bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Human genetic diversity modifies therapeutic gene editing off-target potential

Samuele Cancellieri[1,*], Jing Zeng[2,*], Linda Yingqi Lin[2,*], Francesco Masillo[1], Amy Nguyen[2], Nicola Bombieri[1], Felicia Ciuculescu[3], Varun Katta[4], Shengdar Q. Tsai[4], Myriam Armant[3], Rosalba Giugno[1,#], Daniel E. Bauer[2,#], Luca Pinello[5,#]

1 University of Verona, Computer Science, Verona, Italy
2 Division of Hematology/Oncology, Boston Children's Hospital, Department of Pediatric Oncology, Dana-Farber Cancer Institute, Harvard Stem Cell Institute, Broad Institute, Department of Pediatrics, Harvard Medical School, Boston, Massachusetts 02115, USA
3 TransLab, Boston Children's Hospital, Boston, Massachusetts 02115, USA
4 Department of Hematology, St. Jude Children's Research Hospital, Memphis, Tennessee 38105, USA
5 Molecular Pathology Unit, Center for Cancer Research, Massachusetts General Hospital, Department of Pathology, Harvard Medical School, Boston, Massachusetts 02129, USA

*** Equal contribution**
# Correspondence to: rosalba.giugno@univr.it, bauer@bloodgroup.tch.harvard.edu, lpinello@mgh.harvard.edu

ABSTRACT

CRISPR gene editing holds great promise to modify somatic genomes to ameliorate disease. In silico prediction of homologous sites coupled with biochemical evaluation of possible genomic off-targets may predict genotoxicity risk of individual gene editing reagents. However, standard computational and biochemical methods focus on reference genomes and do not consider the impact of genetic diversity on off-target potential. Here we developed a web application called CRISPRme that explicitly and efficiently integrates human genetic variant datasets with orthogonal genomic annotations to predict and prioritize off-target sites at scale. The method considers both single-nucleotide variants (SNVs) and indels, accounts for bona fide haplotypes, accepts spacer:protospacer mismatches and bulges, and is suitable for personal genome analyses. We tested the tool with a guide RNA (gRNA) targeting the *BCL11A* erythroid enhancer that has shown therapeutic promise in clinical trials for sickle cell disease (SCD) and β-thalassemia. We find that the top predicted off-target site is produced by a non-reference allele common in African-ancestry populations (rs114518452, minor allele frequency (MAF)=4.5%) that introduces a protospacer adjacent motif (PAM) for SpCas9. We validate that SpCas9 generates indels (~9.6% frequency) and chr2 pericentric inversions in a strictly allele-specific manner in edited CD34+ hematopoietic stem/progenitor cells (HSPCs), although a high-fidelity Cas9 variant mitigates this off-target. This report illustrates how genetic variation may modify the genomic outcomes of therapeutic gene editing and provides a simple tool for comprehensive off-target assessment.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

INTRODUCTION

CRISPR genome editing offers unprecedented opportunities to develop novel therapeutics by introducing targeted genetic or epigenetic modifications to genomic regions of interest. Briefly, CRISPR offers a simple and programmable platform that couples binding to a genomic target sequence of choice with diverse effector proteins through RNA:DNA (spacer:protospacer) complementary sequence interactions mediated by a guide RNA (gRNA) restricted by protospacer adjacent motif (PAM) sequences. Editing effectors may consist of nucleases to introduce targeted double strand breaks leading to short indels and templated repairs (e.g. Cas9), deaminases for precise substitutions (base editors), or chromatin regulators for transcriptional interference or activation (CRISPRi/a) among others to achieve a range of desired biological outcomes[1].

CRISPR-based systems may create unintended off-target modifications posing potential genotoxicity for therapeutic use. Several experimental assays and computational methods are available to uncover or forecast these off-targets[2]. Off-target sites are partially predictable based on homology to the target site protospacer and PAM sequence. Beyond the number of mismatches or bulges, a variety of sequence features, like position of mismatch or bulge with respect to PAM or specific base changes, contribute to off-target potential[2–5]. Computational models can complement experimental approaches to off-target nomination in several respects: to triage gRNAs prior to experiments by predicting the number and cleavage potential of off-target sites; to prioritize target sites for experimental scrutiny; and to supplement experimental approaches in nominating sites for sequence validation. Genetic variants may alter protospacer and PAM sequences and therefore may influence both on-target and off-target potential. Although a variety of in vitro and cell-based experimental methods can be used to empirically nominate off-target sites, these methods either use homology to the reference genome as a criterion to define the search space and/or use a limited set of human donor genomes to evaluate off-target potential[3,6]. Therefore, computational methods may be especially useful to predict the impact of off-target sequences not found in reference genomes.

Prior studies considering gRNAs targeting therapeutically relevant genes and population-based variant databases like the 1000 Genomes Project (1000G) and the Exome Aggregation Consortium have highlighted how genetic variants can significantly alter the off-target landscape by creating novel and personal off-target sites not present in a single reference genome[7,8]. Although these prior studies provide code to reproduce analyses, implementation choices make these tools not suitable to analyze large variant datasets and to consider higher numbers of mismatches. In addition, these methods ignore bulges between RNA:DNA hybrids, cannot efficiently model alternative haplotypes and indels, and require extensive computational skills to utilize.

Several user-friendly websites have been developed to aid the design of gRNAs and to assess their potential off-targets[9–12]. Even though variant-aware prediction is an important problem for genome editing interventions, these scalable graphical user interface (GUI) based tools do not account for genetic variants. In addition, these tools artificially limit the number of mismatches for the search and/or do not support DNA or RNA bulges. Therefore, designing gRNAs for therapeutic intervention using current widely available tools could miss important off-target sites that may lead to unwanted genotoxicity. A complete and exhaustive off-target search with an arbitrary number of mismatches, bulges, and genetic variants that is haplotype-aware is a computationally challenging problem that requires specialized and efficient data structures.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

We have recently developed a command line tool that partially solves these challenges called CRISPRitz[13]. This tool uses optimized data structures to efficiently account for single variants, mismatches and bulges but with significant limitations[13]. Here we substantially extend this work by developing CRISPRme, a tool to aid with the design of gRNAs with added support for haplotype-aware off-target enumeration, short indel variants and a flexible number of mismatches and bulges. CRISPRme is a unified, user-friendly web-based application that provides several reports to prioritize putative off-targets based on their risk in a population or individuals.

CRISPRme is flexible to accept user-defined genomic annotations, which could include empirically identified off-target sites or cell type specific chromatin features. It can integrate population genetic variants from sets of phased individual variants (like those from 1000G[14]), unphased individual variants (like those from the Human Genome Diversity Project, HGDP[15]) and population-level variants (like those from the Genome Aggregation Database, gnomAD[16]). Furthermore, it can accept personal genomes from individual subjects to identify and prioritize private off-targets due to variants specific to a single individual.

Here we demonstrate the utility of CRISPRme by analyzing the off-target potential of a gRNA currently being tested in clinical trials for SCD and β-thalassemia[17–19]. We identify possible off-targets introduced by genetic variants included within and extending beyond 1000G. We predict that the most likely off-target site is introduced by a variant common in African-ancestry individuals (rs114518452, minor allele frequency (MAF)=4.5%) and provide experimental evidence of its off-target potential in gene edited human CD34+ hematopoietic stem and progenitor cells.

RESULTS

CRISPRme is a web-based tool to predict off-target potential of CRISPR gene editing that accounts for genetic variation. It is available online at http://crisprme.di.univr.it/. CRISPRme can also be deployed locally as a web app or used as a command line utility, both of which respect genomic privacy offline. CRISPRme takes as input a genome build, a set of variants (VCF files for populations or individuals), gRNA protospacer sequence(s), a PAM, a user-defined threshold of mismatches and bulges, and optional user-defined genomic annotations (**Fig. 1a, Supplementary Note 1**).

We have designed CRISPRme to be flexible with support for new gene editors with variable and extremely relaxed PAM requirements[20]. Thanks to a PAM encoding based on Aho-Corasick automata and an index based on a ternary search tree, CRISPRme can perform genome-wide exhaustive searches efficiently even with an NNN PAM, extensive mismatches (tested with up to 7) and RNA:DNA bulges (tested with up to 2) (**Supplementary Note 2**).

Notably, a comprehensive search performed with up to 6 mismatches, 2 DNA/RNA bulges and a fully non-restrictive PAM (NNN) takes only 19 hours in a small computational cluster (Intel Xeon CPU E5-2609 v4 clocked at 2.2 GHz and 128 GB RAM). All the 1000G variants, including both SNVs and indels, can be included in the search together with all the available metadata for each individual (sex, super-population and age), and the search operation takes into account observed haplotypes (**Supplementary Note 3**). Importantly, off-target sites that represent alternative alignments to a given genomic region are merged to avoid inflating the number of reported sites. Although several tools exist to enumerate off-targets, to our knowledge only a command line tool called *crispRtool*[7,8] incorporates genetic variants in the search. However, its search

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

operation is limited to only 5 mismatches, cannot include DNA or RNA bulges, does not provide a graphical interface and is orders of magnitude slower than CRISPRme (**Supplementary Note 4**).

CRISPRme generates several general reports (**Supplementary Note 5**). First, it summarizes for each gRNA all the potential off-targets found in the reference or variant genomes based on their mismatches and bulges (**Fig. 1b**). Second, it compares gRNAs to customizable annotations. By default, it classifies possible off-target sites based on GENCODE annotations (genomic features) and ENCODE annotations[21] (candidate cis-regulatory elements, cCREs). It can also incorporate user-defined annotations in BED format, such as empiric off-target scores or cell-type specific chromatin features (**Supplementary Note 3**). Third, by using 1000G[14] and/or HGDP[15]] variants, CRISPRme reports the cumulative distribution of homologous sites based on the reference genome or super-population. These global reports could be used to compare a set of gRNAs. Fourth, it generates reports focused on individual off-target sites to demonstrate the frequency of allele-specific off-target sites across super-populations and how genetic variation impacts the predicted cleavage potential using cutting frequency determination (CFD) scores[5]. Finally, CRISPRme can generate personal genome focused reports called *personal risk cards* (**Supplementary Note 6**). These reports indicate off-target sites modified by private genetic variants not found in population databases.

We tested CRISPRme with a gRNA (#1617) targeting a GATA1 binding motif at the +58 erythroid enhancer of *BCL11A*[17,18]. A recent clinical report described two patients, one with SCD and one with β-thalassemia, each treated with autologous gene modified hematopoietic stem and progenitor cells (HSPCs) edited with Cas9 and this gRNA, who showed sustained increases in fetal hemoglobin, transfusion-independence and absence of vaso-occlusive episodes (in the SCD patient) following therapy[19]. This study as well as prior pre-clinical studies with the same gRNA (#1617) did not reveal evidence of off-target editing in treated cells considering off-target sites nominated by bioinformatic analysis of the human reference genome and empiric analysis of in vitro genomic cleavage potential[18,19,22]. CRISPRme analysis found that the predicted off-target site with both the greatest CFD score and the greatest increase in CFD score from the reference to alternative allele was at an intronic sequence of *CPS1* (**Fig. 1c,d**), a genomic target subject to common genetic variation (modified by a SNP with MAF ≥ 1%). CFD scores range from 0 to 1, where the on-target site has a score of 1. The alternative allele rs114518452-C generates a TGG PAM sequence (that is, the optimal PAM for SpCas9) for a potential off-target site with 3 mismatches and a CFD score (CFD_{alt} 0.95) approaching that of the on-target site. In contrast, the reference allele rs114518452-G disrupts the PAM to TGC, which markedly reduces predicted cleavage potential (CFD_{ref} 0.02). rs114518452-C has combined MAF of 1.33% total allele frequency in gnomAD v3.1[16], with MAF of 4.55% frequency in African/African-American, 0.02% in European (non-Finnish) and 0.00% in East Asian super-populations (**Fig. 1e,f, Supplementary Table 1**).

To consider the off-target potential that could be introduced by personal genetic variation that would not be predicted by 1000G variants, we analyzed HGDP variants identified from whole genome sequences of 929 individuals from 54 diverse human populations[15]. We observed 249 off-targets with CFD ≥0.2 for which the CFD score in HGDP exceeded that found in either the reference genome or due to 1000G variants by at least 0.1 (**Fig. 2a**). These additional variant off-targets not found from 1000G were observed in each super-population, with the greatest frequency in the African super-population (**Fig. 2b**). 229 (91.9%) of these variant off-targets not found by 1000G were unique to a super-population and 172 (75.1%) of these were unique to just one individual (**Fig. 2c**). Single individual focused searches, for example an analysis of HGDP01211, an individual of the Oroqen population within the East Asian super-population, showed that most variant off-

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

targets (with higher CFD score than reference) were due to variants also found in 1000G (n=32369, 90.4%), a subset were due to variants shared with other individuals from HGDP but absent from 1000G (n=3177, 8.9%), and a small fraction were private to the individual (n=234, 0.7%) (**Fig. 2d**). Among these private off-targets was one generated by a variant (rs1191022522, 3-99137613-A-G, gnomAD v3.1 MAF 0.0052%) where the alternative allele produces an NGG PAM that increases the CFD score from 0.14 to 0.54 (**Fig. 2d,e**).

To experimentally test the top predicted off-target from CRISPRme, we identified a CD34+ HSPC donor of African ancestry heterozygous for rs114518452-C (the variant predicted to introduce the greatest increase in off-target cleavage potential; **Fig. 1d-f**). We performed RNP electroporation using a gene editing protocol that preserves engrafting HSC function[18]. Amplicon sequencing analysis showed 92.0% indels at the on-target site and 4.8% indels at the off-target site. Evaluable indels were strictly found at the alternative PAM-creation allele without indels observed at the reference allele (**Fig. 3a-c**), suggesting ~9.6% off-target editing of the alternative allele. In an additional 6 HSPC donors homozygous for the reference allele rs114518452-G/G, no indels were observed at the off-target site, suggesting strict restriction of off-target editing to the alternative allele (**Fig. 3d**).

The on-target *BCL11A* intronic enhancer site is on chr2p and the off-target-rs114518452 site is on chr2q within an intron of a non-canonical transcript of *CPS1*. Inversion PCR demonstrated inversion junctions consistent with the presence of ~150 Mb pericentric inversions between *BCL11A* and the off-target site only in edited HSPCs carrying the alternative allele (**Fig. 4a,b**). Deep sequencing of inversion junctions showed that inversions were restricted to the alternative allele in the heterozygous cells (**Fig. 4c,d**). Gene editing following the same electroporation protocol using a HiFi variant 3xNLS-SpCas9 (R691A)[23] in heterozygous cells revealed 82.3% on-target indels with only 0.1% indels at the rs114518452-C off-target site, i.e. a ~48-fold reduction compared to SpCas9 (**Fig. 3c**). Inversions were not detected following HiFi-3xNLS-SpCas9 editing (**Fig. 4b**).

These results demonstrate how personal genetic variation may influence the off-target potential of therapeutic gene editing. In the case of *BCL11A* enhancer editing, up to ~10% of SCD patients with African ancestry would be expected to carry at least one rs114518452-C allele. In general, therapeutic gene editing clinical trials might consider evaluating the impact of population and private genetic variation on gene editing outcomes including individual patient assessment. CRISPRme offers a simple-to-use tool to comprehensively evaluate off-target potential across diverse populations and within individuals. CRISPRme is available at http://crisprme.di.univr.it and may also be deployed locally to preserve privacy (**Supplementary Note 7**).

Data and software availability

CRISPRme source code is available at https://github.com/pinellolab/crisprme and the webapp is online at http://crisprme.di.univr.it. Sequencing data will be deposited in the NCBI Sequence Read Archive database prior to publication.

