Title: Somatic protospacer adjacent motifs are numerous and selectively targetable in cancers

Authors: Selina Shiqing K. Teh[1†], Kirsten Bowland[1†], Alexis Bennett[1], Eitan Halper-Stromberg[1], Alyza Skaist[2], Aparna Pallavajjala[1],Jacqueline Tang[1], Fidel Cai[1], Antonella Macoretta[1], Hong Liang[1], Sarah Wheelan[2,] ‡, Ming-Tseh Lin[1], Ralph H. Hruban[1,2], Robert B. Scharpf[2], Nicholas J. Roberts[1,2], and James R. Eshleman[1,2]

Affiliations:

[1]Department of Pathology, The Sol Goldman Pancreatic Cancer Research Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA
[2]Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, The Johns Hopkins University School of Medicine, Baltimore, MD, USA

[†]These authors contributed equally to this work.
[‡]Current address: Scientific Review Branch, National Human Genome Research Institute; Bethesda, MD 20817, USA

Running title (49/60 chr)*:* Somatic PAMs are common and targetable in cancers

Keywords: Protospacer adjacent motif (PAM), single base substitution (SBS), CRISPR-Cas9, single guide RNA (sgRNA), gene therapy.

Additional information

- Financial support: The Stringer foundation (JRE), Susan Wojcicki and Dennis Troper (JRE), The Sol Goldman Pancreatic Cancer Research Center (JRE), National Institutes of Health grant P50CA62924 (Dr. Alison P. Klein), National Institutes of Health grant P30CA006973 (Dr. William G. Nelson).
- Corresponding author: James R. Eshleman, MD, PhD. Department of Pathology and Oncology, 1550 Orleans Street, Suite 344, Baltimore, MD, 21231. 410-955-3511 (phone), 410-614-0671 (fax). jeshlem@jhmi.edu.
- Conflict of interest disclosures: Kirsten Bowland and Drs. Teh, Roberts and Eshleman and Johns Hopkins University have filed a provisional patent with the USPTO. Johns Hopkins University owns equity in Delfi Diagnostics. Dr. Scharpf is a founder of and holds equity in Delfi Diagnostics. He also serves as the Head of Data Science. This arrangement has been reviewed and approved by the Johns Hopkins University in accordance with its conflict of interest policies. The rest of the authors declare no conflict of interest.
- Word count: 3294; number of figures: 2; number of table: 1.

Abstract (136/150 words)

Somatic mutations in malignant cells genetically distinguish malignant cells from normal cells. We attempted to determine the type of somatic mutation in cancers that would create the largest number of novel CRISPR-Cas9 target sites. Whole genome sequencing (WGS) of three pancreatic cancers revealed that single base substitutions, mostly in noncoding regions, produced the highest number of novel NGG protospacer adjacent motifs (PAMs; median=494) compared to structural variants (median=37) and single base substitutions in exons (median=4). Using our optimized PAM discovery pipeline, we detected large numbers of somatic PAMs (median=1127/tumor) in 587 individual tumors from ICGC via WGS analyses of different tumor types. Finally, we showed that these PAMs, which were absent in patient-matched normal cells, could be exploited for cancer-specific targeting, with >75% selective cell killing in mixed cultures of human cancer cell lines using CRISPR-Cas9.

Statement of significance (34/50 words)

We developed a highly efficient somatic PAM discovery approach and found that a large number of somatic PAMs exist in individual tumors. These PAMs could serve as novel targets to selectively kill cancer cells.

Introduction

Somatic mutations accumulate as we age and genetically distinguish malignant cells from normal cells(1,2). More specifically, when a normal cell transforms into a cancer initiating cell and starts to clonally expand, the somatic mutations already present in the cancer initiating cell are, in theory, propagated during tumorigenesis and should be present in all resultant neoplastic cells(3,4). Indeed, most mutations in cancers are found within the noncoding regions(5) which make up of 98% of the human genome, and these “passenger mutations” are traditionally thought to be nonconsequential to the cell fitness(6).These somatic mutations genetically set the malignant cells apart from the rest of the normal cells, and could be exploited as anticancer therapeutic targets. This prompted us to invent a strategy to turn this vast amount of supposedly inconsequential mutations into targets of therapeutic value.

CRISPR-Cas9 is a highly programmable endonuclease or "molecular scissor"(7–9). It can be designed to induce double strand breaks (DSBs) at specific locations in the human genome, and in theory, can be highly selective. Several papers have reported cytotoxicity associated with multiple Cas9-induced DSBs(10,11), and have either developed various methods to circumvent the toxicity to increase the success rate of gene editing(12–14), or exploited this toxicity to achieve cell-specific killing(15). However, CRISPR-Cas9 is known to tolerate mismatches at the sgRNA target regions, contributing to off-target effects(16–19).

Our approach is inspired by the bacterial adaptive immune system, in which protospacer adjacent motifs (PAMs) have been evolutionarily selected to differentiate between host and viral DNA, which otherwise contain the exact same sequence, to

achieve selective elimination of the pathogen while leaving the bacterial host intact(20). The 3-5 nucleotide PAM sequence serves as a binding signal for Cas9, in which Cas9 would not bind nor cut the target in its absence(21,22), hence significantly decreasing the risk of off-targeting. Thus, we determined the type of somatic mutations in cancers that would create the largest number of novel CRISPR-Cas9 target sites and developed a highly efficient bioinformatics pipeline to identify single base substitutions (SBSs) that form novel PAMs. We hypothesized that we could identify a large number of novel PAMs in different tumor types, and these novel PAMs could serve as cancer-specific targets to selectively kill the malignant cell population.

