beditor: A computational workflow for designing libraries of guide RNAs for CRISPR base editing

Rohan Dandage*[1,2,3,4], Philippe C Després[1,2,3], Nozomu Yachie[5,6,7], Christian R Landry[1,2,3,4]

1. Département de Biochimie, Microbiologie et Bio-informatique, Faculté de sciences et genie, Université Laval, Québec, Québec, G1V 0A6, Canada

2. PROTEO, The Québec Research Network on Protein Function, Structure and Engineering, Université Laval, Québec, Québec, G1V 0A6, Canada

3. Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, Québec, G1V 0A6, Canada

4. Département de Biologie, Faculté de sciences et Génie, Université Laval, Québec, Québec, G1V 0A6, Canada

5. Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan.

6. Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

7. Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan

8. Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan

*Corresponding author: rohan.dandage.1@bio.ulaval.ca

Abstract

Recently engineered CRISPR base editors have opened unique avenues for scar-free genome-wide mutagenesis. Here, we describe a comprehensive computational workflow called *beditor* that can be broadly adapted for designing guide RNA libraries to be used for CRISPR base editing. The computational framework allows users to assess editing possibilities using a range of CRISPR base editors, PAM recognition sequences and the genome of any species. Additionally, potential editing efficiencies of the designed guides are evaluated in terms of an *a priori* estimates, through a specifically designed *beditor* scoring system.

Keywords

CRISPR, base editing, genome-wide targeted mutagenesis, guide RNA library design

Background

CRISPR base editing is emerging as being a highly efficient method for targeted sequence modification without the need for double-stranded DNA break followed by homology-directed repair (1–3). Base editors (BEs) are engineered by fusing a DNA modifying protein with a nuclease-defective Cas9 (dCas9) protein. Currently, two major types of BEs have been developed – Cytosine base editors that catalyze the conversion of Cytosine to Thymine, Adenine or Guanine (1,4–6) and Adenine base editors that catalyze the conversion of Adenine to Guanine (2,7,8). BEs that are currently in use, for example Target-AID (C•G to T•A) (1) and ABE (A•T to G•C) (2), enable many codon and amino acid substitutions (Fig S1). With the enormous potential for new discoveries and developments, this editing capability could increase with new DNA modifying enzymes and engineered Cas9 with expanded spectra of PAM recognition sites (9).

Although CRISPR base editing has opened new avenues for large-scale genome-wide mutagenesis, there is still no computational platform that allows systematic and flexible assessment of base editing possibilities for any set of genomic mutations, making the design of guide RNA (gRNA) libraries and the estimation of potential off-target effects difficult. Currently available computational tools that deal with CRISPR base editing are either focused on non-sense mutations (10) or only allow a limited set of BEs and Protospacer Adjacent Motifs (PAM) recognition sequences (11). To utilize the full potential of base editing technologies, there is a need for a comprehensive and customizable computational workflow that streamlines the design of gRNA libraries on large scale.

When considering potential applications of base editing technologies in genome-wide mutagenesis screens, essential characteristics for a good gRNA designing tool emerge. First, the ability to fetch and manipulate a wide range of genomes is a foremost requirement. Second, some flexibility regarding the scale of the gRNA libraries and genetic screens would be crucial. Third, customizability to adapt to the continuously evolving base editing technologies is imperative, i.e. it is necessary that the tool can accommodate novel and custom BEs with customizable PAMs. Fourth, *a priori* assessment of efficiency of a gRNA and its target site in terms of editability by a given BE and potential off-target effects would help users in improving the quality of base editing experiments and enhance throughput because higher mutagenesis rates allow for detection of smaller fitness or functional effects (12).

In order to address these needs, we present here a comprehensive computational workflow called *beditor* (Fig. 1a) that is directly compatible with more than 100 genomes hosted in the Ensembl genome database (13). Built-in parallel processing allows users to easily scale up the design of gRNA libraries with minimal computational resource requirements. *beditor* allows full customizability in terms of editing properties of BEs (eg.

range of nucleotides where maximum catalytic activity of BE occurs, henceforth simply referred to as ‘activity window’) and PAM recognition sequence. Prospective efficiencies of gRNAs are determined through a scoring system that accounts for the number and types of off-target alignments across the genome and editing properties of BEs (Fig. 1b). Overall, the *beditor* workflow has broad applicability for the genome editing community and its open source implementation allows for continuous enhancements in the future.

