Article

3D genomic mapping reveals multifocality of human pancreatic precancers

https://doi.org/10.1038/s41586-024-07359-3

Received: 11 January 2023

Accepted: 26 March 2024

Published online: 01 May 2024

Check for updates

Alicia M. Braxton[1,2,14], Ashley L. Kiemen[1,3,4,14], Mia P. Grahn[3], André Forjaz[3], Jeeun Parksong[1], Jaanvi Mahesh Babu[1], Jiaying Lai[5], Lily Zheng[5,6], Noushin Niknafs[4], Liping Jiang[7], Haixia Cheng[7], Qianqian Song[7], Rebecca Reichel[1], Sarah Graham[1], Alexander I. Damanakis[1], Catherine G. Fischer[1], Stephanie Mou[1], Cameron Metz[1], Julie Granger[1], Xiao-Ding Liu[1,8], Niklas Bachmann[1], Yutong Zhu[3], YunZhou Liu[5], Cristina Almagro-Pérez[3], Ann Chenyu Jiang[3], Jeonghyun Yoo[3], Bridgette Kim[3], Scott Du[3], Eli Foster[3], Jocelyn Y. Hsu[3], Paula Andreu Rivera[3], Linda C. Chu[9], Fengze Liu[9], Elliot K. Fishman[9], Alan Yuille[10], Nicholas J. Roberts[1,4], Elizabeth D. Thompson[1], Robert B. Scharpf[4], Toby C. Cornish[11], Yuchen Jiao[7,12]✉, Rachel Karchin[4,6], Ralph H. Hruban[1,4], Pei-Hsun Wu[3], Denis Wirtz[1,3,4,15]✉ & Laura D. Wood[1,4,13,15]✉

Pancreatic intraepithelial neoplasias (PanINs) are the most common precursors of pancreatic cancer, but their small size and inaccessibility in humans make them challenging to study[1]. Critically, the number, dimensions and connectivity of human PanINs remain largely unknown, precluding important insights into early cancer development. Here, we provide a microanatomical survey of human PanINs by analysing 46 large samples of grossly normal human pancreas with a machine-learning pipeline for quantitative 3D histological reconstruction at single-cell resolution. To elucidate genetic relationships between and within PanINs, we developed a workflow in which 3D modelling guides multi-region microdissection and targeted and whole-exome sequencing. From these samples, we calculated a mean burden of 13 PanINs per cm^3 and extrapolated that the normal intact adult pancreas harbours hundreds of PanINs, almost all with oncogenic *KRAS* hotspot mutations. We found that most PanINs originate as independent clones with distinct somatic mutation profiles. Some spatially continuous PanINs were found to contain multiple *KRAS* mutations; computational and in situ analyses demonstrated that different *KRAS* mutations localize to distinct cell subpopulations within these neoplasms, indicating their polyclonal origins. The extensive multifocality and genetic heterogeneity of PanINs raises important questions about mechanisms that drive precancer initiation and confer differential progression risk in the human pancreas. This detailed 3D genomic mapping of molecular alterations in human PanINs provides an empirical foundation for early detection and rational interception of pancreatic cancer.

Pancreatic ductal adenocarcinoma (PDAC) is an aggressive malignancy that is frequently diagnosed at an advanced stage, leading to a 5-year survival rate[2,3] of 11%. PDAC arises from noninvasive precursor lesions, most commonly PanINs, which are curable if detected and treated early[1]. PanINs are challenging to study; owing to their small size (by definition less than 0.5 cm in a standard histological section), they cannot be grossly identified[4,5]. So far, studies of human PanINs have evaluated discrete slides of sectioned pancreatic tissue[6–9]. Although this 2D approach may identify parts of PanINs, their total numbers, size, shape or connectivity within tissue cannot be fully appreciated, precluding full understanding of the earliest steps of pancreatic tumorigenesis. A more complete assessment of these features necessitates 3D analysis of larger slabs of tissue. We recently reported CODA, a machine-learning pipeline for 3D image analysis that can generate quantifiable models of large tissues with single-cell resolution[10]. Using CODA, we are able to quantify the number and connectivity of PanINs through systematic evaluation of large pieces of human pancreatic tissue.

[1]Department of Pathology, Sol Goldman Pancreatic Cancer Research Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [2]Department of Comparative Medicine, Medical University of South Carolina, Charleston, SC, USA. [3]Department of Chemical and Biomolecular Engineering, Johns Hopkins University, Baltimore, MD, USA. [4]Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [5]Institute for Computational Medicine, Johns Hopkins University, Baltimore, MD, USA. [6]McKusick-Nathans Institute for Genetic Medicine, Johns Hopkins University, Baltimore, MD, USA. [7]State Key Laboratory of Molecular Oncology, National Cancer Center, National Clinical Research Center for Cancer and Cancer Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, China. [8]Department of Pathology, Molecular Pathology Research Center, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. [9]Department of Radiology, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [10]Department of Computer Science, Johns Hopkins University, Baltimore, MD, USA. [11]Department of Pathology, University of Colorado School of Medicine, Aurora, CO, USA. [12]Institute of Cancer Research, Henan Academy of Innovations in Medical Science, Zhengzhou, China. [13]Convergence Institute, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [14]These authors contributed equally: Alicia M. Braxton, Ashley L. Kiemen. [15]These authors jointly supervised this work: Denis Wirtz, Laura D. Wood. ✉e-mail: jiaoyuchen@gmail.com; wirtz@jhu.edu; ldwood@jhmi.edu

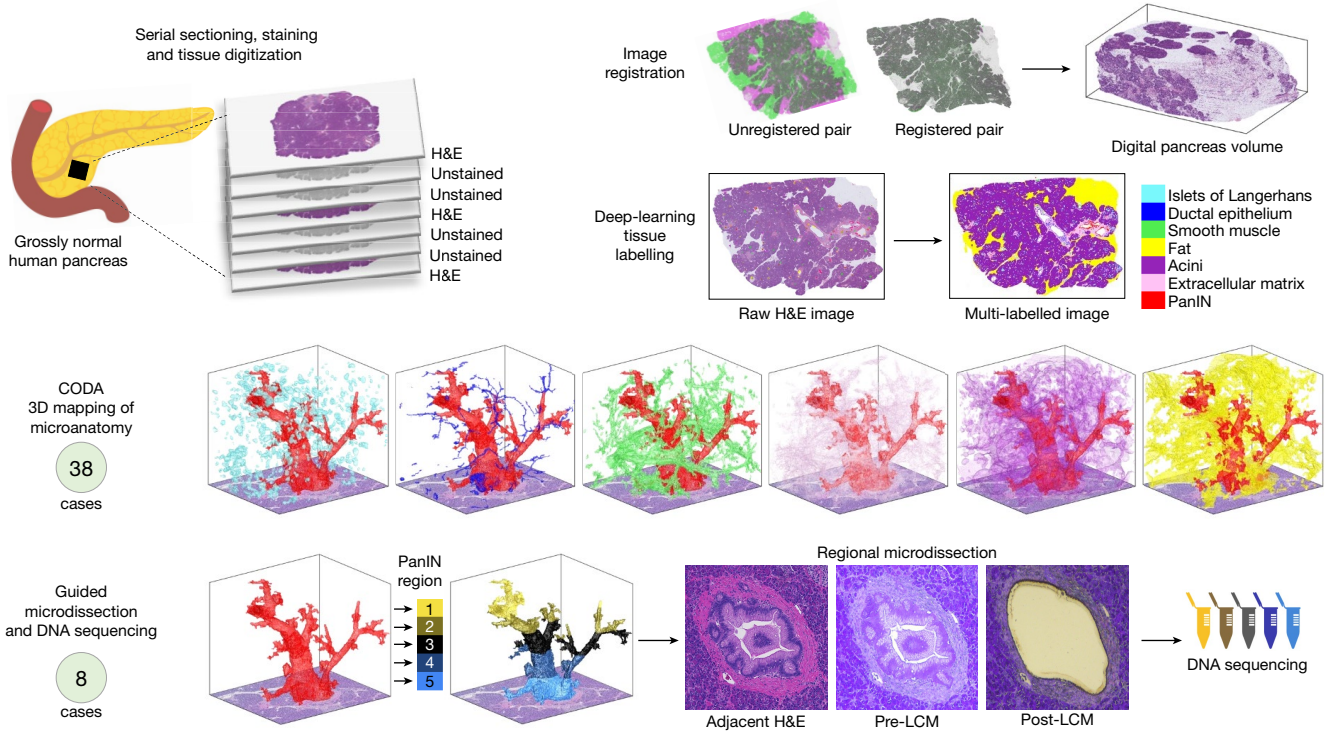


Fig. 1 | Tissue processing workflow and cohorts for CODA 3D modelling and sequencing. Top left, each pancreatic tissue slab was serially sectioned, and every third slide was stained with haematoxylin and eosin (H&E) and imaged for 3D modelling. Top right, with CODA, H&E serial sections are registered to create a digital volume and tissue types are labelled by deep learning, enabling 3D reconstruction of pancreas microanatomy. Middle, thirty-eight slabs were analysed solely for 3D modelling. Bottom, eight additional slabs were 3D modelled, and the intervening unstained slides were used for 3D-guided multi-region microdissection of spatially distinct PanINs for NGS and WES.

The accumulation of somatic mutations drives the initiation and progression of PanINs[1]. These include hotspot mutations in the oncogene *KRAS*, which are found in more than 90% of invasive pancreatic cancers and are thought to initiate pancreatic ductal neoplasia, as well as less prevalent inactivating mutations in tumour suppressor genes such as *CDKN2A* and *TP53*[1,6,11–14]. In intraductal papillary mucinous neoplasms (IPMNs), the larger noninvasive cystic precursors to PDAC, multi-region next-generation sequencing (NGS) has demonstrated driver gene mutations to be heterogeneous, attesting to complex clonal evolution[15–17]. By contrast, driver gene heterogeneity has not been identified in examined primary PDACs or metastases[18,19]. Assessment of genetic heterogeneity in PanINs requires integration of 3D modelling and genetic analysis, so that spatially distinct PanINs can be exhaustively identified and then separately sequenced.

This study describes the 3D microanatomy, cellular features and somatic genetic alterations of human PanINs. We generated 3D models from a large cohort of grossly normal, surgically resected human pancreatic tissue slabs using CODA. We also created a workflow for mapping genetic variation across 3D microanatomy by integrating CODA with multi-region targeted and whole-exome sequencing (WES), enabling rigorous evaluation of both intra-PanIN and inter-PanIN genetic heterogeneity. Together, our data reveal a marked multifocality of PanINs and elucidate their complex molecular origins.

Normal pancreas contains many PanINs

We first used CODA to determine the size, shape and number of PanINs in the human pancreas in 3D at single-cell resolution (Fig. 1 and Extended Data Fig. 1). Thirty-eight slabs of grossly normal pancreatic tissue were obtained from surgical pancreatectomy specimens, including 12 to resect PDAC and 26 to resect other neoplasms not involving the pancreatic ductal system (Supplementary Table 1). We examined a mean of 1,288 slides per slab (range 679–1,703) and mean tissue volume of 2.03 cm^3 (range 0.94–3.62 cm^3) (Supplementary Table 2). Thirty-six slabs contained exclusively low-grade (LG) PanINs, and two slabs included PanINs with regions of high-grade (HG) dysplasia.

Pancreatic ducts were 3D-rendered for each slab to visualize the spatial distribution of PanINs and normal ductal epithelium (Fig. 2a and Extended Data Fig. 2). Across 38 slabs, 889 spatially separate PanINs were modelled, with a median of 21 PanINs per slab (range 4–92) (Supplementary Table 2). PanINs contained a median of 3,515 cells, a mean of 95,021 cells, and a range of 26–7,239,369 cells. Most PanINs (764 out of 889 (86%)) contained fewer than 100,000 cells; 266 (30% of PanINs) contained fewer than 1,000 cells (Fig. 2b,c). Nine slabs (23%) had one or more large PanINs exceeding 1,000,000 cells, and no slab contained more than 3 PanINs of this size. The two PanINs with HG dysplasia occupied the largest volumes in their respective slabs, with cell counts exceeding 200,000 cells (Supplementary Table 2). The vast majority of PanINs (91%) measured less than 0.5 cm in the largest dimension, in keeping with the histopathological definition of PanINs on 2D sections[5]. We next quantified the normal ductal cells in each slab to calculate the percentage of neoplastic cells within the ductal epithelium. We found that a median of 30% of ductal epithelial cells were neoplastic PanIN cells (range 0.2–75%) (Fig. 2d). Although individual PanINs remained small, their cumulative presence occupied more than a quarter of the ductal system in most cases.

The mean overall PanIN burden, calculated as the number of PanINs per slab divided by tissue volume, was 13 PanINs per cm^3 with a range of 1 to 31 PanINs per cm^3 (Fig. 2e). Although not statistically significant, this metric was greater in patients with PDAC elsewhere in their pancreas compared to those with non-ductal disease (Extended Data Fig. 3a). No statistically significant differences in PanIN burden were observed on the basis of sex, age or location of collected tissue (Extended Data Fig. 3b–d). The percentage of the ductal system affected

by PanIN and the number of PanIN cells were similarly higher in patients with PDAC compared to those with non-ductal disease (Extended Data Fig. 3e,f). When relative proportions of cell types constituting a tissue slab were compared, PanIN cells were significantly correlated to normal ductal cells ($P = 0.0015$) and cells in the extracellular matrix ($P < 0.001$) (Extended Data Fig. 3g,h). There was a significant inverse correlation between the relative proportions of PanIN cells and of acinar cells ($P < 0.001$), but no correlation with either fat or islet cells (Extended Data Fig. 3i–l). Together, these data suggest that PanIN is associated with fibrosis and acinar cell dropout, but further work is required to determine causality. One possibility is that inflammation predisposes to PanIN and also results in fibrosis and acinar cell dropout; another possibility is that obstruction of pancreatic ducts by PanINs leads to fibrosis and atrophy.

The above calculations are likely to represent maximum PanIN burdens, as they include all modelled PanINs regardless of their extension beyond the sampled slab. Of the 889 PanINs modelled, 592 were completely enclosed within the analysed tissue, whereas 297 were transected by a sample edge (Supplementary Table 2). The 592 completely contained PanINs represent the minimum number of histologically separate PanINs in our cohort. When only these PanINs are considered, we find a mean of 15 independent PanINs per slab (range 3–64) (Fig. 2e), leading to a more conservative mean PanIN burden of 9 PanINs per cm^3 (range 1–26 PanINs per cm^3).

From these values, we extrapolated the potential PanIN burden of an intact human pancreas. As whole-pancreas volumes for the 3D tissue modelling cohort were unknown, we first determined normal pancreas volumes from the computed tomography (CT) scans of 807 kidney donors without known pancreatic abnormalities (female, 77.51 $cm^3 \pm 17.82$ cm^3; male, 94.18 $cm^3 \pm 20.74$ cm^3 (mean ± s.d.)). All CT scans were confirmed to be radiographically normal without evidence of neoplasia. We used these values to estimate sex-specific normal prior distributions for non-neoplastic whole-pancreas volume, from which 10,000 random deviates were sampled and multiplied by PanIN burdens obtained from our 3D tissue modelling cohort. This generated a prior predictive distribution for extrapolated whole-pancreas PanIN burden in each patient. Our approach led to median extrapolated estimates of 1,021 PanINs (range 138–2,406) for a whole female pancreas and 998 PanINs (range 131–2,902) for a whole male pancreas (Fig. 2f and Supplementary Table 2). The higher estimated burden in the female pancreas mirrors the higher female PanIN burden within our 3D tissue modelling cohort (median 7.80 enclosed PanINs per cm^3 in females versus 6.68 enclosed PanINs per cm^3 in males). Even the lowest PanIN burden in our tissue cohort led to an estimate in excess of 100 PanINs per whole pancreas, underscoring the unexpectedly large number of PanINs in grossly normal pancreata.

CODA-guided NGS enables 3D genomics

Our 3D reconstructions revealed a high multifocality of PanINs within human pancreata. However, anatomical analysis alone cannot distinguish whether these spatially unconnected PanINs arose independently or via intraductal spread of a single PanIN. To assess clonal relationships and genetic heterogeneity between and within PanINs, we integrated our 3D anatomic workflow with multi-region DNA sequencing in eight additional specimens of grossly normal pancreas. Despite smaller average specimen size (mean volume 0.91 cm^3), the mean overall PanIN burden (15 PanINs per cm^3) was similar to that of the previous cohort. These 8 slabs overall yielded 109 spatially separate PanINs (median 10 per slab, range 4–43; Supplementary Table 2), of which 37 were of sufficient size for individual microdissection and NGS. Slabs were divided into five vertical regions along the *z* axis, and every region in each PanIN was microdissected and collected separately. Although most PanINs (34 out of 37) had only LG dysplasia, three PanINs that also contained HG dysplasia had their LG and HG areas isolated separately. We also microdissected regions of PDAC that were identified deep within two slabs (Supplementary Table 3). In total, we sequenced 99 regions from 37 PanINs as well as 5 regions of PDAC using a custom targeted NGS panel of 154 established cancer driver genes, including all major known drivers of pancreatic ductal neoplasia (Supplementary Table 3 and Supplementary Videos 3–10).

