CROTONdb: genome-wide variant effect prediction database for CRISPR/Cas9 editing outcomes

Victoria Li[1], Alicja Tadych[2], Aaron Wong[3], Zijun Zhang[4,5,*]

[1] Harvard University, Cambridge, MA.
[2] Lewis-Sigler Institute of Integrative Genomics, Princeton University, Princeton, NJ.
[3] Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY.
[4] Division of Artificial Intelligence in Medicine, Department of Medicine, Cedars-Sinai Medical Center, Los Angeles, CA
[5] Department of Computational Biomedicine, Cedars-Sinai Medical Center, Los Angeles, CA
[*] Correspondence: zijun.zhang@cshs.org

Abstract

CRISPR/Cas9 is a genome editing tool widely used in biological research and clinical therapeutics. Naturally occurring human genomic variations, through altering the sequence context of CRISPR/Cas9 target regions, can significantly affect its DNA repair outcomes and ultimately lead to different editing efficiencies. However, these effects have not been systematically studied or documented, even as CRISPR/Cas9 is broadly applied to primary cells and patient samples that harbor such genetic diversity. Herein, we present CROTONdb (https://croton.princeton.edu/), a publicly available database that enables fast and comprehensive investigations of single nucleotide variant (SNV) effects on CRISPR/Cas9 outcomes across the human genome. CROTONdb provides predictions for all possible CRISPR/Cas9 target sites in the coding region, spanning over 5.38 million gRNA targets, 32.32 million predicted editing outcomes, and 90.82 million estimated variant effects. Critically, variant effects presented in CROTONdb accurately cover both common and rare/understudied variants. We illustrate the utility of CROTONdb through two case studies, showcasing its impact on both preclinical discoveries and recent clinical trials. Our case studies demonstrate pitfalls of failing to account for genetic variations in Cas9 target selection, and how they can be effectively examined and avoided using our tool. We anticipate CROTONdb having broad clinical utilities in gene and cellular therapies.

Introduction

The clustered regularly interspaced short palindromic repeats (CRISPR)/CRISPR-associated protein 9 (Cas9) gene-editing system is highly effective in modifying specific genomic loci. This state-of-the-art gene-editing technology has shown great promise in the treatment of human diseases, including genetic and infectious diseases as well as cancer (1). For instance, β-thalassemia and sickle cell disease (SCD) are the two most common hemoglobinopathies, genetic disorders resulting from mutations that alter the structure of the oxygen-transporting hemoglobin protein. There are ongoing clinical trials (clinicaltrials.gov NCT03745287 and NCT03655678) using CRISPR/Cas9 *ex-vivo BCL11A* erythroid enhancer-disrupted hematopoietic stem cells for the treatment of these hemoglobinopathies, which afflict individuals with diverse ethnic backgrounds (2,3). Furthermore, CRISPR/Cas9-based T cell editing of human immunodeficiency virus (HIV) co-receptor genes *CCR5* and *CXCR4* could represent an effective cure for acquired immune deficiency syndrome (AIDS), which affects tens of millions of patients worldwide (4,5). CRISPR/Cas9 also has applications in cancer therapeutics, particularly in engineering patient T cells. For example, following the completion of a phase 1 clinical trial (clinicaltrials.gov NCT02793856), deleting the immune inhibitory receptor gene *PDCD1* has been deemed safe and feasible in the treatment of non-

small cell lung cancer, the most common type of lung cancer, which is the leading cause of cancer death worldwide (6,7).

CRISPR/Cas9 editing relies on using guide RNAs (gRNAs) to direct a DNA-cleaving endonuclease to precise genomic loci by matching up with an "NGG" protospacer adjacent motif (PAM) sequence. Following a template-free CRISPR/Cas9-induced DNA double stranded break (DSB), cellular repair pathways are activated, which result in nonrandom repair outcomes. However, the exact biological factors affecting these outcomes remains elusive. Studies have shown these outcomes are determined by local DNA sequences, so researchers have turned to machine learning to predict CRISPR/Cas9 editing outcomes (8).

Approaches to this computational biology modeling problem have included neural networks and k-nearest neighbors, multinomial logistic regression, and gradient-boosting decision trees, in the models inDelphi, FORECasT, and SPROUT, respectively (9–11). These approaches rely on manual feature and model engineering, which restrict their effectiveness to limited existing knowledge of CRISPR/Cas9 DSB repair. Automating feature engineering to overcome this limitation motivated the creation of CROTON, a convolutional neural network (CNN)-based CRISPR/Cas9 editing outcome predictor. CROTON further automated model engineering by leveraging AMBER, a framework built to apply automated machine learning (AutoML)-technique neural architecture search (NAS) to biological CNN design (12). Relative to state-of-the-art inDelphi, FORECasT, and SPROUT, CROTON was more accurate in predicting repair outcomes in a clinically-relevant primary T cell dataset (13).