Results

Development of PAM discovery approach

At the initial conception of this study, we tested two approaches that had the potential to lead to highly selective target cell killing with minimal off-target risk. We used whole genome sequencing (WGS) data from three pancreatic cancer (PC) cell lines and their corresponding normals, as PC is one of the most lethal cancers with a dismal five-year survival rate of only 11.5%(23). We selected the *S. pyogenes* NGG PAM due to its smaller PAM size(20).

We first considered structural variants (SVs), since they could juxtapose a new target DNA sequence next to an existing NGG PAM (Figure 1A). This could theoretically decrease the risk of off-target effects, as the resulting breakpoint is significantly different from the original sequence in the human genome (see Supplementary methods). Single nucleotide polymorphism (SNP) microarray analysis was previously performed on these

three cell lines to identify a small set of SVs (table S1). We used an SV detection software, Trellis(24), to identify SVs comprehensively from WGS data. We confirmed an average of 35 SVs per cell line using the Integrated Genome Viewer (IGV) by comparing tumor to normal, and validated 84.9% of them by PCR amplification across the breakpoint and Sanger sequencing (Table 1, Figure 1A, table S1). We found an average of 22 novel SVs juxtaposed next to an existing PAM per cell line (Table 1, Figure 1A, table S1). Using our sgRNA selection criteria (see Supplementary methods), we obtained an average of 17 good sgRNAs per cell line (Table 1, table S1).

Next, we attempted to discover novel PAMs created from SBSs (Figure 1B). We applied liberal selection criteria to capture an average of 44,019 SBSs per cell line and inspected them on IGV manually by comparing tumor to normal (Table 1, table S2). We identified an average of 488 mutations creating novel PAMs per cell line (Table 1, Figure 1B, table S2). To exclude mutations that might be subclonal or have arisen through *in vitro* culture, we used a variant allele frequency (VAF) cutoff of 30%, which resulted in an average of 345 PAMs per cell line (table S2). 64.4% of these somatic PAMs were located in intergenic regions, while 0.5% were located in exons (Figure 1C, figure S1A-B).

For initial functional testing of sgRNAs, we selected novel PAMs with VAFs >95% (average = 59) as targeting them would, in theory, produced the highest toxicity; and of them, an average of 33 good sgRNAs could be designed using our sgRNA selection criteria (Table 1, table S2). We were able to confirm all the qualifying mutations, except two, using Sanger sequencing (Table 1, table S2). A similar approach using whole exome sequencing (WES) data failed to yield sufficient targetable sites (average = 4;

Table 1, table S3). Thus, we concluded that the WGS-based PAM discovery approach using SBSs was more productive than the SV and WES approaches, and provided hundreds of novel PAMs per cancer as potential CRISPR-Cas9 target sites.

After careful screening and identification of novel PAMs in our three discovery PC cell lines, we developed the software, PAMfinder to detect the prevalence of somatic PAMs in other PC samples and tumor types and to test the generalizability of our PAM discovery approach in a larger tumor population. We compared novel PAMs discovered by PAMfinder with the results obtained from the cell lines used for discovery and found a high concordance between both approaches (Figure 1D, table S4, figure S2), with an average Jaccard Index of 68.7% using a VAF cutoff of 30%, and 82.9% using a VAF cutoff of 95%, demonstrating that the concordance increased as a function of the VAF. The main reasons for discordance of the PAMs identified were differences in read depths and VAFs of the novel PAMs reported by each approach. Based on the PAMfinder results, we measured the proportion of SBSs forming novel PAMs (% PAM) and found an average of 9.2% of SBSs created novel PAMs (table S4). Overall, we found that the PC cell lines contained a large number of novel PAMs, approximately 400 novel PAMs per sample.

High prevalence of novel PAMs in different tumor types

To determine the prevalence of novel PAM in different tumor types, we analyzed VCFs from the ICGC Data Portal(2) using PAMfinder and identified a large number of PAMs in lung cancers (LUCA-KR), esophageal cancers (OCCAMS-GB), and additional PCs (APGI-AU and PACA-CA). To briefly describe the data in these VCFs, WGS data

were aligned to GRCh38 reference genome to produce aligned CRAM files, and these CRAM files were processed through the GATK Mutect2 variant calling(25) workflow as tumor/normal pairs to identify somatic base substitutions. As the WGS on tumors were performed on primary tumor samples, we lowered the VAF cutoff of PAMfinder to account for differences in tumor neoplastic cellularity. We varied the cutoffs and found the majority of mutations detected using a cutoff of 20% were confirmed in IGV. Overall, we found that among the large number of base substitutions (average = 14249/tumor) identified in 587 tumor samples, an average of 14.1% of them produced novel PAM (average = 2394 PAM/tumor; Figure 1E-G, table S5). More specifically, cancers reported in the lung cancer project (LUCA-KR) had an average of 10.0% PAM-generating base substitutions (N=29), those reported in the esophageal cancer project (OCCAMS-GB) had an average of 16.8% (N=386), and the cancers in the two PC projects had an average of 8.9% (APGI-AU; N=45) and 8.7% (PACA-CA; N=127) (table S5).

