bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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 comprehensive Bioconductor ecosystem for the design of CRISPR guide RNAs across nucleases and technologies

Luke Hoberecht[1], Pirunthan Perampalam[2], Aaron Lun[1], and Jean-Philippe Fortin[1,*]

[1]Genentech Research and Early Development, Genentech, Inc., 1 DNA Way, South San Francisco, CA, 94080, USA
[2]ProCogia Inc. under contract to Hoffmann-La Roche Limited

Abstract

The success of CRISPR-mediated gene perturbation studies is highly dependent on the quality of gRNAs, and several tools have been developed to enable optimal gRNA design. However, these tools are not all adaptable to the latest CRISPR modalities or nucleases, nor do they offer comprehensive annotation methods or scalability for advanced CRISPR applications. Here, we present a new ecosystem of R packages that enables efficient gRNA design and annotation for a multitude of CRISPR technologies, including CRISPR knockout (CRISPRko), CRISPR activation (CRISPRa), CRISPR interference (CRISPRi) and CRISPR base editing (CRISPRbe). The core package, *crisprDesign*, offers a comprehensive, user-friendly, and unified interface to add on- and off-target annotations via several alignment methods, rich gene and SNP annotations, and a dozen on- and off-target activity scores. These functionalities are enabled for any RNA- or DNA-targeting nucleases, including Cas9, Cas12, and Cas13. We illustrate the general applicability of our tools by designing optimal gRNAs for three case studies: tiling CRISPRbe library for *BRCA1* using the base editor BE4max, tiling RNA-targeting libraries for *CD46* and *CD55* using CasRx, and activation of *MMP7* using CRISPRa. Our suite of R packages is open-source and deployed through the Bioconductor project to facilitate their use by the CRISPR community.

Keywords: CRISPR, Nuclease, gRNA, functional genomics, Sequence design, Bioconductor, Base Editor, CasRx.

1 Main

The performance of CRISPR-based experiments depends critically on the choice of the guide RNAs (gRNAs) used to guide the CRISPR nuclease to the target site. Variable gRNA on-target activity, as well as unintended off-targeting effects, can lead to inconsistent phenotypic readouts in screening experiments. For the purpose of analyzing pooled screens, many approaches attempt to model gRNA quality in the generation of gene-level scores to improve statistical inference [Meyers et al., 2017, Kim and Hart, 2021, Dempster et al., 2021, Allen et al., 2019, Li et al., 2015]. However, suboptimal gRNA design is only partially mitigated by analysis strategies that sacrifice statistical power for robustness to suboptimal guides. One way to increase the signal-to-noise ratio in screening experiments is to enrich gRNA libraries for gRNAs that have high predicted on-target activity. Predicting on-target activity from the spacer sequence is an extensive area of research, and several algorithms leveraging experimental data have been developed for different nucleases and contexts [Doench et al., 2016b, 2014, 2016a, Wang et al., 2019b, Kim et al., 2018, Hart et al., 2017, Moreno-Mateos et al., 2015, Horlbeck et al., 2016].

In addition to its sequence, the genomic context of the on- and off-target sites for each gRNA is another important consideration for gRNA design. For example, designing gRNAs that uniquely map to the genome can be challenging, especially for genes sharing high homology with other genomic loci, either in coding or non-coding regions [Fortin et al., 2019]. In addition, knowing whether or not an off-target is located in the coding region of another gene can rule out the use of a given gRNA. Finally, genetic variation, such as single-nucleotide

*To whom correspondence should be addressed. Email: `fortin946@gmail.com`

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

polymorphisms (SNPs) and small indels, can have a direct impact on gRNA binding activity and on-target specificity by altering complementarity between spacer sequences and the host cell genomic DNA [Scott and Zhang, 2017, Lessard et al., 2017, Canver et al., 2017, Wang et al., 2018].

The rapid increase of CRISPR-based applications and technologies poses another challenge to gRNA library design. A large variety of nucleases are now available and routinely used, including engineered nucleases that recognize a larger set of PAM sequences [Hu et al., 2018, Nishimasu et al., 2018, Walton et al., 2020, DeWeirdt et al., 2020] and novel classes of nucleases such as the RNA-targeting Cas13 family [Shmakov et al., 2015, Abudayyeh et al., 2016, Konermann et al., 2018]. Each nuclease comes with its own set of gRNA design rules and constraints. In addition, these nucleases can also be mixed and matched with different types of CRISPR applications, increasing the complexity of gRNA design. As an example, CRISPR base editing (CRISPRbe) [Gaudelli et al., 2017, Komor et al., 2016], which requires additional gRNA design functionalities to capture the editing window and prediction of editing outcomes, can be combined with the Cas13 family to perform RNA editing [Cox et al., 2017]. Finally, emerging screening modalities, such as optical pooled CRISPR screening [Feldman et al., 2019], require additional specialized gRNA design considerations.

Given the complexity, heterogeneity, and fast growth of the aforementioned CRISPR modalities and applications, it is paramount to develop and maintain adaptable, modular, and robust software for gRNA design. This ensures that the scientific community can efficiently design first-class CRISPR reagents in a timely manner for both well-established and emerging technologies. An ideal gRNA design framework has the following qualities: (1) it offers multiple cutting-edge methods for on-target scoring and off-target prediction based on gRNA sequences, (2) it provides comprehensive gRNA annotation to enable consideration of the genomic context for all gRNA on-target and off-target sites, (3) it already supports (or can be easily extended to) newer CRISPR technologies, including an arbitrary combination of nucleases and modalities, and (4) it easily scales for designing large-scale gRNA libraries for different screening platforms. While a multitude of web applications and command line interfaces has been developed to enable gene- or other target-specific gRNA design [Heigwer et al., 2014, Moreno-Mateos et al., 2015, Perez et al., 2017, Bae et al., 2014, Montague et al., 2014, Concordet and Haeussler, 2018, Stemmer et al., 2015, McKenna and Shendure, 2018, Heigwer et al., 2016, Bhagwat et al., 2020, Zhu et al., 2014], none of the existing tools completely satisfies the requirements listed above.

In this work, we describe a modular ecosystem of R packages that enable the design of CRISPR gRNAs across a variety of nucleases, genomes, and applications. The *crisprBowtie* and *crisprBwa* packages provide comprehensive on-target and off-target search for reference genomes, transcriptomes, or any custom sequences. The *crisprScore* package provides a harmonized framework to access a large array of R- and Python-based gRNA scoring algorithms developed by the CRISPR community, for both on-target and off-target scoring. The *crisprBase* package implements functionalities to describe and represent DNA- and RNA-targeting CRISPR nucleases and base editors, as well as genomic arithmetic rules that are specific to CRISPR design. Finally, the package *crisprDesign* provides a user-friendly package to design and annotate gRNAs in one place, including gene and TSS annotation, search for SNP overlap, characterization of edited alleles for base editors, sequence-based design rules, and library design functionalities such as ranking and platform-specific considerations. Our ecosystem currently supports five different CRISPR modalities: CRISPR knockout (CRISPRko), CRISPR activation (CRISPRa), CRISPR interference (CRISPRi), CRISPR base editing (CRISPRbe) and CRISPR knockdown (CRISPRkd) using Cas13.

We illustrate the rich functionalities of our ecosystem through three case studies: designing gRNAs to edit *BRCA1* using the base editor BE4max, designing gRNAs to knock down *CD55* and *CD46* using CasRx, and designing optimal gRNAs to activate *MMP7* through CRISPRa for different wildtype and engineered nucleases. Our R packages are open-source and deployed through the Bioconductor project [Gentleman et al., 2004, Huber et al., 2015]. This makes our tools fully interoperable with other packages, and facilitates long-term development and maintenance of our ecosystem.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

2 Results

2.1 *crisprBase* as a core infrastructure package to represent CRISPR nucleases and base editors

The *crisprBase* package implements a common framework for representing and manipulating nucleases and base editors through a set of classes and CRISPR-specific genome arithmetic functions. The *CrisprNuclease* class provides a general representation of a CRISPR nuclease, encoding all of the information necessary to perform gRNA design and other analyses involving CRISPR technologies. This includes the PAM side with respect to the protospacer sequence, recognized PAM sequences with optional tolerance weights, and the relative cut site. Specific *CrisprNuclease* instances can be easily created to represent a diversity of wild-type and engineered CRISPR nucleases (Figure 1). We also implement a *BaseEditor* subclass that provides additional base editing information such as the editing strand and a matrix of editing probabilities for possible nucleotide substitutions.

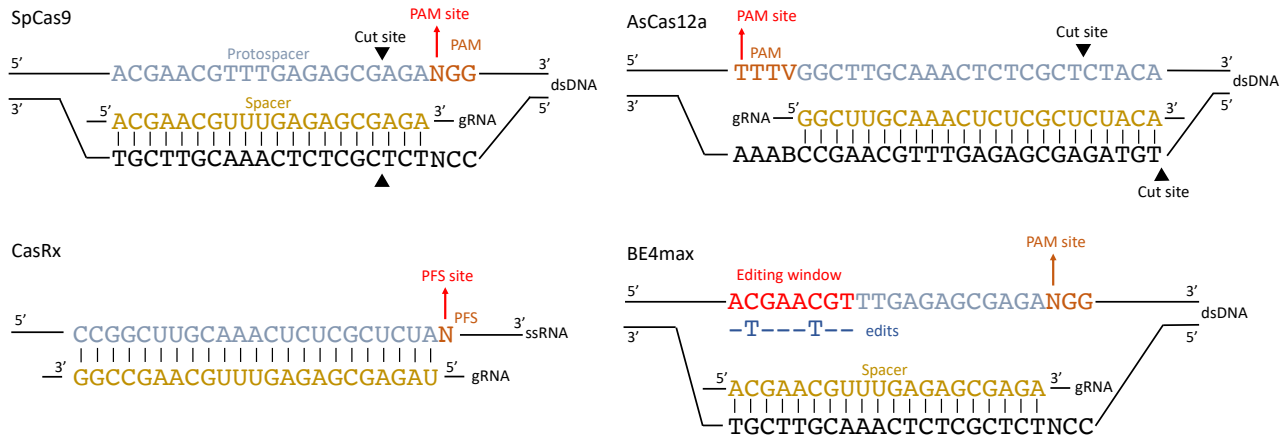


Figure 1. Examples of DNA- and RNA-targeting nucleases represented in *crisprBase*. gRNA spacer sequences are shown in yellow. Target DNA/RNA protospacer sequences are shown in blue. Protospacer adjacent motifs (PAMs) and protospacer flanking sequences (PFSs) are shown in orange. Nuclease-specific cutting sites are represented by black triangles. For the C to T base editor BE4max, on-target editing happens on the DNA strand containing the protospacer sequence. The editing window varies by base editor. The first nucleotide of the PAM/PFS is used as the representative coordinate of a given target sequence.

2.2 *crisprDesign*: a comprehensive tool to perform complex gRNA design

crisprDesign supports a comprehensive suite of methods to design and annotate gRNAs (Table 1). For users, *crisprDesign* provides a centralized and streamlined end-to-end workflow for gRNA design, alleviating the burden of using different tools at different stages of the design process. For developers, *crisprDesign* is built on top of a modular package ecosystem that implements the gRNA design tasks (see Table S1), allowing the same code to be easily re-used outside of CRISPR applications and gRNA design. In the following sections, we describe each of the gRNA design components and functionalities that are available in *crisprDesign*.

2.3 Representation of gRNAs using the *GuideSet* container

The genomic coordinates of gRNA protospacer sequences in a target genome can be represented using genomic ranges. The Bioconductor project [Gentleman et al., 2004, Huber et al., 2015] provides a robust and well-developed core data structure, called *GRanges* [Lawrence et al., 2013, Aboyoun et al., 2012]), to efficiently represent genomic intervals. We provide in *crisprDesign* an extension of the *GRanges* class to represent and annotate gRNA sequences: the *GuideSet* container. Briefly, the container extends the *GRanges* object to store additional project-specific metadata information, such as the CRISPR nuclease employed and target mRNA or DNA sequences (if different from a reference genome), as well as rich gRNA-level annotation columns such

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

		multicrispr	*CRISPRseek*	*crisprDesign*	Function in *crisprDesign*
Nuclease	SpCas9	✓	✓	✓	`crisprBase::SpCas9`
	Any	†	✓	✓	`crisprBase::CrisprNuclease`
On-target search	Reference genome	✓	✓	✓	`findSpacers()`
	Custom sequences		✓	✓	
	Removal of repeats			✓	
Off-target search	Exact string matching	✓	✓	✓	`addSpacerAlignments()`
	Bowtie	††		✓	
	Bwa			✓	
	Iterative alignments			✓	
	Custom sequence		✓	✓	
	Cross-reactivity			✓	
	Minor/major alleles			✓	
On-target scoring	Rule Set 1	✓	✓	✓	`addOnTargetScores()`
	Rule Set 2 / Azimuth	✓	✓	✓	
	CRISPRscan		✓	✓	
	Lindel		✓	✓	
	DeepCpf1		✓	✓	
	DeepHF			✓	
	CRISPRai			✓	
	EnPAM+GB			✓	
	CasRx-RF			✓	
	PAM scoring			✓	
Off-target scoring	MIT		✓	✓	`addOffTargetScores()`
	CFD		✓	✓	
	CasRx			✓	
Library design	Restriction sites		✓	✓	`addRestrictionEnzymes()`
	PolyT signal			✓	`addSequenceFeatures()`
	GC content			✓	
	Hairpin loops			✓	
	SNP annotation			✓	`addSNPAnnotation()`
Gene annotation	Off-target annotation		† † †	✓	`addGeneAnnotation()`
	Isoform specification			✓	
	Reinitiation sites			✓	
	Pfam domains			✓	
	Distance to TSS			✓	
Modalities	CRISPRko	✓	✓	✓	
	CRISPRbe		✓	✓	`addEditedAlleles()`
	CRISPRa			✓	`addTSSAnnotation()`
	CRISPRi			✓	
	RNA editing (Cas13)			✓	
	Optical pooled screening			✓	
	Prime editing	✓	✓	*	

Table 1. gRNA design functionalities for *crisprDesign* and two other Bioconductor packages developed for gRNA design: *multicrispr* and *CRISPRseek*. Check marks indicate which functionalities are present in each package at time of publication. † For *multicrispr*, only CRISPR nucleases with PAM sequences on the 3′ side are accepted. †† For *multicrispr*, the actual off-target alignments are only counted, and not available to the user † † † For *CRISPRseek*, gene annotation of the off-targets is limited to whether or not they overlap exons and introns of known genes. *In progress.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

as on- and off-target alignments tables and gene context annotations. In Figure 2, we show an example of a *GuideSet* storing information about gRNAs targeting the coding sequence of *KRAS* using SpCas9.

