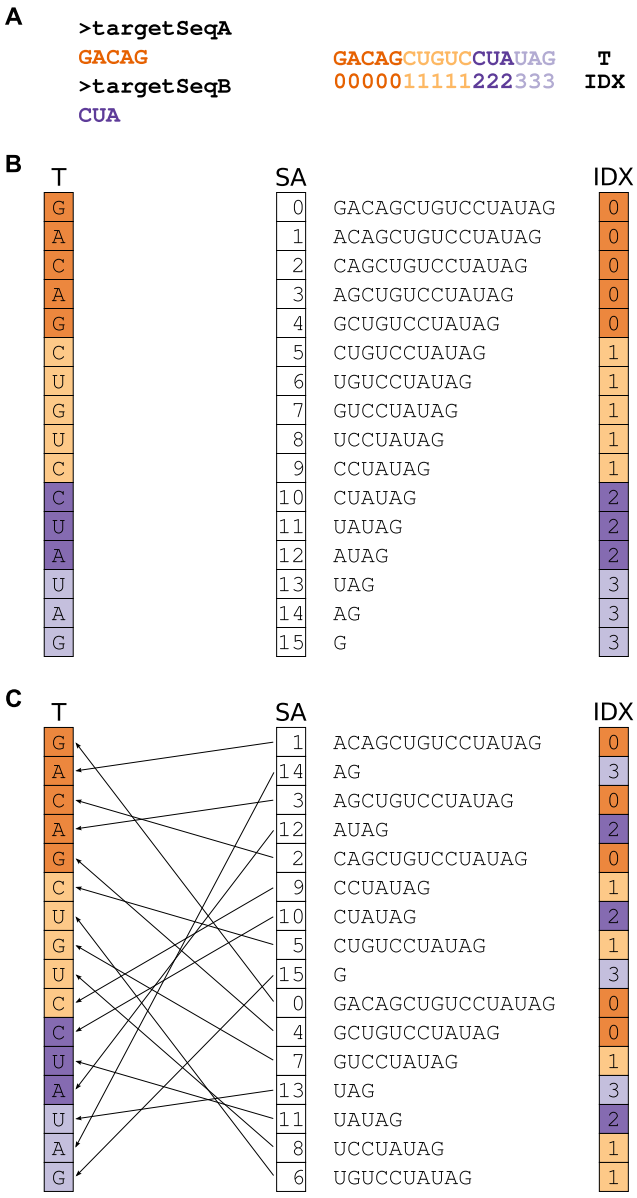
RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets (Supplementary Document)

Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, Anders Frost Rudebeck, Peter F. Stadler, Ivo L. Hofacker, and Jan Gorodkin*

January 17, 2017

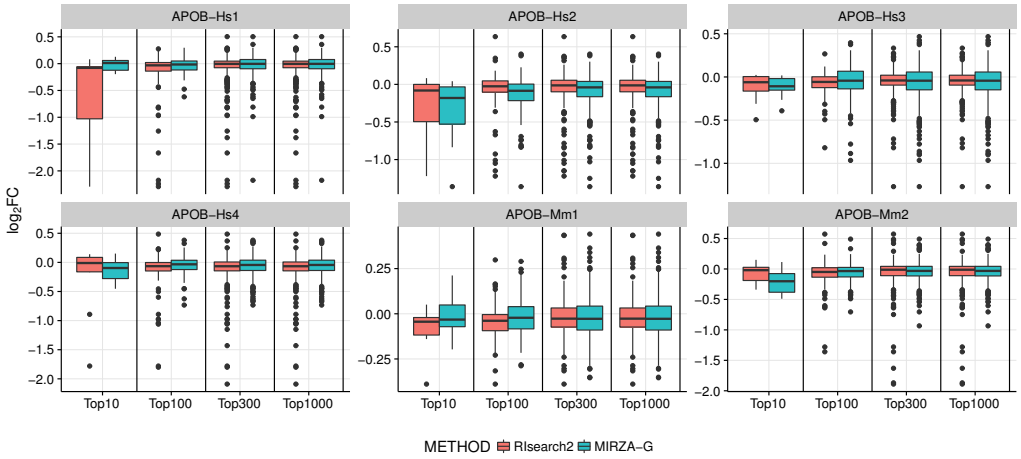
*To whom correspondence should be addressed. Tel: +45 353 33578; Fax: +45 353 34704; Email: gorodkin@rth.dk



Supplementary Figure S1: Building the target suffix array. (A) Given a target FASTA file as shown on the left, the concatenated target string T contains each sequence in original orientation back-to-back with its own reverse complement. Values in IDX identify the originating sequence (where odd numbers mark the reverse complement of the previous sequence). (B) Listed are all 16 suffixes of T with their starting positions (in range $0 \ldots 15$). T and IDX are in order as in (A). (C) To construct the suffix array SA, the suffixes are sorted in lexicographic order. The arrows indicate where the suffix is found in the original string T. IDX is re-arranged according to SA. The three vectors T, SA, and IDX are condensed into one 64-bit array and stored on disc for further use.