The average number of base substitutions (4443 and 5745), somatic PAMs (439 and 569), and percentage of PAMs generated (8.9% and 8.7%) that derived from the two PC projects (APGI-AU and PACA-CA) were comparable to the rates obtained from our three discovery cell lines (average = 4548 SBSs, 417 PAMs, 9.2% PAMs; table S4-5), suggesting the applicability of our findings to the general population of PC tumors. The number of base substitutions and somatic PAMs found in esophageal cancers and lung cancers were much higher than the ones found in PCs, most likely indicating different tissue-specific mechanisms of mutagenicity underlying these neoplasms. Interestingly, the total number of base substitutions and novel PAMs discovered did not

differ significantly between tumor samples that had not undergone treatment and those that had undergone chemotherapy, radiation therapy, or both (Mann-Whitney U tests, $P > 0.05$; figure S3). Thus, we concluded that somatic SBSs in the tumor types we examined yielded hundreds, if not thousands, of novel PAMs in each tumor, and these findings were dependent on tissue type.

Somatic PAMs were maintained in lymph node metastases

Next, we asked whether the target mutations identified in a primary tumor would be maintained in metastases from the patient. For the patient from whom we generated the cell line Panc504, we studied a 6 x 5 mm focus of cancer in one of the regional lymph nodes and documented the presence of all (29 out of 29) PAM generating mutations tested (Figure 2A). A second patient, from whom we generated the cell line Panc1002, had a very small focus (2 x 1 mm) of cancer in one lymph node. After careful macrodissection, we were able to demonstrate the presence of 3 out of 4 PAM generating mutations tested (Figure 2A). Archived material for the third patient (origin of Panc480) was unavailable. While available samples limited our analysis, our data showed that the majority of mutations that created novel PAMs in primary cancers were maintained in regional lymph node metastases.

Selective cell killing with CRISPR-Cas9

Finally, we tested the hypothesis that we could selectively target an individual patient's cancer using sgRNAs designed from our PAM discovery approach. To show proof-of-concept of CRISPR-Cas9 selectivity, we generated Cas9-expressing mouse

and human cell lines and documented Cas9 activity (figure S4A-B). Then, we seeded mouse-human cell line co-cultures, and transduced them with a multi-target sgRNA with 12 target sites in the human genome but none in the mouse genome (table S6). Using both flow cytometry and a human-mouse NGS assay (see Supplementary methods, figure S4C-D), we saw >95% reduction of the human cancer cells in different co-cultures (Figure 2B, figure S4E-F). The human-specific cell killing was dependent on both functional Cas9 and the human-specific sgRNA, showing that the CRISPR-Cas9 system is capable of selectively eliminating cancer cells (figure S4G).

To test selective targeting of a patient's cancer cells while leaving normal cells intact, we selected 7 of the 13 targets that we identified in Panc480 using our novel PAM discovery approach (Figure 1B), confirmed targeting efficiency of individual sgRNAs, and cloned the corresponding sgRNAs into a multiplex sgRNA expression vector (designated MT7; Figure 2C; table S7). After transduction into Panc480 Cas9-expressing cells, we detected cutting activity of all 7 sgRNAs, and not in its controls (Panc1002 Cas9-expressing cell line) and corresponding normal cells from the patient (Onc3286), by deep sequencing at the targeted loci (Figure 2D). As another negative control to check for potential Cas9 off-target activity, Panc1002 Cas9-expressing cells were seeded in cell culture and transduced with Panc480-MT7 which targets mutations unique to Panc480. We performed WGS before transduction (T0) and 14 days post-transduction of MT7 (T14). Using two independent approaches for objective assessment (see Supplementary methods), we found that the indels novel to T14 did not exhibit homology to any of the 7 sgRNAs in 480-MT7 (table S7-8). These indels,

present at low VAF, likely represent background heterogeneity in a bulk cell population or ongoing genomic instability.

We then co-cultured Panc480-Cas9-mApple cells along with Panc10.05-Cas9-EGFP cells and transduced them with MT7. Flow cytometry showed >80% selective reduction of Panc480 cells on day 21 (Figure 2E; paired t test, $P = 0.003$), and this finding was corroborated with STR profiling (Figure 2F; paired t test, $P = 0.03$). Although selective reduction was also seen in Panc480 parental cell line lacking Cas9 (Figure 2F; paired t test, $P = 0.009$), the magnitude of reduction in the presence of Cas9 was larger (76.4% vs 59.6%). This suggests the MT7 expression vector itself was somewhat toxic, but that functional Cas9 was needed to produce the full observed toxicity (Figure 2E-F). Our results demonstrated that the sgRNAs designed via PAM discovery approach were able to yield significant cell death of targeted cells.

Discussion

We propose a highly efficient cancer-specific PAM discovery approach that identifies novel PAMs generated from base substitutions with high confidence. The large number of somatic PAMs discovered in each tumor creates an opportunity for cancer-specific targeting, and novel PAMs that arise during tumorigenesis could create more targets for serial treatment. As cancer is a clonal disease, the distinct set of mutations found in the cancer initiating cell should be found in all primary tumor and metastatic sites, thus making our approach a potential solution to multi-site cancer killing. Since point mutations increase as a function of age(1,2), pediatric tumors might not contain as many somatic PAMs as adult patients. Studies on increasing toxicity with

a small number of CRISPR-Cas9 cuts would be essential to assess the applicability of this approach on pediatric patients. Besides, delivery of CRISPR-Cas therapeutics continues to be an active work of progress, with *in vivo* efficacy demonstrated in clinical settings(35).

