bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

scanMiR: a biochemically-based toolkit for versatile and efficient microRNA target prediction

Michael Soutschek[1,2], Fridolin Gross[1], Gerhard Schratt[1,2,*] and Pierre-Luc Germain[1,3,4,*]

[1] Lab of Systems Neuroscience, D-HEST Institute for Neuroscience, ETH Zürich, Switzerland
[2] Neuroscience Center Zurich, ETH Zurich and University of Zurich, Switzerland
[3] Lab of Statistical Bioinformatics, IMLS, University of Zürich, Switzerland
[4] Swiss Institute of Bioinformatics, Switzerland

* To whom correspondence should be addressed. Email: pierre-luc.germain@hest.ethz.ch.
Correspondence may also be addressed to gerhard.schratt@hest.ethz.ch

ABSTRACT

Despite the importance of microRNAs (miRNAs) in regulating a broad variety of biological processes, accurately predicting the transcripts they repress remains a challenge. Recent research suggests improved miRNA target prediction using a biochemical model combined with empirically-derived affinity predictions across 12mer sequences. Here, we translate this approach into a generally applicable, flexible and user-friendly tool (scanMiR). By compressing and handling miRNA 12mer affinity predictions into lightweight models, *scanMiR* can efficiently scan for both canonical and non-canonical binding sites on transcripts and custom sequences (including circRNAs and lncRNAs). Aggregation of binding sites into predicted transcript repression using a generalized biochemical model correlates better with experimental data than the most accurate alternative publicly available predictions. Moreover, a flexible 3'-supplementary alignment enables *scanMiR* to highlight and visualize unconventional modes of miRNA target mRNA interactions, such as bindings leading to target-directed miRNA degradation (TDMD) and slicing. By specifically scanning for these unconventional binding sites in brain-derived expression data, we provide a first systematic overview of potential TDMD and slicing sites on brain-specific lncRNAs as well as circRNAs. Finally, in addition to the main bioconductor package implementing these functions, we provide a user-friendly web application enabling the scanning of sequences, the visualization of predicted bindings, and the browsing of predicted target repression.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

INTRODUCTION

Since their initial discovery in *C.elegans*, microRNAs (miRNAs) have been shown to act on a plethora of biological pathways in diverse eukaryotic lineages (Bartel, 2018). They derive from short stem-loop precursor sequences, and after maturation associate with Argonaute (Ago) proteins to form the RNA-induced silencing complex (RISC) (Bartel, 2004). As part of this complex, miRNAs bind predominantly to the 3' untranslated region (UTR) of mRNAs to induce post-transcriptional silencing (Jonas & Izaurralde, 2015). In animals, post-transcriptional silencing primarily comprises a combination of translational inhibition and mRNA degradation, with only a few miRNA target transcripts undergoing endonucleolytic cleavage ("slicing") (Ameres & Zamore, 2013). Notable exceptions are target RNA-directed miRNA-degradation (TDMD) sites, which lead to a degradation of the bound miRNA (de la Mata et al., 2015; Kleaveland et al., 2018). Binding specificity is particularly determined by the 5' end of the miRNA, via the so-called seed region that encompasses miRNA nucleotides 2-7 (Bartel, 2009). More specifically, Ago exposes nucleotides (nt) 2-5 of its loaded miRNA to promote initial Watson-Crick pairing with target mRNAs, followed by further potential propagation of the interaction throughout nt 2-8 and the 3'part of the miRNA (Schirle et al., 2014; Sheu-Gruttadauria et al., 2019). The degree of repression that a miRNA imposes on its targets is determined by the time the RISC spends at individual sites, hence by the affinity of a miRNA for its mRNA targets (Denzler et al., 2016; McGeary et al., 2019). Strikingly, due to the influence of Ago on the functional properties of this interaction, the binding kinetics do not resemble conventional RNA:RNA hybridization (Salomon et al., 2015), making it difficult to predict binding efficiency through traditional computational methods.

To date, several bioinformatic approaches have been proposed to identify functional miRNA targets and to approximate their repression efficiency (e.g. Targetscan 7 (Agarwal et al., 2015), DIANA-microT-CDS (Reczko et al., 2012), MIRZA-G (Gumienny & Zavolan, 2015), and MirTarget (Liu & Wang, 2019); see Kern et al. (2020) for a comprehensive review). Most of these methods follow correlative approaches in that they try to infer a repression score from a number of features shown to correlate with repression efficiency of miRNAs, for example structural accessibility, predicted pairing stability, local AU-content, site-type, and 3'UTR length (Agarwal et al., 2015; Grimson et al., 2007; Gumienny & Zavolan, 2015; Krek et al., 2005; Nielsen et al., 2007). Some of the algorithms include the probability of sequence conservation since it has been shown that some of the parameters mentioned before are enriched in the proximity of evolutionary conserved sites (Grimson et al., 2007; Nielsen et al., 2007). One of the publicly-available prediction tools most accurate in explaining observed repression in cellular contexts (Targetscan 7) has been reported to perform as well as recent CLIP experiments (Agarwal et al., 2015). Nevertheless, Targetscan still explains only a minority (~14%) of the variance in logFC in cells perturbed with miRNA mimics (Agarwal et al., 2015).

Recently, several high-throughput studies have attempted to shed light on the biochemical determinants of miRNA-mRNA targeting (Salomon et al., 2015; Becker et al., 2019; McGeary et al., 2019, 2021). In one of these studies, the investigators used purified Ago-miRNA complexes in combination with RNA bind-n-seq (RBNS) to determine the affinities of six miRNAs to any possible

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

12-mer sequence (McGeary et al., 2019). They showed that the affinities and the corresponding dissociation constants (logKD-values) of different canonical site-types correlate very well with repression observed in cellular contexts. The authors went on to construct a biochemical model and showed that miRNA repression predictions based on this model for five of these miRNAs reach an explanatory power of 34%, outperforming a retrained version of Targetscan by more than 20%. In an attempt to extrapolate these findings to other miRNAs, the authors trained a convolutional neural network (CNN) using the empirically obtained data. They demonstrated that repression scores based on affinity values obtained via the CNN substantially outperform Targetscan by ~50% across 12 different miRNAs in another cell-type (McGeary et al., 2019).

While the model and experimental data provided by McGeary et al. (2019) represent major advances in target prediction, they are not available to a broader class of users. Widely used platforms such as TargetScan, instead, are of limited use because they include only a handful of species, cover only a subset of the transcriptome, and are very tedious to apply to custom sequences. Finally, despite the prevalence of R and Bioconductor in bioinformatics, there is a relative paucity of R-based miRNA target prediction tools. We therefore sought to address these needs by integrating and extending the biochemical model of McGeary et al. (2019) within the Bioconductor framework to offer a generally applicable, flexible, and user-friendly package and accompanying web app, *scanMiR* (Fig. 1). *scanMiR* efficiently scans for canonical and noncanonical miRNA binding sites on any custom sequence (incl. circular RNAs) and estimates dissociation constants. We further include a flexible 3'-supplementary alignment, allowing the systematic prediction of TDMD sites as well as sites leading to RNA slicing. Moreover, *scanMiR* entails several novel visualization functions in order to facilitate miRNA-target candidate selection. *scanMiR*-predicted mRNA repression correlates significantly better with measured mRNA fold-changes than Targetscan 7 scores. Finally, transcript-level predictions for human, mouse and rat transcriptomes, as well as the visualization and scanning functions of *scanMiR*, are available to a broader public through a user-friendly webtool.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

