Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Genome Analysis

CRISPRon/off: CRISPR/Cas9 on- and off-target gRNA design

Christian Anthon[1], Giulia Ilaria Corsi[1] and Jan Gorodkin[1,*]

[1]Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen, Thorvaldsensvej 57, 1871 Frederiksberg, Denmark

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: The effectiveness of CRISPR/Cas9-mediated genome editing experiments largely depends on the guide RNA (gRNA) used by the CRISPR/Cas9 system for target recognition and cleavage activation. Careful design is necessary to select a gRNA with high editing efficiency at the on-target site and with minimum off-target potential. Here we present our webserver for gRNA design with a user-friendly graphical interface, which provides interoperability between our on- and off-target prediction tools, CRISPRon and CRISPRoff, for a complete and streamlined gRNA selection.

Availability and implementation: The graphical interface uses the Integrative Genomic Viewer (IGV) JavaScript plugin. The backend tools are implemented in Python and C. The CRISPRon and CRISPRoff webservers and command-line tools are freely available at https://rth.dk/resources/crispr.

Contact: gorodkin@rth.dk

1 Introduction

CRISPR/Cas9, is an RNA-guided DNA endonuclease broadly employed as a genome editing tool. The role of the Cas9 complex in editing is recognizing and cleaving on-target DNA sites, which are subsequently repaired to obtain an edit of interest (Haeussler and Concordet, 2016). To recognize a target, Cas9 binds to a short DNA motif called "protospacer adjacent motif" (PAM) and probes flanking DNA for complementarity with its guide RNA (gRNA) (Anders, et al., 2014). Because mismatches and bulges in the gRNA-DNA hybrid and in the PAM are tolerated, cleavage by Cas9 can also happen at sites other than the on-target, resulting in off-target edits (Fu, et al., 2013). The cleavage efficiency of Cas9 varies at different on- and off-targets, mostly depending on properties of the gRNA and the target site (Doench, et al., 2016; Doench, et al., 2014; Peng, et al., 2018; Wang, et al., 2014; Xiang, et al., 2021; Xu, et al., 2015).

The goal of gRNA design is to select the gRNA with maximum efficiency and minimal off-target potential among the gRNAs that are suitable to cleave a target region. A major computational challenge of gRNA design is the identification and scoring of potential off-targets (pOTs). This process requires a time- and resource-consuming genome-wide search of gRNA targets and the subsequent scoring of possibly several thousands of pOTs. Training machine and deep learning models for on-target efficiency prediction is also computationally demanding, but once such models are produced and loaded relatively negligible time is required to generate results compared to the off-target search and evaluation.

To allow for on- and off-target aware gRNA design, we make use of two *in silico* tools which we previously have been involved in. These are available as webservers and command-line tools: CRISPRon (Xiang, et al., 2021) for on-target cleavage efficiency prediction, and CRISPRoff (Alkan, et al., 2018) for pOT assessment, which searches for pOTs in the genome using RIsearch2 (Alkan, et al., 2017). CRISPRon is a deep-learning model that predicts Cas9-mediated indel frequencies at gRNA on-target sites with top prediction performance in the field (Xiang, et al., 2021). Compared to other notable on-target cleavage prediction tools available as webservers, such as the Azimuth model (Doench, et al., 2016) used in CRISPOR (Concordet and Haeussler, 2018), CRISPick (https://portals.broadinstitute.org/gppx/crispick/public), and CHOPCHOP (Labun, et al., 2019), CRISPRon has the advantage to be trained on indel frequencies, which is a more direct measure of cleavage efficiency than the loss-of-protein-function outcomes employed in the training of Azimuth. This aspect makes CRISPRon more suitable to design gRNAs for tasks beyond the functional knockout of protein-coding genes (*e.g.*, knock-in of short genomic variants, knockout of non-coding

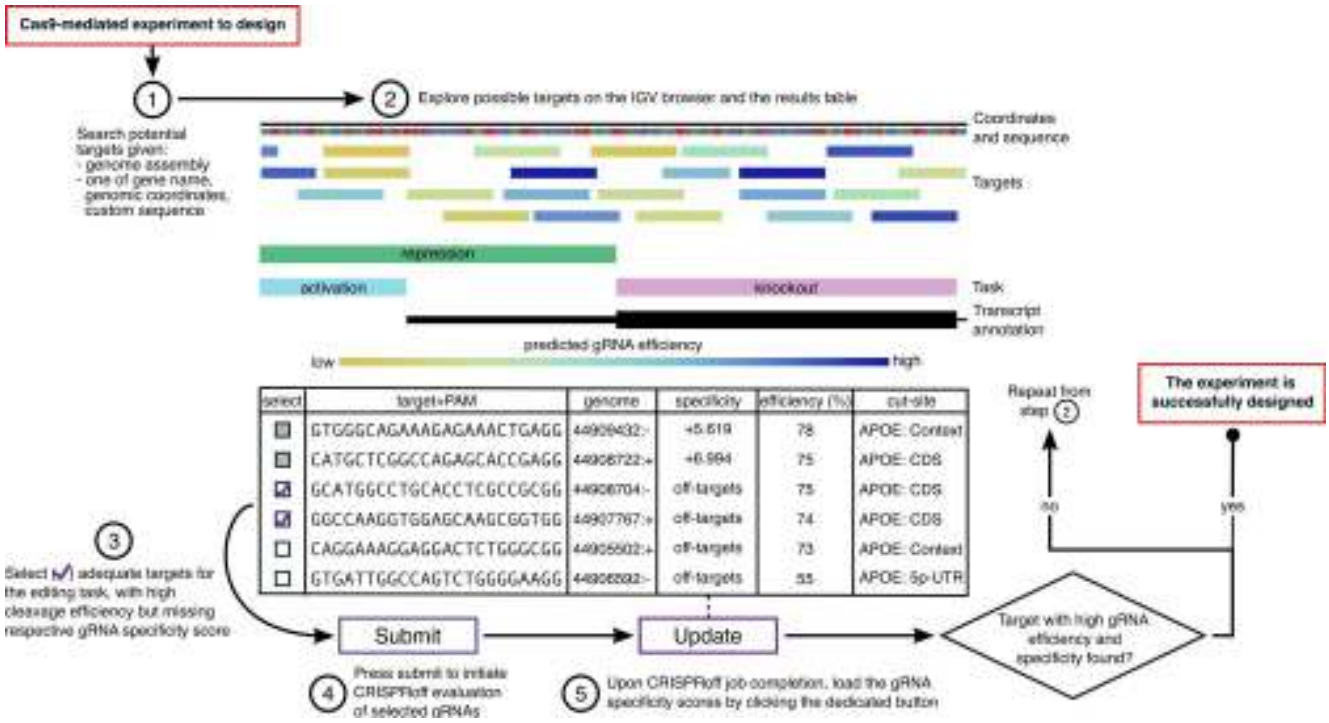


Fig. 1. Design of gRNAs for Cas9 experiments with CRISPRon/off. Sequence of steps to identify targets that can be edited by Cas9 with high efficiency and with minimum off-target potential. The genomic sequence reported in the table rows is that of the DNA target (5'-3' strand) which has the same sequence as the gRNA (which binds to the complimentary DNA) and the PAM. The "efficiency" is the predicted indel frequency. The "specificity" is the log10-scaled probability of binding at the on-target site compared to binding anywhere in the genome. The "genome" coordinate is the start position of the target followed by the strand; all targets are in the same region and the same chromosome, which is not specified. IGV= Integrative Genomic Viewer. CDS=coding sequence. 5p-UTR=5' untranslated region.

RNAs). CRISPRoff is the first computational model for the assessment of pOTs based exclusively on free energy changes, with high prediction accuracy compared to mismatch-based methods and high recall thanks to its consideration for DNA-RNA G-U wobble base pairs (Alkan, et al., 2018). On top of this, CRISPRoff is not trained on species-specific data, which makes it suitable to evaluate pOTs in any species.

The design of gRNAs requires not only accurate on- and off-target evaluation methods, but also an effective user-friendly interface. Here, we present a new interface that allows interoperability between the CRISPRon and CRISPRoff webservers. It is now possible to retrieve genome-wide gRNA off-target information within the CRISPRon platform during gRNA design, as well as obtaining on-target efficiencies for gRNAs tested for off-targets in CRISPRoff. The interoperability of the CRISPRon/off interface enables a complete gRNA design within a single "workflow", speeding up the whole design process and improving usability. In addition to the interoperability, both webservers have been substantially improved in terms of speed and available features compared to the previously published versions. The CRISPRon/off webservers and command line tools are freely available via http://rth.dk/resources/crispr. Via the same link we also provide a pipeline, CRISPRroots (Corsi, et al., 2021), for the post-assessment of on/off-targets in RNA-seq data generated after Cas9-mediated editing experiments.

2 Results and Discussion

The biggest advance in the CRISPRon/off interface is the interoperability between the CRISPRon and CRISPRoff webservers, which provides huge benefits in the whole gRNA selection process and in terms of user experience. In CRISPRon, users can select multiple gRNA-target candidates based on properties such as promoters and CDS, as well as predicted cleavage efficiency scores (indel frequencies at targets). These can be easily inspected either in the built-in Integrative Genomics Viewer browser (Robinson, et al., 2020; Robinson, et al., 2011) or in an interactive table, both provided in the results page. The gRNAs selected in CRISPRon can then be sent directly to CRISPRoff for the evaluation of their off-target potential. CRISPRoff calculates the binding free energy at all the pOTs of each gRNA and summarizes the ability of the Cas9-gRNA complex to bind at the on-target site while accounting for genome-wide pOTs in a single gRNA specificity score (also referred to as CRISPRspec, see Alkan *et al.* 2018 for details (Alkan, et al., 2018)). The results of the CRISPRoff assessment are then imported in the CRISPRon results page, for a final off-target aware gRNA selection (Fig. 1).

The webservers have been updated so that mouse, rat, pig, zebrafish, and fruit fly in addition to human are now available for both on- and off-target search. All genomes and annotations are updated to the latest versions as of March 2022 (human: hg38; mouse: mm39; zebrafish: danRer11; fruit fly: dm6; rat: rn6; pig: susScr11; Ensembl annotations version 104 (Cunningham, et al., 2021)). The results of CRISPRon include annotations of a user selected transcript or, by default, a primary canonical transcript, using either the USCS annotations of canonical transcript (which only exist for human and mouse, http://genome.ucsc.edu (Lee, et al., 2021)) or a simple heuristic for other organisms. The heuristic consists in taking the longest transcript with "gene" biotype and, in case of equally long transcripts, the one with most exons.

The integrated genome browser is enriched with genomic variants for human from dbSNP (Sherry, et al., 2001). The presence of a SNP at the target site is likely to affect the editing efficiency of a gRNA designed to target the reference sequence. Thus, unless the exact sequence of the target is known, gRNAs for all variants of the target should be tested

independently. Moreover, we added indications for target regions suitable for specific editing tasks in protein-coding genes:

- Knockout: 90% of the protein-coding sequence translated from the target's primary canonical transcript starting from the N-terminus, optimal to obtain loss of function of the target protein (Doench, et al., 2016).
- Activation: from 300 nt upstream from the start of the transcription start site (TSS) of the primary canonical transcript, until the start of the TSS.
- Repression: from 200 nt upstream from the start of the TSS of the primary canonical transcript, until 200 nt after the TSS start or until the start of the first CDS for genes with 5' untranslated regions (UTRs) shorter than 200 nt.

The off-target assessment by CRISPRoff is speeded up significantly for human by keeping the indexed genome in the memory and for all organisms by filtering gRNAs for repeat-like sequences prior the extensive off-target search. A gRNA is marked as repeat-like if it maps more than 100 times with up to 2 mismatches in the genome in a bowtie1 search (Langmead, et al., 2009). In the off-target assessment process for gRNAs extraneous to the reference, it is now possible for users to mask-out the input and exclude from the search genomic regions that could potentially interfere with the on/off-target lists and scores. This is useful when the user supplies a sequence which differs from the reference genome, to avoid calling potential off-targets in both the target sequence and the reference genome, leading to incorrect off-targets and a skewed specificity score.

3 Conclusion

The integration of the CRISPR/Cas9 on- and off-target webservers, the additional features, and the speed-ups of the platform will facilitate user navigation and enhance gRNA selection, allowing to better design genome editing experiments maximizing on-target efficiency and simultaneously minimizing potential off-target effects. Future improvements include the pre-calculation of all the on-targets and, for key gRNAs, of the off-targets and the specificity score.

Funding

This work has been supported by the Independent Research Fund Denmark, FTP (9041-00317B) and by the Novo Nordisk Foundation (NNF21OC0068988).

Conflict of Interest: none declared.

References

Alkan, F., *et al.* CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biology* 2018;19(1):177.

Alkan, F., *et al.* RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. *Nucleic Acids Research* 2017;45(8):e60-e60.

Anders, C., *et al.* Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. *Nature* 2014;513(7519):569-573.

Concordet, J.-P. and Haeussler, M. CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research* 2018;46(W1):W242-W245.

Corsi, G.I., *et al.* CRISPRroots: on- and off-target assessment of RNA-seq data in CRISPR–Cas9 edited cells. *Nucleic Acids Research* 2021;50(4):e20-e20.

Cunningham, F., *et al.* Ensembl 2022. *Nucleic Acids Research* 2021;50(D1):D988-D995.

Doench, J.G., *et al.* Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nature Biotechnology* 2016;34(2):184-191.

Doench, J.G., *et al.* Rational design of highly active sgRNAs for CRISPR-Cas9–mediated gene inactivation. *Nature Biotechnology* 2014;32(12):1262-1267.

Fu, Y., *et al.* High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nature Biotechnology* 2013;31(9):822-826.

Haeussler, M. and Concordet, J.-P. Genome Editing with CRISPR-Cas9: Can It Get Any Better? *Journal of Genetics and Genomics* 2016;43(5):239-250.

Labun, K., *et al.* CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Research* 2019;47(W1):W171-W174.

Langmead, B., *et al.* Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009;10(3):R25.

Lee, B.T., *et al.* The UCSC Genome Browser database: 2022 update. *Nucleic Acids Research* 2021;50(D1):D1115-D1122.

Peng, H., *et al.* CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. *Bioinformatics* 2018;34(18):3069-3077.

Robinson, J.T., *et al.* igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *bioRxiv* 2020:2020.2005.2003.075499.

Robinson, J.T., *et al.* Integrative genomics viewer. *Nature Biotechnology* 2011;29(1):24-26.

Sherry, S.T., *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* 2001;29(1):308-311.

Wang, T., *et al.* Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science* 2014;343(6166):80-84.

Xiang, X., *et al.* Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nature Communications* 2021;12(1):3238.

Xu, H., *et al.* Sequence determinants of improved CRISPR sgRNA design. *Genome Res* 2015;25(8):1147-1157.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-31543-6 OPEN

Massively targeted evaluation of therapeutic CRISPR off-targets in cells

Xiaoguang Pan[1,2,12], Kunli Qu[1,2,3,12], Hao Yuan[1,4,12], Xi Xiang[1,3,12], Christian Anthon[5,12], Liubov Pashkova[5], Xue Liang[1,2], Peng Han[1,2], Giulia I. Corsi[5], Fengping Xu[1,4,6], Ping Liu[6,7], Jiayan Zhong[6,7], Yan Zhou[3], Tao Ma[6,7], Hui Jiang[6,7], Junnian Liu[1], Jian Wang[6], Niels Jessen[3,8], Lars Bolund[1,3], Huanming Yang[6,9], Xun Xu[6,10], George M. Church[11]✉, Jan Gorodkin[5]✉, Lin Lin[3,8]✉ & Yonglun Luo[1,3,4,6,8,9]✉

Methods for sensitive and high-throughput evaluation of CRISPR RNA-guided nucleases (RGNs) off-targets (OTs) are essential for advancing RGN-based gene therapies. Here we report SURRO-seq for simultaneously evaluating thousands of therapeutic RGN OTs in cells. SURRO-seq captures RGN-induced indels in cells by pooled lentiviral OTs libraries and deep sequencing, an approach comparable and complementary to OTs detection by T7 endonuclease 1, GUIDE-seq, and CIRCLE-seq. Application of SURRO-seq to 8150 OTs from 110 therapeutic RGNs identifies significantly detectable indels in 783 OTs, of which 37 OTs are found in cancer genes and 23 OTs are further validated in five human cell lines by targeted amplicon sequencing. Finally, SURRO-seq reveals that thermodynamically stable wobble base pair (rG•dT) and free binding energy strongly affect RGN specificity. Our study emphasizes the necessity of thoroughly evaluating therapeutic RGN OTs to minimize inevitable off-target effects.

[1] Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, BGI-Shenzhen, Qingdao, China. [2] Department of Biology, Copenhagen University, Copenhagen, Denmark. [3] Department of Biomedicine, Aarhus University, Aarhus, Denmark. [4] College of Life Sciences, University of Chinese Academy of Sciences, Beijing, China. [5] Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark. [6] BGI-Research, BGI-Shenzhen, Shenzhen, China. [7] MGI, BGI-Shenzhen, Shenzhen, China. [8] Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus, Denmark. [9] IBMC-BGI Center, the Cancer Hospital of the University of Chinese Academy of Sciences (Zhejiang Cancer Hospital), Institute of Basic Medicine and Cancer (IBMC), Chinese Academy of Sciences, Hangzhou, Zhejiang 310022, China. [10] Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, China. [11] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [12] These authors contributed equally: Xiaoguang Pan, Kunli Qu, Hao Yuan, Xi Xiang, Christian Anthon. ✉email: gchurch@genetics.med.harvard.edu; gorodkin@rth.dk; lin.lin@biomed.au.dk; alun@biomed.au.dk

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) RNA-guided nucleases (RGNs) has been used in therapy of several inherited human diseases[1–4]. Major efforts have focused on improving RGN editing efficiency via stabilization of the small guide RNA (sgRNA) thermodynamics[5], modification of the RGNs[6–8], utilization of homology-independent mediated targeted integration (HITI)[9] and optimization of RGN delivery[10,11]. The inevitably adverse effects caused by unspecific RGN editing of cancer genes are major concerns for the clinical application of RGN-based therapies. Improvement of RGN specificity and development of methods for identifying and evaluating the potential off-targets (OT) introduced by RGNs are equally essential to advance RGN-based gene therapy. Several experimental RGN OT identification/quantification methods have been developed (Supplementary Data 1), which can be grouped into three categories (Supplementary Fig. S1). Category One contains genome-wide cell-free biochemical methods which relies on the capture of RGN-induced OT cleavage on either naked DNA or fixed chromatin fibers by sequencing. Examples are CIRCLE-seq (cell-free)[12], Digenome-seq (cell-free)[13], SITE-seq (cell-free)[14], BLISS (ex vivo)[15] and DIG-seq (ex vivo)[16]. Category Two contains methods depending on genome-wide in-cell capturing of RGN-induced off-target cleavage by sequencing, such as GUIDE-seq and IDLV-capture relying on insertion of double strand DNA and IDLV vector to the DNA double strand breaks respectively[17,18], HTGTS and PEM-seq relying on translocation between on-target and off-targets[19,20], and DISCOVER-seq relying on immunoprecipitation of DNA repair protein MRE11 to capture the DNA double strand break (DSB) sites[21]. While cell-free biochemical methods are rapid, conventional, and not depending on reference genomes, they inevitably capture many pseudo off-target sites. In-cell methods (e.g., GUIDE-seq) capture the bona fide RGN off-targets more faithfully as compared to cell-free methods. However, spontaneous DSBs lead to capturing pseudo off-targets independent of RGNs[17]. To complement this, Category Three is composed of targeted in-cell RGN OT validation methods, such as T7 endonuclease 1 (T7E1), targeted deep sequencing, TIDE and CUT-PCR[22,23]. However, current targeted in-cell RGN off-target evaluation methods are greatly limited by their scales. Only a few sites can be evaluated for each RGN in a single study due to their high labor and time cost. A modified targeted amplicon sequencing method based on the rhAmpSeq has thus been reported for simulated and targeted analysis of several CRISPR gRNAs and hundreds of selected off-target sites in a single reaction[24,25]. This method has greatly improved the scale of targeted analysis of CRISPR off-targets by deep sequencing.

Here we introduce and apply SURRO-seq, a high-throughput method for targeted in-cell capture of RGN off-targets based on a pooled lentiviral vectors library encoding gRNA and barcoded surrogate off-target sites, to evaluate therapeutic RGN off-targets in cells. SURRO-seq exhibits high sensitivity and accuracy compared to GUIDE-seq and CIRCLE-seq by evaluating 170 previously investigated OTs from 11 RGNs in HEK293T cells. We then applied SURRO-seq to evaluate 8150 OTs from 110 therapeutic RGNs and identify 783 OTs showing significantly detectable indels. 37 OTs with significantly detectable indels are found in cancer genes, highlighting the clinical significance and great need of pre-assessing RGN OTs with SURRO-seq. The SURRO-seq identified OTs were further validated by targeted deep sequencing of five RGN-edited human cell lines. Analyses of OTs with high indel frequencies revealed that mismatch types leading to thermodynamically stable wobble base pair strongly increase RGN OT effect. We further perform benchmark analyses of latest RGN OT prediction tools with SURRO-seq OT data. The energy-based predictors, which incorporate gRNA and DNA binding energies, give the best performance.

Results

Design of SURRO-seq. Libraries of surrogate vectors have been used in many studies to massively capture on-target efficiencies[26,27]. We and others show that single surrogate episomal vector (or genomic integration site) has been used as a sensitive method to measure RGN off-target activity. But the method is only applicable for evaluating a limited number of RGN OTs[7,28]. Previously, we introduced an optimized high-throughput approach for targeted in-cell evaluation of on-target RGN efficiency using a pool of lentiviral surrogate vectors[29]. Here we introduce site specific barcoding and repurpose the method for high throughput targeted evaluation of RGN off-targets (OTs) in cells. For a given RGN, protospacer sequences of all OTs are very similar and only differ for 1–5 nucleotides (nt). Following RGN editing, deletion indels could erase nucleotides that differ between OTs, making it impossible to uniquely assign the deletion indels to the OTs (Supplementary Fig. S2). To overcome the indel split problem, we introduced a 10-nt barcoding strategy to distinguish indels reads in the ON and OT sites (Supplementary Fig. S2). As showed in Fig. 1, SURRO-seq contains nine major steps (see Supplementary Note 1 for extended description of the method) with three modifications compared to our previous on-target method CRISPRon[29]: (1) The surrogate site contains a 10-nt barcode preceding the 27-nt surrogate OT site, which contains the OT protospacer (20 nt), protospacer adjacent motif (PAM, NGG), 4-nt PAM downstream sequences; (2) Barcode-guided split of indel reads (Supplementary Figs. S2, S3); and (3) Fishers' exact test of OTs with significant indels [a. Two-fold higher indel frequency in the SpCas9 cells as compared to the wild type cells (MOCK); b. Fishers exact test and Benjamini and Hochberg (BH)-adjusted p-value less than 0.05] (Supplementary Note 2).

Validation of previously evaluated RGN OTs with SURRO-seq. First, we sought to assess if SURRO-seq can capture RGN OTs previously evaluated by other methods. We generated a small library (LibA) containing 170 OTs from 11 RGNs (Fig. 2a). These 11 RGNs and 170 OTs had been detected by T7E1[30,31], GUIDE-seq[17] and/or CIRCLE-seq[12]. We transduced SpCas9-overexpressing (SpCas9) and wildtype (MOCK) HEK293T cells with LibA (MOI = 0.3, and 4000-fold coverage of LibA, see methods). Eight days after LibA transduction, indel frequencies introduced in the surrogate OTs were quantified by targeted deep sequencing. Analyses of LibA data (Supplementary Fig. S3–5, Fig. 2b, Supplementary Data 2) showed that SURRO-seq can capture nearly 100% of the T7E1-detected OTs (22 out of 23, Fig. 2b), most (104 out of 149, 70%) of GUIDE-seq-captured OTs (Fig. 2c), and approximately half (78 out of 153, 51%) CIRCLE-seq-captured OTs (Fig. 2d). Five RGNs have been analyzed by GUIDE-seq, CIRCLE-seq and SURRO-seq. Comparison of these 141 OTs from the five RGNs showed that a large subset (82 OTs, 58%) of these 141 OTs are captured by at least two methods (Fig. 2e). There are 1, 18, and 40 OTs only captured by SURRO-seq, GUIDE-seq and CIRCLE-seq respectively (Fig. 2e). CIRCLE-seq is based on CRISPR-Cas9 cleavage of cell-free and histone-free DNA, and GUIDE-seq is relying on repair of DSBs and insertion of the targeted double-strand DNA oligonucleotide in cells. Since the chromatin can inhibit Cas9 off-target effect[16], it is thus not surprising to observe that a subset of these RGN OTs can only be captured by one method[12,13,17]. SURRO-seq offers a complementary in-cell method for targeted validation of RGN OTs identified by genome-wide screening approaches.

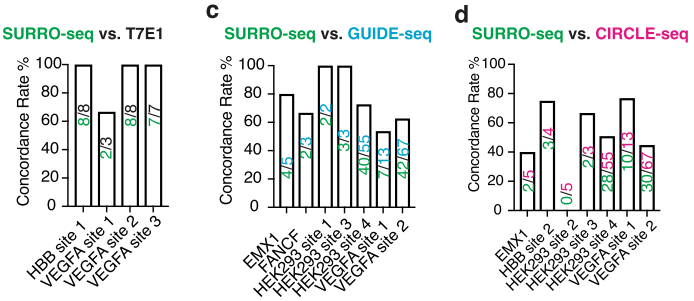
Large-scale evaluation of therapeutic RGNs OTs with SURRO-seq. To investigate if SURRO-seq can be used for high throughput targeted in-cell evaluation of RGN OTs, we selected 110 RGNs

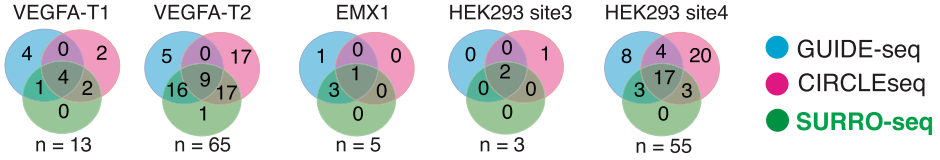
Fig. 1 Design of SURRO-seq. An overview of the nine major steps of SURRO-seq is schematically presented. MM, mismatches; ON, on-target; OT, off-targets; PBS, primer binding sites; BB, BsmBI binding site; Sp, spacer; Scr, SpCas9 gRNA scaffold; BC, barcodes; PS, protospacer; PAM, protospacer adjacent motif; 4r, 4 bp downstream sequences; GGA, Golden Gate Assembly; MOI, multiplexity of infection; WT, wildtype; Results presented in Step 9 are indel frequencies captured by SURRO-seq for RGN *VEGFA* T3.

targeting 21 human genes that have been used in preclinical gene therapy studies (Supplementary Data 3). Given that most RGNs can only tolerate a few (1–3) mismatches between the sgRNA spacer and the protospacer sequences[32–34], we blasted the human reference genome with the 20-nt spacer of each RGN and retrieved all potential OTs with up to 4 mismatches. In total, 8150 OTs from these 110 RGNs were selected and synthesized, cloned into the SURRO-seq vector, and packaged into lentivirus,

a

Method	gRNA	OTs
T7E1	4	26
GUIDEseq	7	148
CIRCLEseq	7	152
Total*	11	170





f

Comparison of VEGFA-T1 OTs detection between T7E1, GUIDEseq, CIRCLEseq and SURRO-seq

protospacer	PAM	T7E1	GUIDE-seq		CIRCLE-seq		SURRO-seq					
GGGTGGGGGGAGTTTGCTCC	TGG	RGI%	RE%	reads	RE%	reads	IF%, SpCas9	I/T reads	IF%, MOCK	I/T reads	p.adj.	
----------	---	10.5	100	2685	100	736	63.29	1236/1953	0.09	2/2264	0E+00	*
----------C	-A-	NA	9.27	249	ND	ND	31.49	137/435	0.6	3/501	9E-47	*
TA----A----C	---	NA	72.07	1935	130.98	964	20.88	3363/16110	0.12	24/20189	0E+00	*
-C--------T	C--	NA	9.16	246	55.71	410	13.46	349/2592	0.08	2/2427	2E-101	*
C--G--A----------	---	3.67	13.97	375	136.68	1006	2.39	73/3058	0.12	3/2581	6E-16	*
---GA-----A------	---	8.54	30.5	819	68.75	506	1.79	34/1897	0.73	9/1227	2E-02	*
-A-G----A-C------	A--	NA	ND	ND	87.23	642	1.14	26/2278	0.04	1/2518	2E-07	*
---G-CA-------A--	---	NA	ND	ND	81.25	598	0.87	33/3814	0.1	4/4049	7E-07	*
CT-GT-----C------	A--	NA	12.07	324	ND	ND	0.65	7/1073	0	0/656	8E-02	NS
---A----T-------	---	8.32	22.23	597	ND	ND	0.27	13/4742	0.62	29/4679	2E-02	NS
-T-G----TA------	A--	NA	8.98	241	ND	ND	0.13	29/22221	0.15	28/18274	7E-01	NS
-A-G-T---------A	---	NA	4.28	115	ND	ND	0.1	6/6051	0.12	6/4818	9E-01	NS
--C----A---GA---	---	NA	ND	ND	58.97	434	0.07	4/6075	0.13	7/5308	5E-01	NS
AT----T-------A-	---	NA	ND	ND	70.38	518	0.04	3/7997	0.07	5/7587	6E-01	NS

g

Comparison of VEGFA-T2 OTs detection between T7E1, GUIDEseq, CIRCLEseq and SURRO-seq

protospacer	PAM	T7E1	GUIDE-seq		CIRCLE-seq		SURRO-seq					
GACCCCCTCCACCCCGCCTC	CGG	RGI%	RE%	reads	RE%	reads	IF%, SpCas9	I/T reads	IF%, MOCK	I/T reads	p.adj.	
----------	---	38.6	100	540	100	294	97	8350/8608	0.75	155/20568	0E+00	*
ATT-----C------	A--	19.4	286.85	1549	448.98	1320	96.38	1492/1548	2.21	74/3352	0E+00	*
ACA-----C------	A--	12.3	218.15	1178	259.18	762	95.53	2458/2573	0.2	17/8481	0E+00	*
CTA----------	---	15	100	540	211.56	622	93.21	851/913	0.4	2/494	8E-306	*
T-------CA------	T--	6.97	148.7	803	ND	ND	90.69	3818/4210	0.46	62/13387	0E+00	*
-GG----------	T--	15.6	89.44	483	682.99	2008	87.57	3495/3991	0.07	8/11623	0E+00	*
---------C---C-	---	33.6	72.41	391	ND	ND	83.79	1685/2011	0.3	8/2693	0E+00	*
-CTT----------A-	---	NA	170.19	919	425.85	1252	77.91	3407/4373	1.4	196/14011	0E+00	*
CCG----------A-	T--	NA	ND	ND	182.99	538	63.37	263/415	1.19	24/2014	3E-209	*
---------TC---A-	---	NA	142.04	767	182.31	536	56.19	608/1082	0.12	3/2547	0E+00	*
AT---------A--C-	T--	NA	22.41	121	ND	ND	55.3	5460/9874	0.32	99/30750	0E+00	*
CTG----C------A-	T--	NA	61.11	330	ND	ND	55.16	540/979	0.56	12/2130	1E-307	*
C-T----C-------A-	C--	NA	62.22	336	ND	ND	52.29	503/962	0.36	10/2784	0E+00	*
C-A----C---------T-	A--	NA	20.56	111	ND	ND	49.48	1085/2193	0.06	3/5328	0E+00	*
CG---T-C----------	---	28.7	82.59	446	ND	ND	48.77	4578/9387	0.41	114/28092	0E+00	*
-T-------T----A---	---	NA	ND	ND	214.29	630	38.94	493/1266	0.04	1/2264	8E-250	*
AG-----A--T-------	G--	NA	ND	ND	304.76	896	37.45	3668/9795	0.17	38/21758	0E+00	*
CCT-----CA-------A-	---	NA	115.74	625	ND	ND	33.24	1950/5867	1.47	167/11357	0E+00	*
AGG-----CA-------	A--	NA	ND	ND	ND	ND	31.91	90/282	0.15	1/664	8E-51	*
C---T---C-------C-	A--	NA	23.7	128	ND	ND	27.44	517/1884	1.47	67/4572	5E-220	*
TC-A----C-------C-	G--	NA	25	135	ND	ND	27.09	1510/5574	0.25	28/11165	0E+00	*
CTT-----CA------C-	A--	NA	47.41	256	ND	ND	24.37	97/398	0.46	7/1524	3E-61	*
AGG-----C--G------	A--	NA	46.67	252	ND	ND	21.15	386/1825	0.2	5/2523	3E-147	*
CT-----AC-------A-	---	29.8	ND	ND	207.48	610	20.01	871/4353	0.24	20/8464	0E+00	*
-C-------A------C-	---	NA	ND	ND	154.42	454	17.9	194/1084	0.72	19/2645	1E-85	*
C--TA----C-------C-	G--	NA	43.52	235	ND	ND	17.73	1172/6610	0.51	87/17133	0E+00	*
C-TT----C------A-	A--	NA	22.78	123	ND	ND	12.92	131/1014	0.39	10/2556	4E-61	*
TG-----TC-------T-	---	NA	146.48	791	308.16	906	11.51	3539/30753	0.18	127/70368	0E+00	*
CC--A--C----------	A--	NA	130.37	704	153.74	452	9.58	377/3934	0.63	60/9515	8E-141	*
C---------C-T-----	A--	NA	ND	ND	200.68	590	9.17	415/4527	0.71	84/11797	2E-151	*
TG------C-------A-	T--	NA	205	1107	ND	ND	5.8	4/69	0	0/197	7E-03	*

(Total VEGFA-T2 OTs data were shown in Extended Tables S2)

hereafter referred to as library B (LibB) (Fig. 3a). We transduced SpCas9 and wildtype (MOCK) HEK293T cells with LibB (MOI = 0.3, 4000-fold coverage) and analyzed indel frequencies from cells eight days after transduction by deep sequencing.

We first analyzed the on-target gRNA efficiency of these 110 RGNs. SURRO-seq successfully captures on-target efficiencies for all 110 RGNs (Fig. 3b). The SpCas9 protein is overexpressed in the HEK293T cells by doxycycline addition. Consistent with our previous observation[29], most ($n = 96$) RGNs exhibited high on-target activity (indel frequencies% (IF%) > 80%, Fig. 3b). A few RGNs ($n = 14$) had relatively low efficiency (IF% < 80%), and these were also significantly ($p < 0.0001$) lower in GC content compared to highly efficient RGNs (IF% > 80%) (Supplementary Fig. S6). Next, we analyzed indel frequencies in the OTs

Fig. 2 Validation of RGN OTs detection between T7E1, GUIDE-seq, and/or CIRCLE-seq by SURRO-seq. a Overview of RGN gRNAs and OTs selected for validated with SURRO-seq. **b**–**d** Comparison of the OT detection concordance rate between SURRO-seq and T7E1 (**b**), GUIDE-seq (**c**) and CIRCLE-seq (**d**). Numbers are total OTs for each RGN (upper) evaluated with the compared method and OTs agreed with SURRO-seq (lower). **e** Venn diagram comparison of OTs with significant significantly detectable off-targets (SURRO-seq) and OTs with deep sequencing reads detected by GUIDE-seq or CIRCLE-seq. Numbers are OT sites. **f**–**g** Comparison of VEGFA-T1 (**f**) and VEGFA-T2 (**g**) OT detections between T7E1, GUIDE-seq, CIRCLE-seq and SURRO-seq. Full results are showed in Supplementary Data 2. RGI, relative gel intensity; RE%, percentage of relative efficiency, calculated by % reads in OT per reads in ON; IF, indel frequency; I/T reads, indel/total reads; *P* values for comparison between SpCas9 and MOCK IF% are calculated with Benjamini and Hochberg (BH)-adjusted Fisher's exact test (two-sided). *, represents OT with significantly detectable indels (adj. *P* value < 0.05, FC (IF% SpCas9/ IF% MOCK) > = 2). NS, represents OTs with not significantly detectable indels.

introduced by RGNs. Surrogate OT sites with low sequencing quality (total clean reads <32 for both MOCK and SpCas9), low synthetic quality (IF% in MOCK > 4%, Supplementary Fig. S7) were filtered. Mutagenesis in essential genes caused by random integration of the SURRO-seq lentiviral vector or by OT targeting could affect cell proliferation. We performed surrogate site enrichment and enrichment analysis (Supplementary Fig. S8) and identified 196 sites (Fold-change between MOCK and SpCas9 > 2). Significantly detectable indels (fold-change (FC) of IF% SpCas9 / IF% MOCK > 2, adj. *p*-value < 0.05) were found in 30 of these RGN OTs and strikingly all depleted (Supplementary Fig. S8C, Supplementary Data 3). Nine RGN OTs (Supplementary Data 3) are known human essential genes by mapping to the human essential gene database DEG10[35]. As these depleted/enriched surrogate sites might affect the later analysis of the effect of DNA context and thermodynamics on RGN OT activity, we also excluded these sites for subsequent analysis. After the filtering, 7140 OTs were retained for downstream analyses (Supplementary Data 3). Quantification of indel frequencies in each OT in SpCas9 and MOCK cells showed that there were not significantly detectable indels for most of these OTs ($n = 6387$, hereafter referred to as NSOT). Significantly detectable indels (fold-change (FC of IF% SpCas9/IF% MOCK) > 2, adj. *p*-value < 0.05) were identified for 753 OTs in SpCas9 cells compared to MOCK (Fig. 3c, Supplementary Data 3). However, the indel frequency of most Sig. OTs were less than 3% (573 out of 753, Fig. 3c, d). We further divided the Sig. OTs into two groups: based on IF% in SpCas9 < 3% (Low Indel Sig. OTs, hereafter referred to as LIOT) and IF% > = 3% (High Indel Sig. OTs, hereafter referred to as HIOT). Notably, most HIOTs contain 1–3 mismatches (Supplementary Fig. S9). Our results demonstrate that the SURRO-seq can be used for high throughput targeted evaluation of RGN-induced indels at surrogate off target sites in cells.

To investigate where were these LIOTs and HIOTs located in the genome and in genes, we annotated their genomic locations according to the presence in intergenic region (IGR) or in genes (2 kb upstream, 5' untranslated region, exon, intron, 3' untranslated region, 2 kb downstream; Supplementary Data 3). Despite that most of the Sig. OTs are found in intron and IGRs, there are still a substantial number of Sig. OTs (nr. of LIOT = 200, nr. of HIOT = 87) found in gene exons and/or regulatory regions that might affect gene expression (Fig. 3d, Supplementary Fig. S10). Notably, 37 Sig. OTs were annotated in cancer-related genes (Supplementary Data 3). The RGN11153 (spacer sequences, CTGCTGCTGCTGCTGCTGGA), which was proposed for Huntington's Disease therapy by targeting the CAG expansion tract[36], exhibit great off-target effect (nr. of LIOT = 43, nr. of HIOT = 35). Two HIOTs of RGN11153 are found in cancer genes *ZFHX3* (exon) and *SOHLH2* (intron). The zinc-finger homeobox 3 (*ZFHX3*) is a tumor suppressor gene and knockout of *ZFHX3* in mouse leads to development of neoplastic lesions. Loss of function mutations in *ZFHX3* are frequently detected in human cancers i.e. high-grade human prostate cancers[37], endometrial cancers[38], urothelial bladder carcinoma[39], lung and brain tumors[40]. This finding emphasizes that carefully evaluating if therapeutic RGNs introduced any off target indels in cancer genes is needed. HIOTs were also found in another two cancer genes *BCOR* (intron) and *NCOR2* (intron) by RGN11155 and RGN11189 respectively. These two RGNs were used for HD (RGN11155) and β-Thalassemia (RGN11189) therapy[41]. For the LIOTs, despite low indel frequency (below 3%), significantly detectable indels were found in the exon of nine cancer genes, such as *U2AF2* and *NKTR* causing Acute myeloid leukemia (AML) (Supplementary Data 3 and Fig. S10). SURRO-seq thus offers a high throughput and targeted approach for in cell evaluation of RGN OTs in cells.

Validation of SURRO-seq identified OTs by targeted deep sequencing of endogenous genomic loci. To validate if OTs captured by SURRO-seq were also presented at the corresponding endogenous sites, we analyzed 23 OTs from seven RGNs in five human cell lines: human embryonic kidney cells (HEK293T), human primary fibroblasts, lung cancer cells (PC-9), ovarian cancer cells (SKOV3), and bone osteosarcoma epithelial cells (U2OS). Of these 23 OTs, 16 and 7 OTs were detected with significant and non-significant indels by SURRO-seq, respectively (Fig. 4a and Supplementary Data 4). These 16 Sig. OTs were selected to cover a broader distribution of indel frequencies, ranging from 2% to 96%. Instead of using lentivirus-based delivery of CRISPRs, we applied an optimized CRISPR delivery approach based on CRISPR ribonucleoprotein (RNP). Highly efficient delivery of CRISPR into various types of cells have been reports by us and many other groups[2,5,42,43]. We also validated that an enhanced green fluorescent protein (*EGFP*) mRNA can be delivered to nearly 100% of cells in all five cell lines (Supplementary Fig. S11). Seven on-target sites and 23 off-target endogenous genomic loci from the five cell lines were analyzed by targeted deep sequencing 48 h after RNP treatments (Fig. 4b). Several Cas9 mutants have been reported with improved specificity[44–46]. In addition to the wild type SpCas9, we also analyzed the indel frequencies in the 7 RGN on-target sites and 23 off-target sites in cells transfected with a high-fidelity Cas9 variant (HiFi-Cas9), of which a single point mutation (p.R691A) was introduced[44].

Analyses of deep sequencing results showed that significant indel frequencies (one-way ANOVA, $p < 0.05$) were detected in all seven RGN on-target sites in the five cell lines (Fig. 4c, Supplementary Fig. S12, Supplementary Data 4). Consistent with early results, HiFi-Cas9 retained similarly high on-target activity as the wild type Cas9 (Fig. 4c), except RGN11208 which is low in GC content (GC% = 20%, Supplementary Fig. S13). Analyses of indel frequencies in the 23 off-target sites showed that there is a good agreement (20 out of 23, 87%) between SURRO-seq and targeted sequencing of endogenous sites. 15 out of 17 (88%) of the SURRO-seq off-target sites with significantly detectable indels were validated by targeted deep sequencing of in Cas9 RNP

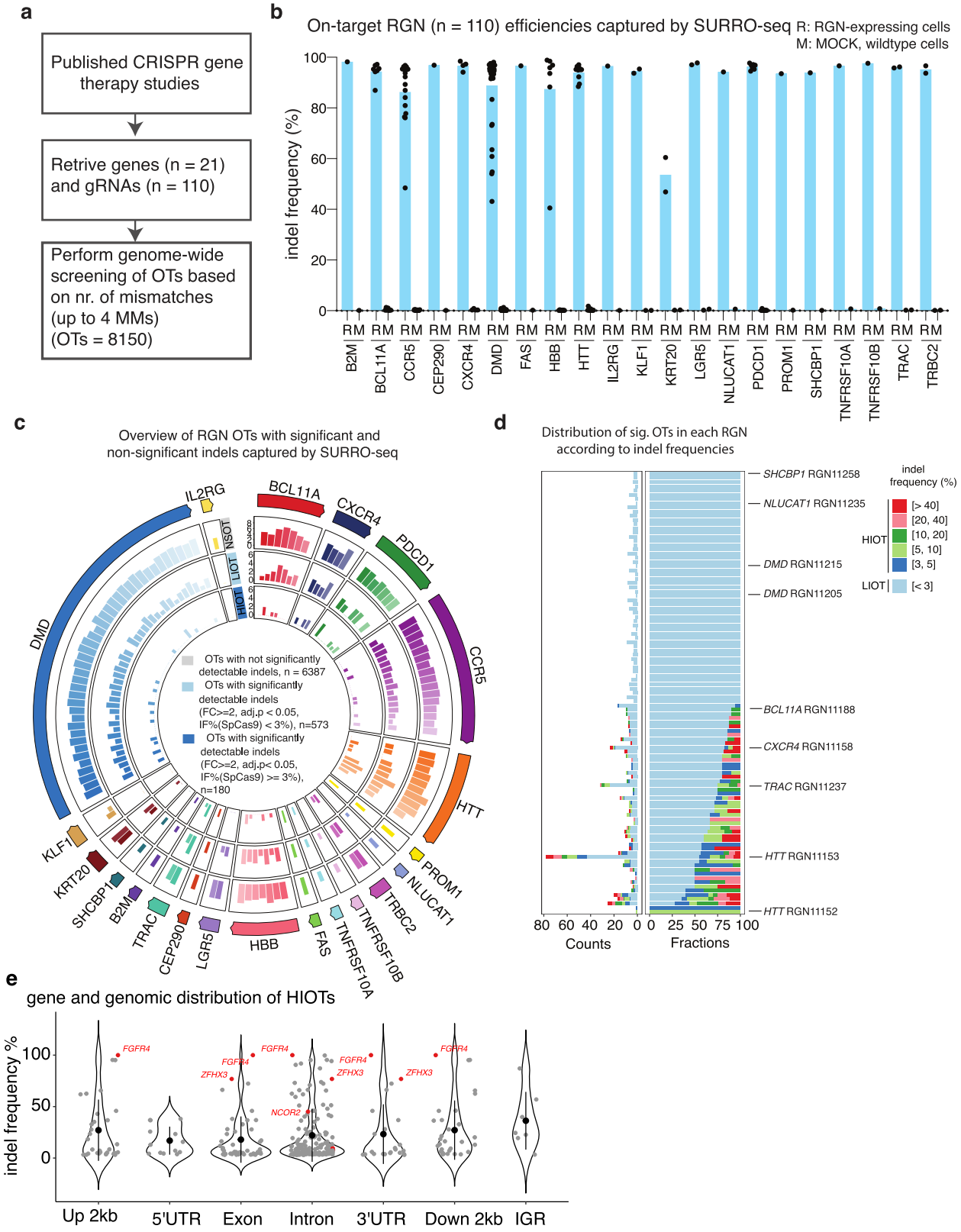
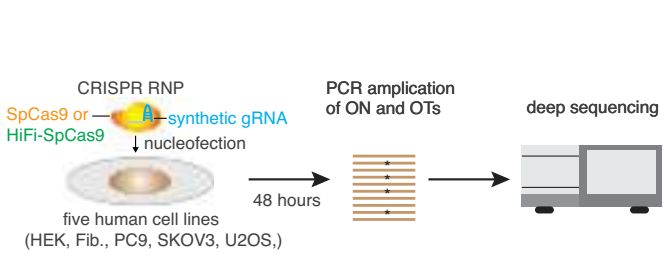
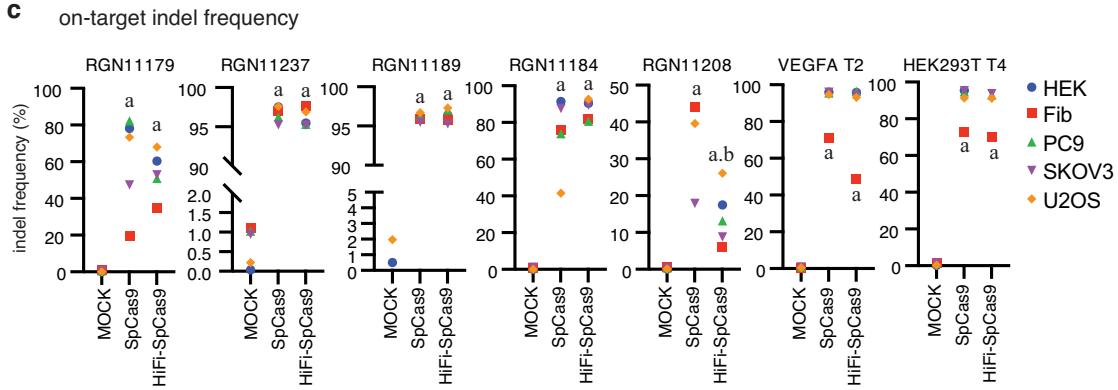


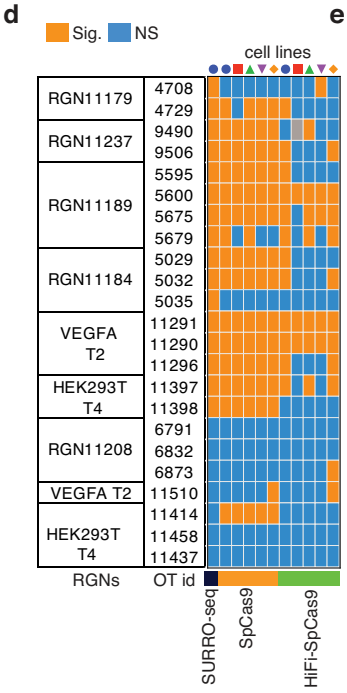
Fig. 3 High throughput evaluation of gene therapy RGN OTs with SURRO-seq. a Overview of gene therapy RGN selection and number of OTs captured. **b** Quantification of indel frequencies for the 110 RGNs by SURRO-seq. R, RGN edited, M, MOCK control. **c** Overview of the number of RGNs OTs with not significantly detectable indel (NSOT, outer circle, adj. *P* value > 0.05 or FC (IF% SpCas9/ID MOCK < 2)), with significantly detectable indels (adj. *P* value < 0.05 and FC (IF% SpCas9/ID MOCK > = 2)) but low indel frequency (<3%, LIOT, middle circle), and significantly detectable indels with high indel frequency (> = 3%, HIOT, inner circle). *P* values are derived from Benjamini and Hochberg (BH)-adjusted Fisher's exact test (two-sided). **d** Bar plot of total number (left) and fraction (right) of LIOTs and HIOTs for the RGNs. **e** Violin plot of the gene and genomics location of the HIOTs and indel frequency. OTs in cancer genes are highlighted in red. Data are presented as mean values +/− SD.

a

RGN	ON	Sig. OTs	NS. OTs
RGN11179	1	2	0
RGN11237	1	2	0
RGN11189	1	4	0
RGN11184	1	3	0
RGN11208	1	0	3
VEGFA T2	1	3	1
HEK293T T4	1	2	3
Total	7	16	7

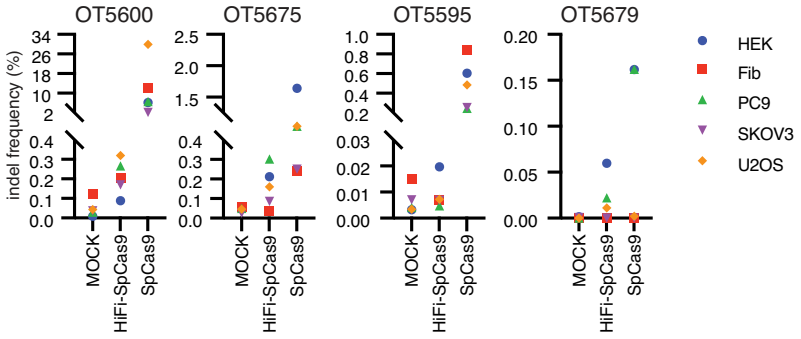


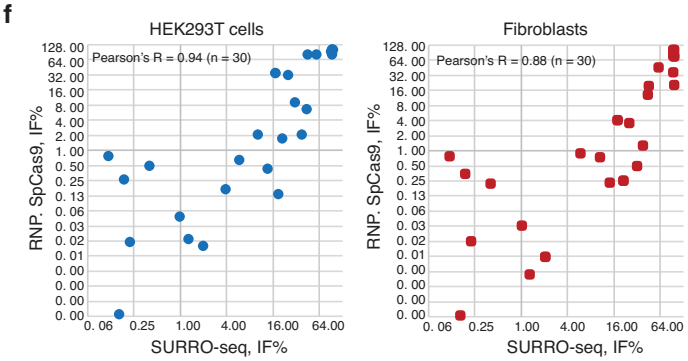




e

SURRO ID	Protospacer																		PAM			SURRO-seq	
5'-3'	A	G	C	C	A	G	T	G	C	C	A	G	A	A	G	A	G	C	C	A	A	G	G	IF%.Sp	IF%.WT
ON	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	98.84	0.18
OT5600	G	C	-	-	-	-	-	-	-	-	-	-	-	-	-	-	A	-	-	T	-	-	-	45.09	0.19
OT5675	C	A	-	-	-	-	-	-	-	-	-	-	G	-	-	-	-	-	-	T	C	-	-	21.38	0.22
OT5595	C	A	-	-	-	-	C	-	-	-	-	-	-	-	-	-	-	-	-	T	G	-	-	5.97	0.23
OT5679	G	C	-	-	-	-	-	-	-	-	-	G	-	A	-	-	-	-	-	-	G	-	-	4.02	0





treated cells (Fig. 4d, e, Supplementary Data 4). Due to differences in RGN delivery, editing time and RGN expression level between SURRO-seq and RNP nucleofection, the indel frequencies from the endogenous off-target sites are much lower than that measured by SURRO-seq (Fig. 4e, Supplementary Data 4). Despite that, there is a generally good correlation between the SURRO-seq and the endogenous editing results (Pearson's R = 0.88–0.94, Fig. 4f, Supplementary Data 4). Most importantly, both the number of OTs with significantly detectable indels and the indel frequency were significantly reduced in cells edited with the high-fidelity HiFi-Cas9, which corroborates with previous finding and highlights the importance and necessity of using high-fidelity Cas9 variants in gene therapy application to minimize the off-target effect[44]. Collectively, we demonstrated

Fig. 4 Validation of endogenous OTs in five human cell lines by deep sequencing. a Overview of RGNs, SURRO-seq identified Sig. OTs and NS. OTs selected for validation. **b** Schematic illustration of the experiments. RNP, ribonucleoprotein; HEK, HEK293T cell; Fib, human skin-derived fibroblasts. **c** Dot plot of on-target indel frequencies (indel reads/total reads %) in the CRISPR RNP edited cells. Indel frequency values were showed in Supplementary Data 4. **d** Heatmap summary of the RGN OTs evaluated by SURRO-seq and deep sequencing in five human cells lines. Indel frequency values were showed in Supplementary Data 4. **e** Example of indel frequencies of four OTs for RGN11189 measured by SURRO-seq and by deep sequencing of RGN edited human cell lines. **f** Scatter plots of indel frequencies for 7 on-target and 23 off-targets (referred to 4a), measured by SURRP-seq and by amplicon sequencing of the corresponding endogenous loci in RNP nucleofected cells (HEK293T and Fibroblasts). Extended plots for all cells can be found in Supplementary Data 4 (sheet 4.9).

that SURRO-seq is a sensitive method for high throughput targeted evaluation of RGN off-targets in cells.

Effect of mismatch positions, mismatch types, and free binding energy on RGN specificity. The RGN off-target data generated by SURRO-seq also allow us to explore how the genomic context affects RGN off-target cleavage. Analysis of indel frequencies of the 753 Sig. OTs (both LIOT and HIOT) showed that indel frequencies were significantly decreased in OTs with more mismatches (Fig. 5a, Supplementary Fig. S10), which corroborates with previous findings and is expected[32,47,48]. While it is generally believed that OTs with 3–4 mismatches are unlikely to be cleaved by RGNs, our results showed that there exists a great heterogeneity in mismatch tolerance among the 110 RGNs and between the OTs with same number of mismatches (Fig. 5a). This phenomenon was also observed for the VEGFA-T2 RGN[17] and validated by SURRO-seq (Fig. 2). We speculated that the heterogeneity of indel frequencies between OTs with same number of mismatches in our dataset is caused by the positions and types of mismatches between the RGN spacer and the target site. It has been well characterized that the CRISPR is less tolerated to mismatches in the PAM-proximal 10–12 nucleotides[49–51].

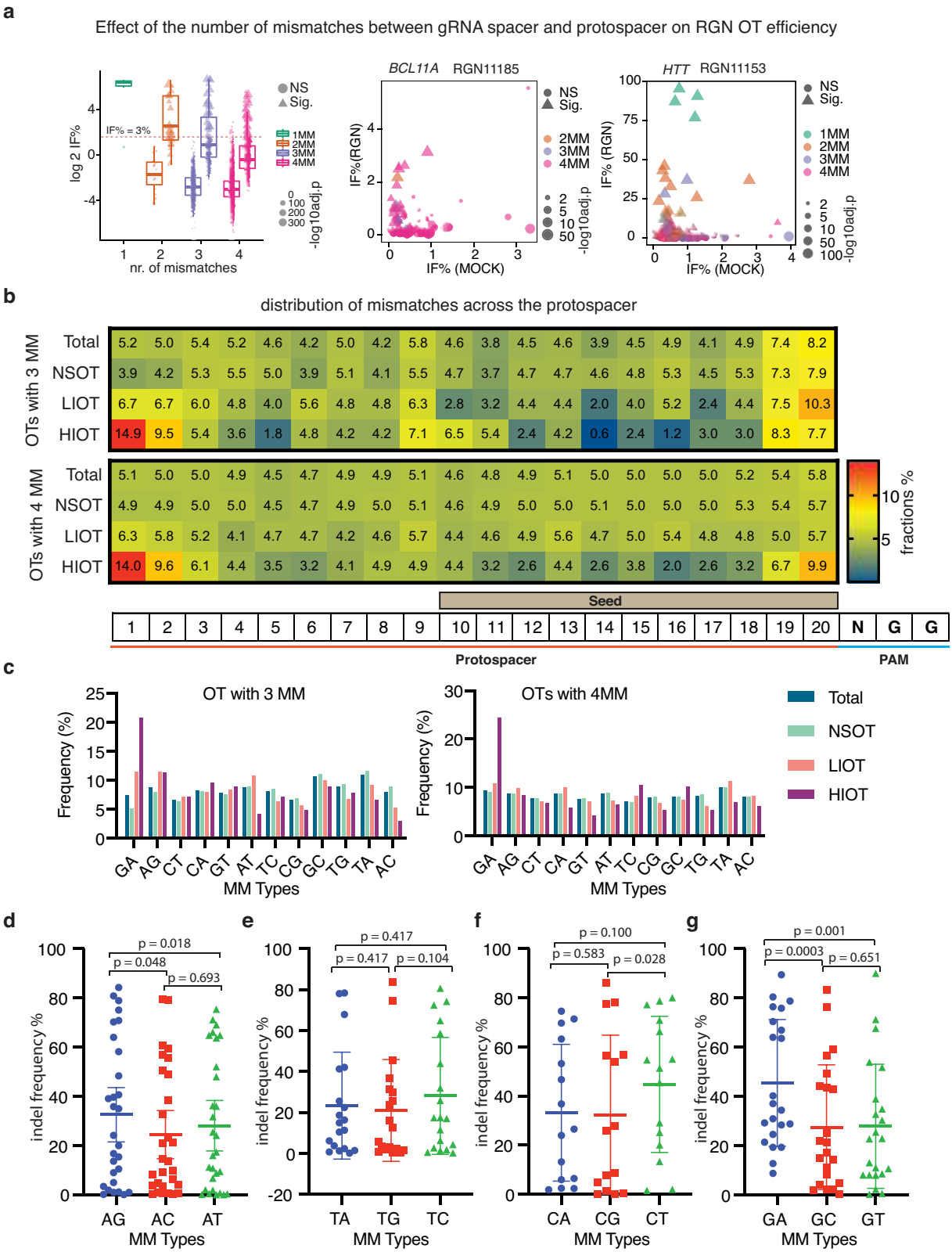
To address if this position-dependent mismatch tolerance contributes to the heterogeneity of OTs, we analyzed the frequency of mismatches occurred in each position of the 20-nt protospacer region for all RGN OTs with 3 or 4 mismatches (Supplementary Data 5). There is a significant (Hypermetric test P value < 0.05) over-representation of mismatches occurred at N1 and N2 positions (the two most PAM-distal nucleotides) and an under-representation of mismatches occurred at the N12-N18 (PAM-proximal seed regions) in OTs with significantly detectable. Interestingly, our analysis also revealed that RGNs seem to be more tolerant to mismatches at the N19 and N20 position as compared to other nucleotides of the seed region (Fig. 5b).

We next analyzed the effect of mismatch types on RGN OTs. Twelve types of mismatches can occur between RGN and off-target sites (Supplementary Fig. S13). To provide a simple description, we only refer to mismatches between the gRNA spacer and the protospacer sequences (the non-targeting strand). Cumulative studies have suggested that RGN exhibits different tolerances to the different types of mismatches. Once such example is the GA mismatch, which generates a wobble base pair (rG:dT) between gRNA spacer and the complementary strand DNA. We analyzed the frequencies of these 12 mismatch types in the OTs from LibB (Supplementary Data 5). Our results showed that GA mismatch (also AG mismatch but to a lesser degree, for OTs with 3MM) was significantly (hypergeometric test p-value < 0.05, compared to NS OTs or total OTs) enriched in the OTs with significantly detectable indel (Fig. 5c). To further validate the effects of mismatch and mismatch type on CRISPR specificity, we generated a small SURRO-seq library (libC) carrying artificially generated OTs with all possible combinations of one mismatch for five RNGs (Supplementary Fig. S14). SURRO-seq-based libC further showed similar findings about the effect of mismatch position and type (GA and AG wobble pairs) on RGN specificity (Fig. 5d–g, Supplementary Fig. S14, Supplementary Data 5). Notably, the RGN 11157, which is low in GC and particularly G content, seems to be less tolerant of mismatches (Supplementary Fig. S14).

Since a large number of regression-based, machine-learning and deep-learning models have already been developed for in silico prediction of RGN off-targets, we benchmarked six RGN OT scoring models: MIT[52], deepCRISPR[53], Cutting Frequency Determination (CFD) score[47], CROP-IT[54], CCtop[55], and CRISPRoff[56] with the LibB Sig. OTs data. Our results showed that CRISPRoff (Pearson R = 0.50, Spearman R = 0.48, p-value < 0.001) outperforms the other four RGN OT scorers (Supplementary Fig. S15). Compared to the other OT prediction scorers, the CRISPRoff has included the free energy feature. We hypothesized that the energy features are the main contributing factors to the RGN OT prediction. To prove that we next analyzed the correlation between the OT indel efficiencies and position-weighted binding energy between gRNA and the (off-) target DNA (ΔG_H), the free energy of the DNA duplex (ΔG_O), or the folding energy of the gRNA only (ΔG_U) as defined by the CRISPRoff energy model[56]. Our results showed that there is significant correlation between indel efficiencies of Sig. OTs and the ΔG_H (Pearson's R = 0.53, Spearman's R = 0.52, p-value < 0.001), ΔG_O (Pearson's R = 0.25, Spearman's R = 0.26, p-value < 0.001), and ΔG_U (Pearson's R = 0.23, Spearman R's = 0.30, p-value < 0.001) (Supplementary Fig. S16). Notably, the feature of gRNA and the (off-)target DNA binding energy (ΔG_H) yields even high correction compared to the seven RGN off-target predicting, corroborating that ΔG_H is the major energy feature determining RGN OT effect[56]. Our data collectively highlighted the importance of mismatch positions (where), mismatch types (which), free binding energy (a combined feature of mismatch positions and types) on RGN off-target effect.

Discussion

In conclusion, we validated and demonstrated that surrogate off-target site-based capturing of RGN cleavage can be used for massively targeted evaluation of SpCas9-based RGN off-targets in cells. Similar to our approach, Fu et al., very recently reported a similar library-based approach of which a pair of on-target and off-target surrogate site was introduced to allow direct comparison of on and off target efficiencies, as well as understanding effect of sequence contexts on RGN specificity[57]. Several generations of CRISPR-derived technologies have successfully reported for gene editing purposes. These include the different classes and types of CRISPR Cas systems and variants, such as SpCas9, SaCas9, NmCas9, Cas-X, Cas-Y, Cas12a, Cas13 (just to mention a few)[51,58–62]. Most importantly, many CRISPR-Cas9 derived genetic and epigenetic editing tools have been developed by fusing the dead Cas9 (dCas9) protein or nickase Cas9 (nCas9) protein to effector proteins or protein domains. By fusing dCas9 or nCas9 to deaminases, the CRISPR-Cas9 system have been repurposed for targeted base editing[62–65], such as A-to-G substitution (ABE), C-to-T substitution (CBE), C-to-G substitution (GBE). For a comprehensive overview of the CRISPR-derived



base editors, we refer readers to the review paper by Porto et al.[66]. Off-targets effects have been reported for these CRISPR base editors. Although not showed in this study, we have demonstrated that high throughput quantification of base editing efficiency can also be achieved using such a surrogate library in cells (BioRxiv. https://doi.org/10.1101/2020.05.20.103614). Recently, we further demonstrated that this surrogate library approach can be used to evaluate PAM compatibility in human cells[67]. We anticipate that SURRO-seq could be adapted to evaluate off-targets of other DNA editing RGN systems, including prime editing[68]. Unlike genome-wide in-cell or cell-free OT screening methods, SURRO-seq is limited by its pre-selected potential OTs for evaluation. In this study, we select potential off-target sites for the therapeutic RGNs based on the number of mismatches

Fig. 5 Effects of mismatch number, position and type on RGN off-target activity. a Box-and-whisker plot of log 2 indel frequency for RGN OTs evaluated by SURRO-seq in LibB. Data are presented as values representing the median (line within the box), the interquartile range (length of the box), the 75 and the 25th percentiles (whiskers above and below the box) of the indel frequencies. Sites were grouped based the number of mismatches (MM), plotted according to significance and log10 adj. *p*-values (Benjamini and Hochberg (BH)-adjusted Fisher's exact test (two-sided)). NS, RGN OTs with not significantly detectable indels; Sig. RGN OTs with significantly detectable indels. One mismatch (NS, $N = 0$ biologically independent RGN OTs; Sig, $N = 6$ biologically independent RGN OTs), two mismatches (NS, $N = 26$ biologically independent RGN OTs; Sig, $N = 49$ biologically independent RGN OTs), three mismatches (NS, $N = 501$ biologically independent RGN OTs; Sig, $N = 140$ biologically independent RGN OTs), and four mismatches (NS, $N = 5860$ biologically independent RGN OTs; Sig, $N = 558$ biologically independent RGN OTs). **b** Heatmap presentation of the fraction of mismatches occurred in each position of the gRNA for OTs in LibB, grouped based on total OTs, NSOTs, LIOTs and HIOTs. **c** Bar plot of appearance frequencies of each type of mismatches occurred in the different groups of RGN OTs in LibB. **d**–**g** Dot plots of indel frequencies for OTs with one mismatch measured in LibC. One-way pair-wise ANOVA analysis was performed for A type mismatches (**d**, $N = 30$ biologically independent mismatch sites), T type mismatches (**e**, $N = 20$ biologically independent mismatch sites), C type mismatches (**f**, $N = 18$ biologically independent mismatch sites), and G type mismatches (**g**, $N = 28$ biologically independent mismatch sites). Data are presented as mean values +/− SD. Indel frequency values can be found in Supplementary Data 5 (5.4–5.7).

(allowing up to 4 mismatches). Previous findings from e.g., GUIDE-seq, CIRCLE-seq, as well as validated by SURRO-seq (Fig. 2g) in this study reveal that significantly detectable indels caused by CRISPR-Cas9 can be found in some off-targets with 5 or 6 mismatches. Furthermore, CRISPR off-targets have also been found in genomic sites containing insertions (DNA bulge) or deletions (RNA bulge) compared to the RGN guide sequences[69]. With the development and improvement of in silico RGN off-target prediction tools, just to mention a few e.g., CRISTA[70], CRISPRoff[56], deepCRISPR[53], CNN_std[71] and Elevation[72], it is advisable that the selection of potential RGN OTs for SURRO-seq evaluation should be predicted and selected with these tools. Conversely, more RGN OT data generated with e.g., SURRO-seq or other comparable methods will facilitate the further improvement of these RGN off-target prediction tools. SURRO-seq offers a sound complementary approach to the genome-wide OT screening methods for further high throughput validation of the RGN OTs.

The CRISPR-Cas9 gene editing technology has been in development for a full decade. We still do not completely understand factors affecting its specificity. These specificity-affecting factors include the gRNA-independent binding of the Cas9 protein to DNA, the number/type/position of mismatches between gRNA spacer and the target site, the epigenetic state (DNA methylation and chromatin accessibility), the expression level and duration of the Cas9 protein and gRNA in cells, and the usage of alternative PAMs. Our results suggest that there is a great heterogeneity in term of the specificity among different RGN gRNAs. Corroborating with previous findings[32,33], CRISPR-Cas9 is less tolerant to mismatches at the seed region (N10-N20). Our data further showed that mismatches at the two upstream PAM proximal position (N19 and N20) were more tolerated than other nucleotides of the seed region. This site-dependent effect could be explained by our recent binding energy model about the effect of sliding PAMs on CRISPR-Cas9 specificity[67]. Indeed, when performing benchmarking of the different CRISPR-Cas9 off-target prediction tools with our data, our results also showed that energy-based predictors out-performed other tools in their accuracy of predicting true off-targets. The energy feature is also in agreement with our finding that Wobble base pair (G-U), which still can provide strong binding between the gRNA and target DNA strand, is tolerated. We therefore recommend the use of energy-based tools for in silico prediction of CRISPR potential off-targets, while future further improvements of their prediction outcome should be achieved with high quality off-target data and the integration of better energy features.

Substantial off target indels were observed for some OTs evaluated in this study when conducted in cell line expressing high level of the wild type SpCas9 protein. However, the level of indels were significantly reduced when the SpCas9 was transiently expressed in cells by RNP delivery. Most importantly, with high fidelity SpCas9 variant (HiFi-SpCas9)[44], our results showed that near all off-target indels could not be significantly detected. Thus, our results strongly indicate that high fidelity SpCas9 variants should be used to its largest extend to avoid any potential adverse effect caused by off-target cleavage. This is particularly important when the CRISPR-Cas9 technology is used for gene therapy, both ex vivo and in vivo deliveries. One remaining major concern of CRISPR gene therapy is the off-target effect leading to oncogenesis due to off-target in cancer genes. Selection of high-fidelity Cas9 variants, carefully design of gRNA with less likelihood of introducing off-target indels in cancer genes, and experimentally validate these potential off-target sites RGN-edited cells are important for lowering the risk of detrimental off-targets in clinical application of RGN. While RGN off-target screening methods, such as GUIDE-seq, DISCOVER-seq, SITE-seq and CIRCLE-seq (also see Supplementary Fig. S1) can be used for genome-wide unbiased detection of RGN off-targets, SURRO-seq overcome the unmet need of high throughput and targeted evaluation of RGN OTs in cells. Our method provides the following four methodological advantages: (1) Scalable. The SURRO-seq library can be generated from a few hundred OTs to over 10,000 OTs. Unlike other methods, SURRO-seq can be used to evaluate hundreds of RGNs in cells simultaneously. (2) Direct evaluation of indels. SURRO-seq directly quantifies the RGN introduced indels at the surrogate off-targets by comparing RGN edited and MOCK cells. (3) High sensitivity. For SURRO-seq, each OT site can be sequenced with a very deep coverage. And direct comparison of indels in RGN and MOCK cells further allow us to sensitively detect OTs with significant indels, and particularly OTs with low indel rate. (4) Clinical significance. SURRO-seq allows us to target evaluate if RGN introduces indels in clinically relevant genes such as cancer genes. However, we also highlight some limitations of SURRO-seq which require further improvement. Each synthetic SURRO-seq oligonucleotide is 170 nt. Synthetic errors introduced in the DNA oligonucleotide library could cause dropout of some OTs after data filtering. Technological improvements in DNA synthesis will overcome this limitation. Alternatively, a two-step cloning strategy can be applied to overcome the length and error-rate limitations of synthetic oligonucleotide pool (Supplementary Fig. S17). First, smaller synthetic oligonucleotides are generated, which contain (a) Two PCR primer binding sites; (b) Two *BsaI* sites (for step1 cloning); (c) RNA spacer; (d) Two *BsmBI* sites (for step 2 cloning) and e. the corresponding surrogate off-target sites. The PCR-amplified surrogate DNAs are cloned to the lentiviral backbone plasmid LentiU6-LacZ-GFP-Puro (BB) (Addgene #170459). Second, the gRNA scaffold stuffer fragment is cloned into the plasmids

generated from Step 1 by Golden-Gate Assembly (*BsmBI*). Another limitation of the technology is caused by the high (average of 1–2%) PCR and/or deep sequencing-induced indel rates observed in wild-type cells. This has limited the detection of potential RGN OTs with low indel frequency. This limitation could be overcome by using improved high-fidelity PCR polymerases and high-fidelity deep sequencing.

As demonstrated in this study, targeted evaluation of RGN OTs in cells with SURRO-seq enables us to investigate the DNA sequences, types of mismatches, and the thermodynamic features on RGN off-target activity. Early studies had shown that other features outside the gRNA (on-target or off-target) sequences such as epigenetic features (e.g., chromatin accessibility, DNA methylation) and gene activity could affect RGN activity and specificity[73–75]. This might partially explain the variations of on-target and off-target activities for those sites observed in the five different cell types (Fig. 4). The surrogate OTs in the SURRO-seq library are randomly inserted in the genome and might not fully capture the epigenetic effects on on-target and off-target RGN activities. Other factors, such as the different presence and preference of DNA double-strand break repair machineries between these cell types (reviewed by Meyenberg M. et al.[76]), could also contributed to RGN activity variations, which however should be addressed in future studies. Although not investigated in this study, we expect that the SURRO-seq method could be used to investigate the effect of e.g., epigenetic factors and DNA repair enzymes on RGN activities (Supplementary Fig. S18). For instance, the SURRO-seq lentiviral library can be stably integrated into wild type (WT) cells. Then, these SURRO-seq WT cells are subjected treatments such as epigenetic modifying molecules (e.g., 5-AZA for DNA demethylation, TSA for histone acetylation), depletion of epigenetic modifying enzymes (e.g., DNA methyltransferases DNMT3A/DNMT3B, histone acetyl transferases P300/CBP), depletion of DNA repair proteins (e.g., DNA ligase 4, XRCC4, MRE11). Thus, SURRO-seq provides an attractive tool for studying factors affecting on-target[29] and off-target RGN activities.

In conclusion, we report a high throughput method for targeted evaluation of CRISPR-Cas9 off-targets in cells. The SURRO-seq offers a great complementary method to the existing tools for CRISPR-Cas9 off-target evaluation, off-target data generation, improvement of prediction, understanding of off-target effect, and facilitate the applications of CRISPR-based gene editing tools in clinical applications.

Methods

Cell culture. Human embryonic kidney (HEK293T), primary human skin-derived fibroblasts (Fib), U2OS, SKOV-3, and PC9 cells were cultured in DMEM media containing 10% fetal bovine serum (FBS) and 1% penicillin–streptomycin in a tissue culture incubator at 37 °C with 5% CO_2. PCR mycoplasma detection kit (cat no. PM008, Shanghai Yise Medical Technology) was routinely used to test the mycoplasma contamination. The cells used in this study have given negative results in mycoplasma contamination test. SpCas9-expressing HEK293T (HEK293T-SpCas9) cells were generated by a PiggyBac transposon system followed by selection in the presence of 50 μg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were transient transduced with pPB-TRE-spCas9-Hygromycin vector and pCMV-hybase vector with a 9:1 ratio to generate SpCas9-expressing HEK293T.

Vector construction. The LentiU6-LacZ-GFP-Puro (BB) vector was generated by our group previously (Addgene ID: 170459). This plasmid can also be acquired from the Luo lab (https://dream.au.dk/tools-and-resources).

SURRO-seq library design. Each SURRO-seq oligo consists of a BsmBI recognition site "cgtctc" with 4 bp specific nucleotides "acca" upstream, following the GGA cloning linker "aCACC", one bp "g" for initiating transcription from U6 promoter, 20 bp gRNA sequences of "gN20", 82 bp gRNA scaffold sequence, 37 bp surrogate target sequences (10 bp barcode sequences, 20 bp protospacer and 3 bp PAM sequences, 4 bp downstream sequences), the downstream linker "GTTTg" and another BsmBI binding site and its downstream flanking sequences "acgg".

The SURRO-seq pool was designed as follows: (1) LibA contains 11 on target and corresponding 170 off target gRNAs from three published off-target detection methods (T7E1, GUIDE-seq, CIRCLE-seq); (2) LibB contains 110 gRNAs retrieved from published studies, which we expect to have sequence characteristics representative of gRNAs in gene therapy applications (cancers, PD-1, DMD, β-hemoglobinopathies, SCD, CCR5, HTT, CEP290). (3) We predicted off-target sites of each gRNA with FlashFry (v 1.80) and retrieved potential off-target with up to 4 bp mismatches in human genome hg19. (4) For each surrogate site, we added 10 bp barcode (fixed "AC" for the first two nucleotides + 8 bp Unique molecular identifiers (UMIs) sequences) to the upstream sequence of each selected gRNA, constructed the surrogate target sequence as 10 bp barcode + 23 bp gRNA (include PAM) + 4 bp downstream = 37 bp; (5) Off target sites with BsmBI recognition site were discarded, because of GGA cloning; (6) LibC contains surrogate sites with all possible 1 bp mismatch for five RGNs. The oligo pools were synthesized in Genscript® (Nanjing, China), and all sgRNA sequences and their oligos are provided in the Supplementary Data 2, 3, and 5.

Construction of SURRO-seq plasmid library. PCR amplification was used to amplify the 170-nt oligonucleotide pool. Firstly, the SURRO-seq oligos diluted to 1 ng/μl followed by PCR amplifications using the primers: SURRO (BsmBI GGA)-F and SURRO (BsmBI GGA)-R (Supplementary Data 6). The PCR reaction was carried out using PrimeSTAR HS DNA Polymerase (Takara, Japan) following the manufacturer's instruction.

The PCR products of SURRO-seq oligos were then used for Golden Gate Assembly (GGA) to generate the plasmids library. 36 parallel GGA reactions were performed, and the ligation products were pooled into one tube. Transformation was then carried out using chemically competent DH5α cells. For each reaction, 10 μl GGA ligation product was transformed in to 50 μl competent cells and all the transformed cells were plated on one LB plate (15 cm dish in diameter) with Xgal, IPTG and Amp selection. High ligation efficiency was determined by the presence of very few blue colonies. To ensure that there was sufficient coverage of each surrogate vector in the oligonucleotide library. For one library containing 12,000 synthetic oligos, 42 parallel transformations were performed, and all the bacterial colonies were scraped off and pooled together for plasmids midi-prep (PureLink™ HiPure Plasmid DNA Midiprep Kit, ThermoFisher Scientific). For small library, equal ratio reduction can be adjusted accordingly. For NGS-based quality quantification of library coverage, midi-prep plasmids were used as DNA templates for PCR amplifications, followed by gel purification and NGS sequencing.

SURRO-seq plasmid library lentivirus packaging. Supernatants containing lentiviral particles were produced by transient transfection of HEK293T cells using PEI 40,000 (Polyethylenimine Linear, MW 40,000). For 10 cm dish transfection, the DNA/PEI mixture contains 13 μg pLenti-SURRO-seq vectors, 3 μg pRSV-Rev, 3.75 μg pMD.2 G, 13 μg pMDGP-Lg/p-RRE, 100 μg PEI 40,000 solution (1 μg/μl in sterilized ddH_2O) and supplemented by Opti-MEM without phenol red (Invitrogen) to a final volume of 1 mL. The transfection mixture was pipetted up and down gently several times, and further incubated and kept at room temperature (RT) for 20 min. The transfection complex was added to 80%-confluent HEK293T cells in a 10-cm dish containing 10 ml of culture medium. After 48 h viral supernatant was harvested and filtered with a 0.45 μm filter. Polybrene solution (Sigma–Aldrich) was added to the crude virus solution to a final concentration of 8 μg/mL. The crude virus solution was aliquoted into 15 mL tubes (5 mL/tube) and store in −80 °C freezer until used.

Lentivirus titer quantification by flow cytometry (FCM). The LentiU6-LacZ-GFP-Puro (BB) vector expresses an *EGFP* gene. The functional titer of our lentivirus prep was assayed by FCM (Supplementary Fig. S19). Briefly, (1) Day 1: Seed HEK293T cells to a 24-well plate. (2) Day 2: Transduce cells at 60~80% confluence. Before transduction, determine the total number of cells using one well of cells. The remaining wells were changed to fresh culture medium containing 8 μg/mL polybrene. A gradient volume of crude virus was added to each well and mix gently; (3) Day 3: Change to fresh medium without polybrene; (4) Day 4: Transduced cells were harvested with trypsin and washed with PBS twice. The suspended cells were fixed in 4% formalin solution at RT for 20 min. Cell pellet was washed with PBS twice and re-suspended in PBS solution, followed by FCM analysis. FCM was performed using a BD LSRFortessaTM cell analyzer with at least 30,000 events collected for each sample in duplicates. The FCM output data was analyzed by the software Flowjo vX.0.7. Percentage of GFP-positive cells was calculated as: $Y\% = N_{GFP\text{-}positive\ cells} / N_{total\ cells} \times 100\%$. For accurate titter determination, there should be a linear relationship between the GFP positive percentages and crude volume. The titer (Transducing Units (TU/mL) calculation according to this formula: $TU/mL = (N_{initial} \times Y\% \times 1000) / V$. V represents the crude volume (μl) used for initial transduction.

