differences in telomere lengths, the degree of shared clonal origin of the colonies was taken into account. To assess for telomere attrition in relation to mutant *JAK2*, we fit a phylogeny aware mixed model for the mean telomere length with a patient specific intercept using the MCMCglmm library in R (iterations = 1,100,000, burnin = 100,000, thinning interval = 1,000). Further information is provided in Supplementary Note 7.

Testing genes under selection

We searched for genes that exhibit a deviation of ratio of non-synonymous to synonymous variants as evidence of non-neutrality. The SNVs and indels were grouped by individual branches across the cohort and the function dndscv from R package dndscv (version dndscv_0.0.1.0) was executed using the default options. The analysis highlighted that only JAK2, DNMT3A, PPM1D and TET2 were significantly under selection.

Phylofit

As an alternative to ABC method we have developed an efficient MCMC approach that models selection/growth by directly fitting the three parameter deterministic phase population trajectory using the joint probability density of coalescence times given the population size trajectory. The model is essentially a parametric adaptation of the phylodyn model. The starting point is Equation 1 in S. Lan. Et al An Efficient Bayesian Inference Framework for Coalescent-Based Nonparametric Phylodynamics:

$$P(t_1,..,t_n \mid N(t)) = \prod_{k=2}^{n} \binom{k}{2} \frac{1}{N(t_{k-1})} e^{-\int_{t_k}^{t_{k-1}} \binom{k}{2} \frac{1}{N(t_{k-1})} dt}$$

Where $\{t_k | k \in 1..n\}$ are the timings of the time ordered coalescences belonging to the mutant clade where t_1 is the first split of the expansion and t_n is the sampling time where these times are expressed as the gap between the event and the sampling time (here assumed isochronous). Substituting our formula for the aberrant cell count $N(t)$ and performing the integral and eliminating terms that do not depend on overall population size, N, the trajectory midpoint, $t^{(m)}$, and the selection coefficient, $\hat{s} = \alpha s$, we arrive at the following log-likelihood:

$$\begin{aligned} &L(t_1,..,t_n|N(t)) \\ &= (n-1)\log(N) + \sum_{k=2}^{n} (\log(1+\exp(\hat{s}(t_{k-1}-T+t^{(m)})))) \\ &-\frac{1}{\hat{s}N} \sum_{k=2}^{n} \left(\binom{k}{2} \exp(\hat{s}(t_{k-1}-T+t^{(m)}))(\exp(\hat{s}(t_{k-1}-t_k))-1) \right) \\ &+\frac{1}{N} \sum_{k=2}^{n} \left(\binom{k}{2} \hat{s}(t_{k-1}-t_k) \right) \end{aligned}$$

Where recall the annualised selection coefficient is $S = \exp(\alpha s) - 1 = \exp(\hat{s}) - 1$.

We incorporate this central likelihood equation into a Bayesian model with uniform priors on $\log_{10}(N)$, $\hat{s}$ and $t^{(m)}$.

$$\hat{s} \sim U(0.001, 2)$$

$$t^{(m)} \sim U(a, b)$$

$$\log_{10}(N) \sim U(4, 7)$$

$$\boldsymbol{t} \sim \text{Phylo}(\hat{s}, t^{(m)}, N)$$

Where Phylo is the distribution defined by the coalescence based log-likelihood equation above.

Additionally, assuming unbiased sampling, we can optionally incorporate the number of sampled mutant-type colonies n_{mut} out of n_{tot} total colonies as an additional layer in the model

$$n_{\text{mut}} \sim \text{Binomial}\left(n_{\text{tot}}, \frac{1}{1+\exp(-\hat{s}(T-t^{(m)}))}\right)$$

This last addition is akin to including VAF targeting in the ABC estimation. The parameters, a and b, setting the realistic range for the midpoint depend on whether the last component of the model is active. The highest and lowest probable values of the aberrant cell fraction (ACF) at colony sampling is estimated from the 99.0% confidence interval of the ACF based on observing n_{mut} of n_{tot} colonies. We the solve to find $t^{(m)}$ under these extreme ACF values in conjunction with extreme values for $\hat{s}$ giving four possible extreme values of $t^{(m)}$. The parameters a and b are respectively the minimum and maximum of these four values. The model is implemented in RSTAN. The input data for this approach is an ultrametric tree. For our samples we obtain the ultrametric tree using the Bayesian tree model presented in Section 8. Credibility intervals for this method do not consider the uncertainty in the branch lengths of the input ultrametric tree, but the model allows us to straightforwardly incorporate what is likely to be the principal source of additional uncertainty, the timing of the first split. We achieve this by optionally adding a normally distributed common offset to the coalescence timings t_i.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Whole-genome sequencing data in the form of BAM files across all samples reported in this study have been deposited in the European Genome–Phenome Archive (https://www.ebi.ac.uk/ega/home) with accession codes EGAD00001007714 (whole-genome sequencing

colonies) and EGAD00001007715 (targeted-recapture sequencing). Per patient VCF files containing information on somatic mutations identified are available on Mendeley (doi: 10.17632/hrmxybrd2n.1) .

Code availability

Single-nucleotide substitutions (SNV) were called using the cancer variants through expectation maximization (CaVEMan) algorithm, version 1.13.14 (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm as implemented in the cgpPindel workflow, version 3.2.0 (https://github.com/cancerit/cgpPindel). Copy number variants were called using the ASCAT algorithm as implemented in the ascatNgs workflow, version 3.2.0 (https://github.com/cancerit/ascatNgs). Mutational signatures analysis was performed using MutationalPatterns v1.10, available on Github (https://github.com/UMCUGenetics/MutationalPatterns) and SigProfiler (https://github.com/AlexandrovLab). Allele counts at SNV and indel sites were carried out using vafCorrect (https://github.com/cancerit/vafCorrect). Telomere lengths were estimated using telomerecat, version 3.2 (https://github.com/cancerit/telomerecat). Mutations were mapped to phylogenetic branches using Rtreemut developed for this study (https://github.com/NickWilliamsSanger/treemut). Temporal branch lengths and per driver mutation rates were inferred using rtreefit developed for this study (https://github.com/NickWilliamsSanger/rtreefit). Simulation of HSC populations and phylogenies with selection were carried out using rsimpop developed for this study (https://github.com/NickWilliamsSanger/rsimpop). Other analyses were carried out using custom R scripts available at https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution.

49. Nangalia, J. et al. Somatic CALR mutations in myeloproliferative neoplasms with nonmutated JAK2. *N. Engl. J. Med.* **369**, 2391–2405 (2013).
50. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
51. Jones, D. et al. cgpCaVEManWrapper: simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinforma.* **56**, 15.10.1–15.10.18 (2016).
52. Ye, K., Schulz, M. H., Long, Q., Apweiler, R. & Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* **25**, 2865–2871 (2009).
53. Van Loo, P. et al. Allele-specific copy number analysis of tumors. *Proc. Natl Acad. Sci. USA* **107**, 16910–16915 (2010).
54. Hoang, D. T. et al. MPBoot: fast phylogenetic maximum parsimony tree inference and bootstrap approximation. *BMC Evol. Biol.* **18**, 11 (2018).
55. Carpenter, B. et al. Stan: a probabilistic programming language. *J. Stat. Softw.* **76**, 1–32 (2017).
56. Tavaré, S. The linear birthdeath process: An inferential retrospective. *Adv. Appl. Probab.* **50**, 253–269 (2018).

