OXFORD

Briefings in Bioinformatics, 18(6), 2017, 1071–1081

doi: 10.1093/bib/bbw084
Advance Access Publication Date: 26 September 2016
Paper

A novel computational method for inferring competing endogenous interactions

Davide S. Sardina, Salvatore Alaimo, Alfredo Ferro, Alfredo Pulvirenti and Rosalba Giugno

Corresponding author: Rosalba Giugno, Department of Computer Science, University of Verona, Strada le Grazie 15 – 37134, Verona. Tel.: +390458027066; Fax: +390458027068; E-mail: rosalba.giugno@univr.it

Abstract

Posttranscriptional cross talk and communication between genes mediated by microRNA response element (MREs) yield large regulatory competing endogenous RNA (ceRNA) networks. Their inference may improve the understanding of pathologies and shed new light on biological mechanisms. A variety of RNA: messenger RNA, transcribed pseudogenes, noncoding RNA, circular RNA and proteins related to RNA-induced silencing complex complex interacting with RNA transfer and ribosomal RNA have been experimentally proved to be ceRNAs. We retrace the ceRNA hypothesis of posttranscriptional regulation from its original formulation [Salmena L, Poliseno L, Tay Y, *et al. Cell* 2011;146:353–8] to the most recent experimental and computational validations. We experimentally analyze the methods in literature [Li J-H, Liu S, Zhou H, *et al. Nucleic Acids Res* 2013;42:D92–7; Sumazin P, Yang X, Chiu H-S, *et al. Cell* 2011;147:370–81; Sarver AL, Subramanian S. *Bioinformation* 2012;8:731–3] comparing them with a general machine learning approach, called ceRNA predIction Algorithm, evaluating the performance in predicting novel MRE-based ceRNAs.

Key words: computational models; experimental validations; competing endogenous RNA

Introduction

The competing endogenous effect is a posttranscriptional activity in which different RNAs (ceRNAs) compete for shared microRNAs (miRNAs), thus regulating each other [1]. In [2], the authors hypothesized that messenger RNAs (mRNAs) act as competitive inhibitors of miRNA function (i.e. miRNA sponge or miRNA decoy) by preventing miRNAs from binding their authentic targets [3], modulating their activity in a cell-type-specific manner. Such a hypothesis helps explaining the difference in binding sites conservation among species and the reason for *in vivo* detection of a poor miRNA-mediated repression.

A competing effect occurs when one or more miRNA response elements (MREs), targeted by the same pool of miRNAs, lie in a RNA transcript. Using techniques based on immunoprecipitation, such as cross-linking, ligation and sequencing of hybrids (CLASH) [4], several experimentally validated and high-confidence miRNA–target interactions have been discovered [4–7] (Figure 1).

The cross talk between RNAs, mediated by MREs, regulates the relative concentrations of transcripts within the cell, yielding large-scale regulatory networks (Figure 2).

Several types of RNAs, in different species, are involved in ceRNA mechanism. These include miRNAs as mediators, and

Davide Stefano Sardina is PhD student in Computer Science at University of Catania. He works on prediction and functional characterization of biochemical interactions and NGS analysis of ncRNA for lymphoma classification.
Salvatore Alaimo is Post-Doc in Computer Science at University of Catania. He works on computational pathway analysis and simulation, NGS analysis, methodologies for precision medicine and their translational applications.
Alfredo Ferro is Full Professor in Computer Science at University of Catania. He works on methods and models for computational biology, biomedicine and personalized medicine.
Alfredo Pulvirenti is Associate Professor of Computer Science at University of Catania. He studies methods for pathway and biological network analysis, drug repositioning, ncRNAs, subgraph matching and motif finding.
Rosalba Giugno is Associate Professor in Computer Science at University of Verona. She works on biological network modeling/analysis, classification of phenotypes by coding and noncoding expression profiles and drug synergy.
Submitted: 26 May 2016; **Received (in revised form):** 18 August 2016

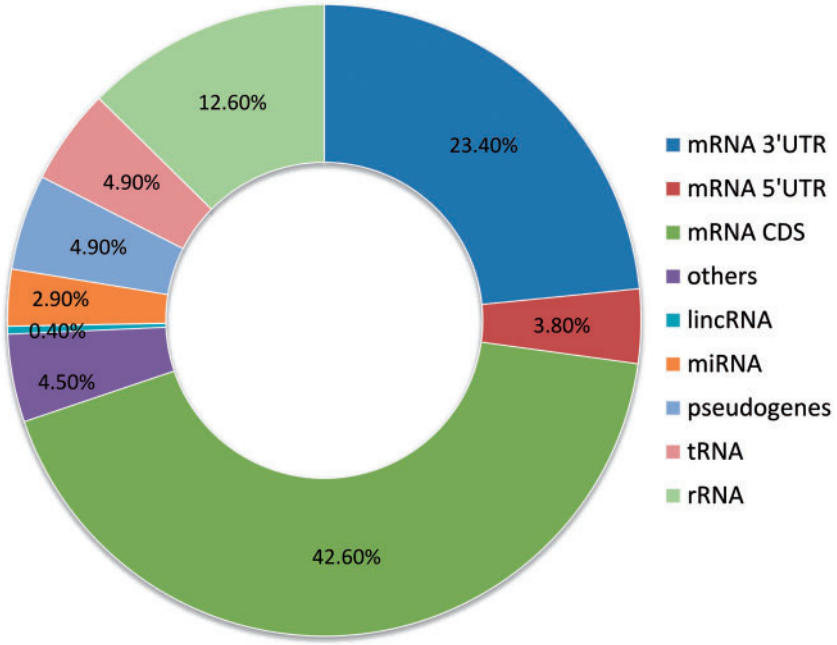

Figure 1. Different miRNA target sites distribution. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

