ARTICLE

https://doi.org/10.1038/s41586-018-0481-8

Precancerous neoplastic cells can move through the pancreatic ductal system

Alvin P. Makohon-Moore[1,15], Karen Matsukuma[2,15], Ming Zhang[3,15], Johannes G. Reiter[4,5,15], Jeffrey M. Gerold[5,15], Yuchen Jiao[6], Lisa Sikkema[1,7], Marc A. Attiyeh[1], Shinichi Yachida[8], Corinne Sandone[9], Ralph H. Hruban[3,10,11], David S. Klimstra[12], Nickolas Papadopoulos[6], Martin A. Nowak[5,13], Kenneth W. Kinzler[6], Bert Vogelstein[3,6,14] & Christine A. Iacobuzio-Donahue[1,12]*

Most adult carcinomas develop from noninvasive precursor lesions, a progression that is supported by genetic analysis. However, the evolutionary and genetic relationships among co-existing lesions are unclear. Here we analysed the somatic variants of pancreatic cancers and precursor lesions sampled from distinct regions of the same pancreas. After inferring evolutionary relationships, we found that the ancestral cell had initiated and clonally expanded to form one or more lesions, and that subsequent driver gene mutations eventually led to invasive pancreatic cancer. We estimate that this multi-step progression generally spans many years. These new data reframe the step-wise progression model of pancreatic cancer by illustrating that independent, high-grade pancreatic precursor lesions observed in a single pancreas often represent a single neoplasm that has colonized the ductal system, accumulating spatial and genetic divergence over time.

The transformation of a normal cell to invasive cancer occurs through the accumulation of genetic and epigenetic changes[1]. Many invasive carcinomas in adults develop from morphologically recognizable noninvasive precursor lesions[2]. The most common precursor lesion associated with pancreatic ductal adenocarcinoma (PDAC) is pancreatic intraepithelial neoplasia (PanIN)[3]. At the morphological level, low-grade PanINs (LG-PanIN, PanIN-1 and PanIN-2) have minimal to moderate cytologic atypia and higher-grade PanINs (HG-PanIN, PanIN-3) have severe cytologic atypia. HG-PanINs exhibit morphological features that are thought to facilitate progression to an infiltrating carcinoma[4].

Aspects of this progression are supported by genetic studies[4–6], but fundamental questions about the development of PDAC remain[7]. The majority of PanINs (regardless of grade) harbour *KRAS* mutations; HG-PanINs and invasive carcinomas are more likely to contain additional driver gene alterations such as those in *TP53*, *CDKN2A*, and *SMAD4*. Moreover, PanINs and neighbouring PDACs often share many genetic alterations in both passenger and driver genes[8,9]. Collectively, these observations suggest that a subset of PDACs arises from adjacent PanINs, just as a colorectal carcinoma can arise from an underlying adenoma[10]. However, in individuals with multiple anatomically distinct PanINs[11], the biological and genetic relationships among these lesions, and their clinical significance, are not fully understood[12]. For instance, cancerization of the pancreatic ducts (an invasive cancer growing back into the duct system and simulating PanINs) by an established PDAC recapitulates lesions with histopathological features that are difficult to distinguish from those of bona fide PanIN precursor lesions[13]. Furthermore, the importance of non-invasive precursor lesions was recently challenged by a whole-genomic-sequence analysis of pancreatic cancers that proposed that tumorigenesis of pancreatic cancer is neither gradual nor slow[14]. We posited that a genomic evaluation of PDAC and matched co-evolving PanINs would provide additional insights into the biology of pancreatic cancer precursors and the dynamics of step-wise progression.

Evolutionary scenarios

Fig. 1a presents the conceptual framework underlying the interpretation of sequencing data generated from one PanIN and PDAC in the same patient, outlining three possible scenarios that in theory might be found. In the first scenario, the PanIN and the PDAC do not share any somatic mutations and arose independently. In the second scenario, the PanIN shares a subset of the somatic passenger and driver gene mutations with the PDAC, but the PDAC contains additional driver or passenger gene alterations not present in the PanIN. Scenario 2 presumes that a common ancestral cell underwent initiation and clonal expansion before seeding the PanIN and PDAC, but neither the common ancestral cell nor the founding PanIN cell had yet acquired all of the genetic events required to generate an invasive neoplasm. In the third scenario, the PanIN and the PDAC share some passenger mutations and all driver gene alterations, and the ancestral cell that seeded both the PDAC and PanIN had already acquired all of the alterations required to form a malignant cancer.

Whole-exome sequencing and phylogenetic analysis

To investigate the progression patterns of pancreatic carcinogenesis, we prospectively screened more than 100 resected pancreases from over a three-year interval to identify those samples in which at least one LG-PanIN (PanIN-2) or HG-PanIN (PanIN-3) was present in a region that was anatomically distinct and far removed from that of the PDAC (see Methods). We excluded any patient with a personal or family history

[1]The David M. Rubenstein Center for Pancreatic Cancer Research, Human Oncology and Pathogenesis Program, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [2]Department of Pathology, University of California, Davis, Sacramento, CA, USA. [3]The Sol Goldman Pancreatic Cancer Research Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [4]Canary Center for Cancer Early Detection, Department of Radiology, Stanford University School of Medicine, Palo Alto, CA, USA. [5]Program for Evolutionary Dynamics, Harvard University, Cambridge, MA, USA. [6]The Ludwig Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [7]VU University Amsterdam, Master's Oncology Program, VU University Medical Center, Amsterdam, The Netherlands. [8]Department of Cancer Genome Informatics, Graduate School of Medicine, Osaka University, Osaka, Japan. [9]Department of Art as Applied to Medicine, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [10]Department of Pathology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [11]Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, MD, USA. [12]Department of Pathology, Memorial Sloan Kettering Cancer Center, New York, NY, USA. [13]Department of Mathematics, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA. [14]Howard Hughes Medical Institute, The Johns Hopkins Kimmel Cancer Center, Baltimore, MD, USA. [15]These authors contributed equally: Alvin P. Makohon-Moore, Karen Matsukuma, Ming Zhang, Johannes G. Reiter, Jeffrey M. Gerold. *e-mail: iacobuzc@mskcc.org

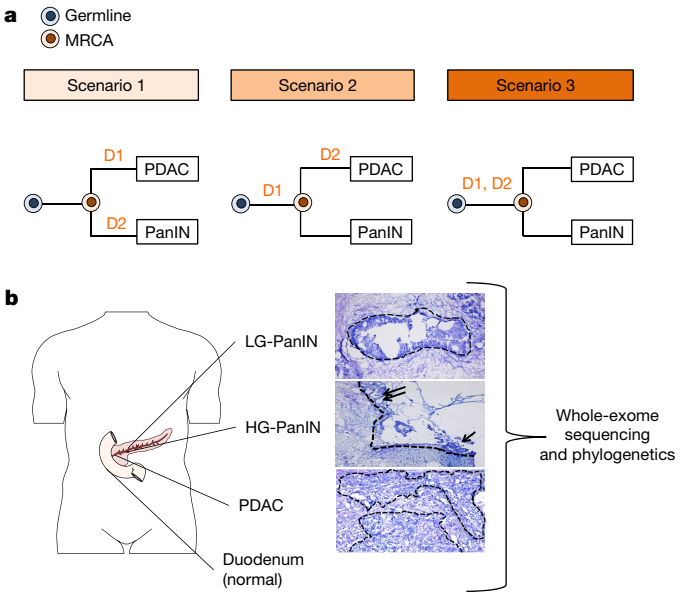


Fig. 1 | Evolutionary scenarios and study strategy of coexistent PanIN(s) and PDAC. a, Evolutionary scenarios of coexistent PanIN(s) and PDAC. For each of the three evolutionary scenarios, D1 and D2 indicate two hypothetical driver gene alterations; blue cells represent the germline (matched normal sample) and orange cells represent the most recent common ancestor (MRCA) for each PanIN–PDAC pair. In scenario 1, none of the somatic gene alterations are shared by the PanIN and PDAC. Mutation D1 is restricted to PDAC and mutation D2 is restricted to the PanIN. In scenario 2, only D1 is shared by the PanIN and PDAC. The mutation in D2 is restricted to the PDAC. In scenario 3, both D1 and D2 are shared by the PanIN and PDAC. **b**, Tissue collection, histological review and microdissection, WES, and phylogenetic analysis of patients. Body diagram adapted from the Motifolio toolkit. Micrographs show examples of PanINs and matched PDAC. The dashed outlines indicate regions that underwent laser capture microdissection for DNA extraction followed by whole-exome sequencing (WES). The LG-PanIN shows well formed papillary structures with nuclear crowding and cytological atypia. The HG-PanIN has regions of pseudopapillary formation (arrows) with a high nucleus–cytoplasm ratio. The matched PDAC shows features of poorly differentiated carcinoma with desmoplasia.

of PDAC from our study, as the dynamics of initiation in patients with germline alterations may be different from that in patients with sporadic pancreatic carcinogenesis[15]. Eight patients were identified, from which twelve PanINs and eight PDACs were sampled for the current study (Supplementary Table 1). All 20 tissue samples were laser-capture microdissected to ensure that a high fraction of the cells within each lesion were neoplastic (Fig. 1b). Despite the microscopic size of the PanINs, we were able to obtain sufficient amounts of DNA to generate high-quality libraries for whole-exome sequencing (WES). Notably, the generation of these libraries did not require whole-genome amplification before WES, thus reducing potential errors in downstream analyses.

We prepared sequencing libraries from each of the lesions as well as from normal tissues from each patient, and used them for massively parallel sequencing on an Illumina HiSeq instrument. We obtained a median canonical exon coverage of 253× across all samples. By comparing each lesion with its matched normal DNA, we identified 2,886 somatic single base substitutions (single nucleotide variations, SNVs) and small insertions or deletions (indels) (Extended Data Fig. 1, Supplementary Table 2). As a group, the PanINs harboured as many SNVs and indels as the PDACs (average of 75 and 80, respectively; Extended Data Fig. 1b). We also analysed somatic copy number alterations (CNAs) and structural variants from the exomic sequencing data (Supplementary Tables 3, 4, Extended Data Figs. 1c, 2). The number of CNAs, unlike the number of SNVs and indels, was higher in PDACs than in PanINs (average of 90 and 68, respectively).

