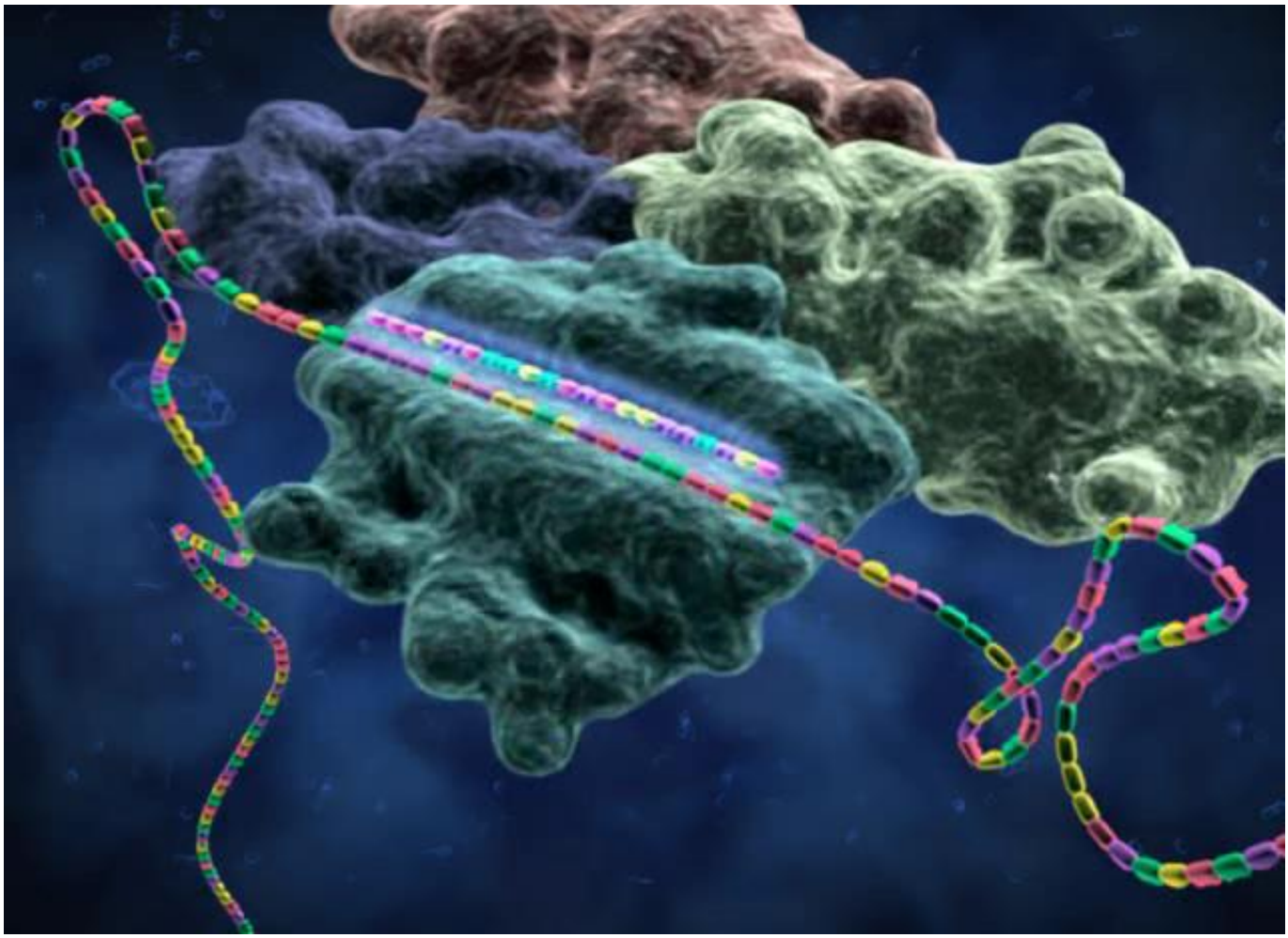
# Bioinformatics Challenges and Advances in RNA interference

