doi:10.1093/nar/gkr1007

doRiNA: a database of RNA interactions in post-transcriptional regulation

Gerd Anders[1], Sebastian D. Mackowiak[2], Marvin Jens[2], Jonas Maaskola[2], Andreas Kuntzagk[1], Nikolaus Rajewsky[2,*], Markus Landthaler[3,*] and Christoph Dieterich[1,*]

[1]Bioinformatics in Quantitative Biology, [2]Systems Biology of Gene Regulatory Elements and [3]RNA Biology and post-transcriptional regulation, Berlin Institute for Medical Systems Biology, Max Delbrück Centre for Molecular Medicine, Robert-Rössle-Straße 10, 13125 Berlin, Germany

Received August 29, 2011; Revised October 7, 2011; Accepted October 19, 2011

ABSTRACT

In animals, RNA binding proteins (RBPs) and microRNAs (miRNAs) post-transcriptionally regulate the expression of virtually all genes by binding to RNA. Recent advances in experimental and computational methods facilitate transcriptome-wide mapping of these interactions. It is thought that the combinatorial action of RBPs and miRNAs on target mRNAs form a post-transcriptional regulatory code. We provide a database that supports the quest for deciphering this regulatory code. Within doRiNA, we are systematically curating, storing and integrating binding site data for RBPs and miRNAs. Users are free to take a target (mRNA) or regulator (RBP and/or miRNA) centric view on the data. We have implemented a database framework with short query response times for complex searches (e.g. asking for all targets of a particular combination of regulators). All search results can be browsed, inspected and analyzed in conjunction with a huge selection of other genome-wide data, because our database is directly linked to a local copy of the UCSC genome browser. At the time of writing, doRiNA encompasses RBP data for the human, mouse and worm genomes. For computational miRNA target site predictions, we provide an update of PicTar predictions.

INTRODUCTION

The regulation of gene activity on the RNA level has been at the heart of intensive research efforts since the description of the operon (1). Post-transcriptional regulation is highly versatile and adaptable by controlling RNA availability in cellular time and space. Messenger RNA stability, transport, storage and translation are largely determined by the interaction of mRNA with microRNAs (miRNAs) and RNA-binding proteins (RBPs). We have just begun to understand the extent and dynamics of transcriptome-wide binding events that lead to the temporal formation of functional ribonucleoprotein complexes.

Within doRiNA, we focus on two key players of post-transcriptional regulation: miRNAs and RBPs.

microRNAs

miRNAs originate from long stem–loop containing primary transcripts (pri-miRNAs) that are generally transcribed by RNA Polymerase II. pri-miRNAs are substrates of the RNAse III enzyme Drosha and its binding partner, the dsRNA-binding protein DGCR8/Pasha. In the nucleus, a complex of Drosha and DGCR8 cleaves pri-miRNAs into ~70 nt precursor hairpins (pre-miRNA), which are exported to the cytoplasm. In the cytoplasm, the pre-miRNA is further cleaved by another RNAse III enzyme Dicer into a mature miRNA and its partner strand, the miRNA* (microRNA star). The mature miRNA is defined as the strand, which is loaded into the RNA-Induced Silencing Complex (RISC) complex. Krol *et al.* (2) give an excellent overview on miRNA biogenesis.

The mature miRNA identifies its mRNA target by binding to partially complementary sites within 3′UTRs (3), resulting in mRNA degradation and translational repression of the RNA target (4). This drastically differs from the short-interfering RNA mechanism, which requires perfect complementarity, and leads to RNA-directed cleavage of the target transcript.

*To whom correspondence should be addressed. Tel: + 49 30 9406 4235; Fax: 49 30 9406 3068; Email: christoph.dieterich@mdc-berlin.de
Correspondence may also be addressed to Markus Landthaler. Tel: +49 30 9406 3026; Fax: +49 30 9406 3068; Email: markus.landthaler@mdc-berlin.de
Correspondence may also be addressed to Nikolaus Rajewsky. Tel: +49 30 9406 2999; Fax: +49 30 9406 3068; Email: rajewsky@mdc-berlin.de

The doRiNA database offers computational miRNA target site predictions for man, mouse and worm. These predictions constitute the long awaited update of PicTar predictions (5–7).

RNA-binding proteins

Nascent RNAs are co-transcriptionally bound by RBPs leading to the formation of ribonucleoprotein complexes. RBPs are characterized by containing one or multiple RNA recognition domains (RBDs), which cooperate to recognize RNA sequences (8). RBPs do not only recognize simple RNA sequence motifs, but can also integrate the structural context into the recognition process. This becomes evident for the simple case of double-strand binding as opposed to single-strand binding RBPs.

The *in silico* prediction of RBP target sites is still in its infancy (9). That is why we decided to exclude computational predictions and constrain our data set to RBP target sites from high-resolution, transcriptome-wide cross-linking and immunoprecipitation (CLIP) experiments (10). One variant of CLIP, called PAR-CLIP (photoactivatable-ribonucleoside enhanced cross-linking and immunoprecipitation), relies on the incorporation of photo-reactive nucleotide analoga into newly synthesized RNA (11). Successful incorporation and cross-linking induces characteristic base substitutions in the sequenced cDNA reads. These base substitutions support target site identification at nucleotide-level resolution. For example, an incorporation of 4-thiouridine into RNA and subsequent cross-linking yields characteristic T → C base transitions in sequencing reads.

