Article

Convergent somatic mutations in metabolism genes in chronic liver disease

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

Received: 17 June 2020

Accepted: 31 August 2021

Published online: 13 October 2021

Check for updates

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[6], Matthew Hoare[3,8]✉ & Peter J. Campbell[1,9]✉

The progression of chronic liver disease to hepatocellular carcinoma is caused by the acquisition of somatic mutations that affect 20–30 cancer genes[1–8]. Burdens of somatic mutations are higher and clonal expansions larger in chronic liver disease[9–13] than in normal liver[13–16], which enables positive selection to shape the genomic landscape[9–13]. Here we analysed somatic mutations from 1,590 genomes across 34 liver samples, including healthy controls, alcohol-related liver disease and non-alcoholic fatty liver disease. Seven of the 29 patients with liver disease had mutations in *FOXO1*, the major transcription factor in insulin signalling. These mutations affected a single hotspot within the gene, impairing the insulin-mediated nuclear export of FOXO1. Notably, six of the seven patients with $FOXO1^{S22W}$ hotspot mutations showed convergent evolution, with variants acquired independently by up to nine distinct hepatocyte clones per patient. *CIDEB*, which regulates lipid droplet metabolism in hepatocytes[17–19], and *GPAM*, which produces storage triacylglycerol from free fatty acids[20,21], also had a significant excess of mutations. We again observed frequent convergent evolution: up to fourteen independent clones per patient with *CIDEB* mutations and up to seven clones per patient with *GPAM* mutations. Mutations in metabolism genes were distributed across multiple anatomical segments of the liver, increased clone size and were seen in both alcohol-related liver disease and non-alcoholic fatty liver disease, but rarely in hepatocellular carcinoma. Master regulators of metabolic pathways are a frequent target of convergent somatic mutation in alcohol-related and non-alcoholic fatty liver disease.

The most common causes of chronic liver disease are chronic alcohol consumption, non-alcoholic fatty liver disease (NAFLD) and viral hepatitis. NAFLD and alcohol-related liver disease (ARLD) have an overlapping pathological spectrum, with fat accumulation in hepatocytes (fatty liver disease) being prominent in both. Chronic alcohol consumption[22] and caloric excess[23] disrupt lipid handling in the liver, with decreased fatty acid oxidation, increased lipogenesis and impaired triglyceride export resulting in the accumulation of both storage and toxic lipid species in hepatocytes[24,25].

Extended cohort of patients with NAFLD and ARLD

We previously sequenced 482 whole genomes from healthy and diseased liver[13], but lacked statistical power for definitive identification of genes under selective pressure. We extended this previous study with an additional 1,108 whole-genome sequences from 20 liver samples, focusing predominantly on NAFLD. We used a hierarchical experimental design: for each sample, comprising around 1 cm^3 of liver tissue, we sequenced 21–52 separate microdissections (Fig. 1a, Supplementary Note 1). In two patients with NAFLD, we took samples from all eight Couinaud anatomical segments of their explanted livers, and sequenced 22–28 microdissections from each segment.

When combined with the previous study, the expanded dataset comprised 1,590 genomes from 34 liver samples, including 5 healthy liver controls, 10 samples from patients with ARLD and 19 samples from patients with NAFLD (Supplementary Table 1). Overall, nine samples were from patients who had a synchronous hepatocellular carcinoma (HCC) and underlying cirrhosis; a further eight samples

[1]Cancer Genome Project, Wellcome Sanger Institute, Hinxton, UK. [2]Department of Surgery, Addenbrooke's Hospital, Cambridge, UK. [3]CRUK Cambridge Institute, Cambridge, UK. [4]Department of Pathology, Addenbrooke's Hospital, Cambridge, UK. [5]MRC Toxicology Unit, University of Cambridge, Cambridge, UK. [6]MRC Cancer Unit, University of Cambridge, Cambridge, UK. [7]Department of Hematology, Erasmus University Medical Center, Rotterdam, The Netherlands. [8]Department of Medicine, University of Cambridge, Addenbrooke's Hospital, Cambridge, UK. [9]Stem Cell Institute, University of Cambridge, Cambridge, UK. ✉e-mail: Matthew.Hoare@cruk.cam.ac.uk; pc8@sanger.ac.uk

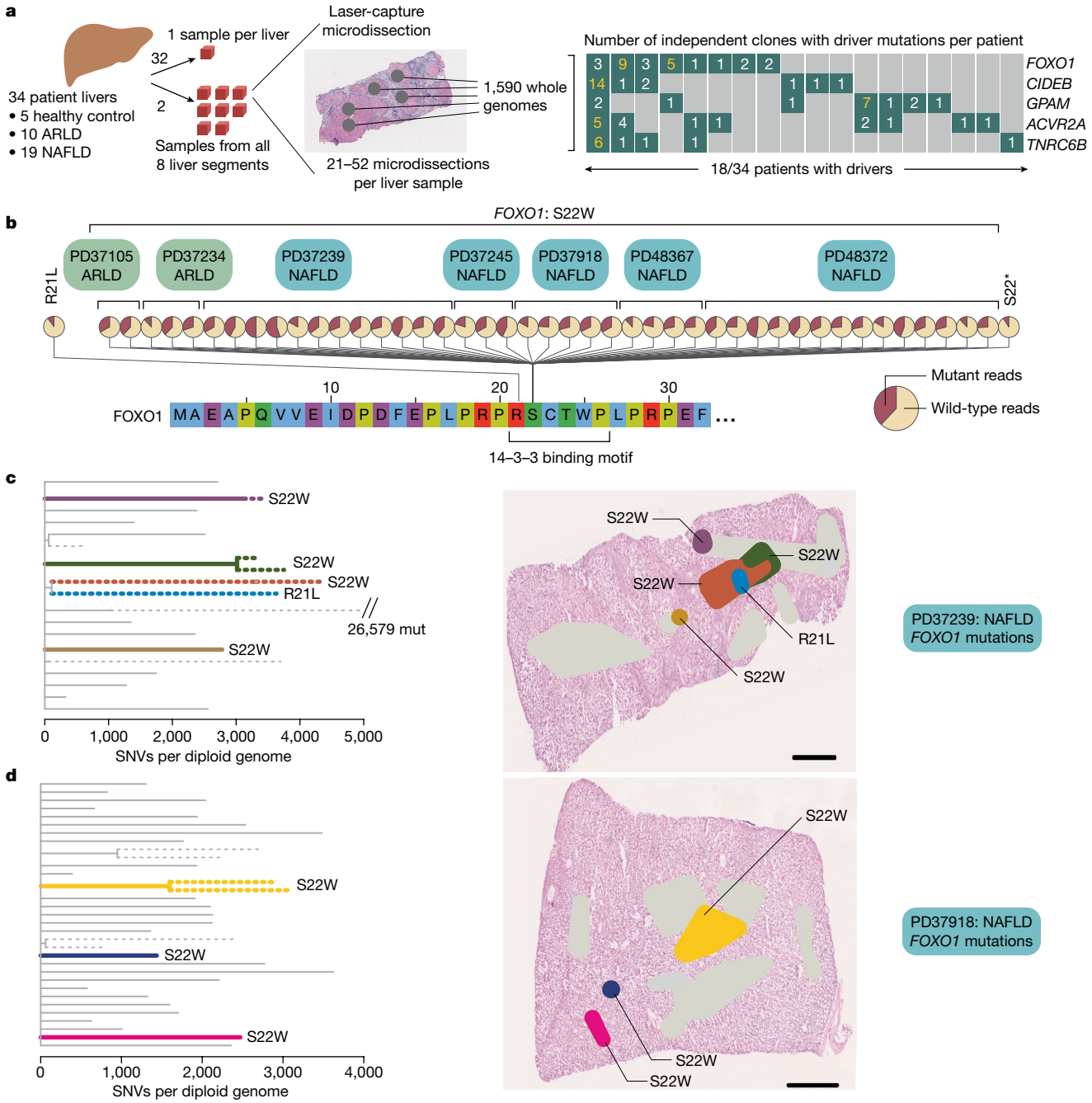


Fig. 1 | Convergent *FOXO1* mutations in chronic liver disease. a, Overview of the experimental design. **b**, Somatic mutations in *FOXO1* grouped by microdissections from affected patients. Pie charts show the fraction of sequencing reads reporting the mutant allele in each microdissection. **c**, **d**, *FOXO1* mutations in patients PD37239 (**c**) and PD37918 (**d**). SNVs, single-nucleotide variants; mut, mutations. Left, the phylogenetic tree, with coloured branches showing independently acquired *FOXO1* mutations. Right, clones from the phylogenetic tree mapped onto a haematoxylin and eosin (H&E)-stained light micrograph of the patient's liver biopsy. Scale bars, 1 mm.

had HCC without underlying cirrhosis. All samples were reviewed by a specialist hepatopathologist. Microdissections were sequenced to an average depth of 31× (Supplementary Table 2).

Driver mutations

Across all protein-coding genes, we identified six genes with a significant recurrence of mutations ($q < 0.05$) after correction for multiple hypothesis testing: *FOXO1* ($q < 2 \times 10^{-16}$), *CIDEB* ($q < 2 \times 10^{-16}$), *ACVR2A* ($q = 7 \times 10^{-9}$), *ALB* ($q = 8 \times 10^{-10}$), *GPAM* ($q = 1 \times 10^{-5}$) and *TNRC6B* ($q = 0.04$; Supplementary Tables 3, 4).

One of these genes, *ACVR2A*, a receptor for activin A in the TGF-β superfamily, is mutated in 5–10% of HCCs[1–6,8]. We observed thirteen missense mutations, two nonsense and one splice-site indel in *ACVR2A* ($q = 7 \times 10^{-9}$), as well as four large-scale structural variants (Extended Data Fig. 1, Supplementary Tables 4, 5).

Four genes identified as significant have not, to our knowledge, previously been reported in HCC, of which *FOXO1*, *CIDEB* and *GPAM* are discussed further below. *TNRC6B* encodes a protein involved in microRNA processing[26]. We observed three nonsense, two essential splice site and one large in-frame deletion as well as three missense mutations in *TNRC6B* ($q = 0.04$) (Extended Data Fig. 2a). This predominance of

protein-truncating variants suggests that inactivation of the gene confers a positive selective advantage on hepatocytes. Notably, one patient with NAFLD had five different mutations in *TNRC6B*, consistent with convergent evolution in independent hepatocyte clones.

We also screened for non-coding driver mutations[27] (Supplementary Table 4). A lncRNA, *NEAT1*, showed a significant excess of mutations compared to the background expectation ($q < 1 \times 10^{-10}$) (Extended Data Fig. 2b). This gene is recurrently mutated in a range of human cancers, including HCC[7], but this is believed to be due to a localized hypermutation process rather than positive selection[27].

FOXO1 hotspot mutations

We found a highly significant excess of missense mutations in *FOXO1* ($q < 2 \times 10^{-16}$), which encodes the major transcription factor in insulin signalling. Overall, we identified 26 clones that had acquired independent *FOXO1* mutations; these were distributed among 45 individual microdissections from 8 patients. Of these, 24 clones contained an identical base change that is predicted to generate an S22W amino acid substitution (Fig. 1b). The other two mutations would generate an R21L substitution and an S22* nonsense mutation. The latter was in a single microdissection from a healthy control liver sample, and we are uncertain of its biological significance—we only saw S22W mutations in patients with ARLD or NAFLD. A structural variant within the first intron of *FOXO1* was observed in a patient with NAFLD (Extended Data Fig. 3).

Of the seven patients with *FOXO1* S22W mutations, six had clear evidence for multiple independent acquisitions of the mutation in different clones; that is, convergent evolution (Extended Data Fig. 4, Supplementary Note 2). In the two patients in whom we sampled all eight anatomical segments of the liver, we found *FOXO1* S22W mutations in three different segments in one patient and in four segments in the other. Furthermore, even within a single segment, there were multiple, independently acquired *FOXO1* mutations, such that one of these two patients had nine independent clones with *FOXO1* S22W among regions sampled. Of the patients in whom we analysed samples from a single segment, one had five independent acquisitions of *FOXO1* missense mutations within around 1 cm^3; a further patient had three separate occurrences; and two patients had two separate acquisitions (Fig. 1c, d, Extended Data Figs. 4, 5).

Mutations impair the nuclear export of FOXO1

FOXO1 is the key transcription factor downstream of insulin signalling. In the fasting state, without insulin, FOXO1 is active in the nucleus of hepatocytes, and upregulates the expression of genes in the gluconeogenesis, glycogenolysis and lipolysis pathways. Upon activation of insulin signalling, activated AKT phosphorylates FOXO1, with threonine 24 (Thr24) being one of three known phosphorylation targets. Thr24 phosphorylation triggers 14-3-3 protein binding and the export of FOXO1 to the cytoplasm[28], where it undergoes ubiquitination and degradation. The Ser22 residue is itself phosphorylated by AMPK, inhibiting the export of FOXO1[29]. We hypothesized that substitution of a bulky tryptophan residue for Ser22 would similarly inhibit the nuclear export of FOXO1.

We transduced the HepG2 (Fig. 2a, b), Hep3B and PLC/PRF/5 (Extended Data Figs. 6, 7) HCC cell lines with retroviral constructs of *FOXO1* containing wild-type, R21L or S22W mutations, fused to C-terminal green fluorescent protein (GFP). Under serum starvation conditions, both wild-type and mutant FOXO1–GFP were predominantly localized to the nucleus, as expected without insulin. With the addition of insulin or serum, wild-type FOXO1–GFP underwent rapid nuclear export. However, even in the presence of insulin or serum, cells with mutant FOXO1–GFP maintained substantial levels of nuclear protein, with high nuclear-to-cytoplasmic ratios. An antibody to phosphorylated Thr24 in FOXO1 showed no binding to mutant constructs (Extended Data Fig. 6b).

We measured the levels of 105 metabolites in 5 independent replicates for HepG2 cells with wild-type FOXO1–GFP or FOXO1(S22W)–GFP, with and without insulin (Fig. 2c). Overall, 43 metabolites were significantly different between S22W and wild-type constructs, with many intermediates in glycolysis, gluconeogenesis and pentose phosphate pathways exhibiting increased levels in cells with mutant FOXO1–GFP. RNA sequencing of transduced HepG2 cell lines revealed significant upregulation of gene sets that are involved in the cell cycle ($q < 0.0001$), lipid catabolism ($q < 0.0001$) and FOXO-mediated transcription targets ($q = 0.008$); and downregulation of gene sets associated with pro-apoptotic processes ($q = 0.0004$) and canonical glycolysis ($q < 0.0001$) (Extended Data Fig. 8, Supplementary Tables 6, 7).

Mutations in *CIDEB*

We observed a significant excess of somatic mutations in *CIDEB* ($q < 2 \times 10^{-16}$). *CIDEB* is the major CIDE-family member that is active in hepatocytes, and it regulates the fusion of intracellular lipid droplets, mediated by the formation of homodimers between CIDEB proteins[17,18]. Homodimerization occurs through electrostatic contacts between positively charged residues on the CIDEB protein from one lipid droplet and negatively charged residues on the other.

In addition to 2 nonsense and 1 stop-loss mutation, we observed 18 missense mutations in *CIDEB* (Fig. 3a). Missense mutations were predominantly located in the two domains implicated in homodimerization of CIDE proteins, and many of them either switched a charged residue for a neutral one or reversed the charge. Previous in vitro mutagenesis studies have shown that altering the charge on key conserved residues, including some of those mutated in our patients, abrogates homodimerization, preventing the fusion and growth of lipid droplets within the cell[17,18].

As for *FOXO1*, mutations in *CIDEB* were frequently acquired in multiple independent clones within the liver of one patient. For example, in one patient with NAFLD in whom we sampled all 8 Couinaud segments, we found 14 clones with non-synonymous mutations in *CIDEB*, distributed over 6 of the 8 segments (Fig. 3b, Extended Data Fig. 9).

GPAM mutations

Another significantly mutated gene was *GPAM*, which encodes mitochondrial glycerol-3-phosphate acyltransferase. This enzyme catalyses the rate-limiting step in triacylglycerol synthesis—namely, the esterification of long chain acyl-CoAs with glycerol-3-phosphate[20,21].

We observed 12 missense and 3 protein-truncating mutations in *GPAM* ($q = 1 \times 10^{-5}$), affecting 7 patients (Fig. 3c). We also observed a tandem duplication 20 kb upstream of the gene in one microdissection (Extended Data Fig. 3b). The clustering of missense mutations in the acyltransferase domain, coupled with the nonsense and frameshift mutations, suggests that the likely consequence of these mutations is impairment of protein function. As we saw for *FOXO1* and *CIDEB*, there was evidence for convergent acquisition of somatic mutations in *GPAM* in different clones from the liver sample of the same patient. For example, one patient had seven separate events affecting *GPAM* (Fig. 3d), and another patient had two separate events (Extended Data Fig. 10).