2.4 *crisprScore* implements state-of-the-art scoring methods

Predicting on-target binding and cutting efficiency of gRNAs is an extensive area of research. Many algorithms have been developed to tackle this problem, basing their prediction on a variety of features: sequence composition of the spacer sequence and flanking regions, including nucleotide content and melting temperature, cell type-specific chromatin accessibility data, and distance to transcription starting site (TSS). Unfortunately, from a user's perspective, the heterogeneity in the algorithm implementations hinders the practical use of those algorithms: some methods are implemented in Python 2 [Doench et al., 2016a, Kim et al., 2018], in Python 3 [Chen et al., 2019, Wang et al., 2019b, DeWeirdt et al., 2020], or in R [Doench et al., 2014, Wessels et al., 2020, Moreno-Mateos et al., 2015]. In addition, the required inputs, data structures, and terminology are not consistent across software and algorithms, increasing the likelihood of user error. Finally, several of the algorithms are currently not bundled up into easy-to-use packages, limiting their accessibility, and therefore their usage.

To resolve this, we created a general and harmonized framework for on-target and off-target prediction of gRNAs, implemented in our R package *crisprScore*. The philosophy behind *crisprScore* is to abstract away from the user the language, implementation, and complexity of the different algorithms used for prediction. It uses the Bioconductor package *basilisk* [Lun, 2021] to seamlessly integrate and manage incompatible Python modules in one user session. This enables *crisprScore* to centralize all Python-based scoring algorithms together with R-based prediction algorithms, reporting all scores in a single data frame for convenient inspection.

We note that while the package provides a harmonized framework from a user perspective, it also allows each scoring algorithm to be implemented with its own sets of parameters and inputs. We have included as many methods as possible (Table 2), with the goal of democratizing the use of different scoring algorithms in an unbiased manner. Developers can easily contribute new methods to the *crisprScore* package as they become available. While this is beyond the scope of our work, our framework also facilitates methods comparison by providing a harmonized user interface.

2.5 *crisprDesign* enables fast characterization and annotation of off-targets

Off-targeting effects occur when a spacer sequence maps with perfect or imperfect complementarity to a genomic locus other than the primary on-target. Given that nucleases can still bind and cut in the presence of nucleotide mismatches between spacer sequences and target DNA sequences [Fu et al., 2013, Hsu et al., 2013b, Pattanayak et al., 2013], it is paramount to obtain and characterize all putative mismatch-mediated off-targets.

The off-target functionalities in *crisprDesign* are divided into two parts: off-target search (alignment) and off-target characterization (genomic context and scoring). For the off-target search, we offer three different alignment methods: *bowtie* [Langmead et al., 2009], the BWA-backtrack algorithm in *BWA* [Li and Durbin, 2009] and the Aho-Corasick exact string matching method implemented in *Biostrings* [Aho and Corasick, 1975, Pages et al., 2016]. We developed two independent R packages to implement the *bowtie* and *BWA* alignment methods: *crisprBowtie* and *crisprBwa*. Notably, the packages were developed to work with any nucleases, and for both DNA and RNA target spaces (reference genomes and transcriptomes). While the maximum number of mismatches for *bowtie* is limited to 3, there is no limit for *BWA*.

Given the short nature of gRNA spacer sequences, both *bowtie* and *BWA* are ideal tools for off-target search and provide ultrafast results. On the other hand, the alignment method based on the Bioconductor package *Biostrings* does not need the creation of a genome index, and is particularly useful for off-target search in short custom sequences. All methods can be invoked via the *addSpacerAlignments* function, which returns the on- and off-target alignments as a *GRanges* object in the *GuideSet* metadata.

To add genomic context to the on- and off-targets, a *TxDb* object can be provided to the *addSpacerAlignments* function. The *TxDb* object is a standard Bioconductor object to store information about a gene model, and can easily be made from transcript annotations available as GFF3 or GTF files. Gene annotation columns are added to the off-target table for different contexts: 5' UTRs, 3' UTRs, CDS, exons, and introns. Finally, users

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Nuclease	Variant	Method	Type	Reference
SpCas9	WT	RuleSet1	On-target efficiency	Doench et al. [2014]
	WT	Azimuth	On-target efficiency	Doench et al. [2016a]
	WT	CRISPRscan	On-target efficiency	Moreno-Mateos et al. [2015]
	WT	CRISPRai	On-target efficiency	Horlbeck et al. [2016]
	WT	DeepHF	On-target efficiency	Wang et al. [2019b]
	HiFi	DeepHF	On-target efficiency	Wang et al. [2019b]
	WT	Lindel	On-target efficiency	Chen et al. [2019]
	WT	MIT	Off-target cutting	Hsu et al. [2013a]
	WT	CFD	Off-target cutting	Doench et al. [2016b]
AsCas12a	WT	DeepCpf1	On-target efficiency	Kim et al. [2018]
	Enhanced	enPAM+GB	On-target efficiency	DeWeirdt et al. [2020]
RfxCas13d	WT	CasRx-RF	On-target efficiency	Wessels et al. [2020]
	WT	CasRx-CFD	Off-target cutting	Fortin and Lun [2022]

Table 2. On-target and off-target scoring methods currently available in *crisprScore*.

can add the MIT and CDS off-target specificity scores [Hsu et al., 2013a, Doench et al., 2016a] implemented in *crisprScore* to characterize the likelihood of nuclease cleavage at the off-targets.

Comparison of the off-target annotation methods

We compared the performance of the two short-read aligners implemented in *crisprDesign*. We note that the *bowtie* alignment is used by the tools *CHOPCHOP* [Montague et al., 2014], *CCTop* [Stemmer et al., 2015], and *multicrispr* [Bhagwat et al., 2020] and that the *BWA* method is used by *CRISPOR* [Concordet and Haeussler, 2018]. The *Biostrings* method is used by both *CRISPRseek* [Zhu et al., 2014] and *multicrispr*.

Bhagwat et al. [2020] reported that *bowtie* misses a large number of double-mismatch and triple-mismatch off-targets in comparison to the gold-standard complete string matching algorithm. To investigate this, we repeated the PAM-agnostic on- and off-target alignment of the 10 spacer sequences described in Bhagwat et al. [2020] to the GRCh38 reference genome using all three alignment methods. In contrast to Bhagwat et al. [2020], all three alignment methods implemented in *crisprDesign* return an identical list of off-targets (see Table S2). This indicates that, contrary to previous reports, both *BWA* and *bowtie* provide a complete on- and off-target search. It appears that the missing off-targets in Bhagwat et al. [2020] are located on unlocalized and unplaced GRCh38 sequences.

Next, we evaluated the run times of four configurations offered in *crisprDesign* for alignment: *bowtie*, *BWA*, an iterative version of *bowtie* (bowtie-int) and an iterative version of *BWA* (bwa-int). We developed iterative versions of the bowtie and BWA alignments to avoid situations where gRNAs are mapping to hundreds of loci in the genome, considerably slowing down the off-target search when a higher number of mismatches is allowed. The iterative strategy starts by aligning spacer sequences with no mismatches allowed. Then, it sequentially performs the alignment with a higher number of mismatches only for sequences that have a low number of off-targets at the previous step. We performed the evaluation on three sets of gRNAs targeting the human genome, each with a different size (see Methods). For all three sets, the *bowtie* and BWA gRNA alignments have comparable run times. (Figure S1). The iterative version of the index-based aligners shows substantial gain in speed for the set of gRNAs targeting *ZNF101*, which contains several non-specific guides overlapping a repeat element.

2.6 Accounting for human genetic variation by adding SNP annotation

Genetic variation, such as SNPs and small indels, can have a direct impact on gRNA binding productivity and on-target specificity by altering complementarity between spacer sequences and the target DNA or RNA [Scott and Zhang, 2017, Lessard et al., 2017, Canver et al., 2017, Wang et al., 2018]. In *crisprDesign*, users can apply the function *addSNPAnnotation* to annotate gRNAs for which the target protospacer sequence overlaps a SNP. This enables users to discard or flag undesirable gRNAs that are likely to have variable activity across different human genomes.

Given that the current human reference genome was built using only a small number of individuals, the allele

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

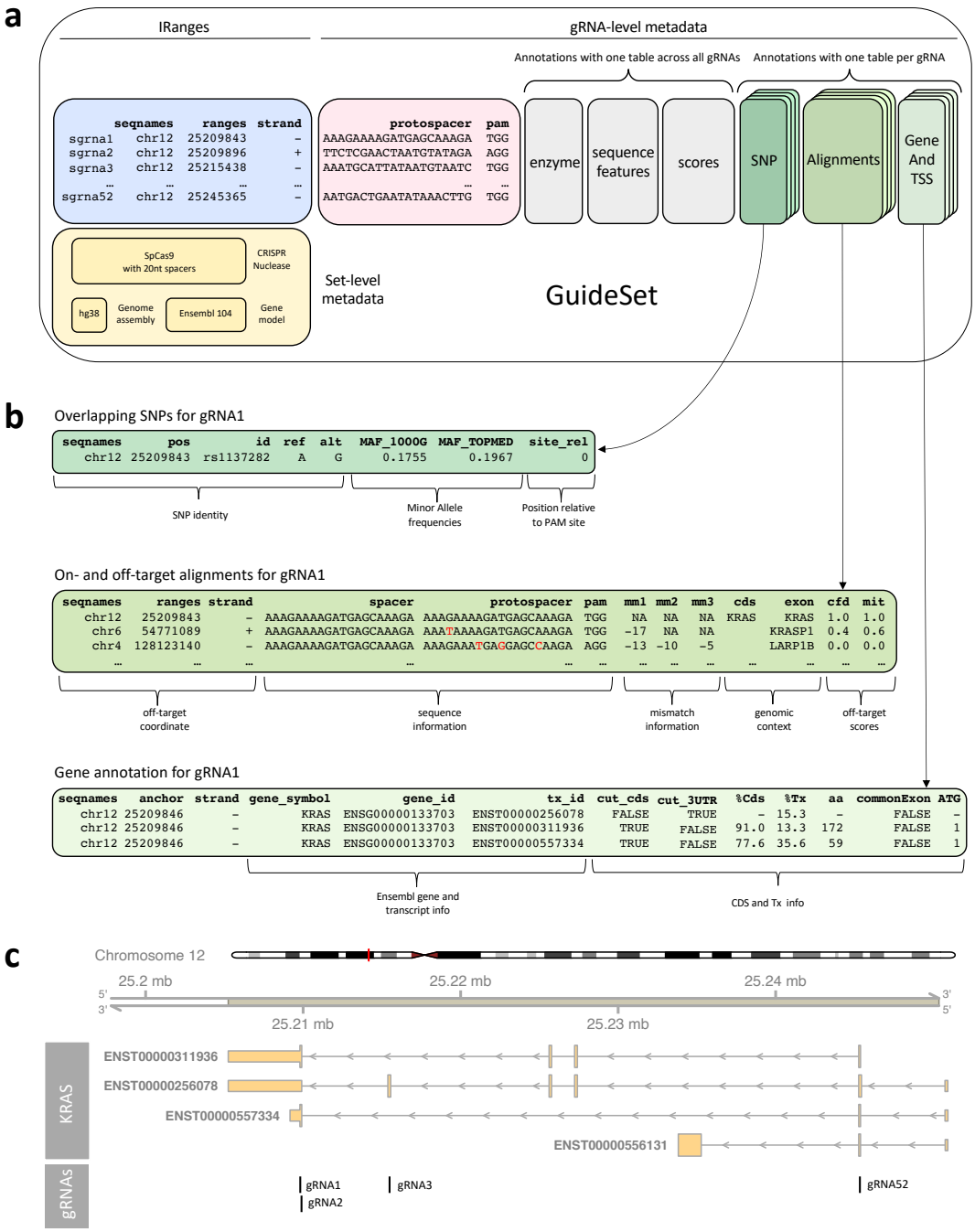


Figure 2. Example of a GuideSet container for gRNAs targeting *KRAS* using SpCas9. a The blue box stores the genomic coordinates in GRCh38 to represent the target protospacer sequences using a GRanges object. By convention, we use the first nucleotide of the PAM sequence (in the 5′ to 3′ direction) as the representative genomic coordinate of protospacer sequences. The pink box stores sequence information of the protospacers and PAMs. The yellow box represents global metadata used for creating the GuideSet, including a formal *CrisprNuclease* object, the reference genome of the target protospacers, and gene model used for annotation. The grey boxes are examples of optional gRNA-level metadata columns that store information about enzyme restriction sites, spacer sequence features such as GC content, and on- and off-target scores. The green boxes represent optional 3d-array annotations for SNP overlap, on- and off-target alignments, and gene context; each annotation stores a detailed table (2 dimensions) for each gRNA (3rd dimension). **b** Selected annotations for gRNA 1 corresponding to the row highlighted in the green boxes of **a**. **c** The first genomic track represents the four annotated protein-coding isoforms of human gene *KRAS* in GRCh38 coordinates. The second track shows the 4 gRNAs shown in the blue box of **a**.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

represented in the human reference genome at a particular locus does not always correspond to the major allele in a population of interest. Inspired by the major-allele reference genome indices provided by the bowtie team (see `https://github.com/BenLangmead/bowtie-majref`), we created two new human genomes to be used throughout our ecosystem that represent the major allele and the minor allele using dbSNP151 (see Methods). Both genomes are available in Bioconductor as *BSgenome* packages. Both packages can be used in our ecosystem to improve gRNA design by designing gRNAs against either the minor or major allele genome, and searching for off-targets in both the major and minor allele genomes.

