

crisprSQL: a novel database platform for CRISPR/Cas off-target cleavage assays

Florian Störtz* and Peter Minary

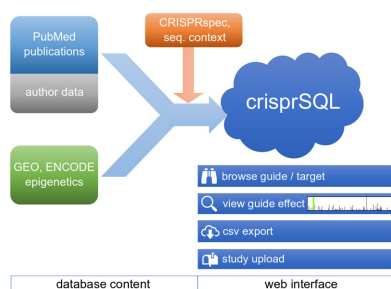
Department of Computer Science, University of Oxford, Parks Road, Oxford OX1 3QD, UK

Received August 14, 2020; Revised September 23, 2020; Editorial Decision September 26, 2020; Accepted October 17, 2020

ABSTRACT

With ongoing development of the CRISPR/Cas programmable nuclease system, applications in the area of *in vivo* therapeutic gene editing are increasingly within reach. However, non-negligible off-target effects remain a major concern for clinical applications. Even though a multitude of off-target cleavage datasets have been published, a comprehensive, transparent overview tool has not yet been established. Here, we present **crisprSQL** (<http://www.crisprsql.com>), an interactive and bioinformatically enhanced collection of CRISPR/Cas9 off-target cleavage studies aimed at enriching the fields of cleavage profiling, gene editing safety analysis and transcriptomics. The current version of **crisprSQL** contains cleavage data from 144 guide RNAs on 25,632 guide-target pairs from human and rodent cell lines, with interaction-specific references to epigenetic markers and gene names. The first curated database of this standard, it promises to enhance safety quantification research, inform experiment design and fuel development of computational off-target prediction algorithms.

GRAPHICAL ABSTRACT



INTRODUCTION

Several clinical applications of the CRISPR/Cas gene editing technology have been realised to date, such as person-

alised cancer treatment with genetically modified chimeric antigen receptor (CAR) T cell-based therapies (1). A variety of possible therapeutic applications in the fields of genetic diseases, infectious diseases and cancers are being investigated (2), as exemplified by 38 clinical studies referring to CRISPR currently registered on <http://clinicaltrials.gov> (accessed 7 August 2020). As Dai *et al.* state, 'one of the major hurdles to the clinical translation of CRISPR/Cas9 is its off-target effects' (3), referring to the unwanted nucleotide insertions or deletions created by CRISPR outside of the targeted genomic locus. Electing an appropriate guide RNA (gRNA) for the desired DNA target sequence is considered a 'crucial first step in avoiding off-target effects' (4). In order to mitigate off-target effects, 'a set of validated off-target sites should be compiled and used to create a reagent set to quantitatively study off-target effects via NGS methods' (5). With off-target data spread across supplementary data of various publications, this remains a tedious and time-consuming task.

Besides the design and execution of medical studies, a comprehensive database of well-known off-target interactions is required for the validation of novel off-target detection methods (6) and new implementations of existing methods in a laboratory setting.

The development of computational off-target prediction algorithms (7,8) also relies on the presence of a comprehensive dataset quantifying off-target effects of gRNAs on DNA from certain cell lines, which is usually annotated with a selection of epigenetic markers by the respective authors. The varying choice of annotation sources, training data and test procedures can result in issues of reproducibility (9).

We contribute an online database which promises to enhance clinical, laboratory and computational research by providing the first one-stop source of off-target resolved CRISPR/Cas cleavage data to date, accompanied by epigenetic annotations and visualisations. In order to provide maximum value for a range of fields including transcriptomics, gene knockout experiments, editing safety-driven guide design and cleavage efficiency prediction, **crisprSQL** offers sequence-resolved data whilst also bridging the gap to gene resolution by attaching GENCODE gene names to interaction targets.

*To whom correspondence should be addressed. Tel: +44 1865 273838; Email: florian.stortz@cs.ox.ac.uk

MATERIALS AND METHODS

Data acquisition

Cleavage rate sources & formats. There have been numerous publications scrutinising the off-target effects of *Streptococcus pyogenes* CRISPR/Cas9 over the past decade (see Table 1), most of which rely on second-generation whole-genome sequencing tools. First-generation approaches such as the T7 endonuclease assay only find off-targets among a previously *in silico* defined set, therefore only yielding limited insight into the full off-target profile.

In order to provide the most detailed insight into cleavage processes, we have chosen to include only those studies which offer base pair-resolved on- and off-target, as well as an unambiguous assignment of which specific guide caused a given cleavage event. This excludes pure on-target studies, pure gene knock-out studies and most studies done on whole gRNA libraries. To our best knowledge, this is an objective which has not been realised before. The use of ENCODE Tier 1 immortalised cell lines (10) facilitates comparison between experiments and annotation with cell-specific metadata.