SURRO-seq library lentivirus transduction. HEK293T-SpCas9 cells were cultured in D10 medium with 50 μg/ml hygromycin throughout the whole

experiment. For SURRO-seq library transduction, at Day −1: 2.5×10^6 cells per 10 cm dish were seeded. For a 12 K SURRO-seq library, transductions were performed in 10 replicates to reach 4000X coverage. For each group, one plate was used for cell number determination before transduction and another plate was used for drug-resistance (puromycin) test control. The remaining 10 plates were used for the SURRO-seq lentivirus library transduction (transduction coverage per gRNA exceeds 4000X of a 12 K library); 2) Day 0: We first determined the approximate cell number per dish. This was used to determine the volume of crude lentivirus used for transduction using a multiplicity of infection (MOI) of 0.3. The low MOI (0.3) ensured that most infected cells receive only 1 copy of the lentivirus construct with high probability. The calculation formula is V = N × 0.3 / TU. V = volume of crude lentivirus used for infection (ml); *N* = cell number in the dish before infection; TU = the titter of crude lentivirus (IFU/mL). The infected cells were cultured in a 37 °C incubator; 3) Day 1: 24 h after transduction, the cell was passaged at a ratio of 3 folds. 4) Day 2: The transduced cells were cultured in D10 medium containing 50 μg/ml hygromycin, 1 μg/mL puromycin, and 1 μg/mL doxycycline to induce Cas9 overexpression. 5) The transduced cells were spitted every 2~3 days when cell confluence reaches up to 90% at a ratio of 1:3. Cells from day 10 were harvested for further genomic DNA extraction. Parallel experiments were performed using wildtype HEK293T cells as MOCK controls.

PCR amplicons of surrogate sites from cells. Genomic DNA was extracted using the phenol-chloroform method. Then the genomic DNA was purified and subjected to SURRO-seq PCR. The PCR primers were SURRO-NGS-F and SURRO-NGS-R1 (Supplementary Data 6). In this study, 5 μg genomic DNA was used as template in one PCR reaction which contained approximately 7.6×10^5 copies of surrogate construct which covered about 63 times coverage of a 12 K SURRO-seq library. For each PCR reaction, briefly, 50 μl PCR reaction system consists of 5 μg genomic DNA, 0.5 μl PrimeSTAR polymerase (2.5 U/μl, R010A, Takara Bio), 4 μl dNTP (2.5 mM each), 10 μl PrimeSTAR buffer (5X, Mg^{2+} Plus), 2.5 μl SURRO-NGS-F primer (10 μM), 2.5 μl SURRO-NGS-R1 primer (10 μM), and supplemented with ddH2O to a final volume of 50 μl. PCR reaction was carried out using a PCR thermal cycle using the following PCR program: 1 cycle at 94 °C for 2 min; 25 cycles of 94 °C for 20 s, 58 °C for 30 s, 68 °C for 45 s; and 1 cycle at 68 °C for 7 min. It is important that the PCR cycles were kept below 25 according to our optimizations. For a 12 K library, 32 parallel PCR reactions were performed to achieve approximately 2016 times coverage of each surrogate construct. Then the PCR products were purified by 1.5% gel and mixed with equal amounts and deep sequenced with a DNBseq sequencer.

Synthetic gRNAs. All synthetic gRNAs used for validation of OTs were chemically modified to increase stability in cells and synthesized by Synthego Co. (California).

RNP nucleofection. The CRISPR RNP was delivered into cells by nucleofection. For one nucleofection, 6 μg SpCas9 protein (Cat# 1081059, IDT) and the 3.2 μg synthetic gRNA (Synthego) was mixed in a PCR tube by pipetting and incubated at room temperature for at least 10 min and maximum 1 h. Then 200,000 suspended cells were gently resuspended cells in 20 uL nucleofection buffer (OptiMEM) by pipetting up and down. The cells and RNP complex were then transferred to a 4D-Nucleofector 16-well nucleocuvette strip (Catalog #: AXP-1004, Lonza). The samples should cover the bottom of the wells, and any presence of air bubble must be avoided. Nucleofection was performed with program CM-138. Immediately after electroporation, prewarmed culture media was added to the cells (180 μL per well of the Nucleocuvette strip). The cells were then transferred into one well of a 12-well cell culture plate with prewarmed medium. 48 h after transfection, cells were harvested for amplicon PCR and deep sequencing.

Deep amplicon sequencing. The on-target and off-target sites were amplified by PCRs. All primers used for PCR were showed in Supplementary Data 6. The amplicons were subjected to deep sequencing on the MGISEQ-2000 sequencer (MGI, China). All the samples were subjected to pair-ended 150 bp deep-sequencing on MGISEQ-2000 platform.

Raw data processing. FastaQC-v0.11.3 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and fastp-v0.19.6 (https://github.com/OpenGene/fastp) with default options were used for data quality control and filtering with the default parameters. The pair-end data was assembled using FLASH-v1.2.11 (http://www.cbcb.umd.edu/software/flash). BWA-MEM-v0.7.17 with default options was used to map the assembled data to the designed oligos sequence to preliminarily distinguish the data of each surrogate site.

Data filtering. The pysam module of Python-3.8 was used to split the aligned data according to the site number of the chip, and the reads of different sites were obtained. Then, we used three steps of strictly controlling parameters to filter the data of each site. Firstly, according to the structure of the chip, g + gRNA (20 bp) + scaffold (82 bp) + barcode (10 bp) + GTTT should remain unchanged at the beginning and end of each site. Then, to remove the chip synthesis errors, the pseudo editing sequences found in WT group were removed from spcas9 group. Finally, to remove the interference of sequencing errors on the data, the extracted sequence of each site was re-aligned to the reference sequence, and the 1 bp indel on N1-N14 and N22-N27 of surrogate (27 bp) sequence were removed. The above three filtering steps were completed with julia-1.5.3 language.

Fisher's exact test and statistical analysis. To obtain stable and effective off-target efficiency, false positive results must be excluded. We used the number of reads of indel and no indel in spcas9 group and WT group to form a 2 × 2 matrix. Fisher's exact test was used to confirm whether the editing of each site was effective. To reduce False Discovery Rate (FDR), all *p*-values were corrected by BH (Benjamin and Hochberg) method. Next, we used strict parameters (Total read numbers(spCas9) ≥ 32, Indel read numbers (spCas9) ≥ 5, Indel Frequency (IF%) (WT) ≤ 25) to filter off-target efficiency with bias. Then we used parameters (Fold Change (FC) > 2, *p*-value (adjusted by BH) < 0.05) to divide the off-target data set into two parts for downstream analysis. The calculation formula of indel efficiency is as follows:

$$Indel\ Frequency(\%) = \frac{Indel\ read\ numbers}{Total\ read\ numbers} \times 100$$

And fold change is as follows:

$$FC = \frac{Indel\ efficiency\,[spCas9]}{Indel\ efficiency\,[WT]}$$

Fisher's exact test (two-sided, adjusted by BH) and other statistical analysis were performed in R-4.0.3. Visualization was completed by R and excel.

Statistics & reproducibility. In this study, Fisher's exact test (two-sided, adjusted by Benjamin and Hochberg) was used for testing the significance of RGN OT indels captured by SURRO-seq. One-way pair-wise ANOVA, unpaired T test (two-sided), paired T test (two-sided), Pearson and Spearman correlations (t-distribution testing for coefficient) were used for other statistical testing and indicated in figure legends. A *p* value less than 0.05 is considered statistically significant. For all RGN OTs, we filtered OTs with low read coverage (total NGS reads less than 32). In SURRO-seq LibB, we filtered OTs potentially affecting cell growth based on enrichment: fold changes (SpCas9 NGS reads/MOCK NGS reads) > = 2 or depletion: fold changes (MOCK NGS reads/SpCas9 NGS reads) > = 2. In this study, no statistical method was used to predetermine sample size. The experiments were not randomized. The Investigators were not blinded to allocation during experiments and outcome assessment.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All NGS data generated by this study have been shared via the CNGB public data depository with the following accession numbers: CNP0001979 and CNP0002648, and the Gene Expression Omnibus (GEO) data depository with the following accession number: GSE206347. A complete list of 702 NGS samples were summarized in Supplementary Data 7. All other relevant data supporting the key findings of this study are available within the article and its Supplementary Information files. Source data are provided with this paper.

Code availability

The computing codes for analyzing indel frequencies with the deep sequencing from SURRO-seq has been deposited to GitHub[77]. URL to the codes: https://github.com/panxiaoguang/Massively_RGN_OTs/tree/v1.0.0.

Received: 7 March 2022; Accepted: 20 June 2022;

References

1. Doudna, J. A. The promise and challenge of therapeutic genome editing. *Nature* **578**, 229–236 (2020).
2. Xiang, X. et al. Efficient correction of Duchenne muscular dystrophy mutations by SpCas9 and dual gRNAs. *Mol. Ther. Nucleic Acids* **24**, 403–415 (2021).
3. Frangoul, H. et al. CRISPR-Cas9 gene editing for sickle cell disease and beta-Thalassemia. *N. Engl. J. Med.* **384**, 252–260 (2021).
4. Esrick, E. B. et al. Post-transcriptional genetic silencing of BCL11A to treat sickle cell disease. *N. Engl. J. Med.* **384**, 205–215 (2021).

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-022-30515-0 OPEN

CRISPR/Cas9 gRNA activity depends on free energy changes and on the target PAM context

Giulia I. Corsi[1,8], Kunli Qu[2,3,8], Ferhat Alkan[1,4], Xiaoguang Pan[2], Yonglun Luo[2,5,6,7✉] & Jan Gorodkin[1✉]

A major challenge of CRISPR/Cas9-mediated genome engineering is that not all guide RNAs (gRNAs) cleave the DNA efficiently. Although the heterogeneity of gRNA activity is well recognized, the current understanding of how CRISPR/Cas9 activity is regulated remains incomplete. Here, we identify a sweet spot range of binding free energy change for optimal efficiency which largely explains why gRNAs display changes in efficiency at on- and off-target sites, including why gRNAs can cleave an off-target with higher efficiency than the on-target. Using an energy-based model, we show that local gRNA-DNA interactions resulting from Cas9 "sliding" on overlapping protospacer adjacent motifs (PAMs) profoundly impact gRNA activities. Combining the effects of local sliding for a given PAM context with global off-targets allows us to better identify highly specific, and thus efficient, gRNAs. We validate the effects of local sliding on gRNA efficiency using both public data and in-house data generated by measuring SpCas9 cleavage efficiency at 1024 sites designed to cover all possible combinations of 4-nt PAM and context sequences of 4 gRNAs. Our results provide insights into the mechanisms of Cas9-PAM compatibility and cleavage activation, underlining the importance of accounting for local sliding in gRNA design.

[1] Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen, Thorvaldsensvej 57, 1871 Frederiksberg, Denmark. [2] Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, Qingdao 266555, China. [3] Department of Biology, University of Copenhagen, Copenhagen 2200, Denmark. [4] Division of Oncogenomics, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. [5] BGI-Shenzhen, Shenzhen 518083, China. [6] Department of Biomedicine, Aarhus University, Aarhus 8000, Denmark. [7] Steno Diabetes Center Aarhus, Aarhus University Hospital, Aarhus 8200, Denmark. [8] These authors contributed equally: Giulia I. Corsi, Kunli Qu. ✉email: alun@biomed.au.dk; gorodkin@rth.dk

The bacterial CRISPR (clustered regularly interspaced short palindromic repeats)-Cas9 endonuclease has been transformed into a powerful genome-editing tool that is nowadays broadly applied in biological, agricultural and medical research[1]. Cleavage by the widely studied *Streptococcus pyogenes* Cas9 ortholog (SpCas9), hereafter referred to as Cas9, is mediated by a 20-nt segment of a guide RNA (gRNA) complementary to a target DNA sequence preceding a protospacer adjacent motif (PAM), where the Cas9 is recruited[2]. The canonical PAM of SpCas9 is 5′- 3′ "NGG". In this study we refer to this canonical PAM sequence as 5′-$N_{-1}GGN_{+1}$-3′, thus including both the GG binding motif and its 1-bp flanking sequence context (N_{-1} and N_{+1}) in the definition of PAM. Once a stable gRNA–DNA heteroduplex is formed, double-strand breaks (DSBs) are produced on the DNA by the Cas9 nuclease domains. In the case of the most broadly used SpCas9 protein, DSBs are introduced by the HNH and RuvC nuclease domains 3 nt upstream from the PAM[3].

A large number of in silico methods for gRNA design report that the cleavage efficiency of Cas9 can largely vary due to the sequence and structural properties of the gRNA[4–17]. However, despite the immense activity in the field, a comprehensive analysis linking gRNA binding patterns, free binding energy changes and unfolding free energy changes to cleavage efficiency is surprisingly still lacking.

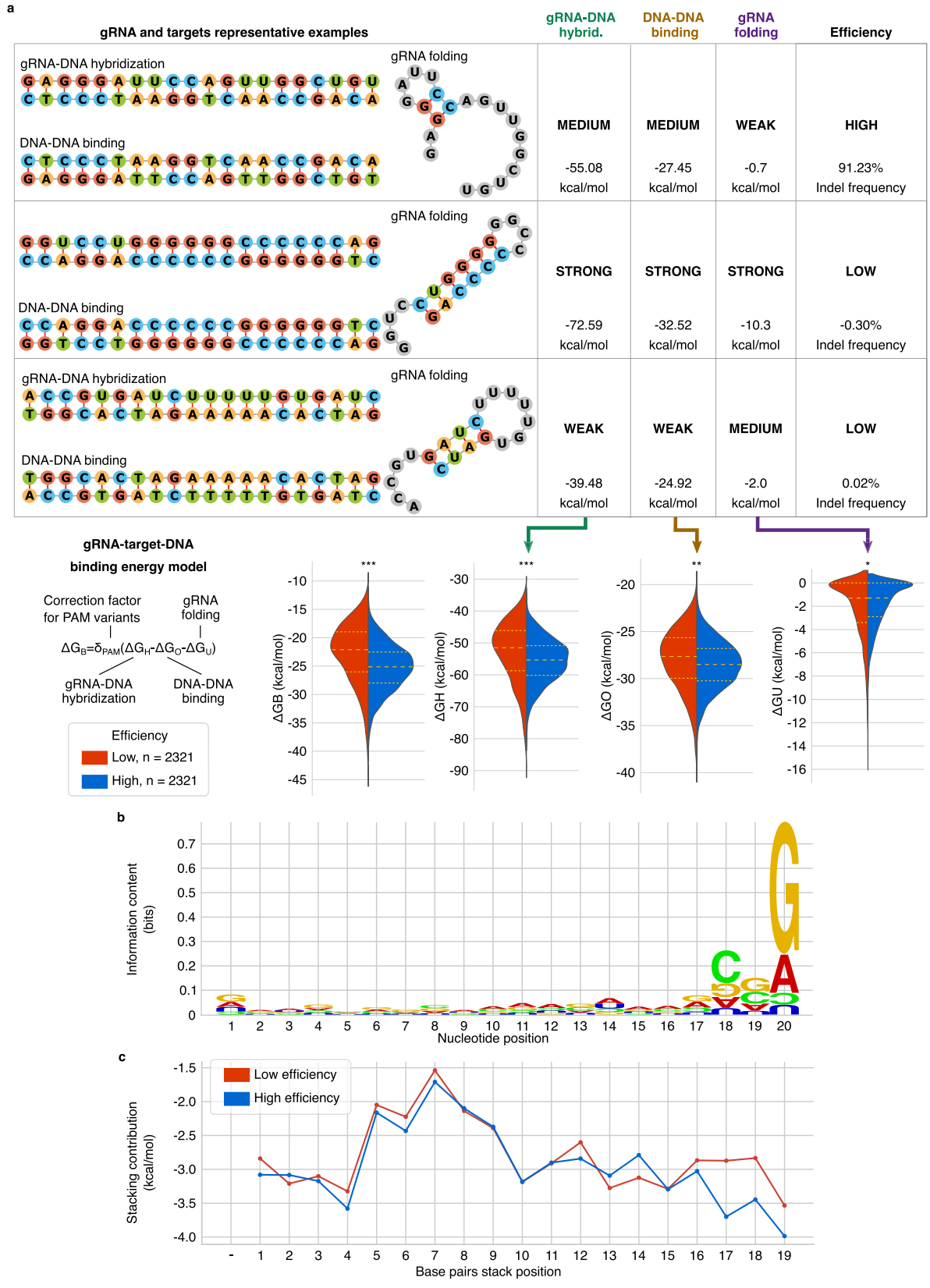
A major advance in the understanding of Cas9-gRNA binding was given by Globyte et al., who revealed that DNA interrogation by Cas9 occurs not only by random collisions with the DNA, as previously reported[18], but also by lateral diffusion. During this process, Cas9 'slides' along the DNA in a local region (≈20 nt)[19] as part of its search for a target on which the Cas9-gRNA complex can bind firmly. The study of Globyte et al. was, however, limited to the Cas9 sliding dynamics on short-distanced PAMs. Hence, the effect of Cas9 binding at sites overlapping the on-target PAM on cleavage activity remains unexplained.

Previously, we introduced the first energy-based model for Cas9-gRNA-target binding. We applied it to predict Cas9 off-target[20] and, recently, on-target activities[17], improving over other available methods. Here, we employ this energy-based model to systematically analyse the relationship between nucleotide bindings and cleavage potential on a large dataset of 11,602 experimentally validated gRNA efficiencies[14,17]. Next, this energy-based model is applied to explain the cleavage activity of Cas9 at gRNA bindings that originate from local overlapping PAMs on which Cas9 slides (local sliding PAMs). Introducing local sliding PAMs in the computation of the CRISPRspec gRNA specificity score, which is a 'competition score' that measures the ability of Cas9 to bind at the on-target while accounting for possible off-targets in the genome, leads to a better identification of gRNAs with high efficiency and low off-target potential. We validated the effect of local sliding PAMs on the cleavage efficiency with public data and in-house generated data from HEK293T cells by altering the sequence and context of Cas9 binding sites at the targets of four gRNAs. Our results further show that local sliding broadens the PAM compatibility of both wild-type and Cas9-derived variants.

Results

Cas9 on-target efficiency modelled with binding free energy changes. Modelling gRNA–DNA bindings in terms of free energy changes is a convenient approach that can be applied to evaluate both on-targets and off-targets with mismatches or bulges. Sufficient gRNA-target-DNA complementarity is essential to trigger the HNH conformational changes that activate DNA cleavage by both HNH and RuvC[21,22]. To test if binding free energy changes can explain this activation we consider the contributions from our previously established CRISPR/Cas9 binding model: $\Delta G_B = \delta_{PAM}(\Delta G_H - \Delta G_U - \Delta G_O)$[20] with $\delta_{PAM} = 1$ for 5′-$N_{-1}GGN_{+1}$-3′ PAMs. We analysed the relationship between binding free energy changes and on-target cleavage efficiency, measured as indel frequency, in a dataset of 11,602 Cas9 gRNAs binding at targets with 5′-$N_{-1}GGN_{+1}$-3′ PAMs from the datasets of Kim et al.[14] and Xiang et al.[17], merged following our previously established protocol[17]. During the pre-processing of the original 23,902 gRNAs, instances with low specificity[20] were removed to avoid biases due to global off-targets competition, and part of the data was isolated as an independent test set for later usage, accounting for data similarity (dataset 'Xiang 2021 test', see Methods and Supplementary Table 1). The gRNA–DNA hybridisation free energy change (ΔG_H) is calculated by using stacked gRNA–DNA base pairs weighted based on biochemical profiling of nuclease-dead Cas9 binding kinetics[20,23]. The ΔG_H of highly efficient gRNAs is mostly confined in a sweet spot narrow ΔG_H interval ranging between −64.53 and −47.09 kcal/mol, while more extreme values are observed for inefficient gRNAs (Fig. 1a and Supplementary Table 2). The relation of ΔG_H with the cleavage efficiency is substantially more profound compared to that of the GC-content, despite the high correlation between these two properties (Supplementary Fig. 1). A similar but less pronounced trend emerges for the target DNA–DNA opening free energy change, ΔG_O, regarded as a penalty for unwinding the DNA (Fig. 1a and Supplementary Table 3). The gRNA self-folding free energy change, ΔG_U, is included in the binding model as a penalty for unfolding gRNA structures, a process required for subsequent target recognition. Our results show that more stable gRNA self-folding structures negatively affect cleavage activity (Fig. 1a and Supplementary Table 4), complementing previous observations[24,25] within a large-scale scenario. Before binding to Cas9, the scaffold sequence attached to a gRNA can interact with the bases in the gRNA sequence. This can disrupt the optimal structure of the scaffold, whose correct folding is required to form a complex with Cas9[26]. Hence, spacers that interact with the scaffold were removed during data pre-processing whenever the structure of the full gRNA, spacer and scaffold, displayed reduced accessibility of the loops and bulges necessary to bind with Cas9 (see Supplementary Fig. 2 and Methods). The ΔG_B binding free energy change combines ΔG_H, ΔG_O and ΔG_U, to estimate the residual gRNA–DNA binding free energy change after accounting for the DNA unwinding and gRNA unfolding penalties. The ΔG_B significantly differs between low and high-efficient gRNAs, with the latter obtaining greater benefit in terms of free energy change by binding to the target (Fig. 1a and Supplementary Table 5). However, despite having extremely low (favourable) ΔG_B, gRNAs highly rich in GC content remain disadvantageous.

The ΔG_H was further detailed in the position-specific free energy change contributions of stacking gRNA–DNA base pairs and this free energy profile was compared with a sequence logo highlighting the positional nucleotide preferences of highly efficient gRNAs (Fig. 1b, c). We observe that the 3′ seed region of highly efficient gRNAs is characterised by more stable interactions (lower free energy change) with the DNA. When ignoring the estimated impact of Cas9 as the weight on the stacking free energy changes, we notice bigger discrepancies in the 5′ part of the gRNA despite the general lack of sequence variation (Supplementary Fig. 3 and Fig. 1b), supporting that these weights are important in the energy-based model. This profile of free energy changes is consistent with the notion that the differences between efficient and inefficient gRNAs are attributable to the actual binding properties near the PAM, which otherwise are 'only' observable as position-wise single nucleotide variations. The preferred nucleotides in the 3′ seed end of highly



efficient gRNAs are guanine at N19-N20 and cytosine at N18-N19, where NX refers to the position in the spacer from the 5′ end. The aversion to uracil (U) at the gRNA 3′ seed end was previously explained as a transcription deficiency, as multiple Ts on the DNA, combined with the downstream T-rich sequence in the DNA sequence of the gRNA scaffold, might trigger transcription termination by polymerase III[27]. Considering that stacking base pairs containing uracil present the lowest binding free energy change benefit (Supplementary Table 6) an additional explanation for this negative effect is possibly the poor hybridisation stability of U-rich gRNA seeds. Supporting this, the presence of up to 2 Us at the gRNA 3′ seed end, not

Fig. 1 Evaluation of gRNAs free energy change properties and seed preferences. a The binding free energy change ΔG_B is dissected in three main components: the RNA–DNA weighted hybridisation free energy change ΔG_H, the DNA opening free energy change ΔG_O, and the gRNA minimum self-folding free energy change ΔG_U. The relationship between these properties and Cas9 efficiency is presented with violin plots comparing highly efficient (top 20%, blue) and inefficient (bottom 20%, red) gRNAs. Statistical significance is computed via the Kolmogorov–Smirnov two-sample test. *P* values (left to right): 7.0E-85 (one-sided), 3.1E-52 (two-sided), 4.0E-20 (two-sided), 4.7E-05 (one-sided). $^{*}p < 0.01$, $^{**}p < 1E\text{-}10$, $^{***}p < 1E\text{-}20$. The plots are accompanied by representative examples of gRNA and DNA interactions. **b** Sequence logo of highly efficient gRNAs with position-specific background frequencies extracted from low efficient gRNAs. The sequence logo was generated with slogo[67] and styled manually. In the logo, bases are shown upside-down when their frequency in the foreground is lower than the expectation derived from the background. **c** Profile of mean base-pair stacking free energy changes in the gRNA–DNA hybrid of gRNAs with high efficiency (top 20%, blue) and low efficiency (bottom 20%, red). The stacking free energy changes are weighted according to previously defined weights estimated from nuclease-dead Cas9 binding kinetics[20,23]. The position of stacking base pairs is parallel to the corresponding nucleotide position in (**b**). Source data are provided as a Source Data file.

contiguous to the gRNA scaffold and not sufficient to trigger polymerase III termination[28], is also linked to low efficiency (Supplementary Fig. 4). Altogether, these results reveal that binding and folding interactions between nucleic acids have a significant role in Cas9 cleavage activation, which requires energetically favourable interactions with tight binding at the gRNA 3′ seed region.

Low gRNA-target hybridisation free energy change results in increased activity at off-targets. In a study investigating the impact of bulges in gRNA–DNA interactions, Lin et al. revealed that DNA bulges are better tolerated at off-target sites with high GC content and that bulged off-targets can even surpass fully complementary interactions in terms of efficiency[29]. To demonstrate this, Lin et al. tested the indel frequency of four gRNAs with different GC content after creating DNA bulges at each position of the target sequences by systematically removing one base at a time in the gRNAs. To explain the efficiency increase recorded at bulged interactions, we examined the affinity between the key free energy change component ΔG_H and the sweet spot. The remaining two components, ΔG_U and ΔG_O, presented only minor variations. There were 19 gRNA-target bindings with at least 10% cleavage frequency, which we analysed, i.e., interactions with DNA bulges present at positions at which they are tolerated[29]. The gRNA with the highest GC content (=70%) has an indel frequency of 30% at the fully complementary target, and the ΔG_H of this binding (−66.85 kcal/mol) falls outside of the sweet spot. Strikingly, six bulged off-targets of this gRNA have increased efficiency (indel frequency = 44.77 ± 8.33%) and in five of these the free energy change penalty caused by the bulge (between +6.58 and +15.38 kcal/mol) is sufficient to enter the preferential ΔG_H range (Supplementary Table 7). On the contrary, a gRNA with medium GC content (=50%) and ΔG_H in the sweet spot exhibit lower efficiency in the presence of DNA bulges that push ΔG_H outside of the desired range (Supplementary Table 7). Nothing can be stated for the remaining two gRNAs in the dataset, with 65 and 35% GC content. The former maintains its position in the range of preferential ΔG_H even in the presence of bulges, which reduce the cleavage efficiency. The latter is instead too unstable to tolerate any bulge (high ΔG_H).

Inspired by this result, we analysed the GUIDE-seq off-target dataset of Tsai et al.[30] to understand if mismatches can favour the binding of gRNAs with high GC content by shifting ΔG_H into the sweet spot. Notably, the four gRNAs with the highest GC-content in the dataset (67–80%) have off-targets with mismatches that surpass the on-target in terms of cleavage activity, measured as GUIDE-seq read counts. We focused our analysis on off-targets with no 3-nt PAM (NGG) variation to the on-target and with up to three mismatches to the gRNA, none of which should be in the four PAM-proximal nucleotides, as this could interfere with sequence composition preferences (Fig. 1b). On- and off-target sites (Fig. 2a) were classified by their cleavage activity compared to that of their on-target and by the number of GUIDE-seq read counts, setting a threshold of 300 as low cleavages (Fig. 2b). Of the four on-target sites, three are within the preferential ΔG_H range, while all four off-targets with higher efficiency compared to the on-target reside in the ΔG_H sweet spot. These off-targets are better centered in the preferential ΔG_H range compared to the on-targets, except for the off-target of the gRNA '*VEGFA* site 1 tru-gRNA'. This is a shortened 18-nt gRNA, and as such, it is expected to have higher ΔG_H (less stable binding) compared to the 20-nt gRNAs employed in the rest of the analyses. Instead, all except one of the 18 off-targets with efficiency lower than the on-target fall out of the sweet spot due to their increased ΔG_H (Fig. 2c). Notably, the efficiency at these off-target sites cannot be explained by the GC content of the matching bases at the target site (Fig. 2d). This shows that the position-specific weighted evaluation of matching and mismatching stacking interactions of ΔG_H gives a major benefit in the assessment of gRNA–DNA binding potential compared to a mere nucleotide content calculation.

These results suggest that the sweet spot affinity is a suitable criterion for defining cleavage efficiency at target sites with perfect complementarity, while the combination of the sweet spot range with bulges and mismatches interestingly explains why some off-targets are more efficient than their corresponding on-targets.

The impact of local sliding over canonical PAMs on cleavage efficiency. Based on the finding that Cas9 can search for PAMs by sliding on the DNA[19], we devised an energy-based model to explain the cleavage activity of Cas9 binding at sites with GG motifs overlapping the on-target PAM (Fig. 3a). We reasoned that Cas9 can bind at juxtaposed upstream PAMs forming a gRNA–DNA interaction with a bulge on the first PAM-proximal gRNA nucleotide, and spontaneously slide towards the intended on-target PAM to maximise gRNA–DNA complementarity and increase stability. Conversely, binding at downstream PAMs can anchor the complex in sub-optimal configurations with a DNA bulge between the gRNA–DNA hybrid and the PAM. Except for this single bulge, the hybrid is tightly coupled by 20 base pairs and can thus prevent Cas9 from sliding away or dissociating from the PAM, while keeping the site inaccessible to other Cas9 complexes. However, because a single PAM-proximal DNA bulge can be tolerated[29], cleavage cannot be excluded at these sub-optimal bindings. Considering the flexibility of nucleic acid bindings, we also do not exclude the possibility of sliding from the on-target PAM to an adjacent downstream PAM at which the gRNA and DNA remain fully bound (Fig. 3a). In line with this model, previous studies have shown that the presence of guanine immediately downstream from the on-target PAM is unfavourable for cleavage efficiency[4,8]. In the merged dataset of Xiang et al. used to study free energy change properties (of 11,602 gRNAs), the mean efficiency at targets with a downstream PAM (DNA

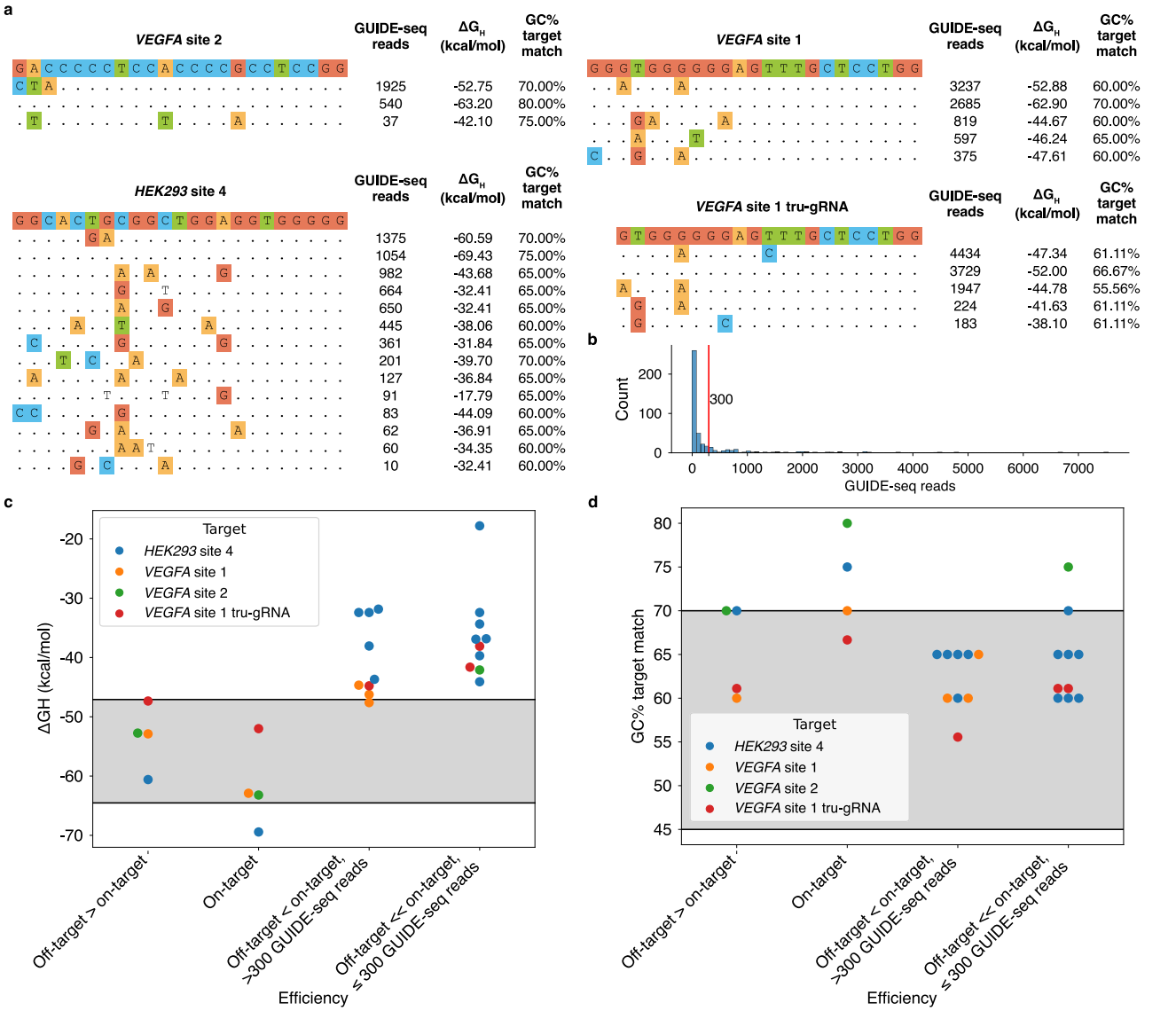


Fig. 2 Cleavage activity at off-targets modelled with binding free energy changes. a Sequences and properties of off-targets identified by GUIDE-seq[30] for four gRNAs that have at least one off-target cleaved more efficiently than the on-target. The sequence of the on-target site is reported on top of a list of the profiles of targets detected by GUIDE-seq. Each sequence in the listed targets is shown as follows: dots indicate bases that are the same as the on-target site, while differences between the on-target sequence are displayed with coloured nucleotides. Off-targets were filtered to avoid sites with >3 mismatches to the gRNA, or mismatches located in the four PAM-proximal nucleotides or the PAM itself. **b** Histogram of GUIDE-seq read counts at target sites in the full dataset of Tsai et al. A red vertical bar indicates the arbitrary threshold of 300, used to distinguish sites with low off-target activity. **c** gRNA-DNA hybridisation free energy change ΔG_H of target sites grouped by cleavage activity. The sweet spot preferential range of hybridisation free energy change ΔG_H, which contains 80% of the highly efficient gRNAs (top 20% efficient) in the merged dataset of Xiang et al., used for free energy change profiling, is highlighted in grey. **d** Percentage of G and C bases at target positions complementary to the gRNA. Target sites are grouped by cleavage activity. The preferential range of GC%, defined as in (**c**), is highlighted in grey. Source data are provided as a Source Data file.

bulge) is 12.64% lower compared to the efficiency at targets with isolated non-overlapping PAMs (Fig. 3b). Instead, a mean increase in efficiency of 7.24% characterises targets with upstream PAMs (gRNA bulge). Finally, targets with both down- and upstream sliding PAMs display a percentage decrease in mean efficiency of 3.96%. These observations show that the context of Cas9 binding sites significantly impacts gRNA efficiency.

Multiple overlapping binding sites distanced more than 1 nt from the on-target one while being interspaced from it only by guanines are rare in the dataset (Supplementary Fig. 5). We analysed adjacent PAMs created by G stretches up to 3 nt upstream or downstream of the on-target PAM binding site. The limit of 3 nt is imposed by the size of the target surrogates used to evaluate efficiency, which are integrated randomly in the genome and have unknown context[14,17]. Multiple downstream sliding PAMs do not penalise gRNA efficiency to a greater extent than single ones. Instead, multiple upstream sliding PAMs are more favourable than single ones (Supplementary Fig. 5). However, these involve position −2 or −3 from the on-target binding site (N19, N20), which are part of the seed region, and therefore the increase in efficiency may be related to the nucleotide preferences in the seed in addition to the sliding effects. This bias is not present at sliding PAMs immediately upstream from the on-target binding site (position −1), which do not overlap the gRNA-target. Therefore, in the following, we focus on overlapping PAMs with up to 1 nt from the on-target Cas9 binding site.

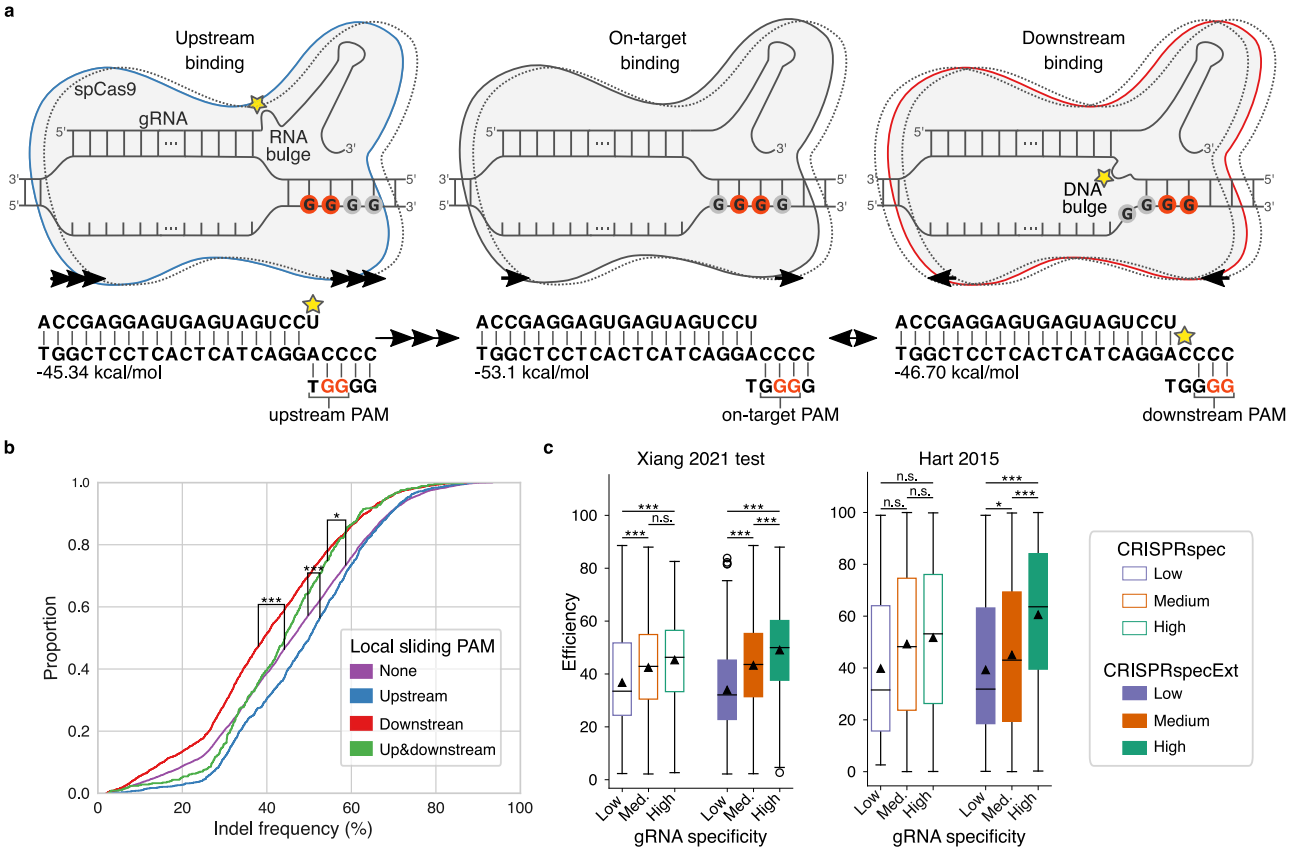


Fig. 3 Impact of local sliding PAMs on gRNA efficiency and specificity. a Cas9-gRNA binding model in which guanine stretches upstream and downstream from the on-target PAM can function as local sliding PAMs at which the gRNA forms a bulged hybrid with the target DNA. The direction of Cas9 sliding is indicated with arrows, and the upstream and downstream sliding complexes are coloured in blue and red, respectively. Bulges are indicated by yellow stars. Stretches of guanine (G), representing possible PAM binding sites, are reported in the drawing and highlighted in red if used by the complex. **b** Cumulative distributions of indel frequencies of gRNAs categorised by the presence of local sliding PAMs at their targets (one-way ANOVA p value = 3.59E-57). Downstream $n = 2239$; upstream $n = 1339$; none $n = 7565$; up and downstream $n = 459$. T-test p value comparing upstream-none = 1.96E-10 (one-sided), downstream-none = 1.50E-40 (one-sided), up and downstream-none = 3.69E-02 (two-sided). **c** Boxplots of gRNA efficiencies from two datasets binned into low, medium and high CRISPRspec or CRISPRspecExt groups before (left) and after (right) the addition of local sliding contributions. Boxes represent the first and third quartiles (Q1 and Q3). The median is shown as a line and the mean by a triangle; whiskers extend up (or down) to the last (or first) value lower (or greater) than Q3+1.5*(Q3-Q1) (or Q1-1.5*(Q3-Q1)). Number of elements: 'Xiang 2021 test', CRISPRspec low $n = 204$, medium $n = 1788$, high $n = 227$, CRISPRspecExt low = 477, medium = 1377, high = 365; 'Hart 2015', CRISPRspec low = 18, medium = 1019, high = 204, CRISPRspecExt low = 79, medium = 774, high = 388. One-sided Kolmogorov–Smirnov two-sample test, Bonferroni-corrected within each dataset: 'Xiang 2021 test', CRISPRspec low-high = 9.61E-08, medium-high = 0.16, low-medium = 1.04E-07, CRISPRspecExt low-high = 6.63E-27, medium-high = 2.49E-06, low-medium = 3.21E-20; 'Hart 2015', CRISPRspec low-high = 0.06, medium-high = 0.36, low-medium = 0.20, CRISPRspecExt low-high = 1.91E-09, medium-high = 1.07E-12, low-medium = 1.69E-2. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$. Source data are provided as a Source Data file.

Local sliding PAMs allow gRNAs to bind with their targets forming bulged interactions. Bulged bindings are often not accounted for by off-target scoring tools, as their computation is highly demanding. However, the bulged interactions at local sliding PAMs are quick to identify and evaluate, and because of their direct competition with the on-target, they are also highly relevant. Hence, considering bulged interactions at local sliding PAMs as pseudo-off-targets, we extend our off-target scoring to explain how sliding PAMs influence gRNA binding competition. Up- and down-stream local sliding PAMs are integrated into our previously defined CRISPRspec global gRNA specificity score[20]. The higher the CRISPRspec score, the less binding competition (off-targets) affects a gRNA. In general, a score above 5 empirically delineates low off-target potential[20]. In brief, CRISPRspec is calculated as $-\log_{10}$ of the fraction between the sum of the Boltzmann-weighted ΔG_B computed at all targets in the genome with up to a certain number of mismatches excluding (numerator) or including (denominator) the on-target[20]. Among competing PAMs, only the most favourable one (lowest ΔG_B) is used. Given a gRNA the effect of local sliding PAMs is integrated by adding to CRISPRspec the gain or penalty in efficiency attributable to the local sliding. To do so, CRISPRspec is firstly linearly re-scaled to the efficiency (see Methods). The parameters that define the local sliding addends are obtained by linearly fitting the deviation of the ΔG_B measured at a local sliding PAM from the median ΔG_B computed at all local sliding PAMs of the same type in the dataset (up- or down-stream) to the efficiency. The deviation is used as extreme ΔG_B is unfavourable (Supplementary Fig. 6). To fit the linear model, the merged dataset of Xiang et al. (excluding the test portion) is used with no prior filter on CRISPRspec (14,981 gRNAs, see Supplementary Table 1). The CRISPRspec score extended with local sliding awareness, CRISPRspecExt, shows extensive downgrading and upgrading, respectively, for gRNA targeting sites with down-stream and upstream local sliding PAMs (Supplementary Fig. 7). A mixed trend is recorded for gRNA targets that possess both types of local sliding PAMs. Interactions at gRNA targets having neither of the two remain unaffected.

To test if the inclusion of local sliding PAMs improves CRISPRspec, we examined the relation between CRISPRspec and

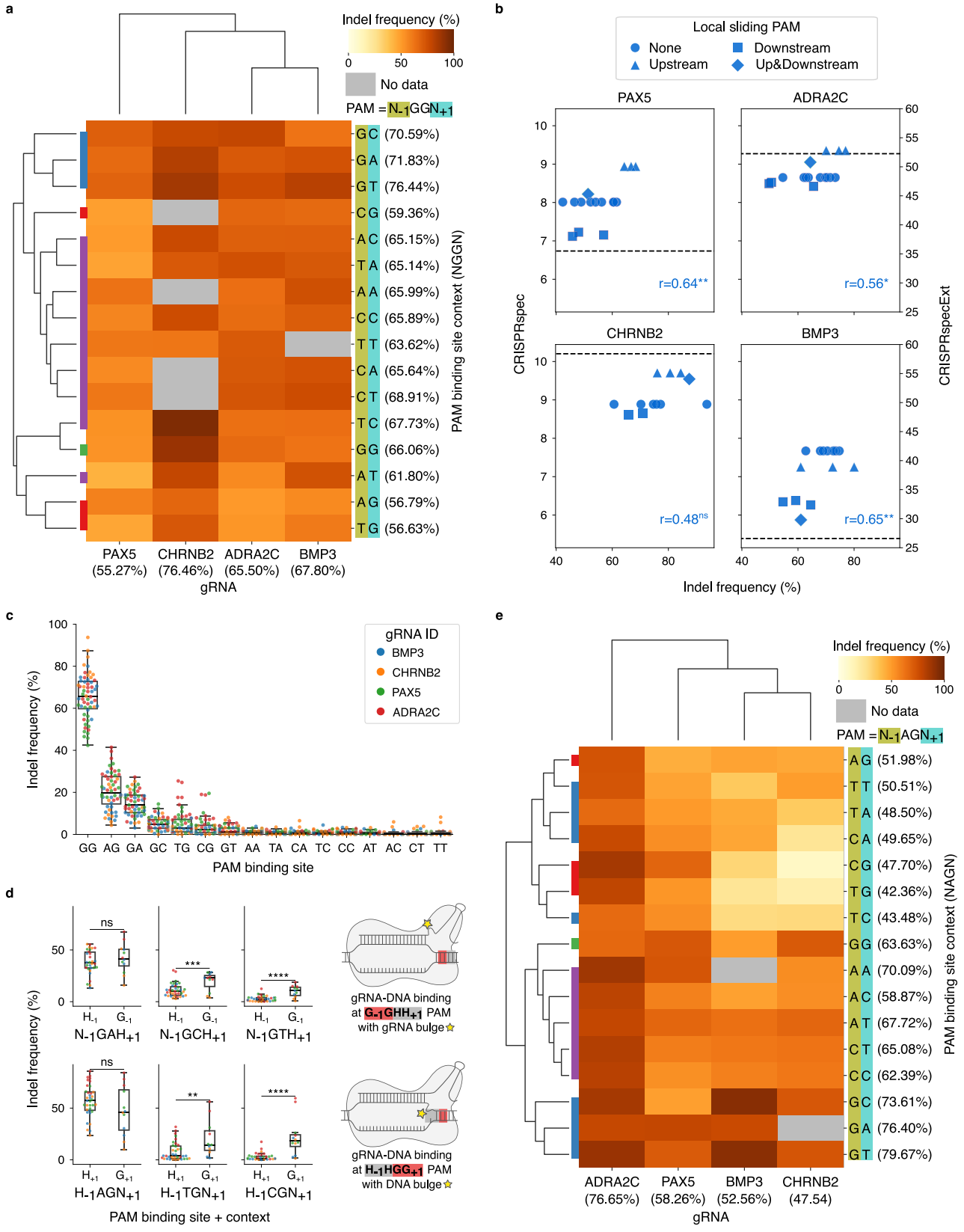
gRNA efficiency before and after the addition of local sliding considerations on two independent test sets. Only gRNAs (20mer) with >3 nt differences to those used for fitting were used (6361 gRNAs for 'Xiang 2021 test', 4026 gRNAs in 'Hart 2015', see Supplementary Table 1). To focus the test on the impact of sliding, the evaluation was performed on gRNAs targeting sites where sliding is possible, i.e., with a downstream or an upstream PAM (2219 gRNAs for 'Xiang 2021 test', 1241 gRNAs in 'Hart 2015'). The gRNAs were partitioned into three equally sized sets of low, medium and high CRISPRspec or CRISPRspecExt. Binning values were estimated on the dataset employed for parameter fitting. In 'Xiang 2021 test' the efficiencies of gRNAs with CRISPRspec medium or high are not different prior to the inclusion of local sliding (Fig. 3c). Other groups, instead, are significantly distinct. After introducing local sliding, the separation between the three groups becomes more profound and all three present significant differences in terms of efficiency (Fig. 3c). On the external independent test set 'Hart 2015', CRISPRspec does not significantly separate gRNAs with different efficiencies, while this task is accomplished by CRISPRspecExt. This analysis shows that CRISPRspecExt can isolate well highly specific, and thus efficient, gRNAs. The positive correlation between our competition score CRISPRspecExt and gRNA efficiency in the 'Xiang 2021 test' and 'Hart 2015' test sets (Pearson's $r = 0.29$ and $r = 0.26$ respectively) is more substantial compared to that of other relevant properties taken individually, such as the ΔG_B and the GC content ($r = -0.26$ and $r = 0.15$ respectively in 'Xiang 2021 test', no significant correlation in 'Hart 2015', Supplementary Fig. 8). These results support the concept that strong binding competition at local and/or global off-target sites impairs cleavage activity.

Validation of sliding effects at canonical PAMs in HEK293T cells. As revealed from the analysis above, the gRNA efficiency is affected by the context flanking the on-target PAM binding site. To further validate this effect, we measured the efficiency of four gRNAs for the genes *PAX5*, *BMP3*, *ADRA2C* and *CHRNB2*. These gRNAs displayed high cleavage efficiency in our previous surrogate-based evaluation of gRNAs and exhibited no trace of enrichment or depletion in the edited cells[17] (see Methods). For each gRNA, target DNA surrogate sites were designed to carry all possible 16 variations of N_{-1} and N_{+1} in the 5′-$N_{-1}GGN_{+1}$-3′ PAM (Supplementary Table 8). The efficiency was measured as indel frequency 6 and 10 days after lentiviral transduction in HEK293T cells, following our previously established protocol[17] (see Methods). The efficiencies obtained at the two time points were strongly correlated (Pearson's $r = 0.98$, Spearman's $R = 0.79$) and were therefore averaged (Supplementary Fig. 9 and Supplementary Data 1). In agreement with the local PAM-sliding model presented above, the cleavage at targets with $G_{-1}GGH_{+1}$ and $H_{-1}GGG_{+1}$ PAMs displays, in this order, percentage increase and decrease in mean efficiency of 11.31 and 12.13% compared to sites with $H_{-1}GGH_{+1}$ PAM, where H symbolises any of A, T or C. Clustering targets based on the $N_{-1}N_{+1}$ context produces an exclusive cluster for $G_{-1}H_{+1}$, which comprises the top three contexts with highest mean efficiency across the four gRNAs, while the lowest 3 belong to the $H_{-1}G_{+1}$ contexts (Fig. 4a).

If local sliding PAMs are not considered, CRISPRspec is constant for different PAM contexts of the gRNA targets. In contrast, CRISPRspecExt is positively correlated to the gRNA efficiency, thanks to the addition of the sliding PAM contributions (Fig. 4b). Thus, our energy-based binding model allows improved identification of highly specific and efficient gRNAs based on the context of their targets. The portion of the variance associated with the differences in the PAM binding site context explained by the sliding activity on adjacent canonical 'GG' binding motifs was assessed as the sum of squares to the grand mean relative to each condition (sliding upstream, downstream, on both sides or on none) and weighted by the size of each group, divided by the total sum of squares (see Methods). The sliding activity explains >50% of the variance in the targets designed for the gRNAs ADRA2C (53.49%), PAX5 (52.44%) and BMP3 (50.58%) while this proportion is lower, 32.44%, in the case of gRNA CHRNB2. Other features of the target sites may be responsible for the remaining, unexplained, variance. For instance, the random integration of the surrogate sequences carrying the targets may impact their accessibility and, therefore, the ability of Cas9 to cleave these sites, which is a limitation of the current surrogate-based approaches[17]. In addition, the effect of the different context nucleotides on PAM-probing and the possible sliding activity of Cas9 on non-canonical PAMs, described in the next section, may also contribute to the remaining variance (i.e., binding and sliding at $G_{-1}GGH_{+1}$ may variate for different H_{+1}). Additional data will be necessary to evaluate this feature and eventually include it in our sliding model.

Local sliding broadens non-canonical PAM compatibility. Cleavage by Cas9 can also verify at target sites flanked by non-canonical PAMs, of which $N_{-1}AGN_{+1}$ is reported as the most active[10]. To test if the concept of local sliding is applicable to non-canonical PAMs, we evaluated the efficiency, measured as indel frequency, of the same four gRNAs described above for target DNA sites bearing all possible variations of the PAM binding motif and its context. The activity of Cas9 at non-canonical PAMs is generally low, therefore we additionally measured gRNA efficiency in HEK293T cells with Doxycycline-induced overexpression of Cas9 (Dox+). The gRNA efficiencies measured in Dox+ and Dox− cells are well correlated (Supplementary Fig. 10). In untreated cells (Doxo−) the mean efficiency at targets flanked by the canonical GG PAM binding site is 65.49% while at the non-canonical AG and GA it is respectively 21.18 and 14.93%. At other non-canonical binding sites, the mean efficiency is below 10% (TG = 5.29%, GC = 5.06%, CG = 3.69%, GT = 2.00%, AA = 1.12%, TA = 0.90%, CC = 0.88%, CA = 0.85%, TC = 0.81%, TT = 0.72%, AT = 0.71%, CT = 0.62%, AC = 0.61%) (Fig. 4c). In the Dox+ group (Supplementary Fig. 11), extremely high efficiency is obtained at bindings with a canonical PAM regardless of their context (mean = 96.79%). The efficiency at the non-canonical AG binding site in Dox+ reaches levels close to those of GG in the Dox− group (mean = 59.54%). The overexpression of Cas9 increases the efficiency at sites flanked by other non-canonical PAMs as well, such as those with binding sites GA, TG, GC and CG, with mean efficiencies of 41.50, 15.09, 12.75 and 8.86%, respectively. Cleavage at targets with other PAMs remains instead rare or null (mean efficiency <5% for all).

Among all possible alternative PAMs those with at least one G in the binding motif are the most efficient ones (Fig. 4c). These sites can incorporate local sliding PAMs with a canonical GG binding motif if a G is present in the context upstream ($G_{-1}GNN_{+1}$) or downstream ($N_{-1}NGG_{+1}$). Following Cas9 binding at such PAMs, gRNAs may form bulged interactions with their targets in lack or excess of 1 nt (Fig. 4d). Indeed, the cleavage efficiency measured in Dox+ HEK293T cells at alternative PAMs with a canonical GG binding motif upstream ($G_{-1}GCH_{+1}$ and $G_{-1}GTH_{+1}$) or downstream ($H_{-1}TGG_{+1}$ and $H_{-1}CGG_{+1}$) is significantly higher compared to that of targets with the same alternative PAM binding sites but no G in the $N_{-1}N_{+1}$ context (Fig. 4d, see also Supplementary Figs. 12–15 for results on all PAM binding sites in Dox- and Dox+). Alternative



PAMs with binding motif AG and GA, which are the closest to the canonical GG in terms of efficiency, do not display a preference for local sliding toward canonical PAMs. This suggests that fully complementary gRNA–DNA interactions at a non-canonical AG or GA are equally tolerated as bulged bindings at the canonical GG PAM binding site.

The efficiencies related to target sites with $G_{-1}AGH_{+1}$ PAMs in Dox+ cells are significantly more efficient (mean increase 33.45%) compared to $H_{-1}AGH_{+1}$ PAMs (Fig. 4e, Supplementary Fig. 14). This trend is similar to the one observed for the canonical $G_{-1}GGH_{+1}$ PAM in Dox−. G_{-1} generates a non-canonical GA binding site, less stable than GG or AG. This could facilitate

Fig. 4 Indel frequency of gRNAs at targets with different PAM sequences in HEK293T cells. a Heatmap of indel frequencies (Dox- cells) of four gRNAs targeting sites with 5′-N$_{-1}$GGN$_{+1}$-3′ PAMs clustered by Euclidean distance (missing values were linearly interpolated from targets with the same N$_{-1}$N$_{+1}$ context). Averages of columns and rows are indicated. Leaves in the left dendrogram are coloured as in Fig. 3b. **b** Pearson's correlation between indel frequency of four gRNAs and their specificity extended with local sliding (CRISPRspecExt, blue colour). *$p < 0.05$, **$p < 0.01$, ns non-significant (gRNA PAX5 $p = 7.05E-03$, ADRA2C $p = 0.02$, CHRNB2 $p = 0.11$, BMP3 $p = 8.73E-3$). The specificity score CRISPRspec, not influenced by PAM contexts, is indicated with a dashed line. **c** Indel frequency (Dox- cells) of gRNAs binding at targets with different PAM binding sites (X-axis). The swarm plot details the indel frequencies of each gRNA. Boxes represent the first and third quartiles (Q1 and Q3). The median is shown as a line; whiskers extend up (or down) to the last (or first) value lower (or greater) than Q3+1.5*(Q3-Q1) (or Q1-1.5*(Q3-Q1)). Number of gRNAs per box, left to right: 59; 56; 53; 56; 60; 54; 62; 60; 58; 61; 59; 62; 58; 57; 62; 63. **d** Indel frequency (Dox+ cells) at targets with alternative PAMs but canonical GG local sliding PAM upstream (top) or downstream (bottom). Boxes and swarm plots are defined as in (**c**). Binding contexts represented in less than three of four gRNAs or with G$_{+1}$ (top) or G$_{-1}$ (bottom) were excluded. See Supplementary Figs. 12–15 for results on all PAMs and cell treatments. One-sided *t*-test *p* values from left to right, top to bottom: 0.17; 9.72E-04; 1.69E-07; 0.94; 1.40E-03; 2.59E-07 (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$, ****$p <$ 1E-04). The alternative hypothesis is that efficiencies at targets with G$_{-1}$ (top) or G$_{+1}$ (bottom) have a larger mean than, respectively, H$_{-1}$ or H$_{+1}$. Number of elements per box, from left to right, top to bottom: 30; 12; 35; 11; 36; 11; 35; 12; 36; 11; 35; 12. **e** Heatmap of indel frequencies (Dox+ cells) of four gRNAs at targets with 5′-N$_{-1}$AGN$_{+1}$-3′ PAMs. See (**a**) for further details. Source data are provided as a Source Data file.

further the search of Cas9 for a more stable binding site, both in terms of protein and gRNA binding. Conversely, the change in mean efficiency (decrease of 17.47%) registered at H$_{-1}$AGG$_{+1}$ compared to H$_{-1}$AGH$_{+1}$ is non-significant (Supplementary Fig. 15). G$_{+1}$ locks Cas9 at a GG binding site, more favourable than AG, but with 1 nt DNA bulge between the PAM and the gRNA–DNA hybrid. Although less pronounced, these effects are also visible in the Dox− group (Supplementary Fig. 16). Similarly, no significant change in efficiency is observed at G$_{-1}$GAH$_{+1}$ (increase in mean efficiency = 10.10%) compared to H$_{-1}$GAH$_{+1}$ (Supplementary Figs. 14, 16). In this case, G$_{-1}$ produces a canonical upstream GG binding site that may antagonise the attempt of Cas9 to maximise the gRNA–DNA binding stability, as this requires sliding towards a less favourable PAM (GA binding site). The possible gRNA–DNA hybrid formed at this site contains a 1-nt gRNA bulge and therefore has limited efficiency. Instead, H$_{-1}$GAG$_{+1}$ PAMs have an increased mean efficiency by 22.22% over H$_{-1}$GAH$_{+}$ (Supplementary Fig. 15). Downstream G$_{+1}$ creates a local PAM sliding with AG binding site, which does not anchor Cas9 as in the previous cases (GG binding site) but may favour the search for a different PAM in the neighbourhood.

Evidence for local sliding in Cas9 variants. To corroborate our observations on the local sliding activity of Cas9, we analysed the data from Kim et al.[16], which measured indel frequencies of gRNAs using 13 SpCas9 variants[16], consisting of the wild-type SpCas9, five high-fidelity variants (eSpCas9(1.1)[31], SpCas9-HF1[32], HypaCas9[33], evoCas9[34] and Sniper-Cas9[35]), five variants with altered PAM compatibility (VQR[36], VRER[36], VRQR[32], QQR1[37] and SpCas9-NG[38]), and two variants with both such properties (VRQR-HF1[32] and xCas9[39]). We and others previously reported that the data presented in Kim et al.[16], in which a modified gRNA scaffold is employed, is not entirely compatible with that of similar recent studies[17,40]. A closer look into the library dedicated to the analysis of PAM contexts (for which cleavage at surrogate targets followed by all combinations of 4-nt PAMs was examined for 30 gRNAs) reveals that in the case of SpCas9, gRNAs with a U at the 3′ seed end (U20) are more efficient (indel frequency mean ± std = 52.46 ± 5.44%, $n = 110$ targets of seven gRNAs) compared to gRNAs with a C20 (mean ± std = 46.23 ± 6.90%, $n = 239$ targets of 15 gRNAs) or an A20 (mean ± std = 47.44 ± 7.64%, $n = 121$ targets of eight gRNAs) (two-sided *t-test p* value C20 compared to U20 = 1.66E-15, A20 compared to U20 = 3.61E-08). Of note, no gRNA with a G20 is present in the dataset. The higher efficiency observed in gRNAs with a U20 is in contrast with our results of nucleotide preferences based on the merged data from Kim et al.[14] and Xiang et al.[17] (Fig. 1b), as well as with other independent reports[4,8,15]. The nucleotide N20 constitutes the gRNA bulge in the upstream sliding and thus plays a central role in the analysis of sliding dynamics. Because of this incongruence, we do not validate local sliding on the canonical PAM using the Kim et al.[16] dataset. The PAM preferences in such dataset are, instead, as expected, with GG, AG and GA being the preferred binding sites (in this order), and other PAM binding sites showing little or no activity (Supplementary Fig. 17). As the efficiency variation linked to different PAM binding sites is much higher compared to that of sliding dynamics on canonical PAMs, we can reliably use the dataset from Kim et al.[16] to further explore our hypothesis on Cas9 sliding from non-canonical toward adjacent canonical PAMs. In the case of the wild-type SpCas9, targets followed by alternative PAMs G$_{-1}$GCH$_{+1}$, G$_{-1}$GTH$_{+1}$ are cleaved more efficiently compared to targets with the same alternative binding site but no G at position −1 (Fig. 5a). Similarly, targets followed by alternative PAMs H$_{-1}$TGG$_{+1}$ and H$_{-1}$CGG$_{+1}$ are cleaved more efficiently than H$_{-1}$TGH$_{+1}$ and H$_{-1}$CGH$_{+1}$ (Fig. 5a). No increase in efficiency is observed for the PAMs G$_{-1}$GAH$_{+1}$ and H$_{-1}$AGG$_{+1}$ compared to H$_{-1}$GAH$_{+1}$ and H$_{-1}$AGG$_{+1}$, due to the higher efficiency at these alternative PAM binding sites (Fig. 5a). These results are congruent with those obtained with the data we generated. Moreover, we examined high-fidelity Cas9 variants that do not contain variations in the PAM-recognition domains after sorting them by efficiency and inverse specificity, previously assessed[16]. The same preferences for up- and down-stream sliding toward canonical PAMs are visible in Sniper-Cas9, which is the variant closest to the wild-type in terms of efficiency and specificity, and gradually disappear in more specific variants (eSPCas9(1.1), SpCas9-HF1, HypaCas9), less tolerant to imperfect gRNA–DNA bindings (Fig. 5a). The variant evoCas9 was shown to be generally poorly active[16] and was thus excluded.

Furthermore, we analysed the efficiency of the Cas9 variant SpCas9-NG, which shows good compatibility with N$_{-1}$GNN$_{+1}$ PAMs, and xCas9, which has increased binding specificity and tolerates N$_{-1}$GNN$_{+1}$ PAMs, although N$_{-1}$GGN$_{+}$1 remains highly preferred[16]. In these variants, upstream sliding is expected at G$_{-1}$HHH$_{+1}$ PAMs, while H$_{-1}$HGH$_{+1}$ and H$_{-1}$HHG$_{+1}$ may result in downstream sliding of, respectively, 1 or 2 nucleotides (Fig. 5b). Consistently with our model, targets followed by GHHH are cleaved more efficiently than those followed by HHHH, in both variants (Fig. 5b top). Downstream sliding shows increased efficiency at targets with either of the PAMs HHGH and HHHG compared to HHHH in SpCas9-NG, while in xCas9 only HHGH is more efficient than HHHH (Fig. 5b bottom). This is expected considering the increased specificity of xCas9, that may tolerate one but not two DNA bulges between the gRNA–DNA binding and the xCas9 PAM binding site. The consistent efficiency increase observed at PAMs that allow for up- or downstream sliding to canonical binding sites and the

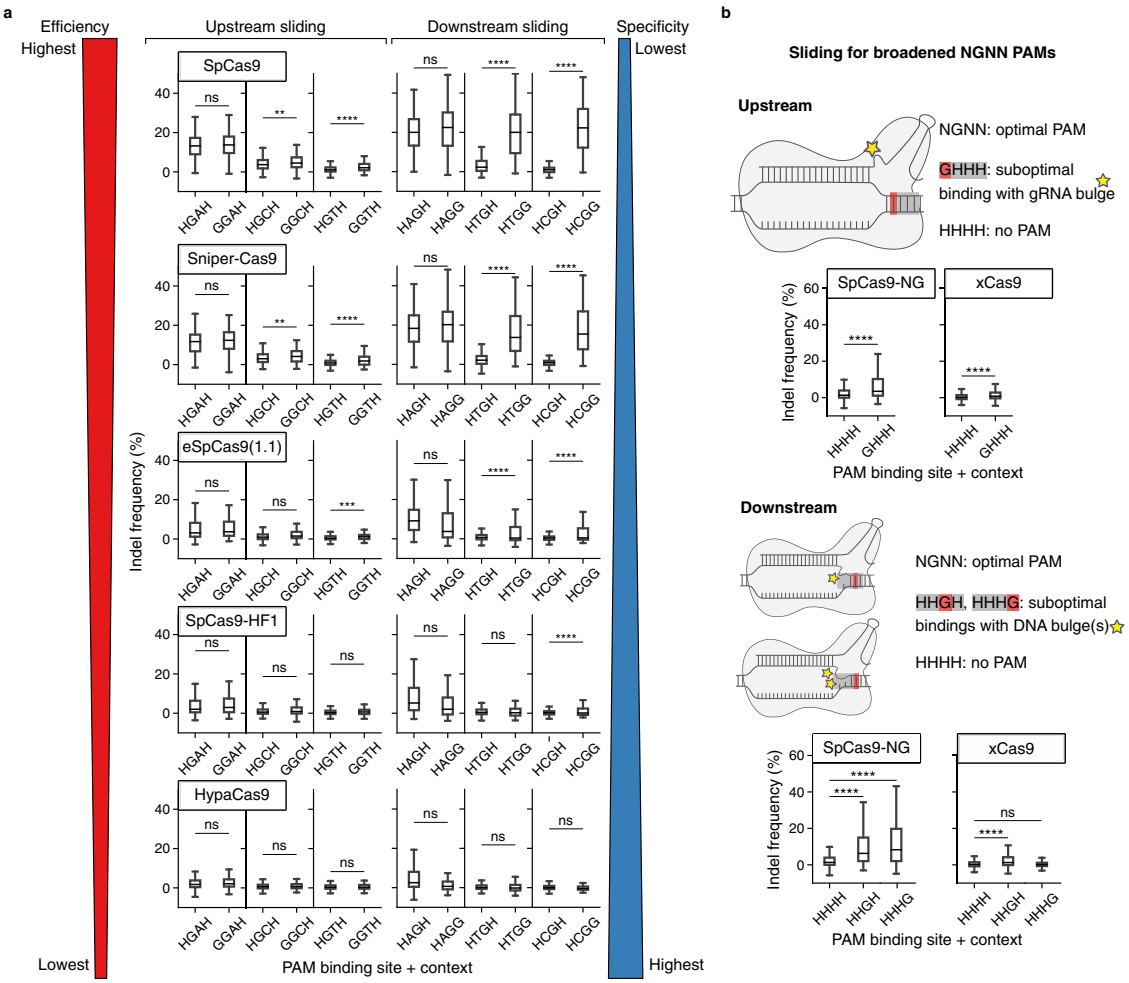


Fig. 5 Sliding affects PAM recognition by Cas9 variants. a Boxplots of indel frequencies produced by Cas9 variants in HEK293T cells at targets followed by different PAMs (X-axis). Each row represents a Cas9 variant, sorted from top to bottom by efficiency and inverse specificity. Boxplots are separated in two groups for upstream (left) and downstream (right) sliding. Boxes represent the first and third quartiles (Q1 and Q3). The median is shown as a line; whiskers extend up (or down) to the last (or first) value lower (or greater) than Q3+1.5*(Q3-Q1) (or Q1-1.5*(Q3-Q1)). One-sided *t*-test *p* values, Bonferroni-corrected by the number of Cas9 variants: from left to right, top to bottom: 1.00; 4.90E-03; 4.15E-09; 0.06; 9.60E-49; 5.53E-84; 0.83; 8.44E-03; 2.43E-07; 0.40; 1.03E-39; 1.78E-63; 1.00; 0.09; 6.87E-04; 1.00; 5.54E-06; 1.08E-12; 0.85; 0.32; 0.13; 1.00; 0.078; 1.16E-05; 0.71; 0.71; 0.32; 1.00; 0.59; 0.06. The alternative hypothesis is that gRNA efficiencies at targets with G_{-1} (left) or G_{+1} (right) have larger mean, respectively, than targets with H_{-1} or H_{+1}. Number of elements in each box, from left to right, top to bottom: 265; 88; 261; 85; 267; 85; 263; 88; 260; 86; 265; 86; 265; 88; 262; 86; 267; 85; 264; 88; 260; 87; 265; 86; 265; 88; 260; 86; 267; 85; 264; 88; 259; 87; 265; 86; 265; 88; 262; 86; 267; 85; 265; 88; 260; 87; 265; 86; 265; 88; 262; 86; 266; 85; 264; 88; 260; 87; 264; 86. **b** Illustration of the sliding mechanism at $N_{-1}GNN_{+1}$ PAMs, and boxplots of indel frequencies in HEK293T cells produced by Cas9 variants with broadened PAM compatibility at targets with different PAMs (X-axis). Boxes are defined as in (**a**). One-sided *t*-test *p* values, Bonferroni-corrected by the number of Cas9 variants: from left to right, top to bottom: 2.30E-59; 8.68E-25; 2.12E-135; 1.94E-186; 7.06E-69; 1.00. The alternative hypothesis used is that gRNA efficiencies at GHHH (upstream case) or HHGH/HHHG (downstream case) targets have larger mean than at HHHH. Number of elements in each box, from left to right, top to bottom: 2322; 799; 2317; 778; 2322; 786; 772; 2317; 786; 770. Source data are provided as a Source Data file.

gradual decrease of this effect in the most specific Cas9 variants strongly support our model for broadened PAM compatibility enabled by sliding on adjacent PAMs.

Discussion

In this work, we describe CRISPR/Cas9 cleavage as an energy-driven process in which efficiency significantly depends on nucleotide hybridisation and folding free energy changes. Our analysis of the free energy change preferences of efficient and inefficient gRNAs underlines the importance of designing gRNAs with low self-folding stability, balanced hybridisation of free energy change and firm target binding at the seed region. In regard to the ongoing discussion on whether Cas9 binding can be modelled in equilibrium[41], here we show that this is the case. Notably, our sweet spot of gRNA–DNA hybridisation free energy change provides an explanation for the higher cleavage activity previously measured at bulged or imperfectly matching off-targets compared to on-targets. Recent studies report that the mismatch tolerance patterns of different gRNAs can highly variate[42], and that sequence composition features such as enrichment of guanine and depletion of uracil characterise gRNAs with high guide-intrinsic mismatch permissiveness[43]. Our results suggest that the fitness to the sweet spot and the free energy changes related to imperfect matches to the gRNA are suitable criteria to evaluate guide-intrinsic mismatch tolerance. By expanding the concept of gRNA-target-DNA hybridisation with the inclusion of imperfect matches containing gRNA or DNA bulges, we explain the activity of gRNAs at targets with adjacent PAMs on which Cas9 can locally slide. In the merged dataset of Xiang et al. used to examine

binding free energy change properties, such local sliding activity affects ≈35% of the gRNAs that have a target site in hg38. We show that sliding activities can affect cleavage efficiency both positively and negatively, depending on the binding free energy change at bulged interactions. The inclusion of local sliding effects, which can be interpreted as local off-targets, improves the definition of gRNA specificity, previously based solely on global off-targets. Although interactions at sliding PAMs can result in DNA cleavage and produce indels near the target site, fulfilling the editing purpose in knockout experiments, their activity nearby the on-target site can lead to undesired effects in applications that require Cas9 to be positioned precisely at the target, such as in base editing. We validate the effect of local sliding on gRNA cleavage efficiency by analysing independent data (Fig. 3b, c) and by generating efficiency data for 4 gRNAs targeting DNA sites that carry all possible alterations of the Cas9 binding site and its immediate context (Fig. 4 and Supplementary Figs. 12–16). Such evaluation of different PAMs and contexts for the same gRNAs eliminates the bias that would derive from measuring the effect of contexts and PAMs using distinct gRNAs. Furthermore, we reveal that cleavage can be obtained more efficiently at targets followed by non-canonical PAMs whenever sliding on adjacent canonical PAMs is possible. This result, supported both by our data and by public gRNA efficiency data from wild-type and engineered Cas9 variants, implies that local sliding broadens Cas9-PAM compatibility. This can be highly useful in those situations in which the choice of sites to be targeted for cleavage is limited (e.g., knock-in of genomic variants). While our study is centred on Cas9-gRNA interactions, the concepts we describe can be extensively applied in relation to other RNA-guided complexes. In conclusion, the local context of Cas9 binding sites can strongly impact Cas9 binding and cleavage efficiency, and this can largely be explained by our binding energy-based model. We also show that the assessment of gRNA specificity is enhanced once local sliding PAMs are considered and that this helps to identify gRNAs with higher specificity and efficiency.

Methods

Cell culture. Human embryonic kidney cells (HEK293T, originally purchased from ATCC catalogue num. CRL-3216) were cultured in DMEM media containing 10% foetal bovine serum (FBS) and 1% penicillin–streptomycin in a tissue culture incubator at 37 °C with 5% CO_2. PCR mycoplasma detection kit (catalogue num. PM008, Shanghai Yise Medical Technology) was routinely used to test the mycoplasma contamination. The cells used in this study gave negative results in the mycoplasma contamination test. SpCas9-expressing HEK293T (HEK293T-SpCas9) cells were generated by a PiggyBac transposon system followed by selection in the presence of 50 μg/ml hygromycin to ensure high Cas9 activity. HEK293T cells were transiently transduced with pPB-TRE-spCas9-Hygromycin vector and pCMV-hybase vector with a 9:1 ratio to generate SpCas9-expressing HEK293T.

Surrogate library design and plasmid library preparation. To construct the gRNA-target-DNA surrogate library we selected four gRNAs with indel efficiency >90% in our previous CRISPRon chip (Dox+)[17] and designed complementary targets flanked by 4-nt PAM sites variated in all possible ways (PAM=NNNN). This resulted in $4^4 = 256$ gRNA-target-DNA surrogates, plus 10 positive sequence surrogates. The total number of sequences we generated was 1062 because two sequences containing BsmBI restriction sites were filtered out. Oligos were structured as follows: 20 bp gRNA, 82 bp gRNA scaffold, 37 bp surrogate target (10 bp barcode, 20 bp protospacer, 4 bp random PAM, 3 bp downstream sequence).

The 1062 PAM library oligo was synthesised in Genscript® (Nanjing, China). All sgRNA sequences and their oligos are listed in Supplementary Data 2. The library was cloned as a pool into a LentiU6-LacZ-GFP-Puro (BB) lentiviral plasmid (Addgene ID:170459) for lentivirus production as previously described in ref. [17]. Briefly, plasmid library cloning started with PCR amplification of the 170-nt oligonucleotide pool. Firstly, the library oligos were diluted to 1 ng/μl and then PCR amplification was performed using the primers: TRAP-oligo (BsmBI GGA)-F(5′-TACAGCTaccacgtct caCACC-3′) and TRAP-oligo (BsmBI GGA)-R (5′-AGCACAAccgtcgtctccAAAC-3′). The PCR reaction was carried out using PrimeSTAR HS DNA Polymerase (Takara, Japan) following the manufacturer's instructions. The PCR products of library oligos were then used for Golden Gate Assembly (GGA) to generate the plasmids library and the ligation products were transformed to chemically-induced competent DH5α cells. About 10 μl GGA ligation product was transformed, for each reaction, into 50 μl competent cells. All transformed cells were plated on one LB plate (Ø15 cm) with Xgal, IPTG, and Amp selection. For one library containing all the synthetic oligos, 12 parallel transformations were performed, and all the bacterial colonies were scraped off and pooled together for plasmids midi-prep (PureLink™ HiPure Plasmid DNA Midiprep Kit).

Lentivirus production and lentivirus titre quantification. The lentivirus was produced in HEK293T cells by co-transfection with packaging plasmids. Briefly, for lentivirus production in 10 cm dish, the DNA/PEI (Polyethylenimine Linear, MW 40000) mixture contains 13 μg pLenti-TRAPseq vectors, 3 μg pRSV-REV, 3.75 μg pMD.2 G, 13 μg pMDGP-Lg/p-RRE, 100 μg PEI 40000 solution (1 μg/μl in sterilised ddH2O) and supplemented by opti-MEM without phenol red (Invitrogen) to a final volume of 1 mL. The transfection mixture was pipetted up and down several times gently, and further incubated for keeping at room temperature (RT) for 20 min. The transfection complex was added to 80%-confluent HEK293T cells in a 10-cm dish containing 10 ml of culture medium. After 48 h viral supernatant was harvested, filtered through a 0.45 μm filter, and polybrene solution (Sigma-Aldrich) was added to the crude virus to a final concentration of 8 μg/mL. The crude virus was aliquoted into 15 mL tubes (5 mL/tube) and stored in a −80 °C freezer. Viral titre was estimated by counting the number of GFP-positive cells in the virus-treated population by flow cytometry (FCM) as follows: (1) HEK293T cells were split and seeded to a 24-well plate on day 1. Generally, gradient volumes of 5, 10, 20, 40, 80, 160 and 320 μl of crude lentivirus were added to the cells, and each volume was assayed in duplicate (Supplementary Fig. 18); (2) On day 2, lentivirus transduction was conducted when cells reached up to 60–80% confluence. Before transduction, the last two wells of cells were detached using 0.05% EDTA-Trypsin to determine the total number of cells in one well (N_{initial}). Then the gradient volume of the crude virus was added to each well and swirled gently; (3) On day 3, we changed to a fresh medium; (4) On day 4, cells were harvested and washed twice in PBS. The cells were fixed in 4% formalin solution at RT for 20 min, then washed with PBS twice. FCM was performed using a BD LSRFortessa™ cell analyzer and the FACSDiva v.9.0 software with at least 50,000 events collected for each sample in replicates. The FCM output data was analysed with the FCM analysis software NovoExpress v.1.5.6. The percentage of GFP-positive cells, for all samples, was calculated as:

$$Y\% = \frac{\mathrm{Num.GFP-positive\ cells}}{\mathrm{Num.total\ cells}} \times 100 \quad (1)$$

For accurate titre determination, there should be a linear relationship between the GFP-positive percentages and crude volume. The titre (transducing units (TU/mL)) was calculated according to the following formula in which V represents the crude volume (μl) used for initial transduction:

$$\frac{\mathrm{TU}}{\mathrm{mL}} = \frac{N_{\mathrm{initial}} \times Y\% \times 1000}{V} \quad (2)$$

Lentivirus library transduction. HEK293T cells stably expressing a low level of SpCas9 were infected at a MOI=0.3 to ensure that most cells receive only one viral construct with high probability. Overexpression of SpCas9 in the HEK293T cell line is induced by doxycycline (Dox). At day 1 (24 h after transduction), transduced cells in each dish were split into two dishes equally. On day 2 (48 h after transduction), the sub-group 1 was changed to a fresh medium containing 50 μg/ml hygromycin + 1 μg/mL puromycin (Dox-free group, or Dox-). The sub-group 2, instead, was changed to medium containing 50 μg/ml hygromycin + 1 μg/mL puromycin + 1 μg/mL doxycycline (Dox-induction group, or Dox+). The transduced cells were split every 2–3 days when cell confluence reaches up to 90%. After 6 and 10 days of selection, cells were harvested for genomic DNA extraction. Parallel experiments were performed using wild-type HEK293T cells.

PCR amplicons of a surrogate library from cells. The genomic DNA was extracted with the phenol-chloroform method. To remove RNA contamination, the genomic DNA was digested with RNase A (OMEGA). Then the genomic DNA was subjected to PCR using PrimeSTAR polymerase (Takara, R045Q). The PCR primers were TRAP-NGS-F (5′-GGACTATCATATGCTTACCGTA-3′) and TRAP-NGS-R1 (5′-ACTCCTTTCAAGACCTAGCTAG-3′). The PCR products were purified by 1.5% gel and mixed with equal amounts and deep sequenced. The amplicons were subjected to deep sequencing on the MGISEQ-2000 (MGI of BGI in China) platform. All the samples were subjected to pair-ended 150 bp deep-sequencing.