The mean distinct coverage for all targeted sequencing samples was 221x. Among PanINs, single-nucleotide variants (SNVs) and/or small insertion–deletion mutations (indels) were identified in well-characterized pancreatic driver genes *KRAS* (36 out of 37), *GNAS* (5 out of 37), *RNF43* (2 out of 37), *TP53* (1 out of 37), and *KDM6A* (1 out of 37). Less prevalent somatic mutations were also identified in *ERBB4* (1 out of 37), *RET* (1 out of 37), *ATRX* (1 out of 37), *STK11* (1 out of 37), *NF1* (2 out of 37), *FLT3* (1 out of 37) and *FGFR3* (2 out of 37). Each PanIN had between one and four somatic mutations identified from the targeted panel, and the two PDACs contained three to five somatic mutations. *KRAS* was the most commonly mutated gene, with all but one PanIN harbouring a mutation (36 out of 37 PanINs; 97%), consistent with previous studies[6,13,20] (Supplementary Table 3). Six PanINs harboured *KRAS* p.Q61H mutations (16%), whereas 31 (86%) had at least one mutation in codon 12 (19 *KRAS* p.G12D; 15 *KRAS* p.G12V; 5 *KRAS* p.G12R; and 1 *KRAS* p.G12C; Extended Data Fig. 4a,b). Fifty-four per cent of PanINs contained only somatic mutations in *KRAS* in the targeted sequencing analysis, with no other gene mutated in the panel.

When sufficient amounts of DNA remained, we complemented our targeted sequencing data with multi-region WES and ultra-deep mutation capsule sequencing of *KRAS* hotspots[21,22]. In total, 57 samples from 24 lesions (20 exclusively LG, 2 with LG and HG dysplasia and 2 PDAC) were analysed further (Supplementary Tables 4 and 5). For cases associated with a PDAC diagnosis, we analysed an additional PDAC sample obtained from archival tissue blocks to examine relationships between the analysed PanINs and co-occurring cancer. The mean distinct WES coverage was 392x (Supplementary Table 5). Between 1 and 4 exomes were analysed per PanIN, with a mean of 13 somatic mutations detected per region. We observed a trend for increasing mutation burden from LG PanIN to HG PanIN to PDAC, with a significantly greater burden in PDAC compared with LG PanINs (Extended Data Fig. 4c).

To characterize patterns of somatic mutations in PanINs, we combined 674 SNVs from 52 WES samples representing 22 distinct PanINs for mutational signature analysis. The number of signatures to be extracted was tested with values between 1 and 25, and the suggested number of de novo signatures by SigProfilerExtracter was 1. This single de novo signature was decomposed into two COSMIC signatures, SBS1 (28.62%) and SBS5 (71.38%). Both SBS1 and SBS5 are clock-like signatures that have been reported in normal cells and neoplastic cells of many cancer types[23,24]. SBS5 was previously identified in normal pancreatic ductal and acinar cells. However, normal pancreatic cells have also been previously reported to harbour SBS18 (damage by reactive oxygen species) and SBS40 (unknown aetiology), signatures not observed in our PanIN samples[23]. In addition to SBS1, mutational signatures such as SBS2 and SBS13 (APOBEC activity), SBS3 (defective homologous recombination DNA damage repair), SBS17A/B and SBS40 (unknown aetiology) and SBS18 (damage by reactive oxygen species) have been identified in PDAC[24]. The mutational signatures of our PanINs overlap with both normal pancreatic tissue and PDAC, but the absence of multiple PDAC-associated signatures, especially those linked to DNA damage and APOBEC activity, suggest that distinct mutational processes may drive early and late stages of pancreatic tumorigenesis.

Copy number analysis was performed for all WES samples. In PanINs, copy number gains and losses were rare. Multiple PanINs in slab 92 had gain of chromosome 8p, whereas loss of chromosome 18p was identified in PanINs from slabs 92 and 117 (Extended Data Fig. 5). Gain of chromosome 1q was present in two different PanINs containing both LG and HG components (slab 92 PanIN D and slab 104 PanIN D; Extended Data Fig. 5), with the highest gain in the HG component of slab

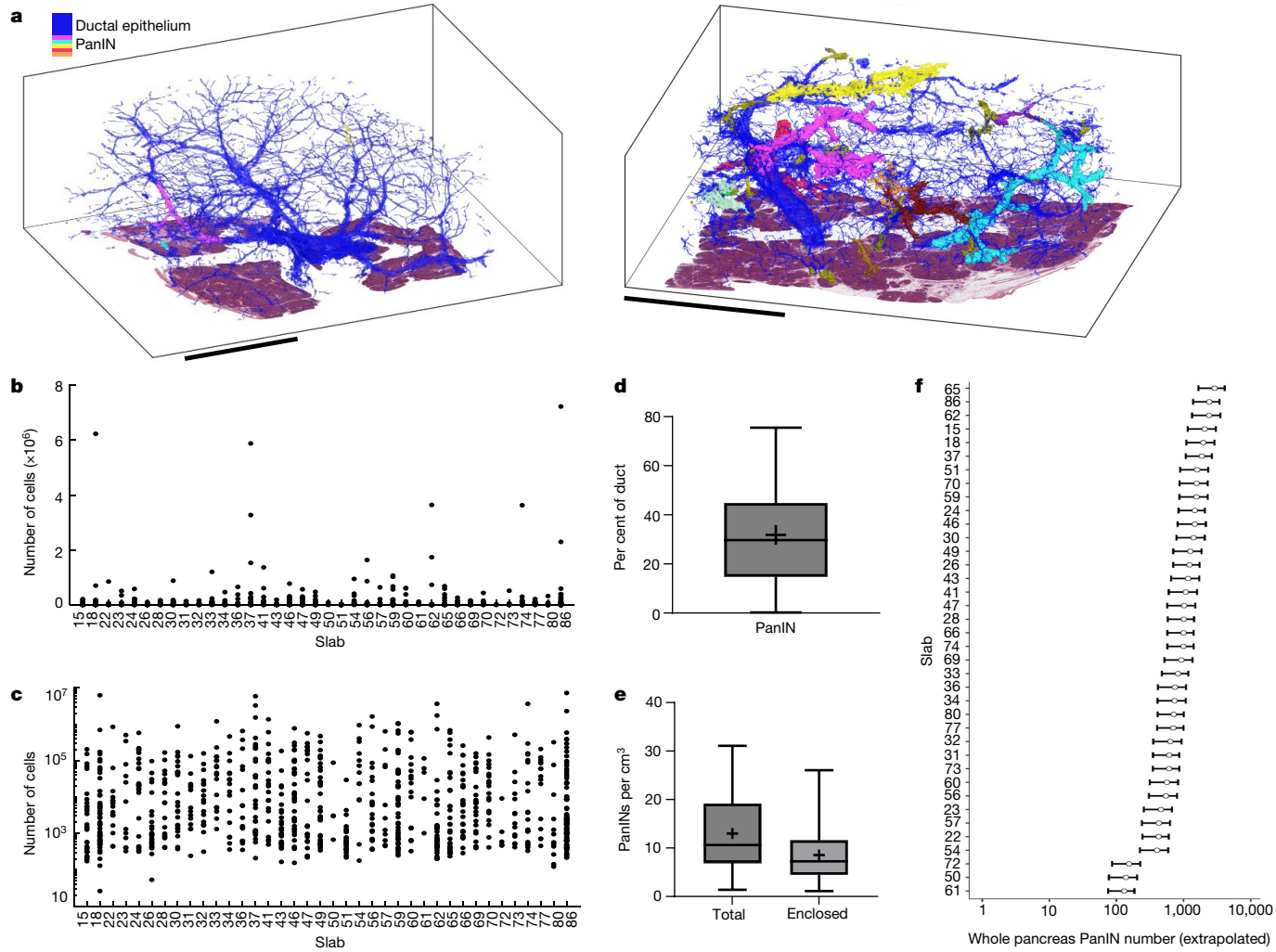


Fig. 2 | Quantified microanatomical features of 3D-rendered PanINs. **a**, Representative 3D models of two human pancreatic tissue slabs rendered by CODA. Blue represents normal pancreatic ducts; each spatially distinct PanIN is depicted in a different colour. Scale bars, 1 cm. **b**, PanIN cell counts for all PanINs. The number of cells in each PanIN is shown for each numbered tissue slab. **c**, PanIN cell counts for the subset of PanINs in **b** that containing fewer than 1,000,000 cells. **d**, Percentage of the ductal system affected by PanINs in each tissue slab. **e**, PanIN burden (the number of PanINs per cm^3 tissue) for each slab, calculated according the total number of PanINs and by PanINs completely enclosed within the analysed tissue. **d**,**e**, The cross indicates the mean of $n = 38$ independent samples. Whiskers extend to minimum and maximum values, box edges show first and third quartiles, and the centre line denotes the median. **f**, Extrapolation of PanIN burden over the whole pancreas, calculated separately for each tissue slab. Error bars correspond to 95% credible intervals of the predictive distribution of the whole-pancreas PanIN numbers.

104. All other copy number alterations were limited to single PanINs. Of note, the HG component of PanIN D in slab 92 harboured loss of *TP53*-containing chromosome 17p as well as a point mutation in *TP53*, indicating bi-allelic alteration of this key tumour suppressor gene. In contrast to the PanIN samples, all PDAC samples contained numerous copy number gains and losses, consistent with previous studies[25–27]. The progression from a relatively low number of copy number alterations in LG PanINs to increasing prevalence in HG PanIN and PDAC suggests that such alterations occur later in carcinogenesis.

PanINs arise via distinct genetic events

All eight pancreatic tissue slabs that underwent 3D genomic analysis contained multiple spatially unconnected PanINs. With our unique experimental approach, comparison of somatic mutations between these PanINs can delineate their shared or independent clonal origin (Figs. 3 and 4 and Extended Data Figs. 6 and 7). For example, multi-region targeted sequencing showed a distinct *KRAS* hotspot mutation for each of the four histologically separate PanINs in slab 104 (Fig. 3a), with no additional shared mutations in the targeted panel (Fig. 3b). In the two PanINs (C and D) with both LG and HG dysplasia, the LG and HG components within a PanIN shared the same *KRAS* hotspot mutation, whereas mutations were not shared between discontinuous PanINs regardless of shared grade of dysplasia. Overall, our results suggested that spatially distinct PanINs arose independently.

With few driver gene mutations in PanINs, drawing robust conclusions from targeted sequencing can be challenging. We therefore compared WES data for 7 samples from 3 PanINs in slab 104. Although we identified a mean of 19 somatic mutations per PanIN, none were shared between lesions that were spatially separate. By contrast, contiguous regions within a single PanIN shared on average 14 somatic mutations, indicating a common clonal origin. Slab 104 also contained a small focus of PDAC that we analysed by targeted and exome sequencing. Whereas its *KRAS* mutation was shared with PanIN D, the PDAC also harboured unshared somatic mutations in *CDKN2A*, *TP53* and *SMAD4*. PanIN D also had an exclusive *ERBB4* mutation that was absent in the PDAC. We hypothesized that PanIN D and PDAC in slab 104 were independent neoplasms that shared a common *KRAS* hotspot mutation by

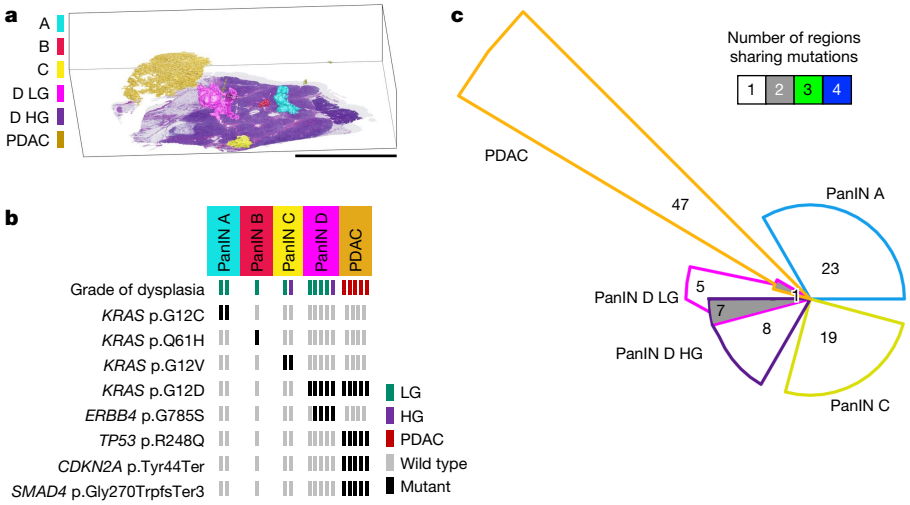


Fig. 3 | 3D model and NGS results for slab 104. Each PanIN is represented by the same colour throughout this figure. **a**, CODA-rendered 3D model of grossly normal pancreatic tissue with multiple spatially distinct PanINs. Scale bar, 1 cm. **b**, Mutation chart with targeted sequencing results. Each row represents a distinct somatic mutation. Columns represent spatially distinct PanINs and bars represent vertical regions within a PanIN. **c**, Chow–Ruskey plot of WES results. Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group.

chance. The WES data, which showed no common somatic mutations between PanIN D and PDAC other than the *KRAS* hotspot, confirmed our hypothesis of distinct clonal origins (Fig. 3c).

In general, most PanINs appear to represent independent clones. Among the cohort, another five PanINs shared no genetic alterations with any other PanIN within the same slabs (Figs. 3–5). Six PanINs shared only their *KRAS* hotspot mutations, with numerous unshared mutations in WES, which we interpret as independent PanINs acquiring common hotspot mutations by chance (Fig. 5 and Extended Data Fig. 7e–h). Even when multiple spatially distinct PanINs existed in close proximity, their mutations reflected independent origins. In addition, whenever LG and HG dysplasia constituted a single PanIN, they shared driver and passenger gene mutations, indicating HG PanINs arose from contiguous LG PanIN (Figs. 3 and 5). Across all three HG PanINs, only one HG-to-LG difference in driver gene mutation was detected—a clonal *TP53* mutation in a HG PanIN that was found to be subclonal in its LG counterpart. This highlights the necessity of 3D modelling to delineate distinct lesions and to discern LG PanINs that are contiguous with HG dysplasia.

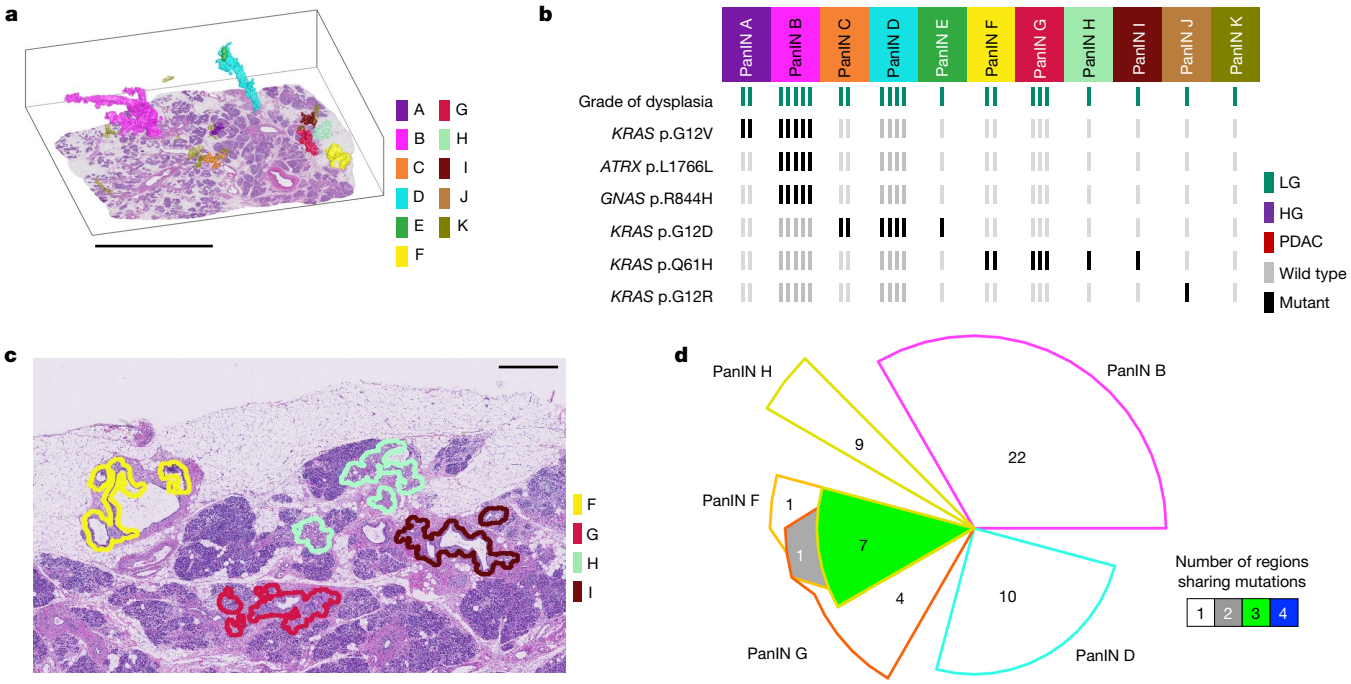


Fig. 4 | 3D model and NGS results for slab 98. Each PanIN is represented by the same colour throughout this figure. **a**, CODA-rendered 3D model from grossly normal pancreatic tissue with multiple spatially distinct PanINs. Scale bar, 1 cm. **b**, Mutation chart with targeted sequencing results. Each row represents a distinct somatic mutation. Columns represent spatially distinct PanINs and bars represent vertical regions within a PanIN. **c**, Representative H&E image of PanINs sharing *KRAS* p.Q61H mutations. Each lesion is anatomically distinct and separated from others by histologically normal pancreatic duct. Scale bar, 1 mm. **d**, Chow–Ruskey plot of WES results. Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group.