Acknowledgements We thank Cambridge Blood and Stem Cell Biobank, funded by the Cambridge Cancer Centre and Wellcome Trust Cambridge Stem Cell Institute, Wellcome Sanger CASM and DNA pipelines for their assistance; and S. Behjati and C. Harrison for valuable discussion. The study was supported by Cancer Research UK (J.N.), EHA Research Award (J.N.), MPN Research Foundation (J.N.) and the Wellcome Trust (P.J.C., A.R.G. and J.L.). Work in the A.R.G. laboratory is supported by the Wellcome Trust, Bloodwise, Cancer Research UK, the Kay Kendall Leukaemia Fund and the Leukaemia and Lymphoma Society of America. J.N. is a CRUK Clinician Scientist fellow. We thank the patients for their participation in the study.

Author contributions J.N., A.R.G. and P.J.C. conceived the study. N.W. performed genomic, phylogenetic and population dynamics analyses with J.N. J.L. assisted with signature, clinical and telomere analyses. E.M. provided genomic data and analyses for normal samples. L.M. assisted with low-input sequencing and mutation signature analysis. A.L.G. assisted with clinical correlation. J.N. and E.J.B. obtained samples. K.J.D. assisted with simulation inferences. A.M. and J.H assisted with computational and laboratory processing pipelines. J.N. directed the study and wrote the manuscript with input from co-authors. All authors reviewed and approved the manuscript.

Competing interests A patent has been filed by the Wellcome Sanger Institute (inventors N.W. and J.N.; Application number PCT/EP2021/071952) covering somatic mutation identification in the context of tumour contamination of the matched germline sample.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-04312-6.

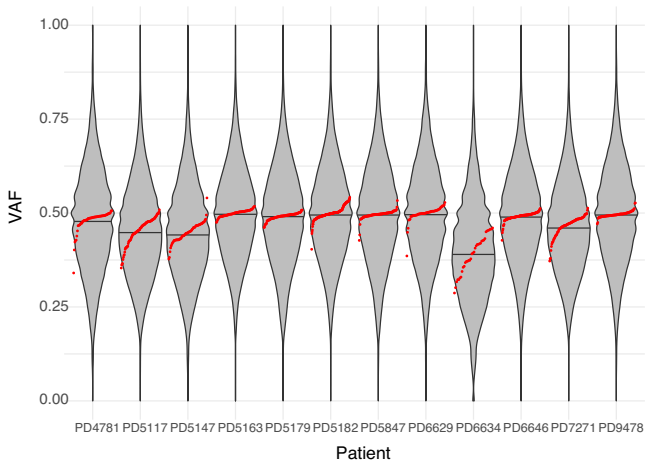
Correspondence and requests for materials should be addressed to Jyoti Nangalia.

Peer review information *Nature* thanks Steven McCarroll, Seishi Ogawa and the other, anonymous reviewers for their contribution to the peer review of this work. Peer review reports are available.

