DATABASE
The Journal of Biological Databases and Curation

Database, 2017, 1–9
doi: 10.1093/database/baw167
Original article

Original article

RAIN: RNA–protein Association and Interaction Networks

Alexander Junge[1,2,†], Jan C. Refsgaard[3,†], Christian Garde[1,4,†], Xiaoyong Pan[1,2,3], Alberto Santos[3], Ferhat Alkan[1,2], Christian Anthon[1,2], Christian von Mering[5], Christopher T. Workman[1,4], Lars Juhl Jensen[1,3,*] and Jan Gorodkin[1,2,*]

[1]Center for Non-coding RNA in Technology and Health, University of Copenhagen, Copenhagen, , Groennegaardsvej 3, DK-1870 Frederiksberg C, Denmark, [2]Department of Veterinary Clinical and Animal Sciences, University of Copenhagen, Groennegaardsvej 3, DK-1870 Frederiksberg C, Denmark, [3]Disease Systems Biology Program, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Building: 06-2-26, Blegdamsvej 3B, DK-2200 Copenhagen N, Denmark, [4]Center for Biological Sequence Analysis, Technical University of Denmark, Kemitorvet, Building 208, DK-2800 Lyngby, Denmark, [5]Institute of Molecular Life Sciences and Swiss Institute of Bioinformatics, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland

*__Corresponding author:__ Jan Gorodkin. Email: gorodkin@rth.dk

Correspondence may also be addressed to Email: lars.juhl.jensen@cpr.ku.dk

Present address: Christian Garde, The Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Building 6.6 Blegdamsvej 3B, 2200 Copenhagen N, Copenhagen, Denmark

[†]These authors contributed equally to this work.

Citation details: Junge,A., Refsgaard,J.C., Garde,C. et al. RAIN: RNA–protein association and interaction networks. *Database* (2016) Vol. 2016: article ID baw167; doi:10.1093/database/baw100

Revised 18 November 2016; Accepted 5 December 2016

Abstract

Protein association networks can be inferred from a range of resources including experimental data, literature mining and computational predictions. These types of evidence are emerging for non-coding RNAs (ncRNAs) as well. However, integration of ncRNAs into protein association networks is challenging due to data heterogeneity. Here, we present a database of ncRNA–RNA and ncRNA–protein interactions and its integration with the STRING database of protein–protein interactions. These ncRNA associations cover four organisms and have been established from curated examples, experimental data, interaction predictions and automatic literature mining. RAIN uses an integrative scoring scheme to assign a confidence score to each interaction. We demonstrate that RAIN outperforms the underlying microRNA-target predictions in inferring ncRNA interactions. RAIN can be operated through an easily accessible web interface and all interaction data can be downloaded.

Database URL: http://rth.dk/resources/rain

Introduction

The study of protein-coding genes and the accumulation of data from expression studies and other complementary methods have helped researchers to generate protein association networks compiled in resources such as the STRING database (1). Using a probabilistic scoring scheme, STRING assigns a score to each physical interaction and functional association (henceforth referred to as *interactions*). The recent version 10 holds interactions for >2000 organisms.

However, interaction networks containing only proteins and their interactions remain incomplete until other important molecular interactions have been included. For this reason, we have focused on complementing protein interaction networks with non-coding RNAs (ncRNAs)—a large class of genes comprising ~16 000 long and ~10 000 short ncRNAs in human [GENCODE version 24 (2)]. Integration of these interactions allows for an analysis of the complex functional interplay of ncRNA–RNA and ncRNA–protein interactions. Data on such interactions, complemented by co-expression and literature mining, are currently emerging (3–5). This led to the generation of databases storing ncRNA interactions such as miRTarBase (6) and TarBase (7) containing microRNA (miRNA)–target interactions. NPInter (5), RAID (8) and StarBase (9) are examples of databases collecting interactions between ncRNAs and proteins.

The analysis of ncRNA interactions is challenged by issues related to data heterogeneity, such as varying quality as well as the usage of different identifiers and interaction scoring schemes. The STRING database, used by thousands of researchers daily, has addressed these challenges for proteins through the use of unified identifiers and calibrated scoring schemes (1). A resource similar to STRING is not available for ncRNAs and their interactions.

Similar to protein interactions, ncRNA interactions are supported by diverse sources of evidence such as expert curation, experiments, text mining and predictions. In order to compare these sources of evidence, a scoring scheme needs to be established that assesses the reliability of each interaction. NcRNAs interacting with either proteins or ncRNAs furthermore affect the pathways these interaction partners are involved in. Hence, an approach that makes it easy to navigate both ncRNA as well as protein association networks promises to benefit the study of cellular interaction networks.

We have used a strategy similar to that of STRING in order to develop RAIN (RNA–protein Association and Interaction Networks), a novel resource that covers ncRNA and their associations with other ncRNAs and proteins. RAIN integrates ncRNA interactions from a diverse set of sources and covers four organisms: human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*) and baker's yeast (*Saccharomyces cerevisiae*). RAIN scores the reliability of each interaction using a scoring scheme based on the comparison to a curated set of interactions. It finally integrates ncRNA–RNA and ncRNA–protein associations with protein–protein associations contained in the STRING database. This enables researchers to explore complex interaction networks in the powerful, yet intuitive interactive STRING user interface.

