CrisPam: SNP-derived PAM analysis web tool and human pathogenic SNPs database for CRISPR allele-specific targeting

Roy Rabinowitz

Tel Aviv University

Roy Darnell

Tel Aviv University

Daniel Offen (✉ danioffen@gmail.com)

Tel Aviv University

Software

Keywords: CrisPam, Allele-specific, SNP-derived PAM, CRISPR tool, pathogenic SNP, SNP targeting, SNP-specific

Posted Date: April 30th, 2019

DOI: <https://doi.org/10.21203/rs.2.9413/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background CRISPR is a promising novel technology for treating genetic conditions. Therefore, it is essential to further develop and promote treatment's safety and specificity. While the guide-RNA offers position-specific DNA targeting, it may tolerate small changes such as single-nucleotide polymorphisms (SNPs). To that end, an allele-specific targeting approach is in need for future treatments of heterozygous patients, suffering from genetic conditions caused by a SNP. The SNP-derived PAM approach allows highly allele-specific DNA cleavage by incorporating a protospacer adjacent motif (PAM) sequence only at the target allele. Description Here we present CrisPam, a tool that detects SNP-derived PAMs for allele-specific targeting by the CRISPR/Cas system. The algorithm scans the generation of each reported PAM for a given DNA sequence and its variations. A successful result is such that at least one PAM is generated by a SNP. Thus, the PAM shall be part of the variant allele only and the Cas protein will therefore be able to exclusively bind the variant allele for gene-editing, while the wildtype allele remains unchanged. Conclusion CrisPam is available online for researchers and also offers access to the CrisPamDB, a database that contains the CrisPam analysis for any reported pathogenic SNP in humans.

Background

The clustered regularly interspaced short palindromic repeat (CRISPR) system enables precise genome editing mediated by a single-guide RNA (sgRNA) that guides the CRISPR associated (Cas) protein to the target DNA in the genome. Cas9, the catalytic unit of the CRISPR system, generates a double-strand break (DSB) in the DNA in the presence of a DNA:sgRNA match and a protospacer-adjacent motif (PAM) in immediate proximity to the target DNA^{1,2}. The diverse of Cas proteins, derived from different bacterial strains, differ in several properties such as PAM sequence, cleavage pattern and position, size, activity in mammalian cells, off-targets and substrate (DNA or RNA). The standard Cas protein has been modified to broaden its applications to base-editing^{3,4}, transcription repression and activation^{5–7}, epigenomic modifications⁸, visualization of genomic loci⁹ and DNA nicking¹⁰ (single-strand cleavage). In an experiment design, the PAM sequence and size of the designated Cas should be taken under consideration; presence of a PAM is a limiting step in targeting unique loci, and the Cas size affects the optional possibilities of delivery systems.

SNP-derived PAM

The CRISPR/Cas system can tolerate some mismatches between the CRISPR RNA (crRNA) and the target DNA. The bases at the positions of 8 to 13 at the 3' end of the spacer (regarding type II Cas proteins) are termed the seed sequence along with the first base at the 5' end. Mismatches at the seed sequence are thought to be not tolerated and abolish DNA cleavage. As for pathogenic single-nucleotide polymorphisms (SNPs), previous studies have shown that targeting an allele caused by a SNP by choosing a gRNA sequence containing the variated nucleotide is seemingly insufficient, resulting in a non-specific knockdown of both the mutant alleles and the wildtype allele in some proportion^{11,12}. A SNP-derived PAM approach overcomes this potential limitation of targeting the disease-causing allele

while leaving the wildtype allele intact. This method dramatically increases the specificity of targeting the mutant allele alone by choosing a PAM sequence that is present only at the mutant sequence. Meaning, the mutant SNP generates the PAM sequence^{12,13}.

When targeting a gene without a particular DNA cleavage location preference, almost all Cas proteins are optional. However, when targeting a SNP in general, or if utilizing the SNP-derived PAM approach in particular, the selection of Cas is limited mostly due to the condition of PAM presence in proximity to the SNP or having a PAM generated by the SNP.

CrisPam

CrisPam is a pythonic code that scans DNA sequences for 30 candidate PAMs from 19 Cas proteins (Table 1). It obtains data of a given SNP, and tests whether it generates a unique PAM sequence in the DNA of the mutation allele only. Thus, CrisPam generates a list of matching Cas proteins for targeting the pathogenic allele. Here we show a database of all known clinically significant pathogenic or likely pathogenic SNPs that generate a PAM. Furthermore, we developed a bioinformatics tool for researchers, available at <http://CRISPR.tau.ac.il>, to detect SNP-derived PAMs at their SNPs of interest regardless of taxonomy and clinical significance.