A major hurdle in therapeutic gene editing is designing appropriate gRNAs for clinical use with efficient on-target editing and effective repair outcomes. In particular, accounting for gRNA off-target effects has been a major research goal, where conventional and interpretable machine learning methods are developed to predict off-target editing sites (14,15). Meanwhile, evaluating editing outcomes following a CRISPR/Cas9-induced DSB has also become increasingly important (11). Furthermore, the prevalence of CRISPR/Cas9 therapeutics-targeted diseases among worldwide populations raises the question of how human genetic variation affects gene-editing outcomes. Notably, single nucleotide variants (SNVs) are the most common type of genetic variation and can be associated with individuals of certain racial or ethnic backgrounds (16).

Currently, most gRNA selection processes are conducted in reference cell line genomes that do not account for genetic variation, which recent studies have shown may affect therapeutic outcomes. It has been revealed that naturally occurring gene variants can lead to potentially cancerous off-target therapeutic CRISPR/Cas9 editing. These findings have led to the development of online tools like CRISPRme to facilitate variant-aware off-target gRNA assessments (17,18). To further minimize the risks of treatment failure or adverse outcomes in CRISPR/Cas9-based therapeutics related to human genetic diversity, we need to broadly assess how genetic diversity affects on-target CRISPR/Cas9 editing outcomes after a DSB.

CROTON is not only accurate, but also variant-aware. Indeed, it was used in the first analysis of the effect of naturally-occurring variants on CRISPR/Cas9 editing outcomes (13). Herein, we present CROTONdb, a database built off CROTON that accounts for SNVs across the human genome in the first systematic evaluation of the effect of human genetic variation on CRISPR/Cas9 gene-editing outcomes. We show the applicability of CROTONdb for therapeutic gRNA design in two case studies — deleting *CD33* and *TTR*

for acute myeloid leukemia (AML) and transthyretin (ATTR) amyloidosis treatment, respectively. Researchers and clinicians can leverage CROTONdb for the design of effective and broadly applicable CRISPR/Cas9 gRNA targets.

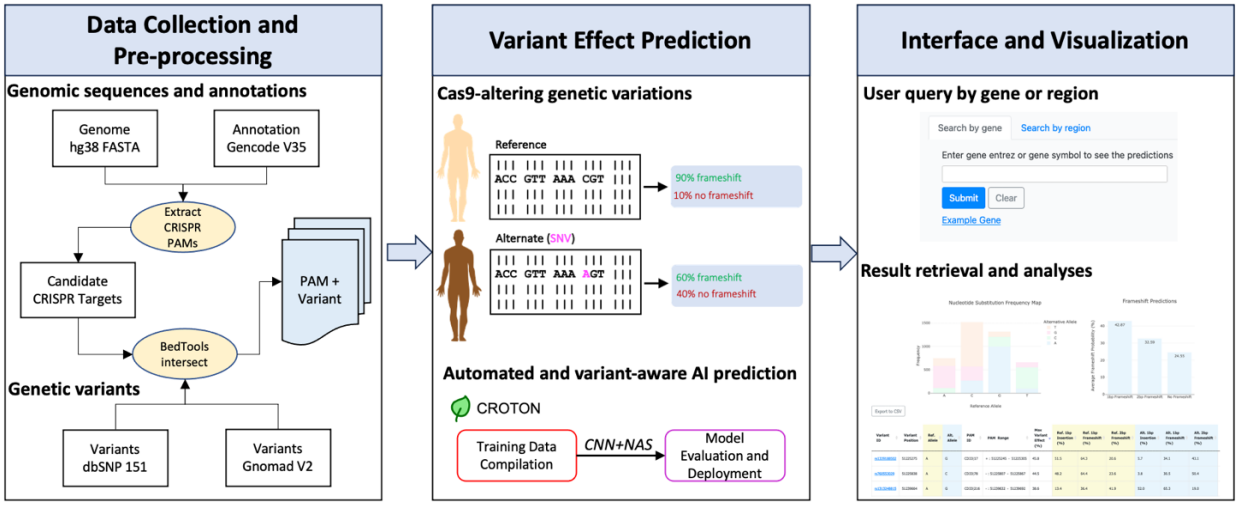


Figure 1: The CROTONdb database workflow. CROTONdb is constructed by comprehensively processing all candidate Cas9 target sites and variant information in the coding region of the human genome (**left**). We leverage CROTON, an automated and variant-aware deep learning model to predict the variant effects for altering Cas9 editing outcomes, which could ultimately lead to differential editing efficiencies (**middle**). CROTONdb provides fast and comprehensive analysis of Cas9-altering genetic variants, enabling users to effectively refine their Cas9 target selection towards efficient and safe genome editing (**right**).

