beditor: A Computational Workflow for Designing Libraries of Guide RNAs for CRISPR-Mediated Base Editing

Rohan Dandage,*,†,‡,§,1 Philippe C. Després,*,†,‡ Nozomu Yachie,,††,‡‡,§§ and Christian R. Landry*,†,‡,§**

*Département de Biochimie, Microbiologie et Bio-informatique and §Département de Biologie, Faculté de Sciences et Génie, Université Laval, Québec, Québec G1V 0A6, Canada, †PROTEO, The Québec Research Network on Protein Function, Structure and Engineering, ‡Centre de Recherche en Données Massives (CRDM), and **Research Center for Advanced Science and Technology, University of Tokyo, 153-8904, Japan, ††Department of Biological Sciences, Graduate School of Science, University of Tokyo, 113-0033, Japan, ‡‡Institute for Advanced Biosciences, Keio University, Tsuruoka, 997-0035, Japan, and §§Graduate School of Media and Governance, Keio University, Fujisawa, 252-0882, Japan

ORCID IDs: 0000-0002-6421-2067 (R.D.); 0000-0003-1407-0643 (P.C.D.); 0000-0003-1582-6027 (N.Y.); 0000-0003-3028-6866 (C.R.L.)

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

KEYWORDS CRISPR; base editing; gene editing; genome-wide mutagenesis; gRNA design

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 (Nishida *et al.* 2016; Gaudelli *et al.* 2017; Hess *et al.* 2017; Kim 2018; Rees and Liu 2018). Currently, two major types of BEs are available: cytosine base editors (CBEs) that enable the conversion of cytosine to uracil by catalysis and then to thymine through replication or repair (Hess *et al.* 2016; Komor *et al.* 2016; Ma *et al.* 2016; Nishida *et al.* 2016), for example BE3 (Komor *et al.* 2016), and Target-AID (C•G to T•A) (Nishida *et al.* 2016). Similarly, adenine base editors (ABEs) enable conversion of adenine to inosine by catalysis and then to guanine through replication or repair (A•T to G•C) (Cox *et al.* 2017; Gaudelli *et al.* 2017; Zafra *et al.* 2018), for example ABE7.10 (Gaudelli *et al.* 2017). Currently BEs enable many codon level substitutions and thus amino acid substitutions (Supplemental Material, Figure S1). Owing to the unique capability of scar-free mutagenesis, BEs have recently found numerous applications in both model and nonmodel organisms (Shimatani *et al.* 2017; Kim 2018; Qin *et al.* 2018; Tang and Liu 2018; Zafra *et al.* 2018) and have substantial promise in therapeutic applications (Rossidis *et al.* 2018; Villiger *et al.* 2018).

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 design of gRNA sequences is a primary requirement for the success of the BE-mediated

doi: https://doi.org/10.1534/genetics.119.302089
Manuscript received December 11, 2018; accepted for publication March 28, 2019; published Early Online April 1, 2019.
Available freely online through the author-supported open access option.

Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7448234.
[1]Corresponding author: Université Laval, 1045, Ave. de la Médecine, Québec City, QC G1R1X6, Canada. E-mail: rohan.dandage.1@bio.ulaval.ca

mutagenesis experiment, the methods and strategies involved in designing gRNA libraries are arguably one of the most important factors in such experiments.

Currently available gRNA designing tools are however either specifically focused on nonsense mutations (Billon *et al.* 2017) or allow limited customization (BioRxiv: https://doi.org/10.1101/373944) (see Table S1), leaving a major roadblock in the applications of the BEs in genome editing. The continuous discoveries of new BEs and the expansion of their existing editing capabilities (Nishimasu *et al.* 2017) demand a complementary development in computational methods that would allow designing gRNAs using new and improved BEs. Moreover, considering the prospective applications across nonmodel organisms, compatibility with diverse genomes is essential. Overall, therefore, a comprehensive computational framework to design gRNA 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 enzymes. 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 >125 genomes hosted in the Ensembl genome database (Zerbino *et al.* 2018) and any annotated custom genomes. Additionally, the *beditor* workflow 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. With its Graphical User Interface (GUI), command line interface, and open source Application programming interface (API), 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 (PyEnsembl 2018), BEDTools (Quinlan and Hall 2010), BWA (Li and Durbin 2009), and SAMtools (Li *et al.* 2009) 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 (Quinlan and Hall 2010). The designed gRNAs are aligned with the reference genome using BWA (Li and Durbin 2009), and alignments are processed using SAMtools (Li *et al.* 2009) 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 Li and Durbin (2009), allowing for a maximum of two mismatches per alignment (Concordet and Haeussler 2018). The *beditor* score is evaluated as follows:

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

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

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

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

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 Doench *et al.* (2016) (Figure S2). This way, penalties increase nonlinearly from $P_{\min}$ to $P_{\max}$, as the distance of nucleotide (i) from PAM sequence increases. Individual penalties assigned for all the mismatched nucleotides in a gRNA (total M_{max}) are then multiplied to estimate a penalty score for a given alignment called P_a (Equation 2). In cases of gRNAs 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 S2). 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)

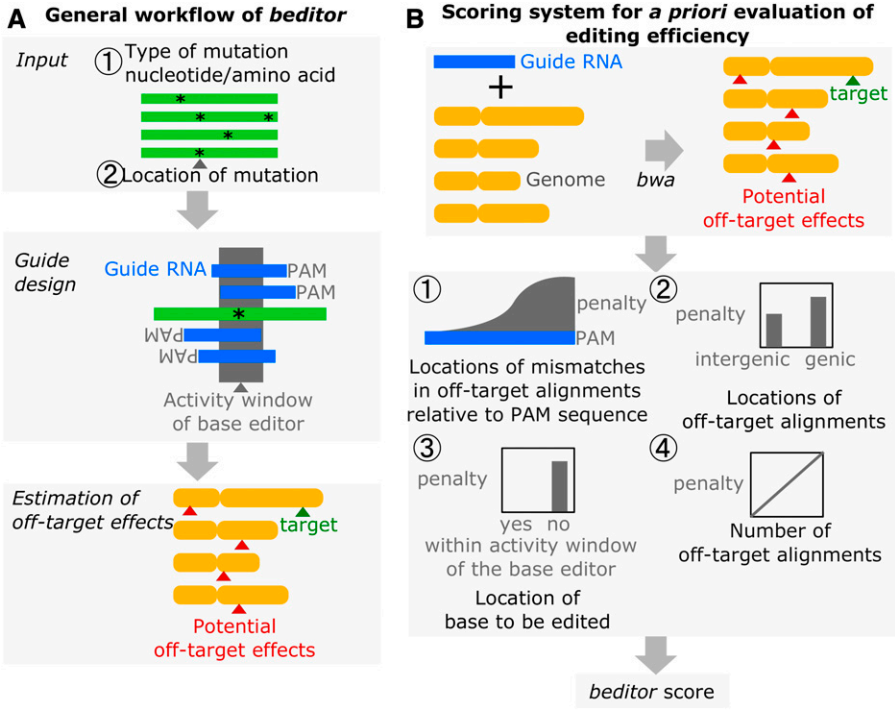

Figure 1 The computational workflow of *beditor* allows for the flexible design of gRNA libraries to be used in CRISPR base editing and offers *a priori* evaluation of mutagenesis potential. (A) Information on the type and location of desired mutations is supplied to the *beditor* workflow as a tab separated 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 the above penalties, a final score is calculated for each gRNA sequence – the *beditor* score.

(Equation 4). Multiplication of individual penalties ensures that if *any* of the criteria are suboptimal, the *beditor* score decreases.

Demonstrative analysis 1: customizability with respect to base editing strategies

As a demonstrative analysis using custom BEs and PAM recognition sequences (Figure 2 and Table S3), 1000 nucleotide and amino acid mutations were randomly assigned across the genome of *Saccharomyces cerevisiae* (https://github.com/rraadd88/test_beditor). Such sets of mutations create uniform data sets ideal for testing features of *beditor*. The input mutation data were created for both mutation formats (either amino acid or nucleotide) and modes of mutagenesis ("model" or "correct"). The command "beditor–cfg params.yml" was executed. Here, params.yml contains input parameters of the analysis (Table S4) 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 S5 and gRNA libraries are included as Data S1.

Demonstrative analysis 2: ability to work with different species

As a demonstrative analysis with different species, sets of 1000 random mutations were created in genomes of 10 representative species (*Bos Taurus*, *Danio rerio*, *Equus caballus*, *Felis catus*, *Gallus gallus*, *Macaca fascicularis*, *Mus musculus*, *Pan paniscus*, *S. cerevisiae*, and *Sus scrofa*) as described in case of demonstrative analysis 1. The input parameters used in this analysis are the same as in Table S4 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 S6, and gRNA libraries are included as Data S2.

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 coordinates 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 (PyEnsembl 2018). The mutation information was provided to the *beditor* workflow as a tab-separated file. The input parameters used in this analysis are the same as Table S4 except for the host name, "homo_sapiens", and genome assembly version, "GrCh38." Summary statistics on the editability are included as Table S7, and gRNA libraries are included as Data S3.