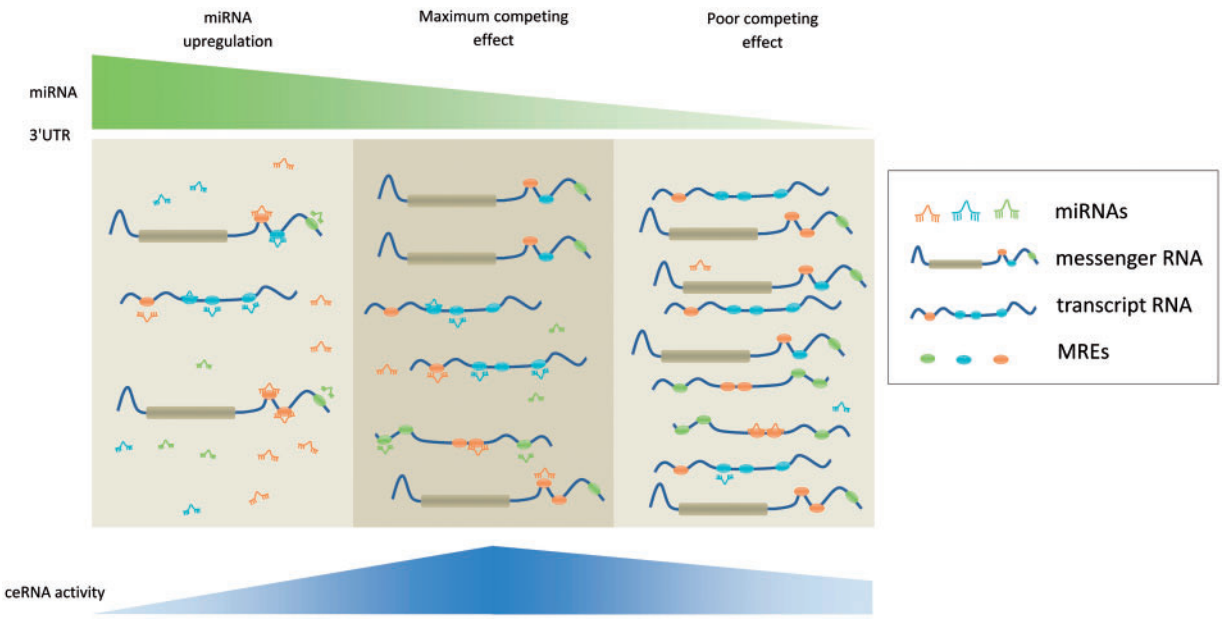

Figure 2. Different scenarios describing the ratio between miRNA and ceRNA expression. The balance between miRNAs and RNAs gives rise to the cross talk. Indeed, in the middle case, the maximum competing effect and cross talk between RNAs occurs. In the leftmost case, miRNAs upregulation causes the downregulation of RNA through quantitative mechanism because MREs are less. Conversely, in the rightmost, the transcript RNAs upregulation causes the downregulation of miRNA. In these cases, less ceRNA effect is reported. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

pseudogenes, long ncRNAs (lncRNA) and circular RNAs (circRNA) as principal actors [8–12].

miRNAs are small (~22 nucleotides), single-stranded ncRNAs. By binding to a MRE on the target RNA, they commonly degrade mRNAs or inhibit the translation of transcripts. Occasionally, miRNAs have been found to enhance gene expression or increase target translation [13, 14]. In [15], the authors introduced a quantitative model describing the nature of regulatory miRNAs. They showed that the target repression activity by miRNA occurs only when their quantity exceeds a certain cellular context-dependent threshold.

A deregulation in tumor suppressor transcripts can disrupt the cellular regulatory network, leading to pathological conditions. Therefore, the analysis of networks induced by ceRNA activities can give insight on the mechanism and the processes underlying the evolution of complex diseases. In [16], the ceRNA regulation mechanism was tested on PTEN, a known tumor suppressor gene and inhibitor of the phosphoinositide 3-kinase/Akt/mammalian target of rapamycin signaling pathway. The authors used a bioinformatics approach, called 'mutually targeted MRE enrichment' (MuTaME), to find putative

ceRNAs of PTEN in a 3′ untranslated region (UTR)-dependent manner, and to demonstrate that mRNAs can cross talk through MREs. To measure the ceRNA effect, MuTaME combines four scores related both to the number of miRNAs shared between two genes and the MREs configuration. Moreover, by selecting their top seven predicted genes, the authors showed coexpression for four of them with PTEN in human tissue.

Pseudogenes are a class of transcripts that have lost their coding function because of the lack of stop codons, or single-nucleotide polymorphisms. They have a regulator role, especially in cancer [17, 18]. Their estimated number (∼19 000) is comparable with that of coding genes. Furthermore, pseudogenes have high sequence similarity with their relative genes. In [10], the authors reported examples of competing gene–pseudogenes in prostate cancer cell lines. In [19], the authors showed that pseudogenes are preserved through evolution. They identified 48 coding loci in the earliest eutherian ancestors, whose ability to encode proteins was lost during rodent evolution. These 'unitary pseudogenes' maintained their tissue-specific expression profile, and MRE sites, suggesting an apparent preservation of their posttranscriptional regulatory role. Moreover, they experimentally verified such a conserved role in the murine Pbcas4 unitary pseudogene, and its human ortholog BCAS4.

lncRNAs are a broad class of non-protein-coding RNAs, whose length is >200 bp. They are transcribed by RNA polymerase II from the reversed strand of a gene, and are subsequently polyadenylated. Long intergenic ncRNAs (lincRNAs) are a subclass of lncRNAs, easier to detect from RNA sequencing (RNA-seq) data because of no overlap with other genes. lncRNAs act as posttranscriptional regulators and cell differentiators [20]. In [21], the authors studied the indirect interactions between linc-MD1 and two transcription factors (TFs), MAML1 and MEF2C, during muscle cells differentiation in both mouse and human. Such interactions appear to be mediated by two miRNAs, miR-133 and miR-135.

circRNAs are a class of ncRNAs whose function is almost unknown. They have been found highly expressed in mouse testis suggesting a role in the tissue differentiation [22]. In [23], Guarnerio *et al*. report that circRNAs were obtained from the fusion of different gene's exons after chromosomal translocations in cancer cells. They showed how one of these fusion circRNAs, f-circM9, is related to drug resistance in leukemia.

In [24], the authors discovered a circRNA, called ciRS-7. It contains >70 MREs acting as a miR-7 sponge, with a consequent increasing of miR-7 targets [12]. In [25], the authors demonstrated the importance to uncover the biological role of ciRS-7, and the relationship between miR-7 and few oncogenes. ciRS-7 was mainly found in neuroblastomas and astrocytomas, renal cells and lung carcinomas. Its close interaction with miR-7 suggests an active role of this miRNA as oncogenes regulator (e.g. EGFR, IRS1, IRS2, Pak1, Raf1, Ack1 and PIK3CD) [25].

In [11], the H. saimiri U-rich RNAs (HSURs) 1 and 2, which are the most conserved and abundant RNA transcripts in Herpesvirus saimiri, have been found related to T-cell activation on the host. Three potential miRNAs (miR-142-3p, miR-27 and miR-16) binding HSURs were experimentally validated *in vivo*. They also identified anticorrelation between an underexpression of miR-27 and the subsequent overexpression of FOXO1 (a known target of miR-27), when both HSUR 1 and 2 are expressed. A detailed list of interactions and experiments regarding miRNA sponges can be found in [26, 27].

Several quantitative *in silico* models to measure ceRNA cross talk have been proposed. Some of them are reviewed in the next section and are experimentally evaluated in this article. We refer to [26] for a complementary review.

We developed a computational method, called ceRNA predIction Algorithm (CERNIA, described in the following Section), which takes into account insights from *in vivo* and *in silico* experiments, such as 5′ UTR and coding region binding sites, and tissue-specific gene expression profiles, to uncover novel ceRNAs, by taking into account both validated and high-confidence miRNA–target interactions [4, 28, 29].

CERNIA makes use of a recommendation system, DT-Hybrid [30, 31], to compute miRNA–target predictions by means of a bipartite network. Then, a network projection step [30, 31] is used to obtain putative ceRNA interactions with associated strengths. Predictions are then filtered through an SVM by using a modified MuTaME scoring, taking into account miRNA-target hybridization energy, shared MREs and RNA expression levels. We compared the prediction power and the characteristics of CERNIA with few state-of-the-art methods. CERNIA is available at https://github.com/dsardina/cernia/.

In silico methods for ceRNA networks and validation

In [32], the authors enstablished the optimal conditions for cross talk, validating their results by means of two known ceRNAs: PTEN and VAPA. They simulated different scenarios in which one or more miRNAs interact with a certain number of targets, considering a number of miRNA–target interaction configurations. The authors also considered indirect interactions, which happen when three ceRNAs compete with themselves (e.g. ceRNA1 competes with ceRNA2, ceRNA2 competes with ceRNA3 and ceRNA1 competes with ceRNA3). Their results highlighted the strong impact of indirect interactions in ceRNA networks, also showing that TF levels are strongly interconnected with ceRNA ones. In [33, 34], the authors proposed two models based on stochastic process and titration mechanism. (Titration is a technique used in chemistry to determine the concentration of an unknown solution. When the molecules interact with each other in a titrative way, there is a threshold above which the phenomenon occurs.) Simulations, corroborating the ceRNA findings, proved that a stronger repression exists when ceRNAs have perfect binding [35].