As our approach exploits the vast number of novel PAMs located in noncoding regions, it requires WGS analyses of both tumor and normal. This becomes a limiting factor to study the applicability of our approach in many different tumor types, as most published data only includes exonic sequencings, and obtaining corresponding normals for each tumor sample is still not a standard practice in many studies. WGS is still expensive and analyses are more computationally demanding than for WES, but the exponential decrease in costs(26) and enhancement of computing power have made these issues less concerning(27). Generating cell lines from primary tumor samples and corresponding normals continues to be challenging; however, this has become much more routine(28–31). We have many cell lines generated from tumors and normals of PCs, which provide us unique opportunities to study them extensively in an *in vitro* setting. This is particularly helpful for tumors with low cellularity (such as PCs(32)), as the detection of mutations in these tumors would be impeded by the limit of detection of most next generation sequencing (NGS) assays. Many studies either utilized laser-based microdissection or flow sorting to increase tumor cellularity, or eliminate samples with low tumor cellularity from their study cohorts, which could potentially lead to sampling bias.

Moreover, as each tumor sample differs in neoplastic cellularity, the VAF cutoff for mutation detection should ideally be customized to each sample in order to decrease

the risk of calling variants that are false positives or false negatives. However, the current clinical workflow in measuring tumor cellularity revolves around two methods: (1) pathology review, which is prone to errors contributed by factors such as sampling error, irregular tumor cell shapes, and heterogeneous stromal cell contamination(33), or (2) quantification of mutant alleles by sequencing, which requires prior knowledge of driver mutations of the cancer type(34). A more robust method to quantify tumor cellularity has been published(33), but it requires SNP microarray data as input. An alternative approach to quantify tumor cellularity directly from NGS data generated from current clinical pipelines would be of importance to detect mutations with high confidence.

Although PAM-finding approaches have been published in a few studies(36,37), our approach is cancer- and, more importantly, patient-specific. Our approach presents a unique opportunity as a new precision medicine-based therapeutic tool that possesses the specificity of a targeted therapy, but without the restriction of a targetable protein. Although more work has to be done to determine the optimal approach of selecting sgRNAs, we envision applicability of our approach to a broad range of cancers.

Acknowledgements

We thank the generosity of the patients who consented to providing their samples. We also acknowledge Drs. Leslie Cope, Chien-Fu Hung, Elizabeth Jaffee, Lei Zheng, Feyruz V. Rassool, Richard Burkhart, Christine Iacobuzio-Donahue, Jessica Gucwa, Qingfeng Zhu, Shiwen Peng, and Toni Seppulla for helpful discussions. We thank Ada Tam, Raluca Yonescu, Emily Adams, Lisa Haley, Anuj Gupta, Lee Blosser, Nicholas Ionta, Suping Chen, Kai Calder, Jormoh Bellepu, Jordyn Winters, and Nori Leybengrub for outstanding technical assistance.

Author contributions:

Conceptualization: SSKT, KB, NJR, JRE

Methodology: SSKT, KB, EHS, SW, MTL, RBS, NJR, JRE

Software: SSKT, KB, EHS, AS, AP, SW, RBS, NJR, JRE

Investigation: SSKT, KB, AB, AS, AP, JT, FC, AM, HL, NJR, RBS

Funding acquisition: RHH, JRE

Project administration: JRE

Supervision: SW, RHH, RBS, NJR, JRE

Writing – original draft: SSKT, KB

Writing – review & editing: SSKT, KB, RHH, NJR, JRE

Methods

WGS target validation and sgRNA design

DNA from tumors and corresponding normals for Panc480, Panc504, and Panc1002 were whole genome sequenced and variants called as previously described(28,38). Putative somatic mutations with a quality score of "PASS", a distinct coverage (DP) > 10, and a genotype quality score (GQ) > 20 were identified using BEDTools(39). Somatic mutations were annotated with region-based (Func.refGene) and gene-based (Gene.refGene) identifications using ANNOVAR(40). Flanking sequences 2 base pairs 5' and 3' to somatic mutation positions were obtained from UCSC table browser(41). The following inclusion criteria were implemented: (1) novel Cs that were adjacent to an existing C, or novel Gs that were adjacent to an existing G; (2) VAF of at least 5% in tumor; (3) a minimum of 18X read depth(42) in both normal and tumor. These mutations were then visually inspected and confirmed on IGV. Somatic mutations with VAF >95% were chosen to put through CRISPOR(43). Somatic mutations that produced sgRNAs with >50 specificity score in CRISPOR were subsequently validated by PCR and Sanger sequencing (Primers Table 1).

PAMfinder

A perl script was written to process VCFs based on their genome builds (hg19 or hg38) to identify somatic variants that produced novel PAMs. Tumor (arrayT) and normal (arrayN) were specified based on column number, read depth were set at 18X(6), and VAF cutoff could be modified based on the purpose of the analysis. For somatic variants that passed through the read depth and VAF filters, the 5' and 3'

genomic sequences flanking the somatic variants were obtained from the FASTA of individual chromosomes to inspect whether novel Cs were adjacent to an existing C or novel Gs were adjacent to an existing G. The output contained information about the somatic variant, the potential sgRNA sequence along with the novel PAM, and specified whether the novel PAM was located on the plus or minus strand of the genome. Script is available on https://github.com/selinateh/PAMfinder.