**Author :** Deepak Anand

**Categories :** Algorithms, Machine Learning, Tools

**Date :** June 25, 2017



**RNA interference** is a post-transcriptional gene regulatory mechanism to down-regulate the gene expression either by mRNA degradation or by mRNA translation inhibition. The mechanism involves a small partially complementary RNA against the target gene. To perform the action, it also requires a class

of dedicated proteins to process these primary RNAs into mature microRNAs. The guide sequence determines the specificity of the miRNA. Therefore, the knowledge of guide sequence is crucial for predicting its targets and also exploiting the sequence to create a new regulatory circuit. In this short review, we will briefly discuss the role and challenges in miRNA research for unveiling the target prediction by bioinformatics and to foster our understanding and applications of RNA interference.

Regulation of gene expression is the key component to any cell or organism. Each process, starting with birth to development and growth requires regulation of genes at a given time. The regulation of promoter activation by post-transcriptional regulation by regulatory RNAs [1], regulatory elements within the UTR [2], mRNA degradation by ribonucleases [3, 4] or by protein degradation [5] provide comprehensive regulation of gene expression. In higher organisms, these mechanisms have evolved into complex and multi-tiered but well-structured systems to regulate gene expression.

RNA interference was discovered by Andrew Fire and Craig C. Mello in 1998 and jointly received the Nobel Prize for physiology and medicine in 2006 for the discovery of “RNA interference” [6]. RNA interference or RNAi is a post-transcriptional gene silencing (PTGS) mechanism in eukaryotes whose role is important in various stages of life or growth conditions. It can modulate the stability of mRNA or inhibit its translation. These effector molecules are named depending on their origin. While short interfering RNA (siRNA) are exogenously synthesized mature RNA or viral encoded RNA, microRNA (miRNAs) are endogenously encoded RNAs [7] and are required for the basic physiological function of the cell. These RNAs have unique secondary structure and exhibit regulatory functions exclusively. While the siRNAs can have perfect complementarity to their target, miRNAs share only partial complementary with their target, therefore, allowing outputs different from that of siRNA. The identification new miRNAs and their targets are important as they not only control basic physiological function but are also associated with various human diseases. For example, miR-600 has been found to control the cell proliferation, migration, and invasion in human CRC cell lines [8]. Besides studies on human cell lines, miRNA investigations are also focused extensively on insects, and plants development [9-11].

**Synthesis and mechanism of action**

While miRNAs are synthesized during various stages of growth, siRNA could be viral or synthetic in origin. Short hairpin RNA (shRNAs) are the class of regulatory RNA similar to miRNA in design, structure, and processing [6, 7, 12], and are expressed from a plasmid or through viral vectors. Although regulatory RNAs could be of any type, their mechanism of action in all cases remains same, because the siRNA or shRNA are designed such a way that they can integrate into the host RNAi pathway.

A miRNA is transcribed as a primary RNA (pri-miRNA) of large size inside the nucleus. pri-miRNA with overhangs at both 5? and 3? sequence forms a hairpin loop and a stem structure (Fig 1). These structures serve as a template for initial processing by a specific ribo-endonucleases, resulting in a 60-70 nucleotide transcript with a hairpin structure known as pre-miRNA. Pre-miRNA is transported from nucleus to the cytoplasm, where a second endonuclease reaction cleaves off the hairpin loop from pre-miRNA, into a short double stranded RNA product (miR-miR*) [6, 13]. The resulting miR-miR* duplex is partially complementary and contains 5? phosphate and 3? overhang of two nucleotides. The strands of this duplex are called “Guide strand” and “Passenger-strand.” Argonaut protein shortens the passenger strand, and the miR-guide strand is retained (Fig1). The guide-miRNA along with Argonaut complex is

called the *RNA interference silencing complex* (RISC), which can recognize complementary target sequence on 3? untranslated region (3'UTR) of the mRNA. A 2-8 nucleotide sequence at 5'end called *seed sequence* provides the specificity to the target mRNA. RISC complex either degrades the bound mRNA or block translation [12, 14, 15] (Fig 1).

