Article

Somatic mutations and clonal dynamics in healthy and cirrhotic human liver

https://doi.org/10.1038/s41586-019-1670-9

Received: 17 November 2018

Accepted: 12 September 2019

Published online: 23 October 2019

Simon F. Brunner[1], Nicola D. Roberts[1], Luke A. Wylie[1], Luiza Moore[1], Sarah J. Aitken[2,3], Susan E. Davies[3], Mathijs A. Sanders[1,4], Pete Ellis[1], Chris Alder[1], Yvette Hooks[1], Federico Abascal[1], Michael R. Stratton[1], Inigo Martincorena[1], Matthew Hoare[2,5]* & Peter J. Campbell[1,6]*

The most common causes of chronic liver disease are excess alcohol intake, viral hepatitis and non-alcoholic fatty liver disease, with the clinical spectrum ranging in severity from hepatic inflammation to cirrhosis, liver failure or hepatocellular carcinoma (HCC). The genome of HCC exhibits diverse mutational signatures, resulting in recurrent mutations across more than 30 cancer genes[1–7]. Stem cells from normal livers have a low mutational burden and limited diversity of signatures[8], which suggests that the complexity of HCC arises during the progression to chronic liver disease and subsequent malignant transformation. Here, by sequencing whole genomes of 482 microdissections of 100–500 hepatocytes from 5 normal and 9 cirrhotic livers, we show that cirrhotic liver has a higher mutational burden than normal liver. Although rare in normal hepatocytes, structural variants, including chromothripsis, were prominent in cirrhosis. Driver mutations, such as point mutations and structural variants, affected 1–5% of clones. Clonal expansions of millimetres in diameter occurred in cirrhosis, with clones sequestered by the bands of fibrosis that surround regenerative nodules. Some mutational signatures were universal and equally active in both non-malignant hepatocytes and HCCs; some were substantially more active in HCCs than chronic liver disease; and others—arising from exogenous exposures—were present in a subset of patients. The activity of exogenous signatures between adjacent cirrhotic nodules varied by up to tenfold within each patient, as a result of clone-specific and microenvironmental forces. Synchronous HCCs exhibited the same mutational signatures as background cirrhotic liver, but with higher burden. Somatic mutations chronicle the exposures, toxicity, regeneration and clonal structure of liver tissue as it progresses from health to disease.

Identifying somatic mutations in non-malignant tissue requires approaches that overcome the polyclonality of this tissue, such as single-cell sequencing[9], cultures of single cells[8,10] or microbiopsy sequencing[11]. The latter relies on local cell division with limited migration leading to a clonal patchwork, which has been observed in liver tissue[12]. We generated whole-genome sequences from 482 laser-capture microdissections (LCMs) of 100–500 hepatocytes (Extended Data Fig. 1a) across 14 individuals: 5 healthy controls; 4 patients with cirrhosis from alcohol-related liver disease (ARLD) and 5 patients with cirrhosis from non-alcoholic fatty liver disease (NAFLD) (Supplementary Tables 1, 2, Extended Data Figs. 4–6). Samples of normal liver were acquired from hepatic resections of colorectal cancer metastases, and samples of cirrhotic liver were taken from patients who underwent liver transplants for synchronous but distant HCC.

To evaluate sensitivity and specificity, we generated independent libraries and sequencing data from different sections of the same biopsy, microdissecting the same x, y-region from adjacent z-stacks separated by around 20 μm. Concordance was high between variants that were called in adjacent sections, but not between distant pairs, suggesting that the specificity of mutation calls was high (Extended Data Fig. 1b). Sensitivity across patients ranged from 50 to 95%, depending on coverage and clonality (Extended Data Fig. 1c–f). As a further check on specificity, targeted deep sequencing of cancer genes from the same library as 96 whole-genome samples confirmed 16 of the 17 mutations that were originally called. In keeping with polyploidy as a late stage of differentiation in the liver[13], 20–25% of mature hepatocytes in microdissected samples were multinuclear (Extended Data Fig. 1g). We therefore deployed copy-number algorithms with an expected ploidy of 4, and report mutational burdens per diploid genome, rather than per cell.

We observed considerable heterogeneity in the burden of somatic substitutions both between and within patients (Fig. 1a, Supplementary Tables 3, 4). Using mixed-effects models, microdissections from

[1]Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, UK. [2]CRUK Cambridge Institute, Cambridge, UK. [3]Department of Pathology, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [4]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. [5]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [6]Department of Haematology and Stem Cell Institute, University of Cambridge, Cambridge, UK. *e-mail: Matthew.Hoare@cruk.cam.ac.uk; pc8@sanger.ac.uk

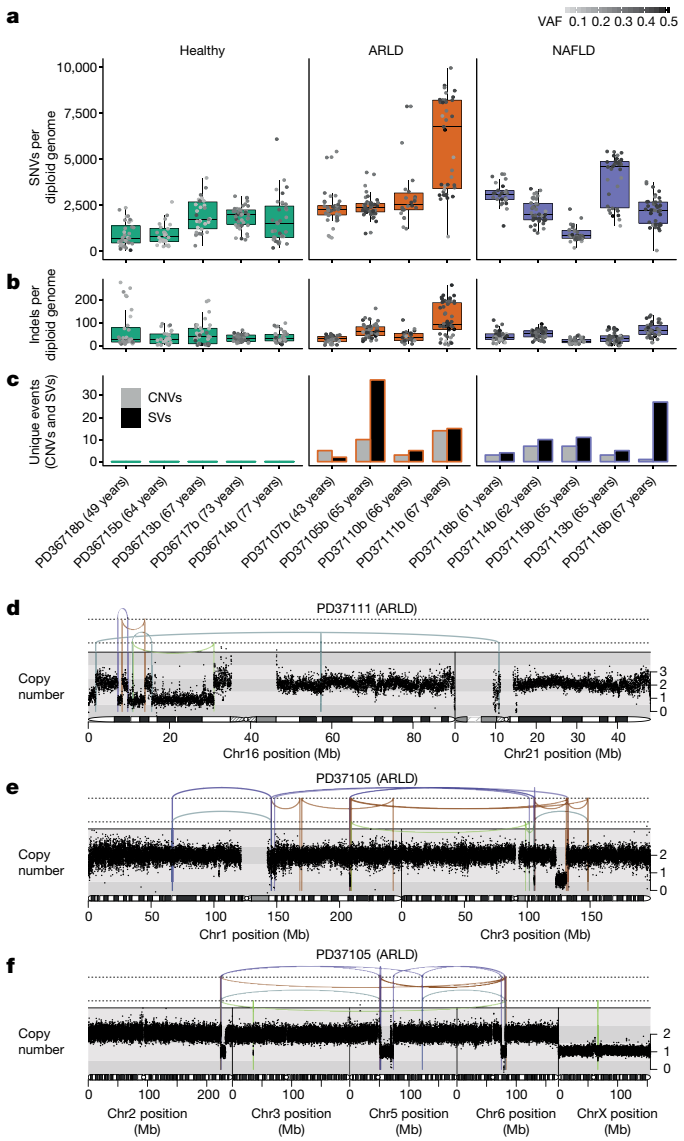


Fig. 1 | Mutational burden observed in non-cancerous hepatocytes. **a**, Burden of single-nucleotide variants (SNVs), corrected by sensitivity of mutation detection. Each box plot represents a patient ($n = 14$ patients; 482 microdissections) and each dot represents one laser-capture microdissected sample. The grey-to-black intensity of the points reflects the median variant allele fraction (VAF) of mutations in each microdissection. Boxes in the box plots indicate median and interquartile range; whiskers denote range. **b**, Burden of indel variants ($n = 14$ patients; 482 microdissections). **c**, Burden of copy-number variants (CNVs) and structural variants (SVs), represented as the number of unique events per patient. **d**, Chromothripsis involving chromosomes 16 and 21, observed in patient PD37111. Black points represent corrected read depth along the chromosome. Lines and arcs represent structural variants, coloured by the orientation of the joined ends (purple, tail-to-tail inverted; brown, head-to-head inverted; turquoise, tandem-duplication-type orientation; green, deletion-type orientation). **e**, Chromothripsis involving chromosomes 1 and 3, observed in patient PD37105. **f**, Chromothripsis involving chromosomes 2, 5 and 6, observed in patient PD37105 (in a separate clone to **e**).

cirrhotic livers had, on average, 1,251 (95% confidence interval, 233–2,268; $P = 0.02$) extra substitutions per diploid genome compared to normal livers, independent of age. In accordance with published values[8], the estimated rate of accumulation of mutations was 33 per year per diploid genome, albeit with wide confidence intervals (95% confidence interval, −17 to 84; $P = 0.18$) and moderate variation between individuals (estimated between-individual s.d., 13 per year). Insertions and deletions (indels) showed the same heterogeneity between and within individuals as substitutions (Fig. 1b).

Structural variants and copy-number alterations occurred in moderate numbers across all nine patients with liver cirrhosis, despite being rare in normal liver (Fig. 1c, Extended Data Fig. 2, Supplementary Tables 3, 4). Occasional aneuploidy at whole-chromosome or arm level occurred, as well as focal events including deletions, tandem duplications and unbalanced translocations (Extended Data Fig. 2). We found five separate clusters of structural variants across three patients, with patterns indicative of chromothripsis[14] (Fig. 1d–f, Extended Data Fig. 2). Chromothripsis—in which multiple rearrangements occur in a single catastrophic mitosis[14]—is a major process of mutation in cancers (occurring in around 5% of HCCs[15]), but is rare in normal somatic cells. Our observation of 1–2% of clones with chromothripsis in chronic liver disease suggests that sustained toxicity and regeneration substantially increases mitotic stress in hepatocytes.

We screened for driver mutations among coding regions, 5′-untranslated regions (UTRs), 3′-UTRs and promoters (Supplementary Tables 5–8). No elements were significant genome-wide after correcting for multiple hypotheses, so we focused on the 30 most-prevalent HCC genes[1–5]. These carried 22 non-synonymous variants that were seen in both normal and cirrhotic samples and included inactivating mutations in the tumour suppressor genes *ACVR2A*, *ARID2*, *ARID1A* and *TSC2* (Extended Data Fig. 3a). When hypothesis testing was restricted to these 30 genes, *ALB* and *ACVR2A* were significant ($q = 0.001$ and $q = 0.001$, respectively). Recurrence in *ALB* (which encodes the protein albumin) probably reflects a mutational process in which indels preferentially occur in highly expressed genes, as reported in HCCs[5,16] (Extended Data Fig. 3b, c). Assuming no negative selection, we can use the ratio of non-synonymous to synonymous substitutions for the 30 HCC genes to estimate the number of driver substitutions among them[17]; this gives a 95% confidence interval of 0.0–13.2 driver mutations in total across 482 microdissections (that is, less than 3%). Among copy-number aberrations of potential importance[1,2,18] (Supplementary Table 9), we found instances of loss of chromosomes 22 and 8p, and gain of chromosome 8q. Two focal deletions in different patients spanned *ACVR2A* (Extended Data Fig. 2c, e). We also found a reciprocal inversion that deleted *CDKN2A* (Extended Data Fig. 2f), the most common focal deletion in HCC, and a deletion that affected *ARID5A*.

We reconstructed phylogenetic trees[19] and layered them onto the histology of the specimens. Samples from healthy controls showed the highly polyclonal nature of normal liver, with little genetic relatedness among even closely located microdissections (Fig. 2a–d, Extended Data Fig. 4). Samples from patients with chronic liver disease showed a clonal structure that was more complex, from which three general inferences can be drawn (Fig. 2e–p, Extended Data Figs. 5, 6). First, we found no sharing of mutations between adjacent liver nodules separated by fibrotic bands. This suggests that the connective tissue that is laid down during cycles of damage and regeneration sequesters clones from the early stages of the disease process. Second, some cirrhotic nodules were monoclonally derived (Fig. 2j, n, for example), whereas others were oligoclonal (Fig. 2f), and shared mutations often extended across microdissections that were millimetres apart. Third, branching structures in phylogenies point to subclonal diversification within nodules. Within such a clone, the proportion of shared, clonal mutations on the trunk relative to those on the subclonal branches gives an estimate in molecular time of when the most-recent common ancestor of the clone emerged. In some patients (for example, patient PD37114; Fig. 2i, j), the common ancestor of individual nodules emerged relatively early in molecular time, whereas in others (for example, patient PD37116; Fig. 2m, n), the common ancestor appeared much more recently. As the majority of liver cells do not have driver mutations, the size and rapidity of clonal expansions observed here demonstrate the considerable intrinsic capacity of hepatocytes to regenerate in response to liver damage.