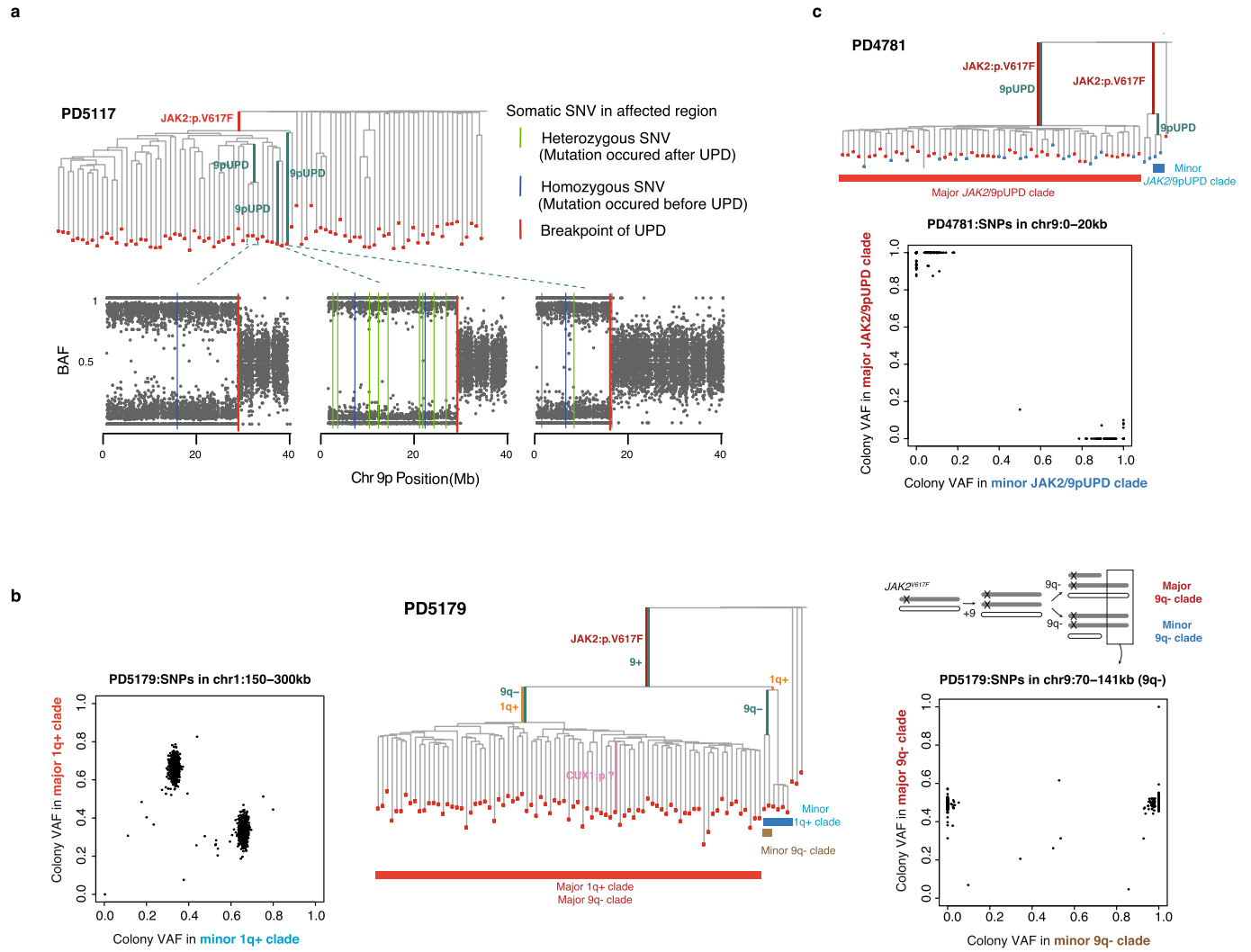
a

Patient	Diagnosis	Gender	Age at Diagnosis Clinical features	Blood counts: diagnosis (top); colony sample (bottom) Hb g/L	Hct	WBC x10^9/L	Plts x10^9/L	Cytoreduction	Thrombosis	Disease progression (age)	Haematological response at colony timepoint	Alive Yes/No (age at death/FU)
PD7271	ET	F	20 Asymptomatic	145 143	0.42	8.9 7.9	923 640	None	No	No	NA	Yes (27)
PD5163	PV	F	31 Portal/splenic vein thrombosis	164 127	0.49	11.6 4.4	435 133	IFN age 32-44; stopped due to cytopenia	Yes	No	Yes	Yes (46)
PD5117	PV	F	64 Asymptomatic	166 136	0.5	10.6 4.1	831 271	HC	No	No	Yes	No (89)
PD5182	PV	M	32 Asymptomatic	179 140	0.54 0.45	10 5	604 397	IFN; refractory age 50, switched to HC	No	No	No	Yes (54)
PD5847	PV	F	44 Portal vein thrombosis	190 190	0.59 0.59	20 20	504 504	IFN	Yes	No	No	Yes (51)
PD9478	PV	F	53 Asymptomatic	201 140	0.6	5.1 15.9	308 111	None (intermittent venesection)	No	PPV-MF (71)	No	Yes (76)
PD4781	PV	F	73 Asymptomatic	151 108	0.46* 	9.5 18	634 193	HC	Yes	No	No	No (86)
PD6646	ET	F	76 Asymptomatic	145 106	0.44 0.32	7.7 2.8	804 409	HC; switched to Pipobroman age 85	No	No	No	No (87)
PD6629	ET	M	54 Asymptomatic	139 149	0.43 	13.5 8.7	842 433	IFN, switched to HC age 59	Yes	No	No	Yes (70)
PD5179	PV	M	34 Budd-Chiari syndrome	182 98	0.55 	64.3 12.5	200 646	HC; switched to Ruxolitinib post MF	Yes	PPV-MF (48)	No	No (51)
PD6634	ET	M	26 Asymptomatic	141 145	0.42 0.45	12.8 5.2	1121 495	HC commenced age 48	No	No	Yes	Yes (63)
PD5147	PV	F	81 Asymptomatic	179 115	0.56 0.35	10.3 4.3	505 134	HC, switched to Pipobroman age 87	No	No	Yes	No (91)