Due to technical limitations, Table 1 is only available as a download in the supplemental files section.

Implementation

The CrisPam tool is web-based. Thus, no software installation effort is required. The CrisPam DB is a .xlsx file and can be opened by Excel. The CrisPam script is written in Python 3.6 and uses standard libraries (xml.etree.ElementTree, csv and time). Biopython is used on the web-based tool²⁹.

Parsing and SNP data analysis

The guiding principle of the SNP-derived PAM concept is having a PAM present in the desired target allele. The following workflow occurs to detect unique PAMs generated by a SNP: Parameters are being parsed from the data (wildtype sequence, mutation sequence, gene name and ID, SNP ID and the chromosome) into a list of SNPs. The code analyses a given SNP by obtaining the DNA sequence upstream and downstream to the SNP, the wildtype nucleotide (reference nucleotide) and the variation nucleotide. The anti-sense strand is analysed to detect unique PAMs generated on the complementary strand as well. 16 PAM sequences of 14 Cas proteins are scanned for, in the DNA sequence (table 1). For each Cas, CrisPam is defined to find its PAM at the position of the SNP (figure 2). Some SNPs have more than one variation nucleotide, thus CrisPam considers any variation of a SNP and scans each one of them. Once a PAM is found, it is accepted as a match only if it exists at a variation allele and not at the wildtype allele. For a given SNP, more than one PAM may be generated, therefore, CrisPam presents all the matches for a given

SNP. The suggested sgRNA sequence for each matching Cas – is the 20-23nt upstream or downstream to the PAM, according to the Cas type (type II or type V, respectively).

The PAM sequences were determined according to previous studies that characterized the unique properties and PAM compatibility for each Cas1-2,14-28.

We obtained a database of all known pathogenic and likely-pathogenic SNPs in humans from NCBI's dbSNP (SNP database). The code is written to analyse dbSNP's data in XML format and each SNP that is found to be PAM generating is represented in a row of a CSV file.

Results

A database of PAM-generating SNPs:

The CrisPam algorithm scanned 49,634 pathogenic SNPs and 14,722 likely pathogenic SNPs (64,356 in total). Successful matches of SNPs that generate at least one PAM were found in 84% of the total SNPs – 41,162 of the pathogenic SNPs and 12,940 of the likely pathogenic SNPs (figure 1).

The SNP-derived PAM targeting approach is highly ideal for heterozygous patients suffering from a disease caused by a SNP. Figure 2 represents a study case SNP (rs63750526 of PSEN1) that generates 7 PAMs. Such SNPs confer the ability to opt the most suitable Cas depending on the application's limitations (vector size, activity efficiency, lab stock etc.).

The full database of PAM generating pathogenic and likely pathogenic SNPs is available at http://crispr.tau.ac.il/DBs/CrisPam_results.xlsx

The CrisPam algorithm is available at <https://github.com/ristlin/CrisPam>

CrisPam – an online SNP-derived PAM finding tool:

We established a web tool that performs CrisPam's SNP-derived PAM targeting abilities on user data. Since many SNPs are yet to be reported and included in NCBI's dbSNP, and for research purposes non-pathogenic SNPs may be of one's interest to target, our web tool offers a platform for researchers to enter their sequences of interest for CrisPam analysis.

CrisPam is available at <http://CRISPR.tau.ac.il>

Discussion

The SNP-derived PAM targeting approach for promoting allele specificity is a promising method in CRISPR based novel therapies to enter the clinic. As most patients suffering from genetic conditions are heterozygous, carrying one copy of a pathogenic allele, developing SNP customized treatments is essential for increasing treatment's safety, by reducing unintended cleavage of the well-functioning

wildtype allele. While many web tools offer gRNA designs for CRISPR based experiments, none of them, to our knowledge, offer an allele-specific gRNA design. Since CrisPam cannot offer scoring and off-targets assessment for now, we strongly suggest further off-target prediction examination of the gRNA of interest. Moreover, gRNA length may vary for different Cas proteins; thus, we strongly recommend using CrisPam as the first step in the experiment design. Further assessments of activity in target organism, gRNA length and off-targets prediction are required. Moreover, for multiple-PAM generating SNPs, considerations such as delivery vector capacity (e.g. AAV or lentivirus) and efficiency may also determine the most suitable Cas protein for the experiment. In the future more features will be added to CrisPam: alternative input options (txt files and rsID) and customized PAMs.