Data pre-processing. Raw sequencing reads were processed to obtain read counts (indel or total) for each surrogate target following our previously published method[17]. Briefly, the method consists of the following passages: quality assessment and cleaning of raw reads with FastQC v.0.11.3 (https://github.com/s-andrews/FastQC) and fastp v.0.19.6[44]; paired reads merge with FLASH v.1.2.1[45]; alignment to the reference library with BWA-MEM v.0.7.17[46] (http://bio-bwa.sourceforge.net/); and reads quantification by extracting the aligned reads and identifying variations in the expected mapping pattern using pysam v.0.15.0[47] (https://github.com/pysam-developers/pysam). For each target site, g + gRNA (20 bp) + scaffold (82 bp) + barcode (10 bp) + GTTT (terminal sequence) are guaranteed to remain unchanged when extracting reads at each random PAM site. To distinguish the variable PAM sites efficiently, we exploited the 10 bp barcode sequence at the beginning of each surrogate target sequence. Relative indel frequencies, expressed in percentage, were obtained separately from the read counts measured on day 6

and day 10 (Dox− or Dox+) respectively as:

$$\text{indel frequency}\,(\%) = \frac{\text{Num. reads with indels}}{\text{Num. total reads}} \times 100 \quad (3)$$

The drop-out of our protocol was minimal, and we obtained the following total unique PAM contexts: gRNA for *BMP3* $n = 251$; *ADRA2C* $n = 255$; *CHRNB2* $n = 250$; *PAX5* $n = 255$. A threshold on the minimum number of total reads was defined by iteratively removing samples with less than n total reads (n from 0 to 200, step size of 5). The value of n with the lowest associated Spearman's correlation p value between the indel frequencies measured on day 6 and day 10 was selected. This procedure was performed separately for Dox− and Dox+, for which the resulting minimum reads thresholds were 90 and 35, respectively. Indel frequencies from day 6 and day 10 were then averaged. This data are available as Supplementary Data 1. For each gRNA, multiple targets of one selected context were evaluated using different barcodes. These presented a mean normalised standard deviation of 0.03 and 0.11 in Dox+ and Dox−, and their efficiencies were averaged (Supplementary Fig. 9).

Collection and pre-processing of external datasets

Merged Xiang et al. dataset (2021). The dataset employed to study the relationship between binding free energy changes and on-target cleavage efficiency was obtained by merging the data of Kim et al.[14] and Xiang et al.[17] as previously described in Xiang et al.[17]. This dataset consists of 23,902 gRNA sequences and corresponding Cas9 efficiencies, measured as indel frequencies, at targets flanked by 5′-N$_{-1}$GGN$_{+1}$-3′ PAMs. The specificity of gRNAs (CRISPRspec) was evaluated with the CRISPRoff pipeline v.1.1.2[20] on the human genome hg38. The following two gRNAs failed to be evaluated by the pipeline within 2 weeks due to the extreme presence of off-targets and were thus excluded: TAAAAATACAAAAAATTAGC, T(x20). We also removed 806 gRNAs with no match to the genome hg38 (randomised gRNAs). Next, the dataset was filtered to remove gRNAs likely to form sub-optimal structures with the scaffold. For this, pairwise binding probabilities of the bases in the sgRNAs (gRNAs + scaffold) were evaluated with RNAfold v.2.2.5[48]. The sgRNAs were compared to the optimal secondary structure of the scaffold, computed by solely folding the scaffold with RNAfold. This structure includes the crRNA:tracrRNA fusion loop and three tail loops[49]. The comparison consisted in measuring the Euclidean distance between binding probabilities for pairs of bases that bind the optimal structure, or the probability to be unbound for those bases that do not pair with any other in the optimal structure (1 – probability of pairing with any other base). For each of the two distance measures, we tested if the 5% sgRNAs with the highest distance to the optimal structure is less efficient than other sgRNAs (Mann–Whitney one-sided test, see Supplementary Fig. 2). A significant result was obtained for the unpaired bases ($p = 3.68E\text{-}15$) while for paired bases the distance was not highly pronounced ($p = 5.31E\text{-}02$). Hence, we removed from the dataset 1155 gRNAs with a high distance to the optimal structure in terms of the probability of bases unpaired in the optimal structure to not pairing with any other base in folded sgRNAs. Additionally, we filtered out gRNAs with indel frequencies at target sites <2% ($n = 539$). The resulting 21,402 gRNAs were split into a training set ($n = 14{,}981$) and a test set ($n = 6421$), keeping sequences with <4 differences in the 30mer (gRNA + target context) in the same subset. The analysis of the free energy change properties was carried out on the training set only, filtered down to 11,602 unique elements by removing gRNAs with low specificity (CRISPRspec <5). This last filter was not applied for the evaluation of the impact of local off-targets on gRNA specificity and efficiency.

Dataset by Lin et al. (2014). The data from Lin et al. consists in 4 gRNAs: R-01, R-30, R-08 and R-25. DNA bulges were generated at any position on the targets by removal of single bases in the gRNAs. Sequences and efficiencies were retrieved from the corresponding publication[29].

Dataset by Tsai et al. (2015). From the GUIDE-seq dataset[30], we analysed four gRNAs with at least one off-target cleaved more efficiently than the on-target. These were processed to remove off-targets with differences to the on-target PAM (3 nt NGG), or mismatches to the gRNA in the PAM-proximal 4 nt, or more than three total mismatches to the gRNA.

Dataset by Hart et al. (2015). The dataset of Hart et al.[50] (4239 gRNAs) was retrieved from the study of Haeussler et al. as dataset Hct1162lib1Avg[51]. In addition to the removal of one gRNA with no match in hg38, gRNAs not overlapping their target gene in the CDS annotations of GENCODE v.32[52] were eliminated, resulting in 4184 gRNAs. The dataset was also filtered to remove gRNAs with high Euclidean distance to the optimal scaffold structure in terms of unpaired bases, using the same threshold identified for the top 5% distant gRNAs described above (Euclidean distance threshold = 3.41). The processed dataset comprises 4066 gRNAs. The efficiencies, measured as fold-change in gRNA abundance, were ranked-normalised with SciPy rankdata[53].

Datasets by Kim et al. (2020). The dataset of Cas9 variants activity and PAM compatibility of 30 gRNAs was downloaded from Kim et al.[16]. All targets were designed to be followed by the same context (TA) after 4-nt PAMs variated in all possible ways (NNNNTA, $n = 4^4 = 256$ contexts), except for 15 AGGVB targets (V ∈ {A, C, G} and B ∈ {C, G, T}), which we excluded. We analysed Cas9 variants compatible with NGNN PAMs and all high-fidelity Cas9 variants except for evoCas9, which showed low mean indel frequency at targets with GG PAM binding site (mean =12.66%). The following number of gRNAs for each Cas9 variant were employed: SpCas9-HF1 $n = 7400$, Sniper-Cas9 $n = 7395$, HypaCas9 $n = 7392$, eSpCas9(1.1) $n = 7388$, wild-type SpCas9 $n = 7382$, SpCas9-NG $n = 7379$, xCas9 $n = 7367$.

Thermodynamic properties of gRNAs and DNA targets. The gRNA folding minimum free energy change (ΔG_U) was computed with RNAfold v.2.2.5[48] using default options. Hybridisation free energy changes of complementary gRNA-target-DNA (ΔG_H) and DNA–DNA interactions (ΔG_O) were computed with the CRISPRoff pipeline v.1.1.2[20]. The pipeline also provides gRNA specificity information (CRISPRspec) if a list of targets in the genome is provided. The off-targets provided to CRISPRoff were searched in the human genome hg38 with RIsearch2 (v.2.1)[54] with the options recommended for usage in combination with CRISPRoff: gRNA seed region from position 1 to 20, the maximum number of mismatches allowed = 6 with no constraint on a minimum number of consecutive matches in the seed, energy upper threshold = 10,000, maximum extension length on the seed = 0, output format = p3. The surrogate DNA targets used in the datasets of Kim et al. (2019–2020), Xiang et al.[17], and in our validation datasets, that represent the surrogate duplicates of on-target sites and have unknown genomic positions, were not included in the search. CRISPRoff was executed with default parameters. Sub-optimal gRNA–DNA interactions were evaluated with an extended version of RIsearch v.1.1[55] developed for this project (RIsearch v.1.2), to allow the scoring of sub-optimal gRNA–DNA interactions flanking sliding PAMs with positional weights on stacking base pairs (see https://rth.dk/resources/risearch). Given a gRNA as query and a DNA sequence as a target, RIsearch v.1.2 was executed with the following options: matrix of RNA–DNA nearest neighbour parameters 'Su95' (same as in CRISPRoff); force-start, to require the interaction to start at the 3′ end of the target; -w CRISPR_20nt_5p_3p, where 'CRISPR_20nt_5p_3p' is the name of the pre-compiled file containing the array of 19 weights for stacking base-pair contributions defined in CRISPRoff[20]. All gRNA–DNA interactions were forced to end at the PAM. The optimal start position was searched by iteratively shortening the DNA target sequence by 1 nt from an initial length of 24 nt upstream of the PAM to a minimum of 2 nt. For targets at PAMs overlapping while sliding 1 nt up/downstream only up to 21 nt were effectively utilised by the algorithm to find optimal gRNA–DNA interactions. This number varies in the case of the data from Lin et al. and GUIDE-seq. The DNA–DNA opening free energy change at bindings with bulges or mismatches was calculated for the sequence stretch involved in the gRNA–DNA interaction, which may be longer or shorter compared to the fully complementary case. For this, the function for DNA–DNA binding free energy change computation of CRISPRoff was used. In the data from Lin et al. and GUIDE-seq, in which gRNA–DNA bindings can have different lengths, out of all possible sub-optimal interactions the one with the lowest ΔG_B was considered the optimal one. For all RNA–RNA, RNA–DNA and DNA–DNA stacking interactions, free energy change parameters were obtained from RIsearch2, which includes parameters for DNA–DNA[56,57], RNA–RNA[58] and RNA–DNA[59–61] stacking interactions, as previously described[20]. Missing parameters, such as those for stacked bulges in RNA–DNA interactions, were obtained by averaging the corresponding RNA–RNA and DNA–DNA parameters, as in previous studies[20,62].

Integration of local sliding PAMs in CRISPRspec. Local sliding PAMs were included in the CRISPRspec score measuring binding competition by considering bindings flanking sliding PAMs as pseudo-off-targets and by including them in the partition function calculation of CRISPRspec. By linearly fitting to the efficiency, we estimate the weighted contributions of bindings attributed to the local sliding PAMs and added it on top of the original CRISPRspec measure obtained from the CRISPRoff pipeline v.1.1.2[20]. Off-targets with up to six mismatches in the human genome hg38 were searched with RIsearch2 (v.2.1)[54] as explained above. Given the set S of all gRNAs, let S_{up} and S_{down} be the subsets of gRNAs in S that have respectively upstream and downstream local sliding PAMs. For a gRNA x, let M be the median function, $|z|$ the absolute value of z, $\triangle G_B\left[x_{up}\right]$ and $\triangle G_B\left[x_{down}\right]$ the ΔG_B free energy changes computed at up- and down-stream binding sites of x, and $y[x]$ the experimentally measured efficiency of x. Then, the extended CRISPRspec, CRISPRspecExt, of gRNA x takes the local sliding PAMs into account as follows:

$$\begin{aligned}\text{CRISPRspecExt}[x] = {} & \alpha_0 + \alpha_1 \text{CRISPRspec}[x] \\ & + \lambda_{\text{up}}[x]\left(\beta_0 + \beta_1 \left|\triangle G_B\left[x_{\text{up}}\right] - M\left(\left\{\triangle G_B[i],\ \forall i \in S_{\text{up}}\right\}\right)\right|\right) \\ & + \lambda_{\text{down}}[x]\left(\gamma_0 + \gamma_1 \left|\triangle G_B\left[x_{\text{down}}\right] - M\left(\left\{\triangle G_B[i],\ \forall i \in S_{\text{down}}\right\}\right)\right|\right)\end{aligned} \quad (4)$$

With:

$$\lambda_{\text{up}}[x] = \begin{cases} 0 & \text{if } x \notin S_{\text{up}} \\ 1 & \text{if } x \in S_{\text{up}} \end{cases} \quad (5)$$

$$\lambda_{\text{down}}[x] = \begin{cases} 0 & \text{if } x \notin S_{\text{down}} \\ 1 & \text{if } x \in S_{\text{down}} \end{cases} \quad (6)$$

The optimal least squares coefficients $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ were estimated as:

$$\hat{\alpha}_0, \hat{\alpha}_1 = \mathrm{argmin}_{\alpha_0,\alpha_1} \sum_{x \in S} \left(y[x] - \alpha_0 - \alpha_1 CRISPRspec[x]\right)^2 \quad (7)$$

$$\hat{\beta}_0, \hat{\beta}_1 = \mathrm{argmin}_{\beta_0,\beta_1} \sum_{x \in S_{\mathrm{up}}} \left(y[x] - \beta_0 - \beta_1 \left|\triangle G_B\left[x_{\mathrm{up}}\right] - M\left(\left\{\triangle G_B[i], \forall i \in S_{\mathrm{up}}\right\}\right)\right|\right)^2 \quad (8)$$

$$\hat{\gamma}_0, \hat{\gamma}_1 = \mathrm{argmin}_{\gamma_0,\gamma_1} \sum_{x \in S_{\mathrm{down}}} \left(y[x] - \gamma_0 - \gamma_1 \left|\triangle G_B\left[x_{\mathrm{down}}\right] - M\left(\left\{\triangle G_B[i], \forall i \in S_{\mathrm{down}}\right\}\right)\right|\right)^2 \quad (9)$$

Hence, parameters for local sliding PAMs are estimated from the absolute deviation of ΔG_B to the median ΔG_B of all local sliding PAMs of the same type (up/down) in the training set.

Statistical analysis. The SciPy v.1.5.0-1.6.3[53], NumPy v.1.18.5[63], pandas v.1.2.0[64] and scikit-learn v.0.23.1[65] modules were applied for data analysis in Python v.3.8.3[66]. Nucleotide strings were managed in Biopython v.1.77. Plots were generated using Matplotlib v.3.2.2 and seaborn v.0.11.1. The impact of gRNA free energy change properties in defining gene knockout efficiency was analysed after sorting and separating gRNAs into two groups, high (top 20%) and low (bottom 20%) efficient, based on their reported indel frequencies (Supplementary Fig. 19). The intervals of preferential values identified for each free energy change property were set to contain 80% of the total highly efficient gRNAs. The portion of efficiency variance associated to the differences in the context of the 'GG' PAM binding site that could be explained by the sliding model was computed as the sum of squares ($SSQ_{sliding}$) between the efficiencies of each sliding condition and the grand mean (GM) divided by the total sum of squares between the efficiencies and the grand mean (SSQ_{tot}):

$$\text{Explained variance} = \frac{SSQ_{sliding}}{SSQ_{tot}} \quad (10)$$

With:

$$SSQ_{sliding} = n_{up}\left(M_{up} - GM\right)^2 + n_{down}\left(M_{down} - GM\right)^2 + + n_{up\&down}\left(M_{up\&down} - GM\right)^2 + n_{none}\left(M_{none} - GM\right)^2 \quad (11)$$

$$SSQ_{tot} = \sum_{s \in S}\left(x_s - GM\right)^2 \quad (12)$$

$$GM = \frac{n_{up}M_{up} + n_{down}M_{down} + n_{up\&down}M_{up\&down} + n_{none}M_{none}}{n_{up} + n_{down} + n_{up\&down} + n_{none}} \quad (13)$$

Where S is the set of all gRNAs with a 'GG' PAM binding site, x_s is the efficiency of a gRNA, and n_{up}, n_{down}, $n_{up\&down}$, n_{none}, M_{up}, M_{down}, $M_{up\&down}$, M_{none} are the size (n) and the mean (M) of respectively the data subsets split by sliding categories: upstream (up); downstream (down); upstream and downstream (up&down); no sliding (none).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The raw sequencing data generated in this study have been deposited in the NCBI Sequence Read Archive under accession code BioProject: PRJNA732236 and in the China National GeneBank under accession code CNP0001874. The primary sequence of the human genome hg38 was downloaded from the NCBI, RefSeq assembly accession GCF_000001405.38. Additional data used in this study are available as supplementary material in the following publications: dataset by Xiang et al.[17] [https://doi.org/10.1038/s41467-021-23576-0], dataset by Kim et al.[14] [https://doi.org/10.1126/sciadv.aax9249], dataset by Lin et al[29]. [https://doi.org/10.1093/nar/gku402], dataset by Tsai et al.[30] [https://doi.org/10.1038/nbt.3117], dataset by Hart et al.[50,51] [https://doi.org/10.1186/s13059-016-1012-2], datasets by Kim et al.[16] [https://doi.org/10.1038/s41587-020-0537-9]. Source data for the figures and supplementary figures are provided as a Source Data file. Source data are provided with this paper.

Code availability

The source code of RIsearch v.1.2 is available under the GNU General Public Licence v.3 via https://github.com/RTH-tools/risearch/ and https://rth.dk/resources/risearch/.

Received: 25 August 2020; Accepted: 27 April 2022;

References

1. Doudna, J. A. & Charpentier, E. Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science* **346**, 1258096 (2014).
2. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
3. Huai, C. et al. Structural insights into DNA cleavage activation of CRISPR-Cas9 system. *Nat. Commun.* **8**, 1375 (2017).
4. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
5. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
6. Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* **16**, 218 (2015).
7. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
8. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
9. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**, 823–826 (2015).
10. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
11. Labuhn, M. et al. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res.* **46**, 1375–1385 (2018).
12. Peng, H., Zheng, Y., Blumenstein, M., Tao, D. & Li, J. CRISPR/Cas9 cleavage efficiency regression through boosting algorithms and Markov sequence profiling. *Bioinformatics* **34**, 3069–3077 (2018).
13. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 80 (2018).
14. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci Adv.* **5**, eaax9249 (2019).
15. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
16. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
17. Xiang, X. et al. Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning. *Nat. Commun.* **12**, 3238 (2021).
18. Sternberg, S. H., Redding, S., Jinek, M., Greene, E. C. & Doudna, J. A. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. *Nature* **507**, 62–67 (2014).
19. Globyte, V., Lee, S. H., Bae, T., Kim, J. S. & Joo, C. CRISPR/Cas9 searches for a protospacer adjacent motif by lateral diffusion. *EMBO J.* **38**, e99466 (2019).
20. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* **19**, 177 (2018).
21. Sternberg, S. H., LaFrance, B., Kaplan, M. & Doudna, J. A. Conformational control of DNA target cleavage by CRISPR-Cas9. *Nature* **527**, 110–113 (2015).
22. Dagdas, Y. S., Chen, J. S., Sternberg, S. H., Doudna, J. A. & Yildiz, A. A conformational checkpoint between DNA binding and cleavage by CRISPR-Cas9. *Sci. Adv.* **3**, eaao0027 (2017).
23. Boyle, E. A. et al. High-throughput biochemical profiling reveals sequence determinants of dCas9 off-target binding and unbinding. *Proc. Natl Acad. Sci. USA* **114**, 5461–5466 (2017).
24. Thyme, S. B., Akhmetova, L., Montague, T. G., Valen, E. & Schier, A. F. Internal guide RNA interactions interfere with Cas9-mediated cleavage. *Nat. Commun.* **7**, 11750 (2016).
25. Jensen, K. T. et al. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.* **591**, 1892–1901 (2017).
26. Nishimasu, H. et al. Crystal structure of Cas9 in complex with guide RNA and target DNA. *Cell* **156**, 935–949 (2014).
27. Wu, X. et al. Genome-wide binding of the CRISPR endonuclease Cas9 in mammalian cells. *Nat. Biotechnol.* **32**, 670–676 (2014).
28. Gao, Z., Herrera-Carrillo, E. & Berkhout, B. Delineation of the exact transcription termination signal for type 3 polymerase III. *Mol. Ther. Nucleic Acids* **10**, 36–44 (2018).
29. Lin, Y. et al. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res.* **42**, 7473–7485 (2014).
30. Tsai, S. Q. et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat. Biotechnol.* **33**, 187–197 (2015).
31. Slaymaker, I. M. et al. Rationally engineered Cas9 nucleases with improved specificity. *Science* **351**, 84–88 (2016).
32. Kleinstiver, B. P. et al. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. *Nature* **529**, 490–495 (2016).
33. Chen, J. S. et al. Enhanced proofreading governs CRISPR–Cas9 targeting accuracy. *Nature* **550**, 407–410 (2017).

Published online 29 November 2021

Nucleic Acids Research, 2022, Vol. 50, No. 4 ***e20***
https://doi.org/10.1093/nar/gkab1131

CRISPRroots: on- and off-target assessment of RNA-seq data in CRISPR–Cas9 edited cells

Giulia I. Corsi, Veerendra P. Gadekar, Jan Gorodkin* and Stefan E. Seemann*

Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen, Thorvaldsensvej 57, 1871 Frederiksberg, Denmark

Received June 16, 2021; Revised October 14, 2021; Editorial Decision October 26, 2021; Accepted October 26, 2021

ABSTRACT

The CRISPR-Cas9 genome editing tool is used to study genomic variants and gene knockouts, and can be combined with transcriptomic analyses to measure the effects of such alterations on gene expression. But how can one be sure that differential gene expression is due to a successful intended edit and not to an off-target event, without performing an often resource-demanding genome-wide sequencing of the edited cell or strain? To address this question we developed CRISPRroots: CRISPR–Cas9-mediated edits with accompanying RNA-seq data assessed for on-target and off-target sites. Our method combines Cas9 and guide RNA binding properties, gene expression changes, and sequence variants between edited and non-edited cells to discover potential off-targets. Applied on seven public datasets, CRISPRroots identified critical off-target candidates that were overlooked in all of the corresponding previous studies. CRISPRroots is available via https://rth.dk/resources/crispr.

INTRODUCTION

The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats)-Cas9 system is an RNA-guided antiviral defense complex capable of cleaving foreign DNA complementary to a short segment of a guide RNA (gRNA) molecule at a DNA site juxtaposed to a motif known as PAM (Protospacer Adjacent Motif) (1). This machinery, originally discovered in prokaryotes, has recently been transformed into a multipurpose genome engineering and visualization technology (2). Among the main applications of CRISPR–Cas9 there are (i) gene knockouts, used to investigate the effects of single or multiple allele losses and (ii) knockins of sequence variants, in which endogenous genes are altered to study genetic disorders. In the former case, gene loss is achieved by mutagenic errors at the cleavage site introduced by error-prone DNA repair pathways such as the non-homologous end-joining (NHEJ) or the microhomology-mediated end joining (MMEJ) (3). In the latter case, a DNA template carrying the genomic variant of interest is delivered to the cell and integrated via homology-directed repair (HDR) after cleavage at a nearby location (4). Following Cas9-mediated editing, cells are sequenced at the targeted locus to examine if the editing was successful. Additionally, few off-target sites predicted by bioinformatics tools based on a gRNA–target sequence-similarity search are typically sequenced to verify the absence of unwanted cleavage events (5).

Genome engineering can be combined with RNA sequencing (RNA-seq) to identify genes whose expression levels are altered as a consequence of the edit (knockin or knockout) (6–13). RNA-seq data can additionally be used to evaluate the presence and abundance of the modified transcript and its possible down-regulation, or total absence, after monoallelic or multiallelic knockout (8). In this regard, RNA-seq can also highlight unwanted editing effects that remain hidden in the sequencing of a short DNA region overlapping the target cleavage site, such as extended loss of heterozygosity and partial or complete loss of a chromosome, all events that have been previously observed in Cas9-edited cells (14–18). In the past few years, RNA-seq data was used in the analysis of potential Cas9 off-target effects either by comparing variants discovered in the transcriptome of edited and non-edited cells (9), or by incorporating off-target predictions with gene expression changes to identify downregulated genes overlapping potential off-targets (7). Although neither of these methods provide a complete assessment, the combination of both allows to prioritize predicted off-targets for validation by pointing to scenarios presenting tangible transcriptome variations. This procedure exploits fully the RNA-seq data, which is instead ignored by generic off-target prediction tools that are based solely on the search for gRNA targets in a given genome while ignoring the transcriptional activity.

Although the literature currently contains a modest number of studies applying CRISPR editing in combination with RNA-seq of at least three replicates of edited and

*To whom correspondence should be addressed. Email: seemann@rth.dk
Correspondence may also be addressed to Jan Gorodkin. Email: gorodkin@rth.dk

wild-type samples (required for statistical significance), we anticipate that the number of such studies will grow rapidly in the future. In 2011, there were only two papers in PubMed (19) combining CRISPR and RNA-seq, while this number has increased to about 300 per year, with a current total of 843 (Supplementary Table S1). Automatizing the screening of such datasets is currently hindered by the lack of details on the gRNA(s) and the edited site(s) in the data repositories. These are usually provided separately (e.g. in a related article), and need to be found manually.

To better exploit the potential of RNA-seq data we developed `CRISPRroots`, a tool that compares RNA-seq reads from Cas9-edited cells and corresponding isogenic controls to evaluate potential off-targets and verify on-target editing outcomes. We assess `CRISPRroots` on seven published RNA-seq datasets with at least three replicates of edited and control samples and show that there are multiple potential off-targets of high relevance that were not taken into account by the corresponding studies. The pipeline around `CRISPRroots` integrates pre-processing, mapping, gene quantification, differential expression, off-target prediction, variant discovery, Cas9-gRNA binding properties, and assessment of genome integrity with cutting-edge tools. The `CRISPRroots` pipeline is made in a user-friendly `Snakemake` (20) workflow that optimizes the handling of computing resources, parallelises tasks, and minimizes software prerequisites via the definition of `Conda` environments (https://docs.anaconda.com/), facilitating re-usability and reproducibility.

MATERIALS AND METHODS

Implementation

`CRISPRroots` is implemented as a pipeline consisting of a number of key modules: (1) RNA-seq read processing and mapping; (2) Somatic variant calling; (3) Variant-based off-target screening; (4) Differential gene expression; (5) Assessment of on-target knockins and knockouts; (6) gRNA off-target prediction; (7) Expression-based off-target screening. Combining these modules as depicted in Figure 1 results in an on/off-target report elucidating whether the on-target edit was successful or not and highlighting possible off-target events found in the RNA-seq data or in promoter regions, which are therefore potentially involved in gene expression regulation. In the following we describe the content of these modules.

(1) **RNA-seq read processing and mapping.** The quality of raw reads is assessed with `FastQC` v.0.11.9 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and summarized with `MultiQC` v.1.9 (21). This process is repeated after each of the subsequent filtering steps. The removal of adapters (provided as FASTA files) is performed with `Cutadapt` v.2.10 (22), which also filters out short reads and low-quality reads. Additional filters can be defined in the configuration file. Reads are cleaned from residual ribosomal RNAs with `Bbduk` v.37.62 (https://sourceforge.net/projects/bbmap/). Clean reads are mapped to the genome with `STAR` v.2.6.1a (2-pass mode) (23), and the resulting mapping files are sorted and indexed with `SAMtools` v.1.9 (24).

(2) **Somatic variant-calling.** Somatic variants between multiple edited and wild-type samples are discovered with the `Mutect2` (25) tool from `GATK` v.4.2.0.0 (26) after processing the reads as follows: (i) mapped reads are sorted by query name with `SortSam` (`Picard` v.2.23.0; http://broadinstitute.github.io/picard); (ii) duplicated read pairs are marked and sorted by coordinates with `MarkDuplicates` (GATK); (iii) split reads are separated with `SplitNCigarReads` (GATK); (iv) short variants are called with `Mutect2` (min base quality=30; minimum callable depth=10); (v) results are filtered with `FilterMutectCalls` (GATK). Step (iii) is specific and necessary to call variants in RNA-seq data, as the splicing of introns results in Ns in the CIGAR string describing the mapping. `Mutect2` is used with default options, and learns unknown parameters in the filter models from the unfiltered data (25). Because step (iv) is highly demanding in terms of computational resources, reads are first grouped by chromosome, and separate instances of `Mutect2` are executed in parallel with `GNUparallel` (27).

(3) **Variant-based off-target screening.** Possible cleaved loci are derived from the coordinates and pattern of somatic short variants (SNVs and indels) as follows (Figure 2A): (i) SNVs: the phospho-diester bonds immediately before or after the variated nucleotide; (ii) insertions: the phospho-diester bond linking the nucleotides in the reference between which the insertion is located; (iii) deletions: the phospho-diester bonds immediately before and after any of the removed bases. Knowing that the cut site is three nucleotides upstream from the PAM, all possible related PAM sites, on any strand, are identified. The search for a PAM can be extended up to n nucleotides (default $n = 2$), to account for possible bulges between the PAM and the cut site. Then, possible gRNA binding regions are defined as the complementary sequences upstream of the cut site that have the same length as the gRNA plus an arbitrary number of m nucleotides (default $m = 2$) to account for possible bulges on the DNA. Bulges on the gRNA are allowed as well. Interactions between the gRNA and its possible targets are evaluated in terms of resulting gRNA-DNA binding energy, ΔG_B, and complementarity in the seed region. The ΔG_B is computed following the `CRISPRoff` v.1.1 (28) binding energy model: $\Delta G_B = \delta_{PAM}(\Delta G_H - \Delta G_O - \Delta G_U)$, where ΔG_H is the gRNA-DNA binding energy, ΔG_O is the energy penalty for opening the target DNA, ΔG_U is the penalty for opening up possible gRNA structures, and δ_{PAM} is a PAM weight (NGG = 1, NAG = 0.9, NGA = 0.8) (28). The weighted gRNA-DNA binding energy, ΔG_H, is computed by `RIsearch1` v.1.2 (29), which allows to force interactions to start at the 3′ end of the target (DNA) and end at the 3′ and 5′ ends of the query (gRNA) and the target, respectively. This is done to penalize interactions with PAM-proximal mismatches more severely compared to `CRISPRoff` (a positive energy is added for mismatches instead of not adding any cost) and to enable the evaluation

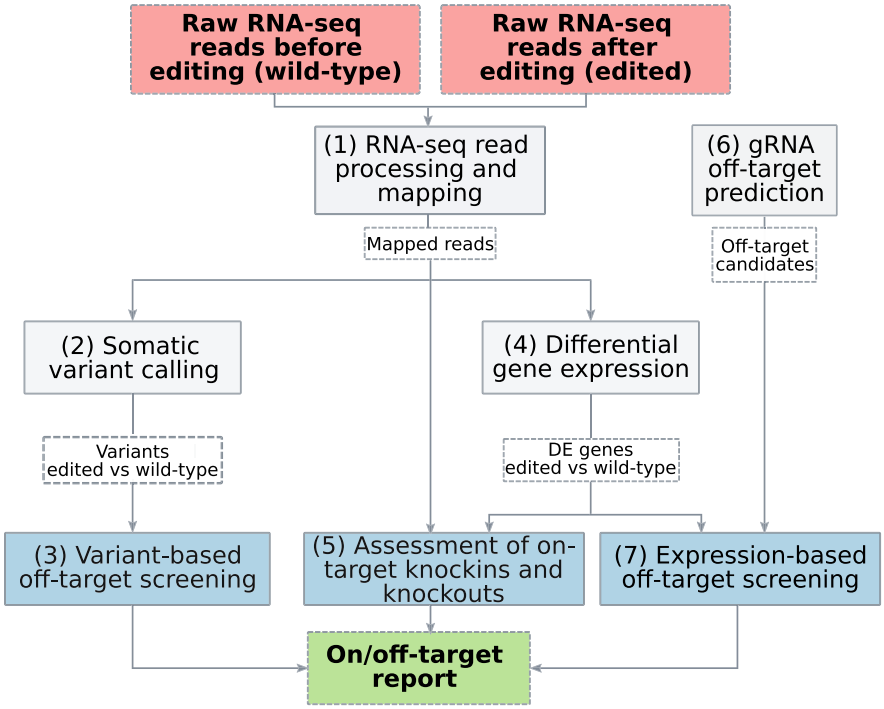


Figure 1. Overview of the `CRISPRroots` pipeline. We implemented the following main external tools in the seven modules: (1) `Cutadapt`, `Bbduk`, `FastQC`, `MultiQC`, `STAR`; (2) `Mutect2`; (3) `RIsearch1`; (4) `featureCounts`, `DESeq2`; (5) `SAMtools`; (6) `RIsearch2`, `CRISPRoff`; and (7) `BEDtools`, `RIsearch1`. The `CRISPRroots` specific modules are colored in blue. Key input/output files are displayed in dashed boxes. As an option, the off-target search and evaluation (modules 3, 6, 7) can run on a variant-aware version of the genome, generated after discovering germline variants with `HaplotypeCaller`.

of potential off-targets to which the gRNA binds forming bulges on the DNA or on the gRNA itself. A positive energy, defined in the `RIsearch1` scoring matrix, is added in the presence of bulges and thus the binding is penalized. In `RIsearch1`, gRNA-target bindings are evaluated using the scoring matrix 'su_95', option '-f' to force the start and end of the interactions, and the same array of Cas9 positional weights defined in `CRISPRoff`. The gRNA minimum free energy, ΔG_U, is obtained with `RNAfold` v.2.2.5 (30). The DNA–DNA opening energy, ΔG_O, is computed with the same function as in `CRISPRoff`, but limiting the calculation to the DNA segment involved in the optimal gRNA–DNA binding to avoid adding energy penalties for unused bases (e.g. Figure 4B shows examples with only a portion of the target DNA being involved in the binding). For each variant, among all possible gRNA bindings starting at any position on the DNA and ending at the PAM site, the one with lowest ΔG_B is retained. Flags are added to the final results to signal repeat-masked regions and known SNPs that were intersected with the variant coordinates using `BEDtools intersectBed` v.2.29.2 (31).

(4) **Differential gene expression.** `featureCounts` from the `Subread` package v.2.0.1 (32) is used to quantify the genes present in a merged set of annotations derived from both GENCODE v.33 (33) and FANTOM-CAT v.1.0.0 (34). Only non-chimeric reads are counted (both mate reads for paired-end sequencing). If known, the library type can be specified directly in the config file. If unknown, the strand specificity can be discovered with `RSeQC` v.4.0.0 (35) within the pipeline's environment. Differential expression analysis is performed on the gene read counts with `DESeq2` from Bioconductor v.3.8 (36). The comparison is done between the conditions 'Edited' and 'Original' (the non-edited wild-type) which are defined in a sample table (see Supplementary Table S2 for an example). A gene is considered differentially expressed (DEG) if its absolute $\log_2$ fold change is >0.5 and the related Benjamini-Hochberg (37) adjusted Wald-test P-value is <0.01. Genes with mean normalized read count across samples <10 are considered as not expressed.

(5) **Assessment of on-target knockins and knockouts.** Expected knockin mutations are defined in the configuration file by their sequence pattern, genomic coordinates, and possible role in splicing (splice donor, splice acceptor, intron). For every edited position `CRISPRroots` summarizes the number of reads reporting the reference nucleotides, variants, skips (which symbolize spliced introns), and other events (insertions or deletions) from a `pileup` of the mapped reads generated with `SAMtools`. Numerical summaries of the read counts for the different alleles and the genotype interpretation are provided. Special attention is given to variations affecting splice sites and introns, which can alter not only the sequence of a transcript but also the way it is spliced (Figure 2B).

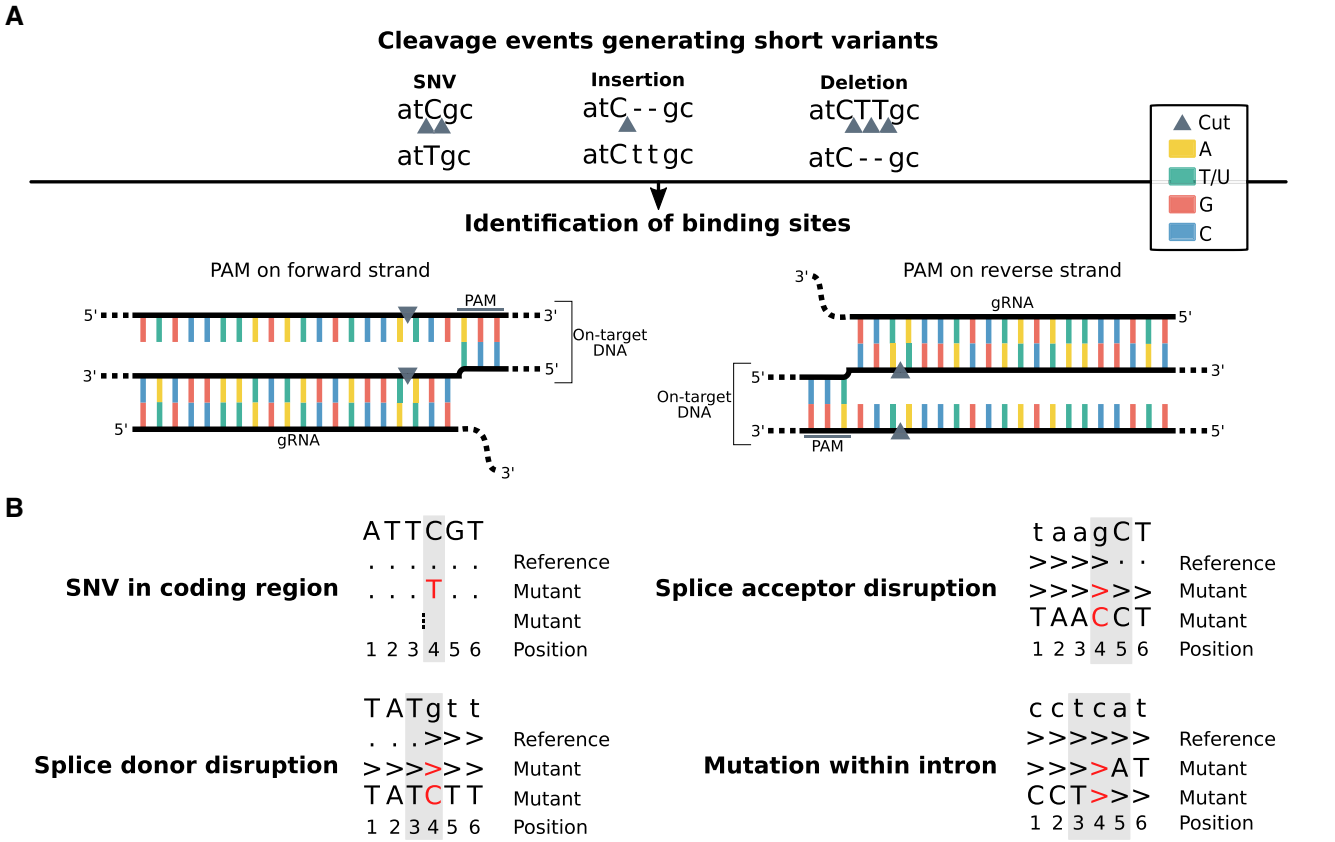


Figure 2. Analysis of sequence variations at possible on-/off-targets. (**A**) Strategy for variant-based off-target screening. Short genomic variants discovered from RNA-seq are screened to find Cas9 binding sites proximal to the possible 'cut' positions associated to the variants. All gRNA–DNA interactions ending at one of the identified binding sites are evaluated, and the energetically most favourable one is retained as most likely off-target for each variant. (**B**) Patterns of on-target single nucleotide variations. Four different types of on-target editing events are shown. For each of them, the reference pileup and examples of other possible mutant pileups (in red) are given. The positions analyzed to evaluate on-target edits are highlighted with grey boxes.

For instance, if a splice acceptor is disrupted, splicing can terminate at a downstream splice acceptor (skipping continuation) or not take place at all (intron retention). Because of this, while SNVs affecting coding loci are assessed at a single genomic position, neighboring nucleotides are included in the evaluation of splice donors, acceptors, and introns. Expression changes at the on-target gene are evaluated with `DESeq2`. Read counts normalized by size factors, the $\log_2$ fold change, and the adjusted *P*-value are summarized in the output.

(6) **gRNA off-target prediction.** To examine potential off-targets that impair expression or that are located in untranscribed regions, and that hence might not be captured by the analysis in (3), we perform a genome-wide search for off-targets with `CRISPRoff`. Following the `CRISPRoff` guidelines, off-targets with up to 6 mismatches to the gRNA are searched with `RIsearch2` v.2.1 (38). This tool enables fast searching of gRNA binding via a suffix arrays approach, but does not allow to constrain and weight gRNA-DNA interactions with any number of bulges as `RIsearch1`. The search is carried out either in the reference genome or, optionally, in its variant-aware version. Potential off-target locations are evaluated with `CRISPRoff`, supplied with `RNAfold` v.2.2.5 (30). Predicted off-targets are filtered to eliminate non-spontaneous bindings ($\Delta G_B > 0$).

(7) **Expression-based off-target screening.** Gene expression changes analyzed with `DESeq2` are employed in concert with off-target predictions to identify candidate off-targets overlapping differentially expressed genes or their promoter regions. The genomic coordinates of DEGs and their promoter regions (by default, 1 kb upstream of the transcription start site) are intersected with the cleavage coordinates of the predicted off-targets with `BEDtools`. The ΔG_B, initially calculated by `CRISPRoff`, is re-evaluated with `RIsearch1` with the same strategy described above for the variant-based screening, to obtain a more precise evaluation of the binding. An exception are potential cleavage events inside an expressed gene or its promoter that is localized on a hemizygous chromosome (e.g. chrX and chrY in male human). If those events are linked to a variant, they are already reported in the output of module (3) and, hence, they are removed from the list of potential expression-based off-targets.

(Optional) Variant-aware reference genome. The search and evaluation of potential off-targets (modules 3, 6 and 7) can be carried out on either the reference genome or on a variant-aware version of it. The variant-aware reference genome includes short genomic variants discovered from RNA-seq with `HaplotypeCaller` (`GATK`) (39). This tool performs local reassembly of haplotypes in regions that differ from a given reference sequence. In contrast to `Mutect2`, which tolerates differences in the ploidy profiles of the detected somatic variants, `HaplotypeCaller` assumes a fixed ploidy as it is designed to call germline variants. The variant-aware genome is generated as follows: (i) split reads are used to call short variants to the reference with `HaplotypeCaller` (minimum phred-scaled confidence for variant calling=20); (ii) results are filtered with `VariantFiltration` (`GATK`) following the `GATK` recommendations (as of 2019, firstly defined in (40)) to remove clusters of SNVs (window size=35, number of SNVs to define a cluster=3) and any variant with either phred-scaled probability of strand bias (FS) > 30 or variance confidence normalized by depth (QD) < 2. Additionally, variants with approximate read depth (DP) <10 are removed. (iii) variants called between non-edited samples and the reference genome are intersected with the `BCFtools` v.1.9 (41) `isec` function keeping only instances carrying identical alleles to produce a solid set of variants to the reference, supported by all samples. (iv) a variant-aware version of the reference genome is generated with `BCFtoolsconsensus`, which also provides a chain file to lift annotation coordinates. The pipeline can be configured to take either the reference (REF) or the alternative (ALT) allele in the presence of heterozygous variants. Although this procedure only provides the union of different haplotypes (non-reference alleles), to our knowledge there is no tool that can insert germline variants in a reference genome while preserving the haplotypes assembled during variant calling. For the test cases presented here, the pipeline was run twice, using in turn the REF and the ALT allele for heterozygous variants. By using the variant-aware genome it may be possible to find potential off-targets that would remain hidden in a reference-based analysis. However, the generation of a variant-aware genome requires significant time and resources, and it did not provide any relevant benefit in the definition of major or critical candidate off-targets concerning our test cases. Thus, the procedure is set as an option in the pipeline.

Datasets

To test the pipeline, seven RNA-seq datasets were retrieved from five public studies. Because one study employed two gRNAs for a single knockout, there are a total number of eight test cases of Cas9-gRNA activity (Table 1).

1-2: QPRT. Haslinger *et al.* generated *QPRT*-homozygous knockout cells by means of two gRNAs targeting different loci, which generated an insertion (QPRT-INS395A) and a deletion (QPRT-DEL268T) at the target sites in SH-SY5Y cells (7). RNA-seq data produced via MACE (massive analysis of cDNA ends) (42) was downloaded in 3 replicates for 3 experimental settings (QPRT-INS395A, QPRT-DEL268T and wild-type empty control (eCtrl)). In the off-target assessment included in the study, no predicted off-target with up to 4 mismatches is reported to overlap a gene downregulated in knockout compared to eCtrl and not downregulated between additionally sequenced wild-type cells and eCtrl.

Table 1. List of test cases and the respective gRNAs and PAMs

Test case	Study	gRNA	PAM
QPRT-INS	(7)	`GCAGCGGGCCAGCGTGTTGA`	`GGG`
QPRT-DEL	(7)	`GCAGTTGAGTTGGGTAAATA`	`TGG`
GRIN2B-FW	(8)	`GATGGCAATGCCATAGCCAG`	`TGG`
GRIN2B-REV	(8)	`AGATTCTGGGTGGAAGCGCC`	`AGG`
APOE	(9)	`CCTCGCCGCGGTACTGCACC`	`AGG`
PIK3CA-HET	(10)	`ATGAATGATGCACATCATGG`	`TGG`
PIK3CA-HOMO	(10)	`ATGAATGATGCACATCATGG`	`TGG`
OGFOD1	(11)	`GGCAGGACGCCGTTCAGTCA`	`CGG`

3-4: GRIN2B. Bell *et al.* generated biallelic *GRIN2B* knockouts with a two-gRNA Cas9-mediated double nickase system, with two gRNAs (GRIN2B-FW and GRIN2B-REV) and differentiated the cells in cortical neurons (8). Of note, the usage of a double nickase system is expected to importantly reduce, but not abolish, off-target activity (43). RNA-seq data was downloaded in 4 replicates for both knockout and control cells.

5: APOE. *APOE3* to *APOE4* induced pluripotent stem cells (iPSCs) were generated by Lin *et al.* (9) and RNA-seq data was sequenced in 3 replicates for both edited and non-edited cells. The study also presents an off-target analysis based on exonic variants between edited *APOE4* iPSCs and parental *APOE3* iPSC, which did not highlight any variation possibly related to off-targets.

6-7: PIK3CA. Heterozygous and homozygous knockins of $PIK3CA^{H1047R}$ in iPSCs were obtained by Madsen *et al.* (10). RNA-seq data in 3 replicates was downloaded for heterozygous (PIK3CA-HET), homozygous (PIK3CA-HOMO), and wild-type iPSCs. The authors confirmed the absence of unwanted edits at 17 off-target locations predicted with http://crispr.mit.edu from the Zhang Lab or Cas-OFFinder (44) by Sanger sequencing.

8: OGFOD1. The effect of Cas9-mediated homozygous knockout of *OGFOD1* in cardiomyocytes was investigated by Stoehr *et al.* (11). The top 20 off-targets predicted by `CRISPOR` (45) were sequenced, without finding mutations attributable to off-target effects (11). RNA-seq data was downloaded in four replicates for both knockout and wild-type cells.

Data pre-processing

As part of the `CRISPRroots` pipeline, raw RNA-seq reads were pre-processed by removing low quality 3′ ends (min phred score = 30), adapters, dangling Ns, and reads shorter than 90% of their original length after cleaning.

RESULTS

Assessment of CRISPR–Cas9 on-target editing activity

We applied the `CRISPRroots` pipeline (version 1.1) on seven public RNA-seq datasets from both Cas9 knockout

Table 2. Properties of the RNA-seq datasets selected for testing the pipeline. The sequencing strategies, the approximate number of reads (or pairs of reads in paired-end sequencing) before and after pre-processing, mapping to the human genome (hg38), and feature-assignment to a set of merged GENCODE (33) and FANTOM-CAT (34) annotations are reported for each of the seven datasets

Dataset	Sequencing protocol, read length (nt)	Raw reads min–max (M)	Pre-proc. reads min–max (M)	Uniquely mapped reads mean ± std (M)	Mapped reads assigned to a feature mean ± std (M)
QPRT-INS (7)	single, <69	5.5–10.5	4.5–8.2	5.2 ± 1.0	4.6 ± 0.9
QPRT-DEL (7)	single, <69	6.0–8.2	4.5–6.4	4.7 ± 0.6	4.2 ± 0.6
GRIN2B-FW/REV (8)	paired, 125	35.1–44.62	27.1–34.0	29.6 ± 2.3	26.4 ± 2.0
APOE (9)	single, 50	12.4–14.5	9.4–12.3	8.5 ± 1.1	7.2 ± 0.8
PIK3CA-HET (10)	single, 50	22.9–32.0	22.5–31.5	20.7 ± 2.5	18.8 ± 2.3
PIK3CA-HOMO (10)	single, 50	23.1–32	22.7–31.5	21.7 ± 3.1	19.7 ± 2.8
OGFOD1 (11)	paired, 50	55.6–83.3	47.4–71.8	51.9 ± 6.8	43.9 ± 5.8

(QPRT-INS, QPRT-DEL, GRIN2B-FW/REV, OGFOD1) (7,8,11) and knockin (APOE, PIK3CA-HET, PIK3CA-HOMO) (9,10) experiments. As mentioned above these seven datasets constitute eight test cases, as two gRNAs (FW and REV) were employed for the knockout of *GRIN2B* in the GRIN2B-FW/REV dataset (Table 1). The datasets are highly heterogeneous in terms of cell types, library preparation, sequencing strategy, and sequencing depth (Table 2). The amount of sequenced reads varies from 5.5–10.5 M in the MACE-sequenced samples (QPRT-INS and QPRT-DEL datasets) to 12.4–83.3 M reads (or paired-end reads) in other samples. The heterogeneity of these datasets allows us to assess the stability of `CRISPRroots` in the presence of input data with different properties.

Distinct strategies are applied on knockout and knockin experiments to assess on-target editing activities, as explained below. Depending on the settings employed for editing, a successful knockout is indicated by a significant loss or complete absence of the target gene in the transcriptome and/or by the presence of loss of function indels at the cleavage locus in aberrant transcripts. We evaluate the knockout effectiveness by comparing the expression level of the target gene in the edited and non-edited cells, and by genotyping target locations on the DNA from mapped RNA-seq reads. We find that three of the four homozygous knockout datasets show a significant downregulation of the respective target genes (Figure 3A): QPRT-INS $\log_2$ fold-change (l2fc) = –3.27, Benjamini-Hochberg adjusted Wald test *P*-value (*P*-adj) = 7.8e-187; QPRT-DEL l2fc = –2.50, *P*-adj = 6.9e-132; and OGFOD1 l2fc = –1.68, *P*-adj = 1.6e-52. In the dataset GRIN2B-FW/REV the expression of the target gene is not downregulated (l2fc = –0.02, *P*-adj = 0.979).

RNA-seq reads mapping at the target cleavage sites of the two gRNAs employed for the knockout of *GRIN2B* (FW and REV) reveal that the edited cells bear in-frame deletions of variable length. These deletions generate 'skips' in the mapping of RNA-seq reads to the genome at the target cleavage sites (Figure 3B). The presence of deletions was also substantiated by Sanger sequencing in the original publication (8), in which these deletions were characterized as frame-shifting. For the QPRT-INS and QPRT-DEL datasets, the status of the on-target edits cannot be assessed from the mapped reads, as the applied MACE sequencing protocol only sequences the 3′ ends of the RNA (7,42). In the OGFOD1 dataset, all the reads fully overlapping the edited locus have deletions of 4 nt, as previously validated by Sanger sequencing in the related study (11).

On-target edits in Cas9-directed knockin datasets are inspected using mapped RNA-seq reads in edited and non-edited lines. Silent mutations introduced to avoid successive Cas9 cleavage are also assessed. Our screening shows that in the APOE, PIK3CA-HET and PIK3CA-HOMO dataset almost all reads (> 99%) at the editing loci map to the wild-type allele in the non-edited cells (Figure 3B). In the APOE edited lines the wild-type allele is substituted with the designed one, and the latter is present in > 96% of the reads covering the edited positions in two of the replicates and in > 88% in a third replicate. The remaining mapped reads contain skips in correspondence to the editing site. In replicate 3, which has the lowest percentage of edited reads, there are 2 skip reads out of 17 total reads covering one edited site (chr19:44908684) and 2 out of 19 at the other (chr19:44908692) (Figure 3B). Reads mapping to the three edits in PIK3CA-HET knockin carry homozygous silent mutations at positions chr3:179234289 and chr3:179234292. The reads covering the third editing site (A>G H1047R target edit at chr3:179234297) are heterozygous in only one replicate, while a second replicate exclusively possesses the reference nucleotide at this locus. However, this is supported by only three reads. No read mapping to the edited loci is found in a third replicate. The homozygous editing at the same coordinates in PIK3CA-HOMO is supported by the exclusive presence of the designed nucleotides at all the edited locations in two of the three replicates, while no read maps to these sites in a third replicate. Of note, the read coverage at these genomic coordinates is low in both wild-type and edited cells (six reads in one replicate and two in the other for all three editing sites). The expression levels of edited genes are also evaluated, as significant downregulation of a gene targeted for editing may signal the partial or complete loss of a chromosome due to Cas9 cleavage. In the analyzed knockin datasets the expression of the edited genes does not change (Figure 3A): APOE l2fc = −0.03, *P*-adj = 1; PIK3CA-HET l2fc = 0.11, *P*-adj = 1; PIK3CA-HOMO l2fc = −0.21, *P*-adj = 0.46.

Identification of potential CRISPR–Cas9 off-target sites

Cas9 off-target activity at sites located within a gene or any genomic feature that affect transcription, such as promoters and enhancers, can produce sequence variations

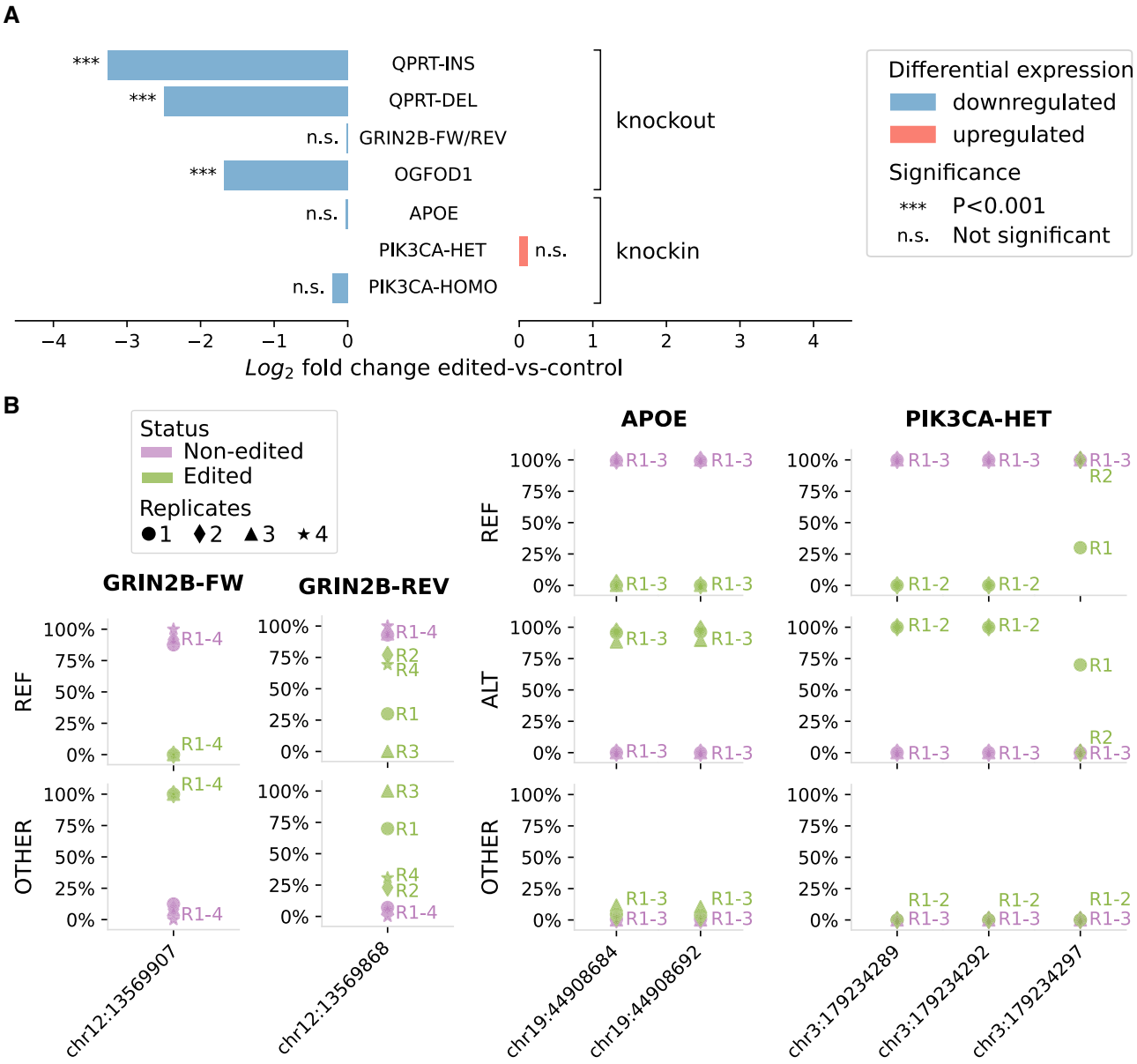


Figure 3. On-target Cas9-mediated edits in test datasets. (**A**) Transcript expression log2 fold change of genes targeted for Cas9-directed knockout or knockin computed by comparing expression levels in edited and non-edited cells with `DESeq2`. Significance determined by the Benjamini-Hochberg adjusted Wald test. (**B**) Fraction of reads mapping to edited nucleotides carrying the reference allele (*REF*), the alternative one (the intended edit) (*ALT*) or anything else (variant/indel/skip) (*OTHER*) in the test datasets GRIN2B, APOE, and PIK3CA-HET. The genomic coordinates of the edit in the human genome (hg38) are reported on the X-axis. Cell replicates are represented with different symbols. Note that only two of the three edited replicates of PIK3CA-HET have reads overlapping the loci described.

that are conveyed to the transcriptome or alter gene expression. To discover potential off-targets from RNA-seq data, we propose a combination of two strategies: the analysis of Cas9 binding sites linked to genomic variants, and the identification of differentially expressed genes harboring predicted off-target sequences. A well supported variant discovered between edited and non-edited cells can delineate off-target events and the corresponding sequence changes introduced by the DNA repair process. Instead, predicted off-targets linked to a differentially expressed gene but without a well supported variant need to be validated by additional DNA sequencing to account for possible silenced alleles (that carry the variant) in an homologous chromosome, whose sequence is unknown in the RNA-seq. An exception for this are predicted off-targets in hemizygous chromosomes (e.g. the Y chromosome; see Methods). The gRNA-DNA interactions of potential off-targets are evaluated in terms of complementarity and the resulting binding free-energy ΔG_B with a modified version of the `CRISPRoff` (28) energy model (see Methods). Given n as the maximum number of mismatches or bulges in the seed, possible off-target interactions are classified into five categories as follows: (i) critical: binding with fully complementary gRNA−DNA seed and linked

to a downregulated gene or a variant; (ii) major type 1: binding with fully complementary gRNA-DNA seed and linked to an upregulated gene; (iii) major type 2: binding with $\leq n$ mismatches or bulges in the seed and linked to a variant or to a differentially expressed gene; (iv) major type 3: predicted off-target with perfect complementarity to the gRNA but overlapping a not expressed gene or an intergenic region; (v) minor: any other potential variant-based or expression-based off-target. Possible off-target interactions that after the re-evaluation of the binding energy with `RIsearch1` (see Methods) are energetically unfavourable ($\Delta G_B > 0$) or that have more than n mismatches or bulges in the seed are flagged. The RNA-DNA base pairs dG·rU and dT·rG, whose contribution to the gRNA−DNA binding energy is limited compared to that of canonical base pairs, are regarded as matches in the binding pattern.

The off-target screening on the test datasets was carried out by searching for either of the PAMs: NGG, NAG and NGA. Up to $n = 1$ mismatches or bulges were tolerated in the seed region (10 nucleotides from the PAM start position (46)). For each dataset, the off-target analysis was performed on a dedicated variant-aware genome in which short variants to the reference discovered from the RNA-seq data of the wild-type samples were introduced. The procedure was repeated twice, by selecting either the reference or the alternative allele in case of heterozygous background mutations.

Our method identifies critical or major predicted off-targets in all of seven test datasets used for testing (eight test cases, Figure 4A), none of which is among those sequenced in the related studies. In seven test cases, `CRISPRroots` identified at least one potential off-target with up to four total mismatches or bulges to the gRNA (Figure 4B). The only exception is the dataset APOE, whose single candidate off-target has a total of five mismatches and one bulge in its binding pattern to the gRNA. Predicted off-targets classified as critical and overlapping the promoter or sequence of a downregulated gene are detected in six of eight test cases (Figure 4A, B). Particularly many critical predicted off-target sequences appear in PIK2CA-HOMO, and 12 of those have high similarity of the gRNA with repeatmasked sequences followed by an AGA non-canonical PAM site (Supplementary Table S3). Three test datasets have one predicted off-target sequence that is fully complementary (with wobble base pairs counted as matches) to the entire length of the applied gRNA. These off-targets are intergenic in PIK3CA-HET (Figure 4B) and PIK3CA-HOMO, or overlap a non-expressed gene in QPRT-DEL (Supplementary Table S3). Potential off-targets linked to variants discovered between edited and non-edited lines are observed in four of eight test cases (Figure 4A). Of these, the only critical one (no mismatch in the seed) is a C>T variant found in the GRIN2B dataset and related to the gRNA GRIN2B-REV. This variant is located on chromosome 19 at position 44908822 in hg38, and it corresponds to a missense SNP in the *APOE* gene (dbSNP (47): rs7412). The co-occurrence of T at position 44908822 and 44908684 in chromosome 19 is referred to as the *APOE2* allele and it is associated with reduced risk of Alzheimer's disease, while the C variant at chr19:44908822 makes a 'neutral' *APOE* (48). All samples have a T at position chr19:44908684, thus the C>T variant at chr19:44908,822 changes the *APOE* of the cortical neurons in the GRIN2B dataset from neutral to protective. This variant is found in all four *GRIN2B* loss of function samples and in two of four controls (Supplementary Figure S1). Although this variant is relevant in the study of cortical neurons, the fact that it is also present in half of the controls makes Cas9 off-target editing unlikely at this position.

The only dataset with no predicted off-target linked to differential expression is APOE, which has also the lowest amount of DEGs and variants (DEGs n=18, variants $n = 1689$), excluding the MACE-sequenced samples. Hence, we checked for the possibility that our pipeline detects possible off-targets in transcriptomic data just by chance based on the size of the search space, i.e. if the number of genomic variants or of binding sites within DEGs correlates with the number of possible off-targets. We did not find a correlation between the number of expression-based potential off-targets (critical, major type 1, or major type 2) and the number of binding sites in DEGs that were identified by searching for the NGG PAM and its reverse complement (Pearson's $r = 0.31$, P-value $= 0.450$; Figure 4C). Also, the number of potential off-targets identified from the variant-based screening is not correlated with the total number of variants discovered between edited and non-edited cells (Pearson's $r = 0.58$, P-value $= 0.132$; Figure 4D).

Running time

The time required to execute the full `CRISPRroots` pipeline (starting from raw reads) launched on a cluster of standard Linux nodes (Intel® Xeon® CPU E5-2650, 60G RAM and 16 cores) varied from 6 to 8 h for test cases with three replicates per condition (QPRT-INS/DEL, APOE and PIK3CA-HET/HOM) to 15-20 h for test cases with four replicates per condition (GRIN2B and OGFOD1). Up to half of the computing time was consumed by the somatic and germline variant calling.

DISCUSSION

Selecting gRNAs with high on-target effectiveness and low off-target potential is the main objective in the design of Cas9-mediated genome engineering. Following the editing, intended on-target modifications and a restricted number of predicted off-targets are usually validated by DNA sequencing. Given the designed gRNA and a reference genome, off-target predictions are identified and scored by computational tools based on the gRNA sequence similarity and/or other binding properties in relation to DNA sites flanked by valid PAMs. Within this process, the information present in eventual RNA-seq data associated with the experiments remains unused. Previous attempts in exploiting RNA-seq to discover off-targets are rather incomplete, as they employed exclusively either variant discovery (9) or expression changes (7). The latter strategy was additionally limited by the parameters employed for off-target prediction, which allowed up

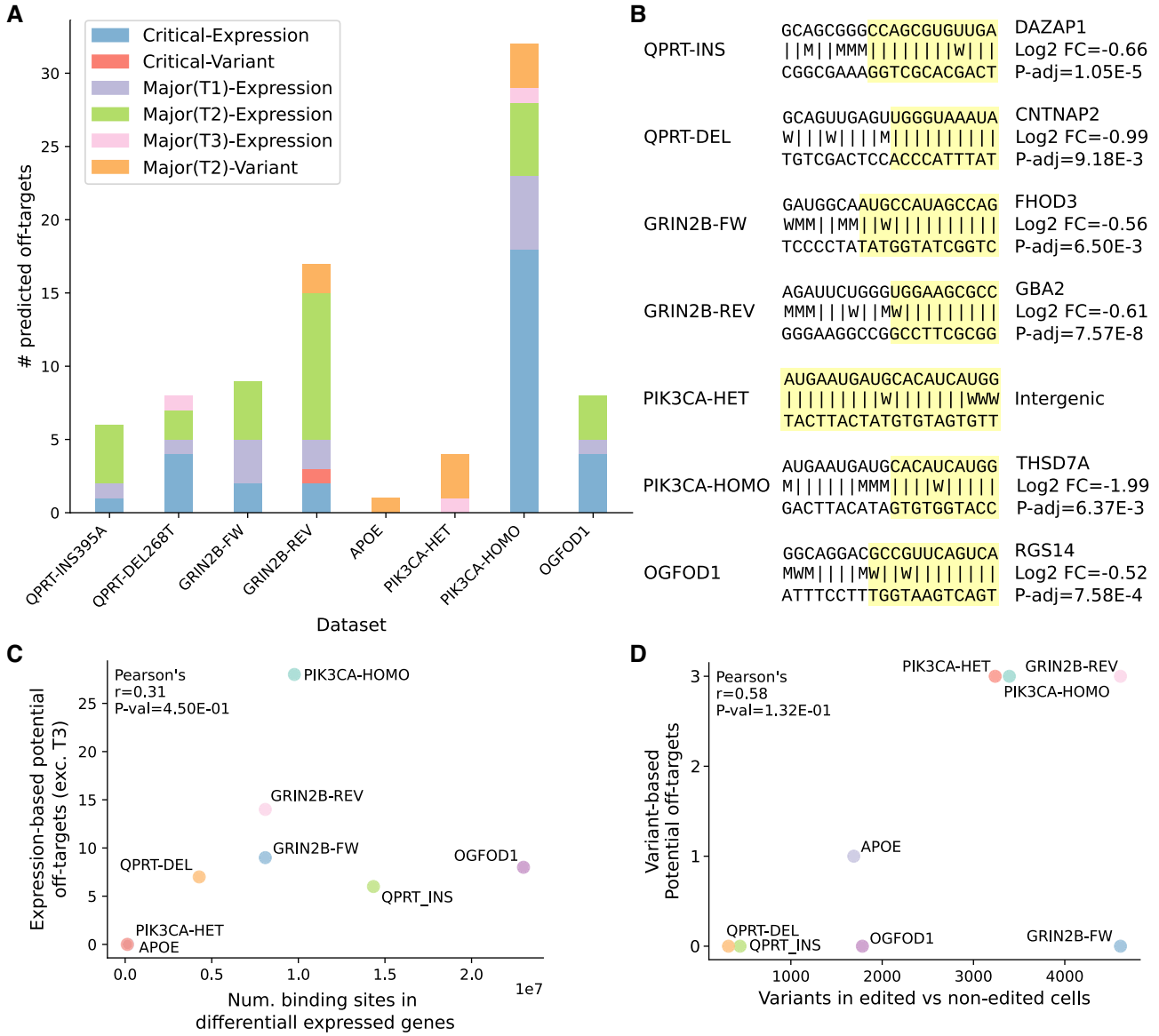


Figure 4. Predicted Cas9 off-target criticalities discovered in test datasets. (**A**) Number of predicted off-targets identified in each dataset split by degree of severity (critical or major) and by discovery method (variant or expression-based screening). Major predicted off-targets related to the expression-based screening are divided in type 1 (T1), type 2 (T2) and type 3 (T3). All major predicted off-targets related to variants are of type 2. (**B**) For each dataset the most favourable (lowest ΔG_B) predicted off-target is reported (preference is given to the critical ones with canonical NGG PAM and not overlapping repeat-masked regions). The gRNA-DNA binding pattern is represented with the following symbols: |, canonical base pair; W, wobble base pair; M, mismatch. The portion of the gRNA-DNA interaction with lowest resulting binding energy ΔG_B is highlighted in yellow, i.e. the region comprising the segment of the DNA target involved in the most energetically favourable binding interaction with the gRNA. Information on the associated downregulated gene(s) is provided (right). Log_2 FC, log_2 fold change; *P*-adj, Benjamini-Hochberg adjusted Wald test *P*-value. (**C**) Correlation between the number of Cas9 binding sites in the differentially expressed genes and the number of potential off-targets discovered by the expression-based `CRISPRroots` analysis. Major type 3 off-targets are excluded because they overlap non-expressed genes or intergenic regions. (**D**) Correlation between the number of short variants discovered from the RNA-seq in Cas9-edited vs controls cells and the number of variant-based potential off-targets.

to four mismatches and did not account for wobble base-pairs.

Here, we introduced a comprehensive method, `CRISPRroots`, to analyse on- and off-targets from RNA-seq data. `CRISPRroots` allows to (i) verify on-target edits intended to affect the transcriptome, (ii) detect off-target events directly visible in RNA-seq reads and (iii) prioritize other potential off-target events based on the evidence provided by gene expression changes.

`CRISPRroots` incorporates knowledge of called variants into a relaxed calculation of gRNA–DNA binding energies to increase off-target sensitivity. The related gRNA binding site might not be found in the `RIsearch2` search that is provided to `CRISPRoff`, for instance because of the presence of a bulge or because of the limit of up to six mismatch/wobble base pairs. Thus, we evaluate possible off-targets related to genomic variants with `RIsearch1`, that also allows for bulges and any number of mismatches.

Furthermore, CRISPRroots allows to search for off-targets either in the reference genome or, optionally, in a variant-aware genome in which short variants discovered from RNA-seq are introduced in the reference. Despite the chance of finding off-targets overlapping germline variants discovered in the transcriptome is undoubtedly limited, their potential occurrence remains a major safety concern in gene therapies.

CRISPRroots prioritizes potential off-targets that affect the transcriptome whereas predicted off-targets without expression- or variant-based support are downgraded. Despite the limitations that come from using RNA-seq for off-target ranking, the prioritized potential off-targets in CRISPRroots are the most interesting ones due to the evidence of related consequences on the transcriptome. Both the potential off-targets related to variants in the transcriptome and those on DEGs are more likely to have direct functional consequences, while those in a non-transcribed region may not result in concrete issues. A limitation of the software is that it cannot distinguish between DEGs altered by the activity of off-targets and those altered by the effects of the intended on-target edit. A downstream analysis with, e.g., the STRING (49) database of protein-protein interactions could provide the information necessary to filter potential off-targets related to DEGs functionally linked to the on-target. However, we believe this practice to be hazardous, in particular for potential off-targets with high gRNA affinity, and in contradiction to the primary goal of the software to select potential off-targets for validation prioritization.

CRISPRroots identified in all of seven test datasets potential off-target criticalities that were not addressed by the original studies. No difference was found in the total critical or major candidate off-targets after running CRISPRroots using either the reference or the alternative allele in heterozygous mutations. Potential off-targets were hidden in previous investigations because of the limited search ability of the chosen off-target prediction tools that did not account for at least one of the following elements: (i) dG·rU and dT·rG wobble base pairs, that are disguised as mismatches in similarity-based off-target searches; (ii) alternative (non-canonical) PAM sites such as NAG and NGA; (iii) high number of mismatches tolerated in gRNA-DNA binding, which is often limited in the off-target searches to 3 or 4. Additionally, some of the potential off-targets highlighted by our method were not selected for validation despite being detected as possible off-targets because of the non-identical sorting of the predictions, ruled by scores differing between off-target prediction tools. The analysis performed by CRISPRroots includes wobble base pairs and up to an arbitrary number of mismatches in a user-defined seed region and adjacent to both canonical and non-canonical PAMs. The scoring system we define is also optimized, as it is based on evidence provided by the sequencing data (variations in the sequence or expression level of genes) which is not accounted for by other off-target predictors. The underlying binding energy model employed in CRISPRroots has high off-target prediction performance (28) and is applicable to any genome.

CRISPRroots evaluates gRNA binding energies and transcriptome changes in RNA-seq data to drastically shorten the list of potential Cas9-gRNA off-target events, making their validation more feasible and impactful. In our test cases, the number of critical off-targets overlapping genes or promoters predicted by CRISPRoff has mean(±std) of 24.25(±15.31), while after the careful re-evaluation of the CRISPRoff results and the inclusion of gene expression change evidence in CRISPRroots the critical potential off-targets are reduced to 3.9(±5.91) (Supplementary Table S4). The filters based on binding patterns and energies also allow to better classify, or rule out, an important number of unfavourable interactions at potential off-target sites related to sequence variants discovered between edited and non-edited cells. Counting potential off-targets with up to 6 mismatches or bulges in the binding to the gRNA and linked to a sequence variant without including binding energy consideration leads to a mean(±std) of 15.75(±17.37) sites, while the additional binding analysis of CRISPRroots allows to highlight the 1.25(±1.50) most suitable events (Supplementary Table S4). Due to the lack of experimentally supported true positive off-targets, we cannot evaluate the false positive rate of our high-scoring candidates. Therefore we emphasize that the putative off-targets should be treated as ranked candidates for experimental validation.

An alternative strategy for off-target control was proposed by Haslinger *et al.* (7). Their study provides RNA-seq data for both a non-targeting empty control vector eCtrl and the wild-type line, and genes differentially expressed between eCtrl and wild-type are excluded from the expression-based off-target analysis presented in the study. The CRISPRroots method currently supports only comparisons between two conditions, edited and wild-type. Thus, only the eCtrl RNA-seq was used in our analysis as non-edited data. The filtering step proposed by Haslinger *et al.* is attractive, but requires the additional sequencing of non-targeting controls, which is not a common practice. Also, while some potential off-targets could be reasonably excluded based on this criterion, others are not as straightforward. For instance, in QPRT-DEL we detect a potential off-target overlapping the gene *RPH3A*, downregulated in the edited cells compared to controls. The gene was also reported to be downregulated in wild-type compared to eCtrl in the original study, but to a lower extent (l2fc = −0.95, P-adj=3.2e-4 in wild-type versus eCtrl; l2fc = −1.31, *P*-adj = 1.8e-6 QPRT-DEL knockout vs eCtrl). Excluding this potential off-target would be incautious, as the change related to QPRT-DEL knockout versus eCtrl is stronger and more significant than that recorded in the wild-type versus eCtrl. Genes harboring predicted off-targets and presenting an increase of expression were also not investigated by Haslinger *et al.* Although we agree that upregulation is a less likely off-target outcome, other types of mRNA misregulation rather than knockdown cannot be excluded (50). In CRISPRroots predicted off-targets related to upregulated genes are classified as major rather than critical, but not eliminated.

In regard to the on-target editing events, we did not observe any evident inconsistency to the reported knockout and knockin events. This is to a certain extent expected, given that all of the edited sites were verified by sequencing

in the original studies. For mapping reads originating from the knockin sequence, partial matches to the reference genome are tolerated with the default parameter settings of `STAR`. In case of knockins of exogenous genes, it is necessary to introduce the knockin sequence in the reference genome before running the pipeline. Even though Cas9 edits are commonly verified by Sanger sequencing in genome engineering experiments, the validation of such edits in RNA-seq is of relevance as it provides the expression levels of the edited alleles. This procedure also excludes possible errors, e.g. in the labeling or sequencing of samples.

In conclusion, we demonstrate our method to be very useful, as it allows for the identification of possible off-target criticalities that were not investigated in published datasets. The method is included in the first comprehensive pipeline for the analysis of RNA-seq data from CRISPR-Cas9 editing experiments, `CRISPRroots`. The pipeline can also be applied in studies involving other RNA-directed endonucleases by adjusting the configuration parameters. We believe that this tool will help saving time and resources in the analysis of genome engineered data, facilitating the advancement in this field.

DATA AVAILABILITY

The datasets analysed in this study are available in GEO with accession numbers: GSE113734 (7), GSE114685 (8), GSE102956 (9), GSE126562 (10), GSE130521 (11). These datasets were derived from the following public domain resource: https://www.ncbi.nlm.nih.gov/geo. The analyses were performed with `CRISPRroots` version 1.1. The `CRISPRroots` software is freely available via https://rth.dk/resources/crispr and on GitHub via https://github.com/RTH-tools/crisprroots. The software comes with a tutorial on how to reproduce the results presented in this article.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Christian Anthon for developing the `CRISPRroots` webpage.

Author contributions: G.I.C. implemented the on-target and off-target assessment method. G.I.C. and V.G. developed and tested the pipeline. G.I.C. drafted the manuscript. All authors edited and reviewed the manuscript. S.E.S. and J.G. supervised the study.

FUNDING

This work (including publication costs) was supported by Innovation Fund Denmark [4108-00008B and 4096-00001B to J.G.] and the Danish Research Council [9041-00317B to J.G.].

Conflict of interest statement. None declared.

REFERENCES

1. Jinek,M., Chylinski,K., Fonfara,I., Hauer,M., Doudna,J.A. and Charpentier,E. (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, **337**, 816–821.
2. Doudna,J.A. and Charpentier,E. (2014) The new frontier of genome engineering with CRISPR-Cas9. *Science*, **346**, 1258096.
3. Chuai,G., Wang,Q.-L. and Liu,Q. (2017) In silico meets in vivo : towards computational CRISPR-based sgRNA design. *Trends Biotechnol.*, **35**, 12–21.
4. Lin,S., Staahl,B.T., Alla,R.K. and Doudna,J.A. (2014) Enhanced homology-directed human genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife*, **3**, e04766.
5. Zischewski,J., Fischer,R. and Bortesi,L. (2017) Detection of on-target and off-target mutations generated by CRISPR/Cas9 and other sequence-specific nucleases. *Biotechnol. Adv.*, **35**, 95–104.
6. Zhang,H., Shi,J., Hachet,M.A., Xue,C., Bauer,R.C., Jiang,H., Li,W., Tohyama,J., Millar,J., Billheimer,J. *et al.* (2017) CRISPR/Cas9-mediated gene editing in human iPSC-derived macrophage reveals lysosomal acid lipase function in human macrophages—brief report. *Arterioscler Thrombosis Vasc. Biol.*, **37**, 2156–2160.
7. Haslinger,D., Waltes,R., Yousaf,A., Lindlar,S., Schneider,I., Lim,C.K., Tsai,M.-M., Garvalov,B.K., Acker-Palmer,A., Krezdorn,N. *et al.* (2018) Loss of the Chr16p11.2 ASD candidate gene QPRT leads to aberrant neuronal differentiation in the SH-SY5Y neuronal cell model. *Mol. Autism*, **9**, 56.
8. Bell,S., Maussion,G., Jefri,M., Peng,H., Theroux,J.-F., Silveira,H., Soubannier,V., Wu,H., Hu,P., Galat,E. *et al.* (2018) Disruption of GRIN2B Impairs Differentiation in Human Neurons. *Stem Cell Rep.*, **11**, 183–196.
9. Lin,Y.-T., Seo,J., Gao,F., Feldman,H.M., Wen,H.-L., Penney,J., Cam,H.P., Gjoneska,E., Raja,W.K., Cheng,J. *et al.* (2018) APOE4 causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types. *Neuron*, **98**, 1141–1154.
10. Madsen,R.R., Knox,R.G., Pearce,W., Lopez,S., Mahler-Araujo,B., McGranahan,N., Vanhaesebroeck,B. and Semple,R.K. (2019) Oncogenic PIK3CA promotes cellular stemness in an allele dose-dependent manner. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 8380–8389.
11. Stoehr,A., Kennedy,L., Yang,Y., Patel,S., Lin,Y., Linask,K.L., Fergusson,M., Zhu,J., Gucek,M., Zou,J. *et al.* (2019) The ribosomal prolyl-hydroxylase OGFOD1 decreases during cardiac differentiation and modulates translation and splicing. *JCI Insight*, **5**, e128496.
12. van der Wel,T., Hilhorst,R., den Dulk,H., van den Hooven,T., Prins,N.M., Wijnakker,J.A.P.M., Florea,B.I., Lenselink,E.B., van Westen,G. J.P., Ruijtenbeek,R. *et al.* (2020) Chemical genetics strategy to profile kinase target engagement reveals role of FES in neutrophil phagocytosis. *Nat. Commun.*, **11**, 3216.
13. Chandrasekaran,A., Dittlau,K.S., Corsi,G.I., Haukedal,H., Doncheva,N.T., Ramakrishna,S., Ambardar,S., Salcedo,C., Schmidt,S., Zhang,Y. *et al.* (2021) Astrocytic reactivity triggered by defective autophagy and metabolic failure causes neurotoxicity in frontotemporal dementia type 3. *Stem cell reports*, **16**, 2736–2751.
14. Lee,H. and Kim,J.-S. (2018) Unexpected CRISPR on-target effects. *Nat. Biotechnol.*, **36**, 703–704.
15. Ledford,H. (2020) CRISPR gene editing in human embryos wreaks chromosomal mayhem. *Nature*, **583**, 17–18.
16. Liang,D., Gutierrez,N.M., Chen,T., Lee,Y., Park,S.-W., Ma,H., Koski,A., Ahmed,R., Darby,H., Li,Y. *et al.* (2020) Frequent gene conversion in human embryos induced by double strand breaks. bioRxiv doi:https://doi.org/10.1101/2020.06.19.162214, 20 June 2020, preprint: not peer reviewed.
17. Zuccaro,M.V., Xu,J., Mitchell,C., Marin,D., Zimmerman,R., Rana,B., Weinstein,E., King,R.T., Palmerola,K.L., Smith,M.E. *et al.* (2020) Allele-specific chromosome removal after Cas9 cleavage in human embryos. *Cell*, **183**, 1650–1664.
18. Alanis-Lobato,G., Zohren,J., McCarthy,A., Fogarty,N.M.E., Kubikova,N., Hardman,E., Greco,M., Wells,D., Turner,J.M.A. and Niakan,K.K. (2021) Frequent loss of heterozygosity in CRISPR-Cas9–edited early human embryos. *Proc. Natl/ Acad. Sci. U.S.A.*, **118**, e2004832117.
19. Coordinators,N.R. (2017) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **46**, D8–D13.
20. Köster,J. and Rahmann,S. (2012) Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, **28**, 2520–2522.