Computational analysis (see Methods) revealed somatic mutations in many well-known driver genes, such as *KRAS*, *CDKN2A*, *TP53*, *SMAD4*, *U2AF1*, and *KMT2D* (Supplementary Table 5). Collectively, the genetic features of this set of PanINs and PDACs were consistent with the results of previous sequencing studies of these tumours[16–20]. To infer evolutionary relationships among the PanINs and PDACs for each patient based on the SNVs and indels, we used Treeomics[21], a phylogenetic method designed specifically for analysing sequencing data from spatially distinct tumours in the same individual[22] (see Methods). Treeomics identified high-confidence phylogenies for the matched samples from each of the eight patients (Fig. 2a–c, Extended Data Figs. 3–5). These analyses allowed us to derive the evolutionary relationships between the coexisting PanINs and the PDAC in each patient.

Evolutionary patterns in PDACs and PanIns

In our cohort, we found two patients (PIN102 and PIN105) in which no passenger gene mutations were shared by the PDAC and PanIN (Fig. 2a, Extended Data Fig. 3). For example, in patient PIN105 both the PanIN and PDAC had a KRAS(G12D) missense mutation. The PDAC exhibited 80 additional point mutations, including a one base-pair frameshift deletion in *TP53*, a missense mutation in *ACVR1B* (ACVR1B(C34Y)), and a 15-base-pair in-frame deletion in *SMAD4*. In addition, the PDAC had acquired CNA losses affecting *CDKN2A*, *MAP2K4*, *TP53*, and *SMAD4* (Extended Data Fig. 3b). The PDAC and PanIN may have arisen independently and by chance accumulated the same *KRAS* mutation (scenario 1), or they may have been initiated by a single KRAS(G12D) mutant clone and subsequently diverged (that is, scenario 2). Scenario 1 may be more likely, given the high frequency of *KRAS* variants in PDAC (more than 90%)[13] and the absence of any other shared somatic variants among the matched PanIN and PDAC samples in both of these patients. Moreover, the PanINs in both of these cases exhibited PanIN-2 histology, and a previous study indicated that LG-PanINs often harbour genetic features that support independent evolution. We note that the previous observation included distinct *KRAS* variants in matched PanINs, in contrast to the two cases presented here[9].

Four of the eight patients showed unequivocal evidence for scenario 2: a common ancestral cell underwent initiation and clonal expansion to form one or more PanINs. Further clonal expansions driven by additional driver gene mutations in a PanIN cell eventually led to a PDAC (Fig. 2b, Extended Data Fig. 4). For example, in patient PIN101, the common ancestor of PanIN lesion A and the PDAC acquired 14 somatic passenger mutations, including KRAS(G12D), as well as losses affecting *ACVR1B*, *MAP2K4*, *TP53*, and *SMAD4* (Fig. 2b, Extended Data Fig. 4a). The PDAC accumulated 28 point mutations, including CDKN2A(A21D) and TP53(R273H) missense mutations, a loss affecting *CDKN2A* and a gain affecting *MYC*. PanIN lesion A accumulated 111 point mutations, including a nonsense mutation in *SDK2*. Similar patterns were found in PIN103, PIN104 and PIN108—driver gene mutations common to all lesions as well as additional driver gene mutations specific to the PDAC (that is, scenario 2; Fig. 2b).

Finally, we observed two patients (PIN106 and PIN107) with phylogenetic patterns consistent with scenario 3, in which all lesions in a single pancreas shared all of the driver gene mutations identified (Fig. 2c, Extended Data Fig. 5). In patient PIN106, the common ancestor of all four samples harboured 47 somatic point mutations, including KRAS(G12D) and TP53(G266E) missense mutations, a mutation affecting the splice region in *ATM*, and GLI3(Q597*) (Fig. 2c). The PDAC subsequently acquired 39 passenger mutations and losses affecting *CDKN2A* and *SMAD4*.

In summary, the lesions in four of these eight patients were unequivocally derived from the same precursor clones, as they shared multiple passenger genes and a subset of driver genes (scenario 2). The presence of these additional driver gene alterations, coupled with phylogenetic analysis, provides persuasive evidence that the PDAC was derived from a PanIN in each case. These results highlight the value

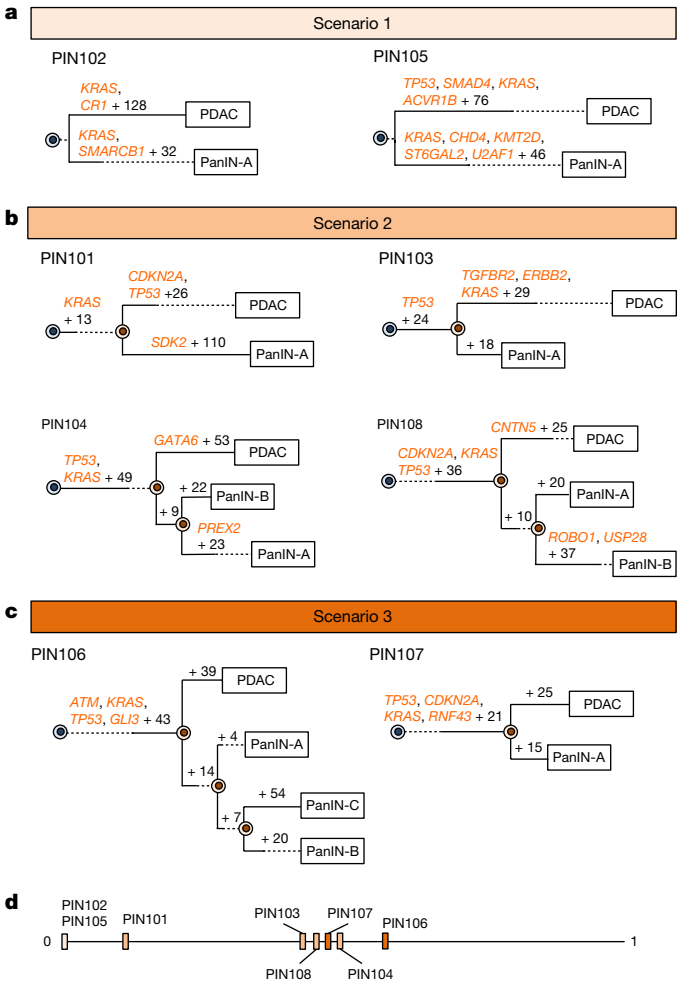


Fig. 2 | Phylogenetics of eight patients. a–c, Phylogenetic trees from SNVs and indels. The sample size is $n = 20$ lesions from eight patients. See Supplementary Table 1 for sample identities. For each phylogeny, gene names are SNVs or indels and the number of additionally acquired mutations is shown. The branch lengths approximate the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. The sequencing data for each driver gene variant was manually reviewed to verify phylogenetic position. **a**, PIN102 and PIN105 are both scenario 1. **b**, PIN101, PIN103, PIN104, and PIN108 are scenario 2. In manual review of the PIN101 sequencing data, a read supporting the presence of the KRAS(G12D) variant was detected in both the PDAC and PanIN-A samples and was thus moved to the trunk of the phylogeny despite the overall low coverage of *KRAS* in PanIN-A. **c**, PIN106 and PIN107 are scenario 3. **d**, Jaccard indices from SNVs and indels. For each evolutionary scenario, the average Jaccard index for each patient was calculated from all driver and passenger variants (see Supplementary Table 6 for values) and plotted on a range from 0 to 1, with values closer to 1 denoting higher genetic similarity.

of genetic evaluation of morphologically distinct legions in revealing the evolutionary dynamics of pancreatic carcinogenesis. Because the PanINs were all anatomically distinct and far removed from the PDAC (see Methods), the data indicate that a single mutant clone had spread through the pancreatic ductal system to generate coexisting neoplastic lesions (Fig. 3a). This situation is similar to what occurs in the bladder, where a single clone can form multiple anatomically distinct neoplasms[23]. Though it would seem much more challenging for a neoplastic cell to travel through the fluid in the pancreatic ductal system than to travel through the urine, this journey has been described in intraductal papillary mucinous neoplasms of the pancreas, and clearly occurred in these four patients as well[24].

To assess genetic relatedness using all somatic variants, we quantified Jaccard similarity coefficients between pairs of lesions within each scenario (Fig. 2d, Supplementary Table 6). Notably, scenario 2 PanIN lesions tended to share fewer somatic variants with the matched PDAC than PanIN lesions in scenario 3 (average Jaccard similarity coefficient of 0.39 and 0.50, respectively), although the range of Jaccard similarity coefficients overlapped between the two scenarios (scenario 2 range, 0.10–0.57; scenario 3 range, 0.44–0.70).

Our phylogenetic analysis also enabled us to estimate the mutational signatures operating in different tumour lineages that led to the PDAC or a coexisting PanIN (Extended Data Figs. 6–8). Some signatures were shared between a PDAC and PanIN, whereas others operated only on a subset of different branches[25].

In PIN106 and PIN107, the PDACs and corresponding PanINs contained the same driver gene SNVs or indels (scenario 3; Fig. 2, Extended Data Fig. 5). In addition to the lost copies of *CDKN2A* and *SMAD4*, several unobserved factors might contribute to the morphological differences between lesions. First, the PDAC may have accumulated additional important genetic events in regions of the genome not assessed by whole-exome sequencing. Second, the PDACs may have acquired epigenetic alterations that were not detectable by the approach we used. Third, the microenvironment may have influenced the progression from a PanIN to a PDAC[13]. Finally, the PanIN lesions in PIN106 and PIN107 may represent cancerization of the ducts as described above. We note that the PDACs in these two patients showed moderate to poorly differentiated histology, thereby decreasing but not fully eliminating this possibility[12].

Modelling time of pancreatic cancer evolution

The WES data allow us to estimate the time required for a cell to progress from a non-invasive, neoplastic clone to an invasive pancreatic cancer[26] (see Methods). We used the number of acquired genetic passenger alterations from a common ancestor to the PanINs and the PDACs, after removing mutations suspected to be drivers or subclonal, to infer the amount of passed time. Because the great majority of the mutations present in any of these lesions are passengers and are not associated with positive or negative growth advantage, these mutations can serve as a molecular clock. On the basis of previously estimated mutation rates[27] and cell division times[28] measured in PanINs, we found that the median time elapsed between the common ancestral cell (Fig. 3b) and the birth of the founder clone of a PanIN was 7.1 years (90% confidence interval (CI) of the median: 3.3–12.2 years). Similarly, the median time elapsed between the common ancestral cell and the founder cell of the PDAC itself was 4.3 years (90% CI 2.3–7.2 years). Because the PanIN samples are monophyletic in all patients, we cannot estimate how long the primary tumour lineage might have existed as a PanIN. Nonetheless, these intervals are conservative underestimates of the times required to develop neoplasia and radiographically detectable cancer, because they do not include any clonal steps before the birth of the common ancestral cell nor the time between the birth of the PDAC founder cell and the multiplication of this cell to form a clinically evident mass. A larger patient cohort is required to assess whether this length of time is characteristic of the population of individuals with PDAC. When the time required for mass development is taken into account, the data suggest that an average of at least 8.1 years elapses between the birth of the common ancestral cell and the presence of a clinically evident mass (see Methods).

