Prolonged persistence of mutagenic DNA lesions in stem cells

Peter Campbell (pc8@sanger.ac.uk)
Wellcome Trust Sanger Institute https://orcid.org/0000-0002-3921-0510

Michael Spencer Chapman
Wellcome Trust Sanger Institute https://orcid.org/0000-0002-5320-8193

Emily Mitchell
Wellcome Trust Sanger Institute

Kenichi Yoshida
The Cancer, Ageing and Somatic Mutation Programme, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire CB10 1SA,

Nicholas Williams
Wellcome Trust Sanger Institute https://orcid.org/0000-0003-3989-9167

Margarete Fabre
Wellcome Sanger Institute

Anna Maria Ranzoni
Nature Research

Philip Robinson
Wellcome Sanger Institute https://orcid.org/0000-0002-6237-7159

Christian Matthias Wilk
University of Zurich and University Hospital Zurich https://orcid.org/0000-0001-5429-7143

Steffen Boettcher
University Hospital Zurich https://orcid.org/0000-0001-9937-0957

Krishna Mahbubani
University of Cambridge https://orcid.org/0000-0002-1327-2334

Kourosh Saeb-Parsy
University of Cambridge https://orcid.org/0000-0002-0633-3696

Kate Gowers
University College London

Sam Janes
University College London https://orcid.org/0000-0002-6634-5939

Stanley Ng
The Cancer Ageing and Somatic Mutation Programme Wellcome Trust Sanger Institute

Matthew Hoare
University of Cambridge https://orcid.org/0000-0001-5990-9604

Anthony Green
University of Cambridge

George Vassiliou
University of Cambridge https://orcid.org/0000-0003-4337-8022

Ana Cvejic
Wellcome Trust–Medical Research Council Cambridge Stem Cell Institute, Cambridge, UK
https://orcid.org/0000-0003-3204-9311

Markus Manz
University and University Hospital Zurich https://orcid.org/0000-0002-4676-7931

Elisa Laurenti
University of Cambridge https://orcid.org/0000-0002-9917-9092

Michael Stratton
Wellcome Sanger Institute https://orcid.org/0000-0001-6035-153X

Jyoti Nangalia
Wellcome Sanger Institute https://orcid.org/0000-0001-7122-4608

Inigo Martincorena
Wellcome Sanger Institute https://orcid.org/0000-0003-1122-4416

Tim Coorens
Broad Institute of MIT and Harvard https://orcid.org/0000-0002-5826-3554

Biological Sciences - Article

Keywords:

Posted Date: November 23rd, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3610927/v1

Additional Declarations: Yes there is potential Competing Interest. PJC, MRS and IM are co-founders, stock-holders and consultants for Quotient Therapeutics Ltd.

Prolonged persistence of mutagenic DNA lesions in stem cells

Authors

Michael Spencer Chapman[1], Emily Mitchell[1,2,3], Kenichi Yoshida[1], Nicholas Williams[1], Margarete A. Fabre[1,2,3], Anna Maria Ranzoni[1], Philip S. Robinson[1], Matthias Wilk[4], Steffen Boetcher[4], Krishnaa Mahbubani[5,6], Kourosh Saeb Parsy[5,6], Kate H. C. Gowers[7], Sam M. Janes[7], Stanley W. K. Ng[1], Matt Hoare[8], Anthony R Green[2,3], George S. Vassiliou[1,2,3], Ana Cvejic[1,2,3], Markus Manz[4], Elisa Laurenti[2,3], Iñigo Martincorena[1], Michael R Stratton[1], Jyoti Nangalia[1,2,3], Tim H. H. Coorens[1,9], Peter J. Campbell[1,2,3]

Affiliations

(1) Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
(2) Wellcome-MRC Cambridge Stem Cell Institute, Cambridge Biomedical Campus, Cambridge, CB2 0AW, UK.
(3) Department of Haematology, University of Cambridge, Cambridge, CB2 2XY, UK.
(4) Department of Medical Oncology and Hematology, University of Zurich and University Hospital Zurich, Zurich, Switzerland
(5) Department of Surgery, University of Cambridge, Cambridge, CB2 0QQ, UK.
(6) Cambridge Biorepository for Translational Medicine, NIHR Cambridge Biomedical Research Centre, University of Cambridge, Cambridge CB2 2XY, UK.
(7) Lungs For Living Research Centre, UCL Respiratory, University College London, London, UK
(8) Early Cancer Institute, University of Cambridge, Cambridge, UK
(9) Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA

Address for correspondence

Dr Peter J Campbell,
Cancer, Ageing & Somatic Mutation Programme,
Wellcome Sanger Institute,
Hinxton CB10 1SA,
United Kingdom
e-mail: pc8@sanger.ac.uk

Abstract

DNA suffers continual damage leaving a cell with thousands of individual DNA lesions at any given moment[1–3]. The efficiency of DNA repair means that most known classes of lesion have a half-life of minutes to hours[3,4], but whether some DNA damage can persist for longer durations remains unknown. Here, using high-resolution phylogenetic trees from 89 donors, we identified mutations arising from 832 DNA lesions that persisted across multiple cell cycles in normal human stem cells from blood, liver and bronchial epithelium[5–12]. Persistent DNA lesions occurred at increased rates, with distinctive mutational signatures, in donors exposed to tobacco or chemotherapy, suggesting that they can arise from exogenous mutagens. In haematopoietic stem cells, persistent DNA lesions, likely from endogenous sources, generated a characteristic mutational signature, so-called SBS19[13]; occurred steadily throughout life, including *in utero*; and endured for 1.5 years on average, with 15% lasting 3+ years. We estimate that a haematopoietic stem cell has, on average, ~4-5 such lesions at any moment in time, half of which will generate a mutation with each cell cycle. Overall, 16% of mutations in blood cells are attributable to SBS19, and similar proportions of driver mutations in blood cancers exhibit this signature. These data imply the existence of a family of DNA lesions, arising from both endogenous and exogenous mutagens, present in low numbers per genome but persisting for months to years, that can generate sizable fractions of cells' mutation burdens.

Main text

A varied set of mechanisms have evolved to repair DNA lesions such as adducted, methylated or oxidised bases[14] – mutations arise when either the DNA repair is erroneous or there is misincorporation opposite an unrepaired lesion during DNA replication. The high rate at which many DNA lesions occur in a genome demands that DNA repair must be equally efficient, meaning that the half-life of an individual lesion is typically much shorter than the time between cell divisions[3,4]. However, a recent study in mice exposed to a single, high dose of the alkylating agent diethylnitrosamine (DEN) has shown that some DNA lesions can persist unrepaired through several cell cycles, generating different mutations at each round of replication[15]. Whether this phenomenon extends to other types of DNA damage, especially endogenously derived lesions in humans, remains unknown.

We hypothesised that high-resolution phylogenetic trees of somatic cells would enable us to infer the persistence of endogenous or exogenous DNA lesions across multiple cell cycles (**Figure 1**). In a phylogenetic tree of somatic cells, each branch-point, formally known as a coalescence, records a historic cell division[16] – successive branch-points tracing a 'line-of-descent' from root to tip record different cell divisions through that clone's ancestry. A given DNA lesion that persisted across several cell divisions would have potential to generate a mutation each time that strand was replicated, and these separate mutations could be detectable in the phylogeny. If different bases were misincorporated opposite the lesion during sequential rounds of DNA replication, closely related clones would carry two alternative mutations at the same position in the genome ('multi-allelic' variants; **Figure 1a,b**), as described in the mouse model of DEN exposure[15]. Furthermore, if the persistent lesion had the same base misincorporated opposite during different rounds of replication, those mutations could, under some circumstances, be recognised through their contravention of the consensus phylogeny ('phylogeny-violating' variants; **Figure 1c**; **Extended Figure 1**).

High-resolution phylogenetic trees of normal stem cells

We collated seven published sets of somatic phylogenies from whole-genome sequencing of single-cell-derived colonies[5–9], organoids[10] or laser-capture microdissections (LCM)[11,12]. The dataset comprises 103 phylogenies from 89 individuals, encompassing a total of 11,429 whole genomes, with a median of 48 samples per individual (range, 11–451; **Table S1**). Each phylogeny was generated from a single tissue type: haematopoietic stem and progenitor cells (HSPCs, n=39), bronchial epithelial cells (n=16) or liver parenchyma (n=48, from 34 individuals, due to separate phylogenies for 8 anatomical segments of the liver in 2 subjects). The HSPC phylogenies were from individuals that fell into five categories: foetal and cord blood (n=4), healthy adults (n=13), stem cell transplant donor/recipient pairs (n=10), patients with myeloproliferative neoplasms (n=10) and chemotherapy-exposed patients (n=2). Variant-calling, filtering and reconstruction of phylogenetic trees were undertaken using established and extensively validated pipelines, as described previously[5–12].

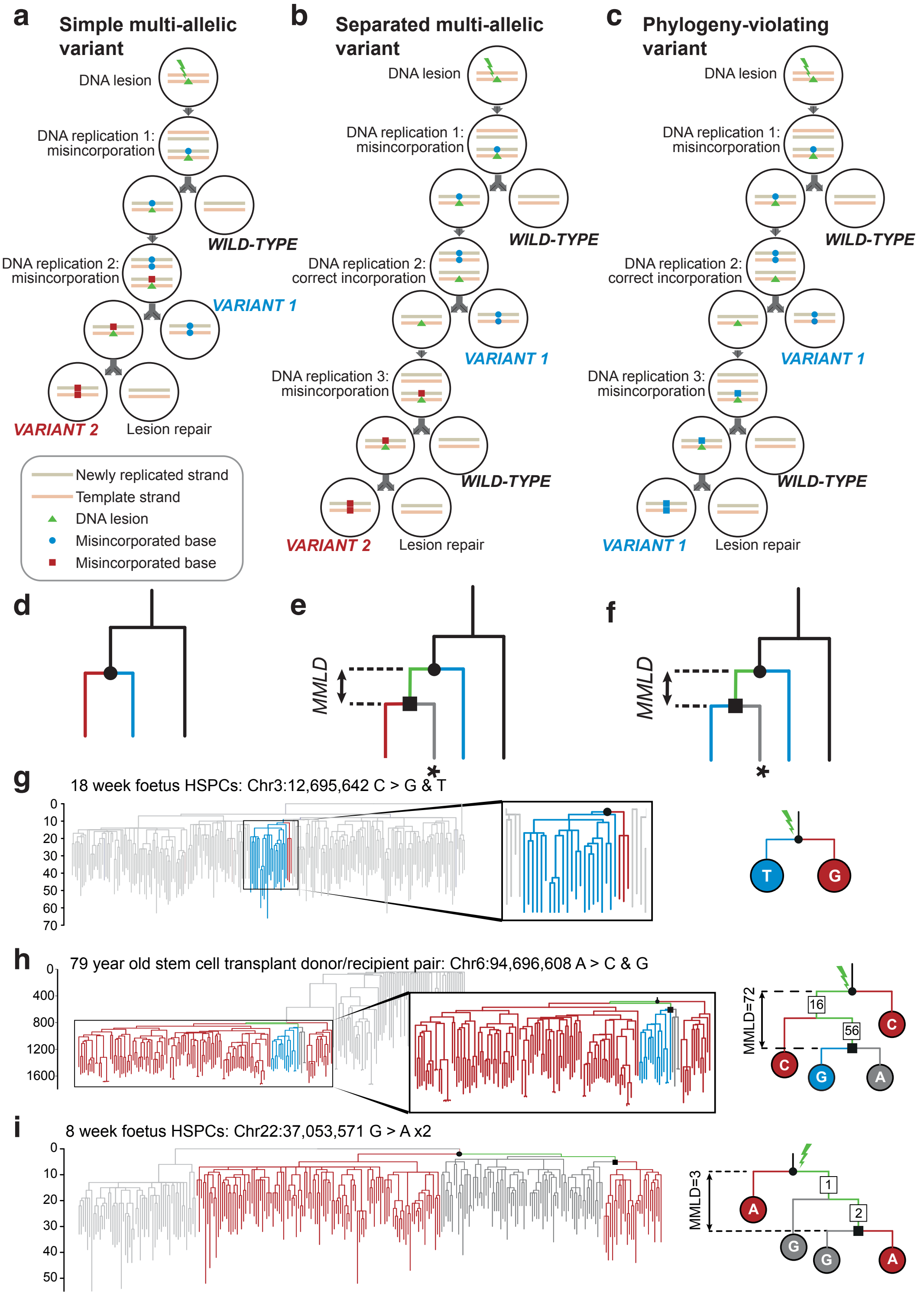

a
Simple multi-allelic variant
DNA lesion
DNA replication 1: misincorporation
WILD-TYPE
DNA replication 2: misincorporation
VARIANT 1
VARIANT 2
Lesion repair
Newly replicated strand
Template strand
DNA lesion
Misincorporated base
Misincorporated base
b
Separated multi-allelic variant
DNA lesion
DNA replication 1: misincorporation
WILD-TYPE
DNA replication 2: correct incorporation
VARIANT 1
DNA replication 3: misincorporation
WILD-TYPE
VARIANT 2
Lesion repair
c
Phylogeny-violating variant
DNA lesion
DNA replication 1: misincorporation
WILD-TYPE
DNA replication 2: correct incorporation
VARIANT 1
DNA replication 3: misincorporation
WILD-TYPE
VARIANT 1
Lesion repair
d
e
MMLD
*
f
MMLD
*
g
18 week foetus HSPCs: Chr3:12,695,642 C > G & T
0
10
20
30
40
50
60
70
T
G
h
79 year old stem cell transplant donor/recipient pair: Chr6:94,696,608 A > C & G
0
400
800
1200
1600
MMLD=72
16
56
C
C
G
A
i
8 week foetus HSPCs: Chr22:37,053,571 G > A x2
0
10
20
30
40
50
MMLD=3
1
2
A
G
G
A

Figure 1

Figure 1. Types of variants resulting from persistent DNA lesions. a, Mechanism of generation of a 'simple' multi-allelic variant (MAV). Orange lines represent the template strand for replication and khaki lines the newly replicated strand. Green triangles represent a DNA lesion generated by a mutagen (lightning symbol). The red squares and blue circles represent different incorrect bases incorporated during DNA replication. **b,** Mechanism of generation of a 'separated' MAV. **c,** Mechanism of generation of a phylogeny-violating variant (PVV). **d,** Appearance of phylogeny resulting from events in **a**. **e,** Appearance of phylogeny resulting from events in **b**. The green line represents the lesion path and the * highlights the subclade that is negative for the mutation resulting from non-mutagenic replication. **f,** Appearance of phylogeny resulting from events in **c**. The green line represents the lesion path and the * highlights the subclade that is negative for the mutation resulting from non-mutagenic replication. **g,** Example of simple MAV from the 18 post-conception week pcw foetal phylogeny. Branches coloured red have progeny harbouring the C>G variant, and blue the C>T variant, both on chromosome 3, position 12695642. The inset shows a zoomed-in view of the relevant portion of the tree. **h,** Example of 'separated' MAV. Red branches have progeny harbouring the A>C variant, and blue the A>G variant, both on chromosome 6, position 94696608. Grey branches have progeny with the reference allele. The inset shows a zoomed-in view of the relevant portion of the tree. **i,** Example of a PVV from the 8pcw foetal phylogeny. Branches coloured red have progeny harbouring the G>A on chromosome 22, position 37053571. MMLD, minimum molecular lesion duration.

Identification of multi-allelic and phylogeny-violating variants

To identify multi-allelic variants (MAVs), we examined all phylogenies for genomic positions recording two or more mutant alleles, revealing 1079 such sites. Such events may occur by chance if the same position happens to mutate in two lineages independently – indeed, for many of these events, the two clades reporting the MAVs were far apart from one another on the tree and did not share a single line-of-descent, suggesting they were not generated from the same persistent DNA lesion (n=727; **Extended Figure 2**). However, 352 MAVs were close together on the phylogenetic tree and within a single line-of-descent, a pattern that would be consistent with a persistent DNA lesion. For MAVs found in phylogenies built from clonal samples (n=293), the precise organisation of mutant clades could be established. In 80% of these cases (233/293), the two mutant clades had the same ancestral node (**Figure 1a,d,g**), whereas in the others (20%, 60/293), the mutant clades could be linked to a single lesion path through the phylogeny (**Figure 1b,e,h**). We refer to these orientations as 'simple' and 'separated' respectively, whereas MAVs that are sufficiently distant on the phylogeny as to be inconsistent with a single DNA lesion we term 'unrelated'.

