doi:10.1038/nature11547

Supplementary methods:

Approvals:

Human research ethical approvals
Australian Pancreatic Cancer Genome Initiative: Sydney South West Area Health Service Human Research Ethics Committee, western zone (protocol number 2006/54); Sydney Local Health District Human Research Ethics Committee (X11-0220); Northern Sydney Central Coast Health Harbour Human Research Ethics Committee (0612-251M); Royal Adelaide Hospital Human Research Ethics Committee (091107a); Metro South Human Research Ethics Committee (09/QPAH/220); South Metropolitan Area Health Service Human Research Ethics Committee (09/324); Southern Adelaide Health Service/Flinders University Human Research Ethics Committee (167/10); Sydney West Area Health Service Human Research Ethics Committee (Westmead campus) (HREC2002/3/4.19); The University of Queensland Medical Research Ethics Committee (2009000745); Greenslopes Private Hospital Ethics Committee (09/34); North Shore Private Hospital Ethics Committee. Johns Hopkins Medical Institutions: Johns Hopkins Medicine Institutional Review Board (NA00026689). Ontario Institute for Cancer Research: University Health Network Research Ethics Board (08-0767-T). Mayo Clinic: Institutional Review Board (PR66-06-05 and PR354-06-08). ARC-NET, University of Verona: approval number 1885 from the Integrated University Hospital Trust (AOUI) Ethics Committee (Comitato Etico Azienda Ospedaliera Universitaria Integrata) approved in their meeting of 17 November 2010 and protocolled [Author: Please reword to avoid 'protocolled'.] by the ethics committee 52070/CE on 22 November 2010 and formalized by the Health Director of the AOUI on the order of the General Manager with protocol 52438 on 23 November 2010. Baylor College of Medicine: Institutional Review Board for Baylor College of Medicine and affiliated hospitals (H-16215).

Animal experiment approvals
Mouse experiments were carried out in compliance with Australian laws on animal welfare. Mouse protocols were approved by the Garvan Institute/St Vincent's Hospital Animal Ethics Committee (09/53 protocol). Wild-type C57BL/6 mice (acinar to ductal metaplasia models) and *Pdx1-Cre; LSL-Kras*^{G12D}*; LSL-Trp53*^{R172H} (PDAC model) mice were all of both sexes and were housed with a 12 h light, 12 h dark cycle, receiving food *ad libitum*.

Methods

Australian Pancreatic Cancer Genome Initiative

Sample acquisition
Samples used were prospectively acquired and restricted to primary operable, non-pretreated pancreatic ductal adenocarcinoma. After ethical approval was granted, individual patients were recruited preoperatively and consented using an ICGC approved process. Immediately following surgical extirpation, a specialist pathologist analyzed specimens macroscopically and samples of the tumor, normal pancreas and duodenal mucosa were snap frozen in liquid nitrogen (for full protocol see APGI website: http://www.pancreaticcancer.net.au/). The remaining resected specimen underwent routine histopathologic processing and examination. Once the diagnosis of pancreatic ductal adenocarcinoma was made, representative sections were reviewed independently by at least 1 other pathologist with specific expertise in pancreatic diseases (AG, DM, RHH, AC), and only those where there was no doubt as to the histopathological diagnosis were

entered into the study. Co-existent Intraductal Papillary Mucinous Neoplasms in the residual specimen were not excluded provided the bulk of the tumor was invasive carcinoma, and the invasive carcinoma samples were used for sequencing. All samples were stored at -80 degrees celcius. Duodenal mucosa or circulating lymphocytes were used for generation of germline DNA. A representative sample of duodenal mucosa was excised and processed in formalin to confirm non-neoplastic histology prior to processing. All participant information and biospecimens were logged and tracked using a purpose built Data and Biospecimen information management system (Cansto Pancreas). Median survival was estimated using the Kaplan-Meier method and the difference was tested using the log-rank Test. *P* values of less than 0.05 were considered statistically significant. Statistical analysis was performed using StatView 5.0 Software (Abacus Systems, Berkeley, CA, USA). Disease-specific survival was used as the primary endpoint.

Sample extraction

Samples were retrieved, and either had full face sectioning performed in OCT or the ends excised and processed in formalin to verify the presence of carcinoma in the sample to be sequenced and to estimate the percentage of malignant epithelial nuclei in the sample relative to stromal nuclei. Macrodissection was performed if required to excise areas of non-malignant tissue. Nucleic acids were then extracted using the Qiagen Allprep® Kit in accordance with the manufacturers instructions with purification of DNA and RNA from the same sample. DNA was quantified using Qubit HS DNA Assay (Invitrogen). Throughout the process, all samples were tracked using unique identifiers.

Exome Library Construction

All libraries for Exome capture were prepared through the standard library generation process of fragmentation, end-repair, and adapter ligation followed by PCR amplification and enrichment for library molecules containing the correct adapter configuration. Two sequencing libraries were generated per exome capture: firstly, 3ug of genomic DNA was used as input and a library was constructed following the SureSelect Whole Exome protocol (Agilent), employing commercially available kits (Applied Biosystems for Fragment Library construction). Secondly, 1ug of genomic DNA was used in construction of libraries utilizing the Beckman Spriworks Kit III system for SOLiD, which automated the process of end-repair, and adapter ligation. In order to obtain the necessary 500ng of library required for Exome capture each library was PCR amplified for a total of 8 to 12 cycles as directed by the respective protocol from Agilent or Beckman to yield approximately 1ug of sequencing library. All libraries were quantified using the Agilent BioAnalyser High Sensitivity Assay.

Exon Capture

Exon capture was completed using the Agilent SureSelect All Exon 50Mb kit (Agilent Technologies; G3370B). Briefly 500ng of library was hybridized for 72 hours at 65 degrees with the Biotintylated RNA Bait Pool, Cot-1 DNA, Sonicated Salmon Sperm and blocking oligos complimentary to the adapter sequences of the library. Following hybridization the RNA-bait annealed libraries were bound to MyOne Streptavidin C1 Dynabeads (Invitrogen; 65001) and washed in pre-heated wash buffer to remove unbound library molecules. The washed and bound library fragments were eluted from the magnetic beads by incubating in 50uL of Elution Buffer for 10 minutes. The eluted Library molecules were neutralized by adding 50uL of Neutralization Buffer and incubating for 10 minutes. The enriched library was purified using the standard AMPure XP protocol (Beckman Coulter; A63882) and subjected to an additional 8 to 12 cycles of PCR amplification. Finally exome enriched libraries were quantified and qualified using the BioAnalyser High Sensitivity Assay (Agilent Technologies; 5067-4626).