Materials and Methods

Genomic data collection and preprocessing

The human reference genome (hg38 release 35) was downloaded from Gencode (https://www.gencodegenes.org/human/) (19). To extract genes and their corresponding coding regions (CDS), we used Gencode comprehensive gene annotation v35 GFF file downloaded from the same website. The python package "gffutils" was employed to analyze the gene annotations and extract corresponding DNA sequences (https://github.com/daler/gffutils). For each CDS in each gene, we extracted all "NGG" sequences from the forward and reverse strands, and extended 30bp in both directions as the candidate Cas9 target region. Within each gene, the candidate Cas9 targets were then sorted based on their smallest genomic coordinates and the corresponding rank was assigned as a numerical PAM ID. For each chromosome, all PAM IDs in this chromosome were compiled into a BED file.

We downloaded the human genomic variations from dbSNP (build 151) from https://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606_b151_GRCh38p7/ and Gnomad Exome (V2-liftover) from https://gnomad.broadinstitute.org/ in the VCF file format (20,21). We then performed an intersection using BedTools (22) between the genome variation VCF file with the PAM BED file generated from the previous step. The resulting BED file contained the information of PAMs and all overlapping genomic variations, which we used to analyze Cas9-altering variant effect predictions.

Variant effect prediction

For each pair of a Cas9 target site (with unique PAM IDs) and a genomic variation, we applied CROTON (13), a state-of-the-art deep CNN to predict the Cas9 repair outcomes for the reference and alternative alleles, respectively. Briefly, CROTON is a fully automatically designed model (12) and is trained using a large-scale synthetic library transfected to the K562 cell line (9). When tested in an independent primary T-cell dataset (10), CROTON achieved superior performance, providing the basis for an unbiased variant effect prediction for Cas9-mediated gene therapy. For a given Cas9 target sequence, we used CROTON to predict 1bp insertion probability, and 1bp and 2bp frameshift frequencies. The VEP was defined as the absolute difference between the predicted probabilities of the reference and alternate sequences.

We compiled an indexed, compressed VCF file for each chromosome to store the VEP results. Each row in the VCF file was a variant-PAM combination where we stored the variant location, reference and alternate alleles and sequences, the predicted repair outcomes for the reference alternate sequences. The VCF files were tabix-indexed to enable efficient random access (23).

