Supplementary Material for

Life histories of myeloproliferative neoplasm inferred from phylogenies

Nicholas Williams, Joe Lee, Emily Mitchell, Luiza Moore, E Joanna Baxter, James Hewinson, Kevin J Dawson, Andrew Menzies, Anna L Godfrey, Anthony R Green, Peter J Campbell, Jyoti Nangalia

Table of Contents:

Supplementary Note 1. Somatic mutation filtering

Following the identification of single nucleotide variant (SNV) using CaVEMan[1] for each colony by comparison to an unmatched normal colorectal crypt sample (PD26636b) that had previously undergone whole genome sequencing, a multi-loci pile up across all colonies was carried out for SNVs identified in individual colonies. Following this, filtering steps were designed to remove germline and artefactual variants. We defined a *Germline SNV Filter (Germline log odds, GLOD)* that was an adaptation of the Mutect germline classifier[2] to identify germline variants whilst also accounting for modest levels of potential contamination in a matched normal sample (Buccal or T-Cell). The method uses just the normal sample and compares the likelihood of the observed read counts assuming the site is a heterozygous SNP compared to the likelihood assuming it is a somatic variant:

$$LOD = \log_{10}\left(\frac{P(Data|Somatic)P(Somatic)}{P(Data|Germline)P(Germline)}\right)$$

Cibulksi et al[7], formulate an expression for the likelihood of observed read counts for a given VAF denoted by M_{VAF} , assuming the following priors: $P(Somatic) = 3 \times 10^{-6}$ and that $P(Germline)$ depends on whether the site is represented in dbSNP or not:

$$P(Germline|\mathrm{dbSNP}) = \frac{P(Germline|\text{Site in dbsnp})P(\text{Site is Germline})}{P(\text{Site in dbSNP})}$$

Giving:

$$P(Germline|\mathrm{dbSNP}) = \frac{\text{Proportion in db snp}(=0.95) \times \text{Number of Germline Mutations Per Human}(=3e6)}{\text{Number of variants in dbSNP}(=30e6)} = 0.095$$

$$P(Germline|\text{not in dbSNP}) = \frac{\text{Number of Germline Mutations Per Human not in dbSNP}(=5\%\text{of } 3e6)}{\text{Number of non dbSNP sites}(\ 3e9)} = 5 \times 10^{-5}$$

They then use the following cut-off:

$$LOD = \log_{10}\left(\frac{L(M_0)P(Somatic)}{L(M_{0.5})P(Germline)}\right) > \log_{10}(\delta)$$

Giving

$$LOD = \log_{10}\left(\frac{L(M_0)}{L(M_{0.5})}\right) > \log_{10}(\delta) + log_{10}(P(Germline)) - log_{10}(P(Somatic)) = \theta$$

This evaluates to $\theta(\text{SNP site}) = 5.5$ and $\theta(\text{Not SNP site}) = 2.2$

We modified this to allow for Tumour contamination,c, of the Normal. Then for somatic variants:

$$VAF(Normal) = c \times VAF(\ Tumour)$$

We conservatively assume a tumour VAF of 0.5 giving:

$$LOD = \log_{10}\left(\frac{L(M_{0.5\times c})P(Somatic)}{L(M_{0.5})P(Germline)}\right) > \log_{10}(\delta)$$

Finally, the read counts have a minimum phred scale quality of 30 – which our filter uses as the error probability rather than the per-read reported base qualities. Cibulksi et al[7] conservatively set $\delta = 10$ so that a variant is only called as Somatic if it is at least 10 times more likely that the variant is Somatic than that it is Germline. We define SNP sites: SNP=[DBSNP or 1000 Genomes or ExAC] and not COSMIC. This gives significantly more sites than

specified above (the union of distinct sites from dbSNP, 1000 genomes and ExAC SNP sites yields 96×10^6 sites - which we approximate as 100 million). This lowers the evidence required for a variant to be classified as Somatic at known SNP sites: $\theta(\text{SNP site}) = 5.0$. In the 1000 Genomes Phase 3[3], it is found that on average GBR samples have around 4 million variant sites per GBR genome and around 10,000 singletons - this corresponds to approximately 99.75% probability of each variant being represented in 1000 Genomes - this considerably lowers the threshold at non-germline sites. Looking at our samples we find that 20,000 (corresponding to 99.5% chance each variant is in our germline resources) is a more reasonable number of germline variants not found in our resources. This gives: $\theta(\text{Not SNP site}) = 1.35$. Following the development of GLOD, we applied the following filters to loci of somatic variants identified across any of the colonies:

-bglod: Remove loci where the GLOD filter applied to the nominated matched normal indicates that the variant is probably a germline variant.

-near_indel: Remove loci within 10 bp of an indel.

-too_close: Remove loci that are within 10 bp of each other.

-max_miss: Remove loci where more than a fifth of the colonies have depth<6.

-max_gmiss: Remove loci where more than a fifth of the colonies have missing genotype.

-count_filter: Remove loci where no samples have a genotype 1 or where all samples have a genotype of 1.

*-vaf_too_low_s: Remove loci where the total mutant read count is significantly less than 0.9*Expected VAF*total depth across all colonies that have genotype=1 (Binomial Test). This implicitly assumes that 10% false positive rate in genotyping is acceptable. The expected VAF accounts for copy number variation and is the depth weight average of the minimum mutant VAF across the mutant colonies at that locus*

*-vaf_too_low_m2: Remove loci where the total mutant read count is significantly less than 0.9*0.5*total depth across all colonies that have at least 2 mutant reads (Binomial Test). Generally, dominates the above.*

*-vaf_zg_too_noisy: Remove loci where total mutant read count at sites with genotype=0 is significantly greater than 0.01*Depth.*

The typical activity and pairwise overlap of these filters is shown below for PD5182 (Figure S1). Further QC plots are provided at https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution.

Figure S1 Activity of somatic variant filters

PD5182

SNV:Pairwise Filter Overlap (% Filtered Variants):Total Fail=39364/Pass=101598

	near_indel	too_close	max_miss	max_gmiss	count	bglod	vaf_too_low_s	vaf_too_low_m2	vaf_zg_too_noisy	manual_exclude
near_indel	**0.2**	0.0	0.0	0.1	0.0	0.1	0.1	0.1	0.1	0.0
too_close	0.0	**2.6**	0.0	0.0	0.8	0.9	0.0	0.0	0.0	0.0
max_miss	0.0	0.0	**0.7**	0.2	0.2	0.4	0.0	0.0	0.0	0.0
max_gmiss	0.1	0.0	0.2	**0.8**	0.1	0.3	0.5	0.6	0.5	0.0
count	0.0	0.8	0.2	0.1	**88.0**	86.0	0.6	1.0	0.8	0.0
bglod	0.1	0.9	0.4	0.3	86.0	**94.9**	0.3	0.5	0.5	0.0
vaf_too_low_s	0.1	0.0	0.0	0.5	0.6	0.3	**1.4**	1.4	1.2	0.0
vaf_too_low_m2	0.1	0.0	0.0	0.6	1.0	0.5	1.4	**2.4**	1.5	0.0
vaf_zg_too_noisy	0.1	0.0	0.0	0.5	0.8	0.5	1.2	1.5	**1.7**	0.0
manual_exclude	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**