2.7 Comprehensive gene and TSS annotations

The genomic context of the on-target sites is paramount for optimally selecting gRNAs in most, if not all, CRISPR applications. In our experience, very few existing tools provide detailed information about the gRNA target sites beyond whether or not they target exons. To address this, *crisprDesign* includes the *addGeneAnnotation* and *addTssAnnotation* functions, which report comprehensive transcript- and promoter-specific context for each gRNA target site, respectively. Users simply need to provide a standard Bioconductor *TxDb* object to specify which gene model should be used to annotate on- and off-targets.

For CRISPRko applications, *addGeneAnnotation* annotates which isoforms of a given gene are targeted, and adds spatial information about the relative cut site within the coding sequence of each isoform. Since translation reinitiation can result in residual protein expression,[Smits et al., 2019], *addGeneAnnotation* reports whether or not the gRNA cut site precedes any downstream in-frame ATG sequences, following the rules of Cohen et al. [2019]. Additionally, to maximize gene knockout based on protein domains [He et al., 2019], we include Pfam domain annotation [Bateman et al., 2004] via the *biomarRt* package [Durinck et al., 2005]. For CRISPRa and CRISPRi applications, *addTssAnnotation* indicates which promoter regions are targeted by each gRNA, as well as the location of the target cut site relative to the TSS. This allows the user to easily select guides in the optimal targeting window.

To further put the gRNA targets into biological context, users can access thousands of genomic annotation datasets through the Bioconductor *AnnotationHub* resource. The resource includes common sources such as Ensembl, ENCODE, dbSNP and UCSC. Where appropriate, those annotations are in the *GenomicRanges* format, which make them directly compatible with the *GuideSet* object used to represent gRNAs in our ecosystem. By leveraging overlap operations on *GenomicRanges*, users can identify which gRNAs are present or absent in a given set of annotated features by using a few lines of code. For example, users can ask *AnnotationHub* whether a gRNA is targeting repeat elements to avoid cutting-induced toxicity, or whether a gRNA targets the region upstream of an annotated Cap Analysis of Gene Expression (CAGE) peak for CRISPRa applications.

2.8 Advanced functionalities for designing gRNA screening libraries

Efficient cleavage can be disrupted by certain features of the gRNA sequence, such as very low or high percent GC content [Chen et al., 2018, Doench et al., 2014, Wang et al., 2014], homopolymers of four nucleotides or longer [Gilbert et al., 2014, Veeneman et al., 2020], and self-complementarity conducive to hairpin formation [Thyme et al., 2016, Labun et al., 2016]. When gRNAs are expressed from a U6 promoter, thymine homopolymers (TTTT) are particularly undesirable as they signal transcription termination. The *addSequenceFeatures* function flags all gRNAs that contain such undesirable sequence features. Another consideration in designing gRNA libraries is to exclude spacer sequences that are not compatible with the oligonucleotide cloning strategy. gRNAs that contain restriction sites of the enzymes used to clone the spacer sequences into a lentiviral vector should be excluded. The *addRestrictionEnzymes* function flags all gRNAs that contain restriction enzyme recognition motifs.

To complement gRNA annotation and assist in library design, we provide a pair of convenient and versatile functions to filter and rank gRNAs within a *GuideSet*: *filterSpacers* and *rankSpacers*. Each function allows users to define their own filtering and ranking criteria to meet their specific needs, respectively. A user could first filter out gRNAs that contain a restriction enzyme recognition site or have more than one single-mismatch off-target located in a CDS region, and then subsequently rank by on-target efficiency score and the distance to the $5'$ end for a given gene. These functions offer a fast and consistent method for flexible gRNA selection using state-of-the-art design rules, automating the library design process in a user-friendly manner.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Optical pooled screening (OPS) is a promising novel screening modality that combines image-based *in situ* sequencing of gRNAs and optical phenotyping on the same physical wells [Feldman et al., 2019]. This enables linking genomic perturbations with high-content imaging at large scale. In such experiments, gRNA spacer sequences are partially sequenced. From a gRNA design perspective, this translates to additional gRNA design constraints to ensure sufficient dissimilarity of the truncated spacer sequences. *crisprDesign* contains a suite of design functions that take into account OPS constraints, while making sure that the final OPS library is enriched for gRNAs with best predicted activity.

3 Case studies

3.1 Designing gRNAs targeting *BRCA1* for the base editor BE4max

CRISPR base editors are deaminases fused to CRISPR nickases to introduce mutations at loci targeted by the gRNAs without introducing double-stranded breaks (DSBs) Gaudelli et al. [2017], Komor et al. [2016]. A recent study characterizing eight popular base editors showed high heterogeneity and complexity of the editing outcomes across base editors, motivating the need of robust but flexible software to design gRNAs for base editing applications [Arbab et al., 2020]. In particular, this includes functionalities for listing and characterizing potential edited alleles introduced by gRNAs to inform the phenotypic readouts created by those gRNAs.

To illustrate the functionalities of our ecosystem for designing base editor gRNAs, we designed and characterized all possible gRNAs targeting the coding sequence of *BRCA1* for the cytidine base editor BE4max [Koblan et al., 2018]. The design workflow is shown in Figure 3.

The first step consisted of designing all possible guides targeting *BRCA1* using the *findSpacers* function in *crisprDesign*. The BE4max *BaseEditor* object from *crisprBase* was used to store nucleotide- and position-specific editing probabilities (see Figure 3a), which inform the editing window of interest for each of the gRNA targets. Next, using the function *getEditedAlleles*, we generated and stored all possible editing events at each gRNA (see Figure 3b). The function also adds a score for each edited allele that quantifies the likelihood of editing to occur based on the editing probabilities stored in the *BaseEditor* object (see Methods). In addition, each edited allele is annotated for its predicted functional consequence: silent, missense, or nonsense mutation. In case several mutations occur in a given edited allele, the most consequential mutation is used to label the allele (nonsense over missense, and missense over silent). For each gRNA, and for each mutation type, we then generated a gRNA-level score by aggregating the likelihood scores across all possible alleles (see Methods). The score represents the likelihood of a gRNA to induce a given mutation type (see Figure 3c, left plot).

To show how our gRNA annotations can be used to understand the phenotypic effects observed in screening data, we obtained data from a negative selection pooled screen performed in MelJuSo using a base editing library tiling the *BRCA1* gene [Hanna et al., 2021]. Given that loss-of-function mutations in *BRCA1* reduce cell fitness [Findlay et al., 2018], gRNAs introducing nonsense mutations are expected to drop out, and can therefore serve as positive controls. We created Receiver Operating Characteristic (ROC) curves to measure how well gRNA dropout can separate positive controls from other gRNAs. We used log fold changes in gRNA abundance between the later time point and the plasmid DNA (pDNA) library as a measure of gRNA dropout (see Methods). We used several thresholds of the nonsense mutation score to label gRNAs as positive controls or not. We observed that gRNA dropout in the screen can separate positive controls well from all other gRNAs, and that performance is improved when using positive controls defined by higher nonsense mutation scores (Figure 3c).

We also characterized gRNAs for off-targeting effects using *crisprBowtie*, added sequence features using *crisprDesign*, and added on-target scores using *crisprScore*. We asked whether or not the magnitude of gRNA dropout in the screen associates with predicted on-target activity for the SpCas9 nuclease. In Figure 3d, we show gRNA dropout as a function of different predicted gRNA efficacy scores: Rule Set 1, Azimuth, and DeepHF. gRNAs predicted to induce nonsense mutations are shown in red, and grey otherwise. Despite the fact that each algorithm was trained on data using a SpCas9 nuclease with intact endonuclease activity, gRNA dropout and predicted gRNA efficacy correlate for all methods ($r = -0.30$ for Rule Set 1, $r = -0.20$ for Azimuth, and $r = -0.17$ for DeepHF). Overall, the different functionalities implemented in our ecosystem provides a set of informative annotations for base editor gRNAs and facilitate the interpretation of experimental data obtained from base editor screens.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

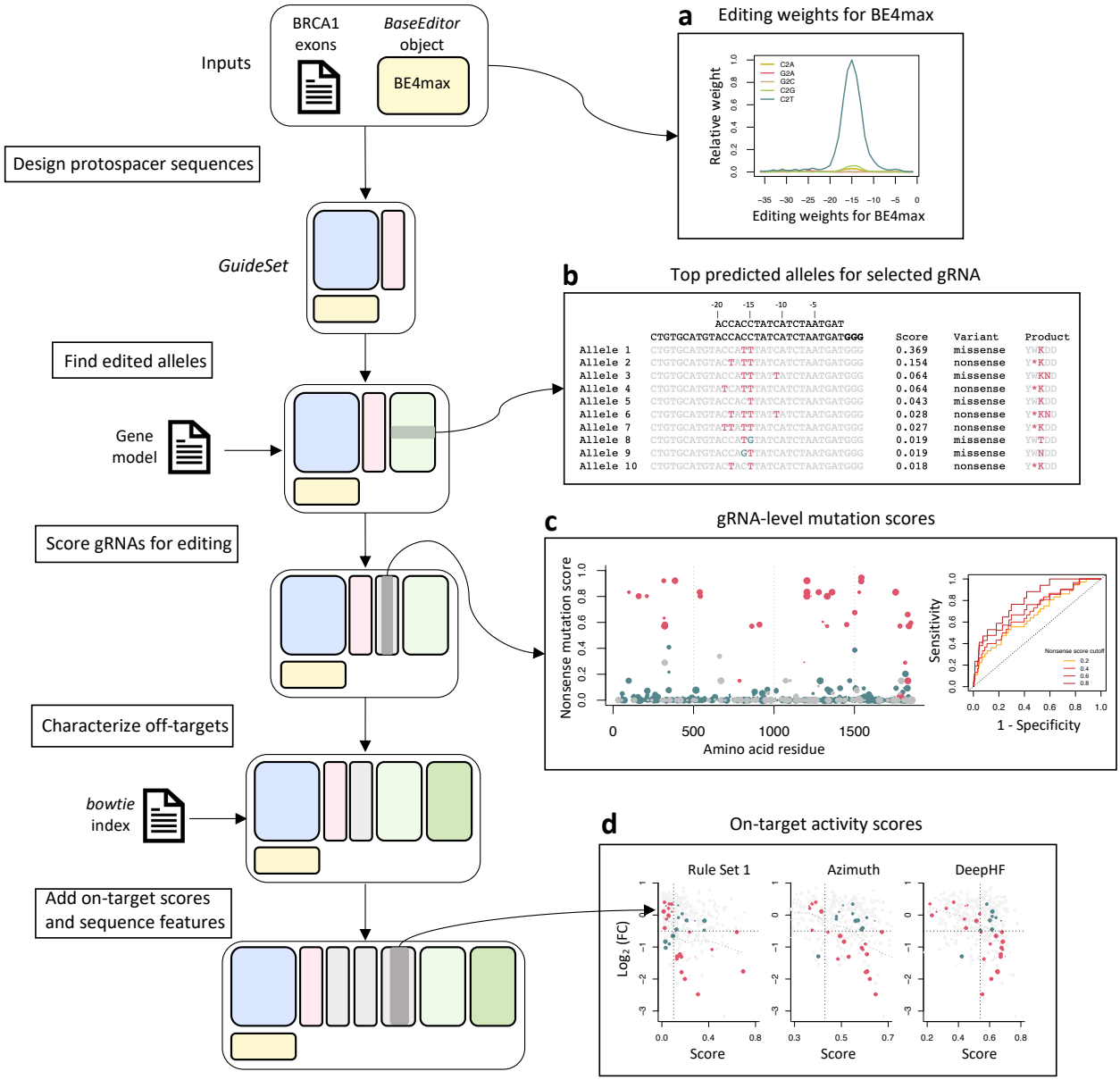


Figure 3. *crisprDesign* workflow to design gRNAs tiling *BRCA1* using the base editor BE4max. On the left: schematic showing the major steps involved in designing BE4max gRNAs targeting *BRCA1*. Two inputs are required: DNA sequences of *BRCA1* exons and a *BaseEditor* object from *crisprBase*. **a** Editing weights for the BE4max base editor from *crisprBase*. **b** 10 top predicted edited alleles for one selected gRNA as returned by *crisprDesign*. The wildtype allele and the protospacer sequence are positioned at the top of the first column, with the PAM sequence highlighted in bold. Edited nucleotides are highlighted in red (C to T) and blue (C to G). Editing scores, variant annotations, and protein product of the edited alleles are also shown. **c** On the left, gRNA-level nonsense mutation score as calculated by *crisprDesign*. Colors represent variant classification: nonsense in red, missense in blue, silent in grey. The size of the dot is proportional to the on-target efficiency *DeepHF* score. On the right, ROC curves for classifying gRNA mutation type (nonsense or not) based on gRNA dropout from the *BRCA1* BE4max dataset (see Methods). Different thresholds of the nonsense score were used to label a gRNA as nonsense or not. **d** Relationships between gRNA dropout from the *BRCA1* BE4max dataset and several on-target activity scores. gRNAs that are not predicted to induce a nonsense mutation are colored in grey, and the size of the dots is proportional to the magnitude of the mutation score. The horizontal dotted lines at -0.5 represent a cutoff to classify a gRNA as active or not. For each method, a score cutoff was determined to classify active versus non-active gRNAs (vertical dotted line). Red and blue dots correspond to gRNAs that are correctly and incorrectly classified, respectively.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

3.2 Annotating and scoring gRNAs for gene knockdown using CasRx

One of the challenges in designing gRNAs specific to RNA-targeting nucleases is to enable on-target and off-target characterization to be performed in a transcriptome space, as opposed to a reference genome. This requires strand-specific functionalities, transcriptome-specific alignment indexes, as well as additional gene annotation functionalities to capture isoform-specific targeting.

