beditor: A computational workflow for designing libraries of guide RNAs for CRISPR-mediated 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

* To whom correspondence should be addressed. Email: rohan.dandage.1@bio.ulaval.ca

ABSTRACT

CRISPR-mediated 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 with a range of CRISPR-mediated base editors, PAM recognition sequences and genomes of many species. Additionally, in order to assist users in selecting the best sets of guide RNAs for their experiments, *a priori* estimates, called *beditor* scores are calculated. These *beditor* scores are intended to select guide RNAs that conform to requirements for optimal base editing: editable base falls within maximum activity window of the CRISPR-mediated base editor and produces non-confounding mutational effects with minimal off-target effects. We demonstrate the utility of the software by designing guide RNAs for base-editing mediated edition of thousands of clinically important human disease mutations.

INTRODUCTION

CRISPR-mediated base editors (BEs) are engineered by fusing a DNA modifying protein with a nuclease-defective Cas9 (dCas9) protein, allowing scar-free targeted mutagenesis (1–5). Currently, two major types of BEs are available – Cytosine base editors (CBEs) that catalyze the conversion of Cytosine to Thymine, Adenine or Guanine (1, 6–8)–for example Target-AID (C•G to T•A) (1)–and Adenine base editors (ABEs) that catalyze the conversion of Adenine to Guanine (A•T to G•C) (2, 9, 10). Currently BEs enable many codon level substitutions and thus amino acid substitutions (Figure S1). Owing to the unique capability of scar-free mutagenesis, BEs have recently found numerous applications in both model and non-model organisms (3, 11–14) and including substantial promise in therapeutics (15, 16).

For designing guide RNAs (gRNA) in BE mediated mutagenesis experiments, several specific requirements for the optimal activities of DNA modifying and Protospacer Adjacent Motif (PAM) sequences need to be taken into consideration. The complexity of this task especially increases when designing gRNA libraries against large sets of targets located across a given genome. Because the success of the BE-mediated mutagenesis experiment is directly dependent on the design of gRNA sequences, the methods and strategies involved in designing gRNA libraries are arguably the most important factors in such experiments.

Currently available gRNA designing tools however are either specifically focused on non-sense mutations (17) or allow very limited capabilities (BioRxiv: https://doi.org/10.1101/373944) (See Table S1), leaving a major roadblock in the applications of the BEs in genome editing. Continuous discoveries of new BEs and expansion of their existing editing capabilities (18) demand a complementary development in computational methods that would allow designing gRNAs using new and improved base editors. Moreover, considering the prospective applications across non-model organisms, compatibility with diverse genomes is essential. Overall, therefore, a comprehensive computational framework to design gRNAs libraries can potentially fuel the progress in CRISPR-base editing technology and its diverse applications in genome editing.

We developed a comprehensive computational workflow called *beditor* (Figure 1a) that can design gRNA libraries with any requirements of DNA modifying enzyme. These include the range of nucleotides where maximum catalytic activity of BE occurs, henceforth simply referred to as ‘activity window’, and PAM recognition sequence. *beditor* is directly compatible with more than 125 genomes hosted in the Ensembl genome database (19) and any annotated custom genomes. Additionally, the *beditor* workflow also provides *a priori* estimates called *beditor* scores for each gRNA that accounts for specific editing requirements of BEs, gRNA binding at off-target sites and the number and

types of off-target effects (Figure 1b). Such estimates will inform researchers in optimizing throughput of their BE-mediated mutagenesis experiments. The *beditor* workflow has broad applicability for the genome editing community and its open source implementation will allow for continuous enhancements in the future.

MATERIALS AND METHODS

Implementation

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. The documentation of the software with API is available at https://www.github.com/rraadd88/beditor/README.md. The *beditor* workflow depends on other open source softwares such as PyEnsembl (20), BEDTools(21), BWA(22) and SAMtools(23) at various steps of the analysis. User provided mutation information is first checked for validity with PyEnsembl (https://github.com/openvax/pyensembl.). Genomic sequences flanking the mutation sites are fetched using BEDTools (21). The designed gRNAs are aligned the reference genome using BWA (22) 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 the DnaFeaturesViewer package (https://github.com/Edinburgh-Genome-Foundry/DnaFeaturesViewer).

beditor scoring system

Alignment of the designed gRNAs (with PAM sequence) with the provided reference genome is carried out using BWA, as used in (22), allowing for a maximum of two mismatches per alignment (24). The *beditor* score is evaluated as follows.

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

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

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

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

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 (25) (Figure S2). 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). In cases of sgRNAs with lengths other than 20, the fitted equation is used to interpolate penalty scores. In case of 5' PAMs, the order of the vector containing position wise penalty scores is reversed. 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 due to the lack of large scale BE editing data, in the current version of *beditor*, penalties are set based on the importance of each requirement (Table S3). Such penalties would be informed from empirical data in the future developments.

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 1: customizability with respect to base editing strategies

For the demonstrative analysis in which custom base editors and PAM recognition sequences were used (Figure 2 and Table S4), sets of 1000 random mutations in *S. cerevisiae* were created as described in case of demonstrative analysis of mutations in different species. The command "beditor --cfg params.yml" was executed. Here, params.yml contains input parameters of the analysis (Table S7) in a user-friendly YAML format. For this analysis, input parameter 'host' was set to 'Saccharomyces_cerevisiae' and parameter 'genomeassembly' was set to 'R64-1-1'. Summary statistics on the editability are included as Table S4 and gRNA libraries are included as Supplementary data 1.

Demonstrative analysis 2: ability to work with different species

For the demonstrative analysis of mutations in different species, 1000 nucleotide and amino acid mutations were randomly assigned across the genomes of example species

(https://github.com/rraadd88/test_beditor). Such sets of mutations create uniform datasets ideal for testing features of *beditor*. The input mutation data was created in combinations of mutation formats (either amino acid or nucleotide) and modes of mutagenesis ("create" or "remove"). The input parameters used in this analysis are the same as Table S7 except for host names and genome assembly versions that were obtained from http://useast.ensembl.org/index.html. The summary statistics on the editability are included as Table S5 and gRNA libraries are included as Supplementary data 2.

Case study analysis

For the case study 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, inputs for nucleotide mutations (reference and mutated nucleotide) and amino acid mutations (transcript ids, amino acid position, reference residue and mutated residue) were identified using PyEnsembl (20). The mutation information was provided to the *beditor* workflow as a tab-separated file. The input parameters used in this analysis are same as Table S7 except for the host name – 'homo_sapiens' and genome assembly version– 'GrCh38'. Summary statistics on the editability are included as Table S6 and gRNA libraries are included as Supplementary data 3.

RESULTS

Design of *beditor* workflow

The *beditor* workflow contains sequential steps that lead from input target sequences and other input parameters to the designed gRNA libraries as output (Figure1a). The user provides information about the desired set of mutations as an input and a library of gRNAs is generated with the corresponding *a priori* estimates called *beditor* scores to help users in selecting the best performing gRNAs. *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. To carry out nucleotide level mutations, the users need to provide genome coordinates and the desired nucleotide after mutagenesis. 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).