Fig 1- Biogenesis of microRNA and target repression.

**Identification of regulatory RNAs/miRNAs**

Initial studies of miRNA discoveries were based on genetic screening [16]. The screening methods were time-consuming and less efficient. In recent years, the most common methods for detection and identification of novel miRNA include RNA sequencing (RNA-seq), transposon mutagenesis [17], Watson-Crick base pairing with probes (northern hybridization, *in-Situ* hybridization, and RT-PCR), cloning and splinted ligation. Sequencing methods use robust and high throughput technology. While RNA sequencing provides direct identification of novel transcripts, other methods of hybridization need prior knowledge of miRNA sequence. For experimental detection, use of locked nucleic acid (LNA) probes improves hybridization efficiency and specificity [18] and helps in the validation of the predicted transcripts. Different methods have their advantages and disadvantages. For example, besides detecting the miRNA using northern hybridization, quantitation can also be performed by direct detection. Additionally, PCR-based detection is favored for low copy templates, which is enabled by many log amplifications.

**miRNA identification and target prediction: the role of bioinformatics**

Although predicting novel RNAi genes through computational approach presents a challenge. However, the methods are good for shortlisting the putative candidates that can be verified through experimental methods. The criteria used for candidate miRNA prediction are based on the comparative genomics with related organisms, genetic loci, the size of the transcript, the structure of primary RNA and free energy of the structure. Report of various strategies used for prediction of candidate micro-RNA-like *micro-conserved elements* (MCE) could be one of the many approaches [19]. The efficiency of prediction can be achieved through several modifications in the algorithms, as well as using the experimental evidence. A list of such tools has been reviewed by Salim and Chandra [20]. Some of these tools are stand alone and can function independently whereas others are based on web services or require other platforms. While all these tools used for prediction and validation of targets bring fewer publications, a huge number of research publications comes from experimental data (Fig 2), suggesting exponential growth in RNA interference research.

Fig 2 Growth in miRNA research as compared to that of miRNA/miRNA target prediction.

Some bioinformatics analysis includes unraveling counterparts and critical components of the RNAi pathway in various organisms. The genes involved in miRNA pathway of many organisms have been experimentally identi?ed, and the information can be accessed from a registry that was constructed by sequencing small RNAs (~22 nt) and mapping them to the genome [13]. The miRNA sequence can occur in intergenic regions, in introns of protein coding regions or exons and introns of noncoding genes. Through bioinformatics tools, the miRNA genes can be identified by using the known biases in targeting strands. This reduces the unwieldy work of analyzing the whole genome of the organism, although, it

could be difficult to achieve easily in the case of miRNA, as there are significant mismatches between the miRNA single strand and the cognate mRNA. For this purpose, there are programs like the MiRscan57 and MiRseeker [21] that use the conserved regions of the stem-loop structures to identify novel miRNA genes across species. Since miRNAs work jointly with other members of pathways, the study of tissue speci?city of genes can help in reviewing miRNA and their targets. Genome sequences also provide ways to predict miRNAs, but verification of expression need to be done by sequencing of RNA species or by other methods of choice.

For RNA interference, a strong binding of the 5? end (the ?rst eight base pairs) of the mature miRNA to the target 3? UTR sequence is essential. Silencing ef?ciency is affected and reduced by G:U wobble pairing in the interacting region. Algorithms and models have been developed for miRNA target interactions, based on experimental evidence. Since the microarray experiments show that miRNAs exhibit tissue-speci?c expression patterns, correlation of mRNA with miRNA expression can help narrow down the predicted targets. Moreover, based on the patterns of expression, the miRNAs and mRNAs can be clustered. Several online analysis tools for miRNA target prediction have been documented in recent past [13, 20, 22-24]. Together with the miRNA identification, the prediction of their promoters is an ever-present challenge, although many underlying assumptions like homologies, UTR location of target sites, can help in identifying good targets. Therefore, caution must be taken for false positive predictions [25] as the coverage of all possible sites, or complete sequence may not be possible, some of which may be unique to certain species and may not follow the rules [22, 23].