Properties of clones and patients with drivers

For the two patients in whom we sampled all eight Couinaud segments of the liver, we found that driver mutations were replicated across multiple regions, suggesting that the findings from a single sample are broadly representative of the whole liver. We therefore extrapolated the total hepatic mass carrying driver mutations for each significant gene (Extended Data Fig. 11a). This revealed, first, that clones with driver mutations accounted for hundreds of grams of liver mass in

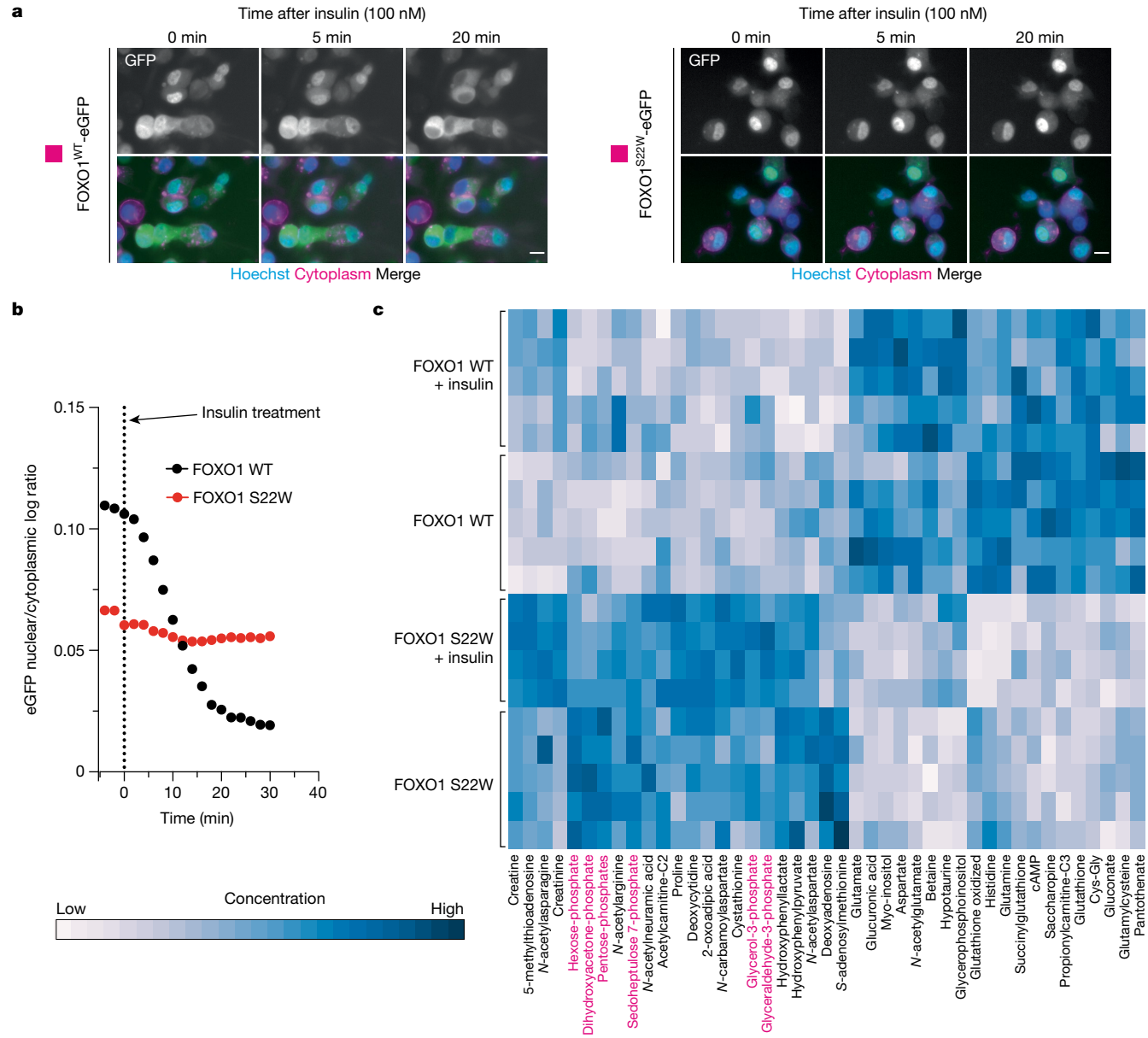


Fig. 2 | Somatic mutations of *FOXO1* lead to impaired nuclear export and insulin resistance in vitro. a, HepG2 cells were transfected with wild-type (WT) or mutant constructs of FOXO1 fused with C-terminal eGFP. Cells were counterstained with nuclear (Hoechst 33342) and cytoplasmic (SPY-555-Actin) markers. Scale bars, 10 µm. **b**, Quantification of eGFP localization, expressed as log nuclear/cytoplasmic fluorescence ratio (mean ± s.e.m.) during live-cell imaging (wild-type FOXO1 cells, $n = 6{,}186$ cells per time point; and $FOXO1^{S22W}$ cells, $n = 7{,}172$ cells per time point). The s.e.m. values were very low, and hence error bars are not easily visible. **c**, Heat map showing concentrations of metabolites (rows) measured in HepG2 cells (expressing wild-type or S22W *FOXO1* construct; with or without insulin) in 5 replicates each (columns). Metabolites that were significant after correction for multiple hypothesis testing ($q < 0.01$) are shown, with intermediates from pentose phosphate and glycolysis–gluconeogenesis pathways in pink.

some patients; and, second, that the distribution of driver mutations showed considerable patient-to-patient variation in which genes were affected and what level of involvement was observed. Notably, clones carrying mutations in *FOXO1* ($P = 0.005$), *CIDEB* ($P = 0.001$) and *ACVR2A* ($P = 0.001$) were larger on average than wild-type clones (Wilcoxon test) (Extended Data Fig. 11b), suggesting that the selective advantage conferred by the mutations enables preferential expansion.

Despite the moderate cohort size, mutations in *FOXO1*, *CIDEB* and *GPAM* were seen across a wide range of patient characteristics: both sexes; broad age span; with and without type 2 diabetes; and variable severity of histological abnormality (Extended Data Fig. 11c–e, Supplementary Code). This suggests that the results from our cohort will generalize across patients with ARLD and patients with NAFLD.

Comparison with HCC

We accessed mutation calls for 1,670 HCCs recorded in the International Cancer Genome Consortium (ICGC) data portal. The three metabolism genes so frequently mutated in our cohort, *FOXO1*, *CIDEB* and *GPAM*, were not significantly mutated in HCC ($q = 1.0$, 1.0 and 0.6, respectively) (Supplementary Tables 7–9). *FOXO1* S22W mutations were found in only 3 of 1,670 HCCs (0.18%; 95% confidence interval = 0.05–0.6%)—a significantly lower fraction than the 24 clones carrying this mutation in our cohort ($P = 2 \times 10^{-5}$, Fisher's exact test). *TNRC6B* exhibited a significant excess of protein-truncating mutations in HCC (24 variants; $q = 0.0001$), suggesting that it is a tumour suppressor gene in malignant hepatocytes.

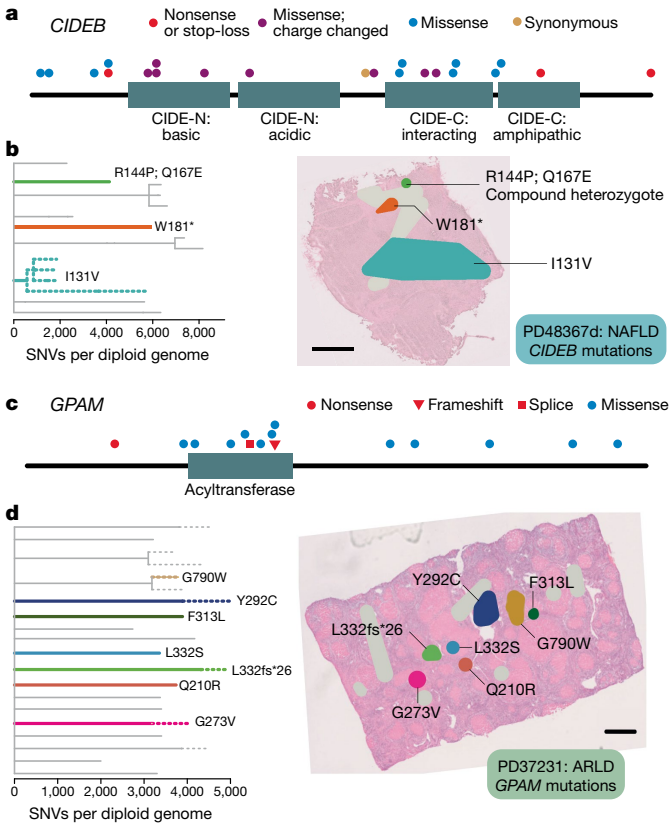


Fig. 3 | Convergent *CIDEB* and *GPAM* mutations in chronic liver disease. a, Distribution of somatic mutations in *CIDEB*. **b**, *CIDEB* mutations in a patient with NAFLD. Left, the phylogenetic tree, with coloured branches showing independently acquired mutations. Right, clones from the phylogenetic tree mapped onto an H&E-stained photomicrograph of the liver. **c**, Distribution of somatic mutations in *GPAM*. **d**, *GPAM* mutations in a patient with ARLD; layout as for **b**. Scale bars, 1 mm (**b**, **d**).

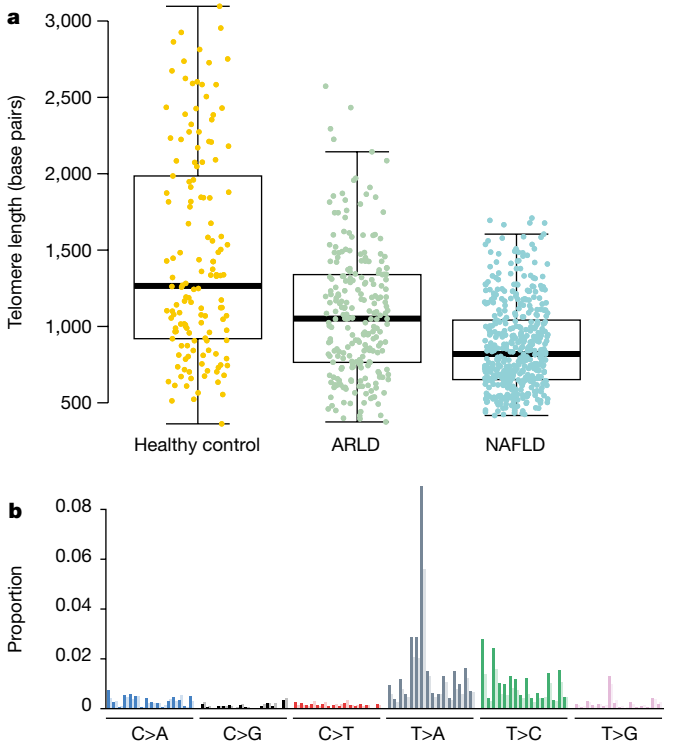


Fig. 4 | Other genomic analyses. a, Distribution of telomere lengths (*y* axis) by disease status (*x* axis). Each point represents the average telomere length estimated from genome sequencing data for a microdissection ($n = 1,202$). Box-and-whisker plots show the median marked with a heavy black line and interquartile range in a thin black box. Whiskers denote the range or 25th and 75th centile plus 1.5× the interquartile range. **b**, Trinucleotide context spectrum on transcribed and untranscribed strands of a new single-base-substitution signature. The trinucleotide context is shown as four sets of eight bars, grouped by whether an A, C, G or T respectively is 5′ to the mutated base, and within each group of eight by whether A, C, G or T is 3′ to the mutated base. The activity of the mutational signature on the untranscribed strand is shown in a pale colour; transcribed strand in a darker colour.

Other genomic analyses

The majority of HCCs have mutations that activate *TERT*, the telomerase gene[2,4,7], but we observed only one mutation affecting the *TERT* promoter in non-cancerous liver. To assess telomere dynamics in our samples, we estimated telomere lengths for each microdissection. We observed considerable between- and within-individual variation in telomere lengths across the cohort, with shorter telomeres in NAFLD and ARLD compared to normal liver (Fig. 4a, Extended Data Fig. 12, Supplementary Code). This suggests that ARLD and NAFLD are associated with a substantial attrition of telomeres, outweighing the relatively minor shortening of telomere lengths with age. Furthermore, telomeres became progressively shorter as the size of a clone increased, reflecting the extra cell divisions associated with hepatocyte regeneration.

We also evaluated mutational signatures in the extended cohort, extracting a signature not previously seen in HCC[2,3]. This signature was characterized by T>A mutations, especially in a CTG context, with transcriptional strand bias suggesting that adenine is the damaged base (Fig. 4b, Extended Data Figs. 13, 14). As we found for other exogenous signatures[13], this new signature showed considerable variability in activity between nearby clones within the same liver sample, accounting for less than 5% of mutations in some nodules, but up to 50% in others, especially on terminal branches of the phylogeny (Extended Data Fig. 14).

Discussion

We hypothesize that the major mechanism that underlies the selective benefit of somatic mutations in *FOXO1*, *CIDEB* and *GPAM*, three master regulators of lipid processing and storage, is that these mutations protect hepatocytes from the lipotoxicity that is common to both NAFLD and ARLD[22–25]. FOXO1 is the critical transcription factor downstream of insulin signalling. We have shown in vitro that the hotspot mutations impair the insulin-mediated nuclear export of FOXO1, which results in insulin resistance, the upregulation of lipid catabolism genes and impaired glucose metabolism. CIDEB regulates the fusion of intracellular lipid droplets in hepatocytes[17,18] and GPAM is the rate-limiting enzyme in the conversion of free fatty acids to storage triglycerides[20,21]. The mutations that we observed in these genes suggest loss-of-function effects, meaning that knockout mouse models would mimic the mutations we observe. Knockout mice for *Cideb* and *Gpam* have a lower hepatic triglyceride content than wild-type controls, and these differences were considerably more pronounced with high-fat diets[19,20], with knockout mice specifically protected against diet-induced steatohepatitis[19]. These phenotypes provide in vivo experimental evidence for our hypothesis that somatic mutations in these genes protect hepatocytes from lipotoxicity.

A major theme emerging from our study is the contrast of within-patient convergence with between-patient divergence. Within the liver of one patient, we observed many independent hepatocyte clones preferentially expanding with mutations in the same metabolism gene—such convergent evolution points to highly specific selective pressures operative in the liver of a given patient. Across different patients, however, there was considerable heterogeneity in both the

frequency of driver mutations and which genes they affected. Further studies in larger clinical cohorts will be required to understand whether this patient-to-patient heterogeneity results from different subtypes of disease; whether it informs on disease severity; and whether it predicts future risk of cancer or liver failure.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-021-03974-6.

1. The 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. 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).
7. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
8. Pinyol, R. et al. Molecular characterization of hepatocellular carcinoma in patients with non-alcoholic steatohepatitis. *J. Hepatol.* **75**, 865–878 (2021).
9. 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).
10. Torrecilla, S. et al. Trunk mutational events present minimal intra- and inter-tumoral heterogeneity in hepatocellular carcinoma. *J. Hepatol.* **67**, 1222–1231 (2017).
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.* **54**, 628–640 (2019).
13. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
14. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
15. Yizhak, K. et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).
16. 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**, eaax2659 (2020).
17. Barneda, D. et al. The brown adipocyte protein CIDEA promotes lipid droplet fusion via a phosphatidic acid-binding amphipathic helix. *Elife* **4**, e07485 (2015).
18. Sun, Z. et al. Perilipin1 promotes unilocular lipid droplet formation through the activation of Fsp27 in adipocytes. *Nat. Commun.* **4**, 1594 (2013).
19. Li, J. Z. et al. Cideb regulates diet-induced obesity, liver steatosis, and insulin sensitivity by controlling lipogenesis and fatty acid oxidation. *Diabetes* **56**, 2523–2532 (2007).
20. Hammond, L. E. et al. Mitochondrial glycerol-3-phosphate acyltransferase-1 is essential in liver for the metabolism of excess acyl-CoAs. *J. Biol. Chem.* **280**, 25629–25636 (2005).
21. Wendel, A. A., Cooper, D. E., Ilkayeva, O. R., Muoio, D. M. & Coleman, R. A. Glycerol-3-phosphate acyltransferase (GPAT)–1, but not GPAT4, incorporates newly synthesized fatty acids into triacylglycerol and diminishes fatty acid oxidation. *J. Biol. Chem.* **288**, 27299–27306 (2013).
22. Jeon, S. & Carr, R. Alcohol effects on hepatic lipid metabolism. *J. Lipid Res.* **61**, 470–479 (2020).
23. Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).
24. Clugston, R. D. et al. Altered hepatic lipid metabolism in C57BL/6 mice fed alcohol: a targeted lipidomic and gene expression study. *J. Lipid Res.* **52**, 2021–2031 (2011).
25. Puri, P. et al. A lipidomic analysis of nonalcoholic fatty liver disease. *Hepatology* **46**, 1081–1090 (2007).
26. Meister, G. et al. Identification of novel argonaute-associated proteins. *Curr. Biol.* **15**, 2149–2155 (2005).
27. Rheinbay, E. et al. Analyses of non-coding somatic drivers in 2,658 cancer whole genomes. *Nature* **578**, 102–111 (2020).
28. Yaffe, M. B. et al. The structural basis for 14-3-3:phosphopeptide binding specificity. *Cell* **91**, 961–971 (1997).
29. Saline, M. et al. AMPK and AKT protein kinases hierarchically phosphorylate the N-terminus of the FOXO1 transcription factor, modulating interactions with 14-3-3 proteins. *J. Biol. Chem.* **294**, 13106–13116 (2019).

