A body map of somatic mutagenesis in morphologically normal human tissues

https://doi.org/10.1038/s41586-021-03836-1

Received: 29 November 2020

Accepted: 20 July 2021

Published online: 25 August 2021

Check for updates

Ruoyan Li[1,2,14], Lin Di[1,2,14], Jie Li[3,4,14], Wenyi Fan[5,14], Yachen Liu[5], Wenjia Guo[5], Weiling Liu[5], Lu Liu[1,2], Qiong Li[3,4], Liping Chen[5], Yamei Chen[5], Chuanwang Miao[5], Hongjin Liu[5], Yuqian Wang[5], Yuling Ma[5], Deshu Xu[1,2], Dongxin Lin[5,6,7,8,15]✉, Yanyi Huang[1,2,9,10,11,15]✉, Jianbin Wang[3,4,15]✉, Fan Bai[1,2,12,15]✉ & Chen Wu[5,6,8,13,15]✉

Somatic mutations that accumulate in normal tissues are associated with ageing and disease[1,2]. Here we performed a comprehensive genomic analysis of 1,737 morphologically normal tissue biopsies of 9 organs from 5 donors. We found that somatic mutation accumulations and clonal expansions were widespread, although to variable extents, in morphologically normal human tissues. Somatic copy number alterations were rarely detected, except for in tissues from the oesophagus and cardia. Endogenous mutational processes with the SBS1 and SBS5 mutational signatures are ubiquitous among normal tissues, although they exhibit different relative activities. Exogenous mutational processes operate in multiple tissues from the same donor. We reconstructed the spatial somatic clonal architecture with sub-millimetre resolution. In the oesophagus and cardia, macroscopic somatic clones that expanded to hundreds of micrometres were frequently seen, whereas in tissues such as the colon, rectum and duodenum, somatic clones were microscopic in size and evolved independently, possibly restricted by local tissue microstructures. Our study depicts a body map of somatic mutations and clonal expansions from the same individual.

Somatic mutations occur naturally in normal cells during cell division. Studies have revealed the somatic mutation landscape of different human tissues, including the skin[3,4], oesophagus[5,6], colon and rectum[7], liver[8], endometrial epithelium[9], bronchus[10], brain[11,12], embryo[13], urothelium[14,15] and blood cells[16,17], mostly through deep DNA sequencing of biopsied tissue samples. Other studies have implemented bioinformatic algorithms to detect somatic mutations from RNA-sequencing data of normal tissues[18,19]. Although these studies have contributed greatly to our knowledge of mutation rates, driver genes and mutagenic factors in different normal tissues from human organs, the tissue samples that they analysed usually came from different donors with distinct germline backgrounds and life histories, thus making a cross-organ comparison challenging. Ideally, for such comparisons, we should analyse normal tissues collected from the same individual. Here we combined laser-capture microdissection (LCM) and mini-bulk exome sequencing to systematically investigate somatic mutagenesis in morphologically normal tissues collected from nine anatomic sites of autopsy samples from five donors.

Sequencing and somatic mutations

In 5 deceased organ donors (PN1, PN2, PN7, PN8 and PN9) aged between 85 and 93, we collected approximately 1,800 microbiopsies from 9 anatomic sites of autopsy samples (Fig. 1a, Supplementary Table 1), which included morphologically normal epithelia from the bronchus, oesophagus, cardia (the junction between the lower oesophagus and the stomach), stomach, duodenum, colon and rectum, and normal parenchyma from the liver and pancreas (Extended Data Fig. 1). We applied a consistent sampling strategy: five layers were sectioned from each tissue (with a 200-μm interval between layers). Within each layer, 10 microbiopsies with approximate 600 cells in each were densely collected using LCM. In tissues such as the colon and rectum, multiple tissue microstructures (that is, crypts) were dissected into single samples to enable our dense sampling strategy with a fixed number of cells in each sample. While assuring quality control, we subjected 1,762 biopsies to whole-exome sequencing (WES), excluding 25 samples with a less than 10-fold average coverage depth from further analysis. The remaining 1,737 samples had an average sequencing depth of 56-fold

[1] Biomedical Pioneering Innovation Center (BIOPIC), School of Life Sciences, Peking University (PKU), Beijing, China. [2] Beijing Advanced Innovation Center for Genomics (ICG), Peking University, Beijing, China. [3] School of Life Sciences, Tsinghua-Peking Center for Life Sciences, Tsinghua University, Beijing, China. [4] Beijing Advanced Innovation Center for Structural Biology (ICSB), Tsinghua University, Beijing, China. [5] Department of Etiology and Carcinogenesis, National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences (CAMS) and Peking Union Medical College (PUMC), Beijing, China. [6] Collaborative Innovation Center for Cancer Personalized Medicine, Nanjing Medical University, Nanjing, China. [7] Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Guangzhou, China. [8] CAMS Key Laboratory of Cancer Genomic Biology, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [9] College of Chemistry and Molecular Engineering, Beijing National Laboratory for Molecular Sciences, Peking University, Beijing, China. [10] Peking-Tsinghua Center for Life Sciences, Peking University, Beijing, China. [11] Institute for Cell Analysis, Shenzhen Bay Laboratory, Guangdong, China. [12] Center for Translational Cancer Research, Peking University First Hospital, Beijing, China. [13] CAMS Oxford Institute (COI), CAMS, Beijing, China. [14] These authors contributed equally: Ruoyan Li, Lin Di, Jie Li, Wenyi Fan. [15] These authors jointly supervised this work: Dongxin Lin, Yanyi Huang, Jianbin Wang, Fan Bai, Chen Wu. ✉e-mail: lindx@cicams.ac.cn; yanyi@pku.edu.cn; jianbinwang@tsinghua.edu.cn; fbai@pku.edu.cn; chenwu@cicams.ac.cn

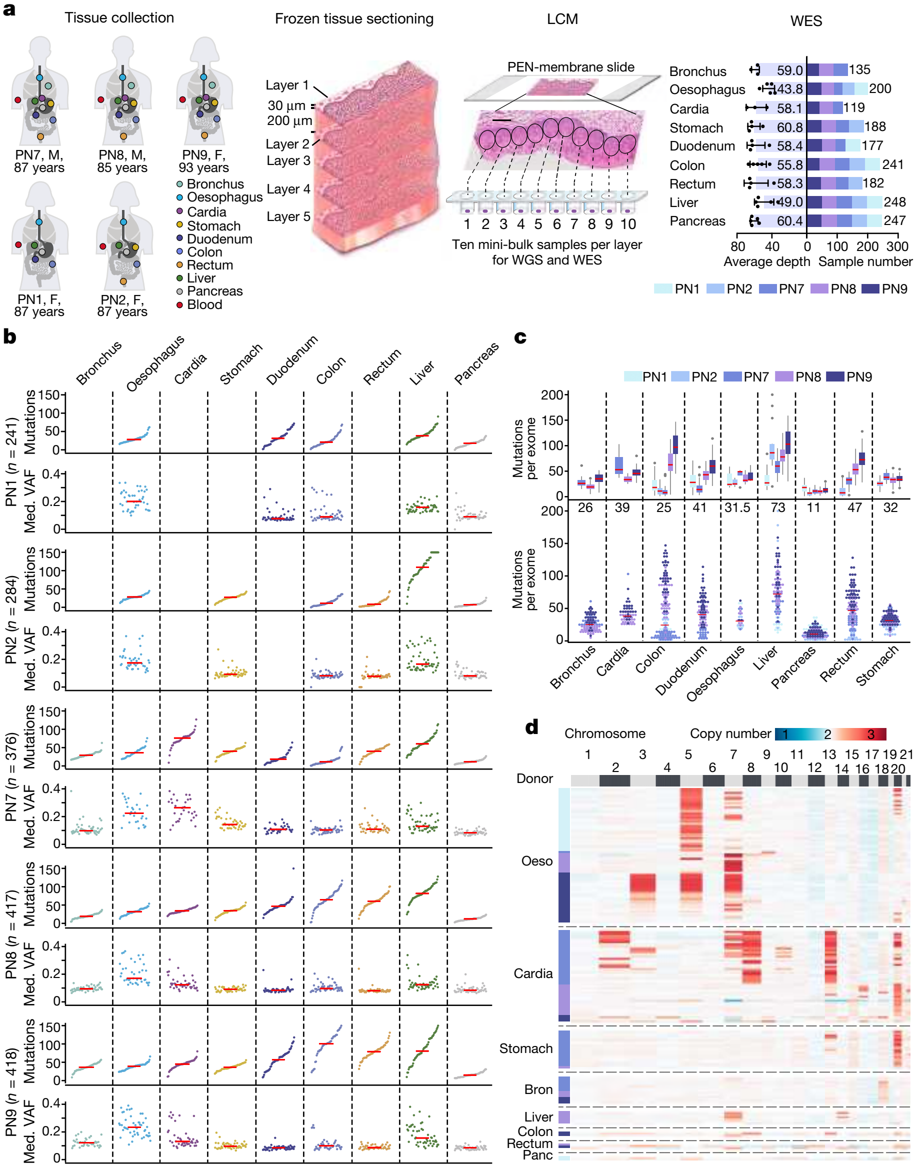


Fig. 1 | Research strategy and summary of genomic alterations detected in normal tissues from five donors. a, LCM and mini-bulk exome sequencing procedure. PEN, polyethylene naphthalate. Scale bar, 400 µm. **b**, Top, number of somatic mutations detected in the coding regions in tissue biopsies. Bottom, median (med.) VAF in tissue biopsies. Each biopsy sample is represented by a coloured dot. Red bars represent medians. **c**, The mutation burden in samples with median VAFs between 0.08 and 0.14. Top, box plots showing the mutation burden in organs from different donors. The lower edge, upper edge and centre of the box represent the 25th (Q1) percentile, 75th (Q3) percentile and the median, respectively. The interquartile range (IQR) is Q3 − Q1. Outliers are values beyond the whiskers (upper, Q3 + 1.5 × IQR; lower, Q1 − 1.5 × IQR). Detailed information about the box plots can be found in Supplementary Table 3. Bottom, dot plots showing the mutation burdens in different organs. The medians are labelled and represented by red bars. **d**, Heat map showing somatic CNAs in biopsy samples. Donor information corresponds to **c**. Oeso, oesophagus; bron, bronchus; panc, pancreas.

(Fig. 1a, Supplementary Table 2). We performed somatic alteration identification using DNA from the peripheral blood cells of each donor as the germline comparators.

Overall, 53,592 unique single-nucleotide variations (SNVs) and 444 small insertions and deletions (indels) were identified (Supplementary Table 3). We performed whole-genome sequencing (WGS) of 43 selected samples with WES data to validate the somatic mutations that were called from WES. The average validation rate was 92.5% (Extended Data Fig. 2a, Supplementary Table 4). Sensitivity corrections were made to the numbers of detected mutations (Extended Data Fig. 2b, c). The numbers of detected somatic mutations and distributions of variant allele frequency (VAF) varied greatly across tissues and donors (Fig. 1b, Extended Data Fig. 2d). In tissues without clear physical microstructures—such as the oesophagus, liver and bronchus—we observed a median VAF of 0.21 in the oesophagus (range 0.1–0.43), 0.14 in the liver (range 0.08–0.38) and 0.1 in the bronchus (range 0.07–0.38), suggesting that there were usually multiple clones in the biopsies and that the degree of clonal expansion in these tissues was different (Fig. 1b). In tissues with clear physical microstructures—such as the colon and rectum—we observed a median VAF of 0.09 in the colon (range 0.06–0.3) and 0.09 in the rectum (range 0.06–0.26), indicating that multiple tissue microstructures were dissected into single samples, thus making these samples polyclonal, although each microstructure was theoretically monoclonal (Fig. 1b).

An interdependence analysis between the VAF distributions and numbers of mutations was performed within and across organs (Extended Data Fig. 3). As the number of detected mutations may be influenced by the clonality of our biopsy samples, we only included samples with comparable median VAFs (between 0.08 and 0.14) for a cross-tissue analysis of mutation burden. Pancreas parenchyma contained the fewest mutations (median and adjusted median: 11 and 12 per exome, respectively), whereas the number of mutations in the liver (median and adjusted median: 73 and 76 per exome) was the greatest among all tissues—substantially higher than the number of mutations in epithelial cells from other organs. Normal epithelial tissues from cardia (median and adjusted median: 39 and 41 per exome, respectively), rectum (median and adjusted median: 47 and 54 per exome) and duodenum (median and adjusted median: 41 and 44 per exome) had higher numbers of mutations (Fig. 1c, Extended Data Fig. 2c). Notably, the number of mutations tended to decrease in highly expressed genes in different tissues (Extended Data Fig. 4a), implying that transcription-coupled repair is more active in highly expressed genes. We also observed varying types of somatic mutations across organs (Extended Data Fig. 2d), which may reflect different underlying mutational processes.

Somatic copy number alterations

We assessed somatic copy number alterations (CNAs) in normal tissues by subjecting 1,764 biopsies to low-depth WGS (Supplementary Table 4). Overall, we observed diploid genomes in most (1,608 out of 1,764; 91.2%) normal samples (Extended Data Fig. 4b, Supplementary Table 4). Sporadic CNA events could be detected in a number of samples (Fig. 1d). Of note, the samples with CNAs exhibited strong organ preferences. Normal oesophageal tissues (31 out of 41 (75.6%) from PN1, 10 out of 41 (24.4%) from PN8 and 24 out of 50 (48.0%) from PN9; $P < 10^{-25}$, hypergeometric test) were found to contain CNAs that were enriched as whole-chromosomal amplifications of chromosomes 3, 5 and 7. Previous studies[6] have reported an occasional amplification of chromosome 3 in normal oesophagus, but not amplifications of chromosomes 5 and 7. In addition, in some cardia samples (27 out of 34 (79.4%) from PN7 and 15 out of 44 (34.1%) from PN8), we detected CNAs that exhibited whole-chromosomal amplifications of chromosomes 2, 7, 8, 13 and 20, which, to our knowledge, have not previously been reported.

Mutational signatures in normal tissues

Through clustering the trinucleotide spectra of somatic mutations, we found that most samples tended to cluster independent of their tissue of origin, whereas some liver samples—mainly from donors PN1 and PN2—distributed separately from the main cluster (Extended Data Fig. 5a, b). Notably, most rectum, colon, stomach, cardia and duodenum samples tended to cluster together ($P < 1 \times 10^{-40}$, hypergeometric test), suggesting that some common major mutational processes actively operate in these tissues (Extended Data Fig. 5b). To further examine the underlying mutational processes, we performed de novo single-base-substitution (SBS) mutational signature extraction based on a Bayesian hierarchical Dirichlet process[7,8] (Methods). In total, we deciphered seven mutational signatures (signatures A to G), each of which mostly conformed to the Catalogue of Somatic Mutations in Cancer (COSMIC) mutational signatures SBS5, SBS1, SBS22, SBS4, SBS45, SBS13 and SBS2, respectively[20,21] (Extended Data Fig. 5c, d, Supplementary Tables 5–7).