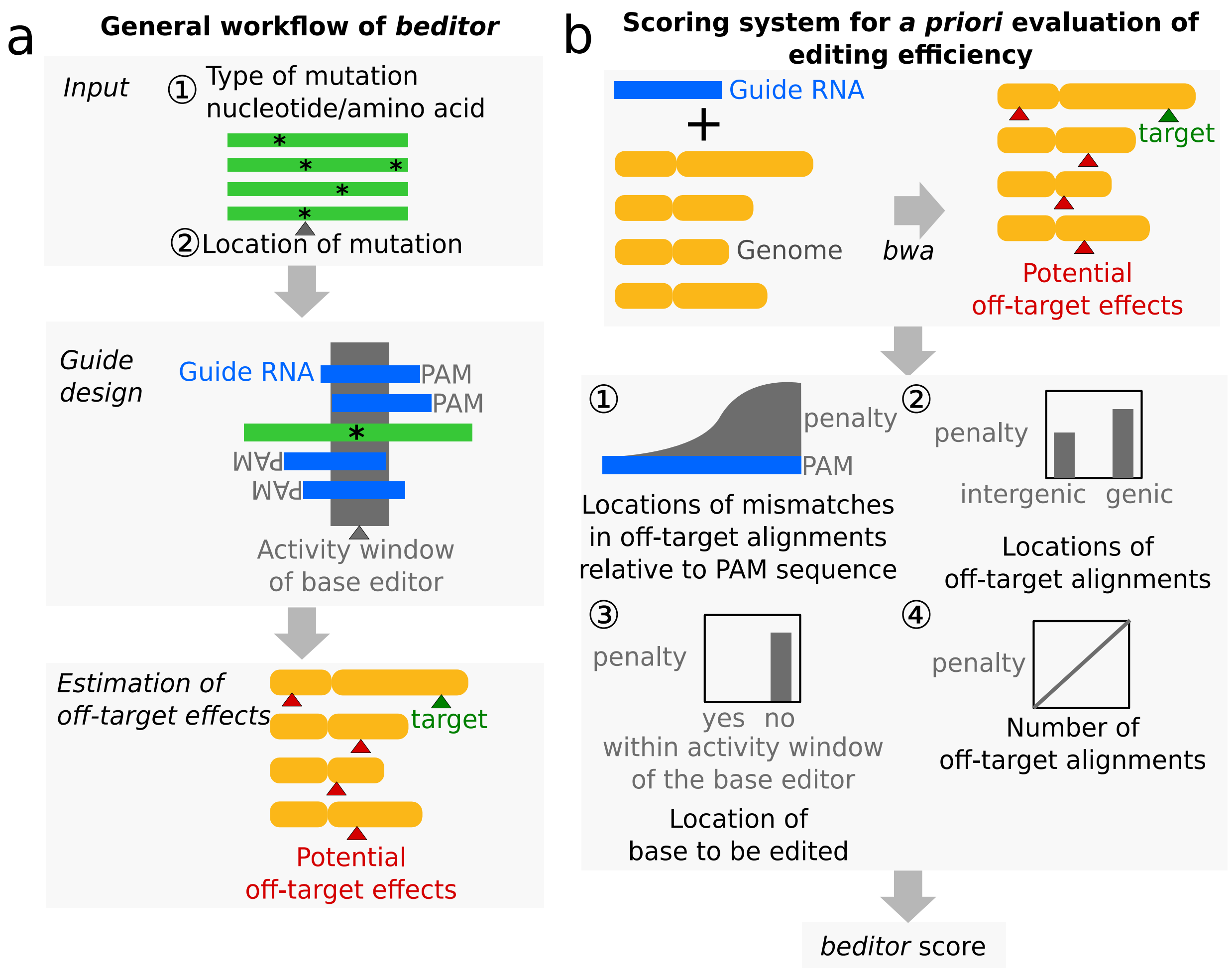
a
General workflow of *beditor*
Input
① Type of mutation nucleotide/amino acid
② Location of mutation
Guide design
Guide RNA
PAM
PAM
PAM
PAM
Activity window of base editor
Estimation of off-target effects
target
Potential off-target effects
b
Scoring system for *a priori* evaluation of editing efficiency
Guide RNA
+
Genome
bwa
target
Potential off-target effects
①
penalty
PAM
Locations of mismatches in off-target alignments relative to PAM sequence
②
penalty
intergenic
genic
Locations of off-target alignments
③
penalty
yes
no
within activity window of the base editor
Location of base to be edited
④
penalty
Number of off-target alignments
beditor score

Fig. 1 The computational workflow of *beditor* allows for the flexible design of gRNA libraries to be used in CRISPR base editing and offers *a priori* evaluation of editing efficiencies.

a Information on the type and location of desired mutations are supplied to the *beditor* workflow as a tab separated file. gRNAs are designed according to the user provided sets of BEs and PAM recognition sequences. Nucleotide windows for maximum activity are considered while designing the gRNAs. Finally, potential off-target effects are assessed.

b A scoring system specifically designed for *a priori* evaluation of gRNA editing efficiencies. Penalties are assigned based on (1) the total number of off-target alignments of gRNAs to the reference genome, (2) positions of the mismatches in the off-target alignments relative to the PAM and (3) genomic locations of off-target alignments and lastly, (4) whether the editable base lies inside the activity window of the BE. Using all of the above penalties, a final score is calculated for each gRNA sequence – the *beditor* score.

Results

In the general workflow of *beditor* (Fig. 1a), the user provides information about the desired set of mutations as an input and a library of gRNAs is generated with corresponding *a priori* efficiency estimates for each gRNA. *beditor* can also be used to execute only a subset of the analysis steps by changing the input parameters or providing inputs for intermediate steps. The standard input of *beditor* depends on the format of mutations i.e. nucleotide or amino acid. For carrying out nucleotide level mutations, the users need to provide genome coordinates and the wanted mutated nucleotides. For carrying out amino acid level mutations, the users provide Ensembl stable transcript ids, the position of the targeted residues and the corresponding mutated residue. The users can also provide inputs to limit the amino acid substitutions to a custom substitution matrix and specify whether only non-synonymous or synonymous substitutions should be carried out. In addition to creating mutations on a wild-type background ('create' mode), the *beditor* workflow also provides an option to design guides that would remove alternative SNPs and to mutate to the reference or wild-type alleles (‘remove' mode).

The *beditor* workflow utilizes a PyEnsembl python API (14) to fetch and work with the genomes of over 125 species and their various assemblies from the Ensembl genome database(13,15), providing a broad utility for researchers across a wide spectrum of fields. *beditor* is also compatible with any customly annotated genome. The ability to carry out parallel processing allows for the design of large gRNA sequence libraries using minimal computational resources. Users can incorporate BEs with varied editing properties and even novel BEs as per requirements. Similarly, users can incorporate any custom PAM sequences in addition to the 16 different PAMs already incorporated in *beditor* (Table S1). Lastly, an *a priori* statistics of efficiency of gRNAs allows the users to select the best set of gRNAs for mutagenesis experiments.

The *beditor* workflow also integrates the Cutting Frequency Determination (CFD) scoring system based on empirical data from genetic screens (16). However, it is only applicable to gRNAs with NGG PAM. Therefore, informed from empirical data from genetic screens (16), we defined *a priori* scoring system to evaluate editing efficiencies of gRNAs. It additionally utilizes features of the alignments of the gRNAs to the reference genome and editing preferences of the BEs. Penalties are assigned to guides based on (1) the total number of off-target alignments, (2) the location of mismatch within gRNA sequences relative to PAMs and (3) whether the gRNA aligns to genic or intergenic regions. Compatibility with the editing preferences of BEs is captured using (4) a penalty that is assigned if the editable base lies outside of the activity window of BE (Fig. 1b, see Methods). The former two penalties account for potential off-target effects while the later ones pertain to increasing editing efficiencies. The third penalty may reduce noise that may occur due to unwanted mutations at off-target sites located in functional genic regions. Optimal gRNA sequences have a *beditor* score of 1, while lower *beditor* score indicates probable lower editing efficiency. Multiplication of the penalties insures that gRNA sequences that carry penalty in *any* of the criteria carries a lower *beditor* score.