We manually extracted the cleavage frequencies from the references in Table 1 using an online PDF-to-csv tool wherever these were not supplied in text form. The included publication (23) chose identical target sequences in both open and closed chromatin regions, thereby promising very telling data on the effect of chromatin accessibility. Publication (28) which is included in the CRISPOR dataset (27) was excluded here because extraction of the relevant data (gRNA, target locus or sequence, binding & cleavage frequency) was not unambiguously possible. For some studies, we additionally obtained unpublished data points from the same measurement series which authors were happy to share on request.

The off-target cleavage assays included in crisprSQL can be divided into two categories. DSB capture approaches in a cellular context (*in situ*) are more physiologically relevant, albeit prone to false positives due to cellular processes which may generate non-Cas9 related DSBs. IDLV (16) and GUIDE-seq (6) rely on the non-homologous end joining-mediated integration of sequence markers into sites upon DSB. IDLV integration does not always happen at the exact DSB location, whilst GUIDE-seq is limited to blunt-end DSBs (29). Typical sensitivities range from 0.1% (6,15) to 1% (16). Other *in situ* techniques include HTGTS (15) (translocation of DSBs to bait DSBs) and BLESS (14) (integration of biotinylated linkers).

In vitro approaches outside of a cellular context do not include effects of epigenetic factors and chromatin on DSB formation, making the subsequent validation of observed off-targets in cells desirable. SITE-Seq (22) has been shown to overestimate off-targets relative to those observed in a cellular context. CIRCLE-seq (20) shows high sensitivity through enrichment of Cas9-cleaved genomic DNA, which leads to a considerable level of background noise overshadowing true positive off-targets (29). DIG-seq (23) uses chromatin-associated DNA and can therefore assess the influence of chromatin on Cas9 cleavage. Experimental protocols for the different assays entail different relative concentrations of Cas9 expression plasmid (100 ng–1 µg), gRNA-

encoding plasmid (50 ng–1 µg), and possible additions such as blasticidin resistance-inducing plasmids. The ratio of Cas9 plasmid : gRNA plasmid ranges from 1:1 to 3:1, and the time between Cas9 transfection and cell harvest for DSB detection ranges from 24 h (14) to 5 days (24).

With the exception of the NucleaSeq assay (26), crisprSQL only includes data points which have been validated in cells. For *in vitro* methods, this is a certain fraction of DSB sites identified by the respective assay, ranging from 62% (24) to 100% (13,22). For these methods, validation data in cellular context has been gained through targeted deep sequencing.

It has become convention to give guide and target sequences as their respective complement in the protospacer strand such that a canonically matched guide-target pair shows as identical base letters. To extract the actual binding sequences from our database, one therefore has to exchange thymine for uracil in the saved gRNA sequence, and take the canonical complement of the saved target sequence.

Data processing. The resulting csv file was imported in Python using pandas. Missing genomic locus information was filled in using the blastn tool. For sequences which were shorter than 23 bp, samtools faidx was used to extend them to this length, making sure the found sequence contains at least the 10 PAM-proximal base pairs of the original, has no uncalled bases (N) and an identical PAM. It has been reported that sequence context plays a significant role for cleavage efficiency (27,30). We therefore provide the 169 bp sequence context around the centre of each target obtained using samtools faidx. This length has been chosen such that the nucleosome core DNA length of 147 bp can be retrieved around any of the 23 bp of each target. To annotate targets with the matching gene name, we retrieve the gff3 files for the respective genomes from the GENCODE database (31), extract the gene annotations and convert them to bed format. Overlap is then checked for each target sequence.

Delivery method. In order to be able to scrutinise the effect of the delivery method of the CRISPR machinery into cells, the database contains a column characterising the mode of delivery (lipofection/electroporation/other), and a column indicating whether a given data point was gained in a cellular context via an unbiased (genome-wide) detection method.

Epigenetic factors. Binding and cleavage as natural events do not only depend on sequence information as acquired in the first step, but also on further atomistic parameters of the guide–target heteroduplex. These epigenetic factors have been obtained experimentally using a variety of techniques. The ENCODE database (32) offers a platform which holds this experimental data and allows to search it. SCREEN is a search tool for the ENCODE database which summarises the presence of certain epigenetic factors / candidate regulatory elements along the human (hg19) and mouse (mm10) genome, for several cell lines or tissues. We obtained bed files from the SCREEN web interface which contain genomic regions in which epigenetic markers are present for four individual markers. This choice has become customary

Table 1. Data sources included in crisprSQL (version 10 July 2020), chronologically ordered by publication date