Sequencing

Enriched libraries were pooled in equi-molar ratios into sets of 4 libraries which were then clonally amplified on 1µm beads using emulsion PCR with input library concentrations of 0.75 to 1.0 pM. Emulsion PCR reactions were generated, amplified and enriched for template positive beads using the EZ-Bead system (Life Technologies; 4453095) and associated commercially available kits. Subsequently enriched beads were processed through the 3' modification protocol to allow binding of the template positive beads to the activated SOLiD XD Sequencing slides (Life technologies; 4456997). End modified, template positive beads were then deposited onto SOLiD XD sequencing slides, targeting approximately 700 million beads per slide. Each slide containing the template positive beads was sequenced using the SOLiD v4 ToP Paired End Sequencing Kit (Life Technologies; 4459181). Initially, a 5bp barcode read was generated to allow for de-convolution of the individual exome capture libraries and subsequently paired-end reads of 35bp and 50bp in the reverse and forward directions, respectively, were generated.

Primary Sequence Analysis

Colour calls and quality scores were generated on the SOLiD sequencers and copied off instrument for processing using Applied Biosystems' BioScope v1.2.1 mapping and analysis software suite. BioScope maps paired reads independently using the Applied Biosystems mapreads colourspace aligner, then pairs the reads including a round of pair rescue where some reads that did not map uniquely can be assigned an unambiguous location based on the unique mapping of the second read in the pair and the nominal insert size of the library. Reads that mapped multiple times were discarded unless a single alignment had a mapping quality that was significantly higher (5 MAPQ units) than all other candidate mappings ("clear zone" method), or they were rescued by their paired read (as described above). All BAMs were processed to identify duplicates using the Picard MarkDuplicates tool from the Broad Institute. In cases where multiple sequencing runs were done on the same library, the BAMs were merged prior to duplicate marking. If multiple libraries were required to achieve desired coverage levels, library-level BAMs were merged to create consolidated tumour and normal exome BAMs for each donor. BAM files and associated metadata in XML format have been uploaded to the European Genome-phenome Archive (EGA; http://www.ebi.ac.uk/ega) under the accession EGAS00001000154.

Reference Sequence

The reference genome used by QCMG is based on the Genome Reference Consortium (http://www.ncbi.nlm.nih.gov/projects/genome/assembly/grc/human/) GRCh37 assembly. There are 3 categories of contigs in the GRCh37 assembly - placed, unlocalised and unplaced – and all GRCh37 contigs are present in the QCMG human reference with the addition of human mitochondrial sequence (NC_012920.1). There are 190 sequences in the placed contig class (GL000001..GL000190) and these are the contigs that make up the 22 somatic and 2 sex chromosomes. These 190 contigs are not included individually in the QCMG reference, rather we use the 24 chromosomes (1-22, X, Y) that are created by conflating the 190 placed contigs. There are 20 sequences in the unplaced contig class (GL000191..GL000210) and these are sequences where the chromosome that the contig belongs to is known but the exact location on that chromosome is unknown. There are 39 sequences in the unplaced contig class (GL000211..GL000249) and these are contigs where the location is unknown even at a chromosomal level.

Somatic Mutation Calling and Annotation

The QCMG somatic mutation calling pipeline calls single nucleotide variants (SNVs) and small indels. All mutation calls except those that did not verify by an orthogonal technology (see section Mutation Verification below) were submitted to the ICGC DCC.

1. Single nucleotide variants. We have developed an automated pipeline for somatic mutation calling (qSNP, manuscript submitted) that was tuned specifically to have a high sensitivity (also see section on Cellularity Estimation and Sensitivity below) in light of germline'contamination' of tumour samples. qSNP takes BAM files from a tumour sample and its matched normal and outputs an annotated list of somatic mutations and germline variants. It builds on the Picard library and samtools v1.1.17 mpileup function to generate pileups at each position of a tumour and matched normal BAM file. Reads for inclusion in the pileup results were filtered on the following criteria: alignment length > 34 or second in a pair and mapped as a proper pair; single mapping quality (SM) > 14; less than 3 mismatches to the reference genome; not a PCR duplicate. Only reads that fulfilled all of the above criteria were included in mutation calling. The samtools mpileup option –A was used to include also non-AAA pairs in the pileup, so long as they also fulfilled the above criteria.

We used a frequency-based approach and the pileups at each position to call putative variant positions. Based on extensive verification, we found that a minimum of 5 mutant reads was a useful lower threshold for somatic mutation calling. For sequence coverage above 50 reads a minimum mutant allele ratio of 0.05 was required. By comparing variant calls in the tumour and matched normal sample, positions were classified as either somatic mutations or germline variants. For somatic mutation calls several additional checks were performed. We excluded positions that had any evidence for the mutation in the normal applying the same read filter as mentioned above. Positions with base changes that were germline variants in another patient were also excluded. Positions that had less than 12 reads coverage in normal were flagged. We found that the verification rate of these positions was much lower, but nevertheless at a rate acceptable to include in verification.

Once somatic mutations were called, their effects on any alternative transcripts were annotated using a local install of the Ensembl database (v61) and the Ensembl Perl API, following the specification of the ICGC data coordination centre. Positions were further annotated with dbSNP, whether a position was inside a Pfam or InterPro protein domain, and whether it had been previously reported in the COSMIC database (v55).

2. *Small indels*. We used the Bioscope (Life Technologies) Small Indel Tool to identify putative indel positions in tumour and normal. Briefly, one read of a read pair is used as an anchor to locally realign its partner. An in-house pipeline was used to compare indel calls in tumour and normal and to classify variants into somatic mutations and germline variants. Similar to the point mutation pipeline several post-processing checks were performed. We excluded somatic indels that had any evidence for the indel in the normal, indels that were germline variants in another patient and flagged those indels that had less than 12 reads coverage in the normal. The Ensembl Perl API was again used to annotate the consequences of indels on any alternative transcripts that may overlap the mutation position.

All mutation calls except those that did not verify by an orthogonal technology (see section Mutation Verification below) were submitted to the DCC.

Baylor College of Medicine

Sample Collection
Blood was directly collected in PAXgene Blood DNA tubes and DNA was isolated using the PAXgene Blood DNA kit (PreAnalytiX, Qiagen, Valencia, CA). The tumor and normal pancreas tissue specimen were collected shortly after resection and stored in a protease inhibitor solution (Roche Applied Science, Indianapolis, IN), RNAlater (Qiagen), or snap frozen, and stored at -80oC. The blood sample specimen was used as matching normal control unless unavailable in which case adjacent pancreatic tissue was used as normal sample. DNA was isolated from 80 mg tissue fragments using the GentraPuregene kit (Qiagen). The quality of the DNA samples were ascertained by electrophoresis and determined to be of high quality (size >23 kb) with no visible degradation in blood or tumor samples.