Supplementary Table S1: Six siRNAs from Burchard dataset.

siRNA ID	siRNA antisense sequence (5' to 3')
APOB-Hs1	AAUUUUUCAAAGUUCCAAU
APOB-Hs2	UAGUUAUUCAGGAAGUCUA
APOB-Hs3	AUUUCAGGAAUUGUUAAAG
APOB-Hs4	UUGGUAUUCAGUGUGAUGA
APOB-Mm1	UUUCAAUUGUAUGUGAGAG
APOB-Mm2	UUUUGCUUCAUUAUAGGAG



Supplementary Figure S2: siRNA-specific $\log_2 FC$ distributions of top transcripts predicted by RIsearch2 (siRNA off-targets pipeline) and MIRZA-G in the Burchard dataset, PLC/PRF/5 cell line. Note that this figure should be evaluated together with Figure 3 in the main document.

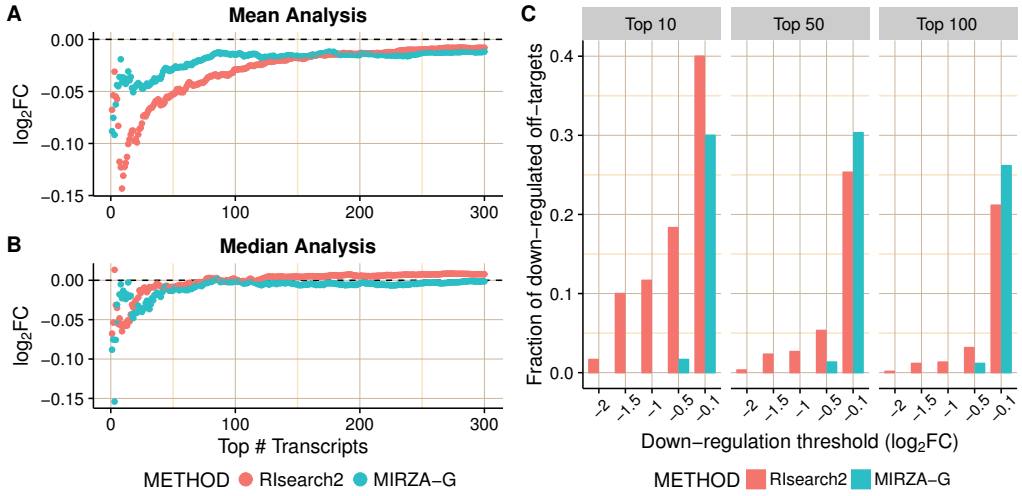
Supplementary Table S2: siRNAs considered for the evaluation of off-targeting potential measures.