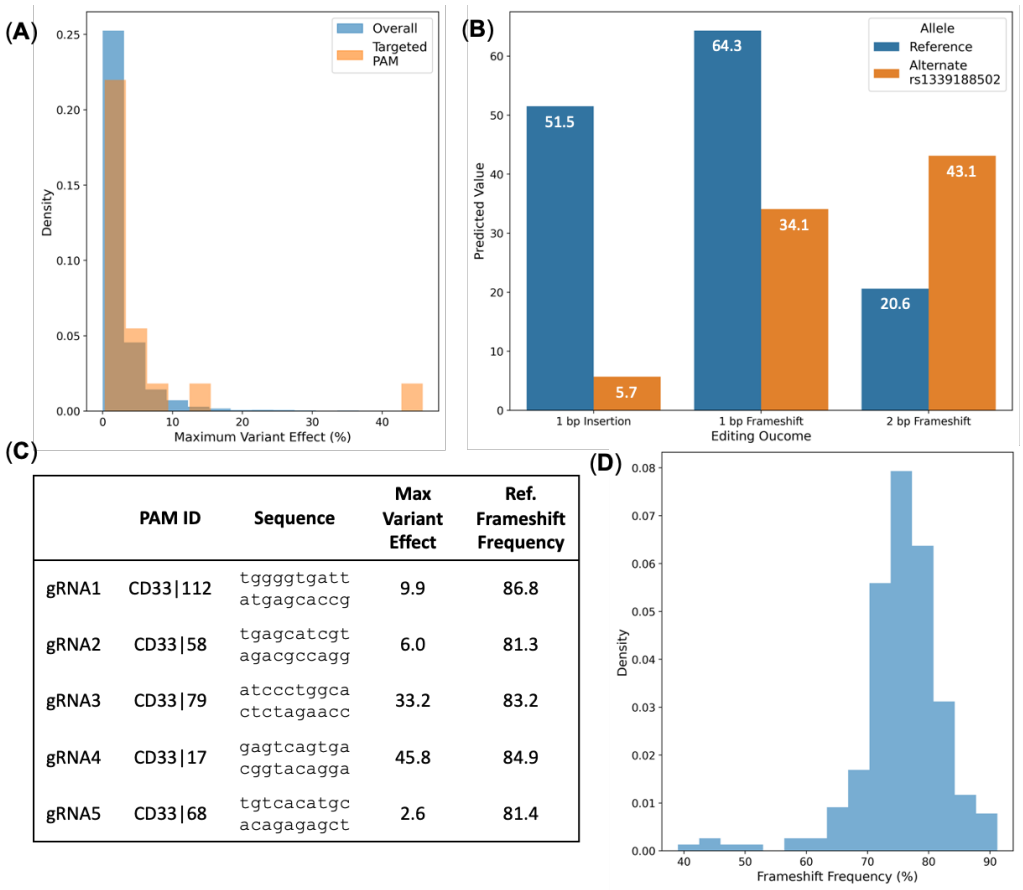
Statistical analysis

For each user-queried genomic region or gene, we performed several statistical analyses to characterize the queried region. We analyzed the nucleotide substitution frequencies for all potential Cas9-altering genomic variations and visualized them in a stacked bar plot. We also visualized the frameshift frequency based on the reference sequences and alternate sequences across all PAM sites in the queried region.

To provide the contexts of a user-queried gene or genomic region for Cas9 efficiency and VEP, we pre-computed the genome-wide statistics for frameshift frequencies, variant load, and variant effects. In particular, we computed the cumulative density function (CDF) of Cas9-induced frameshift frequencies across all possible Cas9 targets in the genome, and compared the average predicted frameshift frequency in the user-queried regions to the genome-wide CDF. Similarly, we computed the genome-wide CDF of variant effects across all target sequences. We computed the average per-base-pair variant count and multiplied by the length of the user-queried genomic region to obtain the expected variant count. We fitted a Poisson distribution with its lambda parameter equal to the expected variant count, and used it to derive the percentile for the observed variant count. We then reported all percentiles in the database interface.

Database implementation

The CROTONdb frontend was implemented in Django, a python-based web framework (https://www.djangoproject.com/). The frontend supports user search by gene name, or by custom genomic region. When searching by gene name, an elastic search engine was implemented to enable fuzzy search and auto completion of gene names. The frontend implementation calls tabix (23) to access the precompiled VCF files in a computational efficient way. Our database is freely available at https://croton.princeton.edu/ .



	PAM ID	Sequence	Max Variant Effect	Ref. Frameshift Frequency
gRNA1	CD33|112	tggggtgatt atgagcaccg	9.9	86.8
gRNA2	CD33|58	tgagcatcgt agacgccagg	6.0	81.3
gRNA3	CD33|79	atccctggca ctctagaacc	33.2	83.2
gRNA4	CD33|17	gagtcagtga cggtacagga	45.8	84.9
gRNA5	CD33|68	tgtcacatgc acagagagct	2.6	81.4

Figure 2: CROTONdb reveals variant effects on a proposed therapeutic *CD33* CRISPR/Cas9 target site. (A) The distribution of the maximum variant effects of SNVs across *CD33* gene and within 60 bp of the proposed gRNA target. **(B)** A comparison of CROTON's predicted 1 bp insertion, 1 bp frameshift, and 2 bp frameshift frequencies on the reference and alternate genome containing the SNV associated with the highest *CD33* variant effect, rs1339188502. **(C)** The CROTONdb PAM IDs, sequences, and CROTON-predicted maximum variant effect and reference frameshift frequencies for identified efficient CD33 gRNA targets. **(D)** The distribution of predicted frameshift frequencies across the *CD33* gene body.

Results

The CROTONdb describes the effect of genetic diversity on CRISPR/Cas9 editing outcomes across the human genome. This comprehensive database includes 5.38 million gRNA targets and 90.82 million estimated variant effects. This platform can be used as a reference for researchers to avoid SNVs that have a large effect on CRISPR/Cas9 gene editing outcomes (**Figure 1**).

Here, we illustrate the application of CROTONdb with two case studies, covering preclinical discoveries of *CD33* knockout in Chimeric Antigen Receptor (CAR)-T cell therapy, and clinical applications of *TTR* inactivation for treating ATTR amyloidosis. These two case studies demonstrate how users can utilize CROTONdb to account for editing outcome variant effects in their gRNA design.