Ref.	Technique	Detects	Guides	Targets	Annot.	Assembly	Cell lines
(11)	T7E1	Heterodupl. DNA	9	88	0	hg19	U2OS, HEK293, K562
(12)	Targeted PCR	Genome change	10	116	11	hg19	K562
(6)	Guide-seq	DSBs	12	575	381	hg19	U2OS, HEK293
(13)	Digenome-seq	Genome change	2	162	7	hg19	HAPI, K562
(14)	BLESS, ChIP	DSBs	6	87	53	hg19, mm9	293FT, N2a
(15)	HTGTS	DNA junctions	3	87	87	hg19	HEK293T
(16)	IDLVs	Viral integration	1	13	13	hg19	HEK293T
(17)	Digenome-seq	Genome change	10	258	234	hg19	HeLa
(18)	Guide-seq	DSBs	7	61	28	hg19	U2OS
(19)	BLESS	DSBs	3	31	31		HEK
(20)	CIRCLE-seq	DSBs	18	7374	3493	hg19	HEK293, U2OS, K562
(21)	Guide-seq	DSBs	5	203	107	hg19	U2OS
(22)	SITE-Seq	DSBs	8	1630	210	hg38	HEK293
(23)	DIG-seq	Genome change	7	141	132	hg19	HeLa
(24)	Guide-seq, WGS	DSBs	31	426	0	mm9, rn5	Mouse & rat embryos
(25)	Guide-seq	DSBs	10	272	160	hg19	U2OS
(26)	NucleaSeq	Cleaved products	2	14108	0		Custom DNA library
		Sum	144	25632	4947		

With 25,632 guide-target pairs, our dataset is more than an order of magnitude larger than both the CRISPOR (27) and DeepCRISPR (7) off-target datasets with less than 680 guide-target pairs each. Note that we have excluded guides with less than two reported off-targets, but kept GC-rich guides and guide-target pairs with low cleavage frequencies. The ‘annotated’ column counts how many of the respective guide-target pairs are annotated with at least one epigenetic marker. The term ‘target’ includes both on-targets, i.e. the intended cutting site which is homologous to the guide sequence, and mismatched off-targets.

in the field of computational off-target prediction (7,33,34): (a) CTCF: CCCTC-binding factor, present at sites which activate gene transcription and are related to chromatin organisation (35), (b) DNase-seq: sites sensitive to the DNase I enzyme correspond to open chromatin (i.e. accessibility of the DNA strand), (c) RRBS: DNA methylation state, obtained by converting unmethylated cytosines to uracil, (d) H3K4me3: trimethylation of the lysine-4 on the histone H3 protein, associated with transcription of nearby genes (36). A full list of the assay files which we used can be found in Supplementary Table S2. We only considered the regions where the marker value is in the 95th percentile across the genome by choosing a cutoff of $Z \geq 1.64$ on the SCREEN web page. Genomic regions present in our database which overlap with a marked region in one of these bed files are annotated with the respective Z score, as well as the respective SCREEN accession string to ensure traceability. Overlap is assessed using the `bedtools intersect` command.

Since SCREEN only offers this data for the hg19 and mm10 gene assemblies and a selection of cell lines, data for other reference genomes was obtained from the ENCODE database itself. The fifth epigenetic marker came from the DRIP-seq assay (37), which detects R-loops between DNA and RNA formed during transcription, exposing single-stranded DNA and thereby possibly causing genome instability. Data files for the DRIP assay had to be converted from bigWig to bed using the BEDOPS suite (38). We downloaded further data gained through the DRIP assay in bed format from the publications (37,39–41) via the GEO database (42). Target loci for which an epigenetic overlap has been found using BamTools (43) were updated using the quality score of the respective bam file region(s) and the unique ENCODE accession string of the bam file. Where two or more overlaps were found for a given target region, only the maximum quality score was kept.

We should note that not all epigenetic markers are available for every cell type and genome assembly (see Table 2). As a workaround, sequences could be realigned to other genome assemblies within species (i.e. hg38 to hg19, mm9 to mm10). As this might introduce impurity into the data, it has not been implemented. However, with the data from Table 2, we are able to check annotation for more than 97% of the off-target loci with at least two of five epigenetic markers.