We found two age-related endogenous mutational signatures[22], SBS1 and SBS5, throughout all normal samples across organs and donors (Fig. 2a, Extended Data Fig. 5e). The relative activities of SBS1 and SBS5 varied across tissues but exhibited a conserved tissue-specific pattern among donors. The duodenum, colon and rectum showed higher SBS1/SBS5 ratios compared with the bronchus, pancreas, oesophagus and liver (Fig. 2b). The preference of SBS1 and SBS5 mutations has been compared among various cancer types[22], but we report the SBS1/SBS5 ratios in normal tissues from different organs. Two other endogenous mutational signatures, SBS2 and SBS13, which have been associated with the activity of APOBEC cytidine deaminases[23], emerged sporadically among normal samples (Fig. 2a).

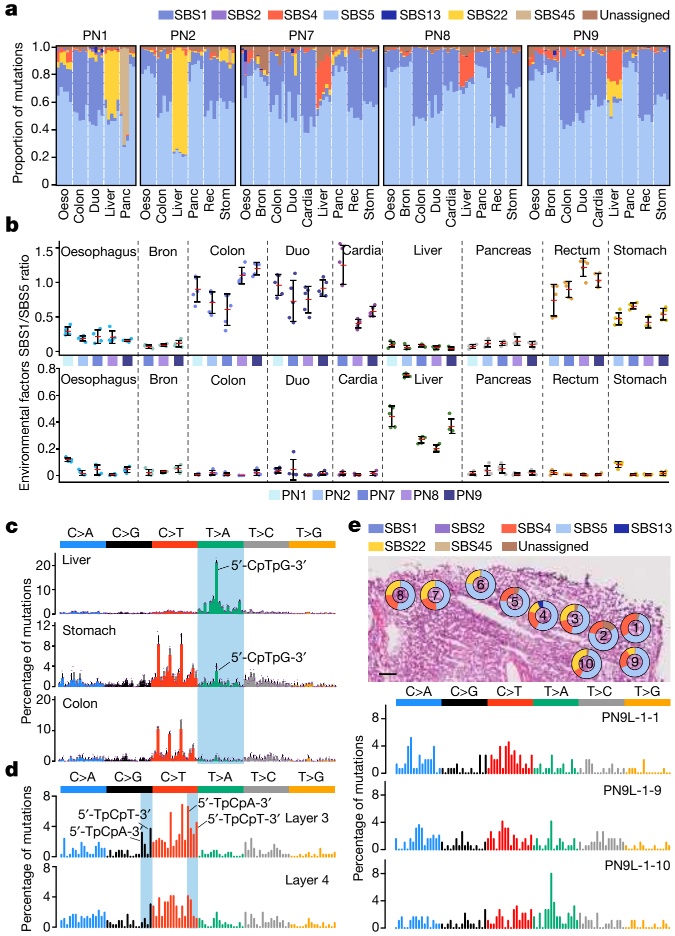


Fig. 2 | Mutational signatures in normal tissues from five donors. a, Stacked bar plots show the proportional contributions of mutational signatures in organs from the five donors. Each stacked bar represents one tissue layer. Duo, duodenum; rec, rectum; stom, stomach. **b**, Top, the ratio of SBS1 to SBS5 across different organs. Bottom, the summed contributions of SBS4 and SBS22 (environmental factors) across different organs. All tissues are with $n = 5$ data points except for oesophagus from donor PN7 ($n = 4$) and bronchus from donor PN8 ($n = 4$). Data are mean ± s.d. Abbreviations of tissues correspond to those in **a**. **c**, Trinucleotide mutational spectra of liver, stomach and colon tissues from donor PN2. Purple dots represent data points of the five tissue layers. Data are mean ± s.d. Typical aristolochic-acid-associated mutational features are shaded in blue. **d**, Trinucleotide mutational spectra of two dissected layers of oesophagus from donor PN7. Typical APOBEC-associated mutational features are shaded in blue. **e**, Top, haematoxylin and eosin (H&E)-stained liver tissue (PN9 layer 1) with superimposed donut charts showing the proportional contributions of mutational signatures, as estimated by deconstructSigs. Scale bar, 200 μm. Bottom, trinucleotide mutational spectra of three samples. Nomenclature of sample IDs: for example, PN9L-1-1 represents the no. 1 liver sample dissected from tissue layer 1 from donor PN9.

We identified two exogenous mutational signatures (SBS4 and SBS22) that exhibited strong transcriptional strand asymmetries (Extended Data Fig. 6a, b). We observed SBS4—the mutational signature that is associated with tobacco smoking—mostly in liver samples but weakly in some bronchus and oesophagus samples (Fig. 2a), consistent with previous findings[5,8,10]. SBS22, for which the underlying aetiological factor is exposure to aristolochic acid, exhibited notable activity in liver samples from three donors (Fig. 2a, Extended Data Fig. 5e). Aristolochic acid mutagenesis has been extensively implicated in liver and bladder cancers in Asian individuals[24–26] and has also been reported in alcohol-related liver disease and normal urothelium[8,15]. Notably, the three donors with obvious SBS22 activity in our study were female (Supplementary Table 1). A sex bias in aristolochic acid mutagenesis has been reported in upper tract urothelial carcinoma, but the underlying mechanism remains unclear[27]. We consistently observed that the liver tissues contained more somatic mutations caused by exogenous mutational processes, suggesting that the liver has a higher risk of exposure to environmental carcinogens than do other organs (Fig. 2b).

Our sampling strategy enabled us to compare mutational signatures across different organs of the same donor, with the assumption that such samples were influenced by the same life history. The potential aristolochic acid exposure history of donors PN1 and PN2 has been found to contribute to the mutagenesis in the liver. Within those donors, we found potential aristolochic acid mutagenesis in other organs, such as the oesophagus and duodenum of donor PN1 and the stomach of donor PN2 (Fig. 2a, c, Extended Data Fig. 6c). To our knowledge, there have been no previous reports of aristolochic acid mutagenesis in normal or cancerous stomach, duodenum or oesophagus tissues.

We observed a marked difference in the mutational spectra between two tissue layers in the oesophagus of donor PN7, with a noticeable APOBEC-associated mutational process in layer 3 but not in layer 4 (Fig. 2d, Extended Data Fig. 7a; $P < 10^{-3}$, multinomial test by Monte-Carlo simulations). Similar intra-tissue heterogeneity of mutational signatures was also exemplified by SBS22 in duodenum samples from donor PN7 (Extended Data Fig. 7b; $P < 10^{-4}$, multinomial test by Monte-Carlo simulations). We also found considerable differences in the mutational spectra and relative activities of SBS4 and SBS22, even between adjacent LCM biopsies (Fig. 2e, Extended Data Fig. 7a, c; $P < 10^{-4}$, multinomial test by Monte-Carlo simulations). This regional variation (both between and within tissue layers) in mutational signature activity may reflect regional activations of different mutagenic driving factors.

Landscape of driver mutations

To identify potential driver genes in normal tissues, we applied the dNdScv algorithm to all detected somatic mutations[28]. With hypothesis testing applied to all coding genes and 126 driver gene candidates (Methods, Supplementary Table 8), we identified 32 potential driver genes, including canonical cancer drivers such as *NOTCH1*, *TP53*, *ARID1A* and *ERBB2* (Fig. 3a, Extended Data Fig. 8a). These 32 genes recapitulated signalling pathways that have been widely implicated in tumorigenesis (Extended Data Fig. 8b, c). In addition, we identified 19 cancer hotspot mutations in 8 driver genes, with the greatest number of hotspot mutations being detected in *TP53* (9 hotspots out of 18 mutations) (Extended Data Fig. 8d, Supplementary Table 9).

The proportion of samples with driver mutations varied across organs. We detected driver mutations in 6.5% (ranging from 2% in PN9 to 12.2% in PN1) of the pancreas parenchyma samples, but 73.8% (ranging from 67.5% in PN2 to 81.6% in PN9) of the oesophageal samples contained at least one driver mutation ($P = 0.0193$, two-sided Wilcoxon rank-sum test) and about 11% (ranging from 4.6% in PN1 to 24.5% in PN9) contained more than three driver mutations (Fig. 3b). On the other hand, mutations in genes such as *ARID1A*, *TP53*, *NOTCH1* and *FAT1* were often shared by multiple samples (Extended Data Fig. 8e), which implies that mutations in those genes were more likely to drive clonal expansions.

Mutations in the 32 potential driver genes were distributed heterogeneously across organs and donors (Fig. 3c). *NOTCH1* was found to be the most frequently mutated gene (65 unique non-silent mutations in 101 samples) (Fig. 3a, c). *NOTCH1* and *TP53* mutations, although widely observed across organs, showed enrichments in oesophageal tissues (ratio of observation to expectation ($R_{O/E}$) = 2.87 and 2.92, respectively; $P < 10^{-4}$, hypergeometric test) (Extended Data Fig. 8f, Supplementary Table 10). *MUC6* was identified as a driver gene that is enriched in normal cardia and stomach ($R_{O/E}$ = 6.59 and 2.14, respectively; $P < 10^{-4}$ and $P = 0.074$) (Extended Data Fig. 8f). It has also been reported as a driver gene with specificity in stomach adenocarcinoma[28–30]. This tissue-specific correspondence of driver genes in cancerous and normal tissues is noteworthy. The prevalence of *MUC6* mutations in normal gastric tissues (cardia and stomach) was significantly higher than that in gastric cancers (Extended Data Fig. 8g; adjusted $P < 10^{-7}$, Fisher's exact test), suggesting that different molecular mechanisms underlie the clonal evolution of normal cells versus that of cancer cells. Similarly, *KMT2D* mutations occurred preferentially in liver tissues ($R_{O/E}$ = 2.76; $P = 0.0003$, hypergeometric test), *ERBB3* mutations occurred preferentially in rectal tissues ($R_{O/E}$ = 3.62; $P = 0.0001$, hypergeometric test) and *SMARCA4* mutations were enriched in duodenal tissues ($R_{O/E}$ = 6.12, $P = 0.005$, hypergeometric test) (Extended Data Fig. 8f, Supplementary Table 10). We also observed heterogeneous driver mutation occurrences between individual donors. For example, four unique *PTCH1* mutations were found in liver samples from donor PN8, but none were found in liver samples from the other four donors (Fig. 3c; $P = 6 \times 10^{-5}$, hypergeometric test).

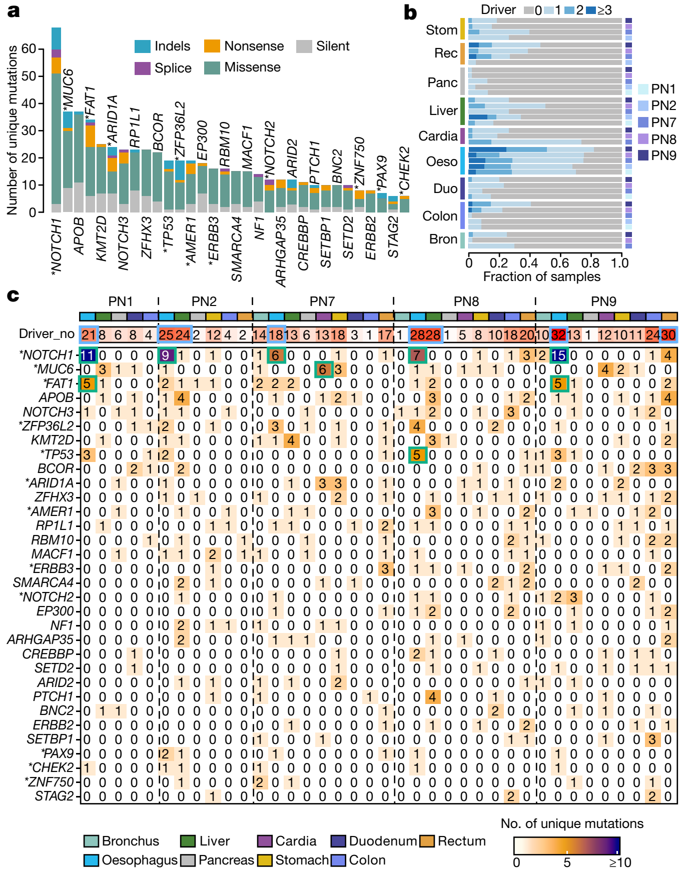


Fig. 3 | Mutational landscape of driver genes across organs. a, Stacked bar plot showing the number of unique mutations in the 32 driver genes. Asterisks indicate genes that are significant ($q < 0.1$) in dNdScv analysis. **b**, Stacked bar plot showing the fraction of biopsy samples with zero, one, two, or three or more driver mutations across the five donors. **c**, Heat map showing the number of non-silent unique mutations in the 32 driver genes across different organs of the 5 donors. Blue boxes indicate organs with high numbers of driver mutations in each donor. Green boxes indicate representative driver genes with high numbers of mutations in different organs and donors. Asterisks indicate genes that are significant ($q < 0.1$) in dNdScv analysis.