Capture Library Construction
SOLiD 4: SOLiD precature libraries were constructed using 5 ug of genomic DNA according to a modified version of the manufacturer's protocol (Applied Biosystems, Inc.). Briefly, the genomic DNA was sheared into fragments of approximately 120 base pairs with the Covaris S2 or E210 system as per manufacturer instruction(Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair (NEBNext End-Repair Module; Cat. No. E6050L) and A-tailing (NEBNext dA-Tailing Module; Cat. No. E6053L). The resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit (Cat. No. M2200L). Solid Phase Reversible Immobilization (SPRI) bead cleanup (Beckman Coulter Genomics, Inc.; Cat. No. A29152) was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR (LM-PCR) was performed using Platinum PCR Supermix HIFi (Invitrogen; Cat. No.12532-016) and 6 cycles of amplification. After bead purification, PCR products were quantified and their size distribution was analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).Primer sequences and a complete library construction protocol are available on the Baylor Human Genome Website:
(http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

HiSeq 2000: After determining DNA concentration and integrity, high molecular weight double strand genomic DNA samples are constructed into Illumina PairEnd precapture libraries according to the manufacturer's protocol (Illumina Inc.) with modification. Briefly, 1ug genomic DNA in 100ul volume was sheared into fragments of approximately 300 base pairs in a Covaris plate with E210 system (Covaris, Inc. Woburn, MA). The setting was 10% Duty cycle, Intensity of 4, 200 Cycles per Burst, for 120 seconds. Fragment size was checked using a 2.2 % Flash Gel DNA Cassette (Lonza, Cat. No.57023). The fragmented DNA was end-repaired in 90ul total reaction volume containing sheared DNA, 9ul 10X buffer, 5ul END Repair Enzyme Mix and H2O (NEBNext End-Repair Module; Cat. No. E6050L) and then incubated at 20°C for 30 minutes. A-tailing was performed in a total reaction volume of 60ul containing end-repaired DNA, 6ul 10X buffer, 3ul Klenow Fragment (NEBNext dA-Tailing Module; Cat. No. E6053L) and H_2O followed by an incubation at 37°C for 30 minutes. Illumina multiplex adapter ligation (NEBNext Quick Ligation Module Cat. No. E6056L) was performed in a total reaction volume of 90ul containing 18ul 5X buffer, 5ul ligase, 0.5ul 100uM adaptor and H2O at room temperature for 30 minutes. After Ligation, PCR with Illumina PE 1.0 and modified barcode primers (manuscript in preparation) was performed in 170µl reactions containing 85 2x Phusion High-Fidelity PCR master mix, adaptor ligated DNA, 1.75ul of 50uM each primer and H2O. The standard thermocycling for PCR was 5' at 95°C for the initial denaturation followed by

6-10 cycles of 15 s at 95°C, 15 s at 60°C and 30 s at 72°C and a final extension for 5 min. at 72°C. Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) was used to purify DNA after each enzymatic reaction. After bead purification, PCR product quantification and size distribution was determined using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).

Exome Capture

SOLiD 4: The pre-capture libraries (2ug) were hybridized in solution to either NimbleGen CCDS (~36 Mb of sequence targets from ~17K genes) or NimbeGen EZ Exome 2.0 (~44 Mb of sequence targets from ~30K genes) Solution Probes according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture LM-PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over the background. The captured libraries were further processed for SOLiD sequencing. Primer sequences and a complete capture protocol are available on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf)

HiSeq 2000: Pre-capture libraries (1ug) were hybridized in solution to VCRome 2.1 exome design (HGSC design, NimbleGen) targeting 43 Mb of sequence from ~30K genes, according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos IHE1, IHE2 and IHE3 (manuscript in preparation) replaced oligos HE1.1 and HE2.1 and post-capture LM-PCR was performed using 14 cycles. Capture libraries were quantified using Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over background.

Sequencing Library and DNA Sequencing

SOLiD 4: The captured libraries were clonally amplified onto 1 um beads using emulsion PCR with a final library concentration of 0.70 to 0.85 pM. Emulsion PCR reactions were generated with 4X bulk reactions in a sealable bag using a Servodyne Electronic Mixer (Cole-Parmer, EW-50008-30, EW-50008-00) at a speed of 720 rpm for 20 min. The 4X bulk reaction was amplified using a Hydrocycler (K-Biosciences, HC-16) with the following cycling conditions, denature for 10 min at 95^{O}C, followed by 40 cycles of 1min at 95^{O}C, 2min at 62^{O}C and 2 min at 72^{O}C with a final extension of 10 min at 72^{O}C. Beads were recovered by centrifugation with 2-butanol and 50 ml conical centrifuge tubes and then enriched and 3' modified according to the Life Technologies Macro-Scale 4 ePCR reaction protocol. The 3' modified template positive beads were deposited on to XD sequencing slides, targeting approximately 300 K beads/panel and sequenced using SOLiD V4 Top reagents. Both barcode fragment and paired end sequencing methods were used in this project. For barcoded methods, capture libraries were individually captured and then pooled in sets of 4 samples after post-capture amplification. The 4 sample barcode library pools were sequenced with SOLiD Barcode Fragment Sequencing Kits (Life Technologies,

4452697). Here the first 5 bp barcode read is utilized to de-convolute the individual capture libraries followed by a 50 bp forward read. Individual capture libraries were sequenced with SOLiD Paired End Sequencing Kits (Life Technologies, 4459179) using a 35bp reverse read followed by a 50bp forward read.

HiSeq 2000: Sequencing was performed in paired-end mode with Illumina HiSeq 2000. Illumina sequencing libraries were amplified by "bridge-amplification" process using Illumina HiSeq pair read cluster generation kits (TruSeq PE Cluster Kit v2.5, Illumina) according to the manufacturer's recommended protocol. Briefly, these libraries were denatured with sodium hydroxide and diluted to 3-4 pM in hybridization buffer for loading onto a single lane of a flow cell in order to achieve 600-700k clusters/mm^2. Barcoded libraries were pooled in sets of 2 for sequencing per lane and all lanes were spiked with 1% phiX control library. Cluster formation, primer hybridization were performed on the flow cell with illumina's cBot cluster generation system.

Sequencing reactions were extended for 202 cycles of SBS using TruSeq SBS Kit on an Illumina's Hiseq 2000 sequencing machine according to the manufacturer's instructions. The Illumina Sequence Control Software (SCS) control the reagent delivery and collect raw images. Real Time Analysis (RTA) software was used to process the image analysis and base calling. On average, about 80-100 million successful reads, consisting of 2x100 bp, were generated on each lane of a flow cell.

Base Calling and Read Mapping

SOLiD 4: Base and quality calling for SOLiD data was performed on-instrument using standard vendor software and settings. Upon completion of a run, read and quality data was copied into our data-center where individual sequence events are split into 10M read bundles and mapped in parallel using BFAST (version 0.6.4). After read bundles are mapped their results are merged back into a single sequence-event-level BAM where read group tags are added. Where necessary, sample-level BAMs are generated by merging using Picard (version 1.7), and duplicate reads are marked at the library level using SAMtools (version 1.7). Variant calling is done using custom filters applied to pileups made at the sample level, also using SAMtools.

HiSeq 2000: Finally, the output of a Illumina HiSeq sequencer are binary bcl files that are processed using the software (BCLConvertor 1.7.1). All reads from the prepared libraries that passed the illumina Chastity filter were formatted into fastq files. The fastq files are aligned to the genome using BWA (bwa-0.5.9rcl) against human reference genome build36. Default parameters are used for alignment except for a 40 bp seed sequence, 2 mismatches in the seed, and a total of 3 mismatches allowed.