Acknowledgement

L.P. received support from the U.S. NIH (R35 HG010717 and RM1 HG009490). D.E.B. was supported by the National Heart, Lung, and Blood Institute (OT2HL154984, P01HL053749), Burroughs Wellcome Fund and the St. Jude Children’s Research Hospital Collaborative Research Consortium. We thank Stuart H. Orkin, Guillaume Lettre, Scot A. Wolfe, J. Keith Joung, Vikram Pattanayak, Karl Petri, and Anne H. Shen for helpful input.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Competing Financial Interests Statement

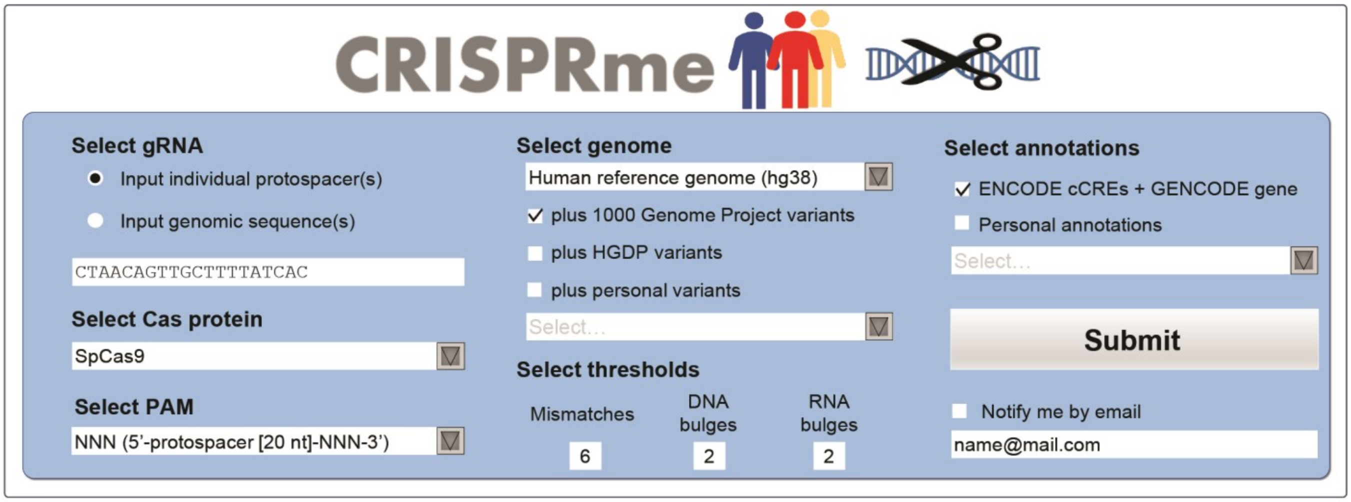
L.P. has financial interests in Edilytics, Inc., Excelsior Genomics, and SeQure Dx, Inc. L.P.'s interests were reviewed and are managed by Massachusetts General Hospital and Partners HealthCare in accordance with their conflict of interest policies.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 1. CRISPRme provides web-based analysis of off-target potential of CRISPR-Cas gene editing reflecting population genetic diversity. a) CRISPRme software takes as input Cas protein type, protospacer and PAM sequence, reference genome, variants, homology threshold and genomic annotations and provides comprehensive, target-focused and individual-focused analyses of off-target potential. It is available as an online webtool and can be deployed locally or used offline as command-line software. **b)** Analysis of *BCL11A*-1617 spacer targeting the +58 erythroid enhancer with SpCas9, NNN PAM, 1000G variants, up to 6 mismatches and up to 2 bulges. **c)** The off-target site with the highest CFD score is created by the minor allele of rs114518452. Coordinates are for hg38 and 0-start. MAF is based on 1000G. **d)** Top 1000 predicted off-target sites ranked by CFD score, indicating the CFD score of the reference and alternative allele if applicable, with allele frequency indicated by circle size. **e)** The top predicted off-target site from CRISPRme is an allele-specific target with 3 mismatches to the *BCL11A*-1617 spacer sequence, where the rs114518452-C minor allele produces a de novo NGG PAM sequence. PAM sequence shown in bold and mismatches to *BCL11A*-1617 shown as lowercase. Coordinates are for hg38 and 1-start. **f)** rs114518452 allele frequencies based on gnomAD v3.1. Coordinates are for hg38 and 1-start.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a



Select gRNA
● Input individual protospacer(s)
○ Input genomic sequence(s)
CTAACAGTTGCTTTTATCAC

Select Cas protein
SpCas9

Select PAM
NNN (5'-protospacer [20 nt]-NNN-3')

Select genome
Human reference genome (hg38)
☑ plus 1000 Genome Project variants
☐ plus HGDP variants
☐ plus personal variants
Select...

Select thresholds

Mismatches	DNA bulges	RNA bulges
6	2	2

Select annotations
☑ ENCODE cCREs + GENCODE gene
☐ Personal annotations
Select...

☐ Notify me by email
name@mail.com

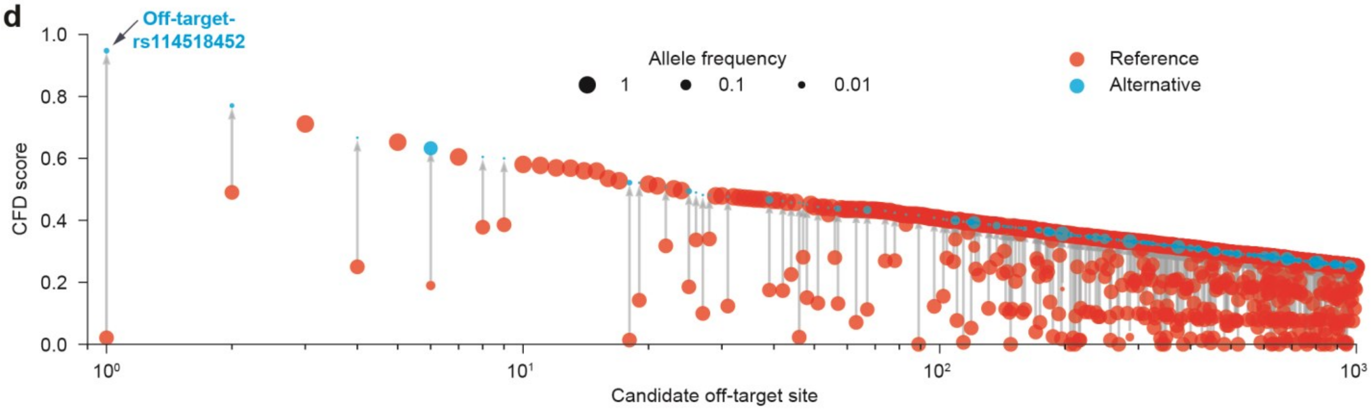
b

Protospacer + PAM	Nuclease	Genome	Bulges	Mismatches						
				0	1	2	3	4	5	6
CTAACAGTTGCTTTTATCACNNN	SpCas9	Reference	0	1	0	7	154	1645	13869	96008
			1	1	13	486	7063	73885	542165	3124545
			2	14	417	9183	121231	1109580	7245109	34849276
		Variant	0	0	0	1	12	109	823	4969
			1	0	2	28	355	2801	18968	278205
			2	2	27	305	3202	26259	188954	8634796

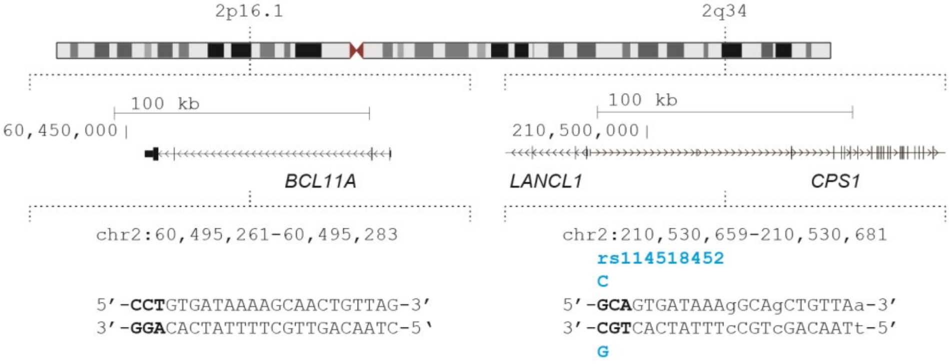
c

Sequence	Alignment	Chr	Position	Strand	Variant ID	CFD	MAF	Annotation
Protospacer+PAM	CTAACAGTTGCTTTTATCACNNN							
Reference	tTAACAGcTGCcTTTATCACTGC	2	210530658	-		0.021		intron:*CPS1*
Variant	tTAACAGcTGCcTTTATCACTGG				**rs114518452** **2-210530659-G-C**	0.947	0.02	

d



e


2p16.1
2q34
100 kb
60,450,000
BCL11A
100 kb
210,500,000
LANCL1
CPS1
chr2:60,495,261-60,495,283
5'-CCTGTGATAAAAGCAACTGTTAG-3'
3'-GGACACTATTTTCGTTGACAATC-5'
chr2:210,530,659-210,530,681
rs114518452
C
5'-GCAGTGATAAAgGCAgCTGTTAa-3'
3'-CGTCACTATTTcCGTcGACAATt-5'
G

f

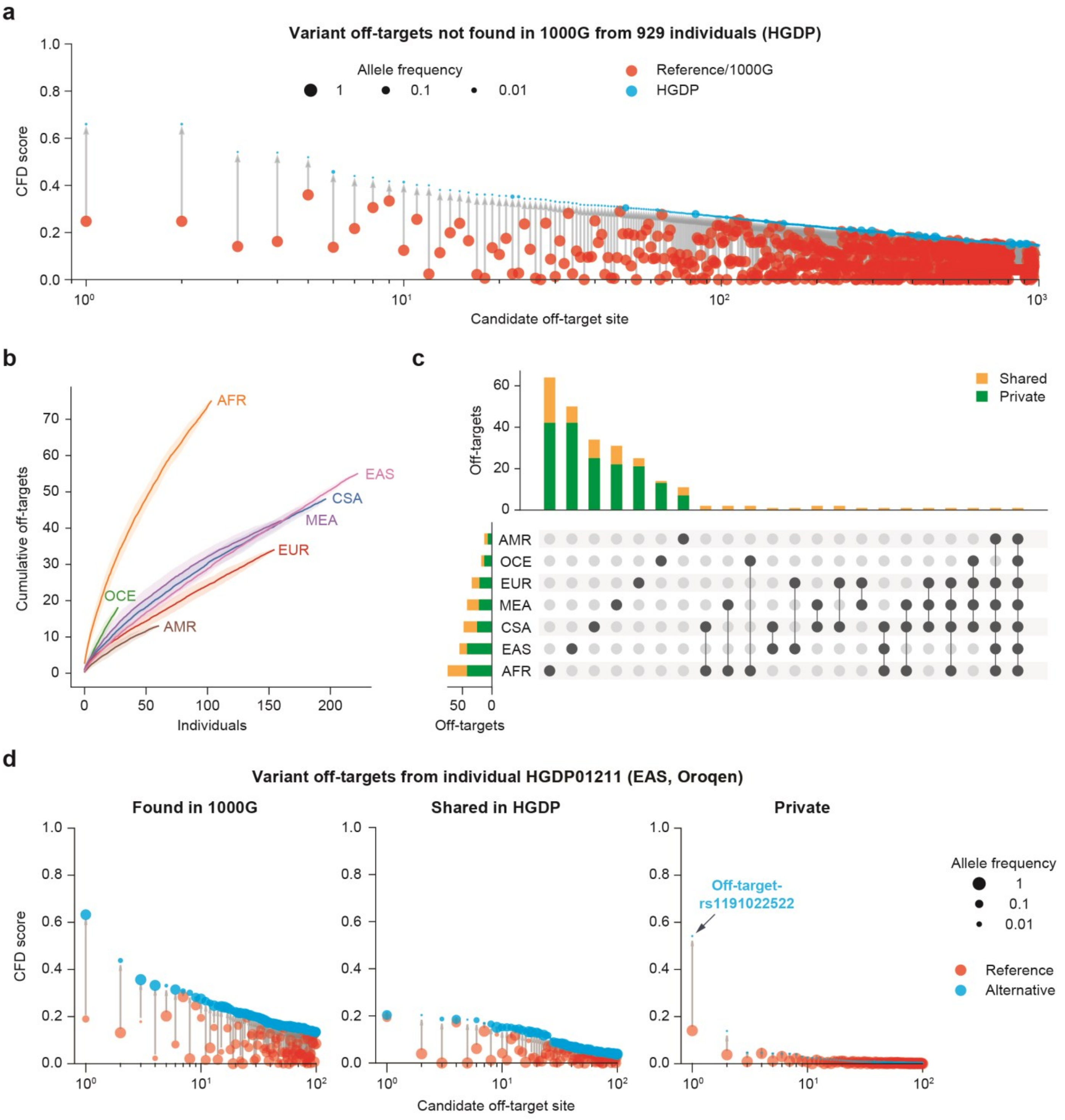
rs114518452 2-210530659-G-C

MAF (%)	Population
4.55	African
0.98	Other
0.07	Latino
0.02	South Asian
0.0002	European (non-Finnish)
0.0	Amish/Ashkenazi Jewish/ East Asian/Finnish

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 2. CRISPRme provides analysis of off-target potential of CRISPR-Cas gene editing reflecting private genetic diversity. a) CRISPRme analysis was conducted with variants from HGDP comprising WGS of 929 individuals from 54 diverse human populations. HGDP variant off-targets with greater CFD score than reference genome or 1000G were plotted and sorted by CFD score, with HGDP variant off-targets shown in blue and reference or 1000G variant off-targets shown in red. **b)** HGDP variant off-targets with CFD≥0.2 and increase in CFD of ≥0.1. Individual samples from each of the seven super-populations were shuffled 100 times to calculate the mean and 95% confidence interval. **c)** Intersection analysis of HGDP variant off-targets with CFD≥0.2 and increase in CFD of ≥0.1. Shared variants were found in 2 or more HGDP samples and private variants limited to a single sample. **d)** CRISPRme analysis of a single individual (HGDP01211) showing the top 100 variant off-targets from each of the following three categories: shared with 1000G variant off-targets (left panel), higher CFD score compared to reference genome and 1000G but shared variant with other HGDP individuals (center panel), and higher CFD score compared to reference genome and 1000G with variant not found in other HGDP individuals (right panel). For the center and right panels, reference refers to CFD score from reference genome or 1000G variants. **e)** The top predicted private off-target site from HGDP01211 is an allele-specific target with 3 mismatches to the *BCL11A*-1617 spacer sequence and 1 RNA bulge where the rs1191022522-G minor allele produces a NGG PAM sequence.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

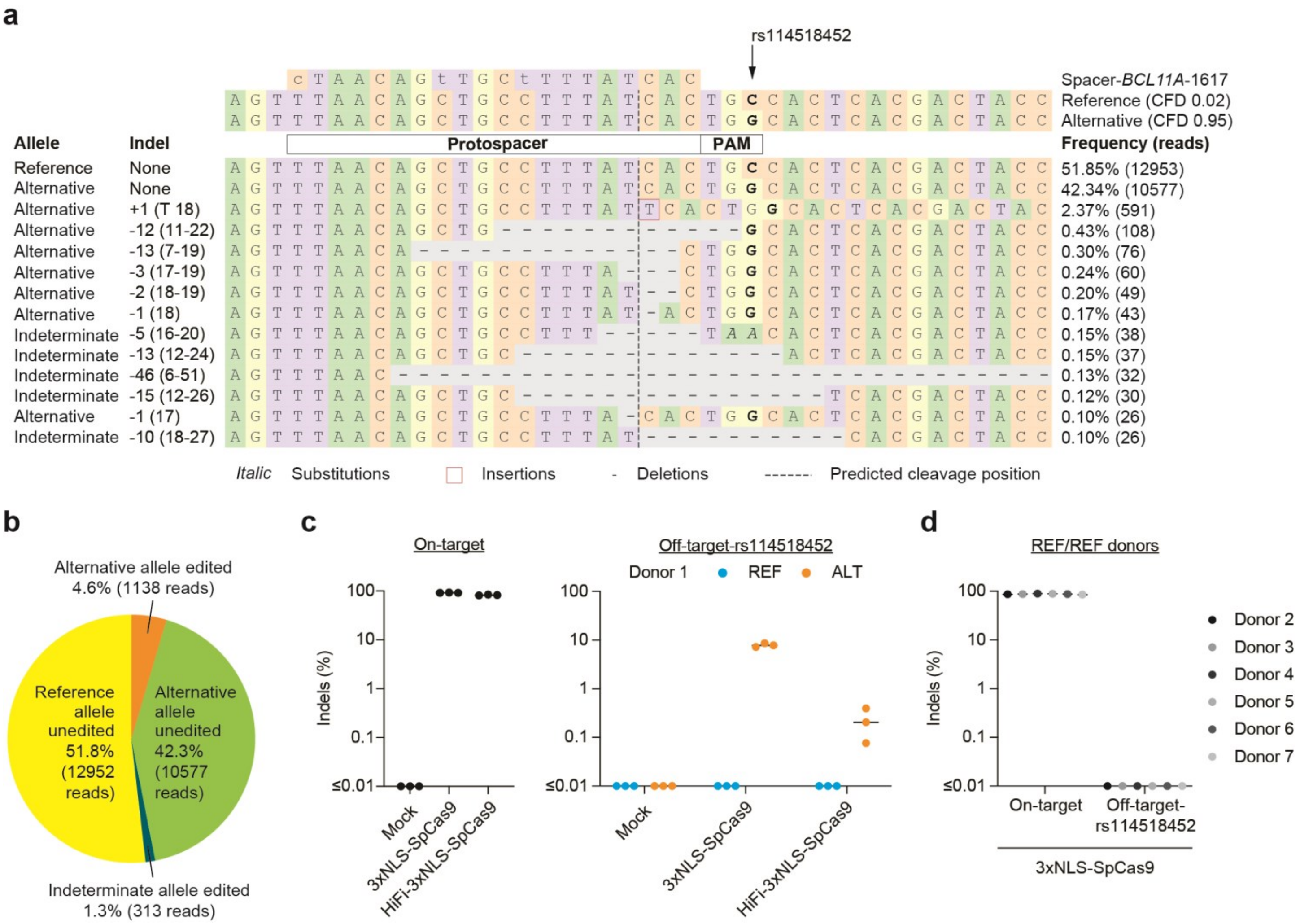


e

Sequence	Alignment	Variant ID	CFD	MAF	Annotation
Protospacer+PAM	CTAACAGTTGCTTTTATCACNNN				
Reference	aT-ACAGcTtaTTTTATCACCAG		0.140		intergenic:*DCBLD2*
Variant	aT-ACAGcTtaTTTTATCACCGG	**rs1191022522 3-99137613-A-G**	0.542	0.00054	