Conclusions

While CRISPR applications have been widely expanded, the SNP-derived PAM approach may be utilized for gene silencing (using inactive Cas), genetic screening and more applications other than allele-specific DNA cleavage. This study emphasizes the emerging importance of broadening PAM compatibility of Cas proteins to enable allele-specific targeting and overcome the PAM limitation. Furthermore, CrisPam offers a simple interface to design an allele-specific targeting experiment using the CRISPR/Cas system.

Availability And Requirements

Project name: CrisPam

Project home page: <http://CRISPR.tau.ac.il>

Programming language: Python 3.6

License: Free for academic end-users solely for non-commercial research purposes.

Any restrictions to use by non-academics: Contact Prof. Dani Offen or Ramot at Tel Aviv University LTD.

List Of Abbreviations

CRISPR Clustered regularly interspaced short palindromic repeat

sgRNA Single-guide RNA

Cas CRISPR associated (protein)

DSB Double-strand break

PAM Protospacer-adjacent motif

crRNA CRISPR RNA

Declarations

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: All data generated or analysed during this study are included in this published article. The SNP data that was used to generate the CrisPam DB for PAM-generating pathogenic \ likely-pathogenic SNPs was obtained from NCBI's dbSNP (<https://www.ncbi.nlm.nih.gov/snp>).

Competing interests: The authors declare that they have no competing interests.

Funding: Not applicable

Authors' contributions: Conceived and designed the study, analysed the data and wrote the paper: RR. Programing: RD and RR. Web integration: RD. Principle Investigator: DO. All authors read and approved the final manuscript

Acknowledgements: The authors would like to thank Oren P. Rabinowitz for his technical support and advice.

References

1. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* (80-.). **339**, 819–23 (2013).
2. Doudna, J. A. & Charpentier, E. The new frontier of genome engineering with CRISPR-Cas9. *Science* (80-.). **346**, 1258096–1258096 (2014).
3. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
4. Gaudelli, N. M. *et al.* Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature* **551**, 464–471 (2017).
5. Gilbert, L. A. *et al.* Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell* **159**, 647–61 (2014).
6. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. *Nature* **517**, 583–588 (2015).

7. Bikard, D. *et al.* Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. *Nucleic Acids Res.* **41**, 7429–37 (2013).
8. Hilton, I. B. *et al.* Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nat. Biotechnol.* **33**, 510–7 (2015).
9. Chen, B. *et al.* Dynamic Imaging of Genomic Loci in Living Human Cells by an Optimized CRISPR/Cas System. *Cell* **155**, 1479–1491 (2013).
10. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).
11. Capon, S. J. *et al.* Utilising polymorphisms to achieve allele-specific genome editing in zebrafish. *Biol. Open* **6**, 125–131 (2017).
12. Christie, K. A. *et al.* Towards personalised allele-specific CRISPR gene editing to treat autosomal dominant disorders. *Sci. Rep.* **7**, 16174 (2017).
13. Courtney, D. G. *et al.* CRISPR/Cas9 DNA cleavage at SNP-derived PAM enables both in vitro and in vivo KRT12 mutation-specific targeting. *Gene Ther.* **23**, 108–12 (2016).
14. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas9 nucleases with altered PAM specificities. *Nature* **523**, 481–485 (2015).
15. Nishimasu, H. *et al.* Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science (80-.).* **361**, 1259–1262 (2018).
16. Hu, J. H. *et al.* Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature* **556**, 57–63 (2018).
17. Ran, F. A. *et al.* In vivo genome editing using Staphylococcus aureus Cas9. *Nature* **520**, 186–191 (2015).
18. Kleinstiver, B. P. *et al.* Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. *Nat. Biotechnol.* **33**, 1293–1298 (2015).
19. Ma, D. *et al.* Engineer chimeric Cas9 to expand PAM recognition based on evolutionary information. *Nat. Commun.* **10**, 560 (2019).
20. Esvelt, K. M. *et al.* Orthogonal Cas9 proteins for RNA-guided gene regulation and editing. *Nat. Methods* **10**, 1116–1121 (2013).
21. Xu, K. *et al.* Efficient genome engineering in eukaryotes using Cas9 from Streptococcus thermophilus. *Cell. Mol. Life Sci.* **72**, 383–399 (2015).

