## Supplementary Methods

### *Selection of cases*

This study was approved by the Institutional Review Board of The Johns Hopkins Hospital. All patients who underwent pancreatic resection at The Johns Hopkins Hospital with a diagnosis of co-occurring IPMN with carcinoma involving the pancreas in the final pathology report were retrospectively identified over a ten year period (2006-2015). We excluded cases in which either neoplastic component was smaller than 0.5cm to ensure that each neoplasm could provide sufficient DNA for molecular analysis. Three different categories of carcinoma co-occurring with IPMN were included: ductal adenocarcinoma of the pancreas, colloid carcinoma of the pancreas, and carcinoma of the ampullary region (ampullary, distal bile duct, and duodenal carcinomas). Five ductal adenocarcinoma cases which met the above criteria were enrolled in a separate study for whole exome sequencing because they were morphologically very suggestive of carcinoma arising from an IPMN. Because these cases underwent comprehensive genetic analysis for another study, they were included in the molecular analyses reported here to validate our sequencing strategy. In addition, the sequencing results from these patients are included in the Discussion to ensure an unbiased cohort. Hematoxylin-and-eosin (H&E) stained slides were prepared from formalin-fixed paraffin-embedded (FFPE) tissue blocks from each case. All included cases were reviewed by an expert pancreatic pathologist to confirm the diagnosis and determine regions for microdissection. Specimens from 87 patients underwent laser capture microdissection and DNA extraction, of which DNA of sufficient quantity and quality (at least 5ng of amplifiable DNA measured by qPCR; see "*DNA extraction*") was obtained from 76 cancers and co-occurring IPMNs. From each patient, we selected two blocks for analysis, if available: one with carcinoma and adjacent IPMN (adj-IPMN) and one with IPMN without adjacent carcinoma (dist-IPMN). In all but one case, the two IPMN samples were obtained from the same grossly identified cyst. The cohort included 56 patients IPMN co-occurring with ductal adenocarcinoma (95 IPMN samples, 56 cancer samples), 13 patients with IPMN co-occurring with colloid carcinoma (20 IPMN samples, 13 cancer samples), and 7 patients with IPMN co-occurring with carcinoma of the ampullary region (11 IPMN samples, 7 cancer samples). Normal tissue slides of each case were also reviewed by a pathologist prior to DNA extraction.

### *Laser capture microdissection*

Ten to twenty 10µm serial tissue sections from FFPE blocks were cut onto membrane slides (Zeiss MembranSlide 1.0 PEN). Sections were de-paraffinized in xylene, rehydrated by descending ethanol solutions and subsequently stained by crystal-violet (Sigma Aldrich; diluted 1:5 in 70% EtOH). Regions of IPMN and carcinoma were identified under the microscope and individually micro-dissected for enrichment of neoplastic cellularity on a Leica LMD7000 instrument. For matched non-neoplastic (normal) samples, five 10µm sections were cut onto regular slides, and tissue was scraped off the slides using a fresh razor blade (Personna) for each sample. After microdissection/scraping, sample tissues were collected in 0.5mL lo-bind tubes (Eppendorf) and immediately processed for subsequent DNA extraction.

### *DNA extraction and quantification*

DNA was extracted from each sample using a combination of QIAamp DNA FFPE Tissue Kit (Qiagen,) and the MagMAX FFPE isolation kit (Applied Biosystems). In brief, 180µL ATL buffer (Qiagen) was applied into sample collection tube together with 20µl of ProteinaseK (Qiagen) and subsequently incubated on an agitating thermomixer for 24h at 55°C and 850rpm. Subsequently, 2µl of MagMAX Protease (Applied Biosystems) together with 15µL MagMAX DNA Digestion Additive (Applied Biodsystems) were added and incubated using the same agitating thermomixer for 2h at 60°C and 350rpm. The enzymatic digestion reaction was quenched at 80°C stationary for 30min. The sample was stabilized using 200ul AL buffer (Qiagen) and 1µg of carrier RNA (Qiagen), followed by incubation at 70°C stationary for 20min. After EtOH (100%) precipitation, sample mix was applied onto QIAamp spin columns (Qiagen). Genomic DNA was EtOH (100%) precipitated and eluted following manufacturer's instructions for QIAamp DNA FFPE tissue kit (Qiagen). Total DNA concentration was measured with the Qubit dsDNA HS assay kit (Invitrogen) according to the manufacturer's instructions on a Qubit 2.0 Fluorometer (Invitrogen) to confirm minimum DNA quantity of 50ng per sample. The quantity of amplifiable DNA was determined using Quantifiler Human DNA Quantification kit for quantitative PCR (Applied Biosystem) according to the manufacturer's instructions with each sample analyzed in duplicate. DNA was diluted to obtain final quantity of 10ng of amplifiable DNA in a volume of 6uL (=1.7ng/uL) for Ion Torrent semiconductor sequencing (see *Library preparation and targeted sequencing*). DNA was stored at -20°C until sequenced.