In [33], the authors designed a synthetic gene circuit in human cells, and fitted the model to experimental data. Their findings show that ceRNA effect strongly relies on the relative concentrations of transcripts and miRNA–target binding strengths. They also observed that their results resembled the conditions found in [36], where a greater concentration of transcripts was needed to lower the miRNA repression on a specific target. See Table 1 for a summary of the methods in [32–34].

Tables 2 and 3 compare the described methods, showing details on data and tools, together with the estimation of the ceRNA effect for gene pairs.

TraceRNA [37] uses both validated miRNA targets from miRTarBase [28] and user-provided ones. Starting from a 'gene of interest' (GoI), the algorithm extracts targeting miRNAs, and predicts their target mRNAs by using both SVMicrO [38] and BCmicrO [39]. Putative genes are filtered on the basis of two criterias: (i) the number of common miRNAs between putative ceRNAs and GoI and (ii) a 'confidence score'. The score can be computed in two ways: by combining the results of the above-mentioned interaction prediction algorithms, or by using a scoring methodology called SiteTest, similar to MuTaME [16]. TraceRNA introduces a *P*-value computed for each putative ceRNA and estimates the relative false discovery rate. It allows

Table 1. Summary of the quantitative models in the literature that describes ceRNA cross talk. We report the goal of the studies, the methodologies used, the results and the validation of the models

	Ala *et al.* [33]	Bosia *et al.* [34]	Yuan *et al.* [35]
Goal	Molecular requirements and extent of ceRNA cross talk	Analyze titration and cross talk within a network of M miRNAs interacting with N mRNA targets	Quantitative analysis of a minimal ceRNA network
Method	Mass action model	Stochastic model	Differential equations
Results	Transcription factor and ceRNA networks are highly linked	Hypersensitivity near the molecular threshold	High expression and MRE affinity of targets counteract RNA interference efficiency
Validated ceRNAs	PTEN-VAMP in five cell lines	–	Genetic gene circuit with iRFP, mKate, EYFP in HEK293 cells

Table 2. ceRNA prediction algorithms characteristics

Name	Interactions types	Data	Tools	Pair estimation	ceRNA classes
Hermes	3′ UTR	TCGA	MINDy, Cupid, snapCGH	Fisher's method MI, CMI	mRNA
ceRDB	3′ UTR	TargetScan	Gene Cluster	MRE-based score	mRNA
TraceRNA	3′ UTR	miRTarBase	SVMicro BCmicrO	Gamma distribution, SiteTest algorithm, Pearson correlation, Borda method	mRNA
Linc2GO	3′ UTR	Human lincRNA Catalog, UCSC, miRBase, KEGG	TargetScan, miRanda, PITA	Hypergeometric distribution	mRNA, lncRNA
starBase	3′UTR 5′ UTR CDS	GEO, GENCODE, circBase, KEGG, PANTHER, MSigDB, miRBase, TargetScan, TarBase, miRecords, miRGator, miRNAMap	miRanda/mirSVR, PITA, Pictar, RNA22	Hypergeometric test	mRNA, lncRNA, circRNA, pseudogene
lnCeDB	Sequence	GENCODE	TargetScan, starBase, miRCode,Smith–Waterman	Hypergeometric test, expression levels analysis	mRNA, lncRNA
HumanViCe	3′ UTR	miRBase, vHot, GENCODE, [14]	TargetScan, miRanda, RNAhybrid, microT, PITA	Hypergeometric test	mRNA, lncRNA, circRNA
CERNIA	3′ UTR 5′ UTR CDS	miRTarBase, starBase, miRecords, [7], TCGA	miRanda, DT-Hybrid	MuTaME, DT-Hybrid, Gene expression correlation, SVM classification	mRNA, lncRNA, pseudogene

users to use gene expression data and take into account only tissue-specific ceRNA interactions.

In [40], the ceRNA hypothesis is exploited to predict the functions of lincRNAs. If a gene and a regulator share some miRNAs, they are likely involved in the same biological process. By using RNA-seq, the method is able to gain insights on the relationship between coding genes and regulatory RNAs. They used TargetScan [3], PITA [35] and miRanda [41, 42] as sources of predicted miRNA–target interactions. lincRNAs are retrieved from [43], and their sequences are downloaded from University of California Santa Cruz Genome Browser. Furthermore, miRNA–lincRNA binding sites were computed with miRanda. Finally, by using a hypergeometric test, the authors estimated the significance of the ceRNA effect for a pair of transcripts on the basis of the number of shared miRNA targets. However, no MREs or relative abundances or biding strengths are used. The final database is called Linc2GO.

Other databases reporting ceRNA interactions include starBase [44], lnCeDB [45] and HumanViCe [46]. The last work has been motivated by two studies that investigated the commonly used pathways used by viruses to infect host cells [47, 48].

Two computational methods based on Pearson correlation are reported in [49, 50]. In [49], the authors focused on ceRNA cross talk between mRNAs and lncRNAs in breast cancer to highlight the regulatory function of the latter. They also compared inferred cancer and normal ceRNA networks by computing sensitivity correlation as the difference between Pearson and partial correlation coefficients and retained only those pairs with sensitivity correlation >0.3.

In [50], the authors performed an analysis of ceRNA mechanism among 20 types of cancer, including a topological analysis of the obtained networks. They use predicted miRNA–target interactions, CLIP-seq data sets from starBase and expressions data from TCGA. They inferred putative ceRNAs by taking into account only pairs with a positive correlation score and a significant adjusted *P*-value. They also computed a sensitivity correlation to examine the role of miRNAs in the ceRNA regulatory network.

Other two known algorithms are ceRDB [51] and Hermes [52], which use CMI, together with Cupid [53], a miRNA–target prediction methodology. In [52], the authors concentrate on large-scale regulatory networks from glioblastoma gene expression

data, whereas in [51], biological meaningful results related to PTEN are reported.

Inferring ceRNA networks through recommendation system: CERNIA

Figure 3 depicts the pipeline of CERNIA.

Data sources

CERNIA uses both validated and high-confidence miRNA–target interactions to create a data set with ∼400 000 interactions, 15 131 target RNAs and 660 miRNAs (see Supplementary Materials, Section *Data sets*).

For each miRNA–target pair in our data set, we computed MREs and their hybridization energy (see Supplementary Materials, Section *MRE extraction*). We obtained ∼7 millions MREs related to all miRNAs and targets of our data set.

Table 3. Comparison of the prediction algorithms on the basis of their features

Name	ceRNA prediction/ validation	Validated miRNA/target interactions	Expression data	CLIP-seq data	MRE
HERMES	✓/✓	×	✓	×	×
ceRDB	✓/×	×	×	×	✓
TraceRNA	✓/×	✓	✓	×	✓
Linc2GO	✓/×	×	×	×	×
starBase	✓/×	×[a]	×	✓	×
lnCeDB	✓/×	×	✓	✓	×
HumanViCe	✓/×	×	×	✓	×
CERNIA	✓/×	✓	✓	×	✓

[a]*Bona fide* miRNA–target interactions. They are overlapping with CLIP-seq data experiments of Ago- and RBP-binding sites, but not specifically validated interactions.

ceRNAs prediction method

To predict ceRNA cross talks, we applied the DT-Hybrid recommendation algorithm [30, 31]. DT-Hybrid needs as input an adjacency matrix A, which represents a bipartite network with m nodes of type M and n of type T, where a_{ij} is 1 if a link between node i and node j *exists*, with $i = 1, 2, \ldots, m$, and $j = 1,2,\ldots,n$. For CERNIA, M is miRNA, T is target and $a_{ij} = 1$ if an interaction among the miRNA i and the target j exists.