Here, we describe a workflow for designing gRNAs targeting *CD46* and *CD55* using the RNA-targeting nuclease CasRx (RfxCas13d) [Konermann et al., 2018] (Figure 4). The workflow takes into consideration the aforementioned issues. To validate our design process, we obtained CasRx pooled screening data performed in HEK 293 cells with gRNA libraries tiling the human genes *CD46* and *CD55* from Wessels et al. [2020]. Since both genes encode for cell-surface proteins, the authors used fluorescence-activated cell sorting (FACS) to sort cells with high and low expression. Their data can be used to investigate gRNA knockdown efficacy based on the change in relative abundance of high- and low-expressing cells for each targeted gene (see Methods).

We first extracted mRNA sequences of both genes using the function *getMrnaSequence* from *crisprDesign*. The mRNA sequences, together with the *CrisprNuclease* object CasRx from *crisprBase*, served as inputs to create a *Guideset*. Next, we predicted on-target activity of the gRNAs using our implementation of the CasRx-RF method [Wessels et al., 2020] available in *crisprScore* (see Methods). The normalized log-fold changes in the screen correlate well with the CasRx-RF score (Figure 4a). We then added a transcript annotation to each gRNA using an Ensembl *TxDb* object as input. This adds a list of targeted isoforms to each gRNA, as well as transcript context (CDS, 5′UTR, or 3′UTR). We observed in the screen that gRNAs targeting a higher number of isoforms, and gRNAs located in CDS, lead to higher activity (Figure 4b, and Figure S2).

We performed an off-target search using *crisprBowtie* to the human transcriptome by providing a bowtie index built on mRNA sequences. We extended the CFD off-target scoring algorithm implemented in *crisprScore* to work with CasRx by estimating mismatch tolerance weights on the GFP tiling screen data from Wessels et al. [2020] (see Methods). The off-target CFD-CasRx score performs well at predicting gRNA activity of single-mismatch and double-mismatch gRNAs in the *CD55* screen (Figure 4c, and see Methods).

Finally, we added sequence features, and ranked gRNAs for targeting *CD55* and *CD46* based on (1) high on-target score, (2) low number of off-targets, and (3) high number of targeted isoforms. If we select gRNAs that target a common exon and that have high on-target score, we enrich for highly active gRNAs in the screening data (Figure 4d).

3.3 Designing optimal gRNAs to activate *MMP7* using CRISPRa using different nucleases

Designing gRNAs for either CRISPRa and CRISPRi applications requires additional considerations. This includes choosing an optimal target region based on chromatin accessibility data and TSS data, and selecting gRNAs based on their positioning with respect to the TSS.

To demonstrate the utility of our ecosystem functionalities for CRISPRa and CRISPRi, we designed gRNAs for CRISPRa using the human gene *MMP7* as an example target (Figure 5). CRISPRi is discussed at the end of this section. One CRISPRa-specific design consideration is the limited number of candidate gRNAs available for a given gene due to the narrow window of optimal activation. Engineered nucleases with less constrained PAM sequences can improve CRISPRa applicability by expanding the set of candidate gRNAs. To investigate this, we designed gRNAs for the promoter region of *MMP7* using four nucleases in parallel: SpCas9, AsCas12a, and the more PAM-flexible versions SpGCas9 [Walton et al., 2020] and enAsCas12a [DeWeirdt et al., 2020].

The first step of the gRNA design was to specify the target region for *MMP7*. We used *AnnotationHub* to find CAGE peaks in the promoter region of *MMP7* to specify the TSS position. We used the CAGE data to identify TSSs instead of RefSeq or Ensembl as the former provides more accurate annotations for designing CRISPRi and CRISPRa gRNAs [Radzisheuskaya et al., 2016]. The 5′ end of the CAGE peak was used as the TSS to define the coordinates of the optimal window of activation (75 and 150 nucleotides upstream of the TSS, as recommended by Sanson et al. [2018]).

Next, we designed all possible gRNAs for the four nucleases using the *findSpacers* function in *crisprDesign*, and stored the gRNAs in four separate *GuideSet* containers. All four nucleases are available in *crisprBase*. We

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

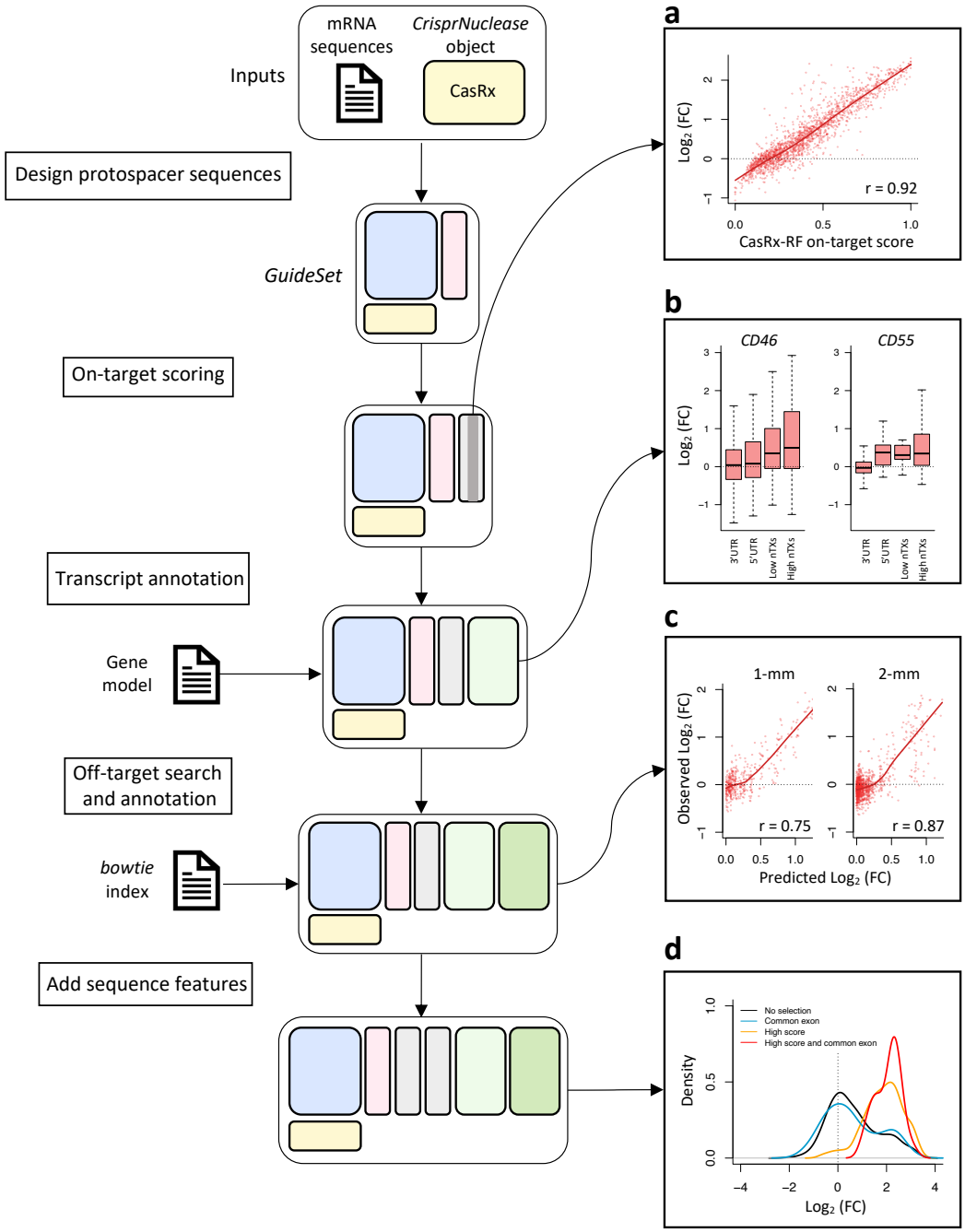


Figure 4. *crisprDesign* workflow to design gRNAs tiling *CD55* and *CD46* using CasRx. On the left: schematic showing the major steps involved in designing CasRx gRNAs targeting *CD55* and *CD46*. Two inputs are required: mRNA sequences of *CD55* and *CD46* and a *CrisprNuclease* object from *crisprBase*. **a** Relationship between on-target CasRx-RF score calculated in *crisprScore* and log-fold changes from the pooled FACS tiling CasRx screening data (see Methods). A higher log-fold change indicates higher gRNA activity. **b** Relationship between log-fold changes from the CasRx screening data and gRNA context for *CD46* and *CD55*: gRNAs targeting 5′ UTR and 3′ UTR for the canonical transcript, and guides targeting a low and high number of isoforms for each of the genes. gRNAs targeting more isoforms show higher enrichment in the screening data. The full isoform annotation is stored in the *GuideSet* objects. **c** Left: relationship between observed log-fold changes of on-target gRNAs in the *CD55* screen and predicted log-fold changes of single-mismatch gRNAs using the off-target CFD-CasRx score implemented in *crisprScore* (see Methods). Right: same as left, but for double-mismatch gRNAs. **d** gRNAs selected in the *CD46* screen for high on-target activity (CasRx-RF score) and targeting a common exon across all protein-coding isoforms enrich for high gRNA activity.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

annotated each *GuideSet* for overlap with DNase I hypersensitivity sites (DHS) from consolidated epigenomes from the Roadmap Epigenomics Project [Kundaje et al., 2015] using *AnnotationHub*. Open-chromatin regions are favorable for the binding of the catalytically inactive Cas9 (dCas9) used in both CRISPRa and CRISPRi [Kuscu et al., 2014, Wu et al., 2014]. We then added sequence features using *crisprDesign*, on-target scores using *crisprScore*, and off-target sites using *crisprBowtie* for each nuclease. Finally, we added overlapping SNPs information using the *addSNPAnnotation* function and using dbSNP151. The end-to-end workflow is presented in Figure 5a.

The designed gRNAs are presented in Figure 5b. With *crisprDesign*, it is straightforward to select candidate gRNAs in the most promising genomic regions - in this case, lying inside both the annotated DHS and the optimal activation window for MMP7. One can immediately appreciate that both nuclease variants (SpGCas9 and enAsCas12a) yield substantially more available gRNAs in the optimal window activation. In particular, enAsCas12a offers several gRNAs with high predicted on-target activity, making it a better candidate for gene activation of *MMP7*. One SNP was also found in the region of interest, and overlapping a gRNA for enAsCas12a, telling users to avoid using that gRNA. Altogether, our ecosystem provides an easy and comprehensive workflow to enable users to design optimal gRNAs for CRISPRa across nucleases.

Designing gRNAs for CRISPRi applications using *crisprDesign* is nearly identical, with the exception that the preferred target region for interference is located downstream of the TSS. The CRISPRai scoring algorithm from Horlbeck et al. [2016], available through *crisprScore*, can be used to select optimal gRNAs for each TSS separately, taking into account both gRNA positioning and sequence content to maximize on-target inhibition. For both CRISPRa and CRISPRi, our gRNA design workflow is also applicable to non-coding regulatory elements, for instance long non-coding RNAs (lncRNAs) as it was done in Liu et al. [2017]. Overall, *crisprDesign* provides end-to-end functionalities that are well-suited for a large array CRISPRa and CRISPRi applications.

4 Discussion

In this work, we introduced a new suite of R packages to perform comprehensive end-to-end gRNA design for a multitude of CRISPR technologies and applications. Our ecosystem enables users to perform gRNA design for diverse nucleases such as PAM-free nucleases and RNA-targeting nucleases, and for several applications beyond CRISPRko such as RNA and DNA base editing and CRISPRa/i. All design functionalities are available from a single package, *crisprDesign*. This eliminates the need to use multiple tools to obtain the necessary information for selecting optimal gRNAs, which is both time consuming and error prone. We demonstrated the diversity of our framework by applying it in three case studies involving different CRISPR technologies with their own specific design considerations.

We were able to show that creating rich gRNA annotations can help investigate gRNA variability and biases observed in experimental data generated from newer CRISPR technologies. To do so, we obtained public pooled screening data from two published studies, a tiling base editor screen of *BRCA1*, and a tiling CasRx screen of *CD46* and *CD55*, and show how some of the gRNA features derived from *crisprDesign* can explain some of the variability in gRNA activity observed in both screens.

The modular architecture of *crisprDesign* enables nucleases, base editors, scoring methods and annotations to be combined depending on the needs of the user. As a result, our design framework can easily adapt to new CRISPR technologies by swapping out the necessary components. For instance, a recent study has shown that the resolution of base editor screens can be greatly increased by combining existing base editors with PAM-extended Cas9 variants [Sangree et al., 2021], while another study shows that RNA-targeting Cas13 nucleases can be combined with deaminases to form RNA base editors [Cox et al., 2017]. Both applications can be readily supported by our ecosystem without the need for further development.