Dataset ID	Array ID	siRNA antisense sequence (5' to 3')	on-target gene	POTS	$POFF_{3'}$	POFF
Dharmacon2008	1595297361	UGUUGCGUUCGGAGAGGCG	ENSG00000166794	25	0.001885	0.004523
Dharmacon2008	1595297366	UUCGUAGGUCAAAAUACAC	ENSG00000166794	23	0.002073	0.005964
Dharmacon2008	1595297373	CACCGUAGAUGCUCUUUCC	ENSG00000166794	29	0.004996	0.44155
Dharmacon2008	1595297378	GUUGCGUUCGGAGAGGCGC	ENSG00000166794	40	0.000135	0.000395
Dharmacon2008	1595297383	UUCAUCUCCAAUUCGUAGG	ENSG00000166794	407	0.001081	0.05085
Dharmacon2008	1595297389	AUGCUCUUUCCUCCUGUGC	ENSG00000166794	410	1.000000	1.000000
Dharmacon2008	1595297394	UUUUGGAACAGUCUUUCCG	ENSG00000166794	510	0.107259	0.766603
Dharmacon2008	1595297399	UUUGGAACAGUCUUUCCGA	ENSG00000166794	410	0.071367	0.454657
Dharmacon2008	1595297422	CCAAACACCACAUGCUUGC	ENSG00000166794	597	0.004381	0.398346
Dharmacon2008	1595297427	AAAUACACCUUGACGGUGA	ENSG00000166794	595	0.986031	0.992469
Dharmacon2008	1595297438	GCAGGAAGAAGACGGACCC	ENSG00000166794	667	0.000884	0.001998
Dharmacon2008	1595297444	CAGGCUGUCUUGACUGUCG	ENSG00000166794	560	0.001269	0.021349
Dharmacon2008	1595297470	UCCGACGCCUGCUUCACCA	ENSG00000111640	15	0.002447	0.785576
Dharmacon2008	1595297477	GGUCGUUGAGGGCAAUGCC	ENSG00000111640	28	0.00336	0.997732
Dharmacon2008	1595297486	UCGUUGAGGGCAAUGCCAG	ENSG00000111640	30	0.009296	0.992592
Dharmacon2008	1595297491	AAGCUUCCCGUUCUCAGCC	ENSG00000111640	308	0.137224	0.956974
Dharmacon2008	1595297496	AAGUCAGAGGAGACCACCU	ENSG00000111640	295	0.068246	0.997708
Dharmacon2008	1595297501	AUGAGCCCCAGCCUUCUCC	ENSG00000111640	336	0.010255	0.434668
Dharmacon2008	1595297507	CAAGCUUCCCGUUCUCAGC	ENSG00000111640	312	0.005592	0.928328
Dharmacon2008	1595297513	UGGCAGUGAUGGCAUGGAC	ENSG00000111640	380	0.011579	0.981844
Dharmacon2008	1595297518	AUUUCCAUUGAUGACAAGC	ENSG00000111640	736	0.000385	0.339128
Dharmacon2008	1595297524	AAAAGCAGCCCUGGUGACC	ENSG00000111640	696	0.091171	0.978453
Dharmacon2008	1595297530	GAGGCUGUUGUCAUACUUC	ENSG00000111640	560	0.366275	0.994911
Dharmacon2008	1595297535	CAUAUUUGGCAGGUUUUUC	ENSG00000111640	848	0.002103	0.431927
Dharmacon2008	1595297546	GAGGCAGGGAUGAUGUUCU	ENSG00000111640	733	0.001725	0.414098
Dharmacon2006	16012097016666	AGUUGCUUCAAAUCUGCUC	ENSG00000169032	253	0.633537	0.927307
Dharmacon2006	16012097016667	ACUUGAUCCAGAGAACCUC	ENSG00000169032	225	0.9472	0.99867
Dharmacon2006	16012097016668	UCAAAUCUGCUCUCUCUGC	ENSG00000169032	380	0.86958	0.9568
Dharmacon2006	16012097016669	AGAACCUCCAUCCAUGUGC	ENSG00000169032	211	0.999987	0.999999
Dharmacon2006	16012097017936	UCACCGUAGAUGCUCUUUC	ENSG00000166794	67	0.043415	0.185472
Dharmacon2006	16012097017939	UUUGUAGCCAAAUCCUUUC	ENSG00000166794	233	0.023665	0.947547
Dharmacon2006	16012097017951	ACACGAUGGAAUUUGCUGU	ENSG00000166794	64	0.002071	0.006507
Dharmacon2006	16012097017953	UUUUUGGAACAGUCUUUCC	ENSG00000166794	480	0.444406	0.929395
Dharmacon2006	16012097018568	UGGUGAAGUCUCCGCCCUG	ENSG00000166794	338	0.009484	0.21155
Dharmacon2006	251209725370	UAUUGGAACAUGUAAACCA	ENSG00000111640	239	0.001483	0.917924
Dharmacon2006	251209725411	CUUGAGGCUGUUGUCAUAC	ENSG00000111640	308	0.002144	0.836531
Dharmacon2006	251209725538	CUCUCCUGUAGCUAAGGCC	ENSG00000166794	638	0.357868	0.99135

