Bioinformatics, 2019, 1–8
doi: 10.1093/bioinformatics/btz811
Advance Access Publication Date: 4 November 2019
Original Paper

Genome analysis

CASPR, an analysis pipeline for single and paired guide RNA CRISPR screens, reveals optimal target selection for long non-coding RNAs

Judith Bergadà-Pijuan[1,2], Carlos Pulido-Quetglas[1,2,3], Adrienne Vancura[1,2] and Rory Johnson[1,2,*]

[1]Department of Medical Oncology, Inselspital, Bern University Hospital, [2]Department for BioMedical Research and [3]Graduate School of Cellular and Biomedical Sciences, University of Bern, Bern 3012, Switzerland

*To whom correspondence should be addressed.

Associate Editor: Jan Gorodkin

Received on May 17, 2019; revised on October 8, 2019; editorial decision on October 27, 2019; accepted on October 30, 2019

Abstract

Motivation: CRISPR-Cas9 loss-of-function (LOF) pooled screening promises to identify which long non-coding RNAs (lncRNAs), amongst the many thousands to have been annotated so far, are capable of mediating cellular functions. The two principal LOF perturbations, CRISPR-inhibition and CRISPR-deletion, employ one and two guide RNAs, respectively. However, no software solution has the versatility to identify hits across both modalities, and the optimal design parameters for such screens remain poorly understood.

Results: Here, we present CRISPR Analysis for Single and Paired RNA-guides (CASPR), a user-friendly, end-to-end screen analysis tool. CASPR is compatible with both CRISPRi and CRISPR-del screens, and balances sensitivity and specificity by generating consensus predictions from multiple algorithms. Benchmarking on ground-truth sets of cancer-associated lncRNAs demonstrates CASPR's improved sensitivity with respect to existing methods. Applying CASPR to published screens, we identify two parameters that predict lncRNA hits: expression and annotation quality of the transcription start site. Thus, CASPR is a versatile and complete solution for lncRNA CRISPR screen analysis, and reveals principles for including lncRNAs in screening libraries.

Availability and implementation: https://judithbergada.github.io/CASPR/

Contact: rory.johnson@dbmr.unibe.ch

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

1 Introduction

CRISPR-Cas9 genome-engineering has been a breakthrough technology by enabling functional screening of non-protein-coding elements. Foremost amongst these are the long non-coding RNAs (lncRNAs), which are challenging to perturb using classical RNA interference (RNAi) technology (Stojic *et al.*, 2018). Deployment of CRISPR in high-throughput pooled screening format promises to discover functional and disease-related genes amongst the tens of thousands of lncRNA gene loci in the latest annotations (Esposito *et al.*, 2019).

Perturbation of lncRNAs requires distinct experimental design compared to protein-coding genes (PCGs). The latter require one single Cas9 protein, targeted by an engineered single guide RNA (sgRNA), to induce an indel mutation in the open reading frame (ORF) and achieve a complete loss-of-function (LOF) frameshift mutation (Esposito *et al.*, 2019). In contrast, lncRNAs have, by definition, no ORF. Thus, two principal LOF approaches have been applied (Fig. 1A). First, CRISPR-inhibition (CRISPRi), where transcriptional repression is achieved by an enzymatically-dead Cas9 (dCas9) fused to a repressor domain, such as KRAB (Liu *et al.*, 2017). Second, CRISPR-deletion (CRISPR-del), where a pair of Cas9 complexes is used to induce simultaneous DNA double-strand breaks flanking the lncRNA and thereby delete it (Aparicio-Prat *et al.*, 2015). CRISPRi is transient and requires one sgRNA; CRISPR-del is permanent and requires two sgRNAs, sometimes referred to as paired guide RNAs (pgRNAs).

In pooled CRISPR screens, functional lncRNAs are identified by the enrichment or depletion of their targeting sgRNAs between two or more populations of phenotypically distinct cells (Esposito *et al.*, 2019). Multiple unique sgRNA constructs are used for every target, and replicated experiments performed, in order to identify hits against a background of technical and biological noise. At the end of experiments, genomically-inserted sgRNA sequences are amplified