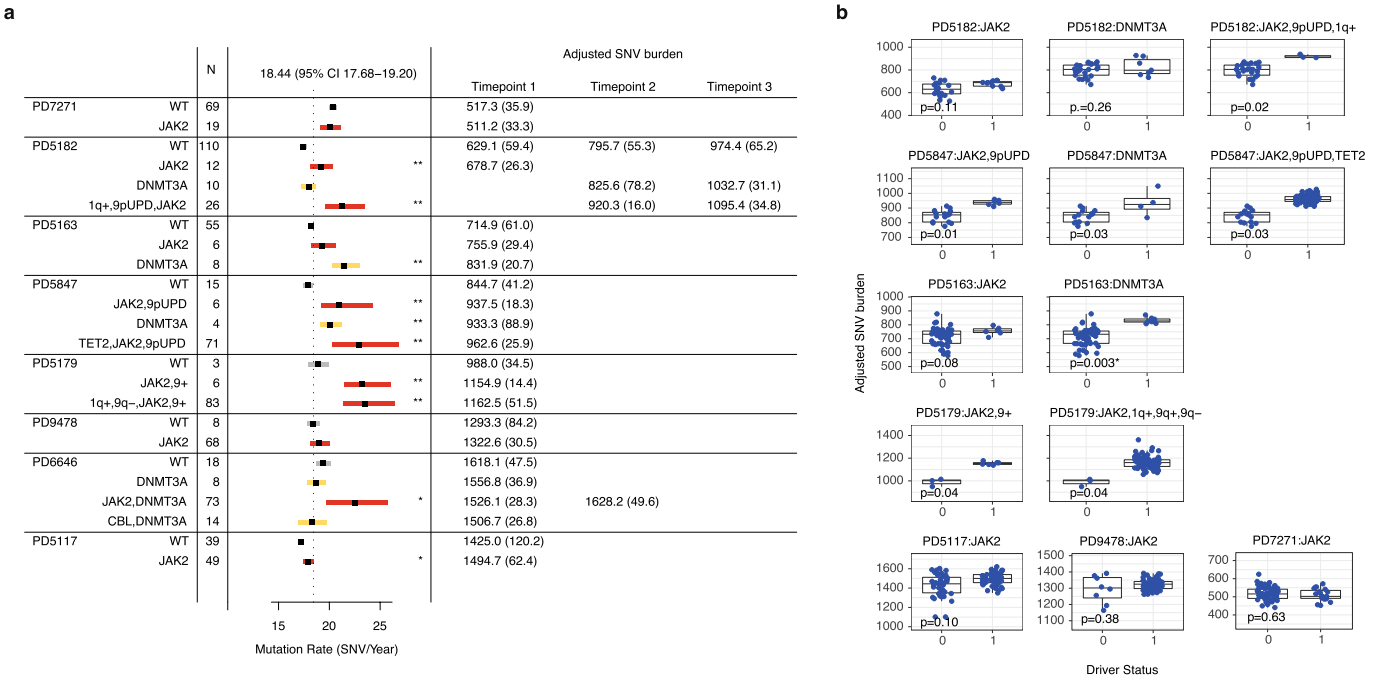
b



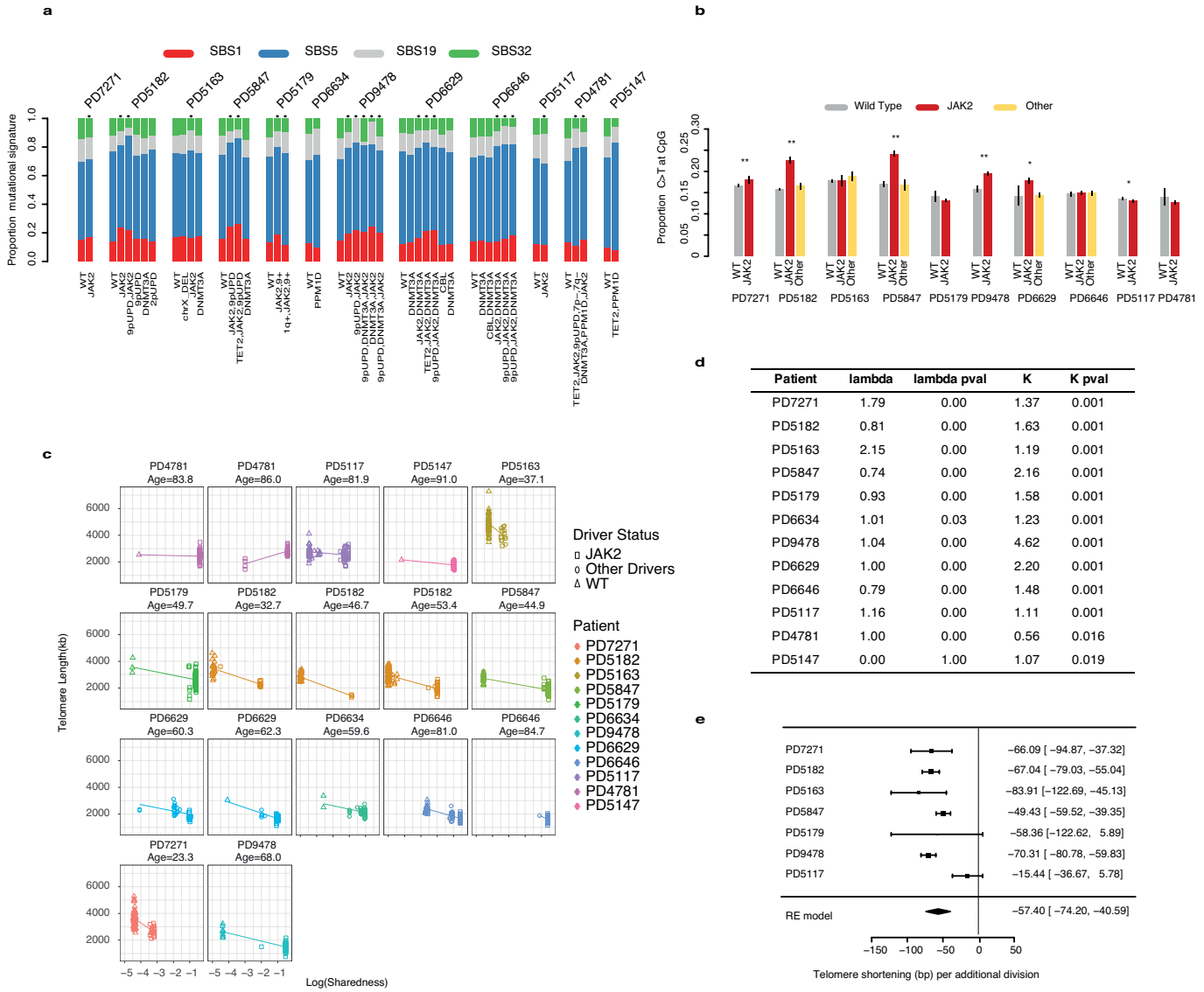
Extended Data Fig. 1 | legend. Patient characteristics and somatic mutation fractions in haematopoietic colonies. a. Patient characteristics. PV, Polycythemia vera; ET, Essential thrombocythaemia; MF, myelofibrosis; HC, Hydroxycarbamide; IFN, Interferon-alpha; FU, follow-up. *PV diagnosed on red cell mass study. **b**. The distribution of variant allele fractions (VAF) for point mutations pooled across colonies per patient. The mean VAF of individual colonies is shown as red dots. Only autosomal somatic mutations are shown, with those in regions with copy-number aberrations and loss-of-heterozygosity excluded. The plot shows that the colony VAFs are close to 0.5 for the majority.



Extended Data Fig. 2 | Legend: Parallel evolution within phylogenetic trees. **a**. Phylogenetic tree of PD5117 depicting 3 separate 9pUPD (UPD, uniparental disomy) acquisitions (blue branches), downstream of *JAK2*^{V617F} (red branch). Below the phylogenetic tree are three B-allele frequency plots showing the regions of 9pUPD in the different clades with vertical red lines showing the boundary of loss of heterozygosity. The event shown on the far right has a distinct breakpoint from the left two events. Blue and green vertical lines show somatic mutations (either prior or subsequent to the UPD event), suggesting that the 9pUPD event depicted in the middle plot occurred first as more mutations have had time to accrue since the copy number aberration. **b**. Phylogenetic tree of PD5179 depicting two separate 1q+ (orange branches) and 9q- (blue branches) acquisitions. Left plot shows the aggregate VAF of germline single nucleotide polymorphisms (SNP) on Chr1 for samples in the 1q+ major clade versus 1q+ minor clade (left plot). SNPs at a VAF = 2/3 in one clade are at 1/3 in the minor clade, and vice-versa, confirming that different parental chromosomes are amplified in each clade. SNPs in the affected 9q- region also exhibit a clear pattern in VAF (right panel), with VAF = 0.5 for samples in the major 9q- clade but VAF = 0 or 1 for samples in the minor 9q- clade. A proposed model of chr9 copy number changes is shown in the upper right. **c**. Phylogenetic tree of PD4781 depicting two separate *JAK2*^{V617F} acquisitions (red branches) each followed by 9pUPD (blue branches). *JAK2*^{V617F} acquisition occurred on different parental alleles in each instance as SNPs on 9p that have a VAF ~1 for samples in the major *JAK2*-mutant clade (horizontal bar coloured red) have a VAF ~0 in samples from the minor *JAK2*-mutant clade (horizontal bar coloured blue) and vice-versa.