PAM discovery on ICGC samples

VCFs containing raw SNV calls from WGS data via the GATK Mutect2 variant calling workflow were downloaded from the ICGC-ARGO Data Portal(44). These VCFs were sourced from four projects: APGI-AU (Australian Pancreatic Cancer Genome Initiative; n=45), LUCA-KR (Personalised Genomic Characterisation of Korean Lung Cancers; n=29), PACA-CA (Pancreatic Cancer Harmonized “Omics” analysis for Personalized Treatment; n=127), and OCCAMS-GB (Oesophageal Cancer Clinical and Molecular Stratification; n=386). Clinical data corresponding to each patient was also downloaded. VCFs were put through SBSfilter.pl and PAMfinder.pl (see Supplementary methods) to identify base substitutions that produced novel PAMs. VAF > 0.2 was used as a cutoff for two reasons: (1) the decreased value from 0.3 to 0.2 was to account for various degrees of tumor cellularity, and (2) we attempted a range of VAF cutoffs (0.1-0.25) and found that the mutations were called at the highest confidence with a VAF > 0.2 based on IGV inspection. Only one sample per patient was included. % PAM was calculated by dividing the number of novel PAM by the total number of base substitutions.

Confirmation of PAMs in regional lymph nodes

FFPE preserved lymph nodes for Panc1002 and Panc504 were cored and DNA extracted using QIAamp DNA Mini Kit (QIAGEN). Novel PAMs previously discovered in WGS of the primary tumor cell lines were PCR amplified (Primers Table 1) and Sanger sequenced. Sequence traces were compared to Sanger of the tumor cell line and patient-matched normal DNA to confirm the presence or absence of the mutation leading to the novel PAM.

Co-culture assays

Cells that expressed either mApple or EGFP fluorescence were co-cultured at different ratios. Proportion of mApple-expressing cells post-transduction of sgRNAs were measured at different time points using Attune NxT Flow Cytometer (ThermoFisher). FCS Express 7 (De Novo Software) was used to analyze the flow cytometry data.

CRISPR multiplex plasmid functional testing

To test the efficacy of multiplex CRISPR arrays expressing multiple sgRNA cassettes, the targeted cell line Panc480 was transduced at a 10:1 MOI with lentivirus expressing a non-targeting sgRNA (NT) or the multiplexed CRISPR array in a lentiGuide-puro backbone. 14 days after transduction and selection with puromycin, cells were harvested and gDNA extracted. The targeted loci were PCR amplified (see “Panc480 mutation validation primers” under Primers Table 1) with NGS adaptors and

sent for amplicon sequencing. The sequencing data was analyzed for the percent of edited reads by CRISPResso2(45). Functional testing was performed in parallel for a non-targeted cell line, Panc1002, and a patient-matched EBV lymph normal cell line for Panc480, Onc3286.

STR analysis

Mixed human DNA samples were PCR amplified using the AmpFLSTR Identifiler PCR Amplification Kit that amplifies 15 microsatellites (Applied Biosystems, Foster City, CA) per manufacturer's instructions, and amplicons resolved on a 3130 capillary electrophoresis instrument (Applied Biosystems). Percentage of a given individual was calculated from on-scale informative peak heights using Chimeranalyzer (https://github.com/young-jon/chimeranalyzer).

Statistical analysis

The appropriate statistical tests were performed in GraphPad Prism (Version 9.2.0). The statistical models used were stated in results and legends of figures. For all statistically significant results, * indicates $P<0.05$, ** indicates $P<0.01$, *** indicates $P<0.001$, and **** indicates $P<0.0001$.

Data availability

The authors confirm that the data supporting the findings of this study are available within the article and its supplementary materials. ICGC data that support the findings of this study are available in the Cancer Genome Collaboratory at

https://cancercollaboratory.org/. These data were derived from the ICGC ARGO platform. Plasmids constructed had been deposited at Addgene. Plasmids expressing various specific sgRNAs are available upon request from the corresponding author. Sequencing data will be uploaded to the recommended archive.

References

1. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538:260–4.
2. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578:82–93.
3. Nowell PC. The clonal evolution of tumor cell populations. Science. 1976;194:23–8.
4. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell. 1990;61:759–67.
5. Khurana E, Fu Y, Chakravarty D, Demichelis F, Rubin MA, Gerstein M. Role of non-coding sequence variants in cancer. Nat Rev Genet. 2016;17:93–108.
6. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A. 2013;110:1999–2004.
7. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337:816–21.
8. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, et al. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339:819–23.
9. Mali P, Yang L, Esvelt KM, Aach J, Guell M, DiCarlo JE, et al. RNA-guided human genome engineering via Cas9. Science. 2013;339:823–6.
10. Aguirre AJ, Meyers RM, Weir BA, Vazquez F, Zhang C-Z, Ben-David U, et al. Genomic Copy Number Dictates a Gene-Independent Cell Response to CRISPR/Cas9 Targeting. Cancer Discov. 2016;6:914–29.
11. Munoz DM, Cassiani PJ, Li L, Billy E, Korn JM, Jones MD, et al. CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer Discov. 2016;6:900–13.
12. Kuscu C, Parlak M, Tufan T, Yang J, Szlachta K, Wei X, et al. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. Nat Methods. 2017;14:710–2.
13. Smith CJ, Castanon O, Said K, Volf V, Khoshakhlagh P, Hornick A, et al. Enabling large-scale genome editing at repetitive elements by reducing DNA nicking. Nucleic Acids Res. 2020;48:5183–95.
14. Niu D, Wei H-J, Lin L, George H, Wang T, Lee I-H, et al. Inactivation of porcine endogenous retrovirus in pigs using CRISPR-Cas9. Science. 2017;357:1303–7.
15. Kwon T, Ra JS, Lee S, Baek I-J, Khim KW, Lee EA, et al. Precision targeting tumor cells using cancer-specific InDel mutations with CRISPR-Cas9. Proc Natl

Acad Sci U S A [Internet]. 2022;119. Available from: http://dx.doi.org/10.1073/pnas.2103532119

16. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, et al. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31:822–6.