Case study 1: Evaluating gRNA targets in preclinical CD33 discovery.
CD33 is a myeloid differentiation antigen that is also expressed on AML blasts and a potential target for cancer immunotherapy. A promising approach to AML treatment is coupling standard bone marrow transplants with CD33-targeted CAR-T Cell therapy. However, since both healthy and malignant myeloid cells are CD33-positive, non-discriminatory therapeutic killing of CD33-expressing cells can compromise immune function and lead to patient death. One approach to addressing this challenge is transplanting CD33-deficient rather than wild type hematopoietic stem and progenitor cells (HSPCs) to patients. These CD33-deficient HSPCs would be spared from CAR T Cell-mediated killing and facilitate functional hematopoiesis (24,25).

A recent study has explored this approach, demonstrating that coupling CRISPR/Cas9 knockout (KO) CD33 donor HSPCs with CD33 CAR T Cell therapy is clinically feasible for AML treatment. The gRNAs used in this study were designed with a reference genome and optimized for KO efficiency (24,26).

On CROTONdb, users can "search by region," inputting their gRNA target range and assessing its corresponding variant effects. For this study's target region-of-interest, searching for the range "chr19: 51225245-51225305" then filtering for PAM ID "CD33|17" reveals that there exist 18 SNVs within 60 bp of this CRISPR/Cas9 target site. CROTONdb produces a table of corresponding search results with "Export as CSV" functionality that facilitates further investigation of target-associated SNVs, allowing users to download and easily return to database search results. Users can also "search by gene" on CROTONdb, enabling use cases like assessing all potential target locations and their corresponding variant effects across a target gene. Using CROTONdb to analyze the selected gRNA region relative to the *CD33* gene body found that CROTON's predicted maximum SNV variant effect distribution is more left skewed on the targeted PAM site. While the overall maximum variant effect of *CD33* is concentrated under 10%, CROTONdb reveals that there are variants with a large effect on CRISPR/Cas9 editing outcomes associated with the study's target site (**Figure 2A**).

After submitting a search query, users can also sort by "Max Variant Effect" to identify which target gene regions are associated with the greatest CROTON-predicted variant effects. Conducting this analysis on *CD33* shows that, in fact, the gRNA chosen by this HSPC KO for AML treatment study is associated with CROTON's highest predicted "Max Variant Effect" on *CD33*. More specifically, CROTON predicts that the SNV "rs1339188502," a missense A to G mutation, decreases the predicted 1 bp insertion probability between the reference and alternate sequences by 45.8%. CROTON's associated predictions for 1 bp and 2 bp frameshift frequency were also altered substantially by this SNV, decreasing by 30.2% and increasing by 22.5%, respectively (**Figure 2B**).

Especially due to the clinical applications of this work, this observation suggests that a different CD33 gRNA target may be more appropriate for *CD33* inactivation. To select their CD33 target site, the CD33-targeting AML study chose between five gRNAs that had been previously identified as having high CD33 KO efficiency experimentally (24). Evaluating these five gRNAs revealed that they have similar reference frameshift frequencies (i.e., the sum of CROTON's 1 bp and 2bp reference frameshift frequency predictions) but varying maximum variant effects. Out of these potential targets, the study's chosen gRNA4 not only had a relatively large reference frameshift frequency of 84.9%, but also the largest maximum variant effect of 45.8% by a wide margin, 12.6% higher than the next largest maximum variant effect. These

CROTONdb findings indicate a different gRNA like gRNA1 which had an even higher predicted reference frameshift frequency of 86.8% and a much lower maximum variant effect of 9.9% may be a better choice for therapeutic CRISPR/Cas9 editing for AML treatment (**Figure 2C**). Notably, compared to the distribution of frameshift frequency across the *CD33* gene body, which is concentrated under 80%, the five identified efficient gRNAs had very high predicted reference frameshift frequencies of above 80%. This result substantiates the applicability of using CROTONdb's frameshift frequencies to identify effective gRNAs to target for gene KO (**Figure 2C** and **2D**)

Overall, this case study indicates that focusing on KO efficiency in only reference genomes is ineffective for broad clinical applications. When choosing between gRNAs like the five potential targets for CD33, although researchers found gRNA4 to be most efficacious experimentally, if they had taken variant effects into account, they would have likely chosen a different gRNA target. Tools like CROTONdb, which allow for both sequence and gene-specific variant effect assessments, should be harnessed to facilitate such variant-aware gRNA selection.