Indels:Pairwise Filter Overlap (% Filtered Variants):Total Fail=7721/Pass=3624

	too_close	max_miss	max_gmiss	count	bglod	vaf_too_low_s	vaf_too_low_m2	vaf_zg_too_noisy	manual_exclude
too_close	**6.3**	5.1	2.5	3.4	5.1	0.2	0.4	0.4	0.0
max_miss	5.1	**31.5**	13.5	16.0	27.4	0.0	0.2	0.3	0.0
max_gmiss	2.5	13.5	**19.3**	2.3	11.9	3.1	5.3	4.8	0.0
count	3.4	16.0	2.3	**56.5**	53.0	0.3	3.0	1.9	0.0
bglod	5.1	27.4	11.9	53.0	**76.7**	1.5	2.4	2.5	0.0
vaf_too_low_s	0.2	0.0	3.1	0.3	1.5	**4.7**	4.7	4.5	0.0
vaf_too_low_m2	0.4	0.2	5.3	3.0	2.4	4.7	**21.2**	15.6	0.0
vaf_zg_too_noisy	0.4	0.3	4.8	1.9	2.5	4.5	15.6	**16.3**	0.0
manual_exclude	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.0**

Figure S1 legend: The diagonal entries show the percentage of filtered candidate variants removed by each filter and the off-diagonal elements show the number of filtered variants removed by both the corresponding filters. The germline filter "bglod" is most active filtering out 94.9% of candidate somatic SNVs in this example.

Supplementary Note 2. Construction and quality assessment of phylogenetic trees

We developed an Expectation Maximisation method (R package "treemut") to soft assign mutations to trees and estimate branch length (https://github.com/NickWilliamsSanger/treemut). This estimates the maximum likelihood probability that each mutation belongs to each branch. Following this, mutations were hard assigned

to branches. We assessed for biases in branch length estimations versus true branch lengths from simulated trees to assess the performance of Rtreemut, as well as assessing the impact of hard assigning mutations to the tree. Simulations indicated that our 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 (Figure S2).

Figure S2 Assessment of accuracy of mutation assignment to branches

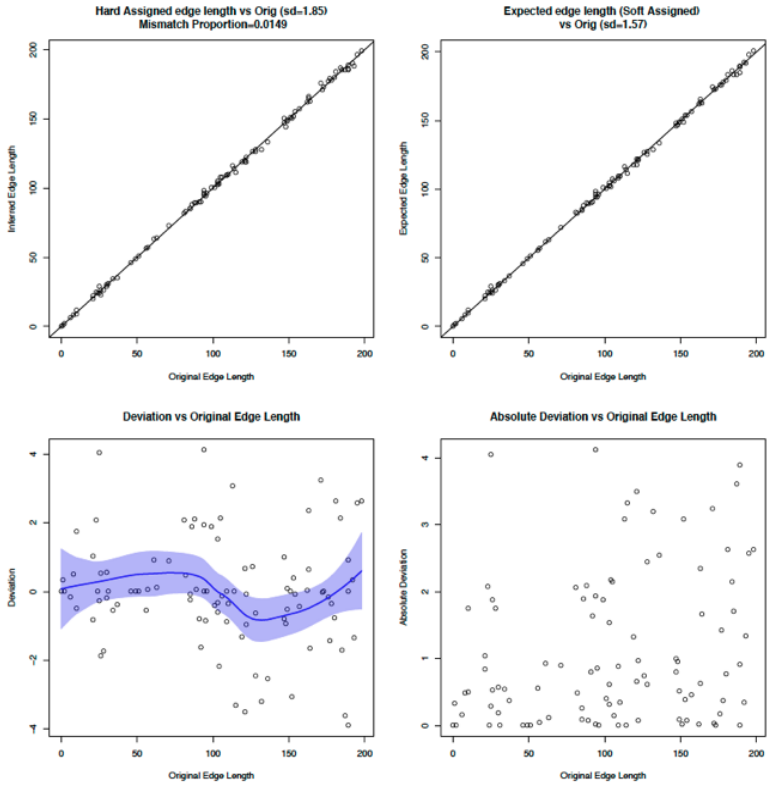


Figure S2 legend: The figure illustrates the accuracy of branch length estimation vs true branch length for a simulated tree with 100 colonies. The graphs in the top row illustrate that using hard assigning rather soft assigning of mutations has minimal effect on the edge length reconstruction. The graphs on the bottom row illustrate that there are no obvious systematic biases in the reconstruction error of short vs long branches.

The length of private branches of low depth colonies will be underestimated because of the limited sensitivity of variant calling. For fully clonal colonies the per colony sensitivity was directly estimated by measuring the proportion of germline sites that were called by CaVEMan, $P_{germline}$. This is a well powered estimate as there are ~ 50,000 such sites for each patient. The proportion of identified germline variants increases with sequencing depth (Figure S3) and can be used to normalise colonies for their sensitivity of somatic mutation detection.

Figure S3 Sensitivity of detection of germline variants per colony

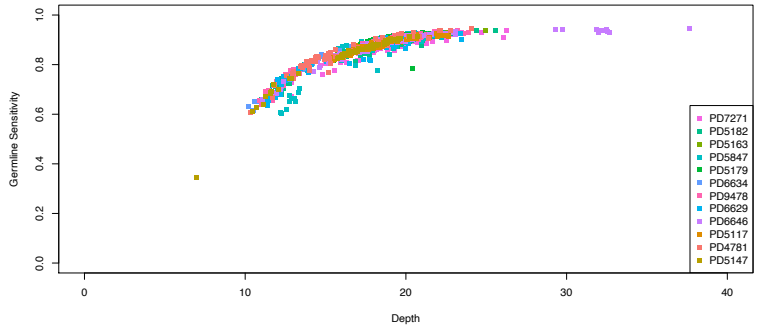


Figure S3 legend: The sensitivity for detecting germline variants (ie. proportion of identified germline variants) versus depth of sequencing per colony.

However, some colonies had a VAF slightly below the expected VAF of a germline heterozygous SNP (0.5) (Extended Data 1b). It will also be noted that the germline sensitivity does not quite asymptote to 1 as depth increases (Figure S3). This is because the CavEMan filtering process will occasionally filter out true variants. We termed this probability the Filter False Positive Rate (FFPR). We modelled VAF dependent sensitivity as follows:

$$P(Calling\ SNV|VAF, depth) = \frac{1 - FFPR}{1 + \exp\{-(a + b \times depth \times VAF)\}}$$

We restrict attention to those branches that are highly shared ($N \geq 5$), so we have a truth set of high confidence somatic mutations, and with a sufficient number of mutations ($M \geq 50$) so that we can accurately estimate the VAF of each colony that shares the branch. This model was fitted using maximum likelihood estimation using the "bbmle" package. To ensure that the sparsely represented low VAF*depth and high VAF*depth branches were well fitted, we down-sampled the over-represented branches in the mid-range of VAF*depth. Similar results were also obtained using weighted asymptotic regression with the "nnls" R package. The non-parametric germline sensitivity approach best captured the idiosyncratic sensitivity of each sample, so we combined the approaches by scaling the germline sensitivity, so that it is unchanged for VAF=0.5 colonies, but modified appropriately for lower VAF colonies:

$$Sensitivity = P_{germline} \frac{P(Calling\ SNV|VAF, depth)}{P(Calling\ SNV|VAF = 0.5, depth)}$$

Private branch lengths were scaled by 1/sensitivity. For shared branches, for the parametric model the most tractable approach was to estimate the probability of there being at least one called sample in the clade:

$$Sensitivity_{para}^{(shared)} = 1 - \prod_{i=1}^{N}(1 - Sensitivity_i)$$

Whereas for the germline based non-parametric approach, $Sensitivity_{germline}^{(shared)}$, the sensitivity was directly estimated as the proportion of germline sites in which at least one of the colonies that share the branch has the variant called. The shared branch sensitivity is then estimated as:

$$Sensitivity^{(shared)} = \frac{1 - \prod_{i=1}^{N}\left(1 - P(Calling\ SNV|VAF_i, depth_i)\right)}{1 - \prod_{i=1}^{N}\left(1 - P(Calling\ SNV|VAF = 0.5, depth_i)\right)} Sensitivity_{germline}^{(shared)}$$