To demonstrate the utility of our computational workflow, we designed a library of gRNAs against a set of clinically associated SNPs in the human genome composed of 61083 nucleotide level and 81819 amino acid level mutations (see Methods). This analysis was carried out with two different BEs: Target-AID and ABE, and two PAM sequences: NGG (17,18) and NG (19,20) and in ‘create’ and ‘reverse’ mode. The output libraries of gRNA sequences (Additional file 1) targets ~25% of the total mutations provided as input (Table S2). On average, ~1.6 guides were designed for each mutation. Among the visualizations produced by *beditor,* the percentage of substitutions that can be edited with the designed guides (% editability) is represented as substitution maps (Fig 1a and b for ‘create’ mode and Figure S2 for ‘remove’ mode). Strategy-wise number of gRNAs designed (Fig 2c and d), nucleotide composition of the library of gRNAs (Figure S3) and alignments of gRNAs with the target sequence (Figure S4) are visualized to aid users in assessing the overall composition of the designed library guide RNAs .

From the demonstrative analysis, *beditor* scores were evaluated for each gRNA sequence in the library. From the distribution of scores (Figure S5), the users may assign a threshold to filter out low efficiency gRNAs. Collectively, by definition, the *beditor* scores are negatively correlated (ρ=-0.94) with the number of off-target alignments (Fig 2e) and penalty assigned for each alignment based on distance of mismatches from the PAM sequence is positively correlated (ρ=0.65) with the distance (Fig 2f). Note that the rank correlation is not perfect because of cases in which there were two mutations in the aligned sequence. Also, purely informed from sequence alignments of gRNAs and the requirements of BEs, the *beditor* score captures the possible empirical efficiencies of gRNAs as apparent from its strong positive correlation (ρ=0.95) with the CFD score (Fig 2g).

The *beditor* workflow is implemented as an open-source python 3.6 package hosted at https://pypi.org/project/beditor. The source code of *beditor* can be accessed at https://www.github.com/rraadd88/beditor.

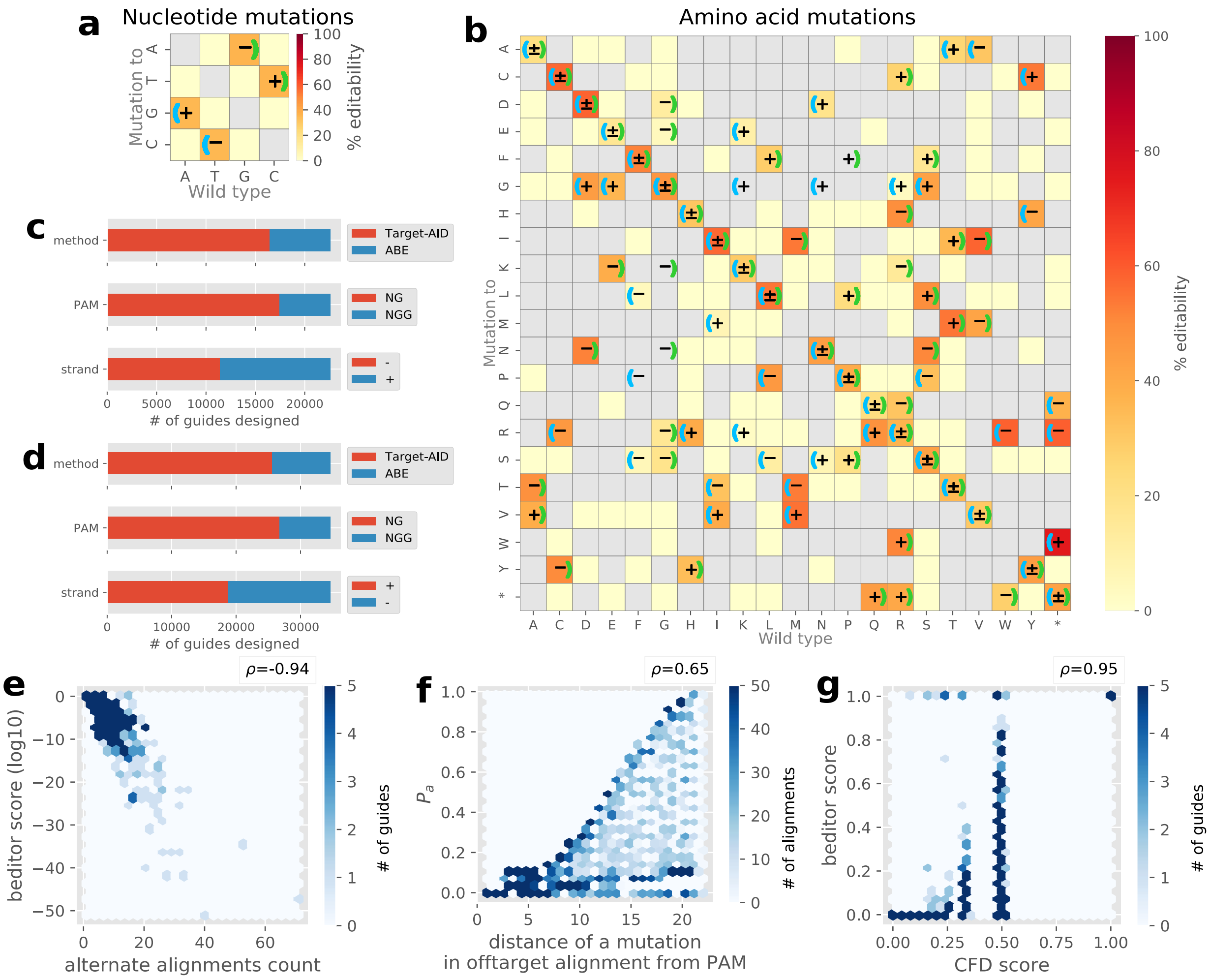

Fig. 2 Demonstrative analysis of clinically associated human SNPs. gRNA libraries were designed to create nucleotide and amino acid level mutations from clinically associated SNPs in the reference human genome. For this analysis, 2 base-editors (Target-AID and ABE) and 2 PAM sequences (NGG and NG) were used.