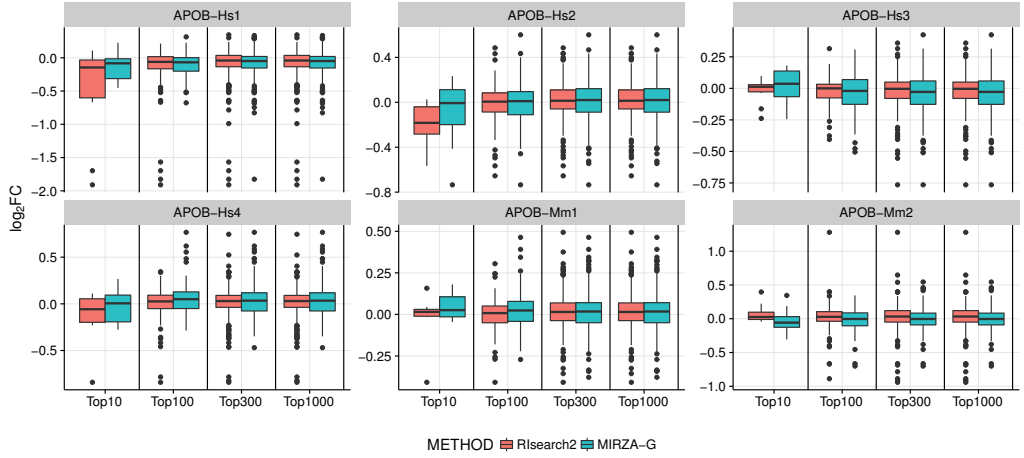
Continued on next page

Supplementary Table S2 – continued from previous page

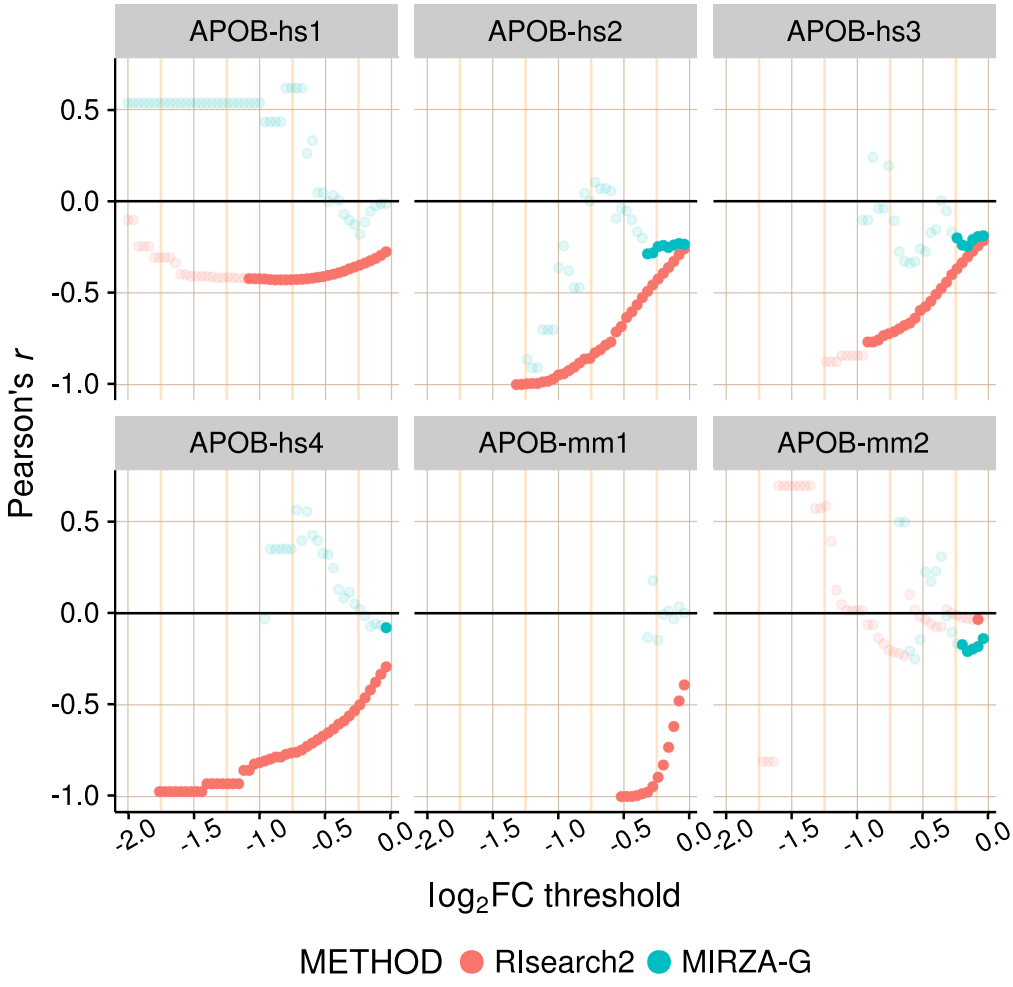
Dataset ID	Array ID	siRNA antisense sequence (5' to 3')	on-target gene	POTS	$POFF_{3'}$	POFF
GSE5769	GSM134317	GGCCAAAGAUUCAAAGCCA	ENSG00000121879	347	0.999903	0.999906
GSE5769	GSM134319	GGUGAUCACUCUCCUUCAU	ENSG00000138182	220	0.999291	0.999514
GSE5769	GSM134321	CAGAAACCUGGUGGUCAAU	ENSG00000115904	500	0.228054	0.700913
GSE5769	GSM134323	CUCAAAUAGGUCGUCCUCA	ENSG00000171132	646	0.895007	0.99331
GSE5769	GSM134325	CUUACCCUUUUGGGUUCUG	ENSG00000134086	176	1.000000	1.000000
GSE5769	GSM134521	AACCGCAGUUCUCUGUAGG	ENSG00000112062	56	0.092327	0.94342
GSE5769	GSM134327	CACACCCUGCCUAUUUCCU	ENSG00000134086	228	1.000000	1.000000
GSE5769	GSM134551	CUAGUGUCAUUCGCAUGUC	ENSG00000138182	189	0.081901	0.120337
GSE5291	GSM119707	GCAAGUCUCCAACAUGCCU	ENSG00000142168	297	0.106708	0.682785
GSE5291	GSM119708	CGCAAGUCUCCAACAUGCC	ENSG00000142168	232	0.584707	0.775272
GSE5291	GSM119709	GCGCAAGUCUCCAACAUGC	ENSG00000142168	37	0.210622	0.30274
GSE5291	GSM119710	UGCGCAAGUCUCCAACAUG	ENSG00000142168	31	0.13633	0.729004
GSE5291	GSM119741	UUGCGCAAGUCUCCAACAU	ENSG00000142168	35	0.029517	0.840096
GSE5291	GSM119742	AUUGCGCAAGUCUCCAACA	ENSG00000142168	26	0.303014	0.602383
GSE5291	GSM119743	CAUUGCGCAAGUCUCCAAC	ENSG00000142168	28	0.02303	0.884827
GSE5291	GSM119744	ACAUUGCGCAAGUCUCCAA	ENSG00000142168	160	0.02209	0.059624
GSE5291	GSM119745	CACAUUGCGCAAGUCUCCA	ENSG00000142168	261	0.003989	0.080062
GSE5291	GSM119746	UCACAUUGCGCAAGUCUCC	ENSG00000142168	471	0.013149	0.125626
GSE5291	GSM119747	GUCACAUUGCGCAAGUCUC	ENSG00000142168	529	0.200849	0.085258
GSE5291	GSM119748	AGUCACAUUGCGCAAGUCU	ENSG00000142168	331	0.155389	0.041397
GSE5291	GSM119749	CAGUCACAUUGCGCAAGUC	ENSG00000142168	323	0.033343	0.021682
GSE5291	GSM119750	GCAGUCACAUUGCGCAAGU	ENSG00000142168	343	0.004085	0.032111
GSE5291	GSM119759	AGCAGUCACAUUGCGCAAG	ENSG00000142168	244	0.002539	0.029897
GSE5291	GSM119761	CAGCAGUCACAUUGCGCAA	ENSG00000142168	287	0.054073	0.253648
GSE5291	GSM119762	GUCAGCAGUCACAUUGCGC	ENSG00000142168	449	0.051946	0.154575
GSE5291	GSM119763	UGUCAGCAGUCACAUUGCG	ENSG00000142168	248	0.109305	0.173462