Customizability for broad utility

The *beditor* workflow utilizes a PyEnsembl python API (20) to fetch and work with the genomes of over 125 species and their various assemblies from the Ensembl genome database(19, 26), providing a broad utility for researchers across a wide spectrum of fields. *beditor* is also compatible with any custom user-made annotated genome. The ability to carry out parallel processing allows for the design of large gRNA sequence libraries using minimal computational resources. The users can incorporate BEs with varied editing properties and even novel BEs as per requirements. Similarly, they can incorporate any custom PAM sequences in addition to the ~16 different PAMs already incorporated in *beditor* (Table S2). Lastly, *beditor* scores allows the users to select the best set of gRNAs for mutagenesis experiments.

Selecting the best performing gRNAs

We defined a novel *beditor* scoring system that can be used to select the best performing gRNAs from designed gRNA library. Due to the lack of large scale, genome-wide base editing data, we relied on few general rules that are applicable to all BEs. These rules pertain to the requirements for optimal mutagenesis. With the penalties assigned to each requirement, the *beditor* score of the optimally performing guide RNA would tend to be higher while that of poorly performing guide RNA would be lower (see Methods). In the future, we wish to determine the values of handcrafted penalty scores from empirical data. The current version of *beditor* scoring system assesses four of the general requirements for optimal mutagenesis. Based on the conformity of the gRNA to the requirements four penalties scores are assigned (Figure 1b, see Methods). (1) While general rules of the BE mediated editing is an active field of research, we utilize the basic rule which is common between all base editors: the editable base should lie within the maximum activity window of the BE. Thereby, the gRNAs are penalized if the editable base does not lie within the maximum activity window of BE. Next, utilizing the alignments of gRNAs to the genome, potential off-target sites are identified. (2) A penalty score is assigned based on the gRNA binding at off-target. It is evaluated by capturing a general trend of mismatch tolerance along the length of a gRNA (25) (See Methods). Given the lack of large scale, genome-wide base editing data, we used empirical data from 'conventional' CRISPR-Cas9–based genetic screens, assuming that the basic principles of gRNA recruitment and binding would be conserved between the two variants of CRISPR based

mutagenesis technologies. (3) Additionally, from alignments of the gRNAs, a penalty is assigned based on the location of the off-target site (genic or intergenic regions). Off-target editing at intergenic regions is less likely to confound the mutational effects compared to functionally important genic regions. Accordingly, penalties are assigned. (4) Lastly, in order to account for number of off-target sites, the 3 penalties are multiplied together to evaluate a *beditor* score per gRNA. Effectively, the optimal gRNAs have a *beditor* score of 1, while lower *beditor* score indicate incompatibility with BE requirements, significant off-targets or off-target effects that confound mutational effects.

Demonstrative analysis 1: customizability with respect to base editing strategies

We demonstrate that the users can incorporate and use custom base editors and PAM recognition sequences by designing gRNA libraries against sets of 1000 randomly assigned mutations (Supplementary methods) with 6 base editors (among which 4 are not yet discovered hence called “hypothetical”) and 16 PAM recognition sequences. With all the combinations of BE specific requirements and all combinations of two mutation formats (nucleotide and amino acid) and two modes of mutagenesis (‘remove’ and ‘create’), gRNA libraries were designed (Figure 2, Table S4 and Supplementary data 1).

Demonstrative analysis 2: ability to work with different species

To show *beditor*’s ability to work with genomes of different species, we designed gRNA libraries for representative 10 Ensembl genomes, for sets of 1000 randomly assigned mutations (Supplementary methods). Genomes of all the species were directly fetched from Ensembl Genome database and gRNA libraries were designed in all combinations of two mutation formats (nucleotide and amino acid) and two modes of mutagenesis (‘remove’ and ‘create’). In case of all the species, through the designed gRNA libraries (Supplementary data 2) reasonable editability (Table S5) was achieved.

Case study: designing a gRNA library against a set clinically relevant SNPs

To demonstrate the utility of our computational workflow, we designed a library of gRNAs against a set of clinically relevant SNPs in the human genome composed of 61083 nucleotide level and 81819 amino acid level mutations (see Supplementary methods). This analysis was carried out with two different BEs: Target-AID and ABE, and two PAM sequences: NGG (27, 28) and NG (29, 30) and in ‘create’ and ‘reverse’ mode. The output libraries of gRNA sequences (Supplementary data 3) targets ~25% of the total mutations provided as input (Table S6). The resulting gRNA libraries were composed of gRNAs designed with each of the input BEs and PAM sequence which targeted both the strands (Figure 3 a and b). On average, ~1.6 guides were designed for each mutation. Summary visualizations of gRNA libraries (Figure 3 c) as well as visualizations of alignments of gRNAs with the target sequence (Figure S3) were generated. Additionally, the percentage of substitutions that can be edited with the designed guides (% editability) is represented

as substitution maps (Figure 4) to indicate the proportion of input mutations that can be edited with the input BEs and PAMs.

Performance evaluation of *beditor* score

From the case study analysis, *beditor* scores were evaluated for each gRNA sequence in the library. From the distribution of scores (Figure S4), the users may assign a threshold to filter out gRNAs with lower *beditor* scores. Collectively, by definition, the *beditor* scores are negatively correlated (ρ=-0.94) with the number of off-target alignments (Figure 5a) and penalty assigned for each alignment based on distance of mismatches from the PAM sequence is positively correlated (ρ=0.65) with the distance (Figure 5b). 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 features of alignments of gRNAs and the requirements of BEs, the *beditor* score recapitulates empirical activity values of gRNAs with a strong positive correlation (ρ=0.95) determined in terms of Cutting Frequency Determination (CFD) score (25) (Figure 5c).

DISCUSSION

CRISPR-mediated base editors are have recently become a new paradigm in genetic editing, owing to their unique ability to carry out precise mutations without the need for DNA breaks (5). Consequently, plethora of applications of BEs have emerged from all corners of the genetic editing community, ranging from study of model and non-model organisms (3, 11–14) to therapeutics (15, 16). However, there has been a lack of robust software that can design gRNA libraries with any specific requirements of CRISPR-mediated base editing experiments. As presented here, the novel computational workflow of *beditor* (Figure 1a) fills in this important gap by allowing comprehensive customizability in terms of requirements of BE, PAM sequence and genome and thus increasing the applicability of the CRISPR-mediated base editing technology to broader community of researchers.

We demonstrate the modularity of *beditor* workflow in terms of the type of BE, PAM and genome by extensively testing them on synthetic sets of mutations. We also show that the workflow can be used to either create a mutation (‘create’ mode) or remove it (‘remove’ mode) with support for both nucleotide and amino acid format of mutations. Collectively, therefore, in terms of integrated customizability alone, *beditor* workflow provides a significant advance over other methods that provided only limited utilities (Table S1). In addition, we also introduce a novel method for *a priori* estimation of mutagenesis potential of gRNAs (Figure 1b), that utilizes features obtained from off-target alignments such as distance between mismatch and PAM recognition sequence. Such estimations would allow users to select a subset of designed gRNA library that would provide optimal mutagenesis in their experiments.