Methods

Samples

All biological samples were collected with informed consent from Addenbrooke's Hospital, Cambridge, UK, according to procedures approved by the East of England Local Research Ethics Committee (16/NI/0196 and 15/EE/0351). All participants consented to publication of research results. The samples were snap-frozen in liquid nitrogen and stored at −80 °C in the Human Research Tissue Bank and the Department of Surgery of the Cambridge University Hospitals NHS Foundation Trust.

Healthy control liver tissues were obtained from patients undergoing hepatic resection of colorectal carcinoma metastases; specimens were obtained distant to the metastases and confirmed free of tumour at histopathological examination; one patient (PD36718) had undergone pre-operative portal vein embolization to the ipsilateral liver, but none had received neoadjuvant chemotherapy before resection. Background diseased liver tissue was obtained from individuals with NAFLD or ARLD, undergoing either hepatic resection for HCC or liver transplantation for HCC or liver failure (Supplementary Table 1). All patients with NAFLD had typical risk factors for the metabolic syndrome and pathological changes compatible with this aetiology: either the presence of steatohepatitis, or steatosis and the pattern of fibrosis; in addition, before transplantation or resection they had extensive clinical investigations excluding other diseases. None had undergone pre-operative locoregional therapy, except PD37118, who underwent a single treatment with trans-arterial chemoembolization. Clinical details, anthropometrics and blood results are detailed in Supplementary Table 1. The *PNPLA3* (rs738409) polymorphism genotype was derived from the whole-genome sequencing (WGS) data.

One patient had three separate samples taken over a five-year timespan, each analysed in this study (PD37918b, PD37915b and PD37910b; Supplementary Fig. 1). Two other patients, who were undergoing liver transplantation for their disease, had samples taken from all eight Couinaud segments of the liver (PD48367a–h and PD48372a–h).

The explant liver histology was reviewed by two specialist liver histopathologists (S.E.D. and S.J.A.), blinded to the other results in the study. All liver specimens were scored according to the Kleiner system on formalin-fixed paraffin-embedded (FFPE) samples away from the fresh-frozen block used for the laser-capture microdissection (LCM). The Kleiner score, developed for NAFLD, assesses the presence of steatosis, lobular inflammation and hepatocyte ballooning to generate a cumulative NAFLD activity score (NAS); artefactual inflammation secondary to surgical handling was excluded. We applied this to the healthy control and ARLD samples, which, in the absence of a validated scoring system for ARLD, allows comparability between all study samples regardless of disease aetiology. Fibrosis was assessed using both the Kleiner[30] and the Ishak[31] scoring systems. 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).

Sample preparation

The protocols for preparing liver tissue sections, LCM and subsequent cell lysis, DNA extraction, and WGS were previously described[13,32]. In brief, 20-µm-thick tissue sections (prepared with a Leica cryotome) were fixed with 70% ethanol and stained with haematoxylin and eosin for subsequent LCM generation using a Leica Microsystems LMD 7000. The micro-dissected samples were then 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 for submission to paired-end WGS. The resultant reads were mapped to the GRCh37d5 human reference genome using the BWA-MEM algorithm.

For RNA sequencing, commercially available human hepatocellular carcinoma HepG2 cell lines (RRID:CVCL_0027) were transduced with: (1) $FOXO1^{R21L}$-eGFP, (2) $FOXO1^{WT}$-eGFP, and (3) $FOXO1^{S22W}$-eGFP. For each of these three lines there were starved and insulin-stimulated conditions for a total of six conditions. For each of these six conditions there were five biological replicates (A to E) for a total of 30 samples. RNA was extracted from 30 HepG2 samples using the Qiagen RNeasy plus kit according to the manufacturer's instructions, and sequenced on the Illumina NovaSeq platform.

SNV calling

Several steps of the SNV calling workflow that was used in this study were previously described. Basic SNV identification used the Cancer Variants through Expectation Maximization (CaVEMan) algorithm[33] to call single-base substitution (SBS) variants, with per-patient bulk biopsies as matched healthy controls. For the two patients with NAFLD that were biopsied from the eight anatomical liver segments, a thyroid follicle LCM sample was used as an unmatched control to call mutations. Duplicate reads and LCM library preparation-specific artefactual variants resulting from the incorrect processing of secondary cruciform DNA structures were removed with bespoke post-processing filtering. The latter filtering step was configured to consider all variants with at least two supporting sequenced DNA fragments. In the current study, the entropy-metric-based variant-filtering step described in our earlier liver paper was replaced with a beta-binomial-based filtering approach as described elsewhere[34], which operates on the principle that authentic somatic mutations are typically over-dispersed (that is, present in only a limited number of genomes in the set of genomes belonging to each patient), whereas systematic artefacts or germline variants are commonly under-dispersed, making them observable across many, if not all, genomes derived from the microdissections from the same patient biopsy. In this study, the number of mutation-bearing and total reads for each SNV was calculated by enumerating raw allele counts for each base (A, C, G, T) per SNV called across all microdissections on a patient-specific basis, in which mutations with a dispersion estimate of ≥ 0.1 were considered to be likely to be true somatic variants. Manual inspection of a subset of final SNV calls using a genome browser was performed to ensure validity. A further sanity check involved checking that spatially proximate microdissections as captured by histology images shared common mutations (that is, within the same vicinity in terms of x–y space on the same tissue section, within the same cirrhotic nodule, or overlapping x–y positions on tissue sections from different z-planes).

Using the hierarchical Dirichlet process for the identification of SNV clusters

The nonparametric Bayesian hierarchical Dirichlet process (HDP) was implemented to cluster SNVs with similar variant allele fractions (VAFs) that were called across multiple microdissections for each patient biopsy. Full mathematical and implementation details of the clustering algorithm are described in a previous publication[13]. This N-dimensional Dirichlet process (NDP) clustering approach was run with 10,000 burn-in iterations, followed by 5,000 posterior Gibbs sampling iterations that were used for clustering. This class of algorithm was chosen for the identification of SNV clusters as there is no requirement to arbitrarily prespecify the number of clusters to find. Instead, at each sampling iteration, there is a defined probability that mutations will be allocated to new clusters that did not exist in the previous iteration. On the other hand, clusters can also be removed in a future iteration in cases in which all member mutations are assigned to other clusters. In this way, the number of SNV clusters are permitted to vary throughout the sampling chain. To avoid overly complex solutions consisting of a large number of clusters, which would increase the chance of creating uninformative ones, an upper limit of 100 SNV clusters per patient was imposed in this study. A multi-threaded version of the ECR algorithm

modified from the label.switching R package[35] was used for rapid label switching correction. Only SNV clusters comprising a minimum of 50 unique mutations were kept for downstream analysis. Input to this algorithm included per-patient data tables consisting of the coverage and counts of each called variant per microdissection. Further details are available in the Supplementary Methods.

Inference of phylogenetic trees

The statistical pigeonhole principle, as described previously[36], was applied to infer phylogenetic clonal relationships between per-patient SNV clusters identified by the NDP algorithm, in which each cluster is represented as a branch of a phylogenetic tree. A given cluster is considered to have strong evidence of being nested within another (that is, sub-clonal relationship) if the fraction of cells carrying the cluster of mutations is lower in all member microdissections relative to the fraction of cells containing another cluster of mutations within the same microdissections, in which the sum of their respective mutant cell fractions (CFs) is also >100%. Otherwise, if the sum of the pairwise mutant CFs is ≤100%, only weak evidence of nesting exists. In cases in which only some microdissections have lower CFs of a given SNV cluster relative to another, the clusters are interpreted to be independent and not nested within one another. In the current study, only clusters with a mutant CF > 0.05 are analysed, and the CF of each SNV cluster is calculated as 2 times the median VAF for each microdissection, which assumes diploidy.

Problematic SNV clusters containing microdissections that do not share any mutations with other member dissections of the same clusters were split up into new independent clusters, and were individually reassessed for phylogenetic relatedness with all other clusters within the same patient biopsy using the pigeonhole principle.

Identification of mutations under selection

To determine whether any coding mutations were under selection in non-tumorous chronic liver disease tissues, the dN/dScv method[37] on the gene, protein domain, codon and hotspot levels was used to identify genes with a higher number of nonsynonymous mutations relative to the expected number from the rate of synonymous mutation acquisition. For this analysis, the unique mutations across the set of all SNV clusters were used as input, while mutations with *q*-values corrected for multiple hypothesis testing of < 0.05 were considered to be under selection. For the identification of non-coding mutations that may be under selection, the NBR algorithm was used[27].

Indel calling

As previously described[13], indel calling was performed using cgpPindel[38]. A naive Bayes algorithm was used to assign each called indel to the SNV clusters identified using the NDP algorithm. As done during SNV calling, the beta-binomial over-dispersion filter was applied to the raw counts of each called indel across the set of microdissections made from each patient biopsy to further filter out artefacts, in which variants with an over-dispersion value of ≥ 0.1 and VAF ≥ 0.025 were considered to probably be real.

We then developed a gradient-boosted regression tree model to accurately separate true from artefactual indels using separate microdissections that were phylogenetically closely related. Full details of our model for calling indels are presented in the Supplementary Methods and Supplementary Fig. 2.

Structural variant calling

Structural variants including deletions, inversions, tandem duplications and translocations affecting large genomic segments were called using the BRASS (breakpoint via assembly) algorithm[39] (https://github.com/cancerit/BRASS). A three-step process was next used to filter out likely artefactual structural variants called by BRASS. First, a custom pipeline was developed that identifies and removes artefactual variants that were introduced by the LCM library preparation protocol, based on comparing the structural variant events detected in each microdissection with those present in a panel of corresponding normal bulk control samples. Second, detailed manual review of all remaining structural variants was conducted using a genome browser and variant annotations. Finally, similar to the sanity checks that SNVs and indels were subjected to, the presence of each structural variant was checked among proximate microdissections where possible, where it is expected that real variants would be shared by such clusters of dissections.

More complex genomic rearrangement events such as chromothripsis[40], in which one or more chromosomes are shattered into as many as thousands of pieces that are subsequently fused back together in a disordered fashion, were additionally identified through detailed manual review of the final set of structural variants.

Calculation of clone areas

The exact spatial positions of 1,202 microdissections were captured in a series of microscopy images taken with the LMD7 laser microdissection in-built camera. The cartesian coordinates of the outer edge of each dissection was extracted using the Canny method as implemented in the edge function from the Image Processing MATLAB toolbox. This resulted in a set of x–y coordinates per microdissection, which were manually annotated to correspond to their respective WGS profiles. Next, the SNV clusters were processed individually starting with the identification of their respective member microdissections that bear mutations assigned to each cluster. For each SNV cluster, the mutant CF (that is, 2 × median VAF) was used to adjust each member dissection's x–y coordinates on the tissue section image to more accurately reflect genetic clone size. A minimal ellipsoid convex hull was subsequently drawn to encompass the adjusted spatial coordinates of each member microdissection of a given SNV cluster, before merging the resultant polygons into a single entity representing the corresponding clone area. Clone area was initially computed in terms of squared pixels, before a pixel to micrometre conversion was applied to translate the units to squared micrometres. For this, multiplicative conversion factors were calculated by first generating images of scale indicators overlaid atop high-resolution scanned histology images of tissue sections. This was done using the NDP.view2 NanoZoomer Digital Pathology slide scanner image viewing software from Hamamatsu Photonics. The scale indicator images were then loaded into the R statistical programming environment using the magick image processing package to determine the exact number of pixels per millimetre for each tissue section image. In this study, only microdissections that contributed mutations with VAF ≥ 0.05 were included in the clone size calculation.

Clone size comparison

The clone areas (μm^2) were compared between hepatocytes that carried driver mutations found in this study to those that did not. Specifically, for each driver, clones wild-type for the driver were uniformly and randomly sampled from each donor bearing the mutation so that the clone areas (weighted by the clone's number of mutations) between the number of mutated clones from each donor could be compared to an equivalent number of wild-type clones that were randomly selected from each corresponding donor. The comparison of clone areas was conducted using the ggstatsplot R package, in which the Bonferroni method of multiple hypothesis testing correction to P values was applied, and the default Mann–Whitney U test for nonparametric pairwise comparisons was used.

Estimation of liver mass containing driver mutations

Several assumptions were made in the calculation of the grams of hepatocytes carrying each of the driver mutations identified in this study: (1) the majority cell type composing samples are hepatocytes, in which driver mutations occurred in diploid genomic regions, and thus mutant hepatocyte fraction = 2 × driver allele frequency; (2) each LCM microdissection was estimated to comprise 100 to 500 hepatocytes;

(3) there are 1.16×10^8 hepatocytes per gram of a typical 1.5-kg human liver[41,42]. For each donor in our study, the liver-wide mass (grams) of mutated hepatocytes was inferred for each driver mutation by first calculating the area (pixels) of all sequenced LCM dissections using histology images. As it was estimated that each dissection contained between 100 to 500 hepatocytes, a linear fit was performed using the R linMap function to map all LCM cut areas within this range, effectively estimating the number of hepatocytes composing each LCM cut. The VAF of each driver mutation was then used to infer the fraction of mutant-bearing cells in each LCM dissection. Next, the proportion of sequenced material per donor containing each driver was calculated by summing estimates from all donor-specific sequenced LCM cuts. These donor-level estimates were then used to approximate the proportion of liver cells carrying each driver on the basis of the estimated number of hepatocytes in a typical human liver. These values were then ultimately used to estimate the number of grams of liver that contained each driver for each donor, assuming that a typical human liver weighs 1.5 kg.

Using HDP for the extraction of mutational signatures

The HDP algorithm as implemented in the HDP R package (https://github.com/nicolaroberts/hdp), was used to extract mutational signatures composing the set of SBSs called in each of the 1,013 SNV clusters identified in healthy control liver and chronic liver disease samples. Input to the algorithm consisted of a matrix of mutation counts per SNV cluster for each of the mutation categories, which in this case consisted of 192 trinucleotide mutational contexts (generated using the SigProfilerMatrixGenerator software[43]) as defined by the six SBS types (C>A, C>G, C>T, T>A, T>C, T>G), with each further defined by all possible combinations of bases (A, C, T, G) flanking the mutated base (3′ and 5′), for the transcribed and un-transcribed strands. A reference catalogue of 65 previously identified 192-context-based mutational signatures from the PanCancer Analysis of Whole Genomes (PCAWG) study was used as prior information[44]. Signatures that had been previously observed in hepatocellular carcinoma (HCC) samples (SBS1, SBS3, SBS4, SBS5, SBS6, SBS9, SBS12, SBS14, SBS16, SBS17a, SBS17b, SBS18, SBS19, SBS22, SBS23, SBS24, SBS26, SBS28, SBS29, SBS30, SBS31, SBS35, SBS37 and SBS40) were assigned the default weighting of 1,000 pseudocounts during analysis to facilitate the extraction of known liver-relevant signatures. The remaining prior signatures were assigned a lower weighting of 100 so as to not rule them out completely in the analysis. By design, HDP allows for a degree of de novo discovery of novel mutational signatures that are dissimilar to the set of known signatures supplied as prior information. To further guide the extraction of liver-related mutational signatures, 314 HCC WGS profiles were also included in the analysis. A burn-in of 50,000 iterations was used, followed by 200 posterior Gibbs sampling iterations that were performed 100 iterations apart, while adjusting the concentration parameter (with shape and rate hyperparameters set to 1 and 20 respectively), which controls the degree of cluster merging versus splitting (lower versus higher values, respectively), a total of five times at each iteration, and starting with 70 clusters in which mutations are initially randomly assigned. A long burn-in combined with widely spaced collection intervals of posterior samples was chosen so as to minimize the chance of violating the assumption of independent posterior sampling. Furthermore, 70 initial clusters were used to ensure that the starting distribution of mutations was spread over all 65 prior reference signatures plus a few additional clusters to promote the extraction of novel mutational signatures beyond the set of given priors. At each iteration, each mutation is assigned to a cluster with a high proportion of mutations in the same mutation category, sample or parent node. Clusters with cosine similarity > 0.9 are merged as per the default settings, whereas residual mutations unassigned to the set of extracted signatures due to uncertain cluster membership are grouped together to represent the percentage of data that is unexplained by the resultant model. A cosine similarity of >0.8 (as computed using the philentropy R package[45]) along with manual inspection was used to determine whether any of the extracted signatures match any of the known priors, in which a slightly lower similarity threshold was used to account for possible variations of the reference signatures. A computational deconvolution method known as the Perturbation model[46] was used to estimate the per cent contribution of PCAWG mutational signatures composing each of the HDP-extracted signatures as a secondary measure of similarity between known and extracted signatures. Extracted signatures that were unique enough such that no close match to any prior can be assigned with reasonable certainty were considered novel. For this analysis, six independent posterior sampling chains were executed concurrently for gauging convergence to stable cluster assignments for all mutations, where random seeds of 1-, 2-, 3-, 4-, 5-, and 6-million were assigned, respectively. The overall HDP node structure including the concentration parameter settings used for signature extraction is outlined in Supplementary Fig. 3.

Using SigProfiler for the extraction of mutational signatures

The SigProfilerExtractor python package[44] (https://github.com/AlexandrovLab/SigProfilerExtractor), which is based on the non-negative matrix factorization algorithm, served as an alternative means for mutational signature identification. The algorithm was configured to identify 15 mutational signatures and run with 1,000 iterations. Comparison of HDP and SigProfiler extracted 192 trinucleotide context signatures was performed by evaluating the cosine similarity metric, in which a value of >0.8 was deemed to indicate that a given pair of signatures were the same or slightly different versions of each other.