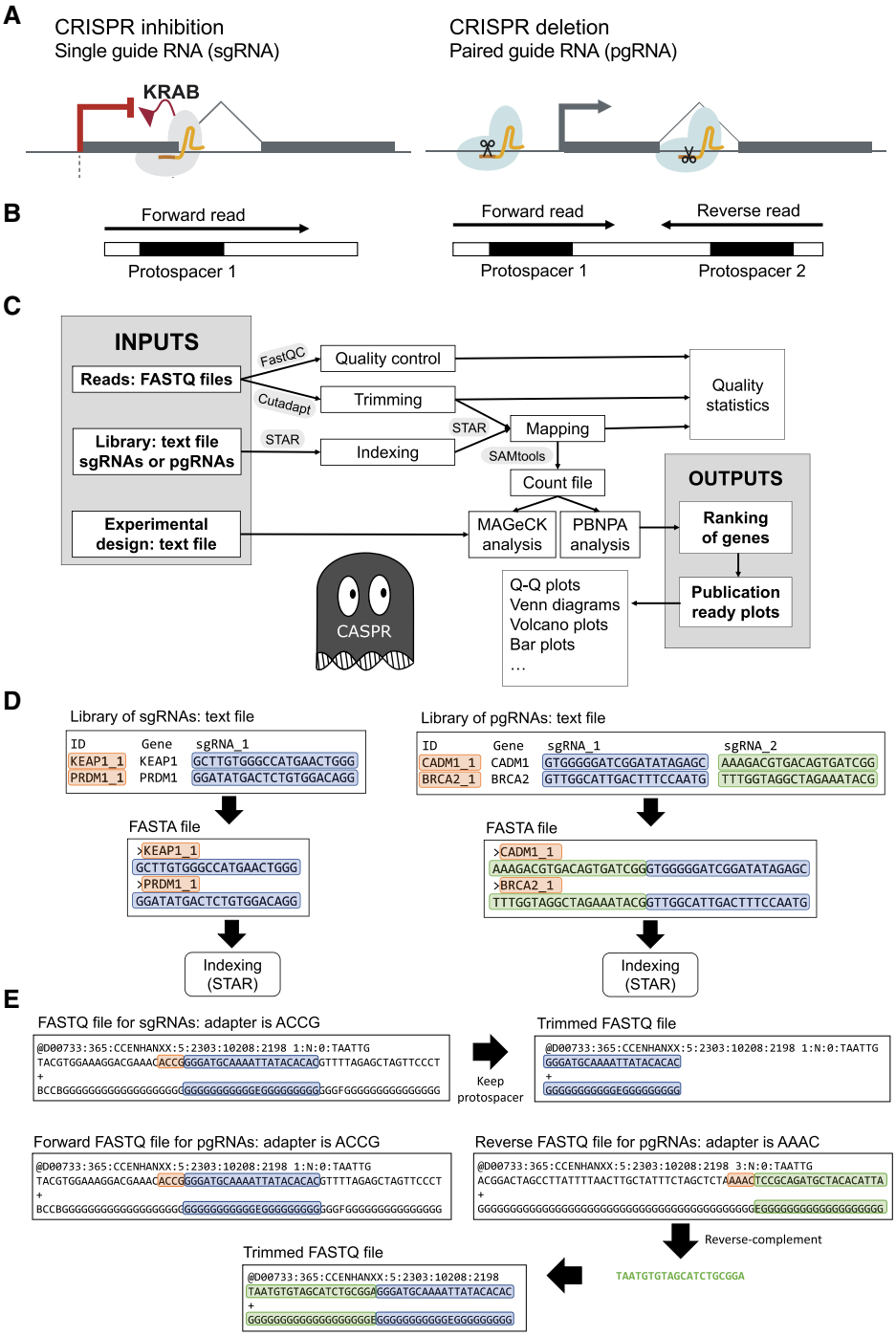

Fig. 1. (**A**) CRISPR loss-of-function perturbations. In CRISPRi, a single sgRNA is used to recruit a chimaeric protein composed of an effector domain (such as KRAB) fused to a catalytically-dead Cas9 (dCas9), and reversibly inhibits transcription of the target gene. In CRISPR-deletion (CRISPR-del), a pair of sgRNAs (pgRNA) recruit wild-type Cas9 endonucleases to sites flanking the target region (here, the gene's transcription start site), creating a genomic deletion and silencing gene expression. (**B**) sgRNAs can be uniquely identified by sequencing their 20 nt protospacer region. For sgRNA libraries, forward reads that contain the 20 nt protospacers are sufficient. For pgRNA libraries, forward and reverse reads are needed. (**C**) The CASPR pipeline. As inputs, it requires the sequencing reads, a design library of sgRNA protospacers, and an experimental design file defining the treatment and control samples. Then, it performs the quality control and trimming of reads, as well as the indexing of the sgRNA library. CASPR maps the reads to the library and provides a count table, which is used to identify gene hits by two different algorithms, MAGeCK and PBNPA. (**D**) Scheme for indexing the library. Protospacers are converted to FASTA format and indexed with STAR. For pgRNA libraries, CASPR concatenates the two protospacers. (**E**) Read trimming of raw sequencing reads to extract protospacer sequences. Resulting FASTQ file will be mapped to the indexed library

by PCR and sequenced by next generation sequencing (Fig. 1B). The unique 20mer protospacer(s) in each sgRNA or pair of sgRNAs, are used as molecular barcodes to track the frequency of each knockout cell population. A growing number of software packages has been created to identify hits from such screens, based on a variety of statistical approaches. These packages may either take as input the raw sequencing reads and provide an 'end-to-end' analysis (such as PinAPL-Py and MAGeCK) (Li *et al.*, 2014; Spahn *et al.*, 2017), or just accept count tables of already processed reads (such as BAGEL, STARS and PBNPA) (Doench *et al.*, 2016; Hart and Moffat, 2016; Jia *et al.*, 2017).

Effective sgRNA library design is critical for the success of lncRNA CRISPR screens. Both CRISPRi and CRISPR-del experiments require the targeting of Cas9 complexes within narrow genomic window of <1 kb around the TSS (Sanson *et al.*, 2018). CRISPR-del studies can delete the lncRNA's promoter and TSS, rather than the whole gene, to minimize off-target effects and maximize deletion efficiency (Aparicio-Prat *et al.*, 2015), although the single available CRISPR-del screen incorporated a mixture of TSS and whole-gene deletions (Zhu et al., 2016). In contrast to PCGs, lncRNA gene catalogues are growing rapidly, their accuracy is poor [particularly in correctly annotating the transcription start site (TSS)], and publicly available screening libraries are highly incomplete (Uszczynska-Ratajczak *et al.*, 2018). Thus, lncRNA screens are at risk of high false negative rates arising from poor annotations.

Although growing numbers of screens are being performed on lncRNAs and other non-coding elements (Diao *et al.*, 2017; Gasperini *et al.*, 2017; Zhu et al., 2016), key resources are lacking. First, no analysis pipeline is capable of handling both single and paired sgRNA experiments. Second, we lack an understanding of the rules by which lncRNAs should be judged as good candidates for inclusion in a screen (Liu *et al.*, 2017). In the present study, we address these issues through the creation of a new CRISPR screen analysis pipeline capable of perturbation-independent, end-to-end analysis. We deploy this pipeline on recently published screen data to better understand the behaviour of such screens and establish guidelines for design of future screens.

2 Materials and Methods

CRISPR Analysis for Single and Paired RNA-guides (CASPR) pipeline: The CASPR pipeline is programmed in Bash. It is based on five subprograms following the workflow presented in Fig. 1C, each of which can be run independently. CASPR is available at https://judithbergada.github.io/CASPR/, with documentation and instructions.

Indexing: The library of single or paired sgRNAs must be provided as a text file with three or four columns, respectively: IDs, target gene names and 20 nt protospacer sequences. CASPR transforms the library into a FASTA file. Importantly, for pgRNAs, this FASTA file concatenates second and first protospacer sequences, in that order (Fig. 1D). The library FASTA file is indexed using STAR (Dobin *et al.*, 2013).