Percentage of substitutions that can be edited by gRNA library designed (% editability) in the demonstrative analysis of nucleotide (**a**) and amino acid mutation (**b**) data (see Methods). Left and right brackets indicate that the substitution is carried out by ABE and Target-AID respectively. +, - and ± indicate substitutions for which guide RNA is designed on +, - and both strands respectively. Shown in gray are substitutions that are absent in the input data. * indicates non-sense mutation.

Number of gRNAs designed by each strategy in case of demonstrative analysis of nucleotide (**c**) and amino acid mutation (**d**) data.

e Relationship between the number of genome-wide off-target alignments and *beditor* score per gRNA. The color of hexbins are scaled according to the number of gRNAs per bin.

f Relationship between the distance of a mutation in off-target alignments and corresponding penalty assigned (P_a). The color of hexbins are scaled according to the number of off-target alignments per bin.

g Relationship between the CFD score and *beditor* score for all the gRNAs carrying NGG PAM sequence. The color of hexbins are scaled according to the number of gRNAs per bin.

ρ is Spearman's correlation coefficient.

Discussion

CRISPR base editing carries an immense potential for scar-free genome-wide removal of non-desired mutations or to study the effect of specific mutations. Therefore, the major application of *beditor* lies in (but is not limited to) designing gRNA libraries for large-scale genetic screens. The modular design of the workflow allows users to customize the computational platform for specific experimental designs. Additionally, its compatibility with the Ensembl database and custom genomes allows broader applications of this resource in the scientific community. With the analysis of clinically important SNPs, we demonstrate that *beditor* is efficient in designing large guide RNA libraries.

Conclusions

The *beditor* workflow is ideally suited for designing library of gRNA sequences for genome-wide mutagenesis screenings using CRISPR base editing. Compatibility with any genome and customizability with respect to the editing make this tool broadly accessible to users from a wide range of disciplines. *beditor* is an open source software and is available at https://github.com/rraadd88/beditor.

Methods

beditor scoring system

Alignment of the designed gRNAs (with PAM sequence) with the provided reference genome is carried out using BWA (21), allowing for a maximum of two mismatches per alignment. The *beditor* score is evaluated based on features of off-target alignments as follows.

$$P_i = \begin{cases} P_{min} & if \text{ mismatch is near PAM} \\ \vdots & \vdots \\ P_{max} & if \text{ mismatch distant from PAM} \end{cases} \qquad ..\ (1)$$

$$P_a = \prod_{i=1}^{M_{max}} P_i \qquad ..\ (2)$$

$$G_a = \begin{cases} G_g & if\ genic \\ G_{ig} & if\ intergenic \end{cases} \qquad ..\ (3)$$

$$B = \left(\prod_{a=1}^{n} P_a * G_a \right) * A \qquad ..\ (4)$$

For an alignment between a gRNA sequence and the genome, P_i is a penalty assigned to a nucleotide in the gRNA sequence based on the position of a mismatch in the aligned sequence relative to the PAM. If the mismatch is near the PAM sequence, a minimal penalty P_{min} is assigned. Conversely, if the mismatch is far from the PAM, a maximum penalty P_{max} is assigned. The relative values of such penalties were determined by fitting a third degree polynomial equation to the mismatch tolerance data from (16). This way, penalties increase non-linearly from P_{min} to P_{max}, as the distance of nucleotide (*i*) from PAM sequence increases. Individual penalties assigned for all the nucleotides in a gRNA are then multiplied to estimate a penalty score for a given alignment called P_a (equation 2). G_a is a penalty defined by whether the off-target alignment lies within a genic or an intergenic region (equation 3). *A* is a penalty based on whether the editable base lies within the activity window of BE (equation 4). Note that the *a priori* penalties are assigned in the current version of *beditor* will be informed by empirical data in the future. The overall *beditor* score *B* for a gRNA is determined by multiplying penalties assigned per alignment (P_a and G_a) for all alignments (n) with a penalty assigned to the gRNA (A) (equation 4). Multiplication of individual penalties insures that if *any* of the criteria is suboptimal, the *beditor* score decreases.

Demonstrative analysis using a set of human mutations

For the demonstrative analysis, a set of clinically associated human mutations were obtained from the Ensembl database in GVF format (ftp://ftp.ensembl.org/pub/release-93/variation/gvf/homo_sapiens/homo_sapiens_clinically_associated.gvf.gz, Date modified: 08/06/2018, 16:13:00). From genomic co-ordinates of SNPs, residue index and types were identified using PyEnsembl (14). Ensembl stable transcript ids, amino acid position, reference residue and mutated residue were used as input to the *beditor* workflow as a tab-separated file. The command "beditor --cfg params.yml" was executed. Here, params.yml contains input parameters of the analysis (Table S3) in a user-friendly YAML format. Output visualizations from this analysis are presented in Fig 2, Figure S3, S4 and S5.

Open-source dependencies

The beditor workflow depends on other open source softwares such as PyEnsembl (14), BEDTools(22), BWA(21) and SAMtools(23) at various steps of the analysis. User provided mutation information is first checked for validity with PyEnsembl (14). Genomic sequences flanking the mutation sites are fetched using BEDTools (22). The designed gRNAs are aligned back to the reference genome using BWA (21) and alignments are processed using SAMtools (23) for evaluation of off-target effects using the beditor scoring system. Visualization of alignments of guide RNAs with genomic DNA are created using DnaFeaturesViewer package (https://github.com/Edinburgh-Genome-Foundry/DnaFeaturesViewer).

List of abbreviations

CRISPR: Clustered Regularly Interspaced Short Palindromic Repeats

gRNA: guide RNA

BE: base editor

PAM: Protospacer adjacent motif

CFD: Cutting Frequency Determination

SNP: Single-Nucleotide Polymorphism

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and material

The dataset analyzed in the study i.e. set of clinically associated human mutations were

obtained from the Ensembl database in GVF format (ftp://ftp.ensembl.org/pub/release-93/variation/gvf/homo_sapiens/homo_sapiens_clinically_associated.gvf.gz, Date modified: 08/06/2018, 16:13:00). Processed data from this study i.e. designed gRNA library is provided in Additional file 1.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the Canadian Institutes of Health Research [299432, 324265 and 387697 to C.R.L. 364920, 384483 and Frederick Banting and Charles Best graduate scholarship to P.C.D.], Laval University [André Darveau Fellowship to P.C.D.] and the Japan Society for the Promotion of Science [S15734 and S17161 to C.R.L. and N.Y.].