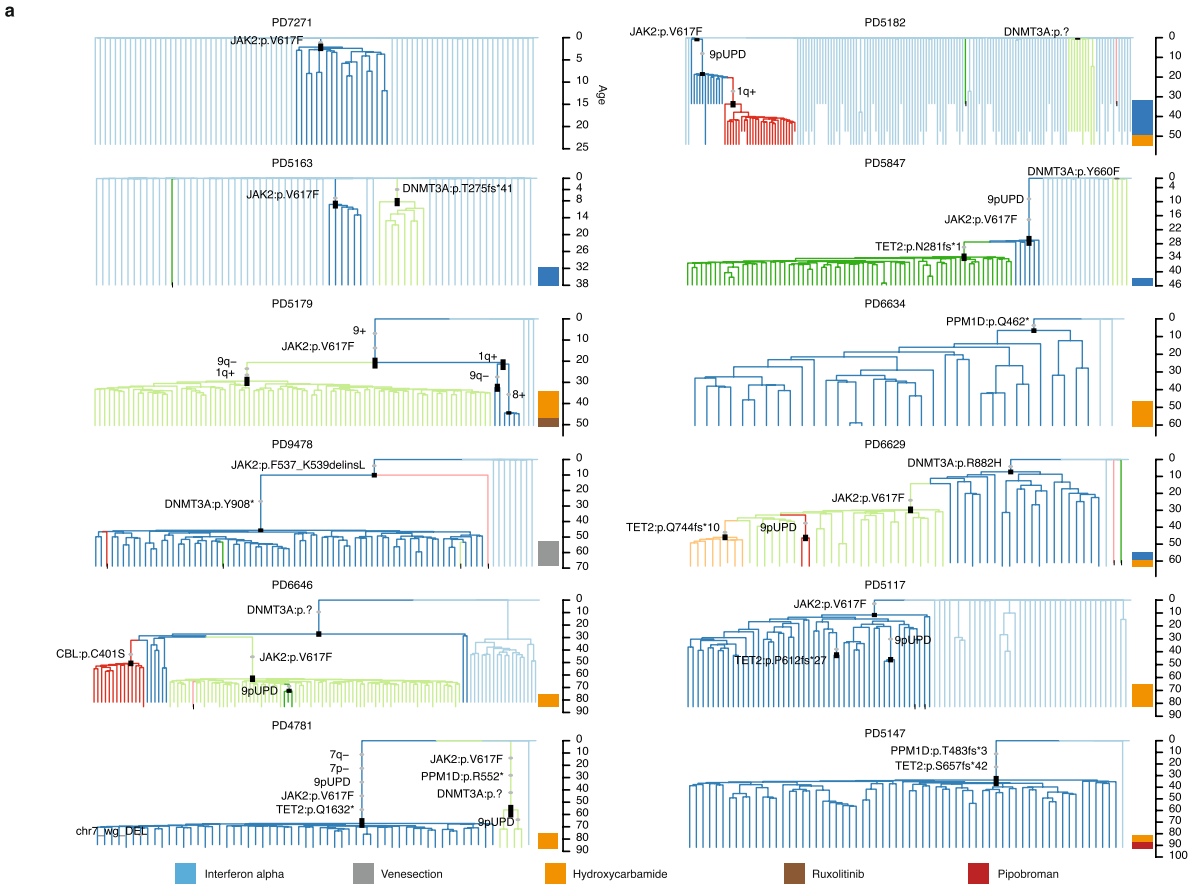


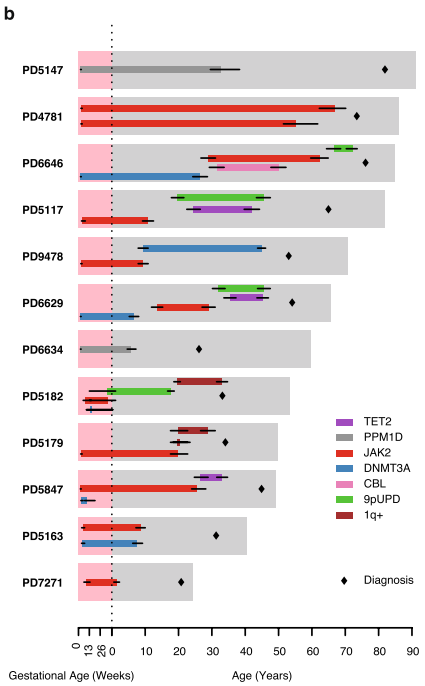
Extended Data Fig. 3 | Legend: Mutation rates and burden following driver mutation acquisition. a. Mutation rate estimates for wildtype and different mutant clades within patients. Mutation acquisition is modelled using Poisson modelling taking into account the timing of transition from wildtype to driver mutation acquisition within mutant clades and an excess mutation rate earlier in life (Methods). Patients and genotypes of clades are shown on the left together with colony number for each clade (N). Wildtype (WT) clades are shown in grey bars, *JAK2*-mutated clades are shown in red and other mutant clades are shown in yellow. The cohort wide estimate for the mutation rate in WT colonies is shown by the dotted black vertical line at the top. *$P < 0.05$, **$P < 0.01$ (** also significant after multiple hypothesis testing; Bonferonni adjusted, two-sided test). Significantly different mutation rates between clades are highlighted only for those significant by both Poisson and Negative Binomial modelling of mutation rates (Methods). Average mutation burdens are shown to the right for the different timepoints of sampling. **b**. Non-parametric comparison of mutation burdens in wildtype versus mutant colonies using limma's rankSumTestWithCorrelation. This accounts for the non-independence of data in mutant colonies but does not account for the timing of driver mutation acquisition. *indicates significance at $P < 0.05$ following Bonferonni multiple hypothesis correction.



Patient	lambda	lambda pval	K	K pval
PD7271	1.79	0.00	1.37	0.001
PD5182	0.81	0.00	1.63	0.001
PD5163	2.15	0.00	1.19	0.001
PD5847	0.74	0.00	2.16	0.001
PD5179	0.93	0.00	1.58	0.001
PD6634	1.01	0.03	1.23	0.001
PD9478	1.04	0.00	4.62	0.001
PD6629	1.00	0.00	2.20	0.001
PD6646	0.79	0.00	1.48	0.001
PD5117	1.16	0.00	1.11	0.001
PD4781	1.00	0.00	0.56	0.016
PD5147	0.00	1.00	1.07	0.019

Extended Data Fig. 4 | Legend: Mutational signatures and telomeres. **a**. Signature contributions of SBS1, SBS5, SBS19 and SBS32 on a per-patient/per-clade basis. Single base substitution mutational signature 5 (SBS5), thought to represent a time-dependent mutational process active in all tissues, was the predominant mutational process in colonies . **b**. The proportion of C>T transitions at CpG dinucleotides across WT, *JAK2*-mutated and colonies with other driver mutations. $^{*}P < 0.05$, $^{**}P < 0.01$ (** also significant after multiple hypothesis testing; Chi-square test). **c**. The relationship between 'sharedness' (see Methods) and telomere length across all phylogenetic trees shows that telomeres shorten in line with increased phylogenetic 'sharedness' in keeping with the increased cell divisions during clonal expansion. **d**. The heritability of telomere length, that is, whether closely related colonies had more similar telomere lengths compared to more distantly related colonies, is assessed using Pagel's Lambda and Blomberg's K, with both values in the vicinity of 1 or above, suggesting that telomere length variation across colonies in a phylogenetic tree follows the expected covariance based on phylogenetic relationship. Power for PD5147 is limited because there is little difference in 'sharedness' in the mutant colonies. **e**. The modelled reduction in telomere length per additional stem cell division in *JAK2* mutant clades is shown per patient, with a cohort wide estimate of −57.4bp (−74.2, −40.59 95% CI). See Supplementary Note 7 for further interpretation.





