RESEARCH

Arena-Idb: a platform to build human non-coding RNA interaction networks

Vincenzo Bonnici[1], Giorgio De Caro[2], Giorgio Constantino[1], Sabino Liuni[2], Domenica D'Elia[2], Nicola Bombieri[1], Flavio Licciulli[2†] and Rosalba Giugno[1*†]

*Correspondence: rosalba.giugno@univr.it
[1]Department of Computer Science,University of Verona, Strada Le Grazie, Verona, Italy
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

Background: High throughput technologies have provided the scientific community an unprecedented opportunity for large-scale analysis of genomes. Non-coding RNAs (ncRNAs), for a long time believed to be non-functional, are emerging as one of the most important and large family of gene regulators and key elements for genome maintenance. Functional studies have been able to assign to ncRNAs a wide spectrum of functions in primary biological processes, and for this reason they are assuming a growing importance as a potential new family of cancer therapeutic targets. Nevertheless, the number of functionally characterized ncRNAs is still too poor if compared to the number of new discovered ncRNAs. Thus platforms able to merge information from available resources addressing data integration issues are necessary and still insufficient to elucidate ncRNAs biological roles.

Results: In this paper, we describe a platform called Arena-Idb for the retrieval of comprehensive and non-redundant annotated ncRNAs interactions. Arena-Idb provides a framework for network reconstruction of ncRNA heterogeneous interactions (i.e., with other type of molecules) and relationships with human diseases which guide the integration of data, extracted from different sources, via mapping of entities and minimization of ambiguity.

Conclusions: Arena-Idb provides a schema and a visualization system to integrate ncRNA interactions that assists in discovering ncRNA functions through the extraction of heterogeneous interaction networks. The Arena-Idb is available at `http://arenaidb.ba.itb.cnr.it`

Keywords: Non-coding RNA; Database; Network; Data integration

Background

The availability of omics repositories represents a powerful resource for the discovery of interactions among non coding RNAs (ncRNAs). The association of metadata to ncRNAs allows researchers to exploit their full potential for inferring new molecular functions. Molecular interactions involve several types of entities including Long non-coding RNAs (lncRNAs) and Small non-coding RNAs (sncRNAs), further divided into subclasses shortly called biotypes. According to HUGO Gene Nomenclature Committee (HGNC) [1], the sncRNAs (see Table 1) are classified into various biotypes of short sequences such as Small interfering RNAs (siRNAs), microRNAs (miRNAs), PIWI-interacting RNAs (piRNAs), small nuclear RNAs (snRNAs), small nucleolar RNAs (snoRNAs), and small cytoplasmic RNAs (scRNAs). The lncRNAs have a broader spectrum of functions [2, 3] such as regulation of

transcription, RNA processing, nuclear-cytoplasmic transport, translation control and modulation of chromatin structure and are, therefore, a potential new class of cancer therapeutic targets [4]. In addition to these classes of ncRNAs there are other different types of ncRNAs whose role is under discovering. The circular RNA are highly active in brain cells and play an important role in neurodegenerative disease and encoding of proteins [5]. The rigorous characterization of the biological functions of extracellular RNAs (exRNAs) in biofluids is a rapidly growing area of research to monitor diseases with a promising use in diagnostic [6].