Variant Discovery

SOLiD 4: Mutations in BAM files generated from SOLiD reads were detected as follows: SamTools Pileup was run to list all variants found in multiple reads at a single locus. The variants were further filtered to remove all those observed fewer than 5 times or were present in less than 0.10 of the reads. At least one variant had to be Q30 or better, and the variant had to lie in the central portion of the read, 15% from the 5' end of the read and 20% from the 3' end. In addition reads harboring the variant must have been observed in both forward and reverse orientations. Finally, the variant base was not observed in the normal tissue. Insertion or deletion variants ("indels") were discovered by similar processing except indels must have been observed in 0.25 of the reads (see below for detection of frameshift indels at microsatellite sites).

HiSeq 2000: BAM files generated from alignment of Illumina sequencing reads were preprocessed using GATK.

Validation of Mutations: Mutations are validated by sequencing PCR amplified DNA from the mutated sample and its matched normal on a 454 instrument. The final mutation file consisted of 498 non-silent mutations and 110 silent mutations. “Non-silent” includes missense, nonsense, splice site or, in-frame and frameshift indels.

Ontario Institute for Cancer Research

Sample acquisition & processing

Samples used were prospectively acquired and restricted to primary operable, non-pretreated pancreatic ductal adenocarcinoma. Patients were recruited preoperatively and consented using a Research Ethics Board approved process consistent with ICGC requirements. A blood sample was obtained from consented individuals for germline DNA isolation. Immediately following surgical resection, specimens were examined by a specialist pathologist. Samples of the tumor were snap frozen in liquid nitrogen. The remaining resected specimen adjacent to the research specimen underwent routine histopathologic processing and examination to confirm the diagnosis of PDAC. Representative sections were reviewed independently by at least 1 other pathologist with specific expertise in pancreatic diseases (AB). All samples were stored at -80 degrees celcius. All participant information and biospecimens were logged and tracked using biospecimen information management systems specific to each collection site (Mayo, Rochester and UHN, Toronto).

Macrodissection was performed if required to excise areas of normal tissue. DNA extractions from tissues were performed using the GentraPuregene Cell kit (Qiagen; 158388). Throughout the process, all samples were tracked using unique identifiers.
Blood extractions were carried out using the GentraPuregene Blood kit (Qiagen; 158467). The blood sample specimen was used as a matching normal control.

The quality of all extracted DNA samples were ascertained by electrophoresis and determined to be of high quality (size >23 kb) with no visible degradation in blood or tumor samples. Appropriate A260:280 ratios were confirmed using a Nanodrop spectrophotometer. STR fingerprinting was also performed on all matched sets of extracted DNA to confirm sample relatedness.

Exome capture

One microgram of genomic DNA from each of the matched tissues (tumor and normal sample) were used for whole exome capture. Genomic DNA was sheared to 300-400 base pair fragments using the Covaris acoustic shearing system as per manufacturer instructions. DNA fragments were then processed with the NEBNext library kit including end-repair and A-tailing. Products were ligated to Illuminaplatform-specific adapters (ordered from IDT) using Quick Ligase (New England Biolabs; E6056B). A final size selection (300-350bp) was performed using agarose/PAGE gels.

Libraries were amplified using the Agilent SureSelect Indexing Pre-Capture PCR primers (barcoded libraries) or PE primer 1.0 & 2.0 (non-barcoded libraries) for 12-15 cycles using NEB Phusion High-Fidelity PCR Master Mix (NEB; M0531L).

DNA was then hybridized in solution using the Sureselect All Exon probes (Agilent Technologies; G3370E). Agilent SureSelect hybrid selection was performed following the manufacturer's protocols (Agilent; G2939AA – Barcoded & Agilent H2103A non-barcoded). In brief, the libraries were adjusted to 250-500ng in 3.4ul of water and then hybridized to the SureSelect probes for a 72 hour incubation period in the presence of SureSelect blocking oligos. The captured DNA fragments were bound to MyOne Streptavidin T1 magnetic beads (Invitrogen; 656-01) and following standard washing protocols to remove non-specifically bound DNA were amplified for 12 cycles of LM-PCR using Herculase II Fusion DNA polymerase (Agilent; 600677). The captured libraries were then quantified using the KAPA SYBR qPCR kit for Illumina libraries (DMark; KK4835) and sequenced on the Illumina GAIIX and HiSeq 2000 sequencing platforms.

Sequencing

Genome Analyzer and HiSeq paired-end flow cells were prepared on the Illumina cluster station and cBot and 2X101 paired end reads were generated on the Illumina GAIIX and HiSeq 2000 platforms following the manufacturer's protocols.

Primary data analysis

Intensities were copied off instrument and CASAVA 1.7.0 was used to generate base calls and sequence files in FASTQ format with Illumina base quality scores. The FASTQ sequences were aligned to the UCSC hg19 reference (including random sequences) using Novoalign v2.07.09. For multiple aligned reads, the top five alignments were retained. The output was generated in SAM format (v1.4) with properly configured read groups. Reads were then sorted and converted to BAM format using Picard 1.40. All reads from individual libraries were merged using Picard v1.40. The reads were filtered using SAMTools v0.1.16 to produce unique aligned reads. PCR duplicates were removed from the library based BAM files and were merged using Picard v1.40 to a single BAM file representing the sample and tissue type.

Variant Calling

Somatic and germline variants were identified using the Genome Analysis Toolkit (GATK v1.0.5083). This comprised of merging all tissue types for a given sample, base quality recalibration and local realignment prior to variant calling. Variants were then annotated using ANNOVAR and the Ensembl (r61) gene reference model.

QCMG / BCM / OICR Common Methods

Mutation Verification using IonTorrent

All coding somatic mutation calls other than silent mutations were selected for verification by targeted IonTorrent sequencing. PCR primers to amplify amplicons (70-150bp) that overlapped the somatic mutation or indel were designed. Tumour and normal DNA was whole-genome amplified prior to PCR using the Illustra GenomiPhi V2 DNA Amplification Kit (GE; 25-6600-30). PCR reactions were set up using a Bravo liquid handler with 10 ng of amplified gDNA and 5 uM of primers mix. A touch down PCR was performed and products were cleaned using Agencourt AMPure XP Beads (Beckman Coulter; A63882). Amplicons were pooled and quantified on an Agilent Bioanalyser High Sensitivity DNA chip (Agilent Technologies; 5067-4626).

Each of the amplicon pools, consisting of either tumor or normal-derived amplicons were processed in parallel, firstly by generating clonally amplified Ion Spheres suitable for deposition and sequencing on the Ion Torrent Personal Genome Machine (PGM). Ion Spheres were generated using the Ion Xpress Template Kit (Life Technologies; 4469001); with approximately 260 million amplicon molecules per emulsion PCR, effectively yielding an emulsion containing 1 amplicon molecule per Ion Sphere. Emulsions were then transferred to standard 96 well PCR plates and amplified using PCR conditions optimized for Ion Torrent emulsion PCR.