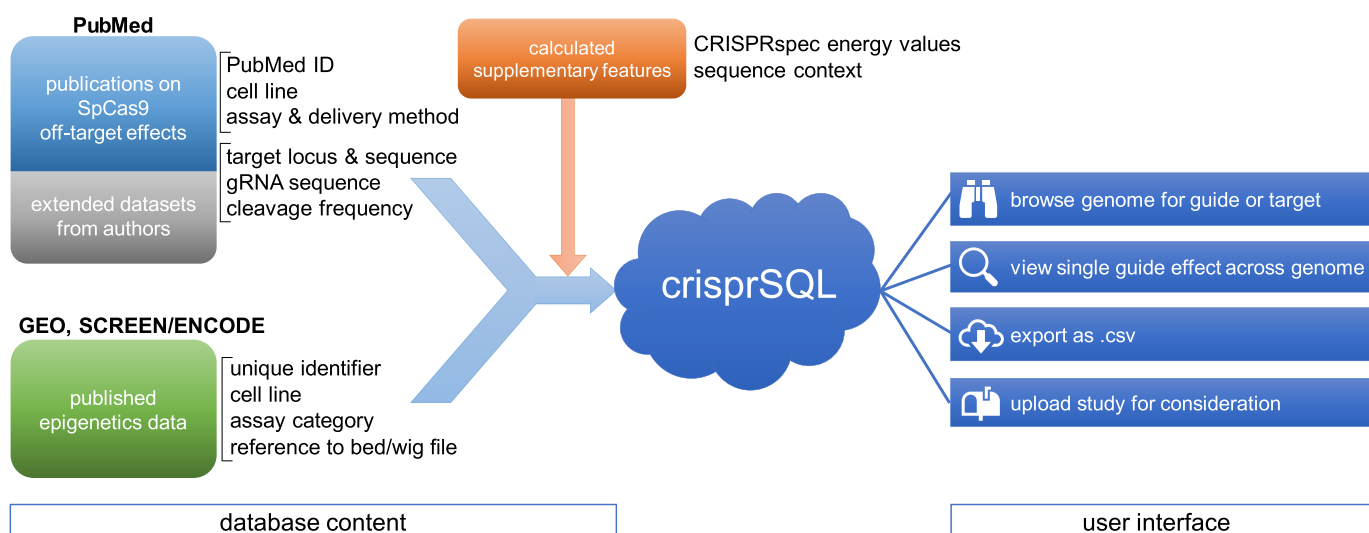
Interaction energies. In reference (44), the authors present an approximate binding energy model (termed CRISPRspec) for the Cas9–gRNA–DNA complex, including contributions from the gRNA–DNA hybridisation and the opening of the DNA–DNA duplex in the target region. RNAfold (45) is used to calculate the approximate free energy of the gRNA folding based on sequence information. We calculate a representative choice of five energy contributions for each gRNA–target sequence pair and annotate the respective data point. The function arguments used to retrieve these can be found in Supplementary Table S1.

Data storage

In order to achieve high enough read and write speeds, we use a Python-implemented SQLite database in local memory for data storage and augmentation. The database is internally separated into three data tables which contain information about the included cleavage assay publications, epigenetics studies datasets and measured guide-target cleavage data points, respectively. This ensures that we are able to save all necessary information in a structured and efficient way which allows full traceability (see Figure 1). From this dataset, we are able to extract convenient views, e.g. in csv format for import into third party

Table 2. Availability of epigenetic markers by cell line in the ENCODE, SCREEN (32) and GEO (42) databases

Assembly	Cell line	Epigenetic marker				
		DNase	CTCF	H3K4me3	RRBS	DRIP
hg19	HEK293	✓	✓	✓	✓	✓
	K562	✓	✓	✓	✓	✓
	HeLa	✓	✓	✓	✓	✓
	U2OS	✓	✓	✓	✓	✓
	HAP1	✓	✓	✓	✓	✓
hg38	HEK293	✓	✓	✓		
	K562	✓	✓	✓		
	HeLa	✓	✓	✓		
	U2OS	✓				
	HAP1	✓				
mm10	Embryonic tissue	✓		✓		
mm9	N2a					
rn5	Embryonic tissue					

**Figure 1.** Overview of the crisprSQL database. Cleavage and epigenetics data (left blue/green fields) are joined together and supplemented by attributes calculated from their properties (top orange field). Additional cleavage data obtained from the authors was included where possible. Brackets indicate single data tables. This forms the crisprSQL database, which has an online user interface supporting data browsing, export and upload of new studies (right).

software, or in MySQL format to provide a website interface.

In conjunction with external tools, our database can be used to suggest optimal guides for cleavage at a certain locus whilst minimising off-target effects.

Website interface

Implementation. In order to make our database easily accessible, we implemented a web-interface using a MySQL-based instance of our database. The frontend and search function are rendered using PHP.

An appropriate subset of database columns was chosen for this so as to not overload the user interface. Columns describing the CRISPRspec energy contributions and the sequence context are not included in the web interface, only in the csv file. Furthermore, the website shows only the presence or absence of a nonzero epigenetic score but not its value. The web interface offers a full-text search for specific guides, target regions or cell lines, and a download of the

full database in csv format. Similar services already exist for high-throughput on-target studies (46), but to our best knowledge, crisprSQL is the first online database for Cas9 off-target assays.

Through use of a state-of-the-art front-end framework, the website is fully functional on phone and tablet screens as well. An overview of the included off-target studies shows the respective detection assays, number of involved guides and targets as well as the gene names which have been predominantly cleaved (Figure 2). It is possible to browse through all included guide sequences grouped by sequence and cell line (Figure 3) as well as search for guide or target sequences, GENCODE gene names or loci. Search results are subsequently shown in a table linking to the original publications, to a genome browser showing the vicinity of the cut site as well as to studies which have demonstrated epigenetic markers at the respective DNA loci. Hyperlinks to all involved study publications and epigenetic data repositories ensure transparency and precise traceability of the included data. The search result is visualised as a barcode-

Studies

The following studies are included in crisprSQL:

No.	study	assay	cell line	total guides	total targets	epigenetically annotated targets	cleaved gene IDs (CF > 1%, on-targets in bold)
1	Kim	Digenome-seq	HAP1	2	162	7	IGDCC3, KRT42P, CITF22-24E5.1, RP11-638F5.1
2	Cho	targeted PCR	K562	10	116	11	CCR5, RP11-24F11.2 , CCR2
3	Ran	BLESS	HEK293	4	53	53	RP11-466M21.1, EMX1
4	Ran	BLESS	N2A	2	34	0	TET2
5	Tsai	Guide-seq	HEK293	4	153	153	KIAA1024, CDH4, CFP89, RP11-115D19.1, RP11-399K21.6

Figure 2. Overview of the included studies, together with metadata such as assay type, the total number of guides and guide-target pairs observed as well as the predominantly cleaved gene names. Studies have been split up according to cell lines.

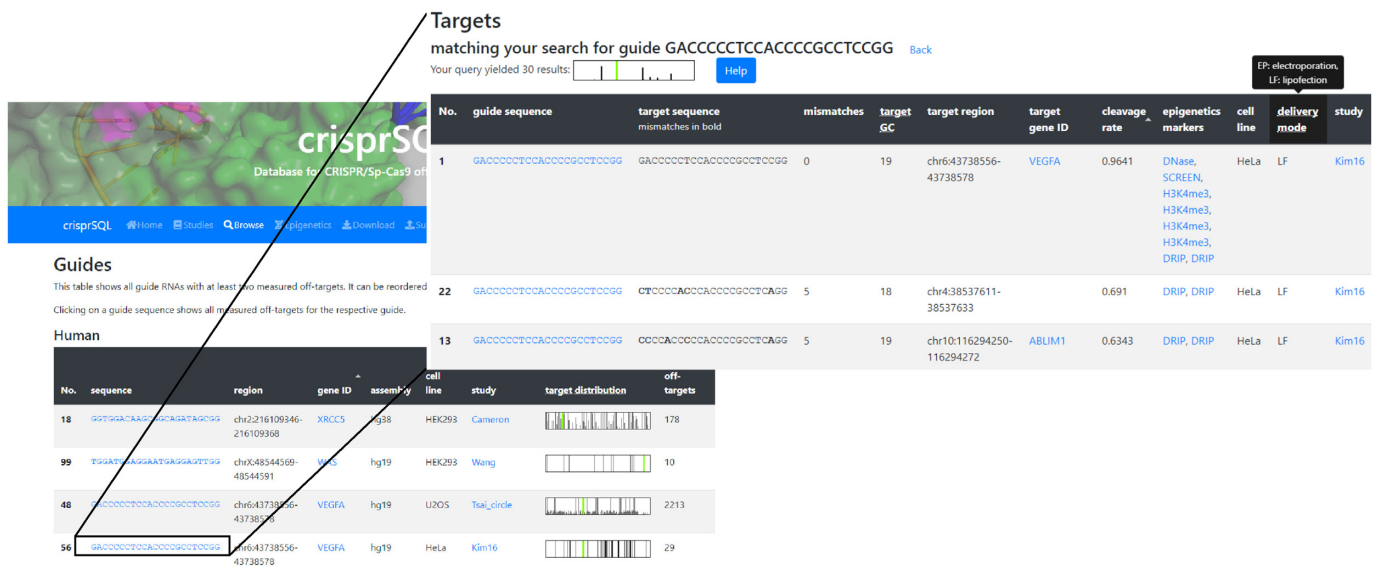


Figure 3. Web interface of the database, showing a list of all included gRNA sequences and which, if any, GENCODE gene name they target. Their respective target distributions across the genome are visualised as a barcode-type histogram, allowing a first assessment of their reported specificity. Upon clicking on or searching for a specific guide (inset), the website shows its full reported off-target profile, i.e. gene names of target loci, cleavage rates and epigenetic markers at the target site, hyperlinked to the respective source. Clicking on a gene name opens the vicinity of the respective cut site in a genome browser.

type histogram plot depicting the locations and observed cleavage rates of found targets across the genome. A similar barcode is given for the editing profile of each guide, allowing a first insight into its specificity. The database therefore lends itself as a powerful in-depth literature research tool for planning off-target studies, such as comparing known off-target performances of specific guides or finding guides whose off-target effects include a specific gene or locus.

The website also allows submission of off-target study results as text files containing guide-target pairs, cell line, measured cleavage rates and publication reference which will then be processed and integrated into the database. In this way, we hope to contribute to data visibility and accessibility in the genome engineering field, as well as model com-

parability and reproducibility in the area of off-target prediction algorithms.

RESULTS AND DISCUSSION

We have established a data processing pipeline and database tool which fill the niche of a thoroughly curated and annotated collection of base-pair resolved CRISPR/Cas9 off-target cleavage assay data. Besides providing a comprehensive search tool for gRNA design on both base pair and gene level, it can act as a data source for both the experimental validation of off-target detection pipelines and for computational off-target cleavage prediction algorithms. This will support experimental research into gRNA design for a variety of applications in the wider gene editing and transcrip-

omics fields, as well as enhance transparency and reproducibility of computational studies.

