nature genetics

Recurrent noncoding regulatory mutations in pancreatic ductal adenocarcinoma

Michael E Feigin[1,2,22], Tyler Garvin[3,22], Peter Bailey[4], Nicola Waddell[5,6], David K Chang[4,7–9], David R Kelley[10], Shimin Shuai[11], Steven Gallinger[12,13], John D McPherson[14,21], Sean M Grimmond[4,6,21], Ekta Khurana[15], Lincoln D Stein[11,16], Andrew V Biankin[4,9,20], Michael C Schatz[1,17,18] & David A Tuveson[1,2,19]

The contributions of coding mutations to tumorigenesis are relatively well known; however, little is known about somatic alterations in noncoding DNA. Here we describe GECCO (Genomic Enrichment Computational Clustering Operation) to analyze somatic noncoding alterations in 308 pancreatic ductal adenocarcinomas (PDAs) and identify commonly mutated regulatory regions. We find recurrent noncoding mutations to be enriched in PDA pathways, including axon guidance and cell adhesion, and newly identified processes, including transcription and homeobox genes. We identified mutations in protein binding sites correlating with differential expression of proximal genes and experimentally validated effects of mutations on expression. We developed an expression modulation score that quantifies the strength of gene regulation imposed by each class of regulatory elements, and found the strongest elements were most frequently mutated, suggesting a selective advantage. Our detailed single-cancer analysis of noncoding alterations identifies regulatory mutations as candidates for diagnostic and prognostic markers, and suggests new mechanisms for tumor evolution.

PDA is a highly lethal malignancy with a 5-year survival rate of 6%, due to therapy resistance and late stage at diagnosis[1]. A detailed understanding of the molecular alterations underlying PDA is required to uncover mechanisms of tumorigenesis and enable development of effective therapies. Exome sequencing efforts have identified genes (*KRAS*, *TP53*, *CDKN2A*, *SMAD4*) and pathways (Wnt/Notch, transforming growth factor-β (TGF-β), axon guidance, cell adhesion) important for PDA progression[2,3]. However, the exome comprises less than 2% of the human genome. Whole-genome sequencing (WGS) analyses have uncovered an average somatic mutation rate of 2.64 mutations per Mb in PDA, indicating that PDA tumors often carry thousands of mutations, most of which are located in noncoding regions and are completely uncharacterized[4].

Relevance of noncoding mutations (NCMs) to cancer development was previously established with the discovery of highly recurrent mutations in the telomerase reverse transcriptase (*TERT*) promoter in sporadic and familial melanoma[5,6]. These mutations create binding motifs for transcription factors in the ETS family and lead to increased *TERT* transcriptional activity[5,7]. Subsequent reports identified *TERT* promoter mutations in a wide-range of human tumors, including glioblastoma and hepatocellular carcinoma[8]. *TERT* promoter mutations are the most common genetic alterations in bladder cancer and correlate with recurrence and survival, demonstrating the potential of NCMs to act as clinical biomarkers[9]. NCMs have also been demonstrated to drive tumor progression from intergenic elements. Somatic mutations in a subset of T-cell acute lymphoblastic leukemia cases generate binding sites for the MYB transcription factor, creating a superenhancer driving expression of the *TAL1* oncogene[10]. Recent analyses have pooled WGS data from multiple cancer types and hundreds of patients, identifying recurrent mutations in regulatory elements of several genes, including *TERT*[11–15]. While multi-cancer studies can identify ubiquitous cancer variants, in-depth analysis of individual cancer subtypes is required for uncovering disease-specific alterations[16].

To detect somatic NCMs in PDA, we developed a computational pipeline to analyze WGS data for 308 PDA tumors from the International Cancer Genome Consortium (ICGC)[17]. We used the FunSeq2 pipeline[18,19] to initiate prioritization of noncoding mutations, which identified hundreds of thousands of noncoding somatic mutations with potential functional implications. To discriminate among this large number of NCMs, we developed GECCO (Genomic Enrichment Computational Clustering Operation) to identify candidate NCMs that drive differential gene expression. This approach reduced the number of putative gene-proximal regulatory regions by three orders of magnitude to a set of high-confidence calls.

Using GECCO, we identify recurrent mutations and interrogate expression data from matched tumors to find variants associated with changes in mRNA levels. We find significant differential expression of 16 genes associated with NCMs. For two of these genes, *PTPRN2* and *SLC12A8*, we uncover previously unidentified clinical relevance in PDA. Specifically, we find that *PTPRN2* expression level is an independent prognostic variable for overall patient survival. Pathway analysis of the genes associated with recurrent NCMs identifies known and new PDA pathways. Furthermore, we find enrichment for mutations in specific regulatory regions, suggesting that NCMs may be acted upon by selection during tumor formation. Our analysis provides a model for tumor evolution via the formation and selection for alterations in noncoding regulatory elements of specific genes as a means of controlling specific biological pathways.

A full list of affiliations appears at the end of the paper.

Received 28 July 2016; accepted 10 April 2017; published online 8 May 2017; doi:10.1038/ng.3861

RESULTS

To analyze NCMs in PDA, we selected all 405 patients with WGS data from the ICGC Pancreatic Cancer Genome Project. We determined the total number of somatic single nucleotide variants (SNVs) and small insertions or deletions (indels) for each patient and retained those with mutation load within 3 s.d. of the mean (mean = 7,937; range = 1–440,471) to exclude the hypermutated tumors with unlocalized replication defects (**Fig. 1a** and **Supplementary Fig. 1**). In total, 2,248,158 SNVs and indels from 308 PDA patient samples were retained for analysis.

General features of GECCO

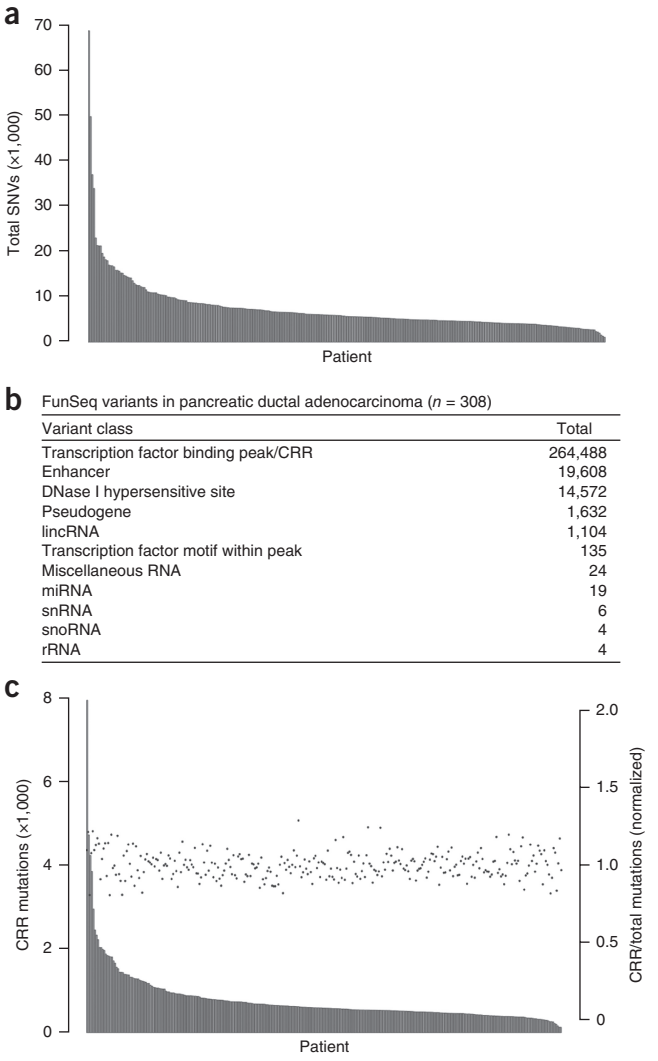
To discover the effect of noncoding mutations on PDA progression and patient outcome, we developed the computational pipeline GECCO (**Fig. 2**). GECCO begins by selecting noncoding mutations falling within Encyclopedia of DNA Elements[20] (ENCODE)-defined transcription factor binding peaks—herein referred to as *cis*-regulatory regions (CRRs), as not all proteins profiled are transcription factors and may be part of larger regulatory complexes. It then proceeds with downstream processing in two parallel modules. We define a "CRR class" to be all CRRs that are bound by the same DNA-binding protein (for example, CTBP2, with 1,781 CRRs across the genome) or proteins involved in DNA-binding complexes (for example, SUZ12, with 1,618 CRRs across the genome). The first module of GECCO associates NCMs with proximal genes and uses permutation testing to identify highly mutated clusters that correlate significantly with changes in gene expression. The second module calculates the mutation rate of each CRR to determine which specific CRR classes are more commonly mutated in PDA.

In the second module, GECCO computes an expression modulation score (EMS) using coupled gene expression data to determine the regulatory impact of each CRR class. The EMS can be used to generate a rank-sorted list of CRRs based on the strength of their relative gene regulatory impact (such that the strongest activators and repressors fall at either end of the list). Taken together, the results generated from GECCO provide information on the impact of NCMs on the expression level of individual genes and identify potential driver transcription factors. Finally, GECCO merges the results of both modules to perform pathway and clinical survival analysis, allowing insights into PDA biology and patterns of somatic mutations in cancer.