From the case study analysis of ~60000 human clinically relevant SNPs, we show that *beditor* workflow provides all round gRNA design capabilities, scanning though combinations of multiple strategies (Figure 3a and b). The validations of *beditor* score from this analysis revealed that rather simple penalty-based evaluations efficiently captured dependence on position of the mismatch in the alignment from the PAM sequence (Figure 5b) and dependence of the off-target effects on number of alignments per gRNA (Figure 5a). Also, the estimations were in strong correlation with empirical data on the off-target effects obtained from 'conventional' CRISPR screens (Figure 5c). Note that because of lack of comparable experimental data, in the current version, *beditor* score attempts capture the general rules using custom rules. In future developments, we wish to improve the specificity and predictability of *beditor* score using data from large-scale base editing experiments. The *beditor* workflow is implemented as an open-source python 3.6 package (hosted at https://pypi.org/project/beditor) allowing possibilities for future developments and improvements.

Together, considering the wide interest in scar-free and precise mutagenesis endowed by CRISPR-mediated base editing, moving ahead, novel and comprehensive gRNA designing workflow of *beditor* is expected to provide applicability to broad community of researchers and possibly become an essential component of the CRISPR-mediated base editing technology itself.

DATA AVAILABILITY

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).

SUPPLEMENTARY DATA

Processed data from this study i.e. gRNA libraries designed for demonstrative analysis 1, 2 and the case study analysis are provided as Supplementary data 1, 2 and 3 respectively.

ACKNOWLEDGEMENT

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 the Yachie Lab for informative discussions.

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.], and the Japan Society for the Promotion of Science [S15734 and S17161 to C.R.L. and N.Y.].

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

REFERENCES

1. Nishida,K., Arazoe,T., Yachie,N., Banno,S., Kakimoto,M., Tabata,M., Mochizuki,M., Miyabe,A., Araki,M., Hara,K.Y., *et al.* (2016) Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. **8729**, 1–14.

2. Gaudelli,N.M., Komor,A.C., Rees,H.A., Packer,M.S., Badran,A.H., Bryson,D.I. and Liu,D.R. (2017) Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. *Nature*, **551**, 464–471.

3. Kim,J.-S. (2018) Precision genome engineering through adenine and cytosine base editing. *Nat. plants*, **4**, 148–151.

4. Hess,G.T., Tycko,J., Yao,D. and Bassik,M.C. (2017) Methods and Applications of CRISPR-Mediated Base Editing in Eukaryotic Genomes. *Mol. Cell*, **68**, 26–43.

5. Rees,H.A. and Liu,D.R. (2018) Base editing: precision chemistry on the genome and transcriptome of living cells. *Nat. Rev. Genet.*, 10.1038/s41576-018-0059-1.

6. Komor,A.C., Kim,Y.B., Packer,M.S., Zuris,J.A. and Liu,D.R. (2016) Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, **533**, 420.

7. Hess,G.T., Frésard,L., Han,K., Lee,C.H., Li,A., Cimprich,K.A., Montgomery,S.B. and Bassik,M.C. (2016) Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. *Nat. Methods*, **13**, 1036–1042.

8. Ma,Y., Zhang,J., Yin,W., Zhang,Z., Song,Y. and Chang,X. (2016) Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. *Nat. Methods*, **13**, 1029–1035.

9. Cox,D.B.T., Gootenberg,J.S., Abudayyeh,O.O., Franklin,B., Kellner,M.J., Joung,J. and Zhang,F. (2017) RNA editing with CRISPR-Cas13 David. *Science (80-.).*, **0180**, 1–15.

10. Zafra,M.P., Schatoff,E.M., Katti,A., Foronda,M., Breinig,M., Schweitzer,A.Y., Simon,A., Han,T., Goswami,S., Montgomery,E., *et al.* (2018) Optimized base editors enable efficient editing in cells, organoids and mice. *Nat. Biotechnol.*, 10.1038/nbt.4194.

11. Zafra,M.P., Schatoff,E.M., Katti,A., Foronda,M., Breinig,M., Schweitzer,A.Y., Simon,A., Han,T., Goswami,S., Montgomery,E., *et al.* (2018) Optimized base editors enable efficient editing in cells, organoids and mice. *Nat. Biotechnol.*, **36**.

12. Shimatani,Z., Kashojiya,S., Takayama,M., Terada,R., Arazoe,T., Ishii,H., Teramura,H., Yamamoto,T., Komatsu,H., Miura,K., *et al.* (2017) Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion. *Nat. Biotechnol.*, 10.1038/nbt.3833.

13. Tang,W. and Liu,D.R. (2018) Rewritable multi-event analog recording in bacterial and

mammalian cells. *Science (80-.).*, **10**, eaap8992.

14. Qin,W., Lu,X. and Lin,S. (2018) Programmable base editing in zebrafish using a modified CRISPR-Cas9 system. *Methods*, 10.1016/j.ymeth.2018.07.010.

15. Rossidis,A.C., Stratigis,J.D., Chadwick,A.C., Hartman,H.A., Ahn,N.J., Li,H., Singh,K., Coons,B.E., Li,L., Lv,W., *et al.* (2018) In utero CRISPR-mediated therapeutic editing of metabolic genes. *Nat. Med.*, **24**, 1513–1518.

16. Villiger,L., Grisch-Chan,H.M., Lindsay,H., Ringnalda,F., Pogliano,C.B., Allegri,G., Fingerhut,R., Häberle,J., Matos,J., Robinson,M.D., *et al.* (2018) Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. *Nat. Med.*, 10.1038/s41591-018-0209-1.

17. Billon,P., Bryant,E.E., Joseph,S.A., Nambiar,T.S., Hayward,S.B., Rothstein,R. and Ciccia,A. (2017) CRISPR-Mediated Base Editing Enables Efficient Disruption of Eukaryotic Genes through Induction of STOP Codons. *Mol. Cell*, **67**, 1068–1079.e4.

18. Nishimasu,H., Yamano,T., Gao,L., Zhang,F., Ishitani,R. and Nureki,O. (2017) Structural Basis for the Altered PAM Recognition by Engineered CRISPR-Cpf1. *Mol. Cell*, 10.1016/j.molcel.2017.04.019.

19. Zerbino,D.R., Achuthan,P., Akanni,W., Amode,M.R., Barrell,D., Bhai,J., Billis,K., Cummins,C., Gall,A., Girón,C.G., *et al.* (2018) Ensembl 2018. *Nucleic Acids Res.*, 10.1093/nar/gkx1098.

20. PyEnsembl (2018) *https://github.com/openvax/pyensembl*.

21. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, 10.1093/bioinformatics/btq033.

22. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 10.1093/bioinformatics/btp324.

23. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

24. Concordet,J.-P. and Haeussler,M. (2018) CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Res.*, **46**, W242–W245.

25. Doench,J.G., Fusi,N., Sullender,M., Hegde,M., Vaimberg,E.W., Donovan,K.F., Smith,I., Tothova,Z., Wilen,C., Orchard,R., *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.

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

27. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O., *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.

28. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science (80-.).*, 10.1126/science.1225829.

29. Hu,J.H., Miller,S.M., Geurts,M.H., Tang,W., Chen,L., Sun,N., Zeina,C.M., Gao,X., Rees,H.A., Lin,Z., *et al.* (2018) Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. *Nature*, **556**, 57–63.