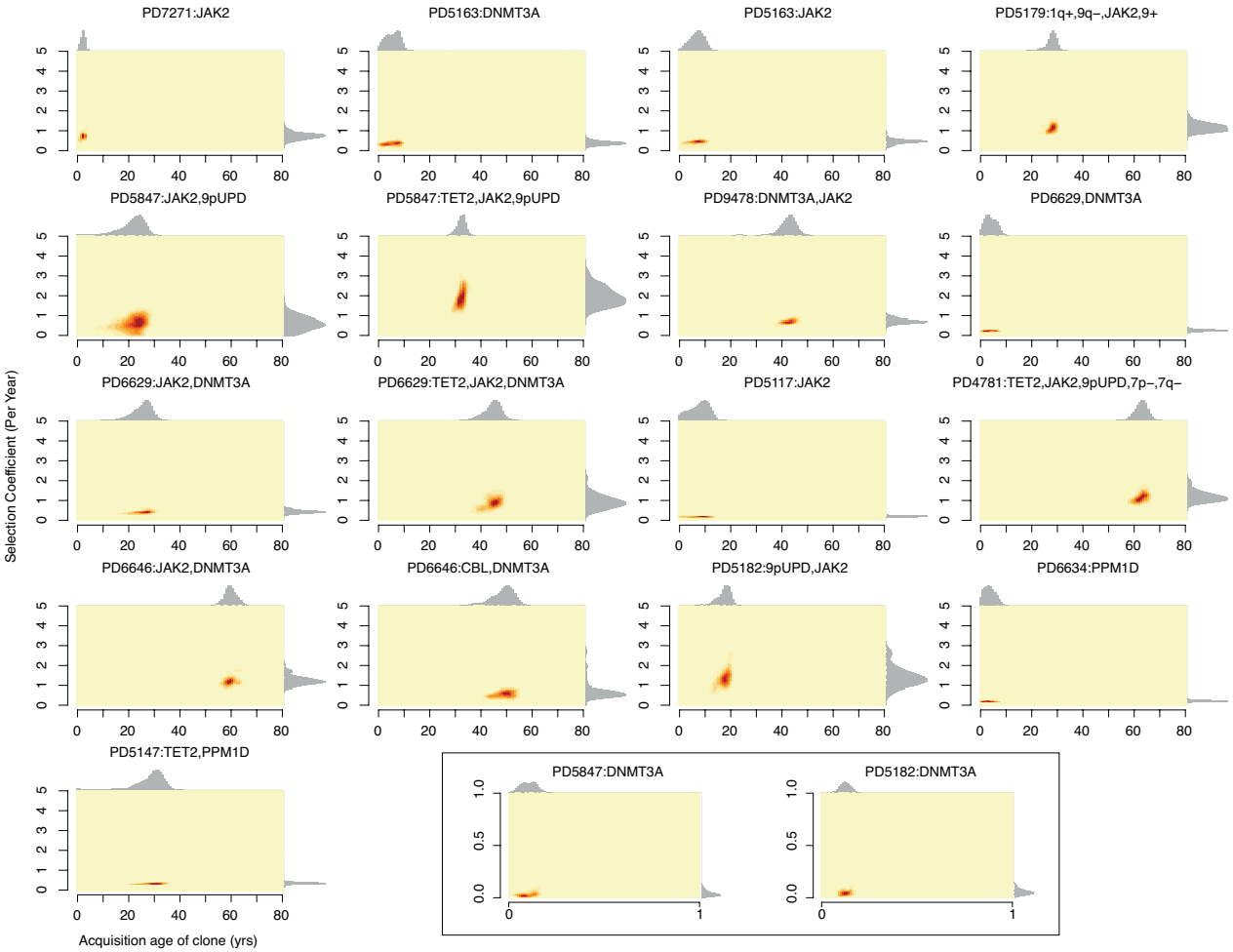
c

Patient	Mutant clone	Start of branch age			End of branch age			Mutation count	
		Median	Lower 95% CI	Upper 95% CI	Median	Lower 95% CI	Upper 95%CI	Start	End
PD5182	DNMT3A	12Wks-pc	7Wks-pc	0.21	14Wks-pc	8Wks-pc	0.29	20	21
PD5847	DNMT3A	2Wks-pc	1Wks-pc	3Wks-pc	9Wks-pc	5Wks-pc	18Wks-pc	2	22
PD5182	JAK2	6Wks-pc	4Wks-pc	13Wks-pc	33Wks-pc	11Wks-pc	1.13	17	44
PD7271	JAK2	7Wks-pc	5Wks-pc	12Wks-pc	1.38	0.6	2.3	18	72
PD6634	PPM1D	0Wks-pc	0Wks-pc	1Wks-pc	5.79	4.6	7.23	0	162
PD6629	DNMT3A	0Wks-pc	0Wks-pc	1Wks-pc	6.54	5.21	8.05	0	162
PD5163	DNMT3A	3Wks-pc	2Wks-pc	5Wks-pc	7.58	6.26	9.2	9	197
PD5163	JAK2	3Wks-pc	2Wks-pc	5Wks-pc	8.55	7.22	10.02	7	204
PD9478	JAK2	1Wks-pc	1Wks-pc	2Wks-pc	9.32	7.89	10.89	1	218
PD5117	JAK2	3Wks-pc	2Wks-pc	6Wks-pc	10.78	9.19	12.5	4	228
PD5182	9pUPD,JAK2	33Wks-pc	11Wks-pc	1.13	17.72	16.71	18.71	44	391
PD5179	JAK2,9+	1Wks-pc	1Wks-pc	2Wks-pc	19.75	17.59	22.75	2	456
PD5179	1q+,JAK2,9+	19.75	17.59	22.75	20.51	18.38	23.45	456	474
PD6646	DNMT3A	1Wks-pc	0Wks-pc	1Wks-pc	26.44	24.27	28.67	0	519
PD5847	JAK2,9pUPD	0Wks-pc	0Wks-pc	1Wks-pc	25.64	23.93	28.18	0	534
PD6629	JAK2,DNMT3A	13.53	11.89	15.26	29	27.05	30.99	289	568
PD5147	TET2,PPM1D	1Wks-pc	0Wks-pc	1Wks-pc	32.8	29.55	38.3	0	628
PD5179	1q+,9q-,JAK2,9+	19.75	17.59	22.75	28.66	26.62	31	456	661
PD5182	1q+,9pUPD,JAK2	19.59	18.6	20.54	32.94	31.4	34.69	428	669
PD5847	TET2,JAK2,9pUPD	26.42	24.74	28.87	32.96	31.45	34.7	550	690
PD5117	TET2,JAK2	24.48	22.48	26.52	41.99	39.68	44.27	472	783
PD5117	9pUPD,JAK2	19.69	17.86	21.6	45.49	43.36	47.53	385	829
PD6629	TET2,JAK2,DNMT3A	35.43	33.59	37.24	45.3	43.51	46.96	685	858
PD6629	9pUPD,JAK2,DNMT3A	32.07	30.17	33.96	45.59	43.7	47.51	624	873
PD9478	DNMT3A,JAK2	9.32	7.89	10.89	44.93	43.7	46.11	218	884
PD6646	CBL,DNMT3A	31.44	29.23	33.67	50	47.69	52.23	611	950
PD4781	DNMT3A,PPM1D,JAK2	1Wks-pc	1Wks-pc	2Wks-pc	55.11	51.46	61.74	1	996
PD6646	JAK2,DNMT3A	28.87	26.71	31.08	62.44	59.63	64.88	563	1118
PD4781	TET2,JAK2,9pUPD,7p-,7q-	1Wks-pc	1Wks-pc	2Wks-pc	66.75	62.21	70.14	1	1159
PD6646	9pUPD,JAK2,DNMT3A	66.66	64.38	68.62	72.06	70.3	73.57	1206	1334

Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | legend: Time based trees and timing driver mutation acquisition. a. Time-based phylogenetic trees. Different coloured branches identify separate clades alongside light blue wild type colonies. The vertical axis represents age post conception with treatment received alongside. Driver mutations are depicted in the middle of the branches but may have occurred at any point between the start and end of the branches. Given the uncertainties in the exact ages at the starts and ends of the branches due to modelling branch lengths from mutation count data (Methods), the credibility intervals for the ends of the branches harbouring driver mutations are shown as black lines and also in b-c. **b**. Each horizontal grey box represents an individual patient from birth until the last colony sampling timepoint. The time before birth is represented on an expanded scale and is shaded pink. Within each grey box is shown the range of ages during which driver mutation and copy number aberrations are estimated to have occurred. The start and ends of each coloured box represent the median lower and upper bounds of time estimates corresponding to the start and end of the shared branches harbouring driver mutations. Thus, the upper bounds (right edge of the coloured boxes) represent the latest time by which mutation acquisition is estimated to have occurred from phylogenetic analysis. Black lines show the 95% credibility intervals for the start and end of the branches carrying the drivers. Mutation timings are inferred from a model where mutation accumulation within branches follows a Poisson distribution but were not substantially different when using a Negative Binomial model. Diamonds show age at diagnosis. **c**. Raw data from a-b is shown with 95% CI intervals around the estimated ages of the starts and ends of branches harbouring driver mutations for different patients, together with adjusted SNV counts for branches.