Prioritization of noncoding mutations

We first identified NCMs in the exact same genomic position in multiple patients and removed common human variants (minor allele frequency > 5% in the 1000 Genomes Project phase I) (**Supplementary Table 1**). This identified several variants reaching over 2% incidence ($n \geq 7$ of 308 patients) in the patient cohort (**Supplementary Table 1**). Among the 11 genes associated with these variants, 6 have been implicated in tumorigenesis: *WASF3* (ref. 21), *BNC2* (ref. 22), *ELMO1* (ref. 23), *GPR98* (ref. 24), *PDE3B* (ref. 25) and *SOX5* (ref. 26). Notably, 10 of 11 of these mutations were found in introns. However, none of the exactly recurrent mutations disrupted, or created, transcription-factor-binding motifs (as defined by the JASPAR transcription factor binding profile database[27]) or fell within known regulatory elements. This analysis is consistent with several pan-cancer analyses that found few exactly recurrent mutations outside the well-characterized *TERT* promoter mutations[11,12].

We extended this analysis by prioritizing NCMs by their association with functional annotations and clustering within regulatory elements. We used the FunSeq2 computational pipeline[18,19] as a high-level filter to remove common variants and identify putative somatic regulatory mutations with functional impact. One important benefit of this approach is that it relies on functional information and thus drastically reduces any biases



Variant class	Total
Transcription factor binding peak/CRR	264,488
Enhancer	19,608
DNase I hypersensitive site	14,572
Pseudogene	1,632
lincRNA	1,104
Transcription factor motif within peak	135
Miscellaneous RNA	24
miRNA	19
snRNA	6
snoRNA	4
rRNA	4

Figure 1 Identification of recurrent noncoding mutations in PDA. (**a**) The total number of SNVs was plotted for each patient. (**b**) FunSeq2 was used to detect and characterize putative somatic noncoding mutations from 308 PDA whole-genome sequences. Mutation counts for each functional category are displayed. miRNA, microRNA; snRNA, small nuclear RNA; snoRNA, small nucleolar RNA. (**c**) The number of *cis*-regulatory region (CRR) mutations (gray bars) and CRR/total SNV ratio (black points) were plotted for each patient.

resulting from inhomogeneous mutation rates across the genome. This initial round of filtering identified 301,596 potential somatic drivers across all 308 patients (mean = 1,988; range = 203–17,902) (**Fig. 1b**); 264,488 of the somatic NCMs fell within ENCODE-defined transcription factor-binding peaks, with most of the remaining mutations within enhancers (19,608) or DNase I hypersensitive sites (14,572) (**Fig. 1b**). We focused our analysis on the 264,488 NCMs within the ENCODE-defined CRRs. There was a direct correlation between CRR mutation rate and total SNVs (**Fig. 1c**). In contrast, we observed no correlations between CRR mutation rate and coding mutations in *KRAS*, *TP53*, *CDKN2A*, *SMAD4* or *ARID1A* (**Supplementary Figs. 2** and **3**).

Analysis of *cis*-regulatory mutations

Starting with 264,488 candidate mutations, we used GECCO to focus our analysis on CRRs within 2 kb of each gene (many of which overlap promoters), seeking to identify clusters of mutations in CRRs that

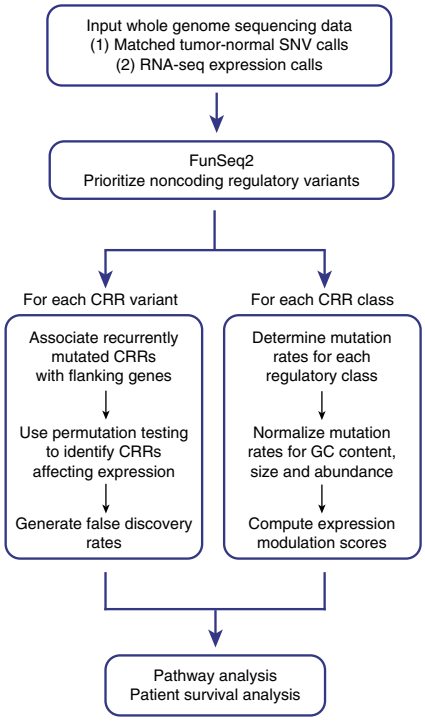


Figure 2 GECCO flowchart. GECCO uses noncoding somatic mutation calls from tumor WGS data to identify clusters of mutations within 2 kb of genes, including those that correlate with changes in gene expression. GECCO also calculates the mutation rate of gene regulatory regions and determines the strength of each regulatory region in terms of the effect on gene expression (EMS). These data can then be used for pathway analysis of genes proximal to noncoding clusters and genes downstream of specific regulatory regions. The gene lists can also be interrogated for patient survival analysis when coupled to outcome data for detection of clinically relevant interactions.

directly affect gene expression (**Fig. 3a**). The requirement to be within 2 kb of a gene excludes many distal enhancer regions but increases the likelihood that a given CRR topologically associates with, and therefore regulates, the expression of its proximal gene. The most frequently mutated CRR (17 patients, 5.52% of cohort) was in a TCF12-binding region proximal to *LHX8* (LIM homeobox 8) (**Fig. 3a**). *LHX8*, a homeobox gene and regulator of craniofacial development, modulates the Hedgehog pathway, a known regulator of PDA pathogenesis[28]. We observed a cluster of mutations in an E2F1-binding region in proximity to *BMP7* (bone morphogenetic protein 7). *BMP7* is a TGF-β family member, with pleiotropic roles in development and cancer progression[29]. GECCO did not detect any recurrent variants in the *TERT* promoter, in concordance with a previous study that failed to detect *TERT* promoter mutations in 24 PDA samples[8]. To determine whether the identified NCMs were within active promoters or enhancers in pancreatic cells, we interrogated histone H3 Lys4 trimethylation (H3K4me3) and Lys27 acetylation (H3K27ac) regions from ENCODE in pancreatic-carcinoma-derived PANC-1 cells. In PANC-1 cells, 37.6% of all transcription factor-binding peaks were found within active predicted promoters or enhancers. In contrast, 58.9% of recurrent NCMs (>5 patients) were found within at least one predicted active promoter or enhancer. The CRRs with recurrent NCMs did not differ significantly in size from those lacking recurrent NCMs. Therefore, recurrent NCMs are enriched in transcriptionally active regions of the genome in pancreatic cancer cells.

We identified clusters of NCMs in regulatory regions of long intergenic non-protein coding RNAs (lncRNAs), including the oncogenic lncRNA metastasis-associated lung adenocarcinoma transcript 1 (MALAT1)[30], and in microRNAs, including the oncogenic miR-21 (ref. 31) (**Fig. 3a**). To infer functional consequences of the most recurrently mutated gene-proximal CRRs, we used data from a published *in vitro* short hairpin RNA (shRNA) screen, which monitored survival in 102 cell lines, of which 13 were pancreatic cancer-derived[32]. Knockdown of 6 of the top 15 genes (*LHX8*, *LMX1B*, *PAX6*, *DMRTA2*, *VAX2* and *CDH15*) was found to decrease cancer cell survival, providing potential functional relevance for these genes as cancer drivers (**Fig. 3a**). Knockdown of two genes, *LMX1B* and *CDH15*, led to selective killing of PDA cell lines among all cancers, suggesting tumor-specific vulnerabilities.

To control for variable CRR size, we calculated a mutational frequency for each cluster harboring at least five mutations, defined as the number of mutations across all patients divided by the number of nucleotides spanning the cluster (**Fig. 3b**). The highest scoring result was an exactly recurrent mutation in the same genomic position in five patients, flanking the acyl-CoA oxidase-like gene *ACOXL*, a known susceptibility locus for chronic lymphocytic leukemia[33]. This mutation was not found to be within a known transcription-factor-binding site as defined by JASPAR. We also identified a cluster of five mutations within 19 nucleotides proximal to the neuronal cell adhesion gene *NRXN3*, a regulator of glioma cell proliferation and migration[34].

While multi-cancer recurrent NCMs have been described[11,12], we lack an understanding of their mutational patterns. For example, it is unknown whether NCMs cluster near the same genes that show recurrent coding mutations for a given disease. Therefore, we looked for clusters of NCMs in association with known PDA genes present in at least five patients (**Supplementary Table 2**). We did not detect any recurrent NCMs in CRRs within 2 kb of *KRAS*, *TP53*, *CDKN2A*, *SMAD4*, *ARID1A* or *MLL3*, in addition to 24 of 26 other PDA genes identified from previous whole-exome analyses (**Supplementary Table 2**)[2,3]. This result is consistent with defects in protein function, rather than alterations in expression, in the pathogenesis of these PDA genes.