RBPs binding sites from our own experiments are all based on the PAR-CLIP method and were processed in the same way (see 'Materials and Methods' section for details). Additionally, we collect and integrate target sites from published HITS-CLIP experiments (12) and other variants into doRiNA as long as they provide precise positional target site information.

Entering doRiNA

The doRiNA database integrates miRNA and RBP target site sets from different species into one framework. We have mainly turned our attention to service availability, query speed and query capability. Service availability is achieved by mirroring the web and database servers (Figure 1). We enable high query speed and complexity by pre-computing several important data characteristics. In doRiNA, users are able to enter the available post-transcriptional regulatory network from a target centric ('Which regulators target gene X?') or regulator centric ('Which genes are regulated by Y?') view. Complex queries using set operations over subregions of genes (e.g. 3′ UTR or 5′ UTR) have been realized without compromising speed. We deem doRiNA a one-stop solution to transcriptome-wide mining of regulatory interactions in post-transcriptional gene regulation.

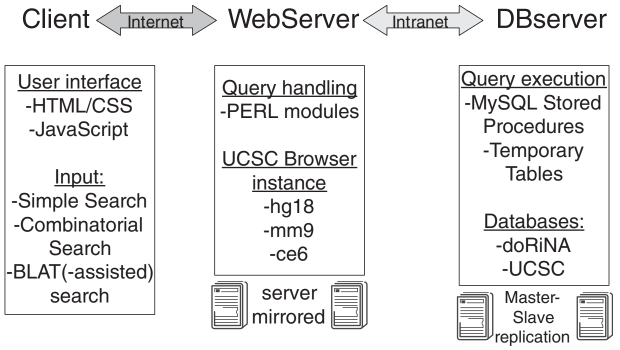


Figure 1. Schematic doRiNA overview. The doRiNA concept is implemented as a linear chain of three: client, web server and database server. We mainly turned our attention to service availability, query speed and query capability.

MATERIALS AND METHODS

The doRiNA infrastructure

To ensure short query response times, we have setup a powerful 12-core web server, which is coupled to a dedicated 8-core MySQL 5.1 database server. A local installation of the UCSC genome browser (13) was directly placed onto the web server. The database server handles requests from the doRiNA user interface as well as from the local UCSC browser installation. Result sets are returned in tabular form (for browsing or download) via the web server and are depicted within the genome browser on a locus-by-locus basis. An overview on the infrastructure is given in Figure 1.

Web server. On the client side, the doRiNA web interface is implemented as a blend of HTML, CSS and Javascript components. The Javascript component utilizes the JQuery library (http://jquery.com/) and the JSON data interchange format (http://www.json.org/). The doRiNA web service is divided into individual species sections (man, mouse and worm). Each species section has a tripartite structure: Simple Search, Combinatorial Search and BLAT(-assisted) search. The user is assisted in formulating a query by several features: autocompletion of gene names, gene name validity checks, conditional activation of web interface control elements and extensive help documents. Search requests can only be submitted if the user input has passed these initial checks. Search requests are processed on the server side by custom PERL modules and scripts. The PERL layer mainly interfaces with the database engine where the actual queries are executed by MySQL Stored Procedures.

Database server. The doRiNA database is built on top of the UCSC genome browser databases. To this end, we have added custom tables to species-specific databases (e.g. hg18). These tables contain precomputed information (e.g. host gene) for each target site to speed up queries. We have put the main work horse of doRiNA, MySQL Stored Procedures in combination with temporary tables, into a separate database (Figure 1). Search requests trigger the execution of cascading stored procedures, which assemble result tables in memory, send them back via the PERL layer to the client and subsequently discard the temporary tables. Concurrent user access is guaranteed by a database inherent session management.

Target sites of RNA-binding proteins

One central question has not been addressed yet: how do we collect and integrate target site information? For RBPs, we follow a 2-fold strategy: first, PAR-CLIP data sets from Hafner *et al.* (11) and data sets that were produced in-house (i.e. by one of the co-authors) are subject to a processing pipeline (PCP, details see below). This pipeline infers target sites based on a nucleotide conversion score and an entropy measure over read stacks in continuously covered transcript regions (read clusters).

Second, other CLIP data sets (HITS-CLIP, PAR-CLIP, iCLIP and variants thereof) are retrieved from external publications and integrated as is. Interested users find details on data acquisition and processing in the corresponding UCSC genome browser track descriptions.

Analysis pipeline for in-house PAR-CLIP data. All in-house PAR-CLIP tracks were produced with our computational pipeline to determine RBP binding sites at an estimated 5% false positive rate (14). The pipeline performs all steps of the PAR-CLIP analysis taking raw reads and producing cluster sets and lists of target genes, in a largely automated and unbiased way. The emphasis is on stringent filtering and controlling the false positive rate in the identification of binding sites.

Briefly, PAR-CLIP reads are aligned to the human transcriptome (mRNAs or pre-mRNAs) or genome (user choice), allowing for up to one mismatch, insertion or deletion. Only uniquely mapping reads are retained.