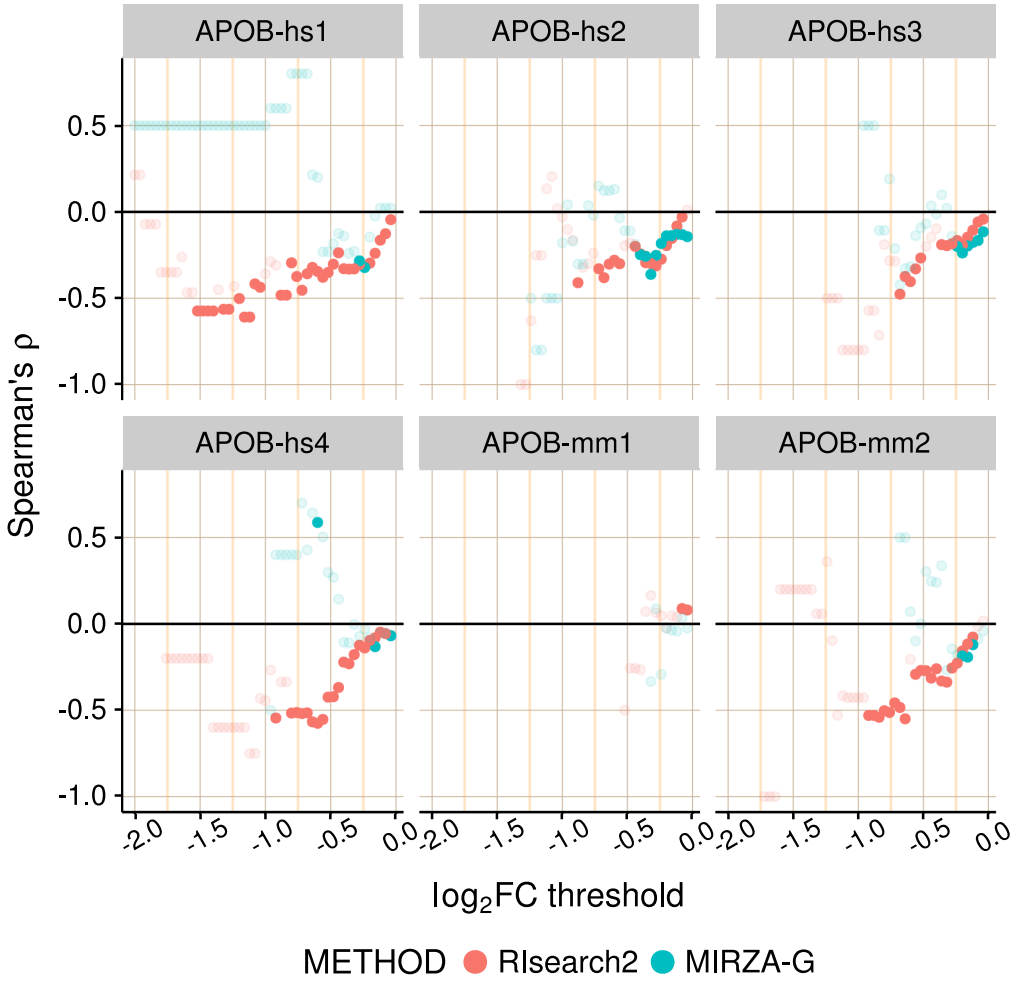
Supplementary Figure S3: Evaluation of individual off-target predictions performed by RIsearch2 (siRNA off-targets pipeline) and MIRZA-G. All analyses are based on the Burchard dataset where six *APOB* siRNAs were transfected into HUH7 cell lines. Top off-targets are selected from method-specific predictions based on the confidence scores generated by either method ($p_{off,j}$ for RIsearch2). In the mean (A) and the median (B) fold changes analyses, mean/median differential expression level (averaged over all siRNAs) of top off-targets are plotted for different numbers of top off-targets considered. (C) Proportion of critical off-targets within the top (10, 50, 100) predictions generated by either method for six siRNAs. Off-targets are considered critical if they are down-regulated, upon siRNA transfection, with a $\log_2 FC$ value lower than the threshold given in the x-axis.



