nature portfolio

https://doi.org/10.1038/s41586-021-03974-6

Supplementary information

Convergent somatic mutations in metabolism genes in chronic liver disease

In the format provided by the authors and unedited

Supplementary Information

Convergent somatic mutations in metabolism genes in chronic liver disease

Stanley W. K. Ng (1); Foad J. Rouhani (1,2); Simon F. Brunner (1); Natalia Brzozowska (1); Sarah J. Aitken (3,4,5); Ming Yang (6); Federico Abascal (1); Luiza Moore (1); Efterpi Nikitopoulou (6); Lia Chappell (1); Daniel Leongamornlert (1); Aleksandra Ivovic (1); Philip Robinson (1); Timothy Butler (1); Mathijs A. Sanders (1,7); Nicholas Williams (1); Tim H. H. Coorens (1); Jon Teague (1); Keiran Raine (1); Adam P. Butler (1); Yvette Hooks (1); Beverley Wilson (1); Natalie Birtchnell (1); Huw Naylor (2); Susan E. Davies (4); Michael R. Stratton (1); Iñigo Martincorena (1); Raheleh Rahbari (1); Christian Frezza (5); Matthew Hoare (3,8)*; Peter J. Campbell (1,9)*.

Table of Contents

Supplementary Note 1: Considerations of experimental design and analysis

Rationale for hierarchical experimental design

A key feature of our experimental design is the choice to perform whole genome sequencing on multiple microdissections per patient sample (typically 30-50 per patient). This is known as a hierarchical design, because we have multiple levels of one variable (here, individual microdissections) nested within levels of another variable (patient samples). For sequencing of normal tissues, there is a strong rationale for adopting such an experimental design, as described below, but statistical analyses of the data emerging need to account for the fact that the repeated measures within an individual are not independent. We discuss these statistical considerations in the next sections.

The rationale for a hierarchical experimental design in sequencing normal tissues is based on the following points –

- *Normal tissue is polyclonal* – As we and others have demonstrated across all normal organ systems studied to date[1–9], including liver[10–13], normal tissue represents a patchwork of independently evolving clones. Indeed, for any given solid tissue, there are typically many millions of independent clones in an adult human. While in either cancer genome sequencing (where all tumour cells derive from a single ancestral cell) or germline genome analyses (where the interest is inherited genetic variation), a single genome per person is sufficient, studying somatic mutations in normal tissues benefits from analysing more than one of the millions of independent clones.
- *Liver disease generates considerable within-patient, clone-to-clone variation* – In our earlier study, we demonstrated that mutation burden and mutational signatures can vary by orders of magnitude among even neighbouring clones within the same patient[10]. In the current study, we show that this clone-to-clone variation extends to include telomere lengths and distribution of driver mutations. Two corollaries follow from this – (1) we fail to characterise a given patient's hepatic landscape unless we assess many clones; (2) we continue to increase statistical power for detecting driver mutations by sequencing multiple samples per patient because they represent independent evolutionary experiments exposed to the same microenvironment.
- *Convergent evolution is frequent in normal tissues* – For a somatic mutation to generate a systemic disease, enough cells within the organ system must carry the mutation. This requires either that a single clone expands to a massive population size (as occurs in cancer) or that the mutation evolves many times independently within the same patient's organ. Several previous studies have demonstrated that parallel evolution commonly occurs in ageing normal tissues[1,4–6,9], and in non-malignant disease[14–16]. Such convergent evolution provides strong clues as to the nature of the selective pressure on the tissue, and the likely systemic phenotypic consequences of somatic mutation for an organ's function.

In summary, then, characterising the landscape of normal tissues requires a different approach than, say, genome sequencing of, say, cancer or inherited predisposition. We do not sacrifice much statistical power for detecting drivers by sequencing multiple microdissections per patient, because clones evolve in parallel. Furthermore, we cannot comprehensively describe a given patient's tissue unless we measure clone-to-clone variation within that patient.

Statistical power in hierarchical designs for sequencing of normal tissues

The statistical power of genomics studies is afforded by the number of samples sequenced – in cancer genome sequencing or germline GWAS-type studies, where one genome per individual is generated, the sample size is simply the number of patients studied. In hierarchical designs such as those undertaken here, the level of within-patient correlation determines how much of a decrease in statistical power results from the repeated measurements. To see this, consider the two extremes – every sample from a patient has exactly the same results versus there is zero correlation among samples from a patient. In the former scenario, no new information accrues from repeated sampling of the patient and therefore the effective sample size becomes only the number of patients analysed, not the number of microdissections. In the latter scenario, every new sample from a given patient adds equivalent amounts of new, independent information and therefore the effective sample size is the number of microdissections sequenced, not the number of patients. In reality, correlation among microdissections from a given patient falls between 0 and 1, and so the effective population size falls between the number of patients analysed and the number of microdissections.

For most solid organs, including liver, cellular mobility is limited, and therefore clones are constrained spatially. This means that the experimentalist can, to an extent, control the level of observed within-patient correlation by varying how close together microdissections are taken. Microdissections that are close together will be more likely to be clonally related (and therefore more correlated). While this decreases statistical power for, say, identifying new genes under positive selection, it allows greater resolution of phylogenetic trees. In the study undertaken here, we deliberately varied the spatial distances between microdissections – some very close together (same x,y-region of adjacent z-sections, separated by ~20μm); others in completely different cirrhotic nodules. The closely adjacent microdissections provided two purposes – first, they allowed us to demonstrate the high sensitivity and precision of mutation calling, as described in our first paper[10]; second, we were able to time key genomic events, such as driver mutations or emergence of mutational signatures, by placing them on high-resolution phylogenetic trees.

As a rule of thumb, we can obtain a rough approximation for our effective sample size for, say, inferring positive selection by calculating the number of unique mutations called as a fraction of the overall number of mutations. This can be derived by summing the branch lengths for all phylogenetic trees. Across all patients, we identified 1,322,612 unique somatic substitutions, with 1,946,613 called overall – this means that from the 1586 microdissections, we have the equivalent of ~1078 (68%) unique samples sequenced. From published power calculations for identifying cancer genes[17], this effective sample size equates to a power of ~90% for detecting a significant excess of mutations in 90% of genes mutated in 2% of clones.

Statistical analyses of the genomic landscape of normal tissues

A key challenge in analysing data from hierarchical designs is to use statistical models that appropriately manage the within-patient correlation. We exemplify this challenge by considering the problem addressed in the manuscript of modelling factors influencing telomere length in chronic liver disease. One facile approach to this question would be to generate a single summary statistic per patient (such as mean telomere length) and then compare these between patients with cirrhosis and

normal controls (with, say, a Student's t test). Such a comparison has a sample size, and therefore statistical power, equal to the number of patients, which as discussed in the previous section means we are discarding much useful information by collapsing the rich clone-by-clone information on telomere lengths into a single patient-level statistic. A second facile approach would be to treat each microdissection as an independent measurement of telomere lengths, and compare the observed telomere lengths for all microdissections from patients with cirrhosis versus all those from healthy controls (with, say, a Student's t test). This has the problem that telomere lengths of microdissections from within a single patient are not independent (see **Extended Figure 12A**), and this approach leads to overly precise estimates of the differences between conditions.

Repeated measures experimental designs are commonly encountered in biological studies, and a range of methods have been developed to explicitly model correlations within levels of 'nuisance' variables while simultaneously estimating effects from variables of interest – so-called 'mixed effects models'. In our scenario of sequencing multiple clones of normal tissue, a particularly useful approach is to base inference around the structure of the phylogenetic tree, since this provides an objective measurement of the expected correlation structure among microdissections. Such models are well established in the field of quantitative genetics, where pedigrees or species trees are studied[18,19].