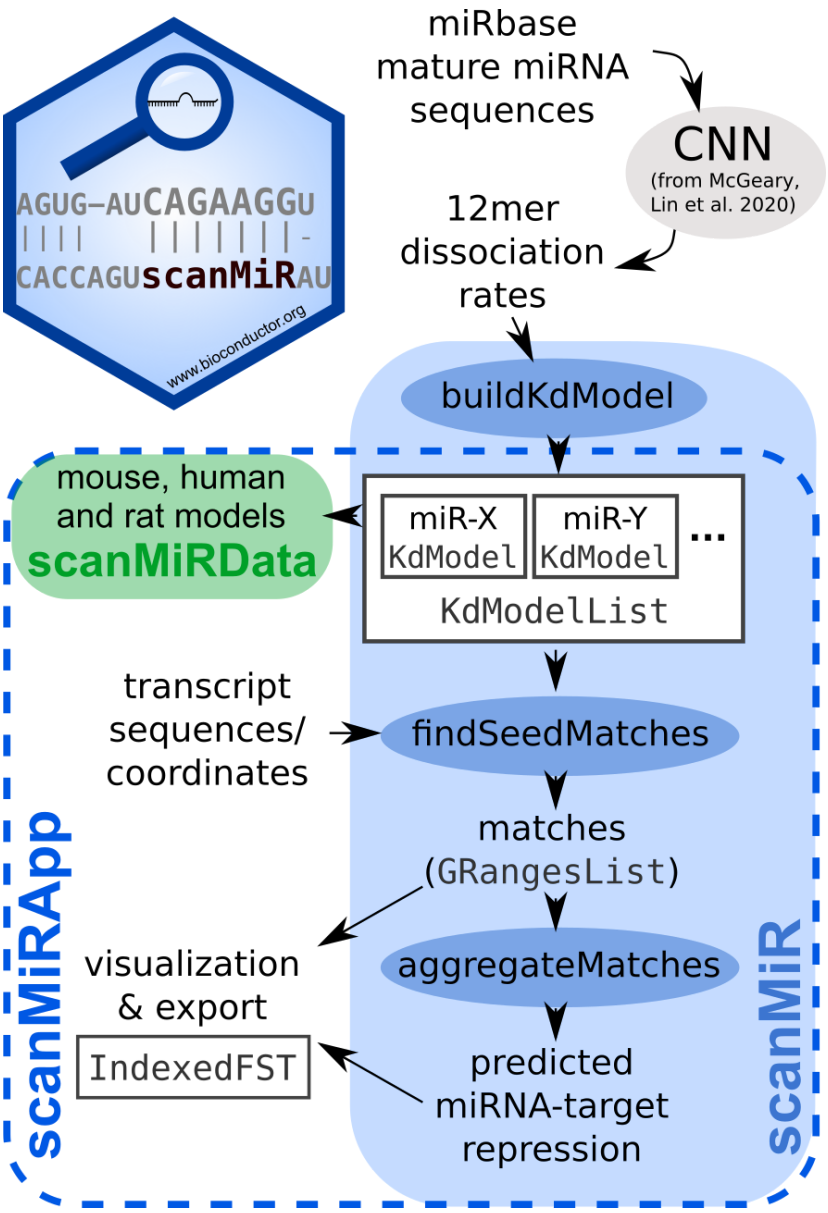


Figure 1: Overview of the *scanMiR* suite. The scanMiR package offers classes and functions to generate and work with miRNA 12-mer affinity models. It enables efficient and flexible site scanning, as well as different methods to visualize alignments. The *scanMiRData* package contains predicted affinity models for all human, mouse and rat miRbase miRNAs. Finally, the *scanMiRApp* package provides an easily deployable web interface to most *scanMiR* functionalities, as well as convenient wrappers to facilitate data handling and integration with annotation packages.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

MATERIAL AND METHODS

For detailed information on the individual methods and algorithms, please consult the Bioconductor packages as well as the additional Git repository containing all scripts used in this paper : https://github.com/ETHZ-INS/scanMiR_paper

Encoding of miRNA dissociation constants in KdModels

Dissociation constants as obtained from the CNN by McGeary et al. (2019) are stored as a deliberately overfitted linear model, with base KDs coefficients for each 1024 partially-matching 8-mers (containing at least four nucleotides (nt) complementary to miRNA positions 1-8, see McGeary et al. (2019)). Subsequently, 8-mer-specific coefficients that are multiplied with a flanking score generated by the flanking di-nucleotides are added to these base 8-mer coefficients. The flanking score is calculated based on the di-nucleotide effects experimentally measured by McGeary et al. (2019). In other words, affinity information for one miRNA can effectively be stored as 2048 integers plus the mature miRNA sequence.

Sequence scanning

Scanning is performed independently for different miRNAs. When including non-canonical sites, sequence scanning is handled by searching for and extending around four-nucleotide seeds. Coordinates are stored as GRanges (Lawrence et al., 2013) and sequences are extracted using Biostrings. Match type, predicted KD, and 3' local alignment are then estimated for each match. For matches within a given distance of each other, the lowest-affinity match is iteratively excluded until no overlap remains.

Since measured miRNA association kinetics of full lengths sequences for miR-21 and let-7a suggest (partial) tolerance for G:U wobble binding in the 3' supplementary region of miRNAs (position 12-17), *scanMiR* allows G:U pairing in the 3' local alignment (Becker et al., 2019). Values for the maximal loop size between the seed and 3' alignment are by default set to 4 (microRNA) and 9 (target) nucleotides based on recent observations of potential extended effective 3'-supplementary binding of microRNAs (Becker et al., 2019; Sheu-Gruttadauria et al., 2019).

To classify miRNA sites as potentially inducing target-directed miRNA degradation (TDMD) sites, we set out the following rules adapted from the empirical observations of de la Mata et al. (2015) and Sheu-Gruttadauria et al. (2019): i) 7mer or 8mer seed-pairing, ii) a 3'-supplementary binding score of at least 6 (to induce the required structural change in the Ago-miRNA complex releasing the 3'-end of the miRNA), iii) a miRNA bulge of at least 1 to avoid central pairing, iv) complementary local alignment of the 3'-end part of the miRNA allowing a maximum of 1 mismatch, v) a maximal target-miRNA bulge difference of 4 (denoted "TDMD?") or 2 ("TDMD") respectively, and vi) a maximal miRNA bulge of 5 ("TDMD?") or 4 ("TDMD") respectively.

Efficient miRNA-induced endonucleolytic cleavage ("slicing") requires central pairing of the miRNA to a target RNA at position 9-11 and is generally thought to be bolstered by complementary pairing of positions 2-15, with single mismatches at specific positions being possibly tolerated (Ameres et al.,

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

2007; Becker et al., 2019). Since a conclusive understanding of miRNA induced target "slicing" remains to be elucidated, we implemented a conservative two-step "slicing" classification: i) requiring complementary binding from positions 2-15 (G:U wobble bindings are allowed outside positions 9-11, though require additional Watson-Crick pairing for each G:U) (denoted "slicing?"), and ii) additionally requiring binding from position 16 towards the 3' end of the miRNA with a maximum of 1 mismatch as well as restricting wobble bindings to miRNA positions beyond the seed ("slicing").

The scans presented here were performed on the GRCm38 ensembl 102 transcriptome.

Aggregation into predicted transcript repression

The aggregation of multiple binding sites into predicted transcript repression is done mostly according to the model by McGeary et al. (2019). Briefly, multiple miRNA binding sites are assumed to have an additive effect on the transcript's decay rate that is proportional to their occupancy. The occupancy of AGO-bound miRNA *g* on a mRNA *m* with *p* binding sites is given by the following equation:

$$N_{m,g} = \sum_{i=1}^{p} \left(\frac{a}{a + c_{\mathrm{ORF}} K_{d,i}^{\mathrm{ORF}}} \right) + \sum_{j=1}^{q} \left(\frac{a}{a + K_{d,j}^{\mathrm{3'UTR}}} \right),$$

where *a* is the relative concentration of unbound AGO-miRNA complexes, and *c* is the penalty factor for sites that are found within the open reading frame (ORF) region. *scanMiR* also includes a coefficient *e* accounting for the effect of the 3' alignment:

$$N_{m,g} = \sum_{i=1}^{p} \left(\frac{a}{a + e_i c_{\mathrm{ORF}} K_{d,i}^{\mathrm{ORF}}} \right) + \sum_{j=1}^{q} \left(\frac{a}{a + e_j K_{d,j}^{\mathrm{3'UTR}}} \right),$$

where *e* is the exponential of the product of the 3' alignment score and a global parameter *p3*. The 3' alignment score roughly corresponds to the number of matched nucleotides. It is by default set to 0 if the number of matches is below 3 (adapted from the observations in Grimson et al. (2007)) and capped to a maximum of 8 in order to account for sites with higher scores probably leading to target-directed miRNA degradation (TDMD).

The repression by miRNA *g* can then be understood as the ratio between its occupancy and a background occupancy term, in which the dissociation constants are set to that of nonspecifically bound sites (i.e. KD=1). More specifically, McGeary et al. (2019) model the repression as:

$$\mathrm{repression} = \log 2(1 + bN_{m,g,\mathrm{background}}) - \log 2(1 + bN_{m,g}),$$

where *b* can be interpreted as the additional repression caused by a single bound AGO. Because UTR and ORF lengths have been reported to influence the efficacy of repression (Agarwal et al., 2015; Hausser et al., 2009), *scanMiR* allows for an additional term to take these effects into account:

$$\mathrm{repression}_{\mathrm{adj}} = \mathrm{repression} \cdot (1 + f \cdot \mathrm{UTR.length} + h \cdot \mathrm{ORF.length})$$

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

UTR.length and *ORF.length* are linearly normalized so that 0 and 1 are respectively the 5% and 95% quantiles of the distribution of lengths (see Garcia et al. (2011) & Agarwal et al. (2015)). This adjustment overall leads to slightly improved correlations with observed repression (Suppl. Fig. 6B-C). The effect being however small except on some extreme transcripts, these parameters are set to 0 by default.