22. Sun, B. *et al.* A CRISPR-Cpf1-assisted Non-homologous End Joining Genome Editing System of *Mycobacterium smegmatis*. *Biotechnol. J.* e1700588 (2018). doi:10.1002/biot.201700588
23. Kim, E. *et al.* In vivo genome editing with a small Cas9 orthologue derived from *Campylobacter jejuni*. *Nat. Commun.* **8**, 14500 (2017).
24. Tan, S. Z., Reisch, C. R. & Prather, K. L. J. A Robust CRISPR Interference Gene Repression System in *Pseudomonas*. *J. Bacteriol.* **200**, e00575-17 (2018).
25. Zetsche, B. *et al.* Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. *Cell* **163**, 759–71 (2015).
26. Gao, L. *et al.* Engineered Cpf1 variants with altered PAM specificities. *Nat. Biotechnol.* **35**, 789–792 (2017).
27. Kleinstiver, B. P. *et al.* Engineered CRISPR-Cas12a variants with increased activities and improved targeting ranges for gene, epigenetic and base editing. *Nat. Biotechnol.* **37**, 276–282 (2019).
28. Fonfara, I., Richter, H., Bratovič, M., Le Rhun, A. & Charpentier, E. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. *Nature* **532**, 517–21 (2016).
29. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–3 (2009).

Figures

Figure 1

The proportion of PAM-generating pathogenic and likely-pathogenic SNPs in humans Representation of PAM generators compared to PAM non-generators. PAM generators are SNPs that generated at least one PAM. While a SNP may be generating more than one PAM, a PAM-generating SNP is counted once regardless of the number of PAMs it generates. 84% of the total SNPs checked where found to be PAM-generators. The average number of PAMs generated by a SNP is 6.97.

Figure 2 A

PSEN1 – WT:

5' -GTACCTCCCTGAATGGACTGCGTGGCTCATCTTGGCTGTGA-3'
3' -CATGGAGGGACTTACCTGACGCACCGAGTAGAACCGACACT-5'

PSEN1 - rs63750526:

(C to A)

5' -GTACCTCCCTGAATGGACTGAGTGGCTCATCTTGGCTGTGA-3'
3' -CATGGAGGGACTTACCTGACTCACCGAGTAGAACCGACACT-5'

Cas	PAM
EQR SpCas9	NGAG
VQR SpCas9	NGAN
SaCas9	NNGRRT
KKH SaCas9	NNNRRT
xCas9	HGA
cCas9	NNVRRN and NNVACT

Figure 2 B

rsID	63750526
Gene ID	5663
Gene name	PSEN1
Clinical Sig.	Pathogenic
WT sequence	GTACCTCCCTGAATGGACTG<C>GTGGCTCATCTTGGCTGTGA
Variant	A
Chr.	14
Base replacements	C to A;
Cas	PAM
SpCas9	NGG
VRER SpCas9	NGCG
EQR SpCas9	NGAG
VQR SpCas9	NGAN
SpCas9-NG	NG
xCas	NGT
SaCas9	NNGRRT
KKH SaCas9	NNNRRT
NmCas9	NNNGATT
StCas9	NNAGAAW
TdCas9	NAAAC
CjCas9	NNNVRYM
SpCas9 (from Streptococcus pasteurianus)	NNGTGA
LbCpf1 & AsCpf1	TTTN
FnCpf1	TTN
AsCas12a RVR	TATV
AsCas12a RR	TYCV
enAsCas12a	TGTV
	VTTV
	TTTT
	TTCN
cCas9	NNVRRN
	NNVACT
	NNVATG
	NNVATT
	NNVGCT
	NNVTG
	NNVGT

Figure 2

A case study - rs63750526 SNP of PSEN1 as an example of a SNP-derived PAM rs63750526 SNP of PSEN1, known as a risk factor for early onset Alzheimer's disease, as an example of multiple PAMs generated by a SNP. A) The variant nucleotide A generated 7 PAMs: EQR SpCas9's NGAG, VQR SpCas9's NGAN, SaCas9's NNGRRT, KKH SaCas9's NNNRRT, xCas9's HGA and cCas9's NNVRRN and NNVACT on

the complement strand. B) The obtained data for rs63750526 from the database generated by CrisPam, showing a pathogenic SNP of PSEN1 and the PAMs it generates.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Table1.pdf](#)