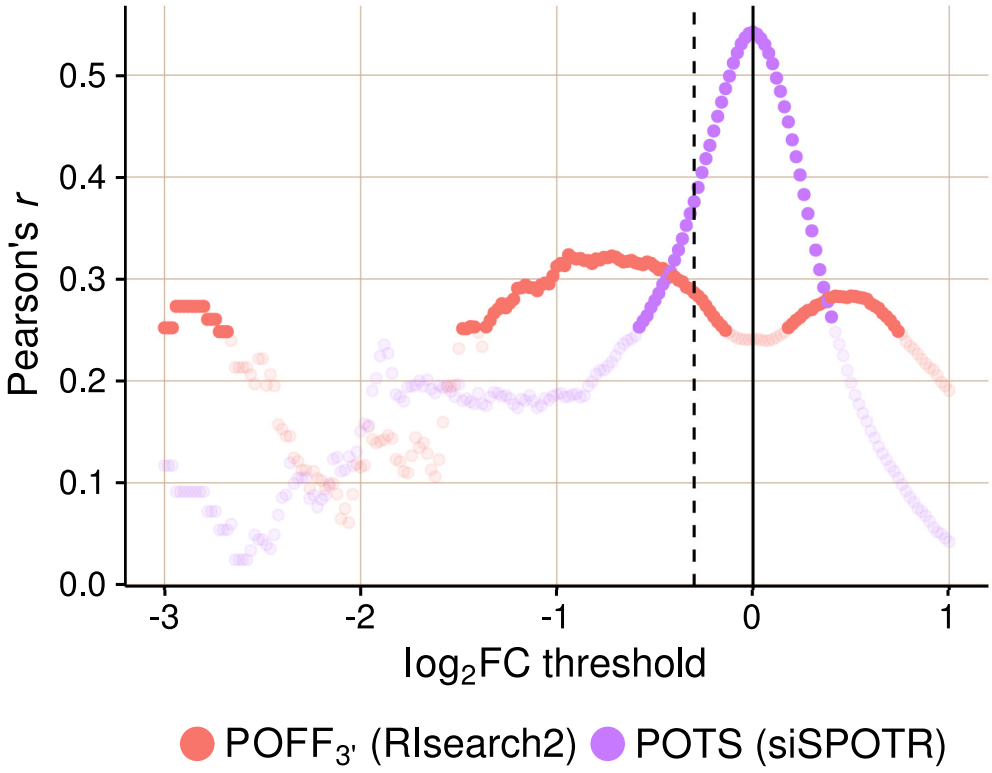
Supplementary Figure S4: siRNA-specific $\log_2 FC$ distributions of top transcripts predicted by RIsearch2 (siRNA off-targets pipeline) and MIRZA-G in the Burchard dataset, HUH7 cell line. Note that this figure should be evaluated together with Supplementarty Figure 3.