Our ecosystem is completely implemented within the Bioconductor project, which provides robust and feature-rich data structures, high-quality documentation and workflows, and seamless interoperability between packages. Data structures defined in *crisprBase* can be reused to facilitate the analysis of CRISPR-based editing events in other packages, such as *ampliCan* [Labun et al., 2019], *GUIDEseq* [Zhu et al., 2017] and *CrisprRVariants* [Lindsay et al., 2016], and *GuideSet* gRNA containers can be integrated with packages that provide analysis workflows for pooled screening data [Wang et al., 2019a, Imkeller et al., 2020, Bainer et al., 2021] to investigate biases and filter out undesirable gRNAs. Finally, the *crisprBowtie* and *crisprBwa* packages provide general

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

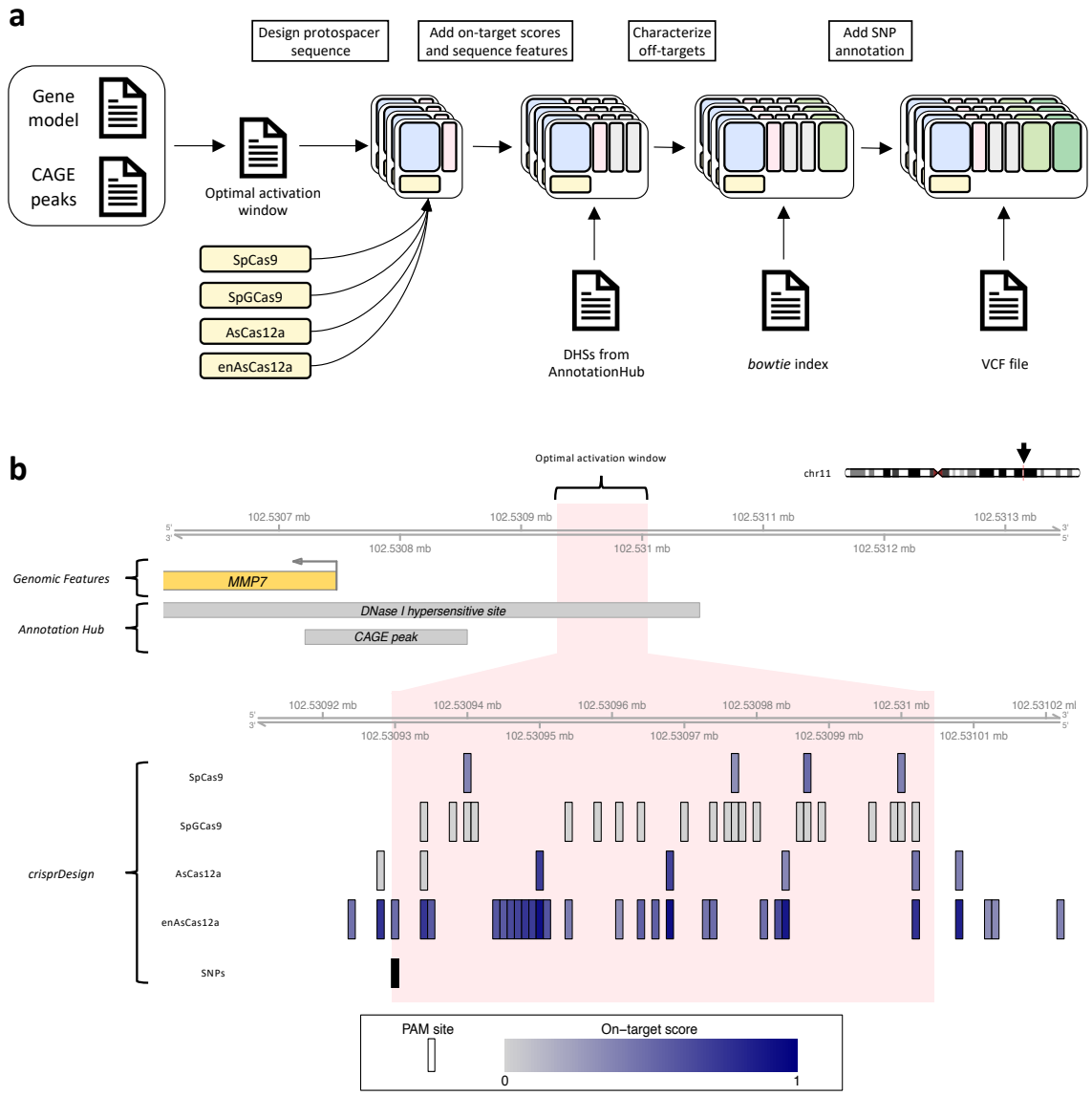


Figure 5. Design of CRISPRa gRNAs for human gene *MMP7* for different CRISPR nucleases. **a** Schematic showing the steps involved in designing CRISPRa gRNAs targeting the promoter region of *MMP7*. A gene model and a list of CAGE peaks are used to define the optimal window for gene activation. A *GuideSet* is created separately for each CRISPR nuclease. DNase I hypersensitive site (DHS) information is obtained from *AnnotationHub* and added to the gRNA annotation. **b** The top track shows the promoter region of human gene *MMP7* on chromosome 11, including part of the 5′ UTR of *MMP7* (yellow). The DHS and CAGE peak grey boxes were obtained using *AnnotationHub* (see Methods). The light pink region corresponds to the optimal region of activation based on Sanson et al. [2018], corresponding to a region [75,150]bp upstream of the 5′ end of the CAGE peak. For each of the four selected nucleases, all canonical PAM sites located within the optimal region are shown. PAM sites are colored by their on-target score: DeepHF for SpCas9, DeepCpf1 for AsCas12a, and enPAM+GB for enAsCas12a. No on-target scoring algorithm was available at time of publication for SpGCas9. The last track corresponds to common SNPs obtained from dbSNP151.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

functions that can be used to map any short sequences, including small-hairpin RNAs (shRNAs) and short-interfering RNAs (siRNAs).

We are continuously extending our suite of tools to make available the latest developments for gRNA design. We are currently extending our work to include design for prime editing [Anzalone et al., 2019]. Prime editing requires the design of template sequences in addition to the usual gRNA spacer sequences. Prime editing tools are rapidly evolving, and our infrastructure provides a robust framework to facilitate the complex design of a variety of prime editing constructs. In addition, we will implement more convenience functions to streamline the design of complex gRNA libraries, such as multiplexing libraries that combine multiple gRNA sequences within one construct. Such libraries can be used to increase on-target efficiency [Replogle et al., 2020], or to simultaneously target pairs of loci in the genome [Han et al., 2017]. We are also committed in making available new scoring algorithms by implemented them in *crisprScore* as they become available. Finally, to make access to our rich ecosystem accessible to all users, we are working on developing an interactive graphical user interface (GUI) using *Shiny* [Chang et al., 2015]. The Shiny application will combine user-friendly characteristics of existing web-based tools with the advanced capabilities of our R-based ecosystem.

5 Methods

Reference genomes, gene models, and genome indexes

The FASTA file for the human reference genome (GRCh38.p13 assembly) was obtained from UCSC to build *bowtie* and *BWA* indexes via the *Rbowtie* [Hahne et al., 2012] and *Rbwa* R packages, respectively. The gene model used throughout the manuscript was obtained from Ensembl (release 104) using the R package *GenomicFeatures*. Common SNPs were obtained from NCBI dbSNP build 151 (`https://ftp.ncbi.nlm.nih.gov/snp/`).

CAGE peak and DNAse I hypersensitivity data

RIKEN/ENCODE CAGE peaks were obtained from *AnnotationHub* using accession number AH5084 [Djebali et al., 2012]. Genomic coordinates were lifted over from hg19 to hg38 using the R package *rtracklayer*. DNAse I hypersensitive sites were obtained from *AnnotationHub* using accession number AH30743. The narrow DNase peaks were obtained using MACS2 on consolidated epigenomes from the Roadmap Epigenomics Project (`E116-DNase.macs2.narrowPeak.gz`) [Kundaje et al., 2015]. Genomic coordinates were lifted over from hg19 to hg38 using the R package *rtracklayer*.

On-target scoring

We implemented several commonly-used algorithms for Cas9, Cas12 and Cas13 nucleases in *crisprScore*. For predicting on-target activity of the wildtype SpCas9 nuclease, we implemented the popular Rule Set 1 [Doench et al., 2014] and Azimuth algorithms [Doench et al., 2016a] (iteration of the popular Rule Set 2 algorithm by the same authors). The package also provides the deep learning-based algorithms DeepWT and DeepHF, developed to predict cutting efficiency of the wildtype SpCas9 and SpCas9-High Fidelity (SpCas9-HF1) nucleases, respectively [Wang et al., 2019b]. We also included the *CRISPRscan* algorithm [Moreno-Mateos et al., 2015] for predicting on-target activity of SpCas9 gRNAs expressed from a T7 promoter. For the wildtype AsCas12a, *crisprScore* offers the deep-learning based prediction method DeepCpf1 [Kim et al., 2018]. For the enhanced AsCas12a (enAsCas12a), *crisprScore* offers the *enPAM+GB* algorithm [DeWeirdt et al., 2020]. For CasRx (RfxCas13d), we adapted the code from the random forest model developed in Wessels et al. [2020]; we referred to the method as CasRx-RF.

For predicting gRNA activity for CRISPRa and CRISPRi, we implemented the prediction method used to design the commonly-used Weissman CRISPRa and CRISPRi v2 genome-wide libraries for human and mouse [Horlbeck et al., 2016]. This method predicts CRISPRa (or CRISPRi) gRNA activity based on the distance to the transcription starting site (TSS), spacer sequence-derived features, as well as chromatin accessibility data and nucleosome positioning using DNase-Seq, MNase-Seq, and FAIRE-Seq data.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

On-target prediction of frameshift-causing indels using *Lindel*

In *crisprScore*, we implemented *Lindel* [Chen et al., 2019], a logistic regression model that was trained to use local sequence context to predict the distribution of mutational outcomes for CRISPR/Cas9. The *Lindel* final score reported in *crisprScore* is the proportion of "frameshifting" indels, that is the frequency of indels predicted to introduce frameshift mutations. By chance, assuming a random distribution of indel lengths, gRNAs should have a frameshifting proportion of 0.66. A *Lindel* score higher than 0.66 indicates that a given gRNA is more likely to cause a frameshift mutation than by chance.

Off-target scoring of individual off-targets

The exact formula that we use to calculate the CFD score in *crisprScore* is

$$\text{CFD} = \prod_{p \in M} w_p(x_{\text{RNA}}, x_{\text{DNA}})$$

where M is the set of positions for which there is a mismatch between the gRNA spacer sequence and the off-target sequence. $w_p(x_{\text{RNA}}, x_{\text{DNA}})$ is an experimentally-derived mismatch tolerance weight at position p depending on the RNA nucleotide x_{RNA} and the DNA nucleotide x_{DNA} (see Doench et al. [2016b] for more details).

The exact formula that we use to calculate the MIT score in *crisprScore* was obtained from the MIT design website (`crispr.mit.edu`):

$$\text{MIT} = \left(\prod_{p \in M} w_p\right) \times \frac{1}{\frac{19-d}{19} \times 4 + 1} \times \frac{1}{m^2}$$

where M is the set of positions for which there is a mismatch between the gRNA spacer sequence and the off-target sequence, w_p is an experimentally-derived mismatch tolerance weight at position p, d is the average distance between mismatches, and m is the total number of mismatches. As the number of mismatches increases, the cutting likelihood decreases.

Composite off-target score for gRNA specificity

To create a gRNA-level composite specificity score, individual off-target cutting scores are aggregated using the following inverse summation formula:

$$\text{Specificity} = \frac{1}{1 + \sum_{i=1}^{n} \text{C}_i}$$

where C_i is the cutting likelihood score (either using the MIT or the CFD method) for the ith putative off-target. A higher composite score indicates higher specificity, which decreases with more off-targets and/or a greater likelihood of cleavage at each off-target. A gRNA with no putative off-targets have a composite score of 1. A gRNA with 2 on-targets, that is a gRNA targeting two genomic loci with perfect complementarity, will have a composite score of 0.5.

Base editing scoring

The behavior of a base editor can be quantified in a 3-dimensional array of editing probabilities. Let p be the genomic position relative to the PAM site; let nuc_u be the original nucleotide; and let nuc_e be the edited nucleotide. Denote $w(p, nuc_u, nuc_e)$ as the probability that nuc_u is edited to nuc_e at position p. Experimental editing weights can be used, possibly after some adequate transformation, to obtain those probabilities.

To score the likelihood of each edited allele, we assume independence of editing events with respect to nucleotide position. Specifically, consider a wildtype allele $U = (u_{p_1}, u_{p_2}, ..., u_{p_n})$ and an edited allele $V = (v_{p_1}, v_{p_2}, ..., v_{p_n})$, where u_{p_i} and v_{p_i} are the nucleotides at position p_i relative to the PAM site for the wildtype and edited allele, respectively. The parameter n is chosen by the user, and should be large enough so that all nucleotides within

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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 editing window of the chosen base editor are represented. We calculate the editing score for the edited allele V (with respect to the wildtype allele U) as follows:

$$S(U,V) = \prod_{i=1}^{n} w(p_i, u_{p_i}, v_{p_j}) \tag{1}$$

For a given edited allele V, we classify the functional consequence of editing as either a silent, missense, or nonsense mutation. We use $f(V)$ to label the mutation. In case an edited allele results in more than one mutation, we choose the most consequential mutation as the label (nonsense over missense, and missense over silent). For a given gRNA targeting the wildtype allele U, and the set of all possible edited alleles V_j, we calculate an aggregated score for each mutation type by summing the editing scores across alleles for each mutation type. For instance, the aggregated score for silent mutations is calculated as follows:

$$S_{silent}(U) = \sum_{j=1}^{N} S(U,V_j)\mathbf{1}(f(V_j) = \text{silent}) \tag{2}$$

where N is the total number of possible edited alleles V_j.

Creation of major and minor allele human genomes

We built major and minor allele genomes for the hg38 build using common SNPs from the dbSNP151 RefSNP database. The "common" category is based on germline origin and a minor allele frequency (MAF) of >= 0.01 in at least one major population, with at least two unrelated individuals having the minor allele. See the dbSNP website `https://www.ncbi.nlm.nih.gov/variation/docs/human_variation_vcf/` for more information. We excluded indels, and only considered SNPs that have MAF greater than 1% in the 1000 Genomes Project population. We then injected major alleles and minor alleles into the reference genome hg38 sequence to create "major allele" and "minor allele" genomes, respectively. Both resulting genomes are provided as standard FASTA files. We generated bowtie and BWA indexes for the two genomes. The two allele genomes are available from Bioconductor via their respective packages:

- `BSgenome.Hsapiens.UCSC.hg38.dbSNP151.major` [Fortin, 2021a]
- `BSgenome.Hsapiens.UCSC.hg38.dbSNP151.minor` [Fortin, 2021b]

Base editing pooled screen data analysis

Dropout screen data in the MelJuSo cell line using a gRNA library tiling *BRCA1* were obtained from the supplementary material of Hanna et al. [2021]. We normalized the raw counts by scaling by the total number of reads, and $\log_2$-transformed the data. We filtered out low-abundance gRNAs that were further than 3 standard deviations below the mean in the plasmid (pDNA) sample. From the later timepoint samples, we subtracted from the pDNA sample log counts to obtain log2 fold changes, and averaged the log2 fold changes across replicates. We filtered out gRNAs targeting multiple loci, and gRNAs with off-targets (with up to 2 mismatches) located in genes other than *BRCA1*.