We used three approaches to assess whether these simple and separated MAVs could plausibly have arisen through two independent events at the same locus. First, we simulated the null model of multi-allelic variants occurring as unrelated mutations to estimate the proportion anticipated to occur in 'simple' and 'separated' orientations by chance. The observed data had a 28-fold higher proportion of simple MAVs and a 3.8-fold higher proportion of separated MAVs than would be predicted by the null model (**Figure 2a,b**; **Extended Figure 3a**). Second, we assessed MAVs with nearby heterozygous germline polymorphisms for phasing (**Extended Figure 3b**). A prerequisite for MAVs being caused by a single lesion is that the phasing is to the same parental copy of the chromosome. As expected, MAVs in an orientation inconsistent with generation by a single persistent lesion (unrelated MAVs) had approximately equal proportions of matching and conflicting phasing (128 matching of 230 total, *p*=0.10, binomial test). In contrast, the phasing was almost universally matched for both simple and separated MAVs (78/81, *p*=7x10^{-20} for simple MAVs; 21/24, *p*=0.0009 for separated MAVs; **Figure 2c**). However, when the mutant clades were separated by two or more nodes with non-mutant clades, they tended towards a more equal distribution of matching and conflicting phasing (**Extended Figure 3c**), implying that a subset of these MAVs were not caused by a persistent DNA lesion. Therefore, these MAVs (n=21), and any others with non-matching phasing (n=3), were excluded from downstream analysis. Third, we compared the distribution of base changes and local sequence context, known as the mutational spectrum, for MAVs against that expected to arise from two independently occurring mutations at the same base. This demonstrated that the unrelated MAVs had a very similar spectrum to that expected for independent mutations (**Extended Figure 4**), whereas the spectrum of simple and separated MAVs was distinct (see following section).

To identify phylogeny-violating variants (PVVs), we developed a statistical approach to detect mutations where there was excessive variability (overdispersion) in read counts reporting the variant either within or outside its assigned branch on the tree (**Extended Figure 1e**; **Methods**). As accurate phylogenies are essential for such inference, we included only phylogenies built from single-cell-derived samples, excluding the liver samples. This identified 847 mutations that violated the phylogeny. For 238 of these, the locations of the different subclones reporting the PVV on the phylogenetic tree were not consistent with generation by a single persistent DNA lesion – as for MAVs, these probably arose through two separate mutations in independent lineages.

The remaining 526 were in an orientation that would be consistent with a persistent DNA lesion. However, these may also occur due to independent acquisition by chance; furthermore, incorrect reconstruction of the phylogenetic tree, loss-of-heterozygosity (LOH) or spontaneous reversion of a somatic mutation within a subclade may also result in a false-positive PVV call. We systematically evaluated each of these possible mechanisms of PVV generation using simulation, phasing, copy number and signatures (**Figure 2d-h**; **Extended Figure 5**; a detailed discussion and analysis of each possible artefactual source of PVVs is reported in **Methods**). Overall, alternative mechanisms accounted for only a small proportion of identified PVVs, which we excluded from further analysis.

In summary, the overwhelming majority of MAVs and PVVs occurring in close proximity on the phylogenetic tree cannot be accounted for by two independent mutational events or other trivial explanations. After excluding those variants that could be explained by alternative mechanisms, we took forward a final dataset of 331 MAVs and 501 PVVs (477 SNVs; 24 indels) for downstream analysis (**Extended Figures 6-7**; **Table S2**).

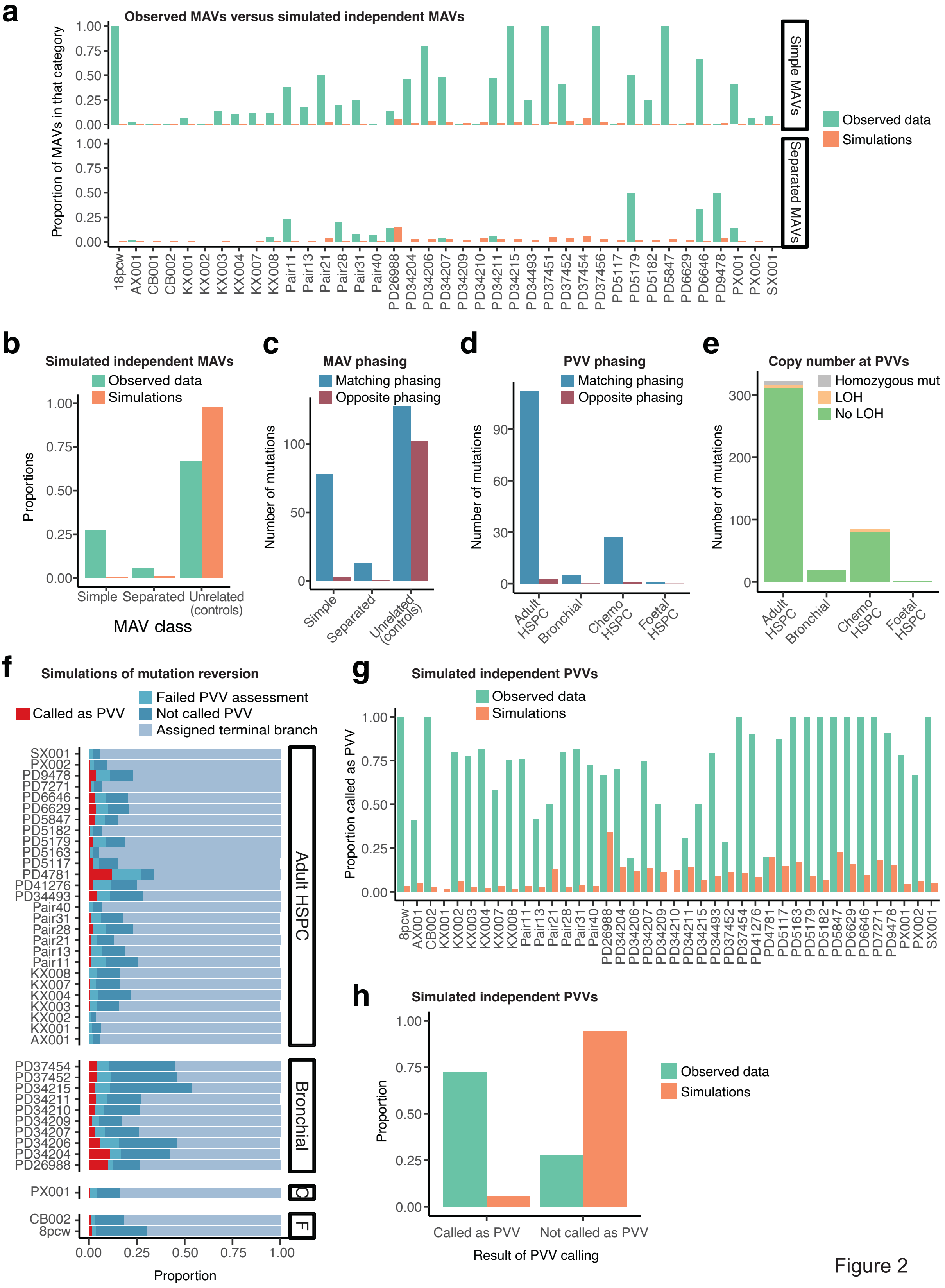


Figure 2

Figure 2. Validation of MAVs and PVVs. a, Bar plot showing a comparison of the proportion of MAVs occurring in 'simple' or 'separated' orientations for each individual, and in the simulated null model of occurrence by independent mutation acquisition. **b,** Overall proportions of simulated independent MAVs that would be classified as 'simple', 'separated' or 'unrelated' compared to those seen in the data. The simulation proportions are weighted by the total number of MAVs called in the actual dataset for each subject in order to reflect their contribution to the MAV dataset. **c,** Phasing comparison of the two mutant alleles of MAVs, including 'unrelated' MAVs (those in an orientation inconsistent with generation by a persistent DNA lesion). **d,** Barplot showing the results of a phasing comparison of the positive subclades of PVV, only including those with two or more positive subclades for which phasing could be confirmed. **e,** Stacked barplot showing the results of a copy number analysis of the PVV negative subclade(s) using the algorithm ASCAT[17] to look for evidence of loss-of-heterozygosity. **f,** Results of simulation of spontaneous mutation reversion, showing that very small proportions would be detected and classified as PVVs using the described approach (red bars), compared to being excluded from downstream analyses for various reasons (different shades of blue). **g,** Results of simulation of apparent PVVs caused by two independent mutations, showing the proportions of detected events that would be classed as PVVs (orange), compared to the proportions in the observed data (green). **h,** Overall proportions of simulated independent PVVs that would be classified as a PVV, compared to the observed data. The simulation proportions are weighted by the total number of PVVs called in the actual data for each subject in order to reflect their contribution to the PVV dataset. LOH, loss of heterozygosity.

Numbers and signatures of MAVs and PVVs

The ability to detect MAVs and PVVs depends on having several nodes in the phylogenetic tree during the timespan of lesion persistence. Phylogenies that have a rich clonal structure therefore provide most statistical power for their detection – as expected, there was a correlation between the number of post-development nodes in the phylogeny and the number of detected MAVs and PVVs (**Extended Figure 8a**). However, the number of MAVs per node varied substantially across tissues, being >10 times higher in bronchial and liver samples than in the HSPC samples (bronchial, 0.36 MAVs/node; liver, 0.22 MAVs/node; HSPCs, 0.02 MAVs/node; p=0.04 for between-group differences, Kruskal-Wallis test; **Figure 3a**). Some bronchial epithelium phylogenies had particularly high numbers of MAVs, typically from current or ex-smokers (p=0.04 for MAVs/node for never-smokers versus smokers with ≥30 pack-years, Kruskal-Wallis test), suggesting that lesions resulting from mutagens in tobacco smoke can persist over multiple cell cycles and lead to variable base incorporation during replication. The MAVs in bronchus and liver phylogenies had a similar spectrum, dominated by T>C/T>A mutation pairs, with some enrichment at ApT dinucleotides (**Figure 3b**). This has most resemblance to the predicted MAV signature that would arise from SBS16 (**Extended Figure 8b-d**), a signature of unknown aetiology that is increased in liver[11,18,19] and tobacco-exposed lung[10].

Most of the PVVs we identified were in the HSPC phylogenies, although this was explicable by the greater statistical power for detection in these trees (**Extended Figure 9**). The PVVs in the adult HSPC phylogenies had a distinctive mutational signature, characterised by C>T transitions particularly at CpT dinucleotides (**Figure 3c**). It most closely matches COSMIC signature SBS19 (cosine similarity, 0.96) and has the same transcriptional strand bias for G on the untranscribed strand (*p*=0.02, two-sided Poisson test; **Figure 3d**), suggesting it is the guanine that carries the lesion. Several previous studies have noted that the characteristic mutational spectrum of HSPCs is different to that seen in other tissues, with more pronounced peaks of C>T at CpT dinucleotides[5,16,20]. Interestingly, we found that this normal HSPC mutational spectrum could be accurately reconstructed from a combination of SBS1, SBS5 and SBS19 (**Extended Figure 10a**), with SBS19 contributing 16% of mutations overall. In contrast, the spectrum of mutations in the normal cells of other tissues[21–24] was effectively reconstructed from SBS1 and SBS5 alone, with SBS19 contributing ≤4% (**Extended Figure 10b**). The aetiology of SBS19 is unknown, originally discovered in a subset of blood cancers, liver carcinomas and pilocytic astrocytomas[13,25].

The chemotherapy-exposed HSPC phylogeny showed elevated numbers of MAVs and PVVs, each with a distinctive spectrum. MAVs showed marked dominance of mutations at T:A pairs, though with minimal context specificity. A notable feature was mixed SNV/indel MAVs, with 10 of 12 representing single nucleotide T deletions at CpT sites combined with T>A or T>G transversions. Of the 90 detected PVVs, 19 (21%) were indels, a far higher proportion than in the rest of the data set (2%). These were all single-nucleotide T deletions at homopolymer tracts of ≥4 T bases mirroring the overall indel signature in this individual (**Figure 3e**). The remaining 67 SNV PVVs were predominantly T>A transversions (78%, **Figure 3c**). These data suggest that the chemotherapy this patient received, which included alkylating agents, generated many DNA lesions that persisted through multiple cell divisions. It is fascinating that the same lesion could generate an indel in one round of replication and a substitution in another (indel/SNV MAVs); two indels interspersed with a correct base incorporation (indel PVVs); or two identical substitutions interspersed with a correct base incorporation (substitution PVVs). This provides *in vivo* evidence for biochemical studies of translesion synthesis showing that a single lesion can result in indels, SNVs or correct base incorporations, depending on whether slippage or extension occurs during lesion bypass[26,27].

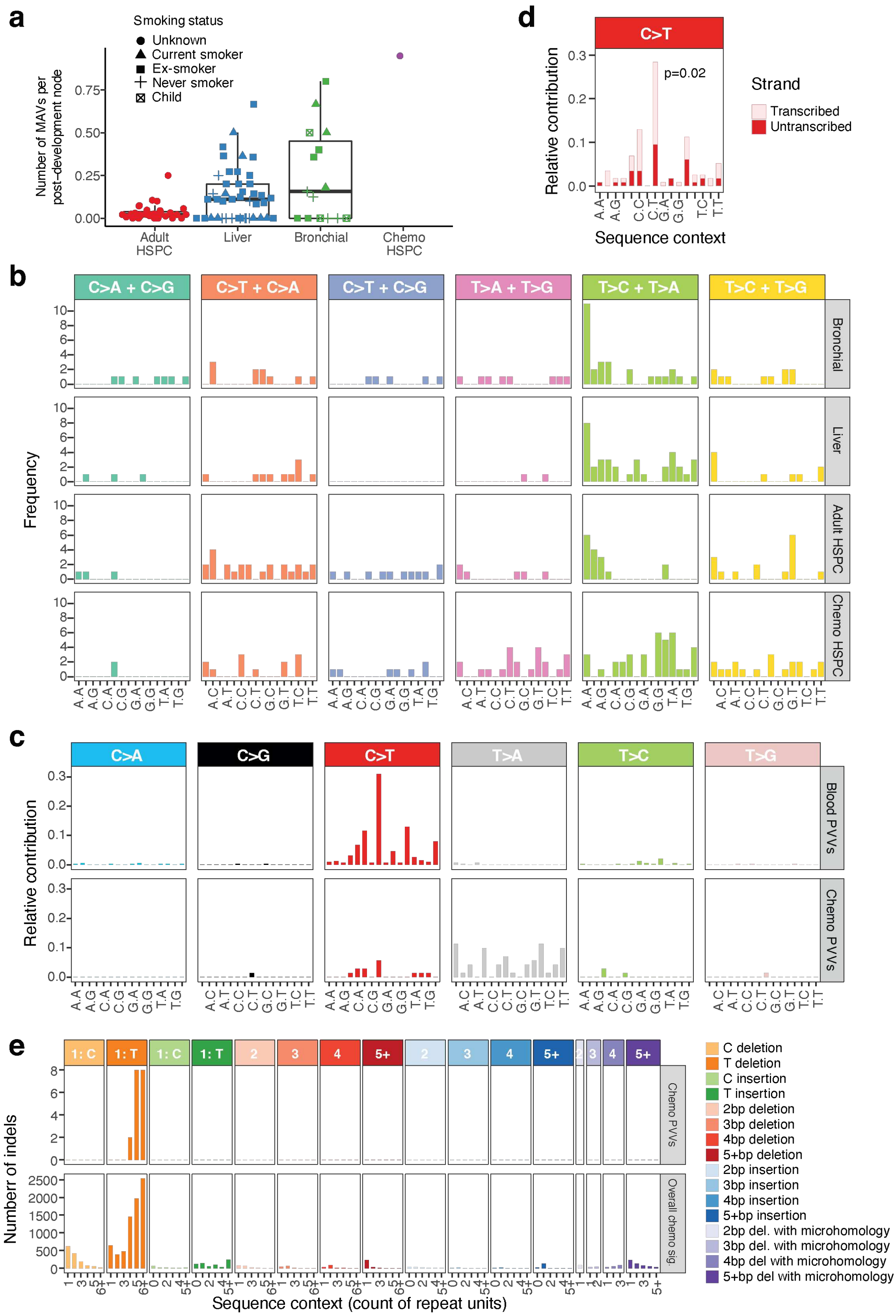


Figure 3

Figure 3. Signatures of MAVs and PVVs. a, Box-and-whisker plot showing the number of MAVs per post-development node for each sample, divided by tissue type, with raw data superimposed. Smoking status, where known, is indicated by shape. The median is marked with a heavy black line

and interquartile range in a thin black box. Whiskers denote the range or 25th and 75th centile plus 1.5× the interquartile range, whichever is less. **b,** Custom 96-profile MAV signatures, displaying the 6 possible SNV combinations that make up an MAV and, for each, the 16 possible combinations of flanking bases around the reference base. These are shown for MAVs in bronchial, liver, adult HSPC and chemotherapy-exposed HSPC phylogenies. MAVs incorporating an indel or multi-nucleotide variant are excluded. **c,** 96-channel spectrum of the observed PVVs in adult HSPC phylogenies, and the chemotherapy-exposed HSPC phylogeny. The trinucleotide context is shown as four sets of four bars, grouped by whether an A, C, G or T respectively is 5′ to the mutated base, and within each group of four by whether A, C, G or T is 3′ to the mutated base. **d,** Transcription strand bias of C>T mutations in the adult HSPC PVVs. **e,** Signature of indel PVVs in the chemotherapy-exposed HSPCs, and the overall indel signature in the same samples. The coloured bars across the top denote the size of the insertion or deletion (described in the legend to the right); within each facet, the x axis denotes the count of repeated units in the reference genome matching the inserted or deleted sequence.