17. Doench JG, Fusi N, Sullender M, Hegde M, Vaimberg EW, Donovan KF, et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol. 2016;34:184–91.

18. Anderson EM, Haupt A, Schiel JA, Chou E, Machado HB, Strezoska Ž, et al. Systematic analysis of CRISPR-Cas9 mismatch tolerance reveals low levels of off-target activity. J Biotechnol. 2015;211:56–65.

19. Lin Y, Cradick TJ, Brown MT, Deshmukh H, Ranjan P, Sarode N, et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. Nucleic Acids Res. 2014;42:7473–85.

20. Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology. 2009;155:733–40.

21. Anders C, Niewoehner O, Duerst A, Jinek M. Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature. 2014;513:569–73.

22. Sternberg SH, Redding S, Jinek M, Greene EC, Doudna JA. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature. 2014;507:62–7.

23. Cancer of the Pancreas - Cancer Stat Facts [Internet]. SEER. [cited 2023 Feb 7]. Available from: https://seer.cancer.gov/statfacts/html/pancreas.html

24. Papp E, Hallberg D, Konecny GE, Bruhm DC, Adleff V, Noë M, et al. Integrated Genomic, Epigenomic, and Expression Analyses of Ovarian Cancer Cell Lines. Cell Rep. 2018;25:2617–33.

25. Van der Auwera GA, O'Connor BD. Genomics in the Cloud. O'Reilly Media, Inc.

Table 1. Number of sgRNAs designed and confirmed using three different methods.

Method	Cell line	No. of somatic mutations	No. of novel PAMs*	No. of good sgRNAs#	No. of Sanger-validated good sgRNAs**
SV	Panc480	38	24	17	17
	Panc504	37	18	15	15
	Panc1002	31	25	18	18
WGS	Panc480	44311	494	13	13
	Panc504	38881	531	48	47
	Panc1002	48866	440	38	37
WES	Panc480	44	7	2	NA
	Panc504	38	1	0	NA
	Panc1002	30	4	0	NA

*For SV approach, the values indicate number of novel junctions flanked by an NGG sequence in which breakpoint sequence has been validated through Sanger sequencing. For WES and WGS approaches, novel PAM indicates a single base substitution of NGN/NNG sequence to NGG (both + and - strands). Only sites with a variant allele frequency (VAF) of at least 5% in tumor and a minimum of 18X read depth in both germline and tumor are counted. Each site was visually inspected and confirmed in IGV.

#"Good sgRNA" is defined as sgRNAs that have >50 specificity score (prediction of how much the sgRNA sequence may lead to off-target cleavage) in CRISPOR. It includes sgRNAs that are inefficient (low knockout frequencies). For SVs all VAFs included. For WES and WGS, only novel PAMs with VAF >95% included.