CasRx pooled screen data analysis

CasRx FACS pooled screening data tiling *CD55*, *CD46* and *GFP* were obtained from Wessels et al. [2020], including processed and normalized log-fold changes for each gRNA `https://gitlab.com/sanjanalab/cas13`. We redesigned all possible gRNAs targeting any of the isoforms of *CD55* and *CD46* using *crisprDesign*, and considered only gRNAs also present in the pooled screening data for downstream analyses. We annotated all gRNAs with gene information (Ensembl release 104) and obtained off-targets with up to 3 mismatches for all gRNAs using *crisprBowtie*. We obtained CasRx-RF on-target activity scores using *crisprScore*. The transcripts annotated as canonical by Ensembl (ENST00000367042 for *CD46*, and ENST00000367064 for *CD55*) were used to visualize log-fold changes.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Component	Link	In Bioc
crisprDesign	https://github.com/Jfortin1/crisprDesign	Yes
crisprDesignData	https://github.com/Jfortin1/crisprDesignData	No
crisprBase	https://github.com/Jfortin1/crisprBase	Yes
crisprBowtie	https://github.com/Jfortin1/crisprBowtie	Yes
crisprBwa	https://github.com/Jfortin1/crisprBwa	Yes
crisprScore	https://github.com/Jfortin1/crisprScore	Yes
crisprScoreData	https://github.com/Jfortin1/crisprScoreData	Yes
Rbwa	https://github.com/Jfortin1/Rbwa	Yes
Analysis code	https://github.com/Jfortin1/crisprDesignPaper	No

Table 3. GitHub repositories

For each gRNA, we quantified the abundance of its target gene by summing transcript per million (TPM) counts in HEK-293 cells for all transcripts targeted by the gRNA. Transcript-level RNA quantification for HEK-293 cells was obtained from the Protein Atlas web portal https://www.proteinatlas.org, on March 5 2022. Data are based on The Human Protein Atlas version 21.0 and Ensembl version 103 . We averaged TPM counts across the two replicates.

We used the single-mismatch (SM) gRNA constructs from the GFP tiling screen to estimate position-dependent probabilities of mismatch tolerance by the CasRx nuclease. To do so, we first calculated differences in log2 fold change (ΔLFC) between SM gRNAs and their corresponding perfect-match (PM) gRNAs. We then fitted a LOESS curve with respect to the nucleotide position to obtain an average ΔLFC at each spacer position (Figure S3a). We transformed the LOESS fitted values to a scale between 0 and 1 to represent them as percentages of activity with respect to the median activity of the PM gRNAs tiling GFP (Figure S3b). We note that given the sparsity of the data, specifying a nucleotide-specific weight at each position was not considered. We adapted in *crisprScore* the CFD off-targeting scoring method to CasRx by using those probabilities as scoring weights. The corresponding scoring algorithm is named CFD-CasRx.

To evaluate the performance of the CFD-CasRx score on an independent dataset, we calculated CFD-CasRx off-target scores on all SM and double-mismatch (DM) gRNAs included in the *CD55* tiling screen. To predict log fold changes of the DM gRNAs, we multiplied their respective PM gRNA log fold changes with the CFD-CasRx on-target scores.

Off-target search comparison

For comparing runtimes of the off-target alignment methods, the following sets of gRNAs were chosen: (1) gRNAs targeting the coding sequence of *KRAS*, for a total of 52 gRNAs; (2) gRNAs targeting the coding sequence of *EGFR*, for a total of 645 gRNAs, and (3) gRNAs targeting the coding sequence of *ZNF101*, for a total of 152 gRNAs. The *KRAS* and *EGFR* cases represent small- and medium-sized sets of gRNAs. For *ZNF101*, a few gRNAs overlap a repeat element, and therefore have a high number of on- and off-targets. Alignment was performed to the GRCh38.p13 genome. The *bowtie* and *Biostrings* alignment methods were evaluated using 0 to 3 mismatches, and the BWA alignment methods were evaluated using 0 to 5 mismatches. Run times were collected on a Macbook Pro with an Intel Core i7 CPU (2.6GHz, 6 cores, 16 GB memory).

Software and code availability

All of our R packages are open-source and available on GitHub (Table 3). Reproducible code of all analyses can be found at https://github.com/Jfortin1/crisprDesignPaper.

Abbreviations

- ABE: adenine base editor
- BWA: Burrows-Wheeler Aligner

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

- CAGE: Cap Analysis of Gene Expression
- CBE: cytosine base editor
- CCLE: cancer cell line encyclopedia
- CDS: coding DNA sequence
- CFD: cutting frequency determination
- CRISPR: clustered regularly interspaced short palindromic repeats
- CRISPRa: CRISPR activation
- CRISPRbe: CRISPR base editing
- CRISPRi: CRISPR interference
- DHS: DNAse I hypersensitive site
- DSB: DNA double-strand break
- FACS: fluorescence-activated cell sorting
- lncRNAs: long non-coding RNAs
- MAF: minor allele frequency
- OPS: optical pooled screening
- PAM: protospacer adjacent motif
- PE: prime editing
- pegRNA: prime editing guide RNA
- PFS: protospacer flanking sequence
- RNAi: RNA interference
- gRNA: single-guide RNA
- shRNA: short-hairpin RNA
- siRNA: small-interfering RNA
- SNP: single-nucleotide polymorphism
- TSS: transcription starting site

Competing interests

The authors declare that they have no competing interests.

Authors contributions

JPF led the software development and supervised the work. JPF conceptualized and wrote the manuscript, with contributions and input from all authors. LH and JPF developed the R packages, with contributions from PP and AL. All authors read and approved the final manuscript.

Acknowledgements

We thank Benjamin Haley, Mike Costa, Amy Heidersbach, Kristel Dorighi, Scott Martin,Rena Yang, Allison Vuong, Oleg Mayba, Sandra Melo Carlos, and Russell Xie for sharing their expertise with us and guiding the development of our software ecosystem. We also thank William Forrest, Hector Corrada Bravo, Michael

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Lawrence, and Benjamin Haley for providing invaluable feedback on the manuscript. Finally, we thank Nitesh Turaga, Lori Shepherd and Marcel Ramos who kindly reviewed our R packages as part of the Bioconductor submission process.

References

Patrick Aboyoun, Herve Pages, and Michael Lawrence. *GenomicRanges: Representation and manipulation of genomic intervals*, 2012.

Omar O Abudayyeh, Jonathan S Gootenberg, Silvana Konermann, Julia Joung, Ian M Slaymaker, David BT Cox, Sergey Shmakov, Kira S Makarova, Ekaterina Semenova, Leonid Minakhin, et al. C2c2 is a single-component programmable rna-guided rna-targeting crispr effector. *Science*, 353(6299), 2016.

Alfred V Aho and Margaret J Corasick. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340, 1975.

Felicity Allen, Fiona Behan, Anton Khodak, Francesco Iorio, Kosuke Yusa, Mathew Garnett, and Leopold Parts. Jacks: joint analysis of crispr/cas9 knockout screens. *Genome research*, 29(3):464–471, 2019.

Andrew V Anzalone, Peyton B Randolph, Jessie R Davis, Alexander A Sousa, Luke W Koblan, Jonathan M Levy, Peter J Chen, Christopher Wilson, Gregory A Newby, Aditya Raguram, et al. Search-and-replace genome editing without double-strand breaks or donor dna. *Nature*, 576(7785):149–157, 2019.

Mandana Arbab, Max W Shen, Beverly Mok, Christopher Wilson, Żaneta Matuszek, Christopher A Cassa, and David R Liu. Determinants of base editing outcomes from target library analysis and machine learning. *Cell*, 182(2):463–480, 2020.

Sangsu Bae, Jeongbin Park, and Jin-Soo Kim. Cas-offinder: a fast and versatile algorithm that searches for potential off-target sites of cas9 rna-guided endonucleases. *Bioinformatics*, 30(10):1473–1475, 2014.

Russell Bainer, Dariusz Ratman, Peter Haverty, and Steve Lianoglou. *gCrisprTools: Suite of Functions for Pooled Crispr Screen QC and Analysis*, 2021. R package version 2.0.0.

Alex Bateman, Lachlan Coin, Richard Durbin, Robert D Finn, Volker Hollich, Sam Griffiths-Jones, Ajay Khanna, Mhairi Marshall, Simon Moxon, Erik LL Sonnhammer, et al. The pfam protein families database. *Nucleic acids research*, 32(suppl_1):D138–D141, 2004.

Aditya M Bhagwat, Johannes Graumann, Rene Wiegandt, Mette Bentsen, Jordan Welker, Carsten Kuenne, Jens Preussner, Thomas Braun, and Mario Looso. multicrispr: grna design for prime editing and parallel targeting of thousands of targets. *Life science alliance*, 3(11), 2020.

Matthew C Canver, Samuel Lessard, Luca Pinello, Yuxuan Wu, Yann Ilboudo, Emily N Stern, Austen J Needleman, Frédéric Galactéros, Carlo Brugnara, Abdullah Kutlar, et al. Variant-aware saturating mutagenesis using multiple cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature genetics*, 49 (4):625, 2017.

Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. Package 'shiny'. *See http://citeseerx. ist. psu. edu/viewdoc/download*, 2015.

Chen-Hao Chen, Tengfei Xiao, Han Xu, Peng Jiang, Clifford A Meyer, Wei Li, Myles Brown, and X Shirley Liu. Improved design and analysis of crispr knockout screens. *Bioinformatics*, 2018.

Wei Chen, Aaron McKenna, Jacob Schreiber, Maximilian Haeussler, Yi Yin, Vikram Agarwal, William Stafford Noble, and Jay Shendure. Massively parallel profiling and predictive modeling of the outcomes of crispr/cas9-mediated double-strand break repair. *Nucleic acids research*, 47(15):7989–8003, 2019.

Sarit Cohen, Lior Kramarski, Shahar Levi, Noa Deshe, Oshrit Ben David, and Eyal Arbely. Nonsense mutation-dependent reinitiation of translation in mammalian cells. *Nucleic acids research*, 47(12):6330–6338, 2019.

Jean-Paul Concordet and Maximilian Haeussler. Crispor: intuitive guide selection for crispr/cas9 genome editing experiments and screens. *Nucleic acids research*, 46(W1):W242–W245, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

David BT Cox, Jonathan S Gootenberg, Omar O Abudayyeh, Brian Franklin, Max J Kellner, Julia Joung, and Feng Zhang. Rna editing with crispr-cas13. *Science*, 358(6366):1019–1027, 2017.

Joshua M Dempster, Isabella Boyle, Francisca Vazquez, David E Root, Jesse S Boehm, William C Hahn, Aviad Tsherniak, and James M McFarland. Chronos: a cell population dynamics model of crispr experiments that improves inference of gene fitness effects. *Genome biology*, 22(1):1–23, 2021.

Peter C DeWeirdt, Kendall R Sanson, Annabel K Sangree, Mudra Hegde, Ruth E Hanna, Marissa N Feeley, Audrey L Griffith, Teng Teng, Samantha M Borys, Christine Strand, et al. Optimization of ascas12a for combinatorial genetic screens in human cells. *Nature Biotechnology*, pages 1–11, 2020.

Sarah Djebali, Carrie A Davis, Angelika Merkel, Alex Dobin, Timo Lassmann, Ali Mortazavi, Andrea Tanzer, Julien Lagarde, Wei Lin, Felix Schlesinger, et al. Landscape of transcription in human cells. *Nature*, 489 (7414):101–108, 2012.

John G Doench, Ella Hartenian, Daniel B Graham, Zuzana Tothova, Mudra Hegde, Ian Smith, Meagan Sullender, Benjamin L Ebert, Ramnik J Xavier, and David E Root. Rational design of highly active sgrnas for crispr-cas9–mediated gene inactivation. *Nature biotechnology*, 32(12):1262, 2014.

John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184, 2016a.

John G Doench, Nicolo Fusi, Meagan Sullender, Mudra Hegde, Emma W Vaimberg, Katherine F Donovan, Ian Smith, Zuzana Tothova, Craig Wilen, Robert Orchard, et al. Optimized sgrna design to maximize activity and minimize off-target effects of crispr-cas9. *Nature biotechnology*, 34(2):184, 2016b.

Steffen Durinck, Yves Moreau, Arek Kasprzyk, Sean Davis, Bart De Moor, Alvis Brazma, and Wolfgang Huber. Biomart and bioconductor: a powerful link between biological databases and microarray data analysis. *Bioinformatics*, 21(16):3439–3440, 2005.

David Feldman, Avtar Singh, Jonathan L Schmid-Burgk, Rebecca J Carlson, Anja Mezger, Anthony J Garrity, Feng Zhang, and Paul C Blainey. Optical pooled screens in human cells. *Cell*, 179(3):787–799, 2019.

Gregory M Findlay, Riza M Daza, Beth Martin, Melissa D Zhang, Anh P Leith, Molly Gasperini, Joseph D Janizek, Xingfan Huang, Lea M Starita, and Jay Shendure. Accurate classification of brca1 variants with saturation genome editing. *Nature*, 562(7726):217–222, 2018.

Jean-Philippe Fortin. *BSgenome.Hsapiens.UCSC.hg38.dbSNP151.major: Full genome sequences for Homo sapiens (UCSC version hg38, based on GRCh38.p12) with injected major alleles (dbSNP151)*, 2021a. R package version 0.0.9999.

Jean-Philippe Fortin. *BSgenome.Hsapiens.UCSC.hg38.dbSNP151.minor: Full genome sequences for Homo sapiens (UCSC version hg38, based on GRCh38.p12) with injected minor alleles (dbSNP151)*, 2021b. R package version 0.0.9999.

Jean-Philippe Fortin and Aaron Lun. *crisprScore: On-Target and Off-Target Scoring Algorithms for CRISPR gRNAs*, 2022. URL `https://github.com/Jfortin1/crisprScore/issues`. R package version 0.99.16.

Jean-Philippe Fortin, Jenille Tan, Karen E Gascoigne, Peter M Haverty, William F Forrest, Michael R Costa, and Scott E Martin. Multiple-gene targeting and mismatch tolerance can confound analysis of genome-wide pooled crispr screens. *Genome biology*, 20(1):21, 2019.