DT-Hybrid performs a two-step projection of the bipartite network according to the concept of resources transfer: the resources present in T nodes are transferred to M nodes and therefore returned back to T nodes. The result of the network projection is a matrix $W = \{w_{ij}\}_{n\times m}$, which contains the weights for each pair of nodes of class T as follows:

$$w_{ij} = \frac{1}{k(t_i)^{1-\lambda}k(t_j)^{\lambda}}\sum_{s=1}^{m}\frac{a_{is}a_{js}}{k(m_s)}$$

where $k(t)$ is the degree of the t node, m_i is the i-th node of class M, t_i is the i-th node of class T, a_{ij} are the entries of A and λ is a tuning parameter, which is optimized to increase prediction

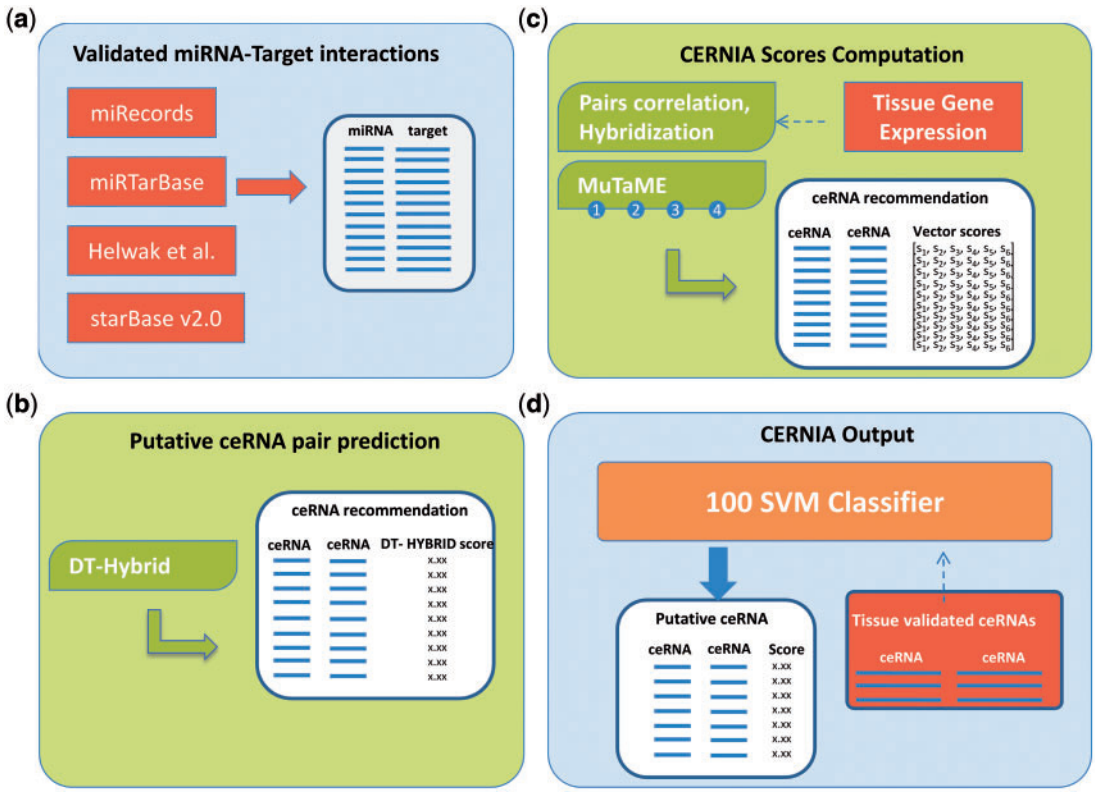

Figure 3. CERNIA pipeline. (**A**) All the validated and high-confidence miRNA–target interactions are collected and merged into a unique data set. The data set is used by the 1st ceRNA pair prediction step performed by a recommendation algorithm called DT-Hybrid (**B**). (**C**) For each pair, we calculated MREs and hybridization energy with miRanda. This score added to the MuTaME score, the DT-Hybrid recommendation score and the correlations between gene expression values for a specific tissue type, form the CERNIA vector of seven scores. (**D**) The tissue validated ceRNAs, available in online repositories, and the CERNIA score vectors are used to classify (using SVMs) a subset of the gene pairs given in (**B**) as the CERNIA putative ceRNAs. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

quality. The λ parameter is related to the quantity and quality of the predictions. A value close to one implies an equal distribution of the weights on the network, resulting in a greater weight for those pairs of targets who share a large pool of miRNAs. Vice versa, a value close to zero implies a distribution of weights based on a nearest-neighbor average, implying a greater number of lower-quality predictions, as the number of shared miRNAs is ignored.

The final scores of the recommendation are as follows:

$$S = W \cdot A$$

Each element in the matrix S represents the suitability of the interaction between a miRNA i and its predicted target j. It indicates the degree of belief of the interaction between a ceRNA pair, obtained on the basis of the number of miRNAs, which simultaneously target both competitors. A higher value implies both a greater number of common miRNAs and a greater prediction.

ceRNAs cross talk score

For each pair of putative ceRNAs, we computed a vector of measures containing (see Supplementary materials, Section *Scoring Function* to have more details):

1. The fraction of common miRNAs;
2. The density of the MREs for all shared miRNAs;
3. The distribution of MREs of the putative ceRNAs;
4. The relation between the overall number of MREs for a putative ceRNA, compared with the number of miRNAs that yield these MREs;
5. The density of the hybridization energies related to MREs for all shared miRNAs;
6. The DT-Hybrid recommendation scores; and
7. The pairwise Pearson correlation between putative ceRNA expressions from selected tissue.

Our score vector extends MuTaME [16], which includes only the points 1, 2, 3 and 4. It is used within the classification step as described below.

CERNIA classification

All putative ceRNA pairs obtained through the previous step are filtered with a classification methodology. To do that, CERNIA uses expression profiles. We keep only ceRNA pairs for which both transcripts are expressed and partition the ceRNA pairs in two subsets: validated/high confidence and predicted. As the size of validated/high-confidence pairs is much smaller than predicted ones, we generated 100 training sets with a comparable number of validated/high-confidence and predicted pairs. The subsets of predicted pairs are considered as non-ceRNAs (i.e. false positives). For each pair, we computed the CERNIA score. Our aim is to identify a thresholding to discern true-positive ceRNA pairs from false-positive ones. CERNIA uses 100 SVMs [54–57] trained on the built data sets. SVMs training and evaluation were used by the e1071 package of the R system. The classifier was trained using a radial basis kernel function, and its parameter was computed as the inverse of the number of dimensions of our training data. The procedure was repeated 10 times. For each putative ceRNA, CERNIA reports a final score that is the percentage of the SVMs that agree in classifying such a pair as ceRNA.

Results on validation and comparison on ceRNA predictions

CERNIA validation

We collected a set of validated and high-confidence miRNA–target interactions from both curated databases (see Supplementary Materials, Section *Data sets*) and CLASH immunoprecipitation experiments on three specific cancer types for which they are known the greatest numbers of validated ceRNAs: breast invasive carcinoma (BRCA), prostate adenocarcinoma (PRAD) and glioblastoma multiforme (GBM).