Timing and duration of persistent lesions

Mutations in HSPCs accumulate at a constant rate throughout postnatal life, with that rate showing minimal cell-to-cell variation either within or between healthy individuals[5,20,22,28]. At birth, blood cells have ~50 mutations each[5,20], and these are acquired at relatively constant rates through the 38 weeks of gestation[7]. Since we know the nodes on the tree at which a persistent DNA lesion must have existed and, for PVVs and separated MAVs, the earliest node at which it was repaired, we can estimate the chronological age at which it occurred and a lower bound on the length of time it persisted unrepaired.

Intriguingly, 24 MAVs and 26 PVVs were likely acquired *in utero* as they could be timed to nodes at <50 mutations of molecular time (the average mutation burden in cord blood cells[5,20]); furthermore, some were identified in the phylogenetic trees of foetal HSPCs (**Figure 1g,i**). In 3 cases, the causative lesion could be traced to the most recent common ancestor of all sequenced cells, probably the fertilised egg[7,29] (**Figure 1i**; **Figure 4a**), and in a further 7 cases, to a cell only 1-2 generations later – a previously published somatic phylogeny also found an MAV present in multiple germ layers[30], consistent with a pre-gastrulation lesion. The rates of mutations arising *in utero* from persistent DNA lesions were lower than seen for post-development nodes – for example, the rate of MAVs in nodes timed to <50 mutations was ~0.004 MAVs/node, a fifth of the rate for adult nodes. Thus, persistent DNA lesions can occur *in utero*, albeit at lower rates than postnatally. Given the shielded environment of the foetus, it seems likely that this DNA damage arises through endogenous processes, although an exogenous mutagen that crosses the placenta cannot be excluded.

In adult blood, MAVs and PVVs occurred steadily throughout the lifespan in numbers commensurate with our power to detect them (**Figure 4b**), consistent with their generation by a clock-like mutational process. As expected, the timing of PVVs and MAVs in tissues exposed to exogenous mutagens, namely the smoking-exposed bronchial epithelium and chemotherapy-exposed HSPCs, varied among individuals and through time, dependent on individual mutagen exposure (**Figure 4c**).

We also estimated lower bounds on the duration of each molecular lesion for all PVVs and separated MAVs, corresponding to the number of mutations acquired elsewhere in the genome while the lesion persisted unrepaired. If the mutations elsewhere in the genome originate from clock-like processes, this minimum molecular duration can be converted to a minimum chronological duration. For PVVs in the adult HSPC phylogenies, most minimum lesion durations ranged between 10 and 100 mutations (median, 21; IQR, 12–37; **Figure 4d**). This corresponds to a median for the minimum chronological duration of 1.3–1.5 years, and suggests that durations >3 years are common (85th centile at 55 mutations of molecular time). The distribution of lesion durations varied depending on the mutation type, with C>T PVVs having shorter minimum durations than T>C PVVs (median, 21 versus 37.5 mutations respectively; p=0.02; Mann-Whitney test; **Figure 4e**). Durations of the separated MAVs were similar to the PVVs (p=0.29; Mann-Whitney test; **Figure 4f**).

The minimum durations of lesions generating PVVs in the bronchial and chemotherapy-exposed HSPC phylogenies were more variable and in many cases much longer when calculated in molecular time (**Figure 4d**). However, since background mutation rates in these settings do not show the clock-like properties seen in unexposed HSPCs, we cannot convert these to chronological time – indeed, it is likely that the longer apparent molecular durations in these settings derives from shorter periods of accelerated mutation rates rather than long real-time durations.

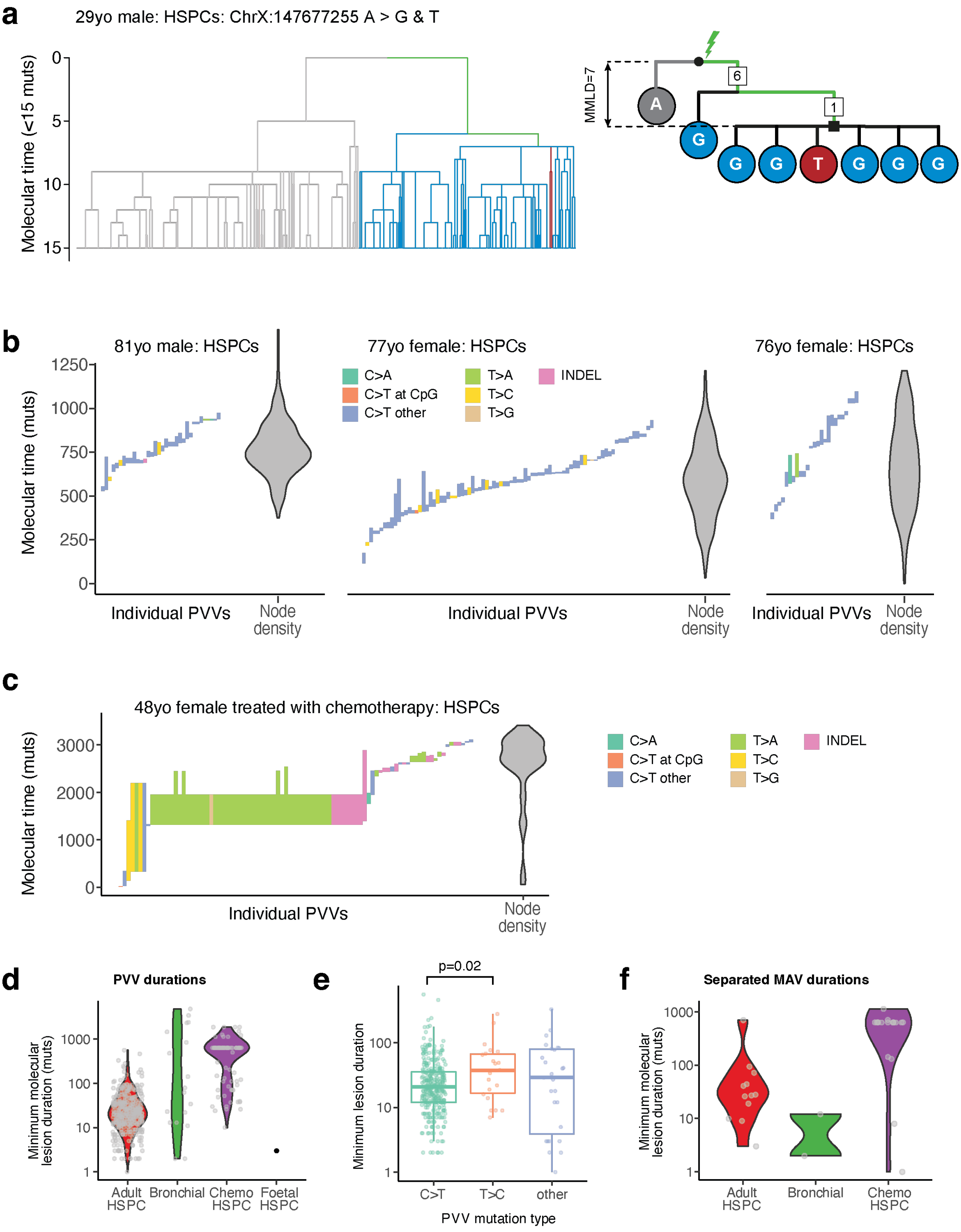


Figure 4

Figure 4. Timing and duration of PVVs. a, Example of a lesion resulting in an MAV that must have been present in the zygote. The phylogenetic tree has been truncated to the first 15 mutations of molecular time to provide sufficient resolution. Branches are coloured by whether descendants of the branch carry the A>G mutation (blue), A>T mutation (red) or reference allele (grey). Right is a schematic showing the clade structure and lesion path. **b**, Plot showing the latest time of lesion acquisition and the earliest time of lesion repair for the three adult HSPC phylogenies with the most PVVs observed.

Each column represents an individual PVV-causing lesion, ordered by the time of lesion acquisition; the bottom of the column is the latest time of lesion acquisition; the top of the column is the earliest time of lesion repair; the height of the column therefore represents the minimum molecular lesion duration (MMLD). Columns are coloured by PVV mutation type. To the right is the 'node density', where the density of post-development internal nodes is shown as a violin plot (kernel-smoothed density plots with vertical mirror symmetry). **c,** as in **b**, but for PVVs for PX001, the chemotherapy-exposed HSPC phylogeny. PVVs are clustered around 500-2500 mutations of molecular time, which corresponds to the timing of chemotherapy in this patient. **d,** Violin plot showing the density of minimum molecular lesion durations (MMLDs) of the detected PVVs, by phylogeny category. Individual data points are superimposed. **e,** Box-and-whisker plot showing the MMLD of PVVs, divided by substitution type. The boxes indicate the median and interquartile range (IQR) and the whiskers extend to the largest/smallest values no more than 1.5xIQR from the box. Individual data points are superimposed. **f,** Violin plot showing density of MMLDs of separated MAVs, by phylogeny category. Individual data points are superimposed.

Frequency and properties of lesions causing PVVs in HSPCs

Our framework for identifying PVVs requires that a lesion must persist across at least 2 nodes in the phylogeny and that the subclones with the PVV are separated on the tree by at least one wild-type subclone. This provides considerable constraint on our power to detect such events – despite this, we called 501 of them across the cohort, suggesting that the underlying lesions must be relatively frequent in the stem cell population. We used approximate Bayesian computation to generate estimates of the distribution of lesion durations and their frequency in stem cells. Across the 3 largest adult HSPC phylogenies, we simulated persistent DNA lesions using a broad, uninformative prior on mean lesion duration, recording (i) if they would result in a detectable PVV and (ii) the measured minimal lesion durations if detected. We then compared the simulations against the observed numbers and durations of C>T PVVs to obtain posterior estimates of their distribution (**Methods**).

The posterior distribution of the mean lesion duration had maximum density at 24.5 mutations of molecular time (95% credible interval, 17.7–34; **Figure 5a**), which is equivalent to 1.5 years, broadly in keeping with the direct estimates calculated above (**Figure 4d**). The proportion of simulated lesions that would have generated detectable PVVs was, as expected, low – the tree structure in the subject KX004, for example, meant that only ~1 in 1200 simulated lesions that lasted more than a year would have been detected as a PVV. In this phylogeny of 451 cells, we actually observed 50 C>T PVVs, implying that there must have been ~60,000 persistent lesions lasting >1 year in the cell ancestries comprising the combined branches of the observed phylogenetic tree. This equates to an estimate of 4-5 such lesions on average in any HSC at any given time. The other phylogenies yielded similar calculations.

We also calculated the base incorporation probabilities opposite these lesions during genome replication. For a basic PVV structure crossing 2 nodes (**Figure 1f**), detection of the PVV requires a fixed tree structure in which the two mutated subclones are separated by a wild-type subclone – such PVVs therefore offer no information on the base-pairing probability. However, 72 PVVs had a lesion path crossing more than 2 nodes, meaning that detection of the PVV does not depend on which base is incorporated for at least one subclone. The frequency of C versus T incorporations in these unbiased subclones was strikingly equal, with 41 C and 49 T base incorporations, giving a T pairing probability of 0.54 ($CI_{95\%}$, 0.43–0.65). Whether these two alternate outcomes reflect stochastic base incorporation by a single DNA polymerase or alternate polymerases with different incorporation preferences is unclear. Both mechanisms have been observed in experimental models of translesion synthesis[31–33].

These estimates of the prevalence and misincorporation rates of PVV lesions accord well with the observed rates of SBS19 in HSPCs. With 16% of clock-like mutations in HSPCs deriving from SBS19 (**Extended Figure 10a-b**), the rate of SBS19 would be ~2-3 mutations/HSPC/year[5,16,20]. The cell division rate of haematopoietic stem cells in humans is estimated to be ~1-2/year[34,35] – a prevalence of 4-5 persistent DNA lesions at any moment in time, half of which would generate a mutation in a given cell division, would therefore generate about 2-4 mutations/cell/year. These calculations suggest that the frequency and persistence of PVV-causing DNA lesions is entirely sufficient to explain the observed rate of SBS19 mutations in haematopoietic stem cells. Following from this, we analysed whether SBS19 causes driver mutations in HSPCs. From genome and exome sequencing data of myeloid cancers, we found that SBS19 was responsible for 10% of coding mutations in myeloid cancer genes, including *DNMT3A*, *TET2*, *ASXL1, TP53*, and up to 16% in some genes (**Extended Figure 10c**).

Strand asymmetry and lesion segregation

A key discovery in the paper reporting persistent DNA lesions in mice exposed to DEN was strand asymmetry of the mutations[15]. With the one-off dose of DEN, adducts were generated on thymines on both strands of a given chromosome in a given cell – at the next cell division, the two daughter cells inherited one each of those strands, but only one daughter cell seeded the eventual liver cancer. This created striking asymmetry of T>N versus A>N mutations on a chromosome-by-chromosome basis in the tumour, noting that this asymmetry depends upon numerous lesions per chromosome generated in a single cell cycle with limited dilution from mutations in other cell cycles. We deployed methods for detecting strand asymmetry of mutations[15] to analyse individual branches from our phylogenies for such lesion segregation. As expected, most phylogenies did not have any branches with evidence of strand asymmetry because of the clock-like properties of most mutational processes in these cells. We did detect lesion segregation in three branches of the chemotherapy-exposed HSPC phylogeny, with positive branches all having large contributions of the chemotherapy mutational signature (**Figure 5b**), consistent with the patterns seen in the DEN mouse model[15].

Interestingly, 6/16 bronchial epithelial phylogenies had at least one branch with significant strand asymmetry. These were not from smokers with many MAVs, but instead derived from branches carrying large proportions of mutations from mutational signatures SBS2 and SBS13 (**Figure 5c**). These signatures are caused by the base-editing activity of the APOBEC enzymes acting on cytosines[36–40]. The distribution of strands affected was such that about half of the chromosomes showed APOBEC mutations equally balanced between forward and reverse strands, with a quarter each showing marked asymmetry to the forward or to the reverse (**Figure 5d**). Such proportions could only occur if APOBEC lesions were generated within a single cell cycle, with skewed patterns arising when the daughter cell inherited, say, the forward strand of both parental copies of the chromosome. Studies of cell lines have reported that APOBEC mutagenesis can happen in episodic short bursts[41] – our data demonstrate that APOBECs can generate many hundreds to thousands of lesions across the whole genome within a single cell cycle. Strand co-ordination of clustered APOBEC mutations near structural variants has been repeatedly observed in cancer genomes[36,39], previously attributed to APOBEC acting on single-stranded DNA[36,40]. Our data suggest that strand coordination need not result from APOBEC modification of ssDNA, but could arise from lesion segregation even with APOBEC acting on double-stranded DNA.