Spatial architecture of mutant clones

We investigated how the accumulation of mutations and the expansion of mutant clones are coordinated in normal tissues. For each donor, we plotted

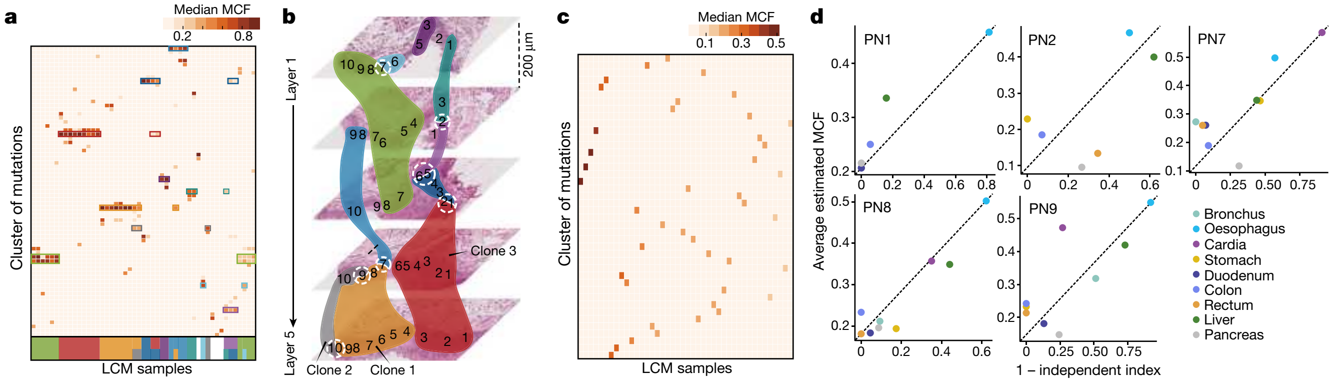


Fig. 4 | Estimation of somatic mutant clonal sizes and construction of spatial clonal expansion maps. a, Heat map showing mutation clustering in biopsy samples in the oesophagus of donor PN9. Each cluster contains mutations with similar MCFs. Different colours are used to indicate clonal sharing events among biopsy samples. **b**, Spatial clonal architecture of oesophagus tissues from donor PN9. The numbers in each layer represent the positions of LCM samples. The overlaid colours correspond to **a** and indicate the ranges of clonal expansions. Intermingling of different clones in single biopsy samples is highlighted using white dashed circles. **c**, Heat map showing the mutation clustering in colon samples from donor PN9. **d**, Correlations of average MCFs (estimated using a Bayesian Dirichlet process) with 1 minus independent index (Methods) of the organs among the five donors.

the distribution of the number of mutations versus the average mutant cell fraction (MCF) of each sample (Extended Data Fig. 9). In the oesophagus and cardia, mutant clones tended to be large, but the numbers of mutations were relatively low. By contrast, normal colonic and rectal tissues accumulated many mutations, although the degree of clonal expansion was low on the spatial scale. In the liver, some samples simultaneously had high mutational burdens and showed substantial clonal expansions.

We observed two major scenarios of somatic clonal evolution across organs: (1) a single mutant clone expands to a macroscopic scale; and (2) competitive mutant clones originate and evolve independently. In the oesophagus of donor PN9, we identified large-scale clonal expansions that covered two to more than ten LCM biopsies and spread across two to three layers (Fig. 4a, b, Extended Data Fig. 10a). Similar circumstances were observed in the oesophagus of other donors (Extended Data Fig. 10b). Mutations such as those that occurred in *NOTCH1*, *TP53* and *ARID1A* may have driven the mutant clonal expansions in the oesophagus of donor PN9 (Extended Data Fig. 10c). For example, clone 1 contained a *NOTCH1* mutation (p.A348D) and expanded to intermix with adjacent clone 2, which had a *FAT1* mutation (p.E3124*). Samples in clone 3 shared no driver mutations but ubiquitously carried CNAs in chromosomes 3, 5 and 7. This potential CNA-driven early clonal expansion in normal oesophagus has not, to our knowledge, been previously reported. The degree of mutant clonal expansion in the liver was comparable to that in the oesophagus, but with fewer driver events (Fig. 3c, Extended Data Fig. 11a, b). By stark contrast, colon samples were found to evolve as independent mutant clones (Fig. 4c, Extended Data Fig. 12a). These two typical scenarios were also observed in samples from other donors (Extended Data Figs. 10–12).

We went on to calculate an independent index for each tissue, which is the ratio of the number of samples that do not share any mutation clusters with others divided by the total number of samples with at least one mutation cluster (Methods). We defined an elevated clonal expansion on the spatial scale as having a low independent index but a high average MCF. Clonal expansions in the oesophagus, cardia and liver tended to be larger, whereas the colon, rectum, and duodenum—which are constrained by tissue physical microstructures—exhibited low degrees of clonal expansions on the spatial scale (Fig. 4d). Mutant clones in the cardia and stomach tissues from donor PN7 were substantially expanded (Extended Data Fig. 11c), which could be associated with the recurrent CNAs that they contained (Fig. 1d). In donor PN2, the degree of clonal expansion in the liver was large and comparable to that in the oesophagus. This could be related to the overwhelming levels of aristolochic acid mutagenesis in the liver tissues of donor PN2 (Fig. 2a).

Discussion

In this study, we used LCM and mini-bulk WES to characterize somatic mutations and clonal expansions in samples from nine anatomic sites that were collected from five donors. We identified varying mutation burdens, CNAs, mutational signatures and degrees of clonal expansion across normal human tissues (Supplementary Discussion). Comparing tissue samples across organs from the same individual could potentially offset the systematic bias that is introduced by differing ages, germline backgrounds or lifestyles when tissue samples from different individuals are compared. For example, we compared the relative activity of SBS1 and SBS5 across different tissues from the same individual.

Our study encompassed various tissue types both with and without physical microstructures, which posed challenges to our LCM experiments (Supplementary Discussion). We kept a consistent sampling strategy across tissues by dissecting a fixed number of cells in each sample, which brought multiple microstructures (for example, crypts) into single samples. Although this research strategy has clear benefits, it could have influenced the estimation of mutation burdens and rates in different tissues and cells and caused the discrepancy between mutation burdens (for example, the mutation burden in colon and rectum) that is reported in our study and other studies[7,31,32]. Comparing the mutation rate of stem cells between tissues with and without physical microstructures remains a challenge, which could potentially be addressed using the nanorate sequencing method in the future[33].

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/s41586-021-03836-1.

1. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
2. Risques, R. A. & Kennedy, S. R. Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet.* **14**, e1007108 (2018).
3. Martincorena, I. *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
4. Tang, J. et al. The genomic landscapes of individual melanocytes from human skin. *Nature* **586**, 600–605 (2020).
5. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).

6. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
7. Lee-Six, H. et al. The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
8. Brunner, S. F. et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. *Nature* **574**, 538–542 (2019).
9. Moore, L. et al. The mutational landscape of normal human endometrial epithelium. *Nature* **580**, 640–646 (2020).
10. Yoshida, K. et al. Tobacco smoking and somatic mutations in human bronchial epithelium. *Nature* **578**, 266–272 (2020).
11. Lodato, M. A. et al. Aging and neurodegeneration are associated with increased mutations in single human neurons. *Science* **359**, 555–559 (2018).
12. Bae, T. et al. Different mutational rates and mechanisms in human cells at pregastrulation and neurogenesis. *Science* **359**, 550–555 (2018).
13. Ju, Y. S. et al. Somatic mutations reveal asymmetric cellular dynamics in the early human embryo. *Nature* **543**, 714–718 (2017).
14. Lawson, A. R. J. et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. *Science* **370**, 75–82 (2020).
15. Li, R. et al. Macroscopic somatic clonal expansion in morphologically normal human urothelium. *Science* **370**, 82–89 (2020).
16. Lee-Six, H. et al. Population dynamics of normal human blood inferred from somatic mutations. *Nature* **561**, 473–478 (2018).
17. Watson, C. J. et al. The evolutionary dynamics and fitness landscape of clonal hematopoiesis. *Science* **367**, 1449–1454 (2020).
18. Yizhak, K. et al. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* **364**, eaaw0726 (2019).
19. Garcia-Nieto, P. E., Morrison, A. J. & Fraser, H. B. The somatic mutation landscape of the human body. *Genome Biol*. **20**, 298 (2019).
20. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
21. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
22. Alexandrov, L. B. et al. Clock-like mutational processes in human somatic cells. *Nat. Genet*. **47**, 1402–1407 (2015).
23. Nik-Zainal, S. et al. Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
24. Poon, S. L. et al. Mutation signatures implicate aristolochic acid in bladder cancer development. *Genome Med*. **7**, 38 (2015).
25. Ng, A. W. T. et al. Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. *Sci. Transl. Med*. **9**, eaan6446 (2017).
26. Du, Y. et al. Mutagenic factors and complex clonal relationship of multifocal urothelial cell carcinoma. *Eur. Urol*. **71**, 841–843 (2017).
27. Chen, C. H. et al. Aristolochic acid-induced upper tract urothelial carcinoma in Taiwan: clinical characteristics and outcomes. *Int. J. Cancer* **133**, 14–20 (2013).
28. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
29. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of gastric adenocarcinoma. *Nature* **513**, 202–209 (2014).
30. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **174**, 1034–1035 (2018).
31. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature*, https://doi.org/10.1038/s41586-021-03822-7 (2021).
32. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
33. Abascal, F. et al. Somatic mutation landscapes at single-molecule resolution. *Nature* **593**, 405–410 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Methods

Data reporting

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

Ethics statement and sample collection

The protocol and informed consent documents of this study were reviewed and approved by the National Cancer Center/Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College Ethics Committee (NCCEC, reference number 20/069-2265). Written informed consent was obtained from all donors. We obtained normal tissue samples from five deceased organ donors who had been recruited at the Body Donation Registration and Receiving Station in Peking Union Medical College, Beijing. None of the donors had undertaken neoadjuvant systemic therapy. Within 16 h of death, we separately collected tissues (lengths ranging from 1 to 5 cm) from nine organs (bronchus, oesophagus, cardia, stomach, duodenum, colon, rectum, liver and pancreas) from each donor. We then opened luminal organs longitudinally and cut each into approximately 0.5 × 0.5-cm pieces. All tissue samples were snap-frozen in liquid nitrogen and stored at −80 °C. The clinical and pathological characteristics of each donor are summarized in Supplementary Table 1.

Histopathological examination

We fixed the tissues in 10% buffered formalin and embedded them in paraffin blocks. Then the formalin-fixed paraffin-embedded (FFPE) tissues were sectioned into 3-μm-thick sections. The specimens were stained with H&E and analysed under light microscopy. Three pathologists independently examined the morphological and histological features of the tissues. The slides were imaged using the Vectra Polaris Automated Quantitative Pathology Imaging System (Perkin Elmer).

Preparation of tissue sections

We embedded tissue sections in optimal cutting temperature (OCT) medium (Thermo Fisher Scientific) at −25 °C. A total of five layers were cut at a thickness of 30 μm using a Leica cryotome, with a 200-μm gap between each layer. We transferred each section to a polyethylene naphthalate membrane slide (Thermo Fisher Scientific) and then incubated the slides in cresyl violet acetate for 1 min and rinsed them twice in water. The remaining, unmounted tissues were used for immunohistochemistry and immunofluorescence analyses.

Laser-capture microdissection

We used an LMD7000 laser microdissection microscope (Leica Microsystems) with 10× magnification and proper laser settings to microdissect the mounted and stained tissues from the previous section. Tissue layers with a target size of 0.06 mm^2, which corresponded to about 600 cells, were microdissected. We placed each 600-cell microdissected isolate into an empty cap of a nuclease-free 0.2-ml Axygen PCR tube (Thermo Fisher Scientific). We took photomicrographs both before and after LCM.

Whole-genome library preparation and sequencing

We lysed the LCM samples using a low-temperature protocol with cold-active protease to reduce DNA-base oxidative deamination, thus eliminating artifacts in somatic mutation calling. Specifically, each biopsy sample was lysed in 8 μl customized lysis buffer (15 μg μl^{-1} native *Bacillus licheniformis* protease (Creative EnzymesNATE-0633), 30 mM Tris-HCl (pH 7.6, Rockland Immunochemicals, MB-003), 10 mM NaCl (Ambion, AM9760G), 5 mM EDTA (Ambion, AM9260G), 0.4% Triton X-100 (Sigma, T9284)) at 6 °C for 1 h. The lysate DNA was further tagmented by 1 μl Tn5 transposome (Vazyme, TTE Mix V50 in TD501) into adaptor-flanked fragments in 20 μl 1× tagmentation buffer (10 mM Tris-HCl, 7 mM MgCl$_2$ (Ambion, AM9530G), 10% *N*,*N*-dimethylformamide (Sigma, D4551), 4× protease inhibitor (Promega, G6521)). After incubating the tagmentation reaction at 55 °C for 1 h, 0.8 μM sequencing index primer and Q5 high-fidelity 1× master mix (New England Biolabs, M0492) were added to perform PCR amplification. The PCR procedure was 10 min at 72 °C for gap-filling; 30 s at 98 °C for pre-denaturation; 21 cycles of 15 s at 98 °C, 30 s at 60 °C and 2 min at 72 °C for denaturation; and 5 min at 72 °C for the last elongation. The purified product was quality-checked and sequenced using the Nextseq 500, Hiseq 4000 or HiSeq XTen sequencers (Illumina). We also performed high-depth WGS of selected samples using the Illumina NovaSeq 6000 sequencer. Image processing from sequencing data was performed using standard Illumina software and pipeline (bcl2fastq v.2.16).

Whole-exome library preparation and sequencing

The sequencing libraries were exome-captured using the SureSelectXT Human All Exon V6 (for oesophagus libraries) (Agilent, 5190-8864) or V7 (for libraries from other tissues) (5191-4005) following the manufacturer's guidelines. The products were quality-checked and sequenced with Illumina HiSeq XTen sequencers (Illumina), generating 2 × 150-bp paired-end reads. Image processing from sequencing data was performed using standard Illumina software and pipeline (bcl2fastq v.2.16).

Copy number analysis based on WGS data

We performed low-depth WGS (about 1.5 × 10^6 uniquely mapped reads) on each sample. For the data analysis, we first used Cutadapt[34] to trim adapters from the paired-end reads. Then, the clean reads were mapped to human reference genome hg19 (University of California) by using Bowtie2[35] with default settings. PCR duplicates were marked using Picard MarkDuplicates (http://broadinstitute.github.io/picard). Unique reads were then tabulated into non-overlapping dynamic bins (500-kb resolution) across the genome. Lowess regression normalization was performed to reduce the GC bias of bin counts. Copy number was called using the R package DNAcopy with the circular binary segmentation (CBS) algorithm. Finally, we calculated median absolute pairwise differences (MAPD) to identify and filter out low-quality samples (MAPD > 0.2).

SNV and indel calling

Paired-end reads from the sequencer were aligned to the human reference genome hg19 u-sing the Burrows–Wheeler Aligner (BWA) with default parameter settings[36]. The aligned BAM files were then sorted and merged (if needed) using SAMtools[37] (v.0.1.19). To call SNVs and indels from the exome sequencing data, we first realigned the mapped reads using the Genome Analysis Toolkit[38] (GATK 2.1–8) based on information of the dbSNP 135 (https://www.ncbi.nlm.nih.gov/snp/). Then, Picard-tools 1.76 was used to fix mate pairs and mark PCR duplicates (http://broadinstitute.github.io/picard). Next, the base quality recalibration was performed with GATK.