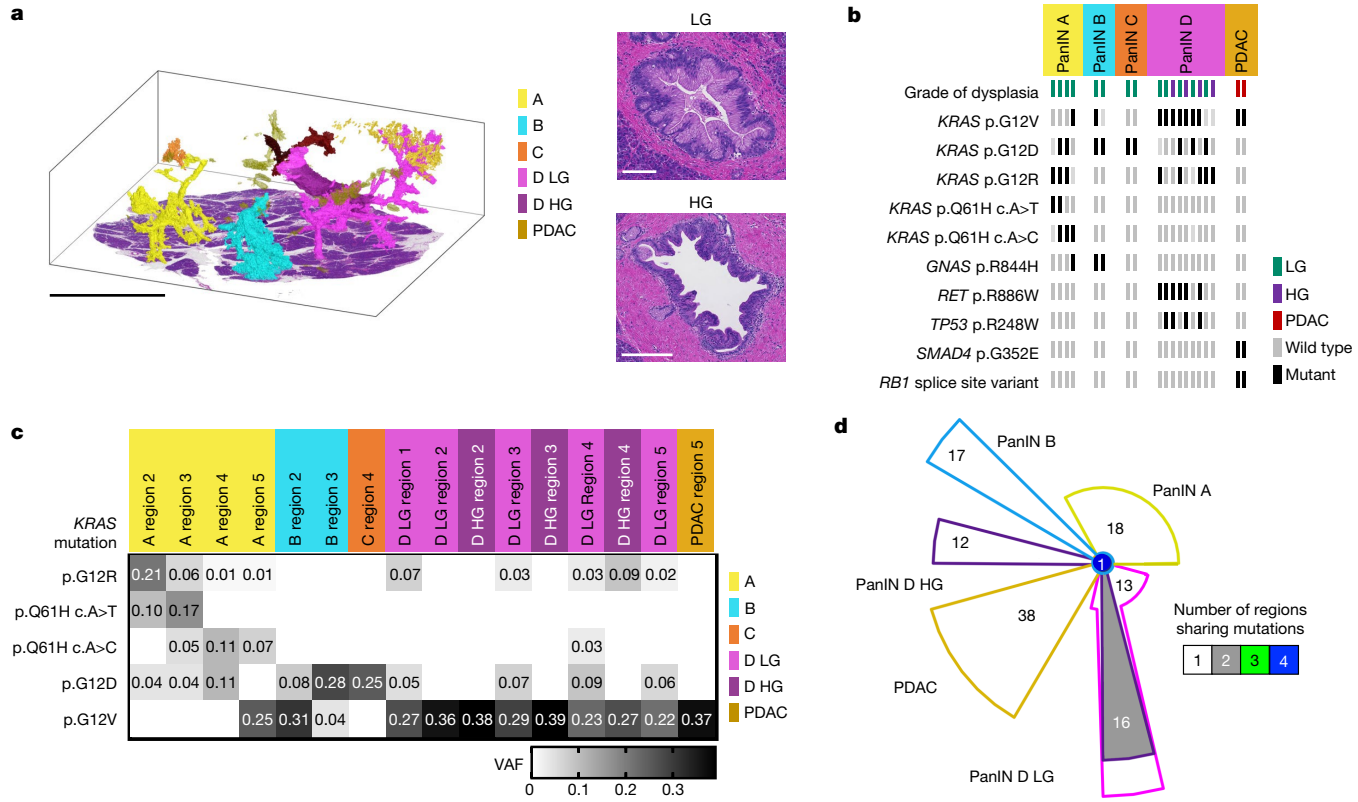


Fig. 5 | 3D model and NGS results for slab 92. Each PanIN is represented by the same colour throughout this figure. **a**, Left, CODA-rendered 3D model of grossly normal pancreatic tissue with multiple spatially distinct PanINs (scale bar, 1 cm). Right, representative H&E images of LG (top) and HG (bottom) PanINs (scale bars, 100 µm). **b**, Mutation chart with targeted sequencing results. Each row represents a distinct somatic mutation. Columns represent spatially distinct PanINs and bars represent vertical regions within a PanIN. **c**, Heat map of VAFs from ultra-deep mutation capsule sequencing of hotspots in *KRAS*. Rows represent distinct *KRAS* oncogenic hotspot mutations. Columns represent PanINs. **d**, Chow–Ruskey plot of WES results. Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group.

Rare intrapancreatic spread of LG PanIN

Of the 11 unconnected PanINs in slab 98, four PanINs (F, G, H and I) that clustered within 4 mm of one another shared *KRAS* p.Q16H mutations (Fig. 4a–c). Notably, PanINs H and I were entirely contained, excluding connection of the lesions outside of the analysed tissue (Fig. 4a). We performed WES on PanINs F, G and H. PanINs F, G and H shared seven somatic mutations, although they were separated by regions of histologically normal non-neoplastic pancreatic duct (Fig. 4d). Similarly, in slab 114, two LG PanINs (A and B) shared seven somatic mutations, including a *KRAS* hotspot mutation. They were at most 1 mm apart but not contiguous; PanIN B was completely contained within the slab and separated from PanIN A by normal non-neoplastic duct (Extended Data Fig. 7a–d). Although rare, our 3D genomic mapping demonstrates that LG PanIN cells are capable of travelling short distances within the pancreas, establishing physically separate but genetically related lesions that share multiple somatic mutations beyond oncogenic hotspots. Considering that PanINs are by definition limited to the ductal epithelium and that all PanINs analysed in this study were found in the ductal system, intraductal spread of LG PanINs is the most likely explanation of our findings. Nonetheless, future studies could assess other possible routes of spread, including via the stroma or vasculature.

Some PanINs may have polyclonal origins

Separate sequencing of multiple regions within individual PanINs enables thorough assessment of intra-PanIN genetic heterogeneity. Whereas 81% of PanINs had 1 clonal *KRAS* mutation, 7 PanINs (19%) from four slabs (slabs 92, 114, 116 and 117) had multiple *KRAS* hotspot mutations. In slab 92 (Fig. 5a), multiple *KRAS* mutations were found in 3 of the 4 PanINs. PanIN A contained 5 different *KRAS* hotspot mutations (2–4 per region); B contained 2. In PanIN D, LG regions possessed 3 different *KRAS* mutations. Of note, this heterogeneity was markedly reduced in the HG components. We note that a *TP53* p.R248W mutation specific to a single LG region was also present in all HG regions, demonstrating expansion of a *TP53*-mutant clone at progression to HG PanIN (Fig. 5b). Loss of 17p was identified only in HG components of PanIN D in copy number analysis (Extended Data Fig. 5), further illustrating the alteration of *TP53* with progression of this PanIN. Intra-PanIN *KRAS* mutational heterogeneity was confirmed by mutation capsule analysis for PanINs A and D, which revealed 5 and 4 distinct *KRAS* mutations, respectively (Fig. 5c). As in other tissue slabs, no somatic mutations were shared by spatially separate PanINs (Fig. 5d). Because *KRAS* mutations are considered very early events in PanIN development, the presence of multiple *KRAS* mutations within a single spatially continuous PanIN suggests a polyclonal origin for some PanINs.

To evaluate the hypothesis that multiple *KRAS* mutations reflect polyclonal origin, we reconstructed the evolutionary history of two PanINs with multiple *KRAS* mutations using PICTograph[28] (Fig. 6a,b). PICTograph is a Bayesian hierarchical model for building clone trees from multi-region sequencing data. By modelling uncertainty in assigning mutations, PICTograph can visualize the most probable ancestral relationships between subclones. To ensure inclusion of passenger mutations from each clone, we analysed the PanINs whose multiple *KRAS* mutations exceeded our WES threshold for mutation calling (10%). In both clone trees, every *KRAS* mutation delineated a distinct

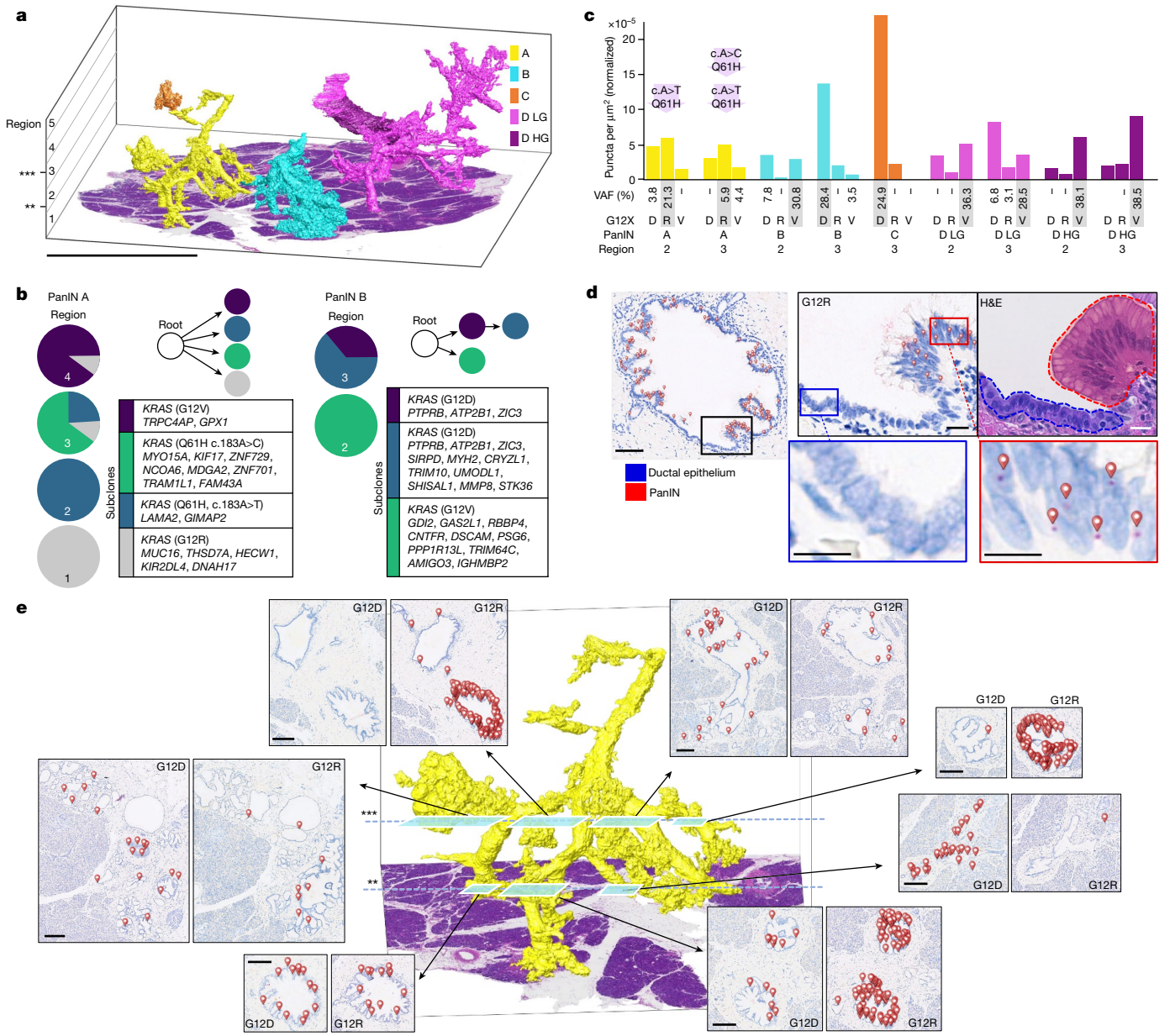


Fig. 6 | Assessment of PanINs with multiple *KRAS* hotspot mutations for slab 92. a, CODA-rendered 3D model of four multi-region sequenced PanINs in slab 92. Asterisks mark the level of sections used for RNA ISH in relation to region boundaries on the *z* axis. Scale bar, 1 cm. **b**, Most probable subclonal evolutionary relationships inferred by PICTograph from multi-region WES data in PanINs A and B. Each colour represents a distinct subclone with assigned mutations in table. Wheels visualize distributions of neoplastic cell fractions of subclones in each PanIN region. **c**, Quantification of *KRAS* G12 point mutation-specific RNA ISH for each PanIN. Signal for the entire PanIN area was normalized to background probe labelling in pancreatic islets on the same slide. One tissue section each from region 2 (containing areas of PanINs A, B and D) and region 3 (containing areas of PanIN A, B, C and D) of slab 92 were analysed for point mutations in *KRAS* G12. For PanIN D, LG and HG areas were quantified and depicted separately. Corresponding VAFs from targeted NGS of each PanIN region are shown below each column, with the most common variant highlighted in grey. Lavender arrows mark regions where *KRAS* Q61H mutations were found by NGS. **d**, Left, example of a histologically normal ductal epithelium continuous with PanIN, found incidentally in a slide labelled by RNA ISH for *KRAS* G12R and adjacent H&E staining (scale bar, 100 µm). Top right, $KRAS^{G12R}$ targets, detected as red puncta and marked with red flags, are present in PanIN cells (outlined in red) but not in adjacent ductal epithelia (outlined in blue); outlined regions are expanded below (scale bars, 10 µm). **e**, Differential spatial distribution of G12 mutations mapped within PanIN A by BaseScope RNA ISH. Red flags mark individual red chromogen puncta. One slide per G12 mutant probe was analysed for the two regions indicated with asterisks: ***slide from region 3; **slide from region 2. Scale bars, 200 µm.

clone that shared no somatic mutations with other clones. The lack of common neoplastic ancestor indicates that each *KRAS* mutation represents a genetically independent clone. We also observed that between some regions of the same PanIN, the variant allele frequencies (VAFs) of different *KRAS* alleles had reciprocal relationships. For example, in PanIN B of slab 92, G12V has the highest VAF (24%) among *KRAS* mutations in region 2, in contrast to a lower G12D (5%). Region 3 shows the inverse: the highest VAF is for G12D (22%) and G12V (2%) is minimal (Fig. 5c). This implies that neither *KRAS* mutant can be a subclonal descendent of the other, furthering the likelihood that they occur in independent clones.

To further confirm that distinct *KRAS* mutations occurred in different populations, we mapped the spatial location of *KRAS*-mutant cells in each PanIN by *KRAS* G12 point mutation-specific RNA in situ

hybridization (ISH). We interrogated one to two regions per slab with individual BaseScope ISH probes for G12D, G12V, G12R and G12C. BaseScope probes discriminate between single-nucleotide substitutions with very high specificity, as shown in previous in situ mutation detection and subclone mapping studies in formalin-fixed and paraffin-embedded (FFPE) cancer tissue[29]. Quantification of ISH signal in each PanIN confirmed the mutations identified via microdissection and sequencing (Fig. 6c,d and Extended Data Fig. 8). Within individual PanINs with multiple *KRAS* mutations identified by sequencing, the spatial distributions of different *KRAS*-mutant probes had very little overlap, as illustrated by PanIN A of slab 92: whereas G12R was the predominant target in several places, another area instead had prominent G12D signal (Fig. 6e). Other parts lacked pronounced labelling for all assayed G12 probes, which we speculate are sites of the two different *KRAS* Q61H mutations also detected in NGS (Fig. 6c). Mapping these areas onto the CODA model revealed them as distant branches in 3D despite their apparent proximity in 2D slides. Similar patterns of differentially localized mutation signals were also apparent in the multi-*KRAS* PanINs B and D of slab 92. Together, evolutionary and in situ analyses favour a polyclonal origin for PanINs with multiple *KRAS* mutations. Of note, our ISH results did not support mutant *KRAS* expression in histologically normal ductal epithelium. Where normal ductal cells were found contiguous with PanIN, we observed a steep decrease in mutant probe signal at the PanIN-normal duct boundary. ISH labelling for all mutant probes in normal ducts was on par with tissue background levels, despite robust labelling for the universal *KRAS* probe (Fig. 6d and Extended Data Fig. 9). This sharp contrast with the prominent mutant *KRAS* signal in the immediately adjacent PanINs indicates that histologically normal ductal cells do not harbour *KRAS* mutations.