21. Ewels,P., Magnusson,M., Lundin,S. and Käller,M. (2016) MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, **32**, 3047–3048.
22. Martin,M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.*, **17**, 10.
23. Dobin,A., Davis,C.A., Schlesinger,F., Drenkow,J., Zaleski,C., Jha,S., Batut,P., Chaisson,M. and Gingeras,T.R. (2012) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.
24. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.
25. Benjamin,D., Sato,T., Cibulskis,K., Getz,G., Stewart,C. and Lichtenstein,L. (2019) Calling somatic SNVs and indels with Mutect2. bioRxiv doi:https://doi.org/10.1101/861054, 02 December 2019, preprint: not peer reviewed.
26. der Auwera,G.A. and O'Connor BD,V. (2020) In: *Genomics in the Cloud: Using Docker, GATK, and WDL in Terra*. 1st edn., O'Reilly Media, Inc, City.
27. Tange,O. (2011) GNU parallel - the command-line power tool. *The USENIX Magazine*, **36**, 42–47.
28. 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.
29. Wenzel,A., Akbaşli,E. and Gorodkin,J. (2012) RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, **28**, 2738–2746.
30. Lorenz,R., Bernhart,S.H., zu Siederdissen,C.H., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, **6**, 26.
31. Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.
32. Liao,Y., Smyth,G.K. and Shi,W. (2013) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, **30**, 923–930.
33. Frankish,A., Diekhans,M., Ferreira,A.-M., Johnson,R., Jungreis,I., Loveland,J., Mudge,J.M., Sisu,C., Wright,J., Armstrong,J. *et al.* (2018) GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.*, **47**, D766–D773.
34. Hon,C.-C., Ramilowski,J.A., Harshbarger,J., Bertin,N., Rackham,O.J.L., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T.M., Severin,J. *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature*, **543**, 199–204.
35. Wang,L., Wang,S. and Li,W. (2012) RSeQC: quality control of RNA-seq experiments. *Bioinformatics*, **28**, 2184–2185.
36. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.*, **15**, 550.
37. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B (Methodological)*, **57**, 289–300.
38. Alkan,F., Wenzel,A., Palasca,O., Kerpedjiev,P., Rudebeck,A., Stadler,P.F., Hofacker,I.L. and Gorodkin,J. (2017) RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. *Nucleic Acids Res.*, **45**, e60.
39. Poplin,R., Ruano-Rubio,V., DePristo,M.A., Fennell,T.J., Carneiro,M.O., der Auwera,G.A.V., Kling,D.E., Gauthier,L.D., Levy-Moonshine,A., Roazen,D. *et al.* (2017) Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv: https://doi.org/10.1101/201178, 24 July 2018, preprint: not peer reviewed.
40. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M. *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
41. Danecek,P., Bonfield,J.K., Liddle,J., Marshall,J., Ohan,V., Pollard,M.O., Whitwham,A., Keane,T., McCarthy,S.A., Davies,R.M. *et al.* (2021) Twelve years of SAMtools and BCFtools. *GigaScience*, **10**, giab008.
42. Zhernakov,A., Rotter,B., Winter,P., Borisov,A., Tikhonovich,I. and Zhukov,V. (2017) Massive analysis of cDNA ends (MACE) for transcript-based marker design in pea (*Pisum sativum* L.). *Genomics Data*, **11**, 75–76.
43. Ran,F., Hsu,P., Lin,C.-Y., Gootenberg,J., Konermann,S., Trevino,A.E., Scott,D., Inoue,A., Matoba,S., Zhang,Y. *et al.* (2013) Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. *Cell*, **154**, 1380–1389.
44. Bae,S., Park,J. and Kim,J.-S. (2014) Cas-OFFinder: a fast and versatile algorithm that searches for potential off-target sites of Cas9 RNA-guided endonucleases. *Bioinformatics*, **30**, 1473–1475.
45. 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.
46. Jiang,F. and Doudna,J.A. (2017) CRISPR–Cas9 structures and mechanisms. *Ann. Rev. Biophys.*, **46**, 505–529.
47. Sherry,S.T., Ward,M.-H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
48. Li,Z., Shue,F., Zhao,N., Shinohara,M. and Bu,G. (2020) APOE2: protective mechanism and therapeutic implications for Alzheimer's disease. *Mol. Neurodegener.*, **15**, 63.
49. Szklarczyk,D., Gable,A.L., Nastou,K.C., Lyon,D., Kirsch,R., Pyysalo,S., Doncheva,N.T., Legeay,M., Fang,T., Bork,P. *et al.* (2020) The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.*, **49**, D605–D612.
50. Tuladhar,R., Yeu,Y., Piazza,J.T., Tan,Z., Clemenceau,J.R., Wu,X., Barrett,Q., Herbert,J., Mathews,D.H., Kim,J. *et al.* (2019) CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *Nat. Commun.*, **10**, 4056.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-23576-0 OPEN

Enhancing CRISPR-Cas9 gRNA efficiency prediction by data integration and deep learning

Xi Xiang[1,2,3,4,12], Giulia I. Corsi[5,12], Christian Anthon[5,12], Kunli Qu[1,6,12], Xiaoguang Pan[1], Xue Liang[1,6], Peng Han[1,6], Zhanying Dong[1], Lijun Liu[1], Jiayan Zhong[7], Tao Ma[7], Jinbao Wang[7], Xiuqing Zhang[3], Hui Jiang[7], Fengping Xu[1,3], Xin Liu[3], Xun Xu[3,8], Jian Wang[3], Huanming Yang[3,9], Lars Bolund[1,3,4], George M. Church[10], Lin Lin[1,4,11], Jan Gorodkin[5,13]✉ & Yonglun Luo[1,3,4,11,13]✉

The design of CRISPR gRNAs requires accurate on-target efficiency predictions, which demand high-quality gRNA activity data and efficient modeling. To advance, we here report on the generation of on-target gRNA activity data for 10,592 SpCas9 gRNAs. Integrating these with complementary published data, we train a deep learning model, CRISPRon, on 23,902 gRNAs. Compared to existing tools, CRISPRon exhibits significantly higher prediction performances on four test datasets not overlapping with training data used for the development of these tools. Furthermore, we present an interactive gRNA design webserver based on the CRISPRon standalone software, both available via https://rth.dk/resources/crispr/. CRISPRon advances CRISPR applications by providing more accurate gRNA efficiency predictions than the existing tools.

[1] Lars Bolund Institute of Regenerative Medicine, Qingdao-Europe Advanced Institute for Life Sciences, BGI-Qingdao, Qingdao, China. [2] BGI Education Center, University of Chinese Academy of Sciences, Shenzhen, China. [3] BGI-Shenzhen, Shenzhen, China. [4] Department of Biomedicine, Aarhus University, Aarhus, Denmark. [5] Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, Frederiksberg, Denmark. [6] Department of Biology, University of Copenhagen, Copenhagen, Denmark. [7] MGI, BGI-Shenzhen, Shenzhen, China. [8] Guangdong Provincial Key Laboratory of Genome Read and Write, BGI-Shenzhen, Shenzhen, China. [9] Guangdong Provincial Academician Workstation of BGI Synthetic Genomics, BGI-Shenzhen, Shenzhen, China. [10] Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [11] Steno Diabetes Center Aarhus, Aarhus University, Aarhus, Denmark. [12] These authors contributed equally: Xi Xiang, Giulia I. Corsi, Christian Anthon, Kunli Qu. [13] These authors jointly supervised this work: Jan Gorodkin, Yonglun Luo. ✉email: gorodkin@rth.dk; alun@biomed.au.dk

Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)-associated protein 9 (Cas9) has been successfully harnessed for programmable RNA-guided genome editing in prokaryotes, humans and many other living organisms[1–5]. A successful CRISPR gene editing application depends greatly on the selection of highly efficient gRNAs. Several machine and deep learning methods have been developed in the past decade to predict on-target gRNA activity[6–16]. However, some of these models exhibit discrepancies in the parameters selected for model validation, and in the data used for testing, which directly impact on the performances reported for such tools (Supplementary Notes 1-2). For instance, the prediction performances of the recent DeepSpCas9variants model[7] appear to be substantially higher when both canonical and noncanonical PAMs are employed for testing compared to an evaluation based solely on canonical PAMs, which are preferred for gRNA designs (Spearman's $R = 0.94$ decreases to $R = 0.70$, Supplementary Fig. 1). While the application of more advanced machine learning strategies has relatively modest impact on gRNA activity prediction performances, a significant improvement can be achieved by increasing the size and the quality of the training data (Supplementary Note 1).

Recent models trained on large-scale data still lack full saturation of their learning curve[9,14], thus leaving space for further data-driven improvement. At present, the amount of gRNA efficiency data suitable to develop machine learning models remains scarce, mostly due to the low homogeneity between studies in terms of experimental design and cleavage evaluation methodologies, which can vary from loss of function, e.g., Xu et al. (2015), Hart et al. (2015), and Doench et al. (2014–2016)[14,17–19], to indels quantification, e.g., Chari et al. (2015), Wang et al. (2019), and Kim et al. (2019–2020)[7–9,20,21]. It is thus essential to produce additional data from gRNA activity compatible with previous studies to develop more accurate prediction methods. To overcome the scarcity of experimental on-target efficiency data previous studies have employed techniques such as data augmentation, widely known in the field of image recognition, creating new input–output pairs by introducing minor alterations in the input sequence of experimentally validated gRNAs while considering their output, the efficiency, unaffected[11]. However, while two mirrored images are encoded by highly different input matrices but maintain the same original meaning, augmented gRNA data are highly redundant and do not guarantee consistency in terms of cleavage efficiency. Thus, data quantity remains the major bottleneck for improving predictors[9,14] (see also Supplementary Note 1).

Here, we show that lentiviral surrogate vectors can faithfully capture gRNA efficiencies at endogenous genomic loci. Using this approach, we generate on-target gRNA activity data for 10,592 SpCas9 gRNAs. After integrating them with complementary published data (resulting in activity data for a total of 23,902 gRNAs), we develop a deep learning prediction model, CRISPRon, which exhibits significantly higher prediction performances on independent test datasets compared to existing tools. The analysis of features governing gRNA efficiency shows that the gRNA-DNA binding energy ΔG_B is a major contributor in predicting the on-target activity of gRNAs. Furthermore, we develop an interactive gRNA design webserver based on the CRISPRon standalone software, both available via https://rth.dk/resources/crispr/. The software may also be downloaded from GitHub on https://github.com/RTH-tools/crispron/[22].

Results and discussion

Massively parallel quantification of gRNA efficiency in cells. To generate further high-quality CRISPR on-target gRNA activity data, we established a high-throughput approach to measure gRNA activity in cells (Fig. 1a) based on a barcoded gRNA oligonucleotide pool strategy as described previously[23,24]. Several optimizations of the original methods[23,24] were introduced to simplify and streamline vector cloning, lentiviral packaging and enrichment of gene edited cells (see Supplementary Note 3, Supplementary Fig. 2). To validate if the indel frequency introduced at the 37 bp surrogate target site could recapitulate that at the corresponding endogenous sites, we analyzed indel frequency at 16 surrogate sites and their corresponding endogenous genomic loci in HEK293T cells by deep sequencing. We obtained a fine correlation between the surrogate and endogenous sites in terms of indel frequencies and profiles (Supplementary Fig. 3, Spearman's $R = 0.72$, p-value $= 0.0016$), in agreement with previous findings[8,9,23,24].

We next generated a large dataset of high-quality CRISPR gRNA activity data in cells using this optimized approach. A pool of 12,000 gRNA oligos, targeting 3834 human protein-coding genes (Supplementary Data 1, Supplementary Note 4), were array-synthesized and selected to avoid large overlap with existing datasets. Targeted amplicon sequencing (depth > 1000) of the surrogate oligo pool, surrogate gRNA plasmid library and transduced wild-type HEK293T cells (multiplexity of infection (MOI) of 0.3) revealed that over 99% of the designed gRNAs were present in the 12 K gRNA plasmid pool and transduced cells (Supplementary Figs. 4-5, source data). We transduced the SpCas9-expressing and wild-type HEK293T cells with the gRNA library with a MOI of 0.3 and a transduction coverage of ~4000 cells per gRNA. A pipeline was established to analyze CRISPR-induced indels and remove sequence variants introduced by oligo-synthesis, PCRs, and deep sequencing, as well as low quantity sites (less than 200 reads, see Methods). Indel frequencies in the cells 2, 8, and 10 days after transduction were analyzed by targeted deep sequencing (Supplementary Fig. 6). Following increased editing time and enrichment of edited cells (puromycin selection), indel frequency rose significantly in cells from day 2 to day 8–10 (Fig. 1b). Overexpression of SpCas9 by doxycycline (Dox) addition leads to a skewed distribution of gRNA efficiency (Supplementary Fig. 7, Supplementary Note 4), thus gRNA efficiencies from Dox-treated SpCas9 cells were excluded for gRNA efficiency prediction model establishment. The indel frequency (on-target activity) of gRNAs from day 8 and 10 were well correlated (Fig. 1c, Spearman's $R = 0.91$). Corroborating previous findings, the indel types introduced by SpCas9 comprise mainly small deletions and 1 bp insertion (Fig. 1d, Supplementary Figs. 7-8) and compared to day 2 the indel types from day 8–10 are better correlated with the indel profiles predicted by inDelphi[24] (Fig. 1e, Supplementary Fig. 7-8, Supplementary Note 5), a machine learning algorithm for predicting CRISPR-induced indels. Our data also revealed that the inserted nucleotide of the most frequent indel type (1 bp insertion) is most frequently the same as N17 nucleotide of the protospacer (4 bp upstream of the PAM) (Fig. 1f, Supplementary Fig. 7, Supplementary Note 5). The average gRNA activity from day 8 and 10 was used for subsequent analyses and model establishment. As a result, we obtained high-quality gRNA activity data for 10,592 gRNAs, of which 10,313 gRNAs are unique for this study (Supplementary Fig. 9, Supplementary Data 1). To independently validate the CRISPR gRNA activity captured by the lentiviral surrogate vector library, we compared gRNA efficiencies commonly measured in our study to those of Kim et al. (2019) and Wang et al. (2019)[8,9] (Fig. 1g). We observed a good correlation (Spearman's $R = 0.67$ to both) between gRNA activities measured by our study and others, higher compared to the agreement between these two existing protocols (Spearman's $R = 0.52$). Our gRNA efficiency data match characteristics of

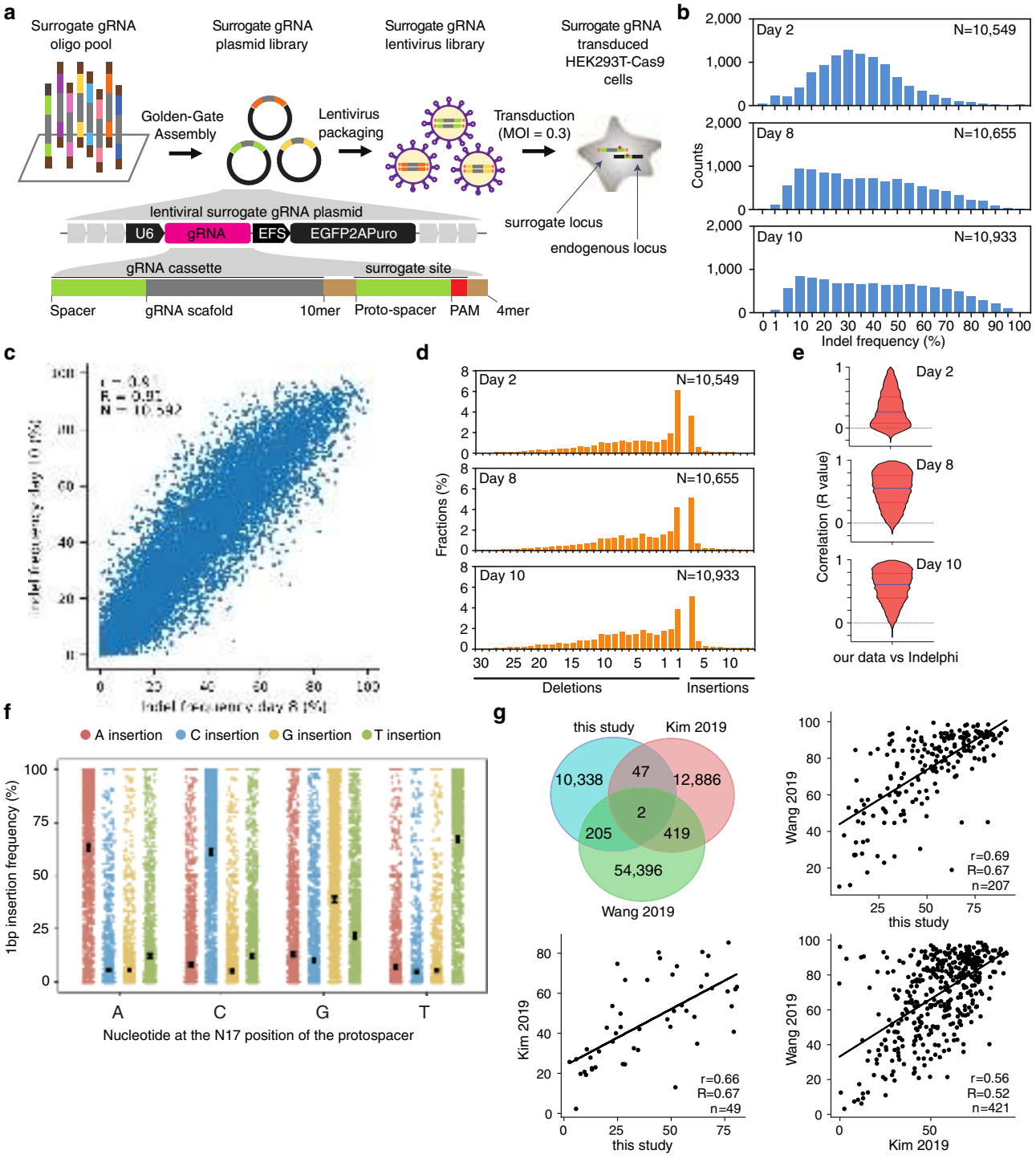


Fig. 1 High-throughput quantification of gRNA efficiency in cells. a Schematic illustration of the lentiviral surrogate vector, oligo pool synthesis, PCR amplification, golden-gate assembly, lentivirus packaging, and transduction. **b** gRNA editing efficiency of all surrogate sites measured by targeted amplicon sequencing. Results are shown for HEK293T-SpCas9 cells at 2, 8, and 10 days after transduction. **c** Correlation between gRNA editing efficiency at 8 and 10 days after transduction. **d** Indel profiles (1–30 bp deletion, 1–10 bp insertion) for all surrogate sites introduced by SpCas9 in HEK293T-SpCas9 cells at 2, 8, and 10 days post transduction. **e** Correlation between the indel profiles measured in cells and those predicted by inDelphi. Data are presented as violin plot with median and quartiles. **f** Dot plot of 1-bp insertion indel frequency (mean ± 95% confidence interval), stratified by the nucleotide at N17 position of the protospacer and the type of nucleotide inserted (see also Supplementary Fig. 7). **g** Correlation between gRNA editing efficiencies measured in this and other major studies for common gRNA + PAM (23 nt) examples, also displayed in a Venn diagram.

previous findings, with a preferential range of GC content between 40 and 90%[25] and stable gRNA structures being unfavorable, in particular for minimum folding energies (MFE) < −7.5 kcal/mol[26] (Supplementary Fig. 10). We conclude that the high-quality gRNA activity dataset of 10,592 gRNAs measured in cells by our study represents a valuable source to further improve the quality of CRISPR-gRNA designs.

Enhanced gRNA efficiency prediction. We developed a deep learning model, which combines sequence and thermodynamic

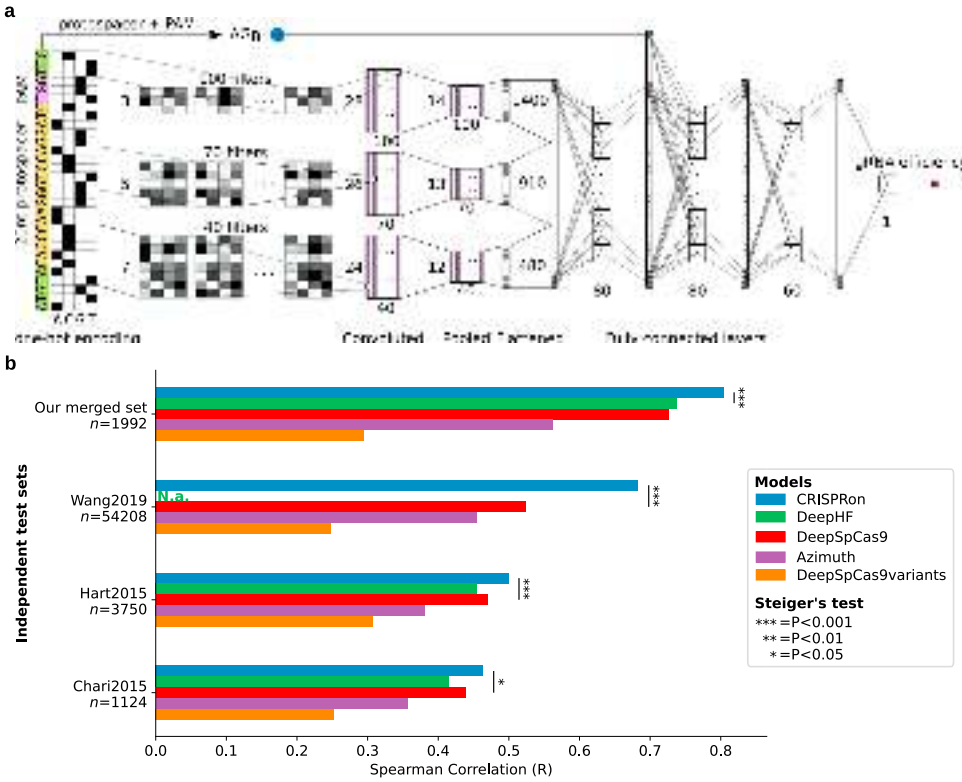


Fig. 2 The CRISPRon model and generalization ability on independent test sets. a Schematic representation of the CRISPRon input DNA sequence and prediction algorithm. The inputs to the deep learning network are the one-hot encoded 30mer and the binding energy (ΔG_B). Note that only the filtering (convolutional) layers and the 3 fully connected layers are shown explicitly and that the thin vertical bars are the output of one layer, which serves as input for the next layer. **b** Performance comparison between CRISPRon and other existing models on independent test sets larger than 1000 gRNAs. N.a. not available (all gRNAs were regarded as training data due to lack of explicit train-test separation). CRISPRon_v0 was employed for testing on the internal independent test set ("Our merged set", including gRNAs from both our study and Kim et al. (2019)). CRISPRon_v1, or simply CRISPRon, was used for the external independent test sets (for a description of the CRISPRon versions, see Supplementary Table 1). The two-sided Steiger's test *P*-values of all comparisons are reported in Supplementary Data 2.

properties automatically extracted out of a 30 nt DNA input sequence constituted of the protospacer, the PAM and neighboring sequences for precise gRNAs activity predictions (Fig. 2a). In addition to the sequence composition, the model embeds the gRNA-target-DNA binding energy ΔG_B, described by the energy model used in CRISPRoff[27], which encapsulates the gRNA-DNA hybridization free energy, and the DNA-DNA opening and RNA unfolding free energy penalties. ΔG_B was observed to be a key feature for predicting on-target gRNA efficiency (see Supplementary Note 6 and feature analysis below). We first trained deep learning models solely on our dataset (Supplementary Table 1) and compared their predictions with those of existing tools on both internal and external independent test datasets. To do that, our CRISPR gRNA activity data were carefully partitioned into six subsets ensuring clustering of the closest gRNA sequences within the same partition (see Methods). The first model, pre-CRISPRon_v0, was trained with a 5-fold cross-validation while using a 6th partition as an internal independent test set solely for measuring the performance. The pre-CRISPRon_v0 and DeepSpCas9 models displayed remarkable and comparable generalization ability when tested on data from the study of one another (Spearman's $R > 0.70$ for both), confirming our data and Kim et al. (2019) data as highly compatible (Supplementary Data 2). The second model (pre-CRISPRon_v1, see Supplementary Table 1 and Supplementary Data 2) was constructed to evaluate on external independent test sets by training on all six partitions with a 6-fold cross-validation. This model displayed performances similar to those of existing tools.

Since pre-CRISPRon_v0 and DeepSpCas9 held comparable performances when trained on their respective datasets, we fused our data with that of Kim et al. (2019) using a linear rescaling based on the 30mer sequences found in both datasets, resulting in a dataset of 23,902 gRNAs (30mer, Supplementary Fig. 9). We did not fuse with the datasets measuring efficiency as indel frequency of Wang et al. (2019) and Kim et al. (2020), because of their scarce coverage of the general gRNA activity landscape (Supplementary Note 2 and 7). After dividing the joint dataset of our study and Kim et al. (2019) into six partitions as explained above, we first developed CRISPRon_v0 with a 5-fold cross-validation to evaluate the model on the internal independent test set. The CRISPRon_v0 increased the performance over pre-CRISPRon_v0 on the Kim et al. (2019) dataset, while only a minor loss (<0.025 in Spearman's R) was observed on our data (Supplementary Table 1 and Supplementary Data 2). On the internal independent test set, CRISPRon_v0 exceeded the performance (Spearman's $R = 0.80$) of notable predictors, such as Azimuth ($R = 0.56$), DeepSpCas9 ($R = 0.73$), DeepHF ($R = 0.74$), and DeepSpCas9variants (Fig. 2b, Supplementary Fig. 11). A final model, CRISPRon_v1 (hereafter called CRISPRon, see Supplementary Table 1 and Supplementary Data 2), was then trained on the full combined dataset with a 6-fold cross-validation. External independent test sets with more than 1000 gRNAs (Fig. 2b) were employed for testing while again ensuring no overlap between what the respective models were trained on (see Methods). On these external independent test sets CRISPRon achieved the highest prediction performance ($R \approx [0.46, 0.68]$)

compared to Azimuth (R ≈ [0.36, 0.45]), DeepSpCas9variants (R ≈ [0.25, 0.31]), and to the so far top-performing models DeepSpCas9 (R ≈ [0.44, 0.52]) and DeepHF (R ≈ [0.42, 0.46]). Additional performance evaluations on datasets with less than 1000 gRNAs confirmed CRISPRon as top-performing model (Supplementary Fig. 11, Supplementary Data 2). A web interface for gRNA on-target efficiency predictions with the CRISPRon model is made available via https://rth.dk/resources/crispr/. The webserver interface utilizes the IGV javascript plugin available from github[28].

Features important for predicting gRNA efficiency. To characterize the gRNA features with the highest impact on gRNA efficiency predictions we trained a gradient boosting regression tree (GBRT) model based on the combined data from our study and Kim et al. (2019) and applied two methods for feature analysis: the Shapley Additive exPlanations (SHAP)[29] and the Gini importance[30] (Supplementary Fig. 12, full details in Supplementary Note 6). Both methods highlight that thermodynamic properties, above all ΔG_B, give a considerable contribution to the learning process. The most notable sequence-composition features include the two nucleotides proximal to the PAM, where G and A are favored over C and T and the presence of the dinucleotide TT, which relates with weak binding free energies and is unfavorable.

Limitations to the study. A few limitations of using the lentiviral surrogate vectors to capture CRISPR gRNA efficiency are highlighted for the need of future improvements. The DSBs generated by CRISPR-Cas9 are predominantly repaired by the non-homologous end joining (NHEJ) and microhomology-mediated end joining (MMEJ) pathways, which leads to the introduction of small indels at the DSB site. However, large deletions or chromosomal rearrangements have also been reported in CRISPR editing as outcomes of repaired mediated by e.g., homology-directed repair (HDR) or single-strand annealing (SSA) in cells[31,32]. The gRNA efficiency quantification approach in this study is based on a 37 bp surrogate target site. Thus, SpCas9 editing outcomes such as large deletions or chromosomal rearrangements are not captured by our method. Earlier, we have discovered that chromatin accessibility at the editing sites affects CRISPR gene editing efficiency[26]. Since the 12 K lentivirus library was randomly inserted in the genome of the targeted cells, the chromatin accessibility state of the surrogate site might be different from the endogenous target site.

Concluding remark. In summary, we report on the generation of on-target gRNA activity data for 10,592 SpCas9 gRNAs and the development of a deep learning model, CRISPRon, which exhibits more accurate gRNA efficiency predictions than other existing tools.

Methods

DNA vectors. The 3rd generation lentiviral vector backbone was generated by synthesis (Gene Universal Inc) and cloning. The human codon-optimized SpCas9 expression vector was based on a PiggyBac transposon vector, carrying a hygromycin selection cassette. All DNA vectors have been Sanger sequenced and can be acquired from the corresponding author YL's lab. The lentiviral vector generated by this study for cloning surrogate oligos has been made available through Addgene (plasmid # 170459). A detail protocol is also made available at the shared protocols platform[33].

Design of the 12 K surrogate oligo pool. Each oligo consists of the BsmBI recognition site "cgtctc" with 4 bp specific nucleotides "acca" upstream, following the GGA cloning linker "aCACC", one bp "g" for initiating transcription from U6 promoter, 20 bp gRNA sequences of "gN20", 82 bp gRNA scaffold sequence, 37 bp surrogate target sequences (10 bp upstream sequences, 20 bp protospacer and 3 bp PAM sequences, 4 bp downstream sequence), the downstream linker "GTTTg", and another BsmBI-binding site and its downstream flanking sequences "acgg".

For the 12 K oligo pool was designed as below: (1) Select ~7000 genes from the drugable gene database (http://dgidb.org)[34]; (2) Discard all the exons which the DNA length is less than 23 bp with filtering; (3) Select the first three coding exons of each gene. If the exons number is less than 3, retain all the exons; (4) Extract all the possible gRNA sequences (including the PAM sequence "NGG") in the filtered exons sequence; (5) Look up off-target sites of each gRNA with FlashFry (v 1.80)[35] and discard gRNAs with potential off-target of 0–3 bp mismatches in human genome hg19 and rank each gRNA based on the number of off-target site in an ascending order; (6) Map and extract the 10 bp upstream and 4 bp downstream flanking sequence of each selected gRNA, construct the surrogate target sequence as 10 bp upstream + 23 bp gRNA (include PAM) + 4 bp downstream = 37 bp; (7) Filter out surrogate sites with BsmBI recognition site, because of GGA cloning; (8) Compare all the selected gRNAs with the database of CRISPR-iSTOP[36]; (9) Construct the full-length sequence of each synthetic oligo, which is 170 bp; In total, the 12 K oligos target 3832 genes. The 12 K oligo pools were synthesized in Genscript® (Nanjing, China), and sequences are given in Supplementary Data 1.

12 K surrogate plasmid library preparation. First, the 12 K oligos were cleaved and harvested from the microarray and diluted to 1 ng/µl. Next, we performed surrogate PCR1 (Supplementary Data 1). The PCR reaction was carried out using PrimeSTAR HS DNA Polymerase (Takara, Japan) following the manufacturer's instruction. Briefly, each PCR reaction contained 1 µl oligo template, 0.2 µl PrimeSTAR polymerase, 1.6 µl dNTP mixture, 4 µl PrimeSTAR buffer, 1 µl forward primer (10 uM), and 1 µl reverse primer (10 uM) and ddH2O to a final volume of 20 µl.

The thermocycle program was 98 °C 2 min, (98 °C/10 s, 55 °C/10 s, 72 °C/30 s) with 21 cycles, then 72 °C for 7 min and 4 °C hold. To avoid amplification bias of oligos introduced by PCR, we conducted gradient thermocycles and performed PCR products gray-intensity analysis to determine the optimal PCR cycles of 21. The best thermocycles should be in the middle of an amplification curve. In this study, the PCR cycle as 21 for oligos amplification. Instead, for PCR amplification of surrogate sites from cells integrated with lentivirus, the PCR cycle was 25. The final PCR product length was 184 bp. We performed 72 parallel PCR reactions for 12 K oligos amplification, then these PCR products were pooled, and gel purified by 2% agarose gel. One microgram purified PCR product were quantified with PCR-free next generation sequencing (MGI Tech).

The PCR products of 12 K oligos were then used for Golden Gate Assembly (GGA) to generate the 12 K plasmids library. For each GGA reaction, the reaction mixture contained 100 ng lentiviral backbone vector, 10 ng purified 12 K oligos-PCR products, 1 µl T4 ligase (NEB), 2 µl T4 ligase buffer (NEB), 1 µl BsmBI restriction enzyme (ThermoFisher Scientific, FastDigestion) and ddH2O to a final volume of 20 µl. The GGA reactions were performed at 37 °C 5 min and 22 °C 10 min for 10 cycles, then 37 °C 30 min and 75 °C 15 min. Thirty six parallel GGA reactions were performed and the ligation products were pooled into one tube.

Transformation was then carried out using chemically competent DH5a cells. For each reaction, 10 µl GGA ligation product was transformed in to 50 µl competent cells and all the transformed cells were spread on one LB plate (15 cm dish in diameter) with Xgal, IPTG and Amp selection. High ligation efficiency was determined by the presence of very few blue colonies (also see Supplementary Fig. 2). To ensure that there is sufficient coverage of each gRNA of the 12 K library, 42 parallel transformations were performed, and all the bacterial colonies were scraped off and pooled together for plasmids midi-prep. For NGS-based quality quantification of the library coverage, midi-prep plasmids were used as DNA templates for surrogate PCR2, followed by gel purification and NGS sequencing. The PCR primers for surrogate PCR2 are showed in Supplementary Data 1.

12 K lentivirus packaging. HEK293T cells were used for lentivirus packaging. All cells were cultured in Dulbecco's modified Eagle's medium (DMEM) (LONZA) supplemented with 10 % fetal bovine serum (FBS) (Gibco), 1% GlutaMAX (Gibco), and penicillin/streptomycin (100 units penicillin and 0.1 mg streptomycin/mL) in a 37 °C incubator with 5% CO_2 atmosphere and maximum humidity. Cells were passaged every 2–3 days when the confluence was ~80–90%.

For lentivirus packaging: (Day 1) Wild-type HEK293T cells were seeded to a 10 cm culture dish, 4×10^6 cells per dish (10 dishes in total); (Day 2) Transfection. Briefly, we refreshed the medium with 7 mL fresh culture medium to 1 h before transfection (gently, as the HEK293T cells are easy to be detached from the bottom of dish); Next, we performed transfection with the PEI 40000 transfection method. For 10 cm dish transfection, the DNA/PEI mixture contains 13 µg lentiviral 12 K plasmid DNA, 3 µg pRSV-REV, 3.75 µg pMD.2 G, 13 µg pMDGP-Lg/p-RRE, 100 µg PEI 40000 solution (1 µg/µl in sterilized ddH2O), and supplemented by serum-free optiMEM without phenol red (Invitrogen) to a final volume of 1 mL. The transfection mixture was pipetted up and down several times gently, then kept at room temperature (RT) for 20 min, then added into cells in a dropwise manner and mix by swirling gently. (Day 3) Changed to fresh medium; (Day 4) Harvest and filter all the culture medium of the 10 cm dish through a 0.45 µm filter, pool the filtered media into one bottle. Each 10 cm dish generated ~7–8 mL lentivirus crude. Add polybrene solution (Sigma–Aldrich) into the crude virus to a final

concentration of 8 μg/mL. Aliquot the crude virus into 15 mL tubes (5 mL/tube) and store in −80 °C freezer.

Lentivirus titer quantification by flow cytometry (FCM). As the 12 K lentiviral vector expresses an EGFP gene, the functional titer of our lentivirus prep was assayed by FCM. Briefly, (1) split and seed HEK293T cells to 24-well plate on day 1, 5×10^4 cells per well. Generally, 18 wells were used to perform the titter detection, a gradient volume of the crude lentivirus was added into the cells and each volume was tested by replicates. In this experiment, the crude virus gradients were 10, 20, 40, 80, and 160 μl for each well (Supplementary Fig. 5). Another two wells of cells were used for cell counting before transduction; (2) Conduct lentivirus transduction when cells reach up to 60–80% confluence on day 2. Before transduction, detach the last two wells of cells using 0.05% EDTA-Trypsin to determine the total number of cells in one well ($N_{initial}$). Then change the remaining wells with fresh culture medium containing 8 μg/mL polybrene, then add the gradient volume of crude virus into each well and swirling gently to mix; (3) On day 3, change to fresh medium without polybrene; (4) On day 4, harvest all the cells and wash them twice in PBS. Fix the cells in 4% formalin solution at RT for 20 min, then spin down the cell pellet at 500 × *g* for 5 min. Discard the supernatant and re-suspend the cell pellet carefully in 600 μl PBS, and conduct FCM analysis immediately. FCM was performed using a BD LSRFortessa™ cell analyzer with at least 30,000 events collected for each sample in replicates.

The FCM output data was analyzed by the software Flowjo vX.0.7. Percentage of GFP-positive cells was calculated as: $Y\% = N_{GFP\text{-}positive\ cells}/N_{total\ cells} \times 100\%$. Calculate the GFP percentage of all samples. For accurate titer determination, there should be a linear relationship between the GFP-positive percentages and crude volume. The titer (Transducing Units (TU/mL) calculation according to this formula: $TU/mL = (N_{initial} \times Y\% \times 1000)/V$. V represents the crude volume (μl) used for initial transduction.

Generation of SpCas9-expressing stable cell lines. SpCas9-expressing HEK293T (HEK293T-SpCas9) cells were generated by a PiggyBac transposon system. HEK293T cells were transfected with pPB-TRE-spCas9-Hygromycin vector and pCMV-hybase with a 9:1 ratio. Briefly, 1×10^5 HEK293T cells were seeded in 24-well plate and transfections were conducted 24 h later using lipofectamine 2000 reagent following the manufacturer's instruction. Briefly, 450 ng pPB-TRE-spCas9-Hygromycin vectors and 50 ng pCMV-hybase were mixed in 25 μl optiMEM (tube A), then 1.5 μl lipofectamine 2000 reagent was added in another 25 μl optiMEM and mix gently (tube B). Incubate tube A and B at RT for 5 min, then add solution A into B gently and allow the mixture incubating at RT for 15 min. Add the AB mixture into cells evenly in a dropwise manner. Cells transfected with pUC19 were acted as negative control. Culture medium was changed to selection medium with 50 μg/ml hygromycin 48 h after transfection. Completion of selection took ~5–7 days until the negative cells were all dead in the untransfected cells. The cells were allowed to grow in 50 μg/ml hygromycin growth medium for 3–5 days for further expansion. PCR-based genotyping was carried out to validate the integration of Cas9 expression cassette (Supplementary Data 1). Although the expression of SpCas9 was controlled by a TRE promoter, we observed significant editing efficiency in cells without addition of doxycycline. Thus, the cells were used as a normal SpCas9-expressing model, while SpCas9 overexpression can be induced by Dox induction.

12 K lentivirus library transduction. HEK293T-SpCas9 cells were cultured in growth medium with 50 μg/ml hygromycin throughout the whole experiment. For 12 K lentivirus library transduction, (1) on Day −1: 2.5×10^6 cells per 10 cm dish were seeded (in 12 dishes). For each group, one dish was used for cell number determination before transduction and one dish for drug-resistance (puromycin) test control and the remaining 10 dishes were used for the 12 K lentivirus library transduction (transduction coverage per gRNA exceeds 4000×); (2) Day 0: We first determined the approximate cell number per dish. This was used to determine the volume of crude lentivirus used for transduction using a multiplicity of infection (MOI) of 0.3. The low MOI (0.3) ensures that most infected cells receive only 1 copy of the lentivirus construct with high probability [41]. The calculation formula is: V = N × 0.3/TU. V = volume of crude lentivirus used for infection (ml); N = cell number in the dish before infection; TU = the titter of crude lentivirus (IFU/mL). The infected cells were cultured in a 37 °C incubator; (3) Day 1: 24 h after transduction, split the transduced cells of each dish to three dishes equally; (4) Day 2: For the three dishes of split (30 dishes in total, three divided into sub-groups), subgroup 1 (10 dishes) were harvested and labeled as the Day 2 after the 12 K lentivirus library transduction. All cells from this subgroup were pooled into one tube and stored in −20 °C freezer for genomic DNA extraction; The subgroup 2 (10 dishes) was changed to fresh D10 medium contains 50 μg/ml hygromycin + 1 μg/mL puromycin (Dox-free group); The subgroup 3 (10 dishes) was changed to D10 medium contains 50 μg/ml hygromycin + 1 μg/mL puromycin + 1 μg/mL doxycycline (Dox-addition group). (5) The transduced cells were spitted every 2–3 days when cell confluence reaches up to 90%. Cells from Day 2, 8, and 10 were harvested and stored in −20 °C for further genomic DNA extraction. Parallel experiments were performed using wild-type HEK293T cells.

PCR amplification of surrogate sites from cells. Genomic DNA was extracted using the phenol-chloroform method. The genomic DNA were digested with RNase A (OMEGA) to remove RNA contamination (In this study, 50 μg RNase A worked well to digest the RNA contamination in 100–200 μg genomic DNA after incubating in 37 °C for 30 min). Then the genomic DNA was purified and subjected to surrogate PCR2 (Supplementary Data 1). In this study, 5 ug genomic DNA was used as temperate in one PCR reaction, which contained ~7.6×10^5 copies of surrogate construct (assuming 1×10^6 cells contain 6.6 μg genomic DNA), which covered about 63 times coverage of the 12 K library. In total, 32 parallel PCR reactions were performed to achieve approximately 2016 times coverage of each gRNA and surrogate site. For each PCR reaction, briefly, 50 μl PCR reaction system consists of 5 μg genomic DNA, 0.5 μg PrimeSTAR polymerase, 4 μl dNTP mixture, 10 μl PrimeSTAR buffer, 2.5 μl forward primer (10 uM), and 2.5 μl reverse primer (10 uM) and supplemented with ddH2O to a final volume of 50 μl. The thermocycle program was 98 °C 2 min, (98 °C for 10 s, 55 °C for 10 s, 72 °C for 30 s) with 25 cycles, then 72 °C for 7 min and 4 °C hold. Then purify all the PCR products by 2% gel, pool the products together and conduct deep amplicon sequencing.

Deep amplicon sequencing. MGISEQ-2000 (DNBseq-G400) was used to perform the amplicons deep sequencing following the standard operation protocol. First, PCR-free library was prepared using MGIeasy FS PCR-free DNA library Prep kit following the manufacturer's instruction. Briefly, measure the concentration of purified PCR products using Qubit 4 ™ fluorometer (Invitrogen) and dilute the concentration of each sample to 10 ng/μl. Ten microliters diluted PCR product was mixed with an A-Tailing reaction which contained A-Tailing enzyme and buffer, incubated at 37 °C for 30 minutes then 65 °C for 15 min to inactive the enzyme. Then the A-Tailed sample was mixed with PCR Free index adapters (MGI.), T4 DNA Ligase and T4 ligase buffer to add index adapter at both 3' and 5' ends of PCR products. The reaction was incubated at 23 °C for 30 min and then purified with XP beads. Then denature the PCR products to be single-strand DNA (ssDNA) by incubating at 95 °C for 3 min and keep on 4 °C for the subsequent step. Transform the ssDNA to be circles using cyclase (MGI) at 37 °C for 30 min and then digested to remove linear DNA using Exo enzyme at 37 °C for 30 min. Purify the products again by XP beads and assay the concentration of library by Qubit 4 ™ fluorometer. The amplicons libraries were subjected to deep sequencing on the MGISEQ-2000 platform. In this study, for each lane four samples (6 ng each) were pooled together for deep sequencing. To avoid sequencing bias induced by base unbalance of surrogate PCR products, 12 ng whole-genome DNA library (balance library) was mixed with the four PCR samples in a final concentration of 1.5 ng/μl and sequenced in one lane. All the samples were subjected to pair-ended 150 bp deep sequencing on MGISEQ-2000 platform.

Data analysis. In order to evaluate the sequencing quality of amplicons and filter the low-quality sequencing data, Fastqc-0.11.3 and fastp-0.19.6[37] were used with default parameters for each sample. The clean sequencing reads of pair-ended segments were merged using FLASh-1.2.1[38] to obtain full-length reads. In order to obtain the amplified fragment reads of each surrogate reference sequence, BsmBI Linker was removed from the surrogate reference sequence. The BWA-MEM algorithm[39] of bwa was used for local alignment, and the reads of all samples were divided into 12,000 independent libraries. Due to the existence of sequencing or oligo-synthesis introduced errors, each library was then filtered. As SpCas9 mainly causes insertions and deletions, the length of surrogate sequence is expected to change from its original 37 bp. We adopt the following steps for data processing and filtering: (1) Obtain the sequence containing gRNA + scaffold fragment as dataset1. (2) Obtain the sequence containing GTTTGAAT in dataset1 as dataset2 (BsmBI linker fragments changed in orientation (GTTTGGAG− > GTTTGAAT)). (3) Extract the intermediate surrogate sequence from dataset2, which removed the length limit. In order to eliminate the interference of background noise before analyzing editing efficiency, all mutations or indels found in WT HEK293T cells group were removed.

The total editing efficiency for each gRNA was calculated according to the following formula:

$$\text{Total editing efficiency} = \frac{(\text{Num. reads with length} \neq 37\ \text{bp})}{(\text{Tot. num. of reads})}\% \quad (1)$$

The average fraction of indels from 30 bp deletion to 10 bp insertion was calculated according to the following formula:

$$\text{Average indels fraction} = \frac{(\text{Num. reads with length range}[7, 47]\ \text{bp})}{(\text{Tot. num. of reads of 12K library})}\% \quad (2)$$

Data collection and preprocessing for machine learning. The 12 K dataset was preprocessed by removing gRNAs supported by less than 200 reads and by intersecting the datasets of gRNAs with efficiencies measured at day 8 ($N = 10{,}655$) and day 10 ($N = 10{,}933$), thus retaining data for 10,592 gRNAs. For training, efficiencies measured at day 8 and day 10, positively correlated (Pearson's $r = 0.91$), were averaged. The following additional datasets were downloaded: Kim (2019–2020)[7,9]; Wang (2019)[8]; Xu (2015)[17,21], Chari (2015, 293 T cells)[20]; and Hart (2015) Hct1162lib1Avg[18] as collected by Haeussler et al.[40]; Doench (2014–2016) from the public repository of the Azimuth project[14,19]. For the dataset

by Doench et al. (2014) only data from human cells was used, while for the later dataset (2016) we filtered for the genes CCDC101, CUL3, HPRT1, MED12, NF1, NF2, TADA1, TADA2B, as previously recommended[14,40], and excluded gRNAs marked for low early time point (ETP). The Wang (2019) dataset was filtered from gRNAs for which no context was defined in the corresponding study[8]. Based on the method used to evaluate gRNA activity, datasets were distinguished into two categories: loss of gene function studies, which comprises Xu (2015), Hart (2015), and Doench (2014–2016) and indel-based, including Kim (2019–2020), Wang (2019), Chari (2015) and this study.

The datasets were preprocessed by removing gRNAs matching one of the following criteria: (1) Not present in hg38 (except for exogenous constructs); (2) No match to the target gene based on GENCODE annotations (v 32); (3) High variance in efficiency between different experimental settings, above the threshold: upper quartile + 1.5× variance interquartile range; (4) Target gene with less than 10 designed gRNAs; (5) Related to a PAM different from 5'-NGG-3' (6) Expressed from a tRNA system; (7) Targeting the last 10% of the merged coding sequences (CDSs) annotated for a target gene (nonsense mediated decay or polymorphic pseudogene transcripts were excluded). Points 2, 4, and 7 were applicable only in the case of loss of function studies. The Kim (2019–2020) datasets were further processed by averaging duplicated 30mer gRNA + context entries (avg. difference between max. and min. indel frequency of replicates = 8.6 and 6.7 in the studies of 2019 and 2020, respectively). Efficiency values not reported as indel frequencies were ranked-normalized with the SciPy rankdata function[41] and normalized efficiencies were averaged between experimental conditions.

After preprocessing, each dataset contained the following number of unique 30mer, gRNA + context sequences: Kim (2019): 13,359; Kim (2020): 8742; Xu (2015): 971; Chari (2015): 1,224; Hart (2015): 4001; Doench (2014): 781; Doench (2016): 2145; Wang (2019): 55,022; this study: 10,592. See Supplementary Table 2 for more details about filtered data. Ours and Kim (2019) datasets were combined by building a linear regression model on overlapping elements (49 pairs) and applying it to scale gRNA efficiencies from our study to those of Kim et al. (2019). Efficiencies were averaged for overlapping 30mers. The merged dataset consisted of 23,902 gRNA + context sequences (30mers).

Generation of gRNA and target DNA features. Features were extracted from a 30mer DNA sequence composed by the target DNA protospacer (20 nt) and the following flanking regions: 4 nt upstream, 3 nt PAM, and 3 nt downstream from the PAM. Position-specific single and di-nucleotides were one-hot encoded, binarizing the presence/absence of a certain nucleotide with the values 0 (absent) or 1 (present). They were denoted as N_X, with N in the set [A,T,G,C] and X being the position on the 30mer. Nucleotides surrounding the "GG" Cas9 binding site were also binarized and denoted as NGGX_YZ, where Y and Z are the nucleotides upstream and downstream from the motif. Sliding windows of 1 and 2 nt were used to count the occurrences of each single and dinucleotide in the 30mer sequences. These position-independent features were labeled by the nucleotide or dinucleotide they account for. The GC content was obtained as the sum of Gs and Cs in the protospacer sequence. The melting temperatures were computed with the Biopython 1.77 Tm_staluc method[42] for three nonoverlapping segments of the protospacer, at positions 3–7, 8–15, and 16–20, referred to as MT_[S,E], where S and E are the start and end positions of the segment. The spacer folding free energy of ensemble and the ΔG_B RNA–DNA binding energy were computed using the energy function in the CRISPRoff pipeline 1.1.1[27], provided with RNAfold 2.2.5[42].

Generation of dataset partitions. The datasets used for training were divided into partitions of approximately equal size (±1 gRNA) accounting for data similarity, to assign highly similar gRNAs to the same partition. This was implemented as follows: (1) we computed the pairwise Hamming distance between all gRNAs based on their on-hot encoded 30mer sequences (gRNA + context) with the SciPy *pdist* function[41] (normalized distances from *pdist* were multiplied by the size of the one-hot encoded array (1×120)); (2) for each gRNA x we stored a list of all gRNAs with Hamming distance ≤ 8 in the one-hot space, which corresponds to a sequence difference ≤ 4 nt; these were regarded as gRNAs "similar" to gRNA x; (3) gRNAs similar to at least one other gRNA in the dataset were the first to be distributed, randomly, in the partitions; whenever a gRNA x was assigned to a partition, all the gRNAs y, z… similar to it (and recursively those similar to y, z, …) were also added to the same partition; (4) once all similar gRNAs were exhausted the remaining gRNAs, not similar to any other, were split into three subsets based on their efficiency (inefficient: up to efficiency percentile 25 (25p), medium-efficient: from 25 to 75p, and highly efficient: above 75p) and the gRNAs in these three subsets were distributed to the partitions pseudo-randomly by assigning a balanced amount of inefficient, medium-efficient and highly efficient gRNAs to each of the partitions until they reached their predetermined size. To preserve gRNAs from the test set of Kim et al. (2019) in a single partition, used as internal independent test set to compare the performances of CRISPRon and DeepSpCas9, the gRNAs in the test set of Kim et al. (2019) were collected in an initial group, which was assigned to the partition destined for usage as internal independent test set prior any other data partitioning. Other gRNAs in the merged dataset similar to any of the gRNAs present in this initial group were added to it during the generations of the partitions, to maintain the internal test set fully independent.

Test settings for the evaluation and comparison of models. Test datasets (both internal and external) were made fully independent by removing all gRNAs highly similar to a gRNA in the training sets of any of the models being compared as follows: (1) the pairwise Hamming distance between the gRNAs in the test and training datasets was computed using the on-hot encoded 20 nt gRNA sequences with the SciPy *cdist* function[41] (normalized distances from *cdist* were multiplied by the size of the one-hot encoded array (1×80)); (2) gRNAs with Hamming distance ≤ 6 in the one-hot space, which corresponds to a sequence difference ≤ 3 nt, were removed. While for the generation of dataset partitions gRNA similarities were computed on 30mer gRNA + context sequences, the sole 20 nt gRNA spacers were employed during the processing of the test datasets because in the dataset of Wang et al. (2019) target contexts are highly different from those in other datasets for identical gRNAs. More restrictive thresholds of similarity (sequence difference ≤ 4 or 5) were also tested. No difference in the general performance of CRISPRon (v0 and v1) and in the comparison with other models were observed, and all of the significant improvements remained as such (Supplementary Data 2). Notably, the fluctuations in performances given by different similarity thresholds were both positive and negative.

Gradient boosting regression trees (GBRTs) for features analysis. Validation hyperparameters were chosen from the following screen: learning rate chosen from [0.08, 0.09, 0.1], maximum tree depth chosen from [3, 5, 7], minimum number of samples to generate a new split chosen from [5, 10, 15, 20], minimum number of samples to be present in a leaf node chosen from [5, 10, 15, 20], total number of trees in the model chosen from [400, 600, 800, 1000]. The validation of hyperparameters was made twice, the first time on five out of six partitions of the dataset, preserving the 6th partition as internal independent test set, and the second time on all six partitions. Selected hyperparameters are reported in Supplementary Table 3. During the validation, each GBRT was initialized five times with different seeds and the best model of the 5 was chosen for each fold/validation set. Predictions were computed by averaging the output of the best GBRTs chosen for each fold. When comparing multiple predictors, independent test datasets were cleaned from gRNAs with ≤3 nt difference on the 20 nt sequence of a gRNA in the training set of any compared predictor.

The CRISPRon deep learning model. The training of our deep learning models uses the Keras/Tensorflow 2.2.0[43] neural network framework with Python 3.8.3. Our strategy takes outset in the deep learning strategies by Wang et al.[8] and Kim et al. (2019, 2020)[7,9]. We employed a one-hot encoding of the input sequence (30mer gRNA + context), which was fed into a number of 3, 5, and 7 sized filters acting directly on the one-hot encoded sequence. The convolutions, which are the outputs of the filters, were flattened and fed into two sequential fully connected layers before giving the gRNA efficiency as the final output (for the full model layout see Supplementary Fig. 13). The number of weights and the layout of the convolutions as well as those of the two final fully connected layers are identical to the architecture used in Kim et al. 2019 and since the hyperparameters and layout of their model was substantially interrogated, we have not attempted further optimizations of this part of our model for CRISPRon. However, the inclusion of an important biological parameter in the deep learning framework was optimized as detailed below.

The partitioning of the data in to 6 subsets used for the GBRT were reused in the training of the deep learning models (see Supplementary Table 1). As in the regular machine learning above, the deep learning models were initially trained on 5-fold cross-validation with the 6th partition set aside for internal independent testing. Each training in the 5-fold cross-validation was repeated 10 times using random seeds and the best model of the 10 was chosen for each fold/validation set. The final output is the average of the output of the best models chosen for each fold. Finally, the process was repeated using all six data partitions for 6-fold cross-validation without an internal independent test set.

The most important biological parameters obtained from the GBRT model Gini and SHAP analysis were ΔG_B, the GC content of the 30mer and the folding energy of the spacer gRNA. Of these, ΔG_B was a far better representative of the on-target efficiency and we therefore decided to include ΔG_B in our deep learning model[27]. Direct inclusion of ΔG_B along-side the convolutions led to an improvement of the mean square error (MSE) from 143.15 to 141.76 on the average of the 5-fold cross-validations of the combined dataset from our study and Kim et al. (2019) (see Supplementary Table 4 and Supplementary Figs. 13-14 for the model layouts, Supplementary Table 5 for the results). Collecting the convolutions in a separate fully connected layer before combining the fully connected layer with ΔG_B led to a further improvement of the average MSE on the 5-fold cross-validation from 144.73 with three fully connected layers but without ΔG_B to 140.83 with ΔG_B (see Supplementary Table 4 and Supplementary Figs. 15-16 for the model layouts, Supplementary Table 5 for the results). The model with convolutions collected in a fully connected layer before combination with ΔG_B thus became our final CRISPRon model as outlined in Fig. 2a with details in Supplementary Fig. 16. This model was trained on the combined dataset from our study and Kim et al. (2019) dataset, split in six partitions, using 6-fold cross-validation. The final CRISPRon-v1.0 output is the average output of the best models obtained from each of the six validation sets after 10 repetitions (see Supplementary Table 6).

All the models were trained and evaluated using the MSE and the training was performed in epochs, where the weights were updated after each batch of 500 examples. The training was stopped when the performance on the validation set did not improve for 100 consecutive epochs and the best performing model by MSE on the validation set was kept. In effect, the training typically ran for 500–1500 epochs in total. The introduction of ΔG_B in the model changed the convergence behavior and we therefore screened for optimal learning rates testing learning as follows. We trained deep learning models on the LK-5 datasets using the layouts with only two fully connected layers and direct inclusion of ΔG_B and tested learning rates of 0.001, 0.0005, 0.0001, and 0.00005 in ten repetitions on each of the 5-fold validation sets (Supplementary Table 7). The optimal learning rate was 0.0001 using ADAM optimization and as above using a batch size of 500. The hyperparameters were used in the further training of the final CRISPRon deep learning model, which includes an extra fully connected layer for collection of the convolutions prior to the inclusion of ΔG_B.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

High-throughput sequencing data have been deposited to the China National GeneBank (accession number CNP0001031) and the GEO repository (accession number GSE173708). The gRNA efficiency data are provided in Supplementary Data 1. The Drugable gene database can be accessed to the link http://dgidb.org. The lentivirus vector used for cloning surrogate oligonucleotides is made available through Addgene (Plasmid #170459). Source data are provided with this paper.

Code availability

CRISPRon website and source via: https://rth.dk/resources/crispr/ and on https://github.com/RTH-tools/crispron[22].

Received: 19 May 2020; Accepted: 6 May 2021;

References

1. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
2. Farboud, B. et al. Enhanced genome editing with Cas9 ribonucleoprotein in diverse cells and organisms. *J. Vis. Exp.* **135**, 57350 (2018).
3. Mali, P. et al. RNA-guided human genome engineering via Cas9. *Science* **339**, 823–826 (2013).
4. Hwang, W. Y. et al. Efficient genome editing in zebrafish using a CRISPR-Cas system. *Nat. Biotechnol.* **31**, 227–229 (2013).
5. Cong, L. et al. Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).
6. Muhammad Rafid, A. H., Toufikuzzaman, M., Rahman, M. S. & Rahman, M. S. CRISPRpred(SEQ): a sequence-based method for sgRNA on target activity prediction using traditional machine learning. *BMC Bioinformatics* **21**, 223 (2020).
7. Kim, N. et al. Prediction of the sequence-specific cleavage activity of Cas9 variants. *Nat. Biotechnol.* **38**, 1328–1336 (2020).
8. Wang, D. et al. Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. *Nat. Commun.* **10**, 4284 (2019).
9. Kim, H. K. et al. SpCas9 activity prediction by DeepSpCas9, a deep learning-based model with high generalization performance. *Sci. Adv.* **5**, eaax9249 (2019).
10. Listgarten, J. et al. Prediction of off-target activities for the end-to-end design of CRISPR guide RNAs. *Nat. Biomed. Eng.* **2**, 38–47 (2018).
11. Chuai, G. et al. DeepCRISPR: optimized CRISPR guide RNA design by deep learning. *Genome Biol.* **19**, 80 (2018).
12. Rahman, M. K. & Rahman, M. S. CRISPRpred: a flexible and efficient tool for sgRNAs on-target activity prediction in CRISPR/Cas9 systems. *PLoS ONE* **12**, e0181943 (2017).
13. Chari, R., Yeo, N. C., Chavez, A. & Church, G. M. sgRNA Scorer 2.0: a species-independent model to predict CRISPR/Cas9 activity. *ACS Synth. Biol.* **6**, 902–904 (2017).
14. Doench, J. G. et al. Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.* **34**, 184–191 (2016).
15. Wong, N., Liu, W. & Wang, X. WU-CRISPR: characteristics of functional guide RNAs for the CRISPR/Cas9 system. *Genome Biol.* **16**, 218 (2015).
16. Moreno-Mateos, M. A. et al. CRISPRscan: designing highly efficient sgRNAs for CRISPR-Cas9 targeting in vivo. *Nat. Methods* **12**, 982–988 (2015).
17. Xu, H. et al. Sequence determinants of improved CRISPR sgRNA design. *Genome Res.* **25**, 1147–1157 (2015).
18. Hart, T. et al. High-resolution CRISPR screens reveal fitness genes and genotype-specific cancer liabilities. *Cell* **163**, 1515–1526 (2015).
19. Doench, J. G. et al. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat. Biotechnol.* **32**, 1262–1267 (2014).
20. Chari, R., Mali, P., Moosburner, M. & Church, G. M. Unraveling CRISPR-Cas9 genome engineering parameters via a library-on-library approach. *Nat. Methods* **12**, 823–826 (2015).
21. Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80–84 (2014).
22. Corsi, G. I., Gorodkin, J. & Anthon, C. *CRISPRon github page.* https://doi.org/10.5281/zenodo.4725572. (2021).
23. Allen, F. et al. Predicting the mutations generated by repair of Cas9-induced double-strand breaks. *Nat. Biotechnol.* **37**, 64–72 (2019).
24. Shen, M. W. et al. Predictable and precise template-free CRISPR editing of pathogenic variants. *Nature* **563**, 646–651 (2018).
25. Labuhn, M. et al. Refined sgRNA efficacy prediction improves large- and small-scale CRISPR-Cas9 applications. *Nucleic Acids Res.* **46**, 1375–1385 (2018).
26. Jensen, K. T. et al. Chromatin accessibility and guide sequence secondary structure affect CRISPR-Cas9 gene editing efficiency. *FEBS Lett.* **591**, 1892–1901 (2017).
27. Alkan, F., Wenzel, A., Anthon, C., Havgaard, J. H. & Gorodkin, J. CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters. *Genome Biol.* **19**, 177 (2018).
28. Robinson, J. T., Thorvaldsdóttir, H., Turner, D. & Mesirov, J. P. igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). *bioRxiv* https://www.biorxiv.org/content/10.1101/2020.05.03.075499v1 (2020).
29. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* Vol. 30. (eds. Guyon, I.) 4765–4774 (Curran Associates Inc., 2017).
30. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
31. Cullot, G. et al. CRISPR-Cas9 genome editing induces megabase-scale chromosomal truncations. *Nat. Commun.* **10**, 1136 (2019).
32. Kosicki, M., Tomberg, K. & Bradley, A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. *Nat. Biotechnol.* **36**, 765–771 (2018).
33. Xiang, X. & Luo, Y. *High throughput quantification of CRISPR gRNA efficiency based on surrogate lentivirus libraries.* https://doi.org/10.17504/protocols.io.bt9jnr4n. (2021).
34. Cotto, K. C. et al. DGIdb 3.0: a redesign and expansion of the drug-gene interaction database. *Nucleic Acids Res.* **46**, D1068–D1073 (2018).
35. McKenna, A. & Shendure, J. FlashFry: a fast and flexible tool for large-scale CRISPR target design. *BMC Biol.* **16**, 74 (2018).
36. Kuscu, C. et al. CRISPR-STOP: gene silencing through base-editing-induced nonsense mutations. *Nat. Methods* **14**, 710–712 (2017).
37. Patel, R. K. & Jain, M. NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. *PLoS ONE* **7**, e30619 (2012).
38. Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957–2963 (2011).
39. Li, H. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).
40. Haeussler, M. et al. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. *Genome Biol.* **17**, 148 (2016).
41. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).
42. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.* **6**, 26 (2011).
43. Abadi, M. et al. *TensorFlow: Large-scale machine learning on heterogeneous systems.* tensorflow.org. (2015).

Acknowledgements

This project was partially supported by the Sanming Project of Medicine in Shenzhen (SZSM201612074, to L.B. and Y.L.), Qingdao-Europe Advanced Institute for Life Sciences Grant (Y.L.), Guangdong Provincial Key Laboratory of Genome Read and Write (No. 2017B030301011 to X.X.), Guangdong Provincial Academician Workstation of BGI Synthetic Genomics (No. 2017B090904014 to X.X.), the Innovation Fund Denmark (4108-00008B, 4096-00001B to J.G.) and the Danish Research Council (9041-00317B to J.G.), Danish Research Council (9041-00317B to Y.L.), European Union's Horizon 2020 research and innovation program under grant agreement No 899417 (Y.L.), the Lundbeck Foundation (R219–2016-1375 to L.L.), the DFF Sapere Aude Starting grant (8048-00072 A to L.L.), and the National Human Genome Research Institute of the National Institutes of Health (RM1HG008525 to G.C.). We thank the China National GeneBank for the support of executing the project under the framework of Genome Read and Write.

scientific reports

Check for updates

OPEN

Novel circRNA discovery in sheep shows evidence of high backsplice junction conservation

Endika Varela-Martínez[1], Giulia I. Corsi[2], Christian Anthon[2], Jan Gorodkin[2]✉ & Begoña M. Jugo[1]✉

Circular RNAs (circRNAs) are covalently closed circular non-coding RNAs. Due to their structure, circRNAs are more stable and have longer half-lives than linear RNAs making them good candidates for disease biomarkers. Despite the scientific relevance of these molecules, the study of circRNAs in non-model organisms is still in its infancy. Here, we analyse total RNA-seq data to identify circRNAs in sheep from peripheral blood mononuclear cells (PBMCs) and parietal lobe cortex. Out of 2510 and 3403 circRNAs detected in parietal lobe cortex and in PBMCs, a total of 1379 novel circRNAs were discovered. Remarkably, around 63% of all detected circRNAs were found to be completely homologous to a circRNA annotated in human. Functional enrichment analysis was conducted for both tissues based on GO terms and KEGG pathways. The enriched terms suggest an important role of circRNAs from encephalon in synaptic functions and the involvement of circRNAs from PBMCs in basic immune system functions. In addition to this, we investigated the role of circRNAs in repetitive vaccination experiments via differential expression analysis and did not detect any significant relationship. At last, our results support both the miRNA sponge and the miRNA shuttle functions of CDR1-AS in sheep brain. To our knowledge, this is the first study on circRNA annotation in sheep PBMCs or parietal lobe cortex samples.

Circular RNAs (circRNAs) are a new class of covalently closed circular non-coding RNAs, formed when a splice donor and upstream acceptor from a linear RNA are linked together, a process also called backsplicing[1]. Due to their circular structure, circRNAs are more stable, resistant to RNAse R and have longer half-lives than linear RNAs[2], making them good candidates for disease biomarkers. Despite being discovered long ago, with the first circular molecules (viroids) revealed by electron microscopy in 1976[3] and the first endogenous circRNA originating from the *DCC* tumour suppressor reported in humans in 1991[4], for a long time circRNAs were thought to be low abundance products derived from splicing errors[5]. With the recent increase in high-throughput sequencing studies, it was shown that these molecules are more common than initially thought and that some of them have important roles in multiple pathways[6,7]. The exact mechanism of circularization is not totally understood, but multiple factors have been related. It has been shown that circRNA biogenesis is positively correlated by RNA polymerase II elongation rate[8]. In addition, multiple reports have shown that reverse complementary sequences in the flanking introns of the backspliced exons brings under close proximity the splice sites[9], allowing for the canonical spliceosomal machinery to be employed. Furthermore, RNA binding proteins such as Quaking (QKI), muscleblind (MBL) and fused in sarcoma (FUS) have also been reported to promote circRNA biogenesis[9].

Although the biological function of most circRNAs remains unknown, some circRNAs have been shown to contain clusters of miRNA binding sites that function as miRNA sponges (e.g., the circRNAs related to CDR1 and SRY sequester miR-7 and miR-138, respectively)[10]. Thus, circRNAs may interfere in the usual miRNA-mRNA binding procedures. Other circRNAs have been shown to contain sequences that can act as internal ribosome entry sites (IRESes), such as circ-ZNF609[11], thus can potentially code for proteins. However, their actual translation in vivo remains to be probed. Last, circRNAs can regulate a number of processes via protein-binding activity (e.g., the circ-FOXO3 forms a ternary complex with p21 and CDK2)[12].

Recent reports have associated circRNA expression with multiple diseases and it has opened a new field for diagnosis and treatment. It has been shown that circRNA levels increase with age in brain, but the same has been shown in age-associated neurological disorders such as Alzheimer's disease and Parkinson's disease[13]. In addition

[1]Department of Genetics, Physical Anthropology and Animal Physiology, Faculty of Science and Technology, University of the Basque Country (UPV/EHU), Bº Sarriena, 48940 Leioa, Spain. [2]Department of Veterinary and Animal Sciences, Center for Non-Coding RNA in Technology and Health, University of Copenhagen, Thorvaldsensvej 57, 1871 Frederiksberg, Denmark. ✉email: gorodkin@rth.dk; begonamarina.jugo@ehu.eus

to evidence of circRNAs playing a role in diseases such as atherosclerotic vascular disease risk, osteoarthritis and diabetes, it has been shown dysregulated expression of circRNAs in multiple types of cancer, including colorectal cancer, hepatocellular carcinoma and breast cancer, among others[14].

More recently, many circRNAs have been reported to be expressed abnormally and play important roles in the progression of autoimmune diseases such as rheumatoid arthritis, systemic lupus erythematosus or multiple sclerosis[15]. Thus, circRNAs may not only serve as potential biomarkers but also act as immune regulators an offer potential opportunities for therapy.

Non-living vaccine antigens, especially purified or recombinant subunit vaccines, are often poorly immunogenic and require additional components to help stimulate protective immunity based on antibodies and effector T cell functions. These additional components, termed adjuvants, are added to vaccines to achieve a better protection, with the aluminium-based ones (especially aluminium hydroxide) being some of the most widely employed adjuvants in human and animal vaccines. Despite its widespread use and its probed safety record, the adjuvant's mechanism of action is not fully understood.

Recently, some concerns regarding the safety of aluminium adjuvants has been raised, due to the possibility for aluminium adjuvants to reach distant organs such as spleen or brain after a long-term exposition. It was shown that after intramuscular injection of the aluminium adjuvant in mice, the material was translocated at a very slow rate in normal conditions to draining lymph nodes (DNL) and thereafter was detected as associated with phagocytes in blood and spleen[16]. In addition, several studies have addressed the translocation of aluminium to the brain[16–18]. However, this remains a subject with much controversy in the scientific community and there is no complete agreement regarding the translocation and biopersistence of this material[17,19,20].

In sheep, a form of the autoimmune/autoinflammatory syndrome induced by aluminium-adjuvants has been described as linked to repetitive inoculation with aluminium-containing vaccines[21]. In this species, a number of circRNAs were previously identified from RNA sequencing data. Li et al. detected 6133 and 10,226 circRNAs in prenatal and postnatal muscle and pituitary glands of sheep, respectively[13,14]. Interestingly, they observed an association of some circRNAs with economically important traits, such as the growth and development of muscle related signaling pathways in the first tissue and the regulation of hormone secretion in the second. In addition to this, the same group identified 9231 circRNAs differentially expressed in the estrus and anestrus pituitary system of sheep[15]. Last, 886 circRNAs were detected in the skeletal muscle by Cao et al., and some of them were reported to be involved in muscle cell development and signaling pathway[16]. Characterizing the circRNA profiles of specific tissues and cell types is a promising way to reveal functional properties of circRNAs.

Until now, there has been no study trying to address the functional role of circRNAs in aluminium adjuvancy through total RNA sequencing data analysis, nor attempts of annotating circRNAs in sheep peripheral blood mononuclear cells (PBMCs) or parietal lobe cortex samples. In this work the circRNAs of these two tissues will be characterized and their expression in animals with different adjuvancy treatments assessed. Characterizing how circRNAs are expressed in different tissues can improve our understanding of the sheep transcriptome and analysing their expression in vaccinated or adjuvanted animals could add information on the role of circRNAs in the immune response to aluminium adjuvants.

Results

CircRNAs characterization and distribution in encephalon and PBMCs. Total RNA-seq data was produced from RNA samples extracted from encephalon and PBMCs. The data have been previously used for in depth differential expression analyses[22,23] and it has been re-analysed for circRNA annotation. Two bioinformatics tools, Segemehl[24] and DCC[25], were selected for circRNA identification, which resulted in 12,475 and 60,375 candidate circRNAs in encephalon and 19,611 and 63,138 candidate circRNAs in PBMC samples by segemehl and DCC, respectively. Out of all the circRNAs detected in the encephalon, 4996 had concordant coordinates in both tools. After filtering circRNAs based on their abundance and expression patterns among samples (see "Material and methods"), 2510 circRNAs were selected for subsequent analyses. In PBMCs, 10,414 circRNAs were concordant between tools. After filtering, 3403 circRNAs were retained for further analysis. Details about filtered circRNAs are available as Supplementary Data S1 and S2 for encephalon and PBMCs, respectively. The naming of circRNAs in each tissue list was performed by assigning sequential unique numeric identifiers. From the 2510 and 3403 circRNAs detected in encephalon and PBMCs, 1236 were present concordantly in both tissues (Fig. 1). The counts from DCC were taken as reference abundance values.

In the available literature a number of studies have described the principal characteristics of circRNAs in human and mouse[10,26]. In our sheep data, in both tissues, we observe that the longer the chromosome, the more circRNAs are detected (Supplementary Fig. S1), and that the circRNAs are most commonly formed by two or three exons, being those composed of two exons the most prevalent ones (Supplementary Fig. S1). This is in accordance with what was previously described in other species[10]. A representation of the location of each circRNA in the reference genome is given in Supplementary Fig. S2 for encephalon and Supplementary Fig. S3 for PBMCs.

Out of the 2510 candidate circRNAs detected in encephalon, 2372 overlap with 1642 annotated sheep genes. Of those circRNAs that originated from an annotated gene, 1927 were concordant with an annotated exon–intron boundary in both ends, while in the other cases, despite the overlap with an annotated gene, at least one end was not concordant with an annotated exon–intron boundary. Concerning the 3403 circRNAs detected in PBMCs, 3249 were found to originate from 2006 annotated sheep genes. Of these, 2597 were concordant with an annotated exon–intron boundary in both ends. In some cases, the cause of the discrepancy between the annotated exon–intron boundaries and the circRNA backspliced junctions could be explained by the incomplete state of the sheep gene annotation. The majority of genes host only one circRNA in both tissues (Supplementary Fig. S1).

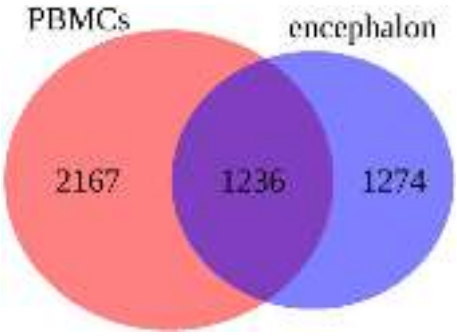


Figure 1. Venn diagram with the number of circRNAs detected in each tissue after filtering for a minimum expression in at least three samples.

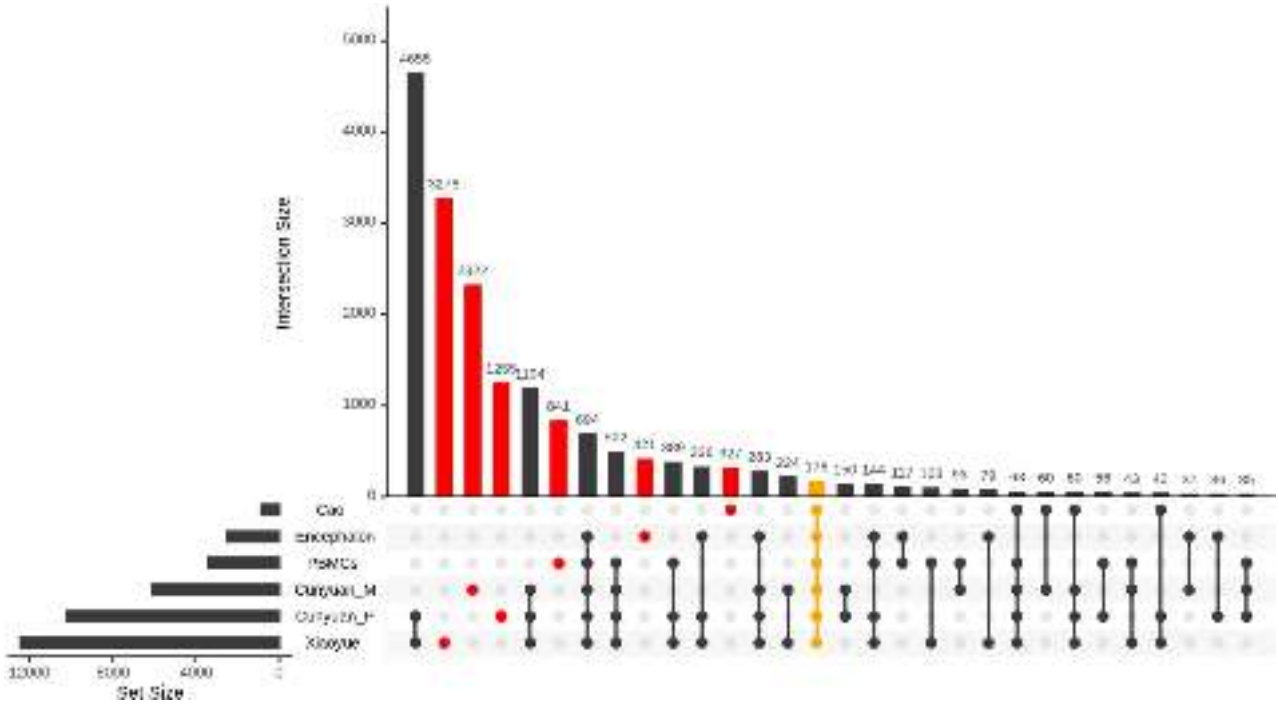


Figure 2. UpSet plot with the comparison of detected circRNAs in different studies. Encephalon and PBMCs refers to the circRNAs detected in this study, while Cunyuan_P (pituitary gland), Xiaoyue (pituitary gland), Cunyuan_M (longissimus dorsi muscle) and Cao (longissimus dorsi muscle) refers to the circRNAs detected in[45–48], respectively. Cells filled with a dot indicate the circRNA is in the corresponding database, while empty cells indicate that the circRNA is not present in the corresponding database. In red the circRNAs that are exclusively expressed in one database and in orange the circRNAs common to all databases. Intersections with less than 30 elements were removed for visualization purposes.

Sheep circRNAs are conserved. CircRNAs have been shown to be tissue specific and to be evolutionary conserved[27]. The circRNAs detected in this study were compared to others previously identified in other tissues (pituitary gland and longissimus dorsi muscle) in sheep. Notably, only 175 circRNAs were consistently detected in all tissues, including ours (Fig. 2). Such low concordance is in agreement with other studies, which showed that the expression of circRNAs is tissue-dependent[8]. In addition, our results showed that 421 and 841 circRNAs were exclusive to the encephalon and PBMCs data, respectively, while the overlap between the two sets is composed of 117 circRNAs (Fig. 2).

In addition to this, the detected circRNAs were compared to the human circRNAs annotated in CIRCpedia[28]. First, sheep circRNA coordinates were translated to human ones with the UCSC liftOver tool[29] and classified based on their backsplice junction conservation. Out of the 2510 detected circRNAs in encephalon, 52 splice sites coordinates could not be lifted. For the rest, nearly all had at least one reported human circRNA utilizing one of the splice sites. A total of 1606 (63.98%) circRNAs were completely homologous to a human circRNA (Fig. 3a).