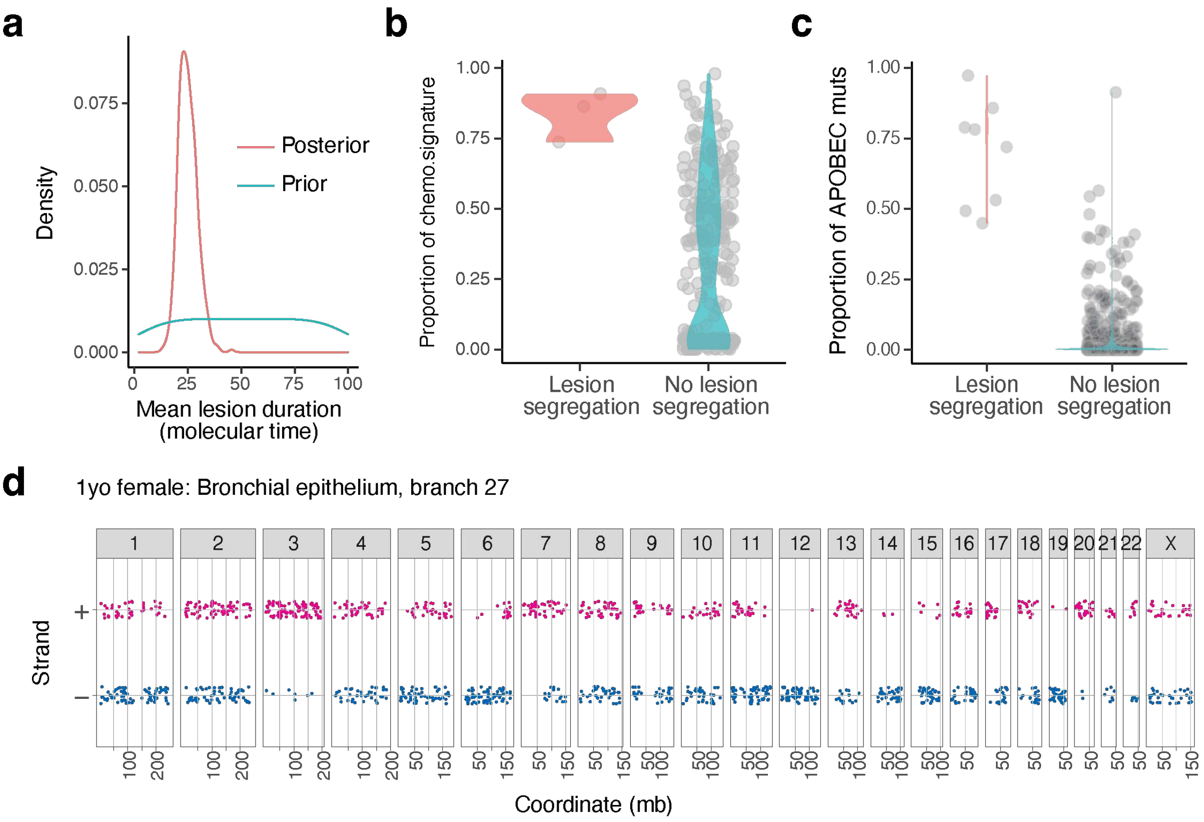


Figure 5

Figure 5. Lesion segregation. a, Density plots showing the prior (green) and posterior (orange) distribution for mean lesion duration in molecular time, measured in numbers of mutations. **b,** Violin plot with overlying individual datapoints showing the proportion of chemotherapy signature 1 in branches of the chemotherapy-exposed HSPC phylogeny PX001, divided by those with significant lesion segregation (n=3) and those without. **c,** Violin plot with overlying individual data points showing the proportion of APOBEC mutations from all bronchial epithelial phylogeny branches, divided by those with significant lesion segregation (n=8) and those without. **d,** The chromosomal strand and position of mutations from a branch affected by APOBEC mutagenesis from bronchial epithelial phylogeny PD37456, demonstrating significant lesion segregation (chromosomes 3, 12 and 14, for example).

Discussion

A vast register of DNA lesions emerges from the quotidian chemistry of life coupled with the rather more elective chemistry of our lifestyles. For example, genotoxic aldehydes can arise from endogenous sources, through innate folic acid metabolism, and exogenous sources, such as hepatic metabolism of dietary ethanol – a sophisticated repertoire of pathways has evolved to either detoxify them or repair the DNA lesions they cause[42–44]. Other pathways replace or repair DNA carrying cross-links, bulky adducts, oxidative damage, ultraviolet light photoproducts or abasic sites[14]. Presumably, the most frequent and most damaging DNA lesions exert the strongest pressure for the evolution of rapid repair mechanisms; the corollary being that there may be a class of lesions that are less prevalent and/or less detrimental for which repair is slower. Such lesions may be invisible to the usual techniques for direct discovery through chemistry and mass spectrometry[45] because of their low prevalence, and their corresponding DNA repair pathways difficult to uncover with standard experimental and knockout approaches.

Here, we used high-resolution phylogenetic trees built from normal human stem cells as an approach to deduce the presence of persistent DNA lesions generating somatic mutations across successive cycles of genome replication. For such an indirect approach, it is exciting how comprehensive a view of the lesion-to-mutation life cycle can be gleaned – for the PVVs in blood, for example, we can infer that lesions occur steadily throughout life, including *in utero*, and are therefore a likely consequence of endogenous cellular processes; that lesions persist in the DNA for months to several years; that lesions preferentially affect guanines in an Ap<u>G</u> context; that they are subject to transcription-coupled nucleotide excision repair; that lesions are present at a density of ~1/billion bases in a given haematopoietic stem cell; and that DNA replication across the lesion has a 50-50 chance of a misincorporation or correct insertion opposite the lesion. These estimates of lesion prevalence and duration are orders of magnitude away from the hundreds to thousands of 8-oxogunanines and methylated bases present in a cell with their associated half-lives of minutes to hours[3,4], data that has informed the high-frequency, rapid-repair model of the lesion-to-mutation life cycle. The signature emerging from persistent DNA lesions accounts for 16% of all mutations in blood cells, with similar proportions among the mutations that drive blood cancers – a fraction that is comparable to that seen for mutations arising from, say, spontaneous deamination of cytosine (SBS1, 14%), a much more frequent lesion in the genome[2]. Thus, DNA damage occurring at low frequency with slow repair also carries considerable threat to genomic integrity.

References

1. Rydberg, B. & Lindahl, T. Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine is a potentially mutagenic reaction. *EMBO J* **1**, 211–216 (1982).

2. Frederico, L. A., Shaw, B. R. & Kunkel, T. A. A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry* **29**, 2532–2537 (1990).

3. Swenberg, J. A. *et al.* Endogenous versus exogenous DNA adducts: Their role in carcinogenesis, epidemiology, and risk assessment. *Toxicological Sciences* **120**, S130–S145 (2011).

4. Hamilton, M. L. *et al.* A reliable assessment of 8-oxo-2-deoxyguanosine levels in nuclear and mitochondrial DNA using the sodium iodide method to isolate DNA. *Nucleic Acids Res* **29**, 2117–2126 (2001).

5. Mitchell, E. *et al.* Clonal dynamics of haematopoiesis across the human lifespan. *Nature* **606**, 343–350 (2022).

6. Williams, N. *et al.* Life histories of myeloproliferative neoplasms inferred from phylogenies. *Nature* **602**, 162–168 (2022).

7. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).

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

9. Campbell, P., Chapman, M. S., Wilk, M. & Boettcher, S. Clonal dynamics after allogeneic haematopoietic cell transplantation using genome-wide somatic mutations. *Preprint (Research Square)* **rs-2868644**, (2023).

10. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).

11. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).

12. Ng, S. W. K. *et al.* Convergent somatic mutations in metabolism genes in chronic liver disease. *Nature* **598**, 473–478 (2021).

13. Alexandrov, L. B. *et al.* The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).

14. Lindahl, T. & Wood, R. D. Quality control by DNA repair. *Science* **286**, 1897–905 (1999).

15. Aitken, S. J. *et al.* Pervasive lesion segregation shapes cancer genome evolution. *Nature* **583**, 265–270 (2020).

16. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).

17. Raine, K. M. *et al.* ascatNgs: Identifying Somatically Acquired Copy-Number Alterations from Whole-Genome Sequencing Data. in *Current Protocols in Bioinformatics* vol. 2016 15.9.1-15.9.17 (John Wiley & Sons, Inc., 2016).

18. Letouzé, E. *et al.* Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat Commun* **8**, (2017).

19. Alexandrov, L. B. *et al.* Mutational signatures associated with tobacco smoking in human cancer. *Science (1979)* **354**, 618–622 (2016).

20. Osorio, F. G. *et al.* Somatic Mutations Reveal Lineage Relationships and Age-Related Mutagenesis in Human Hematopoiesis. *Cell Rep* **25**, 2308-2316.e4 (2018).

21. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).

22. Abascal, F. *et al.* Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

23. Grossmann, S. *et al.* Development, maturation, and maintenance of human prostate inferred from somatic mutations. *Cell Stem Cell* **28**, 1262-1274.e5 (2021).

24. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).

25. Degasperi, A. *et al.* Substitution mutational signatures in whole-genome–sequenced cancers in the UK population. *Science (1979)* **376**, (2022).

26. Pagès, V. & Fuchs, R. P. How DNA lesions are turned into mutations within cells? *Oncogene* **21**, 8957–8966 (2002).

27. Becherel, O. J. & Fuchs, R. P. P. Mechanism of DNA polymerase II-mediated frameshift mutagenesis. *PNAS* **98**, 8566–8571 (2001).

28. Welch, J. S. *et al.* The Origin and Evolution of Mutations in Acute Myeloid Leukemia. *Cell* **150**, 264–278 (2012).

29. Park, S. *et al.* Clonal dynamics in early human embryogenesis inferred from somatic mutation. *Nature* **597**, 393–397 (2021).

30. Coorens, T. H. H. *et al.* Extensive phylogenies of human development inferred from somatic mutations. *Nature* **597**, 387–392 (2021).

31. Baynton, K., Bresson-Roy, A. & Fuchs, R. P. P. Analysis of Damage Tolerance Pathways in Saccharomyces cerevisiae : a Requirement for Rev3 DNA Polymerase in Translesion Synthesis . *Mol Cell Biol* **18**, 960–966 (1998).

32. Napolitano, R., Janel-Bintz, R., Wagner, J. & Fuchs, R. P. P. All three SOS-inducible DNA polymerases (Pol II, Pol IV and Pol V) are involved in induced mutagenesis. *EMBO Journal* **19**, 6259–6265 (2000).

33. Tissier, A. *et al.* Misinsertion and bypass of thymine-thymine dimers by human DNA polymerase ι. *EMBO Journal* **19**, 5259–5266 (2000).

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

35. Rufer, N. *et al. Telomere Fluorescence Measurements in Granulocytes and T Lymphocyte Subsets Point to a High Turnover of Hematopoietic Stem Cells and Memory T Cells in Early Childhood*. *J. Exp. Med* vol. 190 (1999).

36. Roberts, S. A. *et al.* Clustered Mutations in Yeast and in Human Cancers Can Arise from Damaged Long Single-Strand DNA Regions. *Mol Cell* **46**, 424–435 (2012).

37. Chan, K. *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* **47**, 1067–1072 (2015).

38. Law, E. K. *et al.* APOBEC3A catalyzes mutation and drives carcinogenesis in vivo. *Journal of Experimental Medicine* **217**, (2020).

39. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).

40. Taylor, B. J. M. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *Elife* 10.7554-10.7554 (2013) doi:10.7554/eLife.00534.

41. Petljak, M. *et al.* Characterizing Mutational Signatures in Human Cancer Cell Lines Reveals Episodic APOBEC Mutagenesis. *Cell* **176**, 1282-1294.e20 (2019).

42. Garaycoechea, J. I. *et al.* Genotoxic consequences of endogenous aldehydes on mouse haematopoietic stem cell function. *Nature* **489**, 571–575 (2012).

43. Dingler, F. A. *et al.* Two Aldehyde Clearance Systems Are Essential to Prevent Lethal Formaldehyde Accumulation in Mice and Humans. *Mol Cell* **80**, 996–1012 (2020).

44. Burgos-Barragan, G. *et al.* Mammals divert endogenous genotoxic formaldehyde into one-carbon metabolism. *Nature* **548**, 549–554 (2017).

45. Krieger, K. L. *et al.* Spatial mapping of the DNA adducts in cancer. *DNA Repair (Amst)* **128**, 103529 (2023).

46. Ellis, P. *et al.* Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat Protoc* 1–31 (2020) doi:10.1038/s41596-020-00437-6.

47. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).

48. Li, Y. *et al.* Patterns of somatic structural variation in human cancer genomes. *Nature* **578**, 112–121 (2020).

49. Blum, M. G. B. & François, O. Non-linear regression models for Approximate Bayesian Computation. *Stat Comput* **20**, 63–73 (2010).

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

METHODS

Curation of high-resolution phylogenetic trees

We combined data from seven previously published sets of somatic phylogenies. Each phylogeny was made up of cells from a single tissue type – haematopoietic stem and progenitor cells (HSPCs; 39 subjects)[5–9], bronchial epithelial cells (16 subjects)[10] or liver parenchyma (34 subjects)[11,12]. The phylogenies were built from somatic mutations discovered in whole-genome sequencing of single-cell derived colonies (HSPCs), single-cell derived organoids (bronchial epithelium) or laser capture micro-dissection[46] (LCM; liver). Details of the research subjects, sample acquisition, sequencing and variant calling are provided in the original manuscripts, although the clinical and demographic have been summarised here (**Table S1**).

In total, we collected 103 phylogenies from 89 individuals, with two of the research subjects from the liver study providing 8 phylogenies each due to independent sampling from all 8 anatomical segments of the liver in these individuals[12]. There was a median of 48 samples or clones per individual (range: 11–451). For the studies of the haematopoietic system, we defined three categories:

- Foetal and cord blood HSPCs (n=4; 2 from foetal haematopoietic organs, 2 from cord blood);
- Adult HSPCs (n = 28; 10 from deceased donor bone marrow with no known blood disorder, 3 from individuals with known clonal haematopoiesis, 10 from patients with myeloproliferative neoplasms, 10 from donor/recipient pairs of allogeneic haematopoietic stem cell transplant); and
- Chemotherapy-exposed HSPCs (n=2, one treated twice for Hodgkin's lymphoma with alkylating-agent-containing regimens; one treated with R-CVP, a regimen containing cyclophosphamide and vincristine). The R-CVP chemotherapy-exposed blood phylogeny was analysed with the 'adult HSPCs' group, rather than the 'chemotherapy-exposed HSPCs' group as R-CVP did not have significant mutagenic consequences for the HSPC population.

For individuals with single-cell-derived samples, mutations were filtered using similar approaches, with combinations of filters designed to remove germline variants, sequencing artefacts and in vitro-acquired mutations. For individuals from the liver study, collected using laser-capture microdissection, a matched normal was used for mutation calling, with some different downstream filtering steps[11,12]. The numbers of mutations per sample varied considerably across individuals depending primarily on tissue type, donor age, mutagen exposure (such as smoking, alcohol, chemotherapy), and disease status.

For the single-cell-derived colonies, phylogenetic trees were inferred using a maximum parsimony algorithm. For samples from the liver LCM studies, phylogenetic trees were inferred in a two-step procedure – first, the set of n-dimensional vectors of variant allele fractions for each mutation were clustered using a hierarchical Dirichlet process[11]; and, second, the phylogenetic tree describing these clusters was inferred using serial application of the pigeonhole principle[47]. The robustness of variant-calling and phylogenetic tree reconstruction using these methods has been extensively tested, with further details available in the original manuscripts.

Identification of multi-allelic variants

A somatic multi-allelic variant (MAV) occurs if a reference base at the same genomic position is mutated to two different mutant alleles in the same individual. For example, there may be evidence of both a C>A and C>T mutation at exactly the same chromosome and position. In the context of the phylogeny data analysed here, these different mutant alleles will be evident in different clones from the same individual.

Within each subject's phylogeny, we identified mutation pairs with overlap of the mutated locus. For SNV pairs, this is simply identifying SNVs at the same chromosome and position. For deletions and multi-nucleotide variants (MNVs, affecting 2 or more nucleotides), any degree of overlap was classified as a multi-allelic variant. Each MAV was classified as 'simple', 'separated' or 'fail' in a hierarchical manner, based on the orientation of their allocated branches in the phylogeny. If the two mutations had the same parent node (as in **Figure 1d**), the MAV was classed as 'simple', and the parent node classified as the 'lesion node'. If this was not true, but the allocated branch of one mutation fell within the clade defined by the parent node of the other (as in **Figure 1e**), the MAV was classed as 'separated' and the parent node that encompassed both mutations was classified as the 'lesion node'. If neither of these criteria were met, the MAV was classed as 'unrelated'. Separated MAVs where the two mutant alleles were separated by 2 or more subclades with the reference allele had a higher probability of occurring via independent mutation events, as evidenced by the more equal proportion of matching to non-matching phasing comparisons (**Extended Figure 3c**) and simulation. Therefore, these were reclassified as 'unrelated', even though a proportion were likely to be caused by a persistent DNA lesion. 'Unrelated' MAVs were likely to have been caused by independent events occurring at the same genomic locus by chance, a hypothesis supported by their mutational signatures which closely resembled the expected signature for this mechanism (**Extended Figure 4**).