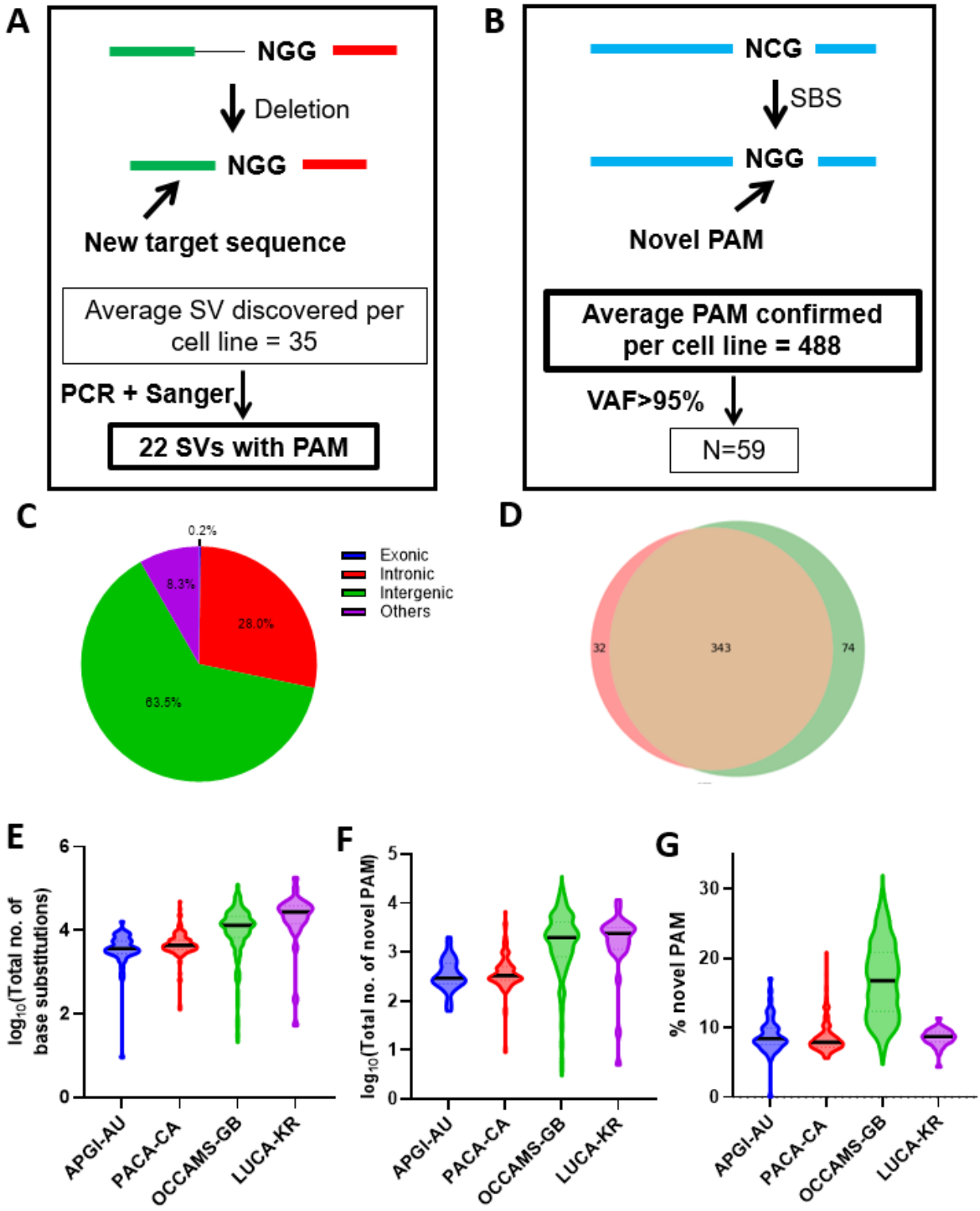
**For WES, Sanger sequencing wasn't performed due to low number of good sgRNAs.

Figure Legends

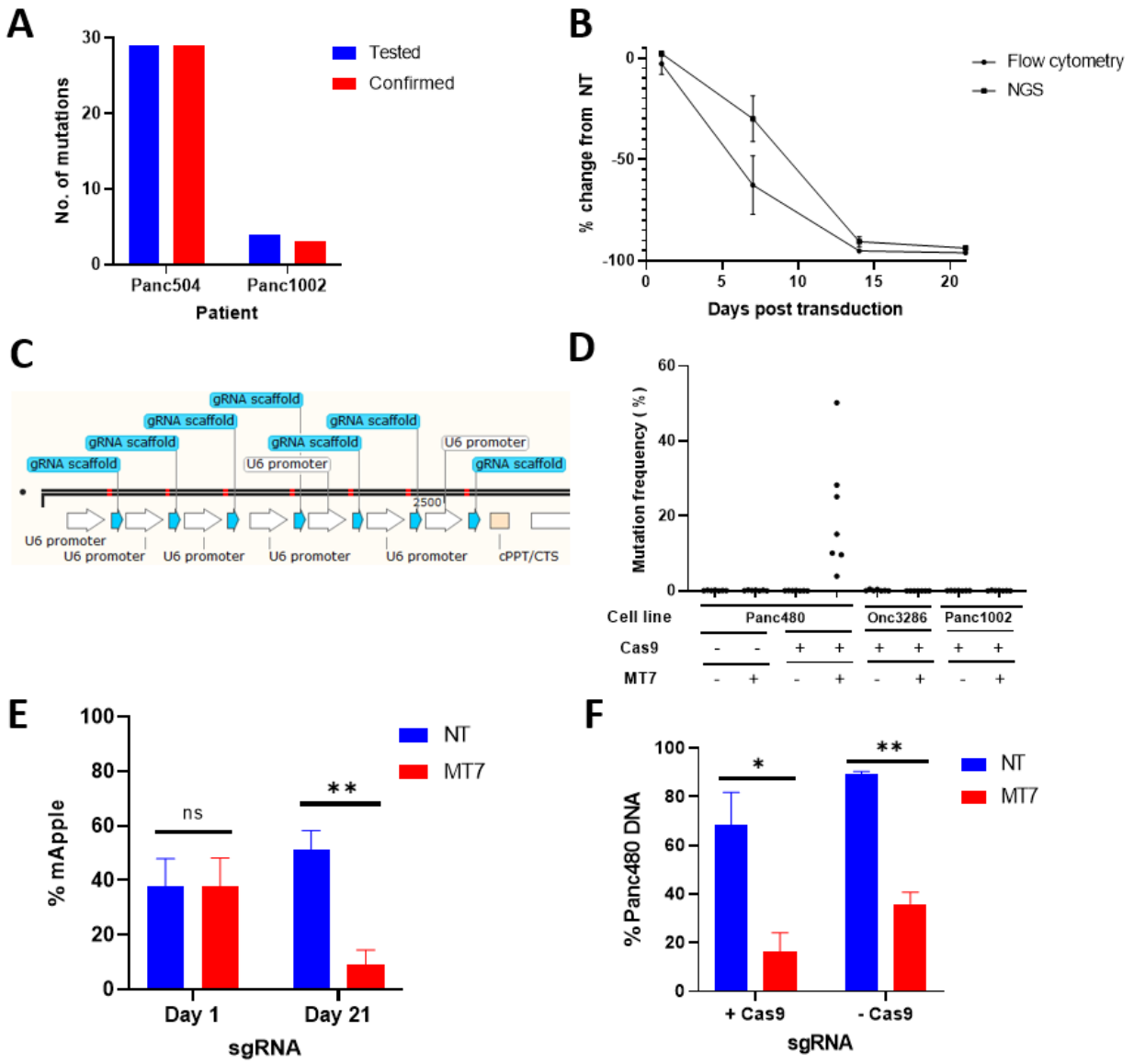
Fig. 1. Large numbers of somatic PAMs were found per tumor. (A-B) Two approaches were developed to identify and design sgRNAs based on novel breakpoints and somatic PAMs found in three PC cell lines. Numbers shown are the averages of three PC cell lines used to develop each approach. (A) Structural variant (SV) discovery approach. For example, with a homozygous deletion, novel DNA sequences (green) are juxtaposed next to a pre-existing NGG PAM. SVs could also theoretically generate a novel NGG (not shown). (B) Single base substitution-based PAM discovery approach. For example, a C→G base substitution mutation can create a novel PAM. (C) Proportion of novel PAMs discovered in Panc504 that were located in different regions of the genome. Others include ncRNA, UTRs, and 1-kb regions upstream/downstream of transcription start/end sites. N=375. (D) Concordance of the number of PAMs discovered in Panc504 between the approach used for discovery (in red) and PAMfinder (in green) was high, with a Jaccard index of 76.4%. The numbers indicate the number of PAMs in each set. (E-G) Frequency of somatic PAMs was examined from WGS analyses of different tumor types: APGI-AU (PCs; N=45), PACA-CA (PCs; N=127), OCCAMS-GB (esophageal cancers; N=386), and LUCA-KR (lung cancers; N=29). Truncated violin plots present the total number of (E) base substitutions (log scale), (F) novel PAMs (log scale), and (G) percentage of base substitutions that contributed to novel PAM in each tumor sample. Detailed statistics could be found in table S5.