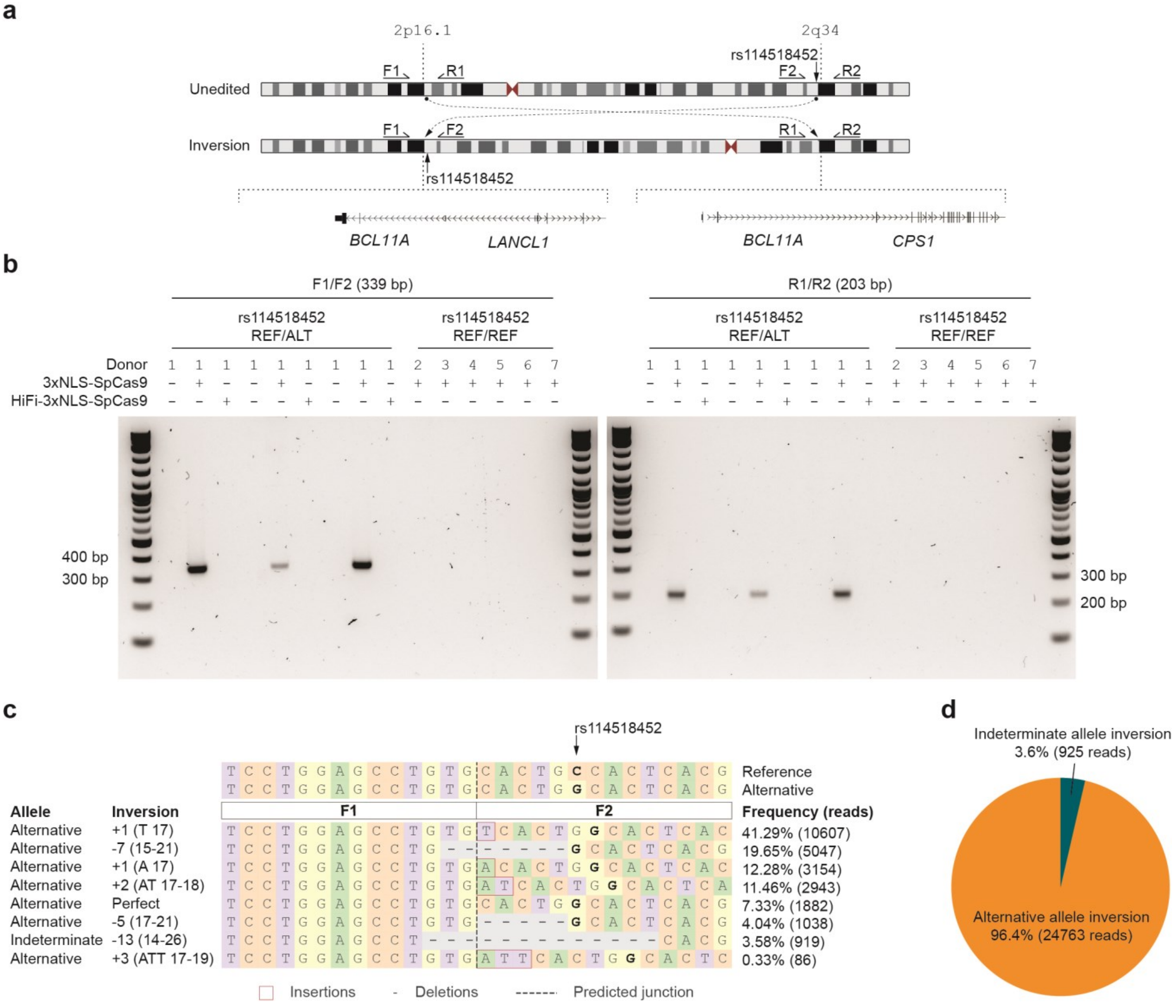
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 3. Allele-specific off-target editing by a *BCL11A* enhancer targeting gRNA associated with a common variant in African-ancestry populations. a) Human CD34+ HSPCs from a donor heterozygous for rs114518452-G/C (Donor 1, REF/ALT) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation followed by amplicon sequencing of the off-target site around chr2:210,530,659-210,530,681 (off-target-rs114518452 in 1-start hg38 coordinates). CFD scores for the reference and alternative alleles are indicated and representative aligned reads are shown. Spacer mismatches are indicated by lowercase and the rs114518452 position is shown in bold. Coordinates are for hg38 and 1-start. **b)** Reads classified based on allele (indeterminate if the rs114518452 position is deleted) and presence or absence of indels (edits). **c)** Human CD34+ HSPCs from a donor heterozygous for rs114518452-G/C (Donor 1) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation, HiFi-3xNLS-SpCas9:sg1617 RNP electroporation, or no electroporation (mock) followed by amplicon sequencing of the on-target and off-target-rs114518452 sites. Each dot represents an independent biological replicate (*n* = 3). Indel frequency was quantified for reads aligning to either the reference or alternative allele. **d)** Human CD34+ HSPCs from 6 donors homozygous for rs114518452-G/G (Donors 2-7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation followed by amplicon sequencing of the on-target and OT-rs114518452 sites.



bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Figure 4. Allele-specific pericentric inversion following *BCL11A* enhancer editing. a) Concurrent cleavage of the on-target and off-target-rs114518452 sites could lead to pericentric inversion of chr2 as depicted. PCR primers F1, R1, F2, and R2 were designed to detect potential inversions. **b)** Human CD34+ HSPCs from a donor heterozygous for rs114518452-G/C (Donor 1) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation, HiFi-3xNLS-SpCas9:sg1617 RNP electroporation, or no electroporation with 3 biological replicates. Human CD34+ HSPCs from 6 donors homozygous for rs114518452-G/G (Donors 2-7, REF/REF) were subject to 3xNLS-SpCas9:sg1617 RNP electroporation with 1 biological replicate per donor. Gel electrophoresis for inversion PCR performed with F1/F2 and R1/R2 primer pairs on left and right respectively with expected sizes of precise inversion PCR products indicated. **c)** Reads from amplicon sequencing of the F1/F2 product (expected to include the rs114518452 position) from 3xNLS-SpCas9:sg1617 RNP treatment were aligned to reference and alternative inversion templates. The rs114518452 position is shown in bold. **d)** Reads classified based on allele (indeterminate if the rs114518452 position deleted).



	Donor	1	1	1	1	1	1	1	1	1	2	3	4	5	6	7
	3xNLS-SpCas9	–	+	–	–	+	–	–	+	–	+	+	+	+	+	+
	HiFi-3xNLS-SpCas9	–	–	+	–	–	+	–	–	+	–	–	–	–	–	–

Allele	Inversion	F1	F2	Frequency (reads)
		TCCTGGAGCCTGTG	CACTG**C**CACTCACG	Reference
		TCCTGGAGCCTGTG	CACTG**G**CACTCACG	Alternative
Alternative	+1 (T 17)	TCCTGGAGCCTGTG	TCACTG**G**CACTCAC	41.29% (10607)
Alternative	-7 (15-21)	TCCTGGAGCCTG--	------**G**CACTCACG	19.65% (5047)
Alternative	+1 (A 17)	TCCTGGAGCCTGTG	ACACTG**G**CACTCAC	12.28% (3154)
Alternative	+2 (AT 17-18)	TCCTGGAGCCTGTG	ATCACTG**G**CACTCA	11.46% (2943)
Alternative	Perfect	TCCTGGAGCCTGTG	CACTG**G**CACTCACG	7.33% (1882)
Alternative	-5 (17-21)	TCCTGGAGCCTGTG	-----**G**CACTCACG	4.04% (1038)
Indeterminate	-13 (14-26)	TCCTGGAGCCT---	----------CACG	3.58% (919)
Alternative	+3 (ATT 17-19)	TCCTGGAGCCTGTG	ATTCACTG**G**CACTC	0.33% (86)

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

REFERENCES

1. Anzalone, A. V., Koblan, L. W. & Liu, D. R. Genome editing with CRISPR-Cas nucleases, base editors, transposases and prime editors. *Nat. Biotechnol.* **38**, 824–844 (2020).

2. Clement, K., Hsu, J. Y., Canver, M. C., Joung, J. K. & Pinello, L. Technologies and Computational Analysis Strategies for CRISPR Applications. *Mol. Cell* **79**, 11–29 (2020).

3. Bao, X. R., Pan, Y., Lee, C. M., Davis, T. H. & Bao, G. Tools for experimental and computational analyses of off-target editing by programmable nucleases. *Nat. Protoc.* **16**, 10–26 (2021).

4. Hsu, P. D. *et al.* DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.* **31**, 827–832 (2013).

5. Doench, J. G. *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 1–12 (2016).

6. Chaudhari, H. G. *et al.* Evaluation of Homology-Independent CRISPR-Cas9 Off-Target Assessment Methods. *CRISPR J* **3**, 440–453 (2020).

7. Lessard, S. *et al.* Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. *Proc. Natl. Acad. Sci. U. S. A.* (2017) doi:10.1073/pnas.1714640114.

8. Scott, D. A. & Zhang, F. Implications of human genetic variation in CRISPR-based therapeutic genome editing. *Nat. Med.* (2017) doi:10.1038/nm.4377.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

10. Listgarten, J. *et al.* Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat Biomed Eng* **2**, 38–47 (2018).

11. Labun, K. *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res.* **47**, W171–W174 (2019).

12. Park, J., Bae, S. & Kim, J.-S. Cas-Designer: a web-based tool for choice of CRISPR-Cas9 target sites. *Bioinformatics* **31**, 4014–4016 (2015).

13. Cancellieri, S., Canver, M. C., Bombieri, N., Giugno, R. & Pinello, L. CRISPRitz: rapid, high-throughput and variant-aware in silico off-target site identification for CRISPR genome editing. *Bioinformatics* 1–8 (2019).

14. Lowy-Gallego, E. *et al.* Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Res* **4**, 50 (2019).

15. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, (2020).

16. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

17. Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192–197 (2015).

18. Wu, Y. *et al.* Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nature Medicine* vol. 25 776–783 (2019).

19. Frangoul, H. *et al.* CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia. *N. Engl. J. Med.* (2020) doi:10.1056/NEJMoa2031054.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

20. Walton, R. T., Christie, K. A., Whittaker, M. N. & Kleinstiver, B. P. Unconstrained genome targeting with near-PAMless engineered CRISPR-Cas9 variants. *Science* **368**, 290–296 (2020).

21. ENCODE Project Consortium *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).

22. Demirci, S. *et al.* Durable and robust fetal globin induction without Anemia in rhesus monkeys following autologous hematopoietic stem cell transplant with BCL11A Erythroid enhancer editing. (2019).

23. Vakulskas, C. A. *et al.* A high-fidelity Cas9 mutant delivered as a ribonucleoprotein complex enables efficient gene editing in human hematopoietic stem and progenitor cells.

24. Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).

25. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

26. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).

27. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

28. Clement, K. *et al.* CRISPResso2 provides accurate and rapid genome editing sequence analysis. *Nat. Biotechnol.* **37**, 224–226 (2019).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Notes

Supplementary Table 1 - Complete population frequencies for rs114518452 from gnomAD v3.1[16]

Population	Allele Count	Allele Number	Number of Homozygotes	Allele Frequency
African/African-American	1882	41386	39	0.04547
Other	19	2090	0	0.009091
Latino/Admixed American	100	15246	0	0.006559
South Asian	6	4830	0	0.001242
European (non-Finnish)	10	67992	0	0.0001471
European (Finnish)	0	10612	0	0
Amish	0	912	0	0
East Asian	0	5170	0	0
Middle Eastern	0	316	0	0
Ashkenazi Jewish	0	3470	0	0
XX	1088	77776	23	0.01399
XY	929	74248	16	0.01251
Total	2017	152024	39	0.01327

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Note 1 - CRISPRme input requirements.

CRISPRme is available as an online web app at http://crisprme.di.univr.it/, or offline as a local web app or a standalone command line package (see **Supplementary Notes 3, 6**). The required inputs to perform an online search are: gRNA protospacer(s), a PAM sequence, a genome build with or without the inclusion of genetic variants (1000G, HGDP and/or personal variants), and thresholds of mismatches and RNA/DNA bulges. We report below a detailed description of each required input.

Protospacer(s)

The protospacer is a genomic target sequence (typically 20 nucleotides) matching the guide RNA (gRNA) spacer sequence. The gRNA spacer directs Cas protein binding to the complementary protospacer genomic target sequence in the presence of a protospacer adjacent motif (PAM). CRISPRme accepts a set of gRNA protospacer(s), one per line, each with the same length.

An example of a gRNA protospacer: `GAGTCCGAGCAGAAGAAGAA`

Genomic sequence(s)

CRISPRme can alternatively take as input a set of genomic coordinates in BED format (chromosome# start end) or DNA sequences in FASTA format. The BED file coordinates will be treated as 0-based. To use this type of input, the user must delimit each entry with a >header. The user can input any number of entries.

An example of BED coordinates:

```
>experiment_1
Chr1 100000 101000
Chr1 120000 120100
Chr10 30000 32000
Chr20 40000 41000
```

An example of DNA sequences:

```
>EMX1
GTATGGTGAGTGTCTAGGGGGCCTGTAGGAACCCCTCCAGAAAAATTCTCACAAGCATTTGAAAATCAGT
GACTTGATCTGGAGAAAAATATAGGGCTGGCATTACAA
>BCL11A
AAGAGGTGAGACTGGCTTTTGGACACCAGCGCGCTCACGGTCAAGTGTGCAGCGGGAGGAAAGTAGTCAT
CCCCACAATA
```

PAM sequence

The PAM is a short (~2-6 nucleotide) DNA sequence adjacent to the protospacer necessary for the Cas protein to bind to a specific DNA target. CRISPRme supports a set of PAMs and users must select one of them in order to perform the search. The software supports both 3’ (e.g. SpCas9) and 5’ (e.g. Cas12a) PAM sequences.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An example of PAM: `NGG`

Genome builds

The genome builds are based on FASTA files from UCSC. The hg38 and hg19 genomic builds are available with the option to incorporate variants from 1000G, HGDP, and/or personal variants in the search.

Search thresholds

CRISPRme allows users to specify the number of mismatches, DNA and RNA bulges tolerated in enumerating potential off-targets. The web-tool allows up to 6 mismatches and up to 2 RNA/DNA bulges. However, for the command line version, these thresholds can be set freely and depend only on the available computational resources (see **Supplementary Note 3**).

Functional annotations

To assess the potential impact of off-target activity, CRISPRme provides a set of functional annotations for coding and non-coding regions. The annotations are based on files obtained from the Encyclopedia of DNA Elements (ENCODE) containing candidate cis regulatory elements[21] and from GENCODE[24] containing annotations for protein coding genes, transcribed but untranslated regions, and introns. In the offline versions of CRISPRme, users can add custom genome annotations, such as cell-type specific chromatin marks or off-target sites nominated by in vitro and/or cellular assays as simple BED files (see **Supplementary Note 3**).

Supplementary Note 2 - Details of the CRISPRme implementation.

CRISPRme software architectures

The CRISPRme web version and front end was developed in Dash, a Python framework to create responsive and interactive web applications (https://plotly.com/dash/). The back end (graphical report generation and data analysis) is based on Python and bash scripts. The search engine is developed in C++ to exploit its speed and stability and to fully leverage parallel computation and compiler optimizations.

CRISPRme genome enrichment with variants

CRISPRme performs the search of potential off-targets based on reference genomic sequences stored in FASTA files. A reference genome can be “enriched” with genetic variants (SNPs and INDELs) encoded in VCF files obtained for example from 1000G, HGDP and/or personal data. Enriched genomes are created by encoding SNPs using IUPAC notation, i.e. nucleotides corresponding to genetic variants can be represented via ambiguous DNA characters. For example, if at a given position the reference allele is G and the alternative allele is A, the tool encodes the two alternatives by using the ambiguous nucleotide R, which corresponds to the IUPAC code for G or A. Based on this procedure, the enriched genome contains all the SNP variants belonging to different samples, including multiallelic sites with three or more observed alleles. CRISPRme treats INDELs differently due to the nature of the variant itself. For each INDEL, it creates a fake chromosome containing the modified DNA sequence and 50 surrounding

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

nucleotides (25 bp on each side). Finally, the association between samples and variants is based on a hash table to allow efficient querying of which sample(s) contain a given SNP/INDEL.

CRISPRme indexing, search and analysis

To perform efficient search operations, CRISPRme creates an index of reference/enriched genomes. This index encodes all the possible candidate off-targets in a tree-based data structure and can be used to efficiently find reference or variant enriched sites[13]. In addition, CRISPRme introduces individual-specific analysis, extracting from the IUPAC encoding of the enriched genome haplotype-specific off-targets and the corresponding samples. For each site, CRISPRme also reports the cutting frequency determination (CFD) score, and if multiple possible alignments overlap at a given site (e.g. for RNA/DNA bulges), the alignment with the highest score is reported. CRISPRme currently adopts the CFD score because it can be efficiently computed for thousands of sites, handles bulges and has been shown to perform reasonably well in predicting off-targets as validated by deep sequencing[3]. However, CRISPRme can be easily extended to support other predictive off-target scores. In addition, a global score called summarized CFD is provided for each guide and defined as follows:

$$summarized\ CDF = 100 * 100/(100 + \sum_{target_i} CFD(target_i))$$

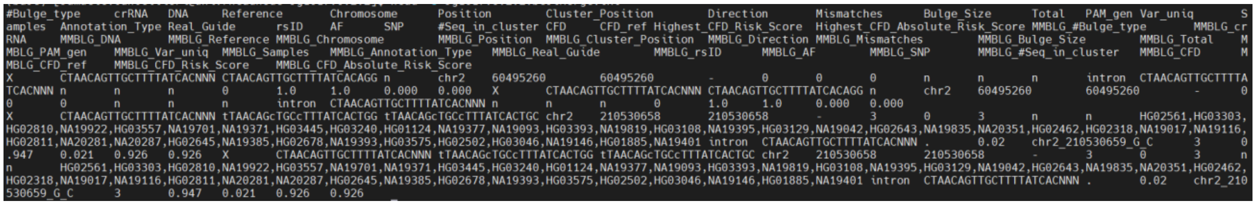
where $target_i$ is one of the enumerated off-targets from the search.