While *b*, *c*, *p3*, *f* and *h* are considered global parameters (i.e. the same for different miRNAs and transcripts and also across experimental contexts), *a* is expected to be different for each miRNA in a given experimental condition. Values for parameters *b* (= 1.77) and *c* (= - 1.71) were therefore obtained from the *Biochemical Plus* model of McGeary et al. (2019), which was optimized with the experimentally determined RBNS data. Parameters *p3*, *f* and *h* were globally fitted to maximize the correlation with miRNA transfection experiments from McGeary et al. (2019).
As shown by McGeary et al. (2019), the performance of the biochemical model is robust to changes in *a* over several orders of magnitude (see also Suppl. Fig. 2B). Furthermore, repression predictions obtained with the globally optimized *a* value outperform Targetscan substantially (see Fig. 3). Therefore, *scanMiR* provides reasonable default parameters, although users can fit and provide their own parameter values.

Comparison of predicted and observed repression

For HeLa and Hek cell lines, we used the transcriptome reconstructions and quantifications provided by the authors in the GEO series (respectively in series GSE140217 and GSE140218). Given the absence of controls, both expression and predicted repression were normalized as described in McGeary et al. (2019).
Targetscan values were obtained by running the Targetscan Python scripts of Kathy Lin (https://github.com/kslin/targetscan) with default parameters and conservation files supplied in the Git repository. The 100way Multi-Species Alignment (MAF) file of the custom reconstructed HEK transcripts of McGeary et al. (2019) was obtained by downloading alignments from the UCSC Genome Browser with *mafFetch* and then further processing these with “Stitch MAF blocks” from *Galaxy* (Blankenberg et al., 2011) and *R*. Common species identifier were obtained from the NCBI Taxonomy Browser. miRNA seed and family information was downloaded from the Targetscan homepage (http://www.targetscan.org/vert_72/) and further processed in R to conform the example input files.
To assess the *scanMiR* performance in another species, we downloaded two datasets of miRNA knockout studies performed in mice (Amin et al., 2015; Eichhorn et al., 2014). Reads were mapped to the GRCm38 genome and subsequently counted with Salmon (Patro et al., 2017). logFC-values were obtained from edgeR (v. 3.32) (Robinson et al., 2010), by filtering for expressed transcripts using the *filtreByExpr*-function and normalizing with a weighted trimmed mean of the logarithmic expression ratios of individual samples (TMM). For the miR-122 knockout dataset from Eichhorn et al. (2014), a common negative binomial dispersion was estimated since the authors performed RNA-sequencing with only one replicate. In order to correlate logFC of these two datasets with *scanMiR* repression predictions, we considered only transcripts that constitute 90% of the expressed transcripts of one

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

gene in the specific setting, that are expressed higher than 10 TPM (Amin et al., 2015) or 0.5 TPM (Eichhorn et al., 2014), respectively, that are reported as representative transcripts in Targetscan, and that are supported with at least five 3p-seq-tags (http://www.targetscan.org/mmu_72/mmu_72_data_download/Gene_info.txt.zip).

Other external datasets used

For the circular RNA scans, we used full circular RNA reconstructions from Zhang et al. (2021). A gtf file of the coordinates was kindly provided by the authors, containing also back-splice junction (BSJ) counts. We extracted the corresponding spliced sequences, appending the first 11 nucleotides at the end to enable the identification of sites spanning the back-splice junction. For the quantification of long non-coding RNAs, we used the average of the GEO samples GSM2094148-GSM2094149, quantified over the ensembl 102 transcriptome using salmon 1.3.0 (Patro et al., 2017).

Web app and indexed random access to large scans

The web interface is a R shiny application which can be deployed through the *scanMiRApp* package. To provide more rapid results, it can optionally use pre-compiled scan and aggregation data. Since the binding site data for all miRNAs and transcripts can amount to a very large size, we implemented the *IndexedFst* S4 class which adds indices to an *fst* file (Mark Klik, 2020) to provide very fast out-of-memory access to specific transcripts/miRNAs. The *IndexedFst* class is not limited to storing binding sites and can be used for any data frame or *GenomicRanges* object.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

RESULTS

Efficient encoding of miRNA dissociation constants

McGeary, Lin and colleagues (2019) developed a convolutional neural network (CNN) predicting the dissociation constants (KD) of given miRNAs on partially-matching 12-nucleotides sequences. Given that there are about half a million possible 12-mers and corresponding KDs for each miRNA, it is impractical to store and use the information across all possible miRNAs. We therefore developed an efficient compression and encoding of the KD predictions (see methods), enabling its rapid application for scanning. As Figure 2A-B shows, the correlation of the reconstituted KD-values to the original ones is high, and above the accuracy of the CNN predictions (McGeary et al., 2019). With this approach, we could for instance store the affinity models for ~2000 miRNAs in approximately 10 MB (instead of over 8 GB for the storage of 8mer-Kd pairs) and could apply them very efficiently for the identification and characterization of putative sites (Figure 3D).

The *KdModel* and *KdModelList* classes defined in *scanMiR* provide an interface for handling these objects. We generated *KdModel* objects for all human, mouse and rat putative miRNAs from miRbase (Griffiths-Jones et al., 2006), which are made available in the accompanying *scanMiRData* package.

In an attempt to draw further biological conclusions from seed properties stored in KdModels, we set out to separate human miRNAs based on patterns of binding specificity (Figure 2C). Specifically, we examined the relationship between the dissociation constants of the top non-canonical seed and the median of canonical 6mers (generally accounting for 7mar-A1 and 6mer sites). miRNAs did not cluster into distinct classes but formed a rough multivariate normal distribution. Nevertheless, we could recognize different paradigmatic examples, such as the highly-specific binding of miR-499a (Figure 2D) and the known strong non-canonical binding of miR-124 (Chi et al., 2012; Wang et al., 2021) (Figure 2E). Based on its proximity to miR-124 in the distribution, we likewise propose the occurrence of previously uncharacterized high-affinity non-canonical binding sites for miR-7.

An important experimental aspect of miRNA research is the mutation of predicted binding sites to investigate their putative function. Although it is known that nucleotide substitutions opposite miRNA positions 2-5 within the seed region most efficiently affect binding (Becker et al., 2019), the importance of single nucleotides within the seed region of different miRNAs has not yet been elucidated. By generating positional weight matrices of miRNAs, based on the KD-predictions of 12mer sequences, *scanMiR* allows the user to directly plot the nucleotide information content of individual miRNAs. Such affinity based nucleotide information content plots suggest for example a uniform importance of nt 2-5 in the seed of miR-499a, whereas for miR-129-5p, nt 3-4 seem to be most relevant for a stable interaction with its targets (Suppl. Fig. 1A). Researchers can easily resort to this information to get a first impression on which nucleotides in the seeds of miRNAs are most required for a stable interaction with their targets, and hence should be primarily mutated to prevent miRNA binding.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