Discussion

Comparison of our results with three recent studies is informative. First, 77% of patients without clinically evident pancreatic neoplasia were found to harbour PanIN-1 lesions at autopsy[11]. In another study, LG-PanINs (PanIN-1 and PanIN-2) from the same patient generally did not share the same genetic alterations, in contrast to our data which show genetic relationships among HG-PanINs (PanIN-3)[9]. When taken together with our results, the data suggest that early neoplastic lesions in the pancreas may represent independent events, and that the success of the neoplastic cells in colonizing the ductal system is achieved only with histological progression and the accrual of additional genetic

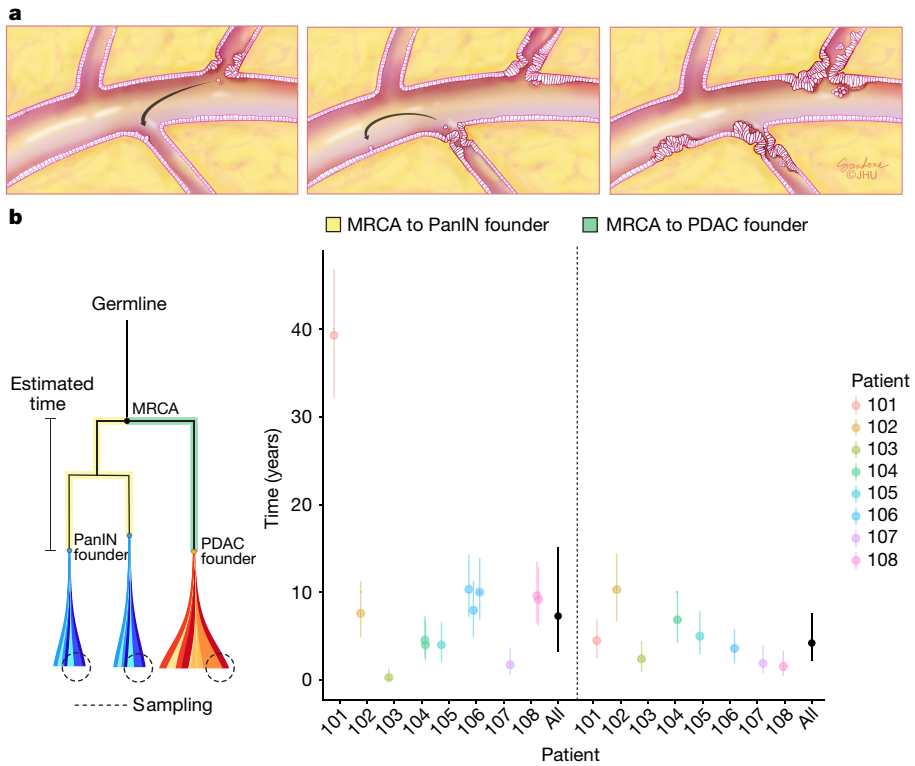


Fig. 3 | Putative growth pattern of coexisting PanIN(s) and PDAC and mathematical model. a, Spatial evolution and PanIN progression in intralobular ducts. LG-PanIN and HG-PanIN lesions represent precursors with differing degrees of nuclear and cytological atypia. An LG-PanIN develops and seeds a cell that travels to a second duct (arrow, left). The first LG-PanIN matures into an HG-PanIN, while an LG-PanIN develops at the second site and a cell subsequently travels to a third duct (arrow, middle). The second site LG-PanIN matures into an HG-PanIN while an LG-PanIN develops at the third site (right). **b**, Estimated progression times. The lineage leading from the MRCA to the PanINs is illustrated in yellow, while the lineage leading from the MRCA to the PDAC is in green. Clonal passenger mutations were used to estimate progression times, shown for each patient with 90% CIs. Overall (black), the inferred median time elapsed between the common ancestral cell and the birth of the founder clone of a PanIN was 7.1 years (90% CI, 3.3–12.2; MRCA to PanIN, $n = 12$). The inferred median time elapsed between the common ancestral cell and the PDAC was 4.3 years (90% CI, 2.3–7.2; MRCA to PDAC, $n = 8$). These estimates assume a mutation rate of 0.0224 per generation and a time per generation of four days (see Methods).

alterations. Notably, the budding off of small clusters of neoplastic epithelial cells into the lumen is one of the pathognomonic morphological features of an HG-PanIN (PanIN-3)[29].

Our data are apparently at odds with the interpretation of a recent study that concluded that PDACs do not arise gradually[14]. This conclusion was based on genetic analyses of microdissected PDACs and did not include an analysis of PanINs, and models were not applied to the data to support such a conclusion. As such, it relied on assumptions about the timing of transition from precursor lesion to invasive carcinoma. By contrast, our data are directly based on genomic analyses of the precursor lesions and their corresponding PDACs. Our step-wise progression model is supported not only by the current data but also by a body of scientific literature[17–20,22,26,30,31] that suggests that single or short base substitutions that gradually accumulate over many years form the great majority of the genetic alterations responsible for this type of tumour. Our findings in no way contradict the observation that multiple chromosome translocations can occur simultaneously (chromothripsis) in a small subset of pancreatic tumours[14,31]. However, they do buttress the idea that PDAC development is a multi-step progression caused by the accumulation of somatic alterations in driver genes, in a process that generally spans many years.

It could be argued that the cases we analysed were unusual in that more than one advanced PanIN was found in each pancreas, and our selection of eight out of approximately one-hundred patients could have introduced an unintended bias in our cohort. However, multiple advanced PanIN lesions are the norm rather than the exception when the entire pancreas is methodically dissected[11]. Furthermore, the mutations in driver genes and distribution of mutational signatures in this cohort are similar to those previously observed in pancreatic cancers. Finally, genomic analysis of a PDAC arising directly from an adjacent HG-PanIN lesion revealed a gradual genetic progression from PanIN to PDAC[8]—similar to our findings for anatomically separate HG-PanIN lesions and their corresponding PDACs.

In summary, we have shown that pancreatic intraepithelial neoplasia (that is, PanIN-2 and PanIN-3) need not be spatially localized lesions; rather, they are a disease that can spread through the entire ductal system. Additional studies—with more patients and a higher density of samples—will be required to determine the frequencies of the evolutionary scenarios we identified and to clarify which features of precursor lesions put them at substantial risk of transformation. Nonetheless, our data suggest that the multiple, apparently discrete PanIN lesions observed in an individual patient often represent a single neoplasm that can spread (contiguously or discontiguously) along the ductal system. This finding provides an explanation for the observation that patients who have had an HG-PanIN or PDAC removed by subtotal pancreatectomy are at high risk for the development of recurrent disease.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0481-8

Received: 1 February 2018; Accepted: 13 July 2018;
Published online 3 September 2018.

1. Vogelstein, B. et al. Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).

2. Vogelstein, B. & Kinzler, K. W. The path to cancer —three strikes and you're out. *N. Engl. J. Med.* **373**, 1895–1898 (2015).
3. Basturk, O. et al. A revised classification system and recommendations from the Baltimore consensus meeting for neoplastic precursor lesions in the pancreas. *Am. J. Surg. Pathol.* **39**, 1730–1741 (2015).
4. Hruban, R. H., Goggins, M., Parsons, J. & Kern, S. E. Progression model for pancreatic cancer. *Clin. Cancer Res.* **6**, 2969–2972 (2000).
5. van Heek, N. T. et al. Telomere shortening is nearly universal in pancreatic intraepithelial neoplasia. *Am. J. Pathol.* **161**, 1541–1547 (2002).
6. Kanda, M. et al. Presence of somatic mutations in most early-stage pancreatic intraepithelial neoplasia. *Gastroenterology* **142**, 730–733.e9 (2012).
7. Makohon-Moore, A. & Iacobuzio-Donahue, C. A. Pancreatic cancer biology and genetics from an evolutionary perspective. *Nat. Rev. Cancer* **16**, 553–565 (2016).
8. Murphy, S. J. et al. Genetic alterations associated with progression from pancreatic intraepithelial neoplasia to invasive pancreatic tumor. *Gastroenterology* **145**, 1098–1109.e1 (2013).
9. 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).
10. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
11. 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).
12. Yamasaki, S., Suda, K., Nobukawa, B. & Sonoue, H. Intraductal spread of pancreatic cancer. Clinicopathologic study of 54 pancreatectomized patients. *Pancreatology* **2**, 407–412 (2002).
13. Kleeff, J. et al. Pancreatic cancer. *Nat. Rev. Dis. Primers* **2**, 16022 (2016).
14. Notta, F. et al. A renewed model of pancreatic cancer evolution based on genomic rearrangement patterns. *Nature* **538**, 378–382 (2016).
15. Roberts, N. J. et al. Whole genome sequencing defines the genetic heterogeneity of familial pancreatic cancer. *Cancer Discov.* **6**, 166–175 (2016).
16. Jones, S. et al. Comparative lesion sequencing provides insights into tumor evolution. *Proc. Natl Acad. Sci. USA* **105**, 4283–4288 (2008).
17. Biankin, A. V. et al. Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
18. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
19. Witkiewicz, A. K. et al. Whole-exome sequencing of pancreatic cancer defines genetic diversity and therapeutic targets. *Nat. Commun.* **6**, 6744 (2015).
20. Bailey, P. et al. Genomic analyses identify molecular subtypes of pancreatic cancer. *Nature* **531**, 47–52 (2016).
21. Reiter, J. G. et al. Reconstructing metastatic seeding patterns of human cancers. *Nat. Commun.* **8**, 14114 (2017).
22. 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).
23. Sanli, O. et al. Bladder cancer. *Nat. Rev. Dis. Primers* **3**, 17022 (2017).
24. Pea, A. et al. Targeted DNA sequencing reveals patterns of local progression in the pancreatic remnant following resection of intraductal papillary mucinous neoplasm (IPMN) of the pancreas. *Ann. Surg.* **266**, 133–141 (2017).
25. Roerink, S. F. et al. Intra-tumour diversification in colorectal cancer at the single-cell level. *Nature* **556**, 457–462 (2018).
26. Yachida, S. et al. Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114–1117 (2010).
27. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc. Natl Acad. Sci. USA* **110**, 1999–2004 (2013).
28. Klein, W. M., Hruban, R. H., Klein-Szanto, A. J. P. & Wilentz, R. E. Direct correlation between proliferative activity and dysplasia in pancreatic intraepithelial neoplasia (PanIN): additional evidence for a recently proposed model of progression. *Mod. Pathol.* **15**, 441–447 (2002).
29. 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).
30. Jones, S. et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
31. Reiter, J. G. & Iacobuzio-Donahue, C. A. Pancreatic cancer: Pancreatic carcinogenesis—several small steps or one giant leap? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 7–8 (2016).