For separated MAVs, the lesion path was defined by starting from the lesion node and working stepwise down the phylogeny along the path containing a mixture of alleles at the genomic position of interest. The last node encompassing two different alleles was classified as the 'lesion repair node'. In most cases this node was only one branch from the lesion node (as in **Figure 1e**), but there were occasional examples of longer lesion paths (**Figure 1h**). The minimum molecular lesion duration (MMLD) was calculated as the sum of branch lengths between the lesion and lesion repair nodes.

Identification of phylogeny-violating variants

A phylogeny-violating variant (PVV) is defined here as a mutation that is discordant with the consensus phylogeny. That is to say, the distribution across the phylogeny of clones carrying the mutant allele is not consistent with a single mutation-acquisition event and consistent inheritance in descendants thereafter.

The mutation assignment algorithm assumes that a single variant results from a single, fixed mutational event, after which all daughter cells carry the mutation. For the vast majority of mutations, these assumptions hold true: a branch exists that forms a clade containing only samples with the variant allele (**Extended Figure 1a**). For PVVs, these assumptions no longer hold true. The maximum-likelihood branch forms a clade that either (i) contains a subset of samples without the mutation (**Extended Figure 1b**), or (ii) does not contain all the samples with the mutation (**Extended Figure 1c**). This can be detected by testing the read counts of all clades that should theoretically be uniformly positive or negative for the variant, and testing for overdispersion of the observed counts. We quantified the overdispersion by assuming the counts to come from a beta-binomial distribution and finding the maximum-likelihood $\boldsymbol{\rho}$ parameter using the optim() and dbetabinom() functions from R packages 'stats' and 'VGAM'. As DNA lesions causing PVVs must occur on internal branches to allow detection, and will then at least partially follow the phylogeny, we reasoned that such mutations would invariably be allocated to internal branches. We therefore assessed only such internal branch mutations for overdispersion (i) within and (ii) outside the clade formed by the assigned node, quantifying the maximum-likelihood $\boldsymbol{\rho}$ parameter for each. PVVs will have a high $\boldsymbol{\rho}$ (empirically set as ≥0.1) for one of these parameters. Additionally, we required strong evidence for either a 'negative' subclade within the assigned clade (no variant reads detected with a minimum depth of 13), or a 'positive' subclade outside the assigned clade (VAF≥0.25 and ≥3 variant reads). Mutations meeting either pair of corresponding criteria (typically ~ 1 in 500 internal branch mutations) were considered phylogeny-violating and taken forward for further assessment.

Next, we assigned the putative 'lesion node' for each phylogeny-violating variant, namely the node containing all samples with the variant. For variants with evidence of overdispersion within the assigned clade (**Extended Figure 1b**), this was the same as the assigned branch. For variants with evidence of overdispersion outside the assigned clade (**Extended Figure 1c**), we iteratively travelled node-by-node up through the phylogeny, until there were no positive clades outside the clade defined by that node.

Finally, we iteratively worked down from the lesion node, attempting to define the 'lesion path' through the phylogeny. If caused by a DNA lesion, PVVs should have a specific orientation: at each cell division, one daughter cell will contain the DNA strand resulting from replication opposite the lesion and will therefore have a fixed genotype at the locus of interest (either the reference or mutant allele) which will be consistently inherited by its progeny. This is therefore a uniform subclade with a consistent genotype throughout. The other daughter cell will inherit the lesion itself, and therefore still has the potential for generation of the two alternate alleles, and therefore seeds a 'mixed' subclade. The lesion path is defined by the path containing mixtures of genotypes, and the outcome of replication at each cell division defined by the genotype of each uniform subclade arising from this lesion path. Once both daughters of the assessed node are uniform (that is, one contains only mutant samples, the other only wild-type), the lesion repair node has been reached (**Extended Figure 1b**). If both daughter nodes of the lesion node contain a mixture of positive and negative clades, this is inconsistent with generation by a persistent DNA lesion (**Extended Figure 1c**). Such mutations were deemed to have been generated by an alternative mechanism – indeed, mutational signature analysis showed that they were predominantly C>T at CpG sites, consistent with their generation by independent mutations at the same site (**Extended Figure 5d,e**).

Phasing of MAVs and PVVs to validate their derivation from a single DNA lesion

To assess if the variants of a multi-allelic variant pair were on the same allele (on the same chromosome copy of a homologous pair), we attempted to phase each variant with proximate heterozygous SNPs. Approximately a third of MAVs had a suitable heterozygous SNP sufficiently nearby for assessment. We extracted heterozygous SNPs within 1kb of the mutation locus using the VCF files of mutations. Each read-pair crossing both the variant and a heterozygous SNP locus was categorised by the base supported at each (**Extended Figure 3b**). For phylogenies built with single-cell-derived samples, matching phasing was confirmed if samples carrying each variant of an MAV pair had read-pairs with either (i) the same SNP base and their respective variant base or (ii) the same SNP base and the reference base. For phylogenies built with inferred clones (the subjects from the liver LCM studies), matching phasing was confirmed only if reads containing each variant base of an MAV pair had the same heterozygous SNP base – that is, the same SNP base phasing with the reference base was insufficient because of the potential inclusion of normal cells in the microdissection. The heterozygosity of the SNP was confirmed in each case. Phasing of the positive subclades of PVVs was similarly assessed. In cases with >2 positive subclades, we considered phasing as confirmed if one or more subclade pair was successfully phased.

Assessment of two independent mutations as an artefactual cause of MAVs

Given that MAVs are rare events, an alternative mechanism for their generation is two independent mutations occurring at the same locus by chance. We formally assessed the proportion of MAVs that would be expected to fall in 'simple', 'separated' or 'fail' orientations from this mechanism. This probability is specific to each phylogeny structure, determined by the number of samples sequenced and their mutation burdens (encoded as branch length in the phylogeny). To generate simulations of MAVs occurring independently by chance, we randomly selected pairs of phylogeny branches with probabilities proportional to their branch length, repeating this 50,000 times for each phylogeny. Each pair was categorised by the orientation of selected branches and compared with the proportions observed in the data. As we did for the observed data, separated MAVs with 2 or more intervening negative subclades were reclassified as 'fail'. To assess the overall degree to which the set of MAVs may be contaminated by those occurring by chance, we calculated a weighted mean of the simulated proportions in each category, using the total number of MAVs detected in each phylogeny as weights.

Assessment of two independent mutations as an artefactual cause of PVVs

As with MAVs, PVVs may result from two independent mutations at the same locus by chance. In addition, PVVs may theoretically result from spontaneous reversion of a somatic mutation in a subclade of the original mutant clade, a phenomenon that has previously been observed in cases where wild-type cells have a selective advantage. However, the orientation of PVVs did not appear to be that expected from either of these mechanisms. To formally test this, we designed simulations of each, testing all phylogenies with at least one PVV.

(1) *Independent mutations at the same site.* Similar to the MAV simulations, we randomly selected pairs of branches with probabilities proportional to their branch lengths. Each sample was assigned

a depth for the simulated PVV locus according to a random draw from a Poisson distribution with the λ parameter being the overall mean depth in that individual. Samples within clades formed by the selected branches, 'positive' samples, were assigned variant counts according to random binomial draws with p=0.5 and n=depth. Other 'negative' samples were assigned variant counts by similar random draws but with p=1x10^{-6} (the error distribution). Analysis then proceeded as with the data: a single branch was assigned with *treemut*; if this branch was a terminal branch, there was no further analysis; if it was a shared branch, the counts within and outside the assigned branch clade were assessed for overdispersion using the beta-binomial distribution and the same ***ρ*** thresholds as for the data; the lesion node was assigned and the mutation classified as 'pass' or 'fail'. For pass mutations, a lesion node, lesion repair node, MMLD and minimum number of cell divisions was calculated and recorded. We repeated this 10,000 times for each phylogeny and recorded the outcome for all. For all individuals, the majority of independent mutations at the same site were assigned to terminal branches and would not have been included among the PVVs in the dataset. A proportion were assigned to shared branches but did not meet the filtering criteria for a PVV. Notably, very low proportions fell in an orientation consistent with generation by a persistent DNA lesion ('pass' PVVs: median, 0.018; range, 0.002–0.12), with the lowest proportions in those phylogenies with the highest numbers of 'pass' PVVs in the data. We then compared the 'pass' PVV numbers as a proportion of the total detectable PVVs in the data and simulations. To assess the overall degree to which the set of PVVs may be contaminated by those occurring by this mechanism, we calculated a weighted mean of the simulated 'pass' or 'fail' proportions, using the total number of PVVs detected in each phylogeny as weights (**Figure 2g**).

(2) *Spontaneous somatic reversion of somatic mutation.* For a somatic reversion event to be evident in a phylogeny, there must first be a somatic mutation, and subsequently a reversion event within the captured lineages of the mutant clade. The probability of a branch giving rise to a captured somatic reversion is therefore proportional to the product of the branch length and the sum of branch lengths within that clade. Intuitively, this can be thought of in these terms: a long branch has lots of mutations with the potential for reversion, and many subsequent long branches within that clade gives much time and many independent lineages for those mutations to revert. Therefore, we selected branches with probabilities weighted by this product. The reversion branch was then chosen from branches within the selected branch clade, again with probabilities weighted by their branch lengths. Simulation then proceeded analogously to the 'independent mutation' simulation, but with positive samples defined as those within the somatic mutation clade, but not in the reversion clade. We repeated this 2,000 times for each phylogeny and recorded the outcome of each (**Figure 2f**). Interestingly, only a minority of somatic reversion events were detected as

PVVs (median proportion, 0.173), as they often result in large positive clades with a single negative sample. These have little impact on the inferred $\boldsymbol{\rho}$ value, as such occasional negative samples are not unexpected with binomial sampling of the variant and wild-type alleles as occurs with sequencing of heterozygous sites. However, as expected, almost all those detected were classified as 'pass' PVVs.

Assessment for loss of heterozygosity as an artefactual cause of PVVs

PVVs may result if a cell containing a somatically acquired mutation loses the mutant allele through say whole chromosome deletion or smaller-scale loss-of-heterozygosity (LOH) mechanisms such as mitotic recombination and focal deletion. To exclude this as the cause of the observed PVVs, we employed two complementary approaches. First, we applied ASCAT, an allele-specific copy number algorithm[17] to each sample in turn, using a phylogenetically unrelated sample from the same individual as the matched normal. For each PVV, we defined the negative subclades and determined the minor allele copy number at the mutant locus for each sample within that clade. The mean value across negative subclade samples was rounded to the nearest integer: a value of 1 was classed as 'No LOH' (heterozygosity is maintained) and 0 as 'LOH'. A small proportion of PVVs (10/426, 2.5%) were shown to be caused by LOH (**Figure 2e**), and were excluded from further analysis.

It is also formally possible that short sub-kilobase LOH events were missed by ASCAT. We therefore used a second approach to directly interrogate sequencing reads and confirm that samples in the negative subclade included reads from the chromosome copy harbouring the mutation. To do this, we first phased the mutation with nearby heterozygous SNPs by interrogating the reads from samples with the mutation. This was possible in approximately one third of cases (**Extended Figure 5a**). We then interrogated colonies from the negative subclade to confirm the presence of reads reporting the SNP allele from the parental chromosome that carried the mutation.

Sequencing of only the wild-type allele by chance is another potential cause of a PVV, particularly if the negative subclade is a single sample with low depth (although a minimum depth of 12x was required in the PVV identification stage). This read-based assessment also excludes this mechanism. Where assessable, the read-based LOH assessment agreed with the result from ASCAT in all but one case (**Extended Figure 5b**).

Assessment of incorrect phylogeny structure as an artefactual cause of PVVs

For all PVVs, there is an alternative phylogeny structure with which the PVV would in fact be consistent (**Extended Figure 1d**). To confirm that PVVs did not simply result from phylogeny inference errors, we counted the mutations in conflict with the structure suggested by the PVV (in most cases this is the same as the MMLD), and confirmed that they were robust. This required ensuring that the mutation calls themselves were not false positives (manual inspection of sequencing alignments) and that the set of colonies reporting the mutations was correctly assigned (**Extended Figure 1e**). In addition, we manually checked up to five such mutations for a large number of PVVs, confirming that they validated the consensus phylogeny as expected. In all but one case, there was more than 1 somatic mutation that convincingly confirmed the consensus phylogeny – thus, for these branches, there was considerably more evidence for the original phylogeny than for the alternative phylogeny suggested by the single PVV. However, in one case, there was only a single mutation confirming the consensus phylogeny, meaning that there was equal evidence for the two phylogeny structures – for this case, it was unclear which mutation confirmed the 'true' phylogeny, and which variant is caused by a persistent DNA lesion.

Inference of expected MAV mutational signatures

We aimed to identify the most likely mutagenic processes causing the MAVs observed in the bronchial epithelium and PX001 (chemotherapy-exposed HSPC) phylogenies. We started with the SNV mutational signatures that were inferred as present in each tissue from the original studies[5–12] (**Extended Figure 6b-d**, left). For the bronchial epithelium and liver, these signatures were previously extracted using a Bayesian hierarchical Dirichlet process as implemented in R package 'HDP' (https://github.com/nicolaroberts/hdp)[48]. We used the same package to extract mutational signatures from phylogeny PX001, the chemotherapy-exposed case, using mutations on individual branches as single samples. For each signature, an expected MAV signature could then be inferred as proportional to the product of the context-specific relative likelihoods of the two SNVs in each MAV, weighted by the abundance of that context in the human genome. Accordingly, we calculated the expected MAV signatures resulting from each extracted mutational signature in bronchial epithelium, liver and blood (**Extended Figure 6b-d**, right) and compared this to the observed MAVs.

Inference of expected PVV mutational signatures

For each potential alternative mechanism that may generate a PVV, a specific, predictable mutational signature is expected. Going through each in turn:

(1) *Two independent mutations at the same site.* The expected signature reflects the square of the likelihood of a given mutation at a specific trinucleotide context, weighted by the frequency of that context in the human genome. Raw mutational signatures can be converted into likelihood signatures by correcting for the trinucleotide frequency. Due to the low abundance of CpG sites in the human genome, this predominantly has the effect of demonstrating the high likelihood of C>T mutations at CpG. This likelihood signature is squared to reflect the fact that the same mutation has to occur twice at that site, before being multiplied by the trinucleotide frequencies to get the expected abundance of such events across the genome. For adult HSPC signature, this results in a signature dominated by C>T mutations at CpG sites, particularly AC̲G trinucleotides (**Extended Figure 5d(i)**). This signature is very similar to that observed for the 109 adult HSPC PVVs that are not in an orientation consistent with generation by a persistent DNA lesion, suggesting that these are predominantly caused by this mechanism (‘cosine similarity, 0.9; **Extended Figure 5e**).

(2) *Spontaneous reversion of a somatic mutation.* For a spontaneous reversion to occur, a mutation at a given trinucleotide context must, at a later time point, be followed by the reverse mutation (at the same context) e.g. C>T at AC̲G must be followed by a T>C at AT̲G. The likelihood of this occurring is proportional to the product of the likelihood of the original mutation and the likelihood of the reversion mutation. The expected signature is therefore this likelihood, multiplied by the trinucleotide frequencies. For the adult HSPC signature, this reveals a signature with a dominant peak at T>C mutations at AT̲G sites, reflecting the high likelihood of the reversion mutation (**Extended Figure 5d(ii)**).

(3) *Loss of heterozygosity, biased allele sequencing or incorrect phylogeny.* All of these mechanisms of PVV generation are agnostic to the identity of the original mutation. Therefore the expected mutational signature of such PVVs should reflect that of the overall tissue signature (**Extended Figure 5d(iii)**).

The observed PVV signature (**Figure 3c**) is clearly distinct from any of these predicted signatures (cosine similarities: 0.15, 0.3, 0.62 for independent mutations, spontaneous reversion and other mechanisms respectively). This supports the premise that these are caused by a specific mutational process.

Inference of mean lesion duration

The C>T PVVs in the HSPC phylogenies had a consistent signature and broadly consistent minimum molecular lesion durations (MMLD). We therefore hypothesised that these PVVs are caused by a single underlying lesion. The MMLD of the C>T PVVs has a median value of 21, corresponding in chronological time to ~1.3–1.5 years. However, this value may not be a good estimate of the true lesion durations for two main reasons:

(1) For each PVV, the MMLD is a *minimum* of the true lesion duration.

(2) The duration of detected PVVs is dependent on the phylogeny structure. For example, if there was only one branching region of the phylogeny suitable for capturing PVVs, any detected PVVs would have the same MMLD, defined by the phylogeny.

We designed an approximate Bayesian computation (ABC) model to account for these factors and estimate the true mean duration of the underlying lesion.