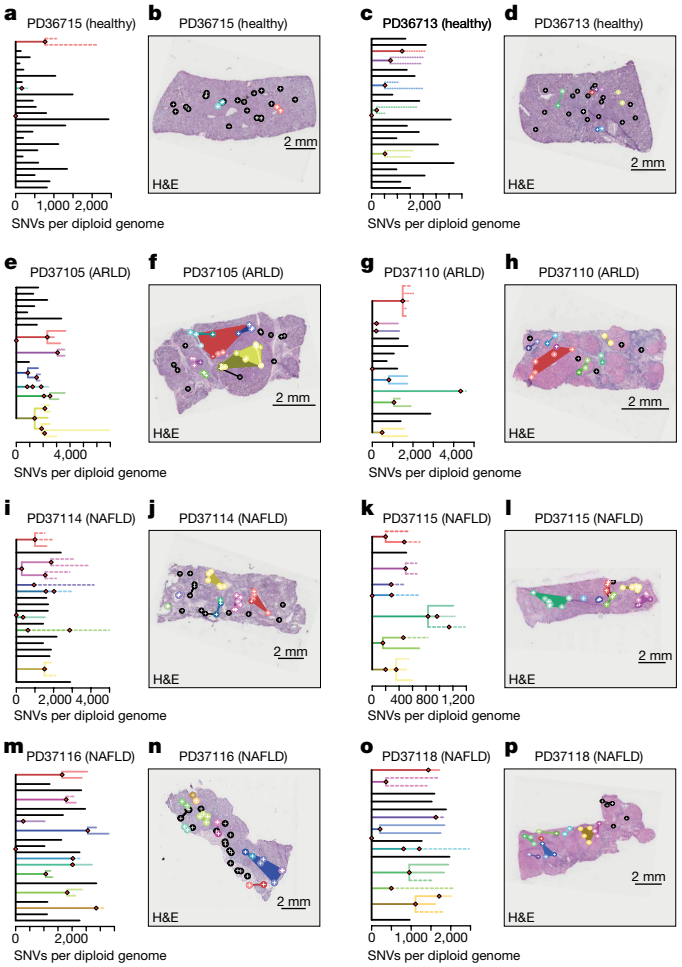


Fig. 2 | Phylogenetic reconstruction of hepatocyte clones. a, Phylogenetic tree constructed from clustering of mutations across microdissected samples in a healthy individual (PD36715). Lengths of branches (*x* axis) indicate the numbers of mutations assigned to that branch. Solid lines indicate that nesting is in accordance with the pigeonhole principle; dashed lines indicate that nesting is in accordance with the pigeonhole principle, assuming that hepatocytes represent 70% of cells; dotted lines indicate that nesting is only based on clustering (clones are assigned as nested if the VAFs of constituent microdissections are lower than those in the parental clone). **b**, Representation of branches from the phylogenetic tree in **a** according to their physical coordinates, overlaid onto a haematoxylin and eosin (H&E)-stained section. Black points represent branches of the tree that share no mutations with any other samples; coloured points represent branches with shared clonal relationships ($n = 26$ microdissections). **c**, **d**, A second healthy individual (PD36713; $n = 30$ microdissections). **e**, **f**, Patient with ARLD (PD37105; $n = 31$ microdissections). **g**, **h**, Patient with ARLD (PD37110; $n = 22$ microdissections). **i**, **j**, Patient with NAFLD (PD37114; $n = 41$ microdissections). **k**, l, Patient with NAFLD (PD37115; $n = 34$ microdissections) **m**, **n**, Patient with NAFLD (PD37116; 43 microdissections). **o**, **p**, Patient with NAFLD (PD37118; 26 microdissections).

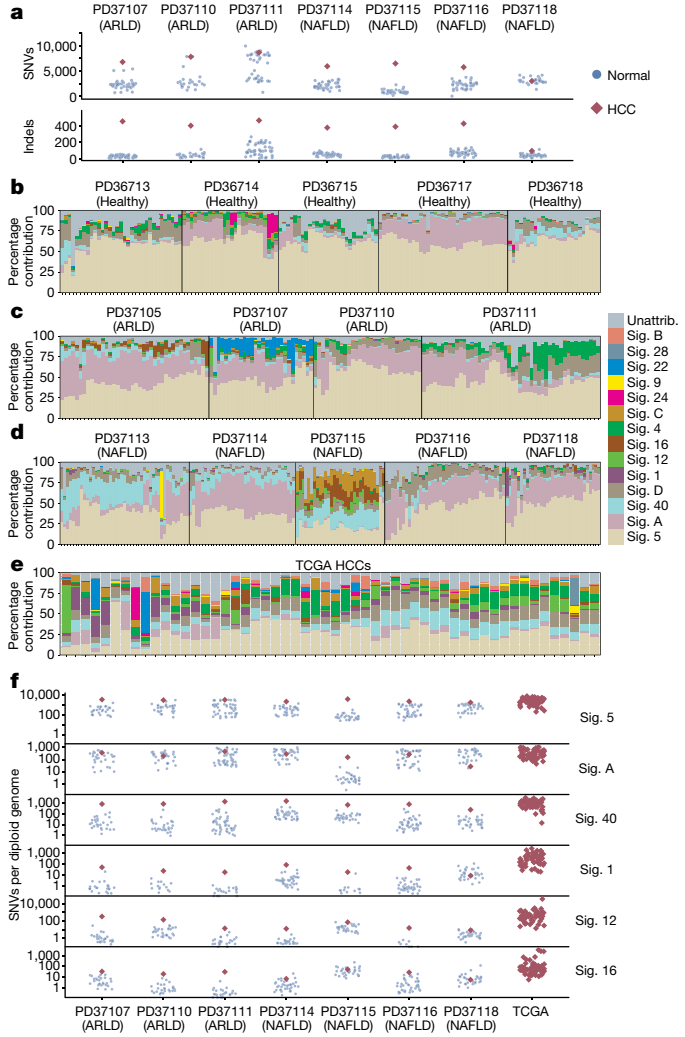


Fig. 3 | Mutational signatures in normal liver, cirrhotic liver and HCC. **a**, Number of somatic substitutions (SNVs; sensitivity-corrected for non-cancerous samples) and indels in each non-cancer microdissection sample (blue circles) and associated synchronous HCCs (red diamonds). **b**–**e**, Estimated proportional contributions of each mutational signature to each phylogenetically defined cluster of somatic substitutions. Data were generated using a Bayesian hierarchical Dirichlet process. Unattrib., unattributed. Stacked bar plots show proportional contributions of signatures in healthy individuals (**b**), patients with ARLD (**c**), patients with NAFLD (**d**) and 54 cases of HCC from TCGA[1] (**e**). **f**, Number of SNVs attributed to prevalent mutational signatures in each non-cancer microdissection sample (blue circles) and synchronous HCCs (red diamonds). Contributions for the TCGA samples are shown on the right. The *y* axis is on a logarithmic scale.

A major debate in the modelling of cancer development is whether cancers need higher rates of mutation to acquire sufficient driver mutations. We compared the mutational burden in cirrhotic liver to synchronous, clonally unrelated HCCs from seven patients. Synchronous HCCs carried, on average, 4,600 more mutations than matched cirrhotic liver (95% confidence interval, 3,600–5,500; $P < 10^{-18}$ using linear mixed-effect models; Fig. 3a). This indicates that rates of mutation increase during malignant transformation, either through cancer-specific mutational processes or through greater activity in cancers of ubiquitous mutational processes.

To assess which mutational processes are active in cirrhosis, we extracted mutational signatures from our 482 microdissections, as well as from the 7 synchronous HCCs and 54 HCC genomes from The Cancer Genome Atlas (TCGA)[1], using two independent algorithms (Fig. 3b–e, Extended Data Figs. 7, 8). Three major groups of mutational signatures emerged: first, those that are ubiquitous and similarly active across cirrhosis and HCC; second, those that are minor contributors in cirrhosis but universally more active in HCC; and third, those that are active in some patients but absent in others, including signatures that arise from exogenous stimuli.

In normal and cirrhotic liver, ubiquitous mutational signatures (5 and A) were prevalent across clones and, in combination, typically accounted for more than 75% of mutations. Signature 5 is widespread across cancers—including HCCs[2,4,20]—and accumulates linearly with age, suggesting that it arises from endogenous mutational processes. Signature A is the dominant cause of mutations in normal blood stem

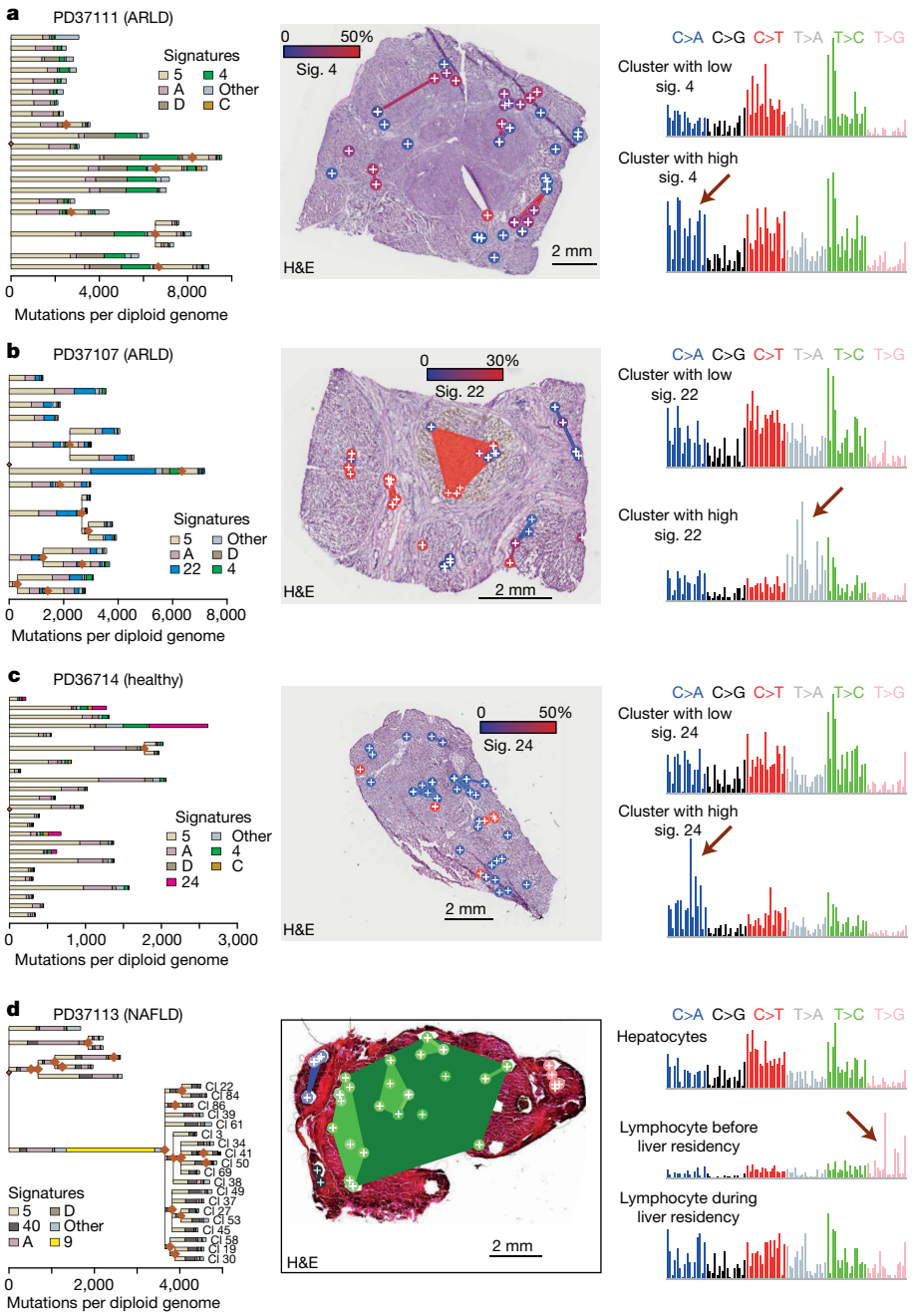


Fig. 4 | Links between exogenous factors and mutational signatures. a, Left, phylogenetic tree of clones in patient PD37111. Each branch is coloured by the proportion of mutations in that branch that are assigned to the different mutational signatures. Middle, overlay of the clones in **a** onto an H&E-stained liver section of patient PD37111 ($n = 39$ microdissections). Clones are coloured according to the proportion of mutations attributed to signature 4, which is linked to tobacco exposure (blue, low activity of signature 4; red, high activity of signature 4). Right, representative mutation spectra for samples with low (top) or high (bottom) burden of signature 4. The six types of substitution are labelled across the top. Within each type of substitution, the contributions from the trinucleotide context are shown as 16 bars. The 16 bars are divided into 4 sets of 4 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 an A, C, G or T is 3′ to the mutated base. Brown arrows indicate parts of the mutation spectrum that are characteristic of the relevant mutational signatures. **b**, Overlay of mutational signatures onto phylogenetic tree of clones in patient PD37107 ($n = 41$ microdissections). The colouring of clones in the middle H&E-stained section is according to signature 22, which is linked to the carcinogen aristolochic acid. **c**, Overlay of mutational signatures onto phylogenetic tree of clones in patient PD36714 ($n = 35$ microdissections). The colouring of clones in the middle H&E-stained section is according to signature 24, which is linked to the carcinogen aflatoxin-B_1. **d**, Overlay of mutational signatures onto phylogenetic tree of clones in patient PD37113 ($n = 37$ microdissections). Cluster 10 has many mutations attributed to signature 9, which is linked to the process of somatic hypermutation in B lymphocytes. Cl, cluster.

cells[10,21] and leukaemias[21], which indicates that it too arises endogenously. In HCCs, although signature A accounted for a lower proportion of mutations than in normal or cirrhotic liver, the absolute numbers of mutations attributed to signature A were comparable (difference between cancer and non-cancer, 60 mutations; 95% confidence interval, −80 to 200; $P = 0.4$; Fig. 3f, Supplementary Table 10). This suggests that signature A is active in hepatocytes throughout life, but is outstripped in HCC by mutational processes that emerge during malignant transformation.

A second group of mutational signatures comprises processes that are relatively quiet in cirrhotic liver but universally more active in HCC (signatures 1, 12, 16, 40 and a new signature, D; Supplementary Table 10). One of these, signature 16, consists of T to C mutations in the ApT context and has a known transcriptional-strand bias, which includes both the preferential repair of damaged adenines on transcribed strands and increased damage on non-transcribed strands[22]. Although this signature is more active in HCCs, we do see its characteristic transcriptional-strand bias in cirrhotic liver (Extended Data Fig. 9a). Signature 1, which is caused by spontaneous deamination of methylated cytosine to thymine, is also much more active in HCC than non-malignant liver. The acceleration and universality of these signatures in HCC suggests that they reflect inbuilt DNA damage and repair processes in hepatocytes that are unmasked during malignant transformation.