Next, we identify clusters of aligned PAR-CLIP reads continuously covering regions of reference sequence and assign two quality scores based on the characteristics of the PAR-CLIP protocol. Efficient cross-linking leads to specific nucleotide conversion events during reverse transcription and next-generation sequencing of RNA from each experiment: cross-linked 4-thiouridine (4SU) and 6-thioguanosine (6SG) residues are converted into C and A, respectively. These conversions mark the RBP binding site on the target RNA (11). The number of these mismatches therefore serves as a cross-link score. The other score addresses problems that may be encountered in a sequencing-based assay: we assign an entropy score based on the number and positions of distinct reads contributing to the cluster to guard against PCR or mapping artifacts.

Finally, the pipeline automatically selects cutoffs on both quality scores by using the reverse complement of the annotated transcripts as a decoy. As PAR-CLIP reads should originate from RBP-bound transcripts, we may regard clusters aligning antisense to the annotated direction of transcription as false positives. We are thus able to select cutoffs on the estimated false positive rate. After filtering by these cutoffs, remaining antisense clusters are dropped. We expect each retained cluster to harbor at least one RBP binding site with a false positive probability $\leq 5\%$.

Additional details can be found in Supplementary data.

External target site data. We have collected several published CLIP data sets from the literature (see web site for details). Target sites were either extracted from Supplementary data or obtained from the corresponding author. Some authors did not assign a score to each target site, which does not allow a score-base ranking of these sites. In that case, target sites are assigned a default score and the rank is set to N/A.

PicTar miRNA target site predictions

We have updated the PicTar miRNA target site predictions to the respective UCSC genome releases of man (hg18), mouse (mm9) and worm (ce6). PicTar 2.0 (7) predicts miRNA target sites in 3′ UTRs and utilizes multiple genome sequence alignments to boost its precision. Briefly, all 3′ UTR alignments for a given species set are scanned for perfect and imperfect seed sequences. Perfect seeds consist of a 7 nt perfect match starting at position 1 or 2 from the 5′-end of a mature miRNA. Imperfect seeds contain one insertion/deletion or mismatches to the 3′ UTR sequence. All candidate sites are subject to probabilistic scoring by an Hidden Markov Model (HMM).

For example, human miRNA targets for mature and star sequences from Mirbase v16 were predicted based on UCSC's 44-way Vertebrate Genome alignment. We have incorporated three conservation levels for human target sites into doRiNA: (i) Mammals, chicken and fish—seed conservation across *Pan troglodytes*, *Mus musculus*, *Rattus norvegicus*, *Canis lupus*, *Gallus gallus*, *Fugu rubripes* and *Danio rerio*. (ii) Mammals, chicken—seed conservation is not required in *Fugu rubripes* and *Danio rerio*. (iii) Mammals—seed conservation is not required in *Gallus gallus*, *Fugu rubripes* and *Danio rerio*.

These conservation levels provide a convenient way to choose the optimal sensitivity level while controlling for false positives.

Integration with the UCSC Genome Browser

All target site information for miRNAs or RBPs are integrated into our local installation of the UCSC genome browser as additional local tracks. This guarantees full access to all genome browser features and simultaneous availability of other genome browser tracks (Variation, Regulation and other tracks). In addition, the genome browser interface is commonly used by biologists world-wide and does not require any additional training.

EXAMPLE APPLICATIONS

In the following section, we will present a few example applications of doRiNA. These examples serve as an entry point to doRiNA and outline three main use cases.

(1) Target centric queries—retrieve all regulators of a predefined gene set.
(2) Regulator centric queries—retrieve all genes that are targeted by a predefined set of regulators.
(3) Complex / Combinatorial queries—set operations on regulator target gene sets.

Target centric queries

Several questions in biology focus on a particular gene or gene set of interest. Frequently, questions like 'which regulators target my gene or gene set of interest?' arise in scientific discussions. We denote these kind of queries as 'target centric'.

Setting up the query. Generally, doRiNA accepts gene symbols and NCBI RefSeq identifiers to define target gene sets. The Simple Search Function (Figure 2A), which is used in this context, offers two different approaches to compile candidate gene lists. The user could either manually define a subset of genes/transcripts (Option 2 in Step 1) or upload a list of gene identifiers (Option 3 in Step 1). For completeness, Option 1 selects the complete available gene set in the corresponding species databases. By using one of these options, the user defines a gene set of interest and subsequently selects post-transcriptional regulators (RBPs and/or miRNAs) to match against (Step 2). All available regulators are conveniently selected by the 'All RBPs in database' and 'All miRNAs in database' records. A score-based ranking cutoff for RBP target sites is finally set in the last step (Step 3) of the user interface. The search submission button becomes activated if all input passes the online syntax checks.

Interpretation of results. Search results are reported back in tabular format. A summary on the number of found target sites and genes is shown at the top of the results page. Each table row corresponds to one target site and contains information in a self-explanatory format. Please note that each column can be used to sort the entire tables. If target site scores are provided for a CLIP experiment, we use them to order the table output via the column (top-percent value). Otherwise, the score is set to a default value. Each row offers links to the UCSC genome browser for the entire gene locus (gene symbol location) or the corresponding target site (target site location).