All of the statistical analyses reported in the manuscript rigorously model and correct for the dependence structure imposed by the phylogenetic tree. The specific details for each analysis are as follows –

- *Reconstruction of phylogenetic trees* – Examples of high-resolution phylogenetic trees for each patient's liver sample are shown in **Figures 1 and 3**. To generate the trees, we used a multidimensional Dirichlet process clustering algorithm that essentially takes the observed variant allele fraction of every mutation in every microdissection from that patient, and identifies the unique clusters that emerge. From these clusters, we then built phylogenetic trees – the logic underlying the tree construction has been described and formalised in our previous papers[10,20], and has been adopted by other groups including the pan-cancer PCAWG consortium[21]. Having reconstructed the phylogenetic trees, every mutation was then placed onto a single branch through its most likely cluster membership. This means that any given mutation, which might be found in multiple microdissections because those regions shared a common ancestor, will be located in one, unique position in the phylogenetic tree, and therefore only 'counted' once.
- *Inference of recurrently mutated genes* – For this, we used an extensively tested and validated algorithm we originally built for inference of driver genes in cancer[22,23]. While normally this would take as input somatic mutation calls from whole genomes, we needed to avoid double-counting of mutations. The input we therefore used were the sets of mutations assigned to each branch of the phylogenetic tree – this means that each mutation is only counted once, in the ancestor in which it occurred, rather than the multiple times it is called in different, clonally related microdissections.
- *Inference of factors influencing telomere lengths* – Here, the clonal relationships among different microdissections make inference particularly challenging, largely because telomere lengths are heritable (the telomere length of the mother cell will be passed on to the daughter cells with some drift). To address this, we used a Bayesian linear mixed model specifically designed for inference on phylogenetic trees[18]. In these models, the random effect design matrix is a 'relatedness' matrix that captures the degree of shared ancestry between all pairs of microdissections. The analysis then estimates and controls for the extent to which variance in telomere lengths can be explained

by clonal relationships, while simultaneously estimating the contributions of other effects such as age and presence of disease.

- *Inference of mutational signatures* – Here, we used Hierarchical Dirichlet processes[24]. As for inferring recurrently mutated genes, each mutation was placed onto a specific branch of the phylogenetic tree so that it was only counted once in the analysis. The 'hierarchical' feature of the hierarchical Dirichlet process then enabled us to group these individual branches as coming from specific samples. The statistical analysis then explicitly allows for different liver samples to have different intensities of mutational signatures, and individual microdissections within each sample to vary around that patient's average distribution.

Supplementary Note 2: Inference of parallel evolution of hotspot mutations

For the genes other than *FOXO1* reported here, including *CIDEB*, *GPAM* and *ACVR2A*, the mutations observed were different from one another. For this reason, the inference of both the number of such variants and where they should be placed on the phylogenetic tree is straightforward and would follow the same logic for placing passenger mutations uniquely (described in Materials and Methods).

A major assumption in the logic for building somatic mutation phylogenies in cancer (and normal tissues) is the 'infinite sites' assumption[25] – that each mutation only occurs once during the somatic evolution of clones under assessment. For passenger mutations, and for non-hotspot driver mutations, this assumption is robust. For example, in the liver microdissections analysed here, we typically observed 1000-3000 somatic mutations per clone – these are scattered reasonably randomly across a genome ~$3x10^9$ base-pairs in size, meaning that the chance that any specific base is mutated is <1/1,000,000. Essentially, 3 gigabases can be considered an infinite number of sites at the mutation burdens we observe to validate this assumption.

The assumption is no longer valid for hotspot mutations, however. Here, the position of the mutation is not random, and there is strong selective pressure to generate the identical base change in different clones independently. For the *FOXO1* mutations described in the manuscript, the predominant variant is the S22W hotspot mutation, which is generated by an identical base change. We have to decide among four possible explanations for this pattern:

1. The base-position is subject to high rates of sequencing or alignment error and its apparent high prevalence is an artefact;
2. The hotspot mutation occurred once during the somatic evolution, and the observed VAFs across microdissections are consistent with it being placed on a single branch of the phylogenetic tree;
3. The hotspot mutation occurred early in embryogenesis and the patient shows congenital mosaicism for the variant;
4. The mutation has evolved in several clones independently (convergent evolution).

From the arguments outlined below, the first three explanations are not plausible, leading to the conclusion that the hotspot mutations evolved independently many times over across different hepatocyte clones within our patients.

<u>Assessment of sequencing and alignment errors at hotspot</u>

The first possible explanation, that high rates of sequencing or alignment errors have generated the variant in multiple samples, is not plausible. Although massively parallel sequencing is subject to characteristic error profiles, the algorithms we have used in this manuscript for mutation-calling are well-established, well-validated pipelines. They have been built to explicitly correct for variation in sequencing error profiles across the genome. In the recent Pan-Cancer Analysis of Whole Genomes (PCAWG) consortium, a head-to-head validation of >10 variant-calling algorithms was performed at Washington University[26]. This independent, blinded analysis revealed that the algorithms used here had very high precision for calling somatic substitutions (>99%). Visual inspection of reads and alignments reporting

the *FOXO1* S22W show that the reads are high quality with few other mismatches and high mapping quality scores (**Extended Figure 4**).

In our earlier publication describing somatic mutations in non-cancerous liver[10], we also undertook an extensive assessment of our mutation-calling approaches. It is less straightforward than for cancer because of the polyclonal nature of normal tissue, but we essentially used microdissections of the same x-y co-ordinates of adjacent z-sections from the same liver biopsy. These microdissections were then processed independently through library production, sequencing and mutation-calling. Because clones occupy regions in 3-dimensional space, this was equivalent to sequencing the same clone multiple times, and we could estimate precision of somatic mutation calls as >95%, even in this more challenging scenario[10]. This is indeed why 43 microdissections report the S22W mutation, but we count only 24 clones – an average of ~2 microdissections per clone report the same mutation. The microdissections reporting the variant are always clustered in physical space (**Extended Figure 4**). This is an accurate form of validation, because of course we have the additional supporting evidence from the hundreds to thousands of other shared passenger mutations on the same branch of the phylogenetic tree. To spell the argument out, while it might be conceivable for some sequencing artefact to manifest as a shared false positive call at one genome position in adjacent microdissections, it is inconceivable that hundreds of such sporadic sequencing artefacts would so exactly match.

Finally, we can evaluate the error profile of sequencing reads across this particular base in the genome in tissues other than liver. In our recently published paper using whole genomes to study normal endometrium, we used exactly the same microdissection, sequencing and mutation-calling pipeline as used here[6]. The reference base for the *FOXO1* mutation is a G; the S22W mutation changes this to a C. In the ~400 genomes from the endometrium paper, there were 10,803 reads reporting the G reference, versus zero reads reporting a C at this genomic position. Compared to the 43 separate microdissections reporting the S22W mutation, typically with variant allele fractions of 10-40%, it is clear that sequencing or alignment errors cannot explain our finding.