The third group of mutational processes represents signatures that are seen sporadically across the cohort and that are frequently caused by exogenous factors. One, signature 4, is found in lung cancers from smokers[20], and also in HCCs, albeit with a less clear-cut relationship to tobacco[2]. Of our 14 patients, 4 had more than 10% of microdissections in which more than 5% of mutations were attributed to signature 4, demonstrating the expected transcriptional-strand bias of this signature on guanines (Extended Data Fig. 9b). Not only did signature

4 show considerable patient-to-patient heterogeneity, but there was also unexpectedly high clone-to-clone and nodule-to-nodule variability within individual livers. In one patient, for example, about half the clones we sequenced had 2,000–4,000 mutations, whereas the other half had 8,000–12,000; these differences were driven by the presence or absence of signature 4 (PD37111; Fig. 4a).

This within-patient regional variability extended to other exogenous factors. In one patient (PD37107), 20–35% of mutations were derived from signature 22 (Fig. 4b, Extended Data Fig. 9c), which is characteristic of exposure to aristolochic acid[23]. This patient grew up in Poland and spent time on holiday in Balkan states where exposure to aristolochic acid is common[24]. In a different patient (PD36714), a subset of microdissections had 10–20% of mutations that were attributable to signature 24 (Fig. 4c), which is associated with aflatoxin-B_1 exposure[5]. Aflatoxin-B_1 is produced by *Aspergillus* moulds that contaminate crops, and biomarkers of exposure to this toxin are prevalent in arable farmers[25]—the occupation of our patient. In both patients, these carcinogens showed notable variability in mutational activity over short distances, generating few mutations in some clones and hundreds to thousands in others. This regional variation in the activity of exogenous signatures is unexpected, and so far unexplained.

In one patient, we found a large clone that carried more than 2,000 mutations attributable to signature 9 (Fig. 4d)—a result of off-target somatic hypermutation in B lymphocytes[20]. A clonotypic rearrangement of *IGH* was evident, which is consistent with the notion that a single B lymphocyte subclonally diversified as it expanded in the liver (Extended Data Fig. 10). Signature 9 was only present on the ancestral trunk, whereas signatures in the subclones (acquired in the liver) were distributed in a similar manner to hepatocytes, suggesting that the hepatic microenvironment shaped the ongoing mutational processes in the lymphocytes.

In conclusion, then, non-malignant liver has considerably lower proportions of clones (less than 5%) with driver point mutations or structural variants than oesophagus or skin[11,26,27], and those present were seen in both normal and cirrhotic liver. In the cirrhotic liver, fibrosis isolated these clones, either with or without driver mutations, restricting their expansion. Moreover, driver mutations were not shared with distant synchronous HCCs, which suggests that the increased risk of cancer in chronic liver disease arises from a myriad of clones that compete independently to acquire sufficient driver mutations. Mutations in the *TERT* promoter are likely to be key events in the progression to HCC; we did not identify any *TERT* promoter mutations in cirrhotic or normal liver, but they are seen in dysplastic hepatic nodules[18,28]. The low proportion of clones with driver mutations that we observed here, and that has also been shown in exome studies performed elsewhere[29,30], means that much larger sample sizes will be needed to comprehensively map how driver mutations accumulate in the progression from normal liver through regenerative and dysplastic nodules to HCC.

These data reveal the genomic consequences of chronic liver disease—increased rates of mutation, complex structural variation (including chromothripsis and aneuploidies) and a low burden of mutations that target known HCC genes. Genomically, one middle-aged, healthy liver looks much like any other: a community of small, tightly packed clones, each comprising a few hundred cells and containing around 1,000–1,500 mutations that come from a limited palette of signatures. Unhealthy livers diverge from this norm and instead exhibit large dynasties of clones, which are sequestered by bands of fibrosis and have a repertoire of signatures that is more variable, more vigorous and more regionally variegated.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1670-9.

1. Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341 (2017).
2. Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
3. Totoki, Y. et al. Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. *Nat. Genet.* **46**, 1267–1273 (2014).
4. Fujimoto, A. et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
5. Letouzé, E. et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
6. Kan, Z. et al. Whole-genome sequencing identifies recurrent mutations in hepatocellular carcinoma. *Genome Res.* **23**, 1422–1433 (2013).
7. Guichard, C. et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.* **44**, 694–698 (2012).
8. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
9. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
10. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
11. Martincorena, I. et al. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
12. Fellous, T. G. et al. Locating the stem cell niche and tracing hepatocyte lineages in human liver. *Hepatology* **49**, 1655–1663 (2009).
13. Sigal, S. H. et al. Partial hepatectomy-induced polyploidy attenuates hepatocyte replication and activates cell aging events. *Am. J. Physiol.* **276**, G1260–G1272 (1999).
14. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
15. Fernandez-Banet, J. et al. Decoding complex patterns of genomic rearrangement in hepatocellular carcinoma. *Genomics* **103**, 189–203 (2014).
16. Imielinski, M., Guo, G. & Meyerson, M. Insertions and deletions target lineage-defining genes in human cancers. *Cell* **168**, 460–472 (2017).
17. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
18. Torrecilla, S. et al. Trunk mutational events present minimal intra- and inter-tumoral heterogeneity in hepatocellular carcinoma. *J. Hepatol.* **67**, 1222–1231 (2017).
19. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
21. Osorio, F. G. et al. Somatic mutations reveal lineage relationships and age-related mutagenesis in human hematopoiesis. *Cell Rep.* **25**, 2308–2316 (2018).
22. Haradhvala, N. J. et al. Mutational strand asymmetries in cancer genomes reveal mechanisms of DNA damage and repair. *Cell* **164**, 538–549 (2016).
23. Poon, S. L. et al. Genome-wide mutational signatures of aristolochic acid and its application as a screening tool. *Sci. Transl. Med.* **5**, 197ra101 (2013).
24. Scelo, G. et al. Variation in genomic landscape of clear cell renal cell carcinoma across Europe. *Nat. Commun.* **5**, 5135 (2014).
25. Rushing, B. R. & Selim, M. I. Aflatoxin B1: a review on metabolism, toxicity, occurrence in food, occupational exposure, and detoxification methods. *Food Chem. Toxicol.* **124**, 81–100 (2019).
26. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
27. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
28. Nault, J. C. et al. Telomerase reverse transcriptase promoter mutation is an early somatic genetic alteration in the transformation of premalignant nodules in hepatocellular carcinoma on cirrhosis. *Hepatology* **60**, 1983–1992 (2014).
29. Kim, S. K. et al. Comprehensive analysis of genetic aberrations linked to tumorigenesis in regenerative nodules of liver cirrhosis. *J. Gastroenterol.* **54**, 628–640 (2019).
30. Zhu, M. et al. Somatic mutations increase hepatic clonal fitness and regeneration in chronic liver disease. *Cell* **177**, 608–621 (2019).

Methods

Data reporting

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

Samples

Patients recruited at Addenbrooke's Hospital, Cambridge gave written informed consent with approval of the Local Research Ethics Committee (16/NI/0196).

Normal liver samples were obtained from patients with liver metastases from colorectal carcinoma. The liver specimens were obtained from resected liver distal to the metastases, and were confirmed to be free of tumour cells by histology. None of the patients had undergone neoadjuvant systemic therapy; one patient had undergone pre-operative portal vein embolization (PD36718) to the ipsilateral liver lobe. Liver tissue from patients with chronic liver disease was derived from explanted diseased livers at the time of transplantation. All of the patients were identified as having ARLD or NAFLD by their clinical history with the transplant hepatology and addiction psychiatry teams, as well as by explanted liver histology. None of the patients had undergone transarterial chemoembolization or other locoregional therapy on the transplant waiting list, except PD37118, who underwent a single treatment to their HCC with transarterial chemoembolization. All of the patients with chronic liver disease, except one (PD37105), demonstrated substantial pre-operative impairment of liver function as evidenced by a UK model for end-stage liver disease (UKELD) score of higher than 50.

The explant liver histology was reviewed by a specialist liver histopathologist (S.E.D.), blinded to the sequencing results. The normal liver specimens had no fibrosis and no evidence of chronic liver disease; the explanted diseased livers uniformly demonstrated cirrhosis and HCC. The background liver histology was scored according to the Kleiner system[31] on formalin-fixed paraffin-embedded (FFPE) samples away from the HCC and the fresh-frozen block used for the sequencing analysis. The Kleiner score assesses the presence of steatosis, lobular inflammation and hepatocyte ballooning to generate a cumulative NAFLD activity score (NAS). The presence or absence of cellular or nodular dysplasia was assessed globally in clinical FFPE samples (Supplementary Table 1), as well as specifically in the fresh-frozen block used for the LCM and sequencing (Supplementary Table 1). Serial H&E-stained sections from the frozen block did not demonstrate dysplasia in any of the cases (Supplementary Table 1). There was no evidence of CRC or HCC on histological review of the fresh-frozen block used for sequencing.

All tissue samples were snap-frozen in liquid nitrogen and stored at −80 °C in the Human Research Tissue Bank of the Cambridge University Hospitals NHS Foundation Trust.

Preparation of tissue sections

Tissue biopsies were embedded in optimal cutting temperature (OCT) medium (Thermo Fisher Scientific) at −25 °C. Sections were cut at a thickness of 20 μm using a Leica cryotome and transferred onto polyethylene naphthalate (PEN) membrane slides (Thermo Fisher Scientific). For fixation, slides were treated with 70% ethanol at room temperature for 2 min. Slides were washed twice in 10% phosphate-buffered saline (PBS) at room temperature for 10 s. For staining, slides were incubated in haematoxylin for 10 s and rinsed twice in water. Slides were then incubated in eosin for 5 s and rinsed once in water. Slides were washed twice with 70% ethanol for 5 s, twice with 100% ethanol for 5 s and in xylene for 5 s. Storage was at −20 °C. Additional sections were stained for H&E, Masson's trichrome and Oil Red O by standard laboratory techniques. All slides were scanned on a Leica AT2 at ×20 magnification and a resolution of 0.5 μm per pixel.

Laser-capture microdissection

Microdissection was performed using a laser-capture microscope (Leica Microsystems LMD 7000). For each biopsy, 48 microdissections were cut with a target size of 20,000 μm^2, which corresponds to about 400 hepatocyte cells. Images were taken before and after LCM.

Sample lysis and DNA preparation

LCM biopsies were lysed using the Arcturus PicoPure DNA Extraction Kit (Thermo Fisher Scientific) following the manufacturer's instructions. DNA libraries for Illumina sequencing were prepared using a protocol optimized for low input amounts of DNA, as described[32].

Whole-genome sequencing

Paired-end sequencing reads (150 bp) were generated using the Illumina X10 platform for 400 samples, resulting in a target coverage of 30×–70× per sample. To avoid the known index-hopping artefact, we chose to avoid multiplexing samples and instead sequenced one sample per flow-cell lane. To increase coverage for a subset of 96 samples, we used multiplexing and achieved 70× coverage. In addition to the LCM samples, we also sequenced a bulk sample for each biopsy and (where available) associated HCC.

The healthy liver samples came from wide resections of hepatic metastases of colorectal cancer. In each case, we sequenced the metastasis; this did not reveal any mutations that were shared between the colorectal cancer and liver, or any variants that were shared by all liver samples absent from the colorectal cancer (beyond regions of loss of heterozygosity in the cancer). Likewise, for the cirrhotic liver samples, we sequenced the matched HCC, which did not revealing any sharing of mutations. In one case, we sequenced microdissections of the fibrotic tissue, and here we also did not find mutations restricted to all liver cells.

Sequencing data were mapped to the human genome, GRCh37d5, using the BWA-MEM algorithm.

Calling of SNVs

Substitution variants were called using the Cancer Variants through Expectation Maximization (CaVEMan) algorithm[33], using the bulk sample of the liver biopsy as the matched normal. As part of the algorithm, the variants were annotated using VAGrENT[34]. Variant calls for bulk sequencing data of the cancer samples were not further filtered. For sequencing of LCMs, post-filtering was performed in three steps.

1) Removal of duplicate counts. We noticed instances in which variant bases were counted twice, owing to the overlap of paired-end sequencing reads. We removed such double counting and re-evaluated variant calls after taking double counts into account.

2) Removal of variants that were introduced during library preparation. We noticed the presence of variants that were introduced owing to incorrect processing of cruciform DNA. Erroneous variants were often present in inverted repeats and frequently accompanied by another proximal (~1–30 bp distance). These inverted repeats can form cruciform DNA before the isolation of DNA or during library preparation. The library preparation protocol used can incorrectly process these secondary DNA structures and inadvertently introduce one or more erroneous variants. For every variant the standard deviation (s.d.) and median absolute deviation (MAD) of the variant position within the read was separately calculated for positive and negative strand reads.

In the case that the variant was supported by a low number of reads for a particular strand, the filtering was based on the statistics determined from the reads derived from the other strand. It was required that either i) 90% of supporting reads reported the variant within the first 15% of the read, as calculated from the alignment start; or ii) the MAD exceeded 0 and the s.d. exceeded 4. In the case that sufficient reads supporting the variant were available for both strands it was required for both strands separately that i) ≤90% of supporting reads reported the variant within the first 15% of the read as calculated from the alignment start; ii) the

MAD exceeded 2 and the s.d. exceeded 2; or iii) at least one strand fulfilled the criteria of a MAD greater than 1 and s.d. greater than 10.