As noted above, experimental protocols for the different assays differ considerably in the concentration ratios of Cas9- and gRNA-encoding plasmids as well as the timespan between cell transfection and harvest. GUIDE-seq (6), Digenome-seq (13) and CIRCLE-seq (20) require a considerably more intricate computational pipeline than other methods to extract enriched genomic intervals after alignment to a reference genome and check for nuclease-induced edit sites. These represent specific assumptions and normalisation steps, creating the need to further investigate when absolute cleavage frequency values are compared between studies, and to at least normalise the cleavage rate distribution for each single study with a monotone, nonlinear function when relative rankings of cleavage frequencies are desired. An example of such a normalisation is shown in Supplementary Figure S1.

The at times incomplete validation of *in vitro* off-target effects in cells further complicates the comparison of cleavage sites between studies. For a direct comparison of sets of cleavage assays regarding their relative performance, we refer to references (17,20,23).

Future developments

We invite submissions of appropriately resolved cleavage frequency data by experimental authors, which will be run through our annotation pipeline and included in the website in regular intervals. Another promising addition will be to include MNase-seq data (47) as an epigenetic marker gained on the respective cell lines in order to quantify nucleosome occupancy, which has been shown to correlate with cleavage frequency (48).

We further envision to extend the database by appropriately annotated studies targeting high-fidelity Cas9 nuclease variants (18) and nucleases from different organisms (14), as well as Cas9 off-target studies on different vertebrate organisms.

In order to provide a one-stop experience for off-target effects, we envision the inclusion of a state-of-the-art cleavage prediction algorithm which can provide predicted off-target effects next to measured off-target effects for a given gRNA.

DATA AVAILABILITY

The crisprSQL database is available at <http://www.crisprsql.com>, where the full data set can be downloaded in csv format. Users are not required to log in to access any of the database features.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

Biotechnology and Biological Sciences Research Council [BB/M011224/1, BB/S07593/1]. Funding for open access charge: Oxford University RCUK Open Access Block Grant.

Conflict of interest statement. None declared.