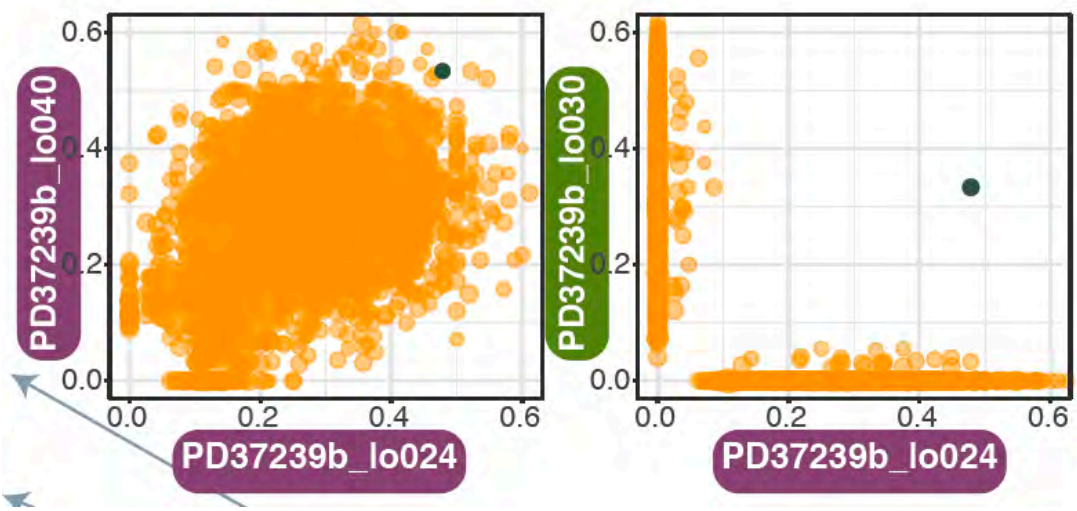
Placement of the *FOXO1* hotspot mutations on the phylogenetic trees

For a hotspot mutation such as the S22W mutation in *FOXO1* (and, say, *PIK3CA* mutations in normal endometrium[6,27]), the only means to distinguish whether a variant is 'identical by descent' (occurred in a single ancestral line) or by 'parallel evolution' (occurred multiple times across independent lines) is to use the somatic mutations in the rest of the genome.

This distinction is illustrated in **Extended Figure 4**, with an insert pasted below (**Supplementary Figure 5**). In a situation where two microdissections share a *FOXO1* S22W mutation (dark green points, **Extended Figure 4**) because they are clonally related, we also see many other passenger mutations shared between the two microdissections (pale orange points). This is shown in the scatterplot on the left in the insert below – many mutations present in the microdissection along the x-axis also have many reads reporting that variant in the microdissection shown on the y-axis.

In contrast, when two clones acquire the same hotspot mutation independently, there will be no other shared mutations since they are clonally unrelated. This is exemplified by the pair of microdissections shown in the scatterplot on the right below – here the *FOXO1* S22W mutation (dark green point) is present at high variant allele fraction in both samples, but *all other mutations* (pale orange points) in the two microdissections lie along the x- or y-axes (allowing for occasional sequencing errors), indicating that they are clonally unrelated.

Further evidence for the independent acquisition of *FOXO1* S22W mutations in multiple clones within the same patient comes from the sampling of all 8 anatomical segments of the liver in two patients included in the study. We found *FOXO1* S22W mutations in 3 separate segments in one of these patients and in 4 separate segments in the other – given the many centimetres of liver separating these segments and their very early foetal divergence, this argues strongly for convergent acquisition of the mutations.



Supplementary Figure 5. Insert from **Extended Figure 4** showing scatterplots of variant allele fraction for somatic mutations called in pairs of microdissections from the same patient. Passenger mutations are shown in pale orange, while the *FOXO1* S22W mutation is shown in dark green. The scatterplot on the left represents two clonally related microdissections, where the majority of mutations present in one microdissection have a high fraction of reads reporting the mutation in the other. In the scatterplot on the right, the only shared variant is the *FOXO1* mutation, indicating that the two microdissections are clonally unrelated.

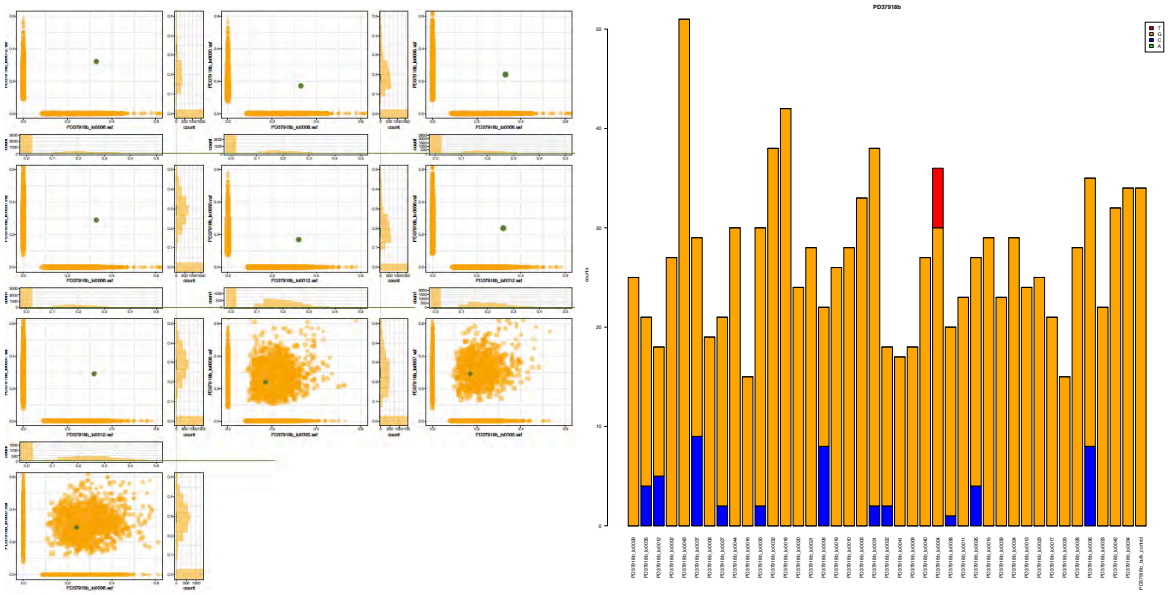
Assessment of congenital mosaicism as an explanation for shared hotspot mutation

A third possible explanation for the hotspot variant being shared across so many microdissections is that it occurred as an early embryonic mutation, and therefore the patient has congenital mosaicism for the variant. In humans, somatic mutations begin to accrue with the first cell division of the zygote, with the mutation rate estimated to be 0.8-1 mutations/genome/cell division on average[2,28]. In theory, if the *FOXO1* S22W mutation occurred early enough in foetal development, it could explain the patchy distribution throughout the liver and the emergence in more than one anatomical segment of the liver.

The difficulty with this explanation is that in many of the pairs of microdissections that share a *FOXO1* mutation, this is literally the only shared mutation (see, for example, the right-hand scatterplot in the insert above). This would imply that the *FOXO1* mutation would have to have been the first somatic mutation to occur during embryogenesis. We know from previous sequencing studies of embryonic mutations that the first mutation is acquired in the first one or two cell divisions, long before gastrulation[2,28]. As a result, these mutations are found in all germ layers and in microdissection sequencing such as that deployed here, where samples are not pure hepatocytes, they tend to be present at variable VAFs across all microdissections This can be seen, for example, in the VAF distributions for the earliest embryonic mutations in our study of lineage tracing from foetal blood[29].