**siRNA/shRNA design and target prediction**

The siRNAs originate from transgenes, viruses, and transposons and form perfect complementary double strand RNA precursors (dsRNAs) [6, 21]. Due to the perfect complementarity, they can silence the same gene from which they originate. The siRNA precursor generates multiple siRNAs that can be transmitted among cells [23, 24]. Processing of siRNA requires their dicing and sorting by Argonaut proteins. The siRNAs get incorporated into RISC and silence genes that are complementary to the RISC incorporated strand. Because of the strand bias showed by RISC, the strand which is more unstable, or whose end is easier to open, is taken up by the RISC [12, 14]. Also, efficient silencing also depends on the base composition along the siRNA active strand. However, the siRNAs can also show off-target effects and silence unintended targets. Thus, a rational design of siRNA is critical.

A unique siRNA selection program, siRNA Scanner, automatically selects siRNA from the given RNA sequences. This program is developed in PERL (Practical Extract Report Language, 5.8.8.6 Build 820), uses a fuzzy logic-based system to calculate siRNA qualities and is accessible in a command line interface [26]. Another program, siRNA scanner ensures minimum user interaction and faster algorithm, thus, is fit for selecting siRNA for gene expression studies. The program also evaluates each possible siRNA candidate extensively and checks for all potential candidates from the given template sequence.

siDirect (http://design.RNAi.jp/) is an online software for computing highly effective siRNA sequences with target-specificity for RNAi in mammalian systems [27]. siDirect software prevents off-target effects that the most BLAST programs may not achieve. siRNAs for some common genes are easily available and are routinely used in experiments to silence genes. To avoid the off-target effects of siRNA constructs, they must be checked rigorously for being unique on non-redundant mRNA database (excluding splice variants). The other pre-requisites for the siRNA design are: Less than 30–60 percent

GC content or the sequences 4 As, 4 Ts, 4 Gs or 4 Cs are eliminated and with TT overhangs on the 3? ends of each strand [4, 6, 12, 14]. There are several websites that follow these criteria and allow the design of siRNAs.

Most commercial siRNA libraries incorporate chemical modifications to the siRNA seed region to help reduce off-target effects (OTEs) [27] but data analysis and experimental follow-up of screen results are essential to identify false-positive hits due to OTEs. For example, software tools that can be used to identify seed-dependent OTEs in RNAi datasets include Genome-wide Enrichment of Seed Sequence matches (GESS) and Haystack [27], which identify putative transcripts that are off-targets, and both produce good results on using larger (genome-wide) data sets.

Short hairpin RNAs or shRNAs are arti?cial constructs that can be inserted into a genome and expressed like endogenous genes [6, 7, 12, 14]. The primary transcripts fold back to form a dsRNA, mimicking native pre-RNA structure. This is directly accepted and used by the RISC to find target mRNA. The context of a known miRNA and siRNA-based designs can be used to construct these shRNAs. Several shRNAs libraries are now commercially available [21, 26, 28]. The libraries allow large-scale studies with siRNAs that are otherwise prohibitively expensive. To start with, all the shRNAs that can putatively silence a gene of choice are picked from the library and are infected into the cell. Following the cell growth for a few generations, the shRNAs that have depleted are investigated. Finally, the mRNA targets of the depleted shRNAs are sought, and further studies are planned.

There are several bioinformatics limitations associated with the use of shRNA. For a more accurate analysis, understanding outcomes of library screening are essential, which can be achieved by simulation of shRNA libraries when used in populations of cells [29]. A central repository of shRNA constructs like the RNAi Codex [29, 30], which can help track results, identify the performance patterns of shRNA facilitate in finding constructs from a variety of sources.