Quality control of the reads: Read qualities are tested by FastQC and outputs are stored for inspection (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/).

Trimming: Sequencing reads (FASTQ files) are trimmed with cutadapt (EMBnet., 1994) based on adapters that are either specified by the users or set by default. Cutadapt identifies the 5′ position of the adapters and removes all nucleotides upstream and downstream of the protospacers (Fig. 1E). Reads are rejected if the adapter is not found or the sequence of the remaining protospacer is shorter than 5 bp. To handle pgRNA libraries, two protospacers must be sequenced using forward and reverse reads (Fig. 1B). Thus, their sequences can be extracted separately as described above. CASPR computes then the reverse complement of the second protospacer by employing the *fastx_reverse_complement* function available at FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), and concatenates the resulting paired sgRNAs to create a new FASTQ file (Fig. 1D). To account for situations in which adapters are not sequenced, CASPR checks if they are placed at the same coordinates in >25% of the reads. Otherwise, protospacers are assumed to start at the first 5′ base pair.

Mapping: Trimmed reads are mapped to the indexed protospacer library using STAR (Dobin *et al.*, 2013). CASPR allows users to tune mapping parameters, in terms of mismatches and minimum number of matched nucleotides. Reads that map to >1 library sequence are discarded to avoid confounding effects. In contrast to other software, CASPR affords flexibility during the mapping step, while providing a proper quantification for both sgRNA and pgRNA libraries.

Test of gene significance: CASPR uses SAMtools (Li *et al.*, 2009) to convert the BAM files containing mapped reads into a count table, which is taken as input to perform the assessment of gene significance. PBNPA (Jia *et al.*, 2017) and MAGeCK in the adjusted Robust Rank Aggregation mode (α-RRA) (Li *et al.*, 2014) are employed in parallel. The degree of agreement between the two methods may be inspected with Venn diagrams. The raw gene-level *P*-values of PBNPA and MAGeCK are combined into a consensus value by Fisher's method (Fisher, 1970). Finally, *P*-values are adjusted to a false discovery rate (FDR) by the Benjamini–Hochberg method. To aid visualization of results, CASPR generates multiple plots in R - quantile–quantile plots, volcano plots, box-plots and other scatter plots. It also creates configuration files that can be inspected using VISPR, the web-based interactive framework (Li *et al.*, 2015).

Software and versions: CASPR was tested with STAR 2.6.0c, FastQC v0.11.8, cutadapt 1.18, FASTX-Toolkit 0.0.14, SAMtools 1.4.1, MAGeCK 0.5.8, R 3.5.0, PBNPA 0.0.3 and VISPR 0.4.14.

2.1 Data and accession codes

All analyses were based on GENCODE 19 gene annotations, and any other genes were discarded. Analyses presented here are based on two published lncRNA CRISPR screens in human cells: the CRISPRi study by Liu *et al.* (2017) and the CRISPR-del study by Zhu et al. (2016). From these studies, we extracted the subset of gene targets that belong to GENCODE annotations, leaving 4325 lncRNAs and 666 lncRNAs for the CRISPRi and CRISPR-del screens, respectively. The names of genes targeted by each study were obtained from the original publication, then converted to GENCODE identifiers using BioMart-ENSEMBL (Smedley *et al.*, 2015). For consistency and to allow fair comparisons of the CRISPR-del and CRISPRi studies, human genome assembly hg19/GRCh37 was used in analyses requiring gene coordinates. RNA-sequencing expression data was obtained from ENCODE, under accession codes ENCSR000CPR (HeLa), ENCSR000CPT (MCF7) and ENCSR000BYO (U87). Cell lines and lncRNAs for which data was not available were omitted. Coordinates of FANTOM5 peaks were retrieved from FANTOM database [FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014], considering the 'robust' CAGE peaks in hg19 coordinates. Studies of TSS annotation were not performed with the CRISPR-del dataset, because it was designed to target multiple genomic regions (e.g. introns, exons, promoters and whole genes), and not only TSSs. The entire set of CASPR results across all cells, perturbations and algorithms is provided in Supplementary File S1.

2.2 Benchmark dataset

As a ground-truth dataset for lncRNAs regulating cancer cell growth, we used an updated version of the manually curated Cancer LncRNA Census (Carlevaro-Fita et al., 2019), and the MiTranscriptome set of tumour-dysregulated lncRNAs (Iyer *et al.*, 2015). Throughout the paper, we refer to these as functionally validated (FV) and differentially expressed (DE) datasets, respectively. FV genes are defined as those with experimental or genetic evidence for a causative role in tumourigenesis, and were collected through careful manual curation from the scientific literature up to 2018. This dataset will be published shortly (Adrienne Vancura *et al.*, manuscript in preparation). DE is based on gene expression analyses of human transcripts from diverse cancer types, and contains significantly differentially-expressed lncRNAs between tumour and normal samples in at least one cancer type. Although an absolute ground-truth is lacking for lncRNAs, FV and DE are currently the most reliable set of positive control cancer lncRNAs available. For the analyses, only the subset of GENCODE-annotated lncRNAs are considered. FV comprises a total of 387 lncRNAs (Fig. 2A), of which 206 and 79 overlap genes screened in CRISPRi and CRISPR-del, respectively. Equivalent numbers for DE are 496, 266 and 86 lncRNAs.

3 Results

3.1 CASPR: versatile CRISPR screen analysis

In order to study lncRNA CRISPR screens, we developed a pipeline called CASPR (Fig. 1C). CASPR has several desirable features. First,

it is compatible with the two principal types of CRISPR screen, namely single-end sequencing of single sgRNAs (such as in CRISPRi screens) (Fig. 1A), and paired-end sequencing of paired sgRNAs (CRISPR-del screens, Fig. 1C). Second, CASPR balances the sensitivity and specificity of predictions, by generating a consensus significance estimate from leading yet methodologically distinct methods MAGeCK and PBNPA (Jia *et al.*, 2017; Li *et al.*, 2014). Third, CASPR offers an end-to-end analysis, commencing with raw amplicon sequencing reads and delivering finished analyses and publication-ready plots. All analyses are accompanied by comprehensive quality-control analysis and statistics.