Discussion

Although PanINs are the most common precursors to PDAC, their small size has impeded extensive study of their prevalence, spatial landscape, and genetic heterogeneity. Our 3D methodology enabled quantitative analysis of PanINs in human pancreatic tissue. Our results demonstrate that an intact adult human pancreas can contain hundreds of PanINs. This high burden is particularly striking considering the relatively low incidence of pancreatic cancer, suggesting that individual PanINs have extremely low risks of progression. These results have important implications for early detection and intervention to prevent pancreatic cancer. Several possibilities could explain this profusion of PanINs, including some predisposing event in the pancreas that encouraged PanIN development, or an initial PanIN that potentiated the development of subsequent PanINs. Our current data cannot favour any one scenario, and the mechanism driving multifocal PanIN initiation remains an important future question. Notably, our in situ assessment of mutant *KRAS* did not provide evidence for *KRAS* mutations in histologically normal epithelial cells, though additional analyses using sensitive mutation detection techniques in fresh tissue samples are necessary to confirm this finding.

Such abundance of precancers in the pancreas runs counter to precancers in other frequently studied organs. For example, although colonic adenomas can also be multifocal lesions, typically fewer than five adenomas are identified throughout the entire colon in the absence of inherited cancer predisposition syndromes[30–33]. However, colonic adenomas are initiated by bi-allelic tumour suppressor gene loss whereas PanINs are initiated by mono-allelic oncogene activation. In this respect, acquired naevi of the skin, possible melanoma precursors initiated by mono-allelic oncogene mutation, may be more similar to PanINs[34–36]. Yet despite skin lesions being readily accessible for quantification, the average individual is reported to have only 10–60 benign naevi[37,38]. These relatively low precursor burdens contrast with our extrapolated PanIN numbers in the hundreds. If the high burden of pancreatic precancer is truly unique, our findings may call for new paradigms in early pancreatic neoplasia. However, such questions of organ-specific differences in tumorigenesis cannot be settled without comprehensive 3D assessment of precancer burden and multifocality in other organs. We envision that our workflow, in which CODA is integrated with genetic analysis, will facilitate future 3D analyses of microscopic precursor lesions in various organs. Moreover, our methodology can be further expanded to incorporate high-dimensional spatial profiling platforms, enabling transcriptomic and proteomic assessment of precancers and their microenvironment.

Using our 3D genomic mapping approach, we identified substantial intra-PanIN heterogeneity of driver and passenger gene mutations, including seven PanINs with multiple *KRAS* mutations. Because *KRAS* hotspot mutations occur early in PanIN development, multiple *KRAS* mutations in one lesion suggests a polyclonal origin. We provide multiple lines of evidence that support this polyclonality—both computational evolutionary reconstruction and in situ detection of *KRAS* hotspot mutations indicate that distinct *KRAS* mutations occur in different cell populations. Polyclonality has also been described in pancreatic IPMNs, in which multiple *KRAS* mutations delineated independent clones within the same precancerous cyst[15]. In pancreatic neoplasia, intralesional heterogeneity of driver gene mutations appears to be a feature of precursor lesions, whereas a single *KRAS* mutation is almost universal in PDAC, even in distant metastases[15–19]. Notably, the distribution of *KRAS* hotspot mutations found in PanINs resembled those reported in PDAC (Extended Data Fig. 4), suggesting that selection for *KRAS* hotspot mutations occurs at PanIN initiation, not progression[39]. Although the phenotype providing the critical advantage for *KRAS* mutants in early tumorigenesis remains unclear[40], our data affirm the widespread expansion of *KRAS*-mutant clones throughout otherwise normal pancreas in humans.

Although this study illuminates the multifocality and genetic heterogeneity of human PanINs, there are limitations to our cohort. First, grossly normal tissue collected during tumour resections may not represent truly 'normal' pancreas. Nevertheless, the majority of patients underwent resection of neoplasms that did not involve the ductal system, which may better reflect PanIN burden and genetic features in the general population. Similarly, our patient age range (mean 66, range 45–87) also does not represent the general population, which may account for the lack of correlation between age and PanIN burden in our cohort, contrary to other reports[41]. Assessment of PanIN burden in non-diseased pancreata across the age spectrum is an important future direction. Whereas 38 slabs were 3D modelled, integrated 3D genomic analysis was limited to 8 slabs. Yet with multi-region processing, the resultant 99 NGS samples yielded unprecedented resolution of the mutational landscape of 37 discrete PanINs. Of these, 34 PanINs contained only LG dysplasia, in accordance with the higher prevalence of LG PanINs than HG PanINs in randomly sampled pancreata[41]. Although one of our PanINs suggests a role for *TP53* mutation, a dedicated HG PanIN cohort is necessary to investigate molecular alterations driving PanIN progression. Finally, genetic origins could not be resolved for a few PanINs owing to insufficient DNA quantities, a persisting challenge for sequencing very small precancers.

As one of the most extensive studies of human PanINs to date, we provide new insights into early pancreatic tumorigenesis. Quantitative 3D modelling revealed a surprising abundance of anatomically distinct PanINs. Our powerful integration of 3D modelling and genome analysis demonstrated that most PanINs are truly multifocal lesions arising from independent clones, and that intraductal spread of LG PanINs is a rare phenomenon. Our meticulous multi-region sequencing approach also unveiled rich genetic heterogeneity within individual PanINs, including multiple *KRAS* mutations suggestive of intra-lesion polyclonality. These findings lay a foundation for future investigations towards mechanisms underlying PanIN multifocality and risk stratification of PanIN progression. Together, our work describes the remarkable spatial and

genetic multifocality of human PanINs and underscores the necessity of integrating 3D microanatomy to accurately resolve genetic origins.

Online content

Any methods, additional references, Nature Portfolio 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/s41586-024-07359-3.

1. Hruban, R. H., Goggins, M., Parsons, J. & Kern, S. E. Progression model for pancreatic cancer. *Clin. Cancer Res.* **6**, 2969–2972 (2000).
2. Siegel, R. L., Miller, K. D., Hannah, F. E. & Jemal, A. Cancer statistics, 2022. *CA* **72**, 7–33 (2022).
3. Ryan, D. P., Hong, T. S. & Bardeesy, N. Pancreatic adenocarcinoma. *N. Engl. J. Med.* **371**, 1039–1049 (2014).
4. Takaori, K., Kobashi, Y., Matsusue, S., Matsui, K. & Yamamoto, T. Clinicopathological features of pancreatic intraepithelial neoplasias and their relationship to intraductal papillary-mucinous tumors. *J. Hepatobiliary Pancreat. Surg.* **10**, 125–136 (2003).
5. Hruban, R. H. et al. An illustrated consensus on the classification of pancreatic intraepithelial neoplasia and intraductal papillary mucinous neoplasms. *Am. J. Surg. Pathol.* **28**, 977–987 (2004).
6. Kanda, M. et al. Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. *Gastroenterology* **142**, 730–733.e739 (2012).
7. Hong, S. M. et al. Genome-wide somatic copy number alterations in low-grade PanINs and IPMNs from individuals with a family history of pancreatic cancer. *Clin. Cancer Res.* **18**, 4303–4312 (2012).
8. Andea, A., Sarkar, F. & Adsay, V. N. Clinicopathological correlates of pancreatic intraepithelial neoplasia: a comparative analysis of 82 cases with and 152 cases without pancreatic ductal adenocarcinoma. *Mod. Pathol.* **16**, 996–1006 (2003).
9. Makohon-Moore, A. P. et al. Precancerous neoplastic cells can move through the pancreatic ductal system. *Nature* **561**, 201–205 (2018).
10. Kiemen, A. L. et al. CODA: quantitative 3D reconstruction of large tissues at cellular resolution. *Nat. Methods* **19**, 1490–1499 (2022).
11. Hosoda, W. et al. Genetic analyses of isolated high-grade pancreatic intraepithelial neoplasia (HG-PanIN) reveal paucity of alterations in *TP53* and *SMAD4*. *J. Pathol.* **242**, 16–23 (2017).
12. Opitz, F. V., Haeberle, L., Daum, A. & Esposito, I. Tumor microenvironment in pancreatic intraepithelial neoplasia. *Cancers* **13**, 6188 (2021).
13. Hata, T. et al. Genome-wide somatic copy number alterations and mutations in high-grade pancreatic intraepithelial neoplasia. *Am. J. Pathol.* **188**, 1723–1733 (2018).
14. Chhoda, A., Lu, L., Clerkin, B. M., Risch, H. & Farrell, J. J. Current approaches to pancreatic cancer screening. *Am. J. Pathol.* **189**, 22–35 (2019).
15. Fischer, C. G. et al. Intraductal papillary mucinous neoplasms arise from multiple independent clones, each with distinct mutations. *Gastroenterology* **157**, 1123–1137.e1122 (2019).
16. Wu, J. et al. Recurrent *GNAS* mutations define an unexpected pathway for pancreatic cyst development. *Sci. Transl. Med.* **3**, 92ra66 (2011).
17. Felsenstein, M. et al. IPMNs with co-occurring invasive cancers: neighbours but not always relatives. *Gut* **67**, 1652–1662 (2018).
18. Connor, A. A. et al. Integration of genomic and transcriptional features in pancreatic cancer reveals increased cell cycle progression in metastases. *Cancer Cell* **35**, 267–282.e267 (2019).
19. Makohon-Moore, A. P. et al. Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer. *Nat. Genet.* **49**, 358–366 (2017).
20. Shi, C. et al. *KRAS2* mutations in human pancreatic acinar-ductal metaplastic lesions are limited to those with PanIN: implications for the human pancreatic cancer cell of origin. *Mol. Cancer Res.* **7**, 230–236 (2009).
21. Qu, C. et al. Detection of early-stage hepatocellular carcinoma in asymptomatic HBsAg-seropositive individuals by liquid biopsy. *Proc. Natl Acad. Sci. USA* **116**, 6308–6312 (2019).
22. Wang, P. et al. Simultaneous analysis of mutations and methylations in circulating cell-free DNA for hepatocellular carcinoma detection. *Sci. Transl. Med.* **14**, eabp8704 (2022).
23. Alexandrov, L. B. et al. The repertoire of mutational signatures in human cancer. *Nature* **578**, 94–101 (2020).
24. Moore, L. et al. The mutational landscape of human somatic and germline cells. *Nature* **597**, 381–386 (2021).
25. Aguirre, A. J. et al. High-resolution characterization of the pancreatic adenocarcinoma genome. *Proc. Natl Acad. Sci. USA* **101**, 9067–9072 (2004).
26. Murphy, S. J. et al. Integrated genomic analysis of pancreatic ductal adenocarcinomas reveals genomic rearrangement events as significant drivers of disease. *Cancer Res.* **76**, 749–761 (2016).
27. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
28. Zheng, L., Niknafs, N., Wood, L. D., Karchin, R. & Scharpf, R. B. Estimation of cancer cell fractions and clone trees from multi-region sequencing of tumors. *Bioinformatics* **38**, 3677–3683 (2022).
29. Baker, A.-M. et al. Robust RNA-based in situ mutation detection delineates colorectal cancer subclonal evolution. *Nat. Commun.* **8**, 1998 (2017).
30. Amano, T. et al. Number of polyps detected is a useful indicator of quality of clinical colonoscopy. *Endosc. Int. Open* **6**, E878–E884 (2018).
31. Müller, A. D. & Sonnenberg, A. Prevention of colorectal cancer by flexible endoscopy and polypectomy. A case-control study of 32,702 veterans. *Ann. Intern. Med.* **123**, 904–910 (1995).
32. Rohan, T. E., Henson, D. E., Franco, E. L. & Albores-Saavedra, J. in *Cancer Epidemiology and Prevention* (eds Schottenfeld, D. & Fraumeni, J. F.) 21–46 (Oxford Univ. Press, 2006).
33. Williams, A. R., Balasooriya, B. A. & Day, D. W. Polyps and cancer of the large bowel: a necropsy study in Liverpool. *Gut* **23**, 835–842 (1982).
34. Pollock, P. M. et al. High frequency of BRAF mutations in nevi. *Nat. Genet.* **33**, 19–20 (2003).
35. Kumar, R., Angelini, S., Snellman, E. & Hemminki, K. BRAF mutations are common somatic events in melanocytic nevi. *J. Invest. Dermatol.* **122**, 342–348 (2004).
36. Ichii-Nakato, N. et al. High frequency of BRAFV600E mutation in acquired nevi and small congenital nevi, but low frequency of mutation in medium-sized congenital nevi. *J. Invest. Dermatol.* **126**, 2111–2118 (2006).
37. Cooke, K. R., Spears, G. F. & Skegg, D. C. Frequency of moles in a defined population. *J. Epidemiol. Community Health* **39**, 48–52 (1985).
38. Schäfer, T., Merkl, J., Klemm, E., Wichmann, H. E. & Ring, J. The epidemiology of nevi and signs of skin aging in the adult general population: results of the KORA-survey 2000. *J. Invest. Dermatol.* **126**, 1490–1496 (2006).
39. Bryant, K. L., Mancias, J. D., Kimmelman, A. C. & Der, C. J. KRAS: feeding pancreatic cancer proliferation. *Trends Biochem. Sci* **39**, 91–100 (2014).
40. Chen, Z., Chen, M., Fu, Y. & Zhang, J. The KRAS signaling pathway's impact on the characteristics of pancreatic cancer cells. *Pathol. Res. Pract.* **248**, 154603 (2023).
41. Matsuda, Y. et al. The prevalence and clinicopathological characteristics of high-grade pancreatic intraepithelial neoplasia: autopsy study evaluating the entire pancreatic parenchyma. *Pancreas* **46**, 658–664 (2017).

Methods

Specimen acquisition

This study was approved by the Institutional Review Board at The Johns Hopkins Hospital, and informed consent was obtained from all patients. Thick slabs of grossly normal pancreatic tissue were collected from surgical pancreatectomy specimens and consecutively assigned a slab number. For 3D modelling, a cohort was selected to investigate PanINs arising in otherwise histologically normal blocks of pancreatic parenchyma. Following histologic review by an expert pancreatic pathologist, slabs containing significant fibrosis, atrophy or PDAC in the analysed tissue slab were excluded. In addition, slabs with a diagnosis of IPMN were excluded due to the histological similarities between PanIN and IPMN and the propensity of IPMN to involve smaller ducts, complicating the reliable distinction of the two precursor lesions[5]. With these criteria, we selected 38 slabs of pancreas tissue from 38 patients for 3D anatomical modelling. Twelve of these patients had PDAC elsewhere in their pancreas. The remaining 26 patients had other pancreatic or small intestinal neoplasms elsewhere in their resection specimens, including pancreatic neuroendocrine tumours, serous cystadenomas, distal common bile duct adenocarcinomas, metastatic carcinomas from other organ sites, mucinous cystic neoplasms, tubulovillous adenomas of the duodenum, ampullary tumours and lymphoepithelial cysts (Supplementary Table 1).

A separate cohort of grossly normal pancreatic tissue slabs underwent combined anatomic and genomic analyses. Because of the unique sectioning scheme (details below) used for 3D modelling and NGS, which included large numbers of specialized membrane slides for microdissection, inclusion criteria for this cohort were based on the first histologic slide cut from the surface of each slab, which was assessed prior to full sectioning. Specifically, PanIN was to be present, and PDAC and significant fibrosis were to be absent on the first slide sectioned. Any patients with a clinical diagnosis of IPMN were excluded. Using these criteria, 8 additional slabs of tissue were selected from eight patients for 3D modelling and NGS—these patients did not overlap with the 38 patients analysed by anatomic modelling only. Three of these eight patients had a pathological diagnosis of PDAC elsewhere in their pancreas. The remaining five of the eight patients had neoplasms not affecting the pancreatic ductal system, including well-differentiated pancreatic neuroendocrine tumours (2), serous cystadenoma (1), distal common bile duct adenocarcinoma (1) and colon cancer metastatic to the pancreas (1).

Sample processing

Each of the 46 slabs of collected tissue were processed as for FFPE, followed by complete serial sectioning at 5 μm. Every third slide was stained with H&E and digitized at 20× magnification for 3D modelling. As previously described, skipping staining on two out of every three slides does not lead to any significant loss in microanatomical information[10]. Sample processing yielded a mean of 1,288 slides per block (range 679–1,703) and mean tissue volume of 2.03 cm^3 (range 0.94–3.62 cm^3). All PanIN lesions present on every 50th slide were manually annotated using Aperio ImageScope to verify accuracy of the generated 3D models[10]. For the eight slabs undergoing NGS, every third slide was cut onto membrane slides (Zeiss Membrane Slide 1.0 PEN; Carl Zeiss) for laser capture microdissection (LCM).