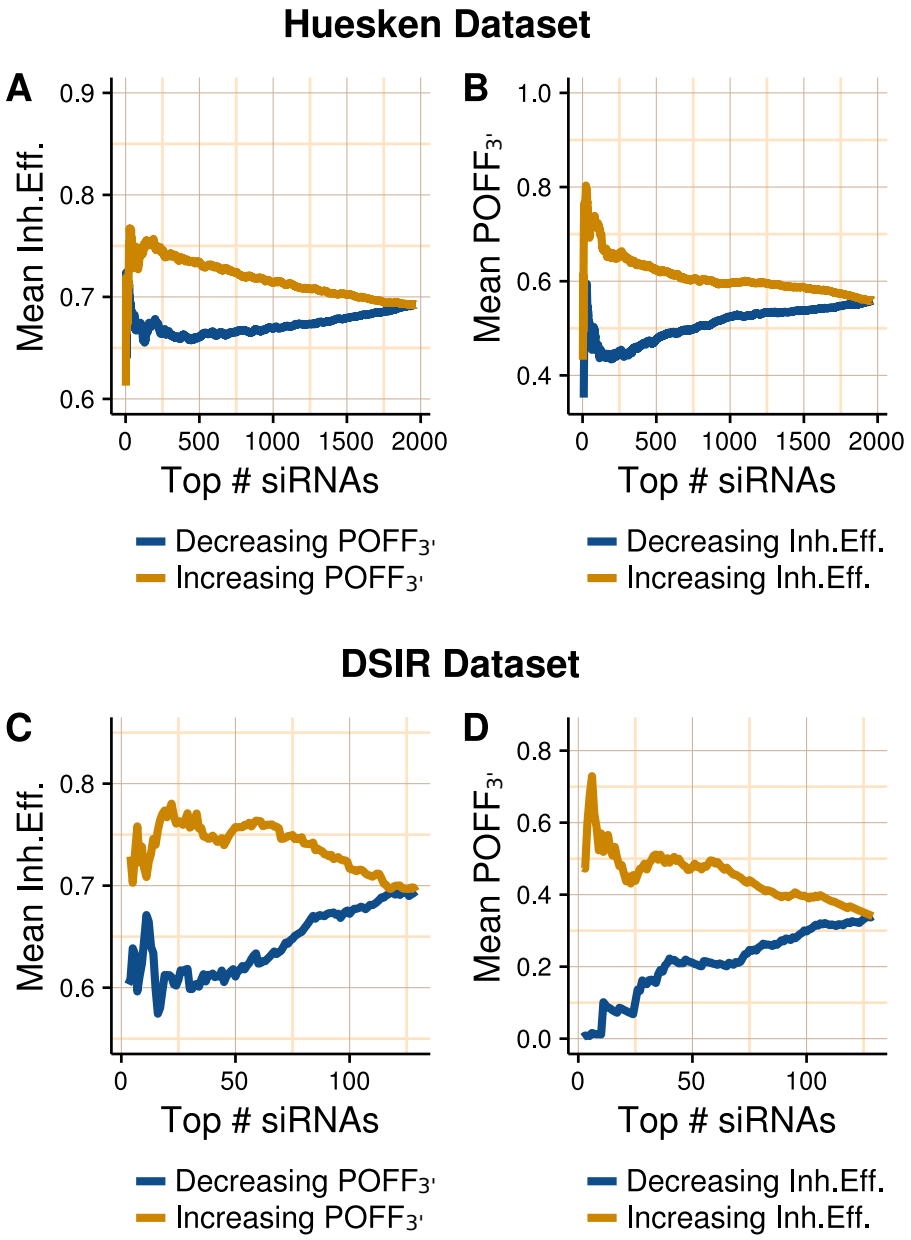
Supplementary Figure S5: Pearson correlation analysis between transcript-specific down-regulation measure ($\log_2 FC$) and off-target confidence scores generated by either method ($p_{off,j}$ for RIsearch2). Each point corresponds to a method- and threshold-specific Pearson's correlation coefficient (y-axis) between method-based off-target confidence score and down-regulation level ($\log_2 FC$ value). The set of transcripts considered in each of them is limited to those with down-regulation lower than the $\log_2 FC$ threshold given in the x-axis and that are predicted as an off-target by the corresponding method. Results are given separately for six siRNA transfection experiments in PLC/PRF/5 cell line. Only significant correlations (p-value < 0.05) are highlighted.



Supplementary Figure S6: Rank correlation analysis between transcript-specific down-regulation measure ($\log_2 FC$) and off-target confidence scores generated by either method ($p_{off,\,j}$ for RIsearch2). Each point corresponds to a method- and threshold-specific Spearman's rank correlation coefficient (y-axis) between method-based off-target confidence score and down-regulation level ($\log_2 FC$ value). The set of transcripts considered in each of them is limited to those with down-regulation lower than the $\log_2 FC$ threshold given in the x-axis and that are predicted as an off-target by the corresponding method. Results are given separately for six siRNA transfection experiments in PLC/PRF/5 cell line. Only significant correlations (p-value < 0.05) are highlighted.