Validated tissue-specific ceRNA interactions are retrived from [27, 58] and miRSponge [59], a manually curated database containing miRNA-mediated interactions. We found 13 interactions for GBM, 5 for BRCA and 6for PRAD (see Supplementary Materials, Section *Validated ceRNAs*, Table S1).

We selected a set of 5641 GBM expressed genes from [52]. We extended it by adding the tissue-specific validated ceRNAs obtaining, respectively, 5642 BRCA expressed genes and 5643 PRAD expressed genes. We refer to such sets as 'gene sets'.

We calculated the scores for all possible pairs of genes in the BRCA, PRAD and GBM gene sets, and performed the classification procedure described in previous Section. These scores represent the percentage of SVMs that agree in classifying a pair as ceRNA. We estimated a 'classification threshold' for each tissue, separately, by computing receiver operating characteristic (ROC) curves for our ensemble of classifiers [60] and chose the third quantile of the best cutoffs distribution that maximize the area under curve (AUC). The estimated average areas (AUCs) are 0.96, 0.89 and 0.91 for BRCA, PRAD and GBM (Figure 4). We classified as ceRNA 1 348 392 pairs for BRCA, 6357 for PRAD and 1 035 222 for GBM.

CERNIA reaches the best sensitivity for BRCA compared with the other two tissue types. In the Supplementary File S2, we report the complete list of validated pairs, whereas in the Supplementary Materials, Section *CERNIA predictions*, Supplementary Figure S1–S3, we report the distributions of the CERNIA scores together with the scores for the validated pairs for each tissue type.

ceRNA predictors comparison

In Figure 5, we report a comparison of CERNIA with starBase and ceRDB. To perform a fair comparison, we used in CERNIA only the scores 1–6 (see section *ceRNAs cross talk score*) because both starBase and ceRDB do not use tissue-specific expression data. We used the validated pairs in BRCA, PRAD and GBM to classify the putative ceRNA pairs. Whereas in Figure 6, CERNIA is compared with a general conditional mutual information methodology (CMI) by applying all the seven scores in the classification step and by using 100 permutations to calculate the *P*-value. We used starBase putative ceRNA pairs, and the restricted set of PTEN ceRNA interactions for CMI and ceRDB comparison. (We focused on PTEN interactions because the computation of CMI on the whole set of pairs was not feasible, and ceRDB web interface accepted only single-gene queries.)

CERNIA predicts 10 validated ceRNA cross talk out of 13 in GBM, 5 out of 5 in BRCA and 3 out of 6 in PRAD. We used the hypergeometric distribution to compute the probability of obtaining by chance the number of validated ceRNAs within the set of predicted interactions. For each data set, we generated the number of possible ceRNA pairs. Such a value was therefore used within the hypergeometric distribution. For all data sets, we got low probabilities of obtaining by chance the validated

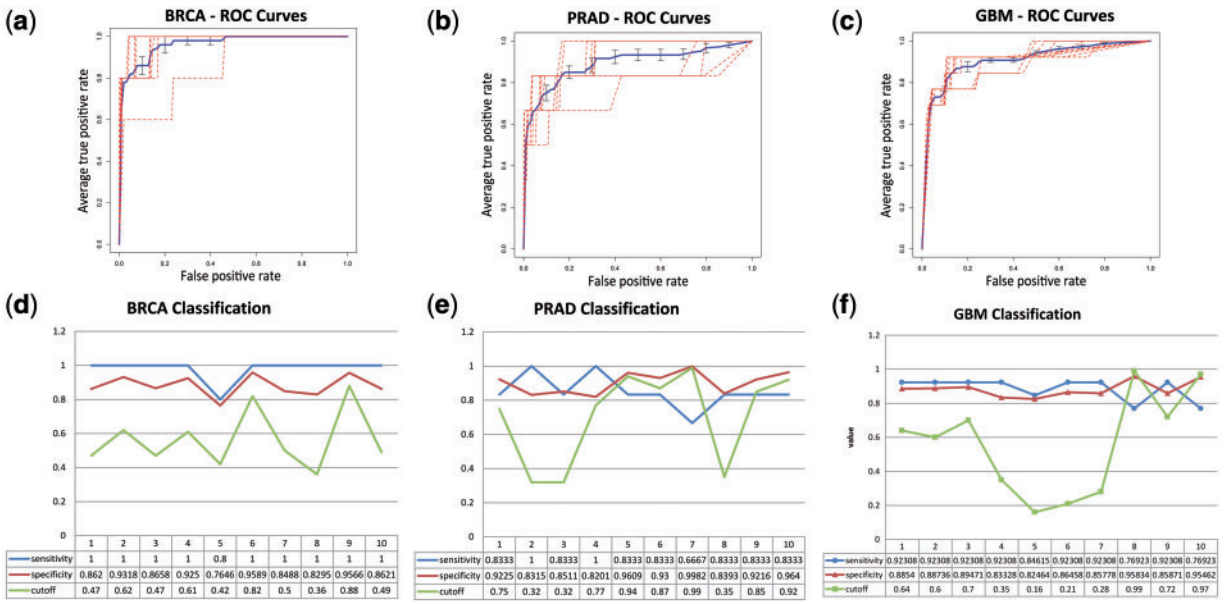

Figure 4. (**A–C**) are ROC curves computed for BRCA, GBM and PRAD, respectively. Curves are obtained by averaging the 10 repeats on the 100 SVMs. In dotted red, we report the original curves. In the (**D–F**), we report curves for sensitivity (blue), specificity (red) and tissue-specific cutoff (green). A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

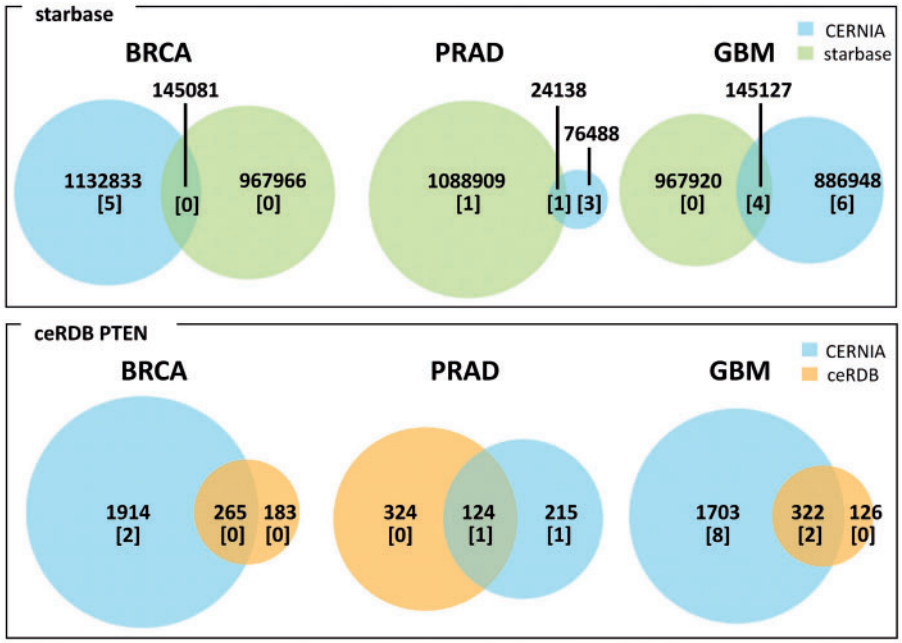

Figure 5. Comparisons between CERNIA, starBase and ceRDB. starBase has the largest number of predictions, whereas we focused on ceRDB's PTEN ceRNA interactions. starBase and ceRDB do not distinguish between tissue type; therefore, we trained CERNIA without using correlations between tissue-specific gene expressions, and carried out the comparison using BRCA, PRAD and GBM validated ceRNA pairs. Reported in square brackets is the number of validated ceRNA interactions for each set. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

interactions: for GBM, the probability is $3.3397{\cdot}10^{-6}$; for BRCA, it is $4.9562{\cdot}10^{-6}$; and for PRAD, it is $3.0904{\cdot}10^{-10}$. In Supplementary Materials, Section *CERNIA predictions*, Table S2 reports the probability of obtaining by chance the number of validated ceRNAs within the set of predicted interactions as function of the scoring threshold. The threshold chosen by ROC curve analysis is the most accurate. Furthermore, we can observe that obtaining by chance such validated pairs is improbable.