Data availability

The authors affirm that all data necessary for confirming the conclusions of this article are represented fully within the article and its tables and figures.

Processed data from this study *i.e.*, gRNA libraries designed for demonstrative analysis 1, 2 and the case study analysis are provided as Data S1–S3, respectively. This data set has been deposited using the GSA Figshare portal.

The *beditor* software is available at https://github.com/rraadd88/beditor under the GNU General Public License

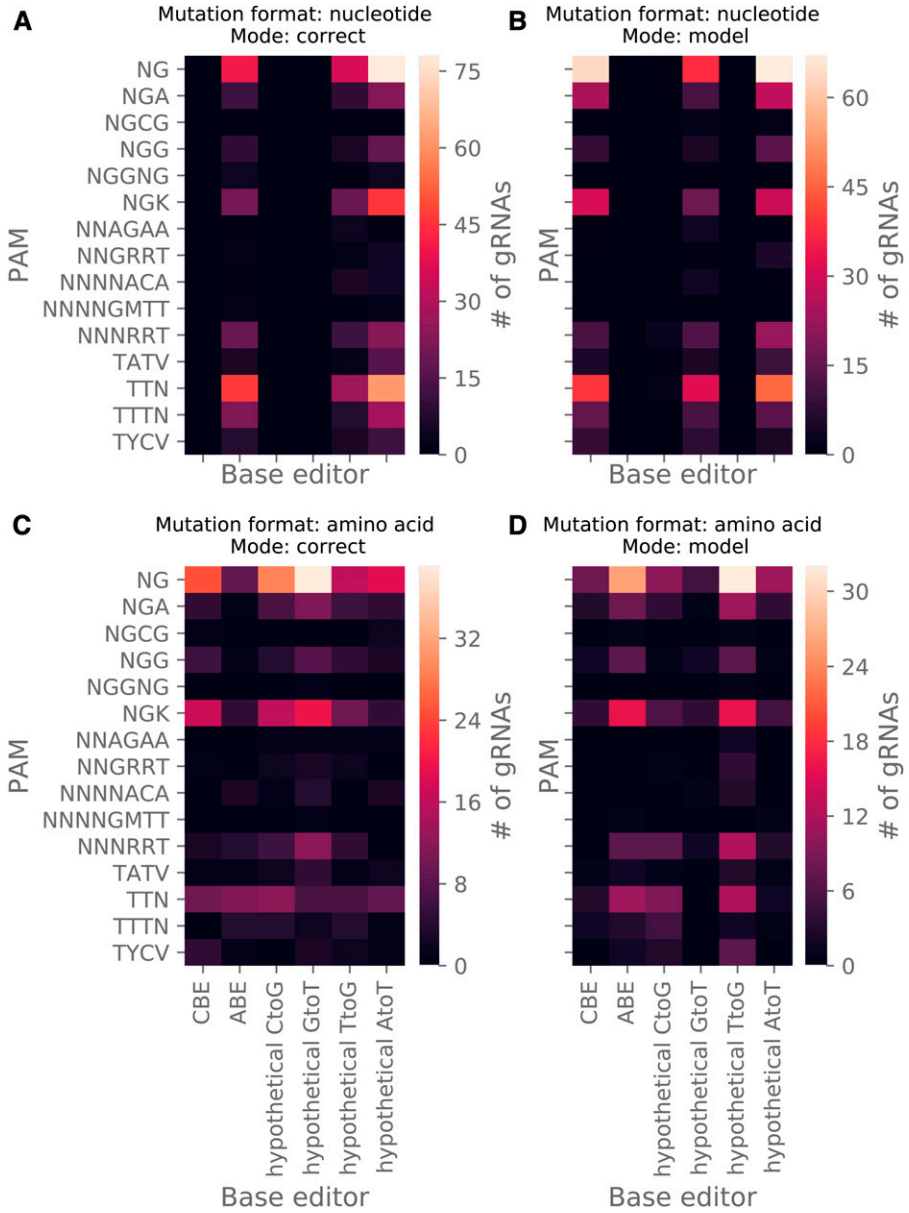

Figure 2 Demonstrative analysis of gRNAs designed with custom base editors and PAM recognition sequences. To demonstrate the utility of *beditor* in utilizing custom base editors and PAM recognition sequences, sets of 1000 randomly assigned mutations in *S. cerevisiae* (see Supplemental methods) were analyzed in two mutation formats (nucleotide or amino acid) and two modes of mutagenesis ("model" and "correct") (each combination is shown in parts A–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.

