Genome-wide exactly recurrent (n≥6) noncoding mutations in PDA

Nearest gene	Patients (%)	Sequence	Gene name/protein function
COX7B2	7 (2.27)	GTCA(T)TA	cytochrome c oxidase subunit
OSBPL9	7 (2.27)	ATTA(T)AT	oxysterol binding protein-like 9; cholesterol transfer protein
WASF3	7 (2.27)	TTTT(A)AA	Wiskott-Aldrich syndrome protein family
ZNF81	7 (2.27)	AATA(T)AA	zinc finger protein; transcription factor
BNC2	6 (1.95)	TTTA(T)AA	basonuclin 2; zinc finger transcription factor
ELMO1	6 (1.95)	TTTA(T)AA	engulfment and cell motility 1; cytoskeletal rearrangement
GPR98	6 (1.95)	TCTC(A)TC	G protein-coupled receptor; central nervous system development
MYO16	6 (1.95)	GCTT(C)GC	myosin XVI; actin-based motor with ATPase activity
PDE3B	6 (1.95)	ATAG(T)AG	phosphodiesterase 3B; regulates cAMP binding of RAPGEF3
SOX5	6 (1.95)	ATAG(T)AG	SRY (sex determining region Y)-box 5; transcription factor
TMEM232	6 (1.95)	ATAG(T)AG	transmembrane protein 232

Supplementary Table 1 - Exactly recurrent mutations in PDA.
The most common exactly recurrent mutations across the patient cohort. Sequence of mutant allele in parenthesis.

NCM overlap with known PDA genes

PDA gene	CRR (# patients)
KRAS	-
TP53	-
CDKN2A	-
SMAD4	-
ARID1A	-
MLL3	-
PIK3CA	-
MAP2K4	-
BRAF	-
ZIM2	JUND (6)
PEG3	TAF1 (6), FOSL2 (5)
NEB	-
FLG	-
TGFBR2	-
ATM	-
HMCN1	-
ACVR1B	-
XIRP2	-
APC	-
FBXW7	-
RB1	-
USP47	-
BRCA2	-
PALB2	-
LKB1	-
PRSS1	-

Supplementary Table 2 - Distribution of gene-proximal NCMs near known PDA genes.
Analysis of the association of NCM clusters with known PDA genes.

Multivariate Analysis			
	Variable	**Hazard Ratio (95% CI)**	***P* Value**
A. Clinico-pathological and *PTPRN2* (n = 254, Starting model)	Sex (Male)	1.16 (0.83 – 1.62)	0.3933
	Lymph Node Metastases (Positive)	1.08 (0.65 – 1.81)	0.7561
	Grade (G3/4)	1.68 (1.19 – 2.38)	0.0033
	Tumor Size (> 20 mm)	1.90 (1.04 – 3.50)	0.0378
	Margin Involvement (Positive)	1.25 (0.87 – 1.81)	0.2208
	Tumor Location (Body/Tail)	1.71 (1.15 – 2.54)	0.0078
	Perineural Invasion (Positive)	1.48 (0.88 – 2.51)	0.1416
	Vascular Invasion (Positive)	1.65 (1.07 – 2.54)	0.0227
	PTPRN2 Expression (Low)	1.42 (1.00 – 2.01)	0.0505
B. Clinico-pathological and *PTPRN2* (Final model)	Grade (G3/4)	1.69 (1.21 – 2.38)	0.0021
	Tumor Size (> 20 mm)	1.98 (1.10 – 3.60)	0.0239
	Tumor Location (Body/Tail)	1.87 (1.26 – 2.75)	0.0017
	Vascular Invasion (Positive)	2.05 (1.44 – 2.92)	<0.0001
	PTPRN2 Expression (Low)	1.43 (1.00 – 2.02)	0.0453

Supplementary Table 3 - PTPRN2 multivariate analysis.
Multivariate analysis of clinico-pathological variables and PTPRN2 expression in the patient cohort.