Correlations with clinical outcomes from pathway analysis

Pathway analysis of recurrently mutated PDA genes has been used to identify signaling networks and biological processes underlying disease pathogenesis[2,3]. To detect patterns in NCM localization at the pathway level, we used the Database for Annotation, Visualization and Integrated Discovery (DAVID), a functional annotation enrichment algorithm for large-scale biological data sets[35]. Pathway analysis of genes near CRRs containing clusters of mutations (>5 patients) identified significant enrichment of several gene families and regulatory processes, including transcriptional regulation, homeobox genes, axon guidance, cell adhesion and Wnt signaling (**Fig. 3c**). The involvement of three of these pathways (axon guidance, cell adhesion, Wnt signaling) in PDA has been identified from previous exome sequencing studies[2,3]. Furthermore, several homeobox genes and transcription factors have been implicated in PDA pathogenesis, including *PAX6* (ref. 36), *HOXB2* (ref. 37), *HOXB7* (ref. 38) and *RUNX3* (ref. 39). Therefore, NCMs display preferential patterns of localization in the PDA genome and, although not found near canonical PDA genes, may act through modulation of canonical PDA pathways. In addition, we uncover a previously unrecognized localization of NCMs near transcriptional regulators and homeobox genes, suggesting a role for these factors in PDA.

The availability of matched gene expression data from a large number ($n = 96$) of patient samples allowed association studies between specific clusters of mutations and changes in gene expression. For each of the 124,075 CRRs, we determined differential gene expression

a Noncoding gene-proximal mutational clusters in PDA

CRR	Nearest gene	Patients (%)	Gene name/protein function	shRNA
TCF12	*LHX8*	17 (5.52%)	LIM homeobox 8	Yes
JUND	LINC01194	16 (5.19%)	Long intergenic non-protein coding RNA	NA
E2F1	*BMP7*	15 (4.87%)	Bone morphogenetic protein 7	No
SUZ12	*LHX8*	15 (4.87%)	LIM homeobox 8	
WRNIP1	*DUSP22*	15 (4.87%)	Dual specificity phosphatase 22	No
EP300	REREP3	14 (4.55%)	Arginine-glutamic acid dipeptide (RE) repeats pseudogene 3	NA
SUZ12	*LMX1B*	14 (4.55%)	LIM homeobox transcription factor	Yes (P)
SUZ12	*PAX6*	14 (4.55%)	Paired box 6, homeodomain	Yes
TCF12	*ZIC4*	14 (4.55%)	Zinc-finger family member 4	No
HDAC2	*FANK1*	14 (4.55%)	Fibronectin type 3 and ankyrin repeat domains 1	No
FOXA1	REREP3	13 (4.22%)	Arginine-glutamic acid dipeptide (RE) repeats pseudogene 3	NA
NFKB1, POU2F2	*ST8SIA4*	13 (4.22%)	ST8 alpha-N-acetyl-neuraminide alpha-2,8-sialyltransferase 4	No
SIN3A	MIR21	13 (4.22%)	MicroRNA21	NA
SIN3A	*VMP1*	13 (4.22%)	Vacuole membrane protein 1	No
SUZ12	*DMRTA2*	13 (4.22%)	Doublesex-and Mab-3-related transcription factor A2	Yes
SUZ12	*VAX2*	13 (4.22%)	Ventral anterior homeobox 2	Yes
SUZ12	*ZIC4*	13 (4.22%)	Zinc-finger family member 4	
BCLAF1	*DUSP22*	12 (3.90%)	Dual specificity phosphatase 22	
BCLAF1	MALAT1	12 (3.90%)	Metastasis-associated lung adenocarcinoma transcript 1 (lncRNA)	NA
BCLAF1	*VMP1*	12 (3.90%)	Vacuole membrane protein 1	
CDH2, JUND	*ZNF595*	12 (3.90%)	Zinc-finger transcription factor	No
CDH2, JUND	*ZNF718*	12 (3.90%)	Zinc-finger transcription factor	No
FOXA1	*CDH15*	12 (3.90%)	Cadherin 15, type 1, M-cadherin	Yes (P)
HDAC2	*CDH8*	12 (3.90%)	Cadherin 8, type 2	No

b Corrected for bounded gene-proximal CRR

CRR	Nearest gene	Patients (%)	Cluster (bp)	Mutation freq. (%)	Gene name/protein function
BHLHE40	*ACOXL*	5 (1.62%)	1	>100	Acyl-CoA oxidase-like
RAD21	*NRXN3*	5 (1.62%)	19	26.32	Neurexin 3, neuronal cell adhesion
MAFK	*MACROD2*	5 (1.62%)	55	9.09	*O*-acetyl-ADP-ribose deacetylase
EGR1	*ARSD*	5 (1.62%)	65	7.69	Arylsulfatase D
REST	*LILRA5*	5 (1.62%)	81	6.17	Leukocyte immunoglobulin-like receptor
CEBPB	*PDE4B*	6 (1.95%)	129	4.65	Phosphodiesterase 4B, cAMP-specific
NRF1	*ANXA11*	5 (1.62%)	134	3.73	Annexin A11
GATA2	*XKR6*	5 (1.62%)	145	3.45	Kell blood group complex-related
NR3C1	*PXDN*	7 (2.27%)	223	3.14	Phroxidasin Homolog
JUND	*NBPF25P*	5 (1.62%)	162	3.09	Neuroblastoma breakpoint family, pseudogene
STAT3	*SORCS1*	6 (1.95%)	205	2.93	Sortilin-related VPS10 domain containing receptor
USF1	*SCAI*	5 (1.62%)	171	2.92	Suppressor of cancer cell invasion
BRF2	FRG1B	5 (1.62%)	186	2.69	FSHD region gene 1 family, lncRNA
CEBPB	*NRXN1*	5 (1.62%)	227	2.20	Neurexin 1, neuronal cell adhesion
ZNF263	LINC00693	6 (1.95%)	283	2.12	Uncharacterized lncRNA

c Pathways regulated by NCMs in pancreatic ductal adenocarcinoma

Regulatory process/gene family	# Genes altered	*P*-value	Representative altered genes
Regulation of transcription	135	3.9×10^{-15}	*ALX4*, *DMRTA2*, *T*, *TWIST1*, *RUNX3*, *WWTR1*
Homeobox	45	6.2×10^{-26}	*LHX5*, *NKX2-8*, *HOXB4*, *IRX1*, *MSX1*, *VAX2*
Neuron differentiation/axon guidance	53	1.1×10^{-19}	*ROBO1*, *SLIT2*, *NRXN1*, *CTNNA2*, *NCAM2*, *BDNF*
Cell adhesion	24	2.8×10^{-4}	*CDH15*, *CDH8*, *CADM1*, *ITGB2*, *LAMA5*, *CNTN4*
Wnt signaling pathway	18	4.3×10^{-2}	*FZD10*, *FBXW11*, *NKD1*, *TCF7L1*, *EN2*

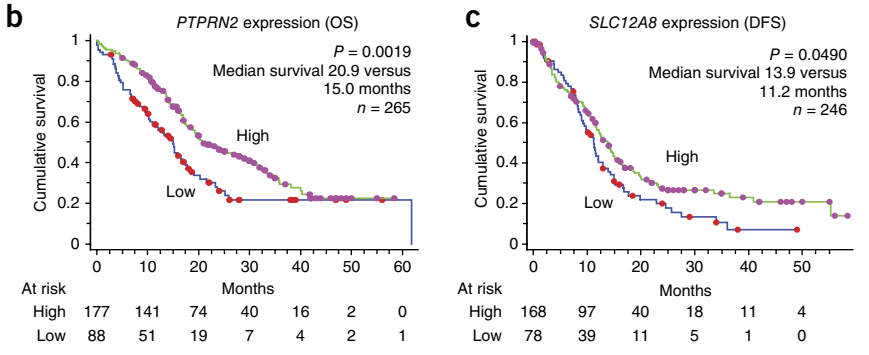
Figure 3 Clustered gene-proximal mutations and pathways in PDA. (**a**) The most common mutational clusters across the patient cohort as determined by GECCO, with associated genes. "Yes" indicates that knockdown promoted cell death in shRNA cancer cell line screen (P denotes PDA-specific). "No" indicates that there was no evidence for effect on cell death in shRNA cancer cell line screen. (**b**) The most significant clusters when corrected for cluster size, as determined by GECCO. (**c**) DAVID pathway analysis was used to identify regulatory processes and pathways from genes associated with recurrent NCMs.

between patients with mutations in a proximal CRR compared to patients without mutations. Using permutation testing, we identified NCMs that significantly modulated expression of their proximal gene and calculated their false discovery rates (FDRs; see Online Methods). Many of the genes with the greatest number of mutations (**Fig. 3a**) did not show significant changes in gene expression. However, this analysis yielded 16 NCMs associated with significant changes in gene expression (at least 3 patients, $P < 0.05$, FDR < 0.25) (**Fig. 4a**). Eight of the 16 NCMs were present in regions marked by H3K4me3 and H3K27ac in PANC-1 cells. None of the statistically significant mutations were associated with increases in gene expression. Three of the genes with statistically significant decreases in expression (*KCNQ1*, *IKZF1*, *TUSC7*) have been implicated as tumor suppressors[40,41], while two (*PTPRN2*, *SNRPN*) are frequently hypermethylated[42,43]. Next we looked for correlations between NCM-associated differential expression and clinical correlates in PDA. The small sample size precluded identification of specific NCMs associated with differences in patient outcome. Therefore, we looked for associations between expression

of these 16 genes and patient outcome. Low mRNA expression of the phosphatase *PTPRN2* and the ion transporter *SLC12A8* were associated with decreased overall survival and decreased disease-free survival, respectively, in a univariate analysis (**Fig. 4b**,**c**). Furthermore, a multivariate analysis identified *PTPRN2* as an independent prognostic variable for overall survival (**Supplementary Table 3**).