Acknowledgements Supported by the V Foundation for Cancer Research, NIH grants F31 CA180682, 2T32 CA160001-06 and 5T32 CA067751-13, an Erwin Schrödinger fellowship (Austrian Science Fund FWF J-3996), SPORE grant P50 CA062924, the Michael Rolfe Foundation, The Lustgarten Foundation for Cancer Research, the Sol Goldman Center for Pancreatic Cancer Research, The Virginia and D.K. Ludwig Fund for Cancer Research and D. Troper and S. Wojcicki.

Reviewer information *Nature* thanks A. V. Biankin, S. J. Chanock and F. Markowetz for their contribution to the peer review of this work.

Author contributions A.P.M.-M., K.M., Y.J., N.P., K.K., B.V., and C.I.-D. designed the study; K.M., S.Y., R.H.H., and C.I.-D. selected the samples; S.Y., K.M., R.H.H., D.S.K., and C.I.-D. reviewed pathology; S.Y., K.M. and M.Z. prepared the DNA samples; A.P.M.-M., K.M., M.Z., Y.J., N.P., K.K., B.V., and C.I.-D. performed sequencing, alignment and mutation calling; A.P.M.-M., J.G.R., J.M.G., and M.A. derived the phylogenies; A.P.M.-M., J.G.R., J.M.G., M.A., and L.S. analysed the structural variants; J.G.R., J.M.G., and M.A.N. performed mathematical modelling; C.S. illustrated the spatial evolution of the lesions; and all authors wrote the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0481-8.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0481-8.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to C.I.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

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

Patient selection. Human tissues were collected with the approval of the Johns Hopkins Hospital Institutional Review Board (protocols NA_00001584 and NA_00017879) after informed and written consent was obtained, following all relevant ethical regulations. Fresh-frozen samples from eight patients who underwent surgical resection of pancreatic cancer at Johns Hopkins Hospital (January 2009–December 2011) with pathological confirmation of PDAC and geographically distinct PanIN-2 or PanIN-3 lesions were selected for study. For inclusion in the study, PDAC, PanIN lesion(s), and normal duodenum tissue were required for each patient. To minimize the possibility of studying cancerization of normal ducts, we included only PanINs in which at least 1.0 cm of uninvolved lobular parenchyma was present between the PanIN and the cancer, or in which the PanINs were present in a block that contained no cancer.

Processing of tissue samples. For each tissue sample, multiple sequential 5-μm-thick cryosections were mounted on polyethylene naphthalate (PEN) membrane slides and stained with cresyl violet for visualization of histological features and confirmation of adequate cellularity. Neoplastic epithelium was laser-microdissected using the Leica LMD7 laser microdissection system.

DNA extraction and quantification. Genomic DNA (gDNA) was extracted from each normal, PanIN, or tumour piece using a standard phenol and chloroform extraction followed by precipitation in ethanol. The gDNA was quantified by LINE assay (counting long interspersed elements (LINEs) using real-time PCR). The LINE forward primer was 5′-AAAGCCGCTCAACTACATGG-3′ and the reverse primer was 5′-TGCTTTGAATGCGTCCCAGAG-3′. The real-time PCR protocol was 50 °C for 2 min, 95 °C for 2 min, 40 cycles of 94 °C for 10 s, 58 °C for 15 s, and 70 °C for 30 s, 95 °C for 15 s, and 60 °C for 30 s. The PCR reactions were carried out using Platinum SYBR Green qPCR mastermix (Invitrogen).

WES and alignment. WES was performed on an Illumina HiSeq 2000 platform for a target coverage of 150×. Upon the completion of WES, the data were analysed in silico to determine overall quality and coverage. Sequencing reads were aligned to the hg19 human reference genome using BWA[32]. Read de-duplication, base quality recalibration, and multiple sequence realignment were performed using the Picard Suite and GATK version 3.1[33,34]. SNVs were called using Mutect version 1.1.6 and indels were detected using HaplotypeCaller version 2.4[33,35].

Filtering of WES data. WES generated a large list of potential mutations, and we evaluated these data to identify high-quality mutations while removing sequence artefacts. Each mutant must have been observed with at least 5% variant allele frequency with 20× coverage in at least one neoplastic sample; each mutant must have been observed in less than 2% of the reads (or three reads total) of the matched normal sample with 10× coverage. This filtering yielded a total of 2,886 mutations for subsequent analysis (Supplementary Table 2).

Driver gene and mutation analysis. All somatic variants causing a frameshift deletion, frameshift insertion, in-frame deletion, in-frame insertion, missense, nonsense, nonstop or splice site/region mutation, or a translation start site were considered. If a variant was a missense or nonsense mutation, we required the variant to have a CHASM *P* value ≤ 0.05 and an FDR ≤ 0.25. In combination with manual review, driver gene mutations were identified if the gene was supported by at least three of the following four methods: 20/20+[36], TUSON[37], MutSigCV[38] (see Supplementary Table 1 in ref. [36] for gene list), and a hotspot analysis[39]. In addition, we also considered genes significantly mutated in large PDAC sequencing studies[17,18,20,40]. Further, we required that each somatic variant have a variant allele frequency of <2% in the patient-matched normal tissue as well as any normal tissue from another patient. If a deleterious variant was detected in a driver gene as described above, and was not detected abundantly in any normal tissue, it was considered a driver gene variant.

CNAs. Allele-specific copy number analysis was performed using FACETS[41]. In brief, FACETS performs a complete analysis that includes library size and GC-normalization, and segmentation of total and allele-specific signals, using coverage and genotypes of single nucleotide polymorphisms simultaneously across the exome[41]. The resulting segments accurately identify points of change in the exome, accounting for diploidy, purity, and average ploidy for each sample. A maximum likelihood approach then assigns each segment with a major and minor integer copy number.

Evolutionary analysis. We derived phylogenies for each set of samples using Treeomics 1.7.9[21]. Each phylogeny was rooted at the matched patient's normal sample and the leaves represented the PanIN or tumour samples. Treeomics employs a Bayesian inference model to account for error-prone sequencing and varying neoplastic cell content to calculate the probability that a specific variant is present or absent. Treeomics infers the global optimal tree based on Mixed Integer Linear programming. For Extended Data Figs. 3–5, the CNAs were not directly used to infer phylogenies in order to prevent bias from potential false-negatives or false-positives, given that CNA calls from multiple samples within a patient are particularly sensitive to varying neoplastic cell content and depth of sequencing. Moreover, WES data usually do not capture the exact breakpoints of CNAs, further complicating phylogenetic analysis. Nevertheless, the common PDAC driver genes *KRAS*, *MYC*, *GATA6*, and *CDK6* were manually reviewed in the CNA data for evidence of gains, while *CDKN2A*, *SMAD4*, *TP53*, *MAP2K4*, *TGFBR2*, and *ACVR1B* were queried for losses. Allelic losses were defined as total copy number (tcn) = 1 or 0, and gains were defined as tcn $\geq$ 4. Given the CNA status of a given driver gene in each sample, the driver gene with the CNA status was manually placed on the corresponding position edge in the phylogeny (previously derived using SNVs/indels). This approach was used with each PDAC driver gene affected by a CNA as defined above.

Our classification of each patient into one of three evolutionary scenarios was based on SNVs/indels that affect key driver genes in PDAC (for example, KRAS(G12D)). Such alterations represent driver gene variants that are readily interpretable with respect to function as well as position on the phylogenetic tree. Nonetheless, CNAs can also affect driver genes involved in pancreatic cancer (for example, *CDKN2A* deletion). If we reclassify the eight patients using both SNVs/indels and CNAs affecting driver genes (Extended Data Figs. 3–5), we find that the evolutionary scenario does not change for six patients. For two patients (PIN106 and PIN107), the scenario changes from scenario 3 to scenario 2, indicating a step-wise progression of PanINs and PDACs for all eight patients. As noted above, the identification and placement of CNAs on a phylogenetic tree remains challenging. Nonetheless, we note that the SNV/indel phylogenies represent a minimum number of evolutionary steps: including additional CNAs would either confirm or increase the total number of steps in the evolution of the PDAC.

Structural variant analysis. We inferred structural variants (SVs) using DELLY2 (v.0.7.5) to verify the reconstructed phylogenies[42]. As the SVs were called for each sample independently, we merged SVs for which DELLY determined breakpoints differing by at most 250 base pairs among the samples of each patient. In total, we found 154 distinct SVs in the eight subjects. After a comprehensive manual review of the called SVs, we developed additional criteria to minimize the number of false-positives. We required that each SV had to pass one of the following two filters in at least one sample: 1) (a) SV is supported by at least three distinct split reads, (b) the ratio of split reads that support the SV to the total number of split reads at the position of the SV is greater or equal to 0.75, and (c) the number of the SV supporting split reads is greater than the number of split reads in the normal sample; or 2) (d) the SV is supported by at least five discordantly paired (DP) reads, (e) the ratio of DP reads that support the SV to the total number of DP reads at the position of the SV is greater or equal to 0.25, and (f) the number of the SV-supporting DP reads is greater than the number of DP reads in the normal sample. After applying these filters, we obtained 40 SVs (Supplementary Table 4).

To create input files for Treeomics, we used the number of SV-supporting split and DP reads as the number of variant reads. We normalized the coverage of SVs such that on average it approximately matched the median coverage of SNVs. Generally, the inferred phylogenies based on the SVs agreed well with the ones based on SNVs. However, because of the lower number of SVs per subject (median 4; range 0–14; Supplementary Table 4), the confidence in the inferred branches was substantially lower than in the phylogenies based on SNVs. For PIN106 (coverage in sample of PanIN-A was extremely low), we inferred a slightly different phylogeny as PanIN-A diverged before the PDAC, probably owing to many false-negatives resulting from the extremely low coverage and the therefore difficult detection of SVs in this sample. For PIN108, no SVs were shared across multiple samples and hence there were no parsimony-informative SVs such that a phylogeny could be inferred.