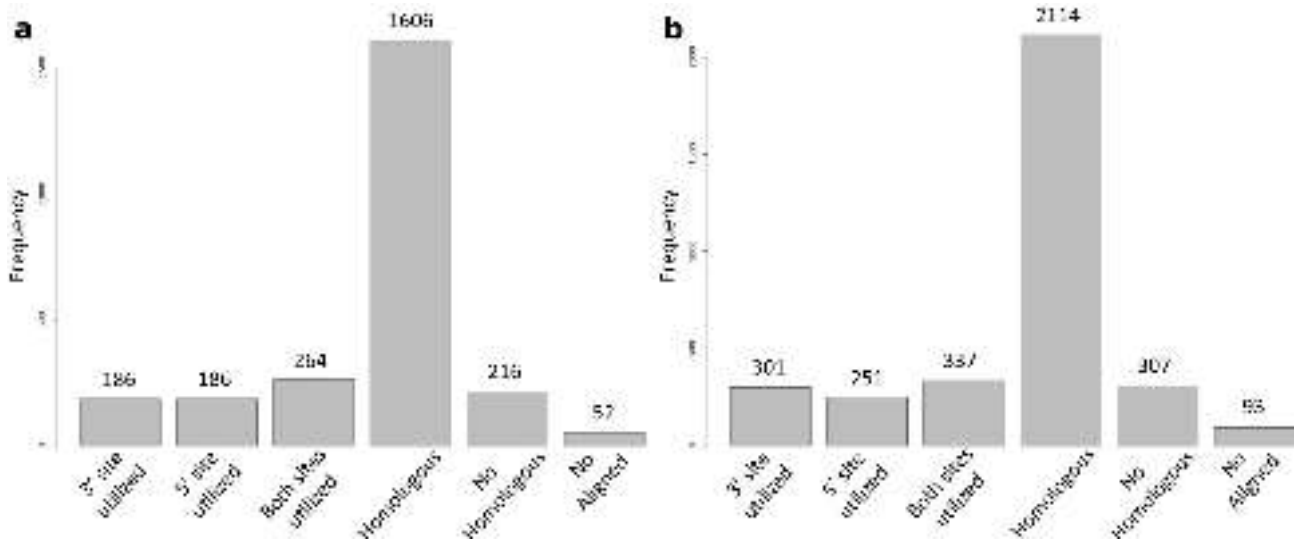


Figure 3. Bar plot with the result of the conservation analysis. In the x-axis the different categories described in "Material and methods" and in the y-axis the number of circRNAs in each category. (**a**) Encephalon; (**b**) PBMCs.

Figure 4. Sub-network from enriched GO terms by g:Profiler in encephalon and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; node colour represents the significance level (FDR).

In PBMCs, out of the 3403 detected circRNAs, 93 splice sites coordinates were not lifted to human, while 2114 (62.12%) circRNAs were found to be completely homologous to a human circRNA (Fig. 3b).

Given that circRNAs include exons of coding genes, sheep circRNAs completely homologous to a human one but lacking a gene annotation in sheep were also screened for possible corresponding genes annotated in human (Supplementary Table S1).

Enrichment analysis. A functional enrichment analysis was conducted with g:Profiler[30] on the GO[31] and KEGG[32] databases for both tissues, by considering the terms annotated for the parental genes of the detected circRNAs and after setting as background all the genes expressed in the corresponding tissue. Terms with an FDR less than 0.05 were selected as significant. The enriched GO terms are represented as networks in Supple-

Figure 5. Sub-network from enriched GO terms by g:Profiler in PBMCs and visualized in Cytoscape after clustering with Autoannotate. Node size correspond to number of genes expressed from the term; edge size represents the number of genes that overlap between different terms; node colour represents the significance level (FDR).

mentary Fig. S4 and S5. Selected highly connected sub-networks of interest are shown in Figs. 4 and 5. The 20 most enriched KEGG pathways are shown in Fig. 6a,b, for encephalon and PBMCs, respectively. Among the GO terms significantly enriched in encephalon, there are a number of terms related to synapse regulation, presynaptic endocytosis, behaviour, brain development and myelination, while among the KEGG pathways glutamatergic synapse, dopaminergic synapse and serotonergic synapse were enriched, suggesting an important role for some circRNAs in synaptic functions. Instead, in PBMCs, we retrieved GO terms related to B- and T-cell proliferation, T-cell differentiation, activation and regulation of immune response and neutrophil degranulation. In both tissues, the KEGG T-cell receptor signaling pathway and B-cell receptor signaling pathway were enriched, suggesting that some circRNAs may be involved in basic immune system functions.

circRNAs acting as sponges. To identify circRNAs which could function as miRNA sponges, we compared all 2510 (encephalon) and 3403 (PBMCs) predicted circRNAs with clusters of miRNA binding sites reported by Pan et al.[33] in the human genome, a dataset that comprises a total of 3673 predicted sponges for 1250 miRNAs. Out of 3 (encephalon) and 4 (PBMCs) sheep circRNAs overlapping one or more candidate sponges-miRNA pairs, we filtered out those entries for which the predicted sponged miRNA does not have a homologous pre-miRNA in sheep. As a result, in the encephalon tissue we identified 1 circRNA (circRNA4960) overlapping predicted sponges for two miRNAs (miR-7 and miR-1224), while in PBMCs we retained two circRNAs, circRNA2342, which overlaps predicted sponges for miR-409, miR-383, miR-370, miR-369 and miR-212, and circRNA8181 for miR-124 (Supplementary Table S2). Then circRNA-target-miRNA pairs were screened for miRNA binding sites in both human and sheep circRNA sequences with RIsearch2[34]. After removing overlapping binding sites as described in Pan et al.[33], 44 and 65 binding sites were respectively found on circRNA4960 for miR-7 and miR-1224. Although the sheep circRNA4960 is shorter than the corresponding cluster of miRNA binding sites detected in human for miR-7 and miR-1224, the per-base binding sites ratio is higher in sheep, further underlying a possible functional role of this molecule in the sheep brain.

One of the most well characterized circRNAs in brain is the one related to the CDR1 gene[35]. Although CDR1 is not annotated in sheep, blasting the human sequence of this gene against the sheep reference genome results in a single hit, matching a region of circRNA4960, detected in our encephalon samples. We lifted the coordinates

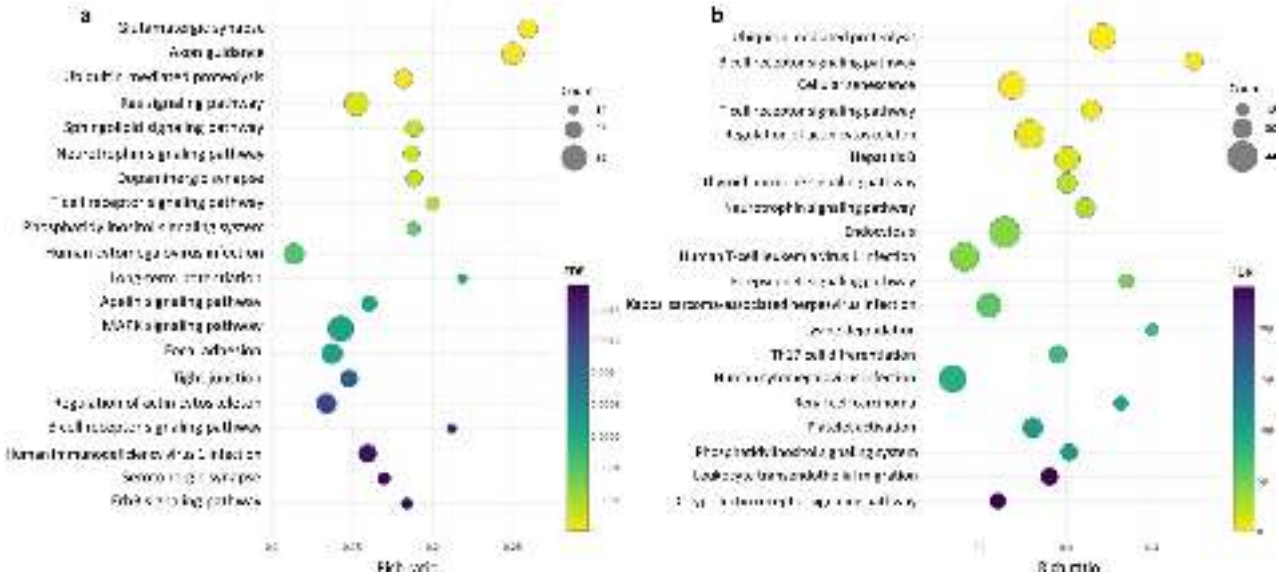


Figure 6. The 20 most enriched KEGG pathways by g:Profiler. The bubble plots show in the Y-axis the enriched KEGG pathways, while in the X-axis the rich ratio is represented (rich ratio = amount of differentially expressed genes in the term/all genes included in the term). Size and colour of the bubble represent the number of differentially expressed genes in the KEGG pathway and enrichment significance (FDR), respectively. (**a**) Encephalon; (**b**) PBMCs.

of the sheep backsplice junctions (sheep genome version Oar_3.1) to the human genome (version hg38) with the UCSC liftOver tool[29] and found that circRNA4960 is homologous to the human CDR1-AS. Interestingly, circRNA4960 was one of the most expressed in our cortex samples (Supplementary Table S3). Among the highly expressed circRNAs detected in encephalon other two were homologous to previously characterized human circRNAs, circRNA4266 and circRNA4357, which originate from HOMER1 and ZNF609 genes, respectively.

In addition, recent studies have shown that miR-671 has sufficient complementarity with CDR1-AS[36]. Interestingly, the binding pattern of miR-671 in sheep is identical to the human one and includes 13 canonical base pairs in the seed region, and only 1 mismatch over the entire sequence (Supplementary Fig. S6). Hence, our results support both the miRNA sponge and the miRNA shuttle functions previously proposed for CDR1-AS in brain and suggest a possible similar mechanism for miR-1224, which is reported as highly expressed in brain according to the Genotype-Tissue Expression (GTEx) Project v8.

The screening of circRNA-target-miRNA pairs identified in PBMCs showed that miRNA binding sites are scattered far away from one another over both the exonic and the much longer intronic regions of circRNA2342 and circRNA8181, with few bindings overlapping with the clusters of miRNA binding sites identified in human, hence we could not infer any sponge activity for these circRNAs. The complete list of binding sites identified for sheep circRNA-miRNA pairs in both encephalon and PBMCs candidate circRNA sponges is available in Supplementary Table S2.

No differential expression due to repetitive vaccination. Our preliminary analysis of transcripts expression showed that the adjuvant sample 116-E derived from the encephalon tissue was an outlier, thus it was removed prior differential expression analysis. A PCA showing clusters of samples is shown in Supplementary Fig. S7. Differential expression analysis was performed with the R package DESeq2[37]. We did not detect any differentially expressed circRNA in any comparison after considering an adjusted p-value < 0.05 as cut-off. We also performed differential expression analysis normalizing the data as spliced reads per billion mapping (SRPBM), and by applying a Kruskal–Wallis test before correcting for multiple comparison with the Benjamini and Hochberg method. Also in this case, there were no significant differences between groups when an adjusted p-value < 0.05 was taken as cut-off.

For the PBMCs samples, the Harman R package[38] was applied to remove any batch effect in the data after normalizing by SRPBM. The PCA with the corrected data is shown in Supplementary Fig. S7. Then, both the limma package[39] and Kruskal–Wallis test were used to test for differential expression, but no circRNA was found to be differentially expressed in any comparison with an adjusted p-value < 0.05.

Discussion

CircRNAs are a novel class of endogenous non-coding RNAs with a cyclic structure formed through a covalent bind of a linear transcript. Lately, circRNAs have gained more attention due to their abundance, their expression levels in specific tissues and their involvement in different biological functions, particularly studied in human and mouse[40–42]. However, studies on circRNAs in non-model organism such as sheep are still lacking, and there is no database recording such data yet. Here, we improved the annotation of circRNAs in sheep by adding a total of 1379 novel circRNAs, combined with relevant information such as conservation and potential function. This

set of robust circRNAs was selected from 2510 and 3403 circRNAs respectively detected in parietal lobe cortex and in PBMCs via in silico analysis of ribo-minus total RNA sequencing data. Most of the identified circRNAs in both tissues are from annotated genes, generally formed by two or three distinct exons, in agreement with what has been previously reported in human and mouse data[43]. In addition, we observe that circRNAs are widely expressed in both of these tissues in sheep, which was somewhat expected since circRNAs are enriched in mammalian brain and human PBMCs[44].

Some circRNAs have a tissue-dependent or developmental stage-dependent expression pattern[8]. The circRNAs detected in this study were compared to other sheep circRNA identified in pituitary gland[45,46] and in longissimus dorsi muscle[47,48]. Only 175 circRNAs were detected in all tissues, while several hundreds of circRNAs were exclusive to each tissue. Furthermore, given that numerous circRNAs have exhibited evolutionary conservation between human and mouse[49], the circRNAs detected in this study were analysed for backsplice site conservation, by comparing them to the human circRNAs available in CIRCpedia. We found that 1606 (63.98%) and 2114 (62.12%) sheep circRNAs have completely conserved backsplice sites between human and sheep in encephalon and PBMCs, respectively. Among the most expressed circRNAs, circRNA4266 and circRNA4357, in order originating from the HOMER1 and ZNF609 genes, had been previously characterized in other species. Consistent with this, it has been shown that the circRNA related to HOMER1 has a regulatory role in cell growth in human bronchial epithelial cells, as its silencing promotes cell proliferation[50]. The circRNA originated from ZNF609 has been shown to adsorb miR-150-5p and to upregulate SP1 transcription factor, promoting the proliferation of nasopharyngeal carcinoma cells[51]. In addition, this circRNA has been related to myoblast proliferation and the fact that its sequence includes an open reading frame and that a fraction of this circRNA is loaded into polysomes indicates that it may encode for proteins[11].

It was previously proposed that the binding activity between circRNAs and RNA binding proteins (RBPs) can have regulatory effects[52], which suggests that circRNAs can impact the same functional processes in which the corresponding linear host gene is involved. Under the assumption that the function of a circRNA may be associated with the known function of its parental gene, GO analysis indicated that the circRNAs identified in encephalon are related to synapse regulation, behaviour, learning process and brain development, while KEGG pathway analysis also related these circRNAs to synapses and to pathways implicated in cell proliferation such MAPK/ERK pathways, the last ones being previously linked to circRNAs[43]. In contrast, in the PBMCs samples, GO terms associated with the immune system such as B- and T-cell proliferation, neutrophil degranulation, the MAPK cascade and the NF-κB signaling were enriched, as well as DNA methylation and histone modification, supporting the possibility that circRNAs could be related to epigenetic alterations, as previously suggested[53]. In both tissues the B- and T-cell receptor signalling pathways were enriched, in addition to Fc epsilon RI signaling pathway, Th17 cell differentiation and platelet activation in PBMCs samples, indicating a potential functional role for circRNAs in the immune system response.

Then, we performed a differential expression analysis to find out if circRNAs could have a role in aluminium adjuvancy in vaccines. We did not detect any differentially expressed circRNAs in any of the two tissues, which indicates that circRNAs may not be connected with aluminium adjuvant effects. Despite this, it should be noted that no differential expression analysis software has been specifically designed to handle circRNA data, in which expression levels are generally lower compared to mRNA and are subjected to greater variability.

Moreover, we screened circRNAs for the presence of clusters of miRNA binding sites, following the concept that circRNAs can act as miRNA sponges. We report that the circRNA CDR1-AS, which corresponds to circRNA4960 in this study, contains numerous binding sites for miR-7 and miR-1224, both reported to be expressed in the mammalian brain. In agreement with our expectations, we observed that this circRNA is highly expressed only in our encephalon samples. In addition, recent studies have shown that miR-671 has sufficient complementarity with CDR1-AS to induce AGO2 endonucleolytic cleavage and, based on this, an alternative function for this circRNA molecule as miRNA shuttle system, releasing its miR-7 cargo upon binding with miR-671, has been proposed[36]. It was shown that the binding sites for miR-671 were retained in sheep, supporting its role in cleavage by AGO2.

In conclusion, a number of circRNAs were identified in sheep encephalon and PBMCs samples, expanding our knowledge on the sheep transcriptome. Moreover, several GO terms and KEGG pathways showed that circRNAs may be involved with synapse regulation and cell proliferation in encephalon and with the immune system response and epigenetic modifications in PBMCs. Furthermore, we showed how circRNA functions associated with the presence of clusters of miRNA binding sites are conserved between sheep and human. This study is a first systematic analysis of circRNAs in sheep parietal lobe cortex and PBMC samples, and it is also a first study of the changes in circRNA expression profiles after an aluminium-based adjuvant vaccine inoculation schedule.

Material and methods

Ethics statement. All experimental procedures were approved and licensed by the Ethical Committee of the University of Zaragoza (ref: PI15/14). Requirements of the Spanish Policy for Animal Protection (RED53/2013) and the European Union Directive 2010/63 on protection of experimental animals were always fulfilled.

Datasets. The data samples used in this work have been previously used for in depth differential expression analyses and detailed information about the experimental design and sequencing can be found in the corresponding articles for both tissues, PBMCs[22] and parietal lobe cortex[23]. Briefly, healthy three-month-old Rasa Aragonesa pure breed lambs from a single pedigree flock, with the condition of not having undergone any kind of vaccination before the experiment, were selected to be placed in the experimental farm of the university of Zaragoza. After a period of two months to acclimatize to the new environment, all lambs were randomly distributed in different treatment groups, each consisting of 7 animals. One of the groups, from now on denominated

Treatment	Time	Samples
Encephalon		
Adjuvant	End (Tf)	114-E, 115-E, 116-E, 117-E
Vaccine	End (Tf)	121-E, 122-E, 124-E, 126-E
Control	End (Tf)	131-E, 135-E, 136-E, 137-E
PBMCs		
Adjuvant	Start (T0)	121-A, 124-A, 125-A
	End (Tf)	121-B, 124-B, 125-B, 125-B[a]
Vaccine	Start (T0)	111-A, 114-A, 116-A
	End (Tf)	111-B, 114-B, 116-B

Table 1. Sample summary. [a]Same RNA sample obtained with a conventional TRIzol extraction method.

vaccine group (Vac), received a subcutaneous treatment with commercial vaccines based on aluminium hydroxide adjuvant. Another group, denominated adjuvant group (Adj), received equivalent doses to the commercial vaccines of aluminium hydroxide only (Alhydrogel, CZ Veterinaria, Spain) diluted in phosphate-buffered saline (PBS). Finally, PBS was administered to the control group. Blood samples were taken at the start (before any vaccination) and at the end of the experiment, while for encephalon (parietal lobe cortex) only samples at the end were taken. In Table 1 there is a summary of the samples used for sequencing.

The complete experiment lasted 475 days, from February 2015 to June 2016. During that period of time, nine different vaccines were administered, which comprises a total of 19 inoculations throughout 16 different inoculation dates. A total amount of 81.29 mg of Al per animal was given in the Vac and Adj groups. A detailed list of the commercial vaccines used in this study can be seen as supplementary material in a previous publication[22].

Out of all the animals, only 12 (four animals per group) and 6 (three animals per group at the start and at the end of the experiment) were selected for sequencing from encephalon and PBMCs, respectively. For both tissues, Illumina Total RNA-seq libraries were used and sequenced with a high sequencing depth.

CircRNA identification. First, a read quality filtering and trimming was performed with Trimmomatic[54] [v0.38] using the following criteria: (1) adaptor removal with the "palindrome" mode for paired-end data; (2) trimming of bases from the start or end of a read if their quality dropped below a Phred value of 20; (3) trimming of reads if the average quality within a sliding window of five nucleotides fell below 20; and (4) read filtering if their length was sorter than 40.

For circRNA identification two tools were selected, segemehl[24] [v0.3.4] and DCC[25] [v0.4.7]. Before running segemehl, quality filtered reads were first aligned to the sheep reference genome (Oar_v3.1) with HISAT2[55] [v2.1.0]. The set of non-aligned reads from the previous step were used to detect circRNAs in segemehl with default parameters. In contrast, for DCC, the quality filtered reads were first aligned to the reference genome with STAR [v2.6.1d][56] following DCC author recommendations. Then, the *chimeric.out.junction* files from the previous alignments and a file with repetitive regions in the sheep genome downloaded from the UCSC table browser (RepeatMasker and Simple Repeats tracks) were passed to DCC. DCC was run with default parameters, except that we require a circRNA had to be expressed with one read in at least one sample to be reported. For further analysis, different filtering criteria were tested for the encephalon and PBMC tissues, as they were subjected to different experimental setups. In both tissues circRNAs needed a minimum of 2 read counts to be considered as expressed. In addition, in encephalon, circRNAs were required to be expressed at least in the same three samples in both tools, while in PBMCs they needed to be expressed at least in the same three samples from one group in both tools. The expression counts for the detected circRNAs and host genes were taken as reference from DCC, focusing mainly in exonic circRNAs for further analysis (still referring them as circRNAs throughout the text).

Conservation analysis. The main databases of circRNA annotation are focused on human, mouse, rat, zebrafish, fly and worm, being sheep circRNA data not submitted to any database to date. A literature search of articles in which circRNAs in sheep are detected and are given at least as supplementary material was done in an attempt to compare the circRNAs annotated in this study. Four studies focusing on two different tissues were found: two from the pituitary gland[45,46] and another two from the longissimus dorsi muscle[47,48].

Then, the detected circRNAs were compared to the ones annotated in CIRCpedia[28] for human. The following steps were performed:

1. The 5′ and 3′ flank coordinates of each circRNA found in sheep were converted to human coordinates with the USCS liftOver tool[29] with default parameters (min. ratio of remapped bases = 0.95).
2. The resulting coordinates were screened for overlap with human annotated circRNAs in CIRCpedia. Splice sites detected in ± 2 nt intervals around the putative human sites were considered homologous.
3. Different categories were assigned to each circRNA: "not-aligned", the sheep coordinates were not translated to human with liftOver; "no homologous", no human circRNA detected near both splice sites; "5′ site utilized", a human circRNA that only uses the 5′ splice site is detected; "3′ site utilized", a human circRNA that

only uses the 3′ splice site is detected; "Both sites utilized", both splice sites are used by different circRNAs in human; and "homologous", a human circRNA using both splice sites is detected.

Enrichment analysis. The detected circRNAs whose origin was in an annotated gene were further analysed as follows. Gene enrichment analysis was conducted using the GO[31] and KEGG[32] databases in g:Profiler[30]. This tool computes p-values for enriched terms using a Fisher's exact test and applies the Benjamini–Hochberg multiple test correction. The set of all expressed genes detected in the total RNA-seq libraries was set as background and related terms associated with the host genes of the circRNAs were tested for enrichment. Terms composed of more than 400 genes, due to limited interpretative value, or composed of less than 5 genes, due to the decrease in statistical power by multiple testing correction, were removed from the analysis. Those terms with an FDR less than 0.05 were selected for further analysis. For visualization purposes, the list of enriched GO term was further analysed with Cytoscape using EnrichmentMap and Autoannotate plugins[57]. EnrichmentMap generates a network in which pathways are visualized as nodes connected between each other if they share many genes. Pathways with common genes often represent similar biological processes and are grouped together as subnetworks. Clusters with less than 3 interconnected nodes were removed for visualization purposes.

circRNAs acting as miRNA sponges. A list of predicted clusters of miRNA binding sites previously reported in the human genome (hg19) was downloaded from Pan et al.[33]. The genomic coordinates of each sponge candidate were converted to hg38 with liftOver (min. ratio of remapped bases = 0.95) and intersected with those of the circRNAs identified in this study, already lifted from the sheep reference genome to the human genome hg38 as explained above, with bedtools (min. fraction overlap = 75%). Results were then filtered by excluding sponges targeting miRNAs for which no high confidence orthologue sequence was reported in sheep according to Ensembl[58] (release 97). All human miRNAs hairpins were screened for similarity with the Oar3.1 genome with BLAST, requiring a minimum sequence identity of 90% on at least 95% of the hairpin. The sequences of the processed miRNAs were downloaded from miRBase[59] (Release 22.1) and the corresponding sheep orthologous were extracted from the alignment provided by Ensembl. CircRNAs were screened for miRNA binding sites with RIsearch2[34], using the following parameters: -s 1:8/6 -e -10 -l 20 -p2. In the same way we re-evaluated the clusters of miRNA binding sites identified in human and noticed almost no difference compared to the binding sites previously reported (Supplementary Table S2). The same criteria were applied to find binding sites of miR-671 on the human CDR1-AS and on the corresponding sheep circRNA4960.

Differential expression analysis. For the encephalon samples, the differential expression analysis was performed via two different methods. First, the analysis was done with DESeq2[37], setting an adjusted p-value < 0.05 as significance cut-off. An alternative method was also applied, given that DESeq2 is not designed to work on circRNA expression data. In this case, for normalization of the circRNA expression data, not only the circRNA counts were taken into consideration to calculate library sizes, but the total amount of reads aligned to the reference annotation was considered. The data was then normalized by SRPBM (Spliced Reads per Billion Mapped Reads)[5]. After normalization, a Kruskal–Wallis test was employed to check for differences between groups, and the resulting p-values were adjusted for multiple comparisons with the Benjamini and Hochberg method. An adjusted p-value < 0.05 was taken as significance cut-off to identify the differentially expressed circRNAs.

For the PBMC samples, a batch effect removal program, harman [v1.12.0][38], was applied after normalizing data by SRPBM. Then, the package limma[39] and the Kruskal–Wallis test were applied to check for differential expression. Those circRNAs with an adjusted p-value < 0.05 were taken as cut-off.

Data availability

RNA-seq data have been deposited in the NCBI Gene Expression Omnibus (GEO) database with experiment accession number GSE128597 for encephalon samples and GSE113899 for PBMCs samples.

Received: 25 September 2020; Accepted: 9 December 2020

References

1. Guo, J. U., Agarwal, V., Guo, H. & Bartel, D. P. Expanded identification and characterization of mammalian circular RNAs. *Genome Biol.* **15**, 409 (2014).
2. Bonizzato, A., Gaffo, E., Te Kronnie, G. & Bortoluzzi, S. CircRNAs in hematopoiesis and hematological malignancies. *Blood Cancer J.* **6**, e483 (2016).
3. Sanger, H. L., Klotz, G., Riesner, D., Gross, H. J. & Kleinschmidt, A. K. Viroids are single stranded covalently closed circular RNA molecules existing as highly base paired rod like structures. *Proc. Natl. Acad. Sci. U.S.A.* **73**, 3852–3856 (1976).
4. Nigro, J. M. *et al.* Scrambled exons. *Cell* **64**, 607–613 (1991).
5. Burd, C. E. *et al.* Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19**, 141–157 (2012).
6. Wang, D., Luo, Y., Wang, G. & Yang, Q. Circular RNA expression profiles and bioinformatics analysis in ovarian endometriosis. *Mol. Genet. Genomic Med.* **7**, e00756 (2019).
7. Sekar, S. & Liang, W. S. Circular RNA expression and function in the brain. *Non-coding RNA Res.* **4**, 23–29 (2019).
8. Ebbesen, K. K., Hansen, T. B. & Kjems, J. Insights into circular RNA biology. *RNA Biol.* **14**, 1035–1045 (2017).
9. Chioccarelli, T. *et al.* Histone post-translational modifications and circRNAs in mouse and human spermatozoa: Potential epigenetic marks to assess human sperm quality. *J. Clin. Med.* **9**, 640 (2020).
10. Ragan, C., Goodall, G. J., Shirokikh, N. E. & Preiss, T. Insights into the biogenesis and potential functions of exonic circular RNA. *Sci. Rep.* **9**, 2048 (2019).
11. Legnini, I. *et al.* Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. *Mol. Cell* **66**, 22-37.e9 (2017).

Published online 1 February 2021

Nucleic Acids Research, 2021, Vol. 49, No. 4 ***1859–1871***
doi: 10.1093/nar/gkab012

Human pathways in animal models: possibilities and limitations

Nadezhda T. Doncheva[1,2,3,*], Oana Palasca[1,2,3], Reza Yarani[4], Thomas Litman[5,6], Christian Anthon[1,2], Martien A. M. Groenen[7], Peter F. Stadler[1,8,9,10,11,12], Flemming Pociot[1,4,13], Lars J. Jensen[1,3,*] and Jan Gorodkin[1,2,*]

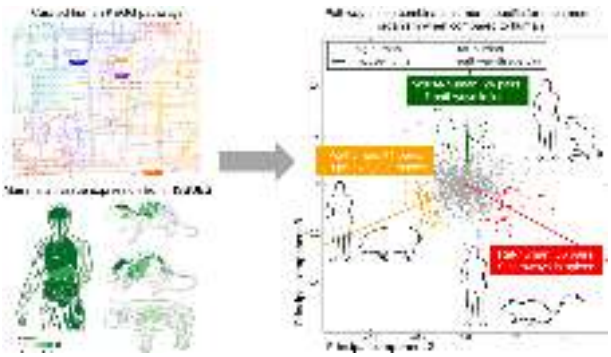
[1]Center for non-coding RNA in Technology and Health, University of Copenhagen, 1871 Frederiksberg, Denmark, [2]Department of Veterinary and Animal Sciences, University of Copenhagen, 1870 Frederiksberg, Denmark, [3]Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark, [4]Translational Type 1 Diabetes Research, Steno Diabetes Center Copenhagen, 2820 Gentofte, Denmark, [5]Department of Immunology and Microbiology, University of Copenhagen, 2200 Copenhagen, Denmark, [6]Exploratory Biology, LEO Pharma A/S, 2750 Ballerup, Denmark, [7]Animal Breeding and Genomics, Wageningen University & Research, 6700 Wageningen, The Netherlands, [8]Bioinformatics Group, Department of Computer Science; Interdisciplinary Center for Bioinformatics; German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig; Competence Center for Scalable Data Services and Solutions Dresden-Leipzig; Leipzig Research Center for Civilization Diseases; and Centre for Biotechnology and Biomedicine, University of Leipzig, 04107 Leipzig, Germany, [9]Max Planck Institute for Mathematics in the Sciences, 04103 Leipzig, Germany, [10]Institute for Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria, [11]Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá D.C., Colombia, [12]The Santa Fe Institute, 87501 Santa Fe, NM, USA and [13]Faculty of Health and Medical Sciences, University of Copenhagen, 2200 Copenhagen, Denmark

Received September 08, 2020; Revised December 08, 2020; Editorial Decision January 04, 2021; Accepted January 07, 2021

ABSTRACT

Animal models are crucial for advancing our knowledge about the molecular pathways involved in human diseases. However, it remains unclear to what extent tissue expression of pathways in healthy individuals is conserved between species. In addition, organism-specific information on pathways in animal models is often lacking. Within these limitations, we explore the possibilities that arise from publicly available data for the animal models mouse, rat, and pig. We approximate the animal pathways activity by integrating the human counterparts of curated pathways with tissue expression data from the models. Specifically, we compare whether the animal orthologs of the human genes are expressed in the same tissue. This is complicated by the lower coverage and worse quality of data in rat and pig as compared to mouse. Despite that, from 203 human KEGG pathways and the seven tissues with best experimental coverage, we identify 95 distinct pathways, for which the tissue expression in one animal model agrees better with human than the others. Our systematic pathway-tissue comparison between human and three animal modes points to specific similarities with human and to distinct differences among the animal models, thereby suggesting the most suitable organism for modeling a human pathway or tissue.

GRAPHICAL ABSTRACT



INTRODUCTION

Animal models play important roles in understanding human diseases. A main concern in using animal models for

*To whom correspondence should be addressed. Tel: +45 35 332204; Email: nadezhda.doncheva@cpr.ku.dk
Correspondence may also be addressed to Lars J. Jensen. Tel: +45 35 325025; Email: lars.juhl.jensen@cpr.ku.dk
Correspondence may also be addressed to Jan Gorodkin. Email: gorodkin@rth.dk

studying human diseases is the fundamental, but not clearly proven, assumption that the genes, pathways and diseases in model organisms are comparable to those of human. No systematic studies have tested this assumption at the tissue expression or pathway levels, although specific tissues or pathways have been compared (1,2). Even with closely related species, such as human and chimpanzee, it might not be trivial to identify the subtle differences in pathway regulation, which may be critical for disease modeling or drug design (3,4). Moreover, the most used animal models, often a rodent such as mouse or rat or an ungulate as the pig, are much more distantly related to human.

Model animals are extensively used to dissect underlying mechanisms of human diseases and to develop new treatments (5), but this is not trivial to do as recently demonstrated (6–9). For example, the most important class of drug metabolizing enzymes, the cytochrome P450 protein family, differs greatly between rodents and humans, both in terms of substrate specificity and multiplicity of the different cytochrome P450 subfamilies (10). For this reason, mice and rats are poor model organisms for testing the effects of drugs that undergo first-pass metabolism in the liver. By contrast, the cytochrome P450 protein family in pig represents a more promising model of human drug metabolism (11). Studies of diseases in animal models cannot be performed without first establishing the physiological pathway regulation in a specific tissue of the healthy animal. It is also important to know to what extent the pathway regulation is the same as in healthy humans, since complex diseases are usually associated with alterations in the activity of one or more pathways (12,13).

Even though many genes between human and model animals are highly similar in both sequence and function, their regulation and interplay can differ. While many databases and resources characterize genes in numerous species (14–16), most pathway annotation efforts focus on human, and most of those available for animal models are thus derived from human pathways (17,18). In the case of primary protein–protein interaction databases that contain experimentally determined physical interactions (19–21), very little data is available for animal models. On the other hand, integrative databases such as STRING (22,23) and IID (24) can provide more comprehensive annotations of the interplay between genes in animal models. These databases combine data from multiple resources, spanning interactions from the primary databases, text mining of biomedical literature, and orthology transfer from other organisms. However, because orthology transfer is used to construct such databases, it is not meaningful to subsequently compare human and animal pathways in order to identify similarities and differences between them. For that, organism-specific data on the pathways is needed, which is missing in current databases. Whereas organism-specific pathway annotations and interactions are scarce, expression data is available for many relevant model organisms.

We present the first systematic comparison of human and animal pathway activity for three specific model organisms (mouse, rat and pig), and aim to facilitate researchers in prioritizing animal models for human disease modeling. Key limitations in establishing disease-specific animal models include incomplete pathway annotation in animals and lack of knowledge of organism-specific pathway regulation. However, we show that it is possible to derive the animal pathways by orthology-based transfer of their human counterparts and study their regulation using organism-specific data such as gene expression. We thus map tissue expression data from healthy individuals onto the established derived animal pathways. Tissue expression data has already been compiled for human, mouse, rat, and pig through the TISSUES database (25). However, comparing human and the animal models is still challenging because of differences in the amount and quality of data available for the different organisms. Although we cannot identify animal pathways that deviate from the human version due to lack of organism-specific pathway data, we can detect similarities and differences in pathway activity between human and the animal models. We highlight several pathways, which, for a given healthy tissue, show better agreement in expression between human and one animal model but not the others.

MATERIALS AND METHODS

Genomes and gene annotations

Gene numbers for the genomes of human, mouse, rat and pig were extracted from the Ensembl release 95 websites (26) for each organism in January 2019 (Table 1). These correspond to genomes GRCh38 (human), GRCm38 (mouse), Rnor_6.0 (rat) and Sscrofa11.1 (pig). GENCODE numbers (27) were reported in Supplementary Table S1 for human GENCODE 30 (08.04.19) and mouse GENCODE M21 (08.04.19).

Orthology relationships

To identify the orthology relationships between the genes of the studied organisms, we made use of the public resource eggNOG v4.5 (evolutionary genealogy of genes: Non-supervised Orthologous Groups) (28). It provides >190 000 orthologous groups (OGs) of proteins for 2031 organisms at different taxonomic levels and is based on Ensembl release 70. We used the 26 253 OGs at the mammalian level (mOGs). When comparing two or more organisms, it is important to keep in mind that there are different sets of OGs that can be used. Thus, depending on the analysis, one or the other option can be more suitable: (a) OGs that have *at least one* protein for *any* organism, (b) OGs that contain *at least one* protein for *each* organism, (c) OGs with *exactly one* protein for *any* organism, (d) OGs with *exactly one* protein for *each* organism.

At the mammalian eggNOG level, there are 8665 mOGs, which contain exactly one protein for each of the four considered organisms (human, mouse, rat and pig), and 11 500 mOGs, which contain more than one protein for each of the four organisms. When we only require pair-wise relationships, the number of mOGs is larger (see Table 2). For the comparison of annotations and the pathway transferability, we used the mOGs with at least one protein for each organism in the compared pairs. For the pathway–tissue analysis, we focused on the set of 11 500 mOGs that contain at least one protein for each of the organisms.

Table 1. Data availability per resource and organism (human, mouse, rat and pig)

Resource/annotation	Human	Mouse	Rat	Pig
Ensembl genome annotations				
Coding genes	20 418	22 619	22 250	22 452
Non-coding genes	22 107	15 795	8 934	3 250
eggNOG mammalian orthology				
Coding genes assigned to orthologous groups	86.7%	84.3%	76.4%	82.2%
Mentions in biomedical literature				
Organism	–	1 824 080	1 629 280	133 937
Gene	–	1 304 170	734 243	57 230
Gene Ontology annotations				
Experimental	107 301	89 360	49 281	817
Author statement	48 894	4 760	3 396	27
Inferred	86 785	170 033	188 718	47 225
Electronic	74 049	44 022	40 559	101 074
High-scoring STRING protein-protein interactions				
Experimental	18 069	1 304	920	1 266
Experimental transferred	12 713	22 030	39 381	32 312
TISSUES expression data				
Experimental datasets	4	4	3	3
Tissues covered by experimental data	20	20	12	20

The number of coding and non-coding genes for the assemblies of human, mouse, rat, and pig in Ensembl release 95 are reported. From the eggNOG v4.5 orthology database, we report the percentage of genes from each organism that are assigned to a mammalian orthologous group. Text mining of all PubMed abstracts and a subset of full text articles available from PMC provided the number of publications that mention each organism and its genes. We grouped the most recent Gene Ontology annotations into four categories based on their evidence codes and counted the number of annotations for each group in each organism. High-scoring protein–protein interactions from the STRING v10.5 database (overall confidence score above 0.7) were counted. For the TISSUES 2.0 database of mammalian expression, the number of experimental datasets supporting the 21 main tissues is reported together with the number of tissues covered by these datasets. See Supplementary Table S1 as well as Methods for more details.

Table 2. Pair-wise overlap of annotations for human–mouse, human–rat, human–pig and mouse–rat

Resource/annotation	Human–mouse	Human–rat	Human–pig	Mouse–rat
eggNOG mammalian orthology				
Common 1-to-1 groups	12 736	11 038	10 916	12 157
Common groups	15 094	13 429	13 573	14 155
Gene Ontology annotations				
Experimental	15 215	4 680	133	3 754
Author statement	1 433	494	2	193
Inferred	50 473	39 740	18 751	64 879
Electronic	25 632	12 430	15 054	13 937
High-scoring STRING protein-protein interactions				
Experimental	537	72	672	66
Experimental transferred	10 422	6 705	9 290	10 859
TISSUES expression data				
Tissues covered by experimental data	19	12	20	12

For each of the selected resources, appropriate features are highlighted. For the eggNOG orthology resources, the number of common (1-to-1) mammalian orthologous groups are reported. For the Gene Ontology annotations and the high-scoring (overall confidence score above 0.7) STRING protein–protein interactions, the number of associations shared between a pair of organisms was determined using the eggNOG mammalian orthology. For the TISSUES 2.0 database, the number of tissues (out of the 21 main tissues) covered by an experimental dataset in both organisms are reported.

Mentions in biomedical literature

Our in-house text mining software tool called *tagger* runs every week on the whole corpus of more than 31 million PubMed abstracts and the Open Access subset of full-text articles available from PMC (29). In order to determine how often each organism and its genes are mentioned in the biomedical literature, we downloaded the corresponding files on 25 May 2020 from http://download.jensenlab.org/ (30). For each organism, we reported the number of unique PubMed entries (abstracts or full-text articles) in which the organism name occurs based on the file *organism_textmining_mentions.tsv*. To count the number of publications, in which genes of a specific organism occur, we downloaded the separate file for mouse, rat and pig (e.g. *mouse_textmining_mentions.tsv*), which contains a list of PubMed identifiers for each gene of this organism. In these files, a publication is assigned to a gene in an organism if both the gene and the organism were mentioned in the same publication according to *tagger*. In Table 1, we report the number of unique PubMed identifiers assigned to each gene. Since researchers often do not explicitly write in a publication that they study human specifically, and this would thus have to be inferred from the context, the numbers for publications mentioning human or human genes would be inaccurate. For this reason, we refrained from including these in Table 1.

Gene Ontology annotations

To investigate the coverage of functional annotations, all Gene Ontology (GO) annotations were retrieved on 25 May 2020 from the GO FTP server (31). Each GO annotation has a code assigned to it, which describes how it was determined. The codes were divided into four different groups: (i) *Experimental*: supported directly by an experiment, including high-throughput methodologies (EXP, HAD, HEP, HMP, IDA, IEP, IPI, IGI, IMP); (ii) *Author statement*: based on statements by the authors in the cited reference (TAS, NAS); (iii) *Inferred*: derived from phylogenetic (IKR, IBA) and computational (ISS, RCA, ISO, ISA, ISM) analysis as well as inferred by curators (IC); (iv) *Electronic*: automatically generated and not reviewed electronic evidence (IEA). For each organism and group, we counted the number of unique pairs of GO terms assigned to a protein (see Table 1). The overlap of GO term annotations between pairs of organisms was determined by mapping the annotated genes to their corresponding mOGs and determining the intersection of pairs of GO term and mOGs between the organisms (see Table 2).

STRING protein–protein interactions

Protein–protein associations were retrieved from STRING, a database of known and predicted protein–protein interactions (22). Raw STRING data (divided into orthology-transferred and original interactions) was downloaded from STRING v10.5. Each file contains the interacting genes (ENSEMBL IDs) and the confidence for each evidence (between 0 and 1), whereas evidences are divided into the original data and the *transferred* interactions (by orthology). To compare the available, high-confidence, not predicted interactions, we considered the *experimental* and *experimental transferred* interactions for each organism. For each interaction type, we counted the number of interactions that have a confidence score ≥ 0.7 (Table 1) or ≥ 0.4 (Supplementary Table S1). The overlap of interactions between each pair of organisms was determined by mapping each interacting gene to the corresponding mOG. An interaction was considered overlapping between two organisms if the pairs of interacting genes were in the same mOGs.

Tissue expression data

For this analysis, we used data from the TISSUES database, which contains gene–tissue associations for human, mouse, rat and pig (25). The database integrates multiple sources of evidence, including transcriptomics data covering all four species, proteomics data only for human, manually curated annotations from UniProt and associations mined from the scientific literature. Importantly, the expression data has been processed such that it is comparable across all sources of evidence and across organisms through a scoring scheme. For each gene–tissue association in each organism, there is an integrated confidence score based on all evidence types. For consistency, the tissue evidence is further summarized into tissue labels, which are based on Brenda Tissue Ontology terms (32).

We retrieved all gene–tissue associations with their corresponding experimental and integrated confidence scores on 26 January 2018. From the 21 tissues, we focused on the seven tissues that are covered by at least two transcriptomics datasets: heart, kidney, liver, nervous system, muscle, lung and spleen. Even though the TISSUES database provides unified confidence scores, the amount and quality of available tissue data varies a lot between organisms due to, for example, study bias. This influences the range of confidence scores as can be seen in the distributions of confidence scores for each tissue (Supplementary Figures S1 and S2). In order to define whether a gene is expressed or not, we calculated organism- and tissue-specific cutoffs based on the 50 percentile of confidence scores (median) for each organism and tissue (Supplementary Table S2). We specifically used the percentile instead of a fixed cutoff such that we have a comparable number of genes for each tissue and organism irrespective of the differences in the distribution of the scores. Furthermore, we chose exactly the 50 percentile in order to better approximate the expected number of expressed genes in a given tissue (33). For completeness, we also performed the analysis using the 25, 40, 60 and 75 percentiles of confidence scores as cutoffs (see Supplementary Results).

Orthology-based pathway transferability

KEGG is one of the most well-known and widely used pathway databases (18). It contains manually drawn pathway maps representing molecular interaction and reaction network diagrams. For our analysis, we obtained the set of 216 human KEGG pathways from the STRING v10.5 KEGG benchmark dataset (22). Pathways with less than five genes matched to OGs in either organism were omitted from the analysis, which resulted in a set of 205 human KEGG pathways.

To assess the transferability of each of these pathways from human to another organism, we used the eggNOG mammalian orthology. For each pathway, we mapped each human gene in this pathway to the mOG it belongs to and thereby converted the pathway–gene association to a pathway–OG association. Then, the pathway transferability from human to another organism was calculated as the proportion of pathway genes in the other organism that have orthologs in the same OGs that contain the human genes. This means that a limiting factor of the pathway transfer is the number of mOGs shared between human and the respective organism. In the special case of pathway transferability from human to *all* three analyzed organisms, we only considered the 11 500 mOGs that cover all four organisms.

Integration of tissue expression data

Given the set of human and orthology-transferred pathways and the tissue expression data from TISSUES, we performed a pathway–tissue analysis, in which we considered for each organism, which pathway genes are expressed in each tissue and compared these among the four analyzed organisms. For each organism, for each tissue and pathway, we calculated the fraction between all pathway genes expressed above the chosen confidence cutoffs (from here on called *expressed pathway genes*) and all genes with any expression information in this pathway (Supplementary Tables S3 and S4). When at least 85% of the orthologous

pathway genes with tissue information were above the chosen tissue confidence cutoff, we considered these pathways *expressed* in the given tissue and organism. We chose the cutoff of 85% after inspecting the proportion of expressed genes in several known pathways (*Citrate cycle (TCA cycle)*, *Spliceosome*, *Ribosome*, *Proteasome*, *Oxidative phosphorylation* and *Propanoate metabolism*) such that these pathways were expressed in most tissues (Supplementary Figures S3–S6). Note that there is a connection between this cutoff and the TISSUES confidence cutoff, which we chose for defining whether a gene is expressed in a tissue or not. The fewer genes are considered as expressed, e.g. only the genes in the 75 percentile of confidence scores, the lower we need to put the cutoff for an expressed pathway (for example to 75%) to have an appropriate result (see Supplementary Results and Supplementary Figures S7–S10).

To compare between organisms, we used the eggNOG mOGs that contain at least one orthologous gene for each of the considered organisms. The TISSUES expression confidence scores of two genes from different organisms were considered comparable, if these two genes belong to the same mOG. If several genes from the same organism belong to the same mOG, the highest confidence score was used in the comparison. In addition, we omitted pathways with less than five genes with expression information in either organism, which resulted in 203 pathways suitable for analysis. Note that this number is slightly lower than the 205 transferred pathways due to the restriction of mOGs to have coverage in all four compared organisms.

Comparison of tissue–pathway combinations across organisms

In order to compare the pathway expression in each tissue between human and the model organisms, we computed the Jaccard index (JI) for each tissue–pathway combination (203 pathways and seven tissues). We defined the JI for a given pathway and tissue between two organisms as the intersection of expressed pathway genes of the two organisms divided by the union of expressed pathway genes in any of the two organisms. A pathway gene is considered expressed if it has a TISSUES confidence score above the chosen organism- and tissue-specific cutoff.

The principal component analysis (PCA) on the JIs for the comparison of human–mouse, human–rat and human–pig was computed using the *scikit-learn* Python package (34) for all pathway–tissue pairs with at least five pathway genes expressed in the given tissue. We used PCA not to reduce the dimensionality but purely as a visualization technique. The loadings for each of the considered variables, which correspond to the JIs for each pair of compared organisms, were computed as the product of the PCA component and the square root of the explained variance for each principal component. Based on the plot of principal components 2 (PC2) and 3 (PC3), we identified a set of pathway–tissue pairs, which are more consistent between human and a specific model organism, by calculating the Euclidean distance of each pathway–tissue pair to the center of the PC2 & PC3 plot. Based on their distance to the loadings, we also grouped the pathway–tissue pairs into six different groups: *mouse, rat, pig, mouse & rat, mouse & pig, rat & pig*. These groups contain pathway–tissue combinations, for which one or two of the model organisms agree more with human than the other(s).

RESULTS

Several limitations and possibilities about modelling human pathways in animal models arise from publicly available data for the two well-established animal models mouse and rat as well as for the emerging one, pig. The quality of genome assemblies and orthology mapping between organisms have significantly improved in the last years, and increasingly more tissue expression data is becoming available. In contrast, annotations in terms of functions, pathways, and protein interactions are still lacking high-quality experimental data to allow detection of differences between animal models and human. Therefore, we derive animal pathways from curated human pathways using reliable orthology relationships and further integrate these pathways with tissue gene expression data from the animal models. Despite the better coverage and quality of data in mouse and human as compared to rat and pig, we can identify several pathways in specific tissues that agree better with human in one animal model compared to the other two.

Available functional annotations for animal models are limited compared to human

For the systematic comparison of animal models to human, we need to answer the following important questions: Which data is publicly available and what is the quality of this data? Here, we analyze and compare the following resources: Ensembl for quality of genome assembly and gene annotation (26), eggNOG for orthology relationships (28), text mining of genes and organisms in the biomedical literature, Gene Ontology (GO) functional annotations of the genes (31), the TISSUES database for gene expression (25), and the STRING database of known protein interactions (22). For each of these resources, representative numbers are listed in Table 1 (for further details, see Supplementary Table S1). The amount of data available varies greatly across resources and organisms; for example, mouse is very well covered by most resources, while rat and pig are covered to a lesser extent and their coverage is different for the different resources. Each resource's content and limitations are presented in more detail below.

Genome annotations. Based on the overall statistics available from Ensembl release 95 for each of the most recent assemblies, we conclude that there is good annotation of coding genes, while annotation of non-coding RNA genes still needs improvement, especially for rat and pig (Table 1). The corresponding numbers of coding and non-coding genes for human and mouse in GENCODE (27) are very similar to the ones in Ensembl (see Supplementary Table S1). A comprehensive genome quality assessment of human and 20 domesticated animals was performed by Seemann *et al.* (35). At that time, the mouse assembly ranked very high based on its quality as opposed to pig and many of the other animals' assemblies.

eggNOG mammalian orthology. To compare annotations for human, mouse, rat, and pig, we used the eggNOG database 4.5.1, which provides orthologous groups (OGs) at different taxonomic levels. We chose the OGs at mammalian level (mOGs) as they are the most fine-grained OGs that contain all four organisms of interest. The number of genes of each organism that are assigned to mOGs is given in Table 1. The coverage is best for human (86.7%), closely followed by mouse (84.3%) and pig (82.2%), whereas rat has only 76.4%. From the 26 253 mOGs in eggNOG, 11 500 cover all four organisms, e.g. contain at least one gene from each organism, and 8665 of them have one-to-one orthologs for human and the three animal models.

Biomedical publications. To approximate the popularity of the three model organisms of interest, we counted how often they are mentioned in PubMed entries (abstracts and publications) using the *tagger* text-mining software (29), which also generates the text-mining associations for the STRING database (23). We considered two different measures: (i) how many PubMed entries mention the organism itself and (ii) how many PubMed entries mention a gene from this organism. The latter number is based on identifying both the organism and the gene in the same PubMed entry and allows us to distinguish publications, which discuss the animal models, especially pig, in connection with veterinary treatments, from the publications, which actually study the molecular biology of the organisms as given by the mentions of genes. As shown in Table 1, mouse and rat are mentioned at least 10 times more often than pig. However, 72% of the entries that mention mouse and only ∼45% for rat and pig appear to specifically study their genes.

Gene Ontology annotations. The Gene Ontology is one the most used resources for functional annotation of genes and proteins. GO annotations can be supported by different types of evidence, including experimental, author statements, computationally inferred such as based on phylogeny, as well as non-curated electronic annotation. For each of these categories, we listed the number of annotations for each organism. As for other resources, there is an imbalance between the different types of evidence and the different organisms, human having most annotations with experimental (107 301) or author statement support (48 894). By contrast, even mouse has a huge proportion of annotations inferred computationally (170 033), in addition to many experimentally supported ones (89 360). For rat, and especially for pig, most annotations are supported only by computationally inferred evidence or electronic annotations (Table 1).

Protein–protein interactions. To assess the availability of molecular interaction data for each organism, we counted the high-confidence (confidence score $\geq$ 0.7) experimental protein–protein interaction data in STRING v10.5. The lack of such data in most considered organisms is evident with only ∼900 interactions for rat and ∼1300 for mouse and pig, which is surprising considering how well studied mouse and rat are (e.g. as indicated by their mentions in the literature). In all four organisms, however, many protein–protein interactions can be transferred by orthology from the other organisms in STRING (Table 1, *experimental transferred* interactions) due to the good quality annotation of protein-coding genes.

Tissue expression data. An important aspect of studying and comparing animal models is the availability and accessibility of tissue expression data. TISSUES 2.0 integrates evidence on mammalian tissue expression from manually curated literature, proteomics, and transcriptomics screens, and automatic text mining. The numbers in Table 1 clearly demonstrate that only a few large-scale experimental datasets cover several tissues. This is especially the case in pig and rat, which generally have poor coverage in terms of tissue expression data. Having sufficient experimental evidence is also a challenge for less studied tissues in human as shown in Supplementary Table S1 and by Palasca *et al.* (25).

Annotation similarity between organisms is mainly determined by data availability

Although the available functional annotations for animals are limited compared to human, it is still possible to perform a direct comparison between human and the animal models. Our goal is to assess the extent to which the overlap is driven by data availability, as opposed to evolution. Thus, we determined the pairwise overlap of annotations between human and the three model animals and compared these to the overlap between mouse and rat (see Table 2). Assuming data with good quality and coverage, we would, due to the evolutionary relationship, expect the agreement between mouse and rat to be better than between human and mouse. However, it appears that the difference in data availability between organisms impacts the overlap more than the evolutionary relationship does.

For one-to-one orthology mapping we get comparable numbers for each pair of organisms, reflecting the good annotation quality of protein-coding genes in all four genomes. This is also the case when broadening the orthologous groups to contain one-to-many and many-to-many orthology assignments (*common groups*). Using the mammalian-level orthology assignments to compare between organisms, we further analyzed GO annotations based on experimental and author statement evidence type. We observe by far the highest overlap between human and mouse, reflecting that these are the two most studied organisms. The high similarity of inferred GO terms between mouse and rat can be attributed to database curators annotating GO terms based on sequence similarity to the same experimentally characterized human genes (14,16). Finally, the GO terms in pig come from inferred or electronic annotations (Table 1), which is reflected in the large overlap between human and pig in these categories.

When comparing protein–protein interactions from STRING, we observe that the overlap of *experimental* interactions is more heavily influenced by the availability of data than is the overlap of *experimental transferred* interactions. The small number of overlapping experimental interactions between rat and both human and mouse is in part explained by the fewer experimental rat interactions (Table 1). In the case of *experimental transferred*, we see a good overlap of ∼10 000 interactions for human with mouse and pig as well

as between mouse and rat, and a somewhat smaller, but still considerable overlap of 6705 interactions between rat and human.

For the analysis of the tissue expression data, we consider the number of tissues and their coverage by experimental datasets. There are at least 19 tissues covered by *at least one* experimental dataset for each of the pairs human–mouse, human–pig and mouse–rat but only 12 tissues for human–rat, which is consistent with the poor tissue coverage for rat (Table 1). In the case of pig, there is data for 20 of the tissues, however, for 13 of them, the evidence originates only from one experimental dataset. As a result, there are only seven tissues (heart, kidney, liver, nervous system, muscle, lung, spleen) that are covered by *at least two* datasets in all pairs of organisms.

In conclusion, we observe that the extent to which the available annotations overlap between pairs of human and animal models depends more on data availability than on how closely related the organisms are. Given the current data, mouse is better annotated than pig and rat and thus has a better overlap with human than with rat.

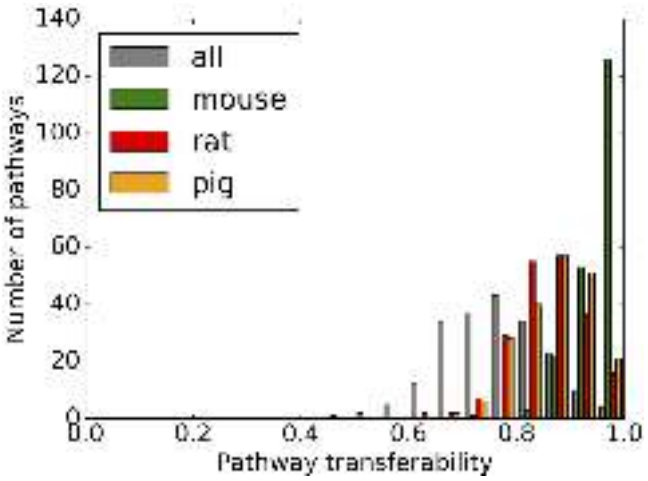


Figure 1. Transferability of 205 KEGG pathways from human to mouse, rat, and pig. Each bar represents the number of pathways, for which a given fraction of genes can be transferred. Transferability to each organism is shown in different colors: mouse in green, rat in red, pig in orange and all organisms in grey.

Quantification of orthology-based pathway transfer from human to animal models

Our observations so far provide an estimate of how well human and animal models are covered and agree with each other for individual resources and types of annotations. However, this comparison is limited to individual genes or at most, pairs of genes in the case of STRING interactions. Here, we analyze how consistent pathways can be in terms of their gene content at the organism level.

Most pathway databases focus their annotation efforts on human and thus, even for popular model organisms such as mouse and rat, contain only very few experimentally determined and curated pathway interactions. Instead, they resort to using orthology transfer from human to derive pathways for other organisms. This is the case for popular pathway databases such as Reactome (17) and KEGG (18). Similarly, integrative protein interaction databases such as STRING (23) and IID (24) include orthology transfer of interactions as an information source. However, the exact methodology of pathway transfer differs between databases and even between different organisms within the same database. This can easily cause inconsistencies both between and within the databases, which makes it very difficult to make a meaningful pathway comparison of the organisms.

To meet these challenges, we started with a set of human KEGG pathways and, for each of these pathways, we assessed how well it can be transferred to mouse, rat and pig using the eggNOG orthology relationships within the mammalian taxonomic level. We quantified the *transferability* of each pathway from human to a model organism as the fraction of genes in the human pathway that could be transferred to the model organism in question (Figure 1).

The limiting factor in this comparison is the number of orthologous groups that contain genes from both organisms being compared; this number is lower for rat and pig than for mouse (Table 2). Thus, the pathway transferability from human to mouse has the highest coverage, in terms of both the number of pathways and the number of genes within a pathway. Out of the 205 considered human pathways, 55 can be transferred to mouse completely, compared to only 8 to rat and 12 to pig. The distribution of pathway transferability for mouse has a mean of 95% and ranges between 73% and 100%. In contrast, the distributions of rat and pig are shifted to lower transferability values with a mean of 85% and 87% (minimum of 61% and 65%), respectively.

We are also interested in how well a human pathway can be transferred to all selected animal models at the same time, i.e. what is their overlap (grey colored bars in Figure 1). In this specific case, we consider only orthologous groups that contain at least one gene from each of the four organisms. Overall, there is a good agreement between human and the three animal models with a pathway transferability range between 48% and 100% and a mean of 76%. This range means that for all pathways (but one), more than half of the pathway genes can be transferred from human to mouse, rat, and pig.

The pathways that can be transferred best to all organisms are mainly pathways in the KEGG categories *Metabolism, Replication and repair* or *Human diseases.* A complete list of pathways and number of genes transferred for each of them for each model organism can be found in Supplementary Table S5. For example, the largest pathways among the ones, which are 100% transferable between human and mouse, include *Prostate cancer* (81 genes), *TGF-beta* (67 genes) and *Adipocytokine signaling* pathways (61 genes), for rat they belong to the *Glycan biosynthesis and metabolism* KEGG subcategory (15 pathway genes on average), and for pig they include *RNA polymerase* (29 genes), *Mismatch repair* (22 genes) and *Steroid biosynthesis* (19 genes). The least transferable pathways (50–60% of genes transferred) relate to the nervous system (*Long-term potentiation* and *Dopaminergic synapse*), certain signaling pathways (such as *Notch* and *VEGF signaling* pathways), and the *Circadian rhythm* pathway. The latter is consistent with rat and mouse being nocturnal animals.

Detection of tissue-specific and broadly-expressed pathways through data integration

In the last couple of years, more and more healthy tissue expression data for animal has been produced and made available in public repositories (36–39). However, this data is difficult to compare across datasets or organisms. Thus, in a previous study, we introduced the TISSUES database (25), which contains tissue expression evidence for human, mouse, rat, and pig in the form of confidence scores that are designed to simplify comparison across datasets and organisms.

In order to explore the differences between the human and animal models at both pathway and tissue level, we integrated the orthology-transferred KEGG pathways with expression data from TISSUES for each organism. We define a pathway gene to be *expressed* in a tissue if it has a confidence score above the chosen cutoff (see Methods for details). In this analysis, we considered the confidence scores based on experimental evidence for seven tissues with good coverage, i.e. at least two experimental datasets available for each organism. We also performed the analysis using the scores that integrate all types of evidence in the TISSUES database as well as different cutoffs for the tissue confidence scores (Supplementary Results and Supplementary Tables S12 and S13).

Out of the 205 pathways, we analyzed only those containing at least five orthologous genes for each of the four compared organisms, resulting in a set of 203 KEGG pathways. Supplementary Table S3 provides the complete list of pathways and, for each of them, the proportion of genes expressed in each tissue and organism. On average in all tissues and organisms, 59.7% of the pathway genes are expressed (Supplementary Table S4). In human, the average across tissues and pathways is 62.1%, while for mouse it is 59.9%, for rat 57.7% and for pig 59.3%. Tissue-wise, we observe that the most pathway genes are expressed in the kidney and liver (>62%), closely followed by lung and spleen with ∼62%, and the least are expressed in the nervous system, heart and muscle (∼57%).

Furthermore, we specifically considered the almost completely *expressed* pathways, which we define as those having at least 85% of the orthologous pathway genes expressed in a specific tissue for each organism (Table 3 and Supplementary Figure S4A). Note that the number of expressed pathways for each tissue and organism is affected both by the requirement of 85% pathway genes as well as the tissue confidence cutoffs, which were chosen such that only the genes with a confidence score above the median value for each tissue were considered expressed. We also performed a more detailed analysis on the connection between these two cutoffs and the robustness of the findings using different cutoffs (see Supplementary Results). Overall, the numbers vary among tissues and organisms, but there are some specific trends. For example, liver has the highest number of expressed pathways (between 29 and 37) in all four organisms, followed by kidney with 26 expressed pathways in human, 25 in mouse and rat, and 19 in pig. For the remaining tissues, we observe a range between 8 and 16 expressed pathways depending on the specific organism and tissue.

Table 3. Number of pathways expressed in each tissue and organism

Tissue/organism	Human	Mouse	Rat	Pig
Heart	12	16	14	8
Kidney	26	25	25	19
Liver	**36**	**37**	**29**	**32**
Lung	11	12	13	14
Muscle	12	14	13	12
Nervous system	12	15	10	12
Spleen	12	13	10	13

A pathway is considered expressed if 85% of the pathway genes are above the chosen tissue confidence cutoff. The analysis was done on the 203 human KEGG pathways and the same number of transferred pathways for mouse, rat and pig using the experimental confidence scores from TISSUES for the seven tissues with support by at least two experimental datasets. The highest number of pathways for each organism (each column) is indicated by a bold font.

Based on the number of tissues, in which a pathway is expressed (Supplementary Figure S5A and Supplementary Table S6), we can divide the pathways into broadly expressed and tissue-specific pathways. The *Citrate cycle (TCA cycle)* is an example of a broadly expressed KEGG pathway with >92% of pathway genes expressed in each organism in all seven tissues (except for lung in mouse) as shown in Supplementary Figure S6A. By contrast, the *Axon guidance* KEGG pathway (Supplementary Figure S6B) is – not surprisingly – much more expressed in the nervous system in all organisms (average of 63%) compared to all other tissues (average of 47% over tissues and organisms). Among the 203 pathways, we find 16, 18, 16 and 13 to be expressed in at least three tissues in human, mouse, rat and pig, respectively (Supplementary Table S7). Of these, 10 pathways are expressed in at least three tissues in all four organisms, namely *Citrate cycle (TCA cycle)*, *Spliceosome*, *Ribosome*, *Proteasome*, *Oxidative phosphorylation*, *Protein processing in endoplasmic reticulum*, *Propanoate metabolism*, *Pyruvate metabolism*, *2-Oxocarboxylic acid metabolism* and *Valine, leucine and isoleucine degradation*.

Evaluation of pathway–tissue agreement between human and animal models

To evaluate which of the pathways expressed in human tissues agree with those in mouse, rat and pig, we assessed how many genes from a pathway are expressed in the same tissue for each pair of organisms (human–mouse, human–rat, human–pig). For each tissue and pathway, we calculated the Jaccard index (JI) as the overlap of expressed pathway genes divided by the union of all expressed pathway genes. A pathway gene is considered expressed if it has a TISSUES confidence score above the chosen cutoff. As a result, for each pathway–tissue combination, we have three JIs of how well this pathway agrees between human and one of the model organisms in the given tissue (Supplementary Table S8).

The average JI over all tissues between human and mouse is 0.63, followed by 0.62 for human–pig, and 0.60 for human–rat (Supplementary Table S9). When we compare the average agreement (over all pathways) between human and the model organisms for each tissue separately, the liver stands out as the tissue with the best agreement for all comparisons, while the remaining tissues are ordered differently

depending on the model organism. For example, the tissue with the lowest average JI for the comparisons human–mouse and human-pig is heart (JI of 0.61 and 0.56, respectively), while, for rat, both lung and muscle have the lowest JI (0.57). The distributions of tissue-wise JIs for each of the three comparisons are shown in Supplementary Figure S11 and further confirm that the agreement between the organisms can strongly vary between the tissues.

To further analyze the similarities and differences between human and the three model organisms on pathway–tissue level, we performed a principal component analysis (PCA) on the JIs for all pathway–tissue pairs, where at least 5 pathway genes are expressed in the given tissue (data shown in Supplementary Table S8). We also plotted the PCA loadings, which show the weight that each model organism has in each principal component (Figure 2). Principal component (PC) 1 accounts for the most variability of the data (82.5%) and highlights the difference between pathway–tissue combinations with high JI and those with low JI, capturing the general agreement between human and all the animal models. From the 34 pathway–tissue combinations that are right-most according to PC1 (PC1 > 0.5), 23 are broadly expressed house-keeping pathways, such as *Citrate cycle* and *Proteasome*. Of those, the highest number of pathways is associated with liver tissue and none of them with the lung. The 43 left-most pathway–tissue pairs according to PC1 (PC1 < -0.5) are mostly small pathways (average size of 8.6 human genes) with low JI (average JI of 0.24 for rat and 0.3 for mouse and pig) and they are distributed across all tissues.

Based on the PCA analysis, PC2 and PC3 separate the three animal models from each other with explained variance of 10.3% and 7.1%, respectively. The loadings of PC2 separate pathway–tissue pairs that show good agreement between human and rat, but poor agreement between human and pig from those showing the opposite behavior. Meanwhile, the loadings of PC3 separate the pathway–tissue pairs based on whether they specifically (or specifically not) show better agreement between human and mouse. We thus used the PC2 & PC3 plot to further identify combinations of pathways and tissues, for which the agreement between human and one of the model organisms is better than with the others. Pathway–tissue pairs that are close to the center of the PC2 & PC3 plot show consistent agreement between human and any of the model organisms; this agreement can be consistently good, if all the JIs are high, or consistently bad, if the JIs are low. In contrast, any point that is very far from the center of the PC2 & PC3 plot represents a specific pathway–tissue combination, for which one of the model organisms has better agreement with human than the others (for a distance distribution see Supplementary Figure S12A). If such a pair is close to one of the three loadings, we assign it to that model organism (mouse, rat, or pig) and consider this pathway–tissue pair to agree more with human and the model organism given by the loading than the other two model organisms (see Supplementary Figure S12B and Methods section). In the cases, where a pathway–tissue combination is between two loadings and furthest away from the third one, we assign it to a category shared between two of the model organisms: mouse & rat (opposite of pig), mouse & pig (opposite of rat), and rat & pig (opposite of mouse). We observed consistent results for the top 100, 200 and 500 pathway–tissue pairs, which are located furthest away from the center (Supplementary Table S10). We observed similar trends when applying the analysis on all data in the TISSUES database (Supplementary Figure S13) and when varying the tissue confidence cutoff (Supplementary Figure S14). Here, we show and discuss only the top 200 in more detail (Table 4).

Among the top 200 in the three compared organisms (Supplementary Table S11), pig has the largest number of 41 pathway–tissue pairs, which show agreement with human for this organism only. The 24 pathway–tissue combinations, which are more consistent between mouse and human, have an average JI of 0.72 and cover all analyzed tissues. Specifically, lung tissue has the largest coverage of seven pathways, while spleen has the lowest with one. For the 35 pathway–tissue pairs, for which rat specifically agrees more with human, the average JI is 0.74 and the largest number of nine pathways is associated with spleen, while the lowest of one with muscle. The 41 pairs, which show good agreement specifically between pig and human, have an average JI of 0.72 and also cover the seven analyzed tissues. Whereas only two of these pathways are associated with nervous system, 10 are with muscle.

Half of the top 200 pathway–tissue combinations were not assigned to one specific organism, but instead to two organisms, which show similar, higher consistency with human than the third organism does. Mouse & rat have the highest number of 44 such shared pathway–tissue pairs, with 10 pathways assigned to heart, nine to muscle and only two to spleen. Since mouse and rat are closely related to each other, this result is not surprising. The high number of 36 pathway–tissue pairs, for which mouse & pig are consistent with human, comes as a close second. Out of these, lung has the largest coverage of 12 pathways, while heart and spleen are represented only by two and three pathways, respectively. For the 20 pathway–tissue pairs, for which rat & pig agree with human more than mouse, six are associated with nervous system, and only one pathway is selected for heart, kidney, and muscle.

Overall, the pathway–tissue pairs that show distinct consistency between human and certain model organisms distribute as follows. The two biggest groups are the 41 pathway–tissue pairs unique to pig and the 44 shared by mouse & rat, i.e. not seen in pig. Of these, 19 pathways relate to muscle. For pig, these pathways are in the KEGG category *Organismal Systems*, while for rodents they are in the *Metabolism* category. The next two groups are those unique to rat (35 pathways) and those shared by mouse & pig (36 pathways), i.e. not seen in rat. For the latter, lung stands out by the highest number of 12 pathways. Finally, the liver stands out as the tissue with the most expressed pathways and the one for which all three animal models agree equally well with human.

Although there are limitations due to varying data availability for each organism, our findings indicate that we can successfully approximate the tissue-specific pathway activity and identify similarities and differences between the three considered model organisms and human.

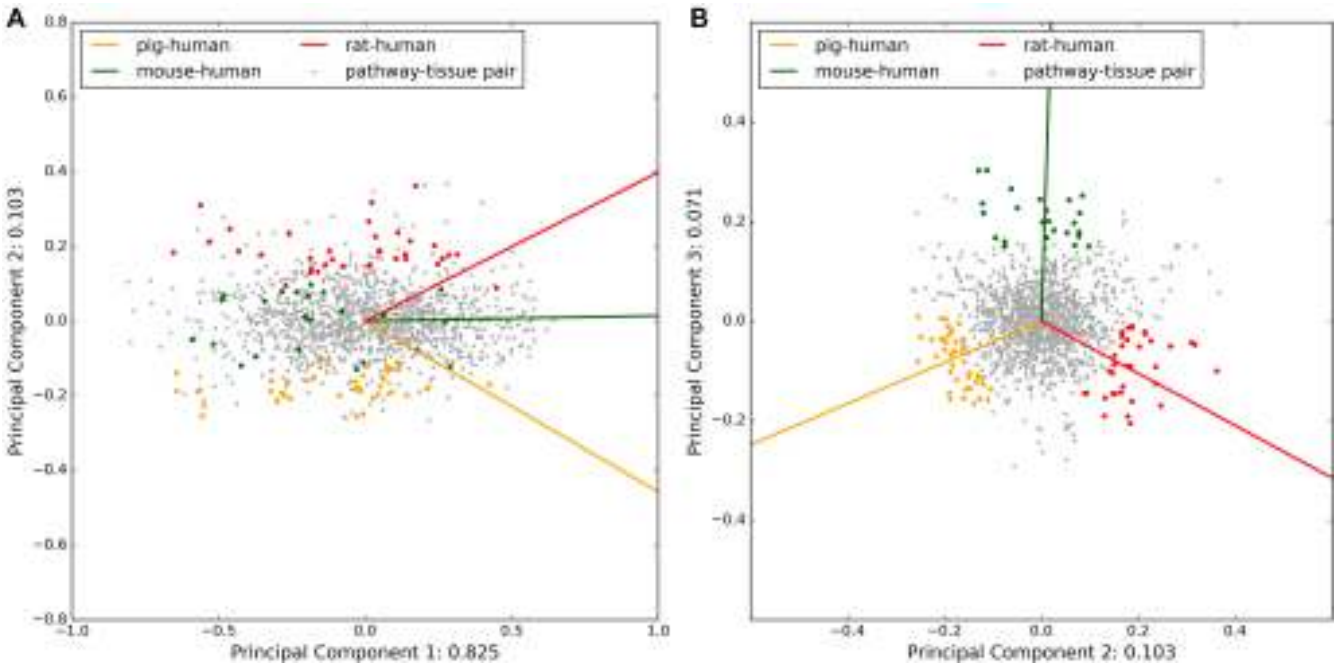


Figure 2. Principal component analysis of the pathway–tissue agreement between human and animal models. PCA was performed on the Jaccard indices (JIs) for all pathway–tissue pairs (grey dots) with at least five expressed pathway genes, where the JIs represent the comparisons human–mouse, human–rat, and human–pig. The PCA loadings are shown as solid lines and colored by the model organism responsible for their direction. While the PC1 & PC2 plot (panel **A**) shows a clear separation between pathway–tissue pairs with high JI and thus, good agreement between human and the respective animal model, the PC2 & PC3 plot (panel **B**) clearly separates the data based on the differences between the animal models. The pathway–tissue pairs located closest to each loading and furthest away from the center of the PC2 & PC3 plot are colored in the same color as the organism loading to indicate that for these pairs, this organism agrees more with human than the others (see Methods section for more details).

Table 4. Top pathway–tissue combinations showing distinct agreement between human and a model organism

	Mouse	Rat	Pig	Mouse & rat	Rat & pig	Mouse & pig
Pathway–tissue pairs	24	35	41	44	20	36
# pathways	20	28	34	34	17	27
Average JI	0.72	0.74	0.72	0.69	0.62	0.70
# pathways by tissue						
Heart	4	7	5	**10**	1	2
Kidney	3	8	4	8	1	5
Liver	2	3	6	6	2	4
Lung	**7**	2	8	4	4	**12**
Muscle	4	1	**10**	9	1	4
Nerv. system	3	5	2	5	**6**	6
Spleen	1	**9**	6	2	5	3

This table shows the top 200 pairs, while Supplementary Table S10 gives an overview of the top 100 and 500 pathway–tissue pairs. The first three columns indicate the numbers of pairs, for which one of the model organisms (mouse, rat or pig) is specifically more consistent with human than the other two. The last three columns refer to the pairs, which are shared between two model organisms and thus consistent with human in a similar way for both organisms. The average Jaccard index (JI) for all pathway–tissue pairs assigned to a group is also listed. For each tissue row, the number of (#) pathways assigned to this tissue is listed. Numbers shown in bold indicate the tissue covered by the highest number of pathways for each column.

DISCUSSION

To summarize our observations, there is an abundance of both experimental and inferred information with good quality for human, including genome quality, orthology relationships, biomedical literature, tissue expression data, gene annotations, and protein associations. Unfortunately, the same is not the case for mouse, rat or pig. While mouse is very often mentioned in the literature and is well covered by tissue expression data and GO annotations, there are very few experimentally determined protein associations reported for it. Meanwhile there is a shortage of most types of annotations and data for both rat and pig. Thus, one of the biggest limitations of the current analysis is the availability of public data for model organisms. This can be improved in the future by encouraging researchers to make publicly available more experimental, curated, high-quality findings generated for organisms other than human, especially when these organisms are popular model animals such as mouse and rat.

Pathway transferability, both in our study and in pathway databases in general, is limited by the data availability and agreement of the individual resources, in particular, the amount of pathway annotations and the quality of

orthology relationships. Nevertheless, the pathway transfer from human works very well for mouse (95% on average) and fairly well for rat or pig (85% and 87% on average, respectively). The pathway transferability also highlights the extent to which the animal models agree with human at a pathway level given the available data. Due to the lack of organism-specific information on pathways, we are not able to detect more pronounced differences between the organisms by using only this type of data. Ideally, we would like to have one single resource with pathways that are curated separately for each organism. Using it would allow us to identify the specific parts of the pathways that are only present in the animal model but not in human. Unfortunately, this is not possible with the current pathway and interaction databases, even though individual resources such as the Mouse and Rat Genome Database (14,16) try to collect and provide organism-specific data. Therefore, we are in practice forced to think of the human curated pathways as more general representations of what is happening in any tissue. We then try to approximate how these pathways behave in specific tissues or model organisms through integration of other types of data such as tissue expression.

The availability of organism-specific tissue expression datasets is considerably better, although still far from ideal. The deposited datasets often come from one individual and tissue and only sometimes cover several tissues in the same individual(s). This gap has been decreasing lately as more and more high-throughput sequencing data is being generated and deposited in public repositories for human (36–38) and farm animals, including pig (39). For example, several large-scale sequencing and annotation efforts have been undertaken by the FAANG (Functional Annotation of ANimal Genomes) consortium with the goal to improve the functional annotation of animal genomes, including pig, goat, sheep, cattle, horse, and chicken (40). These efforts will improve both the genome quality and gene annotation. However, a limitation is still the need to update resources based on the new genome assemblies, which does not always happen quickly enough. Furthermore, there is a clear need for more resources like the TISSUES database, which can calibrate the data from the different technologies and organisms and make it comparable. This also means that the analysis performed here can be significantly improved in the future once more and better quality data becomes available. Another possibility would be to extend the current analysis to include other less popular model animals.

Another important aspect and possible limitation of our analysis is the comparability between organisms. Most importantly, we need well defined orthology relationships between the compared organisms. Identifying orthology between species has improved over the years (41) and allows us to compare even organisms, which are evolutionarily more distant (42). However, orthology assignments are still heavily influenced by the quality of the underlying genomes and their annotations. In our case this means that some of the orthology relationships between human and pig or human and rat might be missing due to the annotation quality of these genomes at the time when the orthology resources were constructed and updated. This lack of complete orthology relationships influences both the pathway transferability and the extent, to which the organisms agree with each other at pathway–tissue level, and thus only allows us to see part of the whole picture now. However, with better annotated genome assemblies and improved orthology, we expect that our framework will reveal an even more complete picture of the similarities and differences between human and different animal models.

The comparison of individual types of data between the selected four organisms indicates that the observed agreement is more driven by the availability of data than by evolutionary relationships. This is likely due to the lack of organism-specific data at various levels. However, we also showed that through integration of pathway data with tissue expression, we can identify both similarities between human and the model organisms and differences with respect to how well the animal models agree with human at a pathway–tissue level. The resulting pathway–tissue–organism associations revealed both expected and unexpected findings as mentioned previously. For example, so-called house-keeping pathways consist primarily of genes that we see expressed in most tissues and organisms, while other pathways were found to be much more tissue-specific. In terms of pathway–tissue differences between the organisms, all tissues except for the liver were associated with more pathways, for which only one or two, but not all three model organisms were consistent with human. With respect to the question, which of these animal models is best suited for modelling a human disease, we can conclude that there is no universal answer and that it depends on the specific tissue and sometimes even the specific pathways involved in the disease.

To make sure that this specific approach of integrating orthology-derived pathways with tissue expression data from human and animal models is robust with respect to the chosen algorithms and cutoffs, we performed a robustness analysis. In order to use the confidence scores for gene–tissue associations from the TISSUES database, we needed to set a cutoff for whether a gene is expressed or not in a given tissue, which is not a straightforward choice. In addition, when identifying which and how many pathways are expressed in a given tissue, we chose a cutoff for the number of expressed pathway genes. The robustness analysis confirmed that, although the absolute numbers change, the trends remain the same, and thus, our findings are consistent and reproducible irrespective of the specific cutoffs chosen.

Our systematic data integration of pathways with tissue expression enables the investigation of mammalian pathway activity in several different healthy tissues of mouse, rat and pig as well as the comparison with the corresponding human tissues. We highlight tissue-specific features of the pathways and point out similarities and differences between human and the model organisms. Ultimately, we identify distinct pathway–tissue combinations, which are specifically more consistent with human for either of the three studied animal models. These findings can support researchers in the decision of which model organism to choose for a human disease of interest.

In the current analysis, we focused only on the three animal models mouse, rat and pig and on the seven tissues, which are well covered by experimental datasets in the TISSUES database. However, if the used resources become

more elaborate in the future, it should be possible to conduct the same type of analysis on more tissues and for more model organisms. This would of course require enough tissue expression data that can be calibrated and made comparable, for example, as done for the TISSUES database. Furthermore, although we based our study on the KEGG pathways database, our workflow for data integration and comparison is applicable to other pathway databases or gene–phenotype and gene–disease associations. The presented framework can also be extended to study the similarity of pathways upon activation or perturbation or to take into account the effect of specific genes, drugs or even diseases on the pathways in the same tissue for different model organisms, given that such a comprehensive collection of data exists and is made publicly available. Thus, future analysis would require the systematic assembly of associations between pathways, tissues and diseases to further aid researchers in choosing the best model organism for studying human diseases.

DATA AVAILABILITY

This study includes no data deposited in external repositories. All results generated in this study are included as supplementary files.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge Helen Cook and Damian Szklarczyk for help with the STRING database and providing the KEGG gene–pathway associations used in STRING, Alberto Santos for help with the TISSUES datasets, and Marie Locard-Paulet for very useful discussions on how to visualize and analyze the data.