What we actually observe in the data is that the *FOXO1* mutation is either seen at high VAF in a microdissection or in no reads at all (at an average depth of 31x). This is shown for one of our patients (PD37918 – see also **Figure 1D**) in the panels below (**Supplementary Figure 6**). This is incompatible with the patterns expected for the first embryonic mutation after fertilisation, especially given that the liver parenchyma represents a patchwork of cells with different foetal origins such as hepatocytes, endothelial cells, inflammatory cells and Kupffer cells.

Finally, there is the additional implausibility that the exact same base change would be the very first embryonic mutation in all 7 patients with multiple independent clones carrying the *FOXO1* S22W variant.



Supplementary Figure 6. Panels showing the distribution of VAFs across PD37918. On the left are scatter plots showing the observed VAF of all mutations for every pair of microdissections carrying the *FOXO1* S22W mutation. The *FOXO1* mutation is coloured in green, while all other mutations are coloured orange. The first 7 scatter plots show that the only shared mutation (VAF>0) is the *FOXO1* variant across the 7 pairs of samples (and hence acquired independently), whereas the last 3 scatter plots show that these pairs of samples are clonally related and the *FOXO1* mutation was acquired in a common ancestor. On the right is a stacked bar-chart showing the count of reads reporting each base call at the *FOXO1* S22W genomic

position (orange is reference base call; blue is mutant base call). These data are incompatible with an early embryonic mutation because these tend to be present at low VAF across all samples from a patient, rather than present at moderate VAF or totally absent as seen here.

Supplementary Methods

Further details on reconstructing phylogenetic trees

A small number of clusters from the HDP model (13 of 966, 1.34%) spanning 11 out of 32 patients (34.3%) were found to be under-split and consisted of microdissections that did not share any mutations. Such problematic microdissections were split off to form additional independent clusters. Clustering of spatially adjacent or proximate microdissections were verified for each set of patient-specific sequenced material by detailed inspection of corresponding histology images.

As the clustering algorithm assigns each mutation to one cluster based on similarity in VAFs across samples by design, the FOXO1 hotspot mutation was manually reassigned to a minimal number of clusters to account for the possibility of multiple independent acquisition events that can occur in several microdissections of a patient biopsy. Specifically, parsimonious hotspot reallocation entailed first identifying hotspot-bearing microdissections (and the SNV clusters of which they are members). Next, a minimal subset of the most phylogenetically parental clusters was identified (see methods for the inference of phylogenetic trees), where each had a maximal number of member microdissections with at least one raw read of support for the hotspot, such that all hotspot-bearing microdissections were accounted for. Importantly, although some hotspot-bearing microdissections were members of multiple clusters, the hotspot was only assigned to the cluster with the maximal number of hotspot-containing dissections in such instances. For a minority of cases, some member dissections of candidate hotspot-bearing clusters did not have any evidence of the hotspot mutation due to low coverage. In these cases, histology images were checked to determine whether spatial positioning of the dissections in question were proximate or distal to hotspot-bearing dissections within the same cluster. As additional support for the likelihood that these candidate dissections carried the hotspot, the sharing of mutations with *bona fide* hotspot-containing dissections was assessed.

Further details of gradient-boosted regression trees for indel calling

To gain a better understanding of the characteristics of authentic indels, a ground truth set of indels that are shared between proximate microdissections per patient was first identified through a detailed review of the set of histology images corresponding to the study cohort. Several features of indels were next considered for statistical modelling including the number of homopolymers, whether an indel occurred in a repetitive HG19 region, the dispersion estimate from the beta-binomial filter, and various cgpPindel VCF fields (i.e., VAF excluding ambiguous and unknown reads, QUAL – indicates the sum of mapping qualities of anchor reads, coverage, coverage_alt – the number of variant supporting reads, the S1 score from Pindel, LEN – the variant length in base pairs, REP – the number of repeats, NB-Tum – the number of reads mapped by the primary aligner to the negative strand showing a similar indel

event in the target sample, ND-Tum – the count of negative strand mapped reads from the primary aligner with or without the indel in the target sample, NP-Tum – the number of reads mapped by Pindel to the negative strand in the target sample, NR-Tum – the unique union of NP-Tum and ND-Tum, NU-Tum – the unique union of NP-Tum and NB-Tum, PB-Tum – the number of reads mapped by the primary aligner to the positive strand showing a similar indel event in the target sample, PD-Tum – the count of positive strand mapped reads from the primary aligner with or without the indel in the target sample, PP-Tum – the number of reads mapped by Pindel to the positive strand in the target sample, PR-Tum – the unique union of PP-Tum and PD-Tum, PU-Tum – the unique union of PP-Tum and PB-Tum, AMB-Tum – Ambiguous reads mapping on both the alleles with the same specificity in the target sample, UNK-Tum – Unknown reads containing mismatches in the variant position and don't align to the reference as the first hit, VT – the variant type either insertion or deletion).

These parameters were used as input to the XGBoost algorithm[30] in R to construct a series of 1,000 gradient boosted regression trees (also known as classification and regression trees (CARTs)) in order to elucidate the typical attributes of likely real indels in the truth set. Manual hyperparameter tuning was performed using a minimal grid-search based approach to arrive at a learning rate (ϵ) of 0.003, a minimum loss reduction (γ) of 20 for leaf node partitioning, a subsampling rate of 60% of the indels and 80% of covariates in the training data for learning regression trees, and a maximum tree depth of three. Furthermore, model training was performed while optimizing a binary logistic objective function, setting the L2 regularization factor (λ) and bias to 0.5, enabling leave-one-out-cross-validation, and setting the model training metric to assess the area under the receiver operator curve (AUROC). All other hyperparameters were set to default or recommended values according to software documentation. At the end of the model training process, a binary ensemble classifier is built in which scores from a large number of shallow regression trees each with low predictive power, are combined to give high predictive power. In the so-called ensemble boosting framework, the set of regression trees are trained additively, with the initial tree $(t = 0)$ trained to directly fit the dependent variable (in this case it is whether or not an indel has features that are similar to those of the patient-specific truth set) such that a logistic loss function is minimized, while subsequent trees are trained to fit the residuals of the overall model learned up until the previous iteration $(t - 1)$. For each indel, a series of scores from the regression tree base learners are then combined in a weighted fashion to arrive at a final score representing the probability of whether or not a given indel is likely a true or artefactual variant. A model score of > 0.7 was used to identify likely real indels. The predictive accuracy of the regression-tree ensemble model was assessed on a per-patient basis by evaluating the AUROC resulting from a comparison between the set of model-predicted and ground truth indels (**Supplementary Figure 2**).

In a boosted regression forest model, N trees (base learner functions $f_t(x_i)$ that map predictive features of indels to model scores) are learned from a training dataset comprising

predictor and dependent variable pairs $\{(x_i, y_i)\}_{i=1}^{n}$, where each tree stores a vector of leaf node scores, and its structure as a mapping function that assigns indel feature values to leaf nodes, while the overall model scores (predictions $\hat{y}_i$) are computed as the sum of scores from each tree:

$$\hat{y}_i^{(N)} = \sum_{t=1}^{N} f_t(x_i) = \hat{y}_i^{(t-1)} + \epsilon f_t(x_i) \quad [1]$$

Here ϵ is the learning rate, which scales each tree's contribution to the overall model score to safeguard against overfitting to the training data.