30. Nishimasu,H., Shi,X., Ishiguro,S., Gao,L., Hirano,S., Okazaki,S., Noda,T., Abudayyeh,O.O., Gootenberg,J.S., Mori,H., *et al.* (2018) Engineered CRISPR-Cas9 nuclease with expanded targeting space. *Science (80-.).*, **9129**, eaas9129.

TABLE AND FIGURES LEGENDS

Figure 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 mutagenesis potential.

a Information on the type and location of desired mutations is 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. Among many base editor and Pam sequence specific requirements, nucleotide windows for maximum activity are considered while designing the gRNAs. Finally, potential off-target effects are estimated.

b A scoring system specifically designed for *a priori* evaluation of mutagenesis potential of gRNAs. 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.

Figure 2. Demonstrative analysis of gRNAs designed with custom base editors and PAM recognition sequences.

In order to demonstrate the utility of *beditor* in utilizing custom base editors and PAM recognition sequences, sets of 1000 randomly assigned mutations in *S. cerevisiae* (see Supplementary methods) were analyzed in 2 mutation formats (nucleotide or amino acid) and 2 modes of mutagenesis ("create" and "remove") (each combination is shown in panel **a** to **d**). In each heatmap, number of gRNAs designed by each combination of a base editor (in columns) and PAM recognition sequence (in rows) is shown.

Figure 3. Case study analysis of clinically relevant human SNPs.

For the case study analysis, 2 base-editors (Target-AID and ABE) and 2 PAM sequences (NGG and NG) were used. Number of gRNAs designed using each mutation format i.e. nucleotide (**a**) and amino acid mutation (**b**) data are shown.

c Representative summary visualization of gRNA libraries designed with Target-AIDbase editor. Nucleotide composition of the gRNAs is shown along the length of the gRNAs. gRNAs are grouped by the position of the editable nucleotides within the activity window of a BE (shown in the rows).

Figure 4. Percentage editability of gRNA libraries designed in case study analysis of clinically relevant human SNPs.

Percentage of substitutions that can be edited by gRNA library (% editability) designed for case study analysis of clinically relevant human SNPs, in the format of nucleotide (**a** and **b**) and amino acid mutations (**c** and **d**) (see Supplementary methods). Also, the gRNA libraries were designed to remove mutations i.e. 'remove' mode (**a** and **c**) and to introduce mutations i.e. 'create' mode (**b** and **d**). Mapped on the heatmaps is a ratio

between number of substitutions that can be edited with the designed gRNAs and the number of substitutions 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 gRNA is designed on +, - and both strands respectively. Shown in gray are substitutions that are absent in the input data. * indicates non-sense mutation.

Figure 5. Performance assessment of beditor scores from case study analysis of clinically relevant human SNPs.

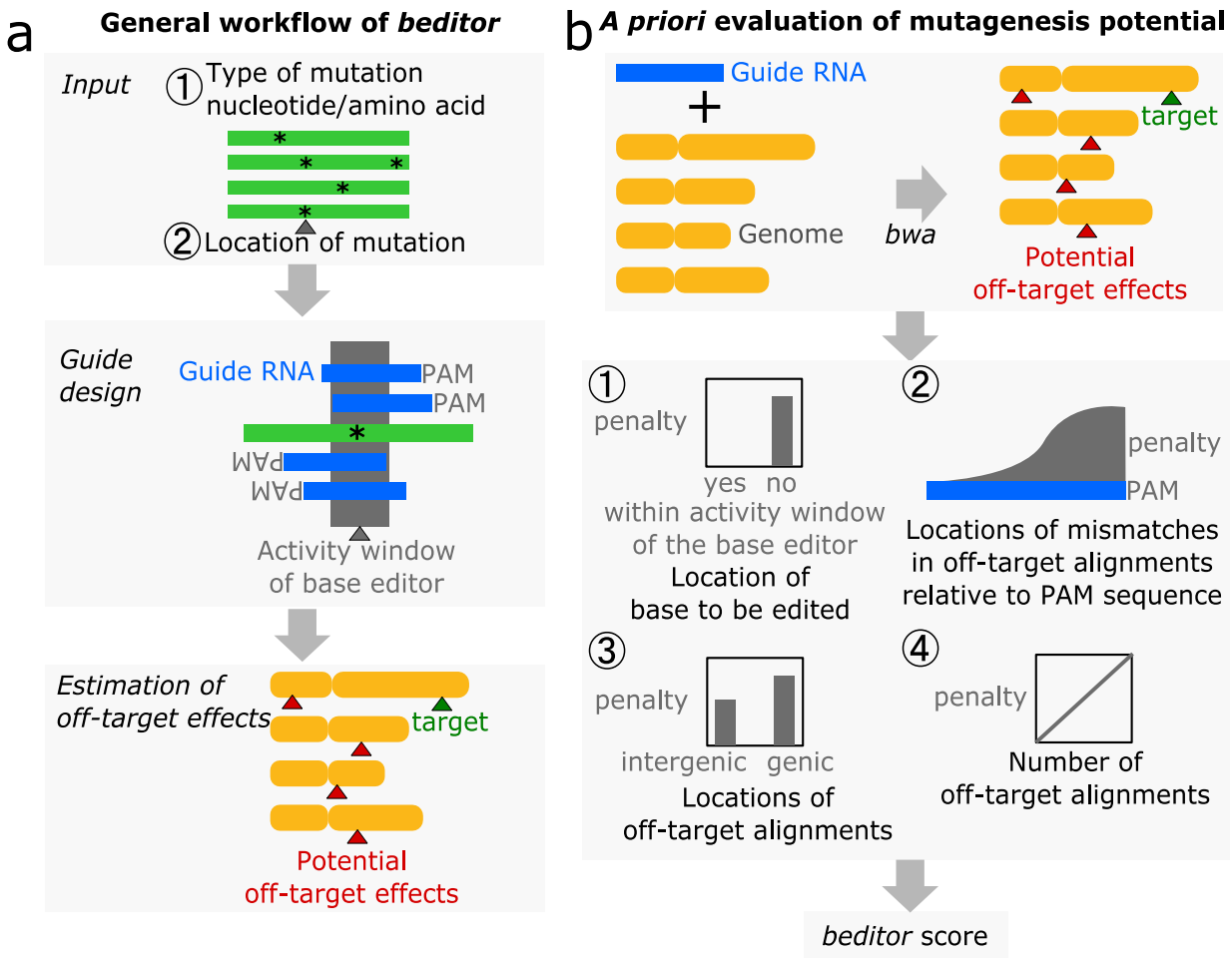
a Relationship between the number of genome-wide off-target alignments and *beditor* score per gRNA. The color of hexbins are scaled according to the number of gRNAs per bin.

b 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.

c 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.