To break and enrich the emulsions a series of centrifugation and washing steps were employed as per the Ion Xpress Template protocol. Enrichment of the template positive Ion Spheres was completed by selectively binding the Ion Spheres containing amplified library fragments to streptavidin coated magnetic beads. A series of three wash steps were employed to remove empty Ion Spheres followed by a single 7 minute incubation with an alkali solution to denature the libraries strands allowing for collection of the template positive Ion Spheres, now containing single stranded libraries molecules.

Samples were sequenced using the Ion Sequencing Kit (Life technologies; 4468997) and the Ion Chip 316 Kit (Life Technologies; 4469496). Following enrichment of the Ion Spheres a single sequencing primer was annealed to the single stranded templates along with binding of the sequencing polymerase. The Ion Spheres, now containing the annealed sequencing primer and bound DNA polymerase where then deposited by centrifugation into Ion 316 Semiconductor Sequencing Chips. The 316 Sequencing chips were then in turn loaded onto the Ion Torrent Personal Genome Machine and subjected to sequential flows of single dNTPs with each successful incorporation of a nucleotide resulting in the release of a hydrogen ion causing a shift in pH and conductivity of the individual well containing the Ion Sphere which was subsequently recorded, collated and translated into a single sequencing read for each Ion Sphere.

Verification of somatic mutations was performed by sequence pileup at each mutant position and a position was considered verified if it had a minimum depth of 100 reads coverage in tumour and normal, a mutant allele frequency of at least 10% in tumour and less than 0.5% in normal. In total, we confirmed by independent IonTorrent amplicon sequencing 1502 mutations and indels as true somatic events.

Copy Number Analysis

Matched tumour and normal patient DNA was assayed with either the HumanOmni1-Quad BeadChip or Omni1M-Duo BeadChips as per manufacturer's instructions (Illumina, San Diego CA). SNP arrays were scanned and data was processed using the Genotyping

module (v1.8.4) in Genomestudio v2010.3 (Illumina, San Diego CA) to calculate B-allele frequencies (BAF) and logR values. GenoCN was used to call somatic regions of copy number change – gain, loss or copy neutral LOH[2]. Recurrent regions of copy number change were determined and genes within these regions were extracted using ENSEMBL v61 annotations. GISTIC2.0 was used to determine significant regions of gain and loss[3].

Gene expression arrays

Total RNA samples (150ng/sample) were amplified and labeled using the Illumina TotalPrep RNA Amplification Kit (LifeTechnologies) as per manufacturer's instructions. Amplified RNA was quantified using a NanoDrop (Thermo Scientific) and distribution of the amplified material was analysed in an 2100 Bioanalyzer (Agilent). Amplified cRNA (750ng) was hybridized onto Illumina human HT-12 arrays (V4), and arrays were scanned on a Bead Array Reader (Illumina). Expression measurements were extracted using the GenomeStudio (version 2009.1) software. Microarray data has been deposited at GEO (accession GSE36924), and in the ICGC Data Coordination Centre (DCC; http://dcc.icgc.org/).

Annotating Mutated Genes

To facilitate the calculation of statistically significantly mutated genes (SMG), a single representative transcript was selected to annotate each somatic mutation based on the significance of the predicted functional effect of the mutation on the transcript, ordered from most significant to least significant as follows: nonsense, frameshift, splice site, in-frame, missense, nonstop/readthrough, silent, and RNA. Splice site mutations were defined as substitutions, deletions, or insertions overlapping the first/last 2bp of an exon or the first/last 2bp of an intron. Mutations affecting 3'UTR, 5'UTR, intronic sequence, and intergenic sequences were discarded for the purposes of downstream analyses of significantly mutated genes. Mutations from the three centres were combined and summarized in MAF format (Supplementary Table 5). Given that the transcript with the most deleterious effect was chosen to represent a given mutation, the overall list of somatic mutations is likely enriched for non-silent compared to silent mutations, especially for cases where multiple transcripts with different reading frames overlap a single mutation site and a silent mutation may have occurred in an alternate reading frame.

Significantly Mutated Genes

The estimate of the statistical significance on mutation rates of recurrently mutated genes depends on the number of mutations observed for each gene and the base coverage across the population and the background mutation rate. The objective is to determine which genes are mutated at a significantly higher rate than would be predicted by the background mutation rate. The mutation rate for each gene was corrected for its length, base composition using the Mutational Significance in Cancer package (MuSiC: http://gmt.genome.wustl.edu/genome-music/current/). Briefly, the coverage status of each target base for each individual was represented by each sequencing center in "wiggle file format"[1]. The coverage across the cohort of patients and the number of mutations at each position were tallied, and a mutation rate for each nucleotide change was calculated (Supp. Table 7). The overall background mutation rate was 0.65 per million bases. This is likely to underestimate the true mutation rate for this cancer due to tumor purity.

MuSiC uses three different statistical tests to evaluate the significance of a given gene's mutation rate: a convolution test, Fisher's combined P-value test, and a liklihood ratio test. We accepted all genes as significantly mutated whose P-value, corrected for multiple tested yielded a False Discovery Rate of less than 0.1 in any of the three tests.

Cellularity Estimation

Tumour cellularity was determined by pathologist review and 2 molecular strategies tarting the analytes subjected to sequencing: deep amplicon sequencing of KRAS and SNP chip based cellularity estimation using a QCMG-developed R tool, qpure (PLoS One submitted).

Deep amplicon based sequencing of exons two and three of the KRAS gene was perfomed with 454 or Ion Torrent using methods described in the mutation verification section. Exons 2 and 3 of the KRAS gene which are known to harbor driver/founder mutations are a hotspot for somatic mutations and are frequently mutated in pancreatic cancer. Amplicons spanning the highly perturbed codons (9,12,13,59,61) of exons 2 and 3 were generated for the 142 clinical cohort of matched tumour and germline samples. Products were subsequently pooled and sequence generated to an average depth of 26608 fold (range 609 to 213544). Identification of somatic mutations was performed by sequence pileup at positions in codon 9,12,13,59,61 and a position was considered mutated if there was a mutant allele frequency in the tumour of at least 0.05 in tumour. Cellularity estimates could not be calculated for samples that were wildtype for KRAS or contained a copy number change of the KRAS gene; instead the qpure cellularity score was used.

qpure Cellularity Estimation: qpure is a software tool written in R that estimates tumor cellularity based on analysis of genotype microarray data from paired tumor and normal DNA samples. A key advantage of qpure over histological estimation of tumour cellularity is that the DNA sample used for sequencing is the same sample used to run the qpure SNP microarrays. In contrast, the DNA sample used in sequencing is unlikely to be derived from the tissue slide used for pathology review.

The principle underlying the method is that heterozygous SNPs in the normal sample should appear homozygous in the tumour when LOH occurs. Any deviation from homozygosity at these SNPs can be attributed to the presence of stromal cells in the tumour. The stromal contamination can be quantified by plotting the degree of deviation from homozygosity (change in B allele frequency) against a standard curve derived from a mixture experiment involving cell line and matched normal DNAs. A panel of 14 mixtures was prepared to cover the following tumour cellularities: 0, 5, 10, 15, 20, 30, 40, 50, 60, 65, 75, 80, 85 and 100%. These were hybridised to HumanOmni1-Quad BeadChip (Illumina, San Diego CA) to create a standard curve of deviations from homozygosity at different tumour cellularites.