SNV calling

We used MuTect[39] (v.1.1.4) to call the SNVs in each biopsy sample, with the genomic DNA of white blood cells (WBCs) from each donor's peripheral blood as the germline comparator. To ensure the accuracy of SNV calling, we applied a series of filtering steps. (1) At least 10-fold coverage was required in the WBC samples containing at most one-fold mutated coverage (one-fold mutated coverage was allowed only when the total local coverage in the WBC was over 50-fold). (2) At least 10-fold total coverage was required in tissue biopsy samples with at least 3-fold mutation coverage. (3) The mutation allele frequency of each SNV was required to be greater than 5%. (4) The minimum value of the maximum mapping quality score of the mutated alleles was 20. (5) Variants both listed in the dbSNP database and (6) reported by the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (http://evs.gs.washington.edu/EVS) were removed.

Indel calling

We used the GATK Unified Genotyper to call indels with a series of filtering steps. (1) At least 10-fold coverage was required in the WBC samples without any mutated reads. (2) At least 10-fold total coverage in tissue biopsy samples and no less than 3-fold mutation coverage was required to support each indel. (3) Variants both listed in the dbSNP database and (4) reported by the NHLBI Exome Sequencing Project were removed. All indels that passed the filtering process were manually reviewed using SAMtools 'tview' to further eliminate those that presented in poorly mapped reads. We used SnpEff v.3.0 (ref. [40]) to annotate all SNVs and indels.

Validation of mutations called from WES on the basis of WGS data

To validate somatic mutations called from WES, we performed high-depth WGS of 43 normal tissues (with WES data) and 5 blood samples. Paired-end reads from WGS were aligned to the human reference genome hg19 (UCSC) using BWA-MEM with default parameters[36]. We then used picard-tools 1.76 to mark PCR duplicates (http://broadinstitute.github.io/picard). Somatic mutations were called using MuTect[39] (v.1.1.4) with the following criteria. (1) At least 10-fold coverage was required in the blood samples bearing at most one-fold mutated coverage (one-fold mutated coverage was allowed only when the total local coverage was over 50-fold in the normal sample). (2) At least 8× total coverage was required in the tissue biopsy samples with at least 2× mutated coverage. (3) The minimum value of the maximum mapping quality score of the mutated alleles was 20. (4) Variants both listed in the dbSNP database and (5) reported by the NHLBI Exome Sequencing Project (http://evs.gs.washington.edu/EVS) were removed. We compared mutations called from WGS (coding regions) with those from WES. In the meanwhile, we adopted a 'force-calling' strategy and examined whether there was evidence of mutations called from WES data appearing in WGS data by performing a pile-up at each mutation locus using WGS BAM files (using SAMtools 'tview' tool).

Sensitivity correction of mutation burden

We performed sensitivity corrections of mutation burdens in different tissues using a method described by two previous studies[41,42]. The sensitivity of mutation calling has a potential influence on the calculation of mutation burden and is related to two factors: the sequencing coverage and the clonality. Therefore, we calculated the sensitivity of mutation calling in each sample with the sequencing coverage and the clonality being taken into consideration. In brief, in each sample, we first generated 10,000 simulated sequencing depths around the observed sequencing depth based on a Poisson distribution (using the observed depth as the lambda). Theoretically, for a specific mutation, mutant reads are drawn from a binomial distribution with the total number of trials being the coverage depth and the probability of success on each trial being the VAF of this mutation. According to this theory, we calculated the probability of observing at least three mutant reads for SNVs and indels (the minimum mutant depth required in our mutation calling process) based on each simulated sequencing depth and the observed median VAF of mutations in the sample. There are 10,000 probabilities generated from this step, and we further calculated the average of all these probabilities as the sensitivity of mutation calling in the sample. Finally, we divided the observed number of mutations in each sample by the estimated sensitivity to correct the mutation burden.

Relationship between the number of mutations and gene expression levels

We downloaded the gene expression matrix from the GTEx project (v.8 data)[43], which includes tissue-specific gene expression for six types of tissues that we collected in this study (colon, oesophagus, liver, stomach, pancreas and lung). Expression levels for each tissue were measured as median transcripts per million (TPM). We ranked genes from low to high according to the $\log_2$-transformed expression levels ($\log_2$(TPM + 1)) and binned them into four quantiles. Mutation numbers in the four quantiles were then calculated.

Mutational signature extraction using the Bayesian hierarchical Dirichlet process

We studied the underlying mutational signatures that operated in normal tissues from different organs on the basis of all SNVs detected in both coding and non-coding regions in the exome sequencing data. To minimize the bias in mutational signature analysis, we combined normal biopsy samples from each dissected layer of each organ into a single sample, thus increasing the number of mutations available for the mutational signature analysis. Only unique mutations in each dissected layer were included to avoid double counting of mutations. We excluded the bronchus layer-1 sample of donor PN8 from the following mutational signature analysis because the mutation number was smaller than 40.

SBS mutational signatures were extracted using the Bayesian hierarchical Dirichlet process (HDP) implemented in the HDP R package[44] (https://github.com/nicolaroberts/hdp). Substitutions (including C>A, C>T, C>G, T>A, T>C and T>G) and their trinucleotide sequence contexts were considered in this analysis. Also, we used all the mutational signatures reported by the Pan-Cancer Analysis of Whole Genomes (PCAWG) project as the prior in this HDP-based analysis, without conditioning on any subsets of these prior signatures. Detailed information about parameter settings are included: (1) Hyperparameters for the α clustering parameter (α and β) were set to 1. (2) The parameter 'initcc' was set to 40 so that the extraction was started with 40 data clusters. (3) The initial 10,000 iterations of the Gibbs sampler (parameter 'burnin') were discarded. (4) After that, we collected 50 posterior samples (parameter 'n') with an interval of 50 iterations (parameter 'space'). (5) After each Gibbs sampling iteration, three iterations of concentration parameter sampling were performed (parameter 'cpiter').

Then, we compared our extracted mutational signatures to those that have been reported and published (COSMIC and PCAWG) based on the cosine similarity (the formula described below). Extracted signatures with a cosine similarity of greater than 0.85 compared to a known signature from either the COSMIC or the PCAWG catalogue of signatures were considered as the known signature with the highest similarity.

Cosine similarities between extracted mutational signatures (A) and known ones (B) were calculated as follows:

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

In this formula, $n = 96$ because we considered all 96 trinucleotide mutation contexts.

In total, we extracted seven mutational signatures that matched known mutational signatures. The relative activities of these signatures were used to generate the bar plot (Fig. 2a). The summed of SBS4 and SBS22 activities was considered as the mutagenic contribution from environmental factors.

Mutational signature analysis using deconstructSigs

We deconstructed the mutational signatures among the individual normal liver samples of donor PN9 using the R package deconstructSigs[45] (v.1.8.0) with default parameters. This approach can identify the closest fit from linear combinations of pre-defined or known mutational signatures and can be used to decipher the relative activity of each signature in each sample using linear decomposition. In our analysis, we restricted the pre-defined mutational signatures to the seven that we de novo extracted (SBS1, SBS2, SBS4, SBS5, SBS13, SBS22 and SBS45) using HDP across normal tissues. We then used the relative weight of each signature to generate donut plots which were subsequently mapped onto histological photographs.

Transcriptional strand bias of SBS4 and SBS22

To investigate transcriptional strand bias, we used the method that was described in a previous study on normal bladder mutagenesis[14]. In brief, gene strand information was extracted from the RefSeq database[46] and mutations were annotated as to whether the pyrimidine base was located on the template or coding strand. The probability, *P*, that a particular mutation, *i*, could be assigned to a given signature, *j*, in genome *k* was calculated as follows:

$$P_{i,j,k} = \frac{W_{j,k} \times F_{i,j}}{\sum_j W_{j,k} \times F_{i,j}}$$

in which $W_{j,k}$ is the proportion of mutations assigned to signature *j* in genome *k* by the HDP method and $F_{i,j}$ is the fraction of mutations in signature *j* that are the same substitution type and occur at the same trinucleotide context as mutation *i*.

Mutation assignment probabilities were used in two different ways for analysing transcriptional strand asymmetries. (1) Mutation assignment probabilities were summed together. (2) Only mutations with an assignment probability greater than 0.5 were included in the analysis.

Detection of potential driver genes under positive selection

To identify potential driver genes in the normal tissues of different organs, we used the dNdScv algorithm in R (https://github.com/im3sanger/dndscv)[28], which calculates the ratio of the rate of non-silent mutations versus silent mutations, while considering the mutation sequence context, the sequence of each gene and the mutation rate variation across genes. As our entire study covered normal samples from nine different organs from five different donors, the driver detection process was complicated. Therefore, we adopted two strategies to detect driver genes using dNdScv. First, we included all somatic mutations detected in all normal tissues in the nine organs from the five donors as input for the dNdScv algorithm. Second, we included all somatic mutations detected in normal tissues in each specific organ from the five donors to do organ-specific dNdScv analysis. In each of those strategies, we used two gene sets (all coding genes and 126 selected driver gene candidates) for the hypothesis testing in dNdScv analysis. The 126-gene list included genes selected from three sources: (1) driver genes that were identified using the dNdScv among cancers in the lung, oesophagus, colorectum, liver and stomach in a previous pan-cancer study[28]; (2) driver genes that were identified in The Cancer Genome Atlas (TCGA) lung[47], oesophageal[48], colorectal[49], liver[50] and stomach[29] cancer studies; (3) driver genes that have been reported in recent normal-tissue-sequencing studies on lung[10], oesophagus[5,6], colorectum[7], liver[8], skin[3] and endometrial[9] tissues. The list of the 126 genes can be found in Supplementary Table 8. Through this analysis, we considered genes that matched one of the following five categories as potential driver genes in our study.

Category 1: genes that were significant ($q < 0.1$) when the analysis involved all tissues from all donors and hypothesis testing was applied across all coding genes. Category 2: genes that were significant ($q < 0.1$) when the analysis involved all tissues from all donors and hypothesis testing was restricted to the 126 selected genes. Category 3: genes that were significant ($q < 0.1$) when the analysis involved normal tissues in certain organs from the 5 donors and hypothesis testing was applied across all coding genes. Category 4: genes that were significant ($q < 0.1$) when the analysis involved normal tissues in certain organs from the 5 donors and hypothesis testing was restricted to the 126 selected genes. Category 5: the genes that did not fit into any of the above four categories but contained at least 8 unique somatic mutations in the 126 selected driver gene candidates.

We excluded *TTN* from the final list of potential driver genes because it mutates frequently in various cancers, most probably because of its large gene size. We conducted pathway enrichment analysis of our 32 putative driver genes using the Reactome FI Cytoscape plug-in[51].

To study the potential preference of *MUC6* mutations in normal cardia and stomach tissues, we listed the top-10 most frequently mutated genes in TCGA gastric cancer studies and investigated the number of mutations in these genes in normal cardia and stomach tissues. In TCGA gastric cancers, there were 13/266 mutations (in a total of 266 mutations in these 10 genes) detected in *MUC6*. In normal cardia and stomach tissues, there were 17/30 mutations (in a total of 30 mutations in these 10 genes) detected in *MUC6*. We used the Fisher's exact test to calculate the statistical significance and the Benjamini–Hochberg method for multiple testing corrections.

To investigate whether there are any hotspot mutations being detected among normal tissues, we searched and downloaded a list of hotspot mutations documented in a previous publication[18]. In this list, hotspot mutations are defined as those in cancer genes (Cancer Gene Census) that appear more than three times in the TCGA consensus somatic mutation call set (https://doi.org/10.7303/syn7214402). We compared these hotspot mutations with mutations detected in driver genes in our study.

Organ preferences of mutated driver genes

To explore the potential organ preferences of mutated driver genes, we compared the observed and expected number of mutations of each driver genes across different organs. The ratio of observation to expectation ($R_{O/E}$) was calculated as follows: $R_{O/E}$ = Observed/Expected, in which the expected mutation numbers in different driver genes across organs were calculated based on the chi-square test. We considered a driver gene both with a $R_{O/E}$ value of greater than 1 in a specific organ and that was detected in that organ in more than one donor as having a potential preference for that organ. We calculated *P* values for the enrichment of driver genes across different normal tissues using the hypergeometric test.

Mutant cell fractions and clone size

The fraction of cells with a somatic mutation in an individual normal biopsy sample is proportional to the VAF of that mutation. However, that estimated fraction can be affected by local CNAs at the mutated site. As described in previous studies[3,5], considering local copy number status, the relationship between the MCF and the VAF of a specific somatic mutation can be described as:

$$\mathrm{MCF} = P \times \mathrm{VAF}/(\mathrm{CN_m} + (P \times \mathrm{VAF}) - (\mathrm{CN_m} + \mathrm{CN_n}) \times \mathrm{VAF})$$

In this equation, *P* represents the average ploidy of cells without this mutation, $\mathrm{CN_m}$ represents the copy number of alleles with the mutation in mutated cells, and $\mathrm{CN_n}$ is the copy number of alleles without the mutation in mutated cells.

In this study, we found that most normal biopsy samples were free of CNAs. Meanwhile, we found only a small number of somatic mutations occurred in the genomic regions with CNVs in some specific normal tissues. Therefore, for mutations in the autosomes and X chromosome of female individuals, the MCF can be simply calculated as follows: MCF = 2 × VAF. For mutations that occurred in the X chromosome of male individuals, MCF can be estimated as follows: MCF = VAF. We only calculated MCFs of SNVs, and not indels, in our study.

Clustering MCFs across multiple samples

To study the clone sharing events across multiple normal biopsy samples in different organs, we clustered somatic mutations from the multiple samples into different clusters based on the Bayesian Dirichlet process. We slightly modified the method described in a previous study[8]. In brief, instead of using VAFs of mutations, we first multiplied MCFs of mutations by 100, which thus represented the percentages of cells with certain mutations. Then we used integer MCF percentages as input for the Bayesian Dirichlet process. As described in that study, the model includes a potential split-merge step at each cycle of the Gibbs sampler, followed by a previously described Metropolis–Hastings proposal for conjugate distributions. We ran the Gibbs sampler for 15,000 iterations, dropping the first 10,000 as a burn-in. We used the Equivalence Classes Representatives (ECR) algorithm[52], implemented in the R package label.switching, to resolve the label-switching problem associated with mixture models. We removed clusters that contained fewer than four somatic mutations.