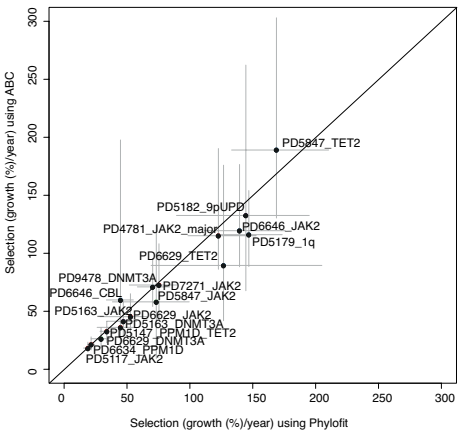
a



b

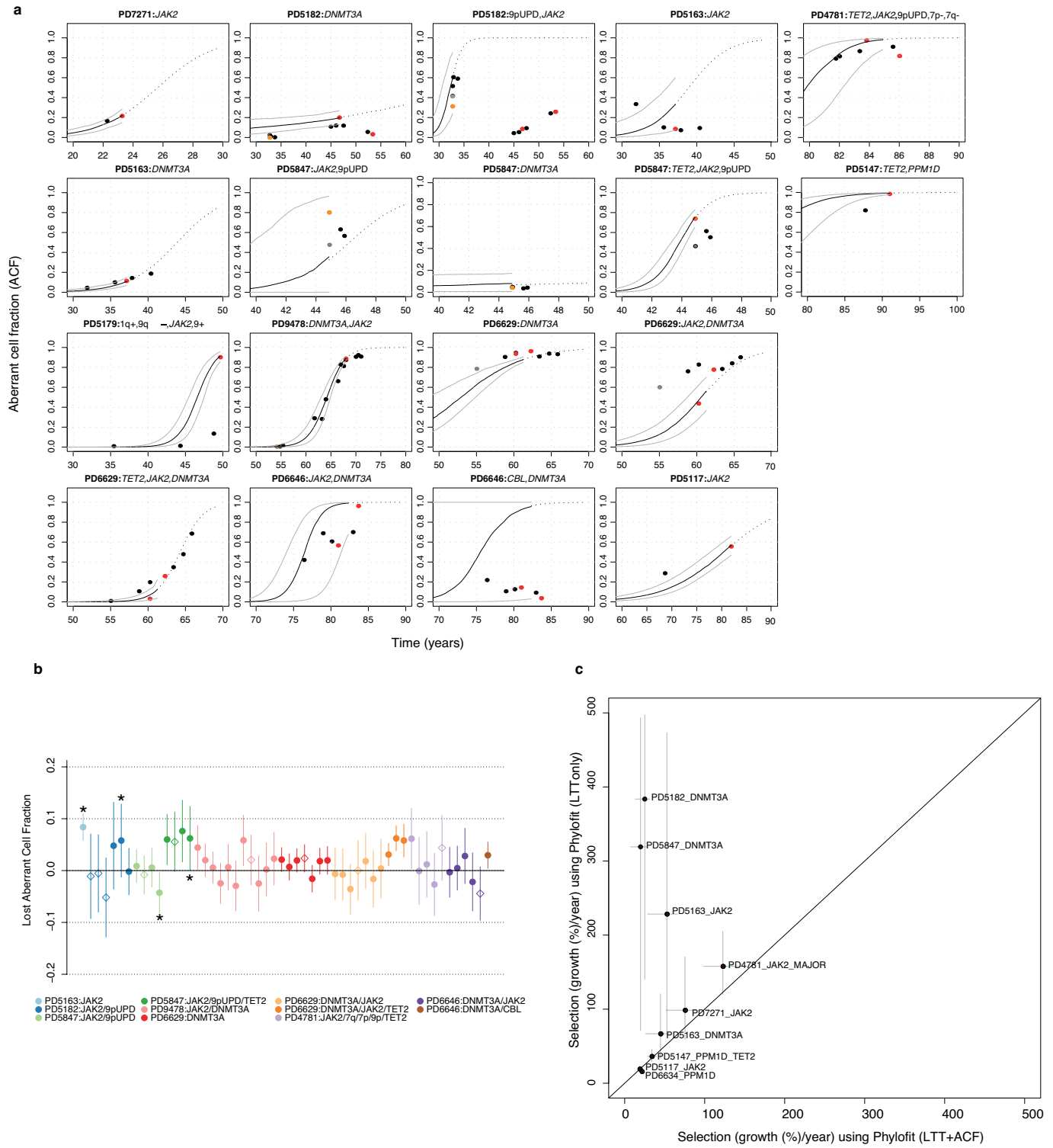
Patient	Clade	Selection (Growth %/year)	95% CI	Clone acquisition age (yrs)	95% CI	N
PD5847	JAK2 , 9pUPD, TET2	189.00	130-303	32.46	28.93-35.03	985822
PD5182	JAK2 9pUPD	133.00	68-262	17.93	12.33-20.35	995907
PD6646	DNMT3A, JAK2	119.00	88-177	59.67	55.98-63.56	996737
PD5179	JAK2 1q+, 9q-, 9+	116.00	89-154	28.02	22.99-30.53	993996
PD4781	JAK2 , 9pUPD, TET2 , 7q-, 7p-	115.00	86-190	62.83	58.58-66.29	997385
PD6629	DNMT3A , JAK2, TET2	90.00	42-176	44.98	37.72-49.93	994935
PD7271	JAK2	73.00	43-108	2.35	0.85-3.49	995243
PD9478	JAK2 , DNMT3A	71.00	54-96	42.77	35.95-47.62	996176
PD6646	DNMT3A , CBL	60.00	32-198	49.54	38.59-54.17	998059
PD5847	JAK2 , 9pUPD	58.00	9-124	22.77	6.91-27.09	994183
PD5163	JAK2	45.00	20-65	7.09	1.53-10.86	996210
PD6629	DNMT3A , JAK2	41.00	28-60	26.4	16.92-29.99	997279
PD5163	DNMT3A	36.00	20-56	5.47	0.90-9.06	996571
PD5147	TET2 , PPM1D	33.00	23-41	30.06	12.81-34.15	951366
PD6629	DNMT3A	26.00	19-36	3.62	0.72-7.59	982503
PD6634	PPM1D	21.00	17-28	3.44	0.50-7.20	982891
PD5117	JAK2	18.00	15-23	8.63	0.73-12.49	960709
PD5182	DNMT3A	5.00	2-11	0.12	0.08-0.17	999534
PD5847	DNMT3A	3.00	1-9	0.1	0.05-0.16	998193