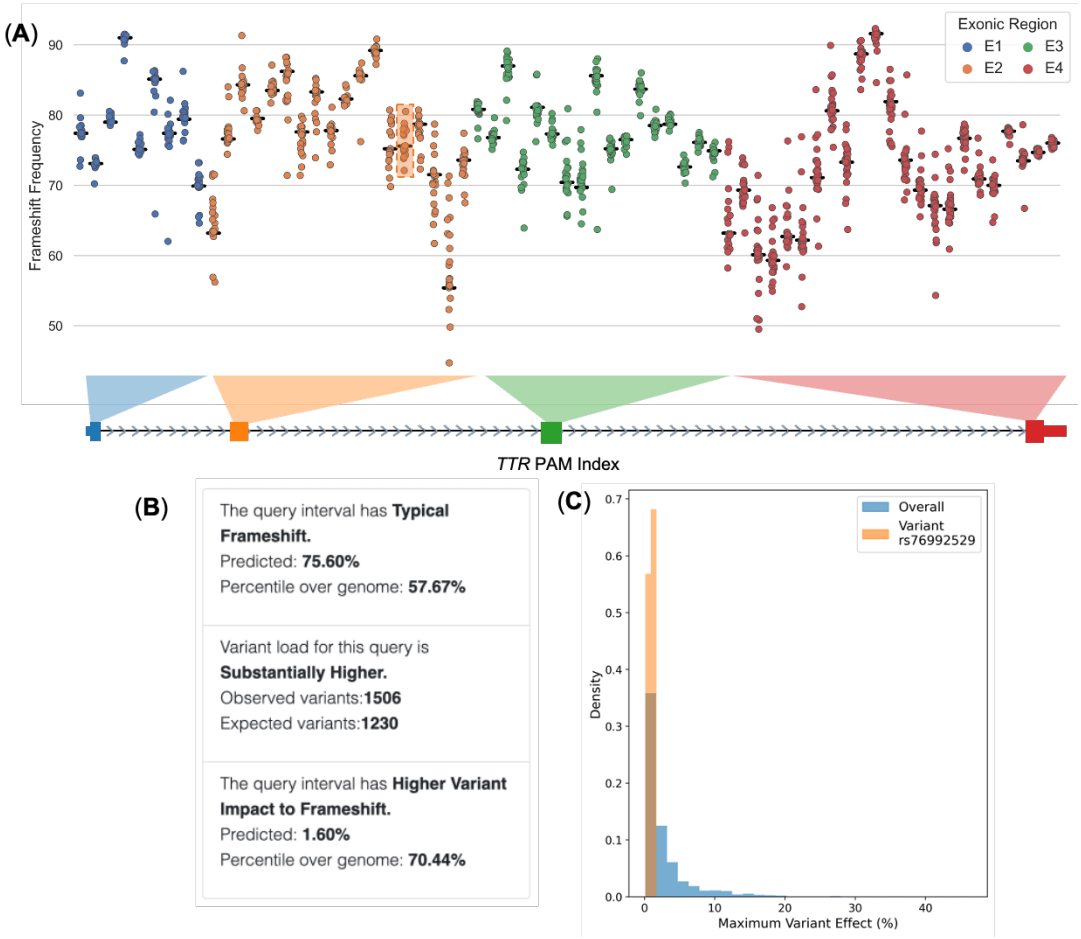


Figure 3: CROTONdb enables therapeutic gRNA target analysis across the *TTR* gene body. (A) The distribution of CROTONdb's frameshift frequency results, the sum of CROTON's 1 bp and 2 bp frameshift frequency predictions, on n=68 PAM sites across the four coding regions of *TTR*. The orange box denotes the PAM site corresponding to the TTR-targeting ATTR amyloidosis clinical trial, black horizontal lines denote the frameshift frequency prediction for the reference genome, and circles color-coded by exon denote the frameshift frequency predictions for alternate

genomes containing SNVs. **(B)** A screenshot of CROTONdb's output for genome-wide comparisons of the *TTR* target region in terms of frameshift frequency, variant load, and variant impact on frameshift frequency. **(C)** The distribution of the maximum variant effect across *TTR* and corresponding to the rs76992529 SNV, an ancestry-associated genetic SNV that predisposes its carriers to ATTR amyloidosis.

Case study 2: Identifying gRNA targets in TTR clinical trial to treat ATTR amyloidosis.

ATTR amyloidosis is a progressive fatal disease that is increasingly recognized as an important cause of cardiomyopathy and heart failure. This disease is characterized by the accumulation of misfolded transthyretin (TTR) protein in tissues, especially the nerves and heart. Notably, ATTR amyloidosis can be acquired or hereditary — indeed, 1 of 25 African Americans carry a hereditary *TTR* point mutation that predisposes them to developing this disorder (rs76992529) (27). Current ATTR amyloidosis therapies include reducing or inhibiting TTR protein synthesis by stabilizing tetrameric *TTR* or degrading *TTR* messenger RNA, respectively. These treatments prolong survival but often fail to prevent disease progression and require long-term therapeutic administration.