Materials and Methods

Sources of evidence

We established four channels of evidence to support the interactions found in RAIN, namely, (i) curated knowledge, (ii) experimental evidence, (iii) miRNA target predictions and (iv) automated literature mining, see Figure 1. Each of the four evidence channels is generated by integrating a number of underlying resources.

(i) *Curated knowledge*. This comprises 867 human molecular interactions that are well established in the scientific literature and/or listed in expert curated databases. The interactions were collected for nine classes of ncRNAs, namely microRNA (miRNA) (3), ribosomal RNA (rRNA) (10), transfer RNA (tRNA) (11), signal recognition particle RNA (SRP RNA) (12), Vault RNA (13–15), Y RNA (16–18), Telomerase RNA (19), small nucleolar RNA (snoRNA) (20) and spliceosomal RNA (U1, U2, U4, U4atac, U6, U6atac, U11, U12) (20). For further

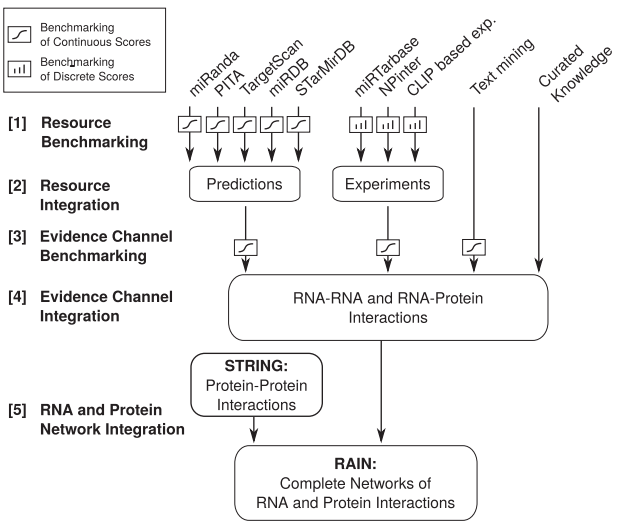

Figure 1. Flow chart illustrating the development of the RAIN database, ranging from establishing scoring schemes for the individual sources of evidence, through integration of resources to evidence channels, to finally defining functional molecular networks.

details on the curated interactions, refer to Supplementary Section 2.

(ii) *Experimental evidence*. This comprises 10 588 interactions supported by experimental data. Cross-linking immunoprecipitation (CLIP) based experiments were retrieved from StarBase (9) and supplemented by interactions identified in CLASH and CRAC experiments (21–23). Furthermore, experimentally supported interactions were extracted from miRTarBase (6) and NPInter (5) and redundancy between the databases was removed. The confidence of the experimental evidence was based on the number of experiments supporting a given interaction.

(iii) *miRNA target predictions*: We ran miRanda (24) and PITA (25) with default settings on all combinations of 3′ UTR sequences of protein-coding genes from Ensembl Biomart (26) and miRNA sequences from miRBase (27). Additionally, we retrieved precomputed predictions for miRDB (28), TargetScan (29, 30) and StarMirDB (31).

(iv) *Text mining*. ncRNA orthology groups were generated using Ensembl Biomart (26) and the miRNA family annotations from miRBase v20 (27). Protein orthology groups retrieved from STRING and these ncRNA orthology groups were supplied to the dictionary-based named entity recognition engine described by Pafilis *et al.* (32) to extract associations between ncRNAs and proteins from MEDLINE abstracts. We refer to Pafilis *et al.* (32) for more details on the named entity recognition software. The subsequent text mining was performed using the same name tagger as used in STRING (33).

A *confidence score* is assigned to each evidence for an association. Curated associations were considered highly reliable and assigned the highest possible confidence score for a single source of evidence, defined as 0.9 in STRING. Experimentally supported associations were assigned confidence scores based on the number of supporting experiments/publications. As in STRING (33), associations derived from text mining were scored based on co-occurrences of gene names. For miRNA target predictions, we used the scoring schemes of the individual predictors, at the outset. To put these heterogeneous scores on a common scale, we converted them to probabilistic scores through benchmarking against the same gold standard set (Figure 1, Step [1]). Assuming independence between the sources of evidence, the combined probability of an association was computed from the resource-specific probabilistic scores (Figure 1, Step [2]). The combined probabilities were subjected to a second round of benchmarking to mitigate violations of the assumption of independence (Figure 1, Step [3]). Finally, the evidence channels were integrated to establish the ncRNA association networks (Figure 1, Step [4]) that interface with STRING to provide a complete ncRNA and protein interaction network (Figure 1, Step [5]). We restricted RAIN to only cover organisms with at least 500 ncRNA interactions with confidence scores >0.15 (the same cutoff is used in STRING) which resulted in the inclusion of human (*Homo sapiens*), mouse (*Mus musculus*), rat (*Rattus norvegicus*) and baker's yeast (*Saccharomyces cerevisiae*).

The gold standard set contained 782 miRNA–mRNA interactions that were deemed to be highly reliable. The interactions involve 171 miRNAs and 437 mRNAs. We defined our gold standard based on the curated miRNA–mRNA interactions from Croft *et al.* (3) as well as miRNA–mRNA interactions from miRTarBase (6) and NPInter (5) that were supported by at least two low-throughput experiments. We defined a low-throughput experiment as one that reports less than five miRNA interactions. To ensure an independent benchmarking of miRTarBase and NPInter, we excluded gold standard interactions originating from miRTarBase and NPInter while establishing the resource-specific probabilistic scoring scheme. Once fitted, this scoring scheme was applied to all interactions, including those defined as gold standard interactions.