Also, microarray probes need to be designed for the more ef?cient study of shRNA expression in cell populations. Finally, techniques that can analyze results of silencing of genes either singly or in groups need to be developed.

## What the future holds

There has been a constant need to develop new algorithms and tools for designing efficient and specific siRNAs [21, 26, 31]. One few of the important on-going developments includes formulating algorithms to ?nd off-targets of siRNAs design for genome-wide studies. It is also essential to maintain a unified data source that presents optimal conditions as well as the design failures of siRNAs used for validation or experimentation. Easy access to such data library would de?nitely help in refining algorithms and would save much of effort that could otherwise go waste by designing non-functional siRNAs.

The information gained by studies on siRNA and miRNA designs can also be used to improvise on the other studies. Using siRNAs in experiments to mimic effects of drugs, a ?eld that exhibits immense potential because it could reduce the cost of large-scale screening.

Many programs have been developed to predict the target of miRNA and design siRNAs. These algorithms assess the sequence similarity between of miRNA and target RNA and the free energy change

that occurs during the interaction. For target predictions, the seed sequences are looked upon the untranslated region (UTR) of mRNA. In humans, most of the seed sequence are found to be located at the 3? end of the mRNA where it can bind and lead to either degradation of translation inhibition of the mRNA. In the list of web applications reviewed by Akhtar et al. [32], one can find a well-categorised list of tools and databases to study or different aspects of miRNA. These are for validation of miRNA (DIANA-TarBase, miRTarBase, miRecords) [33], correlating miRNA expression (StarBase, MiRonTop) metabolic pathways (EiMMO) regulator network (MiGator), transcription factor interactions (TransmiR) and prediction of cellular target of host and viral miRNA are few examples from the list.

Computational analysis can correct some experimental biases that may exist while screening the experimental data. For this purpose, Datta *et al*., [34] have developed a web-based platform named Comprehensive Analysis of RNAi Data (CARD) for integrated analysis and visualization of RNAi screen data. Other tools like R and Python code9 seed analysis predict off-target siRNAs by testing whether siRNAs with shared seed sequence have signi?cantly higher/lower scores compared with the rest [34]. The likely off-targeted genes can be identified by MATLAB code and web-service (http://www.?yrnai.org/gess/) [34]. RNAiCut Identi?es optimal threshold for hit-selection by integrating screen data with PPI networks [35]. Expression ?lter in the algorithm is used for the flagging and ?ltering of the genes based on their expression status.

A completely automated search program for siRNAs, inSiDE, contains a comprehensive set of rules for selecting of siRNA candidates (mentioned earlier in the “siRNAs” section) can be applied and allows for significantly high specificity [36]. After selecting potential siRNA candidates with the optimal functional properties, putative unspeci?c matches, which can cause cross-hybridization, are checked in the databases containing a unique entry for each gene. These truly non-redundant databases are constructed from the genome annotations (Ensembl) which is followed by examining other factors such as the presence of single nucleotide polymorphisms, intron/exon boundaries, transcript specificity, before the design of siRNAs [35, 36].

SPICE is another web-based tool that supports RNAi screening through comprehensive inquiry and automates several challenging tasks during evaluation of the short RNAs obtained by RNAi screening. This tool identifies the shRNA sequence in the input sequence, the target genes as well as the biological information in the database, and finally, prepares search result files that can be accessed locally. The system facilitates RNAi screening, which requires sequence analysis after screening [37].

In addition to carrying out statistical analyses and experimental follow-up studies of RNAi screen data, the application of various bioinformatics approaches can greatly aid in distinguishing between high-confidence and low-confidence screen hits. One common approach is to analyze the data in aggregate using a gene-set enrichment or related algorithms, such as using the DAVID database (the database for annotation, visualization, and integrated discovery) from the US National Institutes of Health, COMPLEAT and other software to detect gene ontology terms, protein complexes or signaling pathways that are enriched in the screen hits compared to controls [38]. These results can be used to confirm low confidence hits, such as hits with borderline statistical scores or hits for which not all RNAi reagents targeting the gene were positive and, conversely, to rule out hits that are the sole representatives of a category, which might suggest that they are false positives.