Figure S4 Sensitivity correction for mutation burden based on depth of sequencing and VAF

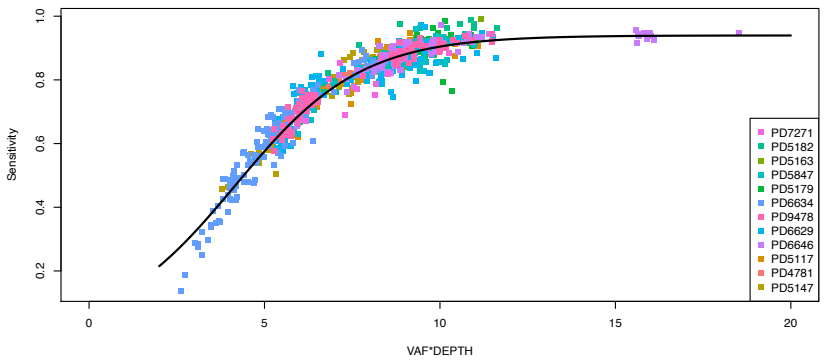


Figure S4 legend: The parametric model for estimating VAF and depth dependent sensitivity is shown. This sensitivity was used to adjust branch lengths in the phylogenetic tree.

Supplementary note 3. Quality assessment of trees and colony filtering

All phylogenetic trees underwent detailed assessment for quality in three subsequent steps:

(a) Assessment of tree topology, branch length and SNV VAF distribution in phylogenetic trees. QC plots were generated to allow visualisation of the quality of tree construction, branch length estimations and assessment for any biases in SNV or copy number aberration calling that might inadvertently affect the tree. QC plots are provided at https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

(b) Tree-by-Colony quality assessment. We next assessed the VAF distribution of the variants assigned to each branch on a per colony basis, creating a VAF plot along the phylogenetic tree on a per colony basis. This allows one to 'walk through' the trees on a per colony basis to visualise both the branch placement and VAF of all the variants present in that single colony with respect to the rest of the tree. This is particularly helpful to ensure

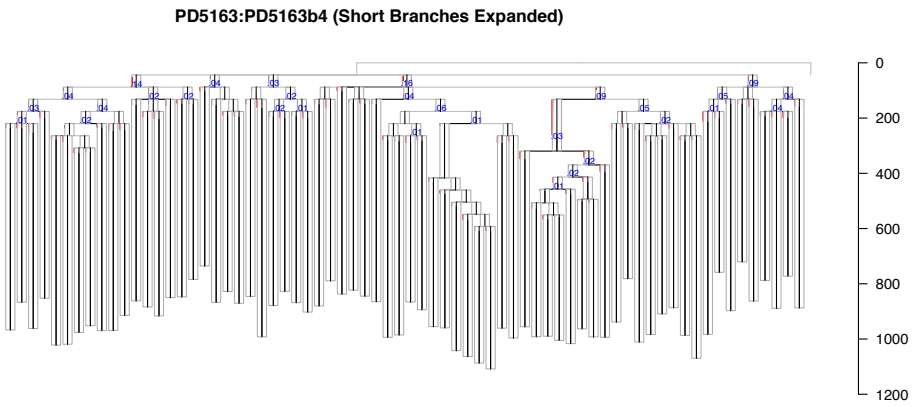
that variants belonging to a single colony are not found in non-ancestral branches whilst also allowing one to assess if other branches in the tree suffer from a lack of sensitivity for picking up specific variants. We performed this assessment for all colonies across all trees. Individual tree-by-colony assessments are provided at https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

(c) Screening for true somatic variants shared between the phylogenetic tree and the germline sample that represent Tumour-in-Normal (TiN) or embryonic variants. Our approach to calling somatic SNVs used first an unmatched analysis (Methods), followed by a matched analysis assessment and further filtering (Supplementary Note 1). This approach also allowed us to identify and rescue true somatic embryonic mutations that occurred prior to the separation of the germline sample tissue and which can be present in other tissues, such as buccal cells or T-cells used as the source of germline material. Indeed, Figure S5c shows that the mutations near the top of the tree are invariably present in buccal epithelium and T-cells. The presence of these mutations in the matched germline sample may also represent contamination of the germline samples with blood cells (TiN) and therefore, it was important not to filter these mutations out as this would result in false shortening of early branches in the phylogenetic tree. Therefore, for all mutations designated as germline by GLOD, we systematically mapped these to the trees to assess whether they were compatible with the tree topology and likely to represent early true somatic mutations that were either embryonic in origin or represented TiN variants. We provide the VAF in the germline sample (highlighted in red) of all somatic mutations in the trees at https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

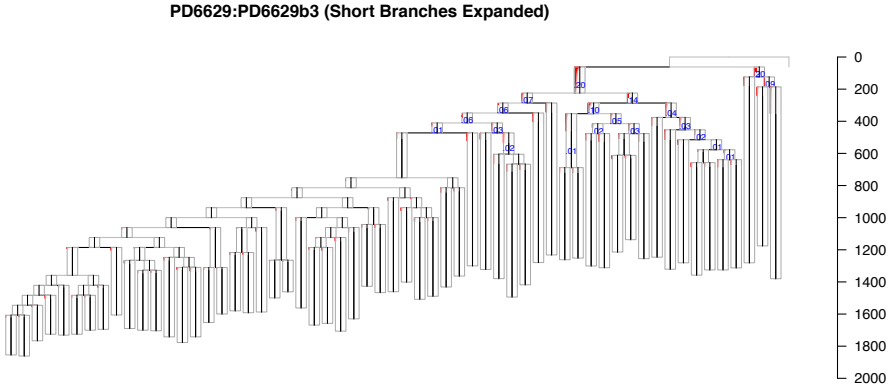
An example from the above file is shown in Supplementary Figure S5. The top panel (A) shows a tree with only modest levels of contamination where genuine somatic mutations present in the germline sample are not falsely filtered out. For some patients with early, highly shared branches, such as PD6629 (B), there was >10% contamination of the germline with somatic mutations from early branches in the phylogenetic tree. Indeed, due to the high level of tumour-in-normal contamination, 40%, exhibited in PD6629, we interrogated the ~31,000 substitutions that were designated germline by the GLOD filter and mapped them to the tree topology. We found that 35 mutations mapped directly to the founding *DNMT3A* shared branch with high confidence and rescued these mutations back to the shared mutant branch. This assessment was performed for all trees.