Defining the clonal independent index of each organ

We defined the clonal independent index of each organ in the five donors on the basis of the results from the Bayesian Dirichlet process that was used to cluster MCFs of multiple biopsy samples. In brief, the Bayesian Dirichlet process generates mutation clusters and their estimated median MCFs across biopsy samples (a matrix with mutation clusters in rows and biopsy samples in columns. The values in the matrix are the estimated median MCFs). We first removed mutation clusters with fewer than four mutations and the remaining clusters were regarded as the valid mutation clusters. Then, we ranked the median MCFs of all valid mutation clusters across all samples and categorized them into 10 bins. We considered that a mutation cluster appeared in a specific biopsy sample when the median MCF of that mutation cluster in that sample was larger than the second lowest MCF bin. After that, we calculated the number of biopsy samples that did not share any mutation clusters with others (number of independent samples). Finally, we defined the independent index of an organ by dividing the number of independent samples by the total number of samples with at least one valid mutation cluster. We also calculated the average MCFs of the estimated median MCFs of valid mutation clusters. Finally, we considered that an elevated clonal expansion of mutant cells in a given organ was indicated simultaneously by high average MCFs and a low independent index for that organ.

Construction of phylogenetic trees and clonal expansion regions

We constructed phylogenetic trees in different organs to depict the clonal relationship of multiple normal biopsy samples using MEGAX[53]. In brief, sequences with a 3 base-pair length surrounding all somatic mutations (including SNVs and indels) were extracted to construct the phylogenetic tree on the basis of the maximum-parsimony algorithm. We regarded both SNVs and indels as single events and indels with different length contributed equally to SNVs in the construction of phylogenetic trees. Phylogenetic trees were further optimized using Adobe Illustrator. Clonal expansion regions in oesophagus and liver samples were defined according to the clonal sharing events revealed by clustering analysis of the MCFs of mutations.

Reporting summary

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

Data availability

The raw WES and WGS data generated in this study have been deposited in the European Genome-phenome Archive (EGA) (https://ega-archive.org) with accession number EGAD00001007859 and the Genome Sequence Archive (GSA) of the Beijing Institute of Genomics with accession number HRA000356 (https://ngdc.cncb.ac.cn/gsa-human). To gain access to the raw sequencing data, please submit requests to the Pan-body Mutagenesis Data Access Committee (EGA accession number EGAC00001002218) or through the GSA online page of this study (https://ngdc.cncb.ac.cn/gsa-human/browse/HRA000356). All somatic mutations detected from WES with functional annotations and allele count information can be found in Supplementary Table 3. RefSeq database: https://www.ncbi.nlm.nih.gov/refseq. NHLBI Exome Sequencing Project: http://evs.gs.washington.edu/EVS. dbSNP database: https://www.ncbi.nlm.nih.gov/snp. COSMIC database: https://cancer.sanger.ac.uk/cosmic. The GTEx project: https://gtexportal.org/home.

Code availability

Mutational signature analysis was performed using the HDP R package v.0.1.5 (https://github.com/nicolaroberts/hdp). Code for mutational signature analysis was adapted from https://github.com/HLee-Six/colon_microbiopsies. Code for the Bayesian Dirichlet process clustering of MCFs was adapted from https://github.com/sfbrunner/liver-pub-repo. Adapted code is available at Zenodo (https://doi.org/10.5281/zenodo.5012918). Driver gene analysis was performed using the dNdScv v0.01 (https://github.com/im3sanger/dndscv).

34. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. J.* **17**, 10–12 (2011).
35. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
36. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
37. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
38. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
40. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* **6**, 80–92 (2012).
41. Coorens, T. H. H. et al. Inherent mosaicism and extensive mutation of human placentas. *Nature* **592**, 80–85 (2021).
42. Olafsson, S. et al. Somatic evolution in non-neoplastic IBD-affected colon. *Cell* **182**, 672–684 (2020).
43. The GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
44. Roberts, N. D. *Patterns of Somatic Genome Rearrangement In Human Cancer*. PhD thesis, Univ. Cambridge (2018).
45. Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S. & Swanton, C. DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biol.* **17**, 31 (2016).
46. O'Leary, N. A. et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–745 (2016).
47. Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
48. The Cancer Genome Atlas Research Network. Integrated genomic characterization of oesophageal carcinoma. *Nature* **541**, 169–175 (2017).
49. The Cancer Genome Atlas Research Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
50. The Cancer Genome Atlas Research Network. Comprehensive and integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341 (2017).
51. Wu, G., Feng, X. & Stein, L. A human functional protein interaction network and its application to cancer data analysis. *Genome Biol.* **11**, R53 (2010).
52. Papastamoulis, P. label. switching: an R package for dealing with the label switching problem in MCMC outputs. *J. Stat. Softw.* **69**, Code Snippet 1 (2016).
53. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

Acknowledgements We thank all the donors and their families for their consent and participation in this study. This project was jointly supported by the National Natural Science Foundation of China (81725015 to C.W., 81988101 to D.L. and C.W., 22050004 to J.W. and 22050002 to Y.H.); Beijing Outstanding Young Scientist Program (BJJWZYJH01201910023027 to C.W.); the Medical and Health Technology Innovation Project of the Chinese Academy of Medical Sciences (2019-I2M-2-001 to D.L. and C.W.); the National Key R&D Program of China (2019YFC1315702, 2018ZX10302205 to F.B. and 2018YFA0108100 to Y.H.); Guangdong Province Key Research and Development Program (2019B020226002 to F.B.); Beijing Municipal Science and Technology Commission (Z201100005320016 to Y.H.); Beijing Advanced Innovation Center for Genomics; and Shenzhen Bay Laboratory.

Author contributions F.B., C.W., Y.H., J.W., D.L. and R.L. conceived the study. R.L., L.D. and W.F. performed data analyses with assistance from L.L., D.X. and Y.W. W.F. performed sample collection, tissue sectioning and LCM with assistance from Y.L., W.G., W.L., L.C., Y.C., C.M., H.L. and Y.M. L.D. performed DNA extraction and library preparation. J.L. performed exome capture with assistance from Q.L. R.L., F.B., W.F., L.D. and J.L. wrote the manuscript with input from D.L., Y.H., J.W. and C.W. D.L., Y.H., J.W., F.B. and C.W. supervised all aspects of this study.

Competing interests The authors declare no competing interests.

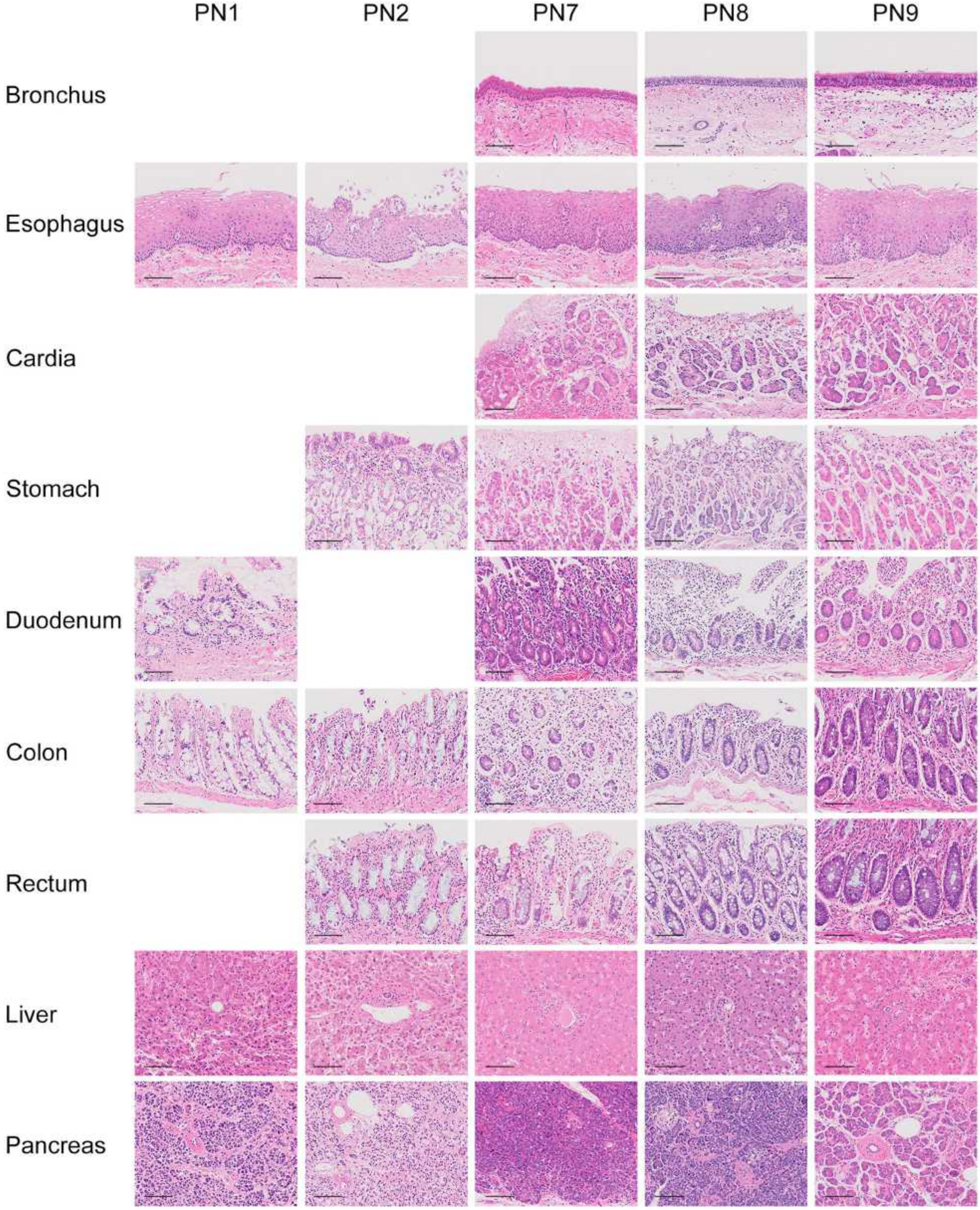
Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-021-03836-1.

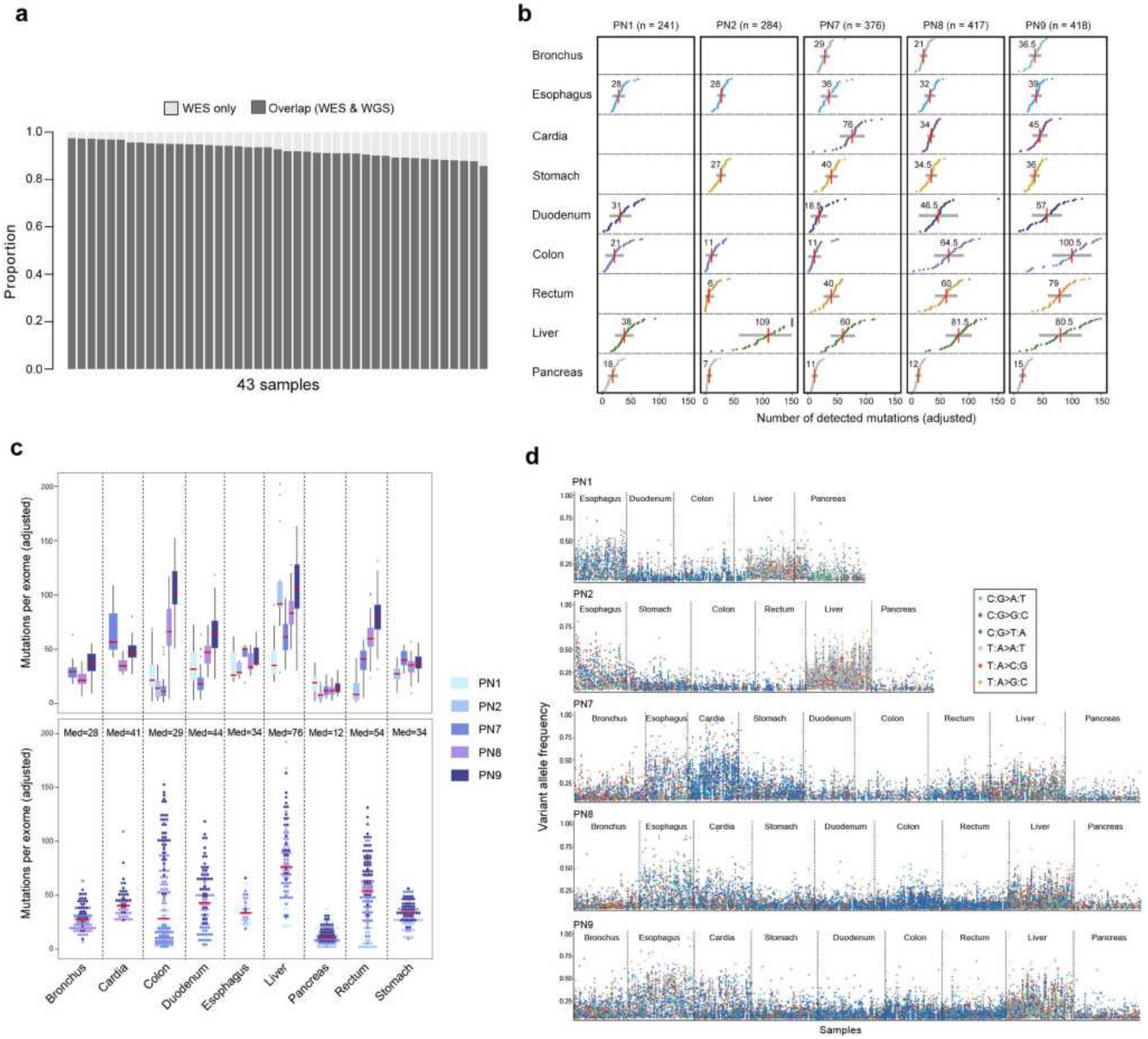
Correspondence and requests for materials should be addressed to D.L., Y.H., J.W., F.B. or C.W.