Authors' contributions

RD and CRL conceived the research. RD designed and developed the software with valuable input and advice from PCD, NY and CRL. RD, PCD, and CRL wrote the manuscript with help from NY. All authors read and approved the final manuscript.

Acknowledgements

We thank members of the Landry and Yachie Labs who provided important suggestions and feedback on the manuscript. We would like to thank Hideto Mori from Yachie Lab for informative discussions.

Authors' information

1. Département de Biochimie, Microbiologie et Bio-informatique, Faculté de sciences et génie, Université Laval, Québec, Québec, G1V 0A6, Canada
2. PROTEO, The Québec Research Network on Protein Function, Structure and Engineering, Université Laval, Québec, Québec, G1V 0A6, Canada
3. Centre de Recherche en Données Massives (CRDM), Université Laval, Québec, Québec, G1V 0A6, Canada
4. Département de Biologie, Faculté de sciences et Génie, Université Laval, Québec, Québec, G1V 0A6, Canada
5. Research Center for Advanced Science and Technology, University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo 153-8904, Japan.
6. Department of Biological Sciences, Graduate School of Science, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
7. Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, Tsuruoka, Yamagata 997-0035, Japan
8. Graduate School of Media and Governance, Keio University, 5322 Endo, Fujisawa, Kanagawa 252-0882, Japan

References

1. Nishida K, Arazoe T, Yachie N, Banno S, Kakimoto M, Tabata M, et al. Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. 2016;8729(August):1–14.

2. Gaudelli NM, Komor AC, Rees HA, Packer MS, Badran AH, Bryson DI, et al. Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature [Internet]. 2017;551(7681):464–71. Available from: http://www.nature.com/doifinder/10.1038/nature24644

3. Kim J-S. Precision genome engineering through adenine and cytosine base editing. Nat plants [Internet]. 2018 Mar;4(3):148–51. Available from: http://dx.doi.org/10.1038/s41477-018-0115-z

4. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016;533(7603):420.

5. Hess GT, Frésard L, Han K, Lee CH, Li A, Cimprich KA, et al. Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat Methods. 2016;13(12):1036–42.

6. Ma Y, Zhang J, Yin W, Zhang Z, Song Y, Chang X. Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat Methods. 2016;13(12):1029–35.

7. Cox DBT, Gootenberg JS, Abudayyeh OO, Franklin B, Kellner MJ, Joung J, et al. RNA editing with CRISPR-Cas13 David. Science (80-). 2017;0180(October):1–15.

8. Zafra MP, Schatoff EM, Katti A, Foronda M, Breinig M, Schweitzer AY, et al. Optimized base editors enable efficient editing in cells, organoids and mice. Nat Biotechnol [Internet]. 2018;(July). Available from: http://www.nature.com/doifinder/10.1038/nbt.4194

9. Nishimasu H, Yamano T, Gao L, Zhang F, Ishitani R, Nureki O. Structural Basis for the Altered PAM Recognition by Engineered CRISPR-Cpf1. Mol Cell. 2017;

10. Billon P, Bryant EE, Joseph SA, Nambiar TS, Hayward SB, Rothstein R, et al. CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes through Induction of STOP Codons. Mol Cell [Internet]. 2017;67(6):1068–1079.e4. Available from: http://dx.doi.org/10.1016/j.molcel.2017.08.008

11. Hwang G, Park J, Lim K, Kim S, Yu J, Eils R, et al. Web-based design and analysis tools for CRISPR base editing. bioRxiv [Internet]. 2018;5:373944. Available from: https://www.biorxiv.org/content/early/2018/07/22/373944?rss=1&utm_source=dlvr.it&utm_medium=twitter

12. Després PC, Dubé AK, Nielly-Thibault L, Yachie N, Landry CR. Double Selection Enhances the Efficiency of Target-AID and Cas9-Based Genome Editing in Yeast. G3 (Bethesda) [Internet]. 2018 Aug 10;g3.200461.2018. Available from: http://www.ncbi.nlm.nih.gov/pubmed/30097473

13. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Res. 2018;

14. PyEnsembl. https://github.com/openvax/pyensembl. 2018;

15. Ensembl: Table of Assemblies. https://useast.ensembl.org/info/website/archives/assembly.html (2018) Accessed on 18 Aug 2018.

16. 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 [Internet]. 2016;34(2):184–91. Available from: http://dx.doi.org/10.1038/nbt.3437

17. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31(9):827–32.

18. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-). 2012;

19. Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature [Internet]. 2018;556(7699):57–63. Available from: http://dx.doi.org/10.1038/nature26155

20. Nishimasu H, Shi X, Ishiguro S, Gao L, Hirano S, Okazaki S, et al. Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science (80-) [Internet]. 2018 Aug 30;9129(August):eaas9129. Available from: http://www.sciencemag.org/lookup/doi/10.1126/science.aas9129

21. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;

22. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;

23. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078–9.