Figure S5 Variant allele fractions of mutations near the top of the trees in bulk DNA samples

A



B



C

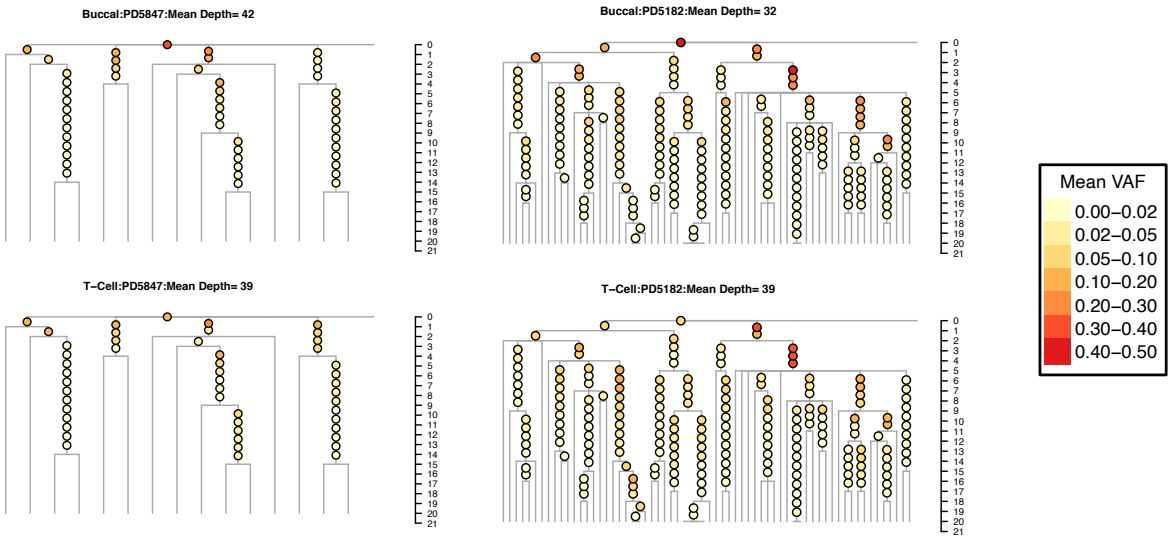


Figure S5 legend: A-B. Phylogenetic trees of two patients where the tops of the trees have been expanded for visualisation. On each branch, the VAF of the somatic mutations in the matched germline sample is indicated in red. On each branch, the VAF of the somatic mutations in the matched germline sample is indicated in red. Branches where the average VAF exceeds 1% are annotated with average VAF in blue. PD5163 (**A**) shows the typical case with modest contamination in the germline. In PD6629 (**B**), there was an early driver mutation acquired and clonal expansion combined with only a few wildtype lineages detected. As a result, variants from earlier branches in the tree are detectable in the germline at >10% VAF. **C.** The top segments (up to 20 mutations of molecular time) of phylogenetic trees from two patients (PD5847, left; PD5182, right). Yellow to red shading shows the corresponding variant allele fraction (VAF) in buccal DNA (top plots) or T-cells (bottom plots) as labelled. The mean depth of sequencing for buccal and T-cell samples is shown in the tree labels.

A summary of the colonies that were removed per patient is shown in Figure S6 and Supplementary Table 1.

Supplementary Table 1. Number of colonies per patient sequenced and taken forward for analysis

	Pass Colonies				Auto QC Fail				Tree QC Fail					
Patient	VAF	Depth	Sens	N	VAF	Depth	Sens	N	VAF	Depth	Sens	N	Pass %	N
PD7271	0.47	19	0.87	88	0.31	18	0.88	6				0	94	94
PD5163	0.50	19	0.89	70				0	0.50	20	0.90	1	99	71
PD5182	0.50	18	0.86	160	0.49	8	0.40	2	0.51	19	0.89	1	99	162
PD5179	0.50	19	0.90	92				0	0.39	18	0.88	1	99	93
PD5847	0.51	18	0.86	96	0.59	10	0.47	17	0.53	15	0.74	2	83	115
PD9478	0.51	17	0.84	80				0	0.54	12	0.76	1	99	81
PD6629	0.51	16	0.81	59				0	0.50	15	0.83	2	97	61
PD5117	0.46	18	0.88	88	0.32	17	0.87	3				0	97	91
PD4781	0.50	16	0.84	61	0.54	9	0.55	2	0.48	13	0.76	1	95	64
PD6646	0.50	19	0.87	117	0.57	9	0.51	1	0.42	18	0.87	2	98	120
PD6634	0.43	14	0.80	36	0.45	10	0.50	3	0.38	14	0.78	3	86	42
PD5147	0.47	17	0.83	66	0.55	8	0.44	23	0.45	11	0.68	1	73	90
N			Passed:	1013								Sequenced:		1084

VAF represents median variant allele fractions of all passed autosomal somatic single nucleotide variant (SNV)loci that are not in regions of copy number aberrations. Sensitivity represents the percentage of known germline SNVs identified by CaVEMan when variant calling was undertaken without a matched germline tissue sample. Depth is calculated as the median coverage across somatic variants.