### *Library preparation and targeted sequencing*

A targeted sequencing approach analyzed 11 PDAC/IPMN related driver genes (*KRAS, GNAS, TP53, SMAD4, CDKN2A, RNF43, TGFBR2, ARID1A, BRAF, MAP2K4,* and *PIK3CA*) on collected DNA samples by Ion Torrent semiconductor sequencing. The custom targeted sequencing panel, described in detail previously, was designed using the Ion AmpliSeq Designer (Pipeline version 4.2; Life Technologies) and consisted of 142 amplicons in 2 primer pools. The panel covers the oncogenic hotspots in *KRAS, GNAS, BRAF,* and *PIK3CA,* along with the entire coding regions of the remaining genes. Library preparation was performed with 5ng of genomic DNA for each primer pool using Ion Ampliseq Library Kit 2.0 (Life Technologies) according to the manufacturer's instructions. The barcoded libraries were loaded onto 318v2 chips and sequenced using an Ion Torrent Personal Genome Machine (Life Technologies) following the manufacturer's instructions.

### *Targeted Sequencing Analysis*

Sequencing data analysis was performed using NextGENe software (v2.4; SoftGenetics). FASTA files were aligned to hg19 human reference genome and genes annotated using the Consensus Coding DNA Sequences (CCDS), RefSeq, and Ensembl databases using following transcript variants: *KRAS*: NM_033360; *GNAS*: NM 001077488; *TP53:* NM_000546; *SMAD4:* NM_005359; CDKN2A: NM_000077 *RNF43:* NM_017763; *TGFBR2:* NM_001024847; *ARID1A:* NM_006015; *BRAF*: NM_004333; MAP2K4:NM_003010; PIK3CA_006218. Tumor samples were compared to normal samples to identify somatic and remove germline alterations. Candidate mutations were identified using the following inclusion criteria: (1) variant allele frequencies of at least 5% of the total reads; (2) ≥100X coverage depth at the variant locus; (3) balance ratio (the ratio of the number of forward reads with the variant to the number of reverse reads with the variant, or vice versa) of the candidate mutation calls was 0.2 or greater; (4) PHRED quality score >25; (5) exclusion of synonymous and non-coding variants. Additionally, each putative somatic mutation was visually inspected using Integrative Genomics Viewer v2.3 (IGV; Broad Institute) or the NextGENe Viewer (v2.4; SoftGenetics) to remove remaining artifactual changes. Variant calls that occurred in truncated (shortened) amplicons were excluded, as these frequently represent false positive calls due mispriming. In addition, mutations at homopolymer or end-read regions (within 20bp adjacent to primer binding sites) were cautiously regarded and more stringent criteria were applied: (1) variant allele frequencies of at least 20% of the total reads; (2) ≥500X coverage depth at the variant locus.

Copy number variation (CNV) of targeted loci at chromosomes 9, 17, and 18 (which include the loci for *CDKN2A*, *TP53*, and *SMAD4*) was assessed using the copy number variation detection algorithm of NextGENe software, as has been thoroughly described previously. Based on specified criteria, we were able to analyze copy number alterations at targeted positions for 62 cancers and 101 IPMNs from the targeted sequencing data.

### *Whole-exome sequencing and analysis*

Whole exome library preparation, capture-based next-generation sequencing and subsequent bioinformatic analyses of neoplastic and normal samples were performed at Personal Genome Diagnostics (Baltimore, MD, USA). In brief, genomic DNA from neoplastic and normal samples was fragmented, followed by end-repair, A-tailing, adapter ligation, and polymerase chain reaction (PCR). PCR product was purified and exonic regions were captured in solution using the Agilent Sure Select kit according to the manufacturer’s instructions (Agilent). Paired-end sequencing, resulting in 100 bases from each end of the fragments, was performed using the HiSeq2500 next-generation sequencing instrument (Illumina). Primary processing of sequence data for both neoplastic and normal samples was performed using Illumina CASAVA software (v1.8). We used 260–1000 ng of input DNA, and a total of 50 353 573 bases of 20 965 coding genes were targeted. The average coverage in the targeted region ranged from 83 to 233, and approximately 94% of targeted bases were represented by at least 10 reads in each sample.

Candidate somatic mutations, consisting of point mutations, small insertions, and small deletions were identified using VariantDx across the regions of interest, as defined by the hg19 human genome build. The specific transcript accession numbers for determining the consequences of somatic mutations are listed in supplementary table S8.