3D reconstruction of serially sectioned H&E sections of human pancreas

Using the previously validated method CODA, we converted the serially sectioned histological slides of human pancreas tissue into digital 3D maps of pancreas microanatomy for all 46 slabs[10]. In brief, the CODA workflow can be split into four steps: image registration, single-cell detection, tissue segmentation, and 3D visualization. For a pair of images, the registration maximizes the 2D cross-correlation of pixel intensity to align all images and correct for tissue rotation, translation, folding, splitting and stretching. A cell-detection algorithm is used to quantify the cellularity of components via detection of 2D intensity peaks in the haematoxylin channel of the H&E images. Deep learning is next used to create microanatomical labels from the histological images. The trained algorithm was used to label, to a resolution of 2 μm, 8 microanatomical structures in histological images of the pancreas: islets of Langerhans, normal ductal epithelium, vasculature, fat, acinar tissue, extracellular matrix (ECM), PanIN and PDAC. By combining the cell detection and the tissue-level semantic segmentation, we estimated the tissue type of all cells in the 3D samples. Samples containing regions of cancer or lymph nodes were labelled using a separate model trained to additionally recognize these components. The image registration, cell detection and tissue segmentation are integrated to create 3D reconstructions of pancreas microanatomy at large scale (up to multi-cm^3), while maintaining single-cell resolution.

All 46 thick slabs of pancreas tissue were reconstructed using CODA. The 3D datasets used for quantification and visualizations were subsampled from the classified resolution of 2 μm per pixel per image to an isometric resolution of 12 × 12 × 12 μm. The cell-detection algorithm and tissue-level deep learning segmentation were validated independently. The cell-detection algorithm was validated through comparison of automatic cell counts to manually generated nuclear coordinates by two independent researchers on five regions of randomly selected images. Cell-detection parameters were adjusted until to maximize the precision and recall obtained for each validation image (see per image precision and recall compared to two sets of manually generated coordinates in Extended Data Fig. 1a). Overall, we achieved an average 90.2% precision and 92.0% recall for the five images. For the deep learning, all available annotated images were collected: 90% of images were used for training and 10% of images were used as an independent testing set of model accuracy across unseen images in the analysed cohort. Models were deemed acceptable when they achieved >90% per class precision and recall on the independent testing images. Per class accuracy, precision, and recall are presented in Extended Data Fig. 1b.

Validation of PanIN detection in samples

All objects labelled as PanIN by the segmentation algorithm were post-processed to ensure that no non-neoplastic regions were counted as PanIN. In the 3D volume matrix saved at a resolution of 12 × 12 × 12 $μm^3$, objects of fewer than 20 voxels and objects present on fewer than three histological sections were eliminated. The histology of all remaining objects was manually assessed to determine whether they were PanIN. The 3D bounding box of each object was determined. The regions of the serial H&E images contained in this bounding box were extracted from the registered 5× magnification image stack and were saved as a separate image stack (Supplementary Video 1). These stacks were manually viewed using FIJI ImageJ[42]. A matrix was created in Microsoft Excel containing the PanIN identifier (labelled A to Z), and the true determined label of the region. All regions determined to be false positive PanIN labelling during manual validation were corrected in the 3D matrices. False negatives were assessed through comparison of the deep learning labelled structures to manual pathologist-guided annotations of neoplastic tissue on one in every 50 sections throughout the samples.

Colour-coded labelling of connected regions of PanIN in 2D H&E images

For the eight slabs that were assessed by the combined 3D anatomic and genomic analyses, videos with exhaustive, colour-coded annotations of PanIN lesions were created using the 3D tissue models so that connected regions of each PanIN could be efficiently identified, microdissected, and collected in separate vials for genomic analysis. For this purpose, PanIN labels were collected from the digital 3D tissue matrix. Independent (that is, discontinuous in 3D) PanINs were

sorted from largest to smallest, and the ten largest PanIN lesions were assigned distinct colours (largest PanIN assigned pink, second largest PanIN assigned light blue, third largest PanIN assigned yellow and so on) for easy visualization. For samples with more than ten identified PanIN lesions, the eleventh through the smallest were assigned the same colour (olive green), as these PanIN were too small to be microdissected and sequenced. Next, we determined which PanIN lesions were present in each *z* plane of the 3D structure. For each plane, the 5× registered H&E image was loaded into MATLAB 2021b. An outline of the regions of each PanIN present in that plane was digitally thickened and was overlaid on the H&E image in the correct location and in the colour assigned to that PanIN. This process created an H&E image with colour outlines marking all PanIN lesions in the section, highlighting connectivity and relative size. These 5× images were saved separately as jpg files, and together as a lower-resolution stacked video (Supplementary Video 2).

Calculation of 3D structural features of PanIN lesions

The 3D digital tissue matrices that defined tissue types and cell coordinates were used to calculate a range of tissue properties. Each PanIN lesion was separately identified using MATLAB 2021b. By summing the number of voxels in each separate PanIN lesion (and converting from voxel to mm^3), the volume of each PanIN lesion was obtained. The volume of a PanIN was defined as the volume of the neoplastic cells, excluding the lumina within the glands. The Regionprops3 command in MATLAB 2021b was used to determine the length of the primary axes containing each PanIN. Through dot multiplication of the PanIN ID matrix with the matrix containing the cell coordinates, the number of cells in each PanIN could be calculated. Cell counts generated from the 2D histological sections were extrapolated to 3D space using a previously developed technique based on cell-type-dependent nuclear measurements, as well as the thickness and spacing of the histological slides[10]. This approach used the following equation:

$$C_{3D} = \sum_{\text{images}} \sum_{\text{subtypes}} C_{\text{image}} \frac{3T}{T + D_{\text{subtype}}}$$

where C_{3D} = total number of cells in 3D space, $\sum_{\text{images}}$ = sum across all images in a pancreas slab, $\sum_{\text{subtypes}}$ = sum across all tissue types classified in the H&E images, C_{image} = total number of cells of one tissue subtype in a single H&E image, T = thickness of a single histological image, D_{subtype} = mean measured diameter of each tissue subtype in H&E images. Finally, the cell density of each PanIN lesion was determined by dividing the number of cells of each lesion by its volume. The total volume, cellularity, and cell density of all other tissue components were similarly calculated using the 3D tissue and cell matrices. The percentage of PanIN cells in the ductal system was calculated for each sample by dividing the total number of PanIN cells by the sum of the total number of PanIN and normal ductal epithelial cells.

Laser capture microdissection

For NGS, membrane slides from 8 of the 46 tissue slabs were deparaffinized in xylenes for 5 min, and dehydrated with 100%, 95% and 70% ethanol for 1 min each. Slides were stained with crystal violet (Sigma Aldrich; diluted 1:4 in 70% ethanol) for 30 s and dehydrated by ascending ethanol solutions. The stained slides were microdissected the same day. For genetic analyses assessing intra- and inter-PanIN heterogeneity, each slab was divided into five regions of equal size along the *z* axis. On average, tissue slabs spanned 778 slides (range 500–1,000) and each region covered 146 slides (range 100–200) (Supplementary Table 3). Each spatially unconnected PanIN lesion identified in the corresponding 3D models was microdissected (Leica LMD7000 instrument) into a separate collection tube (autoclaved 0.5 ml Qubit Assay Tubes, Invitrogen) for each region. Spatially distinct PanIN lesions and PDAC on H&E images were annotated by CODA in colours corresponding to 3D models; these annotated images were used to guide microdissection of distinct lesions into separate collection tubes. LG and HG components of PanIN lesions were isolated separately by microdissection, as was any PDAC identified in the tissue. Matched control samples from each patient were obtained from clinical archives for sequencing, including matched normal tissue (duodenum or spleen) and matched PDAC. For matched control samples, one 5 µm section was cut onto a regular slide, deparaffinized as described above, and scraped into a collection tube utilizing a sterile razor blade.

Determining sample purity

To validate the neoplastic purity of the sequenced PanIN samples, image analysis was performed to determine the percentage of microdissected cells constituting PanIN and non-PanIN tissue. Digital images of stained membrane slides were created before and after microdissection ('pre-LCM' and 'post-LCM' images, respectively). These images were registered to the adjacent H&E slide image. The pre-LCM image was used to determine the cropped region within the adjacent H&E image that was analysed. The post-LCM image allowed for precise determination of the irregularly shaped region microdissected from the slide. After identifying the precise area that was microdissected, the H&E image was segmented and the number of PanIN cell nuclei and non-PanIN cell nuclei were quantified. The neoplastic purity of the microdissected sample was determined by dividing the number of PanIN cells by the total number of cells within the microdissected area.

DNA extraction and quantification

DNA was extracted using the QIAamp DNA FFPE Tissue Kit (Qiagen) following the manufacturer's recommendations with the following modifications: after overnight digestion, samples were sheared to 200–350-bp fragments using the Covaris S220 Sonicator (Covaris). Subsequently, 10 µl of proteinase K was added and samples were digested for 1 h at 56 °C before resuming manufacturer-provided protocol. DNA quantity and fragment size were quantified using an Agilent Bioanalyzer (Agilent) according to the manufacturer's recommendations. Sample processing was the same for matched normal tissue and matched PDAC tissue. Samples were stored at −20 °C until library preparation.

Library preparation and targeted sequencing

All PanIN regions that yielded at least 10 ng of DNA in the 8 slabs used for combined 3D anatomic and genomic analysis were assessed by targeted NGS using a custom panel. The commercially available ClearSeq Comprehensive Cancer (Agilent) targeted panel was selected. Using Agilent SureDesign software, we customized the panel to induce baits for coding regions of additional genes important for pancreatic cancer development. In total, the panel covered 154 well-characterized cancer driver genes, including all major drivers of pancreatic ductal neoplasia (Supplementary Table 3). DNA (10–200 ng) was utilized per sample for library preparation, following the manufacturer's protocol (SureSelectXT HS2 DNA kit; Agilent). Barcoded individual samples were pooled following the manufacturer's recommendations and sequenced on a MiSeq (Illumina) generating 2 × 150 base-paired reads.

Targeted sequencing analysis

Sequencing data were processed by following The Genome Analysis Toolkit (GATK) Best Practices Pipeline[43]. FASTQ reads were converted to an unmapped BAM file and adapter sequences were marked. FASTQ reads were aligned to the human reference genome GRCh38 (GATK resource bundle 4.2.0.0) using Burrows–Wheeler aligner (BWA) MEM software (version 0.7.15) and merged with the unmapped BAM file[44]. The raw mapped reads were then marked for duplicate reads and underwent base quality score recalibration (GATK resource bundle 4.2.0.0). Mutect2 (GATK resource bundle 4.2.0.0) was used to call somatic SNVs and indels[45]. A panel of normals (PON) was created using targeted sequencing data from the eight matched normal samples.

Next, Mutect2 was run for each tumour sample with its matched normal and the PON using default settings. All somatic mutations were annotated with OpenCRAVAT (version 2.2.1). Somatic mutations were subsequently filtered with the following criteria: tumour sample coverage ≥15X; normal sample coverage ≥10X; tumour frequency ≥0.05; ≥5 distinct reads supporting the mutation in tumour sample; normal frequency <0.05. Variants were filtered to remove all noncoding variants and variants present in any normal sample, while retaining coding SNVs, indels, and splice site variants. Mutations meeting these criteria were used for downstream analysis. Candidate mutations were confirmed or rejected via visual inspection in Integrated Genome Viewer (IGV version 2.14.1)[46]. The positions of *KRAS* hotspot mutations (codons 12, 13 and 61) were visually inspected in all samples and included if VAF ≥ 1% and ≥3 distinct reads supported the mutation.

WES and analysis

In the eight slabs used for combined 3D anatomic and genomic analysis, PanIN regions with sufficient DNA were analysed by WES in addition to targeted sequencing. Mutation capsule (MC) technology was applied to profile a FFPE DNA sample with both hybridization-based exome sequencing and amplification-based deep sequencing, as recently described[22]. Approximately 200 ng of the previously fragmented DNA was subjected to end-repair, dA-tailing, and ligation to a customized adapter with random DNA barcodes as a unique identifier (UID) tag[21]. The ligation product was amplified through 10 reaction cycles to generate a whole genome library (MC library) using NEBNext Ultra II DNA library Prep Kit for Illumina (New England Biolabs) and 750 ng MC library was used for exome enrichment. The exome region of the whole genome libraries was enriched with the Agilent SureSelectXT Human All Exon Kit V6 (Agilent; Santa Clara, CA) for slabs 98–155. The PanIN samples of slab 92 were enriched with mixed probes of Agilent SureSelectXT Human All Exon Kit and an Agilent SureSelectXT Custom Kit (Agilent, 5190-4842) targeting 825 genes, including *KRAS*. All libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina) with 2 × 150-bp paired-end reads. FASTQ files were preprocessed to remove the adapter sequences[21]. The low-quality reads were subsequently removed using Trimmomatic (v0.36). FASTQ reads were aligned to the human reference genome hg38 using Burrows–Wheeler aligner (BWA) MEM software (version 0.7.15) for both tumour and matched normal samples. Mutect2 (GATK resource bundle 4.2.0.0) was used to call somatic variants between the tumour–normal pairs, as well as the tumour-only samples against a PON, utilizing default parameters[45]. The somatic mutations were filtered according to the following criteria: (1) the variant coverage in tumour sample was ≥15X, the VAF ≥0.10 and ≥7 distinct reads supporting the mutation in tumour sample; (2) the corresponding variant coverage in normal sample ≥10X, the VAF <0.05. All somatic mutations were then annotated with vcf2maf-1.6.19. Noncoding variants and variants present in matched normal were filtered, retaining coding SNVs and indels and splice site variants. Mutations meeting these criteria were used for downstream analysis. Candidate mutations were confirmed or rejected via visual inspection in Integrated Genome Viewer (IGV version 2.14.1)[46]. The positions of *KRAS* hotspot mutations (codons 12, 13 and 61) were visually inspected in all samples and included if VAF ≥1% and ≥3 distinct reads supported the mutation.

Subclonal Architecture Reconstruction

For PanIN regions with multiple *KRAS* hotspot mutations in WES (slab 92, PanINs A and B), subclonal architecture reconstruction was performed with PICTograph software (version 1.2.0.1)[28]. Mutations meeting the following criteria were included in PICTograph analysis: (1) the variant coverage in tumour sample was ≥15X, the VAF ≥0.10 and ≥7 distinct reads supporting the mutation in tumour sample; (2) the corresponding variant coverage in normal sample ≥10X, the VAF <0.05. PICTograph uses a Bayesian hierarchical model and Markov chain Monte Carlo (MCMC) sampling to jointly infer subclones and cancer cell fractions (CCFs), based on VAF, sample purity, and copy number. Evolutionary trees are built with a modified version of the Gabow–Myers algorithm and scored with the SCHISM tree fitness function[47,48]. MCMC chains were run with 10,000 iterations and burn-in of 1,000. To eliminate single-mutation clusters, the one box parameter was set to TRUE; the number of clusters *k* was set to 4 for slab 92 PanIN A and to 3 for slab 92 PanIN B. Purity was estimated by as described above. CNVkit (version 0.9.10) was used to construct a copy number profile for each PanIN region[49]. Integer copy numbers, adjusted with tumour purity for each PanIN region, were generated using the call function in CNVkit.

Mutational signatures analysis

Mutational signatures were extracted with SigProfiler tools[50–52]. A total of 674 SNVs from WES with VAF of at least 10% in all 52 sequenced PanIN regions were included in the analysis. All SNVs were converted to match the positive strand in the human reference genome GRCh38.p14 (accession number GCA_000001405.29) with a custom script. SNVs were then formatted into ICGC format. SigProfilerMatrixGenerator (version 1.2.17) was used to generate the input matrix for downstream analysis with reference genome GRCh38 and default parameters[51]. SigProfilerExtractor (version 1.1.21) was then used to extract the mutational signatures of the SBS96 matrix with default parameters[52]. The seed for the run is 656921682551684921690464462390680173 99.