Supporting information

beditor: A computational workflow for designing libraries of guide RNAs for CRISPR base editing

Supporting figures

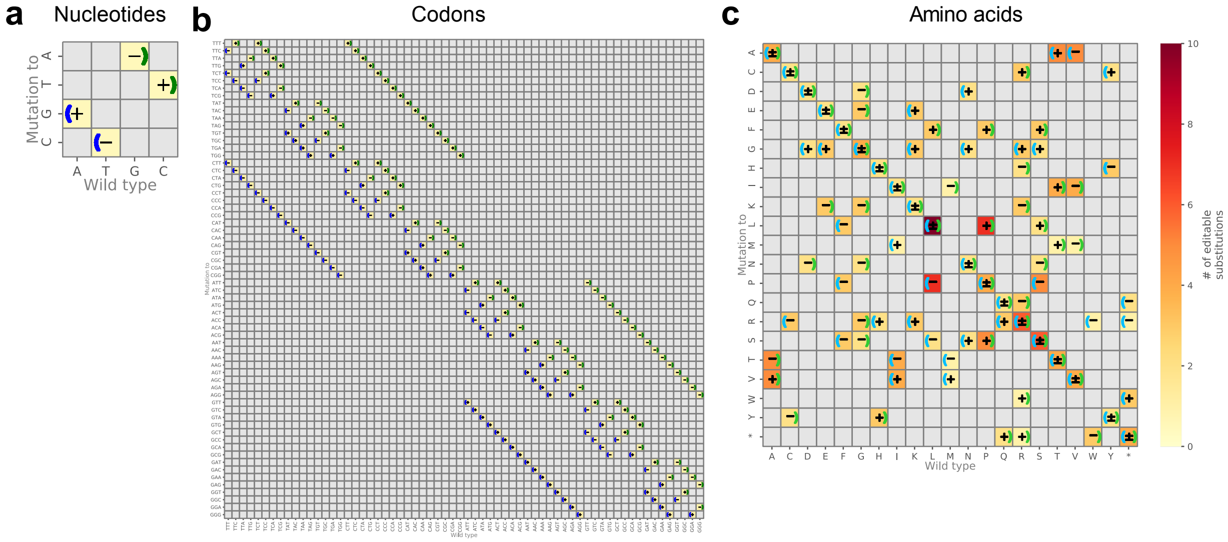

FIg S1 : Editable substitutions by ABE and Target-AID.

a Nucleotide level substitutions.

b Codon level substitutions.

b Amino acid level substitutions.

Mapped on the heatmaps are cumulative number of substitutions that can be edited with ABE or Target-AID. Left and right brackets indicate that the substitution is carried out by ABE and Target-AID respectively. +, - and ± indicate substitutions for which guide RNA is designed on +, - and both the strands respectively. Shown in gray are substitutions that are absent in the input data. * is a non-sense mutation.

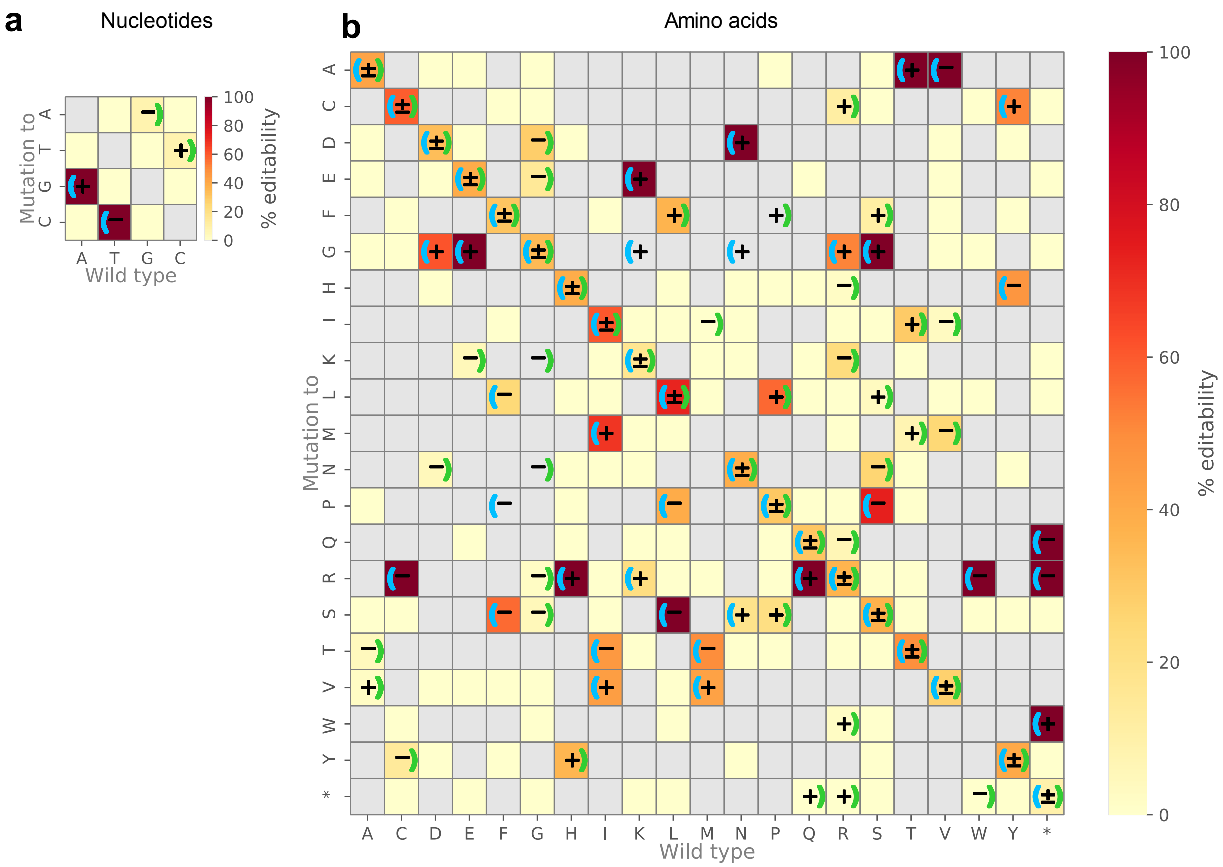

FIg S2: Percentage editability for demonstrative analysis in ‘remove’ mode.

a Nucleotide level substitutions.

b Amino acid level substitutions.

Mapped on the heatmaps is a ratio between number of mutations that can be edited with the designed gRNAs and the number of mutations present in the input data (% editability). Left and right brackets indicate that the substitution is carried out by ABE and Target-AID respectively. +, - and ± indicate substitutions for which guide RNA is designed on +, - and both the strands respectively. Shown in gray are substitutions that are absent in the input data. * is a non-sense mutation.