An additive process of sequential tree structure learning is adopted for building a boosted regression forest, wherein a tree structure $f_t(x_i)$ is learned at each training iteration that minimizes the following objective function:

$$\text{obj}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad [2]$$

$$g_i = \partial_{\hat{y}_i^{(t-1)}} \mathcal{L} = \frac{(1-y_i)e^{\hat{y}_i^{(t-1)}}}{1+e^{\hat{y}_i^{(t-1)}}} - \frac{y_i e^{-\hat{y}_i^{(t-1)}}}{1+e^{-\hat{y}_i^{(t-1)}}} \quad [3]$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} \mathcal{L} = (1-y_i)\left[\frac{e^{\hat{y}_i^{(t-1)}}}{1+e^{\hat{y}_i^{(t-1)}}} - \frac{e^{2\hat{y}_i^{(t-1)}}}{\left(1+e^{\hat{y}_i^{(t-1)}}\right)^2} \right] + y_i \left[\frac{e^{-\hat{y}_i^{(t-1)}}}{1+e^{-\hat{y}_i^{(t-1)}}} - \frac{e^{-2\hat{y}_i^{(t-1)}}}{\left(1+e^{-\hat{y}_i^{(t-1)}}\right)^2} \right] \quad [4]$$

$$\mathcal{L} = y_i ln\left(1 + e^{-\hat{y}_i^{(t-1)}}\right) + (1-y_i) ln\left(1 + e^{\hat{y}_i^{(t-1)}}\right) \quad [5]$$

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} \omega_j^2 \quad [6]$$

Where,

$\text{obj}^{(t)}$ Is the objective function to be optimized for regression tree t,

$f_t(x_i)$ Denotes regression tree t, evaluating the score of feature values x_i,

g_i Is the first order derivative of the loss function with respect to the model score from the previous iteration $\hat{y}_t^{(t-1)}$,

h_i Is the second order derivative of the loss function with respect to the model scores from the previous iteration $\hat{y}_t^{(t-1)}$,

$\mathcal{L}$ Is the logistic loss function, which describes how well the model fits the training data,

$\Omega(f_t)$ Is a regularization term describing the complexity of regression tree t, defined as a function of T leaves and the L2 norm of a vector of scores ω_j of indels assigned to leaf node j.

A logistic loss function was used in this study since the binary classification of likely real or artefactual indels was of interest. In order to approximate the logistic loss, a second order Taylor expansion is used. Thus, each training iteration starts with the calculation of g_i and h_i.

Letting $I_j = \{i|q(x_i) = j\}$ be the set of indels assigned to leaf node j, $f_t(x_i) = \omega_j$ can be used to indicate that assigning indel i to node j will result in a score of ω_j for that indel, which allows for the objective function to take the following form after some algebraic manipulation:

$$\text{obj}^{(t)} = \sum_{j=1}^{T} \left[G_j \omega_j + \frac{1}{2}\left(H_j + \lambda\right)\omega_j^2 \right] + \gamma T \quad [7]$$

$$G_j = \sum_{i \in I_j} g_i \quad [8]$$

$$H_j = \sum_{i \in I_j} h_i \quad [9]$$

Setting $\text{obj}^{(t)} = 0$ and taking the first derivative with respect to ω_j allows us to solve for the optimal leaf node weight ω_j^*:

$$\omega_j^* = -\frac{G_j}{H_j + \lambda} \quad [10]$$

Substituting [10] back into [7] gives the objective function value obj^* as a measure of optimality of a given tree structure, where lower values are more favourable:

$$\text{obj}^* = -\frac{1}{2}\sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \quad [11]$$

As there are often an innumerable number of possible node configurations for any given tree, it is not practical to compute obj^* for every possible tree structure in order to find the one with the optimum value. Instead, a greedy approach is taken to grow a tree one bifurcation at a time, starting with a minimal tree with zero depth. Then for each leaf node in the regression tree t being grown, the following gain metric is calculated, which evaluates the change in the objective function value after each new split is introduced. In this way, a tree is grown such that gain is maximized.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \quad [12]$$

Here, $\frac{G_L^2}{H_L + \lambda}$ and $\frac{G_R^2}{H_R + \lambda}$ denote the scores on the newly added left and right child nodes, respectively, while $\frac{(G_L + G_R)^2}{H_L + H_R + \lambda}$ denotes the score on the parent node where the new split is considered. Together, these three bracketed terms represent the training loss reduction. Lastly, γ represents a regularization term that accounts for the cost of adding complexity to the tree structure when introducing a new split. Moreover, γ also serves as the minimum amount of loss reduction required for partitioning a given leaf node of a given regression tree, where a larger value would result in a tree having fewer new branch points added. This is

because, the following condition would be true more frequently: $\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] < \gamma$, which would result in unfavorable negative gains, indicating that in such cases, branching would lead to a suboptimal tree structure and should therefore not occur. In this way, the calculated gain at each leaf node of a tree can be used to determine which new bifurcations could be generated to maximize loss reduction. To identify an optimal split for a particular feature at a given leaf node, one can systematically explore thresholds for grouping indels according to their feature values in order to calculate the sum of gains for each pair of groupings (i.e., indels with feature values less than or greater than some threshold value). By doing this, it is possible to identify the optimal threshold value that results in maximum loss reduction and therefore gain. Using this approach, optimal splits are identified iteratively for all indel features considered by the model until a prespecified tree depth has been reached, at which point, an optimal tree structure $f_t(x_i)$ will be identified that is then added to the overall model as defined in [1] before proceeding to learn another tree structure in the next training iteration. The algorithm exits only when a predefined number of trees has been learned.

It is notable that the ground truth set does not capture proximately shared indels that are genuinely present at loci where coverage is low, or indels that are private to particular microdissections. Indels in this category were found to often possess attributes that are similar to the ground truth variants and therefore have high model scores, which were validated to likely be real through manual review using a genome browser.

The importance of each model feature was evaluated using Shapley values, which is a concept rooted in game theory that provides a method for the determination of contributions from multiple contributors to a given result in a provably fair manner. In the context of the current study, they are numerical values calculated based on a weighted average of differences in model scores with and without a given predictive feature over the set of models that together accounts for all possible permutations of feature groupings, excluding the feature being evaluated. Thus, the Shapley value of a feature is the mean contribution (effect) of that feature's value to the overall model score, which can be calculated as follows:

$$\phi_i(\mathrm{p}) = \sum_{S\subseteq N/i}\frac{|S|!(n-|S|-1)!}{n!}\big(p(S\cup i)-p(S)\big) \qquad [13]$$

Where,

$\phi_i(\mathrm{p})$ Is the Shapley value of feature i for model prediction p,

$S \subseteq N/i$ Represents all possible sets (S) of feature groupings, excluding feature i,

$|S|!$ Represents the number of permutations of features in set S,

$(n-|S|-1)!$ Represents the number of permutations of features not in set S out of the total of n features considered,

$p(S \cup i)$ Is the model prediction p with feature i,
$p(S)$ Is the model prediction p without feature i.

In the case of regression forests, the Shapley value of a given predictive feature is the weighted sum of Shapley values corresponding to that feature in the ensemble of constituent trees. Here, the SHAP (SHapley Additive exPlanations) implementation designed specifically for regression-tree based models is used, which more rapidly calculates Shapley values based on the subset of features present in each regression tree in the ensemble. For a particular prediction of whether a given indel is likely real or artefactual, the overall model score can be broken down as the sum of Shapley values (contributions to the prediction) of each predictive feature plus the baseline (average) model score calculated from the training data. Presence or absence of features that are highly predictive of whether an indel is likely real or artefactual result in large changes in model scores. High values of such influential features either contribute positively or negatively to the scores, are typically associated with Shapley values of greater or less than zero, and correspond to traits of true or false indels, respectively. Conversely, features with a negligible effect on the overall model scores have near-zero Shapley values.