Somatic copy number alteration analysis

WES data were analysed using CNVKit (version 0.9.9) to identify somatic copy number aberrations in precursor and tumour samples[49]. The depth of coverage in targeted (exons) and off-target (intronic or intergenic) regions were corrected to account for factors including size, GC composition, mappability and spacing. In each sample, the resulting coverage was then normalized with respect to a reference panel consisting of the eight matched normal samples. Next, circular binary segmentation was applied to identify genomic regions at a constant relative copy number level in each sample[53]. To capture large-scale copy number changes, the segmental copy number profile of each sample (normal, precursor, or tumour) was then summarized as an array of chromosome arm-level relative copy number values. Given the limited number of matched normal samples ($n = 8$), we processed another cohort of 14 normal samples previously sequenced by our group following the steps above and defined a composite normal panel ($n = 22$) by adding this new set to the matched normals from the main cohort[54]. The chromosome arm-level relative copy number levels in each precursor or tumour lesion were normalized using the mean and standard deviation in the composite panel of normals to derive arm-level copy number *z*-scores. For the normal samples, the leave-one-out setup is used to calculate the arm-level *z*-scores. Copy number data were visualized using R version 3.6.

KRAS G12 point mutation-specific RNA ISH

KRAS G12 point mutation-specific RNA ISH was performed for the eight NGS slabs. Using CODA 3D models as a guide, consecutive unstained FFPE tissue slides from one *z* region per slab were selected for inclusion of multiple PanIN areas. For three cases (slabs 92, 116 and 117) where NGS detected PanINs with multiple *KRAS* mutations, slides were selected from two *z* regions. Slides from all 11 regions from 8 slabs were evaluated with BaseScope paired double-Z oligonucleotide probes targeting the following single base-pair substitutions in human *KRAS* (NM_033360.3 region 191-232nt): p.G12D c.G35>A, p.G12R c.G34>C, p.G12V c.G35>T. Slides from one case (Slab 104) were also probed for p.G12C c.G34>T. In addition to standard positive (*PPIB*) and negative (*dapB*) control probes, one slide from each slab was also assayed for total KRAS expression with a reference probe targeting an invariant sequence (NM_033360.3 region 89-120nt) in *KRAS*. Manufacturer catalogue numbers for the aforementioned probes are as follows: 705518, 705548, 705558, 705508 and 1084358-C1 (Advanced Cell Diagnostics). The BaseScope Red LS ISH assay (Advanced Cell Diagnostics) was performed as previously

described[29]. In brief, FFPE slides underwent deparaffinization, target retrieval, and protease pre-treatment according to manufacturer guidelines. Following 1zz probe hybridization (2 h at 40 °C) and signal amplification with the LS BaseScope Reagent Kit (Advanced Cell Diagnostics), slides were incubated with Fast Red, counterstained with 50% haematoxylin, then mounted in VectaMount (Vector Laboratories).

For quantitative analysis, slides were scanned at 40× magnification brightfield with a Hamamatsu Nanozoomer S360 (Hamamatsu) or a 3DHistech Panoramic Scan Digital Slide Scanner (3DHistech) and annotated in Omero (Web version 5.17.0; Glencoe Software) and NDP.view2 (version 2.9.29; Hamamatsu). Each PanIN was measured for cross-sectional area and assessed for red punctate ISH signals by a blinded investigator. ISH signal frequency was calculated as number of puncta/section area (in μm^2). To account for varying levels of background probe labelling, values were normalized by the mean signal frequency measured in ten islets of Langerhans within each slide.

Targeted sequencing of *KRAS* mutations

Ultra-deep sequencing of the *KRAS* hotspot positions was performed on the PanIN regions analysed by WES. MC library (200 ng) was used to profile *KRAS* mutations with deep sequencing, as previously described[21]. In brief, the target regions were amplified together with the DNA barcode (UID) in the adapter of the MC library for nine cycles using a target-specific primer and a primer matching the universal sequence in the adapter. A second round of 14 cycles of PCR with one pair of nested primers matching the adapter and the target region was used to further enrich the target region and add the Illumina sequencing adapter. The amplified libraries were sequenced on the NovaSeq 6000 Sequencing System (Illumina) using 2× 150-bp paired-end reads. The target regions were analysed to confirm the mutation status as previously described[21]. In brief, the FASTQ file were preprocessed to extract UID tags[21,55,56]. The residual Illumina adapter sequences and low-quality reads were subsequently removed using Trimmomatic (v0.36). The cleaned reads were mapped to the human reference genome GRCh37 (accession number GCA_000001405.1) using BWA software (BWA, v0.7.15)[55]. BAM files were locally realigned and the base quality scores were recalibrated using Genome Analysis Toolkit (GATK, v3.1). The mpileup command in SAMtools (v0.1.16) was used to identify SNVs and indels[57]. To ensure accuracy, the reads with the same UID tag were grouped into a UID family. If more than 80% of reads in a UID family harboured the same variant and it contained at least two reads, the UID family was defined as an effective unique identifier family (EUID family). The prevalence of each mutation was calculated by dividing the number of mutant EUID families by the total number of the mutant and wild type EUID families. Candidate variants were annotated with the VEP (v83) and Oncotator (version 1.5.0.0)[58,59]. The criteria we adopted for retaining a somatic mutation was that it had an allele fraction of ≥1% and ≥7 UID. The retained mutations were verified manually using IGV (version 2.14.1)[46].

Pancreas CT

An independent group of 807 individuals who were candidates for renal donation were scanned as a part of their routine care. These images were obtained. Patients were scanned on a dual-source Multidetector Computed Tomography (MDCT) scanner (Somatom Definition, Somatom Definition Flash, or Somatom Force, Siemens Healthineers), or a 64-MDCT scanner (Somatom Sensation 64, Siemens Healthineers). Patients were injected with 100–120 ml iohexol (Omnipaque, GE Healthcare) at an injection rate of 4–5 ml s^{-1}. Scan protocols were customized for each patient to minimize dose and included a tube voltage of 100–120 kVp, effective tube current–exposure time product of 250–300 mAs, and pitch of 0.6–0.8. The collimation was 128 × 0.6 mm or 192 × 0.6 mm for the dual-source scanner and 64 × 0.6 mm for the 64-MDCT scanner. Arterial phase imaging was performed with fixed delay or bolus triggering, usually between 30 and 35 s after injection, and venous phase imaging was performed at 60–70 s. All images were reconstructed into thin (0.75-mm slice thickness and 0.5-mm increment) slices. The 3D volume of the pancreas was manually segmented by four trained researchers using commercial segmentation software (Velocity version 4.1, Varian Medical Systems), under the supervision of 3 abdominal radiologists each with between 5 and 35 years of experience. The *x*, *y* and *z* dimensions of each voxel containing pancreas was determined and summed to calculated the total pancreas volume for each scan. The volume of pancreas was then calculated by counting the number of voxels containing pancreas first and then converting this number into the unit of volume according to the voxel spacing of CT scans. To extrapolate the number of PanINs in the whole pancreas, prior distributions for the pancreas volume in women and men were first defined as normal l (77.51, 17.82) and normal 2 (94.18, 20.74), respectively. The extrapolated number of PanINs for an individual in this study was found by multiplying the empirically observed PanIN burden by a random ordinate from the prior distribution. Sampling 10,000 random ordinates and multiplying each by the observed burden, we obtained a predictive distribution for whole-pancreas PanIN numbers. A 95% credible interval for this distribution was derived from the 2.5 and 97.5% quantiles.

Reporting summary

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

Data availability

WES data are available via dbGAP under accession number phs003549.v1.p1 as allowed by the Institutional Review Board based on patient consent. Owing to their large file size (TB scale per slab), raw tissue data will be available from the corresponding authors upon request. Source data are provided with this paper.

Code availability

The 3D rendering software used in this Article is available on GitHub at https://github.com/ashleylk/CODA.

42. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
43. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
45. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
46. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
47. Gabow, H. N. & Myers, E. W. Finding all spanning trees of directed and undirected graphs. *SIAM J. Comput.* **7**, 280–287 (1978).
48. Niknafs, N., Beleva-Guthrie, V., Naiman, D. Q. & Karchin, R. SubClonal hierarchy inference from somatic mutations: automatic reconstruction of cancer evolutionary trees from multi-region next generation sequencing. *PLoS Comput. Biol.* **11**, e1004416 (2015).
49. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Comput. Biol.* **12**, e1004873 (2016).
50. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
51. Bergstrom, E. N. et al. SigProfilerMatrixGenerator: a tool for visualizing and exploring patterns of small mutational events. *BMC Genomics* **20**, 685 (2019).
52. Islam, S. M. A. et al. Uncovering novel mutational signatures by de novo extraction with SigProfilerExtractor. *Cell Genomics* **2**, 100179 (2022).
53. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
54. Fujikura, K. et al. Multiregion whole-exome sequencing of intraductal papillary mucinous neoplasms reveals frequent somatic *KLF4* mutations predominantly in low-grade regions. *Gut* **70**, 928–939 (2021).
55. Zhao, D. et al. Personalized analysis of minimal residual cancer cells in peritoneal lavage fluid predicts peritoneal dissemination of gastric cancer. *J. Hematol. Oncol.* **14**, 164 (2021).
56. Zhao, L. et al. Integrated analysis of circulating tumour cells and circulating tumour DNA to detect minimal residual disease in hepatocellular carcinoma. *Clin. Transl. Med.* **12**, e793 (2022).
57. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

58. McLaren, W. et al. The Ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
59. Ramos, A. H. et al. Oncotator: cancer variant annotation tool. *Hum. Mutat.* **36**, E2423–E2429 (2015).

Acknowledgements The authors acknowledge the following sources of support: NIH/NCI P50 CA62924; NIH/NCI T32 CA153952; NIH/NCI U54 CA268083; NIH/NCI U54 CA274371; NIH/NCI U54 CA143868; NIH/NCI U54 CA21073; NIH/NIAMS U54 AR081774; NIH/NCI U01 CA271273; NIH/NIA U01 AG060903; NIH/NIDDK K08 DK107781; NIH/NCI P30 CA00697; Sol Goldman Pancreatic Cancer Research Center; Lustgarten Foundation; Break Through Cancer; Buffone Family Gastrointestinal Cancer Research Fund; Allegheny Health Network-Johns Hopkins Cancer Research Fund; AACR-Bristol-Myers Squibb Midcareer Female Investigator Grant; Rolfe Pancreatic Cancer Foundation; Joseph C. Monastra Foundation; The Gerald O. Mann Charitable Foundation (Harriet and Allan Wulfstat, Trustees); S. Wojcicki and D. Troper; Lustgarten Foundation-AACR Career Development Award for Pancreatic Cancer Research, in honour of Ruth Bader Ginsburg; the Chinese Academy of Medical Sciences (CAMS) Innovation Fund for Medical Sciences (CIFMS) (2021-I2M-1-067 and 2021-1-I2M-018). The authors thank C. Hruban for assistance with illustration. The authors gratefully acknowledge the staff of Oncology Tissue Services at Johns Hopkins, whose expertise in research histology was essential for these studies.

Author contributions A.M.B. and A.L.K. contributed equally to the work. L.D.W. and D.W. jointly supervised the work. L.D.W., D.W., R.H.H. and P.-H.W. conceived the project. A.M.B. managed the sectioning and scanning of slabs, tissue microdissection, DNA sequencing and analysis of sequencing results. J.M.B., L.Z., L.J., H.C., Q.S., J.L., R.R., S.G., A.I.D., C.G.F., S.M., C.M., Y.L., J.G., X.-D.L., N.B., L.C.C., F.L. and N.N. supported the scanning and sequencing work. J.P. managed the ISH work and performed analysis. A.L.K. managed the computational 3D reconstruction of slabs, anatomical calculations and visualization of results. M.P.G., A.F., Y.Z., C.A.-P., A.C.J., J.Y., B.K., S.D., E.F., J.Y.H. and P.A.R. supported the 3D reconstruction work. E.K.F., A.Y., N.J.R., E.D.T., R.B.S., T.C.C., Y.J., R.K., R.H.H. and P.-H.W. oversaw various aspects of the work. A.M.B., L.D.W., A.L.K. and D.W. created the first draft of the manuscript and figures. J.P., L.D.W., A.L.K. and A.M.B. prepared the revised manuscript and figures, which all authors edited and approved.

Competing interests A pending patent application '*Computational Techniques For Three-Dimensional Reconstruction and Multi-labeling of Serially Sectioned Tissue*' was filed on 24 June 2022 by A.L.K., R.H.H., P.-H.W., D.W. and L.D.W. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41586-024-07359-3.

Correspondence and requests for materials should be addressed to Yuchen Jiao, Denis Wirtz or Laura D. Wood.

Peer review information *Nature* thanks Tim Coorens, Beatrice Knudsen, Marina Pasca di Magliano and Hiroki Ueda for their contribution to the peer review of this work. Peer reviewer reports are available.

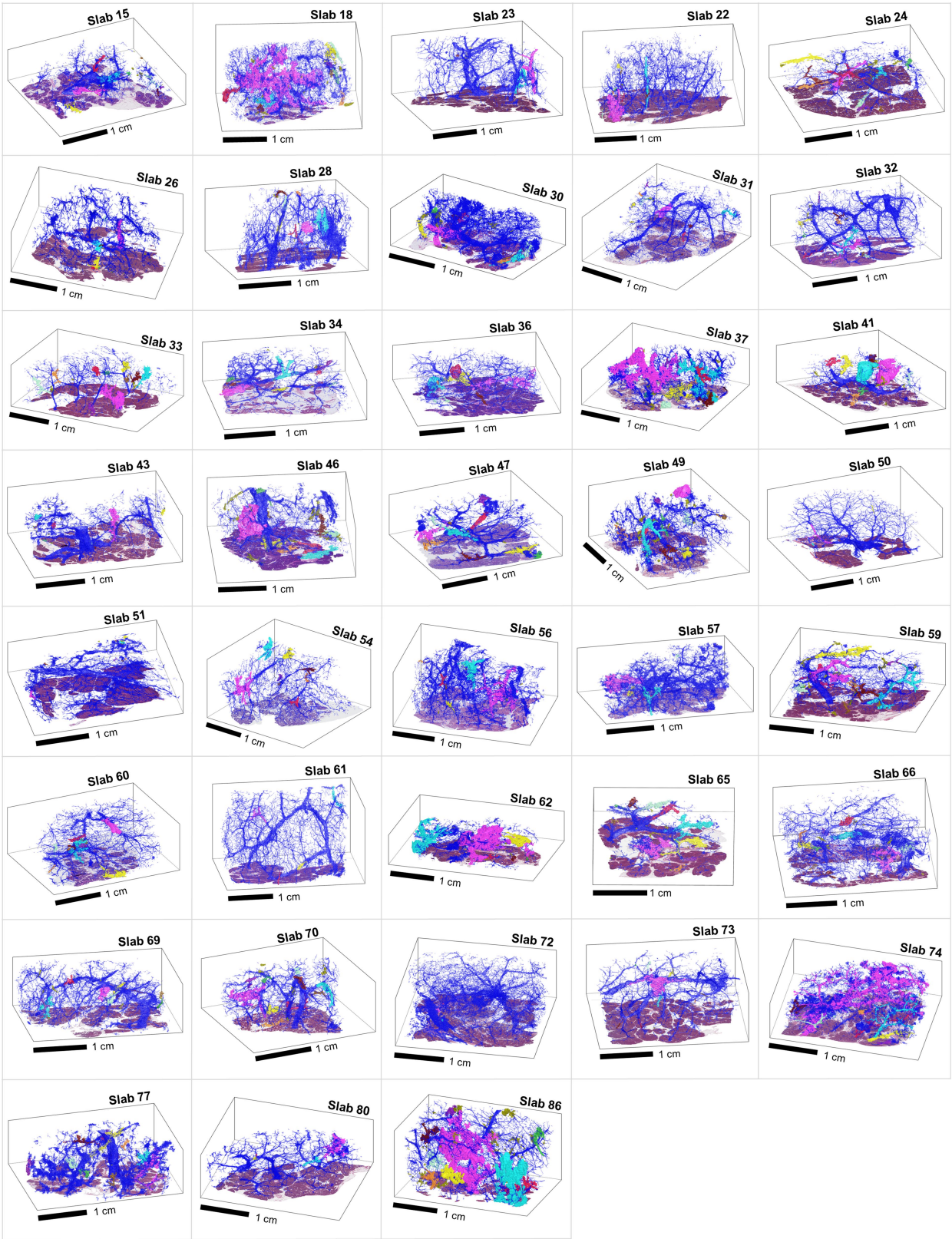
A

	Person 1					Person 2				
Image	TP	FP	FN	Precision	Recall	TP	FP	FN	Precision	Recall
1	4806	260	421	94.9%	91.9%	4681	385	274	92.4%	94.5%
2	8019	754	371	91.4%	95.6%	8106	667	305	92.4%	96.4%
3	3033	278	332	91.6%	90.1%	3070	241	338	92.7%	90.1%
4	4093	295	538	93.3%	88.4%	4121	267	453	93.9%	90.1%
5	6792	477	698	93.4%	90.7%	6578	691	656	90.5%	90.9%