Example: the CDKN1 gene family. Let us assume that we are interested in RBPs as post-transcriptional regulators of the 'Cip/Kip' family, which encompasses cyclin-dependent kinase inhibitor 1 coding genes (CDKN1). Intriguingly, Kedde *et al*. (15) have shown that p27 (CDKN1B) is post-transcriptionally regulated by mir-221 and mir-222 conditional on an Pumillio-induced RNA structure switch. We already know that there are only three gene family members (CDKN1 A to C). That is why we enter these three gene names manually via option 2 in Step 1 of the Simple Search Tab. We are assisted by the autocompletion function of doRiNA. Since we are interested in any regulator of at least one of the three CDKN1s, we select all RBPs and all miRNAs in database in Step 2. We increase search sensitivity to the maximal level by setting the RBP score rank percentile cutoff to 100%. All other settings are left untouched (default values).

The result page summarizes all query results: there are 209 target sites in total of which 194 are RBP target sites and 15 are conserved miRNA target sites (mammals–chicken–fish). Indeed, two mir-221/222 sites and three PUM2 sites have been reported for CDKN1B in the results table. We navigate to the corresponding UCSC view by clicking on one CDKN1B location link, which opens up an UCSC genome browser view that nicely recapitulates the published target site configuration (Figure 2B).

Users may also inspect individual PAR-CLIP target sites at nucleotide-level resolution by clicking on them. This opens a summary page, which links to an in-depth read cluster display where characteristic mutations (e.g. T → C) are indicated in the corresponding sequencing reads.

Regulator centric queries

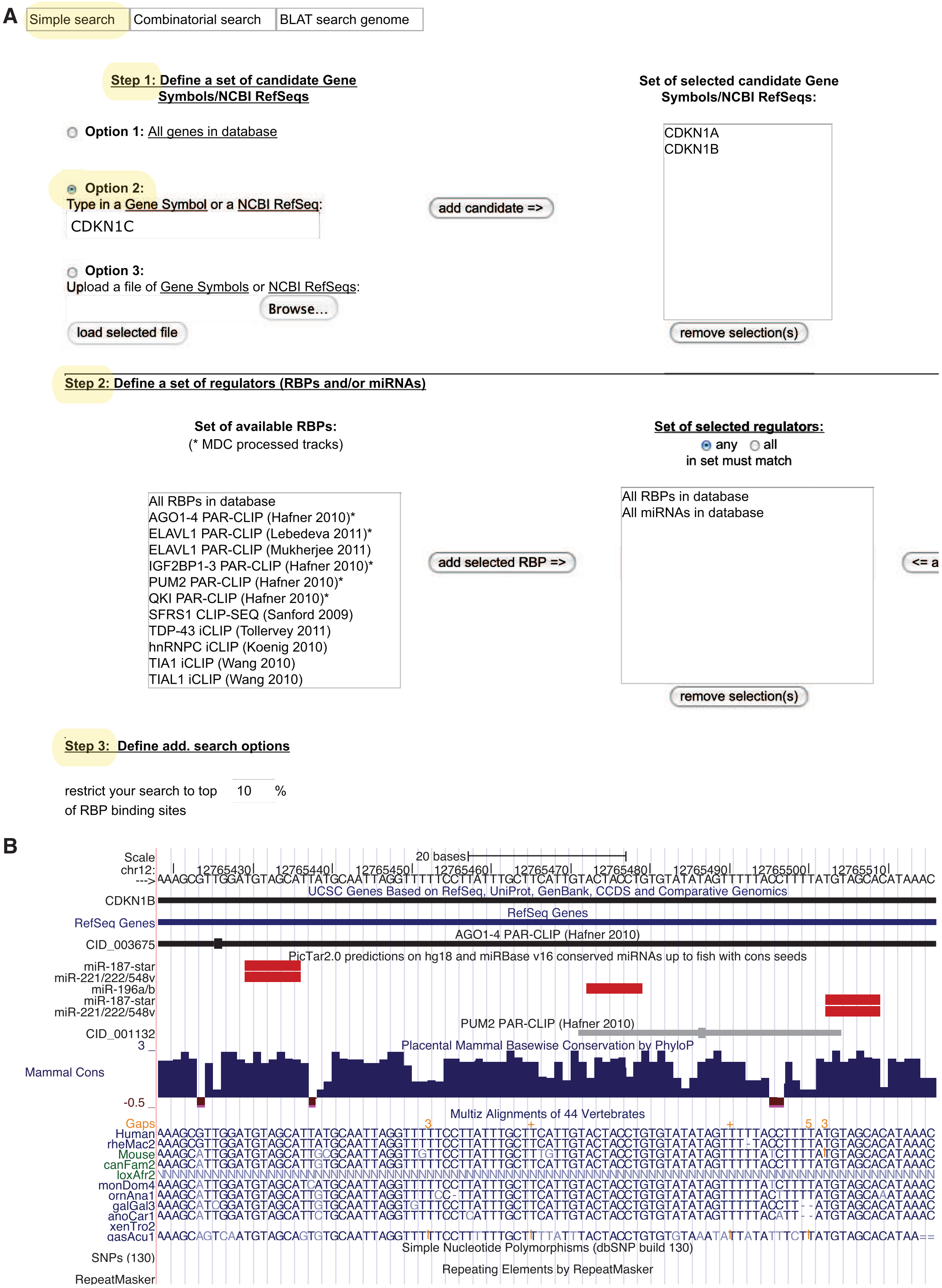
In a different application context, scientists frequently have to define the target gene set of a particular regulator or set of regulators. More specifically, one could be interested in either genes that are co-targeted by all selected regulators (intersection of target sets) or just at least one of the selected regulators (union of target sets). We refer to this view as regulator-centric. The difference in search strategy is mainly to leave the target gene set unconstrained (Option 1) and select a confined set of regulators. The set of regulators is conveniently defined from two available lists: one for RBPs and one for miRNAs. The simple search function provides two set operations (all ≡ intersection and any ≡ union) on the selected list of regulators. A radio button toggles between the intersection and union modes.