Figure S6 Colony quality control

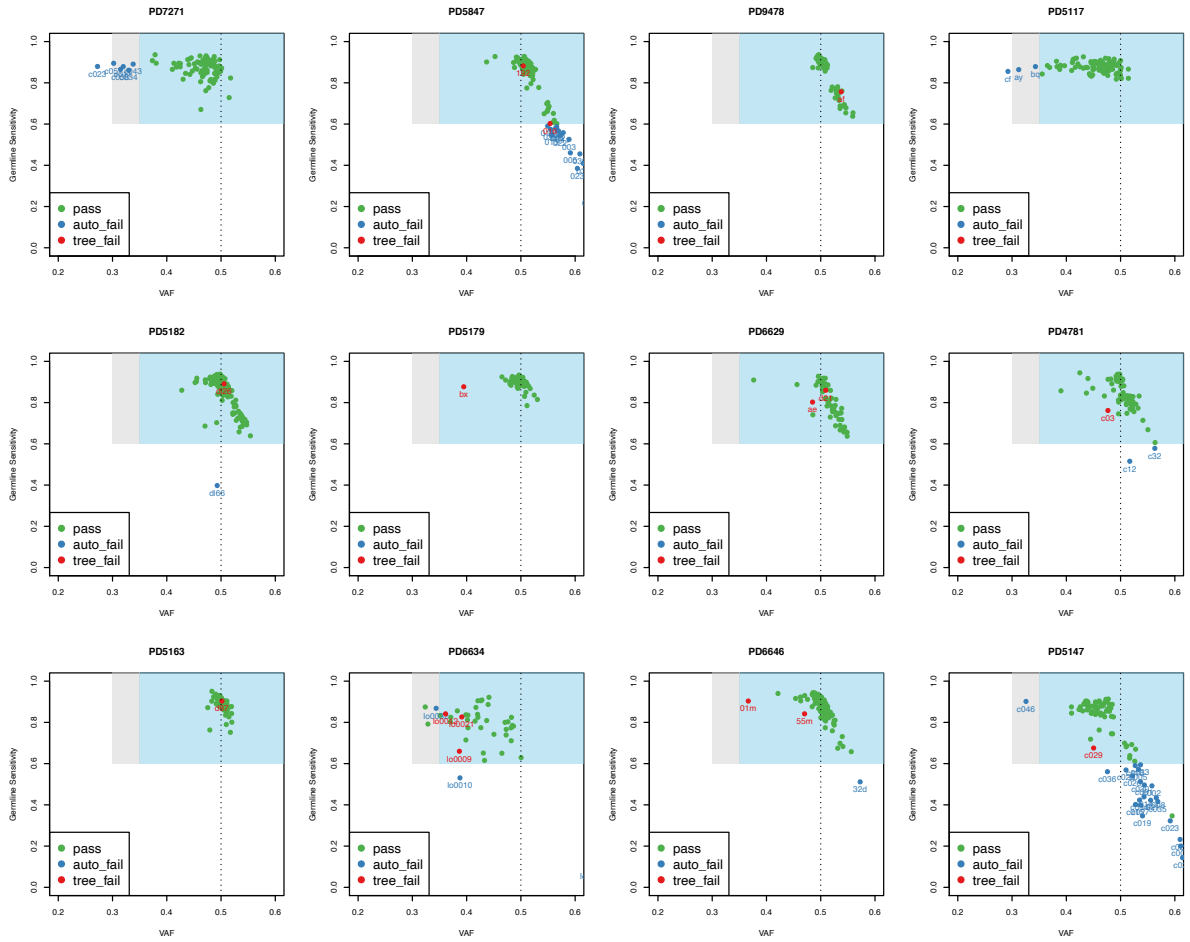


Figure S6 legend. Samples in the blue region pass auto-qc (requiring a sensitivity of calling SNVs of >60%), and where samples pass auto-qc in the grey area this is because the sample's aggregate VAF is not *significantly* less than 0.35 (corresponding to our requirement of a minimum of 70% clonality). The samples highlighted in red were removed during tree QC because they exhibited a degree of cross-contamination with other samples as evidenced by a) exhibiting a higher than expected VAF in other branches not ancestral to the sample or b) exhibiting a lower than expected VAF on any branches that are ancestral to the sample. In addition, samples were removed if they exhibited very close relatedness to another sample and where the segregating mutations had significantly lower depth for one of samples. Such pairs of samples are likely the result of the picking of two colonies as separate samples that were in fact derived from one single cell.

Supplementary Note 4. Interpreting coalescences in a phylogenetic tree

After the first few months of life, the wildtype clades are observed to be entirely independent of each other, as evident by the long terminal branches of the tree. This has previously been shown to be the case in a study of one healthy individual, by Lee-Six et al Nature 2018[4] and has also been directly demonstrated during early human development from phylogenetic trees of haematopoietic lineages in 8 and 16 week old foetuses[5]. Our data are consistent with these studies, and extends it to several more individuals. We see a similar pattern in 'clonal bursts' downstream of a branch harbouring a driver mutation, with an initial high frequency of coalescences that reduces as one moves down the tree.

A branch split (or 'coalescence') within the tree reflects an HSC symmetrically dividing into two long lived daughter HSCs (as each of their progeny has been detected many years later at sampling). Coalescences are more frequent at the top of the tree and at the start of a 'clonal burst' but seem to disappear as one moves down the phylogenetic tree, or down a 'clonal burst'. This is a reflection of the growth of the underlying HSC population and the overall HSC population size compared to what we have sampled. We illustrate this with a few examples. Let's say we were to sample all HSC cells within the patient and draw a phylogenetic tree. In this situation, many wild type HSCs would have an immediate sister cell with a most recent common ancestor dating back one self-

renewal division (and thus, branches near the bottom of the tree). However, because in some cases we sample very few wild type cells from a population of the order of 50,000-200,000 HSCs (eg. ~100 wildtype lineages for PD5182, and less for other patients), then the most recent common ancestor between any two sampled cells tends to date back to when the HSCs population was rapidly growing during embryogenesis. Therefore, the lack of branching in the tree within the wildtype compartment beyond embryogenesis reflects the already large HSC population size in relation to the number of sampled colonies. The fact that the wildtype colonies have long end branches is testament to the fact that cell divisions (and therefore, mutations) have occurred, but that the other progeny of such divisions have not been captured.

We now demonstrate this using first real data and then simulated data. If we randomly sub-sample 7 colonies from Lee-Six et al, Nature 2018[4], then we generally observe no late coalescences (Figure S7) despite repeatedly sub-sampling 7 colonies.

Figure S7 Coalescence capture and number of colonies sampled (real data)

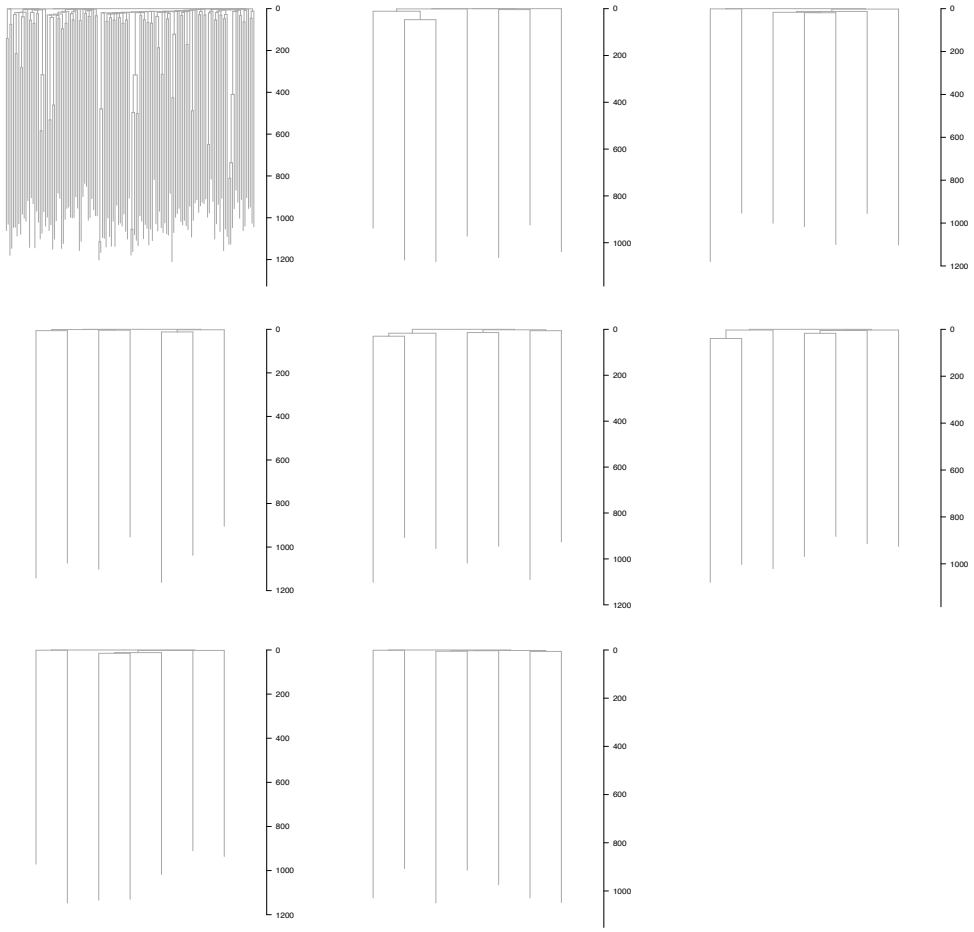


Figure S7 legend: Top left panel shows the phylogenetic tree of the haematopoietic compartment from a 59 year old male from Lee-Six et al, Nature 2018[4]. In subsequent trees, we repeatedly subsample 7 colonies from the original phylogenetic tree (top left) and reconstruct new phylogenetic trees. Subsampling shows that there are no late coalescences captured even though many of these coalescences can be seen in the original tree.

In the above real data example, we do not know the full lineage tree of all extant cells at the time of sampling, but for simulated data we would have this information. To facilitate visualisation, we simulate a modest full population size of 10,000 cells with a cell division rate of 1 per year for 40 years, and then from this full population we sub-sample 10 randomly selected cells. We can plot the resulting full population phylogeny and superimpose the lineage history of the 10 randomly selected cells (Figure S8).