1 **Figure 1**

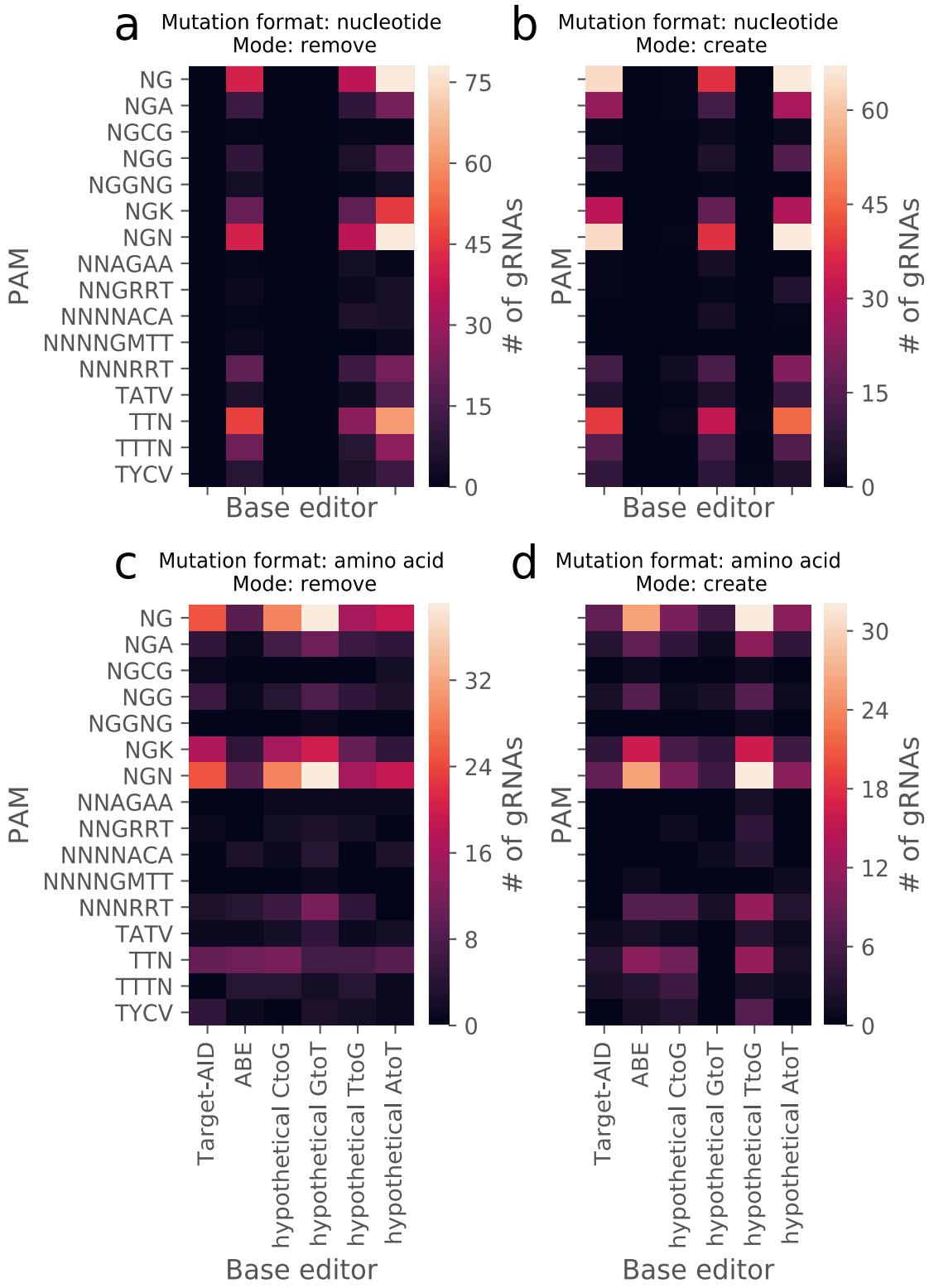


2

3

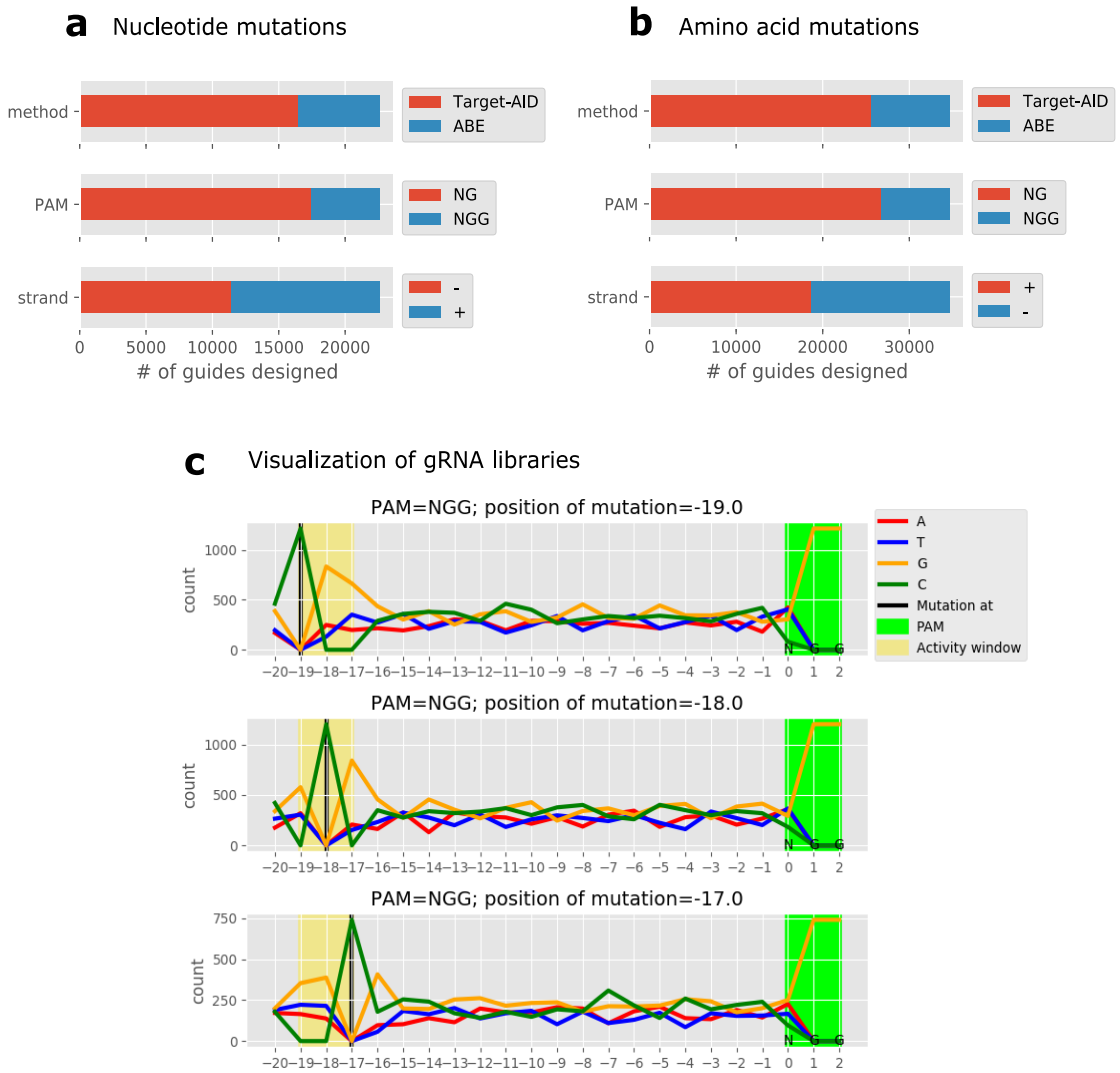
4

1 **Figure 2**

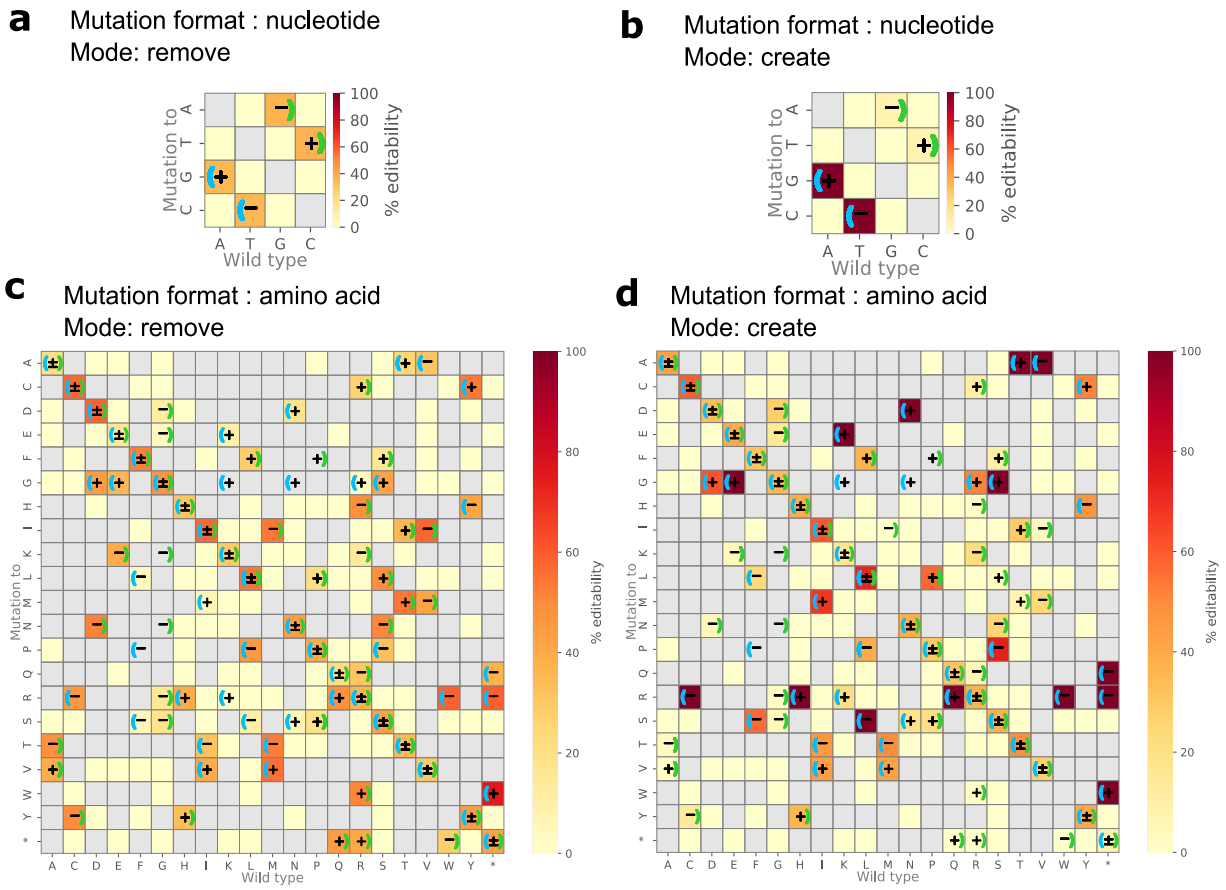

a
Mutation format: nucleotide
Mode: remove
b
Mutation format: nucleotide
Mode: create
c
Mutation format: amino acid
Mode: remove
d
Mutation format: amino acid
Mode: create
PAM
NG
NGA
NGCG
NGG
NGGNG
NGK
NGN
NNAGAA
NNGRRT
NNNNACA
NNNNGMTT
NNNRRT
TATV
TTN
TTTN
TYCV
# of gRNAs
Target-AID
ABE
hypothetical CtoG
hypothetical GtoT
hypothetical TtoG
hypothetical AtoT
Base editor

2

Figure 3


a Nucleotide mutations
method
PAM
strand
Target-AID
ABE
NG
NGG
-
+
0
5000
10000
15000
20000
# of guides designed
b Amino acid mutations
method
PAM
strand
Target-AID
ABE
NG
NGG