Mutation signatures. We assessed the presence of previously identified mutational signatures[43] in each patient. Our phylogenetic analysis enabled us to estimate the signatures operating at different stages of cancer evolution[25]. For SNVs acquired along each phylogenetic branch, we estimated the maximum likelihood signature proportions among 30 previously identified trinucleotide signatures[44] (see https://github.com/mskcc/mutation-signatures). We quantified the uncertainty in these estimates by performing 100 iterations of bootstrap resampling within each branch followed by signature re-estimation. We ignored branches with five or fewer mutations and removed signature 24 because of its similarity to smoking. The maximum likelihood signature estimates and 90% bootstrap confidence intervals for each branch are shown in Extended Data Figs. 6–8. We detect signatures 1, 2, 3, and 6, consistent with previous studies[43]. Additionally, we find evidence for signatures 4 (associated with smoking) and 29 (associated with chewing tobacco). Signatures operating on different branches within a patient were not significantly more similar than those across patients (mean cosine distance similarity 0.62 versus 0.59, $P = 0.21$, one-sided permutation test). We note that signature estimates had large bootstrap uncertainty and the number of patients as well as the number of mutations is limited.

Progression time inference. We assume that the number of passenger mutations n acquired along a lineage during time T (in cell generations) is Poisson-distributed

with rate equal to T times the mutation rate per cell division[16]: $n_\mu \mid T \sim \text{Poisson}(\mu T)$. We assume that a random sample from the population of PanINs or PDACs takes T generations to progress from a previous stage (either MRCA of all sampled PanINs and PDAC in a patient or the MRCA of the most closely related PanIN to the PDAC) to the founder of a particular PanIN or PDAC, and that the mutational clock time μT is gamma-distributed with hyperparameters shape k and scale θ ($k, \theta > 0$) uniform a priori: $\mu T \sim \text{Gamma}(k, \theta)$.

In order to infer the joint distribution of (T, k, θ), we use the following sampling strategy. For each sample i, we update T by sampling directly from the gamma posterior:

$$T_i\mu|n, k, \theta \sim \text{Gamma}\left(k + n, \frac{\theta}{1+\theta}\right)$$

Using the updated values, we jointly update k, θ by Metropolis–Hastings sampling from the posterior:

$$L(k, \theta) \propto \pi(k, \theta) \prod_{T_i} d\text{Gamma}(k, \theta, \mu T_i)$$

where dGamma is the density function for the gamma distribution and $\pi(k,\theta)$ is the prior over the hyperparameters (uniform). This setup pools information about the time to progression for each sample towards the population of progression time estimates, with a flexible structure for the overall distribution of times provided by the gamma distribution.

In order to convert the inferred number of generations to absolute time, we follow a previous method[26] by multiplying by the average time for cell division. To estimate the division time, we again follow the previous method but instead note that 14% of stage II PanINs stain positively for Ki-67[28]. We therefore estimate the generation time of PanIN stage II cells to be 4 days. The mutation rate μ per generation is 0.0224, calculated for 35 Mb of exome sequencing multiplied by a point mutation rate of 6.4×10^{-10} per generation[27].

To calculate the expected time it takes for the PDAC founding cell to grow to a detectable lesion of 1 cm^3 ($\sim 10^9$ cells), we used previously measured PDAC metastasis doubling times of 56 days[45] leading to an exponential growth rate of $r = 0.012$ per day. The probability density function for the time an exponential branching process conditioned on survival takes to reach size $M = 10^9$ is approximately given by:

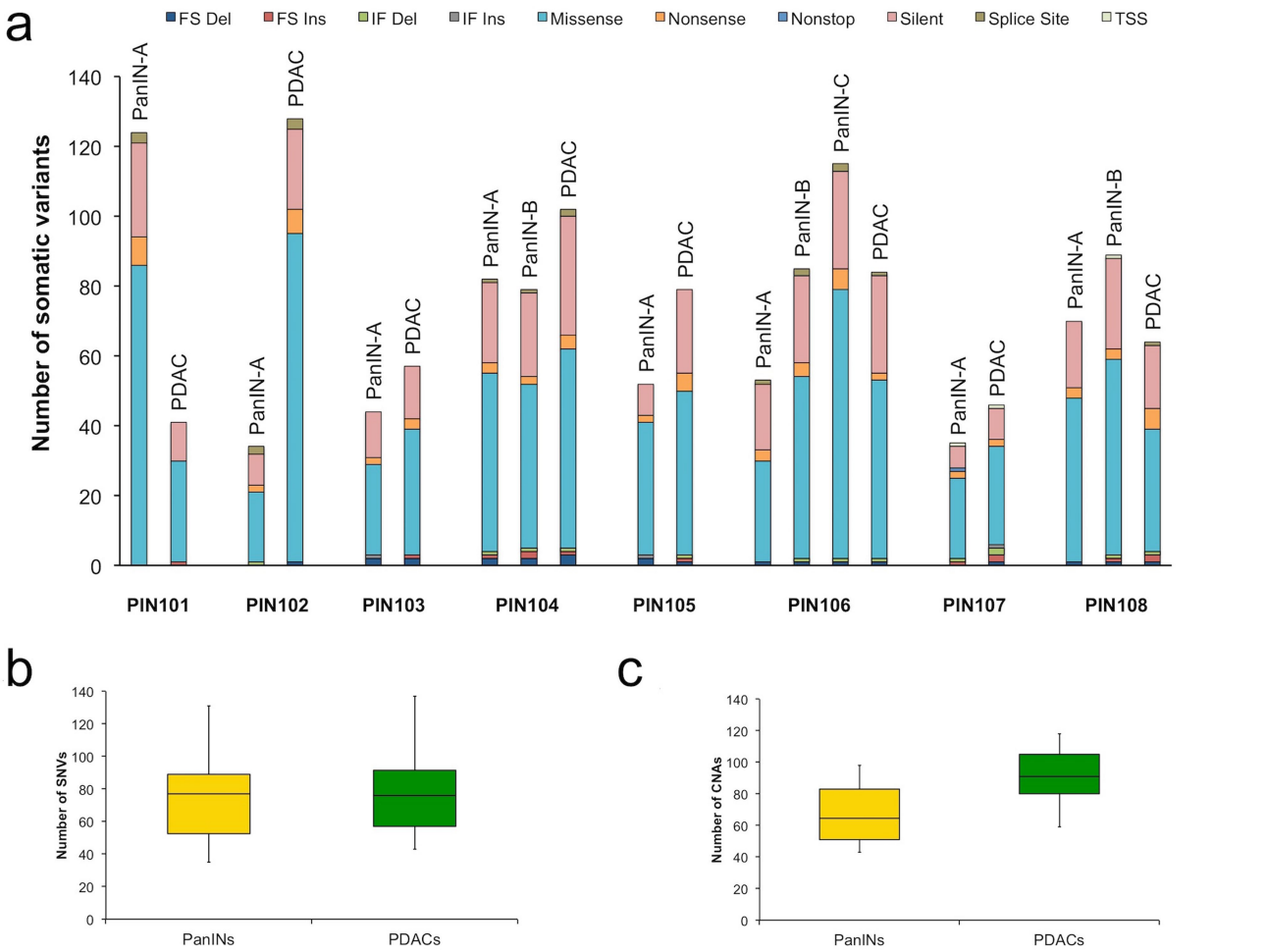
$$f_{t_M}(t) = \exp\left(-\frac{r}{b} M \exp(-rt)\right) \frac{r^2 M}{b} \exp(-rt)$$

Where $b = 1/2.3$ per day is the assumed PDAC cell division rate[26,46].

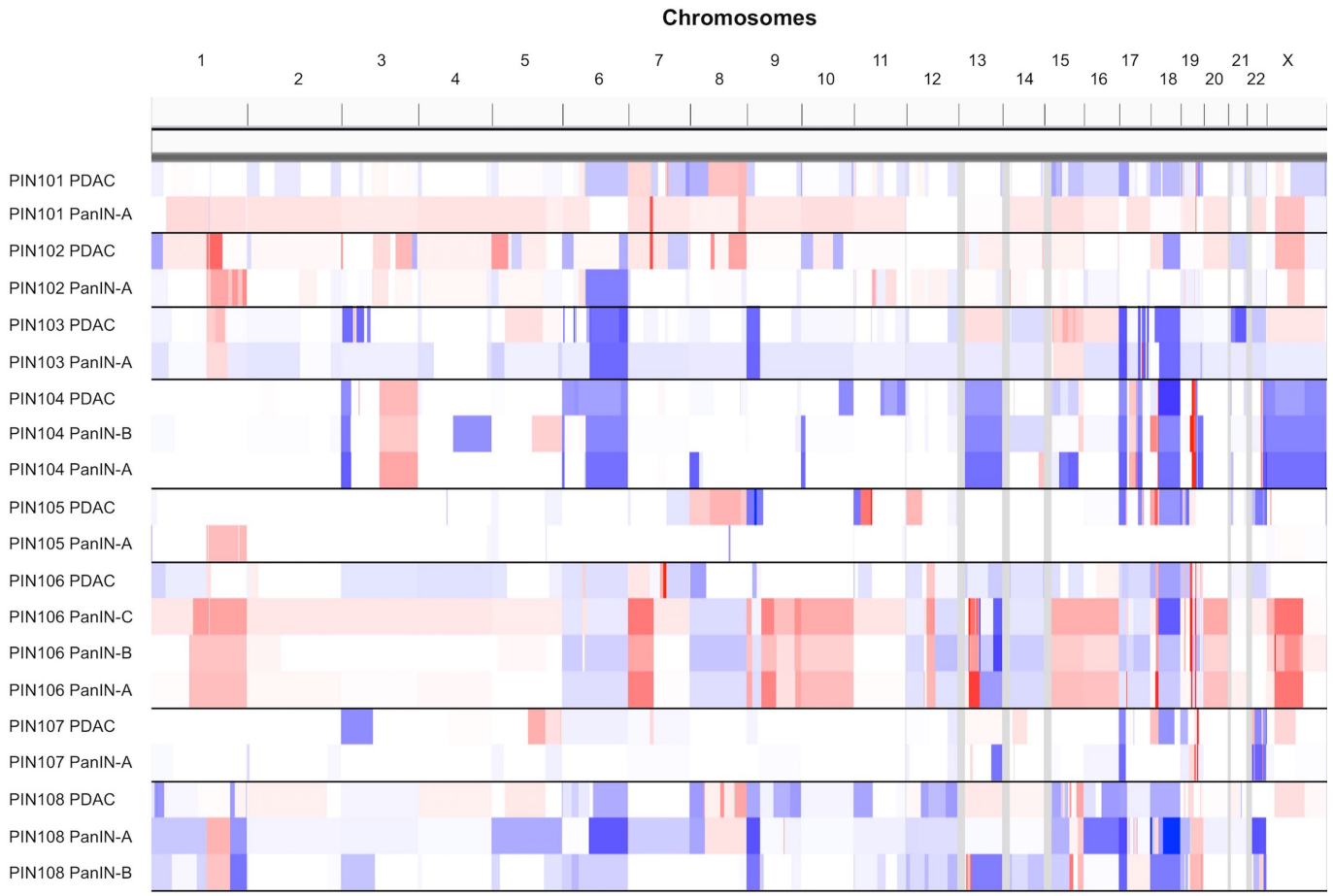
Reporting summary. Further information on experimental design is available in the Nature Research Reporting Summary linked to this paper.