Fig. 2. Selective cell killing was achieved with low number of targets discovered from our novel PAM approach. (A) Novel PAMs arising from mutations in two primary tumors were confirmed of their presence in metastatic sites via Sanger sequencing. (B) Co-cultures of Cas9-expressing human PC (Panc10.05) and mouse fibroblast (NIH 3T3) cell lines transduced with human-specific 230F(12) sgRNA were monitored over time using flow cytometry and a human-mouse polymorphism NGS assay. Error bars indicate mean± SEM; N=3. (C) Tandem CRISPR array with U6 promoter, sgRNA sequence (red line), and sgRNA scaffold targeting 7 novel PAMs in the Panc480 cell line. Diagram was generated by SnapGene. (D) Mutation frequency at 7 Panc480-specific target sites in parental Panc480, Cas9-expressing Panc480, Panc480 patient's Cas9-expressing lymphoblasts (Onc3286), and Panc1002 (negative control) cell lines after treatment with NT (-) or MT7 (+) multiplex sgRNA vector. (E) Flow cytometry analysis of Panc480-Cas9-mApple and Panc10.05-Cas9-EGFP cell mixtures after treatment with NT or MT7 on day 1 and day 21 post transduction of sgRNAs. Paired t tests were performed; ns indicates $P > 0.05$; ** indicates $P < 0.01$. Error bars indicate mean ± SEM; 3 biological replicates with 2 technical replicates each. (F) STR analysis of Panc480 (parental)/Panc10.05-Cas9-EGFP (-Cas9) or Panc480-Cas9-mApple/Panc10.05-Cas9-EGFP (+Cas9) cell line mixtures after treatment with MT7 on day 21. Paired t tests were performed; * indicates $P < 0.05$; ** indicates $P < 0.01$. Error bars indicate mean ± SEM; 3 biological replicates with 2 technical replicates each for +Cas9, 1 technical replicate each for -Cas9.

Figure 1


A
NGG
Deletion
NGG
New target sequence
Average SV discovered per cell line = 35
PCR + Sanger
22 SVs with PAM
B
NCG
SBS
NGG
Novel PAM
Average PAM confirmed per cell line = 488
VAF>95%
N=59
C
0.2%
8.3%
28.0%
63.5%
Exonic
Intronic
Intergenic
Others
D
32
343
74
E
log10(Total no. of base substitutions)
APGI-AU
PACA-CA
OCCAMS-GB
LUCA-KR
F
log10(Total no. of novel PAM)
G
% novel PAM

1 **Figure 2**


A
No. of mutations
30
20
10
0
Tested
Confirmed
Panc504
Panc1002
Patient
B
% change from NT
0
-50
-100
0
5
10
15
20
Days post transduction
Flow cytometry
NGS
C
gRNA scaffold
U6 promoter
2500
cPPT/CTS
D
Mutation frequency (%)
60
40
20
0
Cell line
Panc480
Onc3286
Panc1002
Cas9
MT7
E
% mApple
100
80
60
40
20
0
ns
**
NT
MT7
Day 1
Day 21
sgRNA
F
% Panc480 DNA
100
80
60
40
20
0
*
**
NT
MT7
+ Cas9
- Cas9
sgRNA

2

Supplementary Data

Supplementary Figure S1-5

Supplementary Table S1-8

Supplementary Materials and Methods

Primers Table 1-7

References (26-45)