Naming convention

A consistent naming convention in RAIN was achieved by compiling name and identifier aliases of ncRNA and proteins and generating an alias dictionary that maps these aliases to RAIN identifiers. For proteins and mRNA, RAIN identifiers are equivalent with STRING v10 (1) identifiers, and the alias dictionary is derived from the STRING v10 alias files. Aliases of miRNA were generated from miRBase v20 (27) and the associated miRBase identifiers were used subsequently. Finally, aliases of the remaining ncRNAs were retrieved using Ensembl Biomart v78 (26) and the official name of the given ncRNA was used as the RAIN identifier. The organism-specific database dictated these official names, i.e. HGNC (34) for human, MGI (35) for mouse and rat, and SGD (36) for yeast. All molecular entities were made to conform to the RAIN naming convention prior to establishing the probabilistic scoring schemes.

Probabilistic scoring schemes

For each resource of ncRNA–target interactions integrated into RAIN, a probabilistic scoring scheme was established prior to the process of resource integration. This allowed us to weight the respective resources based on their confidence in the final score integration step, which assigns an easily interpretable confidence score to each interaction.

The probabilistic scoring scheme is established by benchmarking against a gold standard set, Ω, of 782

miRNA–mRNA interactions that are considered to be valid. We denote the ith miRNA–mRNA interaction pair as ω_i and thus $\Omega = \{\omega_1, \ldots, \omega_{782}\}$. Let Ψ denote the set of all possible interactions between the miRNAs, Φ_{mi}, and the mRNAs, Φ_m, contributing to the interactions in Ω. Hence, $\Omega \subset \Psi$. An interaction (mi_j, m_k) between miRNA $mi_j \in \Phi_{mi}$ and mRNA $m_k \in \Phi_m$ contributed by an interaction resource is a true positive (TP) if $(mi_j, m_k) \in \Omega$. Similarly, (mi_j, m_k) is a false positive (FP) if $(mi_j, m_k) \in \Psi \backslash \Omega$.

This is summarized in Figure 2A: Ψ is the benchmark data set consisting of positive examples $\Omega \subset \Psi$ and negative examples $\Psi \backslash \Omega$. In Figure 2A, a white dot represents a TP interaction and a black dot represents a FP interaction. The universe of miRNA–mRNA interactions can be extended beyond Ψ covering miRNAs and miRNAs not present in Φ_{mi} and Φ_m, respectively. We call this set Ψ_{all} and note that $\Psi \subset \Psi_{all}$. Interactions in $\Psi_{all} \backslash \Psi$ are represented by gray dots in Figure 2A.

The following was performed in the interest of estimating whether a gray dot represents a likely interaction or not. Each ncRNA interaction resource had a discrete or continuous raw score assigned to each potential interaction contributed by the resource. To ensure that these scores were comparable, we calibrated them based on their agreement with the benchmark set Ψ. This calibration procedure is described in the following sections.

Scoring schemes for discrete raw scores

When a source of ncRNA interactions provides discrete raw scores, we calibrated by fitting a discrete transfer function as exemplified in Figure 2B. An example of such discrete scores are the interactions extracted from the portion of miRTarBase not overlapping our benchmark set Ψ. Here, the raw score was defined by the number of publications supporting a given interaction.

For each discrete raw score, the fraction of correctly predicted interactions, $TP/(TP + FP)$, was computed for the set of interactions with the given score. This provided a mapping of raw scores to the interval $[0, 1]$ and defined the transfer function from the raw score assigned by a specific interaction resource to its confidence score $\widehat{Cp}$, where $\widehat{Cp}$ estimates the probability that interactions assigned with the raw score are true.

Scoring schemes for continuous raw scores

The interactions contributed by each resource were reduced to those contained in the benchmarking set, Ψ, and sorted in ascending order according to their raw scores. A window containing w interactions was then slid over the interactions using a step size of 1. Supplementary Table S3 lists w empirically chosen for each data set. In each window, the fraction of correctly predicted interactions, Cp, as well as the mean raw score μ was calculated. We estimated the relationship between Cp and μ by fitting a sigmoid transfer function of the form

$$f(x) = \frac{a - d}{1 + \exp(-b \cdot (x - c))} + d,$$

where $lim_{x \to \infty} f(x) = a$ and $lim_{x \to -\infty} f(x) = d$. c shifts the function horizontally and b defines the steepness of the sigmoid function. To achieve the best least squares fit for the transfer function, we defined a number of seeds and boundaries for the parameters $\{a, b, c, d\}$ and used the fit with least mean square error.