Recently, researchers conducted a clinical trial targeting ATTR amyloidosis using CRISPR/Cas9-based therapy to inactivate *TTR* (28). They found that following a single infusion of an *in vivo* CRISPR/Cas9 gene-editing therapy targeting patient hepatocyte *TTR*, TTR levels were greatly reduced with mild adverse events (clinicaltrails.gov NCT04601051). In this study, researchers evaluated knockout efficiency, specificity, and on-target editing to select a gRNA targeting *TTR* on chromosome 18 from base pairs (bps) 31592987 to 31593007. CROTON's predictions on this PAM ID "TTR|24" report a substantial 46% and 30% for 1 bp and 2 bp frameshift frequency, respectively, align with these researcher's selection for editing efficiency. Indeed, across the *TTR* gene body, the researchers' targeted site does have relatively high frameshift frequency (**Figure 3A**, highlighted in orange). Notably, there do exist n=19 potential gRNA target sites on *TTR* with a lower maximum variant effect and higher predicted frameshift frequency, including 7 out of a total of 28 on the first two exons of this gene (**Figure 3A**; **Supplementary Table 1**).

On CROTONdb, users can use the "Search by gene" option and input *TTR* to assess all targetable variant-associated PAM sites on the gene body. Searching by gene also dynamically creates graphs that display statistics like predicted frameshift frequencies and variant substitutions that allow researchers to assess the effectiveness of their gRNA-of-choice against all potential variant-associated target sites in their gene of interest. CROTONdb also outputs statistics comparing the frameshift frequency, variant lower, and variant impact to frameshift for the queried regions against the genome-wide statistics (**Figure 3B**).

The genetic links of ATTR amyloidosis further highlight the pertinence of conducting variant-aware gRNA selection in this case. The hereditary *TTR* point mutation rs76992529 that predisposes individuals of African descent to this disease is not linked with high variant effects on potential PAM-targeting sites (**Figure 3C**). We found that that rs76992529 has an average of only 0.92% max variant effects on all its associated Cas9-targeting sites. This impact is much smaller compared to the genetic variations found across the entire *TTR* gene, suggesting that patients will have a comparable editing efficiency regardless of which allele they carry for rs76992529. Nevertheless, variant-aware gRNA selection is still important in this case, as it can help to identify and avoid gRNAs that are likely to be ineffective due to the presence of variants in the target sequence. Attention must be drawn to the variants that are population-stratified or disease predisposing. CROTONdb streamlines this process via efficient retrieval and analysis of gRNA frameshift

efficiency and all relevant variant information. The users can effectively refine their gRNA selection based on CROTONdb results, in order to enable efficient and safe gene therapies with broad applications.

Discussion

In this study, we developed CROTONdb to search and analyze the potential impacts of naturally occurring genetic variations on CRISPR/Cas9 editing outcomes. Through the *CD33* and *TTR* case studies, we demonstrated CROTONdb's applicability in evaluating PAM sites and discovering additional CRISPR/Cas9 target locations. Notably, CROTONdb's web features like dynamic graphs and downloadable tables facilitate assessments of the frameshift frequencies and variant-effects of inputted regions-of-interest, facilitating gRNA selection.

CROTONdb is broadly useful to researchers or clinicians applying CRISPR/Cas9 editing to human samples. For basic biological research, CRISPR/Cas9 is often utilized in primary cell lines derived from human donors; in the clinic, therapies that involve CRISPR/Cas9-mediated editing of patient-derived cells have been deemed safe and feasible. These biospecimens all harbor genetic diversity that can lead to differential editing outcomes and eventually genome manipulation or therapeutic efficiencies. Whether in the laboratory or the clinic, CROTONdb has broad utility in ensuring CRISPR/Cas9-based studies and treatments are consistent, accurate, and effective.