Data availability. Sequence data have been deposited at the European Genome-phenome Archive (EGA; https://www.ebi.ac.uk/ega/), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAS00001002778. Source data are provided for Fig. 3b and Extended Data Figs. 1, 7, 8. All other relevant data are included within the manuscript or are available upon request from the corresponding author.

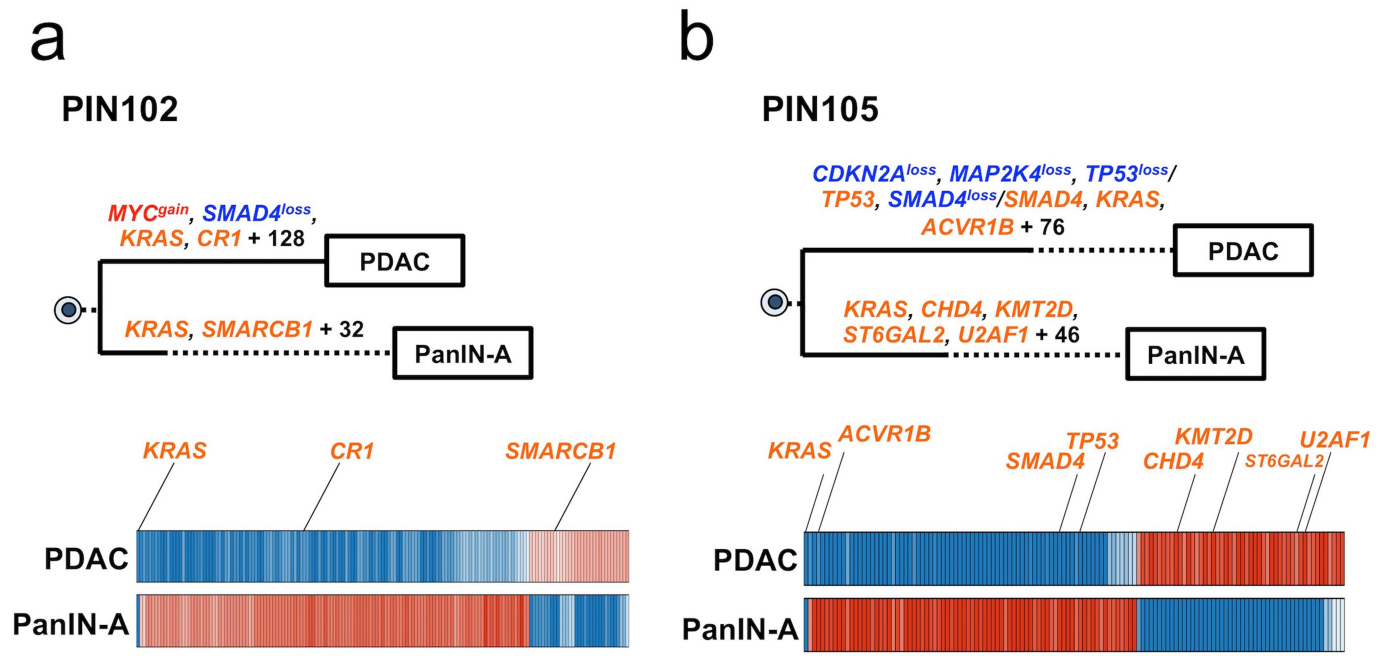
32. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
33. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
34. Mose, L. E., Wilkerson, M. D., Hayes, D. N., Perou, C. M. & Parker, J. S. ABRA: improved coding indel detection via assembly-based realignment. *Bioinformatics* **30**, 2813–2815 (2014).
35. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
36. Tokheim, C. J., Papadopoulos, N., Kinzler, K. W., Vogelstein, B. & Karchin, R. Evaluating the evaluation of cancer driver genes. *Proc. Natl Acad. Sci. USA* **113**, 14330–14335 (2016).
37. Davoli, T. et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
38. Lawrence, M. S. et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
39. Chang, M. T. et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat. Biotechnol.* **34**, 155–163 (2016).
40. Cancer Genome Atlas Research Network Integrated genomic characterization of pancreatic ductal adenocarcinoma. *Cancer Cell* **32**, 185–203.e13 (2017).
41. Shen, R. & Seshan, V. E. FACETS: allele-specific copy number and clonal heterogeneity analysis tool for high-throughput DNA sequencing. *Nucleic Acids Res.* **44**, e131 (2016).
42. Rausch, T. et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics* **28**, i333–i339 (2012).
43. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
44. Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J. & Stratton, M. R. Deciphering signatures of mutational processes operative in human cancer. *Cell Reports* **3**, 246–259 (2013).
45. Amikura, K., Kobari, M. & Matsuno, S. The time of occurrence of liver metastasis in carcinoma of the pancreas. *Int. J. Pancreatol.* **17**, 139–146 (1995).
46. Durrett, R. in *Branching Process Models of Cancer* 1–63 (Springer International Publishing, Switzerland, 2015).



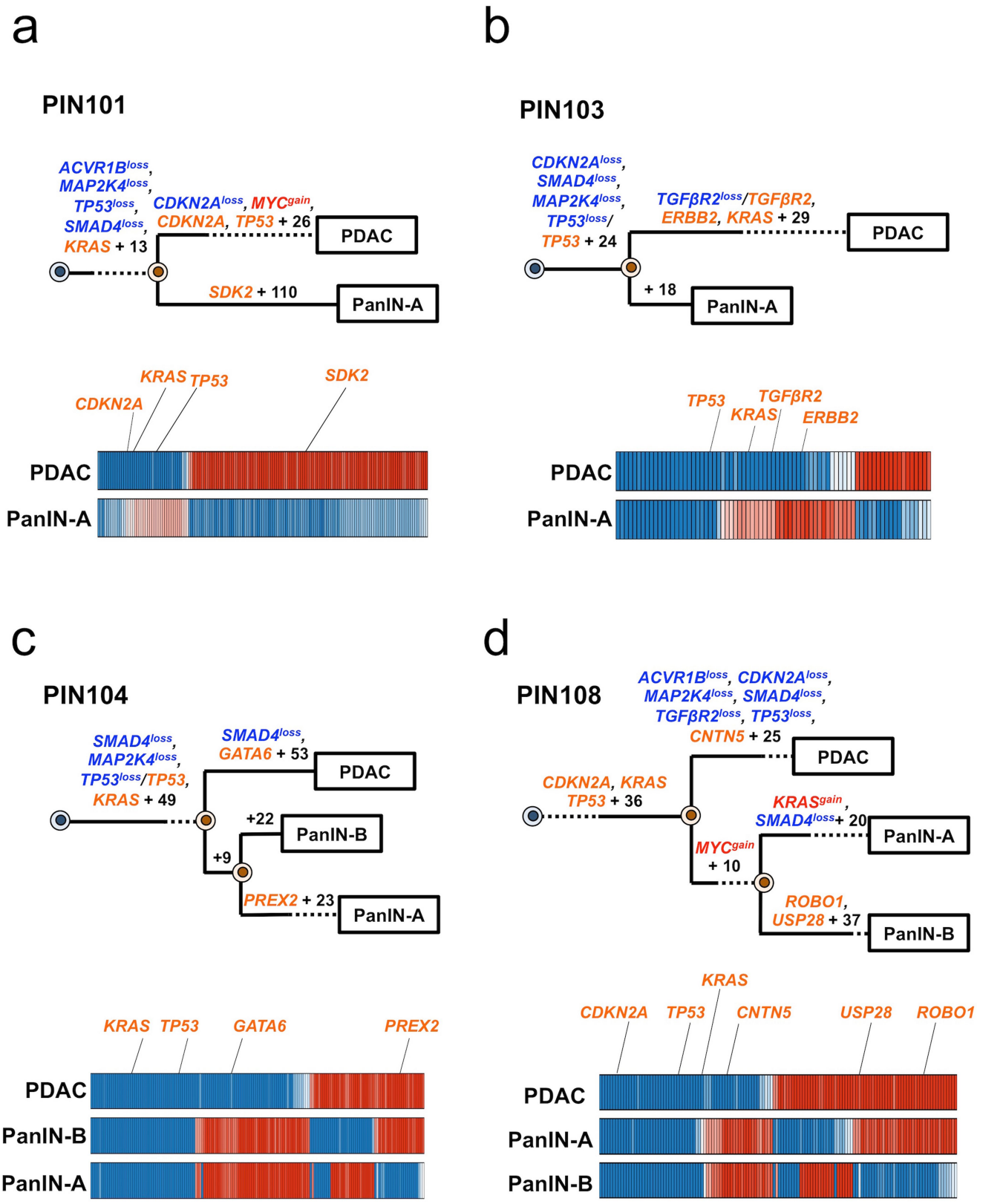
Extended Data Fig. 1 | Mutation counts and features of samples. **a**, Number of somatic mutations detected per sample with clinical features of each patient. The *y*-axis shows mutation counts and the *x*-axis shows patient ID. FS Del, frameshift deletion; FS Ins, frameshift insertion; IF Del, in-frame deletion; IF Ins, in-frame insertion; TSS, transcription start site. **b**, **c**, Box and whisker plots comparing number of somatic SNVs and CNAs between PanINs and PDACs. Yellow, PanINs; green, PDACs. Horizontal line indicates median, whiskers indicate minimum and maximum, and box indicates quartiles. $n = 12$ independent PanIN lesions, $n = 8$ PDACs. **b**, Number of SNVs and indels in PanINs and PDACs. **c**, Number of CNAs in PanINs and PDACs (hisens results only).