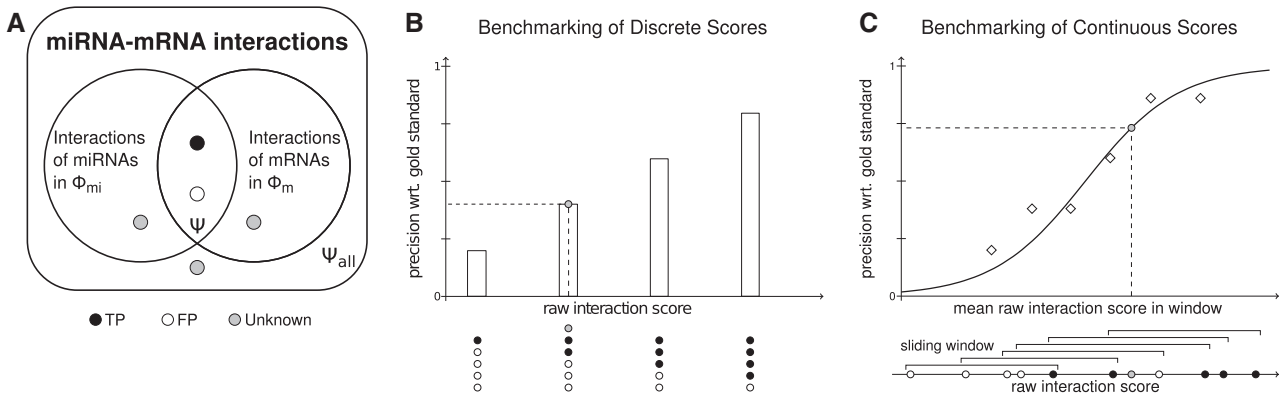

Figure 2. Toy example describing the benchmarking and scoring scheme. **(A)** A true positive (TP) interaction is depicted as a black dot and represents a miRNA–mRNA pair found in the gold standard; a false positive (FP) interaction is depicted as a white dot and comprises interactions where the miRNA and mRNA constituents are in the gold standard, but their pair is not. Interactions where the miRNA or the mRNA were not part of the gold standard are depicted as gray dots. Only TP and FP interactions are used to establish the transfer function, which subsequently is applied to assign confidence scores to all interactions. **(B)** A discrete transfer function is established as the fraction of correctly predicted interactions in each of the discrete raw score bins. **(C)** A continuous transfer function is established based on the TP and FP interactions found in sliding windows. The mean raw interaction score and fraction of correctly predicted interactions were computed for each window, followed by the fitting of a sigmoid transfer function.

After the fitting process, we applied f to map continuous raw scores to confidence scores $\widehat{Cp}$, representing the probability of the interaction being true, as depicted in Figure 2C.

Integration of evidence

Since we wish to integrate the evidence for the interaction i from N sources of interactions belonging to the same evidence channel, we performed the following. Let $\widehat{Cp}_{ji} \in [0, 1]$ be the confidence score of resource j for interaction i. $\widehat{Cp}_{ji}$ can be interpreted as a probability and we are interested in finding the probability that interaction i is true given all available evidence, subsequently denoted by $\tilde{Cp}_i$. Under the assumption of independence between the N source of evidence, we integrated the resource-specific scores using a modified version of the Noisy–Or model, which takes the prior probability, p^*, into account.

$$\frac{1 - \tilde{Cp}_i}{1 - p^*} = \prod_{j=1}^{N} \frac{1 - \widehat{Cp}_{ji}}{1 - p^*}$$

which yield

$$\tilde{Cp}_i = 1 - (1 - p^*)^{1-N} \prod_{j=1}^{N} (1 - \widehat{Cp}_{ji}).$$

The prior probability is defined as the probability of randomly selecting a true positive miRNA–mRNA interaction from all combinations of Φ_m and Φ_{mi}. Given our benchmarking set, the prior used in RAIN is $p^* = 782/(171 \cdot 437) \approx 0.01$. When computing $\tilde{Cp}_i$, accounting for the prior is required to avoid counting the prior for each evidence channel. This prior correction is especially important when dealing with low score close to the prior (see Supplementary Section 5).

Following this integration of the sources of evidence into evidence channels, a second round of calibration was employed to mitigate any violations to the assumption of independence between interaction resources. Note that although each confidence score was computed based on a gold standard only consisting of miRNA–mRNA interactions, the underlying transfer functions mapping raw scores to confidence scores can be applied to score interactions for any class of ncRNAs.

Validation of the integration of miRNA target predictors

For the purpose of evaluating the gain of integrating the respective miRNA target prediction tools into the RAIN prediction channel, we retrieved a list of human and mouse miRNA–mRNA interactions from TarBase (7), that have been tested with functional studies (Luciferase reporter assays). All interactions common between this TarBase set and our gold standard were removed from the TarBase set to establish an independent validation set. This independent validation set comprises a positive set of 1387 confirmed interactions and a negative set of 460 pairs for which the miRNA had no effect on the amount of translated mRNA. The performance was assessed using receiver-operating characteristics (ROC) on the raw scores from the miRNA prediction tools and the combined probabilistic scores for the RAIN prediction channel.

Results and Discussion

RAIN is a novel resource of ncRNA interactions that integrates heterogeneous evidence from experiments, predictions, text mining and expert curation. RAIN comprises a total of 270 242 ncRNA–RNA/protein interactions across four widely investigated organisms: human, mouse, rat and yeast. The number of interactions is summarized in Table 1, with an additional break down of the counts by evidence channel and class of interacting entities in the Supplementary Tables S1 and S2. Furthermore, RAIN interfaces tightly with STRING (1) enabling users to explore networks of ncRNA–RNA, ncRNA–protein and protein–protein associations in an interactive user interface, with the reliability of each interactions represented as a

Table 1. The number of miRNA–mRNA, ncRNA–protein and ncRNA–ncRNA interactions per organism in RAIN with a combined confidence score higher than 0.15

Organism	Number of interactions			
	miRNA–mRNA	ncRNA–protein	ncRNA–ncRNA	Total
H. sapiens (human)	174 853	11 026	2507	188 386
M. musculus (mouse)	77 270	469	35	77 774
R. norvegicus (rat)	19 985	39	1	20 025
S. cerevisiae (Baker's yeast)	0	640	85	725
Total	272 108	12 174	2628	286 910

single easily interpretable confidence score. RAIN is to our knowledge the first resource to offer this.