+
-
0
10000
20000
30000
# of guides designed
c Visualization of gRNA libraries
PAM=NGG; position of mutation=-19.0
PAM=NGG; position of mutation=-18.0
PAM=NGG; position of mutation=-17.0
count
A
T
G
C
Mutation at
PAM
Activity window

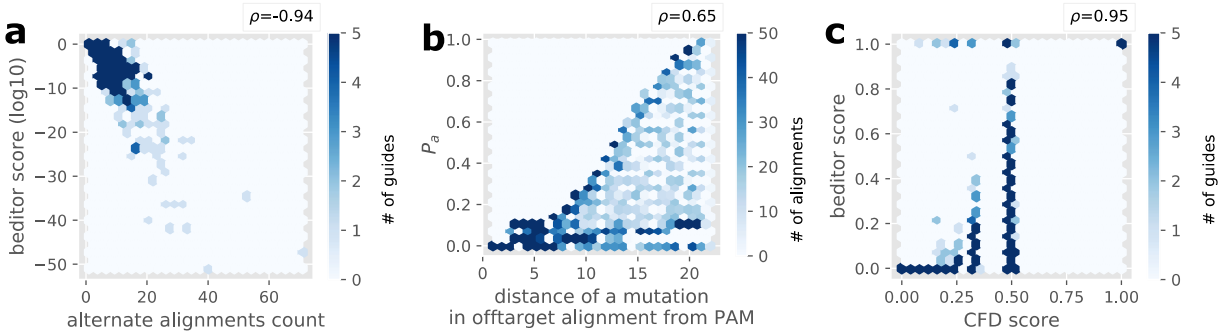
1 **Figure 4**


a
Mutation format : nucleotide
Mode: remove
b
Mutation format : nucleotide
Mode: create
c
Mutation format : amino acid
Mode: remove
d
Mutation format : amino acid
Mode: create
Mutation to
Wild type
% editability

2

3

1 **Figure 5**


a
ρ=-0.94
beditor score (log10)
alternate alignments count
# of guides
b
ρ=0.65
P_a
distance of a mutation
in offtarget alignment from PAM
# of alignments
c
ρ=0.95
beditor score
CFD score
# of guides