Supplementary Figure S7: Evaluation of predicted off-targeting potential measures with siRNA-specific differential expression data from 63 siRNA transfection experiments. Pearson's r is given for each correlation analysis between method-specific off-targeting potential measure, RIsearch2-based $POFF_{3'}$ or siSPOTR-based $POTS$, and threshold-specific expression-based off-targeting potential measure of 63 siRNAs. The expression-based measure is the *total differential expression* (sum of corresponding $-log_2FC$ values) of transcripts that have a complementary heptamer to the siRNA seed region in their 3' UTR and are differentially expressed upon siRNA transfection with a $\log_2 FC$ value lower than the threshold given on the x-axis. Only significant correlations (p-value < 0.05) are highlighted and the $\log_2 FC$ threshold -0.3, which is the evaluation threshold employed in the siSPOTR study, is shown with a dashed line.



Supplementary Figure S8: Relationship between inhibition efficiency (*Inh.Eff.*) and off-targeting potential measure $POFF_{3'}$ of siRNAs from Huesken (A, B) and DSIR (C, D) datasets. A and C show the mean *Inh.Eff.* of top n siRNAs when siRNAs are ranked by $POFF_{3'}$ measure in increasing and decreasing order. B and D show the mean $POFF_{3'}$ measure of top n siRNAs when siRNAs are ranked by *Inh.Eff.* in increasing and decreasing order. Note that y-axes are not in the same scale.

Supplementary Table S3: Execution time and parameter settings of RNA–RNA interaction prediction tools based on a small test set (63 siRNAs vs. 5982 transcripts from human chromosome X and their antisense, all transcripts are longer than 80 nt). The first row, RIsearch2 (precomputation), corresponds to the creation of the target suffix array, which is used by subsequent runs. We timed RIsearch2 with different parameter settings, all with an energy threshold of −10 kcal/mol and seed extension limited to 20 nt up-/downstream. We present running times for the single-threaded version for six different seed settings, as well as running times for a fixed seed setting with increasing number of threads (2, 4, 8, 16). The times for RIsearch2 cover an entire run, including reading the index and writing compressed output files, but excluding the precomputation (first row). Running time with the seed definition `1:12/6`, default for the siRNA off-targets discovery pipeline, is given in bold. For the predecessor version RIsearch, we show the default behaviour in the first row, which only returns the single best interaction per query (siRNA)–target pair, and the variant with score threshold (`-s 1560`) that relates to a maximum hybridisation energy of −10 kcal/mol. We selected a seed length of six for GUUGle and IntaRNA, and parameters `-w 80`, `-L 40` and `-l 30` for IntaRNA in parallel with our siRNA off-target discovery pipeline. For RNAplex, `-e -10` and `-l 30` were also selected in parallel with our pipeline. MIRZA-G, PITA and miRanda were timed with recommended default settings with an additional run for miRanda where predictions were filtered with −10 kcal/mol energy threshold. The PITA and IntaRNA runs have been aborted after running for more than two days. Longer running times of these two tools can in part be explained by the fact that they also compute accessibility profiles on given target transcripts.

Method	**Parameters**	**Execution time**	
		in seconds	equivalent
RIsearch2 (precomputation)	`-c target.fa -o target.suf`	~7 s	~.1 m
RIsearch2	`-s 2:7 -e -10 -l 20 -t 1`	~20 s	~.3 m
RIsearch2	`-s 2:7/5 -e -10 -l 20 -t 1`	~78 s	~1.3 m
RIsearch2	`-s 1:12/6 -e -10 -l 20 -t 1`	**~89 s**	**~1.5 m**
RIsearch2	`-s 1:12/7 -e -10 -l 20 -t 1`	~32 s	~.5 m
RIsearch2	`-s 1:12/8 -e -10 -l 20 -t 1`	~11 s	~.2 m
RIsearch2	`-s 6 -e -10 -l 20 -t 1`	~171 s	~2.8 m
RIsearch2	`-s 6 -e -10 -l 20 -t 2`	~88 s	~1.5 m
RIsearch2	`-s 6 -e -10 -l 20 -t 4`	~50 s	~.8 m
RIsearch2	`-s 6 -e -10 -l 20 -t 8`	~28 s	~.5 m
RIsearch2	`-s 6 -e -10 -l 20 -t 16`	~23 s	~.4 m
RIsearch		~375 s	~6.3 m
RIsearch	`-s 1560 -p1`	~940 s	~15.7 m
GUUGle	`-d 6`	~220 s	~3.7 m
miRanda		~568 s	~9.5 m
miRanda	`-en -10`	~570 s	~9.5 m
MIRZA-G	`-T calculate_per_gene_scores -v 4`	~3669 s	~1 h
RNAplex	`-e -10 -l 30`	~32052 s	~9 h
PITA		>172800 s	>2 d
IntaRNA	`-p 6 -f 1,12 -P -w 80 -L 40 -l 30`	>172800 s	>2 d