3.2 Improved functional maps of lncRNAs in human cells

We evaluated the performance of CASPR by reanalyzing the principal LOF lncRNA CRISPR screens published to date. The Liu study (Liu *et al.*, 2017) utilized CRISPRi with single sgRNAs to knock down 16,401 lncRNAs across seven cell lines. The Zhu study utilized CRISPR-del with paired sgRNAs to target different genomic regions (e.g. introns, exons, promoters and whole genes) of around 700 lncRNAs in two cell lines (Zhu et al., 2016). Both studies screened HeLa cells, where 281 lncRNAs were targeted in common.

We compared the performance of CASPR's two integrated hit-identification algorithms, MAGeCK and PBNPA. MAGeCK employs a negative binomial statistical model to identify significantly enriched or depleted targets, while PBNPA utilizes an empirical model. The two methods identify significantly overlapping but distinct sets of hits (Fig. 2B).

To fairly evaluate screen performance, we used a benchmark dataset of high-confidence cancer lncRNAs, by combining a manually curated set of lncRNAs with functionally validated roles in cancer (FV), with a set of lncRNAs DE in tumours (see Section 2). In previous screens, MAGeCK has tended to be run with very permissive FDR cutoffs, suggesting it has a stringent behaviour (Gasperini et al., 2017; Zhu et al., 2016). Indeed, when run at default settings on CRISPRi data, it showed high precision but low sensitivity (lighter panels in Fig. 2C). PBNPA exhibited similar performance, yet with high variability between cell lines (Fig. 2B). Worryingly, both tools identified almost no hits in the CRISPR-del data for either cell type (darker panels in Fig. 2C and D). Values of performance across cell lines are provided in Supplementary File S2.

We hypothesized that this relatively poor performance of individual methods might be mitigated by integrating the predictions of both. Thus, raw *P*-values were merged by the Fisher method to yield a consensus significance estimate (Fisher, 1970). For the CRISPRi screens, this resulted in improved sensitivity, with only a slight reduction in precision as compared to MAGeCK (Fig. 2C). More importantly, the consensus method identified hits for the CRISPR-del data with acceptable precision, where far fewer were observed by either of the two algorithms alone (Fig. 2C and D). Thanks to its improved sensitivity, the consensus approach yields an increased number of hits over the individual methods across all cell lines and perturbations (Fig. 2D). As expected, the consensus method produced lncRNA hits with significantly enriched and depleted sgRNAs (Fig. 2E).

We next compared CASPR to the state-of-the-art in CRISPR screen analysis, CB^2. CB^2 has been shown to outcompete many