REFERENCES

- Fellmann,C., Gowen,B.G., Lin,P.C., Doudna,J.A. and Corn,J.E. (2017) Cornerstones of CRISPR-Cas in drug discovery and therapy. *Nat. Rev. Drug Discov.*, **16**, 89–100.
- Sun,J., Wang,J., Zheng,D. and Hu,X. (2020) Advances in therapeutic application of CRISPR-Cas9. *Brief. Funct. Genomics*, **19**, 164–174.
- Dai,W.J., Zhu,L.Y., Yan,Z.Y., Xu,Y., Wang,Q.L. and Lu,X.J. (2016) CRISPR-Cas9 for *in vivo* gene therapy: promise and hurdles. *Mol. Ther. - Nucleic Acids*, **5**, e349.
- Han,H.A., Pang,J.K.S. and Soh,B.S. (2020) Mitigating off-target effects in CRISPR/Cas9-mediated *in vivo* gene editing. *J. Mol. Med.*, **98**, 615–632.
- Vakulskas,C.A. and Behlke,M.A. (2019) Evaluation and reduction of CRISPR Off-Target cleavage events. *Nucleic Acid Ther.*, **29**, 167–174.
- Tsai,S.Q., Zheng,Z., Nguyen,N.T., Liebers,M., Topkar,V.V., Thapar,V., Wyvekens,N., Khayter,C., Iafate,A.J., Le,L.P. *et al.* (2015) GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.*, **33**, 187–198.
- Chuai,G., Ma,H., Yan,J., Chen,M., Hong,N., Xue,D., Zhou,C., Zhu,C., Chen,K., Duan,B. *et al.* (2018) DeepCRISPR: Optimized CRISPR guide RNA design by deep learning. *Genome Biol.*, **19**, 80.
- Liu,Q., Cheng,X., Liu,G., Li,B. and Liu,X. (2020) Deep learning improves the ability of sgRNA off-target propensity prediction. *BMC Bioinformatics*, **21**, 51.
- Gao,Y., Chuai,G., Yu,W., Qu,S. and Liu,Q. (2020) Data imbalance in CRISPR off-target prediction. *Brief. Bioinform.*, **21**, 1448–1454.
- Dunham,I., Kundaje,A., Aldred,S.F., Collins,P.J., Davis,C.A., Doyle,F., Epstein,C.B., Frietze,S., Harrow,J., Kaul,R. *et al.* (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
- Fu,Y., Foden,J.A., Khayter,C., Maeder,M.L., Reyon,D., Joung,K. and Sander,J.D. (2013) High frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat. Biotechnol.*, **31**, 822–826.
- Cho,S.W., Kim,S., Kim,Y., Kweon,J., Kim,H.S., Bae,S. and Kim,J.-S. (2014) Analysis of off-target effects of CRISPR/Cas-derived RNA-guided endonucleases and nickases. *Genome Res.*, **24**, 132–141.
- Kim,D., Bae,S., Park,J., Kim,E., Kim,S., Yu,H.R., Hwang,J., Kim,J.-I. and Kim,J.-S. (2015) Digenome-seq: genome-wide profiling of CRISPR-Cas9 off-target effects in human cells. *Nat. Methods*, **12**, 237–243.
- Ran,F.A., Cong,L., Yan,W.X., Scott,D.A., Gootenberg,J.S., Kriz,A.J., Zetsche,B., Shalem,O., Wu,X., Makarova,K.S. *et al.* (2015) *In vivo* genome editing using *Staphylococcus aureus* Cas9. *Nature*, **520**, 186–191.
- Frock,R.L., Hu,J., Meyers,R.M., Ho,Y.-J., Kii,E. and Alt,F.W. (2015) Genome-wide detection of DNA double-stranded breaks induced by engineered nucleases. *Nat. Biotechnol.*, **33**, 179–186.
- Wang,X., Wang,Y., Wu,X., Wang,J., Wang,Y., Qiu,Z., Chang,T., Huang,H., Lin,R.-J. and Yee,J.-K. (2015) Unbiased detection of off-target cleavage by CRISPR-Cas9 and TALENs using integrase-defective lentiviral vectors. *Nat. Biotechnol.*, **33**, 175–178.
- Kim,D., Kim,S., Kim,S., Park,J. and Kim,J.-S. (2016) Genome-wide target specificities of CRISPR-Cas9 nucleases revealed by multiplex Digenome-seq. *Genome Res.*, **26**, 406–415.
- Kleinstiver,B.P., Pattanayak,V., Prew,M.S., Tsai,S.Q., Nguyen,N.T., Zheng,Z. and Joung,J.K. (2016) High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature*, **529**, 490–495.
- Slymaker,I.M., Gao,L., Zetsche,B., Scott,D.A., Yan,W.X. and Zhang,F. (2016) Rationally engineered Cas9 nucleases with improved specificity. *Science*, **351**, 84–88.
- Tsai,S.Q., Nguyen,N.T., Malagon-Lopez,J., Topkar,V.V., Aryee,M.J. and Joung,J.K. (2017) CIRCLE-seq: a highly sensitive *in vitro* screen for genome-wide CRISPR-Cas9 nuclease off-targets. *Nat. Methods*, **14**, 607–614.
- Chen,J.S., Daggas,Y.S., Kleinstiver,B.P., Welch,M.M., Sousa,A.A., Harrington,L.B., Sternberg,S.H., Joung,J.K., Yildiz,A. and Doudna,J.A. (2017) Enhanced proofreading governs CRISPR-Cas9 targeting accuracy. *Nature*, **550**, 407–410.