(GPLv3). The database of gRNAs designed through the *beditor* workflow can be accessed at http://rraadd88.github.io/soft/beditor.

The data set analyzed in the study, *i.e.*, set of clinically associated human mutations, was 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). Supplemental material available at Figshare: https://doi.org/10.25386/genetics.7448234.

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 (Figure 1A). 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. Users can also provide inputs to limit the amino acid substitutions to a custom substitution matrix and specify whether only

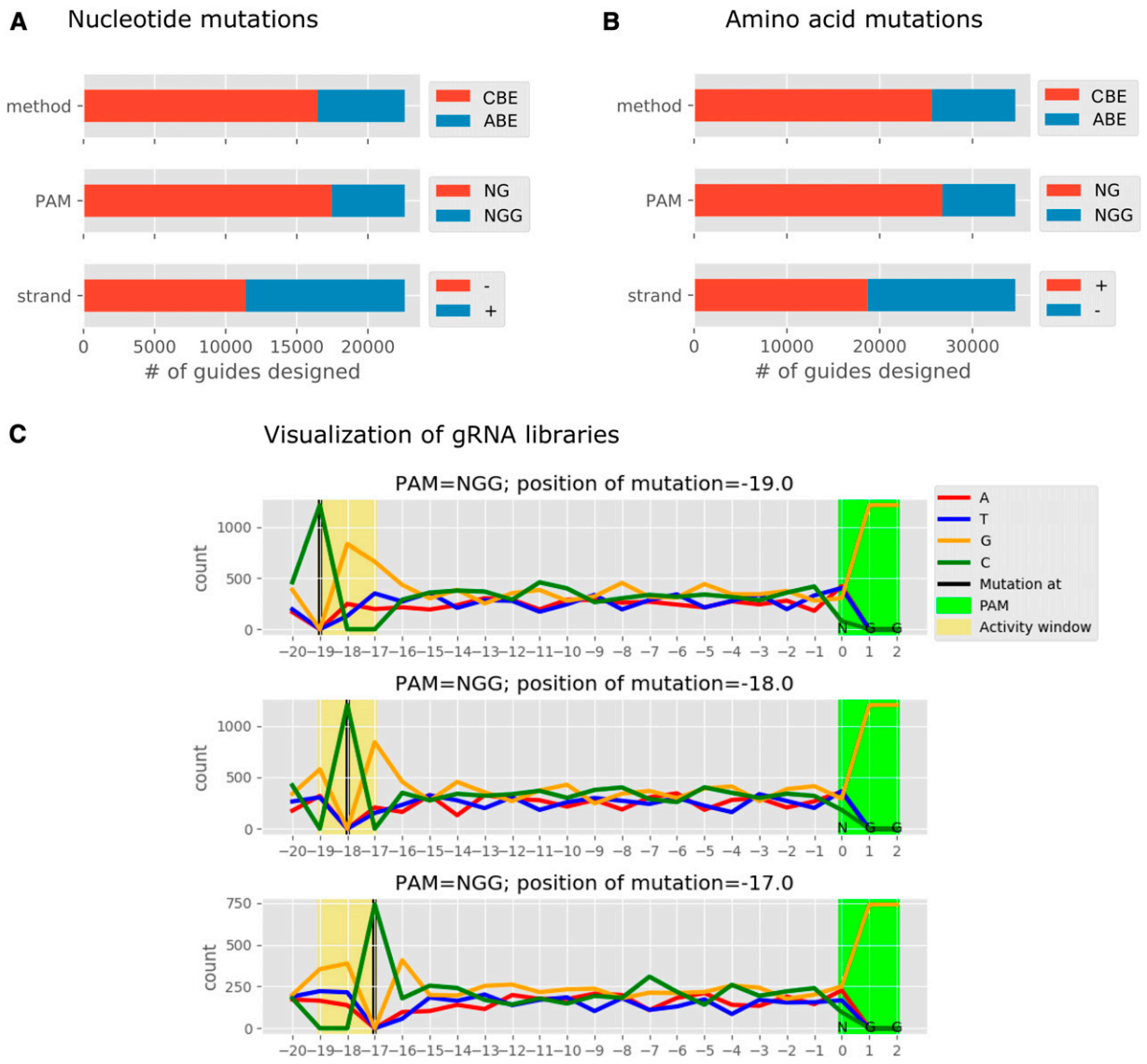

Figure 3 Case study analysis of clinically relevant human SNPs. For the case study analysis, two base editors (Target-AID and ABE) and two PAM sequences (NGG and NG) were used. Number of gRNAs designed using each mutation format, *i.e.*, nucleotide (A) and amino acid mutation (B) data are shown. (C) Representative summary visualization of gRNA libraries designed with the Target-AID base 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).

nonsynonymous or synonymous substitutions should be carried out. In addition to creating mutations on a wild-type background ("model" 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 ("correct" mode). The program can be accessed via GUI (Figure S3), command line, or API. In addition to the gRNAs designed to carry out provided mutations, the *beditor* workflow can also design control gRNAs that are important in the large-scale mutagenesis experiments. The positive control gRNAs designed by *beditor* generate nonsense mutations in the coding region of interest and negative control gRNAs that lack an editable nucleotide in the editing window of the BE, thus expected to have null effect.