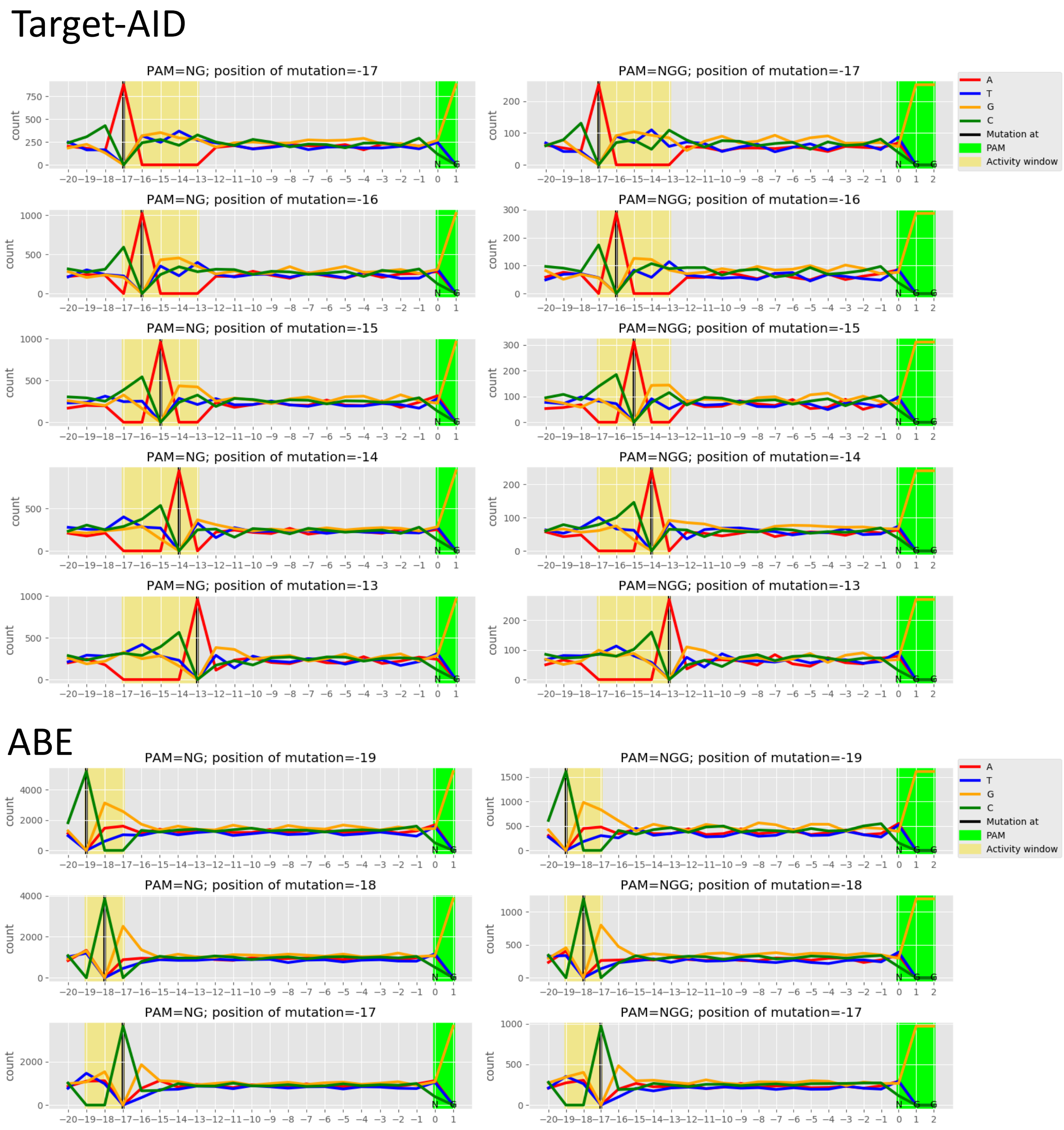

Fig S3: Nucleotide composition of the gRNA library designed for the demonstrative analysis of nucleotide mutations.

gRNA nucleotide composition of the library is represented for Target-AID (top) and ABE (bottom). The gRNAs are subdivided by the types of PAM sequence (shown in columns) and by the position of the editable nucleotides within the activity window of a BE (shown in the rows).

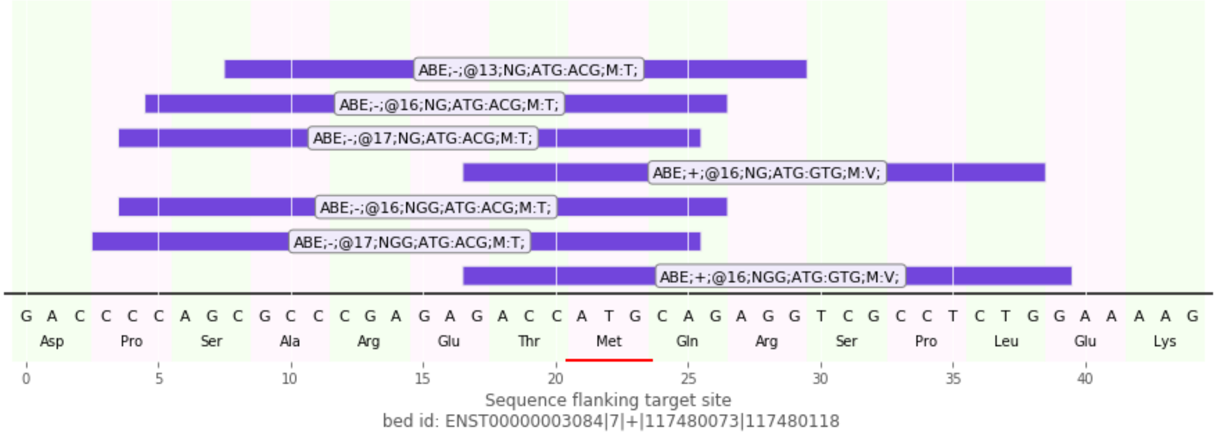

Fig S4: Representative visualizations of alignment between a gRNA and target genomic sequence.

A gRNA is shown in purple and its identity (indicating base editor, strand of mutagenesis, distance from PAM sequence, PAM sequence, reference codon, mutated codon, wild-type amino acid and mutated amino acid) is shown on the guide RNA. The target site is indicated in red color. Reading frames and genomic coordinates of the target DNA are shown below.