3) Comparison with an independent panel. To remove variant calls at badly mapping sites, we compared variant calls in the sequenced samples of each donor biopsy with samples from all unrelated donors in our cohort. For each variant site we expected the reference base to be dominant and conversely expected badly mapping sites to contain frequent non-reference base counts. Thus, we counted the numbers of A, C, G, T indel calls at each variant site across all unrelated samples, resulting in a large 'pileup' table. The dominance of the reference base was evaluated at each variant site using the entropy purity metric E:

$$E = -\sum_i P(x_i)\ln P(x_i)$$

in which x is the count of base $i \in \{A,C,G,T\}$ and the $P(x_i)$ are the fractions of base calls. Values of E close to 0 indicate that almost all reads in the independent panel contain a single base. Higher values of E indicate a mix of base calls at the site. To identify an optimal threshold of E for the filtering of variant sites, we evaluated the entropy metric against a labelled dataset of variant calls. Specifically, during the clustering of variants using the Bayesian Dirichlet process (described below), we identified clusters that had variants with low allele frequency present in all dissections from the same donor. Manual inspection showed that such variants occurred at badly mapping sites. Thus, we labelled variant sites in those clusters as 'badly mapping' and were able to use the area under the receiver operator curve (AUC) to identify a threshold value E_{Thr} of 0.16; this allowed us to separate the two labelled variant groups with an AUC of 0.99.

Bayesian Dirichlet process for clustering VAFs across multiple samples

We extend the model previously developed for clustering VAFs of mutations called in a single sample[19] to mutation data across multiple samples from the same individual. In normal somatic cells, the vast majority of the genome retains its normal, diploid copy number, which means that we can cluster the VAFs directly (excluding mutations on the X and Y chromosomes in males). This has the considerable advantage that the Dirichlet process model we build can rely directly on conjugate prior distributions. The model includes a potential split-merge step at each cycle of the Gibbs sampler, following a previously described Metropolis–Hastings proposal for conjugate distributions[35]. The algorithm could be extended to include a correction for different copy-number states in given samples for a particular mutation through, for example, a Metropolis–Hastings update, but at considerable computational cost. The full mathematical development of the model is detailed in the Supplementary Methods.

We ran the Gibbs sampler for 15,000 iterations, dropping the first 10,000 as a burn-in. We used the ECR algorithm[36], implemented in the R package label.switching, to resolve the label-switching problem associated with mixture models. We dropped clusters that contained more than100 variant sites.

Construction of phylogenetic trees

Phylogenetic trees were constructed manually using the pigeonhole principle, as described previously[19]. In brief, each cluster that was identified using the Bayesian Dirichlet process represented a branch of the phylogenetic tree. Nesting of trees was identified with three different levels of certainty, illustrated on a pair of branches, A and B. 1) In the case that the median VAFs of A and B exceeded 100%, the pigeonhole principle defines that A and B are nested. 2) We can assume that non-hepatocyte cells constitute a sizeable fraction of each LCM sample. Assuming a non-hepatocyte fraction of 30%, we nested branches when the VAFs of A and B exceeded 70%. This non-hepatocyte fraction was chosen as a conservative estimate of the fraction of cells intermixed in our microdissections that are not derived from the hepatocyte clone, on the basis of observed VAF peaks in our data together with single-cell RNA sequencing data from liver tissue. 3) If identical LCMs are members of both A and B, it is highly likely that A and B are nested, rather than independent branches. Thus, we also nested branches in cases in which the LCMs in one branch were a subset of the LCMs in the other (parental) branch.

For each nesting scenario, we defined the parental branch as the one with the higher median VAF in the contained LCMs. We highlighted the evidence level for nesting in each representation of phylogenetic trees, marking branches with evidence level 1 with a solid line, level 2 with a dashed line and level 3 with a dotted line.

Analysis of driver variants

We curated a list of genes that have been found to be significantly mutated in liver cancers in a selection of published studies[1–4,6,7,37–39], as shown in Supplementary Table 5. Using the VAGrENT annotations[34], we counted any regulatory, missense, nonsense, frameshift or essential splice variant as a potential driver variant. To systematically identify genes under mutagenic selection, we used the dN/dS method[17], which screens for genes with an excess of non-synonymous mutations compared to that expected from the synonymous mutation rate.

Sensitivity correction

We identified 138 pairs of LCMs with a midpoint-to-midpoint distance of <500 μm and at least one shared cluster according to the Bayesian Dirichlet process. These LCMs we assumed to represent the same clone, thus providing an opportunity to calculate the sensitivity of calling a variant present in one LCM in the other. If we assume the sensitivity is the same in both samples, then the maximum likelihood estimate for the sensitivity, when mutations not called in either sample are unobserved, is given by:

$$s = \frac{2n_2}{n_1 + 2n_2}$$

in which n_2 is the number of variants called in both LCMs in each pair and n_1 is the number of variants called only in one of the two LCMs. To evaluate the relationship of sensitivity with depth of coverage and VAF, we performed a logistic regression of sensitivity against these two predictors using the lm() function of the R programming language. The model fit was then used to calculate sensitivity for any LCM sample, given the coverage and VAF of the sample.

Analysis of mutational burden

We used a linear mixed effects model to fit the number of variants per LCM sample against the disease aetiology (normal or cirrhotic) and age for each individual. We defined the ID of the individual as a random effect. The slope of the age coefficient was allowed to vary with the random effect. To facilitate the analysis, we used the lmer() function within the lme4 package of the R programming language. To determine the significance of the aetiology and age coefficients, we used an analysis of variance (ANOVA) to perform a χ^2 test that compared our model with models omitting the aetiology and age coefficients, respectively.

Targeted deep sequencing validation of mutation calls

For 96 of the microdissections sequenced by whole-genome sequencing, we performed a targeted deep sequencing validation using an Agilent RNA bait set that covered 350 recurrently mutated cancer genes. Among these genes, a total of 17 mutations were identified in the whole-genome sequencing data from the 96 samples; of these, 16 (94%) were validated, at comparable VAFs, in the targeted deep sequencing data.

Calling of indels

Indels were called using cgpPindel[40]. Variant calls for bulk sequencing data of the cancer samples were not further filtered. To remove

artefactual calls from the LCM-derived data, we performed two post-filtering steps.

1)Assignment to SNV-based clusters. We evaluated how well the VAF distribution of each indel across the LCMs from the same donor compared with the VAF distribution of each SNV-based cluster as identified by the Bayesian Dirichlet process. Given an indel in one LCM sample, we thus counted its occurrence in all related LCMs and assigned the resulting VAF profile to the VAF profiles of the SNV clusters using a Bayes' classifier. We noticed that many indels were assigned to SNV clusters with more than 100 variants, which we had previously removed from the SNV analysis. On closer inspection we noticed that those INDELs had low VAF and occurred frequently in badly mapping regions. We thus discarded indels that were assigned to those clusters.

2) Filtering on the basis of beta-binomial overdispersion parameter. We noticed that many indels occurred with low VAF in a large number of LCMs from the same donor and were, thus, probably artefactual. To systematically identify such indels, we fitted the beta-binomial distribution to the variant counts of each indel across the LCMs from the same donor. The fitted parameter ρ (the overdispersion parameter) was used to filter indel calls. A high value for parameter ρ (overdispersion) occurs when some LCMs have many variant read counts and others few or none. Conversely, a low value occurs when all LCMs have a similar number of variant counts (no overdispersion). On the basis of manual inspection, we removed variant calls with $\rho < 0.02$.

Calling of copy numbers

Copy numbers were called using the ASCAT algorithm[41], assuming an expected ploidy of 4 (to allow for physiologically polyploid hepatocytes) and 60% non-hepatocyte cell contamination for all samples. Testing of robustness around these starting points (different expected ploidy or purity values) found that the specific values used did not materially affect the output. Variant calls for bulk sequencing data of the cancer samples were not further filtered. To remove artefactual variants from the LCM-derived data, we used the SNV-based phylogenetic information. The genome was segmented into 500-bp bins and the ASCAT-based copy number of each bin was calculated. Using the binned copy-number data we calculated the median copy number in each LCM sample and ASCAT event. For each ASCAT event and LCM sample we assigned its absolute deviation from the diploid state. We compared the copy-number profile for each ASCAT event across the LCM samples with the VAF profile of each SNV cluster using cosine similarity (described below) to identify the most similar SNV cluster. Within each SNV cluster we proceeded to merge overlapping ASCAT events. Using manual inspection, we decided to keep ASCAT events if 1) they had a cosine similarity of <0.1 to an SNV cluster; and 2) their assigned SNV cluster was not removed during SNV analysis owing to having more than 100 assigned SNVs.

Calling of structural variants

Structural variants were called using the BRASS algorithm[42] (https://github.com/cancerit/BRASS). Variant calls for bulk sequencing data of the cancer samples were not further filtered. To remove artefactual variants from the LCM-derived data, we used post-processing filters. Manual inspection of the sequencing reads identified for each structural variant showed that many reads were identical except for frameshifts at repetitive sites. We decided that such reads represented duplicates and designed a filter to systematically remove these. We removed structural variants that were supported by more than two reads after removal of duplicates. Each remaining structural variant call was manually inspected.

Calculation of clone size

We determined the midpoint coordinates of each LCM manually from the microscopy images collected during dissection. For each LCM that belonged to a clone as determined by the Bayesian Dirichlet process, we used the chull function of the R programming language to identify the coordinates of the convex hull that included all LCMs. We identified the midpoint of each polygon as the average coordinate of all convex hull vertices. The size of the clone was then assigned to be the Euclidean distance between each convex hull vertex and the midpoint of the polygon. For clones that only consisted of a single LCM, we assigned the minimum clone size discovered across all clones.

Extraction of mutational signatures from SNV contexts using HDP

Mutational signatures were extracted using the HDP package (https://github.com/nicolaroberts/hdp), relying on the Bayesian hierarchical Dirichlet process. The units of signature extraction were mutations assigned to individual branches of the phylogenetic tree, grouped per patient, from the LCM data. In addition, to provide a comparison against signatures extracted in HCCs, we added catalogues of somatic substitutions from 54 whole genomes sequenced by TGCA, analysed using the same core algorithms as were used for the LCM data. The tool was used without defining prior signatures. As hyperparameters we set α and β to 6 for the α clustering parameter. Extraction was started with 40 data clusters (parameter 'initcc'). The Gibbs sampler was run with 10,000 burn-in iterations (parameter 'burnin'). With a spacing of 50 iterations (parameter 'space'), 50 iterations were collected (parameter 'n'). After each Gibbs sampling iteration, three iterations of concentration parameter sampling were performed (parameter 'cpiter'). The resulting signatures were compared to published signatures[20,43] using the cosine similarity metric described below. Extracted signatures with cosine similarity >0.9 compared to a known signature from either the COSMIC[20] or PCAWG[43] catalogue of signatures were assigned the name of the known signature with the highest similarity. Extracted signatures with cosine similarity <0.9 compared to any of the known signatures were assigned new names, which were indexed with the letters A, B and C.

Extraction of mutational signatures from SNV contexts using SigProfiler

We used SigProfiler to extract mutational signatures, relying on the non-negative matrix factorization method[44]. In particular, we report the 'Decomposed Solution' output by the SigProfiler package.

Cosine similarity calculation

To compare two vectors A and B, cosine similarity was calculated as follows:

$$\text{Similarity} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Analysis of the proportion of indels and gene expression

A list of transcribed regions was retrieved from Ensembl using the BioMaRt package[45]. We identified the subset of indel and SNV variants that overlapped with the transcribed regions. The proportion of indels in comparison to the total number of indels and SNVs per gene was calculated. Gene expression was assigned using the liver dataset from the Genotype-Tissue Expression project (GTEx)[46]. To test for the relationship between gene expression and the proportion of indels, we fitted a Poisson regression using the glm function of the R programming language. We modelled the number of indels per gene against an offset of the total number of variants per gene and the expression of the gene.

Analysis of T>C transcriptional-strand bias at transcription start sites

We performed this analysis in a similar way to a published approach[22]. In brief, we retrieved the genomic coordinates of transcription start sites of all the highly expressed genes in the liver from GTEx[46]. We tiled the 10 kb upstream and downstream of the transcription start site into 1,000-bp bins. We overlapped all T>C (transcribed) and A>G (non-transcribed)

variant calls with the tiled regions and summed the number of variants in each tile across all included genes. We also extracted the number of T and A bases in each tile. To test whether the strand bias was significant only in transcribed regions, we fitted a Poisson regression for the number of variant calls against the following predictors: strand (transcribed, non-transcribed), distance from transcription start site (0 for upstream, 1 for downstream) and aetiology (cirrhosis, no cirrhosis), and used the number of T and A bases in each tile as the offset variable.

Analysis of C>A and T>A transcriptional-strand bias

We used the MutationalPatterns package[47] to assign the transcription state for each C>A variant. We retrieved the genomic coordinates of all transcribed regions from Ensembl using the BioMaRt package[45] and extracted the frequencies of C and G nucleotides in these regions. To test for the significance of transcriptional-strand bias, we performed a Poisson regression for the number of C>A variants in each sample and transcription strand against factor variables for the transcription strand, the patient ID and an interaction term for the two factors. We used the C and G nucleotide frequencies as an offset variable. To test for the significance of transcriptional-strand bias for a given donor, we coded the patient ID in a binary fashion: '1' for the target donor, '0' otherwise. We proceeded to test for transcriptional-strand bias of T>A variants in a similar way, using A and T nucleotide frequencies as the offset.