Extended Data Fig. 2 | Allelic CNAs across all patient samples. CNAs were inferred using the FACETS algorithm (Supplementary Table 3, FACETS. purity variants shown). Blue, putative losses; red, putative gains.

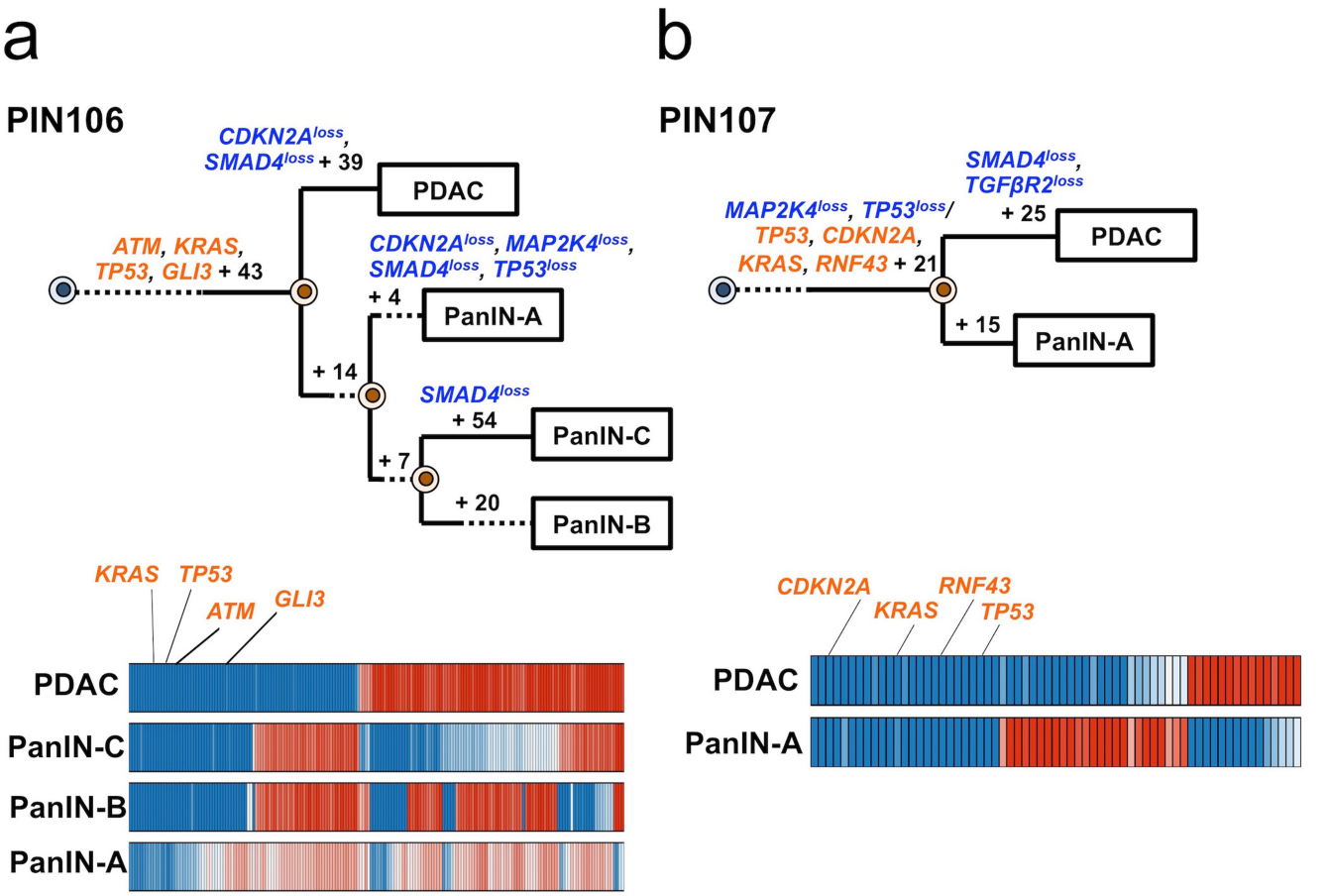


Extended Data Fig. 3 | Phylogenetics of PanINs and the matched primary tumour for patients PIN102 and PIN105. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant was manually reviewed to verify phylogenetic position. For each phylogeny, the number of acquired mutations is in black. The branch lengths are proportional to the number of SNVs and indels. The dashed line indicates the branch from the germline to the PDAC and PanIN-A. For the Bayesian heat maps, samples are indicated on each row while variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, Phylogenetic tree and Bayesian heat map with each variant for PIN102. **b**, Phylogenetic tree and Bayesian heat map with each variant for PIN105.

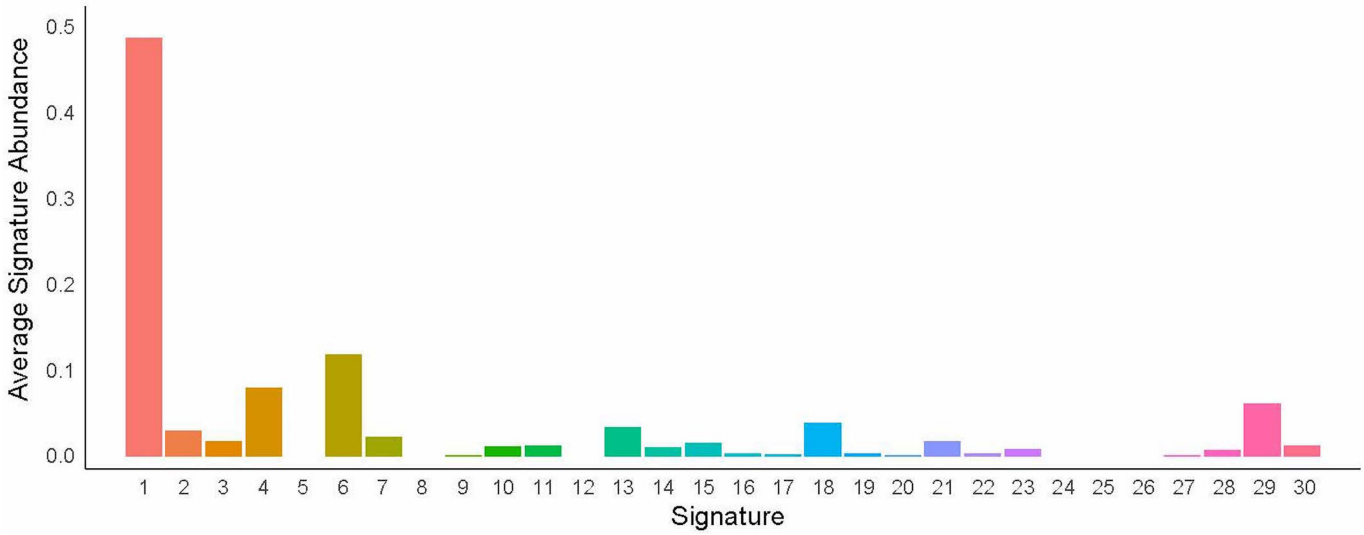


Extended Data Fig. 4 | See next page for caption.

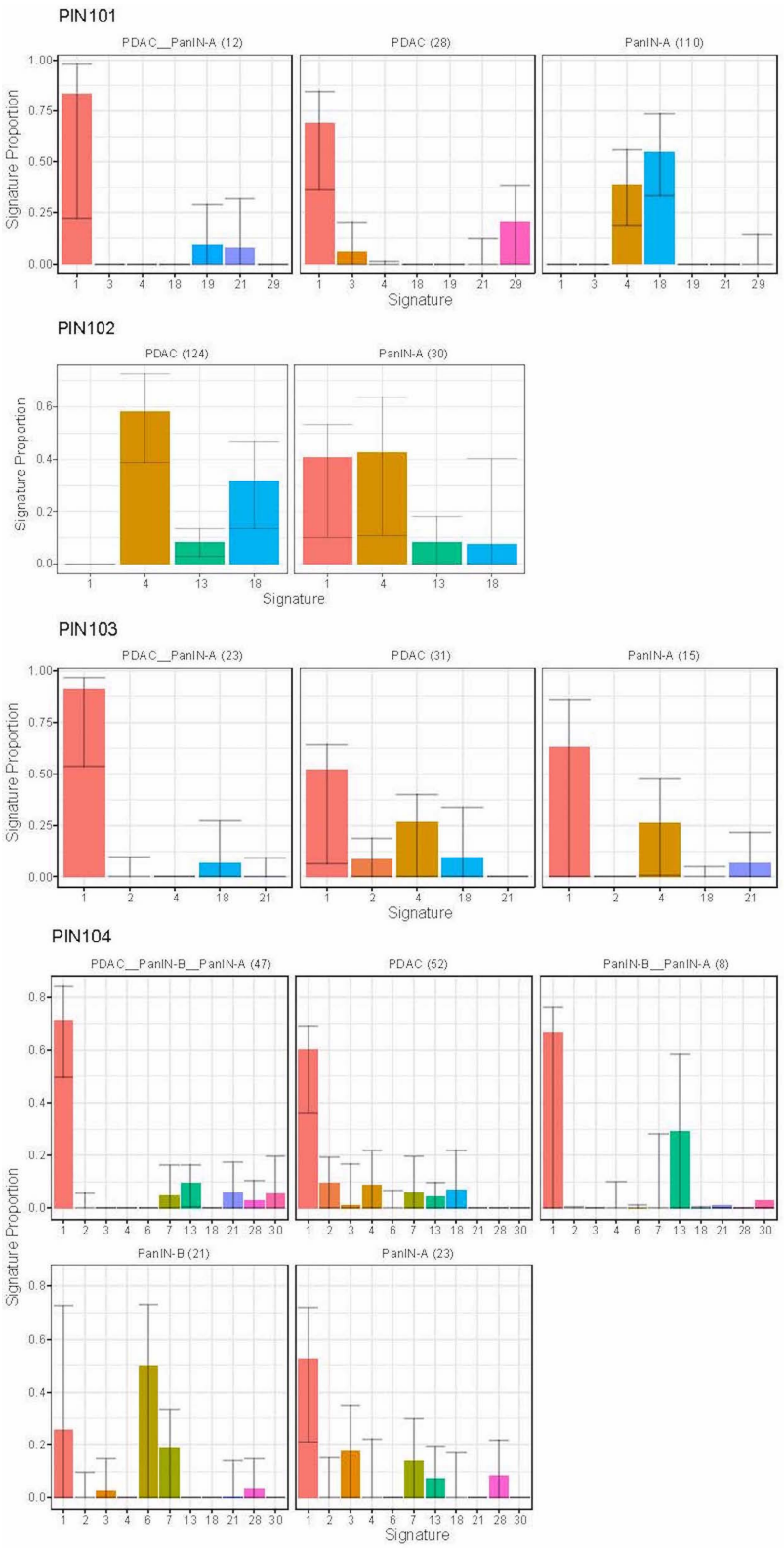
Extended Data Fig. 4 | Phylogenetics of PanINs and matched primary tumours for patients PIN101, PIN103, PIN104, and PIN108. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant were manually reviewed to verify phylogenetic position. For each phylogenetic tree, the numbers of acquired mutations are in black. The branch lengths are proportional to the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. For the Bayesian heatmaps, samples are indicated on each row and variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, PIN101. In manual review of the sequencing data, a read supporting the presence of the KRAS(G12D) variant was detected in both the PDAC and PanIN-A samples and was thus moved to the trunk of the phylogeny despite the overall low coverage of *KRAS* in PanIN-A. **b**, PIN103. **c**, PIN104. The node leading from the first MRCA to the second MRCA has a confidence value of >99%. **d**, PIN108. The node leading from the first MRCA to the second MRCA has a confidence value of >99%.