The human interactions constitute ~66% of the total RAIN interactions. This likely reflects a research bias towards investigating and annotating ncRNA in human relative to mouse and rat. *Saccharomyces cerevisiae* does not harbor miRNAs and the other constituents of the RNAi pathway, thus miRNA target predictions cannot be provided for this yeast species. The *S. cerevisiae* genome does, however, encode a wide range of RNA binding proteins and various classes of ncRNA, that have been investigated in the literature, e.g. by pull-down studies (21, 23). Hence, the interactions of several players in transcriptional and post-transcriptional regulation have been integrated and are available in RAIN for all four organisms. We expect that RAIN will be a valuable tool to facilitate the understanding of the molecular regulatory mechanisms. In addition to aiding the researcher in the process of generating hypotheses to be tested, RAIN also allows researcher to advance differential high-throughput studies with a layer of regulatory network biology.

To demonstrate the gain of integrating the individual sources of evidence, we benchmarked the RAIN prediction channel and the respective miRNA target prediction tools. We performed ROC calculations (Figure 3) on the validation set of 1387 positive and 460 negative miRNA–mRNA pairs as described in Section Validation of the integration of miRNA target predictors. The respective miRNA target predictors impose a score threshold for reporting miRNA targets. Hence, despite subjecting all pairs of miRNA and mRNA 3′ UTR to target prediction, only a subset is reported along with a prediction score from each tool (see Supplementary Section 4 for validation set coverage by each tool). Consequently, the ROC curves are truncated as not all positive and negative pairs in the benchmarking sets are reported by the respective prediction tools. This is especially pronounced for tools that rely on conservation of the miRNA target site and 3′ UTR as is the case for TargetScan. The ROC analyses demonstrate that in addition to improving the coverage of the miRNA interactome, integration of the miRNA target predictors also yields an improved predictive performance.

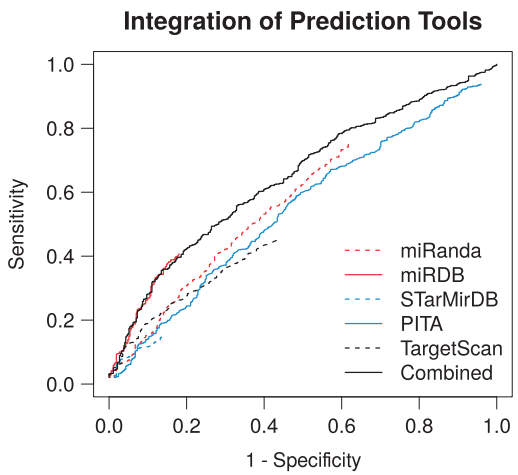

Figure 3. Receiver-operating characteristics of the RAIN prediction channel and the respective miRNA target prediction tools benchmarked against an independent validation set of miRNA–mRNA interactions. The integration of the respective prediction tools yields improved predictive performance. Where specificity = TN/N, sensitivity = TP/P, P is the number of positive and n the number of negative miRNA–mRNA pairs.

We restrained the benchmarking of RAIN to the prediction channel, i.e. the integration of miRNA target predictors. The reason is that the publications underlying the validation set are likely overlapping with the literature evidence underlying RAIN text mining, experiments and curated knowledge evidence. The true performance of RAIN is thus underestimated here as it is only based on the weakest of the four evidence channels.

Utility of the Database

This section describes the RAIN website and user interface. An example use case concludes the section and illustrates the utility of the database.

Query interface

Querying RAIN for a single ncRNA or protein identifier returns interactions for this entity; querying for multiple identifiers returns interactions between these entities. After searching RAIN, an identifier disambiguation page allows the user to choose desired query entities among all ncRNA and protein identifiers in RAIN that match the query. RAIN uses STRING v10 (1) protein identifiers for input protein and mRNA identifiers. miRNAs are mapped to miRBase v20 (27) identifiers. For other ncRNAs, RAIN accepts Ensembl (37) and RefSeq (38) identifiers as well as identifiers from four organism-specific main databases obtained from Ensembl BioMart v78 (26): HGNC (34) for human, MGI (35) for mouse, RGD (39) for rat and SGD (36) for yeast.

Network view

After querying RAIN, a search results page featuring a static image of the resulting interaction network is shown. Associations adjacent to ncRNAs obtained from RAIN and protein–protein interactions from STRING are shown in the same network. Sources of evidence supporting an association are indicated by different edge colors. If interactions for a protein were searched, the number of interacting

ncRNAs and proteins displayed can be adjusted. Furthermore, the interaction network may be downloaded as files in tab-separated and PSI-MITAB format.

Clicking the network image redirects to an interactive network view in STRING allowing users to adjust the confidence score cutoff and network size as well as centering the network on different nodes. Clicking a node or an edge in the interactive network view lists additional information. For each edge, the confidence score computed for each evidence channel as well as the combined confidence score are shown. Clicking a node shows basic information about the corresponding molecular entity. Information about ncRNAs nodes and adjacent interactions are contributed by RAIN while information about proteins and and protein–protein interaction are provided by STRING.