We will continue with our previous example of the PUM2 & mir-221/222 module and search for all its co-targets.

Example: co-targets of PUM2 and miR-221 / miR-222. We retrieve the requested co-targets by selecting the entire gene set in the database (Option 1). The regulator set is constrained to PUM2 and mir-221/222 in step 2. Since we are looking for co-targets, we switch to the intersection mode by choosing the 'all' radio button. Subsequently, we set the RBP score rank percentile cutoff to 100% and leave all other settings untouched. Our query returns 13 target genes (data not shown, one is CDKN1B). We repeat this search with relaxed miRNA conservation criteria and either obtain 60 co-target genes for mammal–chicken conserved miRNA sites or 141 co-targets for mammal-only conserved sites, respectively.

Combinatorial search options

The power of doRiNA becomes eminent in the case of combinatorial search option. This option differs from



Step 1: Define a set of candidate Gene Symbols/NCBI RefSeqs

○ **Option 1:** All genes in database

● **Option 2:** Type in a Gene Symbol or a NCBI RefSeq: CDKN1C

○ **Option 3:** Upload a file of Gene Symbols or NCBI RefSeqs:

Set of selected candidate Gene Symbols/NCBI RefSeqs:

CDKN1A
CDKN1B

Step 2: Define a set of regulators (RBPs and/or miRNAs)

Set of available RBPs:
(* MDC processed tracks)

All RBPs in database
AGO1-4 PAR-CLIP (Hafner 2010)*
ELAVL1 PAR-CLIP (Lebedeva 2011)*
ELAVL1 PAR-CLIP (Mukherjee 2011)
IGF2BP1-3 PAR-CLIP (Hafner 2010)*
PUM2 PAR-CLIP (Hafner 2010)*
QKI PAR-CLIP (Hafner 2010)*
SFRS1 CLIP-SEQ (Sanford 2009)
TDP-43 iCLIP (Tollervey 2011)
hnRNPC iCLIP (Koenig 2010)
TIA1 iCLIP (Wang 2010)
TIAL1 iCLIP (Wang 2010)

Set of selected regulators:
● any ○ all
in set must match

All RBPs in database
All miRNAs in database

Step 3: Define add. search options

restrict your search to top 10 %
of RBP binding sites

B
Scale
20 bases
chr12:
12765430
12765440
12765450
12765460
12765470
12765480
12765490
12765500
12765510
--->
AAAGCGTTGGATGTAGCATTATGCAATTAGGTTTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACCTTTTATGTAGCACATAAAC
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
CDKN1B
RefSeq Genes
RefSeq Genes
AGO1-4 PAR-CLIP (Hafner 2010)
CID_003675
PicTar2.0 predictions on hg18 and miRBase v16 conserved miRNAs up to fish with cons seeds
miR-187-star
miR-221/222/548v
miR-196a/b
miR-187-star
miR-221/222/548v
PUM2 PAR-CLIP (Hafner 2010)
CID_001132
Placental Mammal Basewise Conservation by PhyloP
3
Mammal Cons
-0.5
Multiz Alignments of 44 Vertebrates
Gaps
3
5
3
Human
rheMac2
Mouse
canFam2
loxAfr2
monDom4
ornAna1
galGal3
anoCar1
xenTro2
gasAcu1
Simple Nucleotide Polymorphisms (dbSNP build 130)
SNPs (130)
Repeating Elements by RepeatMasker
RepeatMasker

```
Human    AAAGCGTTGGATGTAGCATTATGCAATTAGGTTTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACCTTTTATGTAGCACATAAAC
rheMac2  AAAGCGTTGGATGTAGCATTATGCAATTAGGTTTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTT-TACCTTTTATGTAGCACATAAAC
Mouse    AAAGCATTGGATGTAGCATTGCGCAATTAGGTTGTTCCTTATTTGCTTTGTTGTACTACCTGTGTATATAGTTTTTATCTTTTATTGTAGCACATAAAC
canFam2  AAAGCATTGGATGTAGCATTGTGCAATTAGGTTTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACCTTTTATGTAGCACATAAAC
loxAfr2  NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
monDom4  AAAGCATTGGATGTAGCATCATGCAATTAGGTTTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTATCTTTTATGTAGCACATAAAC
ornAna1  AAAGCATTGGATGTAGCATTGTGCAATTAGGTTTTCC-TTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACTTTTTATGTAGCAAATAAAC
galGal3  AAAGCATCGGATGTAGCATTGTGCAATTAGGTGTTTCCTTATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACCTT--ATGTAGCACATAAAC
anoCar1  AAAGCATTGGATGTAGCATTCTGCAATTAGGTTTTTCCTCATTTGCTTCATTGTACTACCTGTGTATATAGTTTTTACATT--ATGTAGCACATAAAC
xenTro2
gasAcu1  AAAGCAGTCAATGTAGCAGTGTGCAATTAGGTTTTTCCTTTTTTGCTTTTATTTACTACCTGTGTGTAAATATTTATATTTCTTTATGTAGCACATAA==
```