a NCMs correlate with gene expression changes

CRR (MUT No.)	Nearest gene	MUT allele	WT allele	Fold change	*P*-value	*q*-value
MAX (5)	*PTPRN2*	0.82	10.92	0.075	0.00593	0.09689
FOSL2 (7)	*KCNQ1*	0.85	6.39	0.133	0.02456	0.18212
TAF7 (9)	*SNRPN*	0.46	3.4	0.135	0.00818	0.11818
NFKB1 (7)	*GYPC*	1.08	7.29	0.148	0.01845	0.15157
TAF1 (6)	*PDPN*	2.09	13.08	0.160	0.03544	0.22016
BCLAF1 (5)	*PRSS12*	1.07	6.46	0.166	0.01107	0.14144
MAFK (3)	*SOX5*	0.29	1.63	0.178	0.02851	0.20379
POU2F2 (6)	*MIR4420*	8.16	40.24	0.203	0.01773	0.15157
WRNIP1 (3)	*IKZF1*	0.64	3.15	0.203	0.01811	0.15157
GATA3 (3)	*PCLO*	0.35	1.67	0.210	0.01113	0.14144
JUND (3)	*TUSC7*	0.98	4.53	0.216	0.02909	0.20560
REST (3)	*MTERF4*	1.46	5.78	0.253	0.02209	0.16542
GATA1 (3)	*FNIP2*	7.59	18.32	0.414	0.02588	0.18929
CEBPB (3)	*PNPLA8*	5.69	13.62	0.418	0.01726	0.15157
EGR1 (5)	*SLC12A8*	4.34	7.99	0.542	0.04185	0.23823
SIN3A (3)	*FAM192A*	20.31	30.48	0.666	0.01788	0.15157



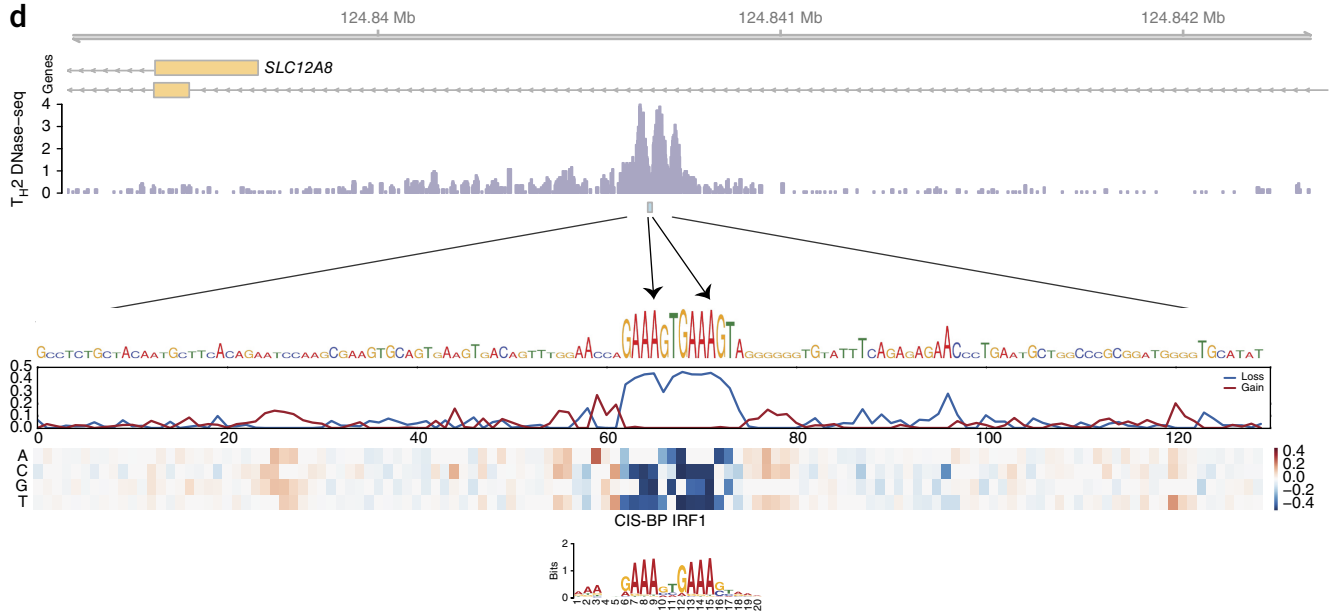


Figure 4 Recurrent gene-proximal mutations correlate with gene expression changes in PDA. (**a**) GECCO used gene expression data from matched PDA patients to correlate NCMs with changes in gene expression. "MUT allele" represents mean expression of linked gene in patients with associated CRR mutations. "WT allele" represents mean expression of linked gene in patients without associated CRR mutations. (**b**) Analysis of overall survival (OS) in PDA patients expressing high (upper 2/3) and low (lower 1/3) levels of *PTPRN2*. Purple dots represent patients with high expression of *PTPRN2* 'at risk' (alive). Red dots represent patients with low expression of *PTPRN2* at risk. (**c**) Analysis of disease-free survival (DFS) in PDA patients expressing high (upper 2/3) and low (lower 1/3) levels of *SLC12A8*. (**d**) Two A→C mutations in a regulatory site on chromosome 3 at positions 124,840,671 and 124,840,678 alter critical nucleotides in an IRF1 and/or PRDM1 binding site. The regulatory site lies in an intron of one isoform and promoter of an alternative isoform of *SLC12A8*. Bottom: heat map displays predicted change in accessibility, considered here as DNase-seq (DNase I hypersensitive site sequencing) signal in the genomic sequence GM12865. CIS-BP, Catalog of Inferred Sequence Binding Preferences; T_H2, type 2 helper T cells. The line plots above measure the maximum (gain) and minimum (loss) predicted change; the loss highlights nucleotides that significantly alter the overall signal upon mutation, as both of these mutations do.

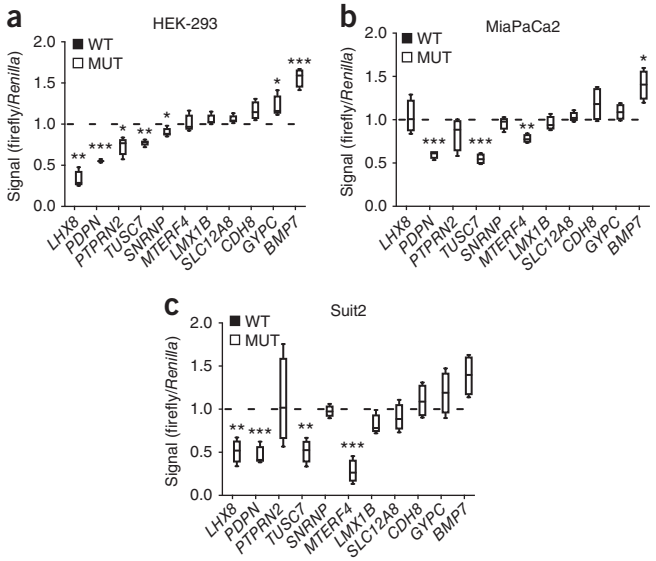


Figure 5 Noncoding mutations modulate luciferase gene expression. (**a**–**c**) Luciferase reporter assays of WT (black) and MUT sequences (white bars) are shown for selected NCMs associated with named genes. For each box-and-whisker plot, center line is the mean, box limits are minimum and maximum values, and whiskers are s.d. Data from a representative experiment ($n = 3$ replicates) with a total of $n = 4$ independent transfected cultures for each cell line are shown. P values were calculated by two-tailed unpaired t test. ($^{*}P < 0.05$; $^{**}P < 0.01$; $^{***}P < 0.001$).

Mechanisms of NCM-modulated expression

To uncover mechanisms by which expression-correlated SNPs may influence transcription, we annotated mutations with their predicted influence on local DNase I hypersensitivity using the software Basset[44] (see Online Methods). The predicted influences of these 55 SNPs were significantly greater in magnitude after Bonferroni correction than a null model of sampling from the full set in 160 of 164 examined cell types. For example, two different mutations in IRF1 and PRDM1 motifs altered critical positions that are likely to affect binding within an intron of *SLC12A8* (**Fig. 4d**). Additional mutations modulate an NRF1 motif in the promoter of *SNRPN* and a GATA motif adjacent to a PU.1 binding site in an intron of *LSAMP* (**Supplementary Fig. 4**). Therefore, GECCO enriches for NCMs with predicted effects on DNase hypersensitivity and transcription factor binding.

While the Basset analysis identified NCMs predicted to affect DNase hypersensitivity, we sought to uncover NCMs directly modulating gene expression. To determine the functional relevance of specific NCMs, we performed luciferase reporter assays in untransformed HEK-293 cells and the MiaPaCa2 and Suit2 PDA cell lines, comparing gene expression driven by wild-type (WT) and mutated (MUT) sequences (**Fig. 5**). Among 11 regions tested, 7 (HEK-293) and 4 (MiaPaCa2, Suit2) mutations significantly altered luciferase expression. Notably, NCMs associated with *PTPRN2*, *PDPN*, *TUSC7*, *SNRNP* and *MTERF4* significantly decreased luciferase expression in one or multiple cell lines, consistent with decreased expression of these genes associated with NCMs in patient samples (**Fig. 4a**). Our validation rate was greater than or comparable in terms of hit rate, and greater in terms of fold change, than that of other recent attempts to identify NCMs driving differential expression[15,16], highlighting the power of GECCO to identify functionally significant NCMs from millions of candidate mutations.