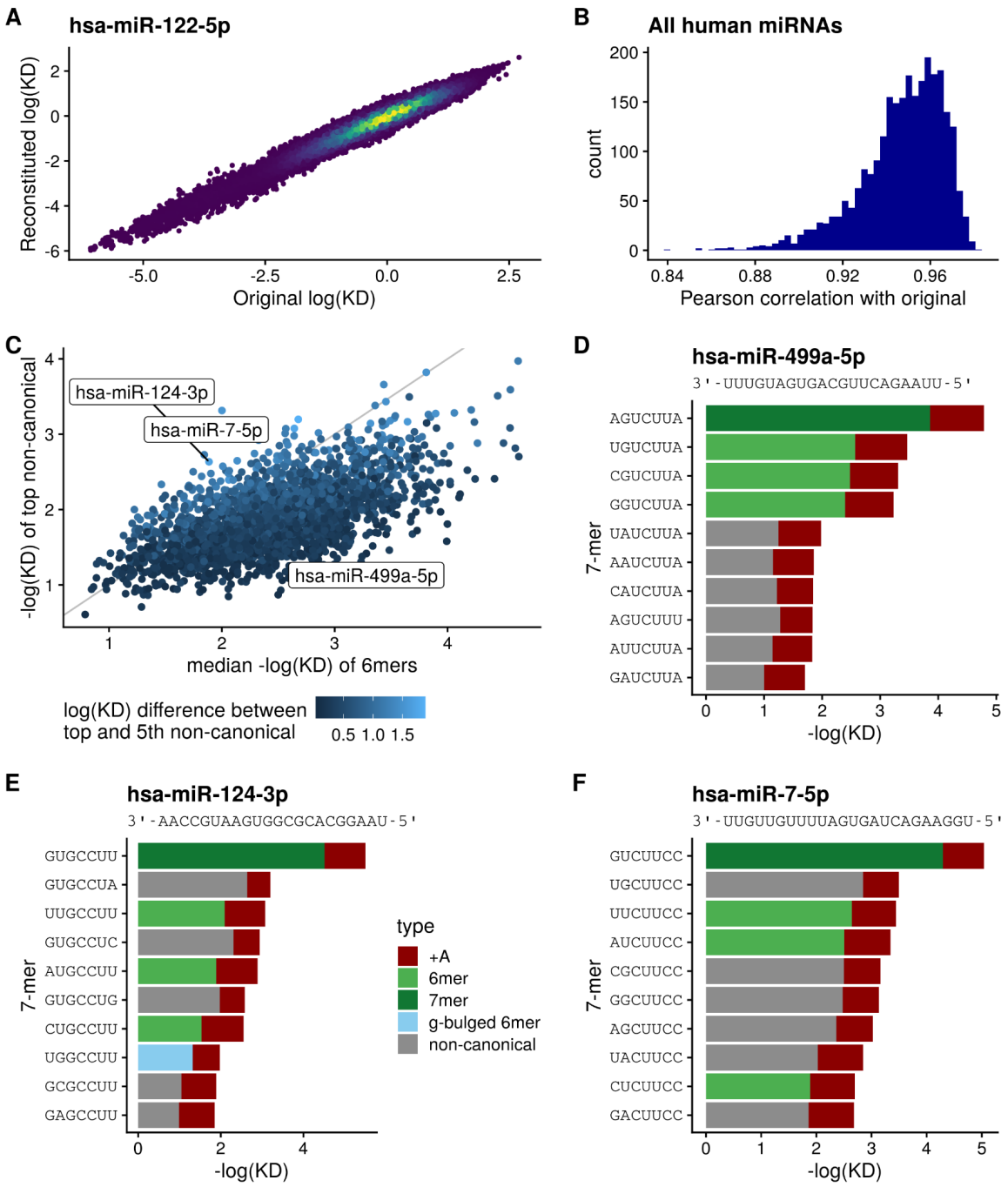


Figure 2: Efficient encoding of miRNA dissociation rates. A: Correlation between log(KD) values before and after compression for an example miRNA. **B:** Distribution of correlations across all human mirBase putative miRNAs. **C:** Relationship between the dissociation constants of the 6-mers and that of the top non-canonical sites, highlighting miRNAs with radically different binding behaviors. The log(KD) difference of the top and 5th highest ranking non-canonical seed illustrates the uniformity of high ranking non-canonical seeds. **D-F:** Example affinity plots showing the top ten miRNA seeds ranked by the average dissociation constant. MiR-499a-5p displays a very specific canonical binding (**D**), whereas miR-124-3p and miR-7-5p include highly ranked non-canonical seeds (**E-F**).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Biochemically-based prediction of miRNA target repression

To assess the accuracy of compressed KdModels for miRNA repression prediction, we scanned HeLa mRNAs for putative binding sites of the five miRNAs that McGeary and colleagues (2019) had initially used for their RBNS experiments and determined site affinity based on the compressed models (see Methods). Aggregating these values using the biochemical model and correlating them to mRNA fold-changes observed in Hela cells following miRNA-mimic transfections revealed similar results as with uncompressed dissociation constants (Fig. 3A-B, Suppl. Fig.2C). This holds likewise true for 12 further transfection experiments that were conducted in another cell type (HEK293) with miRNAs that were not used in the direct RBNS-measurements (Fig. 3C). Importantly, for most of these miRNAs, the correlation values are considerably better than those obtained with Targetscan 7 repression scores (Fig. 3B-C).

One of the constraints of the biochemical model is the necessity to fit the relative concentration of unbound AGO-miRNA complexes (*a*-value) for each miRNA and experiment individually (see Methods, (McGeary et al., 2019)). Bolstered by the observation that even a 100-fold deviation of the optimal *a*-value still leads to a repression prediction outperforming Targetscan (Suppl. Fig. 2B) (McGeary et al., 2019), we went on to fit a globally optimized *a*-value with the 12 HEK-cell transfection datasets. Repression predictions using this global value are slightly lower than those with *a* optimized separately for each dataset, but still outperform Targetscan substantially (Fig. 3C). To further assess the general validity of our approach, we tested predicted miRNA repression on published mouse miRNA knockout datasets. *scanMiR* similarly outperforms Targetscan in this experimental setup, although the overall signal-to-noise ratio and resulting explanatory power was lower (Fig. Suppl. 3). In conclusion, *scanMiR* repression predictions exceed the performance of the latest Targetscan 7 scores across several experimental settings, and the repression predictions are roughly generalizable even without fitting the biochemical parameters to individual experimental settings.

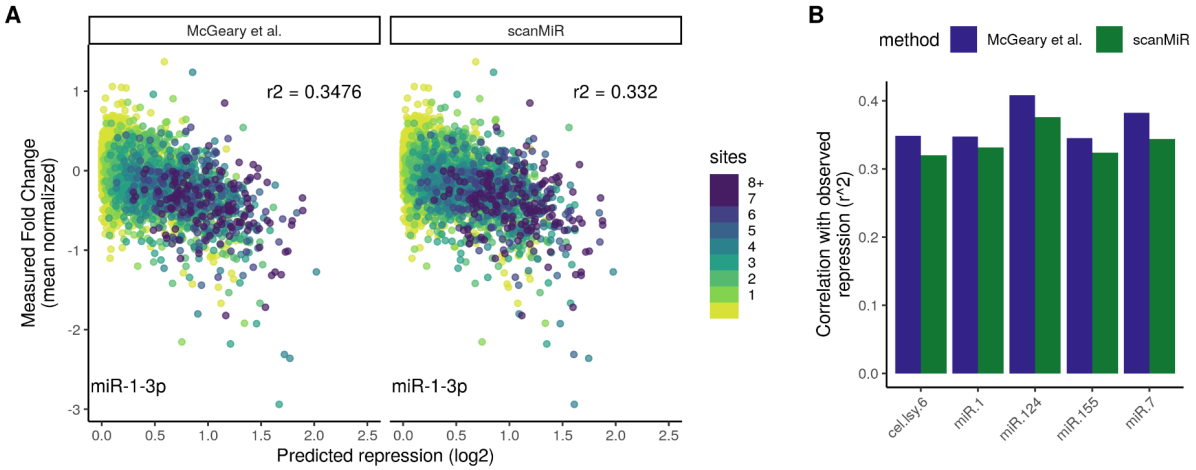
A further key finding of McGeary et al. (2019) was the identification of high-affinity non-canonical miRNA binding sites. By harnessing the whole set of available 12mer sequences and corresponding KD-values to identify and rank miRNA binding sites, *scanMiR* readily detects these presumably efficient non-canonical sites. However, including also the vast number of non-canonical sites with low predicted target affinities significantly exacerbates the scanning and storage of miRNA binding sites. We therefore studied the relation between KD-values of different binding sites and their explanatory power. For 11 out of the 12 miRNA-transfection experiments in HEK-cells, a KD-cutoff value of -0.3 lowered the explanatory power by less than 1%, while at the same time allowing a reduction in the number of stored sites by almost 40% (Suppl. Fig. 4C).

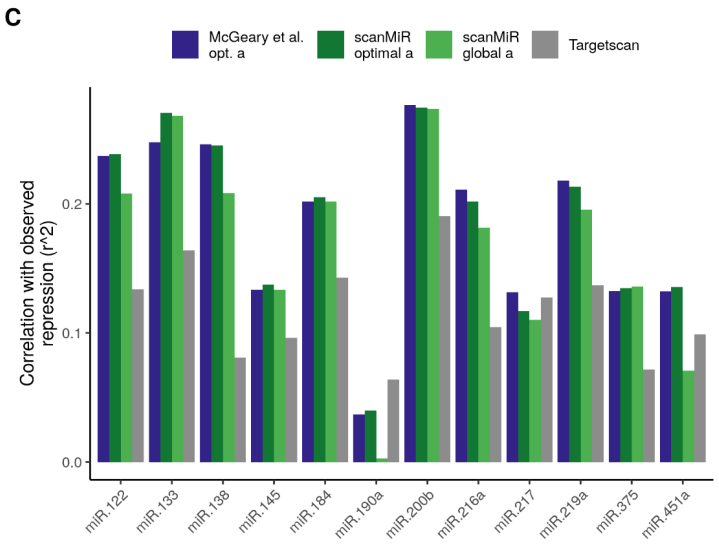
As an additional *scanMiR* feature, we implemented the characterization of g-bulged as well as wobbled (G:U) miRNA binding sites and flag them in the scanning results as well as in affinity plots (Fig. 2E, Fig. 3E, Suppl. Fig 1B). G-bulged non-canonical binding sites comprise a bulged out “G” opposing Nt6 of the miRNA to allow further seed pairing, and the binding of for example miR-124 to these sites has been shown by high-throughput techniques and the functional repression of candidate targets validated with luciferase experiments (Chi et al., 2012; Wang et al., 2021).