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

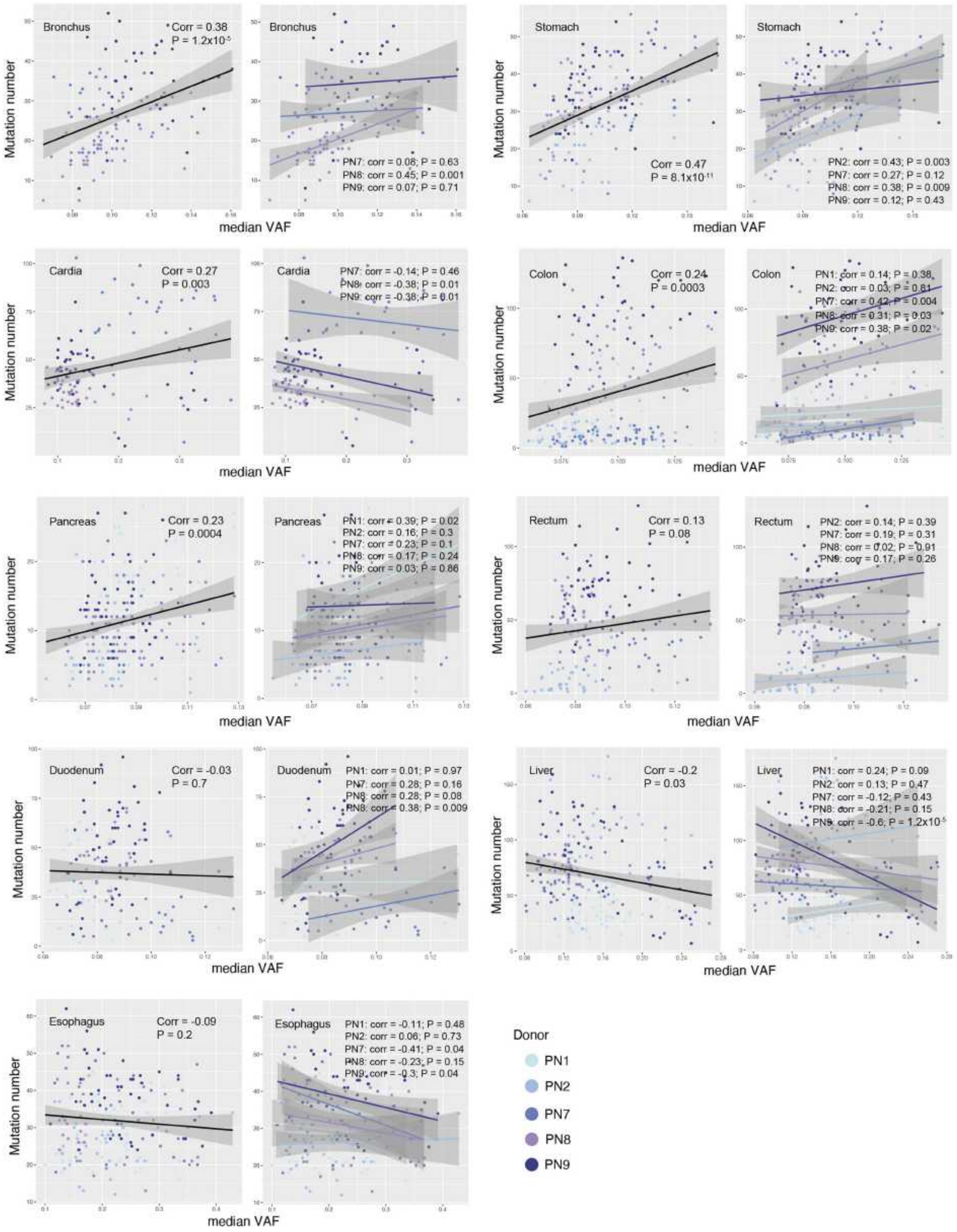
Reprints and permissions information is available at http://www.nature.com/reprints.



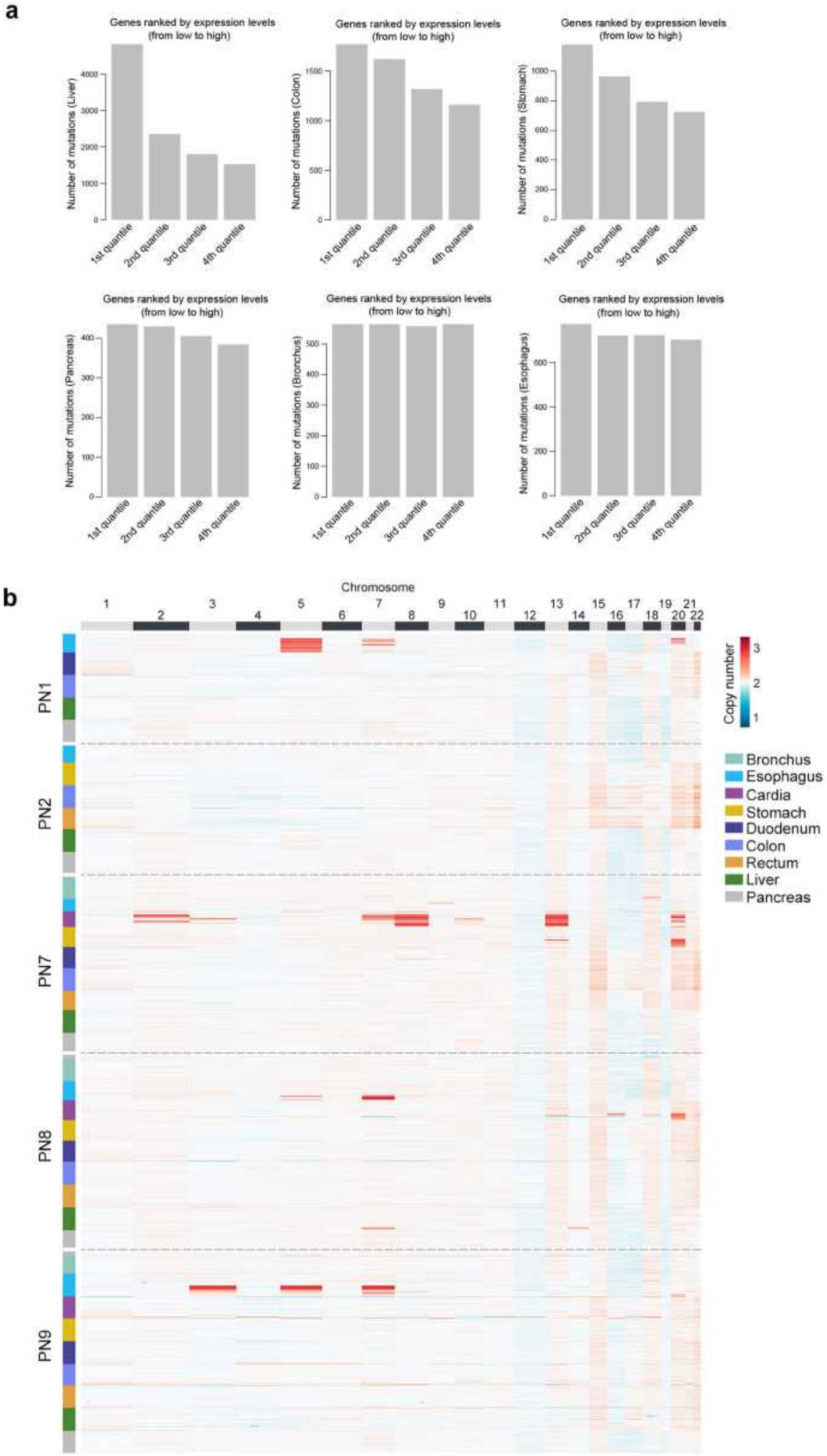
Extended Data Fig. 1 | Normal tissue histology. Representative H&E-stained samples showing the histological features of normal tissues sampled from nine organs from the five donors. Blanks in the figure represent samples that are not available in corresponding organs and donors. Scale bars, 100 μm.



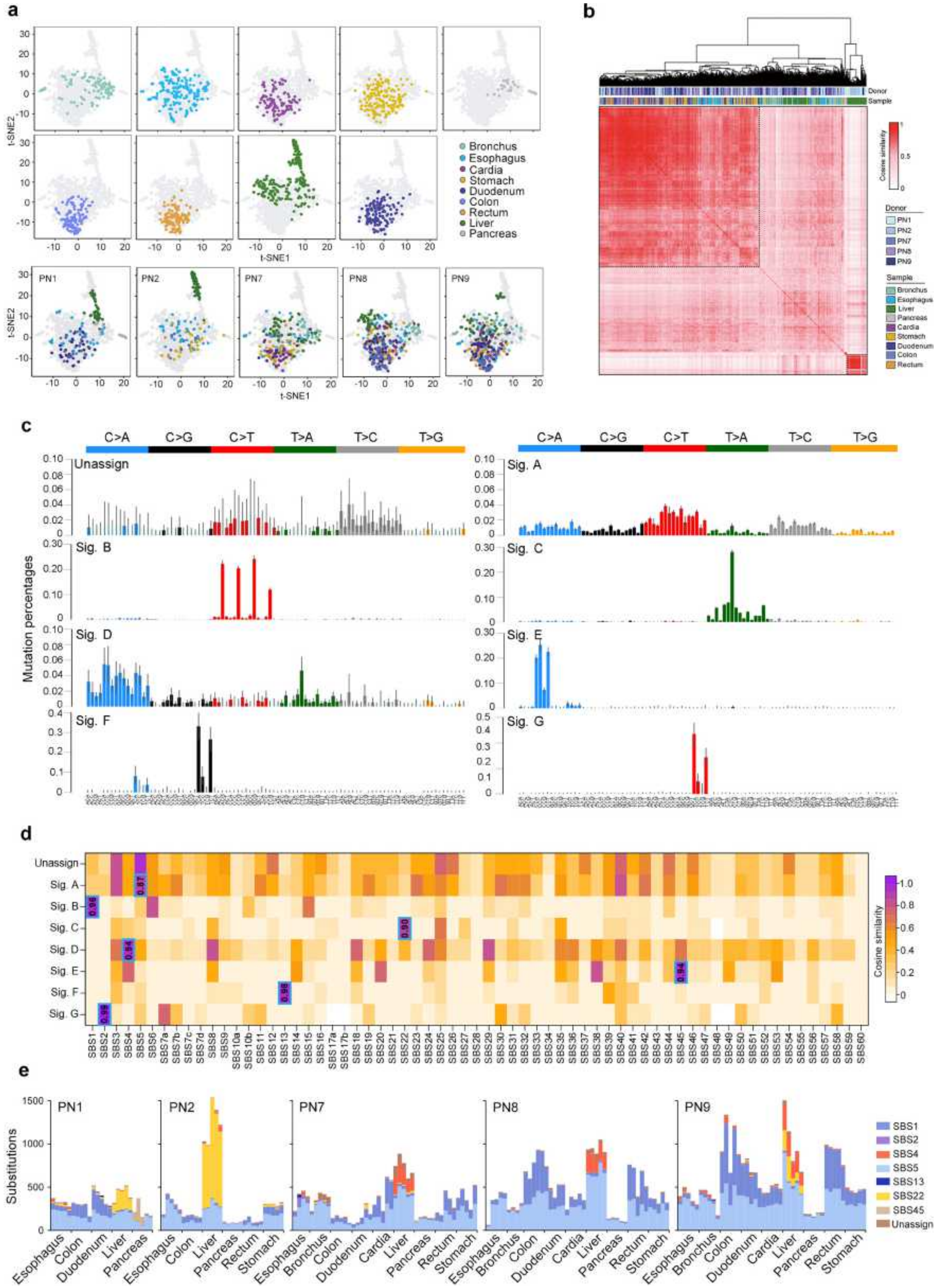
Extended Data Fig. 2 | Detection of somatic mutations. a, Bar plot showing the overlap of mutations detected from WES and WGS of 43 samples. **b**, Adjusted numbers of somatic mutations detected in the coding regions in tissue biopsies from the organs of five donors. Red vertical bars represent median mutation numbers and grey horizontal bars represent standard deviations. **c**, The mutation burdens (after the sensitivity correction) in samples with median VAFs between 0.08 and 0.14. Top, box plots showing the mutation burdens in organs from different donors. The lower edge, upper edge and centre of the box represent the 25th (Q1) percentile, 75th (Q3) percentile and the median, respectively. IQR = Q3 − Q1. Outliers are values beyond the whiskers (upper, Q3 + 1.5 × IQR; lower, Q1 − 1.5 × IQR). Detailed information about the box plots can be found in Supplementary Table 3. Bottom, dot plots showing the adjusted mutation burdens in different organs. Red bars represent the medians. **d**, Scatter plots showing the VAFs of somatic mutations detected in the normal tissues from the nine organs of the five donors. Dots are coloured by mutation type.



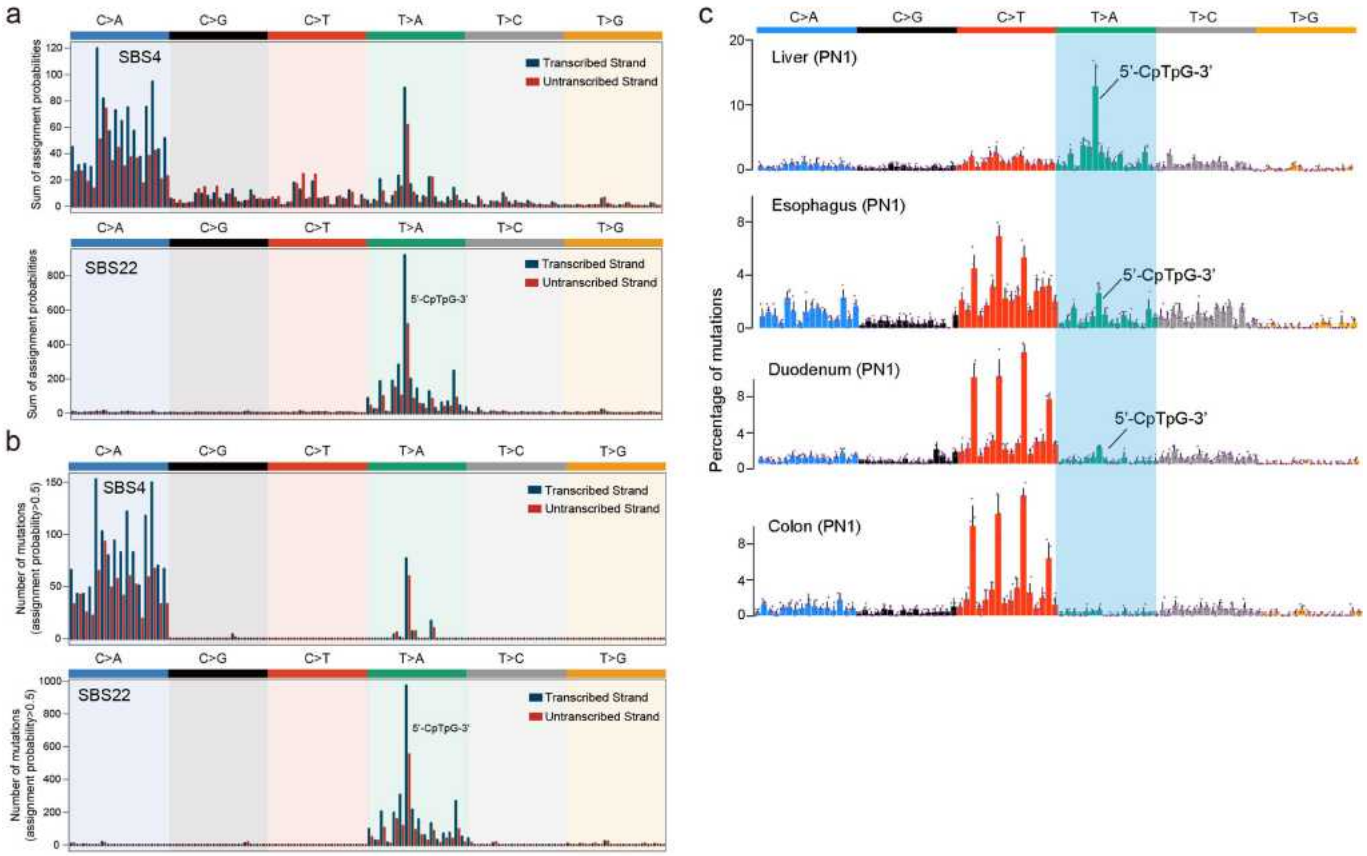
Extended Data Fig. 3 | Correlations and interdependence between VAF distributions and mutation numbers. In each tissue, we calculated the first quantile (Q1) and third quantile (Q3) of the VAF and mutation burden distribution. We defined IQR = Q3 − Q1 and considered samples with a median VAF or mutation burden greater than Q3 + 1.5 × IQR or less than Q1 − 1.5 × IQR as outliers. We excluded these outliers in this analysis. Corr., correlation. The error bands represent the 95% confidence intervals. *P* values are from two-sided correlation tests based on the Pearson correlation coefficient.