First, we created a simulation framework of persistent lesions. In this framework, simulated persistent lesions were introduced at random into an ultrametric version of the phylogeny. Before each simulation, a true mean lesion duration, t_M, was set from a prior distribution:

$t_M \sim Uniform(min=2,max=100)$

From the data, we observed that the causative persistent lesions were primarily generated in HSCs in the bone marrow, and therefore the early embryonic branches of the given phylogeny were not included. Post-embryonic branches were chosen at random, with a probability proportional to the branch length. A 'position' *p* along the selected branch was then also chosen at random from a uniform distribution, therefore defining a lesion acquisition point in the phylogeny. The simulated lesion was then assigned a molecular duration drawn from a gamma distribution with a rate parameter of $1/t_M$ and shape parameter of 1.

From here, a lesion path through the phylogeny was defined, and if the path reached a node, the daughter branch to continue along was chosen at random. At any node, the base paired opposite the lesion during DNA replication was selected as either the alternative base ('Alt') or the reference base ('Ref') with a probability of 0.54 of selecting the alternative, a value defined from the data. At the time of lesion repair, a final base of the clade was similarly decided. From this information, it can be inferred whether or not a detectable PVV has resulted.

A few shortcuts can be made: (1) no lesion acquired on a private branch will ever result in a PVV; (2) the lesion path must cross at least 2 nodes, otherwise a PVV cannot result. Therefore, in either case, the simulation is terminated, and the outcome is determined as 'No PVV'. If a lesion path does cross

≥2 nodes, the resulting base order is examined. Any base order whereby a 'Ref' base is incorporated between two 'Alt' bases will result in a PVV. Any other order will not result in a PVV. Below are some examples. As the order of the last two base incorporations cannot be determined from the phylogeny, they are enclosed within square brackets.

- Ref-[Ref-Ref] -> No PVV
- Alt-[Alt-Alt] -> No PVV
- Ref-[Alt-Ref] -> No PVV
- Alt-[Ref-Alt] -> detectable **PVV**
- Ref-Alt-[Ref-Alt] -> detectable **PVV**
- Alt-Ref-[Ref-Alt] -> detectable **PVV**

Where a PVV does result, the MMLD of the PVV is determined. For each run of the simulation, 5 million PVVs were simulated into each of the three oldest HSPC phylogenies (KX004, KX007 and KX009). These three phylogenies were selected as those with the richest clonal structure, and therefore the most detected PVVs, making them the richest in information regarding the lesion duration. At the end of each simulation, three summary statistics were recorded for each phylogeny:

1. the mean MMLD of detected PVVs,
2. the shape parameter of the best fit gamma distribution for the detected MMLDs.
3. the rate parameter of the best fit gamma distribution for the detected MMLDs.

For the actual approximate Bayesian computation inference, the summary statistics for the data were determined as follows:

- The MMLDs of all C>T PVVs in each of the three phylogenies were collated.
- Outlying values of >100 mutations were removed, as they were considered potentially caused by alternative mechanisms/ alternative lesions.
- For each phylogeny, the mean MMLD, and the shape and rate parameters of the best fit gamma distribution were calculated.

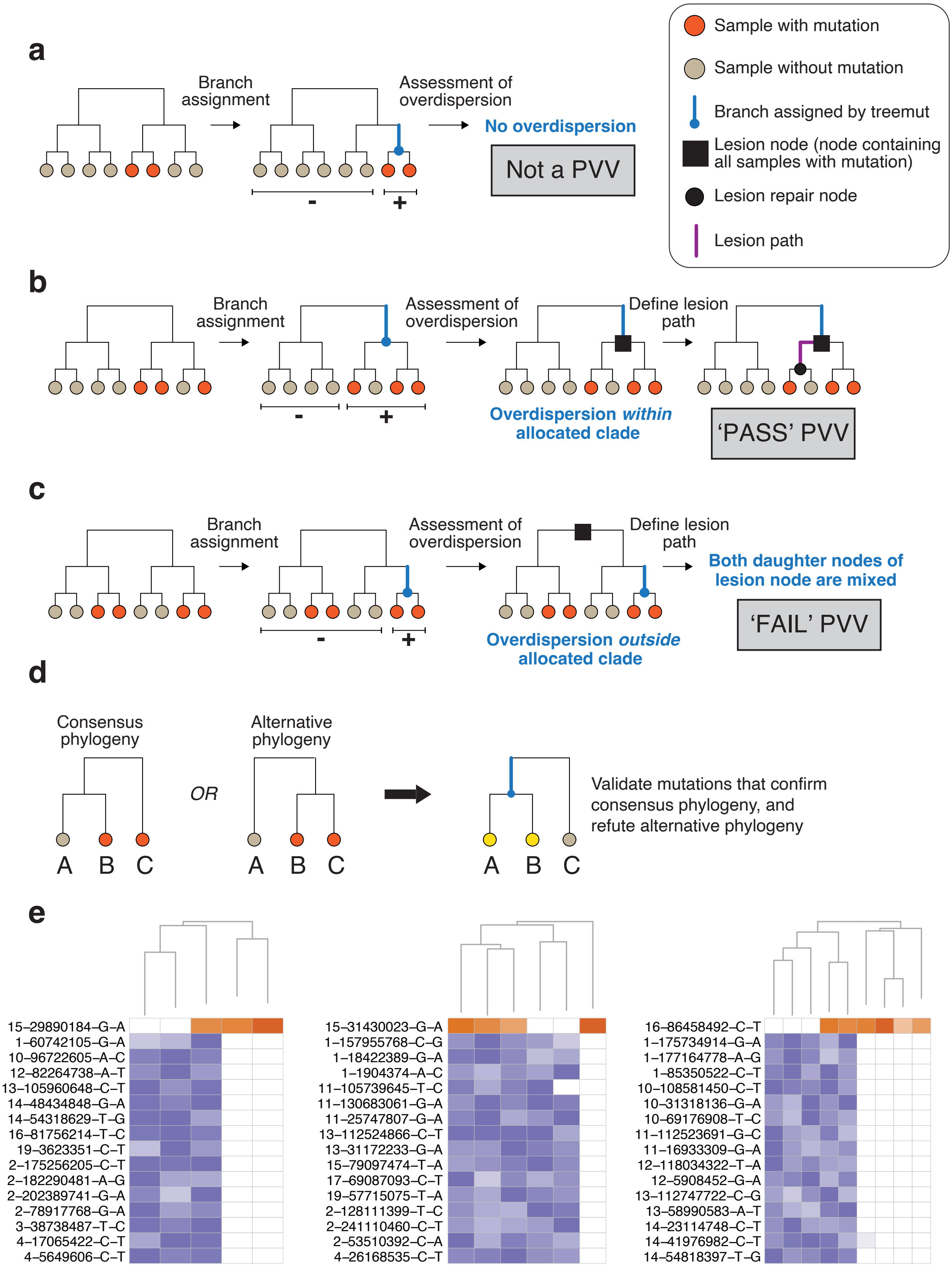
This gave a total of 9 summary statistics across the three phylogenies. We deployed the abc() function from the *R* package ‘abc’, using a neural network regression method, to infer the posterior distribution for the true mean lesion duration (**Figure 5a**). Within this method, each summary statistic is standardised by a robust estimate of the standard deviation (the median absolute deviation), and a Euclidean distance is then calculated for each set of summary statistics (for a given simulation) compared to the true set of summary statistics. The closest 2% of simulations are accepted as a first approximation of the posterior distribution. The neural network regression is an additional step to correct for the imperfect match between the accepted summary statistics and observed summary statistics, by giving greater weight to simulations that match the data more closely[49].

Detection of lesion segregation

We used the ‘calculate_lesion_segregation’ function from the R package MutationalPatterns v3.01 (https://bioconductor.org/packages/MutationalPatterns)[50] to assess for lesion segregation in each individual branch from each phylogeny. As reported in the DEN-exposed mouse model study[15], we used the ‘binomial’, ‘Wald-Waldowitz’, and ‘rl_{20}’ tests – branches were considered positive if the rl_{20} was ≥6 and at least one of the binomial or Wald-Waldowitz tests had a Benjamin-Hochberg-adjusted P <0.05.

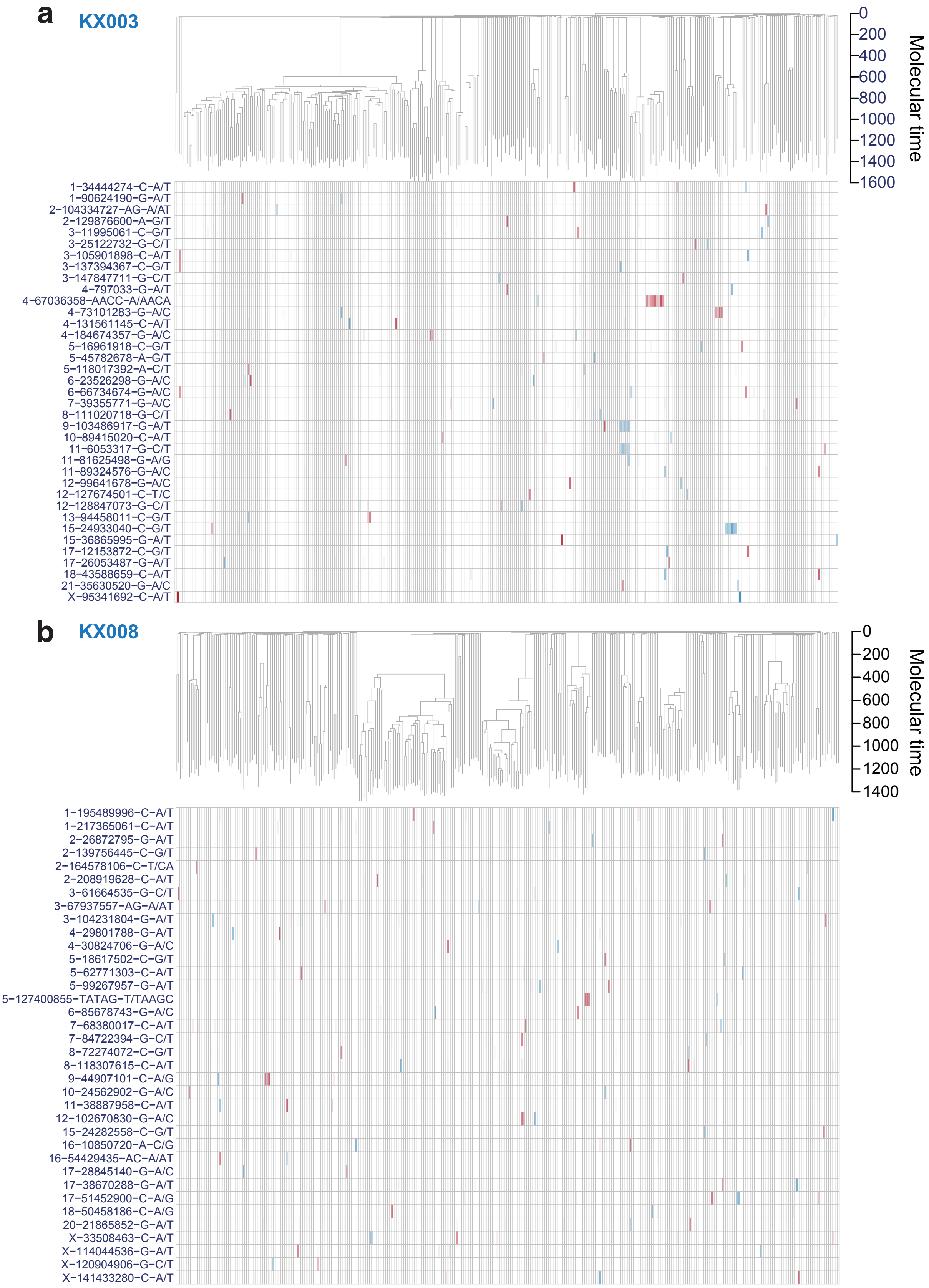
COMPETING INTERESTS

PJC, MRS and IM are co-founders, stock-holders and consultants for Quotient Therapeutics Ltd.



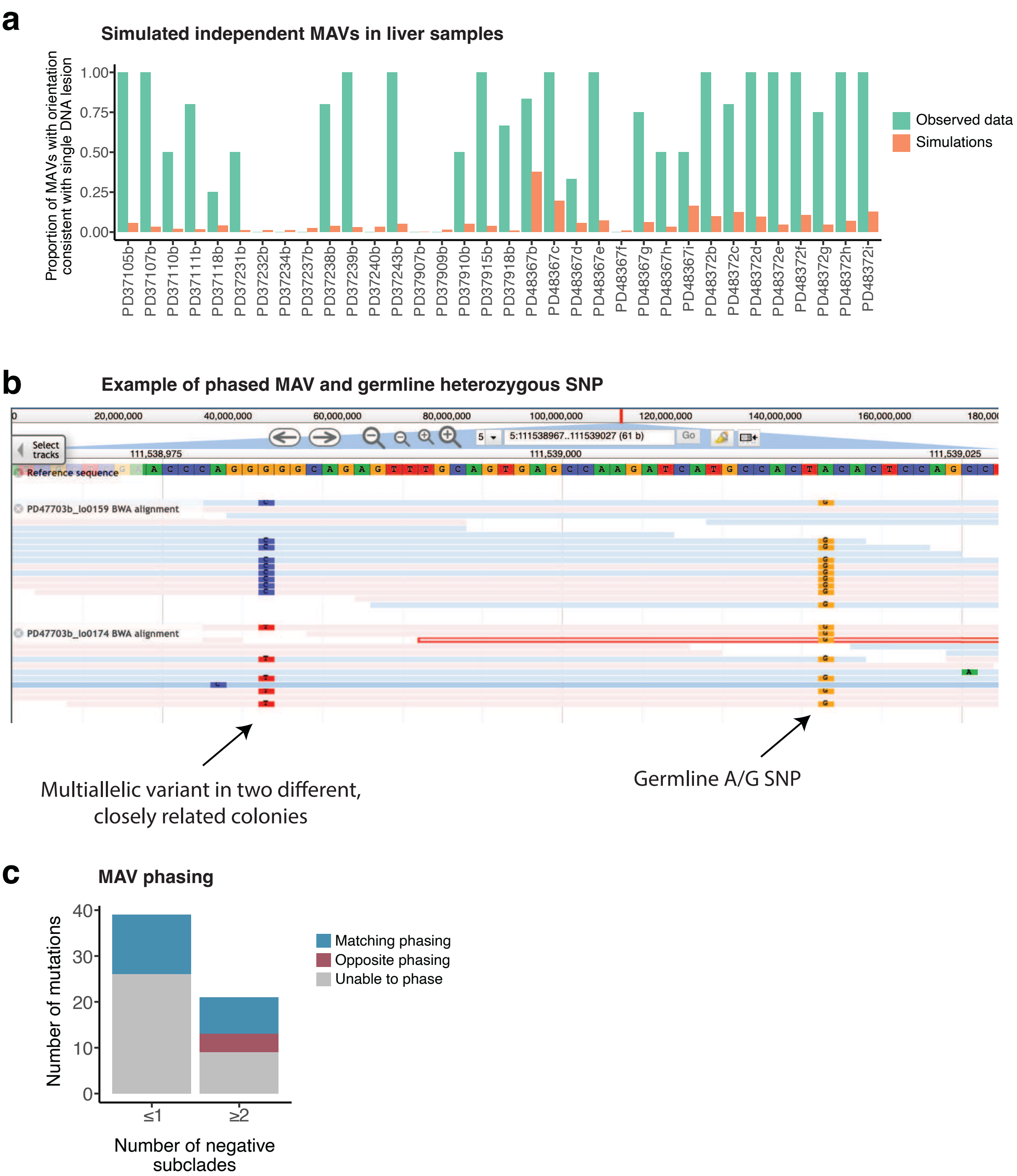
Extended Figure 1

Extended Figure 1. Detection of phylogeny-violating variants (PVVs). a, A mutation that fits the consensus phylogeny will have no overdispersion within the positive clade, or outside the positive clade. **b**, In this example, treemut() attempts to assign a branch for the mutation, even though there is no single branch assignment that fits the data. There is therefore a negative sample within the clade that should be positive with a single mutation. This orientation is consistent with a persistent DNA lesion, and a single ‘lesion path’ can be defined with a ‘lesion node’ and ‘lesion repair node’ assigned. This example would be classified as a ‘PASS PVV’. **c,** An example of a variant where no single branch assignment fits the data, but in this example the significant overdispersion is detected outside the allocated clade. However, this variant cannot be explained by a persistent DNA lesion because both daughter nodes of the lesion node show a mix of mutant and wild-type descendants – this lesion is therefore classed as a ‘FAIL’ PVV. **d,** For any given PVV, there is a reordering of the branches from the consensus phylogeny such that the PVV no longer contradicts the tree structure (alternative phylogeny). We can then assess the distribution of reads reporting the mutations that distinguish the two phylogenies across the descendant colonies for whether they support the consensus or alternative phylogeny. **e,** Three examples of PVVs showing the validation of the consensus phylogeny inferred for that patient. For each, a zoomed-in subsection of the tree is shown above the heatmap; the PVV is shown in the red colour scale; and the mutations confirming the consensus phylogeny are shown in the blue colour scale. The colour scales of the heatmap denote the variant allele fraction, with white as VAF=0. The mutations are annotated as ‘chromosome-position-reference-variant’.