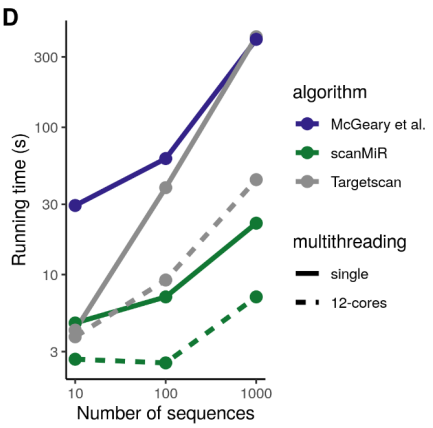
bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Moreover, *scanMiR* enables the user to specify a minimum distance between binding sites. It has been shown previously that binding sites located closer than 7nt to each other exert a repelling function onto each other (Grimson et al., 2007; Sætrom et al., 2007). By default, *scanMiR* therefore removes binding sites within the specified minimal distance of a higher-affinity site (Suppl. Fig. 5, see Methods).

Of note, despite evaluating a considerably larger number of sites (i.e. including all non-canonical sites) than Targetscan, *scanMiR* nevertheless ran significantly faster. This holds likewise true in comparison to the scripts from McGeary et al. (2019) (Figure 3D), which ran moreover substantially slower once including the default transcript folding (Suppl. Fig. 4D).







	10 Seqs.	*100 Seqs.*	*1000 Seqs.*
canonical	198	2219	23971
all sites	4002	43591	471103

E

miR-124-3p g-bulged 7mer site at nt 723 of Mink1

```
miRNA              3'-AACCGUAAGUGGCGCAC-GGAAU-5'
                                        ||| ||||
target 5'-...CCCCCCACCCCUCCUCAAUGUAGUGGCCUUGGA...-3'
```

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Figure 3: Efficient prediction of repression. A: Correlation between predicted and observed repression in an example miRNA (hsa-miR-1-3p) transfection experiment from McGeary et al. (2019), showing comparable correlations between the original algorithm and log(KD) values and our compressed version. Each point corresponds to a transcript, and is colored according to the number of canonical sites in its UTR. **B:** Correlation (Pearson *r*^2) between predicted and observed repression across the 5 miRNAs from McGeary et al. (2019) with experimental affinity measurements. **C:** Correlation (Pearson *r*^2) between predicted and observed repression across 12 further miRNAs in HEK cells (McGeary et al., 2019). Correlations obtained with a globally optimized a-value are shown in light green, correlations based on Targetscan 7 predictions in grey. **D:** Running times of the different methods scanning for binding sites of 12 miRNAs across growing sets of sequences (the sites indicated are from scanMiR before filtering, including all canonical and non-canonical sites). **E:** scanMiR flags g-bulged miRNA binding sites in scan results and correctly represents target alignments of those sites (here shown for the experimentally validated miR-124-3p binding site on Mink1).

Flexible 3' supplementary alignment flags special miRNA binding sites

Upon finding a partial seed match, *scanMiR* performs a local alignment of the 3' region of the miRNA onto the upstream region of the transcript, thereby enabling a flexible and customizable miRNA-mRNA bulge size (Suppl. Fig. 6A). Including a 3'-supplementary pairing score based on this alignment slightly improves *scanMiR* repression predictions (Suppl. Fig. 6B-C) and enables the prediction of special sites determined by this supplementary pairing. It has been shown previously that some miRNA binding sites with near full-length complementarity can lead to endonucleolytic cleavage of the target sequence (Carthew & Sontheimer, 2009). Experimentally validated examples comprise the slicing of Hoxb8 upon binding of miR-196a-5p (Yekta et al., 2004) or the slicing of the Cdr1as by miR-671-5p (Hansen et al., 2011; Kleaveland et al., 2018). *scanMiR* correctly identifies and represents these sites on mRNAs as well as circular RNAs (Figure 4A-D). Moreover, sufficient target complementary at the 3' end of the miRNA in combination with specific mRNA-miRNA bulges can lead to target-directed miRNA degradation (TDMD) (de la Mata et al., 2015; Sheu-Gruttadauria et al., 2019). By taking these determinants into account, *scanMiR* flags high and low-confidence TDMD sites, as shown for the validated miR-7 TDMD site on the lncRNA Cyrano (Figure 4C-D)

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

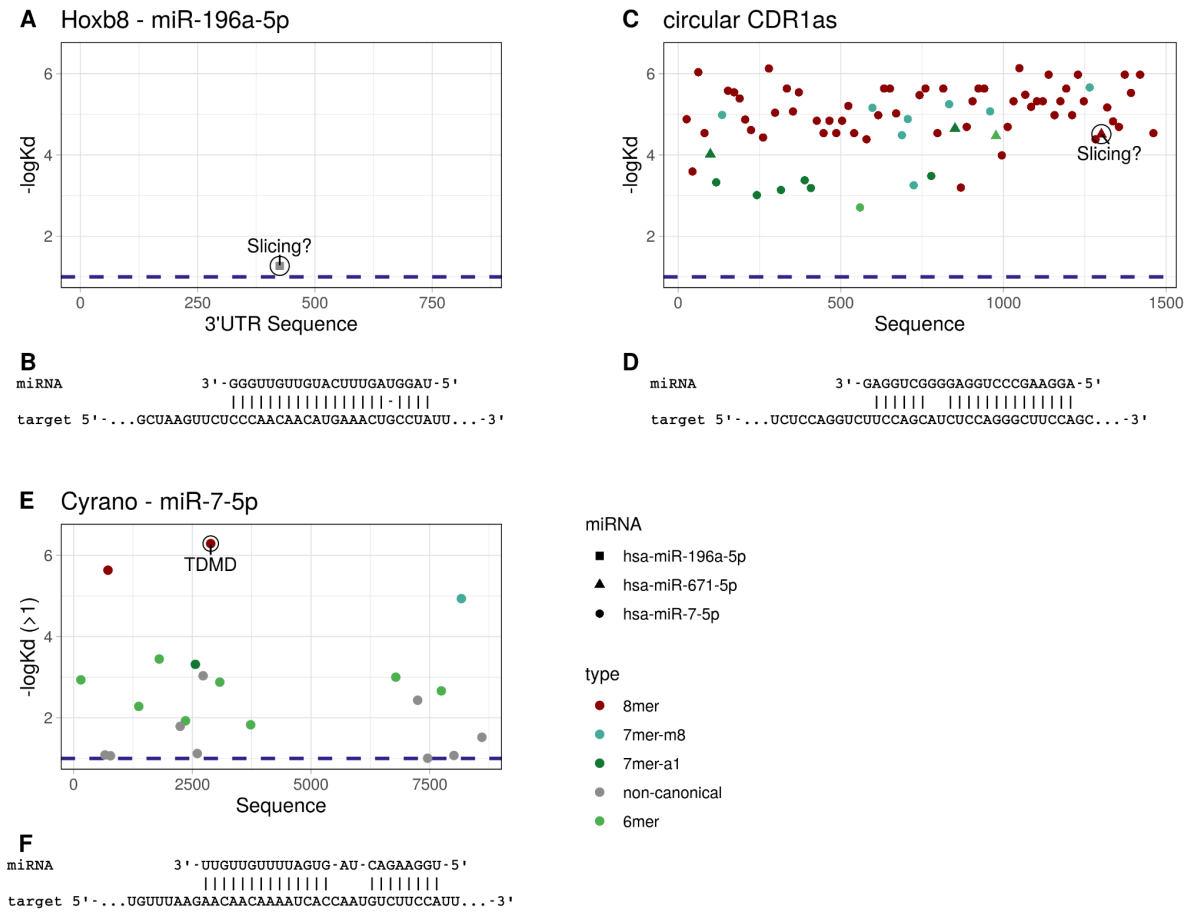


Figure 4: 3' supplementary alignment enables the prediction of special miRNA sites. A: Binding of miR-196a-5p to Hoxb8 leads to endonucleolytic cleavage of the target mRNA. scanMiR flags this site as "Slicing?", despite the low affinity of the seed of miR-196a-5p to Hoxb8, which can be explained by the G:U wobble binding at position five. **B:** miRNA-target alignment shows the almost perfect complement of miR-196a-5p to the Hoxb8 mRNA at this site. **C:** miR-7 and miR-671 binding sites on the Cdr1as circular RNA, revealing especially a very large concentration of miR-7 sites. **B:** The full alignment of the validated miR-671 slicing site on Cdr1as. **E-F:** Validated miR-7 TDMD site on the Cyrano RNA, and its alignment up to the last 3' nucleotides of the miRNA.