Data downloads

All interactions and the benchmarking gold standard can be downloaded as tab-separated files from the RAIN website easing programmatic analyses of RAIN data. Interactions are split by evidence channel and, in contrast to those shown in the RAIN network view, not reduced to those interactions with a combined score larger than 0.15. Furthermore, compiled aliases of ncRNA and protein identifiers are available for download.

Use case

RAIN enables users to study ncRNAs, proteins and their interactions in an intuitive workflow as displayed in Figure 4 and answer complex research questions. Figure 4A exemplifies the *single identifier search* in RAIN while an example for an edge pop-up window with more information about an association of interest is shown in Figure 4B. Listing the sources of evidence and confidence scores for each interaction make the network easily interpretable. Furthermore, the *multiple identifier search* in RAIN is shown in Figure 4C. Finally, a node pop-up window with information about ncRNA nodes is depicted in Figure 4D.

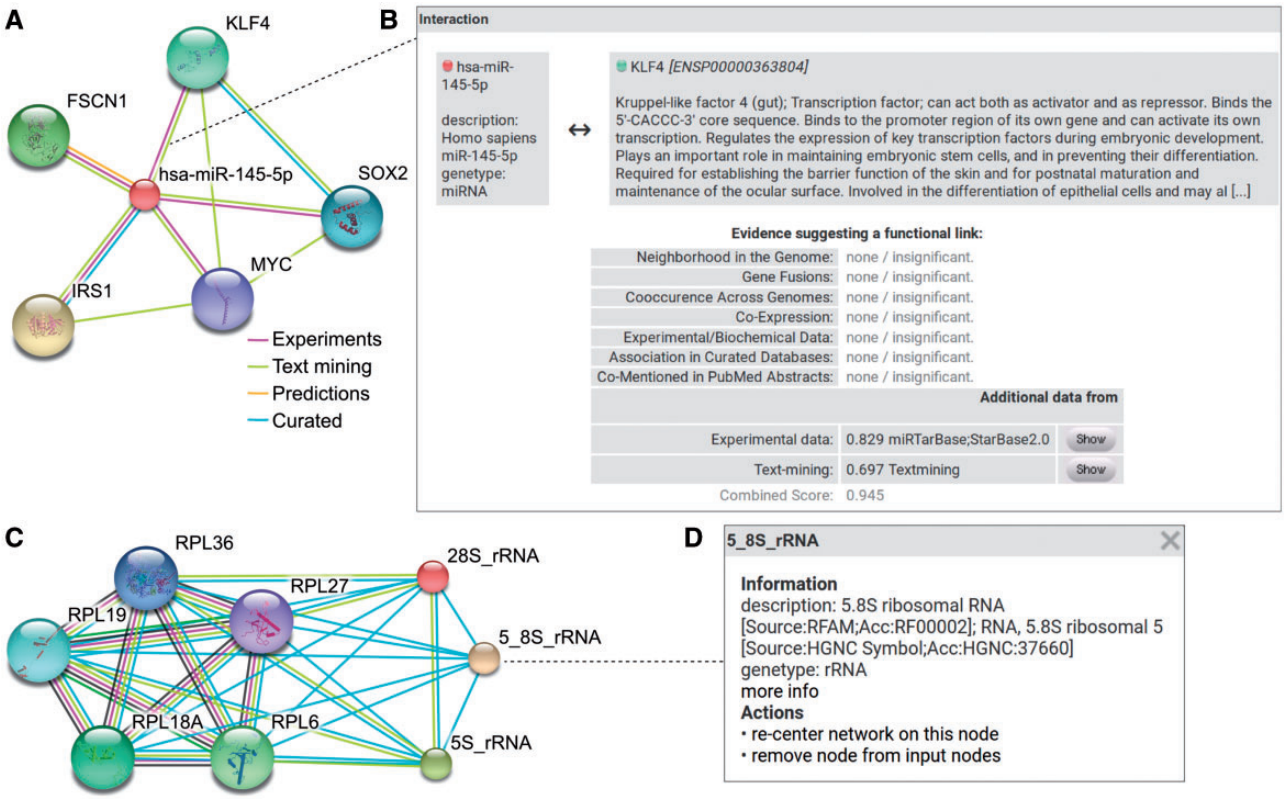

Figure 4. RAIN use case. **(A)** Querying RAIN for human miR-145-5p (miR-145), suggested to act as tumor-suppressor in breast and colon cancer (40, 41), finds multiple oncogenes such as KLF4 and SOX2 (42, 43) as putative targets of miR-145. Evidence channels supporting each interaction are encoded as edge colors. **(B)** Sources of evidence for each association, e.g. between miR-145 and KLF4, are presented in a pop-up opened after clicking an edge in the network. RAIN confidence scores are collected in the 'Additional data' table. Information about KLF4 is provided by STRING. Clicking the 'Show' button leads to a website that links to research articles presenting experimental evidence and displaying detailed text mining evidence, where available. **(C)** In contrast to single identifier search (A), the RAIN multiple identifier search can be used to specifically view interactions between three ribosomal RNAs (28S_rRNA, 5_8S_rRNA, 5S_rRNA) and a subset of five ribosomal proteins part of the large ribosomal subunit. These interactions were extracted from Reactome (10) or found by text mining. **(D)** Clicking an ncRNA node in the network opens a popup with basic information about the ncRNA, e.g. 5.8S rRNA.