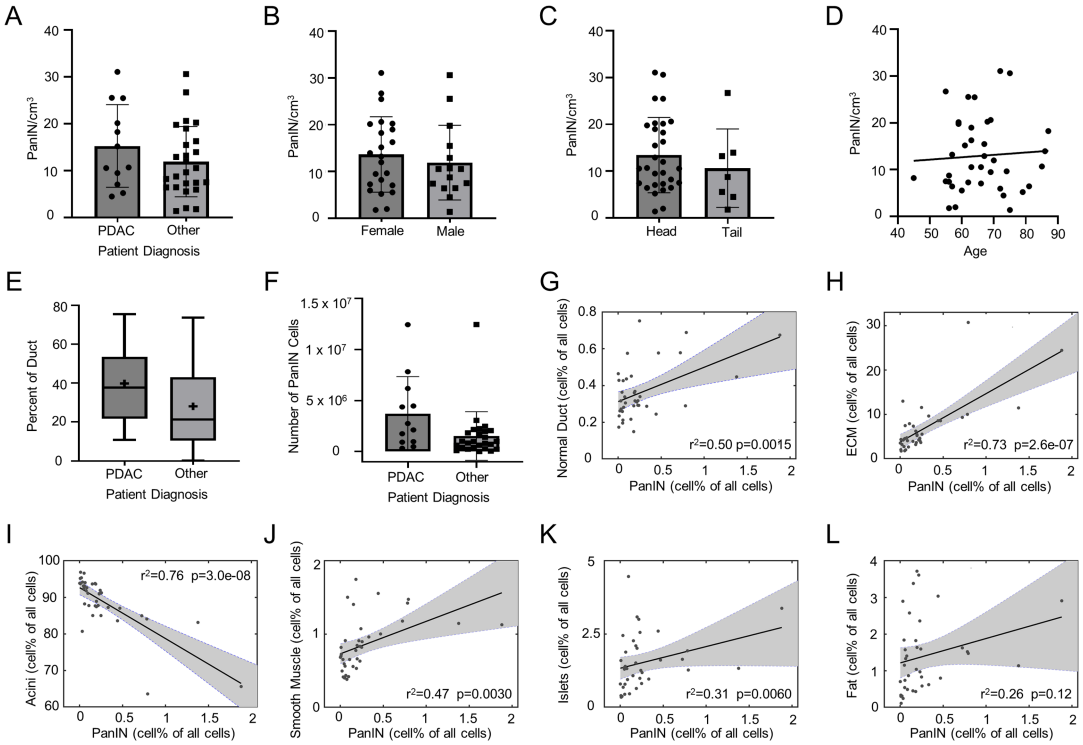
B

Ground Truth \ Predicted	Islets of Lang.	Ductal epithelium	Smooth muscle	Fat	Acini	ECM	Nontissue	PanIN	**Recall**
Islets of Lang.	36431	2	0	0	1506	61	0	0	**95.9%**
Ductal epithelium	0	34224	217	0	529	731	1119	1180	**90.1%**
Smooth muscle	0	0	37418	0	0	498	84	0	**98.5%**
Fat	0	0	0	37827	150	23	0	0	**99.5%**
Acini	40	28	0	0	37883	1	48	0	**99.7%**
ECM	65	0	317	1931	78	35307	175	127	**92.9%**
Nontissue	0	0	11	134	0	10	37829	16	**99.6%**
PanIN	0	5	0	0	0	1006	170	36819	**96.9%**
Precision	**99.7%**	**99.9%**	**98.6%**	**94.8%**	**94.4%**	**93.8%**	**96%**	**96.5%**	**96.6%**

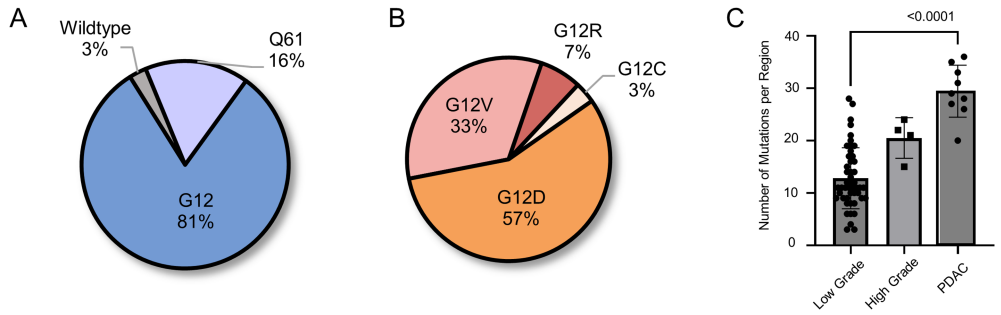
Extended Data Fig. 1 | Validation of CODA segmentation. A. H&E cell detection true positives (TP), false positives (FP), false negatives (FN), precision, and recall compared to manual annotations by two individuals. B. Confusion matrix detailing performance of semantic segmentation algorithm in labelling pancreatic microanatomy in H&E stained tissue images.



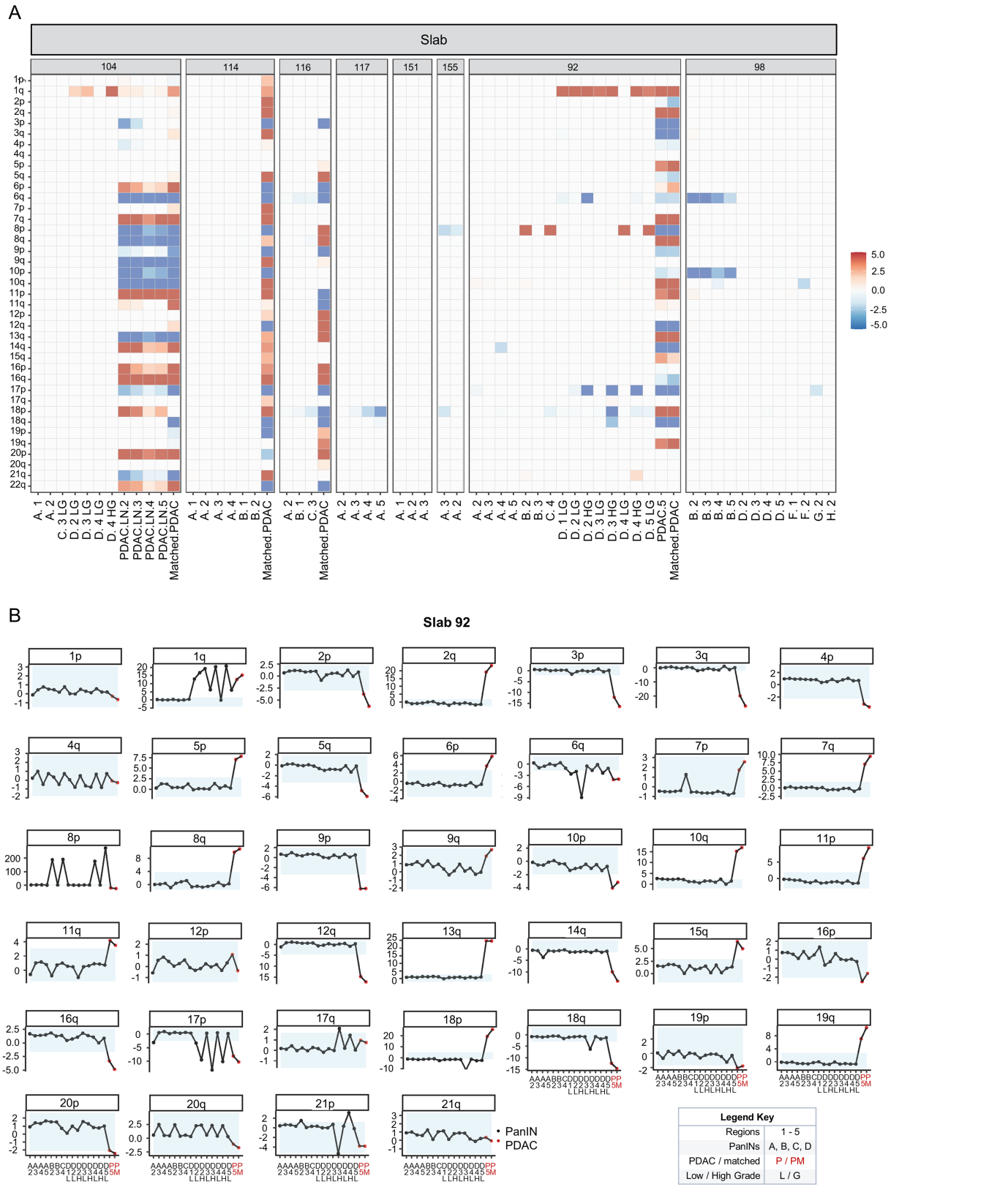
Extended Data Fig. 2 | CODA-rendered 3D models for tissue slabs analyzed. Blue represents normal pancreatic ducts; spatially separate PanINs are indicated with distinct colors.



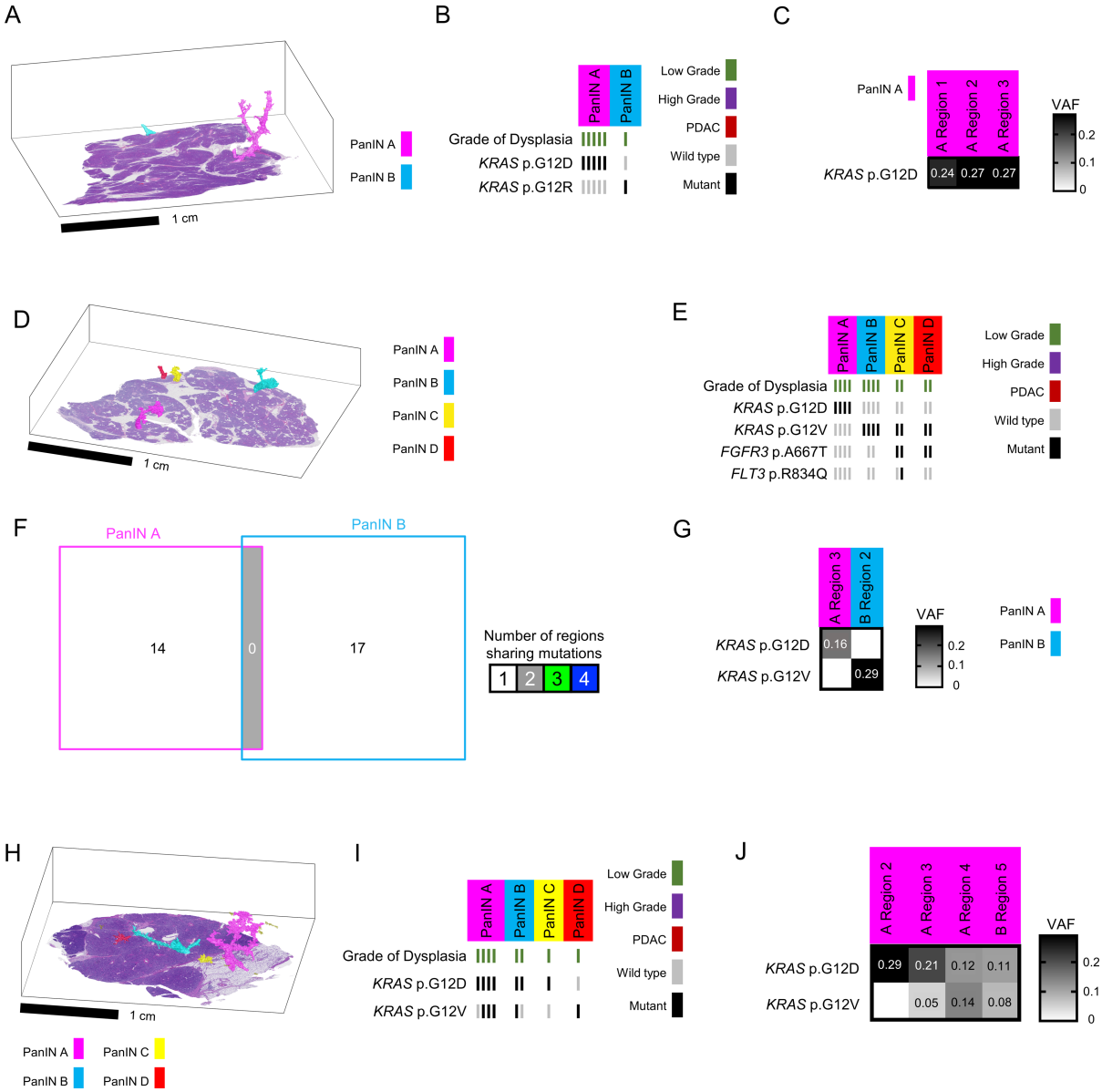
Extended Data Fig. 3 | Quantified features from CODA-rendered 3D models of human pancreatic tissue slabs. A. PanIN burden stratified by disease type. PanINs per cm^3 of analyzed tissue for each slab, not statistically significant. Mean with standard deviations plotted. n = 38 independent samples. B. PanIN burden stratified by sex. PanINs per cm^3 of analyzed tissue for each slab, not statistically significant. Mean with standard deviations plotted. n = 38 independent samples. C. PanIN burden stratified by location of harvested tissue, not statistically significant. Mean with standard deviations plotted. n = 38 independent samples. D. PanIN burden stratified by age, not statistically significant as determined by simple linear regression. E. Percent ductal system affected by PanIN by disease type. Percentage of neoplastic ductal cells was calculated for each slab, not statistically significant by two-tailed Mann-Whitney test. + = mean. n = 38 independent samples. Whiskers denote min-max values while box shows Q1–Q3 with median denoted by a line. F. Number of cells comprising PanINs for each slab by disease type, not statistically significant by two-tailed Mann-Whitney test. n = 38 independent samples. Whiskers show min-max values while box shows Q1–Q3 with median value denoted by a line. G. Correlation of percent PanIN cells (x-axis) to percent normal ductal cells (y-axis). Each point represents a tissue slab. H. Correlation of percent PanIN cells (x-axis) to percent cells in ECM (y-axis). I. Correlation of percent PanIN cells (x-axis) to percent acinar cells (y-axis). J. Correlation of percent PanIN cells (x-axis) to percent smooth muscle cells (y-axis). Correlation of percent PanIN cells (x-axis) to percent islet cells (y-axis). L. Correlation of percent PanIN cells (x-axis) to percent fat cells (y-axis). G-L. r^2 and p-values calculated using the correlation coefficient, and grey shading represents the 95-percent confidence interval.



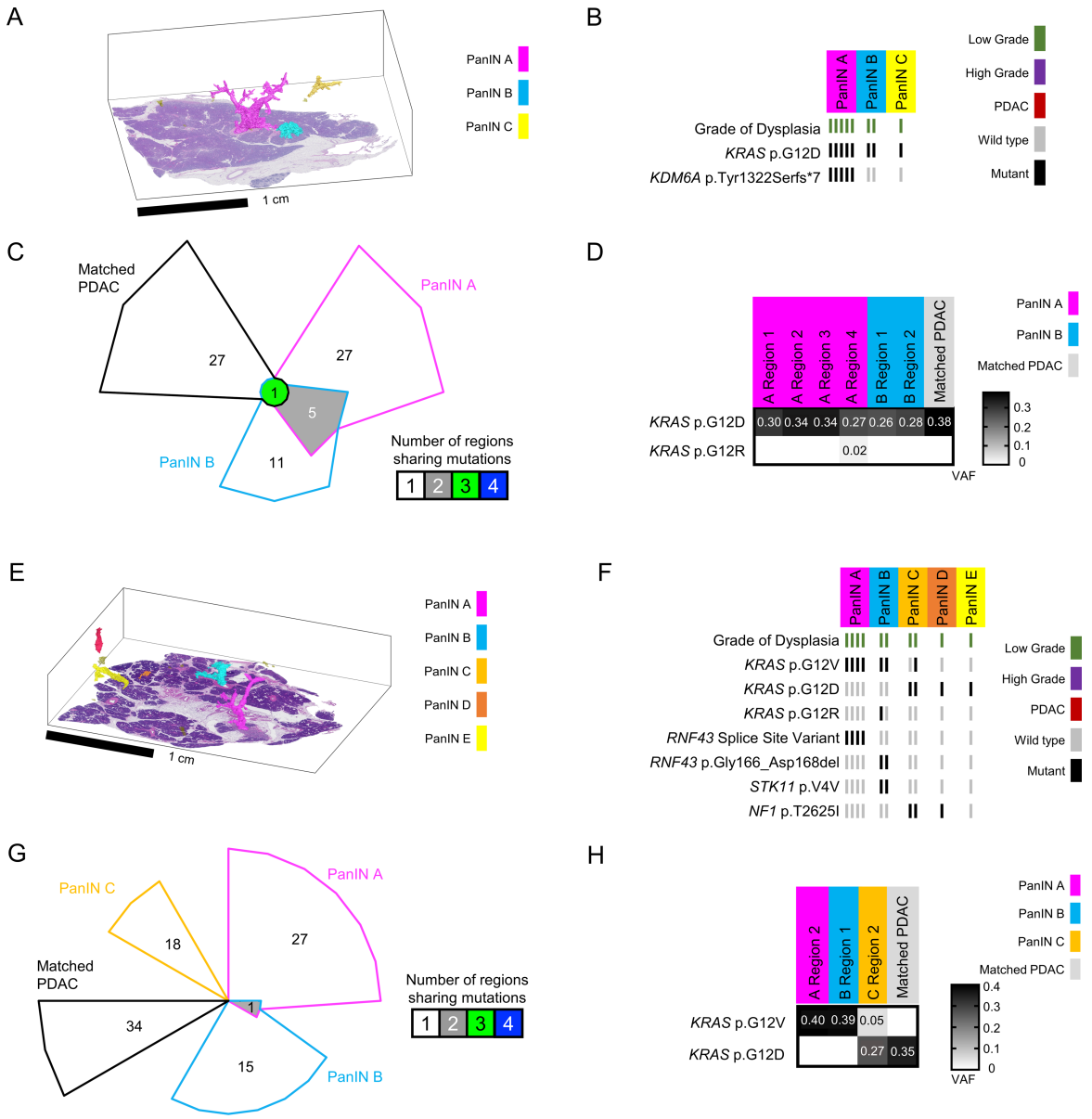
Extended Data Fig. 4 | *KRAS* oncogenic hotspot mutations in PanINs assessed by multiregional targeted NGS of eight tissue slabs and mutation burden by lesion grade from WES. A. Missense mutations occur primarily at G12 and at lower frequencies in Q61. B. At the G12 locus, four different amino acid substitutions were found, with p.G12D being the most common. C. Number of mutations per region stratified by low grade, high grade, and PDAC. Mean with standard deviation plotted. $p < 0.0001$ when comparing mutations in low grade and PDAC samples; determined by Kruskal-Wallis test, corrected for multiple comparisons with Dunn's test. 48 low grade PanINs, 4 high grade PanINs, and 9 PDACs from 38 independent samples were assessed.