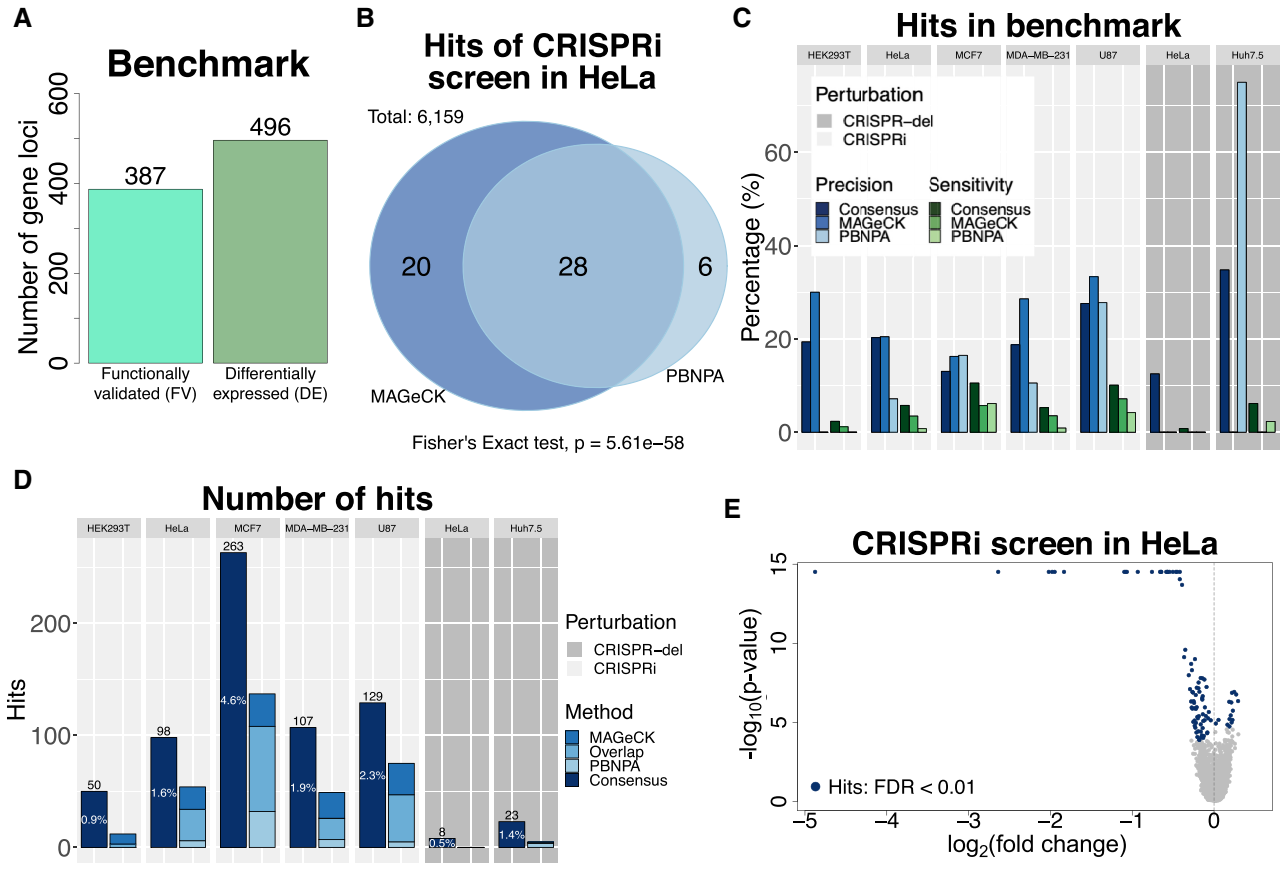

Fig. 2. (**A**) Two benchmark sets of cancer-associated genes were compiled from the literature (see Section 2). Numbers indicate only GENCODE-annotated lncRNA gene loci. (**B**) Overlap of hits identified by MAGeCK and PBNPA algorithms in HeLa CRISPRi screen. (**C**) Sensitivity and precision of MAGeCK, PBNPA and their Fisher consensus, as measured across cell types and perturbations. Accuracy is measured with respect to the union of the benchmark gene sets. The consensus measure shows the highest sensitivity, while maintaining relatively stable precision across conditions. (**D**) Numbers of hits identified by MAGeCK, PBNPA and consensus, separated by cell line and perturbation. (**E**) Volcano plot showing the non-hits (gray) and hits (blue) based on the consensus method in HeLa CRISPRi screen. Log-fold changes of each gene, obtained from the log-fold changes of all sgRNAs, are shown in the *x*-axis; statistical significance is shown on the *y*-axis

existing methods (Jeong *et al.*, 2019). Using CB^2, we reanalyzed the CRISPRi screen in MCF7 and U87, which were the cell lines with the lowest and highest performance of CASPR, respectively. In both screens, a greater number of hits were identified by CASPR compared to CB^2, and CB^2 hits were essentially a subset of CASPR hits (Supplementary Fig. S1A–C). Importantly, CB^2 provided a higher precision at the cost of reducing the sensitivity, and the areas under the curves (AUC) was always lower (Supplementary Fig. S1E and F). It is important to note that performance estimates based on ROC curve are likely to be underestimates, since our benchmarking data is incomplete and many true positives are likely to be interpreted here as false positives.

We also tested the performance of CASPR on CRISPR growth screens of PCGs. Taking known essential and non-essential genes as the benchmark, we evaluated performance of CASPR and CB^2 on two screens: a conventional CRISPR mutation and a CRISPRi screen in RT112 cells (Evers *et al.*, 2016). For CASPR at an FDR <0.01, precision is 100%/100% and sensitivity is 52%/26%, respectively, while optimal performance in each was achieved at FDR <0.75/0.13, respectively. The overall performance of CASPR is slightly below that of CB^2 (Supplementary Fig. S1G and H), indicating that CASPR can be used for analysis of PCG screens.

Thus, CASPR provides an improved sensitivity performance in the analysis of lncRNA CRISPR screens. For the rest of the paper, we use CASPR hits at an FDR cutoff of <0.01.

3.3 Comparing performance of CRISPRi and CRISPR-del

We next sought to compare the performance of the two LOF perturbations, CRISPRi and CRISPR-del. It should be noted that data comes from distinct publications, targeting different (but partially overlapping) sets of lncRNAs.

First, we compared the hits in common between the two methods in the shared cell line, HeLa. Surprisingly, when only considering lncRNAs targeted in both experiments, one observes zero overlap in the hits. These differences are underlined by the fact that no correlation is observed in screening results (Fig. 3A).

We used our benchmark dataset to evaluate the performance of CRISPRi and CRISPR-del across studied cell lines. For CRISPRi, the union of hits is significantly enriched in both functionally validated lncRNAs and DE lncRNAs (Fig. 3B). In contrast, the union of CRISPR-del hits is significantly enriched in neither (Fig. 3C). It should be noted that the lower number of CRISPR-del hits is likely to impact statistical power.

We further evaluated performance by calculating the precision in identifying cancer lncRNAs, as a function of increasing FDR (Fig. 3D and E). Note that precision should be compared to the overall frequency of cancer lncRNAs amongst the screen targets (shown as dotted lines). In the CRISPRi study, the expected behaviour is observed: high precision at low FDR, decreasing as FDR threshold is increased (Fig. 3D). This trend holds for all cell lines, although MCF7 cells show comparatively weak performance. For CRISPR-del, one observes moderate precision at low FDR for Huh7.5 cells, but this tapers off rapidly with increasing FDR cutoff (Fig. 3D). In contrast, hits in HeLa are *depleted* for cancer lncRNAs compared to background expectation (Fig. 3D). Receiver-operator curves (Supplementary Fig. S2A–F) support these observations, although the incompleteness of the benchmarking dataset likely leads to underestimates of the AUC.

Overall, these data show that CRISPRi screens are generally capable of identifying *bona fide* cancer lncRNAs, while the two available CRISPR-del screens display weaker and more variable performance.