Importantly, thanks to the constructed hash table containing the mapping between samples and variants, after the initial search based on the IUPAC code, CRISPRme filters the results by reporting only targets matching existing haplotypes in the populations and the corresponding individuals.

Given that multiple alignments may correspond to a given genomic region, CRISPRme outputs two lists of off-targets. The first file called "bestMerge" includes a single off-target site per genomic region, merging all possible off-targets within 3 bp. The off-targets included in the list are based on the highest CFD score by default, but users can select other criteria such as fewest mismatches and bulges. When the CFD score is identical, the reference alignment is favored over variant alignments. The second file is called "altMerge" (a.k.a. alternative alignments) and includes all the off-targets not included in the first file. This file preserves alternative alignments for off-targets as well as those matching the reference genome or containing other variants with lower CFD scores. A third file created by CRISPRme, called “integrated_results,” is a compact version of the bestMerge file containing only targets ordered by highest CFD score. The file also contains the GENCODE annotations and distance from the nearest gene for each target.

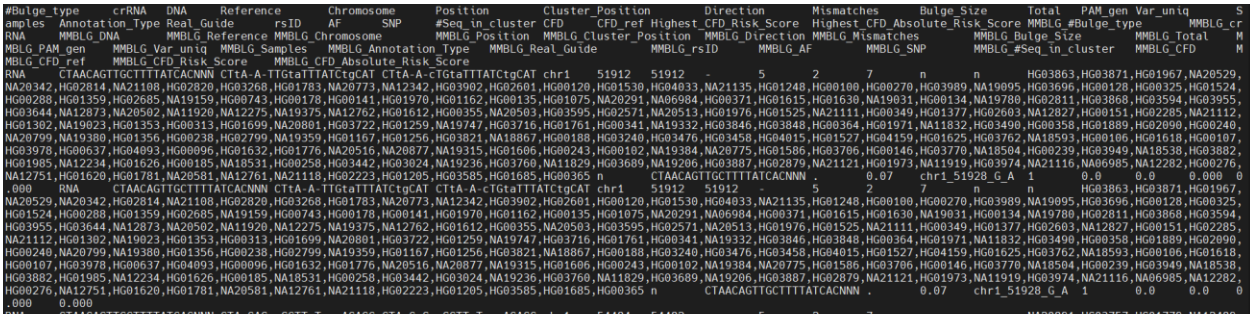
bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

First lines of the bestMerge file:



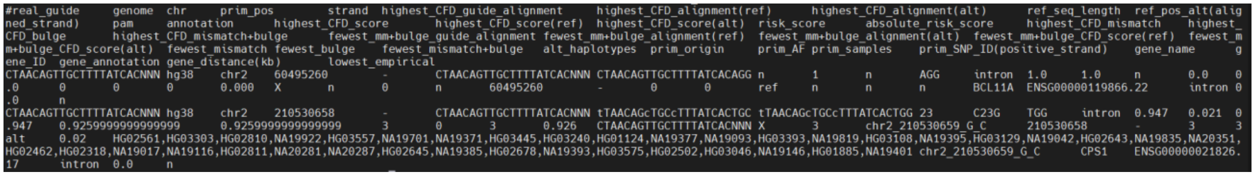
```
#Bulge_type     crRNA   DNA     Reference       Chromosome      Position        Cluster_Position        Direction       Mismatches      Bulge_Size      Total   PAM_gen Var_uniq        S
amples  Annotation_Type Real_Guide      rsID    AF      SNP     #Seq_in_cluster CFD     CFD_ref Highest_CFD_Risk_Score  Highest_CFD_Absolute_Risk_Score MMBLG_#Bulge_type       MMBLG_cr
RNA     MMBLG_DNA       MMBLG_Reference MMBLG_Chromosome        MMBLG_Position  MMBLG_Cluster_Position  MMBLG_Direction MMBLG_Mismatches        MMBLG_Bulge_Size        MMBLG_Total     M
MBLG_PAM_gen    MMBLG_Var_uniq  MMBLG_Samples   MMBLG_Annotation_Type   MMBLG_Real_Guide        MMBLG_rsID      MMBLG_AF        MMBLG_SNP       MMBLG_#Seq_in_cluster   MMBLG_CFD       M
MBLG_CFD_ref    MMBLG_CFD_Risk_Score    MMBLG_CFD_Absolute_Risk_Score
X       CTAACAGTTGCTTTTATCACNNN CTAACAGTTGCTTTTATCACAGG n       chr2    60495260        60495260        -       0       0       0       n       n       n       intron  CTAACAGTTGCTTTTA
TCACNNN n       n       n       0       1.0     1.0     0.000   0.000   X       CTAACAGTTGCTTTTATCACNNN CTAACAGTTGCTTTTATCACAGG n       chr2    60495260        60495260        -       0
0       0       n       n       n       intron  CTAACAGTTGCTTTTATCACNNN n       n       n       0       1.0     1.0     0.000   0.000
X       CTAACAGTTGCTTTTATCACNNN tTAACAGcTGCcTTTATCACTGG tTAACAGcTGCcTTTATCACTGC chr2    210530658       210530658       -       3       0       3       n       n       HG02561,HG03303,
HG02810,NA19922,HG03557,NA19701,NA19371,HG03445,HG03240,HG01124,NA19377,NA19093,HG03393,NA19819,HG03108,NA19395,HG03129,NA19042,HG02643,NA19835,NA20351,HG02462,HG02318,NA19017,NA19116,
HG02811,NA20281,NA20287,HG02645,NA19385,HG02678,NA19393,HG03575,HG02502,HG03046,NA19146,HG01885,NA19401 intron  CTAACAGTTGCTTTTATCACNNN .       0.02    chr2_210530659_G_C      3       0
.947    0.021   0.926   0.926   X       CTAACAGTTGCTTTTATCACNNN tTAACAGcTGCcTTTATCACTGG tTAACAGcTGCcTTTATCACTGC chr2    210530658       210530658       -       3       0       3       n
n       HG02561,HG03303,HG02810,NA19922,HG03557,NA19701,NA19371,HG03445,HG03240,HG01124,NA19377,NA19093,HG03393,NA19819,HG03108,NA19395,HG03129,NA19042,HG02643,NA19835,NA20351,HG02462,
HG02318,NA19017,NA19116,HG02811,NA20281,NA20287,HG02645,NA19385,HG02678,NA19393,HG03575,HG02502,HG03046,NA19146,HG01885,NA19401 intron  CTAACAGTTGCTTTTATCACNNN .       0.02    chr2_210
530659_G_C      3       0.947   0.021   0.926   0.926
```

First lines of the altMerge file:



```
#Bulge_type     crRNA   DNA     Reference       Chromosome      Position        Cluster_Position        Direction       Mismatches      Bulge_Size      Total   PAM_gen Var_uniq        S
amples  Annotation_Type Real_Guide      rsID    AF      SNP     #Seq_in_cluster CFD     CFD_ref Highest_CFD_Risk_Score  Highest_CFD_Absolute_Risk_Score MMBLG_#Bulge_type       MMBLG_cr
RNA     MMBLG_DNA       MMBLG_Reference MMBLG_Chromosome        MMBLG_Position  MMBLG_Cluster_Position  MMBLG_Direction MMBLG_Mismatches        MMBLG_Bulge_Size        MMBLG_Total     M
MBLG_PAM_gen    MMBLG_Var_uniq  MMBLG_Samples   MMBLG_Annotation_Type   MMBLG_Real_Guide        MMBLG_rsID      MMBLG_AF        MMBLG_SNP       MMBLG_#Seq_in_cluster   MMBLG_CFD       M
MBLG_CFD_ref    MMBLG_CFD_Risk_Score    MMBLG_CFD_Absolute_Risk_Score
RNA     CTAACAGTTGCTTTTATCACNNN CTtA-A-TTGtaTTTATCtgCAT CTtA-A-cTGtaTTTATCtgCAT chr1    51912   51912   -       5       2       7       n       n       HG03863,HG03871,HG01967,NA20529,
NA20342,HG02814,NA21108,HG02820,HG03268,HG01783,NA20773,NA12342,HG03902,HG02601,HG00120,HG01530,HG04033,NA21135,HG01248,HG00100,HG00270,HG03989,NA19095,HG03696,HG00128,HG00325,HG01524,
HG00288,HG01359,HG02685,NA19159,HG00743,HG00178,HG00141,HG01970,HG01162,HG00135,HG01075,NA20291,NA06984,HG00371,HG01615,HG01630,NA19031,HG00134,NA19780,HG02811,HG03868,HG03594,HG03955,
HG03644,NA12873,NA20502,NA11920,NA12275,NA19375,NA12762,HG01612,HG00355,NA20503,HG03595,HG02571,NA20513,HG01976,HG01525,NA21111,HG00349,HG01377,HG02603,NA12827,HG00151,HG02285,NA21112,
HG01302,NA19023,HG01353,HG00313,HG01699,NA20801,HG03722,HG01259,NA19747,HG03716,HG01761,HG00341,NA19332,HG03846,HG03848,HG00364,HG01971,NA11832,HG03490,HG00358,HG01889,HG02090,HG00240,
NA20799,NA19380,HG01356,HG00238,HG02799,NA19359,HG01167,HG01256,HG03821,NA18867,HG00188,HG03240,HG03476,HG03458,HG04015,HG01527,HG04159,HG01625,HG03762,NA18593,HG00106,HG01618,HG00107,
HG03978,HG00637,HG04093,HG00096,HG01632,HG01776,NA20516,NA20877,NA19315,HG01606,HG00243,HG00102,NA19384,NA20775,HG01586,HG03706,HG00146,HG03770,NA18504,HG00239,HG03949,NA18538,HG03882,
HG01985,NA12234,HG01626,HG00185,NA18531,HG00258,HG03442,HG03024,NA19236,HG03760,NA11829,HG03689,NA19206,HG03887,HG02879,NA21121,HG01973,NA11919,HG03974,NA21116,NA06985,NA12282,HG00276,
NA12751,HG01620,HG01781,NA20581,NA12761,NA21118,HG02223,HG01205,HG03585,HG01685,HG00365 n       CTAACAGTTGCTTTTATCACNNN .       0.07    chr1_51928_G_A  1       0.0     0.0     0.000   0
.000    RNA     CTAACAGTTGCTTTTATCACNNN CTtA-A-TTGtaTTTATCtgCAT CTtA-A-cTGtaTTTATCtgCAT chr1    51912   51912   -       5       2       7       n       n       HG03863,HG03871,HG01967,
NA20529,NA20342,HG02814,NA21108,HG02820,HG03268,HG01783,NA20773,NA12342,HG03902,HG02601,HG00120,HG01530,HG04033,NA21135,HG01248,HG00100,HG00270,HG03989,NA19095,HG03696,HG00128,HG00325,
HG01524,HG00288,HG01359,HG02685,NA19159,HG00743,HG00178,HG00141,HG01970,HG01162,HG00135,HG01075,NA20291,NA06984,HG00371,HG01615,HG01630,NA19031,HG00134,NA19780,HG02811,HG03868,HG03594,
HG03955,HG03644,NA12873,NA20502,NA11920,NA12275,NA19375,NA12762,HG01612,HG00355,NA20503,HG03595,HG02571,NA20513,HG01976,HG01525,NA21111,HG00349,HG01377,HG02603,NA12827,HG00151,HG02285,
NA21112,HG01302,NA19023,HG01353,HG00313,HG01699,NA20801,HG03722,HG01259,NA19747,HG03716,HG01761,HG00341,NA19332,HG03846,HG03848,HG00364,HG01971,NA11832,HG03490,HG00358,HG01889,HG02090,
HG00240,NA20799,NA19380,HG01356,HG00238,HG02799,NA19359,HG01167,HG01256,HG03821,NA18867,HG00188,HG03240,HG03476,HG03458,HG04015,HG01527,HG04159,HG01625,HG03762,NA18593,HG00106,HG01618,
HG00107,HG03978,HG00637,HG04093,HG00096,HG01632,HG01776,NA20516,NA20877,NA19315,HG01606,HG00243,HG00102,NA19384,NA20775,HG01586,HG03706,HG00146,HG03770,NA18504,HG00239,HG03949,NA18538,
HG03882,HG01985,NA12234,HG01626,HG00185,NA18531,HG00258,HG03442,HG03024,NA19236,HG03760,NA11829,HG03689,NA19206,HG03887,HG02879,NA21121,HG01973,NA11919,HG03974,NA21116,NA06985,NA12282,
HG00276,NA12751,HG01620,HG01781,NA20581,NA12761,NA21118,HG02223,HG01205,HG03585,HG01685,HG00365 n       CTAACAGTTGCTTTTATCACNNN .       0.07    chr1_51928_G_A  1       0.0     0.0     0
.000    0.000
```

First lines of the integrated_result files:



```
#real_guide     genome  chr     prim_pos        strand  highest_CFD_guide_alignment     highest_CFD_alignment(ref)      highest_CFD_alignment(alt)      ref_seq_length  ref_pos_alt(alig
ned_strand)     pam     annotation      highest_CFD_score       highest_CFD_score(ref)  highest_CFD_score(alt)  risk_score      absolute_risk_score     highest_CFD_mismatch    highest_
CFD_bulge       highest_CFD_mismatch+bulge      fewest_mm+bulge_guide_alignment fewest_mm+bulge_alignment(ref)  fewest_mm+bulge_alignment(alt)  fewest_mm+bulge_CFD_score(ref)  fewest_m
m+bulge_CFD_score(alt)  fewest_mismatch fewest_bulge    fewest_mismatch+bulge   alt_haplotypes  prim_origin     prim_AF prim_samples    prim_SNP_ID(positive_strand)    gene_name       g
ene_ID  gene_annotation gene_distance(kb)       lowest_empirical
CTAACAGTTGCTTTTATCACNNN hg38    chr2    60495260        -       CTAACAGTTGCTTTTATCACNNN CTAACAGTTGCTTTTATCACAGG n       1       n       AGG     intron  1.0     1.0     n       0.0     0
.0      0       0       0       0.000   X       n       0       n       60495260        -       0       0       ref     n       n       n       BCL11A  ENSG00000119866.22      intron  0
.0      n
CTAACAGTTGCTTTTATCACNNN hg38    chr2    210530658       -       CTAACAGTTGCTTTTATCACNNN tTAACAGcTGCcTTTATCACTGC tTAACAGcTGCcTTTATCACTGG 23      C23G    TGG     intron  0.947   0.021   0
.947    0.9259999999999999      0.9259999999999999      3       0       3       0.926   CTAACAGTTGCTTTTATCACNNN X       3       chr2_210530659_G_C      210530658       -       3       3
alt     0.02    HG02561,HG03303,HG02810,NA19922,HG03557,NA19701,NA19371,HG03445,HG03240,HG01124,NA19377,NA19093,HG03393,NA19819,HG03108,NA19395,HG03129,NA19042,HG02643,NA19835,NA20351,
HG02462,HG02318,NA19017,NA19116,HG02811,NA20281,NA20287,HG02645,NA19385,HG02678,NA19393,HG03575,HG02502,HG03046,NA19146,HG01885,NA19401 chr2_210530659_G_C      CPS1    ENSG00000021826.
17      intron  0.0     n
```

Supplementary Note 3 - Search with custom personal genomes, VCFs, annotation files and PAMs.

The CRISPRme webapp can be also deployed to any private server (**Supplementary Note 7**) to enable the inclusion of personal genomes, VCF files and custom annotations. The required inputs are similar to the online version (see **Supplementary Note 1**) but the user can upload any personal data as long as they follow the format described below.