Customizability for broad utility

The *beditor* workflow utilizes a PyEnsembl python API (PyEnsembl 2018) to fetch and work with the genomes of over 125 species and their various assemblies from the Ensembl genome database (Zerbino *et al.* 2018), 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 (Figure S4). The users can incorporate BEs with varied editing properties and even novel BEs as per requirements. Similarly, they can incorporate any custom PAM sequences (*e.g.*, Table S3) in addition to the experimentally validated PAMs already incorporated in the current version of *beditor* (listed in

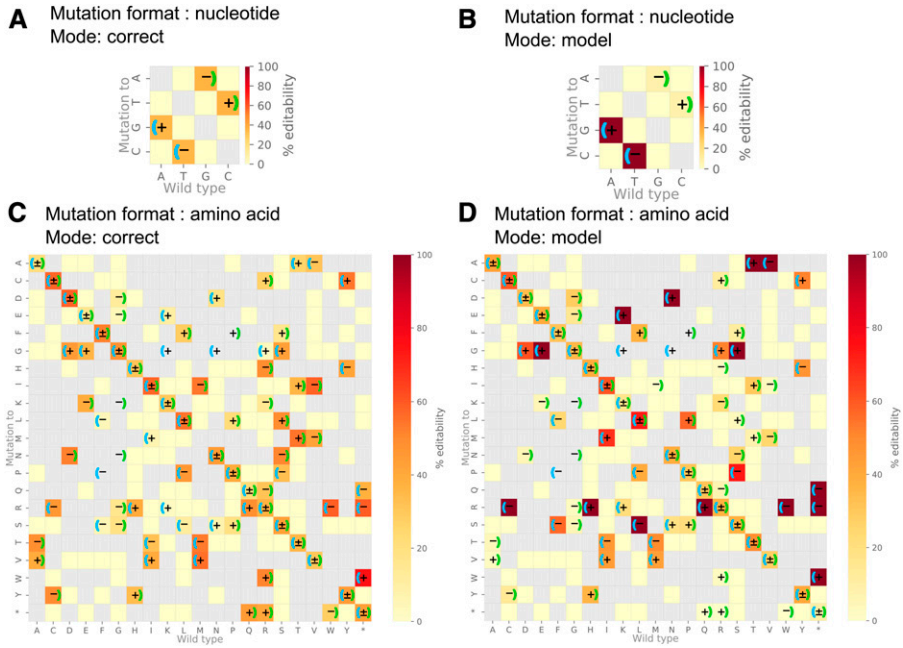

Figure 4 Percentage editability of gRNA libraries designed in case study analysis of clinically relevant human SNPs. Percentage of substitutions that can be edited by gRNA library (% editability) designed for case study analysis of clinically relevant human SNPs, in the format of nucleotide (A and B) and amino acid mutations (C and D) (see Supplemental methods). Also, the gRNA libraries were designed to remove mutations *i.e.*, "correct" mode (A and C) and to introduce mutations *i.e.*, "model" 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 nonsense mutation.

Table S8). Additionally, the location of the PAM with respect to the gRNA (upstream or downstream) and the provided length of gRNA are taken into consideration in the analysis. Lastly, *beditor* scores allow 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 a designed gRNA library. Due to the lack of large-scale, genome-wide base editing data, we relied on a 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 *Materials and Methods*). In the future, we wish to determine 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 penalty scores are assigned (Figure 1B, see *Materials and 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 (Doench *et al.* 2016) (see *Materials and Methods*). Given the lack of large-scale, 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, to account for the number of off-target sites, the three penalties are multiplied together to evaluate a *beditor* score per gRNA. Effectively, the optimal gRNAs have a *beditor* score of 1, while a lower *beditor* score indicates incompatibility with BE requirements, significant off-targets, or off-target effects that confound mutational effects. Finally, to filter out the gRNAs containing putative RNA polymerase III transcriptional terminators (Gao *et al.* 2018), the length of the poly-T stretch per individual gRNA is indicated in the output of *beditor*.