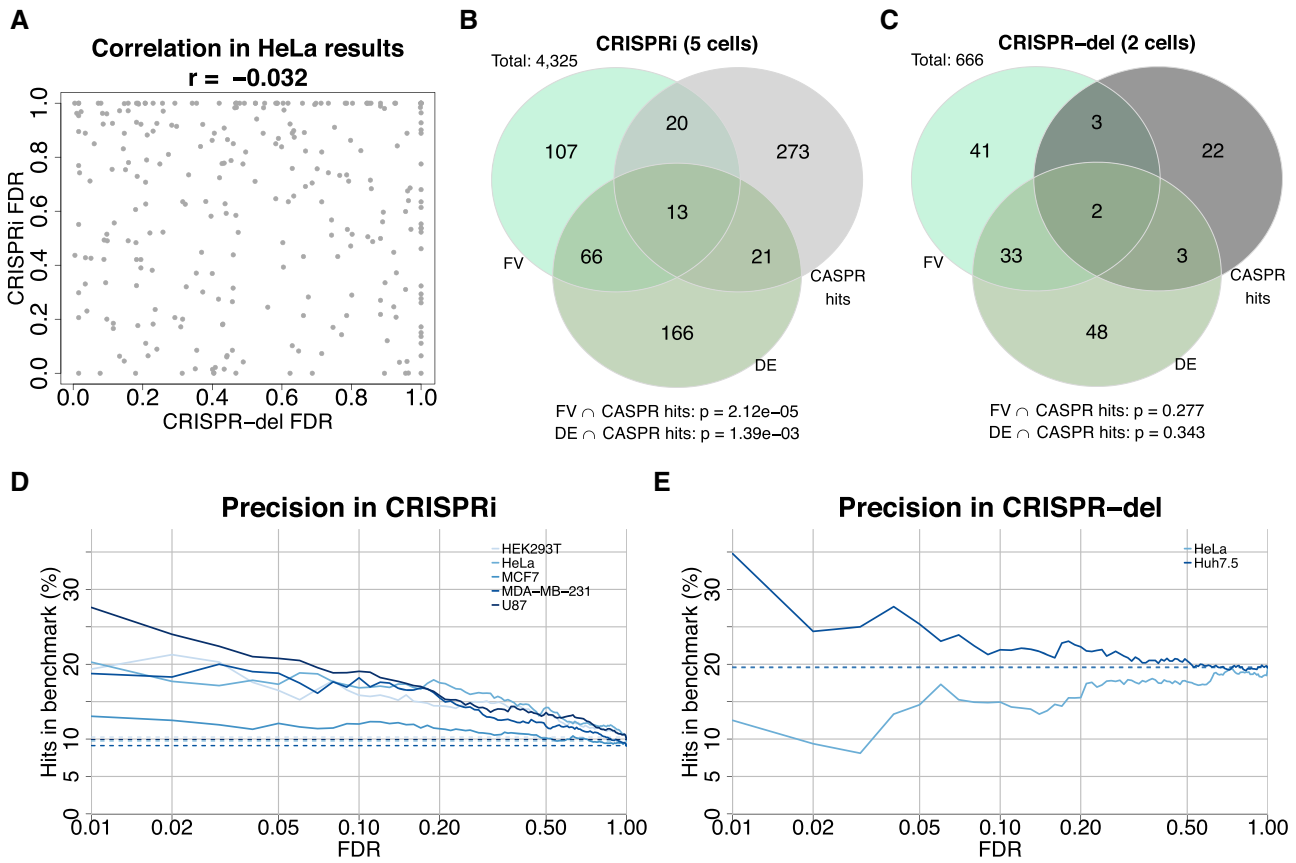

Fig. 3. (**A**) Pearson's correlation between CRISPRi and CRISPR-del FDR in HeLa. (**B** and **C**) Overlap of consensus hits with benchmark datasets. CRISPRi hits (light gray) show a significant overlap with experimentally validated (light green) and DE (dark green) lncRNA sets, while CRISPR-del hits (dark gray) show a weaker and non-significant overlap with both sets. (**D** and **E**) Precision of CASPR consensus method with increasing FDR threshold, for each cell line. Dashed lines indicate the expected background precision