Author contributions: N.T.D., J.G. and L.J.J. designed the project. N.T.D. performed the analysis and wrote the manuscript. O.P. assisted with the analysis and interpretation of the data. R.Y. contributed to the interpretation of the results. C.A. provided support for the data analysis. T.L. and F.P. advised on the project design and together with M.G. and P.F.S. gave input on the interpretation of the results. J.G. and L.J.J. supervised the project and contributed to the manuscript. All authors read and approved the final manuscript.

FUNDING

Danish Council for Independent Research [DFF-4005–00443]; Novo Nordisk Foundation [NNF14CC0001]. Funding for open access charge: University of Copenhagen.

Conflict of interest statement. T.L. is employed both by University of Copenhagen and by LEO Pharma A/S. All other authors declare that they have no competing interests.

REFERENCES

1. Young,R.S., Hayashizaki,Y., Andersson,R., Sandelin,A., Kawaji,H., Itoh,M., Lassmann,T., Carninci,P., Consortium,F., Bickmore,W.A. *et al.* (2015) The frequent evolutionary birth and death of functional promoters in mouse and human. *Genome Res.*, **25**, 1546–1557.
2. Simon,M.M., Greenaway,S., White,J.K., Fuchs,H., Gailus-Durner,V., Wells,S., Sorg,T., Wong,K., Bedu,E., Cartwright,E.J. *et al.* (2013) A comparative phenotypic and genomic analysis of C57BL/6J and C57BL/6N mouse strains. *Genome Biol.*, **14**, R82.
3. Pizzollo,J., Nielsen,W.J., Shibata,Y., Safi,A., Crawford,G.E., Wray,G.A. and Babbitt,C.C. (2018) Comparative serum challenges show divergent patterns of gene expression and open chromatin in human and chimpanzee. *Genome Biol. Evol.*, **10**, 826–839.
4. Santpere,G., Lopez-Valenzuela,M., Petit-Marty,N., Navarro,A. and Espinosa-Parrilla,Y. (2016) Differences in molecular evolutionary rates among microRNAs in the human and chimpanzee genomes. *BMC Genomics*, **17**, 528.
5. Aigner,B., Allison,W.T., Andreatini,R., Antonelli,M., Arndt,S.S., Austin,A., Brand,C., Bukowska,J., Caprariello,A.C., Carlisle,R.E. *et al.* (2017) In: Conn,P.M. (ed). *Animal Models for the Study of Human Disease*. Academic Press.
6. Seok,J., Warren,H.S., Cuenca,A.G., Mindrinos,M.N., Baker,H.V., Xu,W., Richards,D.R., McDonald-Smith,G.P., Gao,H., Hennessy,L. *et al.* (2013) Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci. U.S.A.*, **110**, 3507–3512.
7. Groop,L. and Pociot,F. (2014) Genetics of diabetes - are we missing the genes or the disease? *Mol. Cell. Endocrinol.*, **382**, 726–739.
8. Takao,K. and Miyakawa,T. (2015) Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc. Natl. Acad. Sci.*, **112**, 1167–1172.
9. Weidner,C., Steinfath,M., Opitz,E., Oelgeschläger,M. and Schönfelder,G. (2016) Defining the optimal animal model for translational research using gene set enrichment analysis. *EMBO Mol. Med.*, **8**, 831–838.
10. Nelson,D.R., Zeldin,D.C., Hoffman,S.M.G., Maltais,L.J., Wain,H.M. and Nebert,D.W. (2004) Comparison of cytochrome P450 (CYP) genes from the mouse and human genomes, including nomenclature recommendations for genes, pseudogenes and alternative-splice variants. *Pharmacogenetics*, **14**, 1–18.
11. Puccinelli,E., Gervasi,P.G. and Longo,V. (2011) Xenobiotic metabolizing cytochrome P450 in pig, a promising animal model. *Curr. Drug Metab.*, **12**, 507–525.
12. Hu,J.X., Thomas,C.E. and Brunak,S. (2016) Network biology concepts in complex disease comorbidities. *Nat. Rev. Genet.*, **17**, 615–629.
13. Kitsak,M., Sharma,A., Menche,J., Guney,E., Ghiassian,S.D., Loscalzo,J. and Barabási,A.-L. (2016) Tissue specificity of human disease module. *Sci. Rep.*, **6**, 35241.
14. Bult,C.J., Blake,J.A., Smith,C.L., Kadin,J.A., Richardson,J.E. and Mouse Genome Database Group (2019) Mouse Genome Database (MGD) 2019. *Nucleic Acids Res.*, **47**, D801–D806.
15. Groenen,M.A.M., Archibald,A.L., Uenishi,H., Tuggle,C.K., Takeuchi,Y., Rothschild,M.F., Rogel-Gaillard,C., Park,C., Milan,D., Megens,H.-J. *et al.* (2012) Analyses of pig genomes provide insight into porcine demography and evolution. *Nature*, **491**, 393–398.
16. Smith,J.R., Hayman,G.T., Wang,S.-J., Laulederkind,S.J.F., Hoffman,M.J., Kaldunski,M.L., Tutaj,M., Thota,J., Nalabolu,H.S., Ellanki,S.L.R. *et al.* (2020) The year of the rat: the rat genome database at 20: a multi-species knowledgebase and analysis platform. *Nucleic Acids Res.*, **48**, D731–D742.
17. Fabregat,A., Jupe,S., Matthews,L., Sidiropoulos,K., Gillespie,M., Garapati,P., Haw,R., Jassal,B., Korninger,F., May,B. *et al.* (2018) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, D649–D655.
18. Kanehisa,M., Furumichi,M., Tanabe,M., Sato,Y. and Morishima,K. (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, D353–D361.
19. Licata,L., Briganti,L., Peluso,D., Perfetto,L., Iannuccelli,M., Galeota,E., Sacco,F., Palma,A., Nardozza,A.P., Santonico,E. *et al.* (2012) MINT, the molecular interaction database: 2012 update. *Nucleic Acids Res.*, **40**, D857–D861.
20. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.

Alkan *et al. Genome Biology* (2018) 19:177
https://doi.org/10.1186/s13059-018-1534-x

Genome Biology

RESEARCH Open Access

CRISPR-Cas9 off-targeting assessment with nucleic acid duplex energy parameters

Ferhat Alkan, Anne Wenzel, Christian Anthon, Jakob Hull Havgaard and Jan Gorodkin*

Abstract

Background: Recent experimental efforts of CRISPR-Cas9 systems have shown that off-target binding and cleavage are a concern for the system and that this is highly dependent on the selected guide RNA (gRNA) design. Computational predictions of off-targets have been proposed as an attractive and more feasible alternative to tedious experimental efforts. However, accurate scoring of the high number of putative off-targets plays a key role for the success of computational off-targeting assessment.

Results: We present an approximate binding energy model for the Cas9–gRNA–DNA complex, which systematically combines the energy parameters obtained for RNA–RNA, DNA–DNA, and RNA–DNA duplexes. Based on this model, two novel off-target assessment methods for gRNA selection in CRISPR-Cas9 applications are introduced: `CRISPRoff` to assign confidence scores to predicted off-targets and `CRISPRspec` to measure the specificity of the gRNA. We benchmark the methods against current state-of-the-art methods and show that both are in better agreement with experimental results. Furthermore, we show significant evidence supporting the inverse relationship between the on-target cleavage efficiency and specificity of the system, in which introduced binding energies are key components.

Conclusions: The impact of the binding energies provides a direction for further studies of off-targeting mechanisms. The performance of `CRISPRoff` and `CRISPRspec` enables more accurate off-target evaluation for gRNA selections, prior to any CRISPR-Cas9 genome-editing application. For given gRNA sequences or all potential gRNAs in a given target region, `CRISPRoff`-based off-target predictions and `CRISPRspec`-based specificity evaluations can be carried out through our webserver at https://rth.dk/resources/crispr/.

Keywords: CRISPR-Cas9, Off-targets, Off-target scoring, Energy models, gRNA specificity, gRNA design

Background

The CRISPR-Cas9 system, adapted from a bacterial defense mechanism, is a powerful genome-editing tool that recently revolutionized the field of biology, biotechnology, and medicine [1]. The system consists of the Cas9 protein and a guide RNA (gRNA) which together form a riboprotein complex (RNP) that can bind to gRNA-directed location on genomic DNA. Upon binding, Cas9 cleaves the DNA, making a double-stranded break which enables further DNA modifications on the site. As alternative Class II CRISPR systems, there exist variants of the Cas9 protein and other similar proteins with similar genome-editing potential, like Cpf1 [2], C2c1 [3], and C2c2 [4], but each comes with different targeting constraints and efficiency for the intended cleavage. Cas9 is the first CRISPR protein that has been adapted as a genome editing tool in eukaryotes [5] and has been successfully applied numerous times on many genomes such as yeast, human, and mouse. The CRISPR-Cas9 mechanism starts with the RNP complex recognizing the protospacer adjacent motif (PAM) in the target genome and then forming an RNA–DNA interaction duplex between the gRNA and the DNA on the opposite strand of the PAM upstream region [6–8]. However, gRNAs are mostly designed in a way that only the first 20 nt on the 5′ end are capable of forming this duplex. In the following, we by gRNA refer only to this 20-nt DNA binding region. Note that it is the only region that is changed when targeting different regions in the genome. When PAM recognition is supplemented with a stable gRNA–DNA duplex,

*Correspondence: gorodkin@rth.dk
Center for Non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen,Grønnegårdsvej 3, 1870 Frederiksberg, Denmark

Cas9 protein cleaves the DNA on both strands in a PAM-proximal region, usually 3 nt upstream from the PAM sequence. After this cleavage, DNA could be repaired with non-homologous end joining or homologous DNA repair, enabling insertion or deletion of DNA elements in specific regions. This special capability of the CRISPR-Cas9 system promises revolutionary innovations in the field of biology, biotechnology, and medicine, due to its efficiency and practicality as genome-editing tool [9].

For any CRISPR-Cas9 application, the very first step is to select a target region in the genome, which consequently determines the gRNA sequence to be used. Different gRNA selections have varying on-target cleavage efficiencies, and the underlying molecular mechanism is still not fully understood [10]. So far, several factors such as sequence context, stability of the gRNA binding, chromatin accessibility, and PAM sequence have been reported as influential factors, and several on-target efficiency prediction methods have been proposed to be able to predict the efficiency of intended cleavage (see [11] for a thorough discussion). Another design concern for gRNA selection has been the specificity of the intended cleavage. Even though the CRISPR-Cas9 mechanism is believed to be very specific to carry out the intended cleavage on genome, many studies reported that the Cas9 complex also binds to other unintended regions, called off-targets, and performs cleavage at these off-target sites as well [12–21]. It has been shown that off-target regions are gRNA-specific and that they usually are highly homologous to the intended on-target region. When compared with on-target sites, reported off-target regions generally have up to six mismatches and off-targets with fewer mismatches tend to have more prominent binding and cleavage. Several tools have been developed to find potential off-target regions for given gRNA sequences and they mainly focus on finding off-targets in the genome of interest, allowing up to a certain number of mismatches [22]. However, initial analyses on experimentally reported off-targets showed that the type of mismatch and its distance from the PAM sequence also have significant importance. This information enabled the development of several off-target scoring methods and helped researchers to select their gRNAs with information on their off-targeting potential (see [11, 22] for a thorough discussion).

In this study, we developed novel off-target and specificity scoring methods distinctively by using a biophysical interaction model for Cas9–gRNA–DNA binding. There have been recent efforts to develop biophysical models for Cas9 binding [23–25]; however, none of the models actively made use of the free energy and enthalpy change parameters estimated for nucleic acid duplexes from experimental measurements [26–31]. These duplex-specific parameters enable computation of the free energy of nucleic acid duplexes, and they have been proven to be quite useful for intra- and inter-molecular interaction prediction of RNA molecules [32]. The base pair-specific nature of nucleic acid duplex energy models can potentially explain why some mismatches are more common within reported off-target regions and they can be quite helpful to accurately compute the stability of any Cas9 binding. Thorough details about how we obtain these parameters and make use of them within our scoring methods are given in the "Methods" section.

Results

To assess the off-targeting potential of gRNA selections in CRISPR-Cas9 applications, we developed two novel scoring methods, `CRISPRoff` and `CRISPRspec`. The former calculates an off-target score based on our energy model that approximates the free energy of any gRNA–DNA binding, and the latter provides a specificity score by making use of free energies computed for all possible on- and off-target bindings.

Our approximate free energy model is depicted in Fig. 1. It includes calculating a position-weighted binding energy between gRNA and the (off-)target DNA (ΔG_H), the free energy of the DNA duplex (ΔG_O), the folding energy of the gRNA only (ΔG_U), and a correcting factor (δ_{PAM}) corresponding to the type of PAM sequence. As full energy models are not available, we have made approximate models. The details of the model, parameters, and approximation are described in the "Methods" section. In brief, the `CRISPRoff` score is a score for a specific individual off-target binding and is equal to the negative of ΔG_B shown in the figure and Eq. (4) ("Methods"). The `CRISPRspec` score is the ratio of the Boltzmann-weighted energies of all possible but the binding energy ΔG_B over the on-target region, to the Boltzmann-weighted energies of all possible bindings including the on-target binding energy as listed in Eq. (5) ("Methods"). Hence, the `CRISPRoff` score can be considered as a confidence score assigned to predicted off-target sites of a gRNA and `CRISPRspec` score represents the specificity of this gRNA, or conversely its overall off-targeting potential.

In the following, we present our evaluation results for both methods, followed by our findings on the relationship between on-target cleavage efficiency and specificity of different gRNA selections.

Evaluation of off-target scoring methods

There exist a few methods in the literature that assign confidence scores to predicted off-target sites and we benchmarked our novel method `CRISPRoff` with six of them, CCTop [33], CFD [34], Cropit [35], Elevation (Elevation score) [36], MIT [11, 16], and VfoldCAS [24]. We benchmarked these methods under three different evaluation settings. First, we compared the performance of the methods with receiver operating characteristic (ROC) analysis

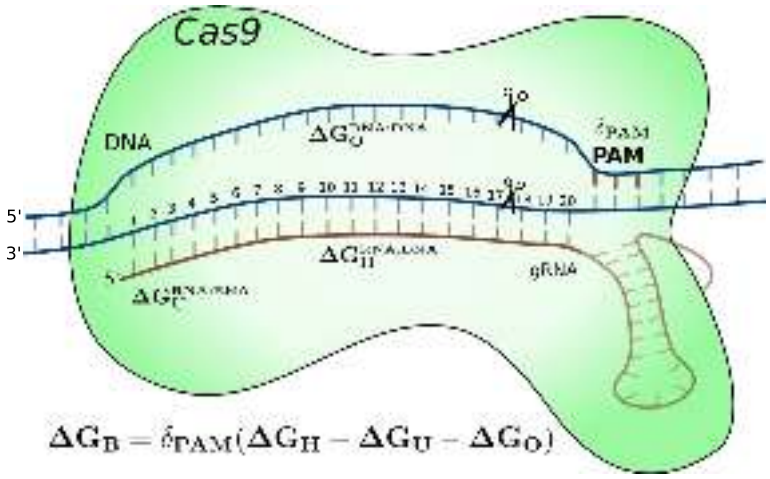


Fig. 1 Players of the energy model that determines the approximate free energy (ΔG_B) of any Cas9–gRNA–DNA binding. In this model, that is posterior to the Cas9–gRNA binding, there are four main contributions to the overall free energy. The first contribution is ΔG_H for the gRNA–DNA hybridization, computed with RNA–DNA duplex energy parameters and weighted by a position-wise estimate of the Cas9 influence in the binding. The second contribution is the ΔG_O penalty to open the DNA–DNA duplex in the target region and it is computed with DNA–DNA duplex energy parameters. The third contribution is the ΔG_U penalty that is the free energy of the gRNA (first 20 nt) folding. This is computed with RNAfold program which incorporates RNA–RNA duplex energy parameters [32]. The fourth contribution is the correction factor δ_{PAM} that is determined by the PAM sequence of the target

using the recently published Haeussler benchmark dataset that evaluated the performance of off-target scoring algorithms in a similar sense [11]. This dataset contains 650 off-target sequences reported for 31 different gRNAs and it is a collection of experimentally supported off-targeting data from 8 different studies [14–21]. Haeussler et al. originally used only a small portion of this data for their evaluation, limiting the ROC analysis to off-target predictions with up to four mismatches, excluding two of the gRNAs which had the highest number of off-targets and two of the assays that use targeted sequencing [14, 16], due to their low sensitivity [11]. In our analysis, assays that are classified as low-sensitivity by Haeussler et al. are also excluded; however, for a more comprehensive evaluation of off-target scoring methods, the two gRNAs with highest number of reported off-targets are included. We assume that the more off-targeting data taken into account, regardless of the volume of off-targets reported for one gRNA, the more comprehensive the performance assessment of off-target scoring methods becomes. We allow up to six mismatches in off-target predictions to include all experimentally supported off-targets (true positives) within the ROC analysis. Note that off-target predictions of the gRNAs in this dataset were also obtained from the benchmark dataset itself. Within the final ROC analysis set, we had 605 true positive (experimentally-supported) off-targets (with PAM sequences of NGG, NAG, or NGA) reported for 26 unique gRNAs, where total number of off-target predictions with up to six mismatches was equal to 1167036.

In Fig. 2, we present our ROC analysis where the true positive rate (TPR) and its corresponding false positive rate (FPR) are reported at method-specific varying thresholds. One can readily see that energy-based off-target score `CRISPRoff` performs better than all other methods with its higher area under the curve. For completeness, the precision-recall (PR) curve of this analysis is given in Additional file 1: Figure S1, where TPR and corresponding positive predictive values (PPV) are reported for each method. The PR curve also supports that `CRISPRoff` is the top performer with its highest area under the curve. A summary of the statistics from the ROC analysis is given in Table 1. In addition to higher area under ROC and PR curves, it is very clear that `CRISPRoff` outperforms all other methods with lower FPR and higher TPR values at given fixed TPR and FPR values, respectively. For example, when `CRISPRoff` score reaches 0.9 TPR, its FPR is 0.06 which is almost two times better than the closest competitors (CFD and Elevation). Note that, at this fixed TPR, the performance gain of `CRISPRoff` over these methods actually corresponds to $>$ 58k fewer FPs in off-target predictions.

In our second benchmark setting, we investigated how well different off-target scoring methods agree with the cleavage efficiency of the experimentally reported off-target regions. In these analyses, the recently published

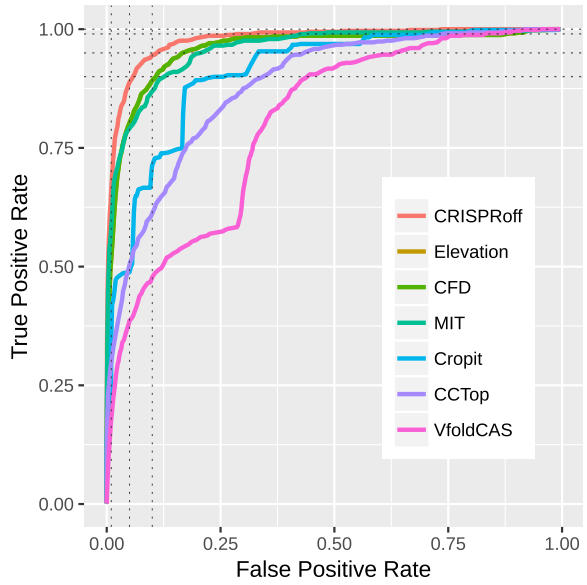


Fig. 2 Receiver operating characteristic (ROC) analysis of off-target scoring methods when benchmarked with the Haeussler dataset [11], allowing up to six mismatches, and NGG, NAG, and NGA PAM sequences for off-targeting. ROC curves for CFD and Elevation methods largely overlap and `CRISPRoff` shows the best performance with the largest area under its ROC curve. FPR and TPR values of the methods at specific points, indicated by dashed lines, are given in Table 1

CIRCLE-seq [37] and SITE-seq [38] experimental datasets were used. In CIRCLE-seq dataset, off-targets are reported in 19 experiments using 11 different gRNAs, whereas this is done for 8 gRNAs at 5 different concentrations within the SITE-seq dataset. Both methodologies detect the gRNA-specific off-targets on a genome-wide level and they provide read counts for cleaved off-target regions in the human genome, representing their cleavage efficiency. In the CIRCLE-seq dataset, some gRNAs are tested multiple times in different cell lines and it is shown that off-targeting is more gRNA-specific than cell-line-specific. In the SITE-seq dataset, experiments at different concentrations show that as the concentration of Cas9 complex increases, the off-targeting effects become more prominent. Within the evaluation, we first made use of the CIRCLE-seq dataset excluding one experiment where the gRNA did not have any perfect complementary target in the human genome (hg38). Each subplot in Fig. 3 indicates the performance of different off-target scoring methods on CIRCLE-seq dataset. In these plots, positive correlation between off-target scores and cleavage efficiencies hints to better performance and it is clear that `CRISPRoff` score is in best agreement with measured off-target activity over all CIRCLE-seq reported off-targets under consideration. This is supported by the `CRISPRoff` score having the highest Pearson correlation coefficient (ρ), which is given in the top-left corner of each plot. Closest to this are the CFD and Elevation scores, which is also in agreement with the ROC analysis above. The analysis with the SITE-seq dataset is however more blurry and does not support this as significantly as the CIRCLE-seq dataset. The correlation between off-target scores and their cleavage efficiency reported by the SITE-seq method is very weak for all methods (see Additional file 1: Figure S2).

In our third benchmark, we evaluated the off-target scoring methods with their accuracy in their top predictions. For every experiment in the CIRCLE-seq dataset, we used the RIsearch2 [39] program to obtain the list of potential off-target sites, up to six mismatches in human

Table 1 Area under ROC (TPR vs. FPR) and precision-recall (PPV vs. TPR) curves for off-target scoring methods when benchmarked with the Haeussler dataset [11], allowing up to six mismatches, and NGG, NAG, and NGA PAM sequences for off-targeting

		Off-target scoring method						
Area		`CRISPRoff`	Elevation	CFD	MIT	Cropit	CCTop	VfoldCAS
ROC		**.98**	.96	.96	.96	.91	.88	.80
PR		**.18**	.08	.08	.12	.05	.06	.01
TPR								
.9		**.06**	.11	.11	.13	.27	.34	.44
.95	FPR	**.11**	.17	.17	.21	.33	.44	.63
.99		**.32**	.88	.88	.44	.71	.74	.84
1		**.73**	.97	.97	.96	.99	.91	.96
FPR								
.01		**.67**	.52	.52	.59	.36	.31	.18
.05	TPR	**.89**	.80	.80	.79	.49	.50	.39
.1		**.94**	.89	.89	.87	.71	.61	.48

Corresponding TPR and FPR performance of the methods are also given for some fixed FPR and TPR values. Best performances are given in bold

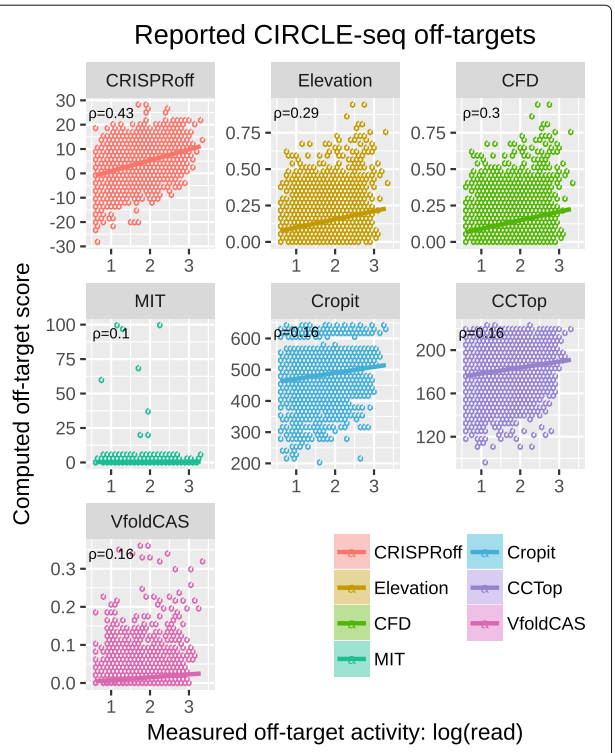


Fig. 3 Method-specific off-target score vs. off-target activity scatterplots (hexagonal binned) with all reported off-targets from CIRCLE-seq dataset. Measured off-target activity, given on the *x*-axis, corresponds to the logarithm of read counts reported for that specific off-target region. Fitted lines are shown together with the Pearson correlation coefficient between *x*- and *y*-axis variables in the top left corner of each subplot

(hg38) genome (see the "Methods" section for details), and filtered them with PAM sequences of NGG, NAG, or NGA. These were then ranked by each of the off-target scoring method. Focusing solely on the top 10 off-target predictions of each method for all 18 experiments (180 predictions in total), the distribution of measured off-target activities was compared in Fig. 4. One can see that top off-targets identified with the `CRISPRoff` and MIT methods have the lowest number of false positives since more than half of their top predictions have cleavage support from the CIRCLE-seq experimental dataset. The median measured off-target activity values of the top off-targets from the CFD, Elevation, Cropit, CCTop, and VfoldCAS methods are equal to 0, indicating more than half of their top predictions have no experimental support. The median values of $\sim$ 1.0 for `CRISPRoff` and MIT methods, suggest similar outperformance of all the other methods for both of these. The corresponding analysis on the SITE-seq data set is presented in Additional file 1: Figure S3. However, in this analysis, the methods show closer performances, except the poor performance of VfoldCAS, Elevation, and CFD.

All in all, our findings from all the benchmarks presented above suggest that the `CRISPRoff` method consistently outperforms the other off-target scoring methods when assigning confidence scores to predicted off-target regions. This is supported by its stronger agreement with experimentally reported off-targets, especially in the CIRCLE-seq dataset, not only in classification but also at cleavage efficiency correlation level.

Evaluation of gRNA specificity scores

Apart from assigning confidence scores to the off-target predictions of a gRNA, another challenge for Cas9 off-targeting assessment is to assign specificity scores to different gRNA selections. To the best of our knowledge, there exist two methods in the literature that can perform this task, namely the MIT [16] and Elevation (Elevation-aggregate) [36] methods. With this study, we propose a novel approach, `CRISPRspec`, to measure the specificity of any given gRNA targeting a selected genome. For more accurate evaluation of the `CRISPRspec`, Elevation, and MIT methods, we use two versions of the MIT specificity score, indicated as MIT and MIT*. The former MIT score is computed by the CRISPOR webserver [11] where off-target space is limited with four mismatches as default and the recommended threshold is 50 to bin the gRNAs into high or low specificity groups. The latter MIT score, MIT*, is computed using the code from the Haeussler benchmarking study [11] with a different off-target prediction set given as input, that is the set used for computing the `CRISPRspec` score. For any given gRNA, this set is generated by using RIsearch2 [39], allowing up to six mismatches between gRNAs and their targets in the human genome (hg38), followed by post-filtering with the PAM sequences of NGG, NAG, and NGA. On the other hand, Elevation score is computed using its own off-target prediction set which also allows up to six mismatches and same PAM sequences.

Performances of the `CRISPRspec`, Elevation, MIT, and MIT* scores are compared using the SITE-seq and CIRCLE-seq datasets. However, evaluation with the SITE-seq dataset is our primary focus since all experiments from this dataset are performed in the same type of cell line. We assume that in this way, we can minimize the potential evaluation error that is caused by different chromatin accessibility patterns of the cells, a parameter that is not taken into account in all methodologies. Besides, the SITE-seq dataset enables assessing the accuracy of specificity scores at different concentrations.

In our evaluation with any of the datasets, we first compute the specificity of gRNAs in that group with all three methods and analyze its agreement with the experimentally measured specificity. The latter is represented by the fraction of off-target read counts within the total read count reported for that gRNA in that dataset. Evaluation

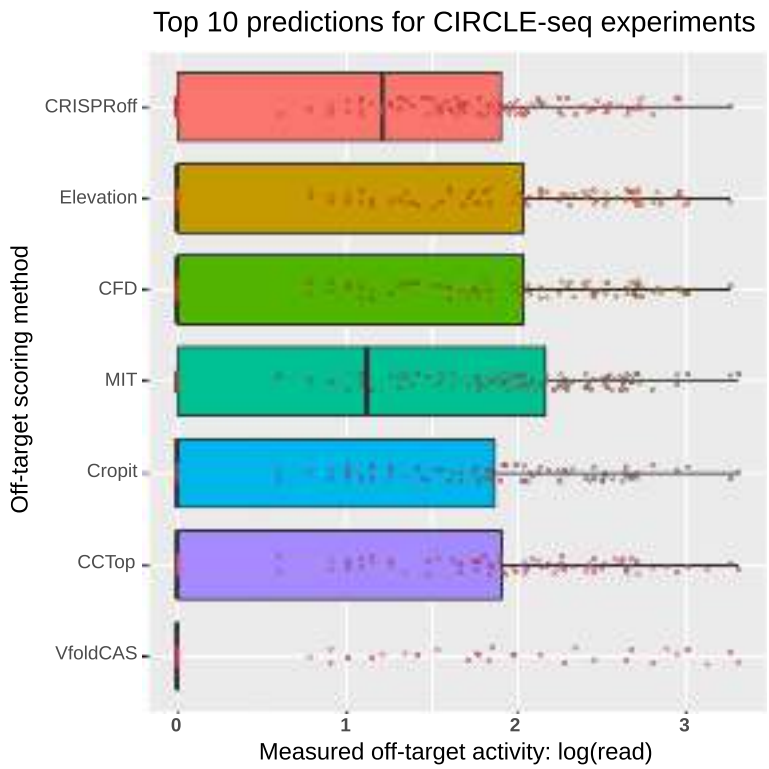


Fig. 4 CIRCLE-seq measured off-target activity distributions of method-specific top predictions (180 in total, top 10 for all 18 experiments). Distributions are given separately for each method in box plot format combined with log(read) values for each off-target prediction as dot plots. Value 0 in *x*-axis corresponds to no experimental support for that off-target prediction

results with the SITE-seq dataset at four different concentrations are shown in Fig. 5 where the x-axis indicates the predicted specificities and the y-axis shows the experimentally measured specificities of the gRNAs. It is expected that gRNAs with higher specificity have a lower fraction of off-target read counts, and therefore, stronger negative correlation between the two measures hints to better performance for that method. Focusing on the first row in Fig. 5, the lowest concentration experiments in the SITE-seq dataset, one can see that `CRISPRspec` specificity score is in best agreement with experimental results due to lower off-targeting activity for highly specific gRNAs and higher off-targeting activity for the low specificity ones. However, agreement with the experimentally measured specificity is much weaker for MIT and MIT* scores and weakest for Elevation method. For the results in the other concentration levels (rows 2–4 in Fig. 5), it is clear that the experimental evidence for specificity differences between gRNAs disappears at higher concentrations so as the agreement between experimental and predicted specificity measures.

The results concerning the CIRCLE-seq dataset are given in Additional file 1: Figure S4, which also suggests that `CRISPRspec` is the top performer ($\rho = -0.72$) when compared to MIT ($\rho = -0.49$), MIT* ($\rho = -0.05$) and Elevation ($\rho = 0.20$) methods.

Specificity and on-target efficiency interplay for gRNAs

On-target cleavage efficiency of a gRNA is influenced by various factors, from gRNA/target sequence context to genomic location of the target, and there are several tools with varying performance that take these factors into account for efficiency prediction of the selected gRNA [11]. However, predicted specificity measure of different gRNA selections is usually not part of on-target efficiency scoring schemes since this relationship is believed to be insignificant. Here, we reanalyze this potential interplay using both numerical (specificity measure) and experimental (cleavage efficiency) data for two groups of gRNAs, Doench2015 [40] (881 gRNAs) and Wang2015 [41] (2921 gRNAs). Firstly, the `CRISPRspec` and MIT* specificity score of these gRNAs are computed and they are assigned into low, medium, and high specificity groups within the respective data sets. The binning thresholds for `CRISPRspec` and MIT* scores are selected in a way that they would create three equal-sized specificity groups for 57980 unique gRNAs that target 16322 different genes in the human

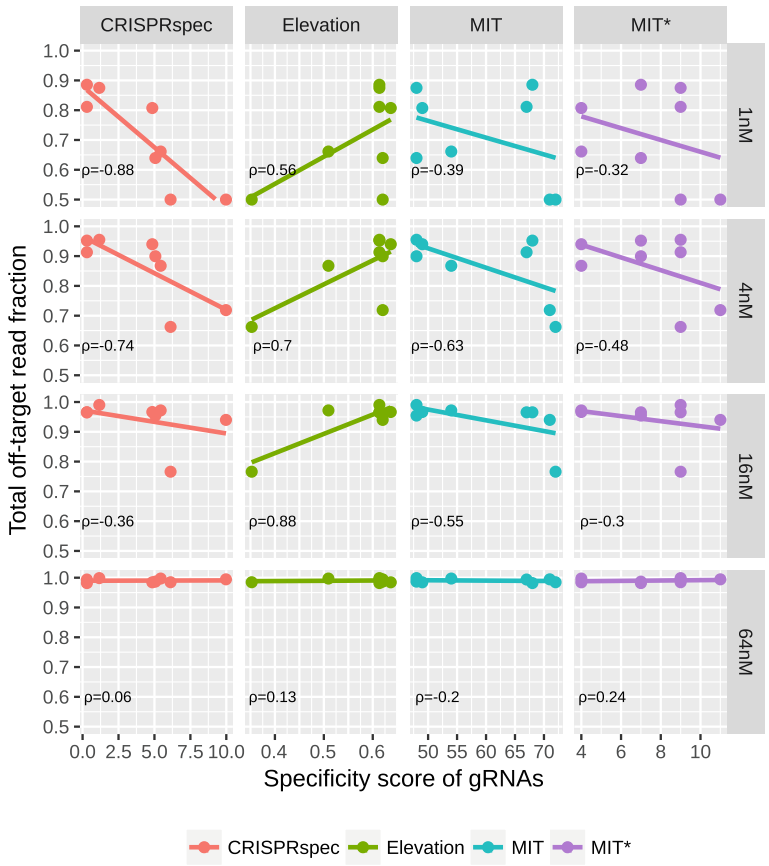


Fig. 5 Total off-target activity reported by the SITE-seq experiments vs. method-specific specificity scores for eight unique gRNAs. For each gRNA, the `CRISPRspec` and MIT* scores have been computed with the same set of off-target predictions allowing up to six mismatches, whereas Elevation scores are based on its own prediction set (up to six mismatches) and MIT score has been computed with CRISPOR tool [11] allowing up to four mismatches in off-target predictions by default. Results regarding to different concentration levels in the SITE-seq dataset are given separately at each row. Fitted lines are shown together with the Pearson correlation coefficient between *x*- and *y*-axis variables in the bottom left corner of each subplot

genome [42]. Secondly, we compare the distribution of experimentally measured on-target cleavage efficiencies of the gRNAs that are binned into different specificity groups.

In Fig. 6, one can see that efficiency distribution of low and high specificity groups are skewed towards opposite ends, indicating that low specificity gRNAs are more likely to have less on-target efficiency and highly specific gRNAs are more likely to be more potent for their intended cleavage. This is supported by pairwise Kolmogorov–Smirnov (K–S) tests within each dataset, indicating significant differences (p value< 0.05) between the on-target modulation frequency distribution of gRNAs from different specificity groups (except the test between low and medium specificity group for Doench2015 dataset). When using the MIT* score instead of the `CRISPRspec` score for the specificity grouping of gRNAs, this interplay, with higher confidence on Doench2015 dataset (lower p values in K–S tests), is still supported. However, this is not the case for MIT* score with the Wang2015 dataset (see Additional file 1: Figure S5). Out of the three pairwise K–S tests within the Wang2015 dataset, K–S tests for low-vs-medium and medium-vs-high specificity groups yield to p values larger than 0.05, whereas the low-vs-high K–S test yields a p value equal to 0.045. Failure of these two K–S tests with MIT* scores in Wang2015 dataset could also be interpreted as a sign of `CRISPRspec` outperforming MIT* score.

Over all, these findings provide a considerable support for the parallel relationship between the specificity and the on-target efficiency of gRNAs and suggests that off-target volume of gRNAs might have negative impact on the efficiency of their on-target cleavage. Therefore, integration of the `CRISPRspec` specificity measure to gRNA

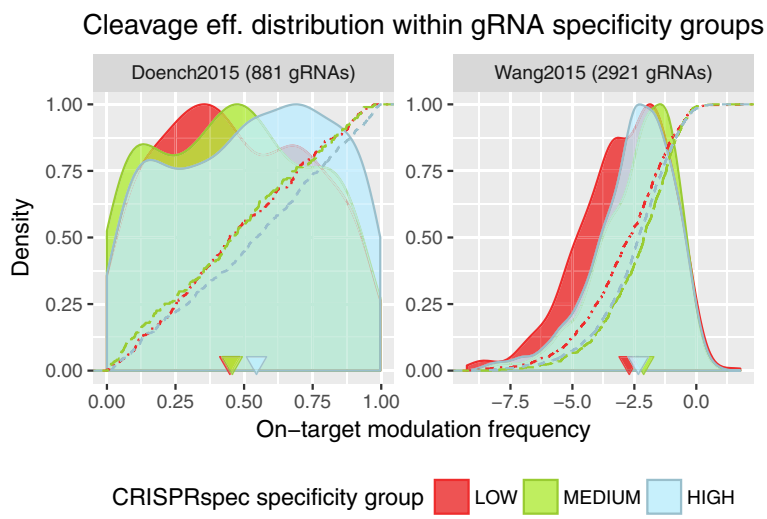


Fig. 6 On-target modulation frequency distribution of gRNAs that are binned into low, medium, and high specificity groups using `CRISPRspec` method. Distributions are given as kernel density estimates (filled curves) together with the cumulative distribution function (dashed lines) of on-target modulation frequencies for each specificity group, separately for each dataset. Given modulation frequencies represents the cleavage efficiency of the intended on-target and are dataset specific. Triangles on the *x*-axis indicates the median values

efficiency prediction tools can potentially improve their performances.

Discussion

Prior to any CRISPR-Cas9 genome-editing application, computational on- or off-targeting assessment of gRNAs is a crucial step to be able to select the most efficient gRNAs with minimum off-targeting effect. With this study, we proposed two novel methods for computational off-targeting assessment, `CRISPRoff` and `CRISPRspec`. The `CRISPRoff` off-targeting score can be interpreted as a confidence score that is assigned to the predicted off-targets of a gRNA and the `CRISPRspec` specificity score is a measure for the specificity/off-targeting potential of a gRNA. Both of the methods are based on an approximate energy model for Cas9–gRNA–DNA binding which is another novel outcome of this study. The model proposed here uses the nucleic acid duplex energy parameters for free energy computation, taking all RNA–RNA, RNA–DNA, and DNA–DNA interactions into account.

In our benchmark analysis with the latest experimental off-target screening datasets, we showed that `CRISPRoff` and `CRISPRspec` scores are more accurate than other available off-target and specificity scoring methods, making them the new state-of-the-art methods for computational off-targeting assessment of CRISPR-Cas9 gRNAs. Their strong agreement with the experimental off-target screens shows that they hold great potential to serve as gRNA design criteria prior to all Cas9 genome-editing applications. For the selection of gRNAs, `CRISPRoff` score can help with accurate ranking of predicted off-target regions, whereby gRNAs with high confidence off-targets on important regions of the target genome could be discarded in the first place. In addition, when the volume of off-targeting is a bigger concern than the individual off-target regions, `CRISPRspec` specificity score can help with pre-filtering of the gRNA selections based on their measured specificity on the target genome. Due to the potential interplay we have shown between the specificity and on-target cleavage efficiency of gRNA selections, selecting highly specific gRNAs can also increase the chances of successful on-target cleavage for Cas9 applications. As a result, these two novel methods, `CRISPRoff` and `CRISPRspec`, provide more accurate off-targeting assessment of gRNA selections and can help researchers to use the CRISPR-Cas9 system with higher efficiency and security.

All benchmarks given in this study are focused on the human genome, simply due to the number of datasets available for human. However, more off-targeting data is becoming available for other organisms as well and we consider the benchmarks on other genomes as part of our future work. The starting point for such benchmarks could be the Anderson2018 dataset, where a few thousand off-target regions are tested for over hundred gRNAs in mouse and rat genomes [43].

As more future work, our free energy-based approach applied here could provide further understanding about the details of the Cas9 binding and cleavage machinery, whether it is on- or off-target. Moreover, our analysis on the specificity-efficiency interplay suggests that predicted

specificity measure of gRNAs, like CRISPRspec, could be incorporated into gRNA design tools and this might enhance the efficiency prediction for gRNA selections.

The methods proposed here solely focus on CRISPR-Cas9 system; however, they can easily be adapted to other CRISPR proteins as well. This would require minor reformulations in the approximate energy model and some of the Cas9-related weights would need to be retrained for the CRISPR protein of interest. These weights could be trained using protein-specific experimental off-targeting and/or biochemical profiling data, as we did here using a biochemical profiling dataset [25] for Cas9 off-target interactions (see "Methods" section). Additionally, our partition function-based approach can incorporate the abundance information of targets as well. This also holds great potential to be applied to off-targeting assessment of RNA-targeting CRISPR proteins, like Cas13 [44]. This approach has been successfully applied to siRNA off-target predictions before [39] and transforming this approach into CRISPR applications is part of the future work.

Conclusions

The performance of the CRISPRoff off-target scoring method and the CRISPRspec gRNA specificity measure not only enables more accurate off-target evaluation of gRNA selections. They imply that the binding energies have a substantial impact on off-targeting mechanisms, which also provides a direction for further studies. Prior to any CRISPR-Cas9 genome-editing application, the CRISPRoff-based off-target predictions and the CRISPRspec-based specificity evaluations can be carried out through our webserver at https://rth.dk/resources/crispr/.

Methods

Approximate free energy model for Cas9 binding

Our observations, along with recent studies [23], support that the binding affinity of the Cas9–gRNA–DNA complex controls not only the occupancy of the target DNA but also influences the cleavage rate of it. Denoting any Cas9 complex binding with $B[g,t]$ and its free energy with $\Delta G_{B[g,t]}$, for a gRNA g and a target DNA t, our approximate free energy computation consists of four components: (i) the free energy contribution of gRNA–DNA hybridization ($\Delta G_H[g,t]$), (ii) the energy penalty for unfolding the gRNA itself ($\Delta G_U[g]$), (iii) another penalty for opening (melting) the double-stranded DNA ($\Delta G_O[t]$), and (iv) a final energy correction $\delta_{PAM[t]}$ based on the PAM sequence of the target t. These components make up the full energy model illustrated in Fig. 1, and the equation in the figure summarizes the free energy approximation of any binding site t for a given gRNA g.

To be able to compute all the ΔG free energy contributors, we made use of the Turner [26] and SantaLucia [27] nearest neighbor energy models for RNA–RNA and DNA–DNA duplexes, respectively. Note that we also used the parameters from the Allawi energy model [30] to complement some of the missing parameters of the SantaLucia model for DNA–DNA duplexes, e.g., G-T mismatches. A summary of these models can be found in the Additional file 1: Section 2. For the RNA–DNA duplex energy model, we primarily used the Sugimoto [28, 29] and Watkins [31] energy models to obtain the free energy parameters for stacked base pairs and some specific single mismatches. Due to the lack of the full energy parameters [23], we simply averaged the DNA–DNA and RNA–RNA parameters to complete the missing parameters of this model. The same approach was also used in the ViennaRNA package [45]. Our resulting nearest neighbor energy models for all three duplexes include base pair stacking energy contributions, penalties for mismatches within internal loops, and specific energy contributions of the internal loops at varying lengths. Further details about the nucleic acid duplex parameters are given in Additional file 1: Section 2. Note that, within the current models, we ignore the energy parameters for bulges since we only score mismatched off-target predictions. This is a common limitation for all off-target scoring methods; however, it is not a concern since bulged off-targets have been rarely reported at very low cleavage rates.

Each of the four contributions to our energy model mentioned above are determined as follows.

(i) $\Delta G_H[g,t]$: This contribution is obtained by summing up the estimated RNA–DNA interaction parameters. However, due to the influence of the Cas9 protein, we weight these for each position i in the interaction ($1 \leq i \leq 19$), by a factor $\Gamma_{Cas9}[i]$ explained below. Thus we compute $\Delta G_H[g,t]$ as

$$\Delta G_H[g,t] = \sum_{i=1}^{19} \Gamma_{Cas9}[i] \times \Delta G^{RNA:DNA}_{g[i,i+1]:t[i,i+1]}, \quad (1)$$

where $\Delta G^{RNA:DNA}_{g[i,i+1]:t[i,i+1]}$ is the estimated free energy contribution of the stacked match (or mismatch) between the gRNA and the target DNA sequence at position i. When Watson–Crick base pair matches are stacked on each other, the free energy contribution of position i depends only on the (i)th and $(i+1)$th bases ($g[i,i+1]$ and $t[i,i+1]$), where the order of i is from 5′ to 3′ end of the gRNA and the other way around (3′ to 5′) for the DNA (see Fig. 1 and Additional file 1: Figure S6). However, interactions formed between gRNAs and off-targets usually contain mismatches and they create interior loops in the RNA–DNA duplex. As explained above, in regions with stacked Watson–Crick base pairs, every stacking pair contributes individually at each position; however, for interior

loops, we compute the overall energy of the interior loop and divide it equally to all positions forming the loop as positional contributions. In Additional file 1: Figure S6, we provide an example gRNA–DNA binding and explain how to compute its positional free energy contributions in Additional file 1: Section 2.1.3.

The influence of the Cas9 protein is modeled heuristically by generating positional weights, $\Gamma_{Cas9}[i]$, for the energy contribution at each position i of the gRNA–DNA binding ($1 \leq i \leq 19$). The base pair stability at different positions of this binding might have different impacts due to the conformation of Cas9 protein and this impact can be trained on biochemical profiling datasets that can measure the kinetics of different gRNA–target bindings. Here, we used a recently published biochemical profiling dataset for Cas9 off-target bindings [25], where association and dissociation rate of nuclease-dead dCas9 interactions are measured with a massively parallel method. Our estimation of $\Gamma_{Cas9}[i]$ parameters are done as follows: For one specific gRNA, denoted with $\hat{g}$, this dataset provides initial association rates across a range of potential off-target sequences. We denote this off-target set with O, every individual off-target with $\hat{o}_n$ and its association rate with $\tilde{a}_n$, where $1 \leq n \leq |O|$. First, for every off-target $\hat{o}_n$, we compute the energy contribution of 19 base pair stackings individually, between the gRNA and that specific off-target. Then, for each position i in the stack, we calculate the W_i position-specific weighted sum of the energy contributions over all off-targets, where the weight is the association rate $\tilde{a}_n$ for every $\hat{o}_n$. Finally, to transform these W_i weighted sums into $\Gamma_{Cas9}[i]$ positional weights, where the lowest positional weight is desired to be 1 with no large deviations from this value, we normalize them with the minimum sum, take its logarithm, and sum it with 1. This computation is formulated in Eq. (2) below and our final set of values have been computed as $\Gamma_{Cas9} = \{1.80, 1.96, 1.90, 2.13, 1.38, 1.46, 1.00, 1.39, 1.51, 1.98, 1.88, 1.72, 2.02, 1.93, 2.08, 1.94, 2.15, 2.04, 2.25\}$. The obtained values show the importance of the PAM-proximal region with consistently higher weights.

$$\Gamma_{Cas9}[i] = \log_{10}\left(W_i / \min_{W_1 \ldots W_{19}}\right) + 1 \qquad (2)$$

$$with\ W_i = \sum_{n=1}^{|O|} \tilde{a}_n \times \Delta G^{RNA:DNA}_{\hat{g}[i,i+1]:\hat{o}_n[i,i+1]}$$

(ii) $\Delta G_U[g]$: For this we use the RNAfold program [32] with gRNA sequence that binds to the target DNA given as input (first 20 nt), and obtain the free energy of predicted MFE structure. Note that for some gRNA sequences, this value is equal to zero due to lack of predicted folded structure.

(iii) $\Delta G_O[t]$: Similar to the RNA-DNA interaction, this is obtained by summing up the estimated DNA-DNA interaction parameters:

$$\Delta G_O[t] = \sum_{i=1}^{19} \Delta G^{DNA:DNA}_{t'[i,i+1]:t[i,i+1]}, \qquad (3)$$

where we note that $\Delta G^{DNA:DNA}_{t'[i,i+1]:t[i,i+1]}$ represents the duplex-specific nearest neighbor energy models as explained above. Since the DNA–DNA duplex (target t and its complement t') at the target site is always perfect-complimentary, we only use the stacking energies of Watson–Crick pairs from DNA–DNA duplex energy parameters, for this computation. As can be seen from the equation above, every stacking position $(i, i+1)$ contributes individually to the overall free energy where the direction for i is from 3′ to 5′ end for target DNA t and the other way around (3′ to 5′) for its complement t'. We provide the stack-specific energy parameters, based on SantaLucia [27] and Allawi [30] energy models, in Additional file 1: Table S2.

(iv) $\delta_{PAM[t]}$: The PAM sequence in the target DNA region is assumed to be responsible for the initial Cas9 recognition but the stability of the Cas9–gRNA–DNA complex is maintained through the RNA–DNA binding. Therefore, we decided to introduce the effect of PAM sequence to the overall binding stability with a parameter δ_{PAM} that influence the computed overall binding free energy. Values for δ_{PAM} have been selected arbitrarily for Cas9, as 1.0, 0.9, and 0.8 for the PAM sequences of NGG, NAG, and NGA, respectively. These values solely reflect our observations in the literature for experimentally validated off-targets of Cas9.

`CRISPRoff` and `CRISPRspec` scores

For a given gRNA g and off-target t_{off}, `CRISPRoff` score is simply equal to the estimated free energy contribution of the off-target binding $\Delta G_B[g, t_{off}]$. However, `CRISPRspec` score computation is more comprehensive since we use a partition function approach from statistical thermodynamics to model the ensemble of all potential interactions. This model has already been proposed for CRISPR applications by Farasat and Salis [23], and it has been successfully applied to siRNA off-targeting assessment before [39]. Through the partition function, we simply compute the summed probability of all potential off-target interactions and propose its negative logarithm as our `CRISPRspec` specificity score. For a given gRNA g, denoting its set of target predictions with $\mathcal{T}_g$ including the intended target t_{on}, and the thermodynamic constant with β, below equations summarize how `CRISPRoff` and `CRISPRspec` scores are computed.

$$\begin{aligned}\texttt{CRISPRoff}[g, t_{off}] &= -\Delta G_B[g, t_{off}] \\ &= -\delta_{PAM}\left(\Delta G_H[g, t_{off}] - \Delta G_O[t_{off}] - \Delta G_U[g]\right)\end{aligned} \quad (4)$$

$$\texttt{CRISPRspec}[g, \mathcal{T}_g] = -\log_{10}\left(\frac{\sum_{\forall t \in \mathcal{T}_g \setminus \{t_{on}\}} e^{-\beta \Delta G_B[g,t]}}{\sum_{\forall t \in \mathcal{T}_g} e^{-\beta \Delta G_B[g,t]}}\right) \quad (5)$$

Other off-target and specificity scoring methods

To compute the other off-target scores that are benchmarked here except the VfoldCAS and Elevation scores (see below), we simply made use of the code implemented in the Haeussler benchmarking study [11]. According to this study, some of these codes were taken from original sources but some were simply implemented by Haeussler et al. according to corresponding papers. For more information about this source code, please see the corresponding benchmark paper [11]. For the VfoldCAS score computation, we used its webserver [24] by uploading the gRNA and off-target sequences when needed.

Elevation scores have been computed using the stand-alone version of the tool (v3.3) that is downloaded through its github page. For any gRNA, both Elevation score (off-targeting) and Elevation-aggregate (specificity) scores have been computed using its own set of off-target predictions since it does not accept user-defined off-target sequences. However, when running the tool, we did not limit the number of off-target predictions and allowed up to six mismatches with NGG, NGA, and NAG PAM sequences (by passing the following arguments: `-forcePamList NGG,NAG,NGA -t 6 -matchSiteCutoff 0`). When benchmarking the off-targeting scores, computed Elevation scores were parsed from the output files of the tool and assigned to corresponding off-target sequences. Note that off-target sequences that we could not compute an Elevation score for have been excluded from the analysis.

Lastly, to compute the original MIT specificity score, we ran the stand-alone version of the CRISPOR tool (v4.2) [11], allowing up to four mismatches between gRNAs and potential off-targets as it is the default option. However, since our `CRISPRspec` score was computed with our in-house predictions, we computed the updated MIT* score using the source code provided by the benchmark study [11].

Benchmarking datasets

For evaluation purposes, we used three different off-targeting datasets. The dataset used for ROC analysis is taken from the benchmarking study [11] through its GIT repository, accessed in June 2017. The downloaded data includes 31 gRNA sequences, 718 reported off-targets, and all off-target predictions with up to four, five, or six mismatches have been generated using the provided code. Note that, as default, NGG, NAG, and NGA were all allowed as PAM sequences in off-target predictions given here. The area under ROC and PR curves were computed using the PRROC [46] package in R environment.

The other two datasets used for benchmarking are the CIRCLE-seq and SITE-seq datasets. For each of the datasets, we downloaded the gRNA sequences (11 in CIRCLE-seq, 8 in SITE-seq) and the reported off-targets (5563 in CIRCLE-seq, 5847 in SITE-seq), along with their read counts from the corresponding supplementary material of the papers. For the off-target predictions of these gRNAs in human genome (hg38), we used the RIsearch2 (v2.1) tool [39]. We allowed up to six mismatches between gRNAs and off-targets that is achieved with following settings: `-s 1:20 -l 0 -m 6:0 -e 1000 -noGUseed -p3`. Then, these predictions were filtered according to valid NGG, NGA, and NAG PAM sequences, and computation of all off-targeting or specificity scores for these datasets was performed as explained above.

For the off-target prediction of gRNAs, we chose the RIsearch2 program due to its high-speed performance and flexibility. It is originally proposed as an RNA–RNA interaction prediction tool that uses a seed-and-extend framework. However, by passing the parameters `-s 1:20 -l 0 -m 6:0`, we have only exploited its suffix array-based seed localization step, finding all off-target regions in the human (hg38) genome that have up to six mismatches with given 20-nt-long gRNA. Note that we ignore all the energies computed by RIsearch2 program and recompute the gRNA–DNA interaction energies within our pipeline.

On-target efficiency datasets

To investigate the relationship between specificity and on-target cleavage efficiency of gRNAs, we used two different datasets, Doench2015 [40] and Wang2015 [41]. However, data for both datasets was also taken from the Haeussler benchmark study [11]. The downloaded data is already processed and includes the gRNA sequences and their cleavage efficiency measured as described in [11]. Doench2015 dataset includes 881 gRNAs with on-target modulation frequencies ranging between 0 and 1, whereas Wang2015 dataset includes 2921 gRNAs with frequencies ranging between −10 and 2. The specificity score computation of these 3802 gRNAs was performed with the same benchmark settings.

Webserver

For the off-targeting assessment of CRISPR-Cas9 gRNAs with `CRISPRoff` and `CRISPRspec` scores, we created a

webserver that meets the needs of different use cases. In the simplest use case, one can upload a gRNA sequence together with its set of predicted off-targets and the webserver returns the computed `CRISPRoff` scores together with the corresponding `CRISPRspec` specificity score of the gRNA, focusing solely on the given set of off-targets. For simplicity, the user can upload the off-target prediction set in different file formats as well, such as RIsearch2 [39] or Cas-OFFinder [47] result files. In this use case, the webserver is not limited to any organisms. Given off-targets can be based on any organism, however, for accurate `CRISPRspec` scorings, given off-target data must be genome-wide and must include the intended on-target sequence as well. Besides, repeated off(on)-target sites in the genome must be given separately as independent target sequences.

In case of missing off-target prediction data for gRNAs or when comparing multiple gRNA designs, the webserver performs the off-target predictions itself, using the RIsearch2 program (v2.1) in the background on a user-selected organism. In this case, the webserver outputs the `CRISPRspec` scores of the gRNAs under consideration together with gRNA-specific links to access the `CRISPRoff` scores of predicted off-target regions. In this use case, on-target and off-target sequences of all potential gRNAs can also be deployed into the UCSC browser [48] with one click for more detailed investigations. The webserver and download links for the scripts that are actively used at the back-end of the webserver are accessible through https://rth.dk/resources/crispr/.

Additional files

Additional file 1: Supplementary document includes Supplementary **Figures S1–S6** and Supplementary **Tables S1–S3.** (PDF 882 kb)

Additional file 2: Source code of CRISPRspec and CRISPRoff. (TAR 12,511 kb)

Acknowledgements
We thank Ivo Hofacker, Stefan Seemann, and all the other members of RTH for fruitful discussions and the anonymous reviewers for their valuable constructive comments.

Funding
This work was supported by The Danish Council for Independent Research (Technology and Production Sciences) and Innovation Fund Denmark (Programme Commission on Strategic Growth Technologies).

Availability of data and materials
Through our webserver at https://rth.dk/resources/crispr/, users can perform the off-target assessment of their gRNAs. This includes off-target predictions with RIsearch2 (v2.1), `CRISPRoff` & `CRISPRspec` score computations and overlapping predictions with known genome annotations. The sourcecode for `CRISPRoff` & `CRISPRspec` score calculation is freely available at https://github.com/rth-tools/crisproff/ [49] and the version from the submission of this article is available as Additional file 2 as well as freely available via the DOI 10.5281/zenodo.1410429. The repositories are released under GNU General Public License v3.0. The generated data used in the publication is also available via DOI 10.5281/zenodo.1410437.

The RIsearch2 program, of which we used version 2.1 for gRNA off-target predictions, is available at https://rth.dk/resources/risearch/.
The sourcecode and data of the Haeussler benchmark study is accessible at https://github.com/maximilianh/crisporPaper [50].
CIRCLE-seq [37] and SITE-seq [38] datasets are accessible through the supplementary material of their corresponding papers.

Authors' contributions
All authors contributed to the project design. FA, CA, and AW wrote the analysis and webserver source code. FA analyzed the data and drafted the full manuscript. All authors critically revised and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

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

Publisher's Note

Received: 5 May 2018 Accepted: 11 September 2018

References
1. Barrangou R. Cas9 Targeting and the CRISPR Revolution. Science. 2014;344(6185):707–8.
2. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, Volz SE, Joung J, van der Oost J, Regev A, Koonin EV, Zhang F. Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System. Cell. 2015;163(3):759–71.
3. Yang H, Gao P, Rajashankar KR, Patel DJ. PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. Cell. 2016;167(7):1814–1828.
4. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DBT, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science. 2016;353(6299).
5. Cong L, Ran FA, Cox D, Lin S, Barretto R, Habib N, Hsu PD, Wu X, Jiang W, Marraffini LA, Zhang F. Multiplex genome engineering using CRISPR/Cas systems. Science. 2013;339(6121):819–23.
6. Deltcheva E, Chylinski K, Sharma CM, Gonzales K, Chao Y, Pirzada ZA, Eckert MR, Vogel J, Charpentier E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature. 2011;471(7340):602–7.
7. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci USA. 2012;109(39):2579–86.
8. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–21.
9. Hsu PD, Lander ES, Zhang F. Development and applications of CRISPR-Cas9 for genome engineering. Cell. 2014;157(6):1262–78.
10. Haeussler M, Concordet JP. Genome Editing with CRISPR-Cas9: Can It Get Any Better? J Genet Genomics. 2016;43(5):239–50.
11. Haeussler M, Schonig K, Eckert H, Eschstruth A, Mianne J, Renaud JB, Schneider-Maunoury S, Shkumatava A, Teboul L, Kent J, Joly JS, Concordet JP. Evaluation of off-target and on-target scoring algorithms and integration into the guide RNA selection tool CRISPOR. Genome Biol. 2016;17(1):148.
12. Fu Y, Foden JA, Khayter C, Maeder ML, Reyon D, Joung JK, Sander JD. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. Nat Biotechnol. 2013;31(9):822–6.
13. Zhang XH, Tee LY, Wang XG, Huang QS, Yang SH. Off-target Effects in CRISPR/Cas9-mediated Genome Engineering. Mol Ther Nucleic Acids. 2015;4:264.

PLOS ONE

RESEARCH ARTICLE

Genome-wide identification of clusters of predicted microRNA binding sites as microRNA sponge candidates

Xiaoyong Pan[1,2], Anne Wenzel[1], Lars Juhl Jensen[1,2]*, Jan Gorodkin[1]*

1 Center for non-coding RNA in Technology and Health, Department of Veterinary and Animal Sciences, University of Copenhagen, Frederiksberg, Denmark, **2** Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark

* lars.juhl.jensen@cpr.ku.dk (LJJ); gorodkin@rth.dk (JG)

OPEN ACCESS

Citation: Pan X, Wenzel A, Jensen LJ, Gorodkin J (2018) Genome-wide identification of clusters of predicted microRNA binding sites as microRNA sponge candidates. PLoS ONE 13(8): e0202369. https://doi.org/10.1371/journal.pone.0202369

Editor: Danny Barash, Ben-Gurion University, ISRAEL

Received: May 30, 2018

Accepted: August 1, 2018

Published: August 24, 2018

Data Availability Statement: All data are fully available at http://rth.dk/resources/mirnasponge.

Funding: This project was mainly financed by University of Copenhagen with additional support from the Novo Nordisk Foundation [NNF14CC0001], the Danish Center for Scientific Computing (DCSC/DEiC), and Innovation Fund Denmark (Programme Commission on Strategic Growth Technologies) [0603-00320B].

Competing interests: The authors have declared that no competing interests exist.

Abstract

The number of discovered natural miRNA sponges in plants, viruses, and mammals is increasing steadily. Some sponges like ciRS-7 for miR-7 contain multiple nearby miRNA binding sites. We hypothesize that such clusters of miRNA binding sites on the genome can function together as a sponge. No systematic effort has been made in search for clusters of miRNA targets. Here, we, to our knowledge, make the first genome-wide target site predictions for clusters of mature human miRNAs. For each miRNA, we predict the target sites on a genome-wide scale, build a graph with edge weights based on the pairwise distances between sites, and apply Markov clustering to identify genomic regions with high binding site density. Significant clusters are then extracted based on cluster size difference between real and shuffled genomes preserving local properties such as the GC content. We then use conservation and binding energy to filter a final set of miRNA target site clusters or sponge candidates. Our pipeline predicts 3673 sponge candidates for 1250 miRNAs, including the experimentally verified miR-7 sponge ciRS-7. In addition, we point explicitly to 19 high-confidence candidates overlapping annotated genomic sequence. The full list of candidates is freely available at http://rth.dk/resources/mirnasponge, where detailed properties for individual candidates can be explored, such as alignment details, conservation, accessibility and target profiles, which facilitates selection of sponge candidates for further context specific analysis.

Introduction

MicroRNAs (miRNAs) are a class of endogenous small non-coding RNAs of about 20 nucleotides in length, which play crucial roles in transcriptional and post-transcriptional control of gene expression through interacting with other RNAs [1–4]. To date, more than 2000 human mature miRNAs (miRBase v20) [5] have been discovered. These mature miRNAs are formed from pre-miRNAs, which are processed by DICER in the cytoplasm [6]. They are estimated to regulate more than 60% of all human protein-coding genes (PCGs) [7] and have been

implicated in many human diseases [8, 9]. With more and more miRNAs being discovered, identifying their functions is becoming increasingly important for understanding the molecular mechanisms of diseases [10].

The miRNAs can themselves be regulated by so-called miRNA sponges, which are RNAs with many miRNA binding sites that compete with the target sites for binding of one or more miRNAs of interest. Denzler et al. analyzed the relationship between sponge activity and number of binding sites and found that more binding sites enhance miRNA sponge effect on releasing mRNA target repression regulated by that miRNA [11]. Artificial miRNA sponges have been used to generate loss-of-function phenotypes for miRNAs in cell culture [12] and to discover miRNA functions *in vivo* [13]. It has advantages over genetic knock-outs and antisense oligonucleotide inhibitors by being cheaper and less time consuming [14]. They are also of therapeutic interest [15, 16].

Natural miRNA sponges with many miRNA binding sites separated by linker regions also exist [17] (S1 Table). They have also been called competing endogenous RNA (ceRNA), and the ceRNA hypothesis suggests that RNAs regulate each other by competing for shared miRNAs [18]. Recently, a circular RNA (circRNA) with more than 70 binding sites was shown to function as a sponge for miR-7 [19, 20]. This natural sponge, named ciRS-7 and CDR1as, has been implicated in cancer-related pathways [21]. A circRNA derived from the gene encoding zinc finger protein 91 (circRNA-ZNF91) with 24 miR-23 binding sites has similarly been identified as a possible miRNA sponge [22]. However, some studies consistently mention that only few circRNAs can function as miRNA sponges [22, 23]. Other types of transcripts can also serve as natural miRNA sponges, such as the pseudogene PTENP1 [24] and the long non-coding RNAs (lncRNAs) H19 [25] and lincRNA-RoR [26]. It has been estimated that there are thousands of RNA transcripts functioning as potential miRNA natural sponges [27], but despite increasing evidence for the ceRNA hypothesis, it still attracts some skepticism [28]. Although there exist several compilations of putative ceRNAs derived from predicted miRNA target sites, CLIP-Seq data, or both [29–31], reviewed in [32], none of the studies to date have systematically analyzed the genome for clusters of miRNA binding sites.

In an attempt to shed more light on this, we here analyze the complete human genome for clusters of predicted miRNA target sites, which may represent natural miRNA sponges. To this end, we identify statistically significant clusters by comparing the numbers of binding sites in the clusters obtained from the real genome and from shuffled genomes, retaining the local sequence composition. We further filter the resulting clusters based on evolutionary conservation and binding energies. With this approach we rediscover one known miRNA sponge ciRS-7 for miR-7 and identify 3672 novel sponge candidates.

Materials and methods

Data sources

The repeat-masked human genome sequence (hg19) was downloaded from the UCSC Genome Browser database [33]. All 2578 human mature miRNAs were extracted from miRBase v20 [5]. GENCODE v19 [34] and circBase [35] were used to annotate sponge candidates, which cover protein-coding genes, lincRNAs, circRNAs, antisense (overlaps a protein-coding locus on the opposite strand), pseudogene, and processed_transcript (a transcript without an open reading frame). To get information about binding site conservation, phyloP (phylogenetic P-values) scores [36] were downloaded from ftp://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/.

Pipeline for clustering of miRNA target sites

To identify genomic clusters of predicted miRNA target sites for a given miRNA, we have developed a pipeline as outlined in Fig 1.

RIsearch2 screen for miRNA target sites. The computational screen on hg19 for target sites of the miRBase miRNAs was performed with a preliminary version of RIsearch2 [37]. It is a seed-and-extend approach to predict RNA–RNA interactions, applying suffix arrays in the first stage to locate initial seed matches (allowing for G–U wobble matches) and using dynamic programming (DP) to extend those matches with the simplified energy model as introduced in RIsearch [38]. The seed was specified to require a stretch of six consecutive bases within the first eight bases of the miRNA sequence to be paired. The window for DP extension was set to always include the entire remaining query sequence outside the seed, and the same number of nucleotides extended by five from the target. This parameter from the preliminary version of RIsearch2 used in here has been replaced with a maximum extension length in the released version. The default value of 20 nt should yield comparable results. The maximum hybridization energy was set to −10 kcal/mol. Overlapping target sites were merged in post-processing.

Markov clustering of predicted miRNA target sites. To identify genomic regions with a high density of predicted binding sites for a given miRNA, we used the Markov Cluster (MCL) algorithm [39], which does not need specify the number of clusters in advance. In our pipeline, we run clustering for individual miRNAs, and each miRNA has different number of clusters.

To this end, we represented the predicted binding sites for each miRNA as a weighted network, in which the weight of the edge between two sites on the same strand of the same chromosome is defined based on their nucleotide distance (x) as follows:

$$sim(x) = \begin{cases} C - x & \text{if } x < C \\ 0 & \text{if } x \geq C \end{cases} \quad (1)$$

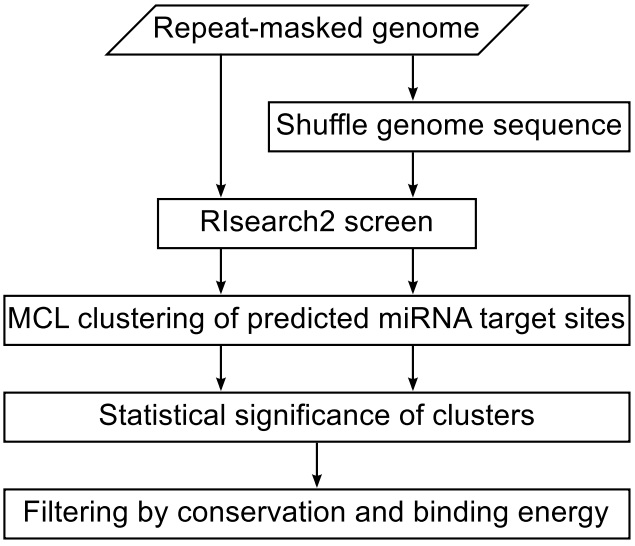


Fig 1. Flowchart of the analysis pipeline. For each mature miRNA in miRBase v20, we ran RIsearch2 against both the real repeat-masked genome and a shuffled version to predict binding sites. We then used the Markov Cluster (MCL) algorithm to identify genomic clusters of binding sites and identified statistically significant clusters by comparing the results for the real and shuffled genomes. Finally, the significant clusters were further filtered by conservation and binding energy.

https://doi.org/10.1371/journal.pone.0202369.g001

where the constant $C = 1000$ determines over which distance the weights decay. The value of C was chosen to allow identification of large clusters while limiting the computational cost. Clusters within this weighted network were then identified using the MCL algorithm with a range of different values for the inflation factor parameter, which influences the size and number of clusters.

Shuffling of genome sequence. We evaluate the statistical significance of the identified clusters by creating a background model from randomized genome sequence and repeating the RIsearch2 and MCL steps described above. To preserve the local dinucleotide content, the non-masked sequence segments of the human genome are shuffled by uShuffle [40] in non-overlapping windows of 120 nt, the typical size of structured RNA [41, 42].

Statistical significance of clusters. For each miRNA, we estimate a cutoff on the number of predicted target sites in a cluster that is required for statistical significance. This is done by fitting the size distribution of the top-10% largest clusters obtained for the miRNA in question on the randomized genome, assuming an exponential tail.

$$\log_{10}(y) = a \cdot x + b \tag{2}$$

where x is the cluster size and y is the number of clusters of a given size. Based on this fit, we extrapolate the largest cluster one would expect to observe in 1000 randomizations.

$$x_{cutoff} = \frac{\log_{10}(1/1000) - b}{a} \tag{3}$$

Only clusters larger than or equal to this cutoff are considered statistically significant. We used this approximation because it would take prohibitively long time to run MCL clustering on the RIsearch2 output for thousands of randomized genomes for every miRNA.

Filtering by conservation and binding energy. To further improve the quality of the predictions, we apply two additional filtering criteria to the statistically significant clusters. First, we extract the evolutionarily conserved subset of target site predictions on the real genome, by requiring that the miRNA seed site has at least five continuous nucleotides with a phyloP score greater than 0.3 [20]. For individual miRNAs, clusters with the percent of conserved sites smaller than 2 times as one would expect to observe in whole genome (the number of conserved binding sites vs the number of all binding sites for this miRNA) are excluded. Second, we filter out statistically significant clusters that are caused by repetitive sequences not masked in the downloaded genome, because RepeatMasker and Tandem Repeats Finder by default only mask simple repeats with a unit length up to 12 and curated repeats from Repbase [43, 44]. Instead of rerunning RepeatMasker with different parameters, we use the binding energies already calculated by RIsearch2 to eliminate clusters for which many sites have the exact same predicted binding energy. To this end, we calculate the normalized Shannon entropy of the binding site energies in each cluster as follows:

$$entropy = -\sum_{i=1}^{m} p_i \cdot \log_2(p_i)/\log_2(n) \tag{4}$$

where m is the number of different binding energy values found in the given cluster, p_i is the relative frequency of a particular energy in there, n is the number of binding sites in the cluster. We used an entropy threshold of 0.6 to filter out clusters that have many sites caused by repetitive sequences.

Characterization of sponge candidates

To characterize the sponge candidates within the web interface, we annotate them with overlapping genes based on their coordinates and strand using genes from GENCODE v19 and circBase. Thus we create a reference annotation in which GENCODE annotation is primary and all annotation from circBase not overlapping that of GENCODE is included as additional annotation. We further calculate a number of properties for each sponge candidate, which are described in the following sections.

SNP density ratio. CircRNAs have significant lower SNP density at miRNA seed sites than in their flanking regions and other sites, suggesting selective pressure to maintain those binding sites [45]. We calculated the SNP density ratio (SDR) between miRNA seed sites and the remaining sequence. For every binding site within a sponge, the seven nucleotides base-pairing with the miRNA seed region (positions 2–8) are defined as miRNA seed site. The gap region between every two seed sites is defined as flanking region, hence also including regions base-pairing with the miRNA outside the seed. The SDR is calculated as ratio between the SNP density in miRNA seed sites and flanking region, shown in Eq (5). SNP data was downloaded from Ensembl 75 [46].

$$SNPdensity = \frac{\#\ of\ SNPs}{\#\ of\ nucleotides}$$
$$SDR = \frac{SNPdensity\ in\ seed\ sites}{SNPdensity\ in\ flanking\ region} \quad (5)$$

Fraction of binding sites within exons. It is furthermore important whether the predicted binding sites are intronic or exonic [22]. We thus calculate the fraction of binding sites in the sponge candidate that fall within exons (FBSE) based on GENCODE v19 as follows:

$$FBSE = \frac{\#\ of\ binding\ sites\ within\ exon}{\#\ of\ binding\ sites\ within\ sponge} \quad (6)$$

Energy and fraction of paired nucleotides for binding sites. For each sponge candidate, we plot the predicted binding site energies and compare them to the distribution of binding site energies for known targets of the same miRNA. We calculated the latter by running RIsearch2 on the 3' UTRs from Ensembl 75 of the known and experimentally identified miRNA targets in the RAIN database [47]. We similarly plot the fraction of paired nucleotides for the binding sites, which is defined as the number of base-pairings between the binding site and the miRNA divided by the length of the mature miRNA.

Visualization of sponge candidates in genome browser. To allow for visualization of sponge candidates in the UCSC browser [33], we display tracks with the position, binding energy, conservation, and accessibility of each binding site. The accessibility track contains the probability of each nucleotide being unpaired within the internal structure of the transcript, estimated using RNAplfold [48] with a maximum base pair span of 120 nt and window size 170 nt.

Results

Clusters of predicted miRNA binding sites

The first step in searching for clusters of predicted miRNA binding sites is to predict the binding sites themselves. To do that several tools can potentially be employed; however, doing a large-scale screen on the complete human genome requires speed. We primarily focus on RIsearch2 as

it was benchmarked against miRNA tools and found to be substantially faster than other RNA–RNA interaction prediction tools [37]. We justify this choice by comparing the RIsearch2-based screen to screens based on a GUUGle search (using a minimum match size of six nucleotides) [49] and a relaxed BLAST search (E-value < 10 000) [50]. For all three tools we subsequently employ the MCL clustering algorithm (with the default inflation factor of 2.0) to identify clusters in the genome with a high density of predicted binding sites for a given miRNA.

We used RIsearch2 to search the 2578 mature miRNAs from miRBase (v20) against the human genome (hg19) and its shuffled counterpart. For each miRNA, we subsequently used MCL to identify clusters of predicted miRNA binding sites and compared the number of clusters obtained on the real and shuffled genomes as function of cluster size (S2 Table). Whereas we were able to perform this analysis for all miRNAs using RIsearch2, this was not feasible for the GUUGle screen due to the large number of predicted miRNA binding sites.

We thus instead, as an example, compare the results from the three methods for miR-7 (Fig 2). For the RIsearch2 screen (Fig 2A) we obtain far more large clusters on the real genome than on the shuffled one. For the BLAST search (Fig 2B) we see a similar but much weaker trend with fewer clusters both on the real and shuffled genome. A possible explanation is that BLAST does not allow G–U base pairing and thus predicts much fewer binding sites. In contrast, GUUGle predicts many more binding sites resulting in more and larger clusters (Fig 2C). However, we observe only very small differences between the real and the shuffled genomes. It should be noted, that GUUGle is intended to be used as a prefilter for more sophisticated but computationally expensive methods. However, as RIsearch2 evaluates the thermodynamic strength of the predicted binding sites and is nonetheless faster [37], we opted to use RIsearch2.

To further illustrate the predictive power of RIsearch2, we compare its predicted miR-7 binding sites to those of TargetScan [51] for the natural sponge ciRS-7 [19]. Both RIsearch2 and TargetScan predict 73 binding sites, of which 72 are in common. The one binding site found by TargetScan, but not by RIsearch2, has a predicted binding energy of −9.87 kcal/mol, which is only slightly above the energy cutoff (−10 kcal/mol). By contrast, the BLAST screen predicts only seven binding sites. While giving comparable results to TargetScan, RIsearch2 runs approximately four times faster.

Whereas RIsearch2 is our preferred miRNA binding sites predictor, we can further improve the results by changing the main parameter in MCL, the inflation factor, which affects the granularity of the clusters. To this end, we ran our pipeline with inflation factors 1.5, 2.0, 2.5, 3.0, 3.5 and 4.0 for each miRNA on both the real and shuffled genomes. These runs took roughly three months to compute on a cluster with 16 nodes, each equipped with two Intel Xeon E5-2650 processors, having a total of 256 cores. For each inflation factor, we pooled the

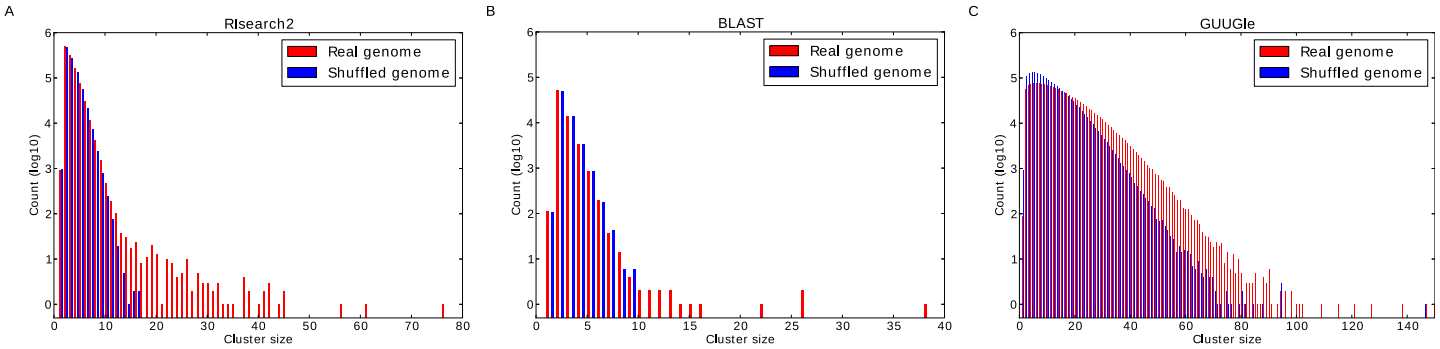


Fig 2. Cluster size distribution for predicted miR-7 binding sites. The plots show the size distributions of the clusters obtained for the real and shuffled genomes when running MCL clustering with an inflation factor of 2.0 on the miR-7 binding sites predicted by (A) RIsearch2, (B) BLAST, and (C) GUUGle.

https://doi.org/10.1371/journal.pone.0202369.g002

resulting clusters for all miRNAs and plotted the cluster size distributions (S1 Fig). As anticipated, higher inflation factors led to smaller clusters. Based on visual inspection of the differences between real and shuffled genomes, we decided to use an inflation factor of 3.5 (Fig 3).

The more predicted miRNA binding sites a cluster contains, the less likely it is to occur in a shuffled genome. We can thus estimate the statistical significance of a cluster based on its size. Since different miRNAs give rise to different cluster size distributions on shuffled genomes (S2 Fig), significance analysis is done separately for each miRNA. In principle, one could estimate the false discovery rate corresponding to a given cutoff on cluster size based on the empirical distribution obtained from thousands of shuffled genomes. However, as this is not computationally feasible, we instead fit the tail of the distribution obtained from a single shuffled genome to find a cluster size cutoff for each miRNA. For example, although the largest miR-7 cluster found in the shuffled genome contains 15 predicted binding sites, we extrapolate that a miR-7 cluster must contain at least 20 binding sites to have less than 0.1% probability of appearing by chance alone. We refer to this as the empirical p-value.