CRISPRme automatically creates the following folders to help organize the data that needs to be provided by the user. See **Supplementary Fig. 1** for an example**.**

- `Genomes`: it contains the genomes in FASTA format. Each genome must be saved into a separate folder. The name of the folder will be used to identify the genome itself and all the linked data such as VCFs and samplesIDs. In **Supplementary Fig. 1** the `Genomes` folder contains:
  - `hg19`: FASTA files for human reference genome build 19.
  - `hg38`: FASTA files for human reference genome build 38.
- `VCFs`: it contains the VCF datasets. Each dataset must consist of chromosome separated VCF files and be saved into a separate folder. The name of the folder is

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

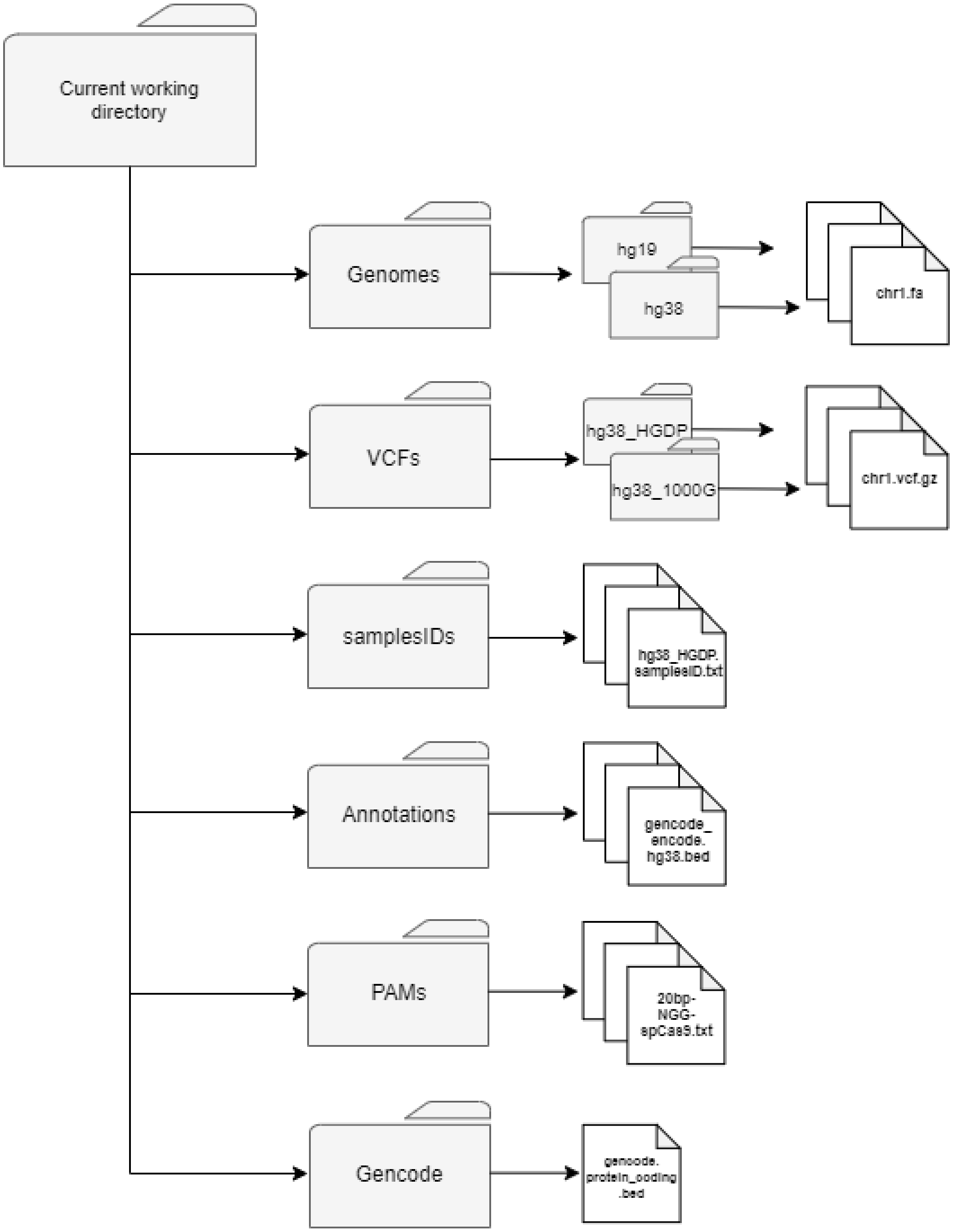
composed of the genome release name followed by the VCF dataset name. In **Supplementary Fig. 1** the `VCFs` folder contains:

  - `hg38_HGDP`: VCF files from HGDP based on hg38.
  - `hg38_1000G`: VCF files from 1000G based on hg38.

- `samplesIDs:` it contains the samplesID files, one for each VCF dataset. The name of the file is composed of the name of the corresponding VCF folder followed by the samplesID suffix. In **Supplementary Fig. 1** the `sampleIDs` folder contains:
  - `hg38_HGDP.samplesID.txt`: tabulated file with header to identify samples for the HGDP dataset.
  - `hg38_1000G.samplesID.txt`: tabulated file with header to identify samples for the 1000G dataset.
- `Annotations:` it contains the annotation BED files. In **Supplementary Fig. 1** the `Annotations` folder contains:
  - `gencode_encode.hg38.bed`: a BED file with annotations from the GENCODE project and ENCODE datasets.
- `PAMs:` it contains the PAM text files, one for each PAM. In **Supplementary Fig. 1** the `PAMs` folder contains:
  - `20bp-NGG-spCas9.txt`: a text file with a single line PAM sequence. The file name contains the position and length of the crRNA (20 bp), followed by the PAM sequence (NGG) and the Cas protein (SpCas9).

Supplementary Figure 1. CRISPRme data storing structure. The directories created by CRISPRme running as an offline web-app or command line, used to upload personal data such as genomes, VCFs, annotations and PAMs.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



The following sections detail the format of the required files in each folder.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Personal genome build

A personal genome can be added as a set of FASTA files (.fa), one for each chromosome (chr1.fa, chr2.fa, chrN.fa), all placed in a single folder. The supported format is based on the specifications for genome assemblies of the UCSC Genome Browser (e.g. hg38).

An example of a personal chromosome in FASTA format:

```
>chr10
ctataatcccagcttgttgggaggccaaggcaggaggatcacttgaagcc
caggagtttgagacgagcctaagcaacatagcaagaccctatctctacaa
TTATAAATATAGTATTTGTTAATATTTGgccaggcgtggtagtacatgcc
Tgtaggcccagctacttggggagaggaggcaggaggatcacttgagggcc
```

VCF files & phasing information

The Variant Call Format (VCF) file stores genetic variant information. CRISPRme accepts compressed VCFs in the 1000G[14]or GATK v3/4 format[25] (VCF v4.1 or newer) in chromosome separated files. VCF files containing variants from multiple chromosomes must be split by chromosome, which can be accomplished using BCFtools[26] with the following command:

```
for chr in chr{1..22} chrX chrY
do
  bcftools view input.vcf.gz --regions $chr | bgzip -c > ${chr}.vcf.gz
done
```

VCF files contain positional information (chromosome and position), reference and alternative nucleotide(s), and may contain sample genotype information (which, if present, can be either phased or unphased). A sample information file must be also provided for CRISPRme, i.e. a tabulated list containing all the samples present in the VCF files with their respective population (e.g. GBR), super-population (e.g. EUR) and sex (e.g. male) information. If the population, super-population, and/or sex information is not available, a placeholder such as 'n' can be used instead. VCF files from 1000G and HGDP are similar in format and report the same data. In 1000G VCFs, each sample column contains the phased genotype. In HGDP VCFs, each sample column contains the unphased genotype if available, along with supplementary data like the read depth and the genotype quality.

An example of VCF file header information from 1000G and HGDP:

- #CHROM – Chromosome
- POS – Position of the variant (1-start)
- ID – rsID or other identifier of the variant
- REF – Reference nucleotide
- ALT – Alternative nucleotide
- QUAL – Phred-scaled quality score for the assertion made in ALT

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- `FILTER` – Testify if the ALT nucleotide passed quality filters. Note that only variant calls that pass all quality filters (denoted with “PASS” in this field) are used for CRISPRme analysis.
- `INFO` – a non-standard field containing details on the variants, including:
  - allele frequency
  - number of total samples
  - total alleles
  - any other possible information on the variants
- `FORMAT` – the format for the variant data reported in the subsequent columns
- `SAMPLE IDs` – a set of columns reporting the IDs of all the samples in the VCF

An example of the tab-delimited sample IDs text file needed:

```
#SAMPLE_ID POPULATION_ID    SUPERPOPULATION_ID     SEX
HG00096    GBR              EUR                    male
HG00097    GBR              EUR                    female
HG00098    GBR              EUR                    male
HG00099    GBR              EUR                    female
HG00100    GBR              EUR                    female
```

- `#SAMPLE_ID` – sample identifier as reported in the VCF file header
- `POPULATION_ID` – population name as reported in the VCF file
- `SUPERPOPULATION_ID` – super-population name
- `SEX` – sex

In addition, CRISPRme supports gnomAD v3.1 VCFs based on an integrated parser. This process converts each gnomAD VCF into a CRISPRme-supported VCF. The parser takes as input a directory containing gnomAD v3.1 VCFs and a pre-generated samplesID file as shown in the following example. The pre-generated file simulates a set of samples, each one belonging to a gnomAD super-population, to be included in the gnomAD VCFs. The file is created by inspecting gnomAD v3.1 VCFs and extracting all the super-populations reported in the files (AFR, AMR, ASJ, EAS, FIN, NFE, MID, SAS and OTH). This file is provided with the test package (see **Supplementary Note 7**) and can be used and extended, if necessary, with any gnomAD v3.1 VCF file.

Example call of the VCF converter:

```
crisprme.py gnomAD-converter --gnomAD_VCFdir VCFs/gnomAD_VCFdir/ --samplesID
samplesIDs/hg38_gnomAD.samplesID.txt --thread 4
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Example of samples ID file to use with gnomAD v3.1 data:

```
#SAMPLE_ID POPULATION_ID    SUPERPOPULATION_ID    SEX
afr        AFR              AFR                   n
ami        AMI              AMI                   n
amr        AMR              AMR                   n
asj        ASJ              ASJ                   n
eas        EAS              EAS                   n
fin        FIN              FIN                   n
nfe        NFE              NFE                   n
mid        MID              MID                   n
sas        SAS              SAS                   n
oth        OTH              OTH                   n
```

CRISPRme supports three types of VCF data:

1. Individual-level, phased VCFs (such as from 1000G) with genotypes for all samples, as well as information on which chromosome in a homologous pair each genotype call came from (delimited with '|'). When phase information is provided, CRISPRme is haplotype-aware and assesses off-target potential only for observed haplotypes.

2. Individual-level, unphased VCFs (such as from HGDP) with genotypes for all samples but lacking phase information (delimited with '/'). When individual-level information is provided but phase information is unavailable, CRISPRme assesses off-target potential for all possible haplotypes for each individual. However, this analysis may include false haplotypes in the case of nearby heterozygous variants present in the same individual.

3. Population-level VCFs (such as from gnomAD) with variant information for overall populations. CRISPRme can assess off-target potential when provided with many known population variants, but note that without individual-level and phase information, many unobserved haplotypes (with variants only observed in distinct individuals) may be included in the analysis.

Custom annotation file and empirical off-target data

The command line version of CRISPRme can be used with a custom annotation file in BED standard format to annotate the potential off-targets. This file can be used in combination with the ENCODE+GENCODE annotation file present in CRISPRme or alone as a substitute file to annotate all the targets. If an off-target is associated with different annotations, each one will be reported.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

An example annotation file, where the columns indicate the chromosome, the genomic start and end coordinates and the annotation description:

```
chr1    790397  790626  CTCF
chr1    869716  870065  DNase
chr1    190865  191071  distal_enhancer
chr1    11869   12227   exon
chr1    12227   12613   intron
chr1    778562  778912  promoter
chr1    181251  181601  proximal_enhancer
```

The coordinates in the custom annotation file are 0-start based to respect the BED format specifications (https://genome.ucsc.edu/FAQ/FAQformat.html#format1).

The command line version of CRISPRme also supports the integration of user-provided empirical off-target results, which can be useful for creating a master summary of candidate off-target sites nominated by in silico, in vitro and cellular methods. Below is an example of the input format, with the data representing several previously nominated off-target sites for gRNA #1617[18].

An example of empirical off-target data for integration:

```
chr10 33753323    33753346    4     CIRCLEseq   OT1 aTtACAGcTGCaTTTATCACAGG
chr21 17537877    17537900    3     CIRCLEseq   OT2 CTAACA-aTGCTTTcATCACGGG
chr1  97697292    97697315    4     CRISPOR     OT21 CaAACAGaTtCTTTTATCtCTGG
chr20 6492313     6492336     4     CRISPOR     OT22 gagACAGTgGCTTTTATCACAGG
```

In order, the columns indicate chromosome, start position, end position, number of mismatches + bulges, user-defined column name for the CRISPRme output file, empirical information to integrate and off-target motif (including PAM). If the motif is not available, it can be substituted with any placeholder (‘n’ for example). The empirical information to integrate can be anything the user desires, such other identifiers (as shown here) or numerical scores.

PAM sequence

In order to use a new PAM, the user must add a new file in the PAMs folder with the following name convention:

##bp_protospacer-PAM_seq-nuclease.txt if the PAM is located at the 3’ end of the targeted sequence
PAM_seq-##bp_protospacer-nuclease.txt if the PAM is located at the 5’ end of the targeted sequence

The content of this file must consist of a series of Ns representing the protospacer and the actual PAM sequence immediately preceding or following it as appropriate. Then, after a whitespace, there must be an integer representing the length of the PAM sequence. E.g. if the PAM considered

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

is NGG → 3. If the PAM sequence is located 5' of the protospacer, then this value must be negative. E.g. if the PAM considered is TTTV → -4.

An example of a PAM file (NGG for SpCas9) with a protospacer length of 20 nt:

```
20bp-NGG-spCas9.txt
NNNNNNNNNNNNNNNNNNNNNGG 3
```

Supplementary Note 4 - Comparison of CRISPRme with available tools.

Although numerous tools are available to enumerate CRISPR-Cas9 off-targets, to our knowledge only two previous studies[7,8] have reported computational strategies to assess off-target potential in presence of genetic variants. Only *crispRtool* from Lessard et al.[7,8] provides a general command line software. Therefore, we decided to focus our comparison by assessing the features and running times of CRISPRme (v1.7.7) and crisprRtool (v2.0.5) on the same hardware (AMD Ryzen Threadripper 3970X 32-Core Processor clocked at 2.2 Ghz with 124 GBs of RAM) to provide a fair assessment. For our tests we used the 1617 sgRNA, NGG PAM, variants from 1000G and a variable number of mismatches and bulges.

Briefly, crisprRtool first adds variants (SNP only) to the reference genome using the IUPAC notation, then searches the input list of sgRNAs on the variant genome and reports a list of putative targets and off-targets with IUPAC nucleotides. The tool also offers the possibility to search each VCF file individually to resolve haplotypes (SNPs and INDELs) of the reported targets. However, for this step the user needs to manually edit and execute a script for each VCF file. In addition, the search operation with crispRtool allows a maximum of 5 mismatches, does not account for bulges, and is not flexible in terms of PAM location relative to the target motif (only 3' is supported).

Using 5 mismatches and the settings described above, crispRtool took 9 hours to complete the non-haplotype resolved search. The haplotype resolved search only on chr1 and using variants from 1000G (6 million SNPs and INDEL variants) took ~37 hours.
Conservatively extrapolating to all other chromosomes, the entire search would take more than 300 hours and will not be as complete as the search CRISPRme offers due to the lack of graphical reports and textual summaries encompassing results from all chromosomes.

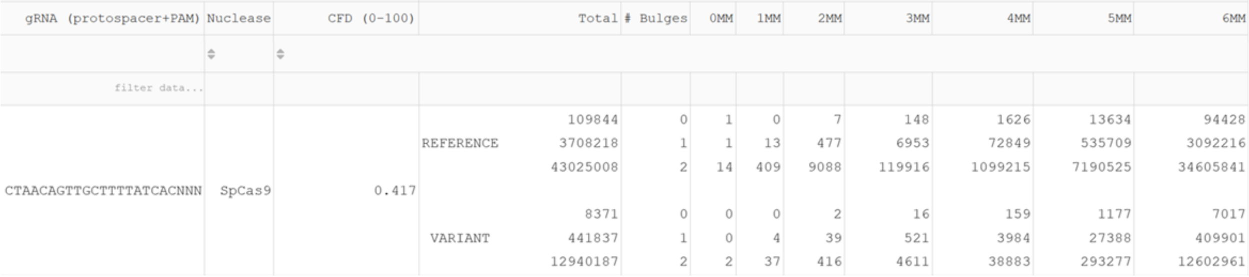
On the other hand, by leveraging an efficient genome index and auxiliary data structures that are constructed only once during the installation (~4 hours), CRISPRme can complete a search of a gRNA with 5 mismatches in ~1 hour. Notably, the haplotype resolved search on the entire genome with up to 6 mismatches and 2 DNA/RNA bulges only takes 2 hours (excluding the guide-independent indexing operation performed during the setup of the tool) and also includes a summary report.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Note 5 - CRISPRme output and graphical reports.

CRISPRme summarizes the results in a table highlighting for each gRNA its CFD score and the count of on-targets and off-targets found in the reference and variant genomes grouped by number of mismatches and bulges (**Supplementary Fig. 2**).

Supplementary Figure 2. CRISPRme result summary. A table summarizing results based on the search with sg1617, NNN PAM, up to 6 mismatches and 2 DNA or RNA bulges on the human reference genome supplemented with the 1000G dataset with 5 super-populations as well as HGDP with 7 super-populations. The table reports the nuclease, the summarized CFD score and the number of targets in each category of mismatches and bulges. In the top left corner there is a “Download General Table”’ button allowing the download of the table as a text file.



gRNA (protospacer+PAM)	Nuclease	CFD (0-100)		Total	# Bulges	0MM	1MM	2MM	3MM	4MM	5MM	6MM
	⇕	⇕										
filter data...												
CTAACAGTTGCTTTTATCACNNN	SpCas9	0.417	REFERENCE	109844	0	1	0	7	148	1626	13634	94428
				3708218	1	1	13	477	6953	72849	535709	3092216
				43025008	2	14	409	9088	119916	1099215	7190525	34605841
			VARIANT	8371	0	0	0	2	16	159	1177	7017
				441837	1	0	4	39	521	3984	27388	409901
				12940187	2	2	37	416	4611	38883	293277	12602961

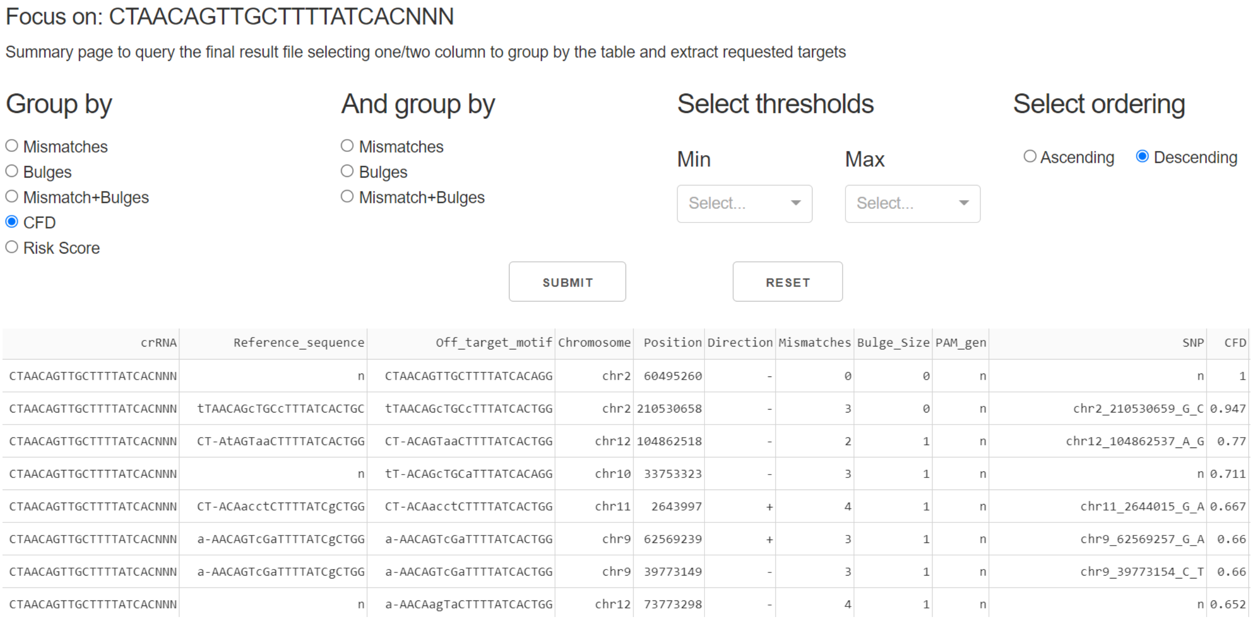
In addition, for each guide, six different interactive reports are generated and are available to be downloaded: *Custom Ranking, Summary by Mismatches/Bulges, Summary by Sample, Query Genomic Region, Graphical Reports* and *Personal Risk Cards* (described in **Supplementary Note 6**).

Custom Ranking

In this report, users can filter and rank potential off-targets based on number of mismatches and/or bulges, CFD score, Risk Score (increase in CFD score due to genetic variants), or a combination of them (**Supplementary Fig. 3**).

Supplementary Figure 3. CRISPRme ranking and filtering of off-targets. Users can define filters, orders and group-by operations to easily retrieve results based on a custom logic suitable for their application. The “DNA” column shows the predicted off-target motif, with the corresponding reference genome sequence shown in the “Reference” column if nominated by a genetic variant. The “PAM_gen” column indicates if a genetic variant generates a PAM not found in the reference genome (‘n’ if false – note that there are no PAM generation events indicated here because a NNN PAM was used in the search).

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Focus on: CTAACAGTTGCTTTTATCACNNN

Summary page to query the final result file selecting one/two column to group by the table and extract requested targets

Group by

○ Mismatches
○ Bulges
○ Mismatch+Bulges
● CFD
○ Risk Score

And group by

○ Mismatches
○ Bulges
○ Mismatch+Bulges

Select thresholds

Min: Select...
Max: Select...

Select ordering

○ Ascending ● Descending

crRNA	Reference_sequence	Off_target_motif	Chromosome	Position	Direction	Mismatches	Bulge_Size	PAM_gen	SNP	CFD
CTAACAGTTGCTTTTATCACNNN	n	CTAACAGTTGCTTTTATCACAGG	chr2	60495260	-	0	0	n	n	1
CTAACAGTTGCTTTTATCACNNN	tTAACAGcTGCcTTTATCACTGC	tTAACAGcTGCcTTTATCACTGG	chr2	210530658	-	3	0	n	chr2_210530659_G_C	0.947
CTAACAGTTGCTTTTATCACNNN	CT-AtAGTaaCTTTTATCACTGG	CT-ACAGTaaCTTTTATCACTGG	chr12	104862518	-	2	1	n	chr12_104862537_A_G	0.77
CTAACAGTTGCTTTTATCACNNN	n	tT-ACAGcTGCaTTTATCACAGG	chr10	33753323	-	3	1	n	n	0.711
CTAACAGTTGCTTTTATCACNNN	CT-ACAacctCTTTTATCgCTGG	CT-ACAacctCTTTTATCACTGG	chr11	2643997	+	4	1	n	chr11_2644015_G_A	0.667
CTAACAGTTGCTTTTATCACNNN	a-AACAGTcGaTTTTATCgCTGG	a-AACAGTcGaTTTTATCACTGG	chr9	62569239	+	3	1	n	chr9_62569257_G_A	0.66
CTAACAGTTGCTTTTATCACNNN	a-AACAGTcGaTTTTATCgCTGG	a-AACAGTcGaTTTTATCACTGG	chr9	39773149	-	3	1	n	chr9_39773154_C_T	0.66
CTAACAGTTGCTTTTATCACNNN	n	a-AACAagTaCTTTTATCACTGG	chr12	73773298	-	4	1	n	n	0.652

Summary by Mismatches/Bulges

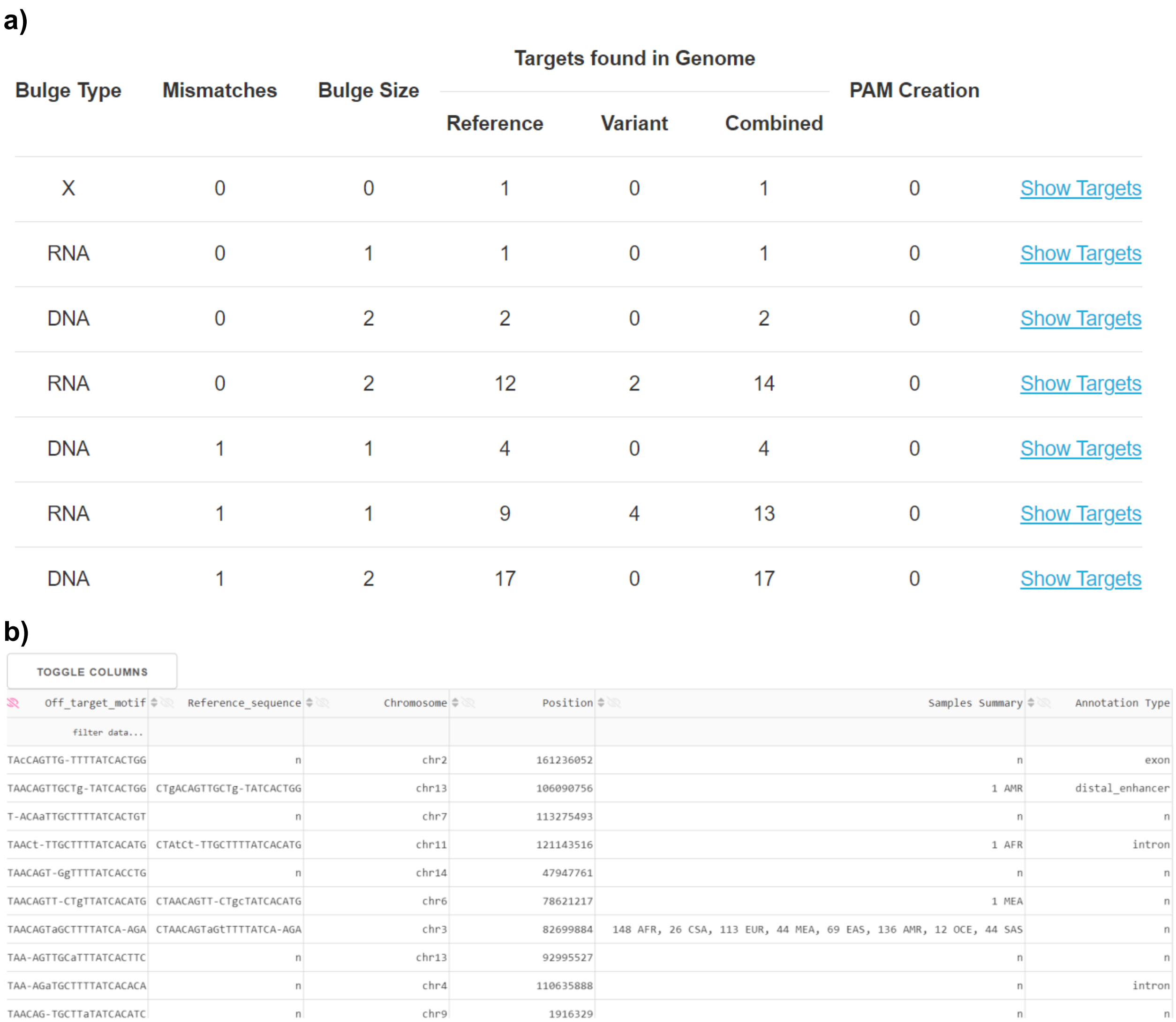
This report shows a matrix separating targets into subgroups based on the type of target, mismatch count and bulge size. “X” targets contain only mismatches, “DNA” targets contain DNA bulges (and may contain mismatches), and “RNA” targets contain RNA bulges (and may contain mismatches) (**Supplementary Fig. 4**).

Supplementary Figure 4. **CRISPRme summary results by Mismatches/Bulges.**

a) Mismatches/Bulges summary table showing the first 7 of 33 rows for a search with up to 6 mismatches and 2 DNA or RNA bulges. The combined column indicates the sum of reference and variant targets.

b) View of the Show Targets with 1 mismatch and 1 RNA bulge. The user can select which column see using the Toggle Columns button on top of the table. In this view we show, Off-target motif, Reference sequence, Chromosome, Position, Samples summary and Annotation type.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Bulge Type	Mismatches	Bulge Size	Targets found in Genome			PAM Creation	
			Reference	Variant	Combined		
X	0	0	1	0	1	0	Show Targets
RNA	0	1	1	0	1	0	Show Targets
DNA	0	2	2	0	2	0	Show Targets
RNA	0	2	12	2	14	0	Show Targets
DNA	1	1	4	0	4	0	Show Targets
RNA	1	1	9	4	13	0	Show Targets
DNA	1	2	17	0	17	0	Show Targets

Off_target_motif	Reference_sequence	Chromosome	Position	Samples Summary	Annotation Type
filter data...					
TAcCAGTTG-TTTTATCACTGG	n	chr2	161236052	n	exon
TAACAGTTGCTg-TATCACTGG	CTgACAGTTGCTg-TATCACTGG	chr13	106090756	1 AMR	distal_enhancer
T-ACAaTTGCTTTTATCACTGT	n	chr7	113275493	n	n
TAACt-TTGCTTTTATCACATG	CTAtCt-TTGCTTTTATCACATG	chr11	121143516	1 AFR	intron
TAACAGT-GgTTTTATCACCTG	n	chr14	47947761	n	n
TAACAGTT-CTgTTATCACATG	CTAACAGTT-CTgcTATCACATG	chr6	78621217	1 MEA	n
TAACAGTaGCTTTTATCA-AGA	CTAACAGTaGtTTTTATCA-AGA	chr3	82699884	148 AFR, 26 CSA, 113 EUR, 44 MEA, 69 EAS, 136 AMR, 12 OCE, 44 SAS	n
TAA-AGTTGCaTTTATCACTTC	n	chr13	92995527	n	n
TAA-AGaTGCTTTTATCACACA	n	chr4	110635888	n	intron
TAACAG-TGCTTaTATCACATC	n	chr9	1916329	n	n

Summary by Sample

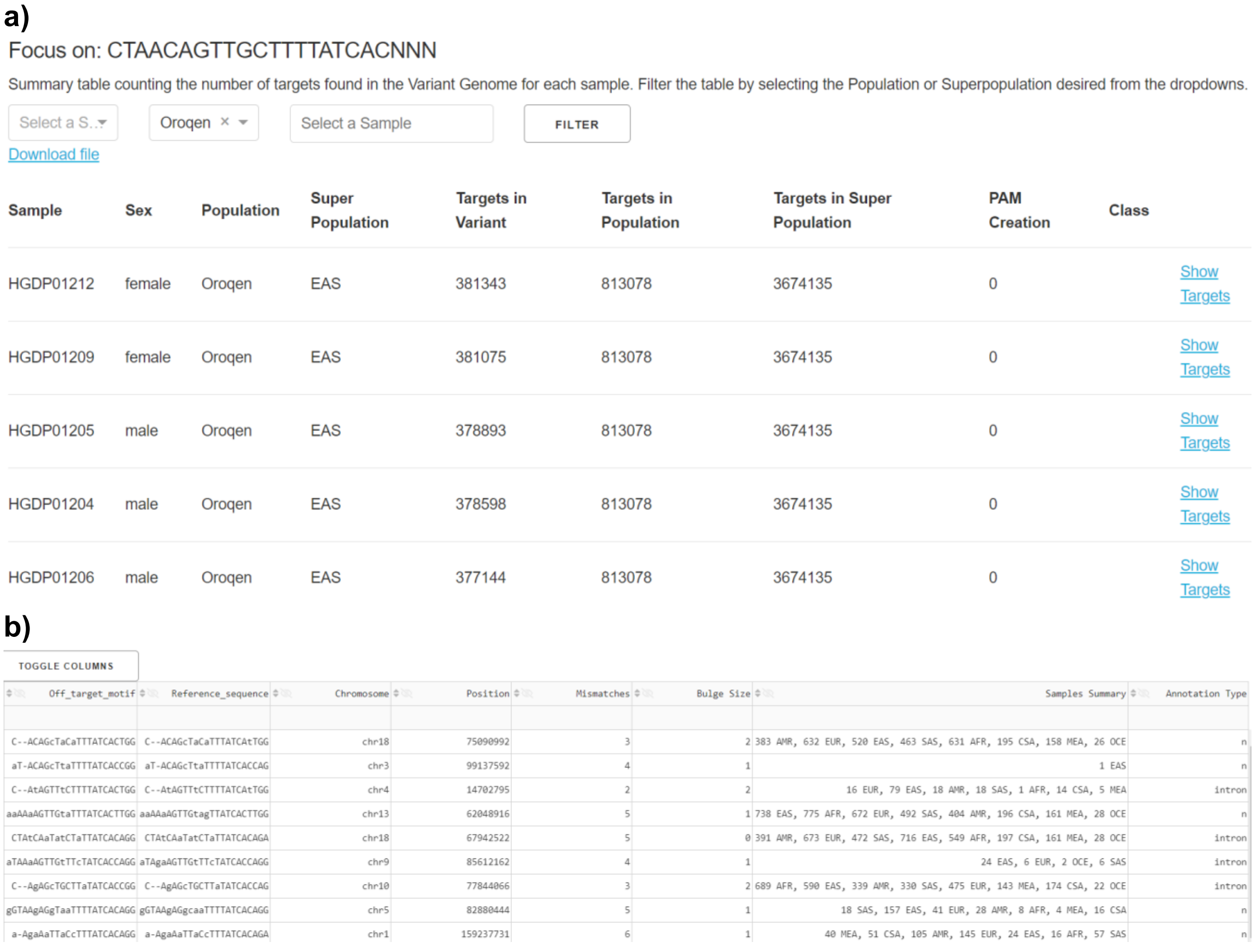
This page shows all the samples present in the VCFs and allows users to extract and visualize targets related to each sample (**Supplementary Fig. 5**).

Supplementary Figure 5. CRISPRme results by sample.

a) Samples, alongside their sex, population and super-population information, are shown in a tabulated list with the count of variant targets for the sample, its population and super-population, as well as the number of PAM creation events for the sample.

b) View of the Show Targets for HGDP01211 sample. The user can select which column see using the Toggle Columns button on top of the table. In this view we show, Off-target motif, Reference sequence, Chromosome, Position, Mismatches, Bulge Size, Samples summary and Annotation type.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



Focus on: CTAACAGTTGCTTTTATCACNNN

Summary table counting the number of targets found in the Variant Genome for each sample. Filter the table by selecting the Population or Superpopulation desired from the dropdowns.

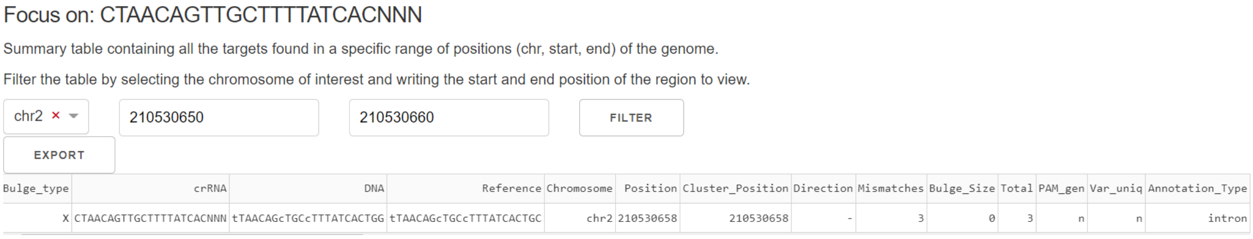
Sample	Sex	Population	Super Population	Targets in Variant	Targets in Population	Targets in Super Population	PAM Creation	Class
HGDP01212	female	Oroqen	EAS	381343	813078	3674135	0	Show Targets
HGDP01209	female	Oroqen	EAS	381075	813078	3674135	0	Show Targets
HGDP01205	male	Oroqen	EAS	378893	813078	3674135	0	Show Targets
HGDP01204	male	Oroqen	EAS	378598	813078	3674135	0	Show Targets
HGDP01206	male	Oroqen	EAS	377144	813078	3674135	0	Show Targets

Off_target_motif	Reference_sequence	Chromosome	Position	Mismatches	Bulge Size	Samples Summary	Annotation Type
C--ACAGcTaCaTTTATCACTGG	C--ACAGcTaCaTTTATCAtTGG	chr18	75090992	3	2	383 AMR, 632 EUR, 520 EAS, 463 SAS, 631 AFR, 195 CSA, 158 MEA, 26 OCE	n
aT-ACAGcTtaTTTTATCACCGG	aT-ACAGcTtaTTTTATCACCAG	chr3	99137592	4	1	1 EAS	n
C--AtAGTTtCTTTTATCACTGG	C--AtAGTTtCTTTTATCAtTGG	chr4	14702795	2	2	16 EUR, 79 EAS, 18 AMR, 18 SAS, 1 AFR, 14 CSA, 5 MEA	intron
aaAAaAGTTGtaTTTATCACTTGG	aaAAaAGTTGtagTTATCACTTGG	chr13	62048916	5	1	738 EAS, 775 AFR, 672 EUR, 492 SAS, 404 AMR, 196 CSA, 161 MEA, 28 OCE	n
CTAtCAaTatCTaTTATCACAGG	CTAtCAaTatCTaTTATCACAGA	chr18	67942522	5	0	391 AMR, 673 EUR, 472 SAS, 716 EAS, 549 AFR, 197 CSA, 161 MEA, 28 OCE	intron
aTAAaAGTTGtTTcTATCACCAGG	aTAgaAGTTGtTTcTATCACCAGG	chr9	85612162	4	1	24 EAS, 6 EUR, 2 OCE, 6 SAS	intron
C--AgAGcTGCTTaTATCACCGG	C--AgAGcTGCTTaTATCACCAG	chr10	77844066	3	2	689 AFR, 590 EAS, 339 AMR, 330 SAS, 475 EUR, 143 MEA, 174 CSA, 22 OCE	intron
gGTAAgAGgTaaTTTTATCACAGG	gGTAAgAGgcaaTTTTATCACAGG	chr5	82880444	5	1	18 SAS, 157 EAS, 41 EUR, 28 AMR, 8 AFR, 4 MEA, 16 CSA	n
a-AgaAaTTaCcTTTATCACAGG	a-AgaAaTTaCcTTTATCACAGA	chr1	159237731	6	1	40 MEA, 51 CSA, 105 AMR, 145 EUR, 24 EAS, 16 AFR, 57 SAS	n

Query Genomic Region

This page allows the user to retrieve targets overlapping a specific genomic region, for example to quickly assess potential off-targets in a given regulatory element or coding region (**Supplementary Fig. 6**).

Supplementary Figure 6. CRISPRme results by genomic region. A table showing the candidate off-target(s) within the region with its motif, position and the position of the first nucleotide of the DNA sequence on the genome without bulges (cluster position), with direction, mismatch count, bulge size, total sum of mismatch and bulges, PAM creation event if present, uniqueness of target in the variant genome, and genomic annotation. Many other columns are reported but not visible here, such as CFD score, AF (allele frequency), rsID (identifier of the variant) and samples sharing the target (if any).



Focus on: CTAACAGTTGCTTTTATCACNNN

Summary table containing all the targets found in a specific range of positions (chr, start, end) of the genome.

Filter the table by selecting the chromosome of interest and writing the start and end position of the region to view.

210530650

210530660