Demonstrative analysis 1: customizability with respect to base editing strategies

We demonstrate that the users can incorporate and use custom BEs and PAM recognition sequences by designing gRNA libraries against sets of 1000 randomly assigned mutations (Supplemental methods) with 6 BEs (among which 4 are not yet developed and are 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 ("model" and "correct"), gRNA libraries were designed (Figure 2 and Data S1).

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 10 representative

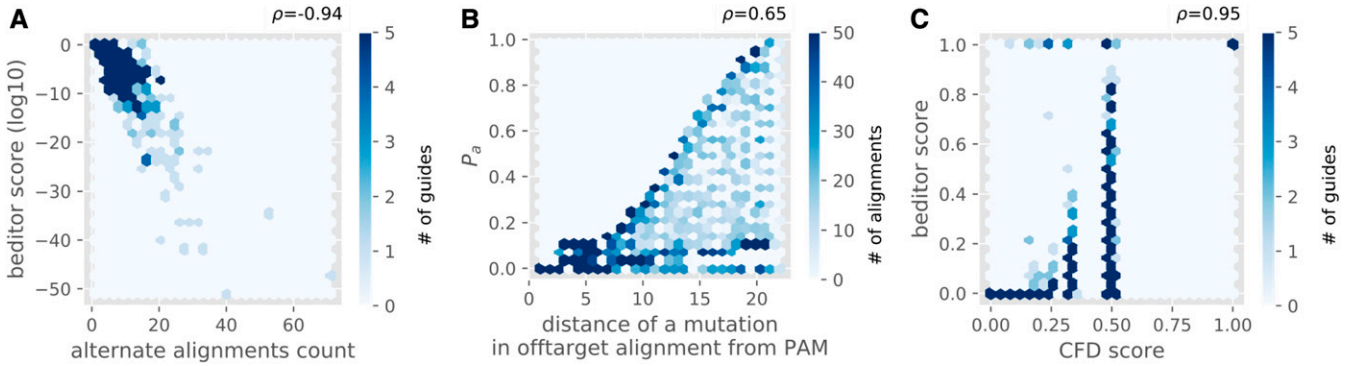

Figure 5 Performance assessment of *beditor* scores from case study analysis of clinically relevant human SNPs. (A) Relationship between the number of genome-wide off-target alignments and *beditor* score per gRNA. The color of hexbins are scaled according to the number of gRNAs per bin. (B) Relationship 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.

Ensembl genomes, for sets of 1000 randomly assigned mutations (Supplemental methods). The genomes of all the species were directly fetched from the Ensembl genome database and gRNA libraries were designed for both mutation formats (nucleotide and amino acid) and two modes of mutagenesis ("model" and "correct"). Through the designed gRNA libraries (Data S2), reasonable editability (Table S6) was achieved for all the species.

Case study: designing a gRNA library against a set of 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 61,083 nucleotide level and 81,819 amino acid level mutations (see Supplemental methods). This analysis was carried out with two different BEs: Target-AID and ABE, and two PAM sequences: NGG (Jinek *et al.* 2012; Hsu *et al.* 2013) and NG (Hu *et al.* 2018; Nishimasu *et al.* 2018) and in "model" and "correct" mode. Note that the purpose of this case study analysis is to provide a demonstration of the functionalities of the *beditor* workflow. For instance, ABE base editor is not known to function in association with NG PAM. Yet, as shown in demonstrative analysis 1, the users may try any combinations of custom BEs and PAM sequences. The output libraries of gRNA sequences (Data S3) target ~25% of the total mutations provided as input (Table S7). 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 3C), as well as visualizations of alignments of gRNAs with the target sequence (Figure S5), 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. The gRNAs designed for this case study analysis, as well as with all the combinations experimentally validated pairs of BEs and PAM sequences (Table S8), are provided as a database with a web interface at http://rraadd88.github.io/soft/beditor.

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 S6), the users may assign a threshold to filter out gRNAs with lower *beditor* scores. Collectively, by definition, the *beditor* scores are negatively correlated ($\rho = -0.94$) with the number of off-target alignments (Figure 5A), and the penalty assigned for each alignment based on distance of mismatches from the PAM sequence is positively correlated ($\rho = 0.65$) 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 ($\rho = 0.95$) determined in terms of Cutting Frequency Determination (CFD) score (Doench *et al.* 2016) (Figure 5C).