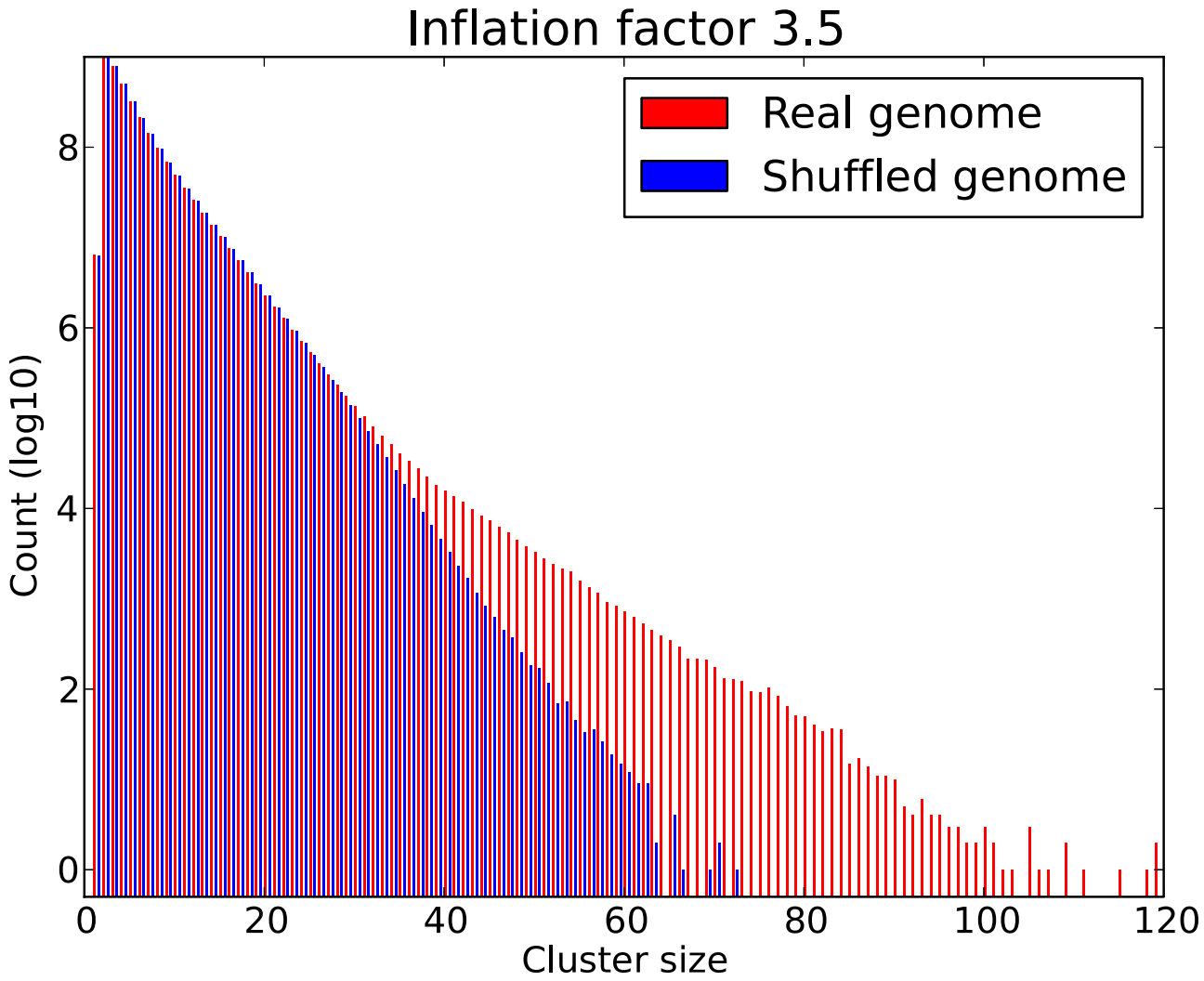


Fig 3. Overall cluster size distribution of miRNA binding sites predicted by RIsearch2. The plot shows the size distributions obtained for real and shuffled genomes when pooling the results for 2578 mature human miRNAs. For each miRNA, we used RIsearch2 to predict binding sites and clustered them using MCL with inflation factor 3.5.

https://doi.org/10.1371/journal.pone.0202369.g003

As a control we also checked if we obtained the same number of predicted miRNA binding sites in the real and shuffled genome. These numbers (at the −10 kcal/mol cutoff) are 3 430 423 948 and 3 386 114 664, with a ratio of 1.013 constituting no bias in the volume of binding sites in the two genomes.

Sponge candidates

Using the pipeline shown in Fig 1 to the second last step, we obtained a total of 71 106 statistically significant ($P < 0.001$) clusters of binding sites for 2543 mature miRNAs (for 35 mature miRNAs we obtained no significant clusters). To identify the ones most likely to be of biological relevance, we employ the last step by filtering for conservation and binding energy (see Methods for details). This reduced the clusters to 3673 sponge candidates for 1250 miRNAs, which can all be viewed and downloaded via our web resource (http://rth.dk/resources/mirnasponge).

To annotate the sponge candidates with presumed genes of origin, we compared the genome coordinates of each sponge to annotated circRNAs from circBase and other genes from GENCODE. We annotated a sponge candidate to a gene if the larger of these two genomic regions covered at least 50% of the smaller one. Given this criterion, the majority of our candidates (2162 out of 3673) fall in unannotated genomic regions, which is not surprising considering that 85.9% of the genome is not covered by either GENCODE or circBase (Fig 4,

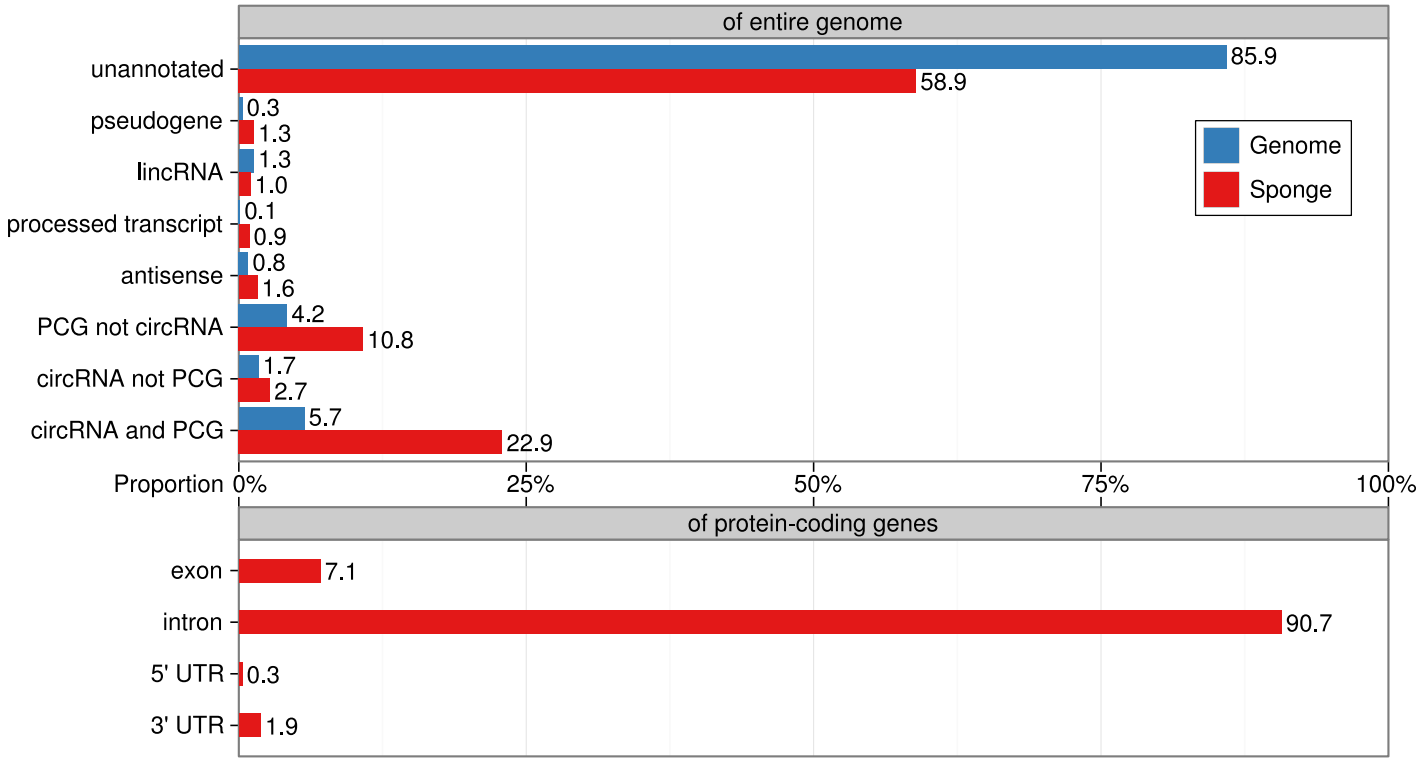


Fig 4. Genomic context of the sponge candidates. The upper bar chart shows the percentage for the different types of transcripts in the genome based on GENCODE and circBase, and their percentage within our sponge predictions are calculated after we assign annotations to the predicted sponge candidates. For each type of transcript, we calculate the percentage of their nucleotides under whole genome and annotated sponges. Then we can evaluate the enrichment via comparing the percent between sponges and whole genome. There is big overlap between PCGs and circRNAs, so we further divide them into "PCG not circRNA", "circRNA not PCG" and "circRNA and PCG". They refer to PCGs not overlapping with circRNAs, circRNAs not overlapping with PCGs, and PCGs overlapping with circRNAs, respectively. The lower bar chart shows the percentage of nucleotides located in intron, exon, 3' UTR, and 5' UTR for all annotated PCG sponge candidates. All percentages are calculated based on the number of nucleotides, excluding masked repeats, and are strand-sensitive.

https://doi.org/10.1371/journal.pone.0202369.g004

Table 1. Filtered miRNA sponge candidates. The table provides an overview of miRNA sponge candidates that have at least 10 binding sites more than what is required for statistical significance, can be annotated with known genes, and have more than 9% of predicted miRNA binding sites within exons. The column *cluster size* lists the number of binding sites in the given cluster for real genome and the cluster size cutoff for statistical significance obtained from shuffling. The sponge candidates are sorted based on the difference between these two cluster sizes.

miRNA	Cluster size real / cutoff		Genomic coordinate chromosome: range (strand)	Annotation
hsa-miR-7-5p	76	20	chrX: 139 865 250–139 866 947 (+)	circRNA: ciRS-7
hsa-miR-4310	54	26	chr22: 50 671 491–50 673 762 (–)	protein_coding: TUBGCP6
hsa-miR-376b-5p	38	18	chr15: 101 093 988–101 095 732 (–)	pseudogene: PRKXP1
hsa-miR-766-5p	70	51	chrX: 139 865 341–139 867 009 (+)	circRNA: ciRS-7
hsa-miR-4295	35	17	chr16: 690 762–691 826 (–)	pseudogene: AL022341.1
hsa-miR-4729	44	27	chr16: 90 060 979–90 062 561 (+)	circRNA: hsa_circ_0041137
hsa-miR-190b	38	21	chr17: 412 328–413 728 (+)	antisense: RP5-1029F21.3
hsa-miR-93-3p	61	46	chr16: 90 060 884–90 062 555 (+)	circRNA: hsa_circ_0041137
hsa-miR-8077	52	38	chr17: 80 211 257–80 213 687 (–)	circRNA: hsa_circ_0046395
hsa-miR-545-3p	32	18	chr10: 133 771 747–133 773 035 (+)	protein_coding: PPP2R2D
hsa-miR-4712-3p	38	24	chr12: 50 745 528–50 747 302 (–)	protein_coding: FAM186A
hsa-miR-433-5p	44	30	chr16: 600 512–601 669 (–)	antisense: LA16c-366D1.3
hsa-miR-649	30	18	chr8: 142 262 432–142 264 762 (–)	circRNA: hsa_circ_0001829
hsa-miR-219b-5p	38	27	chr22: 21 537 410–21 538 848 (+)	processed_transcript: FAM230B
hsa-miR-6761-5p	49	39	chr14: 107 147 088–107 148 903 (–)	circRNA: hsa_circ_0033997
hsa-miR-649	28	18	chr9: 139 996 161–139 997 435 (+)	circRNA: hsa_circ_0089635
hsa-miR-4668-3p	27	17	chrX: 139 865 277–139 866 621 (–)	protein_coding: CDR1
hsa-miR-5692b	20	10	chr3: 195 607 265–195 609 187 (–)	circRNA: hsa_circ_0001377
hsa-miR-34a-3p	31	21	chr10: 133 770 676–133 771 926 (+)	protein_coding: PPP2R2D

https://doi.org/10.1371/journal.pone.0202369.t001

upper panel). Most notably, the predicted sponges are enriched for overlaps with circRNAs and protein-coding genes. In particular the circRNA sponges overlapping with PCGs have four times (22.9%) as many sponges as what one would expect by chance (5.7%) from the genomic annotation. The overlaps of sponge candidates with specific parts of protein-coding genes (intron, exon, 3' UTR, and 5' UTR) are shown in Fig 4 (lower panel).

To identify a subset of sponge candidates of particular interest, we first select the 768 sponge candidates that have at least 10 predicted binding sites more than what is required for the significance cutoff. Of these we focus on the subset that could be annotated with known genes and further require that at least 9% of the predicted binding sites reside within exons (FBSE $>$ 0.09, which is the mean value of FBSEs for sponge candidates annotated with known genes). These sponge candidates are listed in Table 1 and include the known natural sponge ciRS-7.

Web interface

The predicted sponge candidates are freely available through a web interface at http://rth.dk/resources/mirnasponge. The interface provides the ability to search for sponge candidate for a particular miRNA of interest as well as to download the full set of sponge candidates for all miRNAs. An example for miR-7 is shown in Fig 5A. For each sponge candidate, we provide detailed information related to properties of natural miRNA sponges to assist in prioritization, including alignment details from RIsearch2, FBSE, SDR, and accessibility and target profiles with links to the UCSC genome browser (e.g. Fig 5B for miR-7 sponge ciRS-7). The help page provides a detailed explanation of all these properties.

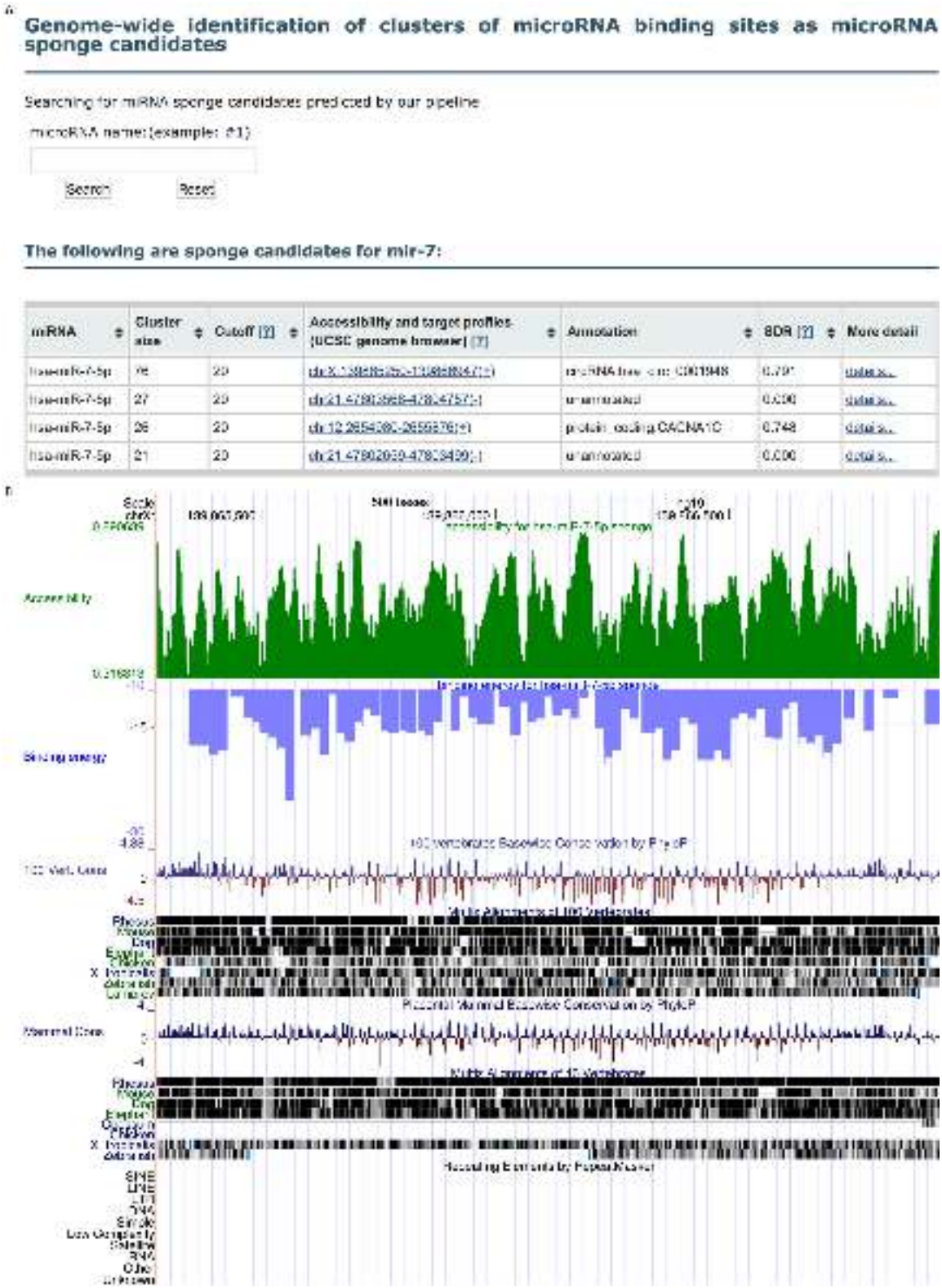


Fig 5. Web resource of miRNA sponge candidates. To illustrate the web resource, we show the results for miR-7. (A) When searching for a miRNA, the user is presented with an overview table of the corresponding miRNA sponge candidates. In case of miR-7, our pipeline suggests four sponge candidates, the top scoring of which is the known sponge ciRS-7 (named hsa_circ_0001946 in circBase). Clicking detail opens a page with detailed properties of this sponge candidate. (B) Clicking the coordinate of a sponge opens the UCSC

genome browser with tracks showing conservation, accessibility profile (probability of the bases being unpaired in the RNA structure), and target profile (binding energies for the miRNA as predicted by RIsearch2). In the example, we show the region chrX: 139 865 280–139 866 947 (+), which corresponds to ciRS-7.

https://doi.org/10.1371/journal.pone.0202369.g005

Discussion

Natural miRNA sponges have recently attracted much attention due to the discovery of increasing numbers of novel miRNA sponges, some of them have been linked to diseases [24, 52]. Experimental verification is still highly time-consuming and prohibitively expensive to perform systematically on a genome-wide scale. For this reason databases such as starBase v2.0 and lnCeDB rely on computational predictions of miRNA sponges; however, because the methods employed focus purely on annotated genes, they are unable to identify other genomic regions that may function as miRNA sponges.

Our study is based on the hypothesis that groups of adjacent binding sites may function together as miRNA sponges. This is consistent with a recent study showing that the number of miRNA binding sites within a sponge correlates with its ability to derepress targets of the miRNA *in vivo* [11]. We have made a genome-wide computational screen that detected potential sponge candidates through clustering of nearby miRNA binding sites. In total, we identified 3673 sponge candidates spanning 1250 miRNAs. The genome-wide analysis was made possible by the RNA–RNA interaction prediction tool RIsearch2, which can predict miRNA binding sites with high speed and with accuracy comparable to that of other methods. Although we compare cluster sizes between real and shuffled genomes and select candidates based on the differences in sizes, further refinement involves calculating p-values over all cluster sizes. However, this will require comparison across clusters based on different miRNAs as some due to their composition might have different clusters sizes than others. To the best of our knowledge, this is the first genome-wide computational approach to predict sponge candidates specifically based on high binding site density in genomic regions.

There are, however, still many limitations to prediction of miRNA sponges. Like all tools for predicting miRNA binding sites, RIsearch2 produces many true and false predictions. The latter can give rise to false predictions of miRNA sponges if they appear clustered in the genome, although our filtering steps do much to alleviate this problem.

In conclusion, we have presented a computational pipeline for discovery of clusters of putative miRNA binding sites. Interestingly, we observe an enrichment ($\sim$2.5-fold) of clusters in protein-coding sequence which is not also annotated as circular RNA. For clusters overlapping sequence annotated both circular RNA and protein-coding sequence we observed an even stronger enrichment ($\sim$4-fold). Both competing endogenous RNA (mRNA) and circular RNA have previously been reported to compete for miRNA binding. Hence, we consider our clusters of miRNA binding sites as miRNA sponge candidates and we in particular obtain intriguing candidates overlapping known genes.

Supporting information

S1 Fig. Overall cluster size distributions of predicted binding sites using different inflation factors. The cluster size distributions shown for both real and shuffled genomes were obtained by pooling results for all 2578 mature human miRNAs. For each miRNA, we used RIsearch2 to predict binding sites and clustered them using MCL with different inflation factors (1.5, 2.0, 2.5, 3.0, 3.5 and 4.0).
(EPS)

Kirsch *et al. BMC Genomics* (2018) 19:899
https://doi.org/10.1186/s12864-018-5234-4

BMC Genomics

RESEARCH ARTICLE Open Access

Identification and characterization of novel conserved RNA structures in Drosophila

Rebecca Kirsch[1,2,7], Stefan E. Seemann[1,2], Walter L. Ruzzo[1,3,4,5], Stephen M. Cohen[6], Peter F. Stadler[7,8,9,10,1,11] and Jan Gorodkin[1,2*]

Abstract

Background: Comparative genomics approaches have facilitated the discovery of many novel non-coding and structured RNAs (ncRNAs). The increasing availability of related genomes now makes it possible to systematically search for compensatory base changes – and thus for conserved secondary structures – even in genomic regions that are poorly alignable in the primary sequence. The wealth of available transcriptome data can add valuable insight into expression and possible function for new ncRNA candidates. Earlier work identifying ncRNAs in *Drosophila melanogaster* made use of sequence-based alignments and employed a sliding window approach, inevitably biasing identification toward RNAs encoded in the more conserved parts of the genome.

Results: To search for conserved RNA structures (CRSs) that may not be highly conserved in sequence and to assess the expression of CRSs, we conducted a genome-wide structural alignment screen of 27 insect genomes including *D. melanogaster* and integrated this with an extensive set of tiling array data. The structural alignment screen revealed ~30,000 novel candidate CRSs at an estimated false discovery rate of less than 10%. With more than one quarter of all individual CRS motifs showing sequence identities below 60%, the predicted CRSs largely complement the findings of sliding window approaches applied previously. While a sixth of the CRSs were ubiquitously expressed, we found that most were expressed in specific developmental stages or cell lines. Notably, most statistically significant enrichment of CRSs were observed in pupae, mainly in exons of untranslated regions, promotors, enhancers, and long ncRNAs. Interestingly, cell lines were found to express a different set of CRSs than were found in vivo. Only a small fraction of intergenic CRSs were co-expressed with the adjacent protein coding genes, which suggests that most intergenic CRSs are independent genetic units.

Conclusions: This study provides a more comprehensive view of the ncRNA transcriptome in fly as well as evidence for differential expression of CRSs during development and in cell lines.

Keywords: Non-coding RNA, RNA secondary structure prediction, *Drosophila melanogaster*, `CMfinder`, Gene expression, Development

Background

Over the last decade our understanding of the functioning of eukaryotic genomes has changed profoundly. The vast majority of the DNA sequence is transcribed into RNA, and protein-coding sequences comprise only a fraction of the informational content encoded by RNA [1, 2]. This is true for mammals as well as for simple model organisms such as yeast [3].

The functions of the vast majority of these transcripts are unknown. The fact that much of the transcriptional output is poorly conserved at the sequence level initially led to doubts that this pervasive transcription was more than just irrelevant "Junk RNA" [4]. A growing body of evidence, however, showed that many non-coding transcripts are under selection acting at the RNA level. One line of evidence is based on the conservation of gene structures [5]. Another traces the evolution of RNA secondary structure elements [6].

*Correspondence: gorodkin@rth.dk
[1]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
[2]Department of Veterinary and Animal Science, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark
Full list of author information is available at the end of the article

Many ncRNAs compiled in the Rfam database exhibit well-conserved RNA secondary structures. Independent ncRNAs such as transfer RNAs (tRNAs), small nuclear RNAs (snRNAs), ribosomal RNAs (rRNAs), small nucleolar RNA (snoRNAs), and microRNAs (miRNAs) constitute only a minute fraction of the genome. However, structured RNA motifs are much more widespread. Regulatory features of mRNAs, such as internal ribosome entry site (IRES) and selenocysteine insertion sequence (SECIS) elements, aptamer domains of riboswitches, or the autoregulatory domains of many messenger RNAs (mRNAs) that encode ribosomal proteins also have recognizable secondary structures [7].

The presence of a stable secondary structure is not in itself a sufficient indication that the RNA has a function: Random RNA sequences typically fold into highly complex secondary structures that are not statistically different from known functional elements [8–11]. Therefore, it is necessary to assess the evolutionary conservation of secondary structures.

A variety of computational methods have been developed to identify negative selection acting on RNA structure. Methods starting from multiple sequence alignments include `qrna` [12], `AlifoldZ` [13], `EvoFold` [14], `RNAz` [15], and `SISSIz` [16]. Their main limitation is the need for reliable sequence-based multiple sequence alignments. This can be partially overcome by methods that align or re-align (presumably) homologous sequences using sequence and structure simultaneously. A widely used tool of this type is `CMfinder` [17, 18]. A pipeline centered around `FoldAlign` [19] uses the same basic logic. We refer to [20] for a review. `REAPR` is an improved method that shares the idea of structure-based re-alignment of regions with the approach pursued here. It achieved a doubling of sensitivity and confirmed a substantial number of its predictions as transcripts [21].

Studies on ncRNA gene families of fruit flies have a long history. Well understood and well conserved ncRNA families, such as miRNAs [22], have frequently been used as model systems to study ncRNA evolution [23, 24]. Several previous experimental as well as computational surveys have suggested that the fruit fly and related insect species still harbor large numbers of unexplored ncRNAs. For example, thousands of long ncRNAs (lncRNAs) were found using deep sequencing [25–27]. A study focusing on 3'-untranslated regions (UTRs) found 184 ncRNA clusters [28]. A quarter of the genomic regions that currently lack annotated genes, i.e., that are currently considered "intergenic", show transcriptional activity (according to the intersection of the current annotation [29] and a genome-wide tiling array study [30]). Most likely, these regions harbor still undescribed transcripts. We consider this value of one quarter as a lower bound since it is unlikely that any individual study captures the complete transcriptome.

A computational screen for structured RNAs using `RNAz` identified about 16,000 candidate RNA elements with an estimated false discovery rate (FDR) of about 40% [31]. However, `RNAz` evaluates only the most conserved parts of genomic sequence alignments and is optimized for specificity. Subsequent computational analyses of mammalian genomes indicate that the number of functional RNAs is most likely considerably higher: Up to a fifth of the genome may be under selection for RNA structure, but only a tenth of these loci show evidence of selection for conservation of nucleic acid sequence [6, 18, 32].

Several computational surveys of structured RNAs [15] have confirmed the presence of large numbers of conserved structured RNA elements in fruit flies, notably a more detailed `RNAz`-based screen [31] and a comparison of several grammar-based methods [33]. `RNAz` likely underestimates the number of conserved RNA structures in flies similar to the situation in mammals. A subsequent study concentrating on coding regions, furthermore, suggests that these also harbor many superimposed RNA structures [34]. `CMfinder` takes this approach a step further by joint folding and structure-based re-alignment of genome sequences [35]. To date, the newest generation of computational ncRNA screening methods have not been applied to fly genomes. We close this gap here and provide a map of conserved RNA structures (CRSs) in the fruit fly *Drosophila melanogaster*. Furthermore, we associate CRSs with expression across all developmental stages in fly as well as expression in cell lines, which has not been done before.

Results

Summary of the `CMfinder` screen

We predicted CRSs on the genomic sequences of 23 drosophilid and four additional insect species extracted from the UCSC Genome Browser (see "Methods" section for details). Multiple alignment blocks shorter than 50 bp or containing fewer than three sequences were removed. Within each alignment block, the sequence-based alignment was ignored and the unaligned sequences were fed to `CMfinder`. A total of 345,285 `CMfinder` predictions passed our filter criteria including a minimum *pscore* [35] $p > 50$ and a minimum element size of 30 nt (see "Methods" section for details). Salient features of these candidates are summarized in Additional file 1: Figure S1: The majority of the predictions had folding energies in the range of about −10 kcal/mol and were shorter than 100 nt.

This fits well with the properties of most of the small structured ncRNA genes and most of the well-known functional RNA elements in mRNAs. Their GC content

lay mainly between 30 and 60% with a tail more pronounced towards 20%, commensurate with the comparatively low overall GC content in drosophilid genomes [36]. A sequence identity of 40 to 80% reflects the lower sequence conservation in structurally conserved RNAs. Most predictions were found in 17 to 23 of the 27 genomes examined. All CRSs are available at https://rth.dk/resources/rnannotator/crs/insect/.

Out of the 345,285 initial predictions, 12,421 overlapped an annotated repeat by at least 50% of their length. These were removed from further processing because the input alignments are unreliable in repetitive regions (see e.g. [37, 38]). While the initial predictions were obtained with a uniform cut-off for `CMfinder`'s *pscore*, previous applications of `CMfinder` to vertebrate genomes have shown that the false discovery rate (FDR) strongly depends in particular on the GC content and the average sequence identity of the input alignments. This is also the case for the fruit fly data (Additional file 1: Figure S2). To evaluate the influence of these two parameters we partitioned the set of repeat-filtered predictions into bins with narrow ranges of both GC content and sequence identity. We independently estimated the FDR for each subset (see "Methods" section for details). Requiring in addition a $pscore > 80$, we observed that the resulting FDR estimates remained below 0.1 in most of the bins (Additional file 1: Figure S3), and predictions with a wide range of sequence identities were included in the remaining set (Additional file 1: Figure S2).

We observed a moderate increase of the FDR with GC content. Given the overall low GC content in drosophilid genomes, this fortunately does not constitute a substantial problem. It is also worth noting that `CMfinder` loses its power at sequence similarities below 40%: In this range, the FDR increased up to 0.5. Computing the FDR for bins depending on GC content, sequence identity, and *pscore* and using an FDR cutoff of 0.1, we retained 46,024 sequences for further analysis. 28% of these showed sequence identities below 60%, constituting promising CRSs candidates. Additional file 1: Figure S2 summarizes the number of CRS predictions as a function of FDR.

To see whether the sequence characteristics of the Rfam elements create a different error profile than seen globally, we re-analyzed a subset of the screen-wide FDR data, namely, the `CMfinder` results from the simulated MAF blocks containing the 527 Rfam elements summarized in Table 2. They yielded only 9 predictions ($pscore > 80$); none corresponded to any of our 93 "positive" Rfam predictions. There were 76 predictions in the native alignments of those regions, yielding an estimated FDR $< 12\%$, in line with our global estimate. For details on FDR estimation we refer to the "Methods" section.

These initial CRS candidates were obtained from independent predictions on both strands. Owing to the near symmetry of RNA secondary structures, it is difficult to distinguish the reading direction of conserved RNA elements [39]. Furthermore, there is no reliable way to identify whether a single predicted element reflects a product from only one strand or if structured functional elements are produced by both strands. The latter has been described for the mir-iab-4 locus [40, 41]. Here, we made a conservative estimate by merging overlapping predictions on opposite strands, so that each genomic locus is assumed to produce one product. Since `CMfinder` searches for local structures and the available genome-wide alignments consist of many often very small blocks, we also merged adjacent elements that are separated by less than 30 nucleotides. This threshold is larger than the usual size of "holes" between consecutive alignment blocks but much smaller than the minimum distance between adjacent known ncRNAs, such as miRNAs in polycistronic clusters (Additional file 1: Figure S5). As a result, we estimated that 30,710 genomic loci encode conserved RNA structures.

Annotation

Half of the predicted motifs were located in introns. This amounts to a slight enrichment compared to a uniform genomic distribution of CRS loci (Table 1). Introns often harbor ncRNAs that are processed from the host transcript [42]. In particular, several vertebrate snoRNAs are encoded in introns of ribosomal genes, allowing the snoRNA and the functionally closely related host gene to be co-expressed efficiently [43]. In addition, choice of splice sites and regulation of alternative splicing frequently involves secondary structures [44–47]. Hence, intronic CRSs constitute interesting candidates for structural elements of novel functional ncRNAs. Both UTRs of coding transcripts and the exonic parts of non-coding transcripts showed significant enrichments: 3.5% of the predictions fell into 5'-UTRs and roughly twice as many predictions in 3'-UTRs, representing a 1.5 and 1.6-fold enrichment, respectively, and the 661 loci (2.2%) in ncRNA exons constituted a 1.5-fold enrichment. In contrast, predicted motifs were under-represented in protein-coding exons (7.5%, 0.38-fold enrichment). This likely reflects the fact that coding exons are more conserved in the primary sequence than in their RNA secondary structure. About a quarter of the CRSs were found in intergenic regions and may belong to yet unknown transcripts.

The `CMfinder` predictions overlapped with 93 of the 527 ncRNAs annotated in Rfam and contained in the input alignments after repeat filtering (Table 2). This yields an estimated sensitivity of about 18% and an FDR of about 10%. We observed strong enrichments for miRNAs, H/ACA-box snoRNAs, composite snoRNAs

Table 1 Overlap of CRSs with the *Drosophila melanogaster* genomic FlyBase annotation (dmel_r6.15, FB2017_02)

Feature	Number of CRSs overlapped	Percentage of CRSs overlapped	Fold enrichment	*P*-value	Total feature number	Number of features overlapped	Percentage of features overlapped
Exon coding	2294	7.5%	0.38	1.0	57906	2375	4.1%
Exon 5′-UTR	1082	3.5%	1.53	$4 \cdot 10^{-12}$	16930	1242	7.3%
Exon 3′-UTR	2409	7.8%	1.63	$6 \cdot 10^{-113}$	11288	1766	15.6%
Exon both UTRs	13	0%	1.13	0.72	415	16	3.9%
Exon ncRNA	661	2.2%	1.50	$4 \cdot 10^{-15}$	4120	588	14.3%
Intron	15565	50.7%	1.24	$3 \cdot 10^{-221}$	52410	6507	12.4%
Intergenic	8639	28.1%	1.12	$5 \cdot 10^{-26}$	12348	2924	23.7%
Unmapped alignment blocks	47	0.2%	–	–	1152	–	–

Annotation tracks were unified to avoid overlapping annotation elements and thereby ambiguous assignment of annotation categories to CRSs. In this context, annotation positions with overlapping 5′- and 3′-UTR exons have been collected in the "Exon both UTRs" category (see "Methods" for details). Predictions overlap the unified annotation feature by at least 1 nt, not considering strands. Prediction counts are given as rounded fractions according to the number of unified annotation features they overlap with. Percentages give the fraction of overlapping from total predictions. Fold enrichments and significance were calculated based on the annotation features contained in the `CMfinder` input alignments

(scaRNAs), snRNAs, as well as cis-regulatory elements. tRNAs showed moderate enrichment. Especially retroelements and stable intronic sequence RNAs (sisRNAs) as well as some tRNAs and H/ACA-box snoRNAs are located in short alignment blocks that had been removed prior to the `CMfinder` run and hence have been excluded as not contained in the input. rRNAs in addition often overlap repeats and also have been filtered out based on this criterion. The enrichment within the remaining lncRNA, C/D-box snoRNA and the histone 3′-UTR stem-loop annotations was not as strong as for other ncRNA classes, fitting the notion of these RNAs being less structured. Of the two ribozymes annotated in Drosophila, we recovered the nuclear Ribonuclease P (RNase P). Some of the predicted motifs may be associated with ncRNAs that are not completely contained in input alignment blocks and thus are not included in the list of known RNAs. The overlap thus is likely a bit higher than reported

Table 2 Overlap of CRSs with the *Drosophila melanogaster* Rfam annotation (v.12.2)

Feature	Total feature number	Filtered feature number	Number of features overlapped	Percentage of filtered features overlapped	Fold enrichment	*P*-value	Number of CRSs overlapped
tRNA	294	247	32	12.9%	4.53	$7 \cdot 10^{-30}$	32
miRNA	92	85	18	21.1%	7.57	$7 \cdot 10^{-28}$	17
rRNA	156	6	1	16.6%	0	0.11	1
C/D-box snoRNA	45	41	2	4.8%	1.76	0.16	2
H/ACA-box snoRNA	27	14	4	28.5%	7.73	$1 \cdot 10^{-5}$	4
scaRNA	6	6	3	50.0%	18.04	$6 \cdot 10^{-8}$	3
snRNA	33	29	20	68.9%	21.15	$2 \cdot 10^{-55}$	20
lncRNA	15	15	2	13.3%	4.81	0.01	2
Cis-regulatory element	20	15	6	40.0%	12.03	$8 \cdot 10^{-14}$	7
Signal recognition particle RNA	4	4	0	0%	–	–	0
Histone 3′-UTR stem-loop	71	62	4	6.4%	2.33	0.03	4
Ribozyme	2	2	1	50.0%	18.04	0.04	1
Retroelements	121	1	0	0%	–	–	0
All	886	527	93	17.6%	5.89	$2 \cdot 10^{-102}$	93

Annotations with a base pair content of less than 30% were excluded. Predictions overlap the annotation feature by at least 50% of the prediction or the annotation feature size. *Filtered features* were filtered for features lying at least 50% of their size within the `CMfinder` input alignment blocks and overlapping a repeat by less than 50% of their size. The `CMfinder` input alignments did not contain sisRNAs, hence these are not listed here

here. In any case, the overlap with Rfam is highly statistically significant overall, and in all but the smallest Rfam sub-categories (Table 2).

Overlap with other ncRNA screens

We compared the results of the CMfinder screen with previous surveys for drosophilid ncRNAs using EvoFold [48], REAPR [21], and RNAz [31, 49] in Table 3. The Sandmann RNAz data were filtered more stringently in order to identify specifically miRNAs and are therefore much more sparse than the predictions from the other screens. Considering only the less restricted screens using RNAz, EvoFold and CMfinder, the proportion of the overlaps is similar to what was observed using these tools in human [18, 50].

The overlaps between surveys conducted with different methods are surprisingly small. However, assuming that the amount of sequence covered by predictions is small compared to the size of the genome, the expected overlap of two independent surveys of the same genome is the product of their sensitivities: $0.18 \times 0.65 = 0.12$ for our CMfinder screen and the Rose RNAz survey. However, both screens were performed using different genome releases, annotation versions and criteria with different FDRs. Therefore, the expected and the actual overlap between the screens are not directly comparable. A large overlap is observed only between the Rose RNAz screen and the REAPR predictions, which, however, are not independent of each other.

Figure 1 shows that the CMfinder predictions are more similar to the RNAz predictions than to EvoFold data in terms of GC content and sequence conservation. The predictions of both methods cover a broad range of sequence conservation, while the phylogeny-based EvoFold method shows a strong preference for highly conserved predictions. However, alignment blocks with low sequence conservation are much less prevalent among the CMfinder predictions than among the RNAz predictions. An explanation for this difference can be inferred from a comparison of the situation in drosophilids to the one in vertebrates. In a genome-wide CMfinder screen in vertebrates [6], most of the predictions had a sequence identity between 60 and 70%, comparable with the drosophilid CMfinder predictions reported here (Additional file 1: Figure S1). However, the input alignments used in the vertebrate and drosophilid CMfinder screens differ greatly. In the vertebrate screen, only 10% of the input aligments overlapped annotated phastCons highly conserved elements [51]. Still, this small fraction of the input gave rise to 50% of all predicted CRSs [6]. In contrast, in fruitflies about 65% of the input alignments overlapped phastCons conserved elements. Hence it is not surprising that the vast majority of the Drosophila CRS predictions are located in highly conserved regions. The larger sequence variation in RNAz predictions might be explained by the higher false discovery rate of the tool. Specifically, the predictions with low phastCons scores may contain more false positives.

Table 3 Pairwise overlaps between predictions of the CMfinder and four additional screens for ncRNAs in drosophilids [21, 31, 48, 49]

	CMfinder	EvoFold	REAPR	RNAz(R)	RNAz(S)
CMfinder	**30710**	1618	3355	3967	410
EvoFold	1655	**22682**	2893	3583	331
REAPR	3340	2807	**30478**	19119	687
RNAz(R)	3993	3499	19358	**42479**	905
RNAz(S)	408	325	686	896	**2469**

RNAz(R) and *RNAz(S)* refer to the RNAz-based screens by Rose et al. [31] and Sandmann et al. [49], respectively. Given are the numbers of predictions in screen A (rows) that overlap predictions of screen B (columns) by at least 1 bp. Boldface values in the diagonal state the number of predictions in each dataset

Expression

To assess whether the predicted structures are likely to represent transcripts with real functions, we used expression data as a filter. Tissue and developmental stage-specific expression may be a good indication of biological function. We employed the modENCODE genome-wide tiling array dataset, which has a resolution of 38 bp, an exon expression score threshold of 300 (median of probe intensities for all probes found within that exon, normalized for cell lines), and consists of samples from 30 developmental stages and several Drosophila cell lines (both polyA+ and total RNA) [30, 52]. In the following, a CRS or genomic feature is categorized as expressed if it overlaps any tiling array transcript region by at least 50% of its size.

Of all CRSs expressed in at least one experiment (20,184), approximately a sixth showed expression throughout most stages and cell lines (Fig. 2). In contrast, the majority of CRSs are expressed in specific contexts. Expression patterns formed two clusters, separating cell line data from expression in flies. While developmental stages were not perfectly clustered together, there were some clear groupings: The six prepupal stages (yellow color in the stage annotation line) fell into an almost separate group. Five of the adult stages (red color in the stage annotation line) were grouped together, with similar CRS patterns in the adult female sample five days after eclosion and the mated ovary (see Additional file 1: Figure S6). The embryonic stages fell into several distinct clusters but were in general separate from other developmental stages (blue color in the stage annotation line).

CRSs that overlap annotated ncRNAs did not fall into obvious clusters with ncRNA classes. It is worth noting that annotated ncRNAs that overlap CRSs preferentially showed nearly ubiquitous expression. Indeed,

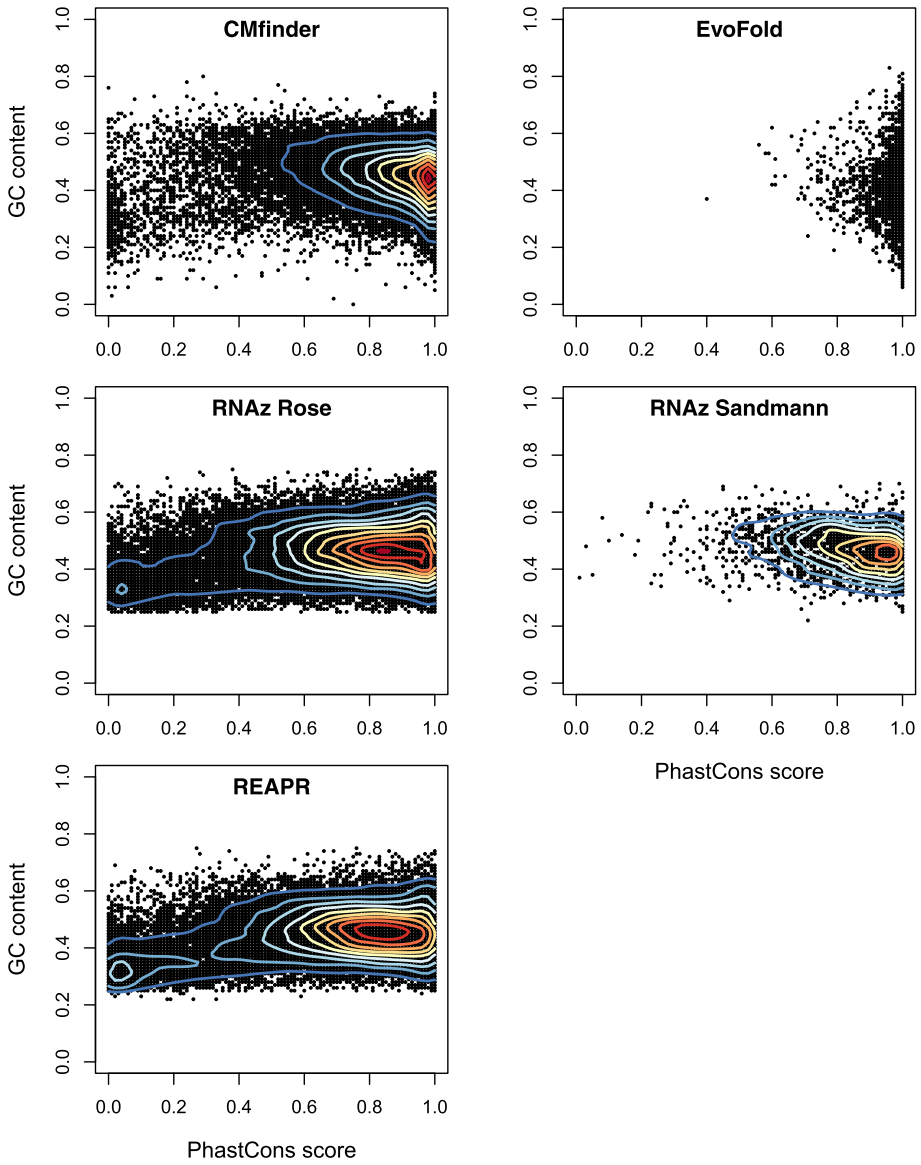


Fig. 1 GC contents and sequence conservation measured in terms of `phastCons` scores [51] of the structured RNA predictions from `CMfinder`, `EvoFold`, `REAPR`, and `RNAz` (Rose, Sandmann) screens [21, 31, 48, 49]

well-expressed transcripts are expected to be found and annotated more easily than sparsely transcribed genes.

Since CRSs are by definition expected to function at the level of RNA, we expect that CRSs are preferentially associated with expressed genomic regions. To test this hypothesis, we used the modENCODE tiling array data to assess the association of CRSs and expression in 100-bp windows sampled from the *D. melanogaster* genome. To avoid a bias due to the more abundant expression of protein coding loci, we removed all loci overlapping coding as well as UTR exons from the analysis. We did not exclude intronic loci, however, because intronic regions are often expressed as independent transcriptional units [53–56]. We observed a systematic enrichment of expression among CRS predictions ($p < 0.05$, Fisher's exact test). This result was independent of whether "expressed" was defined as a tiling array signal in a single experiment or whether a minimum number of positive tiling array data were required (Fig. 3).

Co-expression of intergenic CRSs and adjacent genes

Of particular interest are predictions of motifs for which there has been no functional evidence so far, i.e., in regions annotated as intergenic, but for which expression signals are observed. If a motif shows co-expression with its closest annotated gene, this might suggest a functional

Fig. 2 Expression of CRSs according to publicly available modENCODE tiling array experiments with an exon expression score threshold of 300 (median of probe intensities for all probes found within that exon, normalized for cell lines) [30, 52]. Only CRSs showing at least 50% overlap with at least one transcript region are considered. Non-coding annotations were obtained from Rfam (v.12.2). A subset of the samples is derived from cell lines, which almost form an individual, although heterogenous, cluster (first annotation line). All other samples are derived from flies of one of two different strains (second annotation line) and different developmental stages (third annotation line). Some of the fly samples are derived from specific compartments (fourth annotation line). The fly samples form much more homogenous clusters according to their stages and compartments. *Cell lines:* Blue: Embryo-derived. Red: Derived from prothoracic, mesothoracic, imaginal, antenna or eye-antenna disc. Yellow: Derived from central nervous system. *Strains:* Blue: Yellow cinnabar brown speck. Red: Oregon-R-modENCODE. *Stages:* Blue: Embryos. Green: Larvae. Yellow: Prepupae. Red: Adults. *Compartments:* Green: Gut. Red: Mated ovary. Yellow: Virgin Head

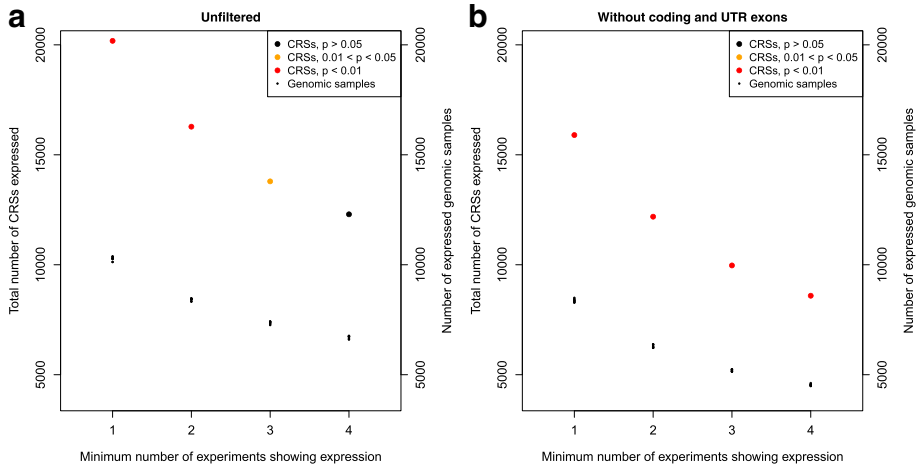


Fig. 3 Significant enrichment for expression among CRS predictions. Expression is defined by a minimum number of modENCODE tiling experiments that show expression (x axis). As background we used 10 samples of randomly selected genomic loci of the same size and number as CRS predictions. The analysis was performed unfiltered (**a**) and filtered to exclude CRSs and genomic samples overlapping protein-coding and UTR exons ($\geq$ 1 bp) to avoid mRNA exon bias (**b**). Significance of the enrichment of expressed CRSs is determined by the highest *p*-value from 10 samples calculated by Fisher's exact test

relationship. One possibility is that the predicted motif could be part of an incompletely annotated UTR. Alternatively it might reflect a novel transcript. As a measure of co-expression based on the modENCODE tiling array data we used a co-expression score that compares the number of tiling experiments in which a `CMfinder` prediction and its closest gene element (UTR or ncRNA exon) are expressed together or individually. The score E_{co}, see Eq. (1) (see "Methods" section), is the difference between the following two ratios: The number of experiments in which both CRS and closest gene (independent of the genomic distance) are expressed normalized by the total number of experiments with CRS expression (Ratio 1); and the number of experiments where the closest gene but not the CRS is expressed normalized by the total number of experiments without CRS expression (Ratio 2). With the help of this score we can determine whether co-expression suggested by the tiling array data is positive or negative. For perfect positive co-expression, Ratio 1 equals 1 (the CRS is exclusively expressed together with its closest gene element), and Ratio 2 equals 0 (the gene element is never expressed without the respective CRS). As a consequence, the difference of both ratios is 1. For negative co-expression the situation is the converse, resulting in a co-expression score of -1. To raise the reliability of the scores, only CRSs that are expressed in at least 3 tiling experiments were considered here.

The majority of CRSs showed a co-expression score of exactly 0, indicating that their expression was not related to that of their closest gene elements (Fig. 4). Distinguishing co-expression signals from noise is a challenge especially for co-expression scores close to 0. In theory, we would assume functional positive co-expression only at a perfect co-expression score of 1 since the adjacent gene can only be expressed if the activating CRS is present as well. However, due to known biases of tiling arrays against sequences with low GC contents and very stable secondary structures [18, 50] we cannot expect complete detection of all expressed transcripts. Therefore, we empirically chose score cutoffs of ≥ 0.5 and ≤ -0.5 for positive and negative co-expression, respectively (also see Additional file 1: Figure S7). While 55 out of 1540 CRSs had scores $\geq$ 0.5, negative co-expression was observed rarely: Only two CRSs had a co-expression score below -0.5. All CRSs are available at https://rth.dk/resources/rnannotator/crs/insect/.

Most CRSs with relevant co-expression scores are expressed under few conditions. Of the 57 CRSs meeting our co-expression criteria, only 15 are expressed under more than five conditions. This is in agreement with the fact that most of the CRSs are expressed in specific contexts. One CRS (see "Examples of novel structures" section) was expressed in 29 experiments and in all of these together with its closest gene at a distance of only 12 bp of the annotated 5'-UTR (in case CRS and the closest gene element overlap 100% with tiling array transcript regions, instead of 50% as applied during the co-expression analysis). This strongly suggests that the current UTR annotation is incomplete and the CRS is in fact a structured UTR element.

As our assessment of co-expression is based on adjacency, it is conceivable that the physical order of co-expressed CRS and closest gene is also preserved in other species. We therefore compared the genes that

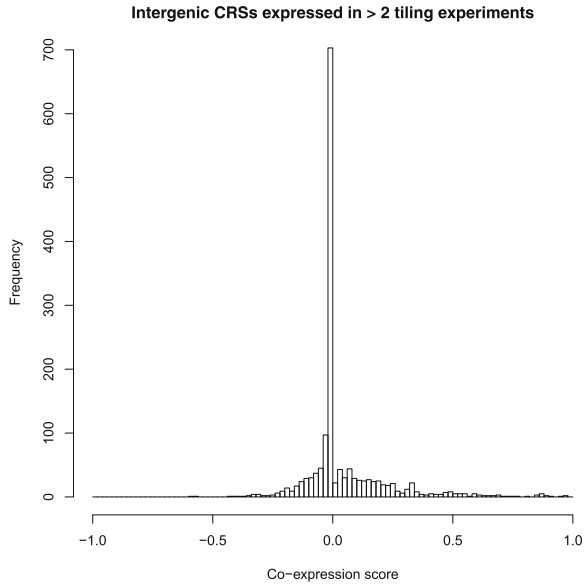


Fig. 4 Co-expression of intergenic CRSs with their closest annotated gene element according to modENCODE tiling array experiments [30]. A co-expression score of 1 indicates perfect positive co-expression, i.e. both CRS and closest gene are exclusively expressed together, −1 indicates perfect negative co-expression, i.e. both elements are only expressed individually, and a score of 0 indicates that the expression of CRS and closest gene were observed to be independent of each other

are adjacent to CRSs in *D. melanogaster* and in other species present in the CRS alignments. Of the 11 currently annotated non-*melanogaster* drosophilid species (FlyBase release FB2018_04), two thirds were required to fulfil the respective synteny criteria in the following analyses. For 13 of the 57 CRSs co-expressed in *D. melanogaster*, the closest genes in the other species were the orthologs of the closest *D. melanogaster* gene. However, this number is likely too conservative as we cannot expect all annotations to be complete and all orthologous relationships between genes of different species to be resolved entirely. More importantly, especially in more distant relatives of *D. melanogaster*, genes can be inserted between the CRS and the ortholog of the *D. melanogaster* gene. Hence, we also considered more relaxed criteria to define synteny: Looking for the ortholog of the *D. melanogaster* gene in each species, regardless of its distance from the CRS, we found 32 CRS-gene pairs to be in the same orientation in other species as in *D. melanogaster*, i.e. both the *D. melanogaster* gene and its ortholog were located either upstream or downstream of the CRS. When we applied an empirical maximal distance of CRS and closest *D. melanogaster* gene (or its ortholog) of 20,000 bp (see Additional file 1: Figure S8), still 18 of these 32 CRS-gene pairs passed the synteny filter. Finally, assuming that phylogenetic distance and quality of the annotation vary between species, we compared the closest genes in *D. melanogaster* and other species in a pairwise manner. In the species most closely related to *D. melanogaster*, *D. simulans* and *D. sechellia*, the genes neighboring 25 and 26 CRSs were orthologs of the gene closest to the CRS in *D. melanogaster*, respectively. The syntenic relationships of a subset of co-expressed CRS-gene pairs in *D. melanogaster* and their orthologous counterparts in other drosophilids provide another level of evidence for functionality of these CRSs.

Developmental stage and cell line specific expression of CRS-containing biotypes

D. melanogaster development is regulated by an orchestra of specific genes, see [57] and the references therein. Here, we connect the expression patterns of CRSs across developmental stages and cell types as a first step towards elucidating their potential roles in fruitfly development. For this as well as the following analysis, we associated genomic locations with a "biotype", i.e., a class of RNAs defined by similar functional and/or structural characteristics, such as miRNA, C/D box snoRNA, or 3'-UTR exon. We asked if expression of CRSs belonging to a particular biotype was statistically over- or underrepresented in a particular developmental stage or cell line (Fig. 5). In order to achieve a fair comparison we normalized the number of instances of a biotype expressed in a particular stage by the number of instances of the same biotype expressed in any of the other stages. For each biotype we then calculated the difference of these ratios for the subsets with CRSs (R_{CRS}) and without CRSs ($R_{\neg CRS}$), see Eq. (2) (see "Methods"). If this difference is positive, there are more instances with CRSs expressed in this stage compared to other stages than is the case for instances without CRSs. A significant difference between R_{CRS} and $R_{\neg CRS}$ may indicate a general role of the CRS-containing biotype instances in differentiating this stage. See Methods for more details of the analysis.

Not surprisingly, CRSs detected by our screen were particularly abundant for ncRNA classes, i.e., the biotypes H/ACA box snoRNAs, scaRNAs, snRNAs, and for cis-regulatory elements. CRSs were relatively rare in highly abundant biotypes such as introns, intergenic and exonic regions. Patterns of stage-specific over- and underrepresentation of CRS-containing biotypes were more similar to each other between pupae and adult stages, and more homogeneous than for the embryonic and larval stages (also see Additional file 1: Figure S9). Cell lines showed different patterns of enrichment and underrepresentation than developmental stages. The pupae stages formed the group with the statistically most significant enrichments of CRS-containing biotypes. Among these were mainly CRSs in 5'- and 3'-UTR exons, introns, promoters, enhancers, and lncRNAs. Adult stages

Fig. 5 Over- and underrepresentation of biotype instances with CRSs compared to instances without CRSs ('ratio difference', color coded) in *Drosophila melanogaster* developmental stages and cell lines. Only instances expressed in at least three tiling array experiments and contained in the `CMfinder` input alignments by at least 50% of the feature size were considered here. Tests of significance (indicated by opacity) assess whether biotype instances with CRSs are expressed more often in a particular stage compared to all other stages than expected by chance. *p*-values have been adjusted for multiple hypothesis testing (Bonferroni). The statistical test has been performed for all stages and cell lines, but in the interest of visibility, a representative subset of stages and cell lines of only total RNA samples has been chosen for this figure. For the full version of the plot, see Additional file 1: Figure S9

shared some of these enrichments, but also exhibited an underrepresentation of expressed CRSs in intergenic regions. Larval and embryonic stages differed from the other stages in that there were fewer stages enriched for CRS-containing UTR exons and lncRNAs and an even stronger underrepresentation of CRS-containing instances of several biotypes, e.g., introns and miRNAs. However, especially H/ACA box snoRNAs with CRSs were enriched in a number of embryonic and larval stages.

In cell lines we observed expression enrichment of CRS-containing UTRs less frequently than in any group of developmental stages. In contrast, ncRNA biotypes, especially snoRNAs, tRNAs, and intergenic regions were more often enriched with CRSs (also see Additional file 1: Figure S9). In summary, CRSs appear to be part of expression patterns that distinguish individual developmental stages from others.

Differential expression of CRSs

In order to elucidate the functional potential of CRSs in development in more detail, we aimed to identify pairs of developmental stages for which CRSs exhibit differential expression correlated with other biotype instances. We calculated a differential expression score $E_{\text{diff}}(i,j)$ as defined in Eq. (3) (see Methods) for each pairwise combination of modENCODE experiments i and j that compares the differential expression of CRS-containing instances and instances without CRSs (Fig. 6). The score can take on values from 0 to 1. The maximal score of 1 means that all structured instances are differentially expressed between experiments i and j whereas none of the unstructured

Fig. 6 Pairwise differential expression scores for introns contained in the `CMfinder` input alignments and expressed in at least three developmental stages or cell lines. The score compares the differential expression of introns containing CRSs and introns without CRSs. The higher the score on a scale of 0 to 1, the more structured introns and the less unstructured introns are differentially expressed

instances is. The product in the equation gives higher impact to situations with high differential expression of structured instances. See Methods for more details of the analysis.

For most biotypes, E_{diff} was small, i.e., the overall expression pattern did not differ much between individual stages. However, there were some structured intronic regions that were differentially expressed between white prepupae and most of the other stages and also cell lines. This could also be explained by the differential expression of the corresponding gene. Hence, in the next step we specifically considered differentially expressed introns in genes of which no exon is expressed in the same experiment. In Fig. 6, one of the most prominent red areas with high differential expression scores exists between intronic loci of white prepupae (two days) and the 12-14 h embryonic stage (both total RNA samples). Of these differentially expressed introns with CRSs, 29

were directly flanked by exons that were not expressed under the same conditions, and 3 were contained in genes of which not a single exon was expressed in the same experiment. One of the overlapping CRSs is referred to in more detail in the "Examples of novel structures" section. This observation suggests that the CRSs could be transcribed independently of the host genes. We note that the differential expression scores in this analysis did not rise above 0.15 for introns. Intronic loci with large differential expression thus are interesting candidates for novel functional transcripts.

Examples of novel structures

Based on the co-expression and differential expression analysis, a number of not previously annotated interesting structured RNA candidates were identified in intergenic or intronic regions. We present three examples representing different kinds of functional evidence: Positive or negative co-expression with the closest annotated gene, very small genomic distance to an annotated UTR, location in an intron showing differing expression from the adjacent exons, and a stable and complex secondary structure. Prediction DC0021109 (Fig. 7a) shows a perfect positive co-expression score with its closest gene *globin 1*, i.e., neither CRS nor *globin 1* are expressed alone in any of the 29 experiments in which expression was observed in this case (in case DC0021109 and the closest exon of *globin 1* overlap 100% with tiling array transcript regions, instead of 50% as applied during the co-expression analysis above). Since the genomic distance between them is only 12 bp, the annotated UTR of *globin 1* is most likely incomplete and DC0021109 is a structured UTR element. An example with a negative co-expression score of -0.57 is DC0018026 (Fig. 7b) with its closest gene *CG12581*, encoding a mostly unknown protein with a phosphotyrosine binding domain, which may be involved in a wide range of processes like neural development, tissue homeostasis or cell growth [58]. DC0018026 folds into a compact, stable consensus structure ($\Delta G = -15.99$ kcal/mol) comprising a multi-branch loop with two hairpins and an external stem. In contrast, DC0013572 (Fig. 7c) is located in an intron of the zinc finger transcription factor gene *CTCF*, which is involved in chromatin organization [59]. The secondary structure of the CRS features two hairpins with a longer conserved single-stranded stretch in between.

Discussion

We conducted a study of evolutionarily conserved RNA structure (CRS) elements in the *D. melanogaster* genome that was designed to assay genomic regions that are only loosely constrained at the sequence level. Therefore, we employed `CMfinder` to leverage structural alignments. Although `CMfinder` can detect CRSs also in highly sequence-conserved regions (unless the sequence identity reaches 100%) we observed that its sensitivity was limited when conservation at sequence level was high. As a consequence, the recall on well-studied RNA classes such as tRNAs, miRNAs and snoRNAs, all of which are very conserved at sequence level, was only moderate. These classes were readily detected in an earlier `RNAz` screen which operated on sequence-based alignments [31].

While `RNAz` performs best in the vicinity of 80% average pairwise sequence identity [31, 50], the majority of the `CMfinder` predictions were observed to lie between 60 and 70% in the screen on vertebrate genomes, and more than one third showed sequence identities below 60% [6]. However, a similar assessment for `CMfinder` on insect genomic alignments has been missing so far. As with `RNAz`, the number of `CMfinder` predictions decreased with sequence similarity. However, the predictions from the present screen have a much smaller FDR than previous `RNAz` results (less than 10% for `CMfinder` compared to up to 50% for `RNAz` [50]). The increased accuracy is ensured by using different cut-offs in different ranges of GC content and sequence conservation, thus controlling the FDR approximately independently of these parameters (Additional file 1: Figure S3). Nevertheless, we find a comparable number of CRSs. Although the overall sensitivity of the `CMfinder` screen was only moderate, it targets CRSs in a different range of conservation than other tools, emphasizing the usefulness of the `CMfinder` approach, in particular to screen in the low conservation range. At present, no single tool is capable of uncovering the entire wealth of RNA structure that is under selective constraints. This calls for research into improved methods for identifying selection pressures on RNA structure that can capitalize on the increasing amount of genome data that are becoming available for comparative genomics approaches.

Comparing the CRSs with genome-wide expression data from a broad range of developmental stages and cell lines, we found that in addition to a large number of nearly ubiquitously expressed loci, there were also sizable groups expressed in specific developmental stages or cell lines. The most statistically significant enrichment of expressed CRSs was observed in pupae, mainly located in UTR exons, promotors, enhancers, and lncRNAs. Interestingly, cell lines express different sets of CRSs than native developmental stages. This is in accordance with the respective modENCODE study of Drosophila cell lines [52]. Only a small fraction of intergenic CRSs was found to be co-expressed with the adjacent protein coding genes, indicating that most intergenic CRSs are independent genetic units.

An unexpected finding from our analysis was the differential association of *detected* CRSs with development type in biotypes such as snoRNAs and miRNAs, which

Fig. 7 Examples of `CMfinder` predictions that are part of putative novel transcripts or possibly incomplete annotations. Example **a** with a particularly high co-expression score and small distance to the closest annotated gene could be part of an incompletely annotated UTR. Example **b** is located much further from the closest annotated gene and hence could be part of a putative novel independent transcript. Example **c** is located in a differentially expressed intron of a gene of which no exons are expressed in the same developmental stage. For more details see description in the main text. Alignment and secondary structure visualization were performed using `RNAalifold` [79, 80]

are known to be conserved in both sequence and structure. This implies that there are differences in the patterns of sequence and/or structural evolution of these ncRNAs that are strong enough to affect how well they are detected by `CMfinder`. A characterization of these differences will require a detailed investigation into possible subclasses of miRNAs and snoRNAs as well as a comparative study of species outside the drosophilid

clade – and hence goes well beyond the scope of the present contribution.

One third of the predicted CRSs were not expressed, according to tiling array expression data. Some of this can be explained by the expected moderate fraction of false positive predictions. However, tiling arrays have several intrinsic biases that may prevent them from measuring CRSs: First, there is a bias against CRS sequences with low GC contents, since these interactions are less stable, and finding optimal stringency parameters to remove randomly bound RNAs but retain all true positives is challenging [18]. Furthermore, very stable secondary structures may not be captured since intramolecular folding will compete effectively with hybridization to the array [50]. Finally, also biological factors are likely to have an impact: Expression of transcripts that are expressed at very specific time points in development can easily be missed if there was no sample taken at exactly this time point. Also, differing external conditions might be necessary to induce expression of specific transcripts which cannot all be covered by any large-scale screen.

As it was observed in the mammalian CRS screen that CRSs hold the potential to bind RNA binding proteins (RBPs) [6], we assume that this is also the case for the fly genomes. Although some studies suggest that some RBPs appear to prefer single-stranded regions [60], other studies suggest that most RBPs prefer structured RNA, such as Staufen [61], Roquin [62] or MLE [63]. Computational surveys [64, 65] strongly suggest that structured binding sites are by no means rare. The interpretation of many CRSs as conserved RBP binding sites not only provides a biologically plausible explanation for the large number of detected loci in otherwise poorly conserved regions, but also suggests that it will be worth while (a) to engage in a large scale clustering of the predicted elements and (b) to compare the detected CRSs also across large phylogenetic distances, in particular with the elements reported in mammals [6, 32].

While the ultimate goal is to understand the role of conserved RNA structures in development, the computational survey reported here has to be content with providing starting points for following research. Our data show that there is a large set of CRSs with specific expression patterns that suggest their involvement in development and differentiation. Of course, such correlational data cannot distinguish between causal regulators and downstream consequences, but they narrow the list of candidates for further studies, both regarding cis-regulatory motifs and presumably independent ncRNA transcripts.

Although beyond the scope of this contribution, it will be relevant to characterize the stability of structures further within their respective biotypes. For example, 75% of the CRSs are located in introns or intergenic regions and can probably be further sub-categorized both by their stability and structural similarity using clustering techniques [66, 67]. It would be interesting to know whether CRSs found in other biotypes show patterns depending on the type of RNA they are part of.

Conclusion

Currently there are approximately 700 structured ncRNAs known in fruitflies [68], as well as thousands of unstructured ncRNAs (mostly lncRNAs [25–27]). While unstructured RNAs are generally easier to identify based on a certain level of sequence conservation, functional RNA structures are more hidden, and dealing with conservation on a structural level requires more elaborate and computationally expensive approaches. Hence, the true number of ncRNAs, especially of structured ones, is expected to be larger than the set we currently know. Accordingly, we found a large number of structurally conserved putative candidates in intergenic and intronic regions, many of which are likely to be functional according to evidence from expression analyses.

Due to the strong tendency of most RNAs, be they functional or not, to take on secondary structures, computational screens for CRSs need to deal with a certain trade-off between sensitivity and specificity as well as rather high false discovery rates, although we believe the latter to be lowered considerably in `CMfinder` screens. As a consequence, different tools for the prediction of conserved RNA structures yield only moderate overlaps when applied to the same genome. Screens conducted with alternative methods on previously investigated genomes therefore are a useful endeavor that contributes complementary data. In conclusion, our study has substantially expanded the repertoire of conserved RNA structures in fly genomes and in contrast to previous studies uncovered CRSs within the context of expression throughout all developmental stages and many cell lines.

Methods

Computational screen for CRSs

The 27-way `MULTIZ` [69] alignment consisting of 23 drosophilids and four additional insect species was downloaded from the UCSC Genome Browser (*Drosophila melanogaster* genome dm6, Aug. 2014, BDGP Release 6 [70]). The MAF (multiple alignment format) files contain sequences for chromosome arms 2L, 2R, 3L, 3R and chromosomes 4, X, Y, as well as mitochondrial sequences (M) and sequences in unassembled scaffolds. MAF blocks containing fewer than three sequences or that are shorter than 50 bp were removed. For all remaining MAF blocks, the reverse complement was generated in addition to be able to make predictions on both strands. Gaps were

removed from the alignments and the sequences were fed in their unaligned form into `CMfinder`.

We ran `CMfinder` (version 0.2.1) with default settings separately on the forward and the reverse strand of the native genome alignment. Default settings are as follows: The maximum number of candidates predicted in each sequence (i.e. MAF block) is 40. At most, 5 single stem-loop motifs with a base pair span between 30 and 100 bp and 5 double stem-loop motifs with a base pair span between 40 and 100 bp are returned. Motifs on the same strand are merged by `CMfinder` if the predicted structure is consistent in both overlapping motifs. The prior for the expected fraction of sequences containing the motif is 0.8. The `CMfinder`-specific *pscore* [35], (`CMfinder` version 0.2.2) was computed for all predicted motifs. It is fundamentally similar to a general time reversible model of sequence evolution extended to include both single-stranded and base-paired regions. Some model parameters were trained using vertebrate Rfam alignments, but we scored our candidate motifs with respect to a phylogenetic tree having topology and branch lengths as estimated for drosophilids (dm6.27way.nh from ref. [70]; `CMfinder`'s `-t` option). As shown in Additional file 1: Figure S2, a good *pscore* is well-correlated with lower estimated FDR across the spectrum of sequence identity and GC content. Sequence identity and GC content were calculated for all `CMfinder` output alignments. As a reference sequence for all predictions we used our species of interest, *Drosophila melanogaster*, and therefore only considered predictions containing this species. All genome coordinates used in the following were derived from the reference genome.

Background model

We estimated the false-discovery rate among the `CMfinder` predictions by synthesizing "background" alignments using `SISSIz` (version 2.0 [16]). Specifically, for each input MAF block, one companion randomized alignment was produced using `SISSIz` with the following options: `--simulate --tstv --maf -n 1`. This simulates sequence evolution from an ancestral sequence derived from the given MAF block using an evolutionary model that preserves mono- and di-nucleotide frequencies in expectation, while exactly preserving the input's gap- and local conservation patterns. Transition and transversion rates are estimated from the input data (`--tstv`), and one random alignment in MAF format is generated per input (`--maf -n 1`). `CMfinder` was run on both strands of the shuffled genome alignment in the same way as on the native alignment.

False discovery rate

In order to find a threshold to filter out the most unreliable predictions, *pscore* lower boundaries from 50 to 150 were applied and the distributions for *pscore*, minimum free energy, GC content, sequence identity, length and number of species of the motif alignments as well as the number of predictions remaining were visually inspected. Based on this, we applied a *pscore* cutoff of $p > 50$ to reduce the number of predictions to a manageable amount.

All predicted motifs with a *pscore* > 50 were filtered for overlap with annotated repeats (as provided through the UCSC genome browser [71]) using `bedtools intersect` [72], removing all predictions that overlap a repeat by at least 50%.

To estimate the false discovery rate (FDR), GC content and sequence identity were categorized so that each bin comprises comparable numbers of predicted motifs with these features. `CMfinder` input MAF blocks were categorized into the same bins. Since their number is sufficiently high in each bin (i.e. more than 100 MAF blocks), all bins were considered, even in case the numbers of predictions in a bin were low. The FDR was estimated for all motifs in each bin with a particular *pscore* cutoff as

$$\begin{aligned}\text{FDR estimate} &= \frac{\text{False Positives}}{\text{False Positives} + \text{True Positives}} \\ &= \frac{\#\text{ predictions on the shuffled alignment}}{\#\text{ predictions on the native alignment}}\end{aligned}$$

Based on the number of predictions left, the individual FDR heatmaps and relationship between mean FDR and sequence identity depending on pscore cutoff, cutoffs of both *pscore* > 80 and FDR ≤ 0.1 were taken to filter out the most unreliable predictions. FDR estimates were assigned to each motif according to the FDR of the respective bin and pscore cutoff.

Annotation

For the overlap with the existing annotation, the most recent *Drosophila melanogaster* annotation data for the dm6 genome release were obtained from FlyBase (dmel_r6.15, FlyBase release FB2017_02) for the genomic annotation (exon type, intron or intergenic region) [29] and from Rfam 12.2 for the non-coding RNA annotation [68].

When converting the FlyBase annotation fasta files into bed files, split entries were converted to a single bed entry without considering splicing. In the rare cases of genes derived from both strands such as trans-spliced mod(mdg4) [73], separate entries were created for both strands.

In order to annotate each prediction unambiguously, the FlyBase genomic annotation tracks were unified such that each nucleotide has only one annotation category assigned. For this purpose, all annotated coding sequences as well as genes were merged using `bedtools merge`. Each annotated exonic region was categorized either as coding exon if located within coding sequence boundaries

or otherwise as non-coding exonic region. Drosophilid Rfam annotations were added to the non-coding exonic regions. To determine if a non-coding region belongs to a ncRNA or to a UTR, each exon's gene parent was checked for the presence of a coding sequence. In case ncRNA and UTR exons overlap (this was observed for approximately 5% of all UTR exons), the ncRNA-exonic character was prioritized and the region of overlap annotated as ncRNA-exonic region. The regions in which 5'- and 3'-UTR exons overlap are categorized as exons of both UTRs because in this case no meaningful prioritization of one UTR type over the other can be made. Then, each so generated annotation bed file was merged using `bedtools merge` and all resulting exons were subtracted from the list of all merged genes to obtain all introns. All exons and introns were subtracted from the complete genomic sequence to obtain all intergenic regions.

For the annotation and all subsequent analyses (unless explicitely mentioned otherwise), individual predictions were merged strand-independently up to a distance of 30 nt using `bedtools merge`.

For the genomic annotation, we counted how many annotation elements (individual exonic, intronic or intergenic regions) overlap a given prediction ($\geq$ 1 bp) without considering strands and then assigned the respective fraction, i.e. 0.5 in case a prediction overlaps two genomic classes. This approach was chosen because the unified exons and introns can be very short. Therefore, a CRS might overlap a number of different categories, and an overlap of at least 50% of the CRS size is less meaningful in these cases.

For the overlap with the non-coding annotation, only ncRNAs lying by at least 50% of their size within individual `CMfinder` input MAF blocks were considered since annotated structures that are not covered by the alignment cannot be predicted. The minimum overlap of 50% takes into account that many ncRNAs consist of several shorter structured motifs, which still can be predicted by `CMfinder` even if only a part of the complete sequence is contained within a MAF block. We only included Rfam annotations with a minimum base pair content of 30%, which means at least 30% of the positions of a sequence must be involved in base pairing. In order to identify known ncRNA elements covered by the predictions, `CMfinder` predictions and Rfam annotation were intersected using `bedtools intersect` with a minimal overlap size of at least 50% of the prediction or the annotation element size.

For each intersection of `CMfinder` predictions and genomic or ncRNA annotation, the fold enrichment FE was calculated as

$$\mathrm{FE} = \frac{\frac{\#\text{ merged overlapping queries}}{\#\text{ queries}}}{\frac{\text{target size (nt)}}{\text{background size (nt)}}}$$

The background size is computed as the total number of columns in the `CMfinder` input MAF blocks. The target size is defined as the total size of all annotation elements under consideration that overlap a MAF block by at least 1 nt. The significance of each enrichment was calculated using the `pnorm` function in `R` as previously described [18]. Specifically, the number of observations was the number of overlaps and the mean was calculated as the product of the total number of CRS candidates and the fraction of the input covered by the annotation.

The FDRs for the recovered and not recovered fractions of the Rfam annotation were estimated in a similar manner as the genome-wide FDR, but only including individual predictions (*pscore* $>$ 80, repeat-filtered) from native and shuffled alignments that overlap (recovered or not recovered) Rfam annotations, without considering strand, GC content, or sequence identity.

Comparison with other ncRNA screens

We compare our predictions to four other genome-wide screens for ncRNAs in drosophilids: 42,482 predictions from an `RNAz` screen [31], 2469 predictions from a more restrictive `RNAz` screen aimed at finding miRNAs [49], 30,478 predictions from a `REAPR` screen [21], and 22,682 predictions from an `EvoFold` screen [48]. Site-specific `phastCons` scores based on the `MULTIZ` 27-way insect alignment were averaged for each predicted motif or CRS, respectively. GC contents were calculated for each *D. melanogaster* sequence. For the prediction overlaps, the coordinates of all three screens were transformed from dm2 to dm6 genome release using the UCSC `LiftOver` utility (https://genome.ucsc.edu/cgi-bin/hgLiftOver). Predictions were intersected using `bedtools intersect` with a minimal overlap of 1 bp.

Expression data set

Tiling array data were obtained from the modENCODE database (v32 [74]), comprising 3,665,935 transcript regions. Each of the 80 experiments corresponds to expression in one cell line or in one developmental stage of one of two fly strains, either total, polyA, or nuclear RNA-sequenced. In a minority of the experiments expression was evaluated specifically in virgin heads, mated ovaries, or the larval gut. Throughout this study, if not stated otherwise, a CRS or any annotation instance is defined as expressed if it overlaps a merged transcript region by at least 50% of its size. For the expression heatmaps, only CRSs showing expression in at least one experiment were considered. The non-coding RNA annotation was obtained from Rfam (v.12.2).

Expression enrichment

To obtain a random background for Fisher's exact test, the *D. melanogaster* genome was divided into 100-bp windows (approximately the average size of a CRS), and 20,184 of the windows were sampled randomly. Samples expressed in at least one to four modENCODE experiments were intersected with the CRSs (overlap at least 50% of the CRS or genomic sample size). The resulting contingency table for each minimum number of experiments consists of the numbers of genomic samples expressed and not expressed and these overlapping CRSs or not. Sampling and Fisher's exact test were carried out 10 times for each minimum number of experiments. In a second test, windows and CRSs overlapping coding or UTR exons were removed in order to avoid a potential mRNA exon bias. Sampling and Fisher's exact test were carried out as previously.

Co-expression with adjacent genes

To evaluate co-expression of a prediction with its closest gene element we define the co-expression score E_{co} as

$$E_{co} = \frac{E_{cg}}{E_c} - \frac{E_{g\neg c}}{E_{\neg c}}, \tag{1}$$

where E_{cg} is the number of experiments in which both the CRS and its closest gene element are expressed, E_c is the number of experiments in which the CRS is expressed, $E_{g\neg c}$ is the number of experiments in which the closest gene element is expressed but not the respective CRS, and $E_{\neg c}$ is the number of experiments in which the CRS is not expressed. $\frac{E_{cg}}{E_c}$ is also referred to as Ratio 1 and $\frac{E_{g\neg c}}{E_{\neg c}}$ as Ratio 2.

For the analysis of synteny of co-expressed CRS-gene pairs, orthologs of *D. melanogaster* genes in all 11 annotated non-*melanogaster* species [75] were obtained from FlyBase (FlyBase release FB2014_06, the most recent release with all genome releases corresponding to the genome releases used in the `MULTIZ` alignment), as well as the corresponding gene annotations for each species. Where necessary, FlyBase chromosome/scaffold identifiers were transformed into UCSC identifiers with the help of the respective assembly reports and GenBank accession numbers [76]. In case of ties when determining neighboring genes of CRSs, i.e., multiple genes with the same distance to the CRS in *D. melanogaster* or any other species, at least one ortholog had to fulfil the respective synteny criterion (being the ortholog to a *D. melanogaster* closest gene, being in the correct orientation with respect to the CRS, or being within the maximum distance, 20,000 bp, of the CRS).