Conclusion

We presented RAIN, a novel database for ncRNAs and their interactions with other ncRNAs and proteins. Associations in RAIN are obtained from a set of resources based on expert curation, experiments, text mining and interaction predictions. RAIN uses a probabilistic scoring scheme to assign a single confidence score to each interaction allowing users to integrate support from all sources of evidence in RAIN in a single number.

RAIN is tightly integrated with the STRING database for protein–protein interactions and adds ncRNAs together with their interactions to the existing protein–protein interaction networks in STRING. RAIN is implemented using the STRING payload mechanism. This allows RAIN users to use interactive and accessible STRING network visualizations. Additionally, potential RAIN users may already be familiar with the STRING interface, further reducing the effort needed to start exploring RAIN.

Future work includes expanding the gold standard to improve the accuracy of the RAIN confidence scores. Furthermore, additional sources of ncRNA interactions such as expert curated interactions from TarBase, which was not included due to current licensing restrictions, could be included. The curated knowledge evidence channel will be expanded to other ncRNA classes and updated according to future literature evidence while maintaining the same high inclusion criteria. The integration of RNA–protein binding site prediction approaches such as RNAcontext (44) or GraphProt (45) would also be of interest. This would, however, require an extension of the gold standard to include this type of interaction. After expanding RAIN to cover more organisms and establishing a comprehensive definition of orthologous groups for ncRNAs, similar to eggNOG (46) for proteins, RAIN interaction evidence could furthermore be transferred between organisms. Finally, further annotations in the node pop-up, e.g. disease association, tissue specificity, and species conservation could prove to be useful. We plan to address these points in future versions of RAIN.

RAIN facilitates the understanding of complex molecular networks through the integration of ncRNA interactions and protein–protein association networks. The graphical web interface provides the researcher with intuitive access to the interactions of ncRNAs and proteins of interest and assigns a confidence score to each association. The incorporation of ncRNAs, including intensely investigated miRNAs and long ncRNAs, makes RAIN a powerful tool to answer current research questions.

Supplementary Data

Supplementary data are available at *Database* Online.

Funding

The Danish Council for Independent Research (Technology and Production Sciences); The Danish Center for Scientific Computing (DCSC/DEiC); Innovation Fund Denmark (Programme Commission on Strategic Growth Technologies); The Novo Nordisk Foundation (NNF14CC0001).

Conflict of interest. None declared.

References

1. Szklarczyk,D., Franceschini,A., Wyder,S. *et al*. (2015) STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Res*., 43, D447–D452.
2. Harrow,J., Frankish,A., Gonzalez,J.M. *et al*. (2012) GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res*., 22, 1760–1774.
3. Croft,L., Szklarczyk,D., Jensen,L.J., and Gorodkin,J. (2012) Multiple independent analyses reveal only transcription factors as an enriched functional class associated with microRNAs. *BMC Syst Biol*., 6, 90.
4. Chen,G., Wang,Z., Wang,D. *et al*. (2013) LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*., 41, D983–D986.
5. Yuan,J., Wu,W., Xie,C. *et al*. (2014) NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Res*., 42, D104–D108.
6. Chou,C.H., Chang,N.W., Shrestha,S. *et al*. (Jan 2016) miRTarBase 2016: updates to the experimentally validated miRNA-target interactions database. *Nucleic Acids Res*., 44, D239–D247.
7. Vlachos,I.S., Paraskevopoulou,M.D., Karagkouni,D. *et al*. (Jan 2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res*., 43, D153–D159.
8. Zhang,X., Wu,D., Chen,L. *et al*. (2014) RAID: a comprehensive resource for human RNA-associated (RNA–RNA/RNA–protein) interaction. *rna*, 20, 989–993.
9. Li,J.H., Liu,S., Zhou,H. *et al*. (Jan 2014) starBase v2.0: decoding miRNA–ceRNA, miRNA–ncRNA and protein–RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res*., 42, D92–D97.
10. Croft,D., O'kelly,G., Wu,G. *et al*. (Jan 2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res*., 39, D691–D697.
11. Lowe,T.M., and Eddy,S.R. (Mar 1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*., 25, 955–964.
12. Rosenblad,M.A., Larsen,N., Samuelsson,T. and Zwieb,C. (2009) Kinship in the SRP RNA family. *RNA Biol*., 6, 508–516.
13. Kickhoefer,V.A., Stephen,A.G., Harrington,L. *et al*. (Nov 1999) Vaults and telomerase share a common subunit, TEP1. *J Biol Chem*., 274, 32712–32717.
14. Kickhoefer,V.A., Liu,Y., Kong,L.B. *et al*. (Jan 2001) The telomerase/vault-associated protein TEP1 is required for vault RNA stability and its association with the vault particle. *J Cell Biol*., 152, 157–164.
15. van Zon,A., Mossink,M.H., Schoester,M. *et al*. (Oct 2001) Multiple human vault RNAs. expression and association with the vault complex. *J Biol Chem*., 276, 37715–37721.