CRR	EM Score	CRR	EM Score	CRR	EM Score
SUZ12	-0.686694944	HDAC2	0.150381608	RXRA	0.273722674
CTBP2	-0.674670553	E2F6	0.150409791	HMGN3	0.281526015
POU5F1	-0.56033248	BHLHE40	0.151537078	NFKB1	0.288817568
ZNF274	-0.245849532	POLR3A	0.159325596	ZEB1	0.303120139
ZZZ3	-0.161998971	JUN	0.162188074	ETS1	0.310574829
BATF	-0.135543815	EBF1	0.163190758	TAF7	0.313174867
MAFF	-0.075863388	TRIM28	0.163945818	CHD2	0.315050091
ESR1	-0.039131089	TFAP2A	0.164300299	JUNB	0.325497894
NANOG	-0.038471214	ZNF143	0.169254105	ATF3	0.326109056
ZBTB33	-0.030552156	STAT1	0.171479031	NRF1	0.326352606
ZNF263	-0.025572293	GABPA	0.172425715	POU2F2	0.327706868
MAFK	-0.02294752	HSF1	0.176973018	SRF	0.337389028
REST	-0.003580494	PBX3	0.177027193	NFE2	0.342471413
SIRT6	0.002156973	FOS	0.17863575	SMARCB1	0.345768127
RAD21	0.009059509	NR3C1	0.178665134	SIN3A	0.354604782
HNF4A	0.02493997	YY1	0.179418835	TAF1	0.363204782
NR2C2	0.030842166	SMARCA4	0.187416973	BRF1	0.364973559
FOXA2	0.035231355	SP1	0.189263356	RFX5	0.372691206
FOXA1	0.036020516	IRF4	0.189407346	SREBF2	0.380044338
BCL11A	0.040986649	ELF1	0.190484821	BDP1	0.396235351
MEF2C	0.046893332	SMARCC1	0.196306055	SIX5	0.402922065
HNF4G	0.047321602	TCF4	0.196772009	SP2	0.411373792
GATA3	0.056356258	EP300	0.198545703	BRCA1	0.420438253
CTCF	0.057809323	PAX5	0.199920331	PRDM1	0.421602184
CEBPB	0.069849107	NFYB	0.20623068	E2F4	0.421865085
TAL1	0.071199973	GATA1	0.207466124	IRF1	0.433584837
SPI1	0.085668185	E2F1	0.209186604	BCLAF1	0.433812837
SETDB1	0.093669622	FOSL2	0.220324569	MXI1	0.436922274
ZBTB7A	0.097578103	USF2	0.221126568	KAT2A	0.451063271
GATA2	0.098239781	TBP	0.227528399	IRF3	0.475299075
SMC3	0.103732676	ESRRA	0.230647673	SMARCC2	0.479021415
MYC	0.103926411	TFAP2C	0.231536652	ELK4	0.490643603
BCL3	0.112807799	SREBF1	0.240511397	THAP1	0.493514238
MAX	0.116546688	HDAC8	0.241186695	STAT2	0.524018361
EGR1	0.119910439	TCF12	0.251270827	GTF2B	0.544751967
USF1	0.120445295	CCNT2	0.263950123	FAM48A	0.567152197
CTCFL	0.129630002	NFYA	0.267597423	GTF2F1	0.669284919
MEF2A	0.130585096	STAT3	0.268609945	WRNIP1	0.693109157
JUND	0.132604078	FOSL1	0.268875227	RDBP	1.123049853

Supplementary Table 4 - CRR expression modulation scores.
Effect of CRR on activity of neighboring gene compared with all other genes in the genome (see Online Methods for analysis details). EM Score, expression modulation score.

Supplementary Note

Results

Somatic mutation calling

SNVs were called using BWA and GATK as previously described[1]. The rates and distribution of coding mutations in commonly mutated PDA genes (*KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *ARID1A*) in the patient cohort was consistent with previous reports (**Supplementary Fig. 2**). We confirmed somatic status of the variants by searching for any evidence of the putative tumor variant in whole genome sequences of matched normal tissue for each patient.

Depletion of SNPs

Cancer mutations are depleted in accessible regulatory regions, particularly in those of the originating cell type[2]. Our set of SNPs was similarly depleted in DHSs from 164 cell types mapped by ENCODE and the Roadmap Epigenomics projects. The top ten most depleted DHS sets were from blood cells, for which >1.5 times fewer SNPs were present than after shuffling. Mapped cell types related to the pancreas were also depleted but inconspicuous in the broader context of many other cell types.

General feature of FunSeq2

The FunSeq2 pipeline filters cancer variants to exclude common polymorphisms from the 1000 Genomes project and retain those in noncoding regions. Further filters select for non-coding mutations in “sensitive” regions (those under strong negative selection), regions of high centrality in the protein-protein interaction network, ENCODE-defined regions captured by chromatin immunoprecipitation (ChIP), and mutations disrupting transcription factor binding motifs. We confirmed the somatic status of the mutations by comparing with matched normal DNA for each patient.

Enrichment of NCMs in CRRs

Noncoding mutations were found to be specifically enriched in certain classes of gene-proximal CRRs, including binding regions for the RNA Polymerase III Transcription Initiation Factors *BRF1* and *BDP1*, the Polycomb Repressive Complex 2 component *SUZ12*, the lysine acetyltransferase *KAT2A*, the negative elongation factor of RNA Polymerase II *RDBP*, and the transcriptional repressor *CTBP2*.

Discussion

The number of NCM-gene expression associations we uncover in this study is higher than that of similar whole genome cancer analyses. Several differences may account for this finding. First, we focused exclusively on a large number of samples from a single cancer type, rather than including a diverse array of cancers. As recurrent somatic NCMs are relatively uncommon (as are most coding mutations in PDA), reducing the heterogeneity of the samples allows detection of rare events. Second, we used GECCO to select those NCMs that are most likely to cause alterations in gene expression and focused on clusters of mutations within specific regulatory regions in close proximity to genes.

We provide evidence that NCMs in specific regulatory element classes are selected for during tumor evolution. These highly mutated regulatory element classes are predominantly those with the greatest impact on gene expression. Further research will

be required to uncover if these regions are actively promoting or repressing gene expression in PDA, or if they are independently associated with highly expressed or repressed genes.

Supplementary References

1. Waddell, N.*, et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495-501 (2015).
2. Polak, P.*, et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360-364 (2015).