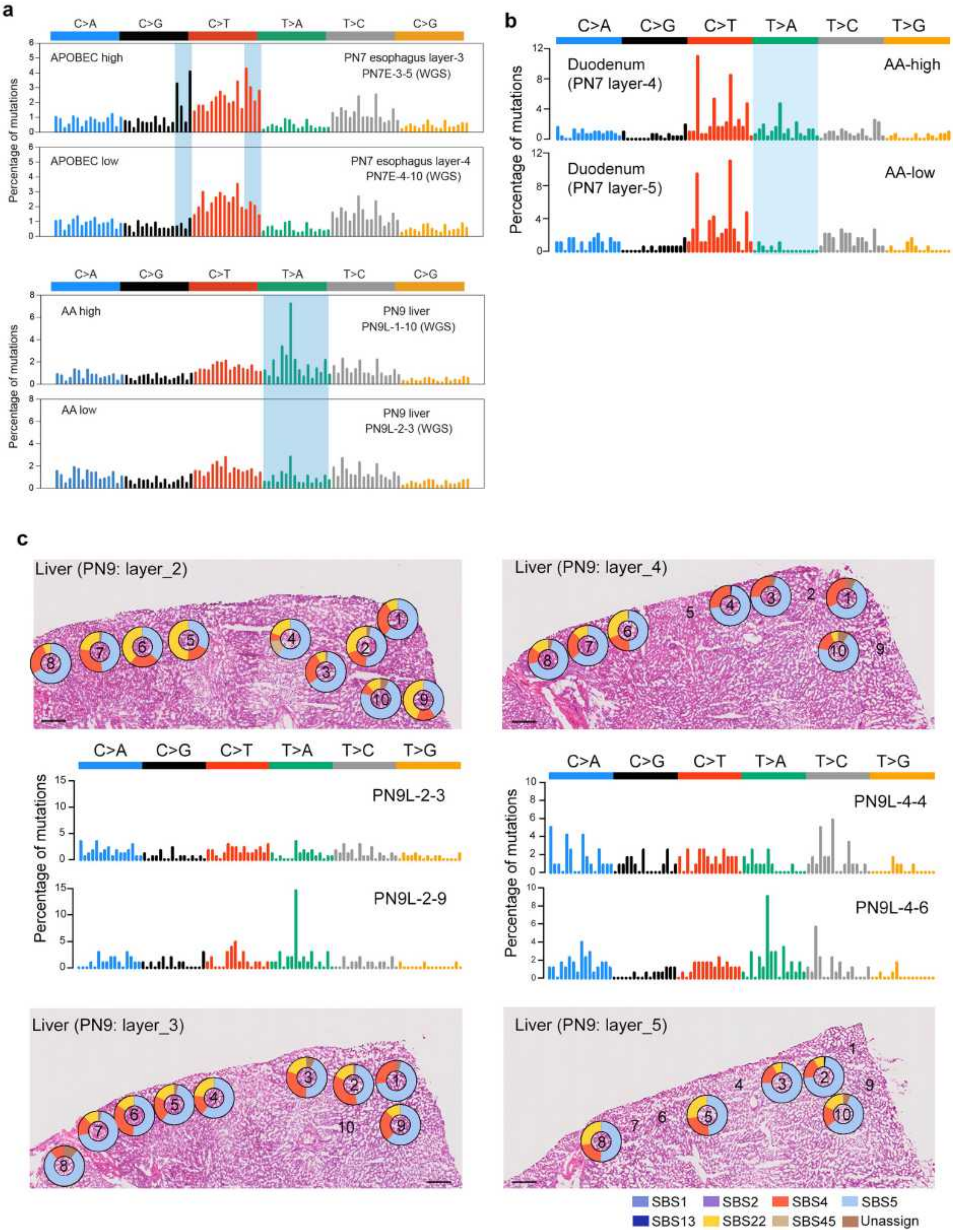
Extended Data Fig. 4 | Mutation numbers and somatic CNAs. a, Bar plots showing the number of mutations among the four intervals. Genes are divided into four intervals according to the tissue-specific gene expression levels. **b**, Heat maps showing somatic CNAs detected in the normal tissues from the nine organs of the five donors. Sex chromosomes were excluded.



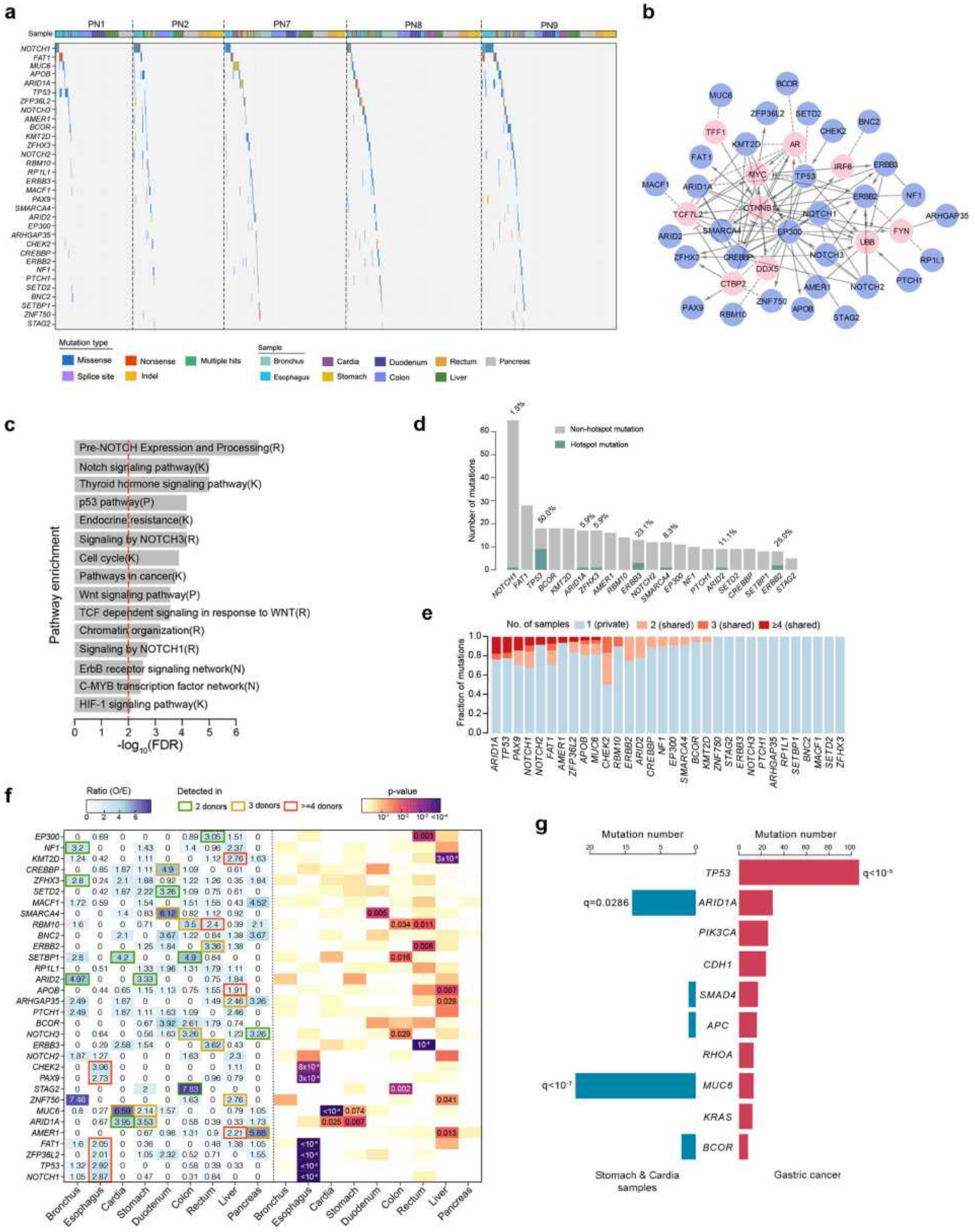
Extended Data Fig. 5 | Mutational spectra and signature analysis. **a**, *t*-stochastic neighbour embedding (t-SNE) plots of the trinucleotide mutational spectra of biopsy samples from each donor, broken down by organ and donor. Only biopsy samples with more than 30 SNVs were included. **b**, Heat map showing the clustering of cosine similarities of the trinucleotide mutational context in different samples. Colour bars above indicate information of donors and tissue types. **c**, Trinucleotide mutational spectra for the unassigned signature and the seven signatures extracted using a Bayesian hierarchical Dirichlet process. The bars represent means (95% credible intervals) of the 96 trinucleotide contexts. **d**, Heat map depicting the cosine similarities between extracted mutational signatures and mutational signatures from COSMIC and PCAWG catalogues. Cosine similarities between the seven extracted mutational signatures and their most similar comparators are highlighted. **e**, Stacked bar plots showing the number of mutations that are caused by different mutational signatures.



Extended Data Fig. 6 | Mutational signature analysis. a, Transcriptional strand asymmetries across 96 mutation contexts for SBS4 and SBS22. Bar plots show the sum of assignment probabilities across trinucleotide contexts, split by whether the pyrimidine is on the template or coding strand. **b**, Transcriptional strand asymmetries across 96 mutation contexts for SBS4 and SBS22. Only mutations with an assignment probability greater than 0.5 are included. **c**, Trinucleotide mutational spectra of liver, oesophagus, duodenum and colon from donor PN1. Purple dots represent data points of the five tissue layers. Data are mean + s.d. Typical aristolochic-acid -associated mutational features are shaded in blue.

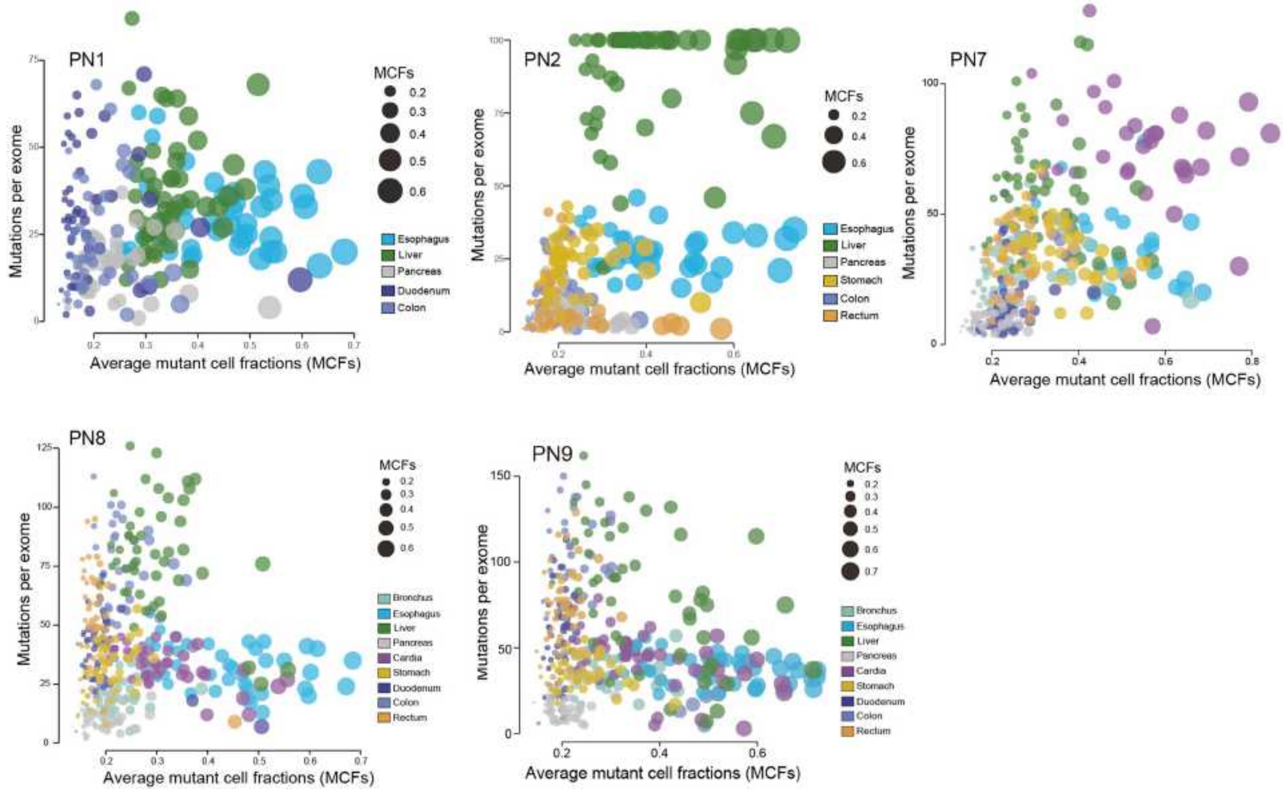


Extended Data Fig. 7 | Intra-donor comparisons of mutational signatures. **a**, The 96 mutation context profiles in two oesophagus samples from donor PN7 (top) and two liver samples from donor PN9 (bottom) based on somatic mutations detected from WGS. **b**, Trinucleotide mutational spectra of two dissected duodenum layers from donor PN7. Typical aristolochic-acid -associated mutational features are shaded in blue. **c**, H&E stained liver tissue (PN9 layer 2 to 4) with superimposed donut charts showing the proportional contributions of mutational signatures, as estimated by deconstructSigs. Scale bars, 200 μm.