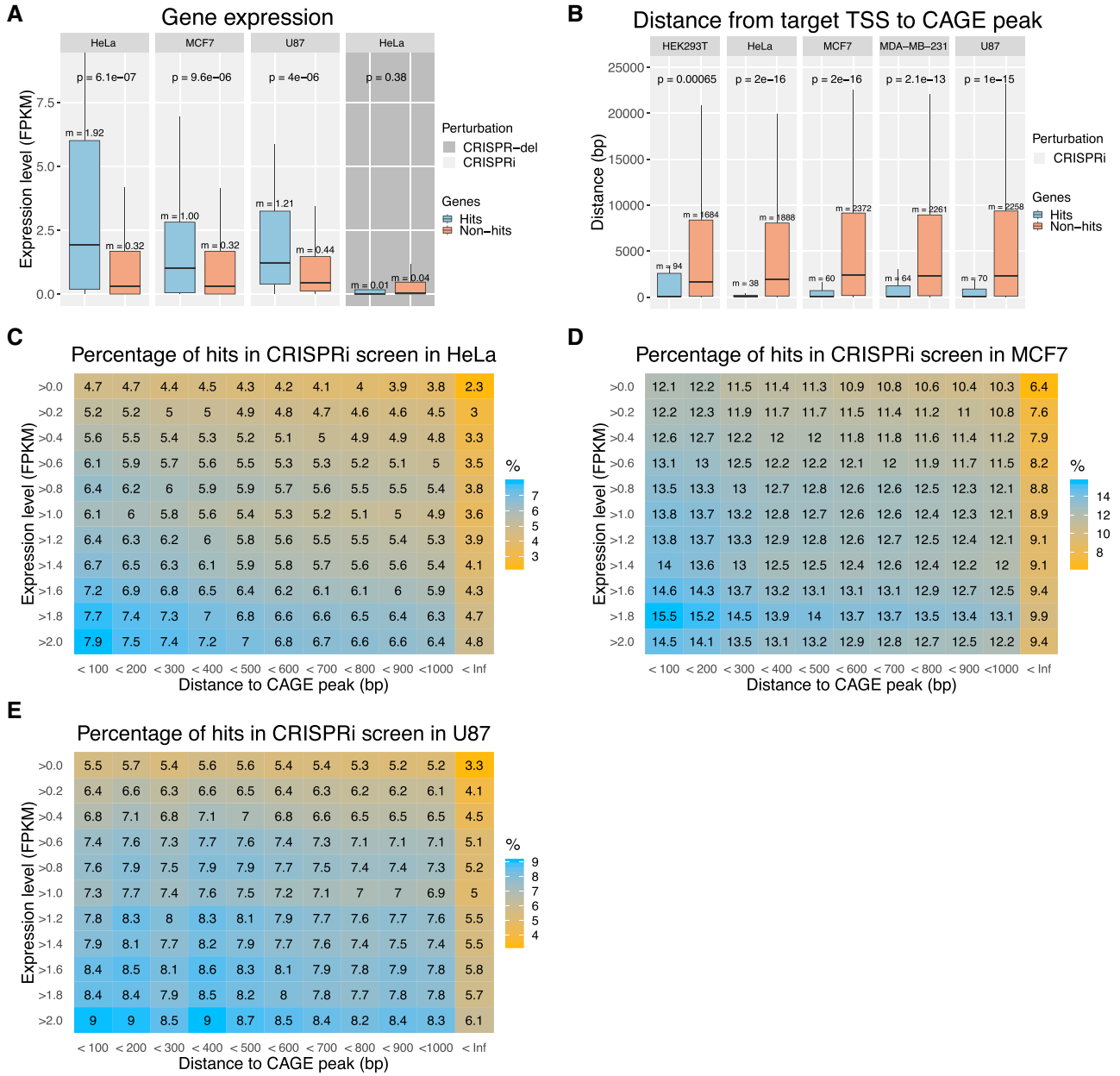

Fig. 4. (**A**) Gene expression levels between consensus screen hits (blue) and non-hits (orange). Expression is measured by RNA-sequencing from the ENCODE consortium. (**B**) The distance from targeted TSS to the true TSS, the latter inferred from CAGE peaks mapped by the FANTOM consortium. (**C–E**) Selection criteria for lncRNA inclusion future screens. Each cell holds the hit rate (percent of screened lncRNAs that are a hit, as defined by consensus method). Axes indicate increasing thresholds for gene expression (*y*-axis) and distance from targeted TSS to nearest CAGE peaks (*x*-axis). High hit rates (blue) are identified at high expression levels and low distances

3.4 Parameters correlating with success of CRISPR screens

Factors correlating with phenotypic hits in CRISPR screens could be used in future to optimize screen design. Liu *et al.* (2017) showed that target gene expression is the single factor with greatest predictive power for identifying hits. We tested this in the CRISPRi data for cells where ENCODE public RNA-sequencing data is also available. In agreement, we found that on average, lncRNA hits have between 3- and 6-fold greater expression than non-hits in CRISPRi data (lighter panels in Fig. 4A). A different behaviour was observed for CRISPR-del screen in HeLa, in which no significant difference is observed in the expression between hits and non-hits, again suggesting that this screen produced few true positives (darker panels in Fig. 4A).

For effective perturbation, sgRNAs must be recruited within a narrow genomic window around the target's TSS (Sanson *et al.*, 2018). The accuracy of present lncRNA transcript annotations is rather poor, and a large fraction of annotated transcript 5′ ends probably lie kilobases or more from the true TSS (Uszczynska-Ratajczak *et al.*, 2018). Thus, we hypothesized that lncRNA hits should have annotated start sites closer to true TSS, compared to non-hits.

To test this, we compared the start position of lncRNA annotations to a map of true TSS, as defined by Cap Analysis of Gene Expression (CAGE) [FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014]. We observed that the average hit's annotated TSS is within 100 bp of a CAGE peak, whereas average non-hits are >1 kb away (Fig. 4B). This is observed in all CRISPRi experiments. For CRISPR-del screens, these analyses were not feasible, due to the mixture of TSS and whole-gene targeting.

We considered the possibility of a confounding interaction between gene expression and CAGE peak presence. However, linear models trained with CAGE peak distance and gene expression

showed that interaction was not significant, while each factor individually contributed to the probability of being a screen hit (*P*-values are shown in Fig. 4A and B).

To facilitate future screen designs, we integrated the above insights into a scheme for target selection (Fig. 4C–E). For given thresholds of annotated TSS to CAGE distance (*x*-axis) and expression (*y*-axis), one can look up the hit rates in each cell line. In HeLa (Fig. 4C), hit rates range from a baseline of 2.3% with no filtering, to around 8% when only considering lncRNAs with TSS <100 bp from a CAGE peak and expression >2 FPKM. Similar trends were observed for other cell lines, considering only lncRNAs annotated in GENCODE and for which expression data was available. These values should be a useful guide in the selection of targets for future lncRNA screens.

4 Discussion

Here, we have presented CASPR, a pipeline for CRISPR screen analysis that is characterized by being an end-to-end solution that can equally handle single or paired sgRNA datasets, and balances sensitivity and specificity through consensus prediction from two leading algorithms. We anticipate that CASPR will be useful for the growing number of groups worldwide who are applying CRISPR functional screening to lncRNAs and other non-coding genomic elements including enhancers (Diao *et al.*, 2017; Gasperini et al., 2017).

In terms of performance, CASPR displays improved sensitivity compared to leading methods MAGeCK and PBNPA, while maintaining similar precision. Critically, CASPR maintains competitive performance in both single (CRISPRi) and paired (CRISPR-del) sgRNA experiments, compared to either algorithm alone. Thus, CASPR is suitable for both main CRISPR screen types. CASPR has been mainly designed and applied to study lncRNA datasets, but it also performs well on CRISPR screens of PCGs with either mutation or CRISPRi perturbations.

The growing interest in CRISPR screening highlights the need for library design guidelines. This is particularly challenging for lncRNAs, due to our ignorance of which ones are functional, and the incomplete state of their gene annotations (Kopp and Mendell, 2018; Uszczynska-Ratajczak *et al.*, 2018). Previous work suggested that steady-state RNA levels were a useful guide to predicting lncRNA hits in CRISPRi screens (Liu *et al.*, 2017). We have corroborated this, and identified a new criterion for target selection in the form of TSS annotations. Combining these two measures, we have produced guidelines for selection of lncRNAs for inclusion in CRISPR libraries. These guidelines should improve future projects by enabling researchers to create smaller libraries focussed on more promising lncRNAs.

We also evaluated the performance of the small number of available CRISPR screens. Overall, CRISPRi data from Liu *et al.* contained a large and statistically significant number of previously identified benchmark gene sets of cancer-promoting or cancer-related lncRNAs. Furthermore, CRISPRi hits tend to be higher expressed and have well-annotated TSS. Altogether, this highlights the quality of the Liu data and the power of CRISPRi in identifying functional lncRNAs. On the other hand, the CRISPR-del data, at least for HeLa cells that could be compared to CRISPRi and for which RNA-seq data were available, displayed no enrichment for known cancer lncRNAs, suggesting that this experiment yielded few genuine hits. It is possible that this low sensitivity arises, in part, due to the fact that in many cases, entire lncRNA genes were targeted, in contrast to their TSS alone. It is likely that these relatively large deletions are less efficient (Canver *et al.*, 2014). Other studies have demonstrated the efficacy of CRISPR-del as a perturbation strategy (Aparicio-Prat *et al.*, 2015; Ho *et al.*, 2015), although a recent study and our own unpublished work, suggests that promoter deletions may give rise to unexpected gene perturbations (Lavalou *et al.*, 2019). At any rate, the Huh7.5 CRISPR-del screen did appear to make true positive predictions. In summary, these findings show that both CRISPRi and CRISPR-del are effective perturbations for pooled screening approaches, although more CRISPR-del data will be necessary to properly compare the performance of these two methods.