Mutational and expression patterns of CRR classes

The second module of GECCO focuses on CRR classes, rather than individual genes, to identify mutational patterns and overall effects on gene expression of each CRR class (**Fig. 6**). We computed the mutation rate for each CRR class, correcting for element size and abundance

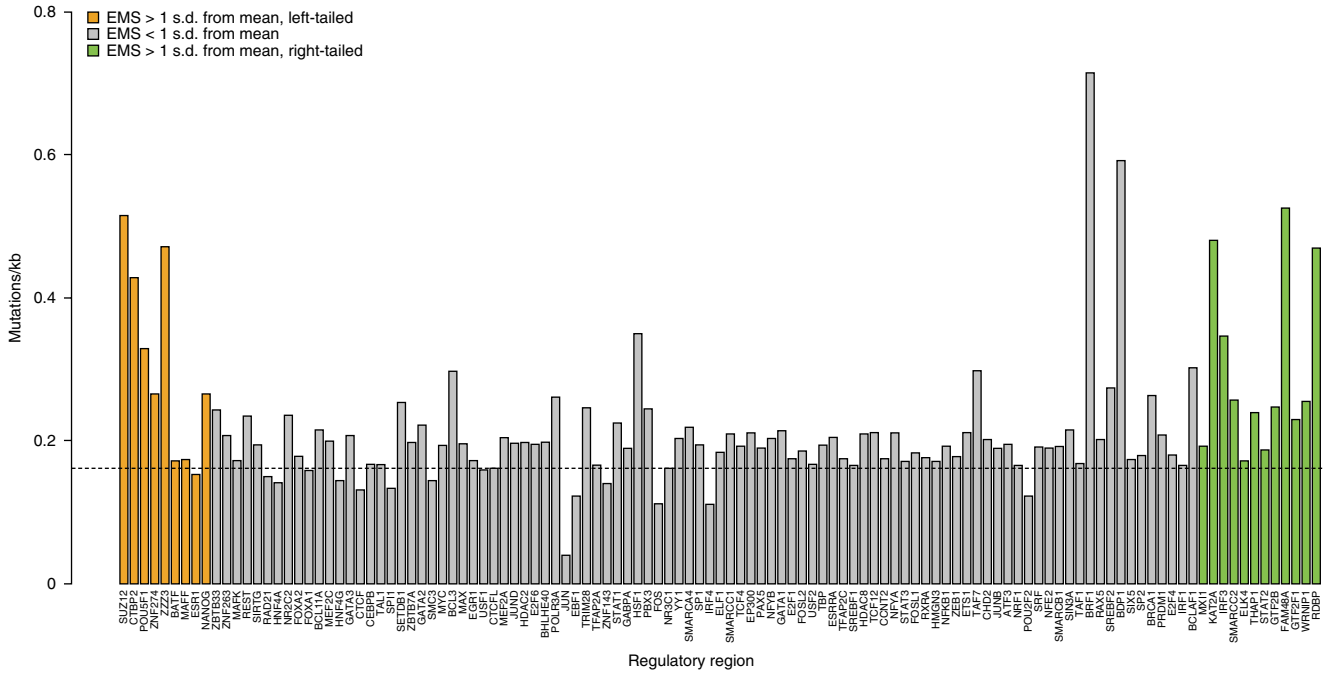


Figure 6 Gene-proximal NCMs are enriched in specific classes of CRRs. Percentage of CRRs with at least two mutations across the patient cohort, corrected for genome abundance and size, ordered from left to right by EMS (most repressive to most active). Dotted line represents mean mutation frequency across all CRRs.

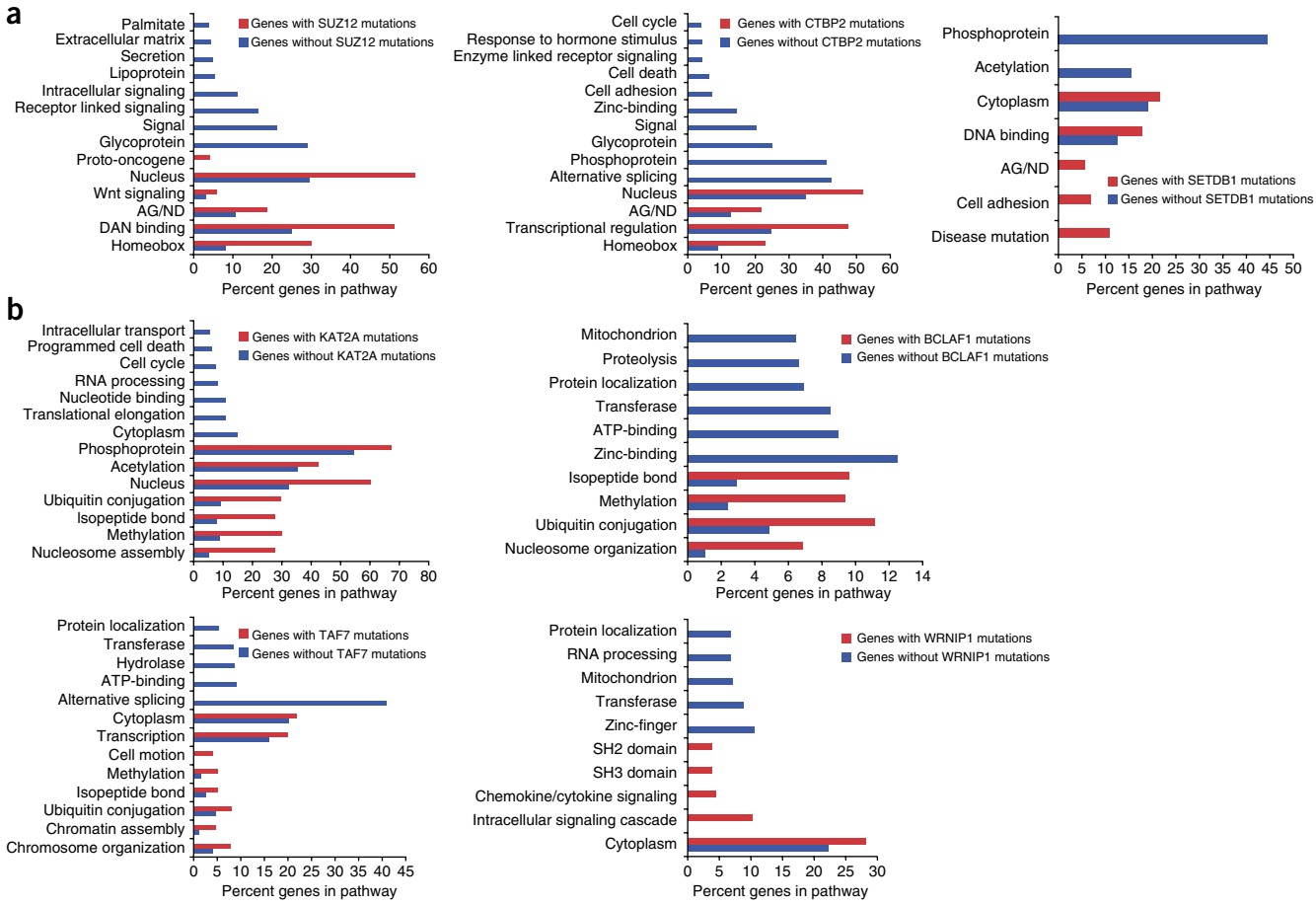


Figure 7 Gene-proximal NCMs in repressors and activators cluster near distinct subsets of genes. (**a**) Pathway analysis of genes associated with recurrently mutated repressive (SUZ12, CTBP2, SETDB1) sites (red bars), versus those never harboring NCMs in those CRRs (blue bars). (**b**) Pathway analysis of genes associated with recurrently mutated activator (KAT2A, BCLAF1, TAF7, WRNIP1) sites (red bars) versus those never harboring NCMs in those CRRs (blue bars). AG/ND, axon guidance/neuronal differentiation.

in the genome. We found no significant effect of GC content on CRR class mutation rate. Noncoding mutations were specifically enriched in certain classes of gene-proximal CRRs (**Supplementary Note**). Next we sought to understand the molecular characteristics of each CRR class in terms of effect on gene expression. We calculated an EMS for each CRR class, reflecting the impact of the presence of that CRR on the expression of the neighboring gene in relation to all other genes. This method compared, for each CRR class, mean expression of genes that are proximal to a CRR to those that are non-proximal. CRRs with strong predicted activating or repressing activity would be proximal to genes with expression levels substantially higher (for activators) or substantially lower (for repressors) than the basal genome expression level (**Supplementary Table 4** and Online Methods). To determine whether the strongest activators and repressors were enriched for those CRRs with the highest mutational frequencies, we considered any activator or repressor that was greater than 1 s.d. from the mean EMS (12 activators, 9 repressors) (**Fig. 6**). The mutational frequencies for each group (activators, repressors, all others with balanced expression) were then calculated and activators and repressors compared to the balanced group ($P = 0.02077$ for activators versus balanced; $P = 0.04982$ for repressors versus balanced). The CRR classes with the highest percentage of mutations across all PDA patients were enriched on either end of the spectrum (most repressive or most active), suggesting that recurrent NCMs are preferentially located in CRR classes with the strongest impact on gene expression. These highly active CRR classes have the largest effect on gene expression and may therefore confer a selective advantage on the cell. In addition, we noted that the six genes identified from the shRNA survival screen (**Fig. 3a**) were all associated with NCMs in highly repressive CRRs. In contrast, every gene that failed to score in the shRNA survival screen was associated with highly active CRRs (**Fig. 3a**).