Telomere lengths and heritability

The telomere length (in units of base pairs) of each microdissection studied was estimated by analysing the corresponding WGS data for telomeric reads (containing TTAGGG and CCCTAA hexamers). To accomplish this, Telomerecat v.3.4.0 software[47] was used, with length correction enabled, while setting the number of simulations to 100 to constrain uncertainties in the length estimates. The samples from PD48367 and PD48372 were unable to have accurate telomere lengths estimated, and are therefore excluded from the analysis—we believe that this is because they were sequenced on the Illumina NovaSeq platform, whereas the other samples were sequencing on the Illumina X10 platform. The different chemistry or base-calling algorithm with NovaSeq apparently interferes with telomere length estimation, possibly because of mis-mapping of poor-quality reads to the ends of chromosomes. Each SNV cluster was assigned the telomere length corresponding to the member microdissection with the highest median VAF.

We modelled telomere lengths using Bayesian mixed effects models—these enabled us to assess the effects of age, clone size and disease on telomere lengths, while concurrently controlling for and quantifying the correlation arising from phylogenetic relationships among clones and within-patient non-independence. The specific algorithm we used was the R package, MCMCglmm[48], and the code and data for the analysis are available in the Supplementary Code. Further details are available in the Supplementary Methods.

Cell culture

HepG2, Hep3B and PLC/PRF/5 cells were obtained from ATCC and cultured in Dulbecco's modified Eagle's medium (DMEM)/10% foetal calf serum (FCS) in a 5% CO_2 atmosphere. Cell identity was confirmed by STR (short tandem repeat) genotyping. Cells were regularly tested for mycoplasma contamination and always found to be negative. Insulin (Sigma) stimulation was performed by culturing the cells in serum-free DMEM for 16 h before adding insulin at a final concentration of 100 nM.

Vectors

Retroviral vectors (pMSCV-hFOXO1-eGFP:P2A:Puromycin) containing wild-type FOXO1 (NM_002015.4) (VB190709-1030pwk), $FOXO1^{R21L}$

(VB190709-1028bjm) or FOXO1^{S22W} (VB190709-1032nwa) were purchased from VectorBuilder.

FOXO1–eGFP imaging and high-content analyses

High-content and live-cell analyses of FOXO1–eGFP expressing cells, counter-stained with Hoechst 33342 and SPY-555-Actin (Spirochrome), were conducted on an Operetta CLS system using a 20× air NA = 0.4 objective. Images of fixed cells were analysed using Harmony software (PerkinElmer). Any non-cellular material (for example, bright areas caused by coverslip edges) were removed; nuclei were segmented from DAPI fluorescence; a 9-pixel-wide cytoplasmic ring from around each nucleus was segmented from GFP fluorescence; and a background region was sampled from any cell-free areas 120–150 pixels away from any nucleus. Nuclei were filtered from fragments or other non-cell small objects by setting thresholds on nuclear area, roundness and width:length ratio. Mean nuclear, cytoplasmic and background GFP fluorescence intensities were measured, and from these the nuclear:cytoplasmic ratio was calculated for each cell using background-subtracted values. The $\log_{10}$ of these values was taken.

For live cells a similar analysis was carried out using CellProfiler. Illumination correction images were calculated for both GFP and Hoechst channels by polynomial fit, and subtracted; nuclei were segmented from the Hoechst images; cytoplasm was segmented from the GFP signal, with a 9-pixel-wide ring around the nucleus used to restrict the measurement to the perinuclear region; mean nuclear and cytoplasmic GFP intensities were measured; and the nuclei were tracked through the time series. Nuclear:cytoplasmic ratios were calculated and the $\log_{10}$ of these values was taken.

Results from the live cells are displayed as the median ± variance of pooled data from four wells, each with 8 fields of view giving 1,000–2,000 cells analysed per well, a total of 6,000–7,500 cells per condition.

Protein expression by immunoblotting

Immunoblotting, on SDS-PAGE gels was performed as previously reported[49] using the following antibodies: anti-β-actin (clone: AC15) (Sigma, A5441, 1:5,000, RRID:AB_476744); anti-AKT (clone: C73H10) (Cell Signaling, 2938, 1:1,000, RRID:AB_915788); anti-phospho-AKT (T308) (clone: 244F9) (Cell Signaling, 4056, 1:1,000, RRID:AB_331163); anti-GFP (Abcam, ab6556, 1:1,000, RRID:AB_305564); anti-FOXO1 (clone: C29H4) (Cell Signaling, 2880, 1:1,000, RRID:AB_2106495); anti-phospho-FOXO1 (T24) (Cell Signaling, (9464, 1:1,000, RRID:AB_329842). Uncropped versions of the blots are shown in Supplementary Fig. 4.

Metabolomics

HepG2 cells expressing either wild-type FOXO1-eGFP or FOXO1^{S22W}-eGFP were cultured overnight in serum-free medium before stimulation with or without 100 nM insulin for 3 h before collection. Cells were washed in PBS, before extraction and lysis in 50% methanol, 30% acetonitrile (both Fisher), 20% ultrapure water and 5 µM Valine d8 (internal control, CK isotopes) on dry ice. The supernatant from the cellular lysate was then stored at −80°C until the stage that metabolomics was to be performed.

A Millipore Sequant ZIC-pHILIC analytical column (5 µm, 2.1 × 150 mm) with a 2.1 × 20 mm guard column (both 5-mm particle size) carrying a binary solvent system was used to perform HILIC chromatographic separation of metabolites. For solvent A, we used 20 mM ammonium carbonate, 0.05% ammonium hydroxide; and for Solvent B, we used acetonitrile. The column oven was maintained at 40 °C and the autosampler tray at 4 °C. A flow rate of 0.200 ml min^{-1} was used for the chromatographic gradient, as follows: 0–2 min: 80% B; 2–17 min: linear gradient from 80% B to 20% B; 17–17.1 min: linear gradient from 20% B to 80% B; 17.1–22.5 min: hold at 80% B. We used randomization to define the order in which samples were processed; analyses with LC–MS were performed blinded to each sample's identity. The injection volume was 5 µl. Pooled samples were generated by equally mixing all of the individual samples; these were interspersed at regular intervals among the samples to provide quality control for the actual test samples.

Metabolites were quantified with a Thermo Scientific Q Exactive Hybrid Quadrupole-Orbitrap Mass spectrometer (HRMS) coupled to a Dionex Ultimate 3000 UHPLC. The full-scan, polarity-switching mode was chosen for the mass spectrometer. The following conditions were used: spray voltage of +4.5 kV/−3.5 kV; heated capillary at 320 °C; the auxiliary gas heater at 280 °C; sheath gas flow was 25 units; the auxiliary gas flow was 15 units; and the sweep gas flow was 0 units. Data from the HRMS were acquired in the range of m/z = 70–900; the resolution was set at 70,000, the AGC target at 1×10^6 and the maximum injection time at 120 ms. For confirming metabolite identities, we used two parameters: (1) precursor ion m/z matched to within 5 ppm of the theoretical mass that would be predicted from its chemical formula; (2) the retention time of metabolites matched the retention time of a purified standard run with the same chromatographic method to within 5% variance. We used the Thermo Fisher software Tracefinder 5.0 Chromatogram to review and undertake peak area integration. To correct variation arising from the analytic process that could arise anywhere from sample handling through to instrument analysis, we normalized the peak area of each metabolite against the total ion count for that sample. These normalized peak areas were then those used in the downstream statistical data analysis (as shown in Fig. 2c).

Statistical analysis of metabolomics data was performed using linear models with insulin (with or without) and *FOXO1* status (mutant or wild type) as the predictive variables, and normalized metabolite levels as the dependent variable. Likelihood ratio tests were used to generate *P* values, which were then corrected for multiple hypothesis testing using the Benjamini–Hochberg method. A threshold of $q < 0.01$ was used for significance. Code and data for this analysis are available in the Supplementary Code.

Preprocessing of RNA-sequencing data from HepG2 cell lines

HepG2 cell line samples ($n = 30$) were subjected to two lanes of 150-base-pair paired-end RNA sequencing using the Illumina HiSeq 4000 platform. The human reference genome used was hs37d5 from the 1000 Genomes Project, with gene annotations based on Ensembl release 75 data. Adaptors and low-quality reads were removed using Trim Galore (https://github.com/FelixKrueger/TrimGalore) with the following parameters: -q 20–fastqc–paired–stringency 1–length 20 -e 0.1. The Spliced Transcripts Alignment to a Reference (STAR) aligner was used to map the raw sequencing reads to the GRCh37 (hg19) human reference genome. Duplicate reads were marked using Picard. Base quality score recalibration was performed using GATK, and substitutions were called using HaplotypeCaller. The featureCounts software[50] was used to summarize gene expression values, and the cpm function from the EdgeR R package[51] was used to normalize the data into the log counts per million scale. All heat maps were generated using the pheatmap R package (https://cran.r-project.org/web/packages/pheatmap/index.html).

Gene set enrichment analysis

Gene set enrichment analysis (GSEA v.3.0) was performed using a pre-ranked list of genes, 2,000 permutations, and all Gene Ontology and Reactome associated gene sets that had at most 500 genes (June_01_2021 version, downloaded from http://download.baderlab.org/EM_Genesets/). Specifically, for each gene, two linear models were built using the lm function in the R statistical programming environment: one that included both *FOXO1* driver and insulin status (that is, either present or absent) as independent variables; and one that only included insulin status. The dependent variable in both models is the expression of the gene in the model. The likelihood ratio test was then used to calculate a *P* value between each pair of nested models per gene. This *P* value was subsequently multiplied by the sign of the regression coefficient for mutation status in the model with the driver for each gene. Finally, the gene list was ranked according to this set of *P* values as follows: -(≈0) … −0.05 … −0.99 … 0.99 … 0.05 … ≈0, wherein genes at the bottom of the list are expected to be the most associated with the presence of the *FOXO1* driver, while accounting for the effects of insulin status.

Reporting summary

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

Data availability

WGS data in the form of BAM files across samples reported in this study have been deposited in the European Genome-Phenome Archive (accession number EGAD00001006255). RNA-sequencing data have been deposited in the European Nucleotide Archive (https://www.ebi.ac.uk/ena/browser/home) with accession number ERP123192.

Code availability

Detailed methods and custom R scripts for the analysis of clinical features, telomere lengths and metabolomics data are available in the Supplementary Code. Other packages used in the analysis are listed below: R: v.3.5.1, Perl: v.5.3.0, Python: v.3.8.5, MATLAB: v.R2019b, BWA-MEM: v.0.7.17 (https://sourceforge.net/projects/bio-bwa/), cgpCaVEMan: v.1.11.2/1.13.14/1.15.1 (https://github.com/cancerit/CaVEMan), cgpPindel: v.2.2.2/2.2.4/2.2.5/3.2.0/3.3.0 (https://github.com/cancerit/cgpPindel), Brass: v.5.4.1/6.0.5/6.1.2/6.2.0/6.3.4 (https://github.com/cancerit/BRASS), ASCAT NGS: v.4.0.1/ 4.1.2/4.2.1 (https://github.com/cancerit/ascatNgs), JBrowse: v.1.16.1 (https://jbrowse.org/), cgpVAF: v.2.4.0 (https://github.com/cancerit/vafCorrect), alleleCount: v.4.1.0 (https://github.com/cancerit/alleleCount), SigProfiler: v.1.0.0-GRCh37 (https://github.com/AlexandrovLab), HDP: v.0.1.5 (https://github.com/nicolaroberts/hdp), dNdScv: v.0.0.1 (https://github.com/im3sanger/dndscv), Telomerecat: v.3.4.0 (https://github.com/jhrf/telomerecat), STAR: v.2.7.6a (https://github.com/alexdobin/STAR), Picard-tools: v.2.20.7 (https://broadinstitute.github.io/picard/), Samtools: v.1.12 (http://www.htslib.org/), TrimGalore: v.0.6.4 (https://github.com/FelixKrueger/TrimGalore), GATK: v.4.1.4.1 (https://gatk.broadinstitute.org/hc/en-us), GSEA: v.3.0 (https://www.gsea-msigdb.org/gsea/index.jsp), XGBoost: v.0.82.1 (https://xgboost.readthedocs.io/en/latest/), NDP.view2 (https://www.hamamatsu.com/eu/en/product/type/U12388-01/index.html), label.switching: v.1.8 (https://cran.r-project.org/web/packages/label.switching/index.html), philentropy: v.0.3.0 (https://cran.r-project.org/web/packages/philentropy/index.html), MCMCglmm: v.2.29 (https://cran.r-project.org/web/packages/MCMCglmm/index.html), Magick: v.2.0 (https://cran.r-project.org/web/packages/magick/index.html), Pheatmap: v.1.0.12 (https://cran.r-project.org/web/packages/pheatmap/index.html), Thermo Fisher software Tracefinder: v.5.0 (https://www.thermofisher.com/uk/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/lc-ms-data-acquisition-software/tracefinder-software.html), CellProfiler: v.4.0.3 (https://cellprofiler.org/), PerkinElmer Harmony: v.4.9 (https://www.perkinelmer.com/category/cellular-imaging-software).

30. Kleiner, D. E. et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1321 (2005).
31. Ishak, K. et al. Histological grading and staging of chronic hepatitis. *J. Hepatol.* **22**, 696–699 (1995).
32. Ellis, P. et al. Reliable detection of somatic mutations in solid tissues by laser-capture microdissection and low-input DNA sequencing. *Nat. Protoc.* **16**, 841–871 (2021).
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. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
35. Papastamoulis, P. label.switching: an R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw.* **69**, Code Snippet 1 (2015).
36. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
37. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
38. 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).
39. 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).
40. Stephens, P. J. et al. Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
41. Sohlenius-Sternbeck, A. K. Determination of the hepatocellularity number for human, dog, rabbit, rat and mouse livers from protein concentration measurements. *Toxicol. Vitr.* **20**, 1582–1586 (2006).
42. Lipscomb, J. C., Fisher, J. W., Confer, P. D. & Byczkowski, J. Z. In vitro to in vivo extrapolation for trichloroethylene metabolism in humans. *Toxicol. Appl. Pharmacol.* **152**, 376–387 (1998).
43. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
44. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
45. Drost, H.-G. Philentropy: information theory and distance quantification with R. *J. Open Source Softw.* **3**, 765 (2018).
46. Qiao, W. et al. PERT: a method for expression deconvolution of human blood samples from varied microenvironmental and developmental conditions. *PLoS Comput. Biol.* **8**, (2012).
47. Farmery, J. H. R. et al. Telomerecat: a ploidy-agnostic method for estimating telomere length from whole genome sequencing data. *Sci. Rep.* **8**, 1300 (2018).
48. Hadfield, J. D. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *J. Stat. Softw.* **33**, v033i02 (2010).
49. Hoare, M. et al. NOTCH1 mediates a switch between two distinct secretomes during senescence. *Nat. Cell Biol.* **18**, 979–992 (2016).
50. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
51. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).

Acknowledgements This work was supported by a Cancer Research UK Grand Challenge Award (C98/A24032) and the Wellcome Trust. S.W.K.N. holds an EMBO Long Term Fellowship (ALTF 721-2019). 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. M.H. is supported by a CRUK Clinician Scientist Fellowship (C52489/A19924) and a CRUK Accelerator Award (C18873/A26813). P.J.C. was supported by a Wellcome Senior Clinical Fellowship until 2020 (WT088340MA).

Author contributions P.J.C., M.H. and S.W.K.N. designed the experiments. S.W.K.N. performed mutation calling and computational analyses including visualization of results for mutation calling; identification of SNV clusters and the inference of phylogenetic relationships between them; assignment of indels and *FOXO1* hotspot mutations to SNV clusters; clone size estimation and comparisons; mutational signature extraction; identification of protein-coding and non-coding drivers; telomere length estimation; processing and normalization of RNA-sequencing data; gene set enrichment analysis; and estimation of the liver-wide mass of driver-mutation-bearing hepatocytes. S.W.K.N. developed software for the refinement of indel calling, phylogenetic inference and visualization of clonal structure, and clone size estimation, visualization and mapping to histological images. P.J.C. assisted with the filtering of structural variants, performed statistical inference of factors that affect telomere length using mixed effects models and supervised all statistical analyses. N. Brzozowska performed telomere length estimation. F.A. and I.M. provided support for running variants of dNdScv. M.R.S. advised on mutational signature extraction. T.H.H.C. provided support for running beta-binomial-based variant filtering. M.A.S. provided support and advice for performing LCM-specific variant-filtering algorithms for SNV and structural variant calls. D.L. and T.B. provided insights into indel filtering associated with homopolymers and problematic genomic loci. F.J.R., S.F.B., Y.H., B.W. and N. Birtchnell performed tissue sectioning, fixing, staining and histology image generation. S.F.B. also performed LCM and submission for WGS, and was responsible for the initial development of source code for producing diagnostic plots to facilitate the manual determination of clonal relationships, and the visualization of phylogenetic tree structures. P.R., A.I. and T.B. provided wet laboratory support. N.W., J.T., K.R. and A.P.B. provided technical support for computational analyses. M.H. and F.J.R. provided biological samples used in this study, and the associated clinical annotations were curated with assistance from S.J.A. and S.E.D. S.J.A. and S.E.D. analysed histology sections of background liver and HCC from all patients in the study, and L.M. supervised microdissection of tissue samples for sequencing. M.H. coordinated all validation experiments relating to *FOXO1* hotspot mutations using HCC cell lines, with additional support from H.N. M.Y., E.N. and C.F. performed analysis of metabolites from HCC cell lines. L.C. and R.R. performed processing and quality control of RNA-sequencing samples and data. P.J.C., S.W.K.N. and M.H. drafted the manuscript with input and guidance from M.R.S. and I.M., and updated the paper after contributions from all authors.