Reporting summary

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

Data availability

Whole-genome sequencing data across the samples reported in this study have been deposited in the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) in the form of BAM files, with accession number EGAD00001004578. Substitution and indel calls have been deposited on Mendeley Data with the identifier https://doi.org/10.17632/ktx7jp8sch.1 ('Somatic mutations and clonal dynamics in healthy and cirrhotic human liver').

Code availability

Single-nucleotide substitutions were called using the CaVEMan algorithm, v.1.11.2 (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm, v.2.2.2 (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm v.5.4.1 (https://github.com/cancerit/BRASS). Miscellaneous scripts for downstream analysis are available on Github (https://github.com/sfbrunner/liver-pub-repo). The analysis of mutational signatures was performed using the HDP hierarchical Dirichlet process package v.0.1.5, which is available on Github (https://github.com/nicolaroberts/hdp).

31. Kleiner, D. E. et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).
32. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* https://doi.org/10.1038/s41586-019-1672-7 (2019).
33. Jones, D. et al. cgpCaVEManWrapper: Simple execution of CaVEMan in order to detect somatic single nucleotide variants in NGS data. *Curr. Protoc. Bioinformatics* **56**, 15.10.1–15.10.18 (2016).
34. Menzies, A. et al. VAGrENT: Variation Annotation Generator. *Curr. Protoc. Bioinformatics* **52**, 15.8.1–15.8.11 (2015).
35. Dahl, D. B. *An Improved Merge-Split Sampler for Conjugate Dirichlet Process Mixture Models. Technical Report No*. 1086 (Univ. Wisconsin–Madison, 2003).
36. Papastamoulis, P. label.switching: An R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw*. **69**, https://doi.org/10.18637/jss.v069.c01 (2016).
37. Fujimoto, A. et al. Whole-genome mutational landscape of liver cancers displaying biliary phenotype reveals hepatitis impact and molecular diversity. *Nat. Commun*. **6**, 6120 (2015).
38. Cleary, S. P. et al. Identification of driver genes in hepatocellular carcinoma by exome sequencing. *Hepatology* **58**, 1693–1702 (2013).
39. Ahn, S.-M. et al. Genomic portrait of resectable hepatocellular carcinomas: implications of *RB1* and *FGF19* aberrations for patient stratification. *Hepatology* **60**, 1972–1982 (2014).
40. Raine, K. M. et al. cgpPindel: Identifying somatically acquired insertion and deletion events from paired end sequencing. *Curr. Protoc. Bioinformatics* **52**, 15.7.1–15.7.12 (2015).
41. Raine, K. M. et al. ascatNgs: Identifying somatically acquired copy-number alterations from whole-genome sequencing data. *Curr. Protoc. Bioinformatics* **56**, 15.9.1–15.9.17 (2016).
42. Campbell, P. J. et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nat. Genet*. **40**, 722–729 (2008).
43. Alexandrov, L. et al. The repertoire of mutational signatures in human cancer. Preprint at https://www.biorxiv.org/content/10.1101/322859v2 (2019).
44. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Rep*. **3**, 246–259 (2013).
45. Durinck, S., Spellman, P. T., Birney, E. & Huber, W. Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat. Protoc*. **4**, 1184–1191 (2009).
46. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
47. Blokzijl, F., Janssen, R., van Boxtel, R. & Cuppen, E. MutationalPatterns: comprehensive genome-wide analysis of mutational processes. *Genome Med*. **10**, 33 (2018).

Acknowledgements This work was supported by a Wellcome Trust and Cancer Research UK (CRUK) Grand Challenge Award (C98/A24032). P.J.C. is a Wellcome Trust Senior Clinical Fellow (WT088340MA); S.F.B. was supported by the Swiss National Science Foundation (P2SKP3-171753 and P400PB-180790); M.A.S. is supported by a Rubicon fellowship from NWO (019.153LW.038); the Cambridge Human Research Tissue Bank is supported by the NIHR Cambridge Biomedical Research Centre; and M.H. is supported by a CRUK Clinician Scientist Fellowship (C52489/A19924).

Author contributions P.J.C., M.H. and S.F.B. designed the experiments; S.F.B. performed the LCM, data curation and statistical analysis, with L.A.W., M.A.S., F.A. and I.M. providing assistance and advice; M.H., S.J.A. and S.E.D. collated and analysed the clinical and histological data from the patients; N.D.R. developed the hierarchical Dirichlet process for extracting mutational signatures; L.M. and P.E. developed the LCM, DNA extraction and library production protocol used; C.A. and Y.H. assisted with sample preparation, processing and tracking; P.J.C., I.M. and M.R.S. oversaw the analysis of mutational signatures and selection analyses; P.J.C., M.H. and S.F.B. wrote the manuscript, with contributions from all authors.

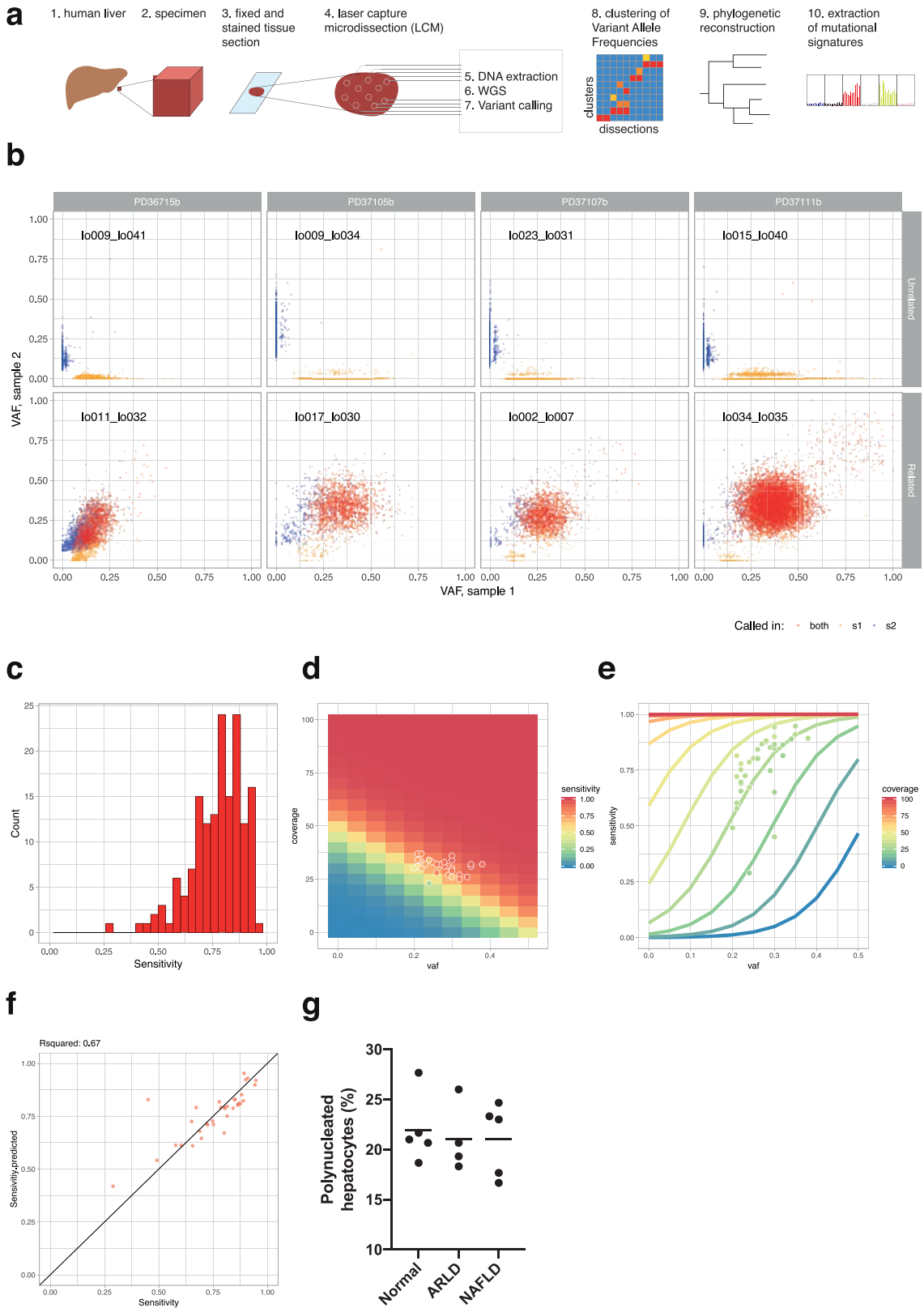
Competing interests The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1670-9.

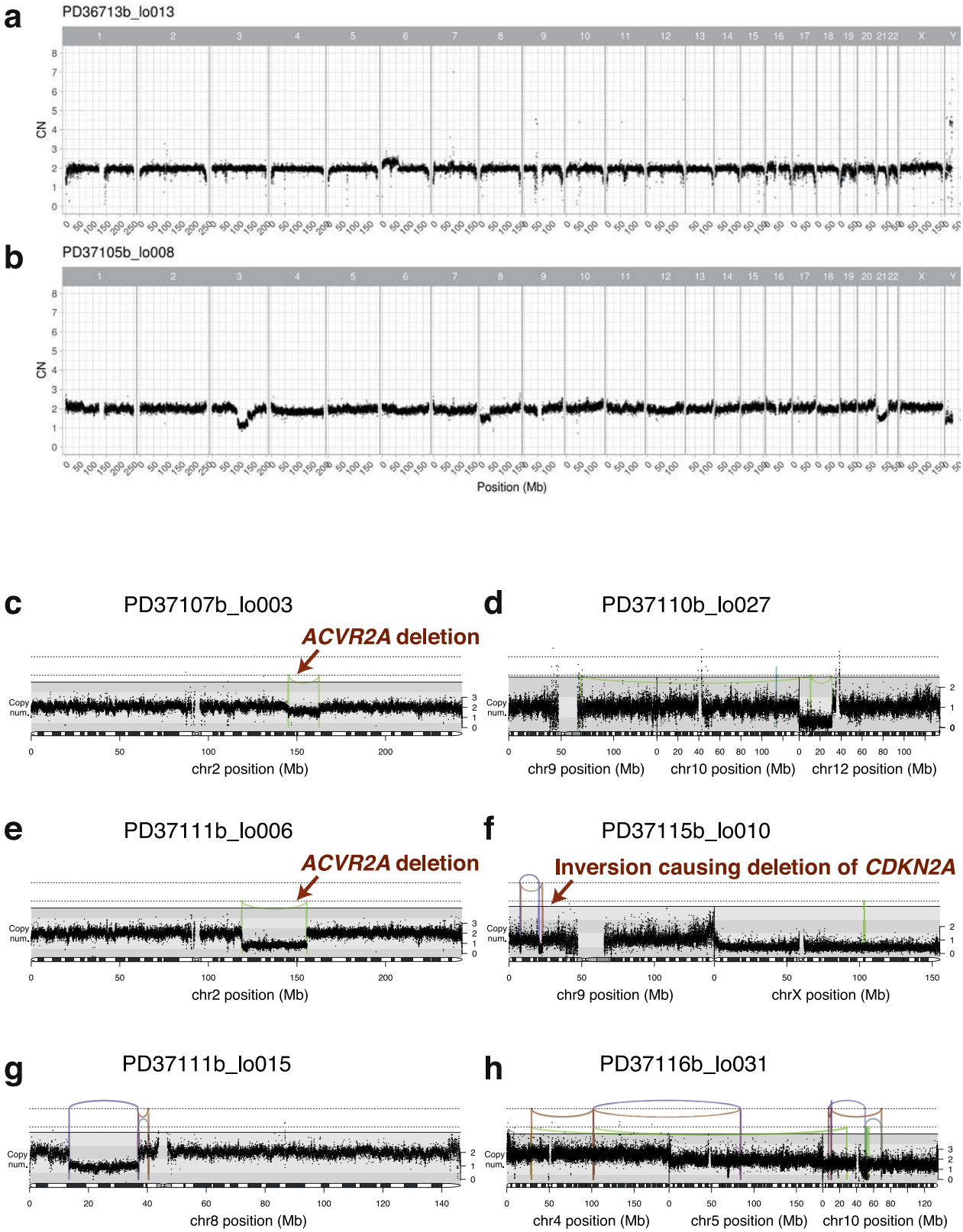
Correspondence and requests for materials should be addressed to M.H. or P.J.C.