Pathway dynamics between activating and repressing CRRs

Next we investigated the patterns of noncoding SUZ12 binding site mutations in our patient cohort, as SUZ12 had the highest repressive score and SUZ12 sites were frequently mutated (**Fig. 6** and **Supplementary Table 4**). We generated two distinct lists of SUZ12-associated genes. The first list contained those genes associated with recurrently mutated SUZ12 sites. The second list contained those genes associated with SUZ12 sites that never harbored recurrent NCMs. We then performed pathway analysis on each gene set to identify differences in biological functions (**Fig. 7a**). We found that genes without recurrent SUZ12 binding site mutations were enriched in glycoproteins, intracellular signaling and the axon guidance/neuron differentiation pathway. In contrast, genes with recurrent SUZ12 binding site mutations were more significantly enriched in homeobox genes, transcription factors, Wnt signaling, proto-oncogenes and the axon guidance/neuron differentiation pathway. Surprisingly,

several categories, including glycoproteins, intracellular signaling and extracellular matrix, were completely absent from the mutant SUZ12 gene set. Therefore, there is specificity for the location of NCMs in PDA, not only for certain CRRs, but also for the corresponding cancer-associated genes and pathways.

To further characterize pathways downstream of commonly mutated repressive CRRs, we performed pathway analysis on genes with and without associated CTBP2 binding site mutations (**Fig. 7a**). Genes without CTBP2 binding site noncoding mutations showed a similar pattern of pathway regulation as SUZ12. These pathways were markedly enriched in the gene set associated with CTBP2 binding site mutations, while alternative splicing and glycoproteins were completely absent. We extended this analysis to another repressive CRR with a high mutational frequency, SETDB1 (**Fig. 6a**). Genes associated with recurrent NCMs in SETDB1 binding sites were enriched in axon guidance/neuron differentiation, cell adhesion and disease mutation pathways. Therefore, mutations in highly repressive CRRs are enriched in PDA and selectively associated with genes regulating a core set of biological processes.

We performed a similar analysis for the commonly mutated activator CRRs, including KAT2A, BCLAF1, TAF7 and WRNIP1 (**Fig. 7b**), and again found specificity for the genes and pathways that are commonly mutated. For all CRRs, there were significant differences in the pathways regulated by genes with or without mutations in a given CRR. KAT2A, BCLAF1 and TAF7 shared a very similar pattern of pathway regulation, with significant increases in nucleosome assembly/organization, methylation and ubiquitin conjugation, all processes involved in chromatin dynamics. This suggests that genes associated with NCMs in transcriptional repressors regulate homeobox genes and PDA-associated pathways, while genes associated with NCMs in transcriptional activators may regulate transcriptional dynamics through modulation of chromatin states.

DISCUSSION

We developed a new computational method, GECCO, to systematically analyze the noncoding genome of PDA to uncover recurrent regulatory somatic mutations. We find patterns of NCMs associated with genes regulating canonical PDA pathways, but not associated with commonly mutated PDA genes. Therefore, NCMs may serve as a new mechanism in cancer cells for regulating pathways critical for tumorigenesis. Furthermore, GECCO uncovers mutations correlated with changes in gene expression, including several known tumor suppressors and aberrantly methylated genes. GECCO produces a set of high-confidence calls that enrich for predicted effects on DNase hypersensitivity and transcription factor binding, as well as functional effects on gene expression, as experimentally demonstrated by luciferase reporter assays. We find enrichment for NCMs in specific CRRs and distinct subsets of pathways associated with NCMs in highly repressive and transcriptionally active CRRs as identified by our EMS algorithm. To our knowledge, this is the first comprehensive analysis of noncoding alterations in PDA, providing insights into PDA pathogenesis and serving as a counterpart to the information gleaned from large-scale exome sequencing projects[2,3].

Mutational analysis of patient tumors is increasingly informing treatment decisions, whereas complementary techniques, including microarray, RNA sequencing, fluorescence *in situ* hybridization and immunohistochemistry, are required to analyze changes in gene or protein expression of cancer drivers that lack coding mutations. As somatic mutations in DNA regulatory elements can alter gene expression of cancer drivers, targeted or whole-genome sequencing may provide clinically useful information for these patients, both in terms of therapeutic decisions and clinical prognosis. Our analysis provides the first collection of NCMs that correlate with changes in gene expression in PDA. Furthermore, we uncover clinical outcome relationships for *PTPRN2* and *SLC12A8*, neither of which has previously been implicated in PDA.

Functional validation of NCM–gene expression associations is a critical step in evaluating the robustness of an analysis pipeline. Our luciferase reporter assay experiments demonstrated that GECCO had a higher validation rate in cancer cell lines than any recent study of NCMs[15,16]. Furthermore, the validation rate in HEK-293 cells, a standard cell line for luciferase assays, was 64%, concordant with the expected false discovery rate. Finally, GECCO accurately predicted the directionality of gene expression changes associated with NCMs. NCMs associated with *PTPRN2*, *PDPN*, *TUSC7*, *SNRNP* and *MTERF4* significantly decreased luciferase expression in one or multiple cells lines, consistent with decreased gene expression of these genes associated with NCMs in patient samples. This is in contrast to a recent report wherein the directionality of gene expression changes in the luciferase assay was not consistent with the predicted response[16]. Therefore, GECCO represents a noteworthy improvement in the ability to identify functionally relevant NCMs.

Pathway analysis of the gene lists generated by GECCO led to several unexpected findings. Strikingly, we found that the most highly recurrent somatic NCMs were located near genes in known PDA-associated pathways, including axon guidance, cell adhesion and Wnt signaling, but not the most commonly mutated PDA genes. This suggests that NCMs may drive tumor progression through modulation of PDA-specific pathways, providing an alternative route for pathway activation and a new mechanism of tumorigenesis. Furthermore, 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. Therefore, clusters of NCMs are enriched in gene-proximal regions with the greatest regulatory impact, again providing evidence for selection during tumorigenesis.

Pathway analysis of genes near NCMs within these highly mutated regulatory regions shows selectivity for PDA pathways. These pathways are not enriched when analyzing genes without associated clusters of NCMs, again arguing in favor of selection. Notably, many transcriptional regulators bind selectively to different regions of the genome in malignant versus non-neoplastic cells[45]. We propose that NCMs found within promoters of PDA pathway genes modify regulatory factor binding to alter gene transcription, thereby providing an additional mechanism promoting cancer.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the members of the Tuveson laboratory, C. Vakoc and A. Siepel for discussions. D.A.T. is a distinguished scholar of the Lustgarten Foundation and Director of the Lustgarten Foundation-designated Laboratory of Pancreatic Cancer Research. D.A.T. is also supported by the Cold Spring Harbor Laboratory Association, the V Foundation, PCUK and the David Rubinstein Center for Pancreatic Cancer Research at MSKCC. In addition, we are grateful for support from the following: the STARR Foundation (I7-A718 for D.A.T.), DOD (W81XWH-13-PRCRP-IA for D.A.T.), Louis Morin Charitable Trust (M.E.F.) and NIH (5P30CA45508-26, 5P50CA101955-07, 1U10CA180944-03, 5U01CA168409-5, 1R01CA188134-01A1 and 1R01CA190092-03 for D.A.T. and R01HG006677 for M.C.S.).