Yanfang Fu, Jennifer A Foden, Cyd Khayter, Morgan L Maeder, Deepak Reyon, J Keith Joung, and Jeffry D Sander. High-frequency off-target mutagenesis induced by crispr-cas nucleases in human cells. *Nature biotechnology*, 31(9):822–826, 2013.

Nicole M Gaudelli, Alexis C Komor, Holly A Rees, Michael S Packer, Ahmed H Badran, David I Bryson, and David R Liu. Programmable base editing of a* t to g* c in genomic dna without dna cleavage. *Nature*, 551 (7681):464–471, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome biology*, 5(10):1–16, 2004.

Luke A Gilbert, Max A Horlbeck, Britt Adamson, Jacqueline E Villalta, Yuwen Chen, Evan H Whitehead, Carla Guimaraes, Barbara Panning, Hidde L Ploegh, Michael C Bassik, Lei S Qi, Martin Kampmann, and Weissman Jonathan S. Genome-scale crispr-mediated control of gene repression and activation. *Cell*, 159(3): 647–661, Oct 2014.

F Hahne, A Lerch, and MB Stadler. Rbowtie: A r wrapper for bowtie and splicemap short read aligners. *URL http://bioconductor. org/packages/release/bioc/html/Rbowtie. html*, 2012.

Kyuho Han, Edwin E Jeng, Gaelen T Hess, David W Morgens, Amy Li, and Michael C Bassik. Synergistic drug combinations for cancer identified in a crispr screen for pairwise genetic interactions. *Nature biotechnology*, 35(5):463–474, 2017.

Ruth E Hanna, Mudra Hegde, Christian R Fagre, Peter C DeWeirdt, Annabel K Sangree, Zsofia Szegletes, Audrey Griffith, Marissa N Feeley, Kendall R Sanson, Yossef Baidi, et al. Massively parallel assessment of human variants with base editor screens. *Cell*, 184(4):1064–1080, 2021.

Traver Hart, Amy Hin Yan Tong, Katie Chan, Jolanda Van Leeuwen, Ashwin Seetharaman, Michael Aregger, Megha Chandrashekhar, Nicole Hustedt, Sahil Seth, Avery Noonan, et al. Evaluation and design of genome-wide crispr/spcas9 knockout screens. *G3: Genes, Genomes, Genetics*, 7(8):2719–2727, 2017.

Wei He, Liang Zhang, Oscar D Villarreal, Rongjie Fu, Ella Bedford, Jingzhuang Dou, Anish Y Patel, Mark T Bedford, Xiaobing Shi, Taiping Chen, et al. De novo identification of essential protein domains from crispr-cas9 tiling-sgrna knockout screens. *Nature communications*, 10(1):1–10, 2019.

Florian Heigwer, Grainne Kerr, and Michael Boutros. E-crisp: fast crispr target site identification. *Nature methods*, 11(2):122–123, 2014.

Florian Heigwer, Tianzuo Zhan, Marco Breinig, Jan Winter, Dirk Brügemann, Svenja Leible, and Michael Boutros. Crispr library designer (cld): software for multispecies design of single guide rna libraries. *Genome biology*, 17(1):1–10, 2016.

Max A Horlbeck, Luke A Gilbert, Jacqueline E Villalta, Britt Adamson, Ryan A Pak, Yuwen Chen, Alexander P Fields, Chong Yon Park, Jacob E Corn, Martin Kampmann, et al. Compact and highly active next-generation libraries for crispr-mediated gene repression and activation. *Elife*, 5:e19760, 2016.

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827, 2013a.

Patrick D Hsu, David A Scott, Joshua A Weinstein, F Ann Ran, Silvana Konermann, Vineeta Agarwala, Yinqing Li, Eli J Fine, Xuebing Wu, Ophir Shalem, et al. Dna targeting specificity of rna-guided cas9 nucleases. *Nature biotechnology*, 31(9):827–832, 2013b.

Johnny H Hu, Shannon M Miller, Maarten H Geurts, Weixin Tang, Liwei Chen, Ning Sun, Christina M Zeina, Xue Gao, Holly A Rees, Zhi Lin, et al. Evolved cas9 variants with broad pam compatibility and high dna specificity. *Nature*, 556(7699):57–63, 2018.

Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with bioconductor. *Nature methods*, 12(2):115–121, 2015.

Katharina Imkeller, Giulia Ambrosi, Michael Boutros, and Wolfgang Huber. gscreend: modelling asymmetric count ratios in crispr screens to decrease experiment size and improve phenotype detection. *Genome biology*, 21(1):1–13, 2020.

Eiru Kim and Traver Hart. Improved analysis of crispr fitness screens and reduced off-target effects with the bagel2 gene essentiality classifier. *Genome medicine*, 13(1):1–11, 2021.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Hui Kwon Kim, Seonwoo Min, Myungjae Song, Soobin Jung, Jae Woo Choi, Younggwang Kim, Sangeun Lee, Sungroh Yoon, and Hyongbum Henry Kim. Deep learning improves prediction of crispr–cpf1 guide rna activity. *Nature biotechnology*, 36(3):239, 2018.

Luke W Koblan, Jordan L Doman, Christopher Wilson, Jonathan M Levy, Tristan Tay, Gregory A Newby, Juan Pablo Maianti, Aditya Raguram, and David R Liu. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nature biotechnology*, 36(9):843–846, 2018.

Alexis C Komor, Yongjoo B Kim, Michael S Packer, John A Zuris, and David R Liu. Programmable editing of a target base in genomic dna without double-stranded dna cleavage. *Nature*, 533(7603):420–424, 2016.

Silvana Konermann, Peter Lotfy, Nicholas J Brideau, Jennifer Oki, Maxim N Shokhirev, and Patrick D Hsu. Transcriptome engineering with rna-targeting type vi-d crispr effectors. *Cell*, 173(3):665–676, 2018.

Anshul Kundaje, Wouter Meuleman, Jason Ernst, Misha Bilenky, Angela Yen, Alireza Heravi-Moussavi, Pouya Kheradpour, Zhizhuo Zhang, Jianrong Wang, Michael J Ziller, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015.

Cem Kuscu, Sevki Arslan, Ritambhara Singh, Jeremy Thorpe, and Mazhar Adli. Genome-wide analysis reveals characteristics of off-target sites bound by the cas9 endonuclease. *Nature biotechnology*, 32(7):677–683, 2014.

Kornel Labun, Tessa G Montague, James A Gagnon, Summer B Thyme, and Eivind Valen. Chopchop v2: a web tool for the next generation of crispr genome engineering. *Nucleic acids research*, 44(W1):W272–W276, 2016.

Kornel Labun, Xiaoge Guo, Alejandro Chavez, George Church, James A Gagnon, and Eivind Valen. Accurate analysis of genuine crispr editing events with amplican. *Genome research*, 29(5):843–847, 2019.

Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome biology*, 10(3):R25, 2009.

Michael Lawrence, Wolfgang Huber, Hervé Pages, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

Samuel Lessard, Laurent Francioli, Jessica Alfoldi, Jean-Claude Tardif, Patrick T Ellinor, Daniel G MacArthur, Guillaume Lettre, Stuart H Orkin, and Matthew C Canver. Human genetic variation alters crispr-cas9 on-and off-targeting specificity at therapeutically implicated loci. *Proceedings of the National Academy of Sciences*, page 201714640, 2017.

Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009.

Wei Li, Johannes Köster, Han Xu, Chen-Hao Chen, Tengfei Xiao, Jun S Liu, Myles Brown, and X Shirley Liu. Quality control, modeling, and visualization of crispr screens with mageck-vispr. *Genome biology*, 16(1):1–13, 2015.

Helen Lindsay, Alexa Burger, Berthin Biyong, Anastasia Felker, Christopher Hess, Jonas Zaugg, Elena Chiavacci, Carolin Anders, Martin Jinek, Christian Mosimann, et al. Crisprvariants charts the mutation spectrum of genome engineering experiments. *Nature biotechnology*, 34(7):701–702, 2016.

S John Liu, Max A Horlbeck, Seung Woo Cho, Harjus S Birk, Martina Malatesta, Daniel He, Frank J Attenello, Jacqueline E Villalta, Min Y Cho, Yuwen Chen, et al. Crispri-based genome-scale identification of functional long noncoding rna loci in human cells. *Science*, 355(6320):eaah7111, 2017.

Aaron Lun. *basilisk: Freezing Python Dependencies Inside Bioconductor Packages*, 2021. R package version 1.3.5.

Aaron McKenna and Jay Shendure. Flashfry: a fast and flexible tool for large-scale crispr target design. *BMC biology*, 16(1):1–6, 2018.

Robin M Meyers, Jordan G Bryan, James M McFarland, Barbara A Weir, Ann E Sizemore, Han Xu, Neekesh V Dharia, Phillip G Montgomery, Glenn S Cowley, Sasha Pantel, et al. Computational correction of copy

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

number effect improves specificity of crispr–cas9 essentiality screens in cancer cells. *Nature genetics*, 49(12): 1779–1784, 2017.

Tessa G Montague, José M Cruz, James A Gagnon, George M Church, and Eivind Valen. Chopchop: a crispr/cas9 and talen web tool for genome editing. *Nucleic acids research*, 42(W1):W401–W407, 2014.

Miguel A Moreno-Mateos, Charles E Vejnar, Jean-Denis Beaudoin, Juan P Fernandez, Emily K Mis, Mustafa K Khokha, and Antonio J Giraldez. Crisprscan: designing highly efficient sgrnas for crispr-cas9 targeting in vivo. *Nature methods*, 12(10):982–988, 2015.

Hiroshi Nishimasu, Xi Shi, Soh Ishiguro, Linyi Gao, Seiichi Hirano, Sae Okazaki, Taichi Noda, Omar O Abudayyeh, Jonathan S Gootenberg, Hideto Mori, et al. Engineered crispr-cas9 nuclease with expanded targeting space. *Science*, 361(6408):1259–1262, 2018.

Hervé Pages, Patrick Aboyoun, Robert Gentleman, and Saikat DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms. *R package version*, 2(0):10–18129, 2016.

Vikram Pattanayak, Steven Lin, John P Guilinger, Enbo Ma, Jennifer A Doudna, and David R Liu. High-throughput profiling of off-target dna cleavage reveals rna-programmed cas9 nuclease specificity. *Nature biotechnology*, 31(9):839–843, 2013.

Alexendar R Perez, Yuri Pritykin, Joana A Vidigal, Sagar Chhangawala, Lee Zamparo, Christina S Leslie, and Andrea Ventura. Guidescan software for improved single and paired crispr guide rna design. *Nature biotechnology*, 35(4):347–349, 2017.

Aliaksandra Radzisheuskaya, Daria Shlyueva, Iris Müller, and Kristian Helin. Optimizing sgrna position markedly improves the efficiency of crispr/dcas9-mediated transcriptional repression. *Nucleic acids research*, 44(18):e141–e141, 2016.

Joseph M Replogle, Thomas M Norman, Albert Xu, Jeffrey A Hussmann, Jin Chen, J Zachery Cogan, Elliott J Meer, Jessica M Terry, Daniel P Riordan, Niranjan Srinivas, et al. Combinatorial single-cell crispr screens by direct guide rna capture and targeted sequencing. *Nature biotechnology*, 38(8):954–961, 2020.

Annabel K Sangree, Audrey L Griffith, Zsofia M Szegletesh, Priyanka Roy, Peter C DeWeirdt, Mudra Hegde, Abby V McGee, Ruth E Hanna, and John G Doench. Benchmarking of spcas9 variants enables deeper base editor screens of brca1 and bcl2. *bioRxiv*, 2021.

Kendall R Sanson, Ruth E Hanna, Mudra Hegde, Katherine F Donovan, Christine Strand, Meagan E Sullender, Emma W Vaimberg, Amy Goodale, David E Root, Federica Piccioni, et al. Optimized libraries for crispr-cas9 genetic screens with multiple modalities. *Nature communications*, 9(1):1–15, 2018.

David A Scott and Feng Zhang. Implications of human genetic variation in crispr-based therapeutic genome editing. *Nature medicine*, 23(9):1095, 2017.

Sergey Shmakov, Omar O Abudayyeh, Kira S Makarova, Yuri I Wolf, Jonathan S Gootenberg, Ekaterina Semenova, Leonid Minakhin, Julia Joung, Silvana Konermann, Konstantin Severinov, et al. Discovery and functional characterization of diverse class 2 crispr-cas systems. *Molecular cell*, 60(3):385–397, 2015.

Arne H Smits, Frederik Ziebell, Gerard Joberty, Nico Zinn, William F Mueller, Sandra Clauder-Münster, Dirk Eberhard, Maria Fälth Savitski, Paola Grandi, Petra Jakob, et al. Biological plasticity rescues target activity in crispr knock outs. *Nature methods*, 16(11):1087–1093, 2019.

Manuel Stemmer, Thomas Thumberger, Maria del Sol Keyer, Joachim Wittbrodt, and Juan L Mateo. Cctop: an intuitive, flexible and reliable crispr/cas9 target prediction tool. *PloS one*, 10(4):e0124633, 2015.

Summer B Thyme, Laila Akhmetova, Tessa G Montague, Eivind Valen, and Alexander F Schier. Internal guide rna interactions interfere with cas9-mediated cleavage. *Nature communications*, 7:11750, 2016.

Brendan Veeneman, Ying Gao, Joy Grant, David Fruhling, James Ahn, Benedikt Bosbach, Jadwiga Bienkowska, Maximillian Follettie, Kim Arndt, Jeremy Myers, and Wenyan Zhong. Pincer: improved crispr/cas9 screening by efficient cleavage at conserved residues. *Nucleic acids research*, 48(17):9462–9477, 2020.

Russell T Walton, Kathleen A Christie, Madelynn N Whittaker, and Benjamin P Kleinstiver. Unconstrained genome targeting with near-pamless engineered crispr-cas9 variants. *Science*, 368(6488):290–296, 2020.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

Binbin Wang, Mei Wang, Wubing Zhang, Tengfei Xiao, Chen-Hao Chen, Alexander Wu, Feizhen Wu, Nicole Traugh, Xiaoqing Wang, Ziyi Li, et al. Integrative analysis of pooled crispr genetic screens using mageckflute. *Nature protocols*, 14(3):756–780, 2019a.