Discussion

CRISPR-mediated BEs have recently become a new paradigm in genome editing, owing to their unique ability to carry out precise mutations without the need for DNA breaks (Rees and Liu 2018). Consequently, a plethora of applications of BEs have emerged from all corners of the genome editing community, ranging from study of model and nonmodel organisms (Shimatani *et al.* 2017; Kim 2018; Qin *et al.* 2018; Tang and Liu 2018; Zafra *et al.* 2018) to therapeutics (Rossidis *et al.* 2018; Villiger *et al.* 2018). However, there has been a lack of robust and customizable 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 a 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 ("model" mode) or remove it ("correct" 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 ~60,000 human clinically relevant SNPs, we show that the *beditor* workflow provides all-round gRNA design capabilities, scanning through combinations of multiple strategies (Figure 3, A 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). This is supported by the currently available data (Kim *et al.* 2017), which suggests that the Cas9-induced off-target effects could be the predictors of off-target effects of BEs (Rees and Liu 2018). However, recent studies have found that the off-target effects of BEs can vary between ABE and CBE (Lee *et al.* 2018; Jin *et al.* 2019; Zuo *et al.* 2019) and could even be largely unpredictable as in the case of CBEs containing APOBEC1 (Jin *et al.* 2019; Zuo *et al.* 2019). Therefore, in future, we wish to update the *beditor* scoring system further, as more information would emerge from large-scale base editing experiments. The future developments of *beditor* workflow would be carried out in an open source manner at https://github.com/rraadd88/beditor.

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

Acknowledgments

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. 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.). The authors declare that they have no competing interests.

Literature Cited

Billon, P., E. E. Bryant, S. A. Joseph, T. S. Nambiar, S. B. Hayward *et al.*, 2017 CRISPR-mediated base editing enables efficient disruption of eukaryotic genes through induction of STOP codons. Mol. Cell 67: 1068–1079.e4. https://doi.org/10.1016/j.molcel.2017.08.008

Concordet, J.-P., and M. Haeussler, 2018 CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. Nucleic Acids Res. 46: W242–W245. https://doi.org/10.1093/nar/gky354

Cox, D. B. T., J. S. Gootenberg, O. O. Abudayyeh, B. Franklin, M. J. Kellner *et al.*, 2017 RNA editing with CRISPR-Cas13. Science 358: 1019–1027. https://doi.org/10.1126/science.aaq0180

Doench, J. G., N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg *et al.*, 2016 Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat. Biotechnol. 34: 184–191. https://doi.org/10.1038/nbt.3437

Gao, Z., E. Herrera-Carrillo, and B. Berkhout, 2018 Delineation of the exact transcription termination signal for type 3 polymerase III. Mol. Ther. Nucleic Acids 10: 36–44. https://doi.org/10.1016/j.omtn.2017.11.006

Gaudelli, N. M., A. C. Komor, H. A. Rees, M. S. Packer, A. H. Badran *et al.*, 2017 Programmable base editing of A•T to G•C in genomic DNA without DNA cleavage. Nature 551: 464–471 [corrigenda: Nature 559: E8 (2018)]. https://doi.org/10.1038/nature24644

Hess, G. T., L. Frésard, K. Han, C. H. Lee, A. Li *et al.*, 2016 Directed evolution using dCas9-targeted somatic hypermutation in mammalian cells. Nat. Methods 13: 1036–1042. https://doi.org/10.1038/nmeth.4038

Hess, G. T., J. Tycko, D. Yao, and M. C. Bassik, 2017 Methods and applications of CRISPR-mediated base editing in eukaryotic genomes. Mol. Cell 68: 26–43. https://doi.org/10.1016/j.molcel.2017.09.029

Hsu, P. D., D. A. Scott, J. A. Weinstein, F. A. Ran, S. Konermann *et al.*, 2013 DNA targeting specificity of RNA-guided Cas9 nucleases. Nat. Biotechnol. 31: 827–832. https://doi.org/10.1038/nbt.2647

Hu, J. H., S. M. Miller, M. H. Geurts, W. Tang, L. Chen *et al.*, 2018 Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature 556: 57–63. https://doi.org/10.1038/nature26155

Jin, S., Y. Zong, Q. Gao, Z. Zhu, Y. Wang *et al.*, 2019 Cytosine, but not adenine, base editors induce genome-wide off-target mutations in rice. Science. .https://doi.org/10.1126/science.aaw7166

Jinek, M., K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna *et al.*, 2012 A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337: 816–821. https://doi.org/10.1126/science.1225829

Kim, D., K. Lim, S. T. Kim, S. H. Yoon, K. Kim *et al.*, 2017 Genome-wide target specificities of CRISPR RNA-guided programmable deaminases. Nat. Biotechnol. 35: 475–480. https://doi.org/10.1038/nbt.3852