## Discussion

While the RNAi effector molecules function in the regulation of a variety of biological processes and can have an immense impact on various applications, much information about their mechanisms with respect to several physiological and regulatory functions still remain to be clarified. High-throughput methods have significantly advanced our knowledge of these small RNAs, and bioinformatics tools play a significant role in parallel to address the different aspects of RNAi research. Today, a vast range of tools are available for prediction of miRNA gene and target, network regulation and for siRNA design. While the tools can deliver to the best of their capacity, the predictions need to be linked and validated with experimental evidence.

**References**

1. Blencowe, B., et al., *Post-transcriptional gene regulation: RNA-protein interactions, RNA processing, mRNA stability and localization.* Pac Symp Biocomput, 2009: p. 545-8.
2. Xue, S. and M. Barna, *Cis-regulatory RNA elements that regulate specialized ribosome activity.* RNA Biol, 2015. **12**(10): p. 1083-7.
3. Acharya, J.K., et al., *Synaptic defects and compensatory regulation of inositol metabolism in inositol polyphosphate 1-phosphatase mutants.* Neuron, 1998. **20**(6): p. 1219-29.
4. Paroo, Z., Q. Liu, and X. Wang, *Biochemical mechanisms of the RNA-induced silencing complex.* Cell Res, 2007. **17**(3): p. 187-94.
5. Genschik, P., et al., *Selective protein degradation: a rheostat to modulate cell-cycle phase transitions.* J Exp Bot, 2014. **65**(10): p. 2603-15.
6. Fire, A., et al., *Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.* Nature, 1998. **391**(6669): p. 806-11.
7. Wilson, R.C. and J.A. Doudna, *Molecular mechanisms of RNA interference.* Annu Rev Biophys, 2013. **42**: p. 217-39.
8. Zhang, P., et al., *miR-600 inhibits cell proliferation, migration, and invasion by targeting p53 in mutant p53-expressing human colorectal cancer cell lines.* Oncol Lett, 2017. **13**(3): p. 1789-1796.
9. Shumin, G., D. Yanfei, and Z. Cheng, *Role of miRNA in plant seed development.* Yi Chuan, 2015. **37**(6): p. 554-60.
10. Willmann, M.R. and R.S. Poethig, *Conservation and evolution of miRNA regulatory programs in plant development.* Curr Opin Plant Biol, 2007. **10**(5): p. 503-11.
11. Darnell, D.K., et al., *MicroRNA expression during chick embryo development.* Dev Dyn, 2006. **235**(11): p. 3156-65.
12. Tomari, Y. and P.D. Zamore, *Perspective: machines for RNAi.* Genes Dev, 2005. **19**(5): p. 517-29.
13. Chou, C.H., et al., *A computational approach for identifying microRNA-target interactions using high-throughput CLIP and PAR-CLIP sequencing.* BMC Genomics, 2013. **14 Suppl 1**: p. S2.
14. Meister, G. and T. Tuschl, *Mechanisms of gene silencing by double-stranded RNA.* Nature, 2004. **431**(7006): p. 343-9.
15. Inui, M., G. Martello, and S. Piccolo, *MicroRNA control of signal transduction.* Nat Rev Mol Cell Biol, 2010. **11**(4): p. 252-63.
16. Lee, R.C., R.L. Feinbaum, and V. Ambros, *The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14.* Cell, 1993. **75**(5): p. 843-54.
17. Muerdter, F., et al., *A genome-wide RNAi screen draws a genetic framework for transposon control and primary piRNA biogenesis in Drosophila.* Mol Cell, 2013. **50**(5): p. 736-48.