starBase contains >1 million predicted ceRNA interactions. The results contain 4 of 13 validated ceRNA interactions in GBM, 0 of 5 in BRCA and 2 of 6 in PRAD. For starBase, the probabilities to get by chance the validated predictions are considerably higher than CERNIA: for GBM, the probability is $8.9210{\cdot}10^{-3}$; for BRCA, it is 0.6959; and for PRAD, it is $5.4875{\cdot}10^{-2}$.

CERNIA and starBase give around the same number of predictions, which are ∼145 000 ceRNA interactions in BRCA, ∼24 000 in PRAD and ∼145 000 in GBM. The number of BRCA-specific and GBM-specific interactions predicted by CERNIA is comparable with starBase. However, CERNIA predicts a smaller set of putative ceRNAs for PRAD.

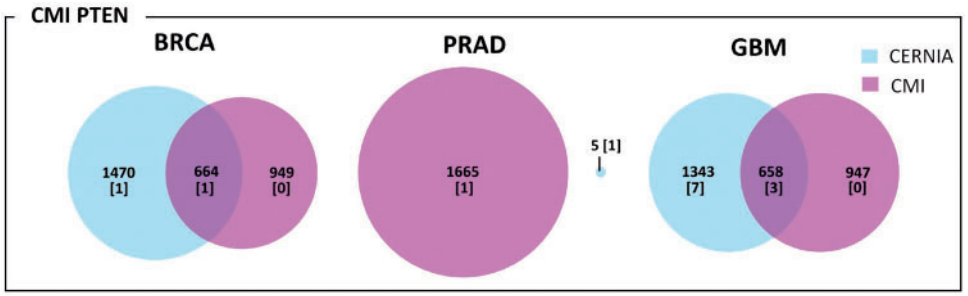

Figure 6. Comparisons between CERNIA and CMI. We focused on PTEN predictions for BRCA, PRAD and GBM. Reported in square brackets is the number of validated ceRNA interactions identified for each set. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

CERNIA is able to predict the majority of PTEN ceRNA interactions reported in literature for GBM and PRAD with respect to ceRDB. Here, CERNIA has again low probabilities (BRCA, 0.1492; PRAD, $1.9097{\cdot}10^{-2}$; GBM, $2.6564{\cdot}10^{-3}$) compared with ceRDB (BRCA, 0.8475; PRAD, 0.2479; GBM, 0.1982).

We compared CERNIA with a general CMI methodology. We used the code implemented in Hermes [52] software. We considered the combination of PTEN and genes in the tissue gene sets obtaining 5641 pairs in BRCA, 5642 in PRAD and 5640 in GBM. We used the same significance threshold reported in [52] with *P*-value $<10^{-4}$. We computed the probabilities to get by chance the PTEN validated interactions, CMI has the following values 0.4084, 0.4136 and 0.2317 in BRCA, PRAD and GBM, respectively, whereas CERNIA achieves the same probabilities: 0.1492 in BRCA, $1.9097{\cdot}10^{-2}$ in PRAD and $2.6564{\cdot}10^{-3}$ in GBM. CERNIA and CMI have in common 664, 0 and 658 predicted ceRNA interactions with PTEN in BRCA, PRAD and GBM, respectively. Both algorithms identify almost an equal number of tissue-validated ceRNA interactions. CERNIA, however, yields better results in GBM.

Finally, we selected BRCA, PRAD and GBM predictions from [50] and compared them with CERNIA. The results showed that their method is not able to predict any of the validated ceRNA interactions. We observed a similar behavior analyzing the BRCA CERNIA's predictions by using the data in Additional file 10 in [49]. These data use 0.3 as threshold that corresponds to the 99th percentile of the sensitivity correlation values distribution. We did not found any common prediction with CERNIA. Furthermore, among their 1103 ceRNA predicted pairs, no validated ceRNA interactions could be found.

The stability of CERNIA performances with respect to ceRDB, starBase and CMI, may be because of the usage of both validated and high-confidence miRNA–target interactions, and a broad set of established ceRNA cross talk rules (MuTaME, DT-Hybrid, gene expression correlation). Furthermore, CERNIA exploits MREs in the whole transcribed sequence.

Functional annotation of CERNIA predictions

We investigate the functional role of CERNIA predictions. For each tissue, we selected the genes within CERNIA-predicted interactions, and performed a functional annotation, Gene Ontology and pathway enrichment of such genes by using MSigDB [61].

The majority of putative ceRNAs (>700) for BRCA are TFs (80 are also oncogenes). In total, 180 are both oncogenes and translocated cancer genes. The same analysis performed for PRAD reported similar results (>20 predicted ceRNAs are tumor suppressors). About 50 genes are both translocated cancer genes and TFs (see Supplementary Materials, Section *GO enrichment*, Tables S3–S5 for a comprehensive description).

MSigDB is used to compute overlaps with known gene sets. We selected the putative ceRNAs in common between BRCA, PRAD and GBM tissues, whose amount is 2505. After, we selected the 'hallmark gene sets' from MSigDB [61] and retained the top 20 most significant overlaps (see Supplementary File S3).

We also performed topological analysis, and looked for cancer-specific genes among putative ceRNAs in the Catalog of Somatic Mutation in Cancer [62]. The results are described in the Supplementary Materials, Section *Topological Analysis* and are reported in Supplementary File S4.

Conclusions and discussion

The competing endogenous (ceRNA) effect is a posttranscriptional regulatory mechanism, which has been studied and acknowledged for many classes of RNA, both protein coding and noncoding. One of the most interesting outcome of the cross talk between such RNAs is that the ceRNA effect is used as a method to easily predict a function for many unknown RNAs in normal and pathological conditions. However, it still remains to be understood if this RNA competition is part of the system biology, and if the ceRNA regulation will help finding suitable druggable targets.

We reviewed the state of the art of *in vivo* and *in silico* ceRNA prediction methods. We developed a new *in silico* method, called CERNIA, which relies on both validated and high-confidence miRNA–target interactions, considering not only 3′ UTR miRNA binding sites but also 5′ UTR and coding sequence (CDS) ones [4, 5]. CERNIA can be used to study the ceRNA competition among different tissue types, and different classes of genes. Compared with other prediction methods and databases, such as Hermes, ceRDB and starBase, CERNIA shares a large amount of putative competing pairs. However, it is also capable to infer novel ceRNA interactions, which could allow extending the current understanding of the competing endogenous effect phenomena.