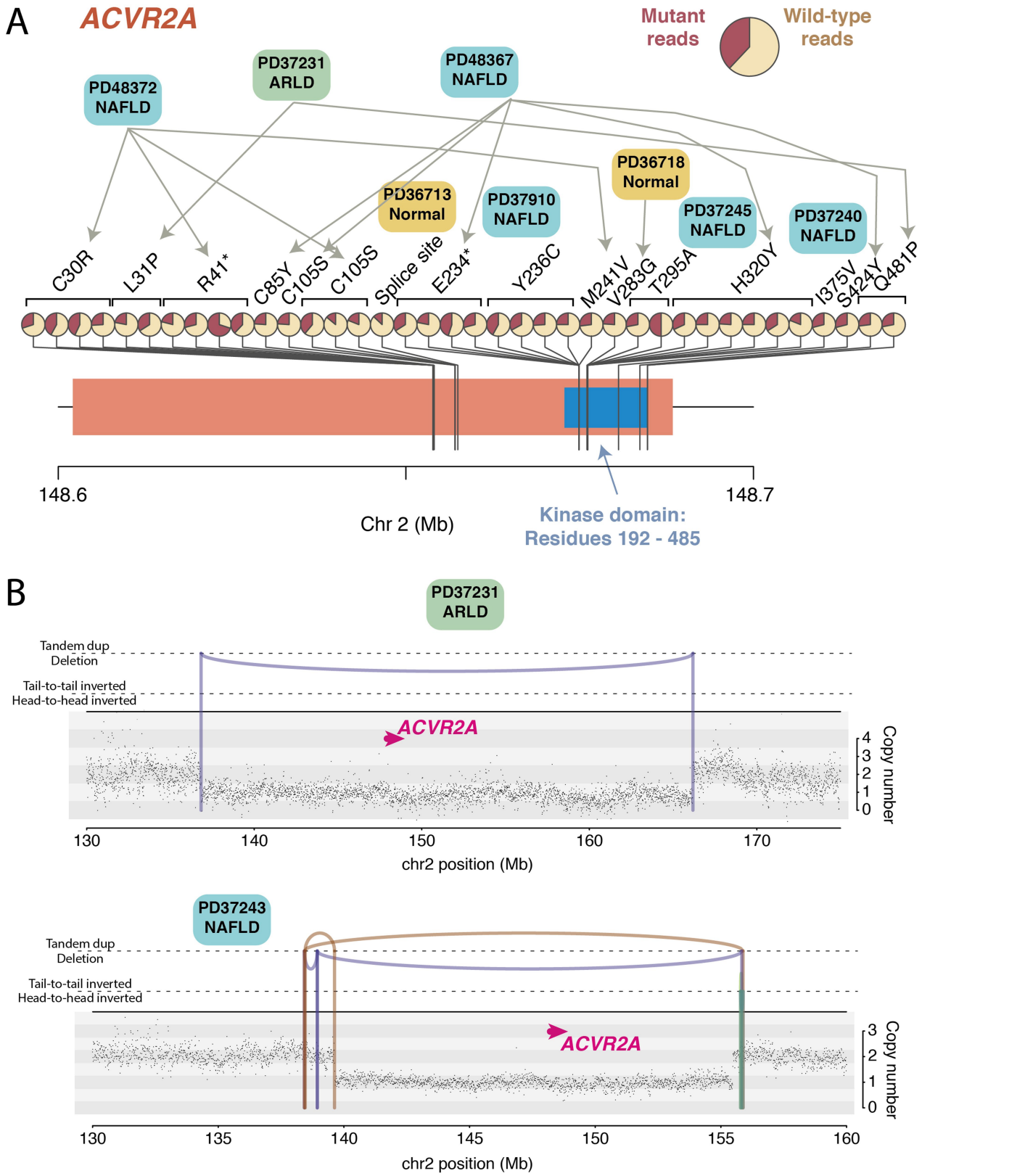
Competing interests A patent has been filed by CRUK's technology transfer office, with support from that of Wellcome Sanger Institute (named inventors: S.W.K.N., M.H. and P.J.C.), covering the use of somatic mutations in liver tissue for stratifying diagnosis and treatment of patients with metabolic diseases.

Additional information

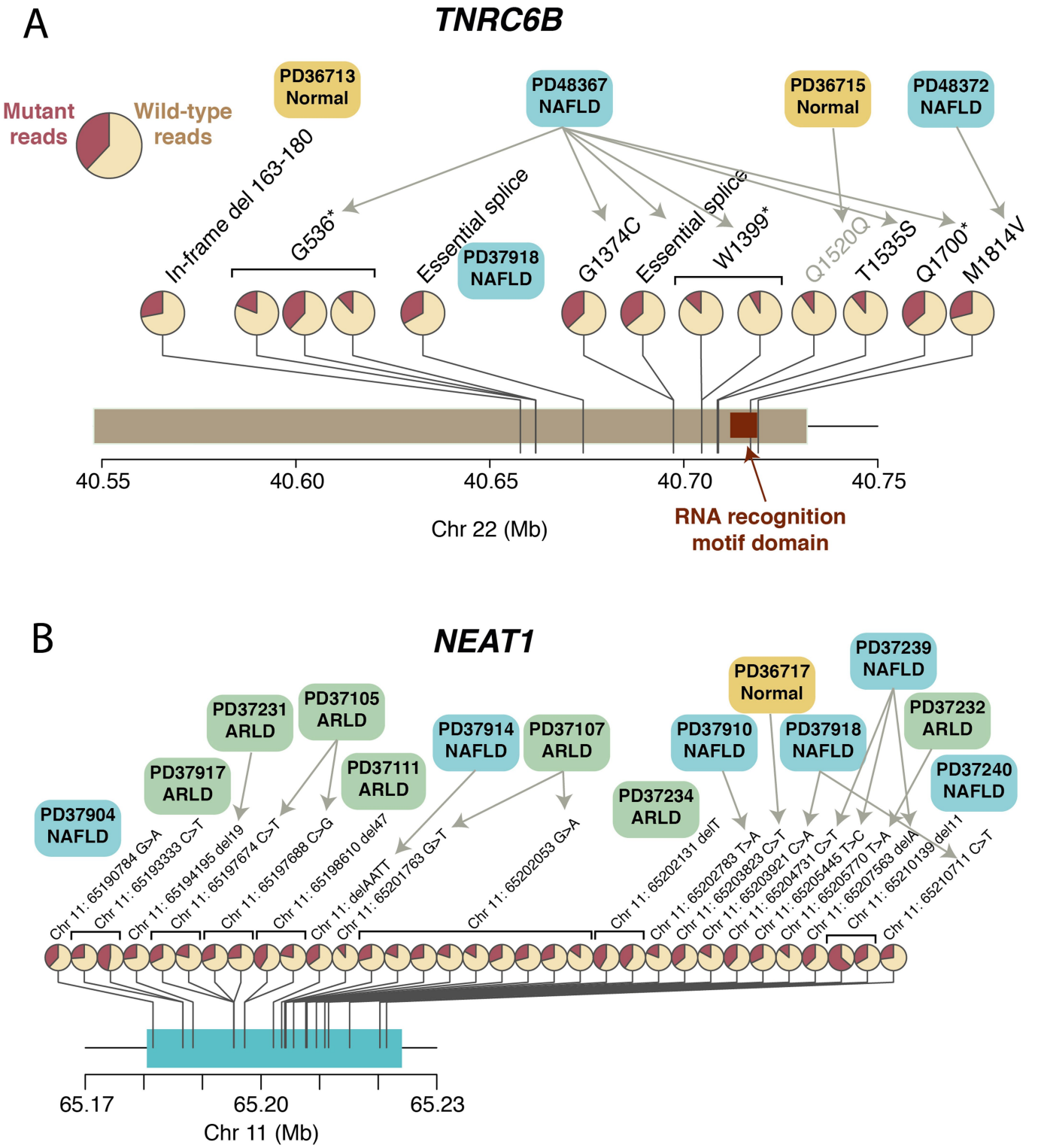
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03974-6.

Correspondence and requests for materials should be addressed to Matthew Hoare or Peter J. Campbell.

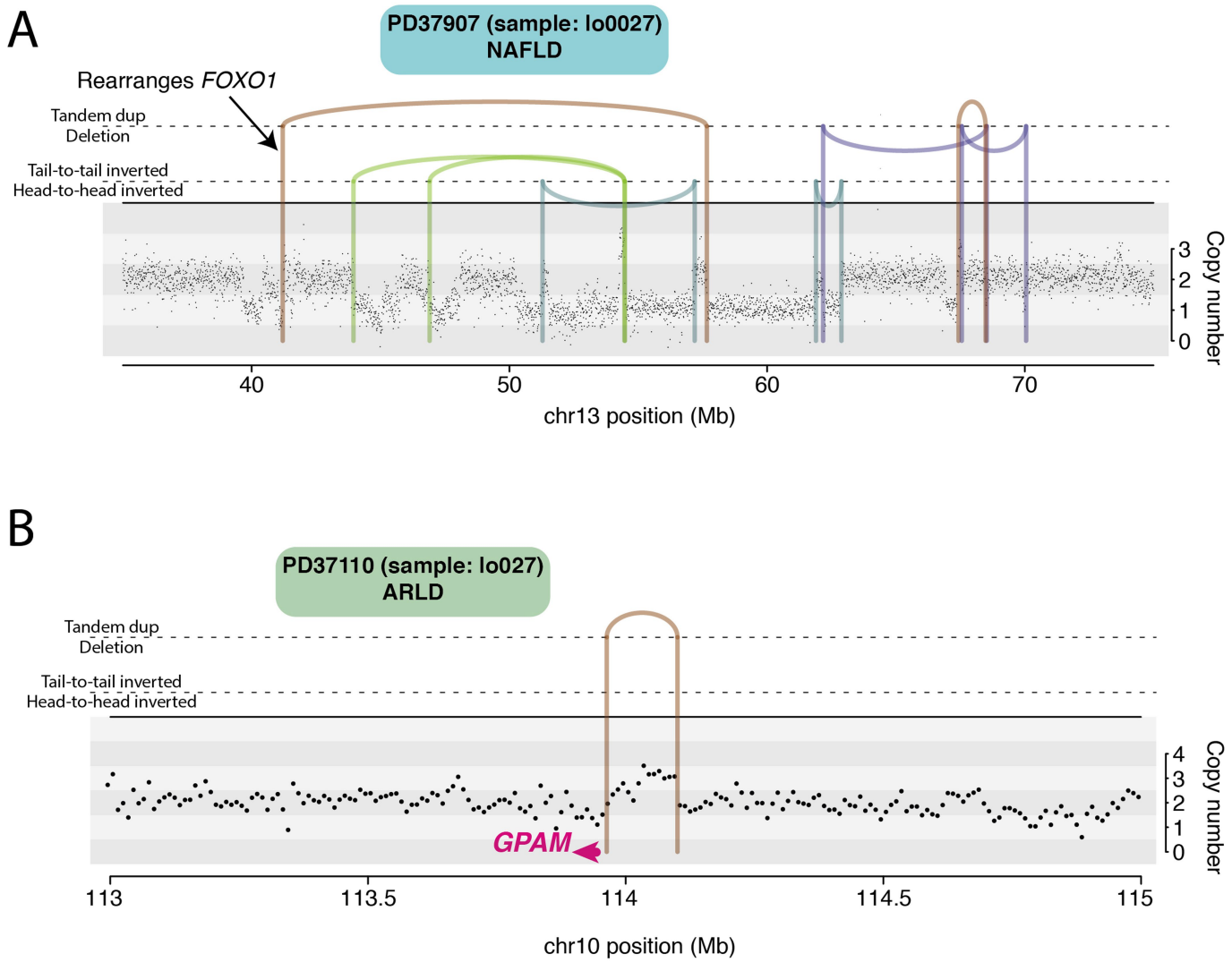
Peer review information *Nature* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.



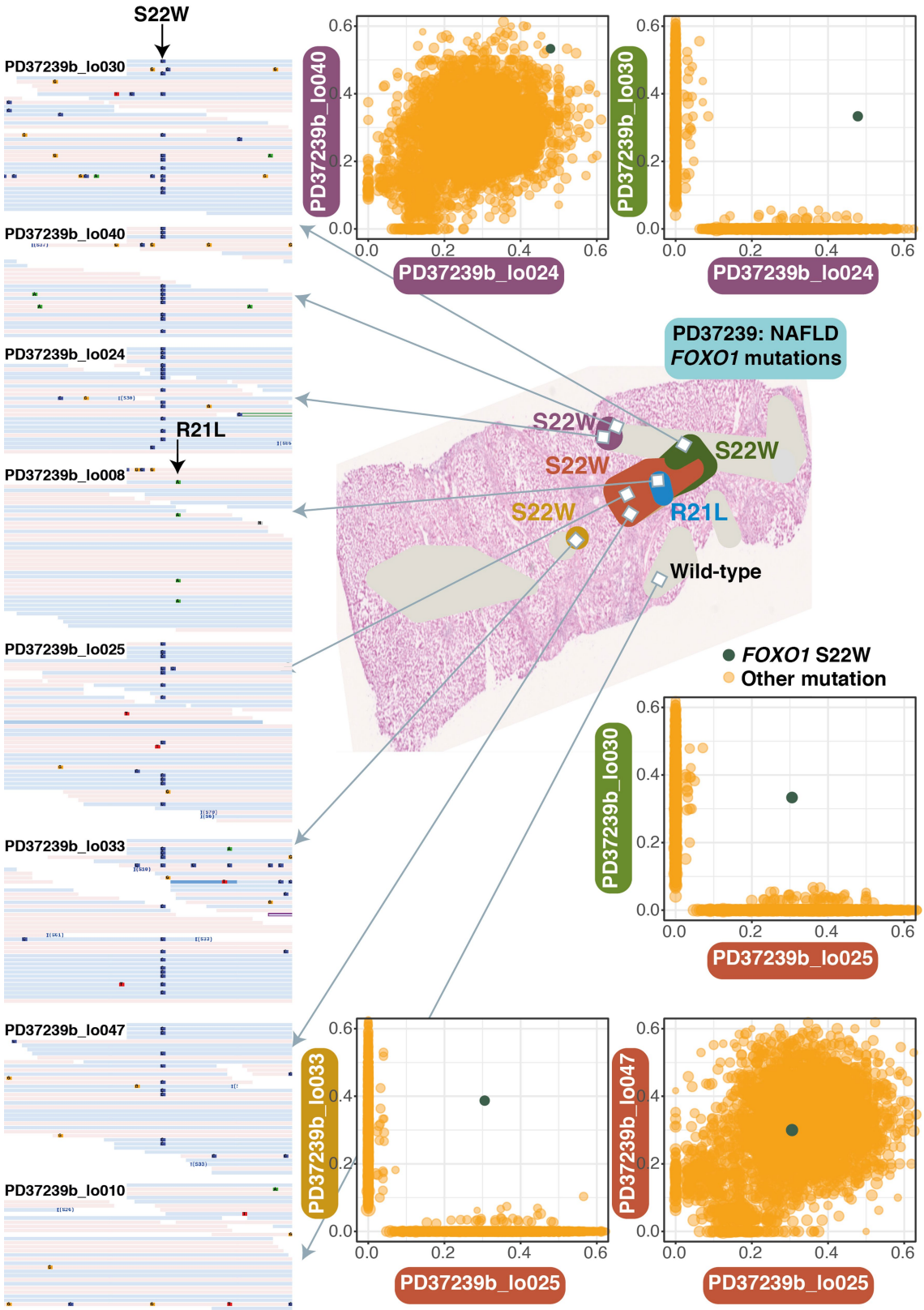
Extended Data Fig. 1 | Mutations in *ACVR2A*. a, Distribution of somatic mutations in *ACVR2A* according to genomic location. Pie charts show fraction of sequencing reads reporting the mutant allele in each microdissection. **b**, Two microdissections in different patients showing structural variants generating copy loss of *ACVR2A*. Black points represent corrected read depth along the chromosome. Lines and arcs represent structural variants, coloured by the orientation of the joined ends (purple, deletion-type orientation; brown, tandem-duplication-type orientation; turquoise, head-to-head inverted; green, tail-to-tail inverted).



Extended Data Fig. 2 | Mutations in *TNRC6B* and *NEAT1*. **a**, Distribution of somatic mutations in *CLCN5* according to genomic location. Pie charts show fraction of sequencing reads reporting the mutant allele in each microdissection. **b**, Distribution of somatic mutations in the long non-coding RNA, *NEAT1*, according to genomic location. Pie charts show fraction of sequencing reads reporting the mutant allele in each microdissection.