Peer review information *Nature* thanks Jessica Zucman-Rossi and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

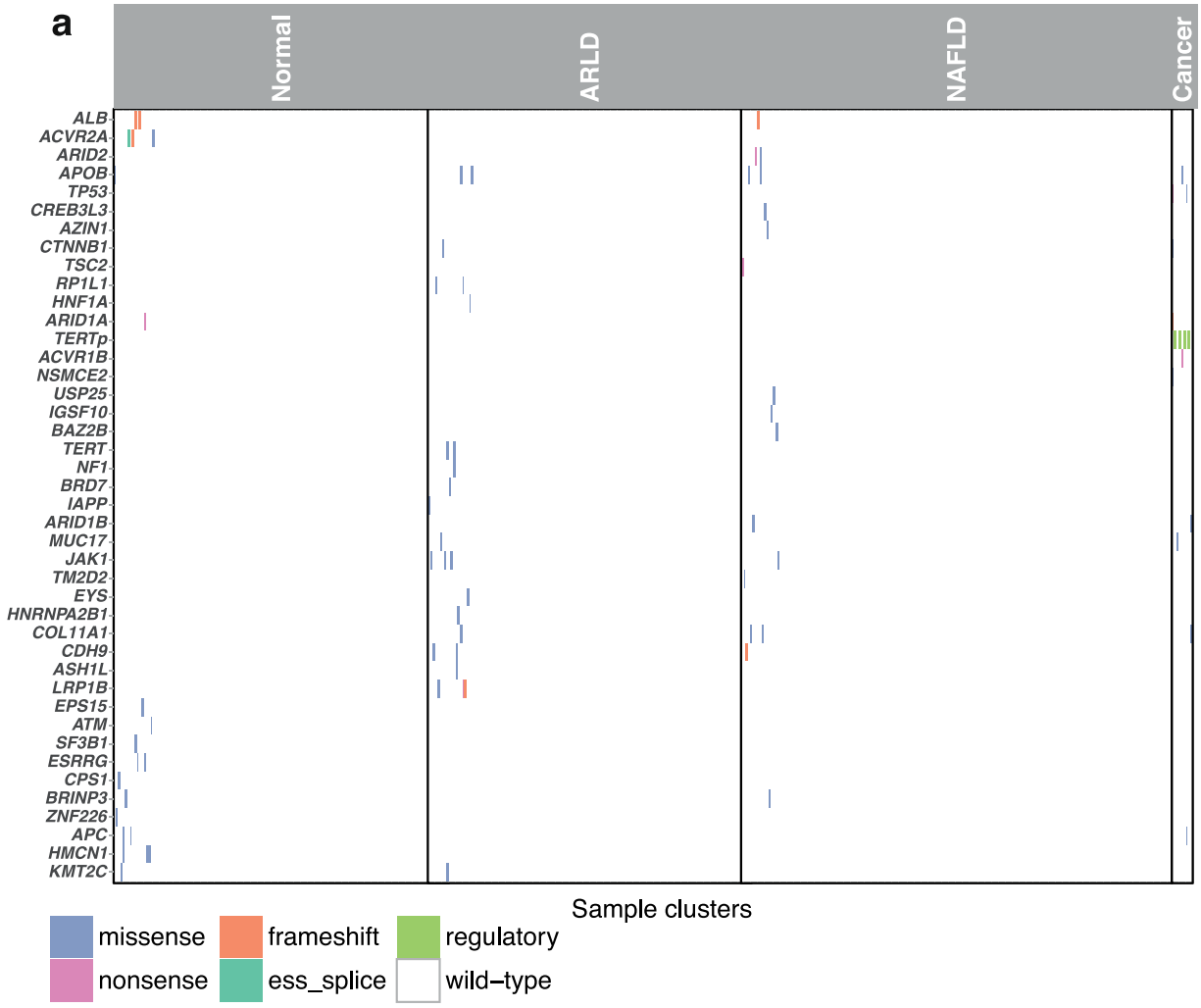


Extended Data Fig. 1 | See next page for caption.

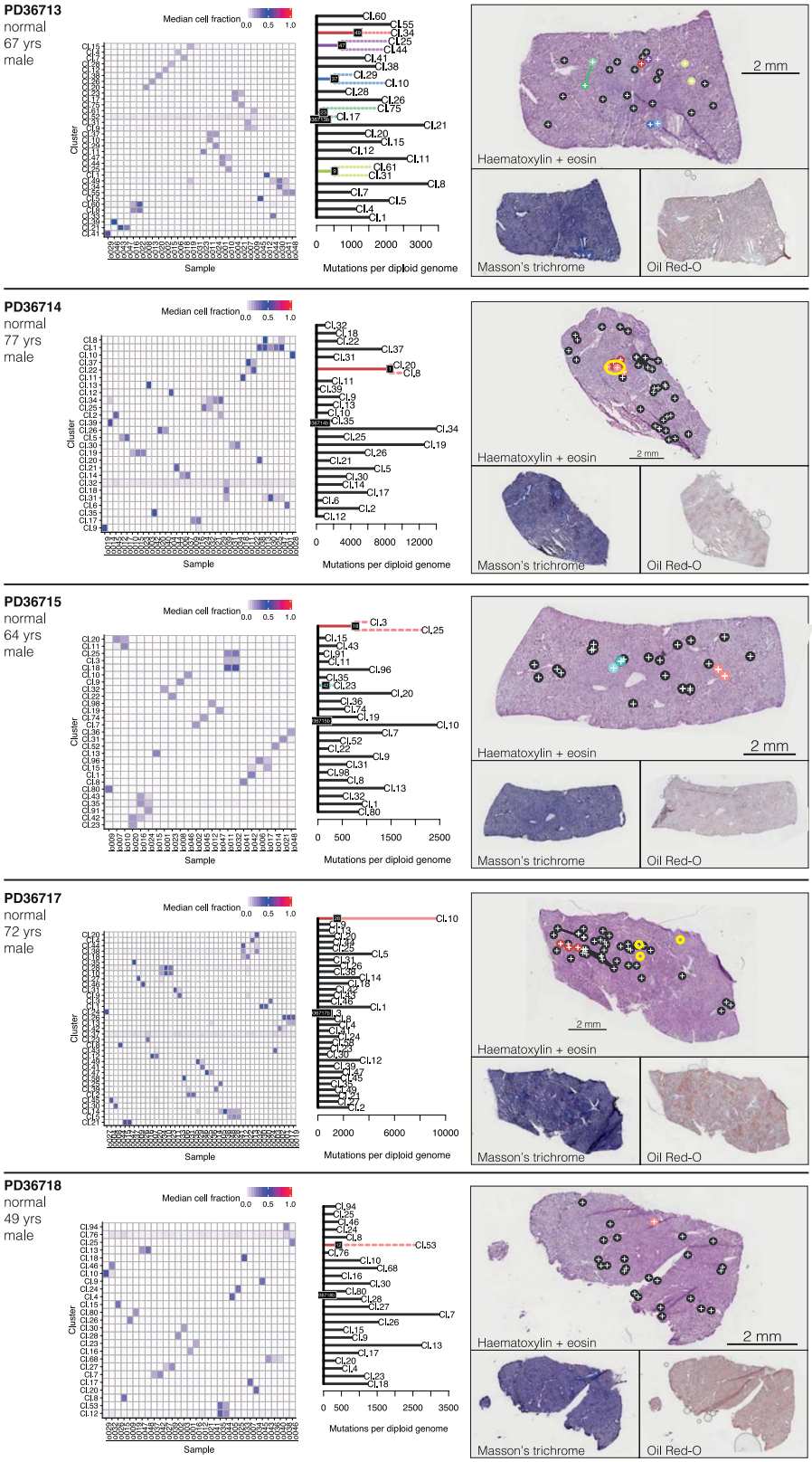
Extended Data Fig. 1 | Sensitivity analysis of SNV calls. a, Overview schematic of the experimental and analytical approach. **b**, Examples of the VAFs of variants from unrelated (top) and related (bottom) pairs of microdissection samples from four donors (left to right). The *x* axis represents the VAF of sample 1 from each pair and the *y* axis represents the VAF of sample 2. Each dot represents one variant. Red, variants called in both samples; yellow, variants called in sample 1; blue, variants called in sample 2. **c**, Histogram of sensitivities calculated for each sample pair. **d**, Heat map of modelled sensitivity at different values of VAF and coverage. The overlaid dots represent the sample pairs that were used to fit the model. **e**, Relationship of VAF, sensitivity and coverage according to the fitted model of sensitivity. The overlaid dots represent the sample pairs that were used to fit the model. **f**, Comparison of calculated (*x* axis) and fitted (*y* axis) sensitivity for each sample pair ($n = 34$ pairs of samples). The R^2 value is the Pearson's correlation coefficient. **g**, Proportion of hepatocytes that are multinucleated in the samples analysed here, estimated by counting 500 cells in each H&E-stained section ($n = 14$ patients). Each point represents the proportion for one patient in the study. The horizontal bars represent the mean for that aetiological group.



Extended Data Fig. 2 | Copy-number and structural variants in chronic liver disease. a, **b**, Genome-wide copy-number profiles for two samples. Black points represent the read depth of discrete windows along the chromosome, corrected to show overall copy number. Arm-level and whole-chromosome gains and losses are evident. **c**–**h**, Focal copy-number changes and structural variants. Black points represent the read depth of discrete windows along the chromosome, corrected to show overall copy number. Lines and arcs represent individual structural variants, coloured by the orientation of the joined ends (purple, tail-to-tail inverted; brown, head-to-head inverted; turquoise, tandem-duplication-type orientation; green, deletion-type orientation). Events that affect known HCC genes are marked with labelled arrows (**c**, **e**, **f**).

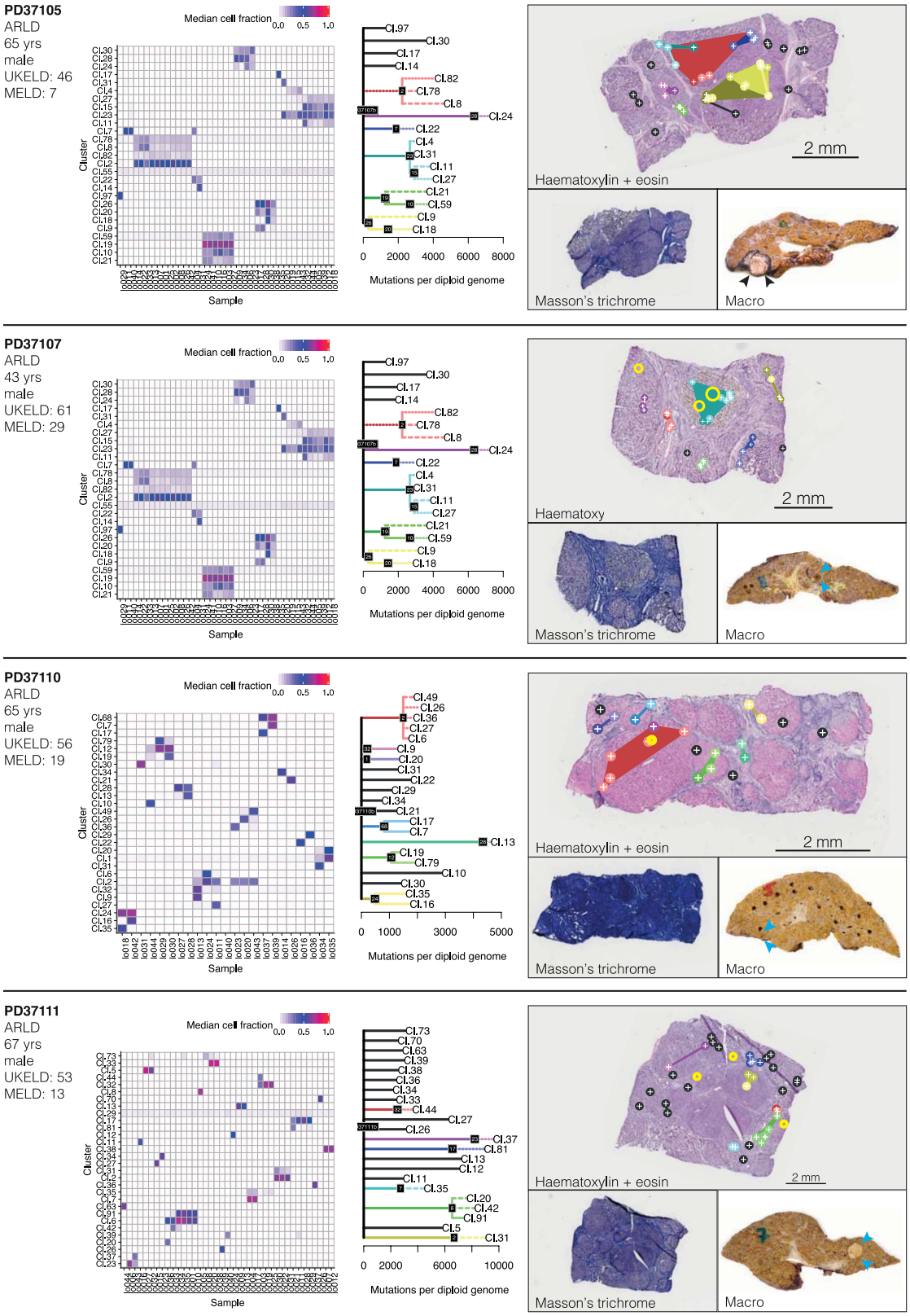


Extended Data Fig. 3 | Events that affect known HCC genes in the cohort. **a**, Distribution of somatic point mutations in individual microdissections (*x* axis) affecting known HCC genes (*y* axis), coloured by class of mutation according to the key underneath the panel. *TERTp*, *TERT* promoter. **b**, Genomic position of SNVs (top; light-blue strip) and indels (bottom; dark-blue strip) detected in *ALB*, the gene encoding albumin. **c**, Relationship of gene expression in liver tissue (*x* axis) and the proportion of indels as a fraction of all point mutations (*y* axis). The grey line represents a Poisson regression model with a significant (two-sided likelihood ratio test; $P < 10^{-16}$) coefficient for gene expression as a predictor for the ratio of indels ($n = 5,458$ genes included in the model). The grey ribbon represents the 99% confidence interval of the parameter estimates.