c



Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | legend: Estimates of clonal expansion rates in patients. a. The figures shows the smoothed posterior density distribution of the selection coefficient (proportion additional growth per year) vs driver timing for all analysed clades from population simulations and approximate Bayesian computation (ABC). Marginal distributions are also shown. The prior distribution for driver timing is clade dependent and is largely determined by the mutation count at the start and end of the associated branch. Both clonal fractions and lineages-through-time were used as summary statistics in the approximate Bayesian computation for estimates of selection. Main plots show driver mutations acquired after birth, and driver mutations pre-birth are shown within the black box, taking into account driver mutation acquisition during a time when the background stem cell population size is modelled to be growing. **b**. Data from **a**. in tabular format. Here, selection coefficients have been converted to clonal expansion (median growth % per year, Selection). The ABC approach gives alternative estimates for ages of driver mutation acquisition as shown. N depicts the number of simulations per clade. Clones with sufficient immediate descendants (>5 coalescences) were included for estimates of selection. **c**. Comparison of estimates of selection of mutant clades (each labelled by patient ID and driver mutation) from ABC versus Phylofit. The grey lines show 95% credibility intervals for estimates from each approach. Correlation coefficient r = 0.96. Note, that the PD5182 and PD5847 *in-utero* DNMT3A expansions from panel **a**. are not shown because, only the ABC approach, and not Phylofit, allowed for modelling selection against a growing background population.



Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | legend: Aberrant cell fractions in bulk blood samples and validation of selection estimates. a. Plots showing aberrant cell fraction (ACF) in colonies and bulk longitudinal mature blood cell samples. Colony samples were derived from peripheral blood (red dots) or bone marrow (orange dots, in PD5182 and PD5847) mononuclear cells. Bulk mature blood cell samples comprised mostly peripheral blood granulocytes (black dots) and occasionally, bone marrow derived (grey dots) granulocytes (in PD5847, PD6629) or mononuclear cells (in PD5182), and whole blood (brown dots, in PD9478, PD6629). ACF in colonies is the clonal fraction proportion of all colonies. In bulk samples, ACF is calculated as twice the mean VAF of variants that map to the shared ancestral branch of the clone. The x-axis is patient age at sample timepoints. Lines depict the inferred ACF trajectories from the top 0.01% of simulations from approximate Bayesian computation. Black lines, median ACF; grey lines, 95% CI; dotted line, inferred future growth trajectory beyond the sampling time using the growth rate *S* and accounting for a sigmoid clonal trajectory as clonal dominance is approached. **b**. 95% confidence intervals for the difference in parent branch and aggregate descendant daughter branch ACFs from phylogenetic tree clades. Confidence intervals are calculated assuming a normal sampling distribution of aggregate mutant read fractions for each branch. Diamonds indicate those recapture samples closest to the colony sampling.* denotes interferon treatment at time of sampling. **c**. Comparison of estimates of selection coefficients for clades with single driver mutations using Phylofit fitted using the branching pattern within the tree (lineage through time, LTT) and ACF (horizontal axis), versus selection coefficients estimates using just the branching pattern of the tree (LTT) and no ACF (vertical axis) to identify clades that show early rapid branching, but smaller than expected final clonal fractions. 95% credibility intervals for selection coefficients are shown as grey lines and the corresponding median estimates as black dots. Possible early faster expansion are seen in two *in utero* mutated-*DNMT3A* clades (PD5182 and PD5847) and the *JAK2*^{V617F} clade in PD5163 prior to Interferon therapy.

nature research

Corresponding author(s): JYOTI NANGALIA

Last updated by author(s): Oct 13, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	FASTQ files were generated by Illumina X10 and Novaseq sequencing machines.
Data analysis	Single-nucleotide substitutions (SNV) were called using the CaVEMan (Cancer Variants through Expectation Maximization) algorithm, version 1.13.14 (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm as implemented in the cgpPindel workflow, version 3.2.0 (https://github.com/cancerit/cgpPindel). Copy number variants were called using the ASCAT algorithm as implemented in the ascatNgs workflow, version 3.2.0 (https://github.com/cancerit/ascatNgs). Mutational signatures analysis was performed using MutationalPatterns v1.10, available on Github (https://github.com/UMCUGenetics/MutationalPatterns) and SigProfiler (https://github.com/AlexandrovLab). Allele counts at SNV and Indel sites were carried out using vafCorrect (https://github.com/cancerit/vafCorrect). Telomere lengths were estimated using telomerecat, version 3.2 (https://github.com/cancerit/telomerecat). Mutations were mapped to phylogenetic branches using treemut (https://github.com/NickWilliamsSanger/treemut). Temporal branch lengths and per driver mutation rates were inferred using rtreefit (https://github.com/NickWilliamsSanger/rtreefit); simulation of HSC populations and phylogenies with selection were carried out using rsimpop (https://github.com/NickWilliamsSanger/rsimpop); General analysis and the estimation of tree based duplication and LOH events were carried out using custom R scripts, these are in a private Github repository (https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution) which will be made publicly accessible prior to publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Whole genome sequencing data in the form of BAM files across samples reported in this study have been deposited in the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home). The accession codes are EGAD00001007714 (Whole genome sequencing colonies) and EGAD00001007715 (Targeted recapture sequencing). All figures were generated
through the analysis of whole genome sequencing data and targeted sequencing data. Clinical data were collected from the Cambridge University Hospital Trust. Final somatic mutations called across all samples will be deposited on Mendeley.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal procedure was used to determine number of single cell derived colonies per patient. We selected greater than 50 colonies per patient to allow capture of a sufficient number of both mutant and wild-type colonies to calculate estimation of mutation burden and tree based HSC population growth parameters. Ten JAK2 positive patients were selected to broadly represent all ages and 2 patients were selected that were negative for JAK2,CALR and MPL driver mutations.
Data exclusions	Single cell derived colonies were excluded from analysis for QC reasons, or because there was evidence that they were grown from the same cell (technical replicate).
Replication	Occasional technical replicates that allowed us to confirm the variant calling and colony positioning within tree topology.
Randomization	Not applicable - this is a descriptive study, not an intervention study.
Blinding	Not relevant - the principal objectives of characterising the timing of drivers and the associated selective coefficients, mutational signatures and burdens were defined at the outset.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	Patients were selected from JAK2-mutated MPN patients attending Cambridge University NHS Trust Hospital, UK, that had undergone previous whole exome sequencing. Apart from ensuring a wide representation of ages, patients were chosen at

	random. In doing so, we found that we captured different MPN subtypes, clinical presentations varying from asymptomatic blood count abnormalities to life threatening thrombosis, different MPN therapies, stable and progressed disease, and a wide mutation spectrum. This allowed us to capture a broad cross-section of MPN.
Recruitment	All participants were enrolled in the study "Causes of Clonal Disorders Study" following fully informed and written consent, and was in line with the Declaration of Helsinki.
Ethics oversight	Cambridgeshire South REC committee, REC ID is 07/MRE05/44

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	N/A
Study protocol	N/A
Data collection	A small amount of clinical data was collected for the 10 participants using the Cambridge Univeristy NHS Trust hospital system.
Outcomes	Outcome events were as documented on the hospital electronic system by the treating physicians.