AUTHOR CONTRIBUTIONS

M.E.F., T.G., M.C.S. and D.A.T. wrote the manuscript. M.C.S. and D.A.T. supervised the study. T.G. performed FunSeq analysis and developed GECCO. M.E.F. performed pathway analysis. M.E.F., T.G., S.M.G., A.V.B., E.K., S.S., L.D.S., S.G. and J.D.M. contributed to data analysis. D.K.C. and P.B. performed patient outcome analysis. D.R.K. performed Basset analysis. N.W. performed germline sequence analysis.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Siegel, R., Naishadham, D. & Jemal, A. Cancer statistics, 2013. *CA Cancer J. Clin.* **63**, 11–30 (2013).
2. Jones, S. *et al.* Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science* **321**, 1801–1806 (2008).
3. Biankin, A.V. *et al.* Pancreatic cancer genomes reveal aberrations in axon guidance pathway genes. *Nature* **491**, 399–405 (2012).
4. Waddell, N. *et al.* Whole genomes redefine the mutational landscape of pancreatic cancer. *Nature* **518**, 495–501 (2015).
5. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
6. Horn, S. *et al. TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
7. Bell, R.J. *et al.* The transcription factor GABP selectively binds and activates the mutant *TERT* promoter in cancer. *Science* **348**, 1036–1039 (2015).
8. Killela, P.J. *et al. TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. USA* **110**, 6021–6026 (2013).
9. Rachakonda, P.S. *et al. TERT* promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc. Natl. Acad. Sci. USA* **110**, 17426–17431 (2013).
10. Mansour, M.R. *et al.* Oncogene regulation. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 1373–1377 (2014).
11. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
12. Fredriksson, N.J., Ny, L., Nilsson, J.A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
13. Melton, C., Reuter, J.A., Spacek, D.V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
14. Mathelier, A. *et al. Cis*-regulatory somatic mutations and gene-expression alteration in B-cell lymphomas. *Genome Biol.* **16**, 84 (2015).
15. Araya, C.L. *et al.* Identification of significantly mutated regions across cancer types highlights a rich landscape of functional molecular alterations. *Nat. Genet.* **48**, 117–125 (2016).
16. Fujimoto, A. *et al.* Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
17. Hudson, T.J. *et al.* International network of cancer genome projects. *Nature* **464**, 993–998 (2010).
18. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
19. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
20. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
21. Teng, Y., Mei, Y., Hawthorn, L. & Cowell, J.K. WASF3 regulates miR-200 inactivation by ZEB1 through suppression of KISS1 leading to increased invasiveness in breast cancer cells. *Oncogene* **33**, 203–211 (2014).
22. Winham, S.J. *et al.* Genome-wide investigation of regional blood-based DNA methylation adjusted for complete blood counts implicates BNC2 in ovarian cancer. *Genet. Epidemiol.* **38**, 457–466 (2014).
23. Dulak, A.M. *et al.* Exome and whole-genome sequencing of esophageal adenocarcinoma identifies recurrent driver events and mutational complexity. *Nat. Genet.* **45**, 478–486 (2013).
24. Sherman, S.K. *et al.* Gastric inhibitory polypeptide receptor (GIPR) is a promising target for imaging and therapy in neuroendocrine tumors. *Surgery* **154**, 1206–1213, discussion 1214 (2013).
25. Uzawa, K. *et al.* Targeting phosphodiesterase 3B enhances cisplatin sensitivity in human cancer cells. *Cancer Med.* **2**, 40–49 (2013).
26. Renjie, W. & Haiqian, L. MiR-132, miR-15a and miR-16 synergistically inhibit pituitary tumor cell proliferation, invasion and migration by targeting Sox5. *Cancer Lett.* **356**, 568–578 (2015).
27. Sandelin, A., Alkema, W., Engström, P., Wasserman, W.W. & Lenhard, B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* **32**, D91–D94 (2004).
28. Flandin, P. *et al.* Lhx6 and Lhx8 coordinately induce neuronal expression of Shh that controls the generation of interneuron progenitors. *Neuron* **70**, 939–950 (2011).
29. Boon, M.R. *et al.* Bone morphogenetic protein 7: a broad-spectrum growth factor with multiple target therapeutic potency. *Cytokine Growth Factor Rev.* **22**, 221–229 (2011).
30. Gutschner, T. *et al.* The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. *Cancer Res.* **73**, 1180–1189 (2013).
31. Moriyama, T. *et al.* MicroRNA-21 modulates biological functions of pancreatic cancer cells including their proliferation, invasion, and chemoresistance. *Mol. Cancer Ther.* **8**, 1067–1074 (2009).
32. 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 (2011).
33. Lan, Q. *et al.* Genetic susceptibility for chronic lymphocytic leukemia among Chinese in Hong Kong. *Eur. J. Haematol.* **85**, 492–495 (2010).
34. Sun, H.T., Cheng, S.X., Tu, Y., Li, X.H. & Zhang, S. FoxQ1 promotes glioma cells proliferation and migration by regulating NRXN3 expression. *PLoS One* **8**, e55693 (2013).
35. Huang, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
36. Mascarenhas, J.B. *et al.* PAX6 is expressed in pancreatic cancer and actively participates in cancer progression through activation of the MET tyrosine kinase receptor gene. *J. Biol. Chem.* **284**, 27524–27532 (2009).
37. Segara, D. *et al.* Expression of HOXB2, a retinoic acid signaling target in pancreatic cancer and pancreatic intraepithelial neoplasia. *Clin. Cancer Res.* **11**, 3587–3596 (2005).
38. Chile, T. *et al. HOXB7* mRNA is overexpressed in pancreatic ductal adenocarcinomas and its knockdown induces cell cycle arrest and apoptosis. *BMC Cancer* **13**, 451 (2013).
39. Whittle, M.C. *et al.* RUNX3 controls a metastatic switch in pancreatic ductal adenocarcinoma. *Cell* **161**, 1345–1360 (2015).
40. Than, B.L. *et al.* The role of KCNQ1 in mouse and human gastrointestinal cancers. *Oncogene* **33**, 3861–3868 (2014).
41. Geimer Le Lay, A.S. *et al.* The tumor suppressor Ikaros shapes the repertoire of notch target genes in T cells. *Sci. Signal.* **7**, ra28 (2014).
42. Anglim, P.P. *et al.* Identification of a panel of sensitive and specific DNA methylation markers for squamous cell lung cancer. *Mol. Cancer* **7**, 62 (2008).
43. Benetatos, L. *et al.* CpG methylation analysis of the MEG3 and SNRPN imprinted genes in acute myeloid leukemia and myelodysplastic syndromes. *Leuk. Res.* **34**, 148–153 (2010).
44. Kelley, D.R., Snoek, J. & Rinn, J.L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res.* **26**, 990–999 (2016).
45. Squazzo, S.L. *et al.* Suz12 binds to silenced regions of the genome in a cell-type-specific manner. *Genome Res.* **16**, 890–900 (2006).

[1]Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [2]Lustgarten Foundation Pancreatic Cancer Research Laboratory, Cold Spring Harbor, New York, USA. [3]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, USA. [4]Wolfson Wohl Cancer Research Centre, Institute of Cancer Sciences, University of Glasgow, Glasgow, Scotland, UK. [5]QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [6]Queensland Centre for Medical Genomics, Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland, Australia. [7]The Kinghorn Cancer Centre, Cancer Research Program, Garvan Institute of Medical Research, Darlinghurst, Sydney, New South Wales, Australia. [8]Department of Surgery, Bankstown Hospital, Bankstown, Sydney, New South Wales, Australia. [9]South Western Sydney Clinical School, Faculty of Medicine, University of New South Wales, Liverpool, New South Wales, Australia. [10]Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA. [11]Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada. [12]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [13]Division of General Surgery, Toronto General Hospital, Toronto, Ontario, Canada. [14]Genome Technologies Program, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [15]Sandra and Edward Meyer Cancer Center, Institute for Computational Biomedicine, Department of Physiology and Biophysics, Weill Medical College of Cornell University, New York, New York, USA. [16]Informatics and Biocomputing, Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [17]Department of Computer Science, Johns Hopkins University, Baltimore, Maryland, USA. [18]Department of Biology, Johns Hopkins University, Baltimore, Maryland, USA. [19]Rubenstein Center for Pancreatic Cancer Research, Memorial Sloan Kettering Cancer Center, New York, New York, USA. [20]West of Scotland Pancreatic Unit, Glasgow Royal Infirmary,Glasgow, Scotland, UK. [21]Present addresses: University of Melbourne Centre for Cancer Research, University of Melbourne, Melbourne, Victoria, Australia (S.M.G.) and Department of Biochemistry and Molecular Medicine, UC Davis Comprehensive Cancer Center, UC Davis School of Medicine, University of California Davis, Sacramento, California, USA (J.D.M.). [22]These authors contributed equally to this work. Correspondence should be addressed to M.C.S. (mschatz@cshl.edu) or D.A.T. (dtuveson@cshl.edu).

ONLINE METHODS

Data acquisition. All data used in this analysis were downloaded from the International Cancer Genome Consortium (IGCG) data portal (https://dcc.icgc.org/projects/). At our last date of access (Feb 11, 2015), simple somatic mutations (SSMs) for 405 pancreatic ductal adenocarcinoma samples were available from the Australian (PACA-AU) and Canadian (PACA-CA) groups. We downloaded the clinical data, SSMs and, when available, sequence-based gene expression (EXP-S) data for all 405 patients.

Preprocessing. The whole-genome sequencing (WGS) required to call SNVs across all 405 patients and the whole genome RNA-sequencing required to calculate gene expression were carried out by two distinct consortia, one Canadian and one Australian. All SNV calls (SSMs) and gene expression calculations (EXP-S) by these two groups were consolidated by ICGC.

SNV calls from whole-genome sequencing. For each of the 405 patients, we extracted the chromosome, start location, end location, somatic allele and mutated allele from the list of SSMs (file: ssm_open.tsv) and converted to bed format. Many of the SNVs were redundant within patients. For each patient, the list of SNVs was sorted by genomic coordinates and consolidated to contain only a single entry for each unique SNV. A subset of patients had extremely low numbers of SNVs (likely due to poor sequencing results) or high numbers of SNVs (likely due to hypermutated regions, unlocalized replication defects or microsatellite instability). Across all 405 patients, the number of unique SNVs ranged from 1 to 440,471, with a mean of 7,937 and a s.d. of 26,224. To remove outliers, we eliminated all patients with less than 100 SNVs (92 patients in total) or an SNV count more than 3 s.d. away from the mean (5 patients in total). This left 308 patients with a mean SNV count of 7,300 and a range from 1,040 to 68,885.