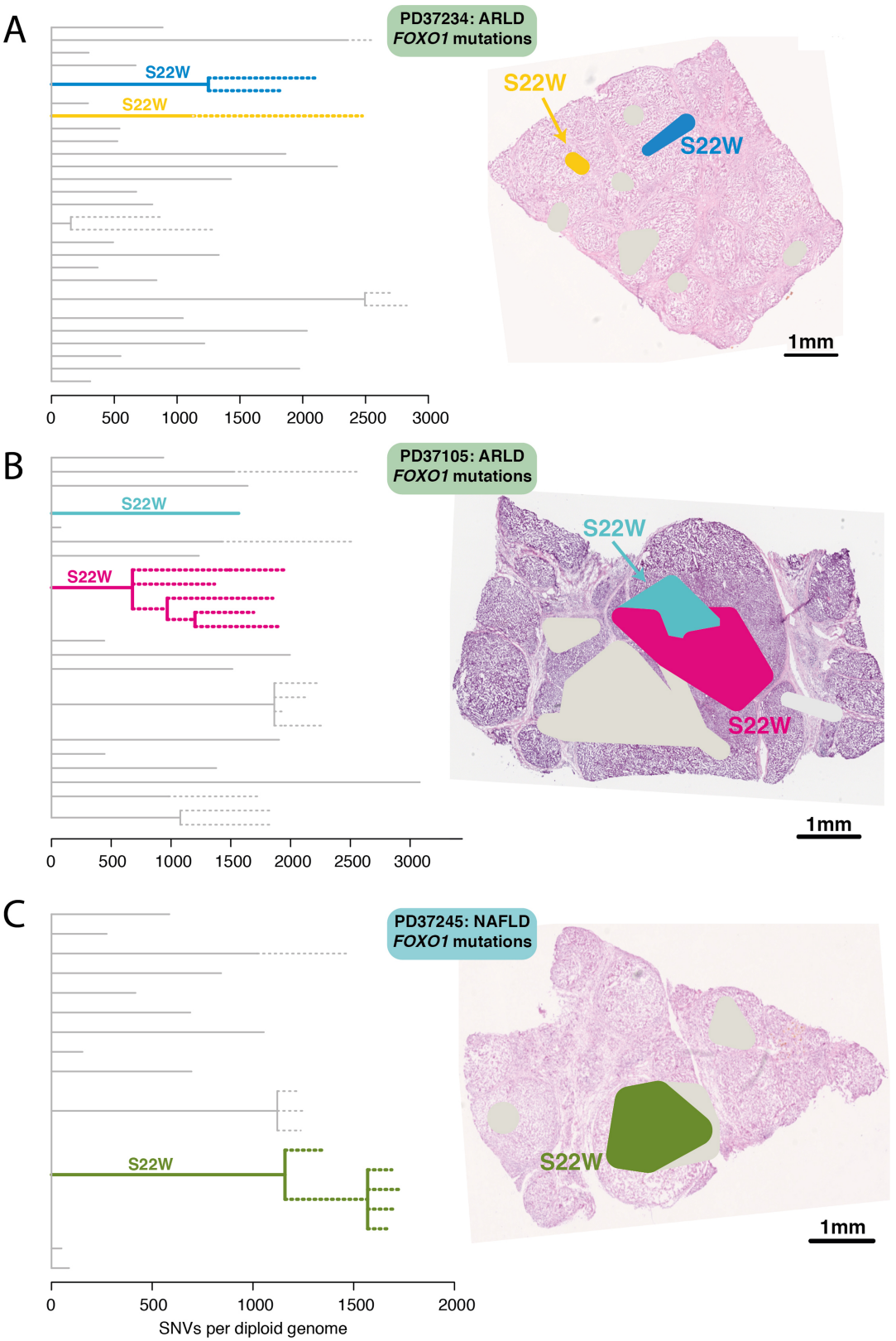


Extended Data Fig. 3 | Structural variants affecting *FOXO1* and *GPAM*. **a**, A chromothripsis event affecting chromosome 13 in one of the microdissections from PD37907, a patient with NAFLD. Black points represent corrected read depth along the chromosome. Lines and arcs represent structural variants, coloured by the orientation of the joined ends (purple, deletion-type orientation; brown, tandem-duplication-type orientation; turquoise, head-to-head inverted; green, tail-to-tail inverted). The structural variant that breaks *FOXO1* is highlighted, and would be predicted to break the gene within the first intron, preserving the first coding exon but deleting the remaining coding exons. **b**, A tandem duplication upstream of *GPAM* in a microdissection from PD37110, a patient with ARLD. *GPAM* is left intact, but the tandem duplication starts 20kb upstream of the gene.

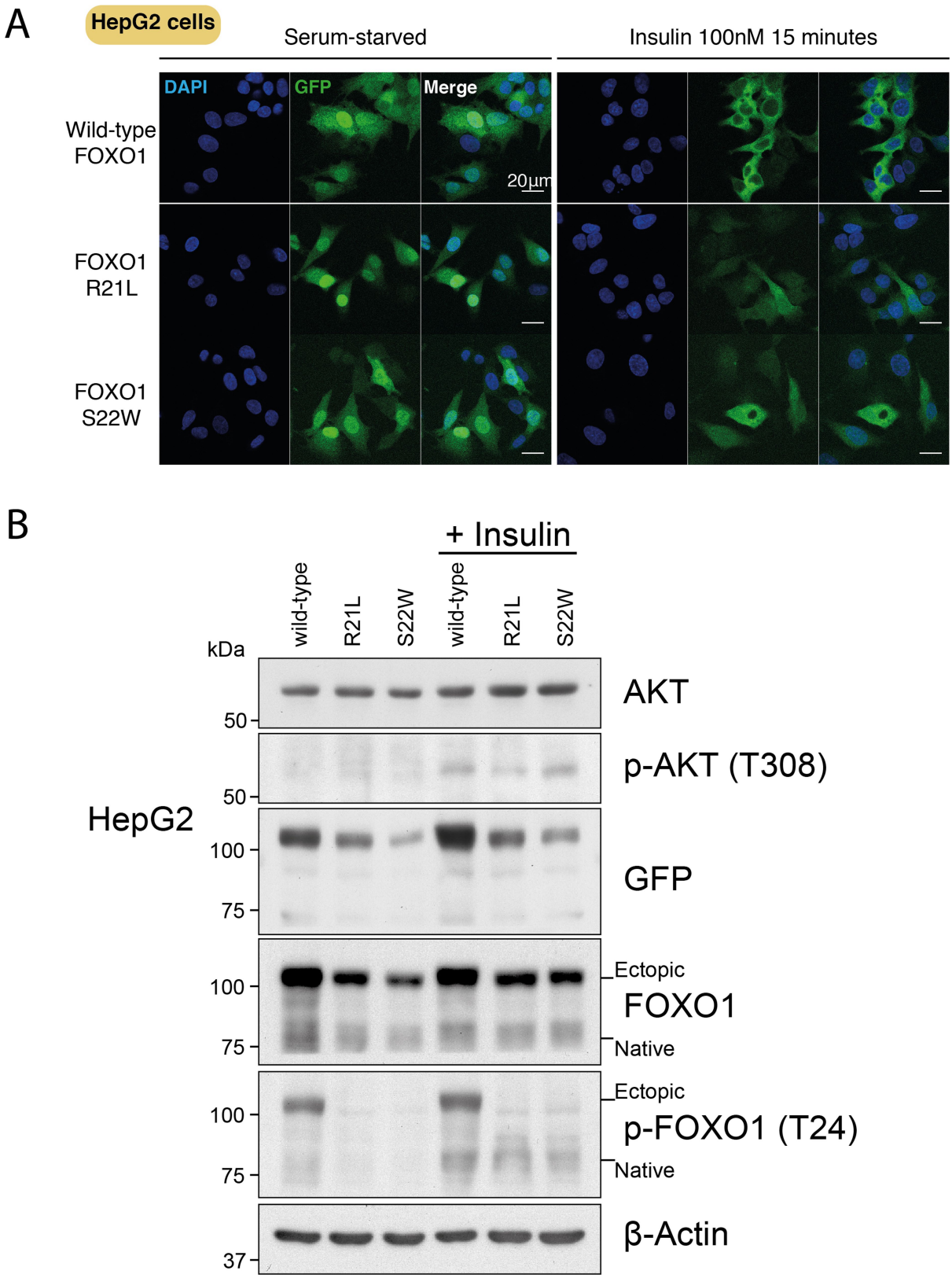


Extended Data Fig. 4 | See next page for caption.

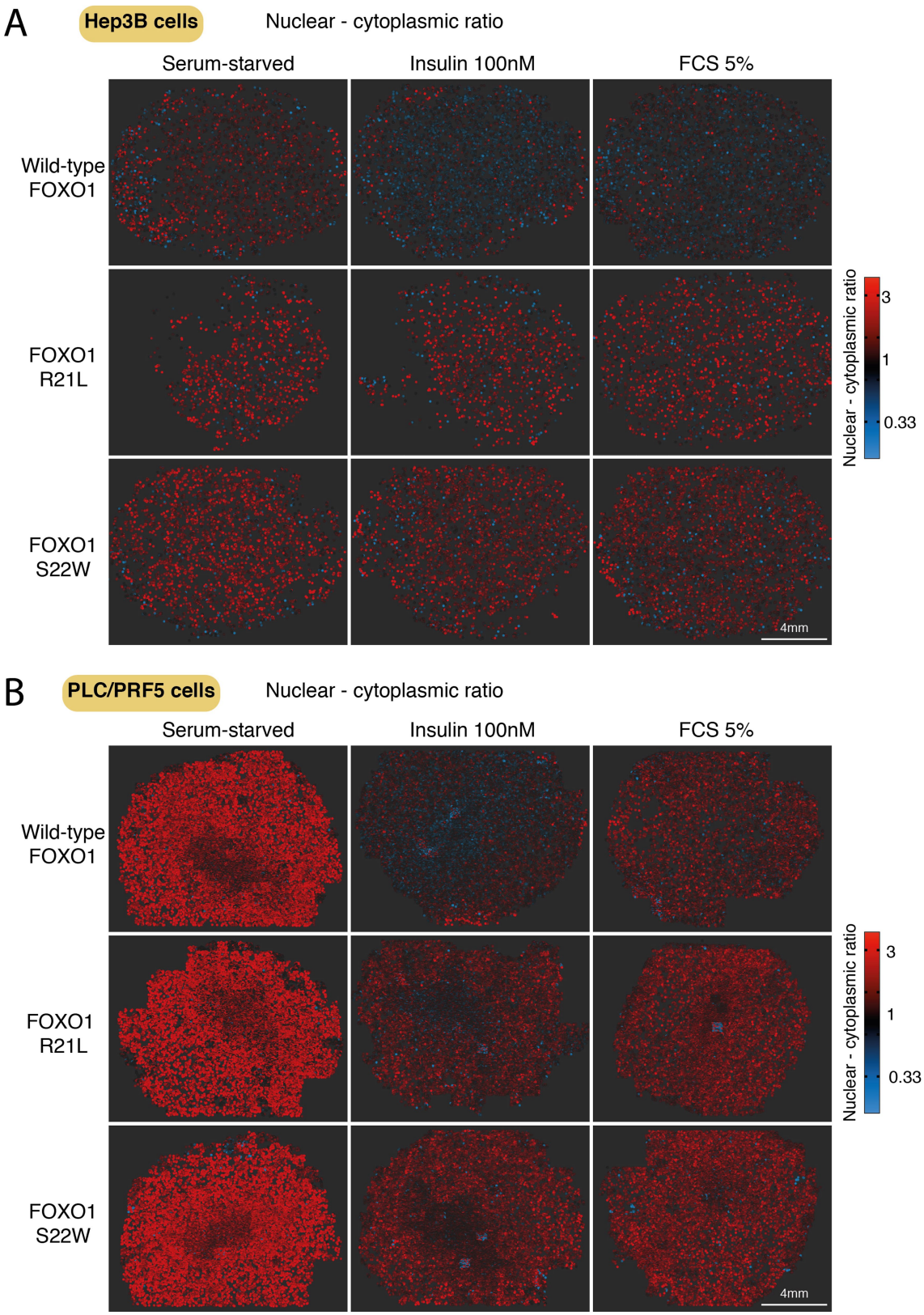
Extended Data Fig. 4 | Multiple independent acquisitions of *FOXO1* mutations in PD37239. The clone map from Fig. 1b is shown, laid onto an H&E-stained section. On the left of the figure, raw sequencing data from representative samples with and without *FOXO1* mutations are shown, with their physical locations on the H&E section shown by the arrows. In the sequencing data, reads mapping to the forward strand of the reference genome are in pink; the reverse strand in blue. Base calls that do not match the reference genome are shown as coloured squares. The locations of the S22W and R21L mutations are marked with arrows. The scatterplots arranged around the H&E section represent VAF plots of mutations in pairs of samples. The colours of the x and y axis titles match the clone map colours of the H&E section. Individual mutations called in either sample are shown in orange, according to their VAF, with the *FOXO1* S22W mutation shown in dark green. In clonally related pairs of samples, most of the mutations are shared by both samples, evident as a cloud of mutations with non-zero VAF. In clonally unrelated samples, the mutations line the *x* and *y* axes, with the one exception being the *FOXO1* mutation, indicating that it is independently acquired in the two clones.



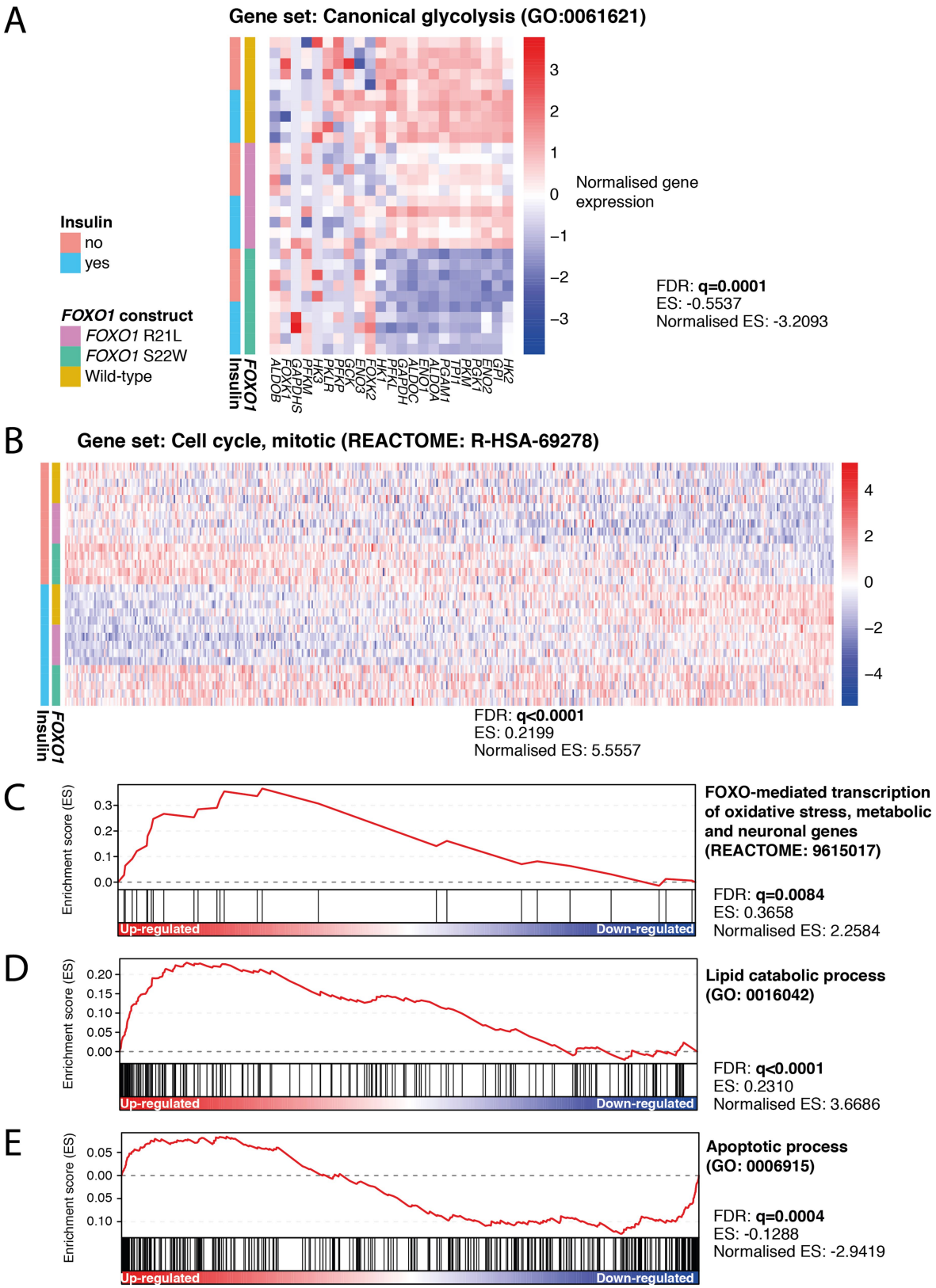
Extended Data Fig. 5 | Further examples of *FOXO1* mutations in patients with chronic liver disease. a–**c**, Phylogenetic trees and clone maps are shown for PD37234 (**a**), PD37105 (**b**) and PD37245 (**c**). The left panel shows the phylogenetic tree, with coloured branches showing independently acquired mutations. 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 < 100% of cells. The right panel shows the clones from the phylogenetic tree mapped onto an H&E-stained photomicrograph of the liver, with *FOXO1*-mutant clones coloured to match the tree.