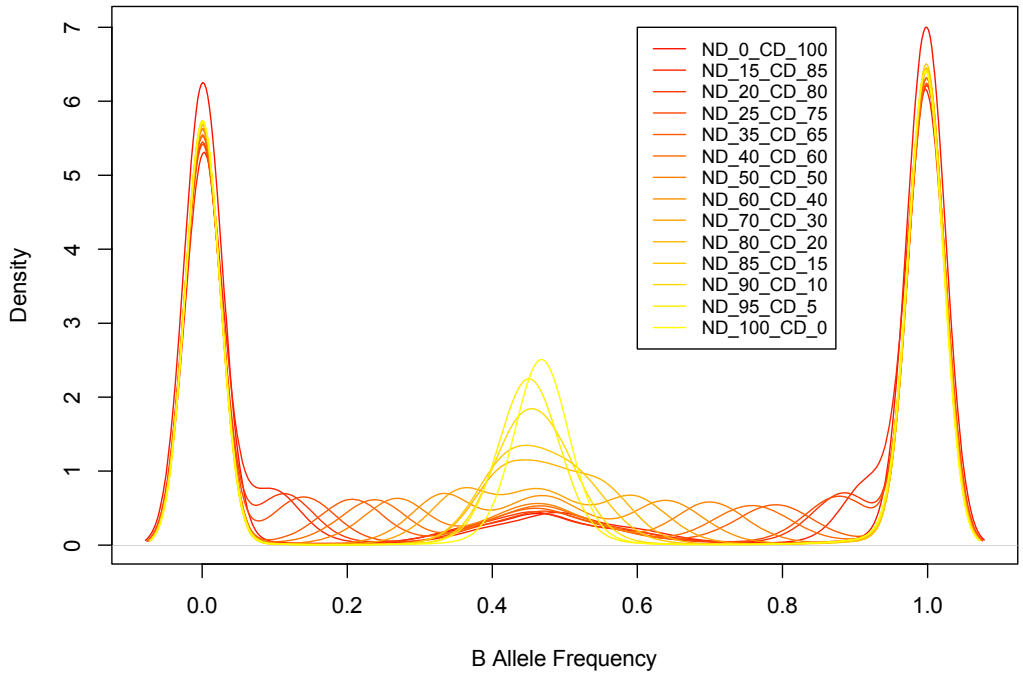

Figure 1. Allele distributions from genotype microarrays for the 14 mixtures at SNPs that are heterozygous in the normal sample and in regions of loss in the tumour. The x-axis shows the B allele frequency ranging from 0 to 1.0 and the y-axis shows kernel density, a measure of the relative proportion of SNPs at a given B allele frequency. The plot for the 100% normal sample (ND_100_CD_0) shows three major peaks at B allele frequencies of 0, 0.5 and 1.0 representing genotypes AA, AB and BB. The 100% tumour sample (ND_0_CD_100) shows only 2 peaks at B allele frequencies of 0 and 1.0 representing genotypes of AA and BB reflecting the loss of heterozygosity in the selected SNPs. The mixtures, where the genotype is a blend of tumour and normal genotypes, show minor peaks at B allele frequencies that indicate the degree of normal admixture.

The analysis process for qpure is:

1. Run SNP arrays on matched normal and tumour samples and identify all SNPs that are heterozygous in the normal sample (using vendor's software);
2. Select all SNPs showing single-copy somatic loss;
3. Plot the B allele frequency (BAF) distribution for these SNPs;
4. Determine the best possible SNP clusters from the BAF distribution from (3)
5. Measure the d-score which is defined as the absolute distance between the two most distant clusters from (4);
6. Determine the cellularity by comparing the d-score from (5) against the standard curve defined from the mixture experiment (described below).

The qpure tool is available for download at https://sourceforge.net/projects/qpure/.

Estimating the impact of tumour cellularity on somatic mutation detection
Patient samples were sequenced at three sites to a mean coverage depth of 65x (APGI), 104x (BCM) and 205x (OICR) to account for the differences in the cellularities of each centre's cohort (42% QCMG, 31% BCM and 28% OICR). In order to gauge the effect of cellularity on calling somatic mutations, we also performed exome sequencing of a series of cell line / matched normal DNA mixtures mimicing different tumour cellularites (0, 10, 20, 40, 60, 80, and 100% tumour DNA) at depths comparable to 65-70x (the lower limit of sequence coverage selected in the study). Variant calling was then run against these data and the decay in detection ability of true positives (independently validated mutations seen in the 100% sample) was recorded. A threshold of 20% cellularity was then applied to the entire cohort (conservatively estimated to have a sensitivity of at least 45%). An additional 7 samples were added with even lower sensitivity due to their much deeper sequencing and the presence of at least 10 independently validated somatic mutations. Finally, the relative range of observed mutations across the entire cellularity range was as follows:

0-19% cellularity (n=14): mean mutations per patient: 20, range 11-36.
20-39% cellularity (n=47): mean mutations per patient: 23, range 1-51.
40-59% cellularity (n=26): mean mutations per patient: 31, range 2-116.
60-83% cellularity (n=12): mean mutations per patient: 36, range 14-70.

INTEGRATING FUNCTIONAL DATA
To investigate the potential functional consequences of mutations and copy number variant genes in our cohort, we integrated our dataset with a large scale in vitro functional screen (data from Cheung et al., 2011)[4], and two recent in vivo sleeping beauty transposon mediated insertion mutagenesis screens(Mann *et al.* and Perez-Mancera *et al*.), that were recently published[5,6].

In the in vitro screen, there are three analytical approaches used to identify 'hits', where high confidence hits are identified by all three methods, and low confidence hits by any of the three methods (see below). Since there were 9 cancer cell line lineages, we defined 4 categories of hits from the in vitroscreen:

1) high/low confidence depleted genes in any cancer cell line lineage
2) high/low confidence enriched genes in any cancer cell line lineage
3) high/low confidence depleted genes in pancreatic cancer only
4) high/low confidence enriched genes in pancreatic cancer only

Depleted genes represented those genes that were "essential" for cell survival, and enriched genes were those there was presumably a survival advantage with knockdown. Since pancreatic cancer specific genes were only those that were 'hits' in pancreatic cancer alone, we compared to genes that were perceived to be functionally relevant across all cancer types.

IN VITRO CELL LINE FUNCTIONAL SCREEN ANALYSIS
Cheung *et al.*[4] screened 102 cancer cell lines from 20 cancer lineages, including Ovarian, Colon, and Pancreas, using a pooled hairpin library. The hairpin library consisted of a pool of 54K hairpins targeting 11,194 genes by ~4 hairpins/gene. Each cell line was infected in quadruplicate, and propagated for at least 16 doublings. Hairpin abundance was assayed

by a custom microarray and those hairpins that are depleted or enriched are predicted to target oncogenes, and tumour suppressors, respectively.