Based on these results, we decided to analyse the occurrence of TDMD sites more systematically. It has been previously suggested that potential TDMD sites are most effective in degrading bound miRNAs when their harboring RNAs are not susceptible to miRNA-mediated decay (de la Mata et al., 2015). Since it has been recently demonstrated that translation is essential for miRNA-mediated decay (Biasini et al., 2021), we specifically screened annotated long non-coding RNAs as well as the sequences obtained from a recent full-length profiling of circRNA isoforms expressed in the mouse brain for the presence of TDMD sites (Zhang et al., 2021). In addition to the experimentally validated miR-7 TDMD site on Cyrano, we predict a number of previously uninvestigated TDMD sites within lncRNAs (Fig. 5A) and circRNAs (Fig. 5B), raising the possibility that a subset of lncRNAs and

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

circRNAs are involved in the regulation of miRNA abundance. An additional screen for slicing sites on circRNAs retrieved much fewer candidates (Supplementary Table 3), among them the known Cdr1as, suggesting that miRNA-dependent slicing plays a minor role in the context of circRNA biology.

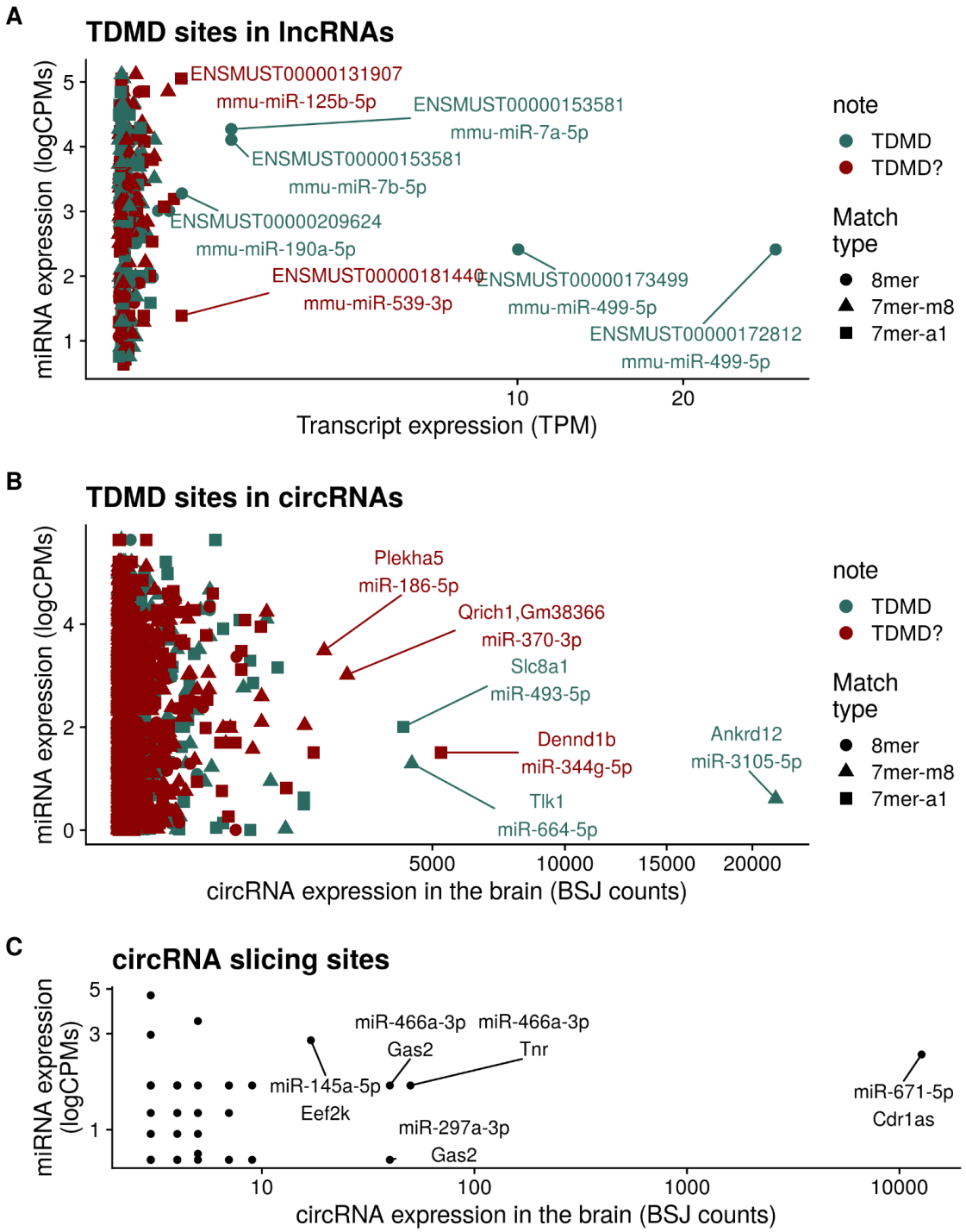


Figure 5: Prediction of non-conventional miRNA sites in non-coding RNAs. A-B: Prediction of TDMD sites in lincRNAs (**A**) and circRNAs (**B**) expressed in the brain. **C:** Prediction of circRNA slicing sites. The by far strongest signal comes from the Cdr1as circRNA.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Prediction of circRNA sponges/transporters in the brain

Due to their stability, circular RNAs (circRNAs) with multiple miRNA binding sites can act as sponges or transporters (Kristensen et al., 2019). The prevalence of such mechanisms however remains unknown. To predict circRNAs that might exert these roles in the mammalian brain, we identified circRNAs that are significantly enriched for canonical binding sites of specific miRNAs using a binomial model conditioned on the total abundance of binding sites for the given miRNA as well as the abundance of background miRNA bindings on the given RNA. Reasoning once more that lowly-expressed circRNAs are unlikely to have significant biological impact, we plotted this enrichment against the abundance of the circRNAs (Figure 6). MiR-7a-5p binding sites on the (by far) most abundant circRNA Cdr1as were most significantly enriched, in line with the previously validated exceptional effectiveness of Cdr1as as a miRNA sponge (Kristensen et al., 2019). Nevertheless, our data suggests that there is quite a number of additional circRNAs containing significantly more miRNA binding sites than expected by chance (see also Supplementary Table 4).

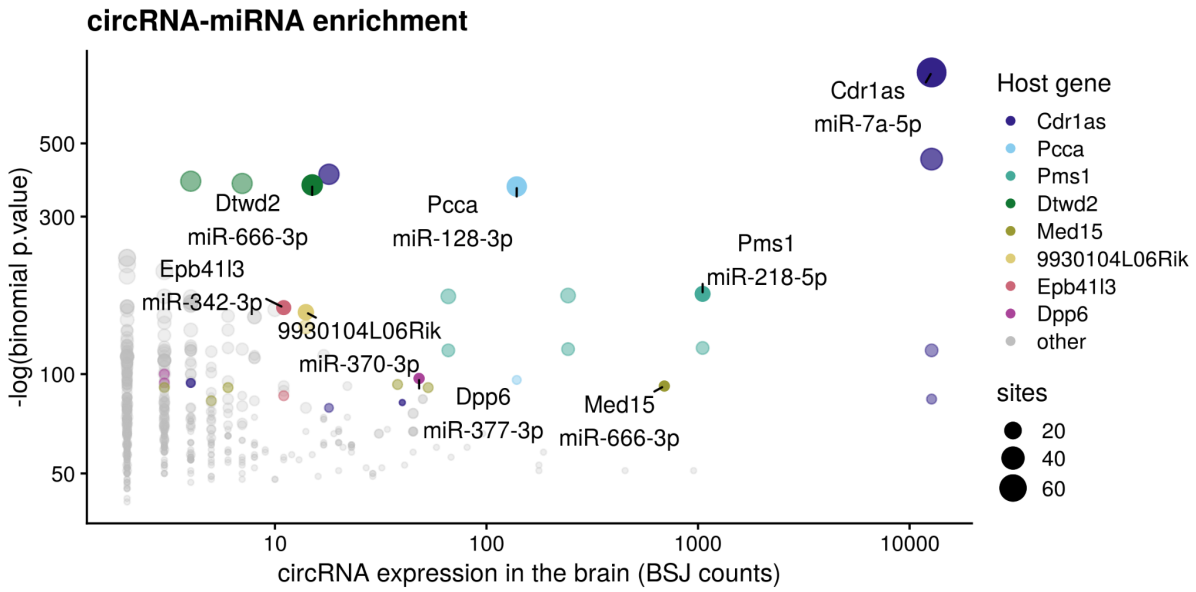


Figure 6: Circular RNAs potentially acting as miRNA sponges. Abundant circRNAs with enrichment for specific miRNA binding sites. Each dot represents a miRNA and circular RNA isoform pair, with only the most significant pair for interesting candidate genes being labeled. The y axis indicates the significance of the circRNA-miRNA enrichment. Only miRNAs with at least 100 reads and circRNAs overlapping genes were considered.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