Figure 2. Example of a target centric query — the 'Cip/Kip' family. (**A**) Web interface — Simple Search Tab. (**B**) Visualization of the 3′ UTR of CDKN1B.

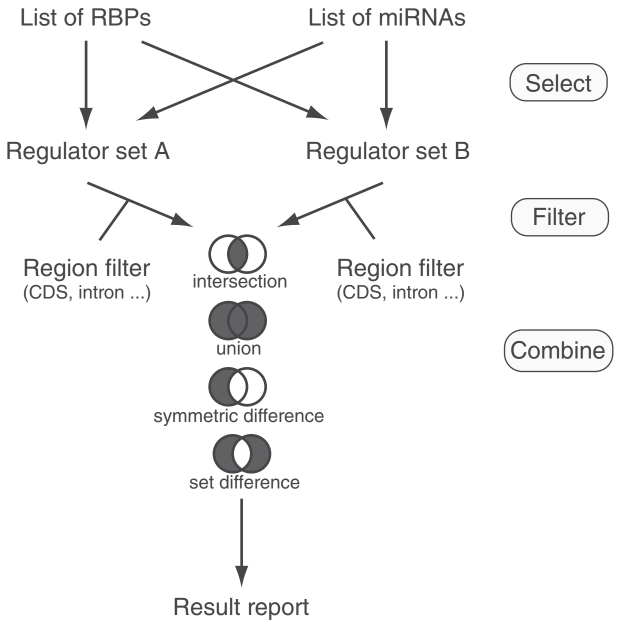


Figure 3. Combinatorial search options.

the aforementioned simple search. It combines the results from two independent simple searches (A and B, Figure 3). Initially, target site positions can be individually confined for sets A and B to a particular gene feature region (CDS, 5′ UTR, 3′ UTR, intron or intergenic). The two filtered sets are subsequently combined by four possible set operations: union, $A \cup B'$ intersection, $A \cap B'$ symmetric difference, $A \Delta B$; and set difference, $A \backslash B$. Please bear in mind that both, the A set and the B set, are themselves the outcome of either a union or intersection step. Figure 3 summarizes the query capabilities of the combinatorial search option.

DISCUSSION

A better understanding and dissection of post-transcriptional regulation is of paramount importance to molecular biology. With the advent of robust high-throughput methods for target site delineation, either computationally or experimentally, we face the challenge of efficient data organization, representation and analysis. doRiNA is our contribution to meet this challenge by providing a biologist-friendly access to the available target site data for miRNA and RBP regulators. Within doRiNA, we consolidate three different needs in data mining: data exploration, querying and retrieval.

We discern three features of doRiNA as especially important: first, the doRiNA database unifies two protagonists of post-transcriptional regulation, RBPs and miRNAs, in one service. Second, the doRiNA web service provides unparalleled query capabilities with minimal response times. Finally, users benefit from doRiNA's integration with other genome-wide data via the UCSC browser (e.g. SNP data could be intersected with miRNA or RBP target sites).

Comparison to related work

The doRiNA database differs from previously published database solutions like starBase (16) and CLIPZ (17) in several aspects. The starBase database is a very data-rich resource but offers only limited query capabilities (e.g. complex set operations are not supported). This is the same for the CLIPZ database, which has been mainly designed as a service for collaborative CLIP data analysis. Moreover, doRiNA contained more genome-wide data sets at the time of writing and is linked to the UCSC genome browser.

CONCLUSIONS

doRiNA does not merely provide rich data sets for browsing and download but empowers users to flexibly specify hypothesis-driven queries. Users may freely define their target site search space by providing gene lists. Complex combinations of regulators may be submitted as search queries. Without loss of speed, doRiNA is able to operate on different data zoom levels ranging from target gene sets down to individual target site nucleotides.

doRiNA benefits from its seamless integration with a local copy of the UCSC browser, which is very popular among computer-affine biologists.

AVAILABILITY

The doRiNA database is freely available at http://dorina.mdc-berlin.de. There are no access restrictions for academic and commercial use. We kindly ask all users to cite the doRiNA manuscript if they employ search results in their publications.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Methods.

ACKNOWLEDGEMENTS

All authors wish to acknowledge fruitful discussions with members of the Berlin Institute for Medical Systems Biology.