Extended Data Fig. 5 | Copy number aberrations in PDAC and PanINs analyzed by WES. A. Chromosome arm-level copy number changes summarized as Z-scores, shown in the precursor and matched tumor samples from each individual. Z-score levels falling within the observed range of the composite panel of normals are depicted in white (copy neutral). For values above or below the normal range, the distance to the maximum or minimum value within the normal range are reported. B-I. Chromosome arm level copy number aberrations in representative sample, slab 92. The blue ribbon marks the range of z-scores observed in the panel of normals.

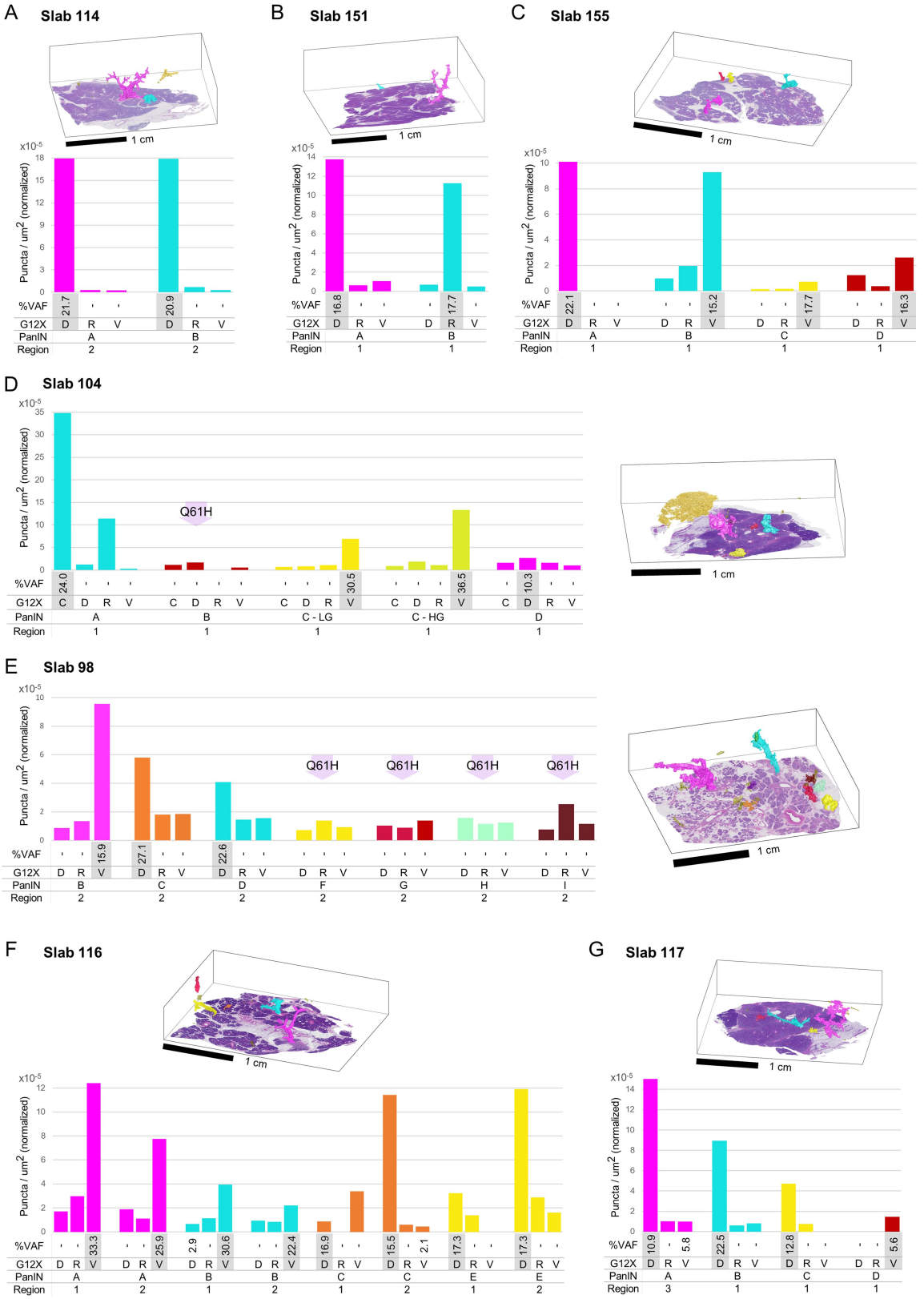


Extended Data Fig. 6 | Slabs 151, 155, 117: 3D model and NGS results. **A-C: slab 151**. A. CODA-rendered 3D model of slab 151, with multiple spatially distinct PanINs (color key at right). B Mutation chart with targeted sequencing results for slab 151. Each row represents a distinct somatic mutation. Columns represent PanINs, colored as in A; bars represent vertical regions within a PanIN. Bar color corresponds to key on right. C. Heatmap of VAFs from Mutation Capsule *KRAS* deep sequencing of for slab 151. Rows represent different *KRAS* oncogenic hotspot mutations. Columns represent PanINs, with colors corresponding to A. **D-G: slab 155**. D. CODA-rendered 3D model of grossly normal pancreatic tissue in slab 155, with multiple spatially distinct PanINs (color key at right). E. Mutation chart with targeted sequencing results for slab 155. Each row represents a distinct somatic mutation. Columns represent PanINs, colored as in D; bars represent vertical regions within a PanIN. Bar color corresponds to key on right. F. Chow-Ruskey plot of WES results for slab 155. Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group. G. Heatmap of VAFs from Mutation Capsule *KRAS* deep sequencing of slab 155. Rows represent different *KRAS* oncogenic hotspot mutations. Columns represent PanINs, with colors corresponding to D. **H-J: slab 117**. H. CODA-rendered 3D model of grossly normal pancreatic tissue in slab 117, with multiple spatially distinct PanINs (color key at right). I. Mutation chart with targeted sequencing results for slab 117. Each row represents a distinct somatic mutation. Columns represent PanINs, with colors as in H; bars represent vertical regions within a PanIN. Bar color corresponds to key on right. J. Heatmap of VAFs from Mutation Capsule *KRAS* deep sequencing. Rows represent different *KRAS* oncogenic hotspot mutations. Columns represent PanINs, with colors as in H.

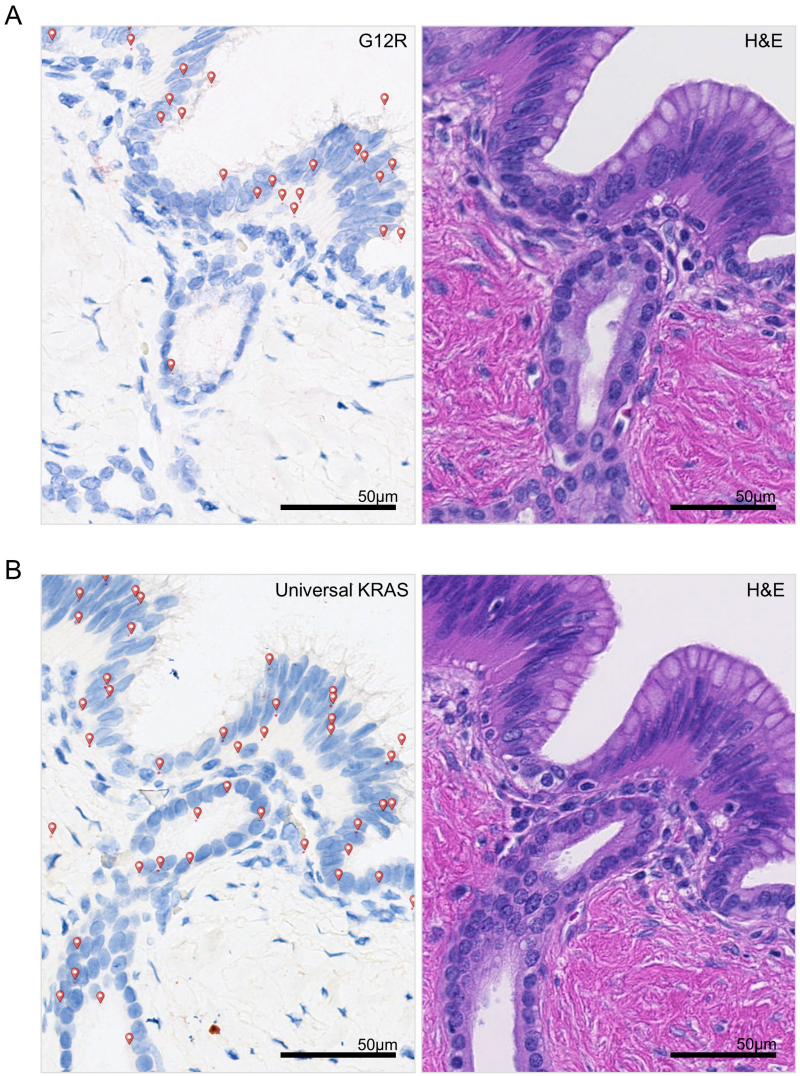


Extended Data Fig. 7 | Slabs 114 and 116: 3D model and NGS results. A-D: slab 114. A. CODA-rendered 3D model of grossly normal pancreatic tissue in slab 114, with multiple spatially distinct PanINs (color key at right). B. Mutation chart with targeted sequencing results for slab 114. Each row represents a distinct somatic mutation. Columns represent PanINs; bars represent vertical regions within a PanIN. Bar color corresponds to key on right. C. Chow-Ruskey plot summarizing WES results for slab 114. Each shape represents a group of mutations. Colors correspond to PanINs in A, except matched PDAC sample sequenced from archival clinical block (grey). Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group. D. Heatmap of VAFs from Mutation Capsule *KRAS* deep sequencing of slab 114. Rows represent different *KRAS* oncogenic hotspot mutations. Columns represent PanINs, colored as in A, except matched PDAC sample sequenced from archival clinical block (grey).

E-H: slab 116. E. CODA-rendered 3D model of grossly normal pancreatic tissue in slab 116, with multiple spatially distinct PanINs (color key at right). F. Mutation chart with targeted sequencing results for slab 116. Each row represents a distinct somatic mutation. Columns represent PanINs; bars represent vertical regions within a PanIN. Bar color corresponds to key on right. G. Chow-Ruskey plot summarizing WES results for slab 116. Each shape represents a group of mutations. Colors correspond to PanINs in E except matched PDAC sample sequenced from archival clinical block (grey). Shapes represent groups of mutations, with overlaps indicating shared somatic mutations. Numbers refer to the number of mutations in each group. H. Heatmap of VAFs from Mutation Capsule *KRAS* deep sequencing of slab 116. Rows represent different *KRAS* oncogenic hotspot mutations. Columns represent PanINs, except matched PDAC sample sequenced from archival clinical block (grey).



Extended Data Fig. 8 | Normalized quantification of KRAS G12 point mutation-specific RNA ISH for each PanIN, shown with CODA-rendered 3D models in corresponding colors. Tissue sections from a single region in each following slab were analyzed for point mutations in KRAS G12. LG and HG areas were quantified and depicted separately when present. Corresponding VAFs from targeted NGS of each PanIN region shown below each column, with most common variant highlighted in grey. Lavender arrows mark PanIN regions where KRAS Q61H mutations were found in NGS. A. Slab 114. B. Slab 151. C. Slab 155. D. Slab 104. E. Slab 98. For the following slabs containing PanINs with multiple KRAS mutations by NGS, tissue sections were analyzed as above from two regions. F. Slab 116. G. Slab 117.



Extended Data Fig. 9 | Representative images of mutant and total *KRAS* RNA ISH staining in histologically normal ductal epithelial cells adjacent to PanIN. A. KRAS G12R mutant RNA ISH probe staining (red puncta) in a PanIN and adjacent normal ductal cells, with consecutive H&E-stained tissue section. B. Universal KRAS reference ISH probe staining of same PanIN and adjacent normal ductal cells, with corresponding H&E image. Scale bars = 50 μm.

nature portfolio

Corresponding author(s): Laura D Wood, Denis Wirtz

Last updated by author(s): Jan 10, 2023

Reporting Summary

Nature Portfolio 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 Portfolio 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	MATLAB 2021b was used for computational microanatomical analysis in this work. All scripts used were written by the authors or are properly cited in the text. Scripts for CODA are available on Github: https://github.com/ashleylk/CODA
Data analysis	GATK version v4.2.0.0 and version 3.1; MuTect2 (GATK resource bundle version 4.2.0.0); OpenCRAVAT version 2.2.1; CNVKit (0.9.10 for PICTograph analysis, 0.9.9 for identification of copy numer alterations); PICTograph software version 1.2.0.1; Trimmomatic version 0.36; Integrated Genome Viewer 2.14.1; Burrows-Wheeler aligner (BWA) MEM software version 0.7.15; SigProfilerMatrixGenerator version 1.2.17; SigProfilerExtractor version 1.1.21; SAMtools version 0.1.16; VEP version 83; VelocityTM, Varian Medical Systems Inc. version 4.1; R version 3.6; Omero Web ver 5.17.0; NDP.view2 ver 2.9.29; Oncotator version 1.5.0.0. Scripts for PICTograph are available on Github: https://github.com/KarchinLab/pictograph/releases/tag/v1.2.0.1

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 Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

WES data is available via dbGAP under accession number phs003549.v1.p1 as allowed by the Institutional Review Board based on patient consent. The 3D rendering software used in this paper is available at the following GitHub page: https://github.com/ashleylk/CODA. Due to their large file size (TB scale per slab), raw tissue data will be available from the corresponding authors upon request.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	The sex of study participants was determined from their medical records, which typically relies on self-reporting. The analyzed cohort includes 27 female participants and 19 male participants. Differences in precancer burden by participant sex were analyzed, with the results reported in the supplementary material.
Population characteristics	All participants underwent clinically indicated pancreatic surgery for a variety of indications detailed in the supplementary material. The mean age of the cohort was approximately 67 years, with a range of 40-90. This age range is in keeping with that of patients typically undergoing pancreatic surgery.
Recruitment	Grossly normal pancreatic tissue was harvested from surgical pancreatectomy specimens at the time of processing in surgical pathology. All participants underwent clinically indicated pancreatic surgery for a variety of indications detailed in the supplementary material. The only self-selection bias is towards patients who chose to undergo pancreatic surgery based on their treating physician's recommendation, which is unlikely to impact the study results. The limitation of the use of surgical pancreatectomy tissue for this study is discussed in detail in the Discussion section of the manuscript.
Ethics oversight	This study was approved by the Institutional Review Board of The Johns Hopkins Hospital

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

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	No sample size calculation was performed. The sample size was determined by the availability of appropriate human tissue samples.
Data exclusions	The inclusion and exclusion criteria for participants are described in detail in the Methods. In participants who met these criteria, no data were excluded from the study.
Replication	Findings may be reproduced using our publicly available computational methods.
Randomization	Randomization is not relevant to our study because it does not contain experimental groups.
Blinding	Blinding is not relevant to our study because it does not contain experimental groups.

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
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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