A web interface to *scanMiR*

To make *scanMiR* accessible to a broader public, we developed a web interface to its main functionalities, available at https://ethz-ins.org/scanMiR/ (or http://130.60.24.171:3838/scanMiR). The interface allows the visualization of mouse, human and rat *KdModels*, the interrogation of top targets for all miRbase RNAs, the easy and flexible scanning of any transcript sequence for miRNA binding sites, and the visualization of binding sites including their supplementary 3' alignment (Suppl. 7). The application can additionally be easily deployed by anybody through the *scanMiRApp* package, enabling for instance groups working on other species to compile their annotation and set up their own portals. *scanMiRApp* additionally includes a number of wrappers (notably connecting with AnnotationHub to easily include updated transcript annotations), as well as features to enable rapid out-of-memory random access to large scans (see methods).

DISCUSSION

The *scanMiR* suite is a fast and flexible way to scan for miRNA binding sites (including TDMD and slicing sites), predict transcript repression, and visualize alignments. In addition, we have attempted to make the package user-friendly: *scanMiR* for instance accepts a variety of inputs (e.g. full miRNA sequences, seeds, KdModels, etc.) and produces as output standard and well-optimized bioconductor classes, allowing interoperability with other packages. Furthermore, the web interface enables its use by biologists without bioinformatic skills. To our knowledge, this is the first user-friendly tool offering i) transcript-level (rather than gene-level) repression predictions, ii) the prediction of TDMD and slicing sites, iii) the ability to scan any custom sequence for high-affinity sites through a user-friendly interface, and iv) the possibility to easily extend its functions for any specific species listed in miRbase. In addition, *scanMiR* enables the easy and flexible application of McGeary et al. (2019)'s biochemical model of miRNA-mediated repression, significantly outperforming the repression predictions of Targetscan.

An important limitation of *scanMiR*'s dissociation constant predictions is that they are based on the CNN from McGeary et al. (2019), which was in turn trained on experimental data from only a handful of miRNAs. As a result, predictions for miRNAs whose 12-mer affinity patterns strongly diverge from those can be expected to be less accurate. Indeed, whereas for some CNN-predicted microRNAs (e.g. miR-138), the accuracy of *scanMiR* repression predictions is more than twice as high as the accuracy obtained with Targetscan score predictions, for others, it is in fact comparable (miR-217) or even lower (miR-190a) (Fig. 3C). Nevertheless, we expect that with time and the accumulation of more experimental data, these predictions can be updated, and used within the same *scanMiR* framework. Furthermore, it has been recently shown for two example miRNAs that RBNS data might

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

also be generated on full miRNA sequences (McGeary et al., 2021). Although the authors were not able to generalize their findings on 3' supplementary binding across miRNAs, a CNN trained on complete miRNA- and full-length transcript-sequences, thereby also including information on structural accessibility, might still further improve miRNA target prediction.

Computational predictions of miRNA-mediated repression primarily focusing on the effectiveness of binding sites can only explain actual mRNA changes in cells to a certain degree. In fact predictions based on the biochemical model can be expected to account for around 50% of the changes in expression attributable to direct miRNA targeting (McGeary et al., 2019). Therefore other critical aspects, such as the accuracy of transcript as well as 3'UTR annotations are likely to improve miRNA target prediction in the future. Studies have suggested that alternative poly-adenylation can have important implications on predicting miRNA-mediated repression in different species and cell types (Nam et al., 2014). Indeed, in the HeLa/HEK cell lines, we observed much lower correlations when using a reference human transcriptome instead of the re-assembled transcriptome. Moreover, it has been shown that cis-acting factors, such as RNA-binding proteins (RBPs), can significantly change the binding of miRNAs to certain transcripts (Kedde et al., 2007, 2010; Mubaid et al., 2019). We hence hypothesize that future advances in computational miRNA target prediction may rather be accomplished by the integration of different high-throughput datasets (e.g. RBP-CLIP data or tissue specific smallRNA and mRNA expression data), than by the improvement of affinity predictions of miRNAs to single binding sites.

Finally, an important feature of *scanMiR* is its capability of identifying effective non-canonical as well as putative TDMD and miRNA slicing sites. We provide to our knowledge the first systematic screen for TDMD sites on non-coding transcripts highly expressed in the brain (see also Supplementary Tables 1-2), highlighting several uninvestigated candidates, which could play an important role in regulating miRNA abundance. Further, the scan for enriched circRNA-miRNA pairs in the brain confirmed the exceptional potency of Cdr1as as a highly-expressed miRNA sponge (Guo et al., 2014). In agreement with previous reports of other circRNAs potentially acting as (less effective) miRNA sponges (Kristensen et al., 2019), a few promising candidates emerged from our analysis which would warrant further investigation. We hope that the resources provided here will empower the community to investigate these, and more generally facilitate miRNA research.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

AVAILABILITY

scanMiR and its source code are available on Bioconductor, as well as in the following repositories:

https://github.com/ETHZ-INS/scanMiR

https://github.com/ETHZ-INS/scanMiRData

https://github.com/ETHZ-INS/scanMiRApp

The *scanMIR* web application is available at https://ethz-ins.org/scanMiR/, and can additionally be deployed anywhere using the *scanMiRApp* package.

All processed data and code underlying the figures is available in the paper's repository, at https://github.com/ETHZ-INS/scanMiR_paper

SUPPLEMENTARY DATA

Supplementary Figures 1-6

Supplementary Tables 1-4

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We thank Winston Becker for sharing the parameters of their model of Ago2 binding affinity, Kathy Lin for providing insights on how normalizations were calculated for the correlations between observed and predicted repression in McGeary et al. (2019), and Jinyang Zhang for providing the assembled circRNAs from Zhang et al. (2021).

FUNDING

This work was supported by the Swiss National Science Foundation [SNF_189486 to GS].

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

REFERENCES

Agarwal, V., Bell, G. W., Nam, J. W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *eLife*, *4*, 1–38. https://doi.org/10.7554/eLife.05005

Ameres, S. L., Martinez, J., & Schroeder, R. (2007). Molecular Basis for Target RNA Recognition and Cleavage by Human RISC. *Cell*, *130*(1), 101–112. https://doi.org/10.1016/j.cell.2007.04.037

Ameres, S. L., & Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nature Reviews Molecular Cell Biology*, *14*(8), 475–488. https://doi.org/10.1038/nrm3611

Amin, N. D., Bai, G., Klug, J. R., Bonanomi, D., Pankratz, M. T., Gifford, W. D., Hinckley, C. A., Sternfeld, M. J., Driscoll, S. P., Dominguez, B., Lee, K. F., Jin, X., & Pfaff, S. L. (2015). Loss of motoneuron-specific microRNA-218 causes systemic neuromuscular failure. *Science*, *350*(6267), 1525–1529. https://doi.org/10.1126/science.aad2509

Bartel, D. P. (2004). MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Bartel, D. P. (2018). Metazoan MicroRNAs. *Cell*, *173*(1), 20–51. https://doi.org/10.1016/j.cell.2018.03.006

Becker, W. R., Ober-Reynolds, B., Jouravleva, K., Jolly, S. M., Zamore, P. D., & Greenleaf, W. J. (2019). High-Throughput Analysis Reveals Rules for Target RNA Binding and Cleavage by AGO2. *Molecular Cell*, 1–15. https://doi.org/10.1016/j.molcel.2019.06.012

Biasini, A., Abdulkarim, B., de Pretis, S., Tan, J. Y., Arora, R., Wischnewski, H., Dreos, R., Pelizzola, M., Ciaudo, C., & Marques, A. C. (2021). Translation is required for miRNA-dependent decay of endogenous transcripts. *The EMBO Journal*, *40*(3), e104569. https://doi.org/10.15252/embj.2020104569

Blankenberg, D., Taylor, J., Nekrutenko, A., & The Galaxy Team. (2011). Making whole genome multiple alignments usable for biologists. *Bioinformatics*, *27*(17), 2426–2428.