2

Supporting information

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

Supporting figures 2
Supporting tables 6
References 14

Supporting figures

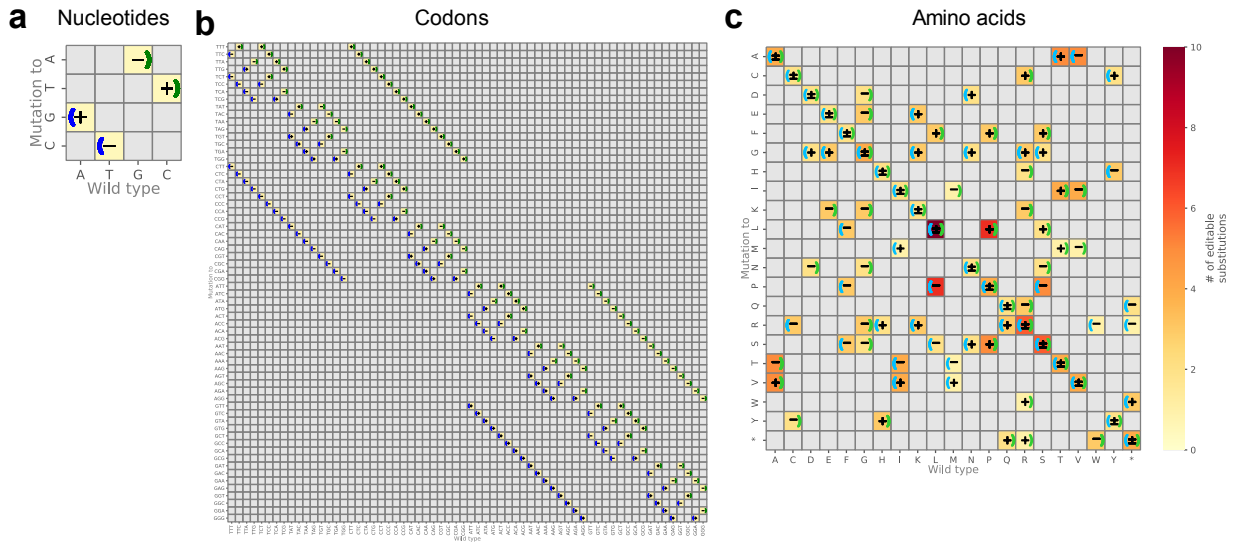


FIg S1 : Editable substitutions by ABE and Target-AID.

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.

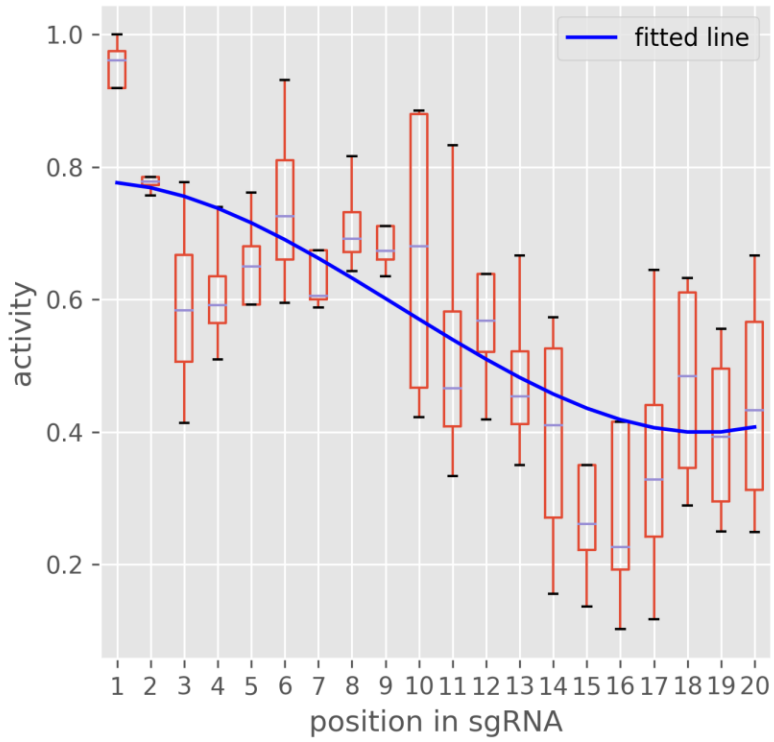


Fig S2: Fitting a third degree polynomial equation to the mismatch tolerance data from Doench et. al (3).

gRNA activity values of and at each mismatch position are shown as boxplots (red). The third degree polynomial equation fitted line is shown in blue.

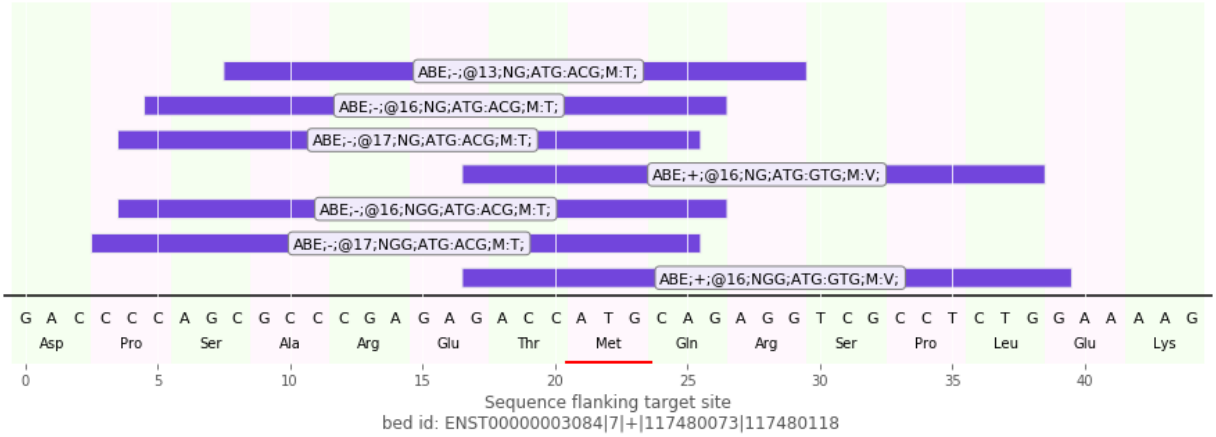


Fig S3: Representative visualizations of alignment between a gRNA and target genomic sequence.

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.

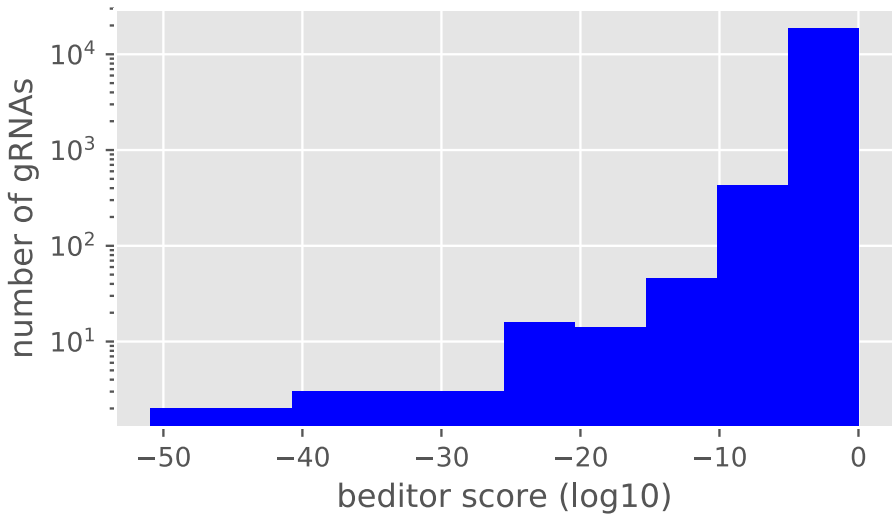


Fig S4: Distribution of beditor scores for the gRNA library designed for the demonstrative analysis of clinically associated human nucleotide mutations in 'create' mode.