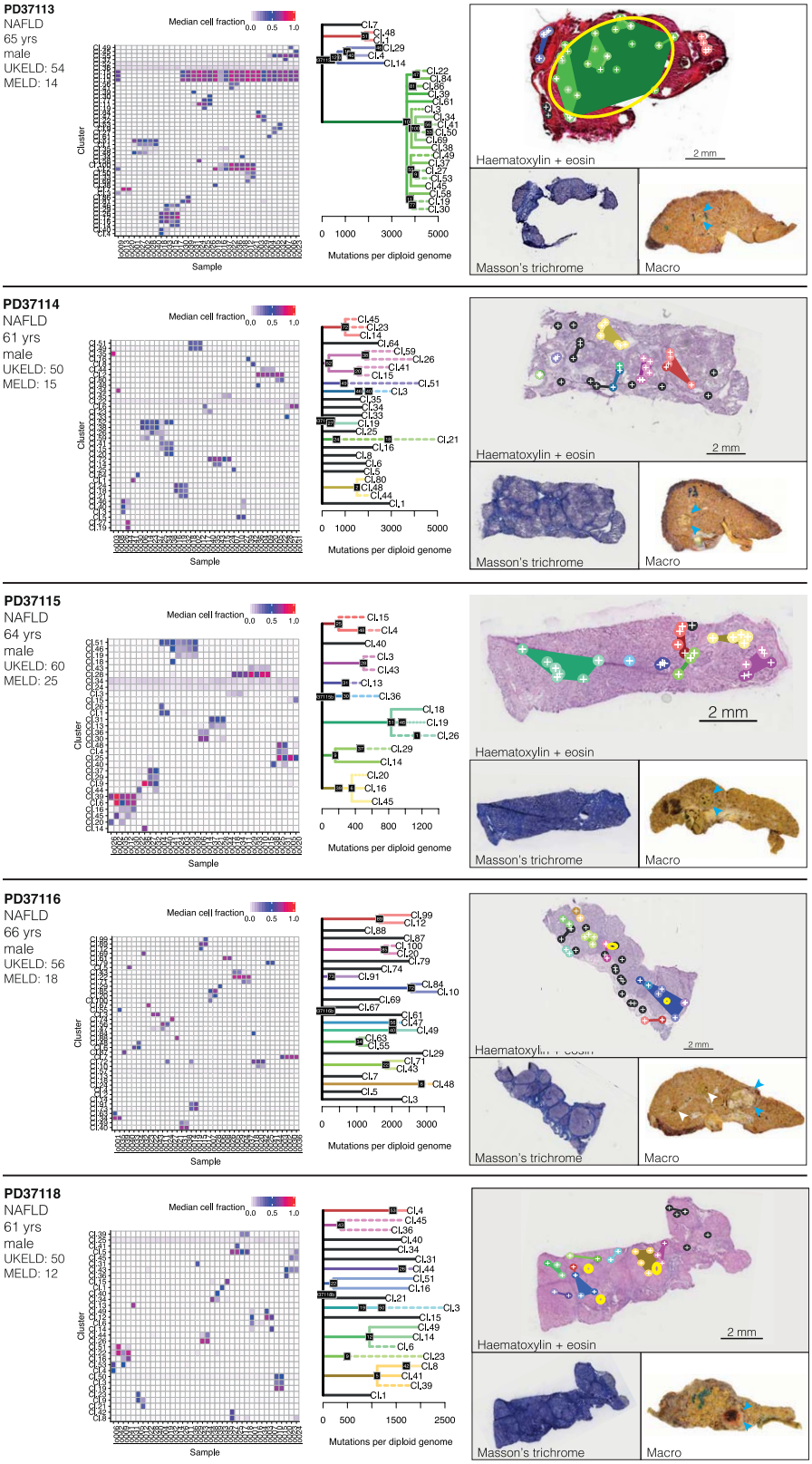


Extended Data Fig. 4 | See next page for caption.

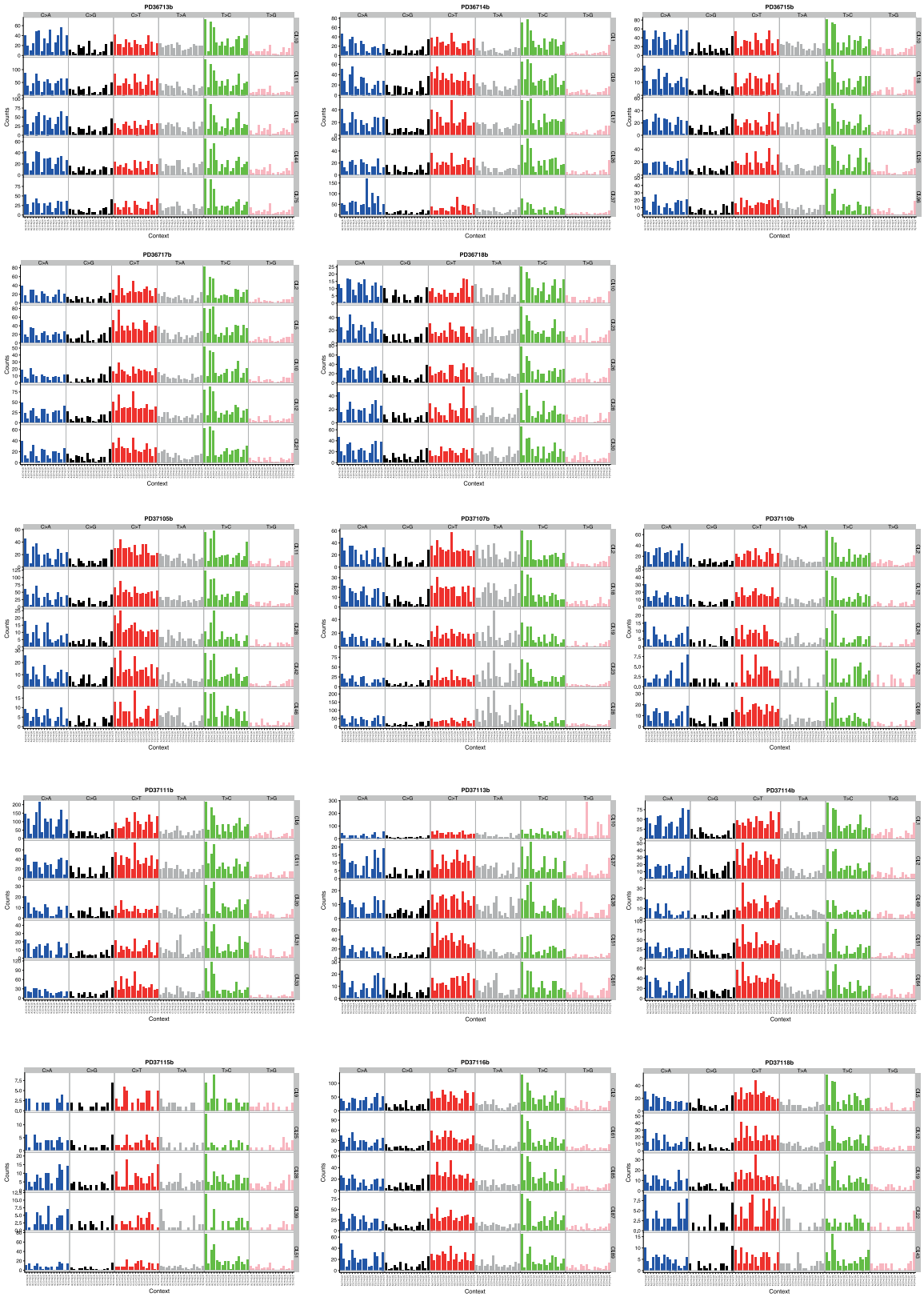
Extended Data Fig. 4 | Phylogenetic reconstruction of hepatocyte clones in non-cirrhotic liver samples. Left, heat maps representing the clustering of the variants observed in each microdissection sample (*x* axis) of the non-cirrhotic livers. Each cluster (*y* axis) contains mutations for which the VAFs across samples are very similar. The colour scale of the boxes represents the estimated mean VAF for that cluster in that sample. Middle, phylogenetic trees constructed from the clustering information. Solid lines indicate that nesting is in accordance with the pigeonhole principle; dashed lines indicate that nesting is in accordance with the pigeonhole principle, assuming that hepatocytes represent 70% of cells; dotted lines indicate that nesting is only based on clustering (a clone is assigned as nested if its constituent LCMs are a subset of LCMs in the parental clone). For details, see Supplementary Methods. Right, representation of clones according to the physical coordinates of the LCM samples, overlaid onto H&E-stained sections (top). Sections stained with Masson's trichrome and Oil Red O are also shown (bottom). Locations of immune or inflammatory cell infiltrates are marked with yellow rings. Sample sizes: PD36713, $n = 30$ microdissections; PD36714, $n = 35$ microdissections; PD36715, $n = 26$ microdissections; PD36717, $n = 42$ microdissections; PD36718, $n = 32$ microdissections.



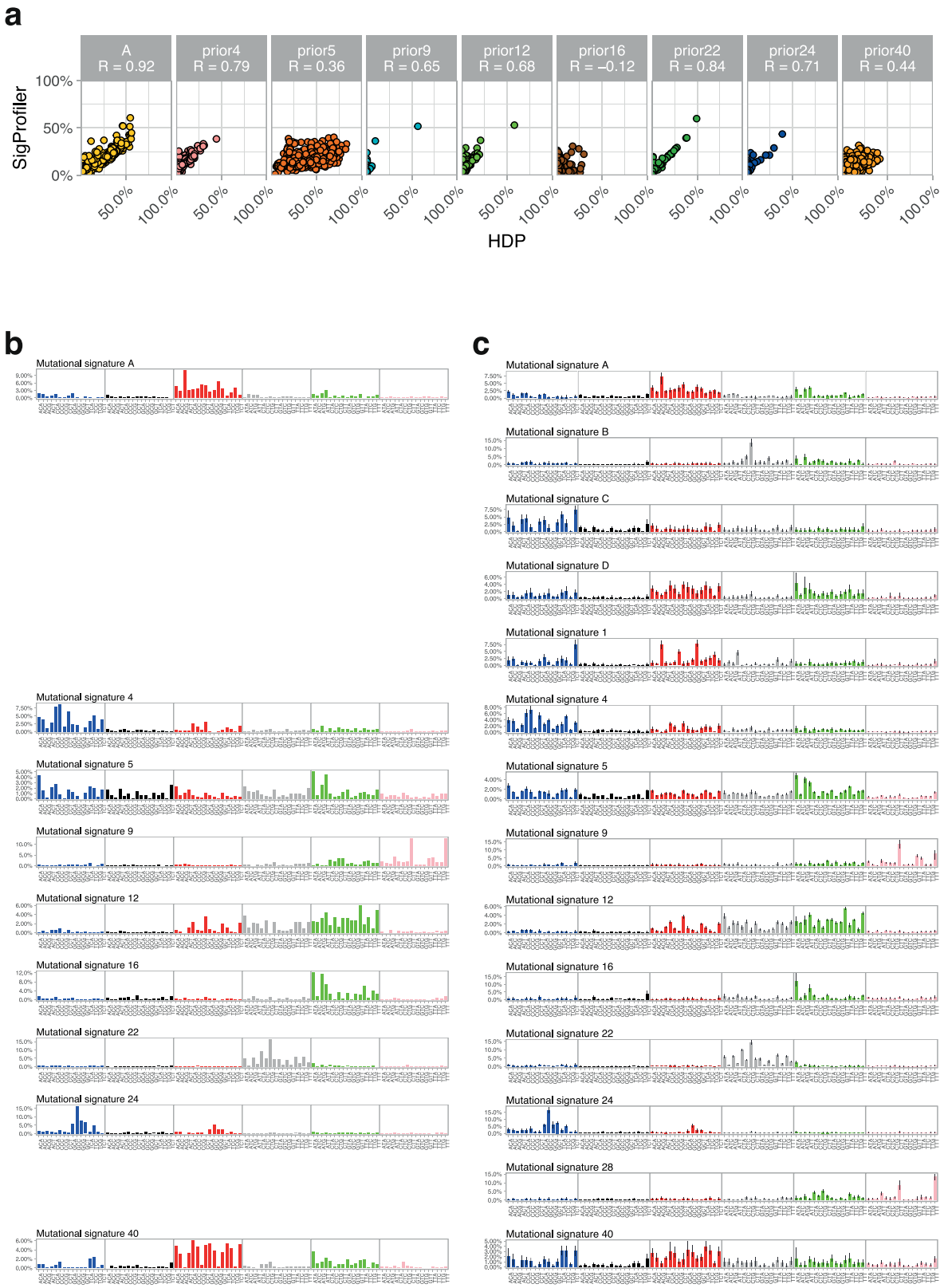
Extended Data Fig. 5 | Phylogenetic reconstruction of hepatocyte clones in alcohol-related cirrhosis. Analogous to Extended Data Fig. 4, but for the cirrhotic livers of donors PD37105, PD37107, PD37110 and PD37111. Right, H&E-stained sections (top); Masson's trichrome-stained sections (bottom left); and macroscopic photographs of the liver, with HCCs indicated by arrows (bottom right). Locations of immune or inflammatory cell infiltrates are marked with yellow rings. Sample sizes: PD37105, $n = 31$ microdissections; PD37107, $n = 41$ microdissections; PD37110, $n = 22$ microdissections; PD37111, $n = 39$ microdissections.