22. Cameron,P., Fuller,C.K., Donohoue,P.D., Jones,B.N., Thompson,M.S., Carter,M.M., Gradia,S., Vidal,B., Garner,E., Slorach,E.M. *et al.* (2017) Mapping the genomic landscape of CRISPR-Cas9 cleavage. *Nat. Methods*, **14**, 600–606.
23. Kim,D. and Kim,J.-S. (2018) DIG-seq: a genome-wide CRISPR off-target profiling method using chromatin DNA. *Genome Res.*, **28**, 1894–1900.
24. Anderson,K.R., Haeussler,M., Watanabe,C., Janakiraman,V., Lund,J., Modrusan,Z., Stinson,J., Bei,Q., Buechler,A., Yu,C. *et al.* (2018) CRISPR off-target analysis in genetically engineered rats and mice. *Nat. Methods*, **15**, 512–514.
25. Listgarten,J., Weinstein,M., Kleinstiver,B.P., Sousa,A.A., Joung,J.K., Crawford,J., Gao,K., Hoang,L., Elibol,M., Doench,J.G. *et al.* (2018) Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.*, **2**, 38–47.
26. Jones,S.K., Hawkins,J.A., Johnson,N.V., Jung,C., Hu,K., Rybarski,J.R., Chen,J.S., Doudna,J.A., Press,W.H. and Finkelstein,I.J. (2020) Massively parallel kinetic profiling of natural and engineered CRISPR nucleases. *Nat. Biotechnol.*, <https://www.nature.com/articles/s41587-020-0646-5>.
27. Haeussler,M., Schönig,K., Eckert,H., Eschstruth,A., Mianné,J., Renaud,J.-B., Schneider-Maunoury,S., Shkumatava,A., Teboul,L., Kent,J. *et al.* (2016) Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.*, **17**, 148.
28. Hsu,P.D., Scott,D.A., Weinstein,J.A., Ran,F.A., Konermann,S., Agarwala,V., Li,Y., Fine,E.J., Wu,X., Shalem,O. *et al.* (2013) DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat. Biotechnol.*, **31**, 827–832.
29. Newman,A., Starrs,L. and Burgio,G. (2020) Cas9 cuts and consequences; detecting, predicting, and mitigating CRISPR/Cas9 on- and off-target damage. *BioEssays*, **47**, 2000047.
30. O’Geen,H., Yu,A.S. and Segal,D.J. (2015) How specific is CRISPR/Cas9 really? *Curr. Opin. Chem. Biol.*, **29**, 72–78.
31. Frankish,A., Diekhans,M., Ferreira,A.M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2019) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
32. Moore,J.E., Purcaro,M.J., Pratt,H.E., Epstein,C.B., Shores,N., Adrian,J., Kawli,T., Davis,C.A., Dobin,A., Kaul,R. *et al.* (2020) Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, **583**, 699–710.
33. Liu,Q., He,D. and Xie,L. (2019) Prediction of off-target specificity and cell-specific fitness of CRISPR-Cas System using attention boosted deep learning and network-based gene feature. *PLoS Comput. Biol.*, **15**, e1007480.
34. Zhang,G., Dai,Z. and Dai,X. (2020) C-RNNCrispr: Prediction of CRISPR/Cas9 sgRNA activity using convolutional and recurrent neural networks. *Comput. Struct. Biotechnol. J.*, **18**, 344–354.
35. Franco,M.M., Prickett,A.R. and Oakey,R.J. (2014) The role of CCCTC-binding factor (CTCF) in genomic imprinting, development, and reproduction. *Biol. Reprod.*, **91**, 125.
36. Sims,R.J., Nishioka,K. and Reinberg,D. (2003) Histone lysine methylation: a signature for chromatin function. *Trends Genet.*, **19**, 629–639.
37. De Magis,A., Manzo,S.G., Russo,M., Marinello,J., Morigi,R., Sordet,O. and Capranico,G. (2019) DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 816–825.
38. Neph,S., Kuehn,M.S., Reynolds,A.P., Haugen,E., Thurman,R.E., Johnson,A.K., Rynes,E., Maurano,M.T., Vierstra,J., Thomas,S. *et al.* (2012) BEDOPS: high-performance genomic feature operations. *Bioinformatics*, **28**, 1919–1920.
39. Nadel,J., Athanasiadou,R., Lemetre,C., Wijetunga,N.A., O Broin,P., Sato,H., Zhang,Z., Jeddloh,J., Montagna,C., Golden,A. *et al.* (2015) RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenet. Chromatin*, **8**, 46.
40. Sanz,L.A., Hartono,S.R., Lim,Y.W., Steyaert,S., Rajpurkar,A., Ginno,P.A., Xu,X. and Chédin,F. (2016) Prevalent, dynamic, and conserved R-Loop structures associate with specific epigenomic signatures in mammals. *Mol. Cell*, **63**, 167–178.
41. Gorthi,A., Romero,J.C., Loranc,E., Cao,L., Lawrence,L.A., Goodale,E., Iniguez,A.B., Bernard,X., Masamsetti,V.P., Roston,S. *et al.* (2018) EWS-FLI1 increases transcription to cause R-Loops and block BRCA1 repair in Ewing sarcoma. *Nature*, **555**, 387–391.
42. Barrett,T., Wilhite,S.E., Ledoux,P., Evangelista,C., Kim,I.F., Tomashevsky,M., Marshall,K.A., Phillippy,K.H., Sherman,P.M., Holko,M. *et al.* (2013) NCBI GEO: archive for functional genomics data sets-update. *Nucleic Acids Res.*, **41**, D991–D995.
43. Barnett,D.W., Garrison,E.K., Quinlan,A.R., Strömberg,M.P. and Marth,G.T. (2011) BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics*, **27**, 1691–1692.
44. Alkan,F., Wenzel,A., Anthon,C., Havgaard,J.H. and Gorodkin,J. (2018) CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.*, **19**, 177.
45. Gruber,A.R., Lorenz,R., Bernhart,S.H., Neubock,R. and Hofacker,I.L. (2008) The Vienna RNA Websuite. *Nucleic Acids Res.*, **36**, W70–W74.
46. Rauscher,B., Heigwer,F., Breinig,M., Winter,J. and Boutros,M. (2017) GenomeCRISPR - a database for high-throughput CRISPR/Cas9 screens. *Nucleic Acids Res.*, **45**, D679–D686.
47. Yuan,G.-C., Liu,Y.-J., Dion,M.F., Slack,M.D., Wu,L.F., Altschuler,S.J. and Rando,O.J. (2005) Genome-Scale identification of nucleosome positions in *S. cerevisiae*. *Science*, **309**, 626–630.
48. Yarrington,R.M., Verma,S., Schwartz,S., Trautman,J.K. and Carroll,D. (2018) Nucleosomes inhibit target cleavage by CRISPR-Cas9 in vivo. *Proc. Natl. Acad. Sci. U.S.A.*, **115**, 9351–9358.