16. Hogg,J.R., and Collins,K. (2007) Human Y5 RNA specializes a Ro ribonucleoprotein for 5S ribosomal RNA quality control. *Genes Dev*., 21, 3067–3072.
17. Stein,A.J., Fuchs,G., Fu,C. *et al*. (2005) Structural insights into RNA quality control: the Ro autoantigen binds misfolded RNAs via its central cavity. *Cell*, 121, 529–539.
18. Green,C.D., Long,K.S., Shi,H., and Wolin,S.L. (1998) Binding of the 60-kDa Ro autoantigen to Y RNAs: evidence for recognition in the major groove of a conserved helix. *rna*, 4, 750–765.
19. Zwieb C., telomdb. http://rnp.uthscsa.edu/rnp/telomDB/telomDB.html (3 March 2015, data last accessed).
20. Jorjani,H., Kehr,S., Jedlinski,D.J. *et al*. (2016) An updated human snoRNAome. *Nucleic Acids Res*., 44, 5068–5082.
21. Kudla,G., Granneman,S., Hahn,D. *et al*. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA–RNA interactions in yeast. *Proc Natl Acad Sci USA*., 108, 10010–10015.
22. Helwak,A., Kudla,G., Dudnakova,T., and Tollervey,D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153, 654–665.
23. Tuck,A.C., and Tollervey,D. (2013) A transcriptome-wide atlas of RNP composition reveals diverse classes of mRNAs and lncRNAs. *Cell*, 154, 996–1009.
24. John,B., Enright,A.J., Aravin,A. *et al*. (2004) Human microRNA targets. *PLoS Biol*., 2, e363.
25. Kertesz,M., Iovino,N., Unnerstall,U. *et al*. (2007) The role of site accessibility in microRNA target recognition. *Nat Genet*., 39, 1278–1284.
26. Kinsella,R.J., Kahari,A., Haider,S. *et al*. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, 2011, article ID bar030.
27. Kozomara,A., and Griffiths-Jones,S. (2014) miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res*., 42, D68–D73.
28. Wong,N., and Wang,X. (2015) miRDB: an online resource for microRNA target prediction and functional annotations. *Nucleic Acids Res*., 43, D146–D152.
29. Garcia,D.M., Baek,D., Shin,C. *et al*. (2011) Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nat Struct Mol Biol*., 18, 1139–1146.
30. Agarwal,V., Bell,G.W., Nam,J.W., and Bartel,D.P. (2015) Predicting effective microRNA target sites in mammalian mRNAs. *Eife*, 4,
31. Rennie,W., Liu,C., Carmack,C.S. *et al*. (2014) STarMir: a web server for prediction of microRNA binding sites. *Nucleic Acids Res*., 42, W114–W118.
32. Pafilis,E., Frankild,S.P., Fanini,L. *et al*. (2013) The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS ONE*, 8, e65390.
33. Franceschini,A., Szklarczyk,D., Frankild,S. *et al*. (2013) STRING v9.1: protein–protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*., 41, D808–D815.
34. Gray,K.A., Yates,B., Seal,R.L. *et al*. (2015) Genenames.org: the HGNC resources in 2015. *Nucleic Acids Res*., 43, D1079–D1085.
35. Eppig,J.T., Blake,J.A., Bult,C.J. *et al*. (2015) The Mouse Genome Database (MGD): facilitating mouse as a model for human biology and disease. *Nucleic Acids Res*., 43, D726–D736.
36. Hirschman,J., Balakrishnan,E.R., Christie,K.R. *et al*. (2006) Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the Saccharomyces cerevisiae genome. *Nucleic Acids Res*., 34, D442–D445.
37. Flicek,P., Amode,M.R., Barrell,D. *et al*. (2014) Ensembl 2014. *Nucleic Acids Res*., 42, D749–D755.
38. Pruitt,K.D., Brown,G.R., Hiatt,S.M. *et al*. (2014) RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res*., 42, D756–D763.
39. Shimoyama,M., De Pons,J., Hayman,G.T. *et al*. (2015) The Rat Genome Database 2015: genomic, phenotypic and environmental variations and disease. *Nucleic Acids Res*., 43, D743–D750.
40. Gotte,M., Mohr,C., Koo,C.Y. *et al*. (2010) miR-145-dependent targeting of junctional adhesion molecule A and modulation of fascin expression are associated with reduced breast cancer cell motility and invasiveness. *Oncogene*, 29, 6569–6580.
41. Zhang,J., Guo,H., Zhang,H. *et al*. (2011) Putative tumor suppressor miR-145 inhibits colon cancer cell growth by targeting oncogene Friend leukemia virus integration 1 gene. *Cancer*, 117, 86–95.
42. Yu,F., Li,J., Chen,H. *et al*. (2011) Kruppel-like factor 4 (KLF4) is required for maintenance of breast cancer stem cells and for cell migration and invasion. *Oncogene*, 30, 2161–2172.
43. Santini,R., Pietrobono,S., Pandolfi,S. *et al*. (2014) SOX2 regulates self-renewal and tumorigenicity of human melanoma-initiating cells. *Oncogene*, 33, 4697–4708.
44. Kazan,H., Ray,D., Chan,E.T. *et al*. (2010) RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*., 6, e1000832.
45. Maticzka,D., Lange,S., Costa,J.F., and Backofen,R. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol*., 15, R17.
46. Huerta-Cepas,J., Szklarczyk,D., Forslund,K. *et al*. (2016) eggnog 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Res*., 44, D286–D293.