Extended Data Fig. 6 | Somatic mutations of FOXO1 impair its phosphorylation and nuclear export. a, HepG2 cells were transfected with the indicated wild-type or mutant constructs of FOXO1 fused with a C-terminal GFP. Cells were counterstained with DAPI to highlight the nucleus, and imaged after overnight serum starvation conditions (left) and after 15 min of exposure to 100 nM insulin (right). Studies were performed in triplicate. **b**, HepG2 cells, expressing ectopic eGFP-tagged wild-type or mutant FOXO1 constructs as indicated and treated for 15 min with vehicle or insulin (100 nM), were analysed for the indicated proteins by immunoblotting. Molecular weight markers (kDa) indicated. Studies were performed in triplicate. Uncropped versions of the blots are shown in Supplementary Fig. 4.

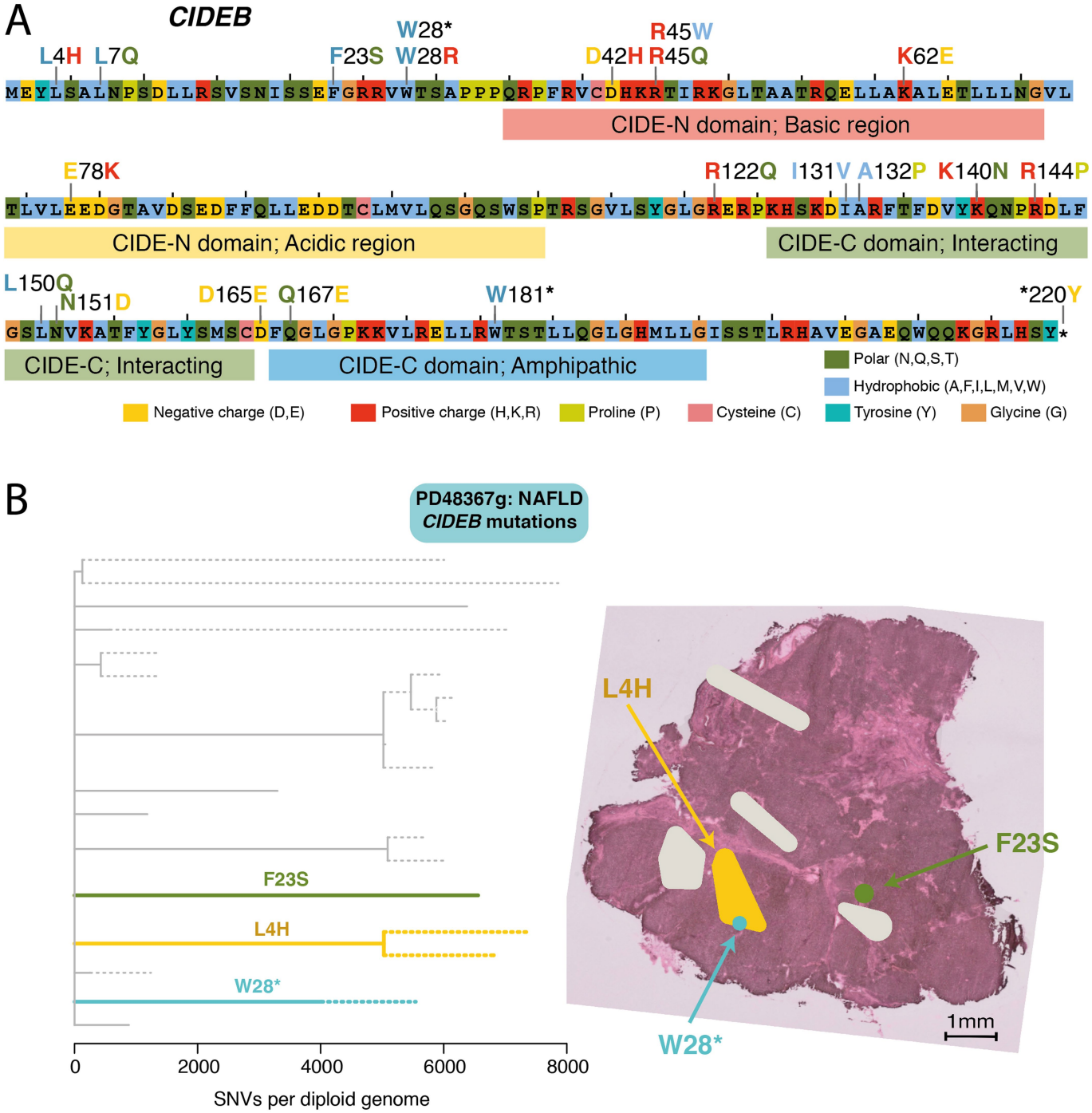


Extended Data Fig. 7 | Nuclear–cytoplasmic ratios for wild-type and mutant FOXO1-GFP constructs in HCC cell lines. a, **b**, Wide-field view of Hep3B (**a**) and PLC/PRF5 (**b**) cells pseudocoloured on a blue-to-red scale by the nuclear-cytoplasmic ratio of FOXO1-GFP. Cells were imaged under conditions of serum starvation (left), after exposure to insulin 100nM for 15 min (middle) or foetal calf serum (FCS) for 15 min (right).

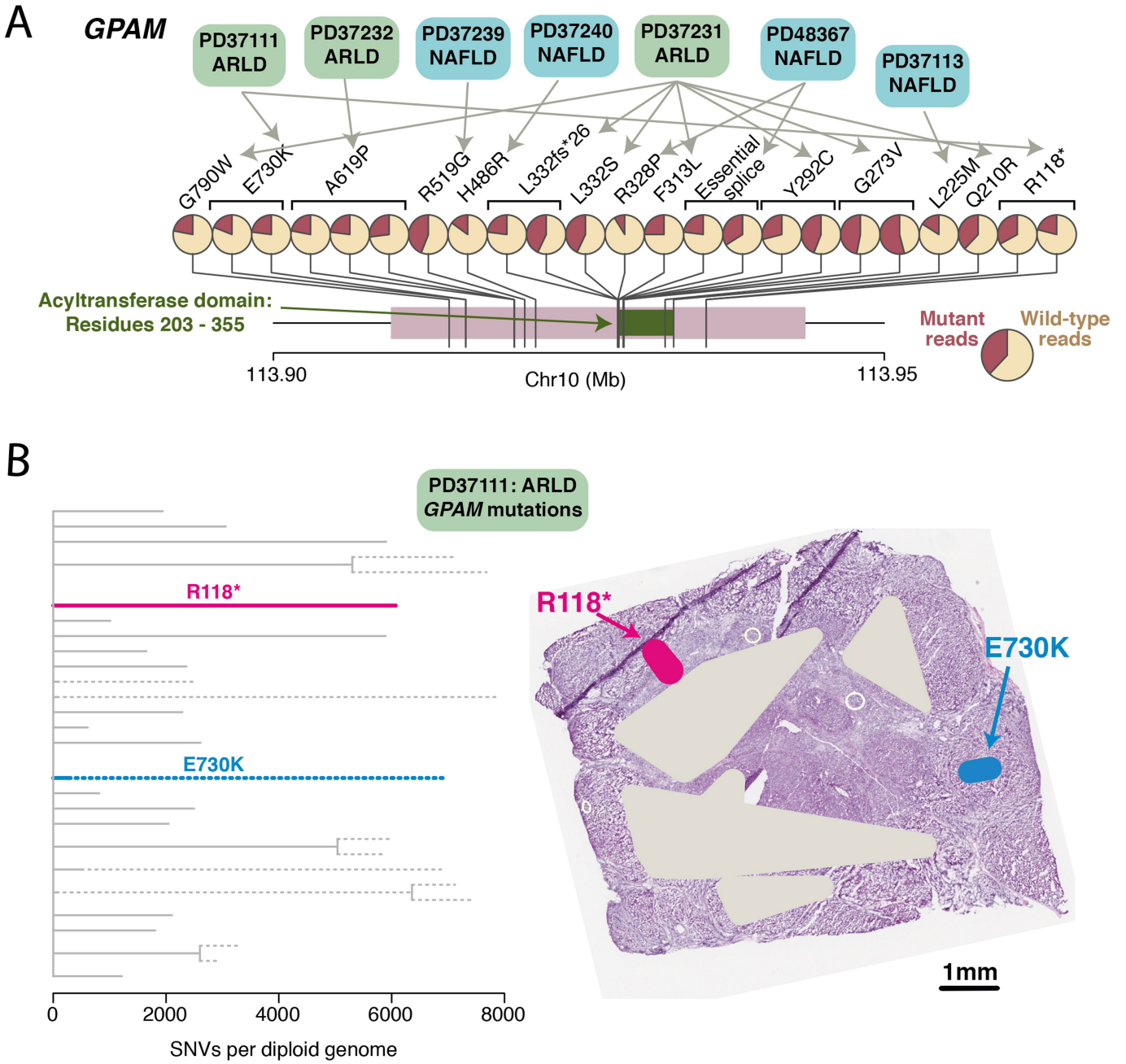


Extended Data Fig. 8 | See next page for caption.

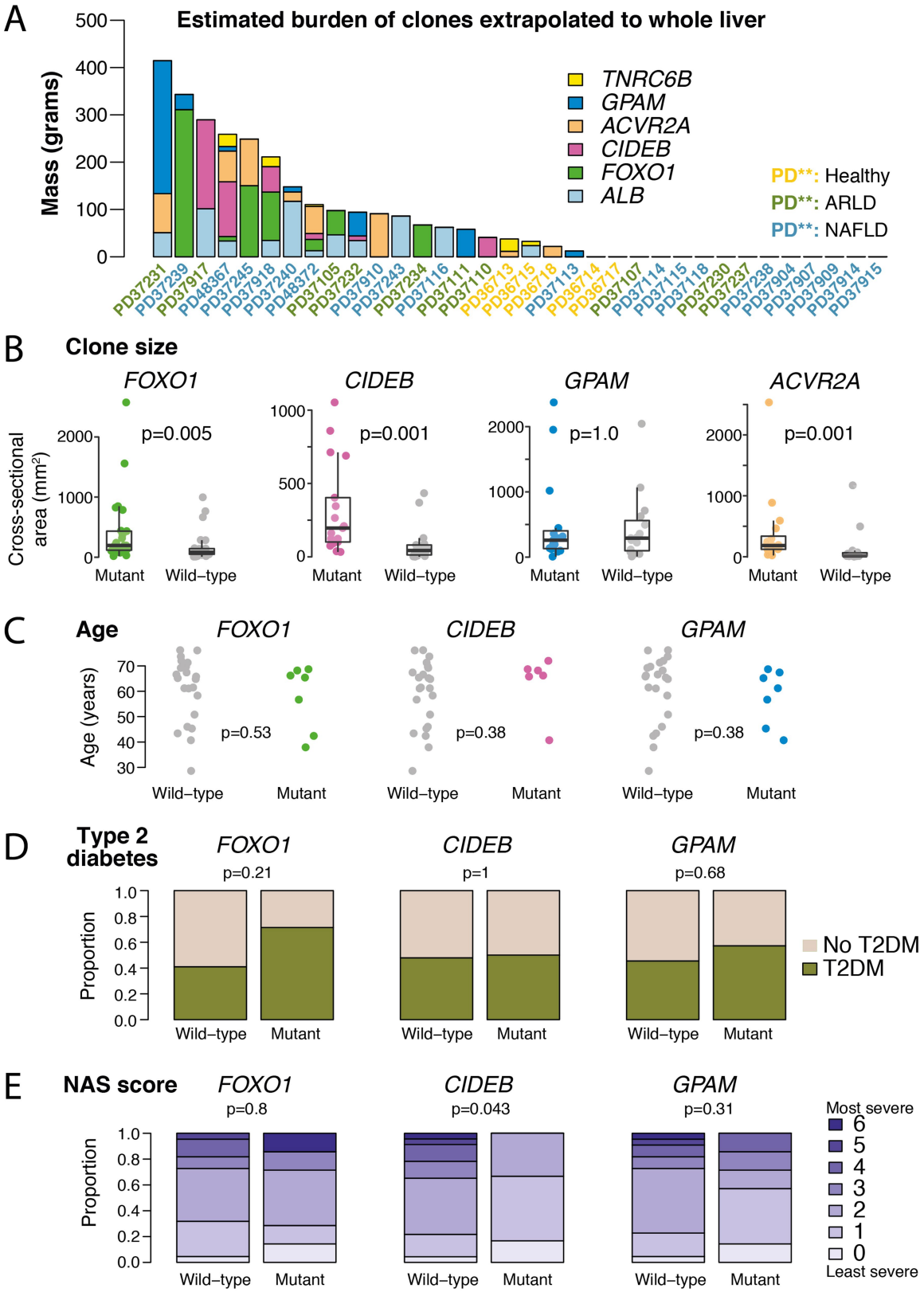
Extended Data Fig. 8 | RNA sequencing from cell lines transduced with either wild-type or mutant FOXO1-GFP constructs. a, Heat map showing gene expression levels for genes in the 'Canonical Glycolysis' gene set from GO (GO:0061621). The order of genes on the x axis is determined by the level of significance (and direction of change) and the order of samples on the y axis is by condition (*FOXO1* status and insulin status). **b**, Heat map showing gene expression levels for genes in the 'Cell cycle, mitotic' gene set from Reactome (R-HSA-69278). The order of genes on the x axis is determined by the level of significance (and direction of change) and the order of samples on the y axis is by condition (*FOXO1* status and insulin status). **c**–**e**, Enrichment plots for the 'FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes' gene set of Reactome (9615017) (**c**); 'Lipid catabolic process' gene set of GO (0016042) (**d**); and 'Apoptotic process' gene set of GO (0006915) (**e**). In each, the top panel reflects the cumulative enrichment score as the gene set is traversed from most up-regulated to most down-regulated in the presence of *FOXO1*-mutant constructs. The bottom panel in each shows the ranking of each gene in the gene set across all genes measured.



Extended Data Fig. 9 | *CIDEB* mutations in patients with chronic liver disease. a, Distribution of somatic mutations in *CIDEB*. Amino acid residues are coloured by type, with observed mutations in chronic liver disease shown above the wild-type protein sequence. **b**, Phylogenetic trees and clone maps are shown for one of the Couinaud segments of PD48367 with *CIDEB* mutations. The left panel shows the phylogenetic tree, with coloured branches showing independently acquired driver mutations. 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 < 100% of cells. The right panel shows the clones from the phylogenetic tree mapped onto an H&E-stained photomicrograph of the liver, with mutant clones coloured to match the tree.

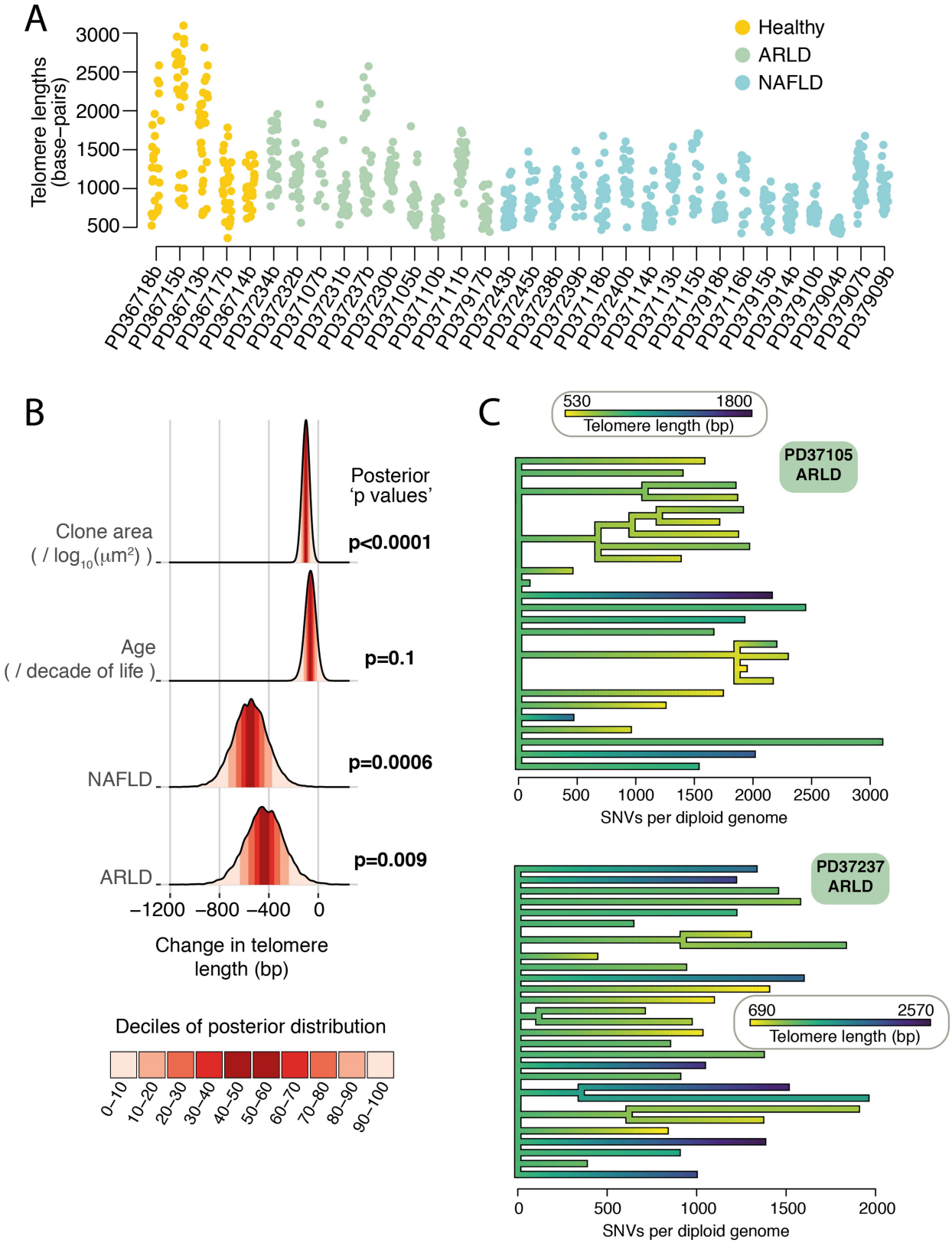


Extended Data Fig. 10 | *GPAM* mutations in patients with chronic liver disease. a, Distribution of somatic mutations in *GPAM* according to genomic location. Pie charts show fraction of sequencing reads reporting the mutant allele in each microdissection. **b**, Phylogenetic trees and clone maps are shown for a biopsy from PD37111 with *GPAM* mutations. The left panel shows the phylogenetic tree, with coloured branches showing independently acquired driver mutations. 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 < 100% of cells. The right panel shows the clones from the phylogenetic tree mapped onto an H&E-stained photomicrograph of the liver, with mutant clones coloured to match the tree.