Bulge_type	crRNA	DNA	Reference	Chromosome	Position	Cluster_Position	Direction	Mismatches	Bulge_Size	Total	PAM_gen	Var_uniq	Annotation_Type
X	CTAACAGTTGCTTTTATCACNNN	tTAACAGcTGCcTTTATCACTGG	tTAACAGcTGCcTTTATCACTGC	chr2	210530658	210530658	-	3	0	3	n	n	intron

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Graphical Reports

This page creates several graphical reports for each selected sgRNA.

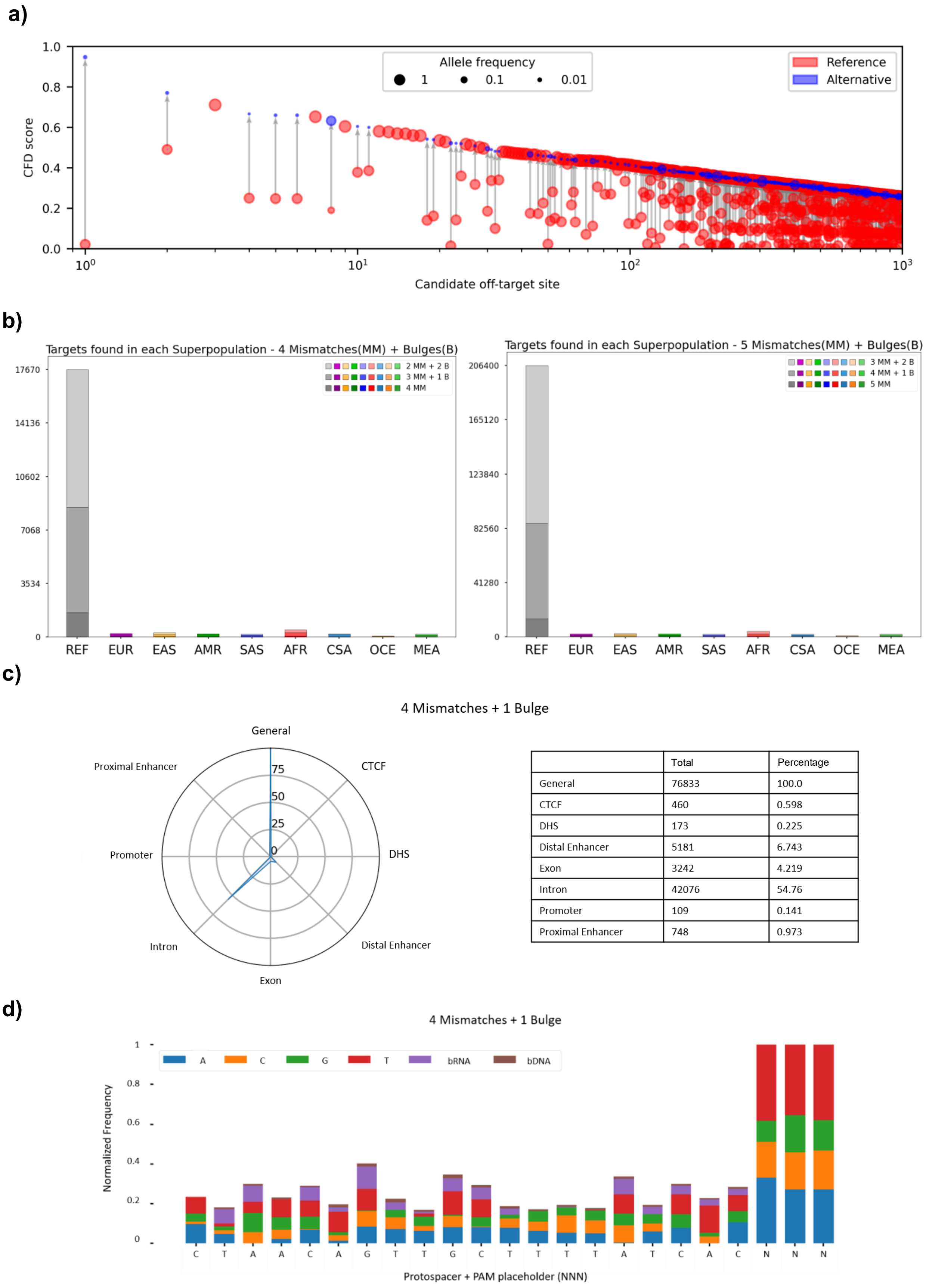
- A stem plot (**Supplementary Fig. 7a**) shows how genetic variants affect predicted off-target potential. The arrow connecting the red (reference allele off-target) and blue (alternative allele off-target) dots shows the increase in predicted cleavage potential due to the variant.
- Bar plots depict how candidate off-targets are distributed across super-populations based on the number of mismatches and bulges (**Supplementary Fig. 7b**).

- A radar chart with the relative specificity of the analyzed guide for each functional region based on annotations from GENCODE and ENCODE. A larger area in the chart represents a gRNA with more potential off-targets falling in annotated regions, possibly representing an undesirable outcome. A summary table provides the count and percentage of off-targets with a given annotation (**Supplementary Fig. 7c**).

- A motif logo summarizing the frequency of mismatches and bulges (b) for each base of the protospacer + PAM (**Supplementary Fig. 7d**).

Supplementary Figure 7. Graphical reports. a) CRISPRme reference/alternative CFD comparison. A stem plot shows the distribution of CFD scores for candidate off-targets ranked in descending order by CFD score. For candidate off-targets for which a genetic variant increases the CFD score, the CFD scores of both the alternative (blue) and reference (red) allele off-targets at the same locus is shown. The area of the circle is proportional to allele frequency. **b)** Stacked bar plots summarizing the number of candidate off-targets for each category of mismatch + bulge in super-populations present in the input variant data. (For example, if mismatch + bulge = 4, the 3 categories are: 4 MM and 0 B, 3 MM and 1 B, and 2 MM and 2 B.) **c)** The proportion of candidate off-targets with each genomic annotation with respect to all those identified. **d)** A motif logo plot showing the distribution of mismatches and bulges across all the enumerated sgRNA putative off-targets.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



	Total	Percentage
General	76833	100.0
CTCF	460	0.598
DHS	173	0.225
Distal Enhancer	5181	6.743
Exon	3242	4.219
Intron	42076	54.76
Promoter	109	0.141
Proximal Enhancer	748	0.973

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Supplementary Note 6 - CRISPRme Personal Risk Cards.

CRISPRme provides a dedicated page to generate reports called *Personal Risk Cards* that summarize potential off-target editing by a particular gRNA in a given individual due to genetic variants. This feature is particularly useful for retrieving and investigating private off-targets.

The report contains two dynamically generated plots depicting the candidate variant off-targets for the sample (**Supplementary Fig. 8a**) and those that are unique to it (**Supplementary Fig. 8b**). These plots highlight how the introduction of genetic variants can change the predicted off-target cleavage potential , thereby demonstrating the importance of variant-aware off-target assessment as in CRISPRme. The report also contains two tables (**Supplementary Fig. 8c**), consisting of a summary (Table 1) and information on each extracted candidate off-target (Table 2) with the following columns:

Table 1:

- `Personal`, count of all the candidate variant off-targets for the selected sample (including variants unique and non-unique to the individual)
- `PAM creation`, count of all the instances where a genetic variant in the sample introduces a new PAM.
- `Private`, count of all the candidate variant off-targets uniquely found in the selected sample.

Table 2:

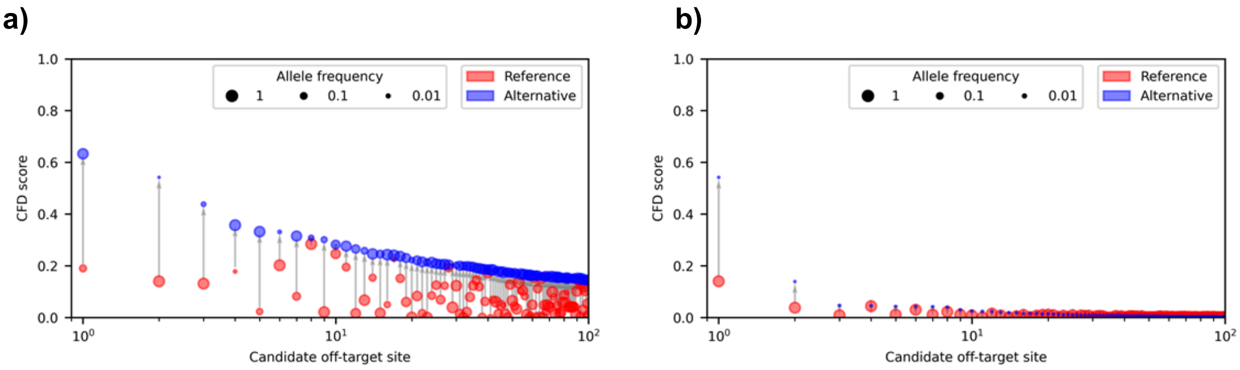
- `Bulge type`, type of target, X, DNA, RNA
- `crRNA`, the sequence of the crRNA
- `DNA`, the sequence of target DNA
- `Reference`, the sequence of the reference target DNA
- `Chromosome`, the chromosome contig containing the target
- `Position`, the relative position (related to bulges) in the reference genome
- `Cluster Position`, the absolute position (related to PAM) in the reference genome
- `Direction`, the target strand
- `Mismatches`, mismatches count of the target
- `Bulge_size`, dimension of bulges in the target or crRNA if DNA
- `Total`, the sum of mismatches and bulge_size

Personal card table file (as shown in **Supplementary Figure 8c**) is downloadable as a file (see **Supplementary File S3**).

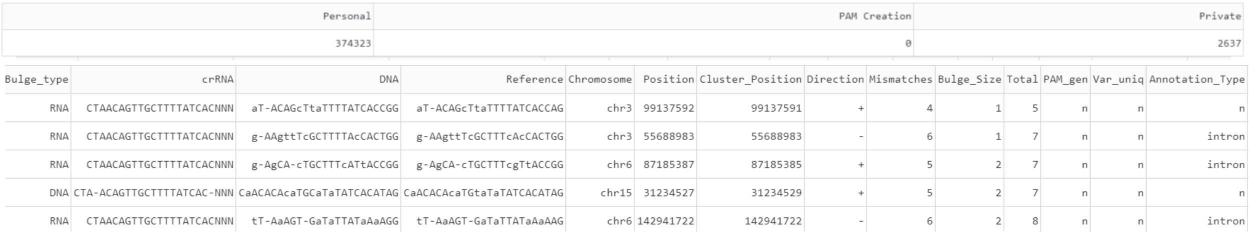
Supplementary Figure 8. CRISPRme personal risk card. Example shown is for HGDP01211. **a)** Plot of potential variant off-targets for the selected sample. **b)** Plot of potential off-targets unique to the selected sample. **c)** The top table reports the number of personal variant off-targets, instances of PAM creation and private off-targets. The bottom table lists for each target the crRNA and DNA sequences, its position and cluster position, its chromosome, strand direction,

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

mismatches, bulge size and total. It also reports the CFD score for the reference and variant sequence, the annotation and possible PAM generation due to substitution or insertion/deletion (not shown in the figure).



c)



Personal	PAM Creation	Private
374323	0	2637

Bulge_type	crRNA	DNA	Reference	Chromosome	Position	Cluster_Position	Direction	Mismatches	Bulge_Size	Total	PAM_gen	Var_uniq	Annotation_Type
RNA	CTAACAGTTGCTTTTATCACNNN	aT-ACAGcTtaTTTTATCACCGG	aT-ACAGcTtaTTTTATCACCAG	chr3	99137592	99137591	+	4	1	5	n	n	n
RNA	CTAACAGTTGCTTTTATCACNNN	g-AAgttTcGCTTTTAcCACTGG	g-AAgttTcGCTTTcAcCACTGG	chr3	55688983	55688983	-	6	1	7	n	n	intron
RNA	CTAACAGTTGCTTTTATCACNNN	g-AgCA-cTGCTTTcATtACCGG	g-AgCA-cTGCTTTcgTtACCGG	chr6	87185387	87185385	+	5	2	7	n	n	intron
DNA	CTA-ACAGTTGCTTTTATCAC-NNN	CaACACAcaTGCaTaTATCACATAG	CaACACAcaTGtaTaTATCACATAG	chr15	31234527	31234529	+	5	2	7	n	n	n
RNA	CTAACAGTTGCTTTTATCACNNN	tT-AaAGT-GaTaTTATaAaAGG	tT-AaAGT-GaTaTTATaAaAAG	chr6	142941722	142941722	-	6	2	8	n	n	intron

Supplementary Note 7 - Details of CRISPRme installation and usage.

Installation

CRISPRme can be used offline and is available as a comprehensive Conda package containing the

1. A command line version of the software and
2. A web based tool with a graphical interface accessible from a local browser.

To install the package:

- Follow this link and install Conda: https://docs.conda.io/en/latest/miniconda.html
- After the installation is complete, add channels to the Conda local installation by typing the following commands into the terminal:
    - `conda config --add channels defaults`
    - `conda config --add channels bioconda`
    - `conda config --add channels conda-forge`
    - `conda install python=3.8`
- Type into the terminal `conda install crisprme -y`

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

To allow use on any platform, CRISPRme is also available as a Docker image:

- Download the latest available CRISPRme image using the command `docker pull pinellolab/crisprme`

- After the pull is complete, a new Docker container can be created any time starting from the clean image

A complete installation guide is available at https://github.com/pinellolab/CRISPRme

Usage

CRISPRme offline can be run as a web-app or a command line tool. The required input files are identical for the two versions (web-app and command line) and are described in **Supplementary Note 3**.

Web-app

1. Download our test package,
   ```
   conda install gdown -y
   gdown https://drive.google.com/uc?id=11wn9pg6AWzDYZ7V_LjBIjGvgx95bnVJ1
   ```
2. Execute the command `crisprme.py web-interface`
3. Open a web browser and visit `127.0.0.1:8080`. The homepage of CRISPRme will open
4. If a remote server is used to host CRISPRme, input the IP address of the server in the web browser, e.g. `192.1.2.3:8080` and the home page will open in the browser.
5. Now you can directly input data and select how you want to perform the search, like genome release, PAM sequence and annotation for genomic regions.

Command line

1. Download our test package,
   ```
   conda install gdown -y
   gdown https://drive.google.com/uc?id=11wn9pg6AWzDYZ7V_LjBIjGvgx95bnVJ1
   ```
2. Write the following command into the terminal `crisprme.py complete-search --help`
3. The above command will show all the mandatory and optional inputs for a CRISPRme complete-search. Please see **Supplementary Note 3** for the expected formats of these inputs.
   a. Mandatory input:
      i. `--genome`, the reference genome folder
      ii. `--guide`, the file that contains the guides used for the search
      iii. `--pam`, the file containing the PAM
      iv. `--annotation`, the file containing annotations of the reference genome
      v. `--bMax`, the maximum number of bulges

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

vi. `--mm`, the maximum number of mismatches
vii. `--output`, the output folder for the results

b. Optional input:

i. `--vcf`, the file with the list of VCF folders to be used
ii. `--samplesID`, the file with the list of sample ID files (must have the same number of lines as the file passed to `--vcf`)
iii. `--gene_annotation`, a GENCODE or custom annotation to find the nearest gene for each off-target found
iv. `--bDNA`, the number of DNA bulges permitted in the search phase
v. `--bRNA`, the number of RNA bulges permitted in the search phase
vi. `--merge`, the threshold to merge nearby off-targets (based on genomic position), using the off-target with the highest CFD score as the pivot [default 3]
vii. `--thread`, the number of threads used in the process (default is ALL available minus 2)

CRISPRme functions:

1. Targets integration function, can be used to generate an integrated result file (see **Supplementary Note 2**) with user defined empirical data.
   Command:
   `crisprme.py targets-integration`
   Input：
   a. `--targets`, bestMerge or altMerge file
   b. `--genome_version`, the genome release used in the search
   c. `--guide`, the file that contains the guides used for the search
   d. `--gencode`, a GENCODE or custom annotation to find the nearest gene for each off-target found
   e. `--output`, the output folder for the results
2. gnomAD converter function, can be used to convert gnomADv3.1 VCFs to CRISPRme compatible VCFs (see **Supplementary Note 3**).
   Command:
   `crisprme.py gnomAD-converter`
   Input:
   a. `--gnomAD_VCFdir`, used to specify the directory containing gnomADv3.1 original VCFs
   b. `--samplesID`, used to specify the pre-generated samplesID file necessary to introduce samples into gnomAD variant
   c. `--thread`, the number of threads used in the process (default is ALL available minus 2)

An example of a complete command to start CRISPRme could be found at https://github.com/pinellolab/CRISPRme

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Input files

CRISPRme input files and directories, each file is described in more details in Supplementary note 3:

--genome, directory containing the genome in fasta format and chromosome separated (e.g. https://hgdownload.soe.ucsc.edu/goldenPath/hg38/bigZips/hg38.chromFa.tar.gz).
--vcf, text file with a list of folders containing VCF files (one folder per line)
--guide, text file containing the protospacer(s) formatted in a single line with prefix or suffix Ns as placeholder for PAM sequence (e.g. for a spCas9 compatible protospacer, CTAACAGTTGCTTTTATCACNNN)
--pam, text file containing a single PAM sequence as explained in Supplementary Section 3
--annotation, BED file containing a set of annotations in BED format as explained in Supplementary Section 3
--samplesID, text file with a list of files containing information about samples (one file per line as in --vcf)
--bMax, set the number of maximal bulges to search with the created index (used to create a searchable index)
--mm, set the number of allowed mismatches for the search
--bDNA, set the number of allowed DNA bulge size
--bRNA, set the number of allowed RNA bulge size
--output, directory to collect all the output files

An example of content for a file given as input for --vcf is the following:

```
hg38_HGDP/
hg38_1000G/
```

Representing the directories containing VCF files saved into the main VCFs folder.

An example of content for a file given as input for --samplesID is the following:

```
hg38_1000G.samplesID.txt
hg38_HGDP.samplesID.txt
```

Representing the files saved into the main samplesIDs/ folder.

Output files

CRISPRme outputs various results files:

- `jobID.bestMerge.txt.integrated_results.tsv` contains candidate off-targets sorted by CFD score with GENCODE annotations and associated empirical data (if in input). This file contains the most comprehensive information on the potential off-targets identified by CRISPRme. Summary files for each guide, containing tabulated information about samples associated targets, counts for result summary table, population distribution data and cumulative CFD score.
    - o `jobID.general_target_count.GUIDE.txt` contains a table showing count of targets grouped by number of mismatches and bulges

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- `jobID.summary_by_guide.GUIDE.txt` contains count of targets divided in reference and variant for every combination of MM+BULGES
- `jobID.summary_by_samples.GUIDE.txt` contains a table that for each sample reports the number of personal and private targets, counting also the number of PAM creation events.

- `jobID.bestMerge.txt` contains the candidate off-targets, with only the off-target having the highest CFD score within each merge window included to reduce the number of nearly redundant off-targets in the results (see **Supplementary Note 2**).
- `jobID.altMerge.txt` contains the other off-targets (with alternative alignments and/or alleles) found in the same genomic regions as those included in the best files.
- `jobID.PopulationDistribution.txt` contains the counts necessary used to produce population bar plots, reporting how many targets belong to each superpopulation and in which subcategory (mismatches + bulges).
- `jobID.acfd.txt` contains the summarized CFD score calculated for each guide used in the search.
- `jobID.bestMerge.txt.empirical_not_found.tsv` contains the reported empirical targets (if in input) that are were not associated to an in -silico predicted target.
- `params file`, contains the parameters of the search.
- `email file`, contains message to send to input email when job is finished.
- `log file`, contains the report of started and terminanted steps for each step in the analysis process.
- `log_verbose`, contains the verbose log of each step executed during the analysis.
- `log_error`, contains the error report of any possible error reported during the analysis.
- `pam file`, contains the PAM sequence used in the search.
- `guides file`, contains the list of guide(s) used in the search.
- `list_vcfs file`, contains the list of directory containing the vcf files.
- `samplesID file`, contains the list of files with original sampleIDs used in the search.
- `sampleID file`, contains all the samplesID from original sampleID files for each dataset used in the search.
- `jobID.guide_dict_GUIDE file`, contains the dictionary to generate in real time radar chart graphical reports ((see **Supplementary Fig. 7c**).
- `jobID.motif_dict_GUIDE file`, contains the dictionary to generate in real time motif dict graphical reports ((see **Supplementary Fig. 7d**).
- `crispritz_targets` directory, contains all the raw target files generated by CRISPRme:
    - `refgenome_pam_guides.targets`, Reference targets file, contains all the raw targets generated from the search.
    - `refgenome+vargenome_pam_guides.targets`, Variant targets file(s), one for each VCF dataset used in the search, contains all the raw targets generated from the search.
    - `indels_refgenome+vargenome_pam_guides.targets`, INDELs targets file(s), one for each VCF dataset used in the search, contains all the raw targets generated from the search.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

- `imgs` directory, contains all the images produced with the result data, population distribution barplots and summary charts for specific guide analysis
    - `CRISPRme_top_1000_log_for_main_text_GUIDE file`, figure reported in graphical summary (see **Supplementary Fig. 7a**) with reference and alternative targets with compared CFD score.
    - `CRISPRme_top_100_linear_annotated_GUIDE file`, figure non reported in graphical report, generated to investigate functional annotations of targets.
    - `populations_distribution_GUIDE_#total file`, figure reported in graphical summary (see **Supplementary Fig. 7b**) with population distribution for each combination of mismatches + bulges.
    - `Summary_single_guide_GUIDE_#mm_#bul_ORIGIN file`, figure reported in graphical summary (see **Supplementary Fig. 7c,d**) containing radar chart, table and motif logo of examined guide for specific combination of mismatches and bulges (can be generated for any sample, population and superpopulation).
- `jobID.bd`, SQL database of bestCFD tabulated file, fast database created to perform real time query on the result file (see **Supplementary Fig. 5b**).
- Different temporary file or run-time generated files, many different text files, one for each search or filtering done in the web application. These files will be stored to avoid redoing computations after the first computation. Each file will be zipped to be also downloadable through the web interface.

Downloadable files

- Summary by gRNA targets, file containing targets found in a specific combination of mismatches and bulges size.
- Summary by sample targets, file containing targets found in a specific sample, these targets can be unique to the sample or shared with other samples.
- Summary by sample table, file containing a table where rows are the samples contained in VCF data and columns are the count of targets in superpopulation, population and sample
- Summary by position targets, file containing targets found in a specific genetic region.
- Graphical summary, page to access all the summary figures and charts collecting information about targets results.
- Personal card, file containing all the ‘private’ targets related to a sample, private means targets only found in the specific sample.

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Methods

Cell culture

Fresh G-CSF mobilized peripheral blood cells from healthy donor 1 were obtained from Miltenyi Biotec (Auburn, CA). CD34+ HSPCs were isolated using CliniMACS® CD34 reagent (Miltenyi, 130-017-501). Cryopreserved human CD34+ HSPCs from mobilized peripheral blood of deidentified healthy donors 2-7 were obtained from the Fred Hutchinson Cancer Research Center (Seattle, Washington). CD34+ HSPCs were cultured into Stem Cell Growth Medium (SCGM) (CellGenix, 20806-0500) supplemented with 100 ng ml-1 human Stem Cell Growth Factor (SCF) (CellGenix, 1418-050), 100 ng ml-1 human thrombopoietin (TPO) (CellGenix, 1417-050) and 100 ng ml-1 recombinant human FMS-like Tyrosine Kinase 3 Ligand (Flt3-L) (CellGenix cat# 1415-050). HSPCs were electroporated with 3xNLS-SpCas9:sg1617 RNP or HiFi-3xNLS-SpCas9:sg1617 RNP 24 h after thawing. Twenty-four hours after electroporation, HSPCs were transferred into erythroid differentiation medium (EDM) consisting of IMDM (LIFE, 12440061) supplemented with 330 μg ml- holo-human transferrin (Sigma, T0665-1G), 10 μg ml- recombinant human insulin (Sigma, 19278-5ML), 2 IU ml- heparin (Sigma, H3149), 5% human solvent detergent pooled plasma AB (Rhode Lisland Blood Center), and 3 IU ml- erythropoietin (Pharmacy). Five days after electroporation, cells were harvested for gDNA extraction.

Protein purification

Hi-Fi-3xNLS-SpCas9 plasmids were transformed into BL21 (DE3) competent cells (MilliporeSigma, 702353) and grown in Terrific Broth (TB) media at 37°C until OD600 2.4-2.8. Cells were induced with 0.5 mM isopropyl ß-d-1-thiogalactopyranoside (IPTG) per liter for 20 hours at 20°C. Lysed pellet in 25 mM Tris, pH 7.6, 500 mM NaCl, 5% glycerol and passed through homogenizer twice and centrifuged at 20,000 rpm for 1 hour at 4°C. Proteins were purified by Nickel-NTA resin and treated with TEV protease (1 mg lab made TEV per 40 mg of protein) and benzonase (100 units/ml, Novagen 70664-3) overnight at 4°C. Subsequently, purified by size exclusion column (Amersham Biosciences HiLoad 26/60 Superdex 200 17-1071-01) and ion exchange with a 5 ml SP HP column (GE 17-1151-01) according to the manufacturer's instructions. Proteins were dialyzed in 20 mM Hepes buffer pH 7.5 containing 400 mM KCl, 10% glycerol, and 1 mM TCEP buffer, and contaminants were removed by Toxin Sensor Chromogenic LAL Endotoxin Assay Kit (GenScript, L00350). Purified proteins were concentrated and filtered using Amicon ultra filter units – 30k NMWL (MilliporeSigma, UFC903008) and ultrafree-MC centrifugal filter (MilliporeSigma, UFC30GV0S). Protein fractions were further assessed on TGX stain free 4-20% SDS-PAGE (Biorad, 5678093) and quantified by BCA assay.

RNP electroporation

Electroporation was performed using Lonza 4D Nucleofector (V4XP-3032 for 20 μl as the manufacturer's instructions). CD34+ HSPCs were thawed 24 h before electroporation. For 20 μl Nucleocuvette Strips, the RNP complex was prepared by mixing 3xNLS-SpCas9 protein (100 pmol, purified as previously described[18]) or HiFi-3xNLS-SpCas9 protein (100 pmol) and sgRNA (300 pmol, IDT) with glycerol (2% of final concentration, Sigma, G2025) and incubating for 15 min at room temperature immediately before electroporation. 50K HSPCs resuspended in 20 μl

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

P3 solution were mixed with RNP and transferred to a cuvette for electroporation with program EO-100. The electroporated cells were resuspended with SCGM medium with cytokines and changed into EDM 24 h after electroporation.

Measurement of +58 BCL11A enhancer on-target and OT40 off-target indel and inversion

Editing frequencies were measured with cells cultured in EDM 5 days after electroporation. Briefly, genomic DNA was extracted using the Qiagen DNeasy Blood and Tissue kit (Qiagen, 69506). The *BCL11A* enhancer DHS +58 on-target site was amplified using forward primer AGAGAGCCTTCCGAAAGAGG and reverse primer GCCAGAAAAGAGATATGGCATC. The off-target-rs114518452 site was amplified using forward primer TAAGATTCTTTTGGTTCTGGCT and reverse primer AGAGAGGCAGTATTTACGATGC. The inversion junction was amplified using +58 forward primer AGAGAGCCTTCCGAAAGAGG (F1) and off-target-rs114518452 forward primer TAAGATTCTTTTGGTTCTGGCT (F2), or +58 reverse primer GCCAGAAAAGAGATATGGCATC (R1) and off-target-rs114518452 reverse primer AGAGAGGCAGTATTTACGATGC (R2). KOD Hot Start DNA Polymerase (EMD-Millipore, 71086-31) was used for PCR and followed cycling conditions: 95 degrees for 3 min; 30 cycles of 95 degrees for 20 s, 60 degrees for 10 s, and 70 degrees for 10 s; 70 degrees for 5 min. 1 µl of locus specific PCR product was used for indexing PCR with KOD Hot Start DNA Polymerase and index primers following cycling conditions: 95 degrees for 3 min; 10 cycles of 95 degrees for 20 s, 60 degrees for 10 s, and 70 degrees for 10 s; 70 degrees for 5 min. Resulting PCR products were subjected to deep sequencing.

Amplicon deep sequencing and analysis

Amplicons were sequenced using paired-end 150 bp reads on an Illumina MiniSeq system with >18,000X coverage per sample for the off-target-rs114518452 site and >3,800X coverage per sample for the on-target site. Reads were trimmed for adapters and quality using Trimmomatic[27] in paired-end mode for the off-target-rs114518452 site and in single-end mode for the on-target site due to a nearby difficult-to-sequence homopolymer region. Editing outcomes were analyzed using CRISPResso v2.1.0[28] by aligning to the expected reference and/or alternative allele amplicons. A Needleman-Wunsch gap opening penalty of -30 (CRISPResso2 default: -20) was used to ensure more accurate alignment of reads to the reference vs. alternative allele amplicons for off-target-rs114518452 since they only differ by a single nucleotide. Only indels overlapping the expected SpCas9 cleavage site (3 bp upstream of the PAM) were counted as gene edits. The median observed indel frequency is reported for samples for which technical replicates were performed (n = 4), which includes all amplicon sequencing at the off-target-rs114518452 site for the donor heterozygous for rs114518452. Representative reads collapsed by allele identity and indel type are presented in the plots.

Inversion PCR

Nested PCR was performed to amplify the inversion junction. First step PCR was amplified using the outer primers on-target +58 forward, CACACGGCATGGCATACAAA, and off-target-rs114518452 forward, AATAGCCAAACTACTGAGCATTGTG; or the outer primers of on-target +58 reverse, CACCCTGGAAAACAGCCTGA, and off-target-rs114518452 reverse, ACTAAGGCAATTGTTGTCCAAGC. KOD Hot Start DNA Polymerase was used for PCR and

bioRxiv preprint doi: https://doi.org/10.1101/2021.05.20.445054; this version posted May 21, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

followed cycling conditions: 95 degrees for 3 min; 30 cycles of 95 degrees for 20 s, 60 degrees for 10 s, and 70 degrees for 10 s; 70 degrees for 5 min. 1 µl of PCR1 product was used for the second step PCR amplifying with inner primers on-target +58 forward (F1) and off-target-rs114518452 forward (F2), or on-target +58 reverse (R1) and off-target-rs114518452 reverse (R2) with cycling conditions: 95 degrees for 3 min; 10 cycles of 95 degrees for 20 s, 60 degrees for 10 s, and 70 degrees for 10 s; 70 degrees for 5 min. Resulting PCR products were loaded on a 2% agarose (VWR, 97062-250) gel. Images were captured by the BioRad ChemiDoc^{TM} MP Imaging System.