Further details on MCMC generalised linear mixed models for telomere lengths

To use these models effectively, we needed to construct a relatedness matrix among clones on phylogenetic trees. This required making trees ultrametric (namely, all tips finish at the same point). Although there are existing methods for making trees ultrametric, they did not especially suit our purposes for somatic mutations in normal tissues. We have more power for estimating the true number of early mutations (since they will be seen in more than one microdissection, and on average at higher variant allele fraction). Our observed estimates of branch length for the late, singleton branches will be less certain. Therefore, we developed a recursive function to place each branch-point (coalescence) on a given fraction of molecular time. Starting at the root of the tree, and progressing towards the tip, the position of each coalescence is estimated as the number of mutations acquired between root and that coalescence divided by that number plus the average of the number of mutations in branches descending from the coalescence.

We used a Bayesian linear mixed effects model to model telomere lengths. The dependent variable was the average telomere length ('Length') of each clone, measured in base-pairs, and expected to follow a Gaussian distribution. The fixed effects were: Aetiology (dummy variables for ARLD and NAFLD); Age of patient (in years); Number of mutations in that clone; and Clone area. We fitted a random effect for each patient and a random effect for the phylogenetic relationships encoded in the form of a block diagonal matrix. The priors were uninformative inverse-Wishart distributions. We ran the MCMC chain for 11,000,000 iterations with 1,000,000 of these as a burn-in, thinned to every 1000 iterations.

We checked the usual diagnostics on the MCMC chain - these showed (1) there were no systematic trends in the variables after the burn-in has completed (the burn-in was of sufficient length); (2) autocorrelation was relatively low among adjacent estimates with the thinning to every 1000th instance of the chain; (3) diagnostic tests of convergence were all passed.

We derived estimates for the 'somatic heritability' of telomere lengths by calculating the component of variance in the random effects that was attributed to the variable describing phylogenetic relationships versus the residual variance and variance attributed to between-patient effects. The estimate of 'somatic heritability' for telomere lengths was 21% (95% posterior interval = 0-44%) – that is, once the effects of age, disease and clone size have been accounted for, ~21% of the residual variation in telomere lengths across the cohort can be explained by the phylogenetic relationships among clones. However, there is some instability in this estimate, and maybe about 26% of iterations of the MCMC chain suggest no somatic heritability. Either way, the confounding effects of this between-sample correlation have been corrected within the modelling framework.

Counting alleles and indels

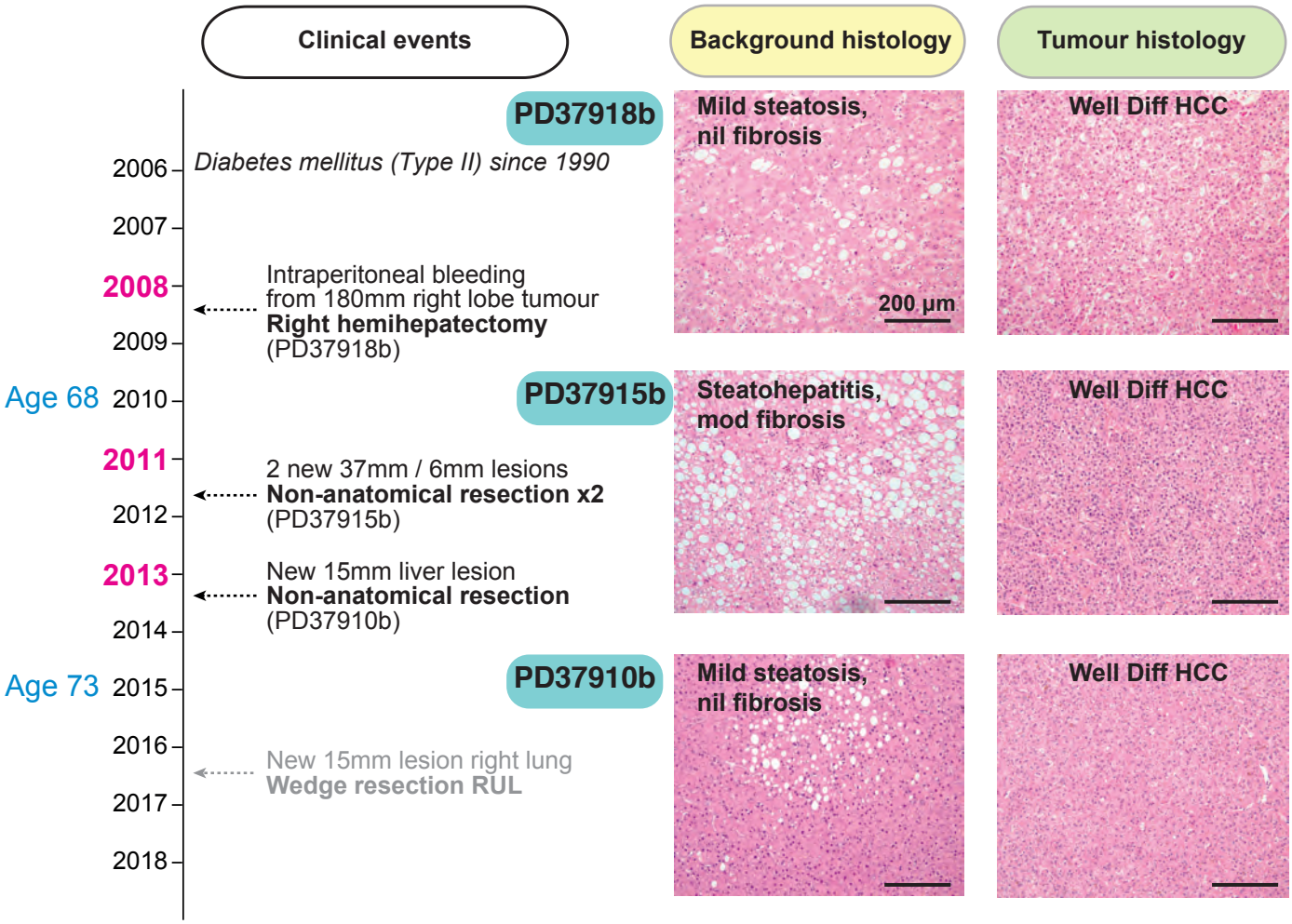
Raw allele counts of each base (A, C, G, T) and total coverage of unfiltered reads were enumerated using alleleCount software for SBS-type variants to generate input for the beta-binomial based filter, while setting the minimum mapping and base quality to 30 and 25, respectively (http://cancerit.github.io/alleleCount/). For input to NDP clustering and for counts of indels, cgpVAF software was used (https://github.com/cancerit/vafCorrect).

Allele counter software was used to determine the number of unfiltered raw counts of each base directly from the bam files of both LCM microdissections and bulk control samples. Stacked barplots were generated from these count data for each patient found to carry the FOXO1 S22W hotspot driver mutation to visualize its distribution of across clones and bulk control (if available).