Bibliography

1. Kang, K., Song, Y., Kim, I., et al. (2022) Therapeutic Applications of the CRISPR-Cas System. *Bioengineering (Basel)*, **9**.
2. Frati, G. and Miccio, A. (2021) Genome Editing for β-Hemoglobinopathies: Advances and Challenges. *J. Clin. Med. Res.*, **10**.
3. Pahan, K. (2019) A Broad Application of CRISPR Cas9 in Infectious, Inflammatory and Neurodegenerative Diseases. *J. Neuroimmune Pharmacol.*, **14**, 534–536.
4. Liu, Z., Chen, S., Jin, X., et al. (2017) Genome editing of the HIV co-receptors CCR5 and CXCR4 by CRISPR-Cas9 protects CD4+ T cells from HIV-1 infection. *Cell Biosci.*, **7**, 47.
5. Pandey, A. and Galvani, A.P. (2019) The global burden of HIV and prospects for control. The global burden of HIV and prospects for control. *Lancet HIV* **(2019)** , *6*, e809–e811.
6. Lu, Y., Xue, J., Deng, T., et al. (2020) Safety and feasibility of CRISPR-edited T cells in patients with refractory non-small-cell lung cancer. *Nat. Med.*, **26**, 732–740.
7. Sung, H., Ferlay, J., Siegel, R.L., et al. (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J. Clin.*, **71**, 209–249.
8. Molla, K.A. and Yang, Y. (2020) Predicting CRISPR/Cas9-Induced Mutations for Precise Genome Editing. *Trends Biotechnol.*, **38**, 136–141.
9. Allen, F., Crepaldi, L., Alsinet, C., et al. (2018) Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.*, **37**, 64–72.
10. Leenay, R.T., Aghazadeh, A., Hiatt, J., et al. (2019) Large dataset enables prediction of repair after CRISPR-Cas9 editing in primary T cells. *Nat. Biotechnol.*, **37**, 1034–1037.
11. Shen, M.W., Arbab, M., Hsu, J.Y., et al. (2018) Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature*, **563**, 646–651.
12. Zhang, Z., Park, C.Y., Theesfeld, C.L., et al. (2021) An automated framework for efficiently designing deep convolutional neural networks in genomics. *Nature Machine Intelligence*, **3**, 392–400.

13. Li, V.R., Zhang, Z. and Troyanskaya, O.G. (2021) CROTON: an automated and variant-aware deep learning framework for predicting CRISPR/Cas9 editing outcomes. *Bioinformatics*, **37**, i342–i348.
14. Zhang, Z., Lamson, A.R., Shelley, M., et al. (2023) Interpretable neural architecture search and transfer learning for understanding sequence dependent enzymatic reactions. Interpretable neural architecture search and transfer learning for understanding sequence dependent enzymatic reactions. *arXiv [q-bio.MN]* **(2023)** .
15. Lin, J., Zhang, Z., Zhang, S., et al. (2020) CRISPR-net: A recurrent convolutional network quantifies CRISPR off-target activities with mismatches and indels. *Adv. Sci.*, **7**, 1903562.
16. Human Genome Structural Variation Working Group, Eichler, E.E., Nickerson, D.A., et al. (2007) Completing the map of human genetic variation. *Nature*, **447**, 161–165.
17. Lessard, S., Francioli, L., Alfoldi, J., et al. (2017) Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. *Proc. Natl. Acad. Sci. U. S. A.*, **114**, E11257–E11266.
18. Cancellieri, S., Zeng, J., Lin, L.Y., et al. (2022) Human genetic diversity alters off-target outcomes of therapeutic gene editing. *Nat. Genet.*, **55**, 34–43.
19. Harrow, J., Denoeud, F., Frankish, A., et al. (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7 Suppl 1**, S4.1-9.
20. Karczewski, K.J., Francioli, L.C., Tiao, G., et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
21. Sherry, S.T., Ward, M.H., Kholodov, M., et al. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
22. Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
23. Li, H. (2011) Tabix: fast retrieval of sequence features from generic TAB-delimited files. *Bioinformatics*, **27**, 718–719.
24. Kim, M.Y., Yu, K.-R., Kenderian, S.S., et al. (2018) Genetic Inactivation of CD33 in Hematopoietic Stem Cells to Enable CAR T Cell Immunotherapy for Acute Myeloid Leukemia. *Cell*, **173**, 1439-1453.e19.
25. Laszlo, G.S., Estey, E.H. and Walter, R.B. (2014) The past and future of CD33 as therapeutic target in acute myeloid leukemia. *Blood Rev.*, **28**, 143–153.
26. Doench, J.G., Hartenian, E., Graham, D.B., et al. (2014) Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.*, **32**, 1262–1267.
27. Gabrovsek, A., Donnelly, J.P., Estep, J., et al. (2019) Transthyretin Cardiac Amyloidosis in the Black Population: A Ten-Year Experience. *J. Card. Fail.*, **25**, S6.
28. Gillmore, J.D., Gane, E., Taubel, J., et al. (2021) CRISPR-Cas9 In Vivo Gene Editing for Transthyretin Amyloidosis. *N. Engl. J. Med.*, **385**, 493–502.