Finally, these results support the existence of *bona fide* functional lncRNAs that regulate cell growth, a fundamental cellular phenotype. The fact that CRISPRi screen hits significantly overlap two independently generated sets of cancer lncRNAs, suggest that significant numbers of functional lncRNAs exist and may be found by CRISPR-based strategies.

In summary, CASPR will be a useful tool for researchers wishing to employ CRISPR screening to map the functional elements within the non-coding genome.

Acknowledgements

The authors are grateful to Rémy Bruggmann for providing the access to the IBU cluster (University of Bern Interfaculty Bioinformatics Unit) upon which bioinformatic analyses were performed, and Andrés Lanzós for his advice, support and expertise concerning the statistical analyses. We acknowledge administrative support from Deborah Re and Silvia Roesselet (DBMR).

Funding

Work in the Johnson laboratory is funded by the Swiss National Science Foundation through the National Center of Competence in Research (NCCR) 'RNA & Disease', by the Medical Faculty of the University and University Hospital of Bern and by the Helmut Horten Stiftung.

Conflict of Interest: none declared.

References

Aparicio-Prat,E. *et al.* (2015) DECKO: single-oligo, dual-CRISPR deletion of genomic elements including long non-coding RNAs. *BMC Genomics*, **16**, 846.

Canver,M.C. *et al.* (2014) Characterization of genomic deletion efficiency mediated by CRISPR/Cas9 in mammalian cells. *J. Biol. Chem.*, **289**, 21312–21324.

Carlevaro-Fita,J. *et al.* (2019) Cancer LncRNA Census: evidence for deeply-conserved roles of long noncoding RNAs in tumorigenesis. *Commun. Biol.*, in press.

Diao,Y. *et al.* (2017) A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nat. Methods*, **14**, 629–635.

Dobin,A. *et al.* (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, **29**, 15–21.

Doench,J.G. *et al.* (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.

EMBnet,M. (1994) *Embnet.news: European Molecular Biology Network Newsletter*. EMBnet Administration Office, Sweden.

Esposito,R. *et al.* (2019) Hacking the cancer genome: profiling therapeutically-actionable long noncoding RNAs using CRISPR-Cas9 screening. *Cancer Cell*, **15**, 545–557.

Evers,B. *et al.* (2016) CRISPR knockout screening outperforms shRNA and CRISPRi in identifying essential genes. *Nat. Biotechnol.*, **34**, 631–633.

FANTOM Consortium and the RIKEN PMI and CLST (DGT). *et al.* (2014) A promoter-level mammalian expression atlas. *Nature*, **507**, 462–470.

Fisher,R.A. (1970) *Statistical Methods for Research Workers*. Hafner Pub. Co., New York.

Gasperini,M. *et al.* (2017) CRISPR/Cas9-mediated scanning for regulatory elements required for HPRT1 expression via thousands of large, programmed genomic deletions. *Am. J. Hum. Genet.*, **101**, 192–205.

Hart,T. and Moffat,J. (2016) BAGEL: a computational framework for identifying essential genes from pooled library screens. *BMC Bioinform.*, **17**, 164.

Ho,T.-T. *et al.* (2015) Targeting non-coding RNAs with the CRISPR/Cas9 system in human cell lines. *Nucleic Acids Res.*, **43**, e17.

Iyer,M.K. *et al.* (2015) The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.*, **47**, 199–208.

Jeong,H.-H. *et al.* (2019) Beta-binomial modeling of CRISPR pooled screen data identifies target genes with greater sensitivity and fewer false negatives. *Genome Res.*, **29**, 999–1008.

Jia,G. *et al.* (2017) A permutation-based non-parametric analysis of CRISPR screen data. *BMC Genomics*, **18**, 545.

Kopp,F. and Mendell,J.T. (2018) Functional classification and experimental dissection of long noncoding RNAs. *Cell*, **172**, 393–407.

Lavalou,P. *et al.* (2019) Strategies for genetic inactivation of long noncoding RNAs in zebrafish. *RNA*, **25**, 897.

Li,H. *et al*. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li,W. *et al*. (2014) MAGeCK enables robust identification of essential genes from genome-scale CRISPR/Cas9 knockout screens. *Genome Biol*., **15**, 554.

Li,W. *et al*. (2015) Quality control, modeling, and visualization of CRISPR screens with MAGeCK-VISPR. *Genome Biol*., **16**, 281.

Liu,S.J. *et al*. (2017) CRISPRi-based genome-scale identification of functional long noncoding RNA loci in human cells. *Science*, **355**, aah7111.

Sanson,K.R. *et al*. (2018) Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. *Nat. Commun*., **9**, 5416.

Smedley,D. *et al*. (2015) The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Res*., **43**, W589–W598.

Spahn,P.N. *et al*. (2017) PinAPL-Py: a comprehensive web-application for the analysis of CRISPR/Cas9 screens. *Sci. Rep*., **7**, 15854.

Stojic,L. *et al*. (2018) Specificity of RNAi, LNA and CRISPRi as loss-of-function methods in transcriptional analysis. *Nucleic Acids Res*., **46**, 5950–5966.

Uszczynska-Ratajczak,B. *et al*. (2018) Towards a complete map of the human long non-coding RNA transcriptome. *Nat. Rev. Genet*., **19**, 535–548.

Zhu,S. *et al*. (2016) Genome-scale deletion screening of human long non-coding RNAs using a paired-guide RNA CRISPR–Cas9 library. *Nat. Biotechnol*., **34**, 1279–1286.