FUNDING

MDC Systems Biology Network (MSBN) as a participant of the Helmholtz-Alliance on Systems Biology (to S.D.M.); Deutsche Forschungsgemeinschaft for a fellowship in the International Research Training Group Genomics and Systems Biology of Molecular Networks (GRK 1360 to M.J.). Federal Ministry for Education and Research (BMBF) and the Senate of Berlin, Berlin, Germany (to M.L. and C.D.). Funding for open access charge: MDC.

Conflict of interest statement. None declared.

REFERENCES

1. Rajewsky,N. (2011) MicroRNAs and the Operon paper. *J. Mol. Biol.*, **409**, 70–75.
2. Krol,J., Loedige,I. and Filipowicz,W. (2010) The widespread regulation of microRNA biogenesis, function and decay. *Nat. Rev. Genet.*, **11**, 597–610.
3. Bartel,D.P. (2009) MicroRNAs: target recognition and regulatory functions. *Cell*, **136**, 215–233.
4. Filipowicz,W., Bhattacharyya,S.N. and Sonenberg,N. (2008) Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat. Rev. Genet.*, **9**, 102–114.
5. Krek,A., Grün,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.* (2005) Combinatorial microRNA target predictions. *Net. Genet.*, **37**, 495–500.
6. Grün,D., Wang,Y.L., Langenberger,D., Gunsalus,K.C. and Rajewsky,N. (2005) microRNA target predictions across seven Drosophila species and comparison to mammalian targets. *PLoS Comput. Biol.*, **1**, e13.
7. Lall,S., Grün,D., Krek,A., Chen,K., Wang,Y.L., Dewey,C.N., Sood,P., Colombo,T., Bray,N., Macmenamin,P. *et al.* (2006) A genome-wide map of conserved microRNA targets in C. elegans. *Curr. Biol.*, **16**, 460–471.
8. Lunde,B.M., Moore,C. and Varani,G. (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.
9. Li,X., Quon,G., Lipshitz,H.D. and Morris,Q. (2010) Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, **16**, 1096–1107.
10. Jensen,K.B. and Darnell,R.B. (2008) CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. *Methods Mol. Biol.*, **488**, 85–98.
11. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
12. Licatalosi,D.D., Mele,A., Fak,J.J., Ule,J., Kayikci,M., Chi,S.W., Clark,T.A., Schweitzer,A.C., Blume,J.E. and Wang,X. (2008) HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, **456**, 464–469.
13. Fujita,P.A., Rhead,B., Zweig,A.S., Hinrichs,A.S., Karolchik,D., Cline,M.S., Goldman,M., Barber,G.P., Clawson,H., Coelho,A. *et al.* (2011) The UCSC Genome Browser database: update 2011. *Nucleic Acids Res.*, **39**, D876–D882.
14. Lebedeva,S., Jens,M., Theil,K., Schwanhäusser,B., Selbach,M., Landthaler,M. and Rajewsky,N. (2011) Transcriptome-wide analysis of regulatory interactions of the RNA-binding protein HuR. *Mol. Cell*, **43**, 340–352.
15. Kedde,M., van Kouwenhove,M., Zwart,W., Oude Vrielink,J.A., Elkon,R. and Agami,R. (2010) A Pumilio-induced RNA structure switch in p27-3′ UTR controls miR-221 and miR-222 accessibility. *Nat. Cell Biol.*, **12**, 1014–1020.
16. Yang,J.H., Li,J.H., Shao,P., Zhou,H., Chen,Y.Q. and Qu,L.H. (2011) starBase: a database for exploring microRNA-mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. *Nucleic Acids Res.*, **39**, D202–D209.
17. Khorshid,M., Rodak,C. and Zavolan,M. (2011) CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Res.*, **39**, D245–D252.

PAR-CLIP computational pipeline

All PAR-CLIP cluster sets, which are flagged with an asterix(*) on the web site, were generated with our own computational pipeline, essentially as described in (Lebedeva et al. 2011).

Read preparation and mapping

Solexa sequenced reads are subjected to adapter-removal by gap-end free overlap alignment with FAR 1.81 (http://sourceforge.net/projects/theflexibleadap/ unpublished). RefSeq gene models (Kent, 2002; Pruitt et al., 2005) were obtained from the UCSC genome data base hg18 (Kent et al., 2002; Fujita et al., 2011) on 09/29/2010 and used to prepare spliced mRNA sequences (RS) and unspliced pre-mRNA (preRS) sequences. The human genome sequence assembly hg18 was obtained via UCSC GoldenPath. Representative human ribosomal RNA sequences were obtained from NCBI (NR_003286.2, NR_023363.1, NR_003285.2, NR_003287.2) and added to the set of reference sequences to monitor the behavior of highly abundant and repetitious (pseudo gene) sequences. The reference sequences thus obtained are:

hg18_RS (spliced RefSeq mRNA + rRNA)

hg18_preRS (unspliced RefSeq mRNA + rRNA)

hg18 (genome + rRNA)

bwa 0.5.8c (Li & Durbin, 2009) was used to compile indices and map each of the libraries. A custom script was used to filter the results for maximum edit-distance and uniqueness (relaxed to allow for isoforms) and furthermore to transform the coordinates back to the hg18 coordinate system. Reads aligning antisense to transcripts were deliberately kept at this step to serve as an estimate for false positives later. Unique alignments in genomic coordinates are collected into a pileup file using SAMtools 0.1.8 (Li & Durbin, 2009).