Experiment-specific expression

We tested for expression enrichment of CRS-containing biotypes in specific experiments, e.g., developmental stages and cell lines. For the kth biotype B_k and the lth modENCODE experiment E_l, we define the CRS-ratio R_{CRS}, a non-CRS ratio $R_{\neg CRS}$, and the ratio-difference R_d as

$$R_{CRS}(B_k, E_l) = \frac{N(B_k, E_l, CRS)}{N(B_k, E, CRS)},$$

$$R_{\neg CRS}(B_k, E_l) = \frac{N(B_k, E_l, \neg CRS)}{N(B_k, E, \neg CRS)},$$

$$R_d(B_k, E_l) = R_{CRS}(B_k, E_l) - R_{\neg CRS}(B_k, E_l), \tag{2}$$

where $N(B_k, E_l, CRS)$ is the number of biotype B_k instances overlapping at least one CRS (minimum overlap of 50% of instance or CRS size) expressed in the currently considered experiment E_l, and $N(B_k, E, CRS)$ is the respective number expressed in any other experiment E. Only instances expressed in at least three experiments and contained in the `CMfinder` input alignments by at least 50% of their size were considered. For each biotype and each modENCODE experiment, a one-sided Student's t-test (coding exons, 5'-UTR exons; normally distributed non-CRS ratios) or Wilcoxon-Mann-Whitney test (all other biotypes; non-CRS ratios not normally distributed) was performed to test the significance of deviations of the CRS-ratio from the mean of all non-CRS ratios for that biotype. Depending on the ratio difference being larger or smaller than 0, the alternative hypothesis for the `R` functions t.test() and wilcox.test() was set to 'less' or 'greater', respectively. All p-values were adjusted for multiple hypothesis testing (Bonferroni). Exon and intron biotypes in this analysis are from the FlyBase annotation (dmel_r6.15, FlyBase release FB2017_02). Promoter annotations were obtained from the EPDnew database [77], enhancer annotations were obtained from the Fly Enhancers database [78], and all non-coding annotations were obtained from FlyBase and, where available, combined with the Rfam annotation (Rfam 12.2). For calculating non-CRS ratios for intergenic regions (FlyBase) we split them into 100 bp long windows and categorized them into bins according to their GC content and sequence identity (in the same way as the `CMfinder` predictions for the FDR calculation). From each of these bins as many intergenic windows were sampled as there are predictions in that bin.

Differential expression

To analyze the correlation of RNA structures and differential expression between developmental stages, for each biotype and each modENCODE experiment two expression vectors (with elements of 1 for expression in this experiment, 0 for not being expressed) were generated: one for all biotype instances containing CRSs, and one for instances without CRS. Only instances that are expressed in at least three experiments were considered. Then, for

Bioinformatics, 35(9), 2019, 1494–1502
doi: 10.1093/bioinformatics/bty859
Advance Access Publication Date: 5 October 2018
Original Paper

Sequence analysis

Inferring disease-associated long non-coding RNAs using genome-wide tissue expression profiles

Xiaoyong Pan[1,2], Lars Juhl Jensen[2,*] and Jan Gorodkin[1,*]

[1]Department of Veterinary and Animal Sciences, Center for Non-coding RNA in Technology and Health, University of Copenhagen, 1870 Frederiksberg C, Denmark and [2]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen N, Denmark

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on April 30, 2018; revised on August 28, 2018; editorial decision on October 3, 2018; accepted on October 4, 2018

Abstract

Motivation: Long non-coding RNAs (lncRNAs) are important regulators in wide variety of biological processes, which are linked to many diseases. Compared to protein-coding genes (PCGs), the association between diseases and lncRNAs is still not well studied. Thus, inferring disease-associated lncRNAs on a genome-wide scale has become imperative.

Results: In this study, we propose a machine learning-based method, DislncRF, which infers disease-associated lncRNAs on a genome-wide scale based on tissue expression profiles. DislncRF uses random forest models trained on expression profiles of known disease-associated PCGs across human tissues to extract general patterns between expression profiles and diseases. These models are then applied to score associations between lncRNAs and diseases. DislncRF was benchmarked against a gold standard dataset and compared to other methods. The results show that DislncRF yields promising performance and outperforms the existing methods. The utility of DislncRF is further substantiated on two diseases in which we find that top scoring candidates are supported by literature or independent datasets.

Availability and implementation: https://github.com/xypan1232/DisIncRF

Contact: gorodkin@rth.dk or lars.juhl.jensen@cpr.ku.dk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Long non-coding RNAs (lncRNAs) play crucial roles in many biological processes and are involved in a variety of diseases (Chen *et al.*, 2013; Mirza *et al.*, 2015). Currently, the role of protein-coding genes in diseases is well investigated and collected in public databases, e.g. DISEASES (Pletscher-Frankild, 2015), DisGeNET (Pinero *et al.*, 2015) and OUGene (Pan and Shen, 2016). However, the vast majority of genetic susceptibility loci related to diseases is located in non-coding DNA regions, intergenic and intronic for PCGs (Hindorff *et al.*, 2009), a substantial fraction of which contain lncRNA genes (Esteller, 2011). Still, most of the associations between diseases and lncRNAs are unknown. A computational approach is thus needed to help identify disease-associated lncRNAs and to provide a comprehensive overview thereof.

Recent studies have shown that genetic disorders usually manifest themselves only in a few tissues (Kitsak *et al.*, 2016; Lage *et al.*, 2008; Magger *et al.*, 2012; Winter *et al.*, 2004). In addition, disease-causing mutations in human PCGs often lead to tissue-specific phenotypes, indicating associations between diseases and tissues (Blokzijl *et al.*, 2016; Bornigen *et al.*, 2013; Magger *et al.*, 2012). Combining text-mined disease–tissue associations from biomedical literature with a tissue expression atlas revealed that the average expression of disease-associated genes is higher in the disease-related tissues than in other tissues (Lage *et al.*, 2008).

Several studies have used tissue expression data to infer lncRNA–disease associations. Guilt-by-association is a widely used strategy that scores candidate genes based on their similarity in expression to known disease-associated genes (Chen *et al.*, 2016a, 2017). For example, the LRLSLDA method infers disease-associated lncRNAs by assuming that lncRNAs involved in the same disease have similar expression patterns (Chen and Yan, 2013). The KATZLDA and FMLNCSIM methods extend this approach by also taking into account functional similarity metrics between the lncRNAs (Chen, 2015a; Chen *et al.*, 2016b). A common limitation of all three methods is that they use only disease associations for lncRNAs thus ignore the well studies PCGs, which are often co-expressed with lncRNAs in diseases (Tsoi *et al.*, 2015). In contrast, LnCaNet integrates experimentally verified associations between cancer types and PCGs with co-expression associations between PCGs and lncRNAs (Liu and Zhao, 2016). However, lncRNA–disease associations are scored based on correlated expression with individual disease-associated PCGs, and the method does not cover other diseases than cancers.

With the objective to reduce the number of false positives and improve the prediction accuracy, some approaches integrate more information. GeneTIER (Antanaviciute *et al.*, 2015) prioritizes candidate disease genes using disease–tissues associations. Similarly, Tissue Specific Expression Analysis (TSEA) identifies genes enriched in disease-associated tissues. The method first defines sets of tissue-enriched genes, and then identifies significant consistency between tissue-enriched genes and disease-associated genes (Wells *et al.*, 2015). Both methods require extra information in advance, such as disease-tissue or gene-tissue associations. Other studies prioritize disease-associated genes using functional interaction networks, where it is assumed that similar diseases may be caused by functionally associated genes. Guan *et al.*, used tissue expression data to construct tissue-specific functional network to identify disease-associated genes instead of using global functional network (Guan *et al.*, 2012). Similarly, NetWAS combines tissue-specific interaction networks and genome-wide association study (GWAS) to infer disease-gene associations (Greene *et al.*, 2015). Tissue-specific network can improve the quality of predictions and reduce the noise compared with only using GWAS. In addition to the methods listed above, several studies have used machine learning to infer genes associated with a specific disease of interest. Recently Cogill and Wang (2016) applied support vector machines (SVM) (Vapnik, 1998) to infer genes associated to autism spectrum disorders (ASD) using expression profiles in healthy brain as features.

In this study, we propose a computational method DislncRF, which takes advantage of available RNA-seq expression profiles across multiple healthy tissues and well-studied disease-associated PCGs to infer disease-associated lncRNAs. With DislncRF, we train multiple balanced Random Forest (RF) models (Breiman, 2001) to learn expression patterns from PCGs involved and not involved in a disease, and apply the learned models to infer disease-associated lncRNAs. In addition, DislncRF is also able to automatically identify disease-tissue associations.

2 Materials and methods

We created a training set consisting of tissue expression profiles from RNA-seq data for PCGs involved and not involved in a specific disease. For each RNA-seq dataset and each disease, we trained RF models to predict PCGs associated to that specific disease from those not associated to that specific disease. Subsequently, these RF models are applied to predict if lncRNAs are associated to that specific disease or not, based on the given RNA-seq dataset.

2.1 Data source

We collect tissue RNA-seq expression data, disease-PCG associations and disease–lncRNA associations as follows:

2.1.1 Tissue RNA-seq data

We use four sources of RNA-seq expression datasets with variable number of tissues to train and evaluate DislncRF:

1. The Genotype-Tissue Expression project, GTEx: we directly use the processed expression profiles (GTEx_Analysis_v6p_RNA-seq_RNA-SeQCv1.1.8) from GTEx portal (GTEx Consortium 2013; GTEx Consortium 2015), which contains expression data in 53 human tissues.
2. The human body map 2.0 project, E-MTAB-513: It contains the raw RNA-seq data for 19 samples across 16 human tissues (Derrien *et al.*, 2012).
3. Gene evolution in tetrapods, GSE43520: This dataset consists of raw RNA-seq data for 11 human samples across four human tissues (Brawand *et al.*, 2011)
4. LncRNAs evolution in mammalians, GSE30352: It consists of raw RNA-seq data from 21 samples across six human tissues (Necsulea *et al.*, 2014)

2.1.2 Disease–PCG associations

For disease-PCG associations, we download the integrated associations from DISEASES database (Pletscher-Frankild, 2015), including knowledge, experiments and text mining channels. This database gives confidence scores to the associations according to the supporting evidence. We only use associations with confidence scores greater or equal to 2. Likewise, we require diseases to have at least 50 associated PCGs with expression profiles to avoid the problem with too few data for machine learning model training. In total, we obtained sufficiently many disease-associated genes for 237 diseases. Supplementary Table S1 shows the number of diseases with different cutoffs for confidence and number of associated PCGs.

For validation of predicting disease-PCG associations, we split the dataset constructed from disease–PCG pairs into two parts: the data for 20% of the PCGs is randomly kept as the independent test set consisting of disease-PCG pairs, and the remaining 80% is used as the training set. We evaluate the classifier performance on PCG expressions for each disease of the 237 diseases.

2.1.3 Disease-lncRNA associations

We collected human experimentally verified lncRNA-disease associations from LncRNADisease (Chen *et al.*, 2013) and Lnc2Cancer (Ning *et al.*, 2016). After mapping the associated diseases into the disease ontology terms and gene names into Ensembl gene identifiers, we build a gold standard set with 735 unique disease-lncRNA associations.

Here, we use the experimentally verified lncRNA–disease associations only to evaluate the performance of DislncRF. As negative examples, we randomly selected the same number of lncRNAs not associated with the disease as we have lncRNAs associated with it, thus resulting in a dataset that is balanced not only overall but also for each disease. Since no disease–lncRNA associations were used for training the models, neither as positive nor negative examples, this dataset is completely independent of the training set.

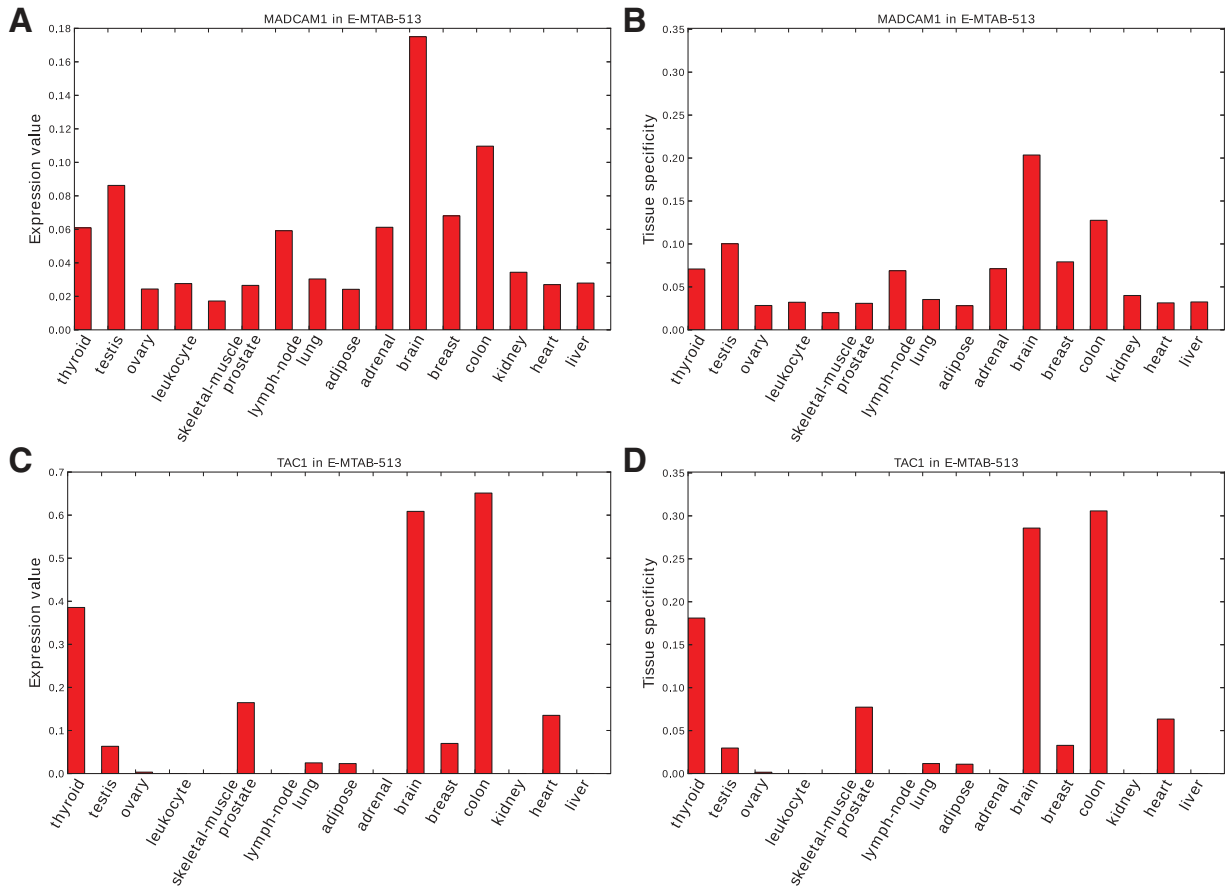


Fig. 1. In E-MTAB-513 (human body map 2.0 dataset) (Derrien *et al.*, 2012), the expression level and tissue specificity of two inflammatory bowel disease associated genes MADCAM1 and TAC1. (**A**) The expression values of gene MADCAM1; (**B**) The tissue specificity value of gene MADCAM1; (**C**) The expression values of gene TAC1; (**D**) The tissue specificity value of gene TAC1

2.2 Data preprocessing

We downloaded the RNA-seq reads for the three datasets E-MTAB-513, GSE43520 and GSE30352 and processed them using the latest version of the TISSUES database (Palasca *et al.*, 2018). The pipeline uses STAR version 2.5.0b (Dobin *et al.*, 2013) to map the raw reads from all datasets to the reference genome with 19 732 PCGs and 13 336 lncRNAs used by GTEx, and quantifies the expression levels of genes using Cufflinks (Trapnell *et al.*, 2010). After obtaining the quantified expression levels, we calculate the median expression value x across the samples for each tissue, which is the same strategy used in GTEx. Finally, we log-transform the values, using $\log_2(1+x)$ to avoid problems with genes for which zero expression was observed in a given tissue.

To remove the impact of different scales of expression levels, we normalized them to tissue expression specificity, which is the fraction of (log-transformed) expression of one gene in one tissue relative to the sum of its expression in all tissues (Tsoi *et al.*, 2015):

$$T_{is} = \frac{e_{is}}{\sum_j e_{ij}}, \quad (1)$$

where s is the tissue index and i is the gene index. Figure 1 illustrates the expression level and tissue specificity of two genes associated to inflammatory bowel disease. The two genes have different scales of expression levels but have similar scale of tissue specificity level.

2.3 Random forest classification

A random forest is an ensemble of multiple decision trees (Breiman, 2001), in which each tree is built from bootstrapping samples and randomly selected feature subset of original features. RFs can be used for classification, as well as for feature importance analysis. During the training process, out-of-bag (OOB) error is calculated for individual feature before and after randomly permuting the values of that feature, the importance score is averaging the above two OOB error difference over all trees.

For selecting the parameters of each RF model, we applied GridSearchCV (Pedregosa *et al.*, 2011) to optimize the parameters by 3-fold cross-validation scheme over a parameter grid (min_samples_leaf: [1, 2, 3], max_features: ['auto', 'sqrt', 'log2'], n_estimators: [5, 10, 20, 50]). The variables are defined as follows: min_samples_leaf is the minimum number of samples in leaf node, max_features is the number of features for splitting a node, and n_estimators is the number of trees.

2.4 DislncRF pipeline

The DislncRF pipeline, while intended to predict disease-associated lncRNAs, consists of RF models trained on PCGs involved/not involved in a disease. This leads to a class imbalance problem, since the number of genes implicated in a disease is much fewer than the number of genes not implicated in this disease (e.g. 50 vs. 19 732 for

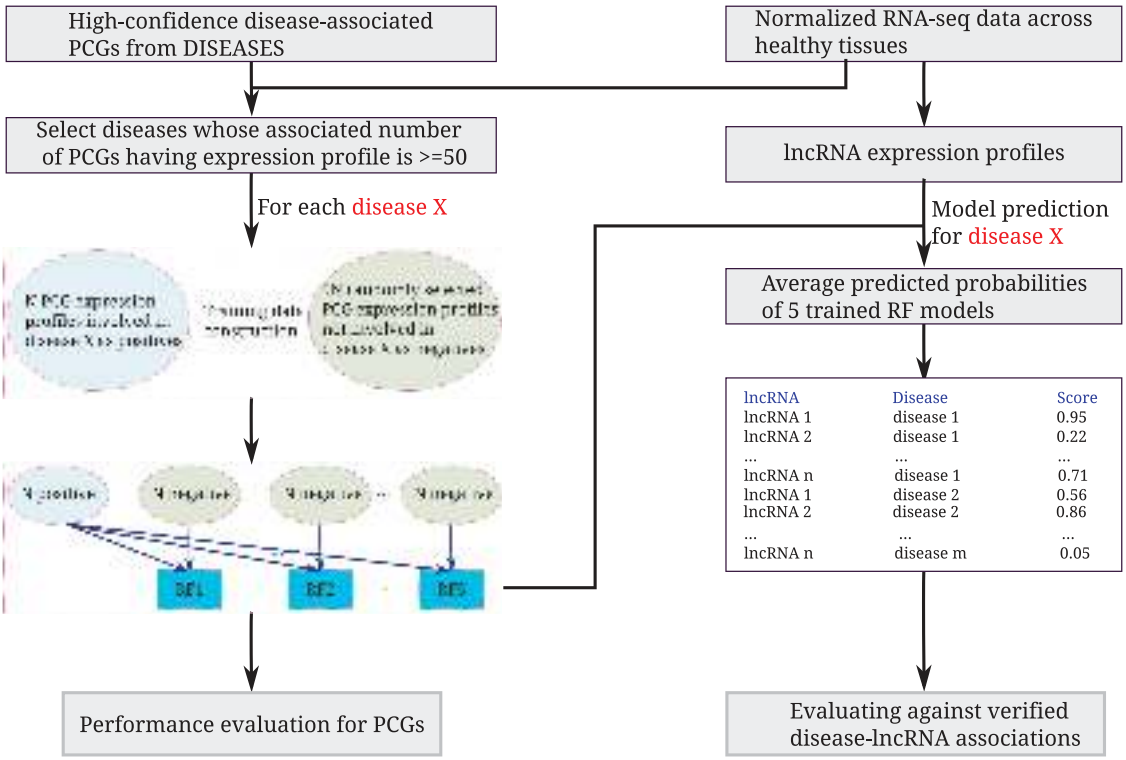


Fig. 2. The flowchart of DisIncRF. For each disease *X*, we firstly extract *N* disease *X* associated protein coding genes (PCGs) from DISEASES database, and randomly sampling 5*N* negative PCGs not involved in disease *X* under three criterion described in text. Then we train five random forest models on the balanced dataset with the same positive set but different negative set. Finally, the five trained RFs are used for inferring disease *X* associated lncRNAs

Autistic disorder). To avoid classifiers becoming biased to the negative class while making use of there being more such training examples, we constructed five balanced RFs, each trained on the same disease-associated PCGs but different subsets of PCGs not involved in disease. The workflow of DisIncRF is shown in Figure 2. For each disease *X* among the total 237 diseases:

1. Construct the positive training set with *N* positive examples of PCGs involved in disease *X*. Randomly select five sets of each *N* negative examples, resulting in five datasets each with the same positive but different negative examples, using the following criteria: (i) The PCGs should be associated with other diseases than *X*, to prevent that the positive examples are biased towards well studied PCGs compared to the negative examples. (ii) The PCGs must have no evidence of association with disease *X*, also not with a DISEASES confidence score below the threshold used to define positive examples. PCGs with a confidence score below 2 for disease *X* are thus neither used as positive nor as negative examples.
2. Train five RF classifiers, one for each dataset, to discriminate positive example PCGs from negative examples from their tissue specificity profile and sum of log-transformed expression values across tissues.
3. Apply the trained RF models for lncRNAs to obtain probability scores of an lncRNA to be associated to disease *X*. The final score is the average of the probabilities from the five trained RF models.

In the end, after obtaining the association scores between the 237 diseases and 13 336 lncRNAs, we evaluate the predicted disease-associated lncRNAs using the gold standard set.

2.5 Baseline methods

To verify the advantage of DisIncRF over existing approaches, we compared its prediction performance with that of the coding-non-coding co-expression (CNC) method (Liao *et al.*, 2011; Liu and Zhao, 2016) under guilt-by-association framework. For each disease *X* and lncRNA, CNC proceeds as follows: (i) Extract PCGs associated with the disease. (ii) Calculate Pearson's correlation coefficients (PCC) between PCGs associated with *X* and the lncRNA. (iii) Keep only those co-expression pairs, which have absolute PCC > 0.3 and *P*-value < 0.01 (similar to Liu and Zhao, 2016). (iv) Assign the mean value of up to *K* largest PCC value, where *K* is the predefined number of largest PCCs with this lncRNA, as the predicted score for the lncRNA with disease *X*.

As a second baseline method, we used the following simple *K* nearest neighbor (KNN) approach, where *K* is the predefined number of neighbors: (i) Calculate the pairwise PCCs between all lncRNAs and all PCGs. (ii) For each lncRNA, obtain the *K* nearest PCGs according to PCC and count how many of them are associated with each disease *X*. This count is used as the score between the lncRNA and disease *X*. We evaluate KNN and calculate the fraction of correct predicted associations for each raw score (Junge *et al.*, 2017).

2.6 Validation of tissue importance

A merit of the RF algorithm is that it can also learn the importance score of input features. We made use of this to analyze the importance of each tissue for prediction of each disease. To validate that the features extracted by the RF models are consistent with biological knowledge of the disease, we compared them to manually curated

and text-mined disease–tissue associations (Binder *et al.*, 2014; Pafilis *et al.*, 2013). To quantify the agreement between the RFs and the two sets of disease–tissue associations, we calculated Pearson correlation coefficients between the tissue importance scores and the confidence scores for the disease–tissue associations.

3 Results

In this study, we report the performance of DislncRF on PCGs and lncRNAs, and compare it with two baseline methods to demonstrate the advantages of DislncRF. Lastly, two case studies are investigated for predicted disease–lncRNA associations from DislncRF.

3.1 Validation of DislncRF for PCGs

For each disease, we optimized parameters for the five RFs using GridSearchCV (see the Supplementary Material). As shown in Table 1, DislncRF achieves high performance on the independent test set for prediction of the disease-associated PCGs on the four RNA-seq datasets (E-MTAB-513, GSE43520, GSE30352 and GTEx). Supplementary Figure S1 shows the box plots of individual measurements of the 237 diseases. For GTEx, the average performance overall diseases is a Matthews correlation coefficient (MCC) of 0.676, which is slightly better than what is obtained for the three other datasets. This is presumably because the GTEx dataset is more complete, covering 53 tissues, as opposed to at most 16 tissues in the other datasets. It is thus likely include additional tissues of relevance for the studied diseases. Conversely, we get the lowest MCC performance on GSE43520, most likely because this dataset only covers four tissues, and many disease-associated tissues are not included. Hence, many informative signals from the associated tissues cannot be captured for the disease-associated genes. We see that DislncRF yield high sensitivity and low precision on these four datasets due to that our test data has five times as many negative examples than positive ones, even though high accuracy and area under the ROC curve (AUC) can be obtained on the balanced test set.

Table 1. The average performance of the disease-associated gene classification using PCG expression profiles on the test set consisting of disease-PCG pairs

Dataset	Sensitivity	Specificity	AUC	MCC	Precision	AUPRC
GTEx	0.902	0.875	0.953	0.676	0.612	0.905
E-MTAB-513	0.899	0.877	0.955	0.675	0.612	0.905
GSE43520	0.887	0.816	0.931	0.577	0.501	0.862
GSE30352	0.891	0.832	0.939	0.604	0.530	0.878

Note: In this study, we trained one model for each of 237 diseases. AUC is the area under ROC curve and AUPRC is the area under Precision-Recall Curve.

3.2 Predicting disease-associated lncRNAs

After training models on PCG profiles, we used the trained models to predict disease-associated lncRNAs. As shown in Figure 3A, DislncRF yields the AUC of 0.687, 0.649, 0.592 and 0.718 on E-MTAB-513, GSE43520, GSE30352 and GTEx, respectively. In addition, DislncRF yields the Area under Precision-Recall Curve (AUPRC) of 0.669, 0.642, 0.575 and 0.687, respectively. We also report other performance measurements in Table 2.

Considering that there are more verified lncRNAs for cancer diseases than non-cancer diseases in our gold standard set, we further separated the in total 237 diseases into two subgroups, namely cancer (44) and non-cancer diseases (193). We performed the same analyses for both subgroups as was done for all diseases. We also constructed the Receiver Operating Characteristic (ROC) curve (Supplementary Fig. S2). The results indicates that DislncRF obtain slightly higher performance for cancer type than non-cancer diseases. The reason might be that the PCGs and lncRNAs are more studied in cancer and the cancer is a more homogeneous collection of tissues than the rest, which provides better training (PCGs) and test (lncRNAs) set.

DislncRF performs better on PCGs than on lncRNAs, which is not surprising as the RF models are trained on PCGs. This is not surprising as the models are trained on PCGs, which in many ways have differences to lncRNAs. In particular, PCGs are typically

Table 2. The performance of the DislncRF for predicting disease-lncRNA associations, which is benchmarked against the independent disease-lncRNA test set (Section 2.1.3)

Dataset	Sensitivity	Specificity	AUC	MCC	Precision	AUPRC
GTEx	0.527	0.765	0.718	0.301	0.692	0.687
E-MTAB-513	0.476	0.759	0.687	0.245	0.664	0.669
GSE43520	0.515	0.681	0.649	0.199	0.617	0.642
GSE30352	0.503	0.611	0.592	0.115	0.564	0.575

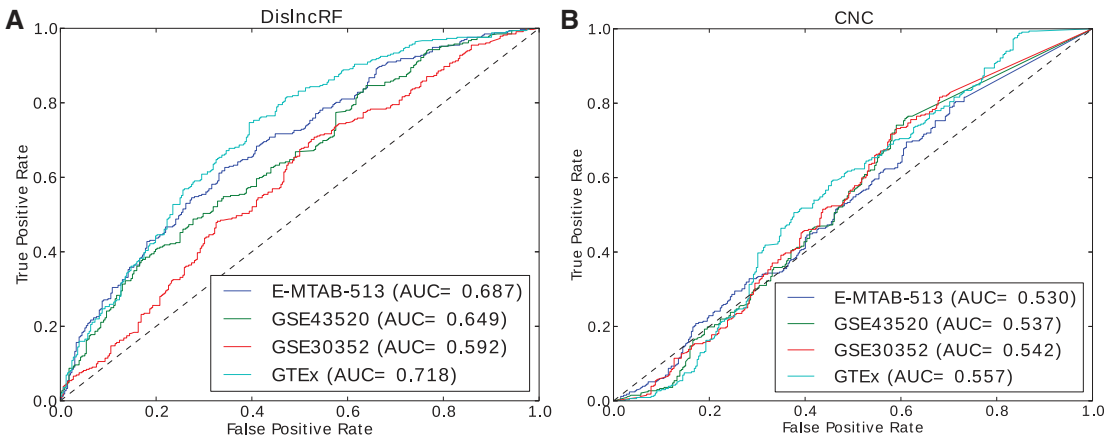


Fig. 3. ROC comparison for inferring disease-associated lncRNAs on E-MTAB-513, GSE43520, GSE30352 and GTEx using DislncRF and CNC. (**A**) The performance of DislncRF; (**B**) The performance of CNC

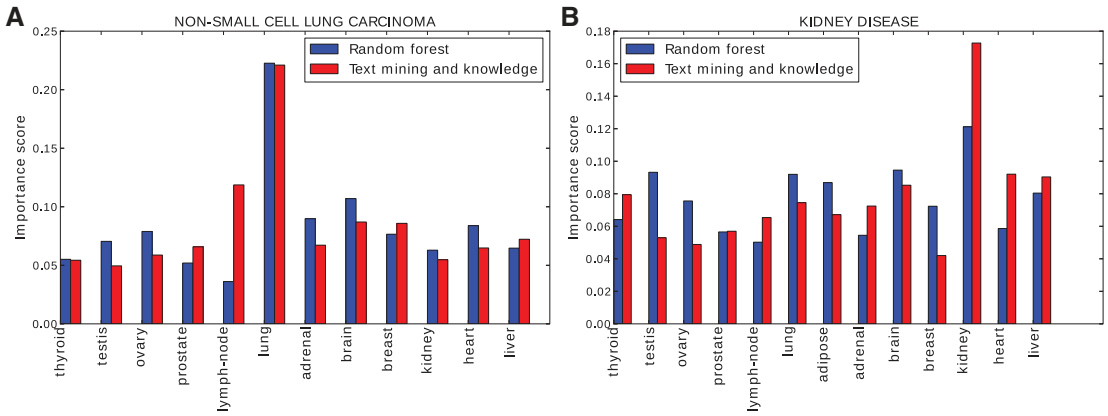


Fig. 4. Tissue importance for disease-associated gene detection are from random forest feature analysis, association scores of text mining and curated knowledge are extracted from DISEASES database. (**A**) The tissue importance for non-small cell lung carcinoma; (**B**) The tissue importance for Kidney disease

higher expressed than lncRNAs and consequently are observed to be expressed in more tissues using any given cutoff on expression level (Supplementary Fig. S3).

While some performance measures depend on the balance between the number of positive and negative examples others do not. In Tables 1 and 2, the first three measure are independent of this balance, whereas the last three all decrease as the fraction of negative examples increases. Since the number of false positive predictions at a given score cutoff is directly proportional to the number of negative examples, the Precision will scale as $1/(1+x)$ where x is the ratio of negative to positive examples. The AUPRC metric will scale identically since recall does not depend on the ratio. In case of MCC, the relationship is a bit more complex but can also be explicitly modelled.

3.3 Comparing with baseline methods CNC and KNN

To put the performance of DislncRF into perspective, we compare its prediction performance with baseline method CNC, which is evaluated as being exactly done for DislncRF. As can be seen in Figure 3B, the performance of CNC is close to random (roughly AUC = 0.54) in all four studied datasets when $K = 3$. We also tested $K = 1$ and $K = 5$ for CNC, the K value has no big impact on the performance (Supplementary Fig. S4). The results indicate that the PCC-based significantly co-expressed disease-associated PCGs are not enough to infer high-confidence disease-associated lncRNAs, and it easily suffers to noises. However, machine learning-based DislncRF use a bunch of positive and negative training data to learn expression patterns, which are more robust to noises. CNC has higher runtime than DislncRF, especially for prediction step, which requires calculating Pearson correlation coefficients between all disease-associated PCGs and candidate lncRNAs.

To evaluate KNN with $K = 50$, we, for each lncRNA, checked the number of disease-associated PCGs belonging to the 50 nearest PCGs by their PCC values. Using a threshold of more than one nearest neighbor to predict the disease association result in no true positive predictions and using the minimum threshold one neighbor result in 6 true positives, 497 false positives and a precision of 1.2%, which is less than the prior of 2.4%. This shows that KNN performs no better than random for this problem, showing that it is not usable as a baseline.

We observe that the supervised strategy of DislncRF outperforms the unsupervised strategy of the two coexpression-based methods, which might not be a surprise since the supervised framework can take much more features than only co-expression into account.

3.4 Linking tissues to diseases

We also analyzed the tissue importance based on the RF feature selection. The RFs rank the importance for individual tissues in the detection of disease-associated genes. As shown in Figure 4, DislncRF identifies the most important tissues as being lung for non-small cell lung carcinoma and kidney for kidney disease. The results agree with the curated and text mining tissue–disease associations and obviously make biological sense. In addition, we calculated the PCC between importance scores from the RF model and the associated scores from curated and text mining for 237 diseases. When averaging over the diseases we obtain PCCs of 0.313, 0.759, 0.440 and 0.348 on E-MTAB-513, GSE43520, GSE30352 and GTEx, respectively. Contrary to the baselines CNC and KNN, the RF-based DislncRF is often able to point to the tissue that is important for a given disease. It shows the RF models are biologically meaningful, which gives more reason to believe their predictions.

3.5 Case studies

To demonstrate the applicability of DislncRF to predict disease–lncRNA associations genome-wide, we present case studies of lncRNAs associated to prostate cancer and inflammatory bowel disease, for which some of disease associations can be confirmed from recent published studies (Mirza *et al.*, 2015; Ning *et al.*, 2016).

3.5.1 Prostate cancer

We extracted experimentally verified prostate cancer associated lncRNAs from Lnc2Cancer database (Ning *et al.*, 2016). In total, 27 are verified lncRNAs associated with prostate cancer. To demonstrate the prediction accuracy for prostate cancer, we pool the predicted scores of prostate cancer from DislncRF for the 27 verified lncRNAs and 27 other randomly selected lncRNAs from the remaining 13 309 lncRNAs. As shown in Figure 5, DislncRF yields the AUCs of 0.750, 0.620, 0.695 and 0.875 on E-MTAB-513, GSE43520, GSE30352 and GTEx, respectively.

For GTEx, we additionally checked the top-20 lncRNAs that were predicted to be associated with prostate cancer but not annotated as such in the Lnc2Cancer database. Of the top-20 lncRNAs predicted by DislincRF to be associated with prostate cancer, MIG205HG has the most compelling literature support. This

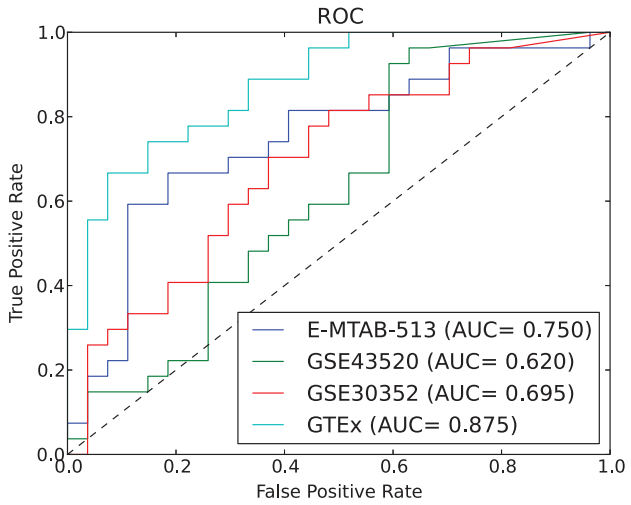


Fig. 5. The prediction performance of DislncRF for prostate cancer

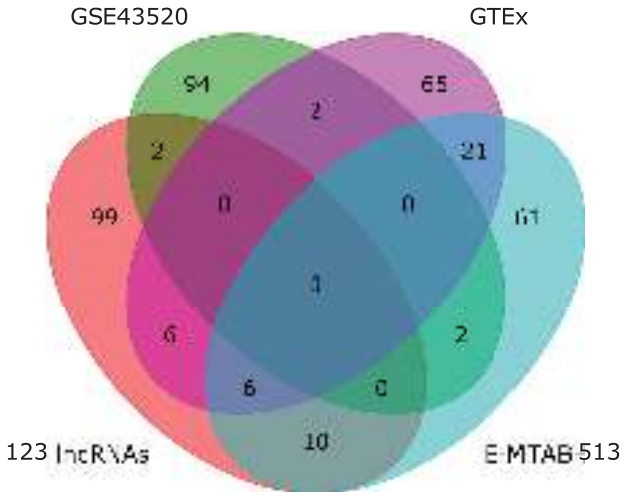


Fig. 6. Venn diagram for the 123 significantly up- and down-regulated lncRNAs associated with IBD and the top-100 predicted candidates from DislncRF on E-MTAB-513, GSE43520, GTEx. There are 28 overlapped lncRNAs between the top 100 lncRNAs of E-MTAB-513 and GTEx. Of these, six are significantly differentially expressed lncRNAs that are associated to IBD. Both E-MTAB-513 and GTEx have only two shared lncRNAs with GSE43520

lncRNA is produced from the same host gene as miR-205-5p, a miRNA that is differentially expressed in prostate cancer (Verdoodt *et al.*, 2013) and associated with prostate cancer risk and progression (Luu *et al.*, 2017). MIR205HG has also been shown to deplete miR-590-3p (Di *et al.*, 2018), which is implicated in prostate cancer (Sun *et al.*, 2017). Very recently, MIG205HG was also found to be differentially regulated in a reanalysis of two published prostate cancer transcriptomics studies (Ye *et al.*, 2018). A second candidate, RP11-7K24, has similarly been shown to be differentially expressed in prostate cancer (Han *et al.*, 2017), but did not have any additional supporting evidence. For most of the remaining candidates, we were unable to find any literature at all, which only emphasizes the need for new tools to study them.

3.5.2 Inflammatory bowel disease (IBD)

In a recent study (Mirza *et al.*, 2015), we performed microarray expression profile for IBD including Crohn disease (CD) and ulcerative colitis (UC). To detect dysregulated lncRNAs in IBD, differential expression analysis was carried out using the LIMMA package (Smyth, 2004) to identify up- and down-regulated lncRNAs based on inflamed UC vs Control and inflamed CD vs control. Here, we use this dataset for further testing of the DislncRF IBD predictions.

From (Mirza *et al.*, 2015), we extracted significantly up- and down-regulated lncRNAs associated with IBD using two criteria: False Discovery Rate (FDR) <0.01 and an absolute fold change (FC) >2. In total, we got 123 differentially expressed lncRNAs. Of these, 16, 2, 1 and 12 lncRNAs belong to the top 100 candidate IBD lncRNAs predicted for E-MTAB-513, GSE43520, GSE30352 and GTEx, respectively. The corresponding Venn diagram is shown in Figure 6.

Furthermore, we obtained 39 IBD loci associated to differentially expressed lncRNAs from Mirza *et al.*, 2015. We evaluated the top 100 lncRNA candidates overlapping with these (Table 3). For GTEx dataset, we can find five verified lncRNAs in top 100 candidates.

We also investigated the ROC curve of the 123 differentially expressed lncRNAs (DELs) and the 39 loci associated lncRNAs (LALs) for IBD. We first obtained the predicted scores between IBD and all lncRNAs using DislncRF. For each lncRNA in the 123 DELs and 39 LALs, we directly extracted the predicted IBD scores, and randomly selected the same number of unique negative lncRNAs not in the 123 DELs and 39 LALs, respectively. Then we pool them to evaluate the performance (Figure 7). The results demonstrate that DislncRF performs the best in GTEx for DELs and LALs, with an AUC of 0.748 and 0.753, respectively. The analysis also indicates that more tissues can provide more informative signals for the IBD-associated gene detections, even if some of them are not considered relevant to IBD.

4 Discussion

We considered the disease-associated gene detection as a supervised learning problem, which requires high-quality training data. We extracted high-confidence disease–PCG associations from the DISEASES database as the positive set. Since the DISEASES database is based on current knowledge, there are bound to be true disease–PCG associations, which have not yet been discovered. Therefore, when benchmarking against a database like DISEASES, even a perfect method would make correct, novel predictions that get counted as false positives. Still, a better method will obviously perform better on the given benchmark data. We can thus conclude that DislncRF outperforms existing methods.

There are several promising avenues to explore and possibilities to improve DislncRF. (i) More biological information should be integrated, such as GO information, SNPs and genomic context. For example, disease-associated lncRNAs are generally related to neighbor disease-associated PCGs on genome (Kumar *et al.*, 2013). This is useful information, which could be integrated into DislncRF. One simple strategy is that the predicted scores of DislncRF are scaled up (multiply a factor greater than 1) for lncRNAs in upstream and downstream of disease-associated PCGs, otherwise scale down (multiply a factor smaller than 1). (ii) We can construct machine learning models on a training set consisting of PCGs profiles, and then apply them for the test set with lncRNAs profiles. This requires that the training and test sets are taken from the same data distribution, while not overlap in the same or have near identical examples, as this will lead to overfitting. However, as indicated in Supplementary Figure S3, there are some differences in tissue

Table 3. The number of 39 IBD loci associated differentially expressed lncRNAs in predicted top 100 candidates among 13 336 lncRNAs. - mean no candidates

Dataset	lncRNAs being in top 100	Rank
E-MTAB-513	SLC12A5-AS1, PSMB8-AS1,RP11-94L15, IHCP5	14th, 24th, 32th, 62th
GSE43520	–	–
GSE30352	SMIM25, AL357060, RAJ009632	27th, 37th, 43th
GTEx	PSMB8-AS1, HCP5, AC245128, RP11-94L15, RP11-290F20	12th, 27th, 32th, 41th, 43th

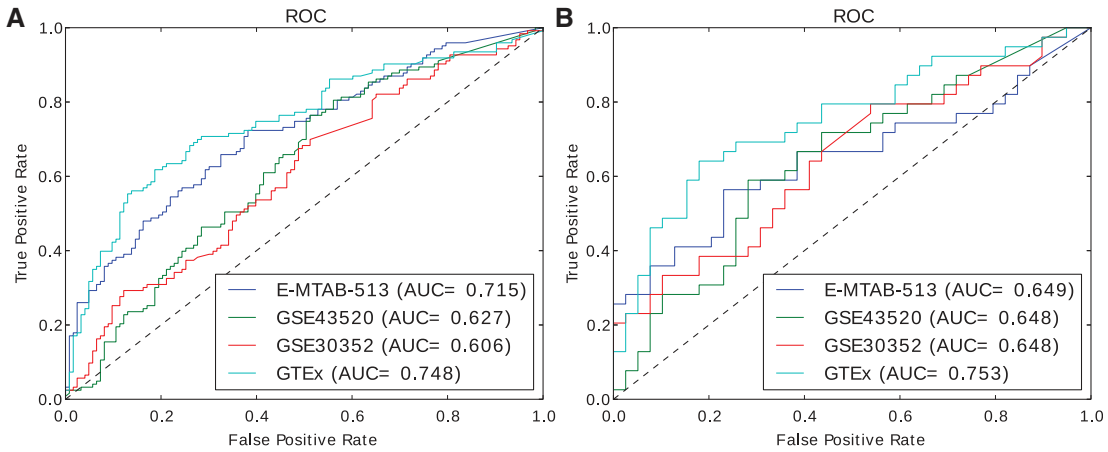


Fig. 7. The ROC curve for (**A**) the 123 differentially expressed lncRNAs and (**B**) the 39 loci associated lncRNAs of IBD

distributions between PCGs and lncRNAs, which leads to a covariate shift (Shimodaira, 2000) between training and test set. However, this is at least partially addressed by how we normalize the expression profiles.

Other future work includes extending the classification scheme from training a binary classifier for each disease to developing a multi-class framework, taking into account the full contingency matrix of how often genes associated with one disease are predicted for another disease. The overall performance of such a framework can be evaluated through a multi-class correlation coefficient (Gorodkin, 2004) and will allow for more detailed assessment of which diseases hard to distinguish from one another.

5 Conclusion

In this study, we presented DislncRF, a method for prediction of disease-associated lncRNAs using genome-wide tissue expression profiles and disease-associated PCGs. In contrast to case–control studies, which require disease-specific datasets to detect differentially expressed genes for diseases, DislncRF is based on training of multiple balanced random forest models on generic tissue expression profiles from disease-associated PCGs. This enabled us to learn expression patterns for disease-associated genes, which were then applied to infer disease-associated lncRNAs. DislncRF yielded promising performance and performed considerably better than baseline methods based on correlation coefficients, such as the coding-non-coding co-expression based method CNC. Analyzing the feature importance of the RF models showed that DislincRF consistently puts most emphasis on the tissues known to be important for each disease. The source code of DislncRF is freely available at https://github.com/xypan1232/DislncRF and http://rth.dk/resources/DislncRF.

Acknowledgements

We thank Alberto Santos Delgado for comments on an earlier version of this manuscript.

Funding

This work was supported by PhD fellowship from University of Copenhagen, the Innovation Fund Denmark (0603-00320B), the Novo Nordisk Foundation (NNF14CC0001) and the Danish Center for Scientific Computing (DCSC, DeIC).

Conflict of Interest: none declared.

References

Antanaviciute,A. *et al.* (2015) GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles. *Bioinformatics*, **31**, 2728–2735.

Binder,J.X. *et al.* (2014) COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database (Oxford)*, ***2014***, bau012.

Blokzijl,F. *et al.* (2016) Tissue-specific mutation accumulation in human adult stem cells during life. *Nature*, **538**, 260–264.

Bornigen,D. *et al.* (2013) Concordance of gene expression in human protein complexes reveals tissue specificity and pathology. *Nucleic Acids Res.*, **41**, e171.

Brawand,D. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.

Chen,G. *et al.* (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res.*, **41**, D983–D986.

Chen,X. (2015) KATZLDA: kATZ measure for the lncRNA-disease association prediction. *Sci. Rep.*, **5**, 16840.

Chen,X. *et al.* (2016a) IRWRLDA: improved random walk with restart for lncRNA-disease association prediction. *Oncotarget*, **7**, 57919–57931.

genes

MDPI

Article

WebCircRNA: Classifying the Circular RNA Potential of Coding and Noncoding RNA

Xiaoyong Pan [1,2,3], Kai Xiong [2,4], Christian Anthon [1,2,4], Poul Hyttel [2,4], Kristine K. Freude [2,4], Lars Juhl Jensen [1,3,*] and Jan Gorodkin [1,2,4,*]

1 Center for Non-Coding RNA in Technology and Health, University of Copenhagen, 1870 Frederiksberg C, Denmark; xypan172436@gmail.com (X.P.); anthon@rth.dk (C.A.)
2 Department of Veterinary and Animal Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark; hpw927@alumni.ku.dk (K.X.); poh@sund.ku.dk (P.H.); kkf@sund.ku.dk (K.K.F.)
3 Department of Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark
4 BrainStem—Stem Cell Center of Excellence in Neurology, University of Copenhagen, 1870 Frederiksberg C, Denmark
* Correspondence: lars.juhl.jensen@cpr.ku.dk (L.J.J.); gorodkin@rth.dk (J.G.)

Received: 7 September 2018; Accepted: 2 November 2018; Published: 6 November 2018

Abstract: Circular RNAs (circRNAs) are increasingly recognized to play crucial roles in post-transcriptional gene regulation including functioning as microRNA (miRNA) sponges or as wide-spread regulators, for example in stem cell differentiation. It is therefore highly relevant to identify if a transcript of interest can also function as a circRNA. Here, we present a user-friendly web server that predicts if coding and noncoding RNAs have circRNA isoforms and whether circRNAs are expressed in stem cells. The predictions are made by random forest models using sequence-derived features as input. The output scores are converted to fractiles, which are used to assess the circRNA and stem cell potential. The performances of the three models are reported as the area under the receiver operating characteristic (ROC) curve and are 0.82 for coding genes, 0.89 for long noncoding RNAs (lncRNAs) and 0.72 for stem cell expression. We present WebCircRNA for quick evaluation of human genes and transcripts for their circRNA potential, which can be essential in several contexts.

Keywords: Circular RNA; random forest; noncoding RNA

1. Introduction

Circular RNAs (circRNAs) were recently discovered to be widespread, abundant, expressed across species, and implicated in several diseases. They are created by non-linear backsplicing between a splice donor and an upstream splice acceptor, and evidence is emerging for them playing functional roles as microRNA (miRNA) sponges [1,2] and in regulation of gene splicing and transcription [3]. Recently, the miR-7 sponge *CDR1as* has been found to be involved in stem cell regulation of periodontal ligament [4]. Other studies suggest that circRNAs can encode proteins [5], and 90% of the 92,375 human circRNAs in the circBase database (v0.1) [6] arise from protein-coding genes (PCGs). The number of discovered circRNAs has been rapidly increasing in recent years due to the development of new high-throughput sequencing technologies, and circBase now contains more than 90,000 circRNA transcripts [6]. In addition, circRNAs are expressed in a cell/tissue-specific manner [2]; for example, 16,017 are expressed in stem cells, and they are especially prominent during embryonic development [7].

Current computational pipelines are focused on identifying presence of backsplicing junction-spanning reads from RNA-seq data [8]. Commonly, pipelines to identify circRNAs map

the RNA-seq reads into a reference genome using mappers such as TopHat [9], and then use the unmapped reads to detect the backsplicing junction spanning reads. This principle is used in circRNA detection programs such as CIRCexplorer [10] and find_circ [2]. As reported by Hansen et al. 2016 [11], these tools suffer from relatively high false positive rates, and dramatic differences are observed between the various tools. In contrast to these tools, which take RNA-seq data as input, we employ a strategy based solely on predictions from the primary sequence.

Our strategy takes outset in learning sequence-derived patterns from accumulated identified circRNAs using machine learning, and apply the trained models to filter out falsely annotated circRNAs as a post-processing step. For a given sequence, our tool outputs three scores: the first two signify the potential of the transcript being a circRNA under the assumption that it is a PCG or a long noncoding RNA (lncRNA), respectively, and the third scores how likely it is to be expressed in stem cells if it is indeed a circRNA. Underlying the three respective types of output scores are three random-forest models: (i) circular RNA potential of PCGs (CP-PCG); (ii) circular RNA potential of lncRNAs (CP-lncRNA), which is based on the work in Pan and Xiong 2015 [12]; and (iii) stem cell potential of circRNAs (SP-circRNA). We introduce calibrated scoring schemes for the three types of predictions and furthermore make the method available as a user-friendly web server, which takes one or more transcripts as input, either in the form of genome coordinates or nucleotide sequences.

2. Materials and Methods

In this study, we present a machine learning based method to classify the circular RNA potential for coding and non-coding RNA (Figure 1). Data from circBase and GENCODE v19 [13] were used to create the training data. From these, we extracted different features such as sequence composition and graph representations of, e.g., RNA secondary structure and conservation. We then trained random forest models to perform the classification based on the extracted features.

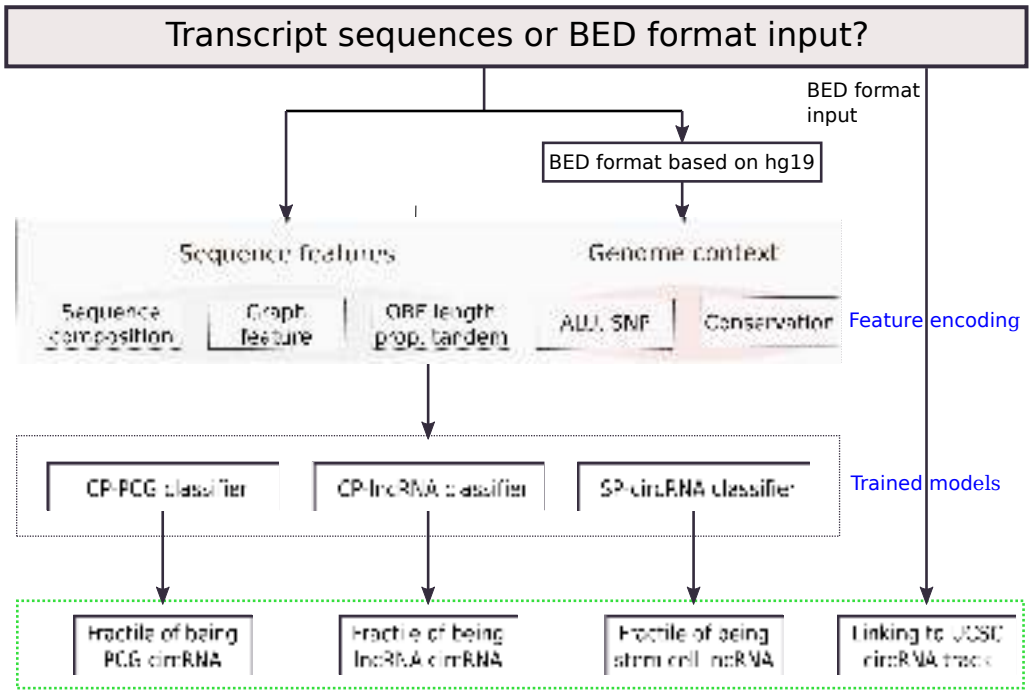


Figure 1. Flowchart of the WebCircRNA framework. BED: browser extensible data; ORF: open reading frame; ALU: transposable element; SNP: single nucleotide polymorphism; CP: circular RNA potential; PCG: protein coding gene; lncRNA: long non-coding RNA; SP: stem cell potential; circRNA: circular RNA; UCSC: University of California, San Diego.

2.1. Construction of Datasets

We downloaded 92,375 circRNA transcripts from circBase [6]. For circRNAs, we removed transcripts shorter than 200 nt and overlapping circRNA transcripts, which resulted in a set of 14,084 circRNAs also used in PredcircRNA [12]. We also collected 20,345 PCGs from GENCODE v19. To ensure a clean dichotomy for training, we removed the PCGs overlapping circRNAs in circBase resulting in 9533 PCGs used in the following. We also extracted the lncRNAs using the same processing scheme as for PCGs, resulting in 19,722 lncRNAs. We randomly selected 10,000 for training lncRNAs and used the remaining 9722 lncRNAs for independent testing. Of the 10,000 training lncRNAs, 3500 are lincRNAs and 75 overlap with PCGs. Next, we constructed the dataset for circRNAs versus PCGs. We randomly selected 10,000 circRNAs and 8000 PCGs for training and used the remaining 4084 circRNA and 1533 PCGs as independent test set. For stem-cell dataset, we obtained 2082 circRNAs only expressed in H1hsec and randomly selected the same number of other circRNAs from other cell lines, not expressed in H1hsec. We used 1800 stem cell circRNAs and 1800 circRNAs not expressed in H1hsec as training dataset and the remaining as independent testing dataset. An overview of how these were used for training and testing is provided in Table 1.

Table 1. The details of training and independent test sets. The table summarizes which sequences were used as positive and negative examples for the respective random forest (RF) models.

Model	Positive Data	Negative Data
circRNA vs PCG	Total: 14,084 circRNAs Training: 10,000 Independent testing: 4084	Total: 9533 PCGs not overlapping with circRNAs Training: 8000 Independent testing: 1533
circRNA vs lncRNA	Total: 14,084 circRNAs Training: 10,000 Independent testing: 4084	Total: 19,722 lncRNAs not overlapping with circRNAs Training: 10,000 Independent testing: 9722
Stem cell vs not	Total: 2082 circRNAs Training: 1800 Independent testing: 282	Total: 2082 circRNAs Training: 1800 Independent testing: 282

To validate the WebCircRNA on non-human data, we collected the circRNA sequences of *Mus musculus* from circBase and the PCGs and lncRNAs of *Mus musculus* from GENCODE. We obtained 5657 mouse circRNAs, 3904 mouse lncRNAs and 16,763 mouse PCGs. We used CD-HIT [14] to remove all sequences that are more than 80% identical to a human sequence. We obtained 5397 mouse circRNAs, 3700 lncRNAs and 16,552 PCGs as mouse test data. The sequences of 5397 circRNAs and 3700 mouse lncRNAs are tested on CP-lncRNA model trained on human data. Similarly, the sequences of 5397 circRNAs and 16,552 mouse PCGs are evaluated on CP-PCG model.

2.2. Feature Encoding

We extracted 178 features as described below, which are summarized in Table 2. The values for each feature were normalized to the interval from 0 to 1.

Table 2. The 178 extracted features divided into four groups.

Feature Group	Feature Names
Basic sequence features	Length; AG, GT, GTAG, AGGT, GC content; 64 trinucleotide frequencies
Graph features	Top 101 graph features from GraphProt 1.0.1
Conservation features	Mean, standard deviation of conservation score
Other features	ALU, tandem, ORF length, ORF prop, SNP density

Basic sequence features. These are made up of the features that are trivially derived from the sequence. They include the gene length, its GC content and the 64 trinucleotide frequencies. They also

comprise the GT, AG, GTAG and AGGT frequencies; the rationale behind this is that the GT/AG signal associates with exon-junction [15] and backsplicing.

Graph features. For each sequence, we generated a set of RNA structure features using GraphProt 1.0.1 [16]. The rationale for this is that RNA structure plays key roles in gene splicing [17] , which has an impact on forming circRNAs, such as backsplicing [18]. The RNA graph uses nodes to represent the nucleotides and edges to represent the bond relationships between the nucleotides. The input for GraphProt is RNA sequences, whose corresponding graphs are created using the script fasta2shrep_gspan.pl (with parameters -seq-graph-t -nostr -stdout -fasta). The resulting graph is used as input to GraphProt (using -a FEATURE) to create a set of more than 30,000 graph features. To reduce the high-dimensional graph features, we applied random forest, implemented in Scikit-learn [19], to rank importance score for graph features based on randomly selected circRNA and non-circRNA subset. By ranking the graph features based on the random forest feature score, measured on 1747 circRNAs and 1747 other RNAs, we chose the Top 101 graph features with high importance score (see Section 2.3). We tried Top 50, 101 and 200 graph features. When combining with 77 other features in Table 2, the Top 50, 101, and 200 yield similar five-fold cross-validation performance on circRNA vs. PCG training data with the area under receiver operating characteristic (AUC) values of 0.798, 0.797 and 0.793, respectively. Hence, we used the Top 101 graph features for the three models.

Conservation features. These were calculated from per-base phyloP conservation score [20]. For each sequence, we calculated the mean and standard deviation of the conservation scores across the length of the sequence.

Other features. ALU repeats can make the splice sites recognize each other, which promote circularization. Therefore, we derived the ALU frequency for each sequence from RepeatMasker track of the UCSC Genome Browser [21]. Similarly, tandem duplications within a gene can promote backsplicing; we detected them using Tandem Repeats Finder [22], and calculated the frequency of tandem repeats over the sequence. To evaluate the protein-coding potential of each sequence, we extracted the longest Open Reading Frame (ORF) using txCdsPredict from the UCSC Browser. From this, we calculated the ORF length and the ORF propensity (ORF prop), which is ORF length divided by the sequence length. Finally, SNP information is also integrated by calculating the SNP density using data from the 1000 Genomes Project [23].

2.3. Random Forest Models

Random forests [24] are constructed from multiple unpruned decision trees, with each tree grown from bootstrap sampling of the training data and using a random subset of the input features. During bootstrap sampling, 2/3 of data were used for decision tree training, and 1/3 of data (out-of-bag data) were used for inner validation. In this study, we used the random forest implementation from Scikit-learn [19]. We optimized the number of trees in the random forest based on the cross-validated performance. We used the following number of trees: 80 for CP-PCG, 100 for CP-lncRNA and 60 for SP-circRNA. The optimal parameters are searched in the range from 10 to 100 with a step size of 10, and we selected the value with the highest performance of a five-fold cross-validation on the training set. It should be noted that feature selection was not involved in the model optimization, as we only ranked graph features in a separate data subset, as also mentioned in Section 2.2.

Random forest models can also be used to rank the input features based on their individual contributions. During the training process, the out-of-bag (OOB) error was kept and averaged over all trees. Then, for one feature at a time, its feature values were randomly shuffled over the OOB samples, and the OOB error was re-computed. The importance score for a feature was obtained by averaging the difference in OOB error over all trees.

2.4. Prediction Scores

The training of the three respective classifiers CP-PCG, CP-lncRNA and SP-circRNA was done using a five-fold cross validation. To calculate a combined score of the five resulting cross-validated

models, we first converted the score from each model to a fractile within the score distribution of the respective validation set. Next, the average of the five fractile scores was calculated (Figure 2). Calculating the average fractile score is equivalent to calculating the average rank, since the respective lists of genes used to train each classifier are equally long within a cross-validation ensemble. What we used to combine the scores to across an ensemble of classifiers is thus rank aggregation, which is an established method in the literature [25]. For CP-PCG and CP-lncRNA, the higher is the score, the more likely the sequence is to be a circRNA. Similarly, a higher fractile score implies that a circRNA is more likely to be expressed in stem cells in the case of SP-circRNA. To calibrate the scores of the three classifiers, we estimated the false positive rate (FPR) as a function of the fractile from the score distribution on the negative data (outlined in Table 1).

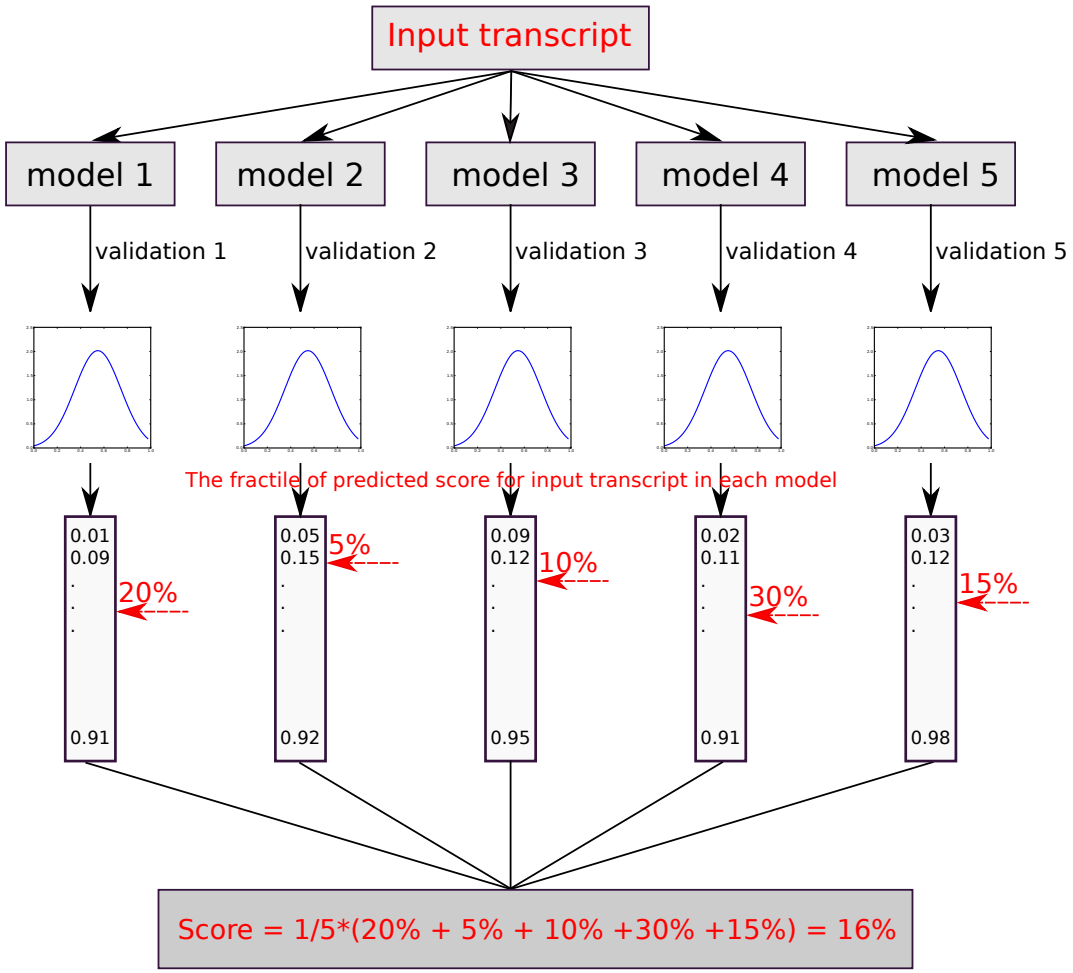


Figure 2. The flowchart illustrates how the final fractile score of the input sequences is obtained. Each model predicts a score which is then converted into a fractile. Novel sequences not in the validation sets are scored relative to the fractile in each model and then averaged over all five models.

3. Results

3.1. Performance Evaluation

To assess the quality of the predictions provided by the webcircRNA web server, we evaluated the performance of each of the three classifiers on independent test sets. Evaluating the performance of CP-PCG on the independent test set, we achieved a sensitivity of 0.736, a specificity of 0.752, and an AUC of 0.821 (Figure 3A). For CP-lncRNA, we obtained similar test set performances with a sensitivity of 0.827, a specificity of 0.806, and an AUC of 0.889 (Figure 3B). Finally, SP-circRNA presented a sensitivity of 0.688, a specificity of 0.673, and an AUC of 0.718 (Figure 3C). Here we do not

compare WebCircRNA with other RNA-seq based methods (e.g., find_circ and CIRCexplore), since they take completely different starting points. WebCircRNA takes sequences of assembled transcripts as inputs, but find_circ and CIRCexplore starts from RNA-seq data. WebCircRNA can be used as a post-processing step to remove incorrectly annotated circRNAs from those RNA-seq based tools.

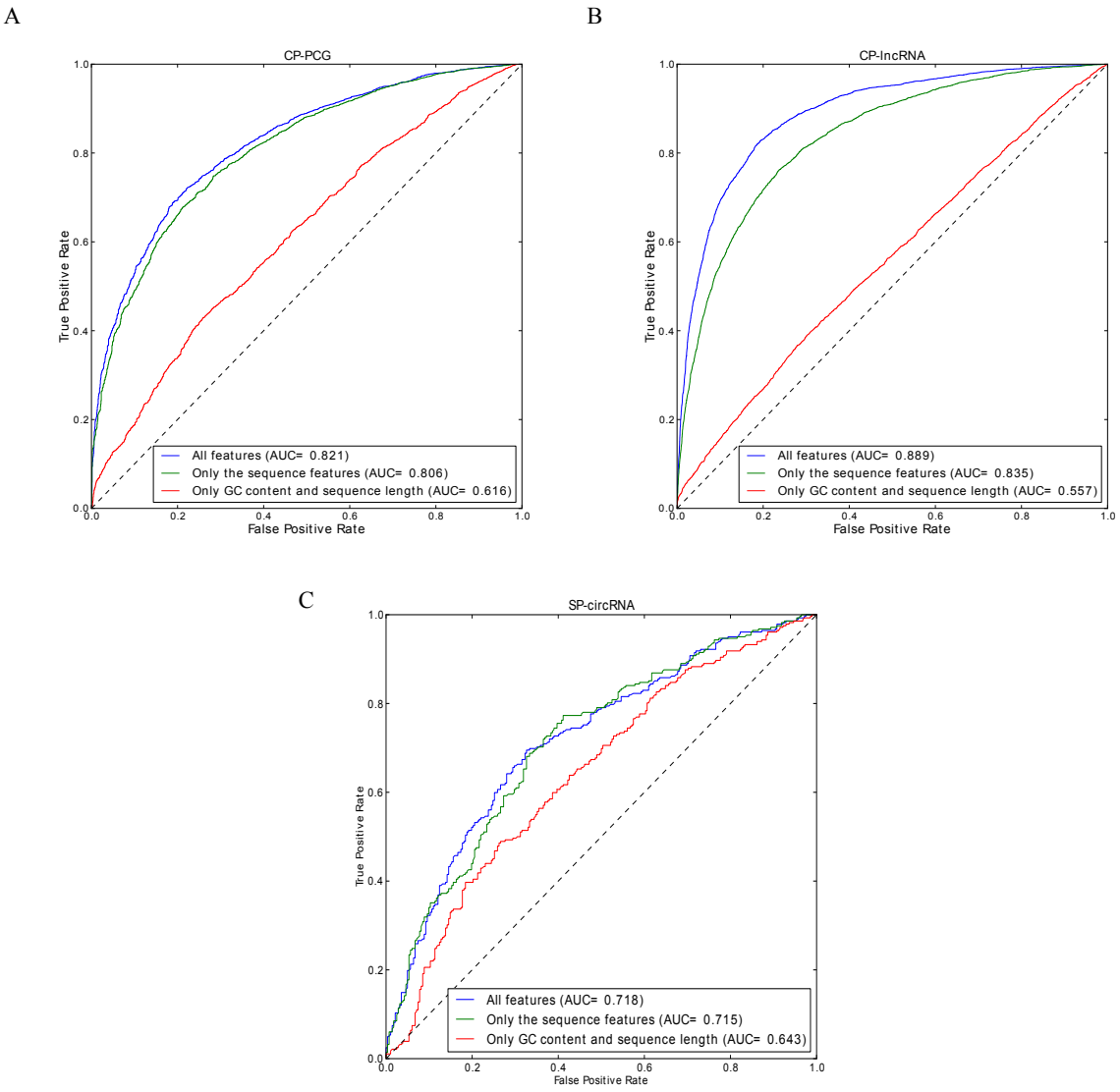


Figure 3. ROC curves for: (**A**) the PCG circRNA model (CP-PCG); (**B**) the lncRNA circRNA model (CP-lncRNA); and (**C**) stem cell circRNA model (SP-circRNA). In these three instances, the ROC curve using all 178 features indicated in Table 2 is compared to models using "only the GC content and sequence length" and "only the sequence features".

To test the CP-PCG and CP-lncRNA models on non-human organisms, we applied them on mouse test data. The receiver operating characteristic (ROC) figure is shown in Figure 4. We obtained performances of AUC 0.811 of CP-PCG and AUC 0.924 of CP-lncRNA, indicating the CP-PCG and CP-lncRNA models trained on human data can also be applied for other organisms.

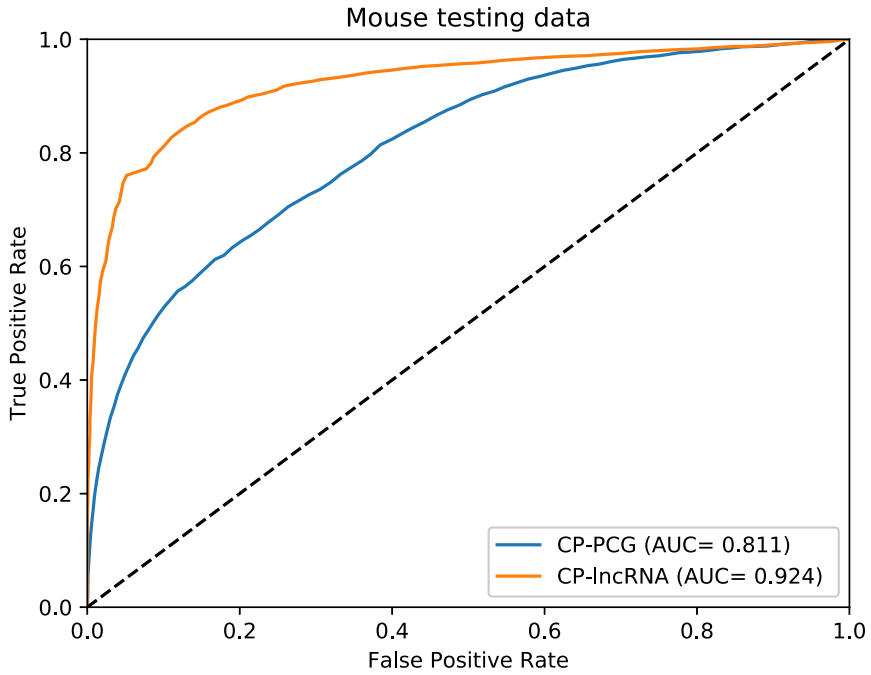


Figure 4. ROC curves for the testing mouse data on the CP-PCG and the CP-lncRNA, which are trained on human data using "only the sequence features", respectively.

3.2. Feature Importance

To test if more basic features can account for the performance, we also trained models on two simpler feature sets: one consisting only of the GC content and the sequence length, and one including all external features that we did not directly derive from the primary sequence (Table 2). The performance of these were compared to that of all features. Whereas the GC content and sequence length performs much worse than the full model (slightly better than random), leaving out the external features leads to only a minor drop in performance for CP-PCG and SP-circRNA but a notable drop for CP-lncRNA, as shown in Figure 3. These results show that it is not trivial to predict circRNAs from sequence alone. The difference in performance observed for external features between the classifiers likely reflects that PCGs are in general much more strongly conserved than lncRNAs [13], since conservation is among the external features. This highlights the importance of training separate circRNA classifiers for PCGs and lncRNAs.

This prompted us to further analyze which features are the most important for the classifiers. We therefore considered the Top-10 highest ranked features according the random forest models (Figure 5). Consistent with our performance observations, conservation features are by far the most important for lncRNAs, while they are less profound for PCGs. This is in agreement with the with observation that circRNAs often are evolutionarily conserved [26]. Conversely, sequence length matters more in the PCG and stem cell circRNA cases.

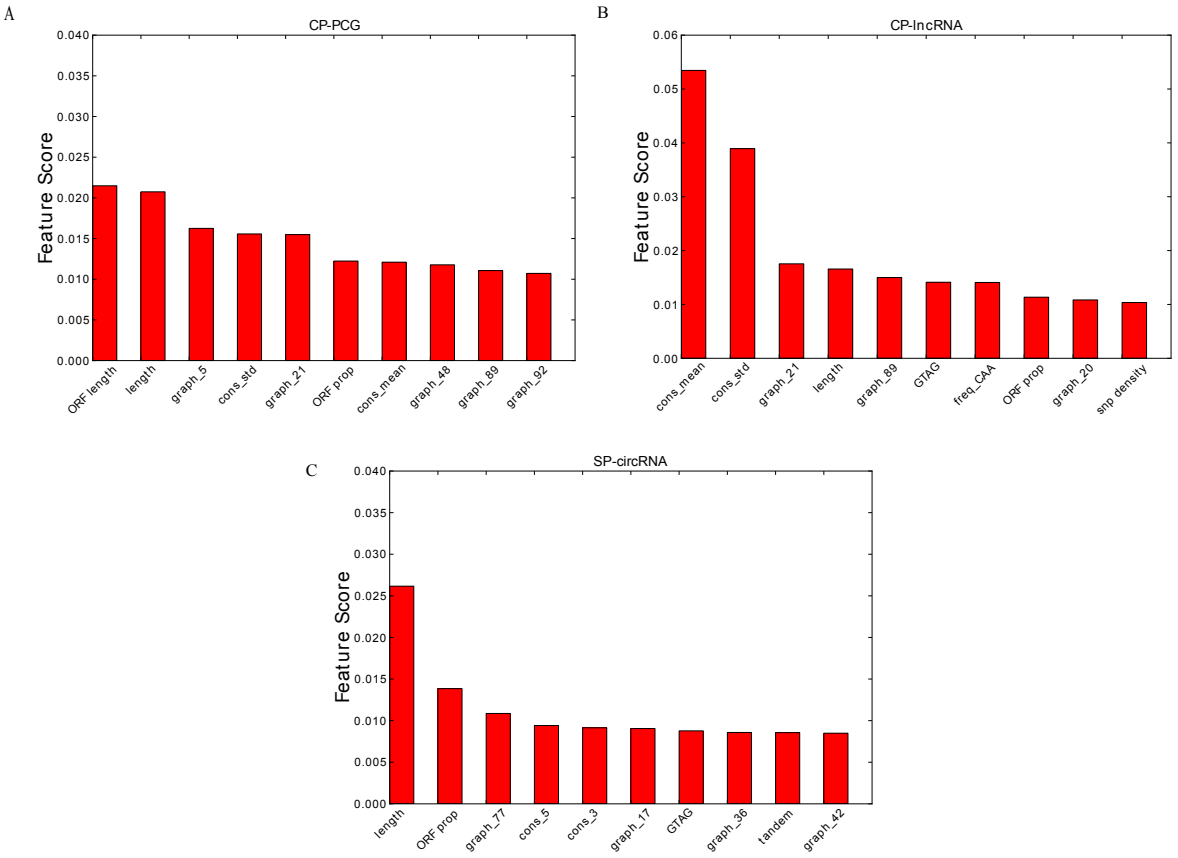


Figure 5. The Top 10 features for: (**A**) the CP-PCG model; (**B**) the CP-lncRNA model; and (**C**) the SP-circRNA model. Prefix graph refers to the 101 graph features, prefix cons refers to conservation feature and prefix freq refers to codon frequency feature.

3.3. The WebCircRNA Web Server

Using the graphical representation of scoring CP-PCG, CP-lncRNA and SP-circRNA, the web server takes two types of input, a BED file with coordinates of the human genome (hg19) or a FASTA file. Each line in the BED file is considered as an entity on which predictions are made. Whereas this provides flexibility in terms of interfacing with genome browsers, the mode of the FASTA file allows for another form of flexibility, such as spliced sequences or sequences from organisms closely related to human. When submitting in the BED format, the sequences must be from the human genome. When submitting in FASTA mode, any sequence will be accepted, however, one should be aware that the models were trained only on human data. The WebCircRNA software can be freely downloaded on the download page of the web site.

As examples for using the web server, we present the analysis of three genes CDKN2B-AS, DHDDS and OCT4, not part of the training data. (Figure 6). CDKN2B-AS is a lncRNA gene, whose circRNA isoform is associated with atherosclerosis via changing INK4/ARF expression [27]. WebCircRNA correctly predict this as circRNA with a score of 72% as seen in the column "lncRNA circRNA". Since we know that this gene encodes a lncRNA, the column "PCG circRNA" should be ignored. Conversely, DHDDS and OCT4 are PCGs; we thus consider the column "PCG circRNA" and not "lncRNAs circRNA". The score of 86% for DHDDS is in agreement with it having a circRNA isoform [28]. For OCT4, we obtain a score 47%, indicating that it is unlikely to produce a circRNA isoform. To our knowledge, there is indeed currently no evidence supporting that OCT4 can be a circRNA. The "stem cell circRNA" predictions suggest that CDKN2B-AS is likely expressed in stem cells, whereas DHDDS is not. Since OCT4 is not predicted to have a circRNA isoform, the stem cell prediction is not applicable to this gene.

Gene name	Position	PCG circRNA	FPR	lncRNA circRNA	FPR	Stem cell circRNA	FPR
CDKN2B-AS	chr9:22046749-22056386 (+)	75	3	72	4	60	27
DHDDS	chr1:26772806-26774151 (+)	86	1	79	2	33	68
OCT4	chr6:31132134-31138427 (-)	47	28	81	2	74	12

Figure 6. WebCircRNA. Example output for the lncRNA CDKN2B-AS and the two PCGs DHDDS and OCT4 is shown. We thus ignore the "PCG circRNA" score for CDKN2B-AS and the "lncRNAs circRNA" scores for DHDDS and OCT4. Because "stem cell circRNA" scores only apply to circRNAs, this too should be disregarded for OCT4, since it is not predicted to have a circRNA isoform. The respective "FPR" shows the estimated false positive rate of the corresponding methods. When submitting a BED file, the genomic context of any prediction can view in the UCSC browser via the link in the "Position" column.

4. Concluding Remarks

In this study, we addressed classifying protein coding gene (PCG) and long non-coding RNA (lncRNA) genes for their circRNA potential by using features mainly encoded from the primary sequence along with a few other features including conservation. We showed that this yielded performance superior to the basic methods using only features such as GC content and sequence length, demonstrating that this problem is non-trivial. Furthermore, as circular RNAs are getting increasing awareness within the context of stem cells, we included a classifier for circRNAs expressed in stem cells or not. Again, we find that, although less profound, sequence features in general improve the performance over a basic model making use of only GC content and sequence length.

In this study, we used random forest (RF) as the classifier. It has been shown that RF can outperform other methods in many classification tasks [29]. The choice of RF vs. some other machine learning algorithm is not what limits performance. The directions of research that could improve performance is work on obtaining better quality datasets on which to train the models, and possibly more work on feature engineering, i.e., adding new relevant input features for the classifiers.

We expect that the method works, e.g., for other mammals as well based on our mouse case. However, some minor improvements can probably be obtained if training is carried out individual organisms. The models trained on human data were tested on mouse data cleaned for high similarity (80%) to human. WebCricRNA still yields a high accuracy (Figure 4). However, it is still difficult to prove that the method works for other mammals, because, as homology to the human dataset makes it very hard—if possible at all—make a sufficiently large independent datasets for another mammals.

We provide a user friendly web interface by WebCircRNA, which allows the user to either upload a sequence (FASTA format) or a BED file (human genome only), which is a PCG, a lncRNA or a circRNA, to retrieve prediction of the circRNA potential (for PCG and lncRNA) or whether the circRNA is expressed in stem cells. The web server returns and visualizes the score and associated false positive rate from three models, namely CP-PCG, CP-lncRNA and SP-circRNA.

Author Contributions: X.P., K.X., L.J.J. and J.G. conceived the study. The prediction methods were developed by X.P. and K.X., The study is supervised by L.J.J., J.G., P.H. and K.K.F. The web server and software was implemented by X.P. and C.A. All authors contributed to writing the paper and approved the manuscript.

Funding: This work was funded by the Innovation Fund Denmark, Danish Council for Independent Research (Technology and Production Sciences), the Novo Nordisk Foundation (NNF14CC0001) and the Danish Center for Scientific Computing (DCSC, DeiC).

Acknowledgments: We thank Nikolai Hecker for useful discussions and advice.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results