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

https://doi.org/10.1093/bioinformatics/btr398

Carthew, R. W., & Sontheimer, E. J. (2009). Origins and Mechanisms of miRNAs and siRNAs. *Cell*, *136*(4), 642–655. https://doi.org/10.1016/j.cell.2009.01.035

Chi, S. W., Hannon, G. J., & Darnell, R. B. (2012). An alternative mode of microRNA target recognition. *Nature Structural and Molecular Biology*, *19*(3), 321–327. https://doi.org/10.1038/nsmb.2230

de la Mata, M., Gaidatzis, D., Vitanescu, M., Stadler, M. B., Wentzel, C., Scheiffele, P., Filipowicz, W., & Grosshans, H. (2015). Potent degradation of neuronal miRNAs induced by highly complementary targets. *EMBO reports*, *16*(4), 500–511. https://doi.org/10.15252/embr.201540078

Denzler, R., McGeary, S. E., Title, A. C., Agarwal, V., Bartel, D. P., & Stoffel, M. (2016). Impact of MicroRNA Levels, Target-Site Complementarity, and Cooperativity on Competing Endogenous RNA-Regulated Gene Expression. *Molecular Cell*, *64*(3), 565–579. https://doi.org/10.1016/j.molcel.2016.09.027

Eichhorn, S. W., Guo, H., McGeary, S. E., Rodriguez-Mias, R. A., Shin, C., Baek, D., Hsu, S., Ghoshal, K., Villén, J., & Bartel, D. P. (2014). MRNA Destabilization Is the Dominant Effect of Mammalian MicroRNAs by the Time Substantial Repression Ensues. *Molecular Cell*, *56*(1), 104–115. https://doi.org/10.1016/j.molcel.2014.08.028

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., & Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, *18*(10), 1139–1146. https://doi.org/10.1038/nsmb.2115

Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A., & Enright, A. J. (2006). miRBase: MicroRNA sequences, targets and gene nomenclature. *Nucleic Acids Research*, *34*(suppl_1), D140–D144. https://doi.org/10.1093/nar/gkj112

Grimson, A., Farh, K. K. H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007).

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, *27*(1), 91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Gumienny, R., & Zavolan, M. (2015). Accurate transcriptome-wide prediction of microRNA targets and small interfering RNA off-targets with MIRZA-G. *Nucleic Acids Research*, *43*(3), 1380–1391. https://doi.org/10.1093/nar/gkv050

Guo, J. U., Agarwal, V., Guo, H., & Bartel, D. P. (2014). Expanded identification and characterization of mammalian circular RNAs. *Genome Biology*, *15*(7), 409. https://doi.org/10.1186/s13059-014-0409-z

Hansen, T. B., Wiklund, E. D., Bramsen, J. B., Villadsen, S. B., Statham, A. L., Clark, S. J., & Kjems, J. (2011). MiRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *The EMBO Journal*, *30*(21), 4414–4422. https://doi.org/10.1038/emboj.2011.359

Hausser, J., Landthaler, M., Jaskiewicz, L., Gaidatzis, D., & Zavolan, M. (2009). Relative contribution of sequence and structure features to the mRNA binding of Argonaute/EIF2C–miRNA complexes and the degradation of miRNA targets. *Genome Research*, *19*(11), 2009–2020. https://doi.org/10.1101/gr.091181.109

Jonas, S., & Izaurralde, E. (2015). Towards a molecular understanding of microRNA-mediated gene silencing. *Nature Reviews Genetics*, *16*(7), 421–433. https://doi.org/10.1038/nrg3965

Kern, F., Backes, C., Hirsch, P., Fehlmann, T., Hart, M., Meese, E., & Keller, A. (2020). What's the target: Understanding two decades of in silico microRNA-target prediction. *Briefings in Bioinformatics*, *21*(6), 1999–2010. https://doi.org/10.1093/bib/bbz111

Kleaveland, B., CY, S., Stefano, J., & DP, B. (2018). *A Network of Noncoding Regulatory RNAs Acts in the Mammalian Brain. LID - S0092-8674(18)30634-2 [pii] LID - 10.1016/j.cell.2018.05.022 [doi]. 1097-4172 (Electronic)*. https://doi.org/10.1101/279687

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M., & Rajewsky, N. (2005). Combinatorial microRNA target

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

predictions. *Nature Genetics*, *37*(5), 495–500. https://doi.org/10.1038/ng1536

Kristensen, L. S., Andersen, M. S., Stagsted, L. V. W., Ebbesen, K. K., Hansen, T. B., & Kjems, J. (2019). The biogenesis, biology and characterization of circular RNAs. *Nature Reviews Genetics*. https://doi.org/10.1038/s41576-019-0158-7

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M. T., & Carey, V. J. (2013). Software for Computing and Annotating Genomic Ranges. *PLOS Computational Biology*, *9*(8), e1003118. https://doi.org/10.1371/journal.pcbi.1003118

Liu, W., & Wang, X. (2019). Prediction of functional microRNA targets by integrative modeling of microRNA binding and target expression data. *Genome Biology*, *20*(1), 18. https://doi.org/10.1186/s13059-019-1629-z

Mark Klik. (2020). *fst: Lightning Fast Serialization of Data Frames* (Version 0.9.4) [R].

McGeary, S. E., Bisaria, N., & Bartel, D. P. (2021). Pairing to the microRNA 3′ region occurs through two alternative binding modes, with affinity shaped by nucleotide identity as well as pairing position. *BioRxiv*, 2021.04.13.439700. https://doi.org/10.1101/2021.04.13.439700

McGeary, S. E., Lin, K. S., Shi, C. Y., Pham, T. M., Bisaria, N., Kelley, G. M., & Bartel, D. P. (2019). The biochemical basis of microRNA targeting efficacy. *Science*, *366*(6472). https://doi.org/10.1126/science.aav1741

Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., & Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, *13*(11), 1894–1910. https://doi.org/10.1261/rna.768207

Patro, R., Duggal, G., Love, M. I., Irizarry, R. A., & Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, *14*(4), 417–419. https://doi.org/10.1038/nmeth.4197

Reczko, M., Maragkakis, M., Alexiou, P., Grosse, I., & Hatzigeorgiou, A. G. (2012). Functional microRNA targets in protein coding sequences. *Bioinformatics*, *28*(6), 771–776. https://doi.org/10.1093/bioinformatics/bts043

bioRxiv preprint doi: https://doi.org/10.1101/2021.06.16.448293; this version posted June 17, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

Sætrom, P., Heale, B. S. E., Snøve, O., Aagaard, L., Alluin, J., & Rossi, J. J. (2007). Distance constraints between microRNA target sites dictate efficacy and cooperativity. *Nucleic Acids Research*, *35*(7), 2333–2342. https://doi.org/10.1093/nar/gkm133

Salomon, W. E., Jolly, S. M., Moore, M. J., Zamore, P. D., & Serebrov, V. (2015). Single-Molecule Imaging Reveals that Argonaute Reshapes the Binding Properties of Its Nucleic Acid Guides. *Cell*, *162*(1), 84–95. https://doi.org/10.1016/j.cell.2015.06.029

Schirle, N. T., Sheu-Gruttadauria, J., & MacRae, I. J. (2014). Structural basis for microRNA targeting. *Science*, *346*(6209), 608–613. https://doi.org/10.1126/science.1258040

Sheu-Gruttadauria, J., Pawlica, P., Klum, S. M., Wang, S., Yario, T. A., Schirle Oakdale, N. T., Steitz, J. A., & MacRae, I. J. (2019). Structural Basis for Target-Directed MicroRNA Degradation. *Molecular Cell*, *75*(6), 1243-1255.e7. https://doi.org/10.1016/j.molcel.2019.06.019

Sheu-Gruttadauria, J., Xiao, Y., Gebert, L. F., & MacRae, I. J. (2019). Beyond the seed: Structural basis for supplementary micro RNA targeting by human Argonaute2. *The EMBO Journal*, *38*(13), 1–14. https://doi.org/10.15252/embj.2018101153

Wang, Y., Soneson, C., Malinowska, A. L., Laski, A., Ghosh, S., Kanitz, A., Gebert, L. F. R., Robinson, M. D., & Hall, J. (2021). MiR-CLIP reveals iso-miR selective regulation in the miR-124 targetome. *Nucleic Acids Research*, *49*(1), 25–37. https://doi.org/10.1093/nar/gkaa1117

Yekta, S., Shih, I. -hung, & Bartel, D. P. (2004). MicroRNA-Directed Cleavage of HOXB8 mRNA. *Science*, *304*(5670), 594–596. https://doi.org/10.1126/science.1097434

Zhang, J., Hou, L., Zuo, Z., Ji, P., Zhang, X., Xue, Y., & Zhao, F. (2021). Comprehensive profiling of circular RNAs with nanopore sequencing and CIRI-long. *Nature Biotechnology*, 1–10. https://doi.org/10.1038/s41587-021-00842-6