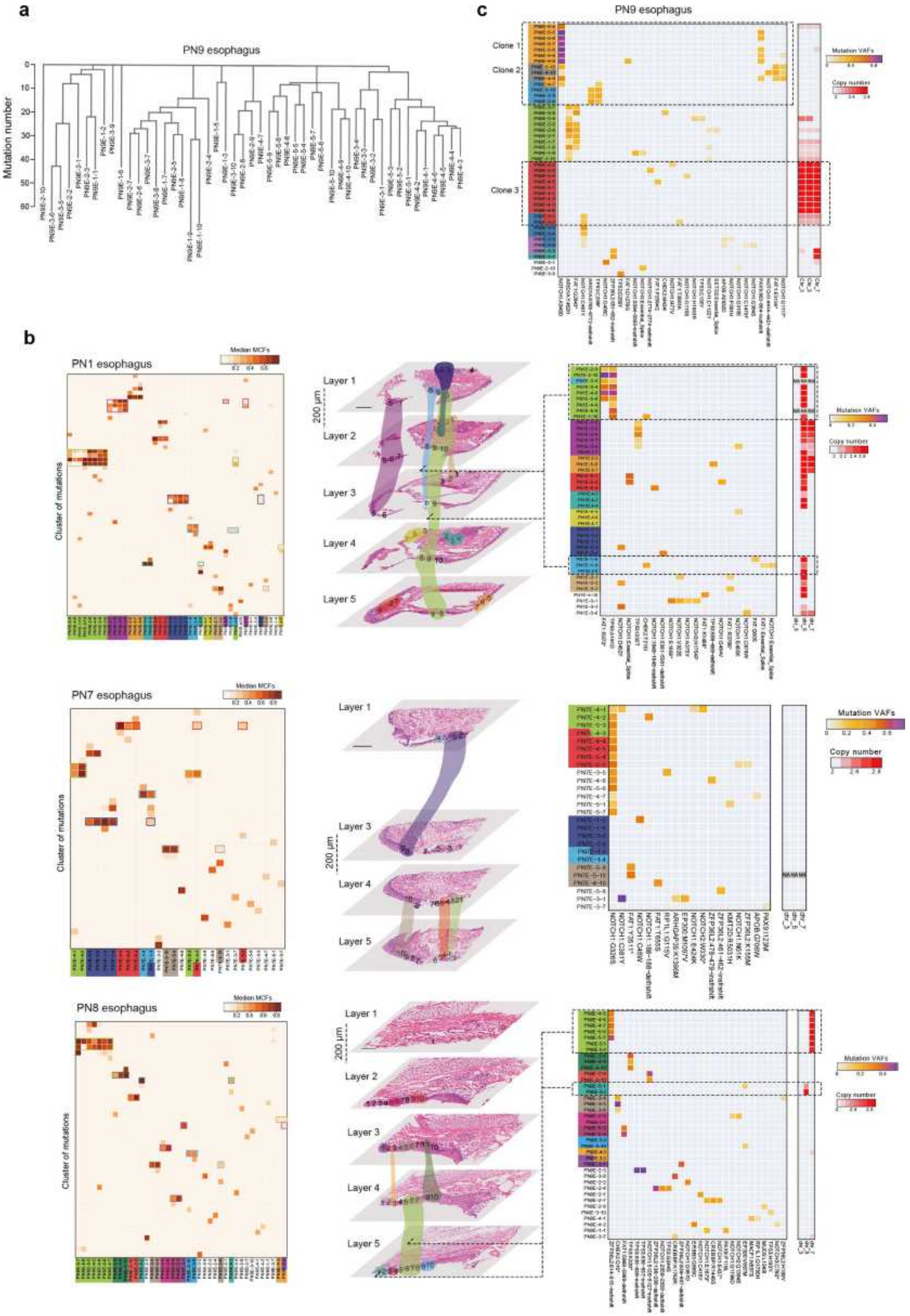


Extended Data Fig. 8 | See next page for caption.

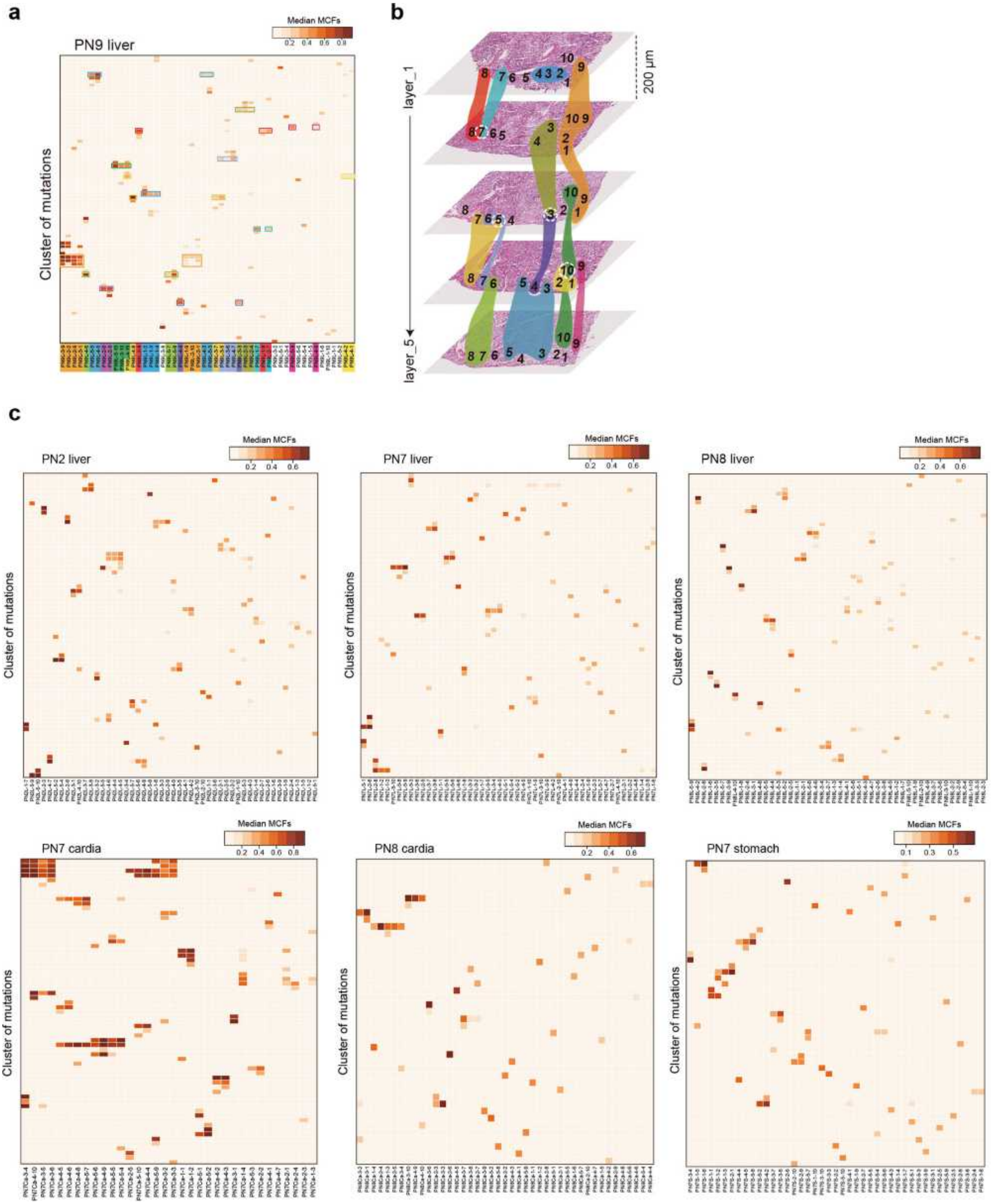
Extended Data Fig. 8 | Landscape of driver mutations. a, Mutational landscape of the 32 putative driver genes across different organs from the 5 donors. **b**, The functional interaction (network of the 32 driver genes. Driver genes are in blue nodes and linker genes (those not significantly mutated but highly connected to driver genes in the network) are in pink nodes. **c**, Significantly enriched pathways of the 32 driver genes. The vertical red line marks a false discovery rate (FDR) of 0.01. **d**, Bar plot showing the numbers of total mutations and cancer hotspot mutations in driver genes. The percentages of hotspot mutations are labelled on the top of the bar plot. **e**, Fraction of driver mutations that are private or shared by more than one biopsy sample. **f**, Heat maps showing the ratio of the numbers of observed to expected (O/E) driver mutations across different organs (left) and the *P* values for the enrichment (right). *P* values from one-sided hypergeometric tests. **g**, Bar plots comparing the number of mutations in gastric cancer top-10 most frequently mutated driver genes in TCGA with normal stomach and cardia samples in this study. Adjustment for multiple comparisons was performed. Adjusted *P* values (*q*-value) are labelled.



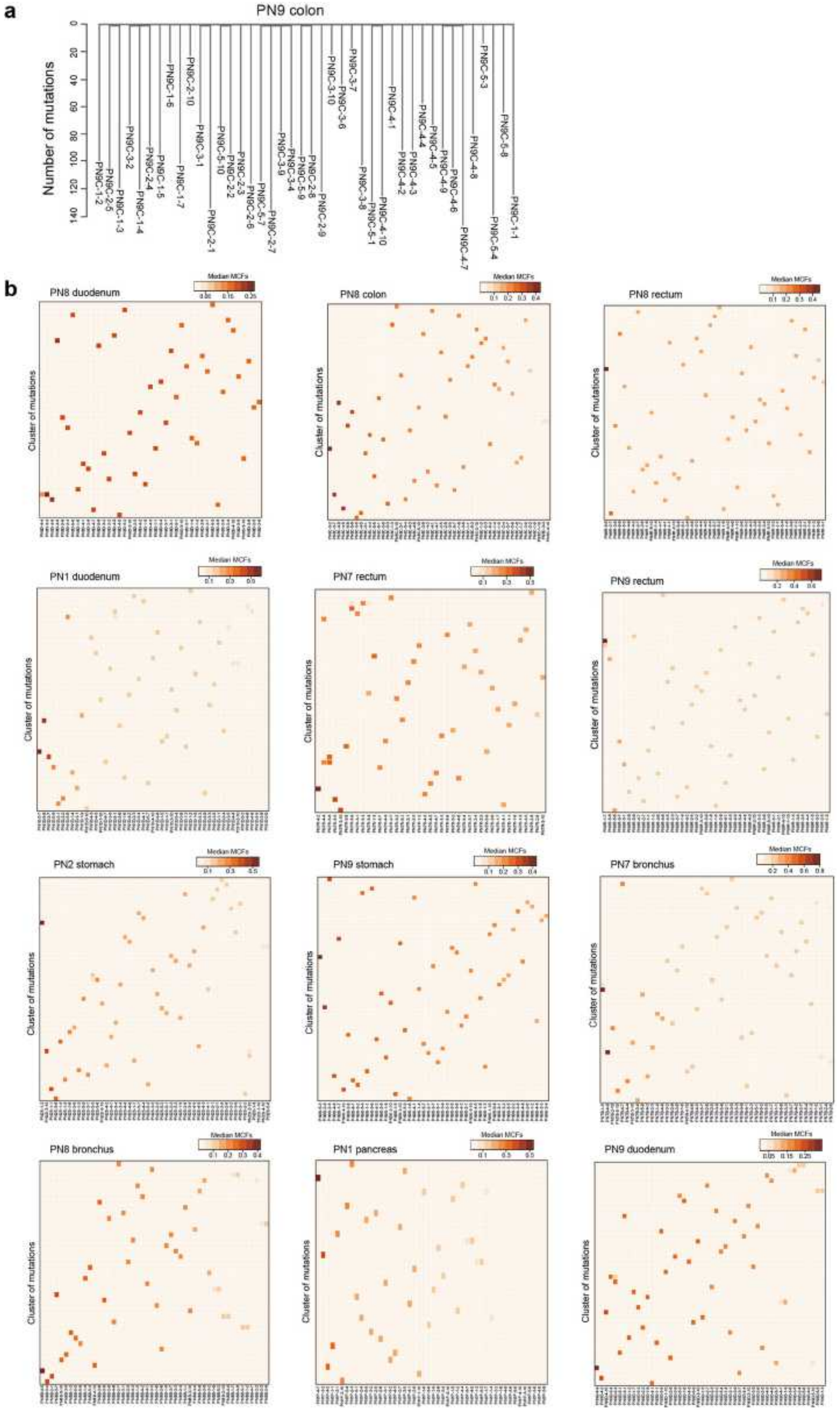
Extended Data Fig. 9 | Relationships between mutational burdens and average mutant MCFs. Bubble plots show the correlations between average MCFs and mutational burdens in biopsy samples across different organs in donors PN1, PN2, PN7, PN8 and PN9.



Extended Data Fig. 10 | Mutant clonal expansion in oesophageal epithelium. a, Phylogenetic tree depicting the clonal relationships of the biopsy samples of the oesophagus of donor PN9. **b**, Heat maps show mutation clustering, spatial clonal architecture and potential driver mutations or CNAs in samples from the oesophagus. Scale bars, 800 μm. **c**, Heat maps showing potential driver mutations and CNAs in oesophagus samples from donor PN9.



Extended Data Fig. 11 | Representative examples of large scale mutant clonal expansion. a, Heat map showing the mutation clustering in liver samples from donor PN9. **b**, Spatial clonal architecture of liver tissue from donor PN9. The numbers in each layer represent the positions of LCM biopsy samples. The overlaid colours correspond to **a** and indicate the ranges of clonal expansions. **c**, Heat maps show mutation clustering in samples from the representative organs. Each cluster contains mutations with similar MCFs.



Extended Data Fig. 12 | Representative examples of independent clonal evolution. **a**, Phylogenetic tree depicting the clonal relationships of colon biopsy samples from donor PN9. **b**, Heat maps showing clustered mutations in samples from representative organs. Each cluster contains mutations with similar MCFs.