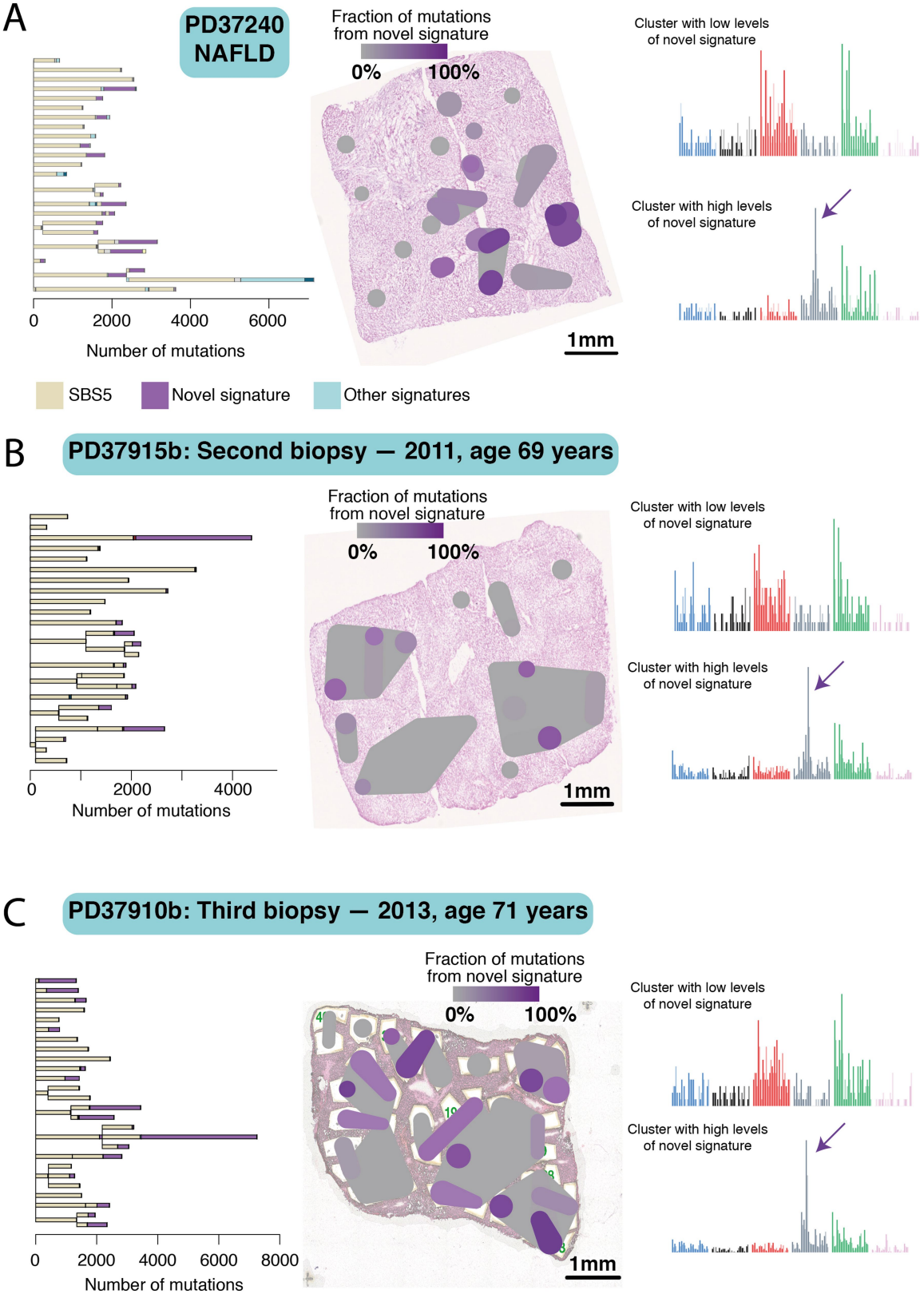
Extended Data Fig. 11 | See next page for caption.

Extended Data Fig. 11 | Properties of clones and patients with driver mutations. a, Stacked bar chart showing the estimated cumulative liver mass carrying driver mutations, extrapolated from samples analysed in each patient. The calculations assume a total liver mass of 1500g for each patient. Bars are coloured for each of the 6 recurrently mutated genes identified in the study, and patient codes on the x axis are coloured for disease status. **b**, Estimated clone size for the 4 most frequently mutated genes compared to wild-type clones. The points are overlaid on box-and-whisker plots where the median is marked with a heavy black line and the interquartile range in a thin black box. The whiskers denote mark the full range of the data or 25^{th}/75^{th} centile plus 1.5x the interquartile range (whichever is smaller). The p values are two-sided, derived from Wilcoxon rank-sum tests and have not been corrected for multiple hypothesis testing. Sample sizes are n = 25 mutant clones for *FOXO1*; n = 17 mutant clones for *CIDEB*; n = 15 mutant clones for *GPAM*; and n = 32 mutant clones for *ACVR2A*. **c**, Scatter plot showing the distribution of ages of patients in the cohort by whether they carried clones with mutations in the specified genes or not. The p values are two-sided, derived from Wilcoxon rank-sum tests and have not been corrected for multiple hypothesis testing. Sample sizes were n = 7 *FOXO1* mutant versus n = 22 *FOXO1* wild-type; n = 6 *CIDEB* mutant versus n = 23 *CIDEB* wild-type; and n = 7 *GPAM* mutant versus n = 22 *GPAM* wild-type. **d**, Stacked bar charts showing the proportion of patients with or without type 2 diabetes by whether they carried driver mutations in each gene. The p values are two-sided, derived from Fisher's exact tests and have not been corrected for multiple hypothesis testing. Sample sizes were as for **c**. **e**, Stacked bar charts showing the distribution of the NAFLD Activity Score (NAS) by whether they carried driver mutations in each gene, with low scores denoting a low degree of histological abnormality. The p values are two-sided, derived from chi-squared tests for trend and have not been corrected for multiple hypothesis testing. Sample sizes were as for **c**.



Extended Data Fig. 12 | Analysis of telomere lengths. a, Scatter plot showing the distribution of telomere lengths for samples grouped by disease status, and ranked from lowest to highest age within each disease category. **b**, Posterior distributions of the effect size of clone size (per $\log_{10}(\mu m^2)$), age (per decade of life) and disease state (NAFLD and ARLD versus normal) on telomere lengths. Density plots are shown from the MCMC sampler, coloured by decile. Posterior 'p values' are calculated from the posterior samples of the MCMC chain and are two-sided and not corrected for multiple hypothesis testing. **c**, Telomere lengths layered onto two representative phylogenetic trees from patients with ARLD. Branches are coloured on a yellow-to-blue scale according to telomere lengths of the sample with the highest VAF assigned to that branch. The internal nodes are estimated using maximum likelihood and colours are interpolated along each branch.

Extended Data Fig. 13 | Distribution of mutational signatures across the phylogenetic trees within the cohort. Estimated proportional contributions of each mutational signature to each phylogenetically defined cluster of somatic substitutions. Stacked bar plots show proportional contributions of signatures in normal controls (top), patients with ARLD (middle), and patients with NAFLD (bottom).



Extended Data Fig. 14 | Distribution of the new T>A signature across three samples. a, Signatures for a sample with high rates of the novel signature (PD37240). The left panel shows phylogenetic trees with each branch coloured by the proportion of mutations in that branch assigned to the different mutational signatures. The contribution from the new signature is coloured purple. The middle panel shows the overlay of clones onto an H&E-stained liver section. Clones are coloured on a grey-to-purple scale according to the proportion of mutations attributed to the novel signature. The right panel shows observed mutation spectra for representative clones with low (top) or high (bottom) burden of the novel signature, laid out as for Fig. 4b. Purple arrows indicate parts of the mutation spectrum that are characteristic of the new mutational signature. **b**, **c**, In one patient with NAFLD, we had three samples from 2008 (not shown as the signature was absent), 2011 (**b**) and 2013 (**c**), with the relative contribution of the signature increasing over time. The photomicrograph of the H&E section in **c** was captured after the microdissections were excised, hence the white gaps in the tissue.

nature research

Corresponding author(s): Campbell, Hoare

Last updated by author(s): Aug 12, 2021

Reporting Summary

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

Statistics

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

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

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

Software and code

Policy information about availability of computer code

Data collection	Image processing from sequencing data using standard Illumina X10 and NovaSeq pipeline
Data analysis	Alignment and variant calling performed using Sanger Institute's custom pipeline. Single-nucleotide substitutions were called using the CaVEMan (cancer variants through expectation maximization) algorithm (https://github.com/cancerit/CaVEMan). Small insertions and deletions were called using the Pindel algorithm (https://github.com/genome/pindel). Rearrangements were called using the BRASS (breakpoint via assembly) algorithm (https://github.com/cancerit/BRASS). List of programs and softwares: • R: version 3.5.1 • Perl: version 5.3.0 • Python: version 3.8.5 • MATLAB: version R2019b • BWA-MEM: version 0.7.17 (https://sourceforge.net/projects/bio-bwa/) • cgpCaVEMan: version 1.11.2/1.13.14/1.15.1 (https://github.com/cancerit/CaVEMan) • cgpPindel: version 2.2.2/2.2.4/2.2.5/3.2.0/3.3.0 (https://github.com/cancerit/cgpPindel) • Brass: version 5.4.1/6.0.5/6.1.2/6.2.0/6.3.4 (https://github.com/cancerit/BRASS) • ASCAT NGS: version 4.0.1/ 4.1.2/4.2.1 (https://github.com/cancerit/ascatNgs) • JBrowse: version 1.16.1 (https://jbrowse.org/) • cgpVAF: version 2.4.0 (https://github.com/cancerit/vafCorrect) • alleleCount: version 4.1.0 (https://github.com/cancerit/alleleCount) • SigProfiler: version 1.0.0-GRCh37 (https://github.com/AlexandrovLab) • HDP: version 0.1.5 (https://github.com/nicolaroberts/hdp) • dNdScv: version 0.0.1 (https://github.com/im3sanger/dndscv)

• Telomerecat: version 3.4.0 (https://github.com/jhrf/telomerecat)
• STAR: version 2.7.6a (https://github.com/alexdobin/STAR)
• Picard-tools: version 2.20.7 (https://broadinstitute.github.io/picard/)
• Samtools: version 1.12 (http://www.htslib.org/)
• TrimGalore: version 0.6.4 (https://github.com/FelixKrueger/TrimGalore)
• GATK: version 4.1.4.1 (https://gatk.broadinstitute.org/hc/en-us)
• GSEA: version 3.0 (https://www.gsea-msigdb.org/gsea/index.jsp)
• XGBoost: version 0.82.1 (https://xgboost.readthedocs.io/en/latest/)
• NDP.view2 (https://www.hamamatsu.com/eu/en/product/type/U12388-01/index.html)
• label.switching: version 1.8 (https://cran.r-project.org/web/packages/label.switching/index.html)
• philentropy: version 0.3.0 (https://cran.r-project.org/web/packages/philentropy/index.html)
• MCMCglmm: version 2.29 (https://cran.r-project.org/web/packages/MCMCglmm/index.html)
• Magick: version 2.0 (https://cran.r-project.org/web/packages/magick/index.html)
• Pheatmap: version 1.0.12 (https://cran.r-project.org/web/packages/pheatmap/index.html)
• Thermo Fisher software Tracefinder: version 5.0 (https://www.thermofisher.com/uk/en/home/industrial/mass-spectrometry/liquid-chromatography-mass-spectrometry-lc-ms/lc-ms-software/lc-ms-data-acquisition-software/tracefinder-software.html)
• CellProfiler: version 4.0.3 (https://cellprofiler.org/)
• PerkinElmer harmony: version 4.9 (https://www.perkinelmer.com/category/cellular-imaging-software)

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

Data

Policy information about availability of data

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

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

Whole genome sequencing data in the form of BAM files across samples reported in this study have been deposited in the European Genome-Phenome Archive (Accession number EGAD00001006255; https://www.ebi.ac.uk/ega/home). RNA-sequencing data has been deposited in the European Nucleotide Archive (Accession number ERP123192; https://www.ebi.ac.uk/ena/browser/home).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	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, 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.
Data exclusions	Samples with low mean coverage (<15x) were excluded due to the inaccuracy of mutation catalogues
Replication	Experiments for metabolomics and RNA-sequencing on cells transfected with wild-type or mutant FOXO1 constructs were performed with 5 replicates. Replicates showed consistent results.
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 prespecified.

Reporting for specific materials, systems and methods

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

Materials & experimental systems

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

Methods

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

Antibodies

Antibodies used

anti-β-Actin (clone: AC15) (Sigma, A5441, 1:5000, RRID:AB_476744); anti-Akt (clone: C73H10) (Cell Signaling, 2938, 1:1000, RRID:AB_915788); anti-phospho-Akt(T308) (clone: 244F9) (Cell Signaling, 4056, 1:1000, RRID:AB_331163); anti-GFP (Abcam, ab6556, 1:1000, RRID:AB_305564); anti-FOXO1 (clone: C29H4) (Cell Signaling, 2880, 1:1000, RRID:AB_2106495); anti-phospho-FOXO1 (T24) (Cell Signaling, (9464, 1:1000, RRID:AB_329842).

Validation

anti-β-Actin (clone: AC15): Validated by supplier with the following notes - species reactivity: pig, Hirudo medicinalis, bovine, rat, canine, feline, human, rabbit, carp, mouse, guinea pig, chicken, sheep; application(s): western blot: 1:5,000-1:10,000 using cultured human or chicken fibroblast cell extracts.
anti-Akt (clone: C73H10): Validated by supplier with the following notes - species reactivity: human, mouse, rat, monkey; application(s): suitable for western blot.
anti-phospho-Akt(T308) (clone: 244F9): Validated by supplier with the following notes - species reactivity: human, mouse, rat, monkey; application(s): suitable for western blot.
anti-GFP (Abcam, ab6556): Validated by supplier with following notes - species reactivity: independent; application(s): suitable for Suitable for: IHC-P, Electron Microscopy, ICC, IP, Flow Cyt, IHC-Fr, western blot
anti-FOXO1 (clone: C29H4): Validated by supplier with the following notes - species reactivity: human, mouse, rat, monkey; application(s): suitable for western blot.
anti-phospho-FOXO1 (T24): Validated by supplier with the following notes - species reactivity: human, mouse, rat, monkey; application(s): suitable for western blot.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	The 3 HCC cell lines (HepG2, Hep3B and PLC/PRF/5) were all obtained from ATCC.
Authentication	Identification confirmed by SNP genotyping
Mycoplasma contamination	All cell lines were confirmed as Mycoplasma negative
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used in this study.

Human research participants

Policy information about studies involving human research participants

Population characteristics	The dataset comprised 1590 genomes from 34 liver samples, including 5 normal liver controls with no prior neoadjuvant therapy, 10 with alcohol-related liver disease (ARLD) and 19 with NAFLD (Supplementary Table S1). All patients with ARLD or NAFLD had HCC, liver failure or both and tissues were derived from hepatic resection or transplantation. Overall, 9 samples were from patients who had a synchronous HCC and underlying cirrhosis; a further 8 samples had HCC without underlying cirrhosis, including 3 hepatic resection samples from one patient over a 5-year timespan (Extended Figure 1). All samples underwent central histological review by specialist hepatopathologists, and the histological and clinical features of the patients matched those expected for the underlying disease processes (Supplementary Table S1). The average age of research subjects was 61 years, and the male:female split was 29:5.
Recruitment	Recruited through Addenbrooke's Hospital, Cambridge, UK. All patient gave written informed consent, and were typically of advanced stage liver disease. Because explanted liver samples were mostly used, there is a recruitment bias towards high severity of disease.
Ethics oversight	East of England Research Ethics Committee: 15/EE/0351 and 16/NI/0196

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