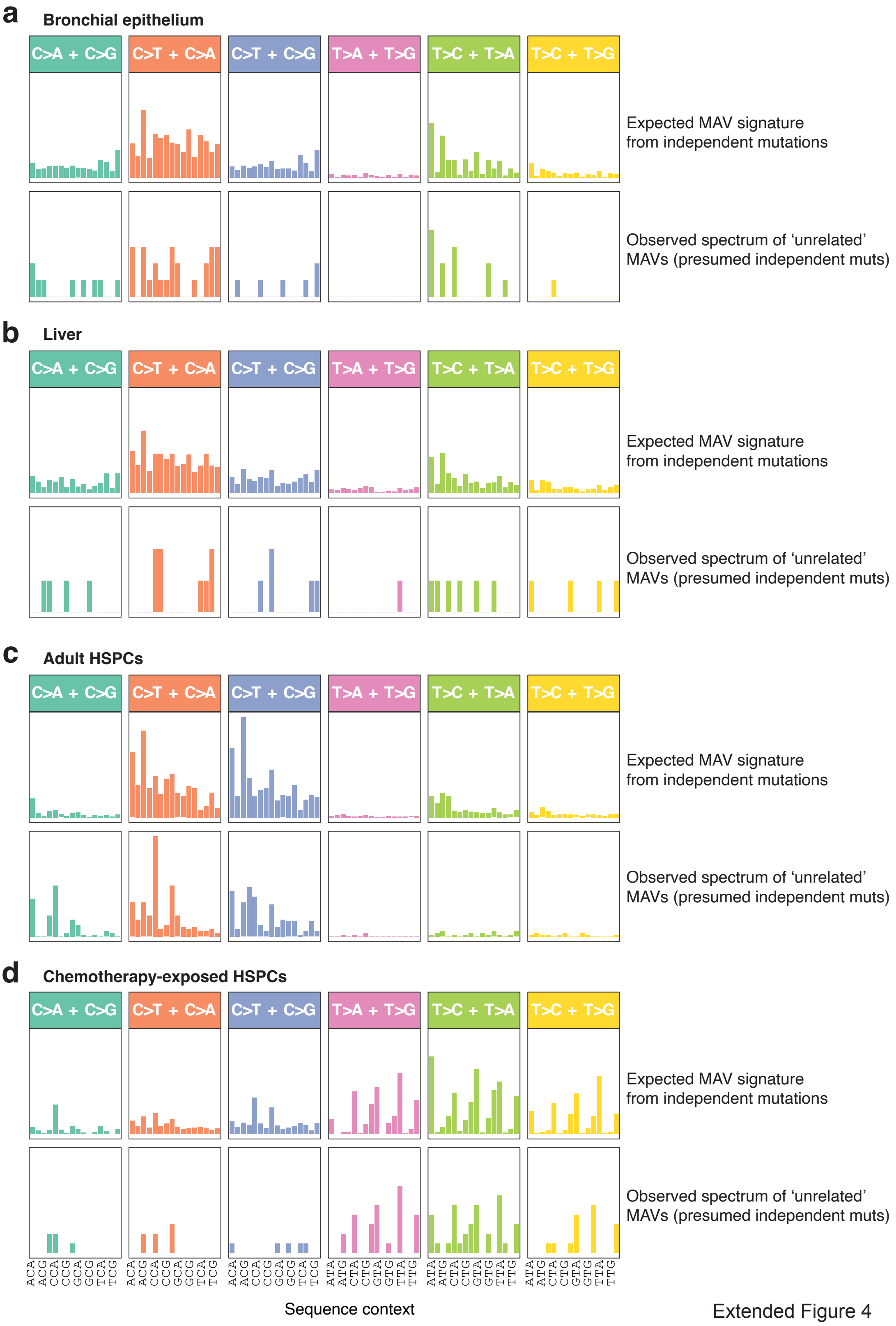
Extended Figure 2

Extended Figure 2. Examples of multi-allelic variants (MAVs) likely to have arisen by chance through independent mutations at the same locus. For KX003 (**a**) and KX008 (**b**), the phylogenetic trees are shown above the heatmap. The two alternate alleles at each locus are coloured blue and red, with colour intensity scaled by variant allele fraction, in the colonies that carry them (corresponding to the tips of the phylogenetic tree). Note that the mutations are typically far apart on the tree, suggesting that they arose through independent events. The mutations are annotated as ‘chromosome-position-reference-variant’.



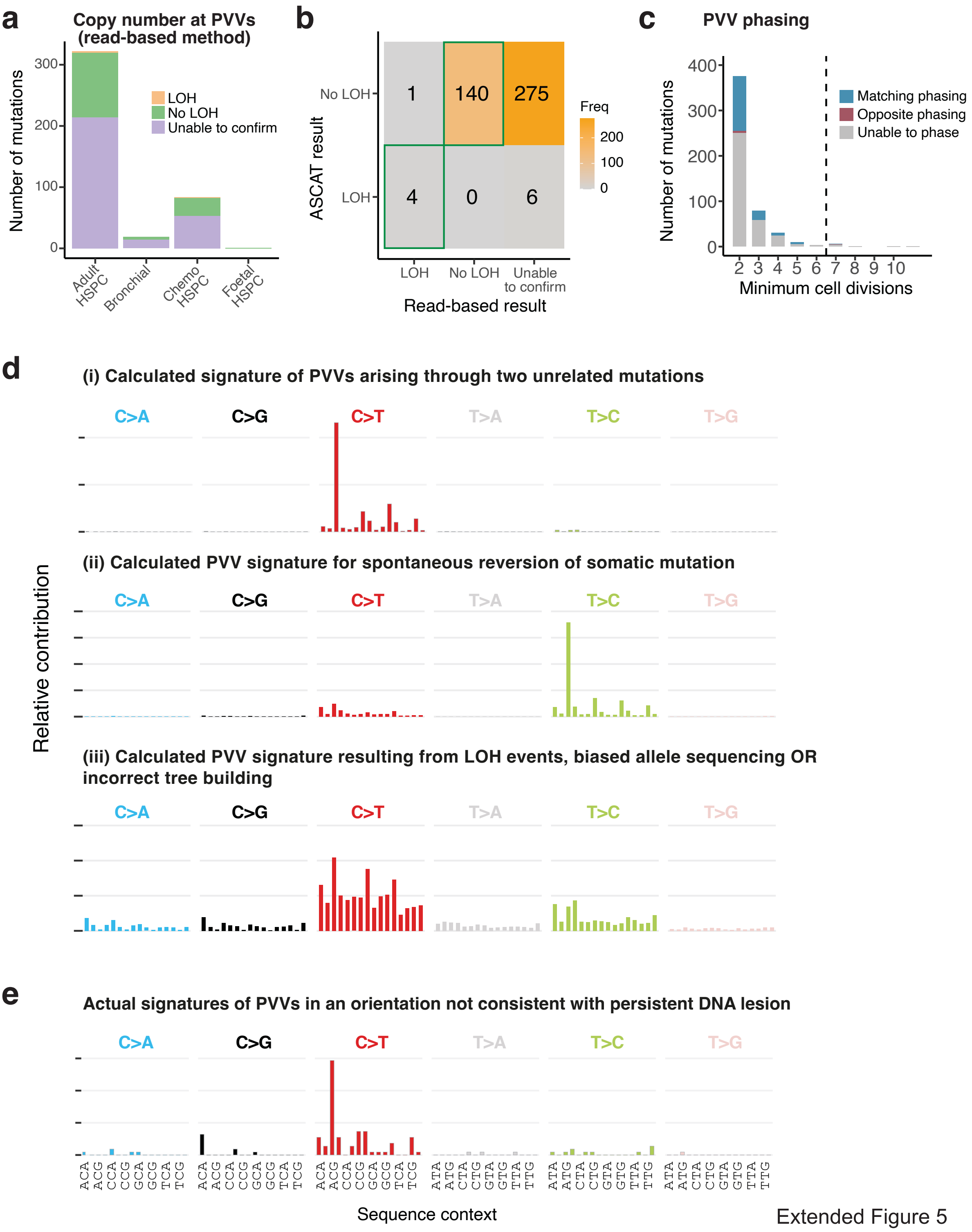
Extended Figure 3

Extended Figure 3. Validation of multi-allelic variants (MAVs). a, Barplot showing a comparison between the observed data (green) and simulated independent MAVs (orange) of the proportion of MAVs occurring in an orientation either consistent or inconsistent with arising from a single DNA lesion. **b,** A JBrowse plot showing an example of a correctly phased MAV with heterozygous germline SNP. Alignments of individual reads to the relevant section of the human genome from two different colonies that are closely related on the phylogenetic tree are shown, with pink reads denoting those mapped to the forward strand; blue reads to the reverse strand. Base calls that do not match the reference genome are shown as coloured rectangles. **c**, Phasing comparison results of 'separated' MAVs with 1 or fewer intervening negative subclades, versus those with 2 or more intervening negative subclades. In the latter group the ratio of matching to non-matching phasing is somewhat more balanced, and therefore these MAVs were excluded from downstream analysis.



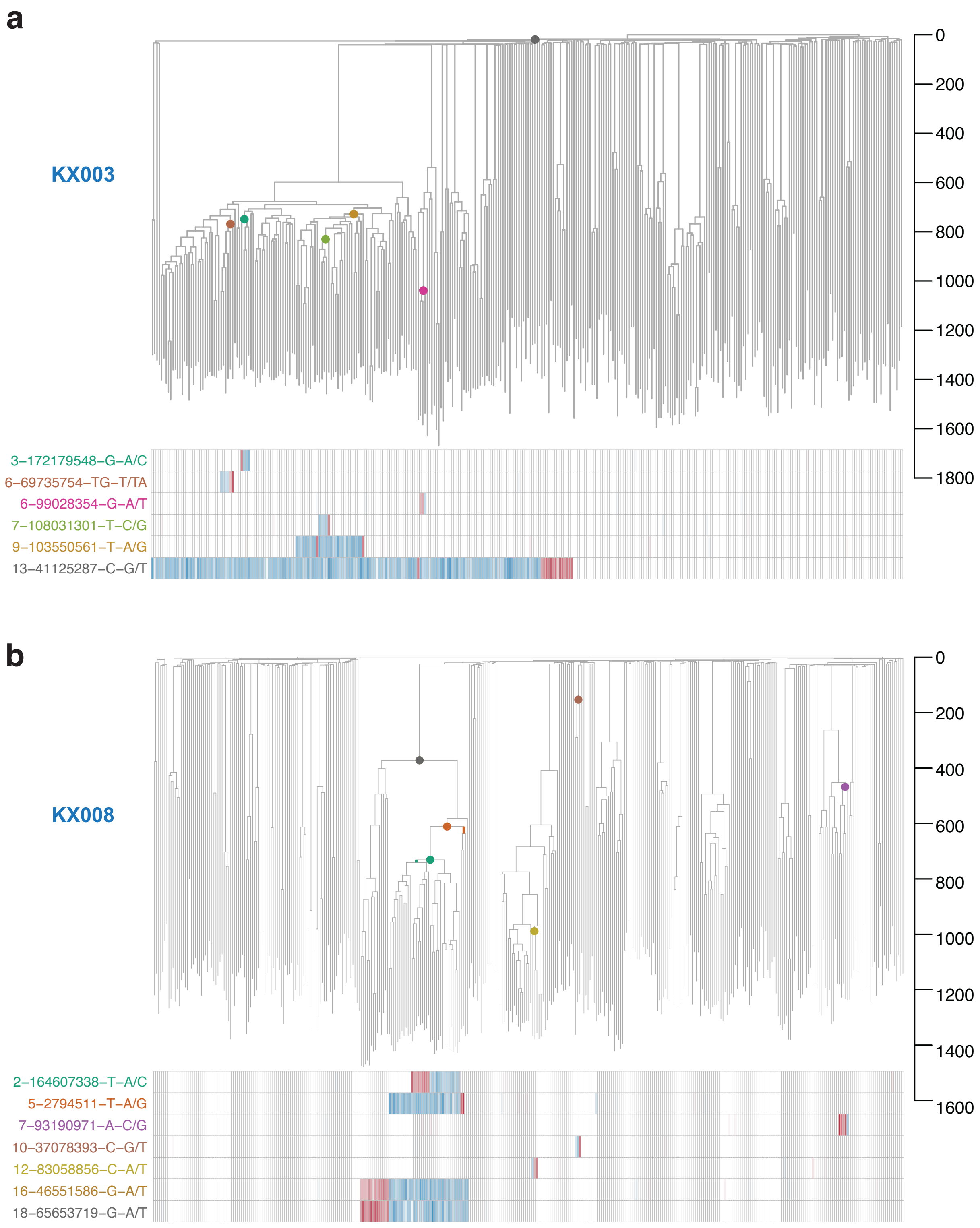
Extended Figure 4

Extended Figure 4. Expected versus observed spectrum of independent MAVs. a, Barplot showing the expected signature of MAVs for bronchial epithelium for two mutations occurring at the same locus independently by chance (top) and the observed spectrum for 'unrelated' MAVs (those that had an orientation on the phylogeny that was incompatible with arising from a single lesion). **b,** As for **a**, but for MAVs in the liver samples. **c**, As for **a**, but for MAVs in the adult HSPCs. **d**, As for **a**, but for MAVs in the chemotherapy-exposed HSPCs.



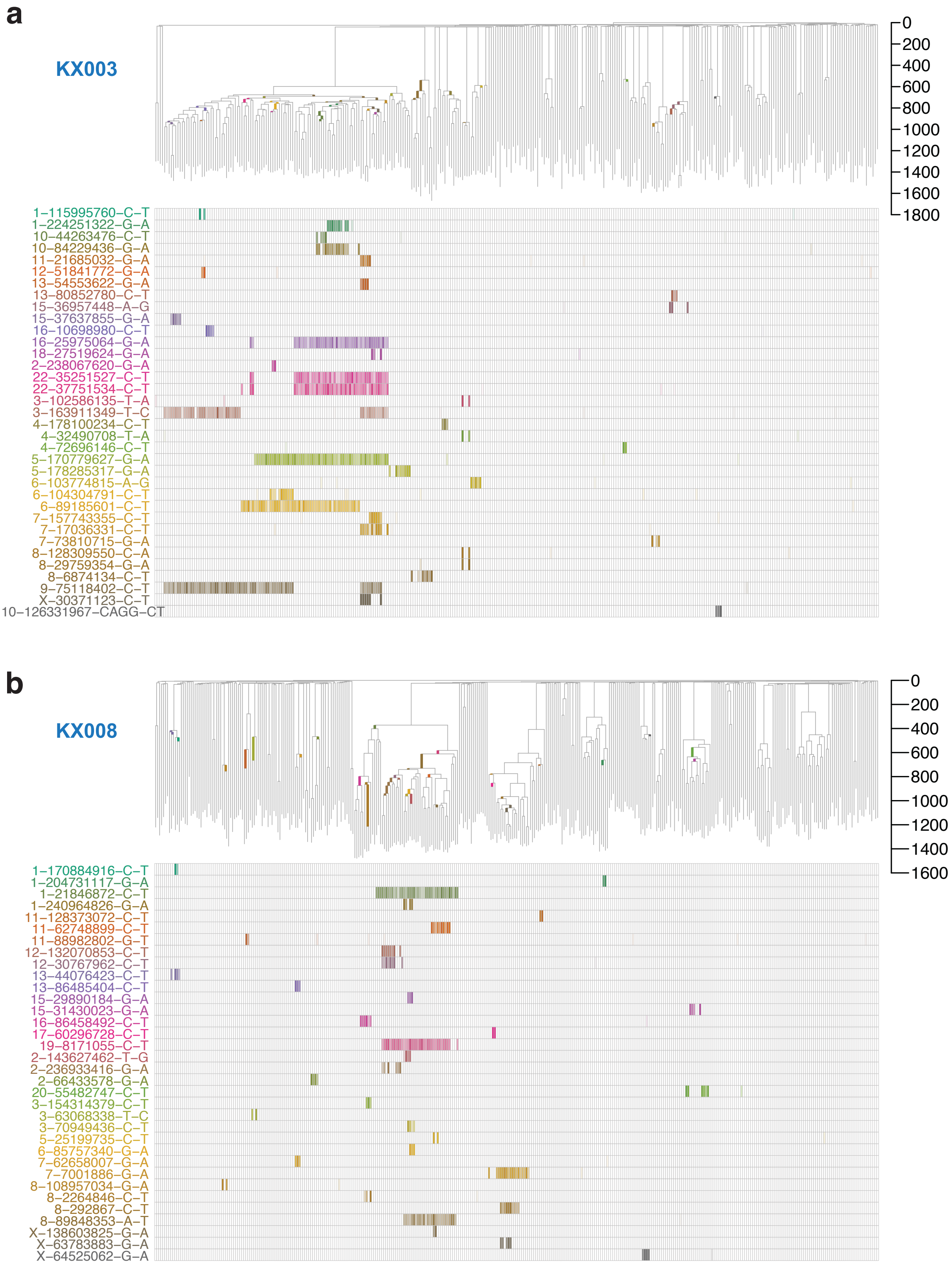
Extended Figure 5

Extended Figure 5. Validation and spectrum of PVVs. a, Barplot showing the results of copy number analysis using a reads-based method to ensure small (<1kb) deletions were not resulting in artefactual PVV calls. **b,** Contingency graph showing comparison of copy number analyses by ASCAT (y axis) versus the reads-based method (x axis). **c**, Stacked barplot showing the minimum number of cell divisions through which the PVV-causing lesion must have persisted unrepaired, coloured by phasing. **d,** Calculated blood PVV mutational signatures expected for those occurring (i) as two independent events, (ii) spontaneous reversion events, or (iii) other alternate mechanisms. **e,** Actual mutational signature of 'FAIL' PVVs from the adult HSPC phylogenies, namely variants that were in an orientation inconsistent with a single DNA lesion – note the resemblance to **d**, panel (i).