In vivo studies on mice are a starting point to uncover the role of specific RNAs in activating cancer pathways *in vivo*. In [63], the authors validated the ceRNA cross talk between BRAFP1 pseudogene and BRAF in human, then performed an *in vivo* study on the murine counterparts, Braf-rs1 and B-raf. Their findings supported the hypothesis that such a pseudogene induces lymphomas in mice. In some human DLBCL cell lines, the overexpression of BRAFP1 caused an increase in BRAF levels and cell proliferation. On the other hand, the ceRNA cross talk has aroused some criticisms [36, 64–66]. In [36], the authors proved the inability of derepressing a target by means of a single miRNA family. This is because of the high amount of endogenous MREs concentration needed to repress the miRNA function, which is impossible to reach for a single

target. Another important aspect in ceRNA cross talk validation refers to the methodology used to quantify the contribution of all endogenous MREs for a specific miRNA. However, the authors did not exclude the possibility that this huge number of binding sites is provided by different lncRNAs, suggesting a viable analysis of the noncoding landscape in diseases. They also concluded that it is more likely to have miRNA-mediated cross talk between mRNAs in conditions of suddenly changes of RNA concentration, such as in cell differentiation or cancer. In [67], the authors uncover a lncRNA called BC032469, which acts as a ceRNA competing for miR-1207-5p with human gene TERT, and another lncRNA called OIP5-AS1/cyrano bound to HuR [68]. The investigation of unknown lncRNAs landscape by using a combination of Ago CLIP-supported or CLASH and RNA-seq experiments can improve the knowledge and the functional characterization of transcripts [27].

The authors in [66] report both the difficulties related to technical, and informatics methodologies, in measuring and validating the ceRNA effects. The first refers to the bias of the argonaute (Ago) HITS-CLIP protocol, which does not take into account the localization of the Ago proteins, the specificity of the RNA binding or uses Dicer-deficient cells. The latter basically refers to miRNA–target prediction algorithm for 3′ UTR-dependent binding sites, and lack of targets gene expression in the prediction model.

The Ago–miRNA complex repression and the ceRNA effect can be quantitatively studied with small RNA-seq, poly-A RNA-seq and iCLIP techniques, as performed in [65]. They also performed transcriptome reconstruction, and transcripts quantification with Cufflink in mouse embryonic stem cell in the presence or absence of Ago protein. This approach, exploiting RNA-seq, is able to find novel RNA isoforms, resulting in a more accurate computation of MREs. The iCLIP experiments were used to find Ago-binding sites, and measure their affinity to vary the concentration. The authors concluded that Ago binding happens when miRNA and target sites have approximately the same amount of copies. In [64], the authors tried to bridge the gap between simulated and experimental results deducing that an mRNA containing a couple of MREs with normal expression within a cell for a miRNA cannot significantly cross talk with other mRNAs, although the same conclusion has to be examined for lncRNAs. More in general, cross talk is likely to occur when the concentration of a specific MRE sequence is comparable with the total amount of competing binding sites in the cell [64].

Finally, we can summarize, by considering the ceRNA cross talk state of art, that an emphasis is needed on environment studies and prediction models regarding miRNA–target predictions (in particular those with noncanonical binding sites) and cellular localization of the actor molecules [miRNA, transcripts RNA, TFs and RNA-binding proteins (RBPs)] to understand the dynamic nature of the cell regulatory network.

Key Points

- ceRNA posttranscriptional cross talk needs to be deepened with *in vivo* experiments, and validated as a widespread biological regulatory mechanism.
- RNA-seq, CLIP-seq and CLASH are powerful techniques, which can be exploited to improve the understanding of the ceRNA effect.
- The number of prediction methods for ceRNA cross talks is increasing, and much more accurate methods are expected in the future together with more validated data.
- Improve ceRNAs prediction methods to accurately predict functions for many unknown RNAs in normal and pathological conditions.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgments

The results shown here are in whole or in part based on data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

Funding

This work was supported by the Gruppo Nazionale per il Calcolo Scientifico (GNCS – INDAM).

References

1. Salmena L, Poliseno L, Tay Y, *et al.* A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language? *Cell* 2011;**146**:353–8.
2. Seitz H. Redefining microRNA targets. *Curr Biol* 2009;**19**:870–3.
3. Friedman RC, Farh KK-H, Burge CB, *et al.* Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res* 2008;**19**:92–105.
4. Helwak A, Kudla G, Dudnakova T, *et al.* Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell* 2013;**153**:654–65.
5. Hausser J, Syed AP, Bilen B, *et al.* Analysis of CDS-located miRNA target sites suggests that they can effectively inhibit translation. *Genome Res* 2013;**23**:604–15.
6. Xie P, Liu Y, Li Y, *et al.* MIROR: a method for cell-type specific microRNA occupancy rate prediction. *Mol Biosyst* 2014;**10**:1377–84.
7. Khorshid M, Hausser J, Zavolan M, *et al.* A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat Methods* 2013;**10**:253–5.
8. Memczak S, Jens M, Elefsinioti A, *et al.* Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 2013;**495**:333–8.
9. Franco-Zorrilla JM, Valli A, Todesco M, *et al.* Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* 2007;**39**:1033–7.
10. Poliseno L, Salmena L, Zhang J, *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* 2010;**465**:1033–8.
11. Cazalla D, Yario T, Steitz JA, *et al.* Down-regulation of a host microRNA by a Herpesvirus saimiri noncoding RNA. *Science* 2010;**328**:1563–6.
12. Hansen TB, Jensen TI, Clausen BH, *et al.* Natural RNA circles function as efficient microRNA sponges. *Nature* 2013;**495**:384–8.
13. Place RF, Li L-C, Pookot D, *et al.* MicroRNA-373 induces expression of genes with complementary promoter sequences. *Proc Natl Acad Sci USA* 2008;**105**:1608–13.