Supporting tables

Table S1: Comparison of features of *beditor* with other existing tools.

Feature	***beditor***	**iSTOP**[1]	**BE-analyzer**[2]
Number of species whose genomes are directly compatible	>125	~70	~40
Searches sites editable by base editors	Yes	Yes	Yes
Designs gRNAs against predefined set of nucleotide mutations	Yes	No	No
Designs gRNAs against predefined set of amino acid mutations	Yes	No	No
Designs gRNAs for mutational scanning using a predefined amino acid substitution matrix	Yes	No	No
Designs gRNAs against non-sense mutations	Yes	Yes	No
Scale of gRNA libraries	multiple target sites	multiple target sites	1 target site at a time
Multithreading/ parallel computing	Yes	No	No
Allows addition of any novel/custom PAM recognition sequence	Yes	No	No
Allows addition of any novel/custom base editors	Yes	No	No
Estimation of off-target effects	Yes	No	No
Integrates independent measures of off-target effects	Yes (CFD scores[3])	No	No

Table S2: Default PAM recognition sequences supported by beditor.

PAM	Description	position	length of guide sequence	Reference
NG	xCas9	3'	20	[4]
NGA	SpCas9 mutant	3'	20	[5]
NGCG	SpCas9 mutant	3'	20	[5]
NGG	SpCas9	3'	20	[6,7]
NGGNG	Cas9 *S. Thermophilus*	3'	20	[8,9]
NGK	xCas9	3'	20	[4]
NGN	xCas9	3'	20	[4]
NNAGAA	Cas9 *S. Thermophilus*	3'	20	[8,9]
NNGRRT	SaCas9	3'	21	[10]
NNNNACA	CjCas9	3'	20	[11]
NNNNGMTT	Cas9 *N. Meningitidis*	3'	20	[12]
NNNRRT	KKH SaCas9	3'	21	[5,10]
TATV	AsCpf1 mutant	5'	23	[13]
TTN	Cpf1 *F. Novicida*	5'	23	[14]
TTTN	Cpf1 *Acidaminococcus / Lachnospiraceae*	5'	23	[14]
TYCV	TYCV AsCpf1 mutant	5'	23	[13]

Table S3: Penalties used in calculation of *beditor* score.

Penalty	Value
P_{min}	0.9
P_{max}	0.1
G_g	0.5
G_{ig}	0.9
A	0.5

Table S4: Summary statistics of demonstrative analysis with custom base editors and PAM recognition sequences.

	nucleotide	amino acid	nucleotide	amino acid
Remove/create mutation	remove	remove	create	create
Total number of mutations in the input data	1000	1000	1000	1000
Total number of mutations edited	263	117	246	105
Total number of guides designed	817	562	791	403
% editability	26.3	11.7	24.6	10.5
number of guides per mutation	3.10646	4.80342	3.21545	3.8381

Table S5: Summary statistics of demonstrative analysis of representative set of species.

Species	metrics	nucleotide	amino acid	nucleotide	amino acid
	Remove/create mutation	remove	remove	create	create
	Total number of mutations in the input data	1000.00	1000.00	1000.00	1000.00
Bos taurus	Total number of mutations edited	76.00	45.00	69.00	36.00
	Total number of guides designed	123.00	83.00	121.00	62.00
	% editability	7.60	4.50	6.90	3.60
	number of guides per mutation	1.62	1.84	1.75	1.72
Danio rerio	Total number of mutations edited	45.00	39.00	57.00	43.00
	Total number of guides designed	71.00	59.00	74.00	72.00
	% editability	4.50	3.90	5.70	4.30
	number of guides per mutation	1.58	1.51	1.30	1.67
Equus caballus	Total number of mutations edited	80.00	53.00	55.00	43.00
	Total number of guides designed	136.00	94.00	92.00	81.00
	% editability	8.00	5.30	5.50	4.30
	number of guides per mutation	1.70	1.77	1.67	1.88
Felis catus	Total number of mutations edited	94.00	55.00	81.00	50.00
	Total number of guides designed	168.00	103.00	160.00	80.00
	% editability	9.40	5.50	8.10	5.00
	number of guides per mutation	1.79	1.87	1.98	1.60
Gallus gallus	Total number of mutations edited	93.00	39.00	74.00	46.00
	Total number of guides designed	180.00	71.00	139.00	77.00
	% editability	9.30	3.90	7.40	4.60
	number of guides per mutation	1.94	1.82	1.88	1.67
Macaca fascicularis	Total number of mutations edited	45.00	51.00	52.00	48.00
	Total number of guides designed	89.00	83.00	90.00	75.00
	% editability	4.50	5.10	5.20	4.80
	number of guides per mutation	1.98	1.63	1.73	1.56
Mus musculus	Total number of mutations edited	81.00	46.00	74.00	47.00
	Total number of guides designed	147.00	93.00	133.00	83.00
	% editability	8.10	4.60	7.40	4.70
	number of guides per mutation	1.81	2.02	1.80	1.77
Pan paniscus	Total number of mutations edited	73.00	57.00	62.00	54.00
	Total number of guides designed	123.00	89.00	98.00	90.00
	% editability	7.30	5.70	6.20	5.40
	number of guides per mutation	1.68	1.56	1.58	1.67
Saccharomyces cerevisiae	Total number of mutations edited	36.00	30.00	36.00	35.00
	Total number of guides designed	52.00	45.00	46.00	52.00

	% editability	3.60	3.00	3.60	3.50
	number of guides per mutation	1.44	1.50	1.28	1.49
Sus scrofa	Total number of mutations edited	68.00	50.00	55.00	45.00
	Total number of guides designed	108.00	106.00	86.00	82.00
	% editability	6.80	5.00	5.50	4.50
	number of guides per mutation	1.59	2.12	1.56	1.82

Table S6: Summary statistics for case study analysis of clinically relevant human SNPs.

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 S7: Input parameters used for the demonstrative analysis of clinically associated human genetic variants

Variable	Input	Description
Host		Name of host organism
genomerelease	93	Ensembl genome release
genomeassembly		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. 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]. Elsevier Inc.; 2017;67:1068–1079.e4. Available from: http://dx.doi.org/10.1016/j.molcel.2017.08.008

2. 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

3. 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]. Nature Publishing Group; 2016;34:184–91. Available from: http://dx.doi.org/10.1038/nbt.3437

4. 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

5. 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;

6. 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.

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 (80-). 2012;

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

9. 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;

10. 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;

11. 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;

12. 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;

13. 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

14. 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;