Figure S8 Coalescence capture and number of colonies sampled (simulated data)

Figure S8 legend: The top panel shows a 10,000 cell lineage tree where the lineages of 10 randomly selected cells have been traced on the tree as highlighted in red. All the coalescences (self-renewal divisions) of the 10 selected cells are highlighted by red dots. The second panel shows the phylogeny of those 10 selected sampled cells where these self-renewal divisions are not seen, but we know to be present from the top panel. A background population of just 1000 cells gives a very different picture (third and fourth panels). Third panel shows a 1,000 cell lineage tree where the lineages of 10 randomly selected cells have been traced on the tree as highlighted in red. All the coalescences (self-renewal divisions) of the 10 selected cells are highlighted by red dots. The fourth panel shows the phylogeny of only those 10 sampled cells. Coalescences are captured because the background population size was smaller.

Supplementary Note 5. Mutation accumulation over age in wildtype colonies and following driver mutations

In order to estimate the wildtype mutation rate, we used two methods. First, the ultrametric time-based trees generated using Rtreefit were created for individual patients (Methods). The model incorporates an excess mutation rate in early life that will add 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 are 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 then combined in a random effects meta-analysis using the R package metafor. This analysis indicated that the per patient wild type mutation rate in mutations per year is drawn from a distribution with mean 18.4 (95% CI 17.7-19.2) and variance 0.97 (Extended Data Fig. 3a).

To orthogonally validate this rate of mutation acquisition in wildtype cells over age, we also applied mixed effects modelling 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. For each patient, we include the timepoint with the most wild-type colonies and we exclude timepoints with <3 wildtype colonies. Using the nlme package the model was specified in R as: lme(nsnv_count ~ age_at_sample_pcy , random=~-1+age_at_sample_pcy|patient)

This fitted model had an intercept of 95.6 (95% CI 17.7 – 173.6), and the per patient wild type mutation rate in mutations per year is drawn from a distribution with mean 17.0 (95% CI 14.8-19.2) and a variance of 0.97 in keeping with more rapid mutation acquisition in early life, reassuringly consistent with the value estimated using the time-based ultrametric trees (Extended Date Fig. 3a). Furthermore, both estimates of the rate of SNV acquisition in wildtype colonies are also in line with those reported recently from haematopoietic colonies using single molecule ultra-low error duplex sequencing[6].

In Figure S9 below, we compare the mutation burden correlation with age for the raw SNV counts per colony, for sensitivity-adjusted mutation burden, and for branch length adjusted mutation burden (see Methods).

Figure S9 Mutation burden versus age for adjusted and unadjusted SNV calls

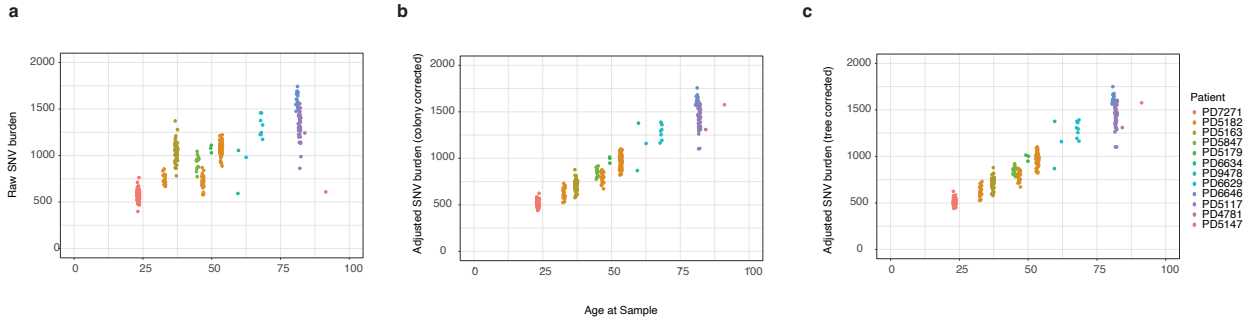


Figure S9 legend: (A) shows the number of raw non-germline CaVEMan SNV calls per colony and patient age. Regions of CNA/LOH are masked and the count has been rescaled to account for the masked proportion of the genome. This graph does not take into account the variability in mutation detection due to colony clonality (as judged by VAF) and depth of sequencing. (B) shows the single sample sensitivity-adjusted mutation burden vs age which shows a strong linear relationship between age and mutation burden. (C) shows the tree based adjusted SNV burden, where a colony's mutation burden is the sum of the sensitivity-adjusted lengths of its ancestral branches. Note the similarity between methods used in (B) and (C) to correct SNV burden.

Figure S9 and Figure 4c shows that whilst mutations accrue steadily across life at a constant rate (~17-18 mutations/ year), there is some patient to patient variability in mutation rate across the cohort. In addition, the phylogenetic trees in Figures 2-4 show that for some patients, mutation burdens in mutant clades may be modestly different to those in wildtype or other clades. Therefore, we modelled patient-specific and clade-specific mutation rates in converting phylogenetic trees to time-based ultrametric trees (Methods). We modelled mutation accumulation using a Poisson distribution, however, results assuming a Negative Binomial distribution were very similar. Several patients in the cohort showed evidence of more mutations in mutant clades compared with wildtype lineages (Extended Data Fig. 3a). Given that comparisons of mutation rates between genotypes relies on assumptions of the underlying distribution of mutation acquisition, we also explored non-parametric testing. The difficulty encountered with non-parametric testing is that the mutation burdens across the colonies in a patient are not independent measures due to their shared ancestry. The limma package's rankSumTestWithCorrelation corrects for non-independence of the samples using a single estimate of the average correlation between samples. A caveat of this approach is that only a single timepoint can be taken into account, and inter-sample correlation can be corrected within only one group. Nevertheless, the trend to increased mutation burden in mutant clades is observed again (Extended Data Fig. 3b), but it is now only nominally significant. Upon multiple hypothesis correction, only the *DNMT3A* clade in PD5163 demonstrates significantly increased mutation burden but note this clade has only 8 colonies. Therefore, overall, we were insufficiently powered to detect a consistent difference in mutation rates between wildtype and mutant colonies in patients with MPN. Per patient analysis of branch timings, mutation rates, copy-number timings and branch VAF distributions for targeted follow-up samples are available in the reports subdirectory of:
https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

Supplementary Note 6. Mutational signature assessment

To identify and compare mutational processes across the genotypes, within patient clades were grouped and treated as individual samples for de-novo signature analysis (Figure S10). No novel signatures were discovered over the standard PCAWG 60 signatures[7]. The identified signatures were SBS1, SBS5, SBS9, SBS19, SBS23, SBS32 and SBS40, and a reduced signature set (SBS1, SBS5, SB19 and SBS32) were taken forward (Figure S11).