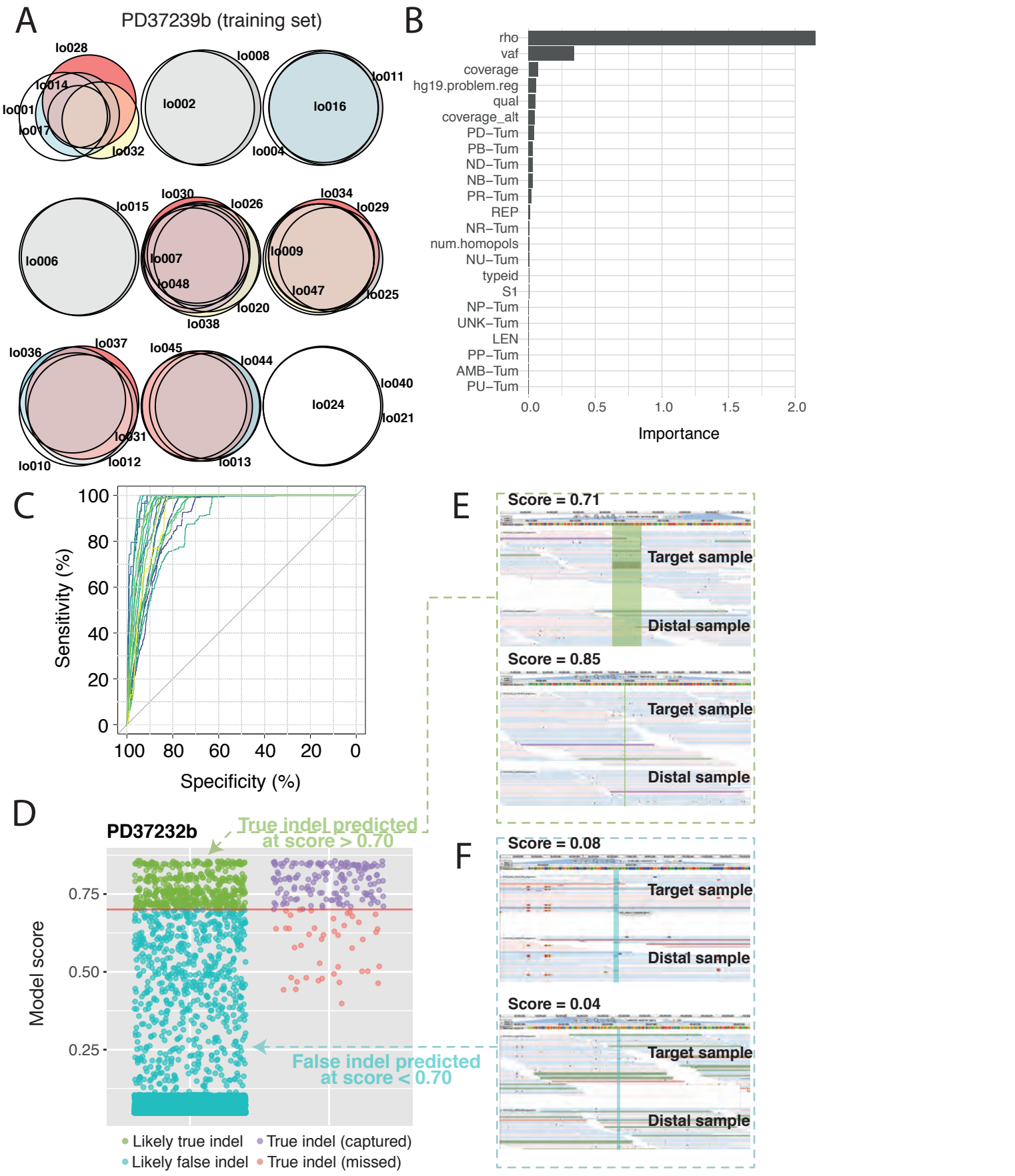
References

1. Martincorena, I. *et al.* High burden and pervasive positive selection of somatic mutations in normal human skin. *Science (80-.).* **348**, 880–886 (2015).
2. Lee-Six, H. *et al.* Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
3. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
4. Martincorena, I. *et al.* Somatic mutant clones colonize the human esophagus with age. *Science (80-.).* **917**, 911–917 (2018).
5. Yokoyama, A. *et al.* Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
6. Moore, L. *et al.* The mutational landscape of normal human endometrial epithelium. *Nature* **580**, (2020).
7. Yoshida, K. *et al.* Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
8. Blokzijl, F. *et al.* Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
9. Yizhak, K. *et al.* RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science (80-.).* **364**, eaaw0726 (2019).
10. Brunner, S. F. *et al.* Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
11. Zhu, M. *et al.* Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. *Cell* **177**, 608–621 (2019).
12. Kim, S. K. *et al.* Comprehensive analysis of genetic aberrations linked to tumorigenesis in regenerative nodules of liver cirrhosis. *J. Gastroenterol.* (2019). doi:10.1007/s00535-019-01555-z
13. Brazhnik, K. *et al.* Single-cell analysis reveals different age-related somatic mutation profiles between stem and differentiated cells in human liver. *Sci. Adv.* **6**, 1–11 (2020).
14. Kakiuchi, N. *et al.* Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis. *Nature* **577**, 260–265 (2020).
15. Nanki, K. *et al.* Somatic inflammatory gene mutations in human ulcerative colitis epithelium. *Nature* **577**, 254–259 (2020).
16. Olafsson, S. *et al.* Somatic Evolution in Non-neoplastic IBD-Affected Colon. *Cell* **182**, 672-684.e11 (2020).
17. Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495–501 (2014).
18. Hadfield, J. D. & Nakagawa, S. General quantitative genetic methods for comparative biology: Phylogenies, taxonomies and multi-trait models for continuous and categorical characters. *J. Evol. Biol.* **23**, 494–508 (2010).
19. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, 1–22 (2010).
20. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
21. Gerstung, M., Jolly, C., Leshchiner, I. & Dentro, S. C. The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).
22. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues.