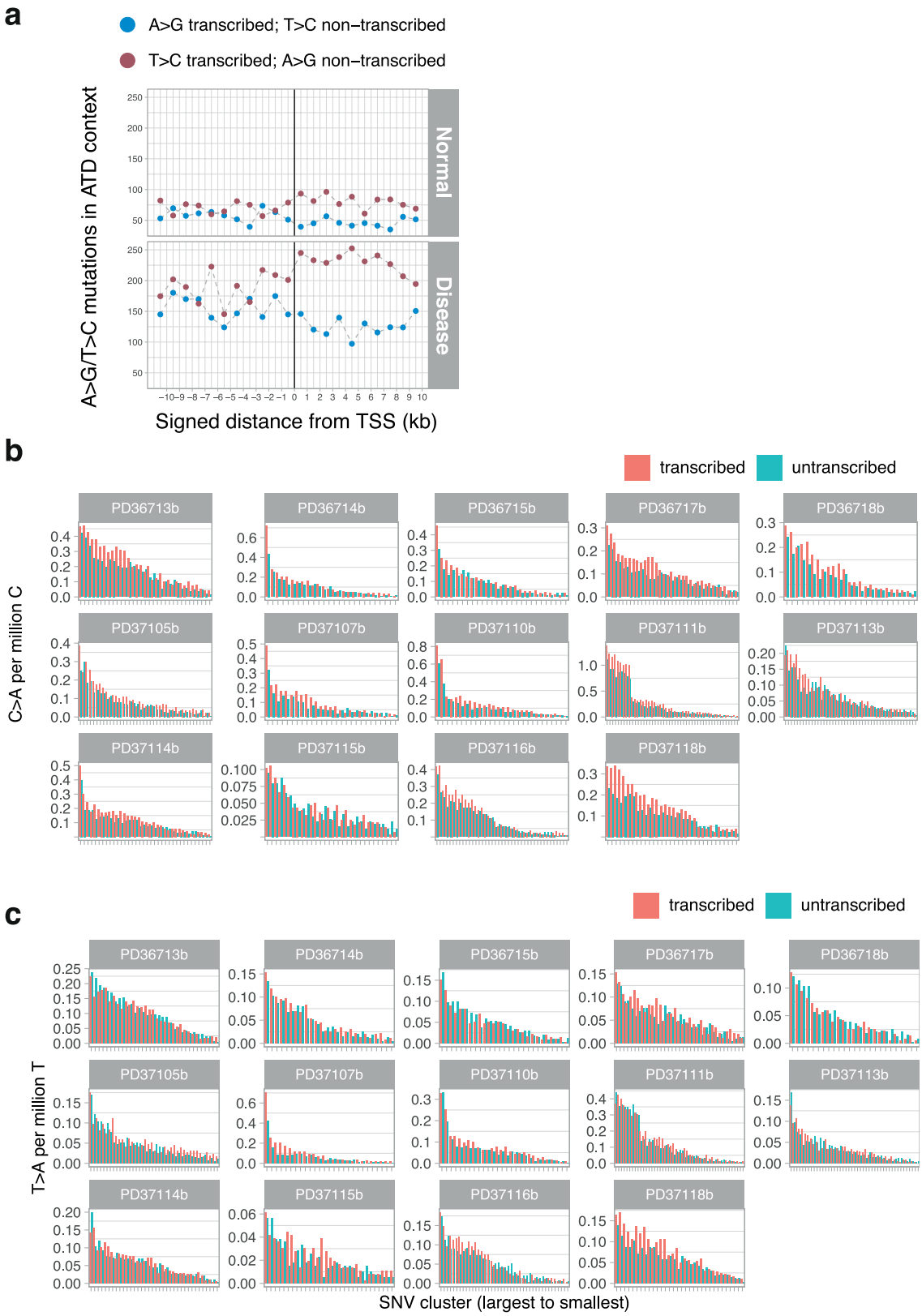
Extended Data Fig. 6 | Phylogenetic reconstruction of hepatocyte clones in non-alcoholic fatty liver disease with cirrhosis. Analogous to Extended Data Fig. 4, but for the cirrhotic livers of donors PD37113, PD37114, PD37115, PD37116 and PD37118. Right, H&E-stained sections (top); Masson's trichrome-stained sections (bottom left); and macroscopic photographs of the liver, with HCCs indicated by arrows (bottom right). Locations of immune or inflammatory cell infiltrates are marked with yellow rings. Sample sizes, PD37113, $n = 37$ microdissections; PD37114, $n = 41$ microdissections; PD37115, $n = 34$ microdissections; PD37116, $n = 43$ microdissections; PD37118, $n = 26$ microdissections.



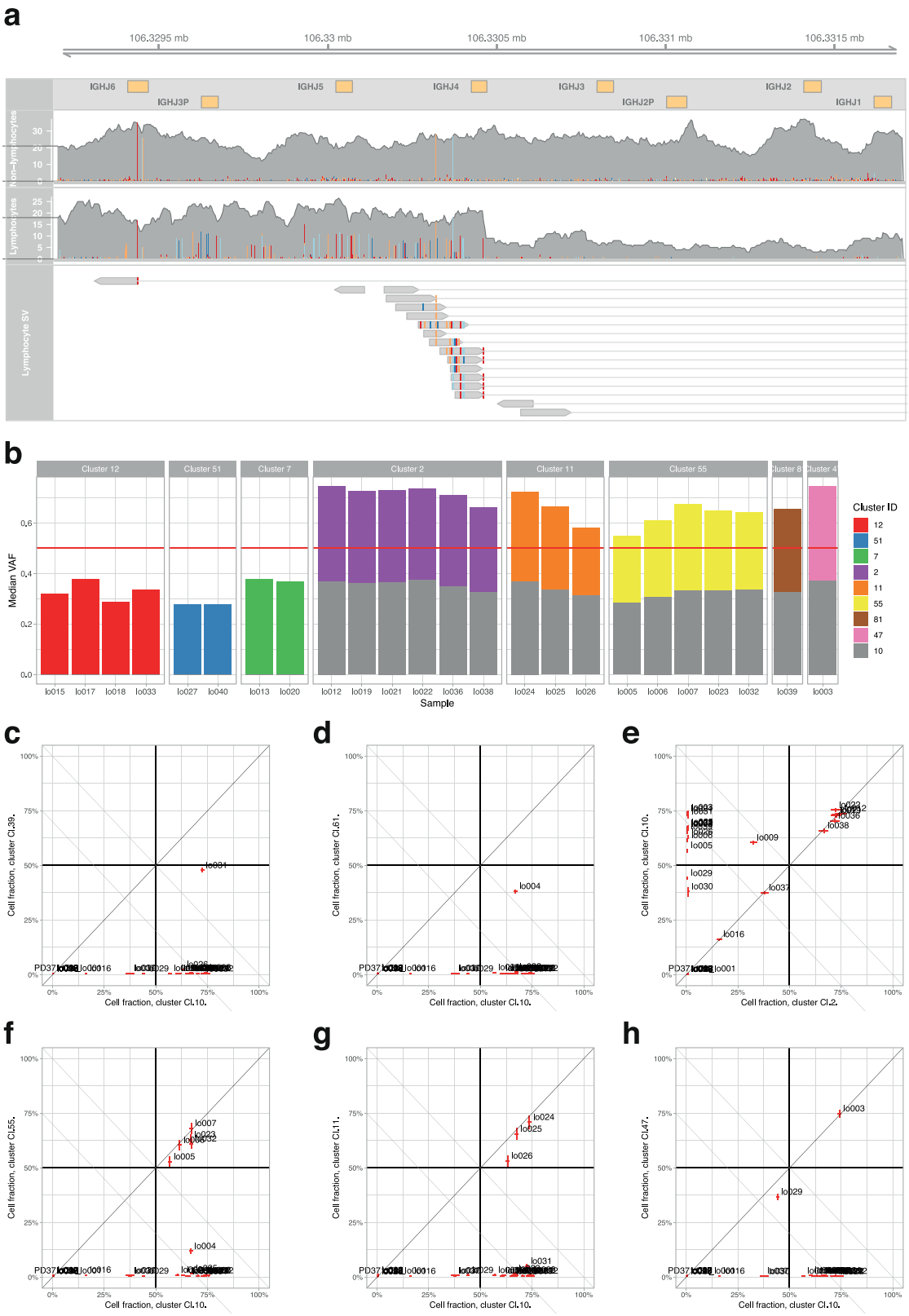
Extended Data Fig. 7 | Mutation spectra for individual microdissections. From each donor, we chose five clones to represent the heterogeneity in mutation spectra in the trinucleotide context. The six types of substitution are labelled across the top. Within each panel, the contributions from the trinucleotide context (bases immediately 5′ and 3′ of the mutated base) are shown.



Extended Data Fig. 8 | Details of the extraction of mutational signatures. **a**, Dot plots showing the concordance for signature attributions between the two signature algorithms ($n = 479$ microdissections). Mutational signatures on the *y* axis were extracted using non-negative matrix factorization and those on the *x* axis were extracted using a Bayesian hierarchical Dirichlet process. The *R* values are Pearson's correlation coefficients. **b**, Signatures extracted by non-negative matrix factorization. The six substitution classes are separated by grey vertical lines, and are presented in the following order: C>A, C>G, C>T, T>A, T>C, T>G. Within each class of mutation, the contributions from the trinucleotide context (bases immediately 5′ and 3′ of the mutated base) are shown. **c**, Signatures extracted by the Bayesian hierarchical Dirichlet process, as for **b**. Where a signature matches one from **b**, it is shown on the same row.



Extended Data Fig. 9 | Transcriptional-strand bias in patterns of mutations. **a**, Transcriptional-strand bias of T>C mutations at the ATD context before and after the transcription start site (TSS) of highly expressed liver genes. **b**, Bar plots representing the numbers of C>A variants on the transcribed and non-transcribed strands. Each hepatocyte clone is represented individually (*x* axis). Note the strand bias in the highly mutated clones of PD37111, in which the tobacco signature is most active; the strand bias indicates that the damaged base is the guanine, as expected for polycyclic aromatic hydrocarbons. **c**, Bar plots representing the numbers of T>A variants on the transcribed and non-transcribed strands. Each hepatocyte clone is represented individually (*x* axis). Note the strand bias in the highly mutated clones of PD37107, in which the aristolochic acid signature is most active; the strand bias indicates that the damaged base is the adenine.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Mutations in a B lymphocyte clone in a cirrhotic liver. **a**, Illustration of a portion of the B cell receptor (*IGH*) region on chromosome 14. Shown are the coverage tracks of an LCM sample that does not belong to the lymphocyte lineage (top) and a sample that belongs to the lymphocyte lineage (middle). In the centre of the displayed region there is a drop of copy number in the lymphocyte track, which indicates a structural rearrangement. The bottom track shows the paired-end reads that contribute to a rearrangement event in the lymphocyte sample, colocalized with the drop in copy number. **b**, Application of the pigeonhole principle: if two clusters of heterozygous mutations in regions of diploid copy number are in different cells, then their median VAFs must sum to ≤0.5 (if they sum to >0.5, equivalent to a combined cellular fraction of >1, then there must be some cells that carry both sets of mutations—hence one cluster would have a subclonal relationship with the other). Cluster 10 is the cluster with the unique VDJ rearrangement of *IGH* that is shown in **a** and the large number of mutations attributed to signature 9. Clearly, samples from clusters 2, 11, 55 and so on have VAFs which, when combined with cluster 10, sum to >0.5. Therefore, they must be subclonal to cluster 10, even though they do show signature 9. **c**–**h**, Representative pairwise decision graphs for clusters of mutations. The median cellular fraction is shown for pairs of clusters across every sample from the patient. Where at least one sample falls above or to the right of the $x + y = 1$ diagonal line, those two clusters must share a nested clonal–subclonal relationship.

nature research

Corresponding author(s): Peter Campbell

Reporting Summary

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

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Image processing from sequencing data using the proprietary Illumina X10 software that is maintained, installed and distributed by Illumnina with their X10 platform.
Data analysis	Single-nucleotide substitutions were called using the CaVEMan (cancer variants through expectation maximization) algorithm, version 1.11.2 (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm, version 2.2.2 (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm version 5.4.1 (https://github.com/cancerit/BRASS). Miscellaneous scripts for downstream analysis are available on Github (https://github.com/sfbrunner/liver-pub-repo). Mutational signatures analysis performed using the HDP hierarchical Dirichlet Process package version 0.1.5, available on Github (https://github.com/nicolaroberts/hdp).

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

Data

Policy information about availability of data

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

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

Sequence data that support the findings of this study have been deposited in the European Genome-Phenome Archive (https://www.ebi.ac.uk/ega/home) with accession number EGAD00001004578.

Field-specific reporting

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

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

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	No formal sample size calculation was performed. Sample size was chosen to give good representation of inter-patient and intra-patient variability in mutation burden, requiring a nested design - 4-5 patients for each of 3 aetiologies (normal, alcohol-related liver disease and non-alcoholic fatty liver disease), with 30-50 microdissections per patient to enable linear mixed models to estimate both within-patient and between-patient variance.
Data exclusions	No data exclusions
Replication	Replication for sequencing and mutation calling was achieved by microdissecting the same x,y regions from adjacent z sections separated by 20 micrometres. Microdissections collected in this was were independently processed, sequenced and variant called. This approach successfully replicated the mutation calls, with quantitative results from the replication described in the manuscript and displayed in Extended Data Figure 1.
Randomization	Not applicable - this is a descriptive study, not an intervention study.
Blinding	Not applicable - all dependent variables were computationally generated (mutation counts, signatures etc) and statistical analyses were pre-specified.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants

Methods

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

Human research participants

Policy information about studies involving human research participants

Population characteristics	We analysed liver biopsies from 5 normal livers and 9 patients with cirrhosis. Clinical characteristics of the cohort are described in Supplementary Table 1. Normal livers were broadly age-matched to patients with cirrhosis, encompassing middle aged to older individuals (43-77 years). In keeping with the demographic distribution of chronic liver disease, the majority of patients were male.

Recruitment	Recruited through the Tissue Bank at Addenbrooke's Hospital, Cambridge, UK. We explicitly studied samples from patients with chronic liver disease who had a synchronous hepatocellular carcinoma, meaning that patients were at advanced stages of disease. Otherwise, we anticipate no recruitment biases affecting the sample mix.