Kim, J.-S., 2018 Precision genome engineering through adenine and cytosine base editing. Nat. Plants 4: 148–151. https://doi.org/10.1038/s41477-018-0115-z

Komor, A. C., Y. B. Kim, M. S. Packer, J. A. Zuris, and D. R. Liu, 2016 Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature 533: 420–424. https://doi.org/10.1038/nature17946

Lee, H. K., M. Willi, S. M. Miller, S. Kim, C. Liu *et al.*, 2018 Targeting fidelity of adenine and cytosine base editors in mouse embryos. Nat. Commun. 9: 4804. https://doi.org/10.1038/s41467-018-07322-7

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Ma, Y., J. Zhang, W. Yin, Z. Zhang, Y. Song *et al.*, 2016 Targeted AID-mediated mutagenesis (TAM) enables efficient genomic diversification in mammalian cells. Nat. Methods 13: 1029–1035. https://doi.org/10.1038/nmeth.4027

Nishida, K., T. Arazoe, N. Yachie, S. Banno, M. Kakimoto *et al.*, 2016 Targeted nucleotide editing using hybrid prokaryotic and vertebrate adaptive immune systems. Science 353: aaf8729. https://doi.org/10.1126/science.aaf8729

Nishimasu, H., T. Yamano, L. Gao, F. Zhang, R. Ishitani *et al.*, 2017 Structural basis for the altered PAM recognition by engineered CRISPR-Cpf1. Mol. Cell 67: 139–147.e2. https://doi.org/10.1016/j.molcel.2017.04.019

Nishimasu, H., X. Shi, S. Ishiguro, L. Gao, S. Hirano *et al.*, 2018 Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science 361: 1259–1262. https://doi.org/10.1126/science.aas9129

PyEnsembl, 2018 PyEnsembl: Python interface to access reference genome features (such as genes, transcripts, and exons) from Ensembl. Accessed: Jul 31, 2018. Available at: https://github.com/openvax/pyensembl.

Qin, W., X. Lu, and S. Lin, 2018 Programmable base editing in zebrafish using a modified CRISPR-Cas9 system. Methods 150: 19–23. https://doi.org/10.1016/j.ymeth.2018.07.010

Quinlan, A. R., and I. M. Hall, 2010 BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842. https://doi.org/10.1093/bioinformatics/btq033

Rees, H. A., and D. R. Liu, 2018 Base editing: precision chemistry on the genome and transcriptome of living cells. Nat. Rev. Genet. 19: 770–788 (erratum: Nat. Rev. Genet. 19: 801). https://doi.org/10.1038/s41576-018-0059-1

Rossidis, A. C., J. D. Stratigis, A. C. Chadwick, H. A. Hartman, N. J. Ahn *et al.*, 2018 In utero CRISPR-mediated therapeutic editing of metabolic genes. Nat. Med. 24: 1513–1518. https://doi.org/10.1038/s41591-018-0184-6

Shimatani, Z., S. Kashojiya, M. Takayama, R. Terada, T. Arazoe *et al.*, 2017 Targeted base editing in rice and tomato using a CRISPR-Cas9 cytidine deaminase fusion. Nat. Biotechnol. 35: 441–443. https://doi.org/10.1038/nbt.3833

Tang, W., and D. R. Liu, 2018 Rewritable multi-event analog recording in bacterial and mammalian cells. Science 360: eaap8992. https://doi.org/10.1126/science.aap8992

Villiger, L., H. M. Grisch-Chan, H. Lindsay, F. Ringnalda, C. B. Pogliano *et al.*, 2018 Treatment of a metabolic liver disease by in vivo genome base editing in adult mice. Nat. Med. 24: 1519–1525. https://doi.org/10.1038/s41591-018-0209-1

Zafra, M. P., E. M. Schatoff, A. Katti, M. Foronda, M. Breinig *et al.*, 2018 Optimized base editors enable efficient editing in cells, organoids and mice. Nat. Biotechnol. 36: 888–893. https://doi.org/10.1038/nbt.4194

Zerbino, D. R., P. Achuthan, W. Akanni, M. R. Amode, D. Barrell *et al.*, 2018 Ensembl 2018. Nucleic Acids Res. 46: D754–D761. https://doi.org/10.1093/nar/gkx1098

Zuo, E., Y. Sun, W. Wei, T. Yuan, W. Ying *et al.*, 2019 Cytosine base editor generates substantial off-target single-nucleotide variants in mouse embryos. Science. https://doi.org/10.1126/science.aav9973

Communicating editor: D. Greenstein