Gene expression (FPKM) from whole genome RNA-sequencing. Of the 308 patients that passed the previous filtering step, 96 had expression data available from ICGC. For each of the 96 patients, we extracted the normalized read count (FPKM) and Ensembl gene ID (file: exp_seq.tsv). While the vast majority of genes had expression data across all 96 patients, there were several thousand Ensembl genes that only contained expression data for a subset of patients. To streamline and simplify downstream analysis, we kept only the 50,861 Ensembl genes that were shared by all 96 patients. In addition, there were three patients (DO33168, DO35098, DO35100) that had gene expression from either two or three independently sequenced samples. For these three patients, the gene expression for each gene was calculated by taking the mean across all samples.

Analyzing noncoding variants with GECCO. To identify potential noncoding cancer drivers, we first used FunSeq2 (v2.1.0) as a high-level filter to prioritize our SNVs. The unique SNVs for each of the 308 patients were converted to bed format and analyzed by FunSeq2 using the command "./run.sh -inf bed -n" to identify only noncoding variants. This analysis pipeline requires a suite of annotation data that is used to make calls and score noncoding variants. These were downloaded from http://funseq2.gersteinlab.org/. One of these files, "ENCODE.annotation.gz", contains the full list of transcription factor binding sites/CRRs used in our analysis along with their exact genomic coordinates.

Processing recurrently mutated cis-regulatory regions (CRRs). FunSeq2 generates a number of output files, including Recur.Summary, which contains a list of all noncoding elements, the genomic coordinates of these elements, the fraction of patients with a mutation in this element and the full list of patient identifiers along with the genomic locations of each mutation. While the ENCODE annotation data provides a number of different noncoding elements (enhancers, transcription factor binding sites, DNase hypersensitivity, etc.), we chose to focus our analysis on transcription factor binding sites—referred to in this manuscript as CRRs—as they were the most highly represented class of elements identified. CRR proximal genes were found by intersecting CRRs with genes that had been expanded by 2 kb at their 5′ and 3′ ends.

Calculating CRR mutation rates. As described above, the full list of CRRs (121 distinct CRR classes in total), including their counts and genomic positions, can be found in "ENCODE.annotation.gz." GECCO makes two separate calculations across all 121 CRR classes using the CRR genomic information: for a given CRR class, it calculates (i) the fraction of distinct CRR sites that are mutated within the class and (ii) the base level mutation rate for each CRR class (the number of mutations in all CRRs of a given class divided by the total number of base pairs of all CRRs in a given class). For an individual CRR, there are three ways in which GECCO calculates the mutational frequency: (i) by summing the number of mutations in a given CRR, (ii) by calculating the fraction of bases in the CRR that are mutated (that is, mutation counts normalized by read length), or (iii) by calculating the fraction of bases in a CRR mutation cluster. Option (iii) is computed by first determining the cluster size within a CRR, the number of bases required to span all mutations in a given CRR. For example, consider a 2-kb CRR with 9 mutations. If the two most distantly separated of the 9 mutations are 100 bp apart, then the length of the mutation cluster is 100 bp. The mutational frequency of the cluster is then computed by dividing the number of mutations in that cluster by the size of the cluster (9/100 = 9.0%). This approach weights exactly recurrent or proximal mutations more strongly than distant mutations.

Pathway analysis. The Database for Annotation, Visualization and Integrated Discovery (DAVID), a functional annotation enrichment algorithm for large-scale biological data sets, was used for pathway analysis, with the following annotation categories: SP_PIR_KEYWORDS, GOTERM_BP_FAT, KEGG_PATHWAY, PANTHER_PATHWAY, SMART. A Bonferroni-corrected *P*-value of 0.05 was used as a cutoff for enrichment significance.

Survival analysis. 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. Clinico-pathologic variables analyzed with a *P* value of less than 0.25 on log-rank test were entered into Cox proportional-hazard multivariate analysis, and redundant variables were eliminated using a backward elimination method. Statistical analysis was performed using StatView 5.0 Software (Abacus Systems). Overall survival (OS) or disease-free survival (DFS) was used as the primary endpoint (*PTPRN2* expression > 4.98 defined as high, and *SLC12A8* expression > 7.03 defined as high).

Computing differential expression. Differential expression was computed for each recurrently mutated CRR that was within 2 kb of an Ensemble gene using permutation testing. For each CRR/gene pair, the 96 patients with mutation data were split into two groups: patients with mutations in the CRR and patients without mutations in the CRR. Using the expression data downloaded from ICGC for the gene of interest, a *t*-test is performed to generate a single *t*-value (the observed *t*-value). The expression values for patients with mutations in CRRs and the expression values for patients without mutations are then permuted 100,000 times to generate 100,000 additional *t*-values (the permuted *t*-values). These *t*-values generally fit a Gaussian distribution, to which the observed *t*-value is then compared to using a two-tailed test. The empirical *P*-value is computed as the fraction of times (x/100,000) that a permuted *t*-value falls further outside the Gaussian distribution than the observed *t*-value. Once *P*-values have been calculated for all recurrently mutated genes proximal to CRRs, GECCO estimates *q*-values (the false discovery rate) for each call. This is done using the "qvalue" package in R and measures the proportion of false positives incurred given the *P*-value distribution.

Luciferase reporter assay and statistics. Sequences of the 150 bp surrounding specific NCMs (wild type, WT; or mutant, MUT) were synthesized (Integrated DNA Technologies) and cloned into pGL4.23 (Promega), containing a minimal promoter driving firefly luciferase. Five thousand cells per well (HEK-293, MiaPaCa2 or Suit2) were cotransfected in 96-well format with the specific WT or MUT vector and pRL-SV40P (*Renilla* luciferase, Addgene #27163) as a normalization control. Luciferase activity was measured 48 h after transfection with the Dual-Luciferase Reporter Assay System (Promega). Values reported are firefly luciferase divided by *Renilla* luciferase. Analytical statistics were generated in Prism 7.0 (GraphPad), and *P* values are from two-tailed unpaired *t* tests. All cell lines were obtained from ATCC and tested for mycoplasma contamination.

Computing expression modulation scores (EMS). Some CRRs bind transcription factors or transcription factor components with well-known expression

doi:10.1038/ng.3861

modulation, including SUZ12 and CTBP2, which act as transcriptional repressors, or BDP1 and BRF1, which act as transcriptional activators. However, many of the 121 CRRs used in this study have unexplored or unvalidated directions of expression modulation. We developed a method to infer the direction and effect of expression modulation for each CRR class by comparing the expression of genes proximal to CRRs in a given CRR class to the mean expression of all other active genes in the genome.

Many genes are inactive in any given tissue, and in a given RNA-seq experiment, ~50% of genes show low to no expression. For all 96 patients with expression data, we found this also to be true, with ~50% of genes showing no expression. When computing the expression modulation for each CRR class, we ignored all genes that showed no expression in at least 90% of patients (86 patients or more). For a given CRR class and for each of the 96 patients, we compute (i) the mean expression of all genes proximal to CRRs in that class and (ii) the mean expression of all genes nonproximal to a CRR in that class. For a given CRR class, we then compute the log of the ratio between (i) and (ii) for each of the 96 patients and then take the mean of the log ratio for all 96 patients to get a single "expression modulation score" for each CRR class. The log of the ratio will be negative if the mean expression of genes proximal to a CRR class is lower than the genome average (repression) and will be positive if the mean expression of genes proximal to a CRR class is higher than the genome average (activation). Note that this calculation is not meant to generate absolute numerical score for the repressive or activating activity of a CRR but is instead used to generate a rank-sorted list of CRR classes based on their expression modulation.

Basset analysis. Basset is a recently introduced method based on convolutional neural networks to accurately predict DNase I hypersensitive sites from DNA sequence, thus enabling annotation of the influence of mutations on accessibility[44]. We trained the Basset deep convolutional neural network on DNase I hypersensitive sites from 164 cell types mapped by ENCODE and the Roadmap Epigenomics projects. From this, we predicted the influence of variants on the presence of DNase hypersensitivity in each cell type by computing the difference between predictions on sequences with each allele. Candidate high impact variants were further analyzed for the ability to interrupt known binding sites by converted Basset-learned first convolution layer filters to probabilistic position weight matrixes by counting nucleotide occurrences in the set of sequences that activate the filter to a value that is more than half of its maximum value. We identified the likely binding protein for the motifs by querying the CIS-BP database[46] (accessed on 12 June 2015) using the TomTom v4.10.1 search tool[47] and requiring an FDR $q < 0.1$.

Code availability. All code can be requested by contacting M.C.S.

Data availability. All data used in this analysis were downloaded from the International Cancer Genome Consortium (IGCG) data portal (https://dcc.icgc.org/projects/). At our last date of access (11 February 2015), simple somatic mutations (SSM) for 405 pancreatic ductal adenocarcinoma samples were available from the Australian (PACA-AU) and Canadian (PACA-CA) groups. We downloaded the clinical data, SSMs and, when available, sequence-based gene expression (EXP-S) data for all 405 patients.

46. Weirauch, M.T. *et al.* Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
47. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. & Noble, W.S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).