18. Wienholds, E., et al., *MicroRNA expression in zebrafish embryonic development.* Science, 2005. **309**(5732): p. 310-1.
19. Gu, P., et al., *Novel microRNA candidates and miRNA-mRNA pairs in embryonic stem (ES) cells.* PLoS One, 2008. **3**(7): p. e2548.
20. Salim, A. and C.S. Vinod, *Computation Prediction of miroRNAs and their Targets.* Journal of Proteomics and Bioinformatics, 2014. **7**(7): p. 193-202.
21. Sachdanandam, R., *RNAi as a bioinformatics consumer.* Briefings in Bioinformatics, 2005. **6**(2): p. 146-162.
22. Rennie, W., et al., *STarMir: a web server for prediction of microRNA binding sites.* Nucleic Acids Res, 2014. **42**(Web Server issue): p. W114-8.
23. Lewis, B.P., C.B. Burge, and D.P. Bartel, *Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets.* Cell, 2005. **120**(1): p. 15-20.
24. Lu, T.P., et al., *miRSystem: an integrated system for characterizing enriched functions and pathways of microRNA targets.* PLoS One, 2012. **7**(8): p. e42390.
25. Pinzon, N., et al., *microRNA target prediction programs predict many false positives.* Genome Res, 2017. **27**(2): p. 234-245.
26. Vijayakumar S and P. S., *siRNA Scanner – A Fuzzy Logic Based Tool for Small Interference RNA Design.* J Proteomics Bioinform, 2008. **1**: p. 154-160.
27. Yilmazel, B., et al., *Online GESS: prediction of miRNA-like off-target effects in large-scale RNAi screen data by seed region analysis.* BMC Bioinformatics, 2014. **15**: p. 192.
28. Das, M.K. and H.K. Dai, *A survey of DNA motif finding algorithms.* BMC Bioinformatics, 2007. **8 Suppl 7**: p. S21.
29. Silva, J.M., et al., *Second-generation shRNA libraries covering the mouse and human genomes.* Nat Genet, 2005. **37**(11): p. 1281-8.
30. Sigoillot, F.D., et al., *A bioinformatics method identifies prominent off-targeted transcripts in RNAi screens.* Nat Methods, 2012. **9**(4): p. 363-6.
31. Olson, A., et al., *RNAi Codex: a portal/database for short-hairpin RNA (shRNA) gene-silencing constructs.* Nucleic Acids Res, 2006. **34**(Database issue): p. D153-7.
32. Akhtar, M.M., et al., *Bioinformatic tools for microRNA dissection.* Nucleic Acids Res, 2016. **44** (1): p. 24-44.
33. Papadopoulos, G.L., et al., *DIANA-mirPath: Integrating human and mouse microRNAs in pathways.* Bioinformatics, 2009. **25**(15): p. 1991-3.
34. Dutta, B., et al., *An interactive web-based application for Comprehensive Analysis of RNAi-screen Data.* Nat Commun, 2016. **7**: p. 10578.
35. Kaplow, I.M., et al., *RNAiCut: automated detection of significant genes from functional genomic screens.* Nat Methods, 2009. **6**(7): p. 476-7.
36. Santoyo, J., J.M. Vaquerizas, and J. Dopazo, *Highly specific and accurate selection of siRNAs for high-throughput functional assays.* Bioinformatics, 2005. **21**(8): p. 1376-82.
37. Kamatuka, K., M. Hattori, and T. Sugiyama, *shRNA target prediction informed by comprehensive enquiry (SPICE): a supporting system for high-throughput screening of shRNA library.* EURASIP J Bioinform Syst Biol, 2016. **2016**(1): p. 7.
38. Mohr, S.E., et al., *RNAi screening comes of age: improved techniques and complementary approaches.* Nat Rev Mol Cell Biol, 2014. **15**(9): p. 591-600.

## Sharing is caring. Spread the love!

- Print
- Email
- LinkedIn
- Twitter
- Facebook
- Google
-