Since Cheung *et al.*focused on only depleted hairpins within the ovarian lineage, we reanalysed this data to identify depleted and enriched genes within the pancreas lineage, as well as all other lineages with at least 3 cell lines.

We obtained the PMAD normalised data (dated 3/3/2011), sample annotation, and GenePattern modules from the Achilles phase 2 section at the Integrated Genome Portal (IGP, www.broadinstitute.org/IGP).

To identify lineage-specific candidate oncogenes and tumour suppressors, we applied the same approach as Cheung *et al.*:

1. Use the *MakeSubsetGctAndCls* GenePattern module to create 13 pairs of GCT/CLS files for each of the 13 lineages with >= 3 cell lines.
2. For each of the 13 lineages, calculate a weight of evidence (*WoE*) statistic for each hairpin, using the *ScorebyClassComp* GenePattern module. The *WoE* is a non-parametric 2-group statistic, discussed in Cheung *et al.*, which assigns a score to each hairpin based on the consistency of change in one lineage vs all other lineages. Extreme positive and negative scores represent those hairpins that were consistently enriched, or depleted across all cell lines from the lineage being tested, with respect to all other cell lines.
3. Identify gene-level hits by 3 methods, looking for *depletion*, that is, hairpins that went down over time:
    a. rank genes based on the best-hairpin (ie the most negative), then select top 150
    b. rank genes based on the second-best hairpin, then select top 300
    c. rank genes based on a GSEA-like Kolmogorov-Smirnov statistic on all hairpins for each gene, then select top 300
    - in all cases, obtaining p-values using 1000 permutations
    - for (a) and (b) we implemented this in R, following the algorithm set forth in Cheung *et al.*. Code available on request.
    - for (c) we used the RIGER program (Luo et al, 2008,[7]http://www.broadinstitute.org/cancer/software/GENE-E).
4. Identify gene-level *enriched* hits as per step 3, looking for hairpins that went up over time.
5. For each lineage, and each direction, plot a 3-way Venn diagram comparison for the hits from the 3 approaches.
    - high or low confidence hits were identified by all 3, or any of the three approaches, respectively
6. The union of the high or low confidence hits across the 13 lineages were combined to obtain the lists of high or low confidence hits in all cancer lineages, respectively.

A heatmap of the high confidence candidate oncogenes (Supp Figure 5), and candidate suppressors (Supp Figure 6) identified from each lineage with ≥ 3 cell lines (~25 genes per lineage and direction). The best-scoring hairpin was selected to represent each gene. The heatmap uses a relative colour scheme, were hairpin data were row-standardized.

IN VIVO SLEEPING BEAUTY MUTAGENESIS SCREENS

Two independent sleeping beauty transposon mutagenesis screens in pancreatic Kras transgenic mouse models were compared to human PC data:

SCREEN 1: Karen Mann *et al.*[5]

Mice

Mice carrying the *LSL-Kras*^{G12D}allele were crossed to mice carrying a pancreatic-specific *Pdx1-Cre* driver to remove the loxP-STOP-loxP (LSL) cassette and activate expression of oncogenic *Kras*^{G12D} in the pancreas[8]. These mice were then crossed to a compound transgenic line containing up to 350 copies of a mutagenic *SB* transposon from a single donor site in the genome, and an inducible *SB*floxed-stop transposase allele knocked into the ubiquitously expressed *Rosa26* locus. We used two transposon transgenic lines, T2Onc2 and T2Onc3, located on different chromosomes, to obtain insertion data from the entire genome (reviewed by Copeland in [9]). Animals were aged for tumor formation. Tumor-free survival was significantly decreased when oncogenic *Kras*^{G12D} was combined with *SB*.

Gaussian Kernel Convolution Method for CIS Determination

Isolation of the *SB* transposon insertion sites from tumor DNAs was performed using splinkerette PCR to produce barcoded products that were pooled and sequenced on the 454 GS-Titanium sequencers (Roche) platform. Reads from sequenced tumors were mapped to the mouse genome assembly NCBI m37 and merged together to identify non-redundant transposon insertion sites. The Gaussian kernel convolution (GKC) statistical framework was used to identify CISs (Common Insertion Sites), regions in the genome that contain more transposon insertions than expected by chance. The GKC method employs multiple kernel scales (widths of 30K, 50K, 75K, 120K and 240K nucleotides) [10]. The outputs for each convolution were merged and insertions contained within the smallest kernel were used to define the CIS. The *P*-value for each CIS was adjusted by chromosome and a cut-off of $P<0.05$ was used.

Gene-Centric Common Insertion Site (gCIS) Computational Method

Gene-centric CISs (gCIS) were analyzed using the methods published by Brett *et al.*[11]with slight modifications. We only considered transposon insertions at uniquely mappable TA dinucleotides within the coding regions of all RefSeq genes. A mappable TA is defined by the presence of a uniquely mapped 40bp junction on either side of the TA in the mouse genome. The *P*-value for each gCIS is based on chi-square analysis and the threshold using the Bonferroni correction (0.05/21508 RefSeq genes) is 2.32x10-6.

SCREEN 2: Perez-Mancera*et al.*[6]

Mice

The KCTSB13 compound mutant mice were generated by crossing the following strains of mice: *LSL-Kras*^{G12D}[12], the pancreatic-specific *Cre* recombinase driver *Pdx1-Cre*[8], the *T2/Onc2* transgenic line, that contains 30 copies of the transposon in the chromosome 1 [13] and the Rosa26-LSL-SB13 strain, that conditionally expresses the SB13 transposase[6].The KCTSB13 compound mutant mice express both the *Kras*^{G12D} and *SB13* alleles in pancreatic cell progenitors. Animals were monitored until clinical signs of tumor formation. Tumor-free survival was significantly decreased when oncogenic *Kras*^{G12D} was combined with *SB13*.

CIS analysis
WA similar approach to SCREEN 1 above was used, analysing the CISs with Multiple kernel scales (widths of 15K, 30K, 50K, 75K, 125K and 250K nucleotides). For highly significant CISs with narrow spatial distributions of insertion sites, the 15K kernel was the scale on which CISs were identified. Additional statistical analysis of insertion sites was performed using a Monte Carlo framework [14].

Annotation of Human Mutation and CNV data
We converted mouse gene symbols to human symbols using custom R scripts based on the NCBI homologene.data file (build65). And annotated the geneset with detected mutations and CNV using the high confidence calls from each of the screens.

SURVIVAL ANALYSIS
We extracted RNA from tumour samples using the Qiagen Allprep® Kit (Qiagen, Valencia, CA) in accordance with the manufacturers instructions, assayed for quality on an Agilent Bioanalyzer 2100 (Agilent Technologies, Palo Alto, CA), and subsequently hybridized to Illumina Human HT-12 V4 microarrays. Raw files (idat format) were processed using *IlluminaGeneExpressionIdatReader* (Cowley et al, manuscript in preparation). Following array quality control, these data were preprocessed by variance stabilization transformation (VST), and then robust spline normalization, using the *lumi* R/Bioconductor package [15]. For each gene, the probe with most variable expression across tumour samples was selected, and expression levels of probes were discretized using the diverse percentile corresponding to the estimated proportion of genomic aberrations in human PC. Disease-specific survival was used as the primary endpoint. Median survival was estimated using the Kaplan-Meier method and the difference tested using the logrank test. Survival analysis was performed using the *survival*R/Bioconductor package[16].