Daqi Wang, Chengdong Zhang, Bei Wang, Bin Li, Qiang Wang, Dong Liu, Hongyan Wang, Yan Zhou, Leming Shi, Feng Lan, et al. Optimized crispr guide rna design for two high-fidelity cas9 variants by deep learning. *Nature communications*, 10(1):1–14, 2019b.

Guanqun Wang, Meijie Du, Jianbin Wang, and Ting F Zhu. Genetic variation may confound analysis of crispr-cas9 off-target mutations. *Cell discovery*, 4(1):18, 2018.

Tim Wang, Jenny J Wei, David M Sabatini, and Eric S Lander. Genetic screens in human cells using the crispr/cas9 system. *Science*, 343(6166):80–84, Jan 2014.

Hans-Hermann Wessels, Alejandro Méndez-Mancilla, Xinyi Guo, Mateusz Legut, Zharko Daniloski, and Neville E Sanjana. Massively parallel cas13 screens reveal principles for guide rna design. *Nature biotechnology*, 38(6):722–727, 2020.

Xuebing Wu, David A Scott, Andrea J Kriz, Anthony C Chiu, Patrick D Hsu, Daniel B Dadon, Albert W Cheng, Alexandro E Trevino, Silvana Konermann, Sidi Chen, et al. Genome-wide binding of the crispr endonuclease cas9 in mammalian cells. *Nature biotechnology*, 32(7):670–676, 2014.

Lihua J Zhu, Benjamin R Holmes, Neil Aronin, and Michael H Brodsky. Crisprseek: a bioconductor package to identify target-specific guide rnas for crispr-cas9 genome-editing systems. *PloS one*, 9(9):e108424, 2014.

Lihua Julie Zhu, Michael Lawrence, Ankit Gupta, Alper Kucukural, Manuel Garber, Scot A Wolfe, et al. Guideseq: a bioconductor package to analyze guide-seq datasets for crispr-cas nucleases. *BMC genomics*, 18 (1):1–10, 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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.

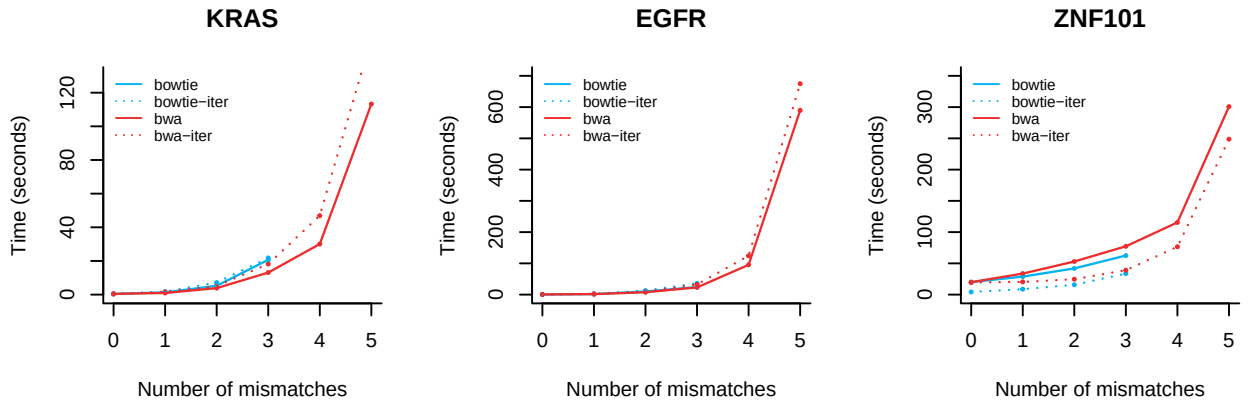
Supplementary Tables and Figures

Name	Type	Description
crisprDesign	Software	Comprehensive one-stop shop for gRNA design
crisprBase	Software	Nuclease specification and gRNA arithmetics
crisprBowtie	Software	gRNA spacer alignment with *bowtie*
crisprBwa	Software	gRNA spacer alignment with *BWA*
crisprScore	Software	On- and off-target scoring algorithms for gRNAs
crisprScoreData	Data	Pre-trained machine learning models for *crisprScore*
Rbwa	Software	R wrapper for *BWA* aligner

Supplementary Table S1. Table of R packages available in our gRNA design ecosystem.

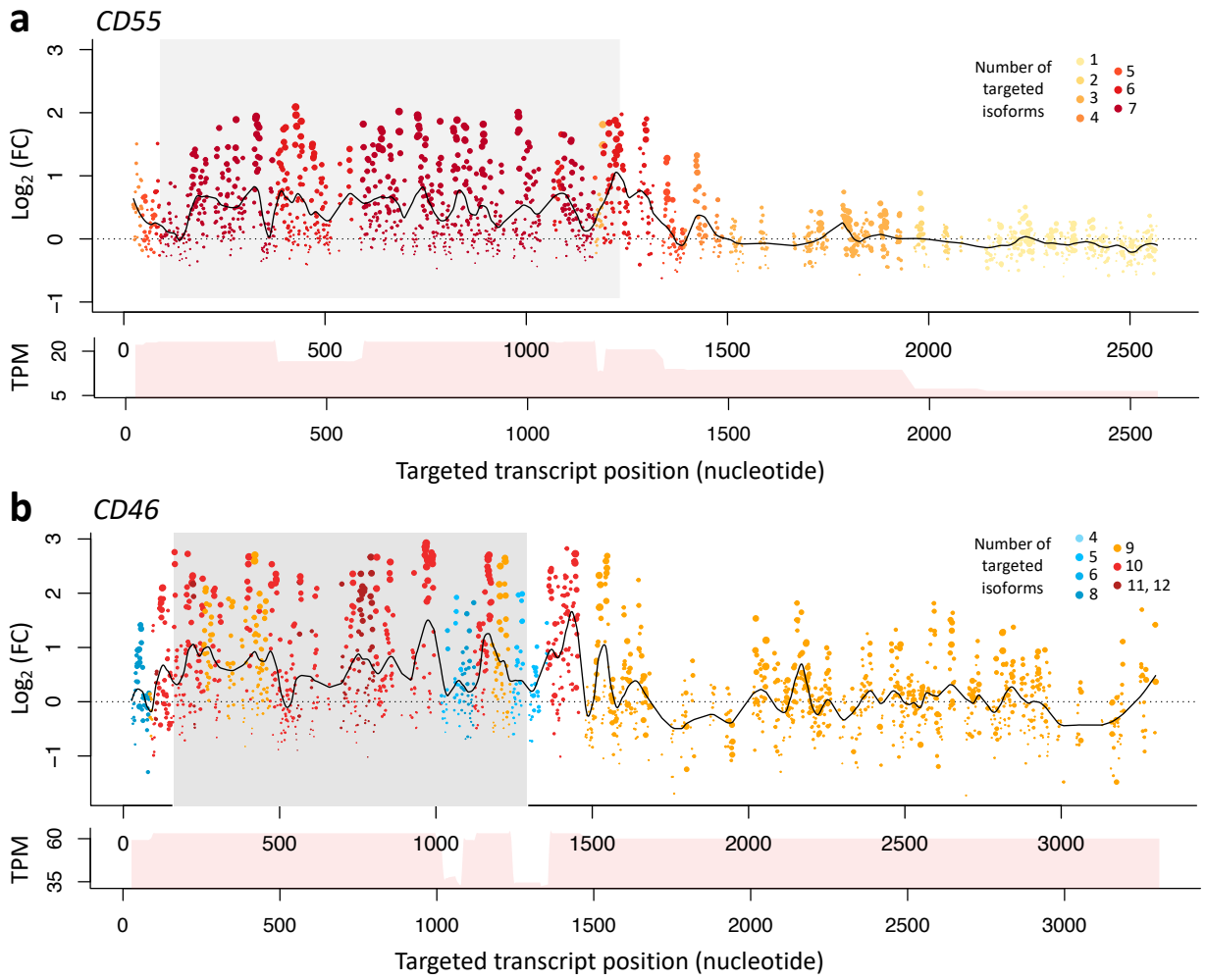
Target	Spacer specification			Number of mismatches											
Gene	Coordinate	Spacer	PAM	0			1			2			3		
				AC	BO	BW	AC	BO	BW	AC	BO	BW	AC	BO	BW
CFTR	chr7:117559595	ATTAAAGAAAATATCATCTT	TGG	1	1	1	7	7	7	145	145	145	2,314	2,314	2,314
	chr7:117559605	TCTGTATCTATATTCATCAT	AGG	1	1	1	7	7	7	125	125	125	1,704	1,704	1,704
HBB	chr11:5227002	CATGGTGCATCTGACTCCTG	AGG	2	2	2	0	0	0	14	14	14	210	210	210
	chr11:5227004	GTAACGGCAGACTTCTCCTC	AGG	1	1	1	0	0	0	7	7	7	83	83	83
HEXA	chr15:72346571	TGTAGAAATCCTTCCAGTCA	GGG	1	1	1	0	0	0	25	25	25	298	298	298
	chr15:72346578	ATCCTTCCAGTCAGGGCCAT	AGG	1	1	1	0	0	0	6	6	6	203	203	203
PRNP	chr20:4699588	AGCAGCTGGGGCAGTGGTGG	GGG	1	1	1	2	2	2	96	96	96	909	909	909
	chr20:4699589	GCAGCTGGGGCAGTGGTGGG	GGG	1	1	1	12	12	12	100	100	100	1,052	1,052	1,052
	chr20:4699595	GGGGCAGTGGTGGGGGGCCT	TGG	1	1	1	2	2	2	56	56	56	860	860	860
	chr20:4699598	GCAGTGGTGGGGGGCCTTGG	CGG	1	1	1	0	0	0	32	32	32	421	421	421

Supplementary Table S2. Table of on- and off-target alignments in the GRCh38.p13 for the 10 spacer sequences reported in Bhagwat et al. [2020] using a PAM-agnostic approach. Number of mismatches between 0 and 3 were considered for 3 different aligners: Aho-Corasick exact string matching as implemented in *Biostrings* (AC), *bowtie* aligner via the *crisprBowtie* package (BO), and *BWA* aligner via the *crisprBwa* package (BW). All 3 alignment methods agree.



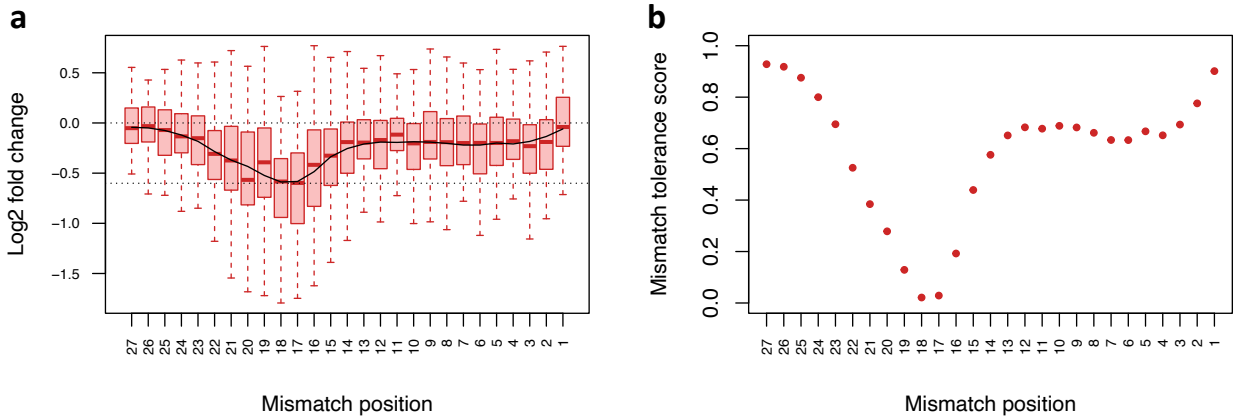
Supplementary Figure S1. Comparison of the off-target annotation methods implemented in *crisprDesign* We compare the four different annotation methods available via the *addSpacerAligmments* function in *crisprDesign*: *bowtie*, via the *crisprBowtie* package (`bowtie`), *BWA*, via the *crisprBwa* package (`bwa`), and an iterative version of both algorithms to diminish the impact of highly non-specific gRNAs (`bowtie-iter` and `bwa-iter`).

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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 Figure S2. CasRx tiling screens of *CD55* and *CD46*. Pooled FACS tiling screening data of genes *CD55* and *CD46* performed in HEK 294 cells using CasRx (*Rfx*Cas13d). Processed and normalized $\log_2$ fold changes were obtained from Wessels et al. [2020]. Both screens are represented using the canonical Ensembl isoforms. We remapped and reannotated all gRNA sequences using *crisprDesign*; isoform annotation, on-target activity score using CasRx-RF as implemented in *crisprScore*, and off-target alignments were added to each gRNA. The color of the dots indicates the number of isoforms targeted by each gRNA. The size of the dots is proportional to the on-target activity score. The coding sequence (CDS) is highlighted in grey. LOESS regression curves are shown as solid lines. For both genes, transcript per million (TPM) counts in HEK 293 cells summed across all isoforms overlapping a given nucleotide position are shown below the log-fold change panels.

bioRxiv preprint doi: https://doi.org/10.1101/2022.04.21.488824; this version posted April 28, 2022. 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 Figure S3. Probability weights used for off-target scoring of CasRx gRNAs a Boxplots of the differences in log2 fold change (ΔLFC) between single-mismatch (SM) gRNAs and their corresponding perfect-match (PM) gRNAs in the GFP tiling screen. X-axis represents the mismatch position within the spacer sequence, with 1 being the position next to the direct repeat. The smooth curve was obtained using LOESS regression. The dotted line represents the average log-fold change of all PM gRNAs after multiplying by -1. **b** CasRx mismatch tolerance probabilities estimated from (a) and used in the CFD scoring method in *crisprScore*.