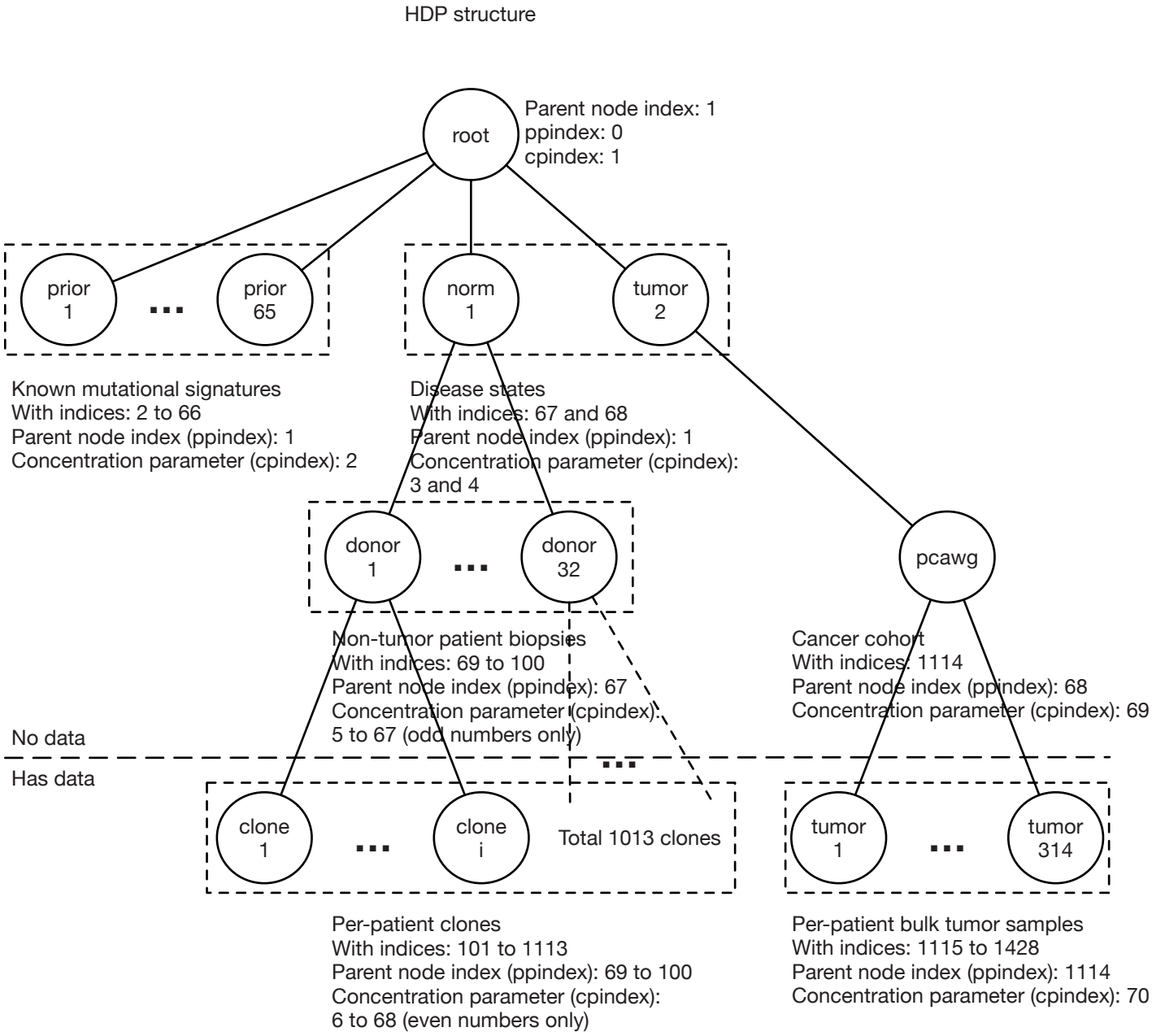
Cell **171**, 1029–1041 (2017).
23. Rheinbay, E. *et al.* Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
24. Neal, R. M. Markov Chain Sampling Methods for Dirichlet Process Mixture Models. *J. Comput. Graph. Stat.* **9**, 249–265 (2000).
25. Jiao, W., Vembu, S., Deshwar, A. G., Stein, L. & Morris, Q. Inferring clonal evolution of tumors from single nucleotide somatic mutations. *BMC Bioinformatics* **15**, 35 (2014).
26. Pan-cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
27. Suda, K. *et al.* Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep.* **24**, 1777–1789 (2018).
28. Ju, Y. S. *et al.* Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
29. Spencer Chapman, M. *et al.* Lineage tracing of human development through somatic mutations. *Nature* **595**, 85–90 (2021).
30. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System*. (2016). doi:10.1145/2939672.2939785



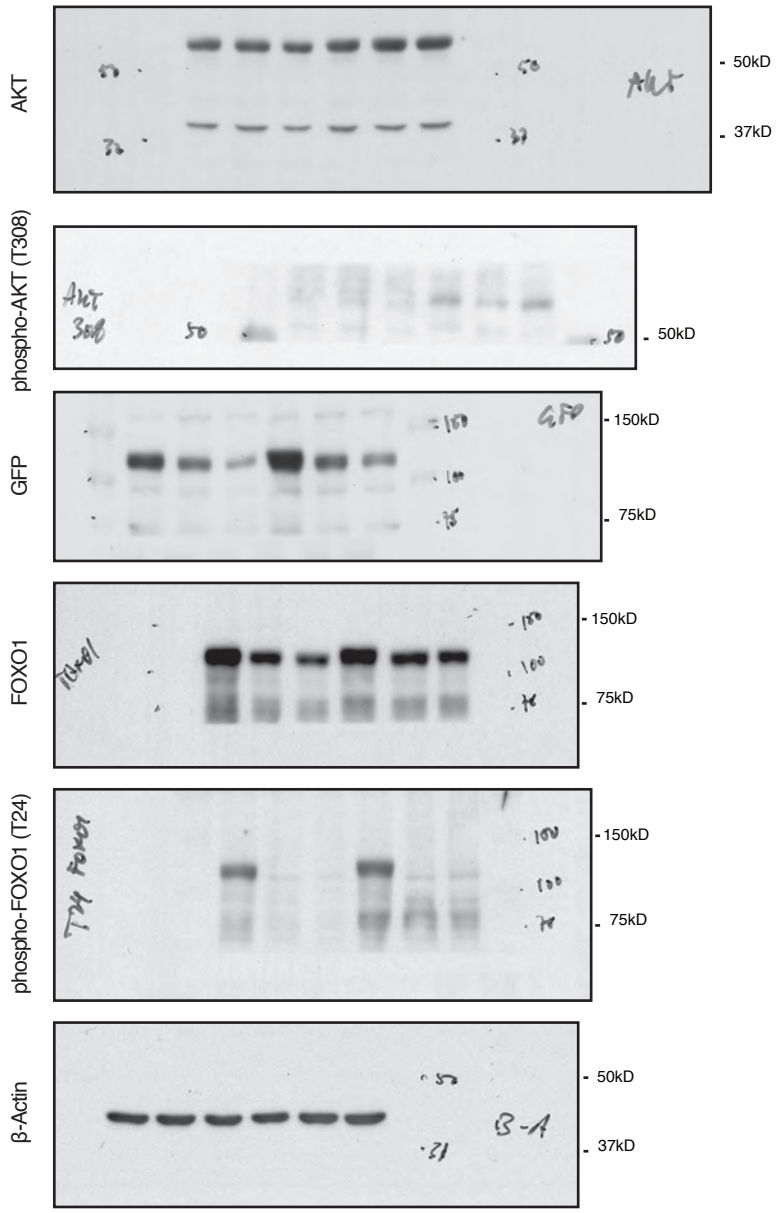
Supplementary Figure 1. Longitudinal clinicopathological study of multiple liver resections from a single patient. Timeline (left) shows the clinical events over time (y axis) from initial presentation with tumour rupture and intraperitoneal bleeding, three liver resections for hepatocellular carcinoma (HCC), and recent lung metastasis. Representative photomicrographs showing haematoxylin and eosin (H&E) stained tissue sections of background liver (centre) and liver tumour (right) for each resection. Right hemihepatectomy (PD37918b) shows mild steatosis and no fibrosis and a well-differentiated HCC. Non-anatomical liver resections (PD37915b) shows steatohepatitis and moderate fibrosis and two well-differentiated HCCs. A further non-anatomical liver resection (PD37910b) shows mild steatosis and no fibrosis and a well differentiated HCC. Original magnification x100. All scale bars = 200 µm.



Supplementary Figure 2. A regression-tree based indel identifier. (A) Euler diagrams representing the approximate sharing of indels between sets of proximate LCM microdissections corresponding to different clones from PD37239b. Each circle represents one microdissection (labelled "lo…"). (B) Indel identification model feature importance as assessed using Shapley values. (C) Receiver operator characteristic (ROC) plot depicting indel identification accuracy relative to the ground truth set of indels shared between proximate microdissections per patient sample. (D) A representative breakdown of indel identification model output for PD37232b, where indels with model scores >0.70 are considered to be real. Green and turquoise data points are not in the ground truth set, whereas purple and pink data points are. (E,F) Genome browser images showing supporting reads for likely real (E) and artefactual (F) indels as identified by the model (score >0.7).



Supplementary Figure 3. The HDP structure used for the extraction of mutational signatures.



Supplementary Figure 4. Uncropped and unprocessed scans of the Western blots shown in **Extended Figure 6b**.