Identification of binding sites and crosslink positions

A set of custom analysis scripts is run on the pileup output to identify read clusters as stretches of continuous read coverage on one strand.

Clusters are discarded if they overlap with repeatMask 3.2.7 (Smit et al., 1996) elements (obtained via UCSC). A cluster of length L is internally represented as a Lx6 matrix with the aggregate number of observed A,C,T,G nucleotides, deletions and insertions at each position. Together with the reference sequence for the cluster this allows to efficiently compute two different quality measures:

- the number of characteristic nucleotide conversion events (T to C for 4SU and G to A for 6SG respectively)
- an information entropy score of the nucleotide observations and read start and end positions

Whereas the nucleotide conversions represent the canonical signature of crosslink events (Hafner et al., 2010), the purpose of the entropy score is to reduce the impact of potential read amplification and mapping artifacts: identical reads aligning in the same positions contribute zero entropy. On the other hand, differing nucleotide observations in overlapping reads and variable start and end positions within a cluster must derive from independent reads and alignments, increasing confidence in the quality of the cluster.

Clusters aligning antisense to known transcripts likely represent false positives due to amplification and mapping artifacts. Therefore, cutoffs on the entropy score and the number of nucleotide conversions were chosen such that the ratio of antisense to sense aligning clusters does not exceed 5%, while retaining the maximal number of sense aligning clusters.

After the cutoff estimation antisense clusters are dropped so that we confidently estimate the ratio of false positives to true positives among the remaining, sense-aligning clusters to be less or equal 5%. We observed that randomly picked high quality antisense clusters often reside in areas of convergent transcription and RNA-Seq coverage indicates they might be true positives. We therefore believe that our cutoffs are conservative.

PAR-CLIP cluster scoring

In addition to counting the characteristic nucleotide conversion events (T to C for 4SU and G to A for 6SG PAR-CLIP) we compute an entropy score for each cluster. The purpose of this score is to rank clusters by the amount of 'independent information' contributed by the constituting reads. By means of a cutoff on this score we deplete mapping- and amplification artifacts with low entropy while retaining high quality clusters. After collecting all reads which belong to a cluster we have a continuous stretch of covered nucleotides with positions i=1...L. We now count, for each position, the number of sequenced nucleotides A,C,G,T plus deletion and insertion events. Together these 6 numbers A_i, C_i, G_i, T_i, D_i, I_i form a vector O_i of 'observation-counts' at position i.

(1) The whole cluster may be thought of as the matrix M of observation-counts at all positions $(O_1,...,O_L)$.

(2) At each position, the sum over all possible observations yields the depth of coverage N_i .

(3) Observation-counts can be normalized to yield observation frequencies o_{ij} on which the Shannon-entropy over all events j (A,C,G,T,D,I) can be computed. The sum of these entropies over the whole cluster H_O captures the amount of sequencing variability within a cluster and forms the first term in our entropy-score.

The second term H_N captures the variability of read start and end positions. The depth of coverage is normalized by the total coverage, representing the shape of the cluster by the 'frequency of coverage' at each position. For a single read (or an amplification artifact) the coverage is constant within the cluster, resulting in a uniform distribution. We therefore compute the Kullback-Leibler divergence between the observed coverage distribution n_i and the uniform distribution $u_i = 1/L$.

Scoring Functions

$$O_i = (A_i, C_i, G_i, T_i, D_i, I_i)^T \tag{1}$$

$$M = (O_1,\ ...\ O_L) \tag{2}$$

$$N_i = \sum_{j=1}^{6} O_{ij} = A_i + C_i + G_i + T_i + D_i + I_i \tag{3}$$

$$o_{ij} = \frac{O_{ij}}{N_i} \tag{4}$$

$$H_O = -\sum_{i=1}^{L} \sum_{j=1}^{6} o_{ij} \log(o_{ij}) \tag{5}$$

$$u_i = \frac{1}{L} \tag{6}$$

$$n_i = \frac{N_i}{\sum_{i=1}^{L} N_i} \tag{7}$$

$$\begin{aligned} H_N &= D_{KL}(n \parallel u) && (8) \\ &= \sum_{i=1}^{L} n_i \ \log\left(\frac{n_i}{u_i}\right) && (9) \\ &= \sum_{i=1}^{L} n_i \ \log(n_i L) && (10) \end{aligned}$$

$$Score = H_O + H_N \tag{11}$$

Crosslink score normalization for the UCSC genome browser

In order to compare the results of PAR-CLIP libraries that have been sequenced at different depth and to enable a useful gray scale representation in the UCSC browser we normalize the crosslink scores *C* in each cluster set to the interval [0,1000], obtaining

$$S = 1000 \cdot \frac{\log(\min(1000, C))}{\log(1000)}$$

The transformation is monotonous and does not change the order of clusters. However, it compresses the dynamic range by capping of the linear score range at very high numbers (*C* > 1000) and a logarithmic scaling of the remaining range. The original numbers of nucleotide conversions and also the entropy score are contained in the GFF files in our download area.