Mouse models and RTPCR of AXON GUIDANCE genes

***In vitro* acinar to ductal metaplasia (ADM) model.** Pancreatic acinar cells were isolated from 8-12 week old C57BL/6 mice, as previously described[17]. Briefly, total mouse pancreas was digested with a collagenase P (Roche) solution and cellular aggregates were washed in Hank's balanced salt solution (HBSS) (Gibco BRL) supplemented with5% fetal bovine serum (Sigma) and filtered over 500μm and 100μm meshes (Spectrum Laboratories). Viable cells were recovered after low speed centrifugation over 30% FBS and cultured for 5 days in suspension in RPMI 1640 glutamax medium (Gibco BRL) supplemented with 10% FBS, penicillin (75 pg/mL), streptomycin (100 pg/mL), geneticin sulphate (25 μg/mL), and soybean trypsin inhibitor (0.1 mg/mL) (all from Sigma) on untreated plastic (Sterilin). These cultures reproduce persistent acinar to ductal metaplasia[17,18].

Pancreatic Injury: In vivo acinar to ductal metaplasia (ADM) model. Acute pancreatitis was induced in 8-12 weeks C57BL/6 mice by eight hourly, intraperitoneal injections of sulphated Caerulein on two consecutive days (2μg caerulein/200μl injection volume, Sigma)[19]. Untreated pancreas was used as a control. Animals were sacrificed at 48h after initiation of treatment, a time point where acinar to ductal metaplasia had occurred and after which acini will recover differentiation[18,19].

Murine PDAC model. A genetically engineered mouse strain with an activating mutation of KRas and a mutant p53 driven by the Pdx1 gene promoter (LSL-KRasG12D; Pdx1-Cre; Trp53^{R172H}) was used as a model that recapitulates the development of human PDAC tumors between 6 and 9 months of age[20]. Established tumours were sampled at the stage of ascites development. PDAC histology was confirmed in paraffin sections.

Quantitative RT-PCR analysis. Total RNA was isolated from cells using the GenElute Mammalian Total RNA kit (Sigma) and from mouse tissue using Trizol Reagent (Invitrogen) according with manufacturer's procedure. RNA integrity was assessed using Agilent-technology (2100 Bioanalyzer; Agilent, Palo Alto, CA). Reverse transcription was performed using Superscript III First Strand Synthesis System (Invitrogen) and 10 ng RNA-equivalent was used for PCR. Real time PCR was performed with Power SYBR Green PCR Master Mix (Applied Biosystems) using the 7900HT Fast Real-Time PCR System (Applied Biosystems) (primer sequences available on request). All analyses were done in triplicate and melting curve analysis was performed to control for product quality and specificity. Expression levels were calculated using the comparative method of relative quantification, with Hprt as normalizer. Data are analysed by Prism 5.0 applying Student's t-test. Results are presented as mean±SEM relative to the internal control (purified acini at start of culture, normal pancreas tissue, and WT normal pancreas, respectively). Statistical significance is considered when P values are <0.05.

References

1 Fujita, P. A. *et al.* The UCSC Genome Browser database: update 2011. *Nucleic Acids Res* **39**, D876-882, doi:10.1093/nar/gkq963 (2011).

2 Sun, W. *et al.* Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res* **37**, 5365-5377, doi:10.1093/nar/gkp493 (2009).

3 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol* **12**, R41, doi:10.1186/gb-2011-12-4-r41 (2011).

4 Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proc Natl Acad Sci USA* **108**, 12372-12377, doi:10.1073/pnas.1109363108 (2011).

5 Mann, K. M. *et al.* Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. *Proc Natl Acad Sci USA* **109**, 5934-5941, doi:10.1073/pnas.1202490109 (2012).

6 Pérez-Mancera, P. A. *et al.* The deubiquitinase USP9X suppresses pancreatic ductal adenocarcinoma. *Nature*, doi:10.1038/nature11114 [Epub ahead of print] (2012).

7 Luo, B. *et al.* Highly parallel identification of essential genes in cancer cells. *Proc Natl Acad Sci USA* **105**, 20380-20385, doi:10.1073/pnas.0810485105 (2008).

8 Hingorani, S. R. *et al.* Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse. *Cancer Cell* **4**, 437-450 (2003).

9 Copeland, N. G. & Jenkins, N. A. Harnessing transposons for cancer gene discovery. *Nat Rev Cancer* **10**, 696-706, doi:10.1038/nrc2916 (2010).

10 March, H. N. *et al.* Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nature genetics* **43**, 1202-1209, doi:10.1038/ng.990 (2011).

11 Brett, B. T. *et al.* Novel molecular and computational methods improve the accuracy of insertion site analysis in Sleeping Beauty-induced tumors. *PLoS One* **6**, e24668, doi:10.1371/journal.pone.0024668 (2011).

12 Jackson, E. L. *et al.* Analysis of lung tumor initiation and progression using conditional expression of oncogenic K-ras. *Genes Dev* **15**, 3243-3248, doi:10.1101/gad.943001 (2001).
13 Collier, L. S., Carlson, C. M., Ravimohan, S., Dupuy, A. J. & Largaespada, D. A. Cancer gene discovery in solid tumours using transposon-based somatic mutagenesis in the mouse. *Nature* **436**, 272-276, doi:10.1038/nature03681 (2005).
14 Keng, V. W. *et al.* A conditional transposon-based insertional mutagenesis screen for genes associated with mouse hepatocellular carcinoma. *Nat Biotechnol* **27**, 264-274, doi:10.1038/nbt.1526 (2009).
15 Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-1548, doi:btn224 [pii] 10.1093/bioinformatics/btn224 (2008).
16 Therneau, T. & Lumley, T. survival: Survival Analysis Including Penalised Likelihood. *R package version 2.36-10*, URL http://cran.r-project.org/web/packages/survival/ (2011).
17 Pinho, A. V. *et al.* Adult pancreatic acinar cells dedifferentiate to an embryonic progenitor phenotype with concomitant activation of a senescence programme that is present in chronic pancreatitis. *Gut* **60**, 958-966, doi:10.1136/gut.2010.225920 (2011).
18 Rooman, I. & Real, F. X. Pancreatic ductal adenocarcinoma and acinar cells: a matter of differentiation and development? *Gut*, doi:10.1136/gut.2010.235804 (2011).
19 Jensen, J. N. *et al.* Recapitulation of elements of embryonic development in adult mouse pancreatic regeneration. *Gastroenterology* **128**, 728-741 (2005).
20 Hingorani, S. R. *et al.* Trp53R172H and KrasG12D cooperate to promote chromosomal instability and widely metastatic pancreatic ductal adenocarcinoma in mice. *Cancer Cell* **7**, 469-483 (2005).