14. Huang V, Place RF, Portnoy V, *et al*. Upregulation of Cyclin B1 by miRNA and its implications in cancer. *Nucleic Acids Res* 2011;**40**:1695–707.
15. Mukherji S, Ebert MS, Zheng GXY, *et al*. MicroRNAs can generate thresholds in target gene expression. *Nat Genet* 2011; **43**:854–9.
16. Tay Y, Kats L, Salmena L, *et al*. Coding-independent regulation of the tumor suppressor PTEN by competing endogenous mRNAs. *Cell* 2011;**147**:344–57.
17. Pink RC, Wicks K, Caley DP, *et al*. Pseudogenes: Pseudofunctional or key regulators in health and disease? *RNA* 2011;**17**:792–98.
18. Poliseno L, Marranci A, Pandolfi PP. Pseudogenes in human cancer. *Front Med* 2015;**2**: 68
19. Marques AC, Tan J, Lee S, *et al*. Evidence for conserved post-transcriptional roles of unitary pseudogenes and for frequent bifunctionality of mRNAs. *Genome Biol* 2012;**13**:R102.
20. Hu W, Alvarez-Dominguez JR, Lodish HF. Regulation of mammalian cell differentiation by long non-coding RNAs. *EMBO Rep* 2012;**13**:971–83.
21. Cesana M, Cacchiarelli D, Legnini I, *et al*. A long noncoding RNA controls muscle differentiation by functioning as a competing endogenous RNA. *Cell* 2011;**147**:358–69.
22. Capel B, Swain A, Nicolis S, *et al*. Circular transcripts of the testis-determining gene Sry in adult mouse testis. *Cell* 1993;**73**:1019–30.
23. Guarnerio J, Bezzi M, Jeong JC, *et al*. Oncogenic role of fusion-circRNAs derived from cancer-associated chromosomal translocations. *Cell* 2016;**165**:289–302.
24. Hansen TB, Wiklund ED, Bramsen JB, *et al*. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J* 2011;**30**:4414–22.
25. Hansen TB, Kjems J, Damgaard CK. Circular RNA and miR-7 in cancer. *Cancer Res* 2013;**73**:5609–12.
26. Le TD, Zhang J, Liu L, *et al*. Computational methods for identifying miRNA sponge interactions. *Brief Bioinform* 2017, 1–14. 2017;**18**:577–90.
27. Tay Y, Rinn J, Pandolfi PP. The multilayered complexity of ceRNA crosstalk and competition. *Nature* 2014;**505**:344–52.
28. Hsu S-D, Tseng Y-T, Shrestha S, *et al*. miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res* 2013;**42**:D78–85.
29. Xiao F, Zuo Z, Cai G, *et al*. miRecords: an integrated resource for microRNA-target interactions. *Nucleic Acids Res* 2008;**37**:D105–10.
30. Alaimo S, Bonnici V, Cancemi D, *et al*. DT-Web: a web-based application for drug-target interaction and drug combination prediction through domain-tuned network-based inference. *BMC Syst Biol* 2015;**9(Suppl 3)**:S4.
31. Alaimo S, Pulvirenti A, Giugno R, *et al*. Drug-target interaction prediction through domain-tuned network-based inference. *Bioinformatics* 2013;**29**:2004–8.
32. Ala U, Karreth FA, Bosia C, *et al*. Integrated transcriptional and competitive endogenous RNA networks are cross-regulated in permissive molecular environments. *Proc Nat Acad Sci USA* 2013;**110**:7154–9.
33. Yuan Y, Liu B, Xie P, *et al*. Model-guided quantitative analysis of microRNA-mediated regulation on competing endogenous RNAs using a synthetic gene circuit. *Proc Nat Acad Sci USA* 2015;**112**:3158–63.
34. Bosia C, Pagnani A, Zecchina R. Modelling competing endogenous RNA Networks. *PLoS One* 2013;**8**:e66609.
35. Kertesz M, Iovino N, Unnerstall U, *et al*. The role of site accessibility in microRNA target recognition. *Nat Genet* 2007;**39**: 1278–84.
36. Denzler R, Agarwal V, Stefano J, *et al*. Assessing the ceRNA hypothesis with quantitative measurements of miRNA and target abundance. *Mol Cell* 2014;**54**:766–76.
37. Flores M, Huang Y. A new algorithm for predicting competing endogenous rnas. In *Proceedings 2012 IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)* 2012, *ie eexplore digital library*. Doubletree by Hilton, Crystal City, VA, Washington, DC, USA, pp 118–121.
38. Liu H, Yue D, Chen Y, *et al*. Improving performance of mammalian microRNA target prediction. *BMC Bioinform* 2010; **11**:476.
39. Yue D, Guo M, Chen Y, *et al*. A Bayesian decision fusion approach for microRNA target prediction. *BMC Genomics* 2012;**13(Suppl 8)**:S13.
40. Liu K, Yan Z, Li Y, *et al*. Linc2GO: a human LincRNA function annotation resource based on ceRNA hypothesis. *Bioinformatics* 2013;**29**:2221–2.
41. Betel D, Wilson M, Gabow A, *et al*. The microRNA.org resource: targets and expression. *Nucleic Acids Res* 2007;**36**:D149–53.
42. Betel D, Koppal A, Agius P, *et al*. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* 2010;**11**:R90.
43. Cabili MN, Trapnell C, Goff L, *et al*. Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011;**25**: 1915–27.
44. Li J-H, Liu S, Zhou H, *et al*. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. *Nucleic Acids Res* 2013;**42**:D92–7.
45. Das S, Ghosal S, Sen R, *et al*. lnCeDB: database of human long noncoding RNA acting as competing endogenous RNA. *PLoS One* 2014;**9**:e98965.
46. Ghosal S, Das S, Sen R, *et al*. HumanViCe: host ceRNA network in virus infected cells in human. *Front Genet* 2014;**5**:249.
47. Shackelford J, Pagano JS. Tumor viruses and cell signaling pathways: deubiquitination versus ubiquitination. *Mol Cell Biol* 2004;**24**:5089–93.
48. Hayward SD. Viral interactions with the Notch pathway. *Semin Cancer Biol* 2004;**14**:387–96.
49. Paci P, Colombo T, Farina L. Computational analysis identifies a sponge interaction network between long non-coding RNAs and messenger RNAs in human breast cancer. *BMC Syst Biol* 2014;**8**:83.
50. Xu J, Li Y, Lu J, *et al*. The mRNA related ceRNA–ceRNA landscape and significance across 20 major cancer types. *Nucleic Acids Res* 2015;**43**:8169–82.
51. Sarver AL, Subramanian S. Competing endogenous RNA database. *Bioinformation* 2012;**8**:731–3.
52. Sumazin P, Yang X, Chiu H-S, *et al*. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma. *Cell* 2011;**147**:370–81.
53. Chiu H-S, Llobet-Navas D, Yang X, *et al*. Cupid: simultaneous reconstruction of microRNA-target and ceRNA networks. *Genome Res* 2014;**25**:257–67.
54. Bennett KP, Campbell C. Support vector machines: Hype or Hallelujah? *ACM SIGKDD Explor Newsl* 2000;**2**:1–13.
55. Cortes C, Vapnik V. Support-vector networks. *Mach Learn* 1995;**20**:273–97.
56. Chang C-C, Lin C-J. LIBSVM: a lybrary for support vector machines. *ACM Trans Intell Syst Technol* 2011;**2**:27.
57. Schölkopf B, Smola AJ, Williamson RC, *et al*. New support vector algorithms. *Neural Comput* 2000;**12**:1207–45.

58. Sanchez-Mejias A, Tay Y. Competing endogenous RNA networks: tying the essential knots for cancer biology and therapeutics. *J Hematol Oncol* 2015;**8**:30.
59. Wang P, Zhi H, Zhang Y, *et al*. miRSponge: a manually curated database for experimentally supported miRNA sponges and ceRNAs. *Database* 2015;**2015**:bav098.
60. Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2008;**37**:1–13.
61. Liberzon A, Subramanian A, Pinchback R, *et al*. Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 2011;**27**:1739–40.
62. Forbes SA, Beare D, Gunasekaran P, *et al*. COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res* 2014;**43**:D805–11.
63. Karreth FA, Reschke M, Ruocco A, *et al*. The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma in vivo. *Cell* 2015;**161**:319–32.
64. Jens M, Rajewsky N. Competition between target sites of regulators shapes post-transcriptional gene regulation. *Nat Rev Genet* 2014;**16**:113–26.
65. Bosson AD, Zamudio JR, Sharp PA. Endogenous miRNA and target concentrations determine susceptibility to potential ceRNA competition. *Mol Cell* 2014;**56**:347–59.
66. Thomson DW, Dinger ME. Endogenous microRNA sponges: evidence and controversy. *Nat Rev Genet* 2016;**17**:272–83.
67. Lü M-H, Tang B, Zeng S, *et al*. Long noncoding RNA BC032469, a novel competing endogenous RNA, upregulates hTERT expression by sponging miR-1207-5p and promotes proliferation in gastric cancer. *Oncogene* 2015;**35**:3524–34.
68. Kim J, Abdelmohsen K, Yang X, *et al*. LncRNA OIP5-AS1/cyrano sponges RNA-binding protein HuR. *Nucleic Acids Res* 2016;**44**:2378–92.