Figure S10 Defined clades in phylogenetic trees for mutation signature analysis

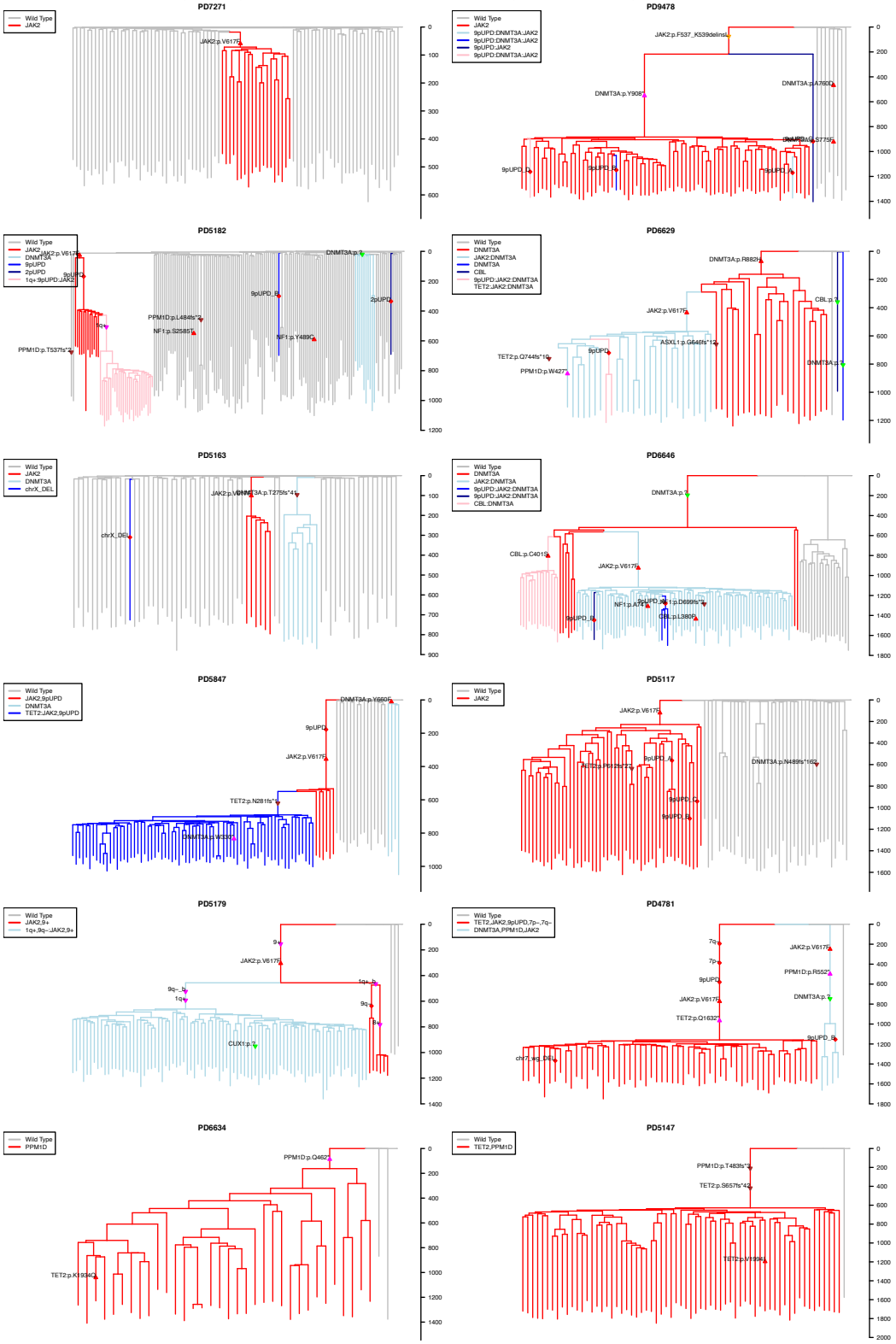


Figure S10 legend: The within-patient distinct branch colourings illustrate the per patient groupings of branches that were treated as individual samples for a de-novo signature analysis.

Figure S11 Mutation signatures

A

B

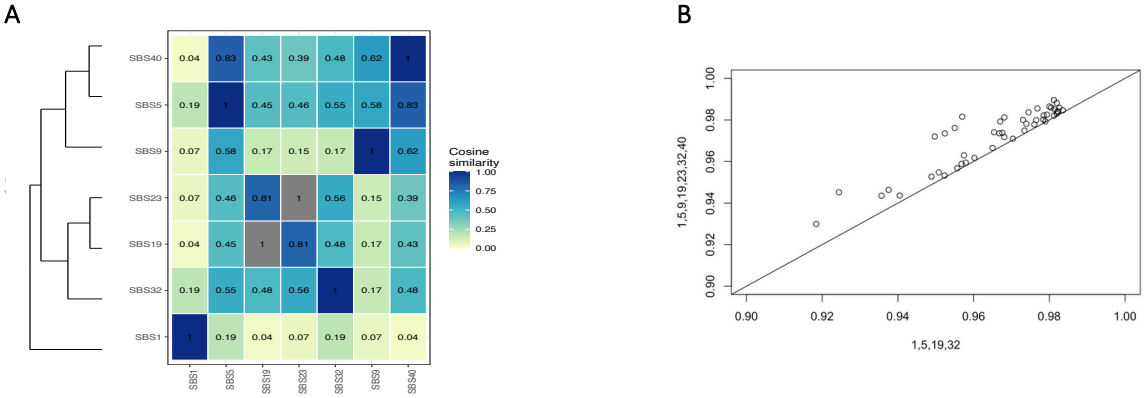


Figure S11 legend: A. Comparison of a reduced signature set (SBS1, SBS5, SB19 and SBS32) versus the set (SBS1, SBS5, SBS9, SBS19, SBS23, SBS32 and SBS40). Pair SB19 and SB23 had a high cosine similarity (0.81) as did SBS5 and SBS40 (0.83) as shown in the left panel. **B.** Removal of SBS9, SBS23 and SBS40 resulted in an acceptable loss in reconstruction accuracy (mean cosine similarity 0.970 vs 0.975) as shown on the right.

Mutation signature analysis and workflow is provided at:
https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

Supplementary Note 7. Phylogenetically aware telomere analysis

Following the observation that *JAK2*-mutant colonies had significantly shorter telomere lengths than wild type colonies (Figure 4d), we wished to control for the fact that *JAK2*-mutant colonies have a more recent shared ancestor and so measures across *JAK2*-mutant colonies within an individual patient are not independent measures of telomere length. Therefore, we quantified and adjusted for the phylogenetic distance between colonies. We defined 'sharedness', that captures the degree of shared history as a weighted average of the proportion of sampled clones that share each mutation. Extended Data Fig. 4c shows that telomere length tends to reduce across all phylogenetic trees, irrespective of genotype, with increased phylogenetic 'sharedness'. To specifically ask how much shorter telomere lengths become as a result of a *JAK2*-mutation, we fitted 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=1000). We found a mean telomere length reduction of 829bp (95% CI -663 to -969bp, p<0.001) as a result of a *JAK2* mutation.

We hypothesised that such telomere attrition could be due to increased cell divisions required for clonal expansion. This relationship between 'sharedness' and telomere length would also suggest that telomere lengths are heritable and therefore, an *in vivo* phenomenon. To formally assess the heritability of telomere length from the phylogenetic trees, we asked if the telomere lengths of more closely related colonies were more similar than those of more distantly related colonies. We formally assessed such heritability of telomere length using two well established metrics, Pagel's Lambda and Blomberg's K, that essentially measure whether the observed covariance in a trait is in line with the expected covariance based on the phylogenetic relationship (both measures are expected to equal 1 if the trait exhibits the expected covariance). We assessed significance using phytool's phylosig function. Extended Data Fig. 4d shows that the lambda values are all in the vicinity of 1 or above and are significantly non-zero for all patients except for PD5147 where power is limited because there is little difference in sharedness in the mutant colonies. This strongly supports the heritability of telomeres as an *in vivo* trait within the phylogenetic trees. This also supports the conclusion that telomere attrition in *JAK2*-mutated colonies occurs *in vivo* although we cannot exclude the possibility that *in vitro* culture would also additionally impact on telomere lengths differentially between *JAK2* and wildtype genotypes.