### *Categorical assessment of relatedness*

Based on shared somatic mutations, we assessed the relatedness of the cancer, adj-IPMN, and dist-IPMN samples in each patient. To do this, we categorized somatic single base substitutions as “hotspots” or “non-hotspots”, with hotspots describing specific amino acid positions at which somatic mutations occur at high prevalence. In PDACs, codons 12, 13, and 61 in *KRAS* are commonly occurring hotspots, while codon 201 in *GNAS* is a hotspot limited to IPMNs. For IPMN/PDAC and IPMN/IPMN sample pairs, we considered codons 12, 13, and 61 in *KRAS* as hotspots, while we considered codon 201 in *GNAS* as a hotspot only for IPMN/IPMN sample pairs.

Two neoplasms were classified as “likely related” if they shared at least two somatic mutations (including oncogenic hotspot mutations in *KRAS)* or at least one non-hotspot mutation. In addition, a shared hotspot *GNAS* mutation classified an IPMN/cancer pair as “likely related”, as *GNAS* mutations are rare in non-IPMN-associated pancreatic ductal adenocarcinomas. Two neoplasms were classified as “likely independent” if they shared no somatic mutations. Neoplasms were designated as “indeterminate” if they shared only a single hotspot mutation in *KRAS* (in codon 12, 13, or 61). In addition, paired IPMN samples were designated as “indeterminate” if they shared only a *GNAS* hotspot mutation (in codon 201).

### *Statistical assessment of relatedness*

We computationally generated 10 million synthetic PDAC-IPMN sample pairs, and we compared the mutational overlap of synthetic sample pairs with that of real sample pairs from this study. Here, a synthetic sample is a collection of site-specific mutations, generated by randomly selecting mutations from a distribution of site-specific mutations that represents the sample type (PDAC or IPMN) being modeled. For instance, a synthetic PDAC was generated by selecting a collection of mutations from the true site-specific mutation distribution of PDACs in this study; the number of mutations in each synthetic sample is a random number drawn from a Gaussian distribution of mutation counts derived from the true PDACs in this study. IPMNs were derived using the same procedure, except that mutation frequency and count distributions were derived from the true IPMNs. After 10 million synthetic PDAC-IPMN sample pairs were generated, each occurrence of identical mutational overlaps was counted and a probability calculated for each specific overlap. For instance, there were 2 synthetic sample pairs that shared a *KRAS* Q61H and *ARIDA* I941M mutation; therefore, the probability that this apparent molecular relatedness would occur by random chance is 0.0000002. Because the true PDAC and adj-IPMN for patient IPP32 each had a *KRAS* Q61H and *ARID1A* I941M mutation (online supplemental Table S3), we estimate that the probability that the apparent relatedness of these lesions occurred by random chance is 0.0000002, suggesting a high probability that these lesions are indeed related. Importantly, approximately 9 million synthetic PDAC-IPMN pairs shared no mutations; therefore, the probability of having no mutational overlap is ~0.9 for the synthetic samples without a matched sample. We applied the above-described approach to PDAC/IPMN pairs and IPMN/IPMN pairs, and used the resulting probabilities as quantitative estimates of relatedness for all intra-patient lesions (online supplementary table S5). We also plotted the estimated probabilities for all inter- and intra-patient lesions as three separate heat maps (online supplementary figure S5). Because lesions from different patients cannot have originated from the same precursor lesion (i.e., they cannot be related), these inter-patient relatedness probabilities help establish relative confidence limits for inferring relatedness from intra-patient lesions.

### *Immunohistochemistry*

Immunohistochemistry was performed using monoclonal antibodies against p53 (mouse anti-human, 2.5ug/ml; Clone Bp-53-11, Ventana Medical Systems) and Smad4 (mouse anti-human, 1:500; clone B8, Santa Cruz Biotechnology). The Ventana BenchMark ULTRA platform was used to process the p53 immunolabelling, and the Leica BOND platform was used for Smad4 immunolabelling according to the manufacturer's protocol. p53 immunolabelling was interpreted as previously described. Briefly, lesions with diffuse nuclear staining (≥ 60% neoplastic cells) were designated as “aberrant – increased”, while those with complete absence of expression were designated as “aberrant – lost”. Smad4 immunolabelling was interpreted as previously described. Smad4 expression was designated as “lost” when all the neoplastic cells showed no staining in cytoplasm and nuclei, and as “retained” when neoplastic cells showed immunolabelling in cytoplasm or in nuclei. Non-neoplastic cells such as islets of Langerhans or stromal cells show sporadic nuclear expression and were utilized as an internal positive control for both p53 and Smad4 expression.