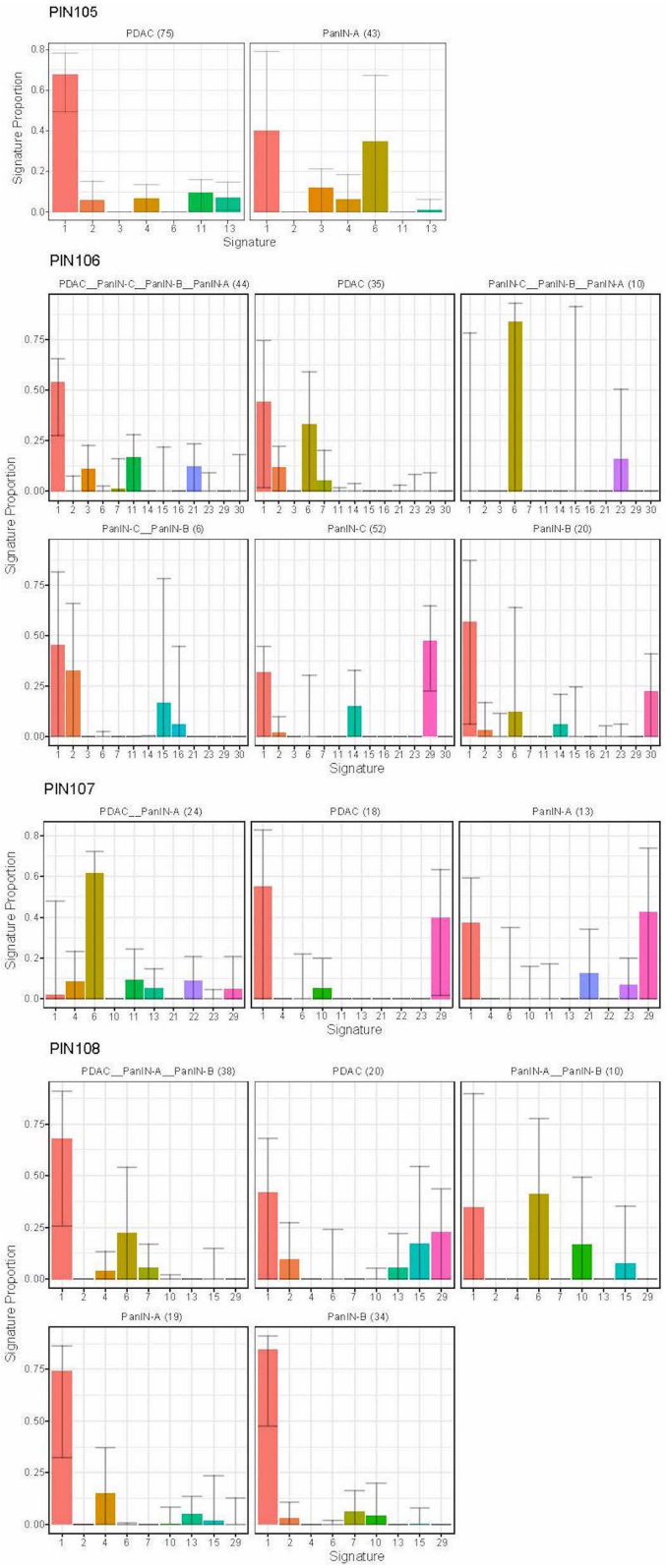
Extended Data Fig. 5 | Phylogenetics of PanINs and matched primary tumours for patients PIN106 and PIN107. See Supplementary Table 1 for sample identities. Gene names in orange text are SNVs or indels, in blue are copy-number losses, and in red are copy-number gains affecting putative driver genes. The sequencing data for each driver gene variant were manually reviewed to verify phylogenetic position. For each phylogeny, the numbers of acquired mutations are in black. The branch lengths are proportional to the number of SNVs or indels. The dashed lines indicate branches that have been extended to accommodate gene annotation and variant numbers. For each Bayesian heat map, samples are indicated on each row while variants are represented by each column. The colour of each tile indicates the probability that the variant is present or absent in the corresponding sample. Dark blue, >99.9% probability of being present; dark red, >99.9% probability of being absent. Light blue and red indicate lower probabilities; white tiles indicate approximately 50% probability. **a**, PIN106. The node leading from the first MRCA to the second MRCA has a confidence value of >99% and the node leading from the second MRCA to the third MRCA has a confidence value of 82%. **b**, PIN107.



Extended Data Fig. 6 | Average signature abundance across samples. Signature numbers 1–30 from Alexandrov et al.[43] are shown on the *x*-axis with signature abundance averaged across phylogenetic branches shown on the *y*-axis. Each histogram is coloured by signature identity.



Extended Data Fig. 7 | The proportion of mutational signatures from Alexandrov et al.[43] estimated in PIN101–PIN104. Signatures are shown on the *x*-axis, with the proportion of each signature shown on the *y*-axis. Each bar is coloured by signature identity. The text on the top of each panel denotes the corresponding phylogenetic branch and the number of mutations acquired along it in parentheses. Error bars depict 90% CIs in the signature proportion estimated by 100 iterations of bootstrap resampling.



Extended Data Fig. 8 | The proportion of mutational signatures from Alexandrov et al.[43] estimated in PIN105–PIN108. Signatures are shown on the *x*-axis, with the proportion of each signature shown on the *y*-axis. Each bar is coloured by signature identity. The text on the top of each panel denotes the corresponding phylogenetic branch and the number of mutations acquired along it in parentheses. Error bars depict 90% CIs in the signature proportion estimated by 100 iterations of bootstrap resampling.

nature research

Corresponding author(s): Christine Iacobuzio-Donahue

Reporting Summary

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

Statistical parameters

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

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

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No previously published data were collected for this study.
Data analysis	WES analysis: BWA was used for alignment, Picard and GATK version 3.1 were used for read de-duplication, base quality recalibration, and multiple sequence realignment, somatic mutations were called using Mutect version 1.1.6 and HaplotypeCaller version 2.4. Treeomics version 1.7.9 was used for deriving phylogenies. CHASM was used to analyzed driver gene variants (http://www.cravat.us.). FACETs was used to call allele specific copy number variants, and DELLY2 (v.0.7.5) was used for detecting structural variants. Mutation signatures were analyzed according to the strategy here: https://github.com/mskcc/mutation-signatures.

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

Data

Policy information about availability of data

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

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

Sequence data have been deposited at the European Genomephenome Archive (EGA), which is hosted by the European Bioinformatics Institute (EBI) and the Centre for Genomic Regulation (CRG), under accession number EGAS00001002778. Further information about EGA can be found at https://ega-archive.org and "The European Genome-phenome Archive of human data consented for biomedical research" (http://www.nature.com/ ng/journal/v47/n7/full/ng.3312.html). All other relevant data are included within the manuscript or are available upon request from the corresponding author (C.I-D.).

Field-specific reporting

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

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

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

Life sciences study design

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

Sample size	The sample size was determined by a review of >100 resected pancreata from patients within a three-year interval (from 2009 to 2011) that met the study criteria: no prior personal or family history of pancreatic cancer, at least one tumor sample collected, at least one distinct precursor lesion collected, and the availability of patient matched normal duodenum sample. This approach yielded eight patients from whom eight pancreatic cancer specimens and 12 independent precursor lesions were collected. Given that multiple lesions were collected per patient, we could reliably detect at least one evolutionary scenario per patient.
Data exclusions	To include only high quality somatic variants in the phylogenetic analysis, pre-determined criteria required that each mutant must have had at least a 5% variant allele frequency with 20x coverage in one or more neoplastic samples within a patient, and that each mutant had 2% or less of the reads (or 3 reads max) in the matched normal sample with 10x coverage.
Replication	All three evolutionary scenarios were detected among multiple pairs of neoplastic samples across the eight patients. In addition, we observed each evolutionary scenario independently in at least two patients.
Randomization	The analyses focused on a cohort of eight pancreatic cancer patients. The study was not an experimental or clinical trial, and thus no randomization was performed. The patients were not subdivided prior to analysis.
Blinding	For sample collection and pathological review, the investigators were not blinded to ensure that the study criteria were met for each case. For data collection and analysis, the investigators were blinded to patient identity. The investigators were not blinded to each sample's identity (precursor or tumor), however pathological status did not contribute to or influence mutation calling, driver gene variant annotation or the phylogenetic analyses.

Reporting for specific materials, systems and methods

Materials & experimental systems

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

Methods

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

Unique biological materials

Policy information about availability of materials

Obtaining unique materials	Given their microscopic size, human tissues and DNA samples were exhausted for this study and are not available.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Supplementary Table 1 provides information regarding patient age, gender, tumor staging, and genotype status. All human patients were diagnosed with pancreatic cancer and underwent surgical resection at Johns Hopkins Hospital. For inclusion, samples from each pancreatic cancer, pancreatic intraepithelial neoplasia, and normal duodenum were required.
Recruitment	No patient recruitment was performed specifically for this study. The tissues analyzed were collected from patients who received surgical treatment and did not have a family or personal history of pancreatic cancer.