We hypothesised that if excessive telomere attrition was occurring due to clonal expansion, then the degree of telomere shortening could reflect the number of additional HSC symmetric divisions that had occurred for clonal expansion. We estimated the number of additional divisions from the terminal aberrant cell fraction assuming a simple exponential growth model and an HSC population size of 100,000. When a single HSC acquires a driver mutation and starts to clonally expand, it initially represents a small proportion of the HSC population: Initial aberrant cell fraction=$1/N_{HSC}$. Every additional symmetric division in the mutant population relative to wild type cell population average increases the mutant population by a factor of 2 and so the final aberrant cell

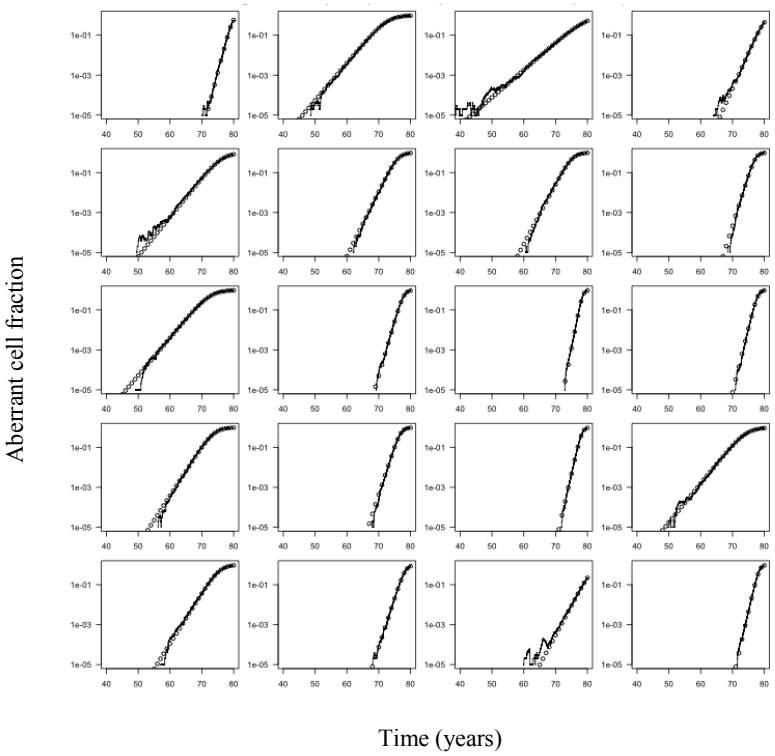
fraction=$2^{Nextra}/N_{HSC}$. Therefore, the number of additional cell divisions ~ $log2(N_{HSC} \times$aberrant cell fraction). A fixed effect regression analysis, accounting for phylogenetic relatedness of colonies, was combined in a random effect meta-analysis for a cohort wide estimate. We examined *JAK2*-mutated patients with > 2 wild-type colonies and excluded PD6646, since in this case, the wild-type colonies had themselves undergone a clonal expansion. We estimated that there is telomere loss of 57bp (SE 8.6, p<0.001) per additional HSC symmetric cell division (Extended Data Fig. 4e), compatible with that reported previously[8]. Telomere analysis including code and workflow is provided in:
https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

Supplementary Note 8. Assessment of rsimpop and approximate Bayesian computation

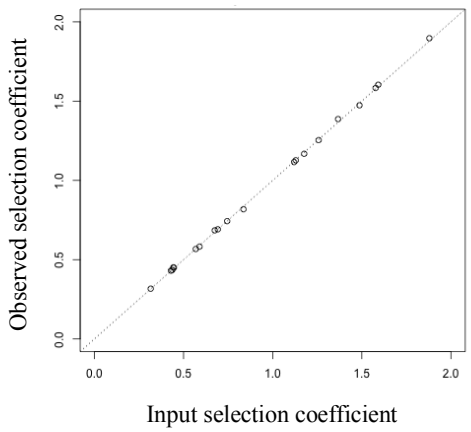
We assessed the performance of simulations run by rsimpop in three areas: (i) the longer term deterministic phase of clonal expansion, (ii) the implementation of selection, and (iii) the short-term stochastic behaviour of cells with driver mutations. For (i), we ran 20 rsimpop simulations for 80 years of life with a population size of 100,000, wild-type cell division rate of one per year, and drivers introduced with selection coefficients uniformly sampled between 0.1 and 2. The population level aberrant cell count trajectories were recorded by the simulator and a logistic curve was fitted to the data using nonlinear regression via the "nls" function in R. So as to capture just the deterministic phase, only timepoints subsequent to initially reaching a total aberrant cell count of 10,000 were included in the regression. Figure S12A shows that the fitted aberrant cell fraction trajectory for the same simulations closely track the actual trajectories. For (ii), Figure S12B shows that the measured selection coefficients closely track the input selection coefficients provided to the simulator. For (iii), we recorded the number of attempts of driver mutation introduction across a total of ~13 million HSC simulations undertaken during the approximate Bayesian computation (ABC) analysis. Simulations with driver acquisition >1 year post conception were binned into selection coefficient and total HSC population size bins. The empirical distribution of the number of driver mutation introduction attempts in each bin was converted into a bin specific maximum likelihood probability of extinction. As shown in Figure S12C, we found that the probability of stochastic extinction of driver mutations precisely tracks the probability expected by theory[9], which provides reassurance that the expected short term stochastic behaviour of clones is correctly simulated. Overall, we assessed and confirmed that the short-term stochastic behaviour, the implementation of selection, and the longer term deterministic phase of growth, all behave as expected from our model, and these fit the real data well. A RMarkdown document that shows how to run the ABC for any particular clade of interest in order to estimate the selection coefficient with confidence intervals and the timing of acquisition of the driver mutation is provided in https://github.com/NickWilliamsSanger/mpn_phylogenies_and_evolution

Figure S12 Assessment of performance of simulations and ABC

A



B



C

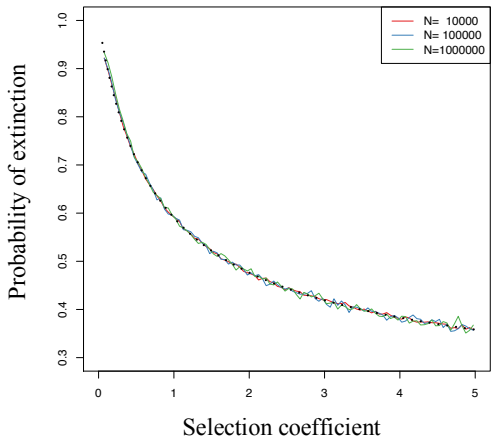


Figure S12 legend: (A) The observed aberrant cell fraction trajectory (lines) overlaid on to the fitted trajectory. The plots show that the asymptotic form of the simulated aberrant cell fraction is consistent with that expected from theory. **(B)** The observed selection coefficient vs the rsimpop input selection coefficient. The y-axis coefficient is inferred using non-linear regression and the dashed line shows the expected relationship between this simulation input and simulation output. **(C)** The graph depicts the relationship between the selection coefficient (or fitness) of a driver mutation harbouring HSC and the likelihood of stochastic extinction after acquisition of the driver mutation. The theoretical extinction probability (dotted line) is overlaid on the chart. S, the selection coefficient (that is, the proportional increase in clone size per year) is modelled as an increase in the rate of symmetrical HSC cell division due to a driver mutation.

References

1. 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).
2. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
3. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
4. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
5. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
6. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* 1–6 (2021) doi:10.1038/s41586-021-03477-4.
7. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
8. Vaziri, H. *et al.* Evidence for a mitotic clock in human hematopoietic stem cells: loss of telomeric DNA with age. *Proc. Natl. Acad. Sci. U. S. A.* **91**, 9857–9860 (1994).
9. Tavaré, S. The linear birthdeath process: An inferential retrospective. *Adv. Appl. Probab.* **50**, 253–269 (2018).