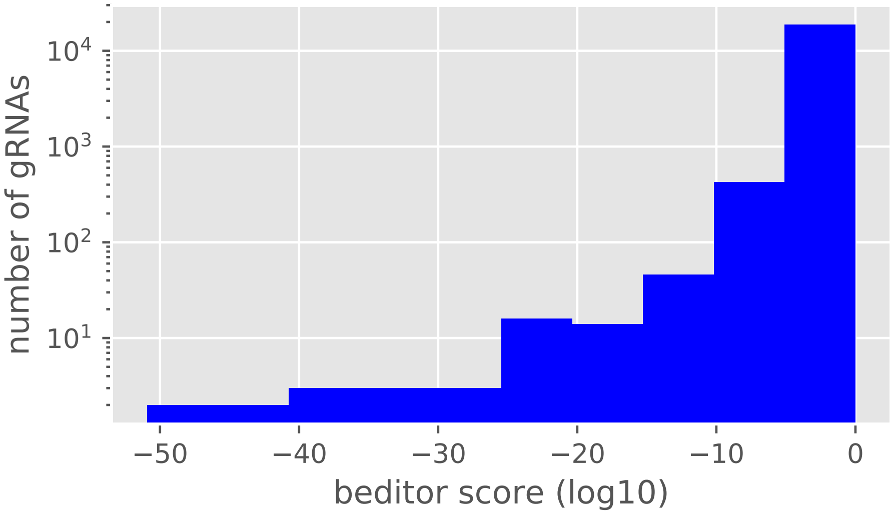

Fig S5: Distribution of *beditor* scores for the gRNA library designed for the demonstrative analysis of nucleotide mutations.

Supporting tables

Table S1: PAM recognition sequences supported by *beditor*.

PAM	Description	position	length of guide sequence	Reference
NG	xCas9	3'	20	[1]
NGA	SpCas9 mutant	3'	20	[2]
NGCG	SpCas9 mutant	3'	20	[2]
NGG	SpCas9	3'	20	[3,4]
NGGNG	Cas9 *S. Thermophilus*	3'	20	[5,6]
NGK	xCas9	3'	20	[1]
NGN	xCas9	3'	20	[1]
NNAGAA	Cas9 *S. Thermophilus*	3'	20	[5,6]
NNGRRT	SaCas9	3'	21	[7]
NNNNACA	CjCas9	3'	20	[8]
NNNNGMTT	Cas9 *N. Meningitidis*	3'	20	[9]
NNNRRT	KKH SaCas9	3'	21	[2,7]
TATV	AsCpf1 mutant	5'	23	[10]
TTN	Cpf1 *F. Novicida*	5'	23	[11]
TTTN	Cpf1 *Acidaminococcus / Lachnospiraceae*	5'	23	[11]
TYCV	TYCV AsCpf1 mutant	5'	23	[10]

Table S2: Summary statistics of demonstrative analysis.

Mutation format	**nucleotide**	**amino acid**	**nucleotide**	**amino acid**
Remove/create mutation	remove	remove	create	create
Total number of mutations in the input data	61083	81819	61083	81819
Total number of mutations edited	13709	19996	13867	20420
Total number of guides designed	23432	35390	22587	33311
% editability	22.44	24.43	22.70	24.95

Table S3: Input parameters used for the demonstrative analysis of clinically associated human variants

Variable	Input	Description
host	homo_sapiens	Name of host organism
genomerelease	93	Ensembl genome release
genomeassembly	GRCh38	Genome assembly version
dinp	din.tsv	File path of input tab-separated file
mutation_format	[aminoacid, nucleotide]	Whether the input data consists of amino acid or nucleotide mutations.
reverse_mutations	[FALSE, TRUE]	FALSE if design guide RNAs to 'create' mutations, TRUE to 'remove' mutations.
mutations	mutations	Information about mutated amino acid is taken from the input file
mutation_type	N	Type of mutations to process, N: non-synonymous, S: synonymous, else: both
keep_mutation_nonsense	FALSE	Whether to process non-sense mutations
pams	[NGG, NG]	List of PAMs to use
BEs	[Target-AID, ABE]	List of Base Editors to use
max_subs_per_codon	1	Maximum number of nucleotides that can be edited in the target codon.
mismatches_max	2	Maximum number of mismatches allowed in the alignment gRNAs against reference genome.
cores	5	Number of processors to use for parallel processing
chunksize	200	Number of mutations to process per processor

References

1. Hu JH, Miller SM, Geurts MH, Tang W, Chen L, Sun N, et al. Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature [Internet]. Nature Publishing Group; 2018;556:57–63. Available from: http://dx.doi.org/10.1038/nature26155

2. Kleinstiver BP, Prew MS, Tsai SQ, Topkar V V., Nguyen NT, Zheng Z, et al. Engineered CRISPR-Cas9 nucleases with altered PAM specificities. Nature. 2015;

3. Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, et al. DNA targeting specificity of RNA-guided Cas9 nucleases. Nat Biotechnol. 2013;31:827–32.

4. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science (80-). 2012;

5. Magadán AH, Dupuis MÈ, Villion M, Moineau S. Cleavage of phage DNA by the Streptococcus thermophilus CRISPR3-Cas system. PLoS One. 2012;

6. Garneau JE, Dupuis MÈ, Villion M, Romero DA, Barrangou R, Boyaval P, et al. The CRISPR/cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature. 2010;

7. Kleinstiver BP, Prew MS, Tsai SQ, Nguyen NT, Topkar V V., Zheng Z, et al. Broadening the targeting range of Staphylococcus aureus CRISPR-Cas9 by modifying PAM recognition. Nat Biotechnol. 2015;

8. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;

9. Hou Z, Zhang Y, Propson NE, Howden SE, Chu L, Sontheimer EJ, et al. Efficient genome engineering in human pluripotent stem cells using Cas9 from Neisseria meningitidis. Proc Natl Acad Sci USA. 2013;

10. Gao L, Cox DBT, Yan WX, Manteiga JC, Schneider MW, Yamano T, et al. Engineered Cpf1 variants with altered PAM specificities. Nat Biotechnol [Internet]. Nature Publishing Group; 2017;35:789–92. Available from: http://dx.doi.org/10.1038/nbt.3900

11. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015;