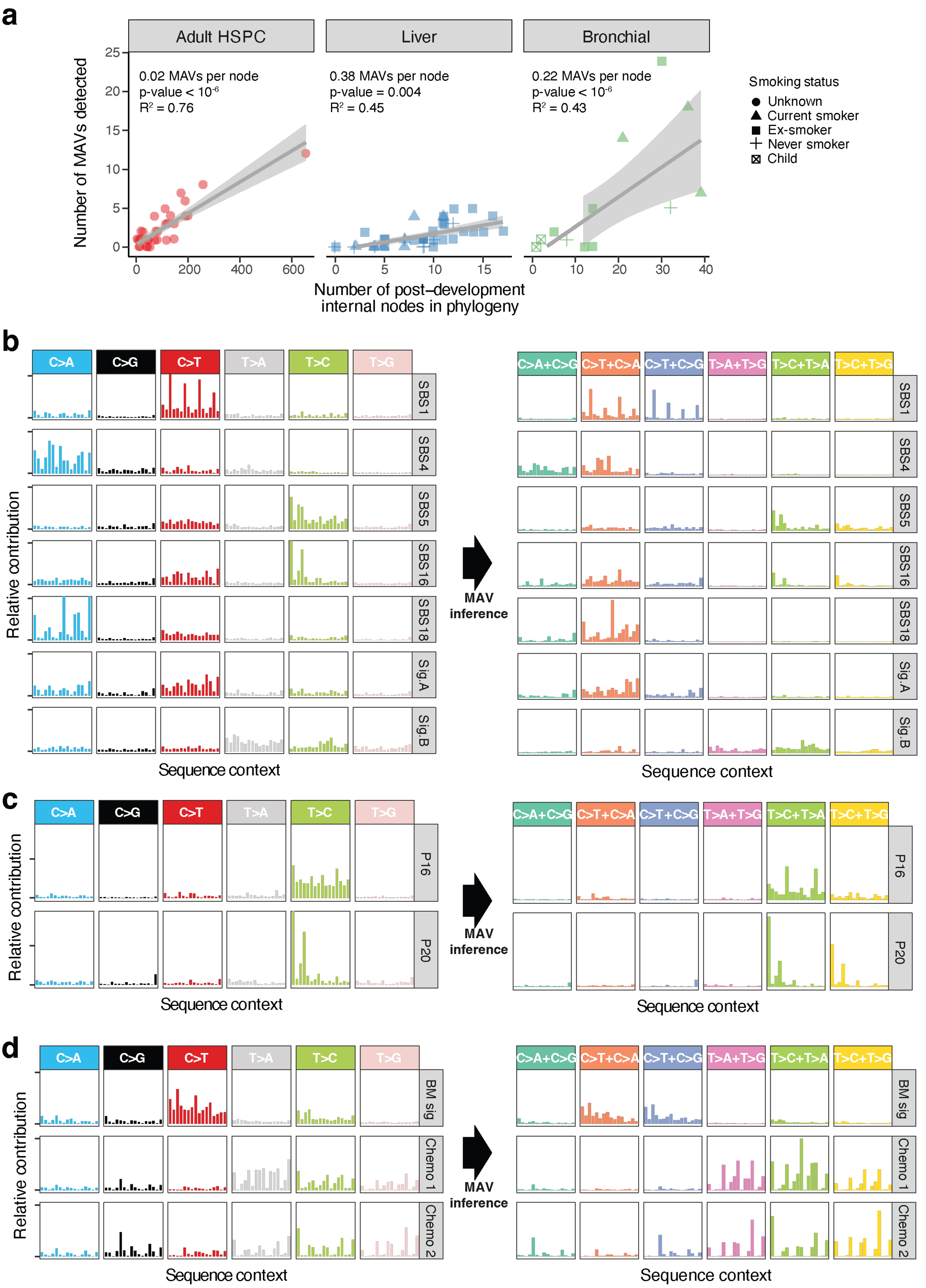
Extended Figure 6

Extended Figure 6. Examples of multi-allelic variants (MAVs) that occurred close together on phylogenetic tree and in an orientation consistent with a persistent DNA lesion. For KX003 (a) and KX008 (b), the phylogenetic trees are shown above the heatmap. The two alternate alleles at each locus are coloured blue and red in the heatmap, with colour intensity scaled by variant allele fraction, in the colonies that carry them (corresponding to the tips of the phylogenetic tree). The node at which the DNA lesion must have existed is marked on the phylogenetic tree with a coloured circle; colours correspond to the colours of the mutation annotation in the heatmap. For separated MAVs, the minimal lesion persistence path is highlighted in the same colour along the relevant branches. Note that the mutations are typically close together on the tree, suggesting that they arose from the same lesion. The mutations are annotated as ‘chromosome-position-reference-variant’.



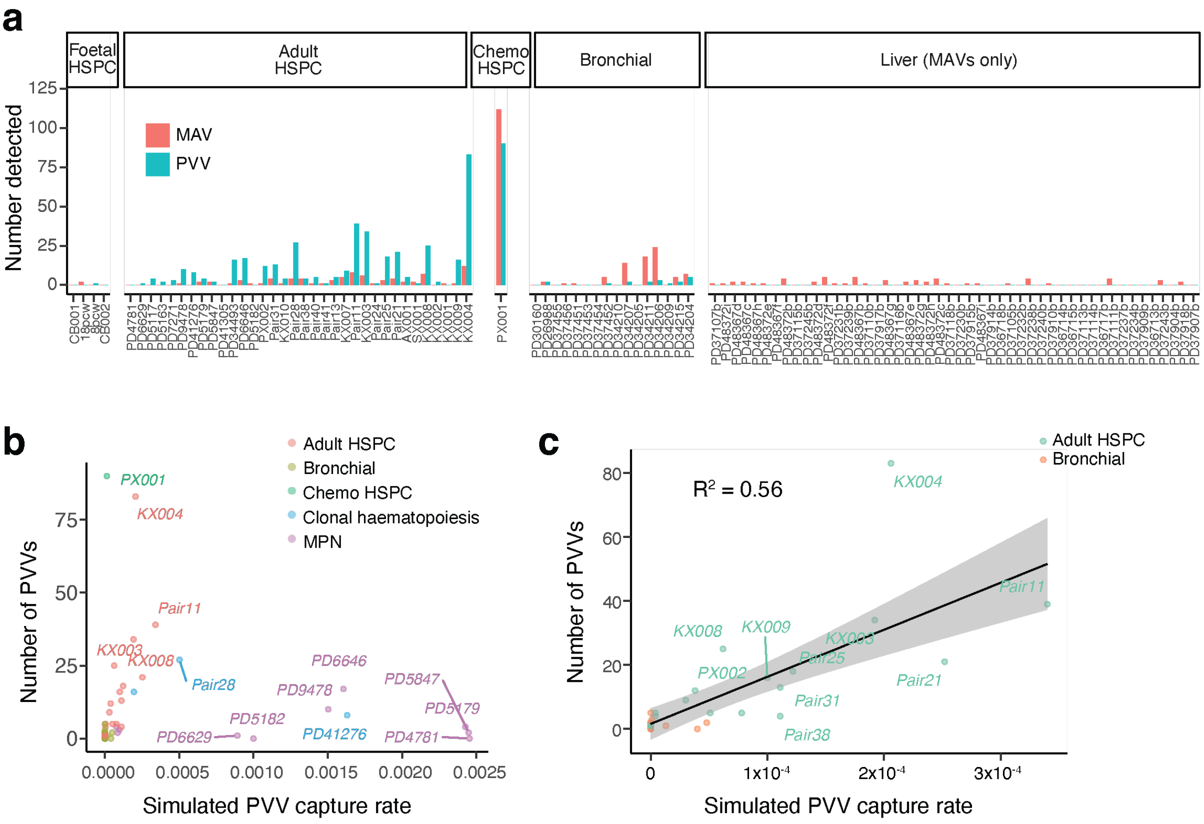
Extended Figure 7

Extended Figure 7. Examples of phylogeny-violating variants (PVVs) that occurred close together on phylogenetic tree and in an orientation consistent with a persistent DNA lesion. For KX003 (a) and KX008 (b), the phylogenetic trees are shown above the heatmap. The individual PVVs each represent a row of the heatmap, with colour intensity scaled by variant allele fraction in the colonies that carry them (corresponding to the tips of the phylogenetic tree). The minimal lesion persistence path is highlighted in the same colour along the relevant branches of the phylogenetic tree. Note that the mutations are typically close together on the tree, but with an interspersed subclade of wild-type colonies that enables identification as a PVV. The mutations are annotated as ‘chromosome-position-reference-variant’.



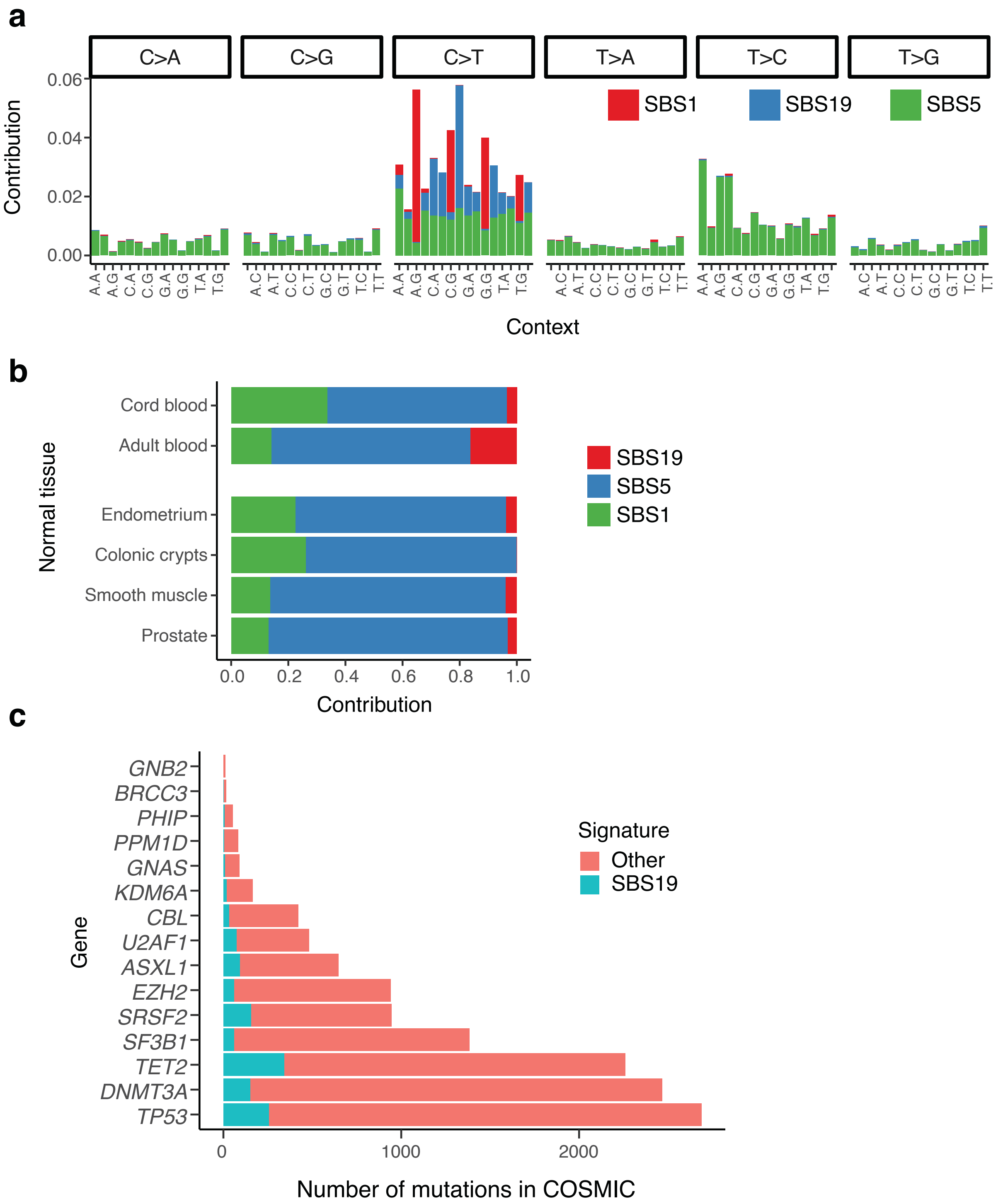
Extended Figure 8

Extended Figure 8. Number and expected spectrum of MAVs. a, Scatter plot showing the numbers of MAVs in each phylogeny against the total number of post-development nodes within that phylogeny, separated by tissue. Smoking status (where known) is indicated by the shape. Grey lines display the correlation between these values by univariate linear regression, with the shaded area showing the 95% confidence interval of this relationship. **b,** On the lefthand side are the extracted 96-profile SBS signatures from the bronchial epithelium as per Yoshida et al[10]. On the righthand side is the calculated MAV signature expected for that signature. **b,** As in **a**, but for the liver samples. Only two of the 30 extracted liver signatures[12] are displayed. These are the signatures that most closely match the observed MAV signature. **c,** As in **a**, but for the chemotherapy-exposed blood samples. Note that the 'BM signature' is the same signature extracted for the normal adult blood phylogenies. BM, bone marrow; SBS, single base substitution.



Extended Figure 9

Extended Figure 9. Statistical power to detect PVVs. a, Barplot showing the number of MAVs and PVVs detected in each phylogeny. Phylogenies are divided by their tissue type, and within each type, phylogenies are ordered by number of samples. The 331 MAVs and 501 PVVs were unevenly distributed across samples. **b,** Relationship between the simulated PVV capture rate and the observed number of PVVs for the whole dataset. Individuals who have a large clonal expansion (MPN, in purple, and those with clonal haematopoiesis, in blue) have lower than expected numbers of PVVs than the simulations would expect. **c,** Scatter plot of the simulated PVV capture rate and observed numbers of PVVs, excluding samples with a single large clonal expansion or chemotherapy exposure. The black line displays the correlation between these values by univariate linear regression, with the shaded area showing the 95% confidence interval of this relationship.



Extended Figure 10

Extended Figure 10. SBS19 causes 19% of mutations in HSPCs, including driver mutations. a, Stacked barplot showing the signature decomposition of mutations in normal HSPCs. Bars are grouped by the 6 mutation types, and within each by the 16 base contexts comprising the base before and the base after the mutation. Each context has three stacked bars, denoting the fractional contribution of each signature to mutations in that specific context. **b,** Stacked barplot showing the fraction of mutations attributable to SBS1, SBS5 and SBS19 in normal cells from different organ systems. **c,** Stacked barplot showing the number of mutations attributable to SBS19 in driver genes for myeloid cancers from the COSMIC database.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.xlsx
- TableS2.xlsx