In physiological conditions, many biological entities interact with each another and are key regulators of many cellular processes and contribute to a multitude of diseases [7]. Understanding a biological interactions system demands understanding the details of its components, and their interactions. Available public biological resources provide narrowed but systematic overviews of relationship schema among biological entities. For example, an individual miRNA may regulate multiple mRNAs, and in contrast, an individual gene may also be regulated by multiple miRNAs, thus representing a complex network of miRNA-mRNA interactions. More recently, other layers of regulation have added further complexity in regulatory networks. It has been proposed that the binding of microRNAs to their targets can be buffered by transcripts mimicking the sequences of the true targets, therefore protecting them from repression; these transcripts have been called competitive endogenous RNAs (ceRNAs) [8, 9]. If these ceRNAs possess many miRNAs response elements (MREs) and are expressed at high enough levels, they act to sequester miRNAs [10]. Many existing databases are unified catalogues of annotations, sequences and expression information for human ncRNAs [11, 12, 13, 14, 15, 16, 17, 18, 19]. These databases are frequently developed only in the contest of one or few biotypes of ncRNAs and without the integration of diseases associations. Tools such as the ones reported in [20, 21, 22] provide an integration procedure which does not verify sequence similarity and is mostly focused on genes, proteins and in some cases miRNAs [23]. Moreover, none of these databases provide an integrated vision of relationships between different ncRNA biotypes and other entities [24, 25]. In this paper, we present a computational framework (Arena-Idb) to realize non-coding RNA-Gene regulatory networks. Arena-Idb addresses the gap of existing methods providing a framework for network reconstruction of ncRNA heterogeneous interactions (i.e., with other type of molecules) and relationships with human diseases which guides the integration of data extracted from different sources via mapping of entities and minimization of ambiguity. Arena-Idb handles knowledge regarding biological products (i.e., information linking transcribed RNA and translated proteins to their corresponding source genes, thus from DNA to RNA or protein, and from RNA to protein) and cross-references (i.e., the mapping between different nomenclature systems). To keep non-redundant sequences it filters the information by comparing cross-link references and sequence similarity using the Cleanup software [26]. Compared to its previous version [27], Arena-Idb provides (i) a mapping procedure for managing entities, (ii) improving the accuracy of the integration process by identifying the sequence entity, (iii) reconstructed data storage and update including seven new sources as Disease Ontology, lnc2cancer, lncACTdb, mir2disease, miRecords, mirSponge, PSMIR, StarBase and TarBase, (iv) a more functional web

interface that provides many new features such as, among others, a browser section that allows users to visualize, filter and download data by different criteria; a search section that enables queries also for chromosomal location; and a network visualization system that also allows the download of data in a readable format for Cytoscape import. The Arena-Idb can be accessed or downloaded as whole integration system at `http://arenaidb.ba.itb.cnr.it` .

Construction and content

The construction of Arena-Idb is realized through a series of sequential steps that go from the collection of data from different ncRNA and interaction databases to the mining and integration of data for the construction of heterogeneous interaction networks. An overview of the process developed for the integration of input data sources is shown in Figure 1. A initial non-redundant collection of ncRNAs is built by performing object recognition via sequence identity. Interaction sources, that also contain other types of objects, are integrated by cross-link identity recognition. The result of the integration contains information about the objects, the interactions between ncRNAs and integrated objects and biological products from genes to ncRNAs. Figures 2, 3, 4, and 5 give the details of the integration process summarized into four steps. We first describe how data are extracted and represented in Arena-Idb, than we describe each integration step sequentially.

Data content

The Arena-Idb data storage is implemented using two different Database Management Systems (DBMS): i) a Relational DBMS, MySQL release 5.5, and ii) a Graph DBMS, neo4J community edition 3.1.3. The MySQL database stores data about names, annotations and sequences and it is used to efficiently query ncRNAs and to optimize the retrieval of associated annotations and sequences information. The Graph DBMS efficiently handles the construction and visualization of the networks of thousands of biological entities (nodes) and relations (edges). We use the relation part of the data storage also to facilitate the integration in Arena-Idb of new data sources (often released as relational DBMSs). We developed specific procedures in Cypher Query Language for the data porting from relational DBMS to Neo4J which automatically ingest relationships and graph information about alias, multi-resources referencing and biological entities interactions.

Table 2 reports the data sources integrated in Arena-Idb together with further information such as the type of extracted biological entities. To gather data from all sources we implemented customized Extract, Transformation and Load (ETL) procedures for data available in different forms: TSV (Tab-separated values), CSV (Comma- separated values), and Biomart/Ensembl instances that are queried and processed by REST API, R procedures and Pentaho Data Integration (Kettle) scripts (http://www.pentaho.com/product/data-integration). Sequence data in Arena-Idb are loaded by using REST Biomart API calls for VEGA/HAVANA and ENSEMBL ncRNAs, by parsing the Genbank entries files (GBFF flat files) downloaded from NCBI FTP using BioJava API calls, and by parsing downloadable fasta formatted files from mirBase, GtRNAdb, and pirnaBank. Tables 3 and 4 report the total amount of entities and interactions, respectively, that result in Arena-Idb at the end of the integration process.

Arena-Idb stores biological entities according to their biological classes (gene, pseudogene, pcRNA, ncRNA, protein, phenotype, other) and biotype. A biotype is a consensus classification of entities by their physical or functional characteristics, for example the distinction between long non-coding RNAs and microRNAs or circulating RNAs (`http://vega.archive.ensembl.org/info/about/gene_and_transcript_types.html`).

Biological entities are often reported in multiple sources. Some of them define an internal nomenclature system, called also namespace, and assign new identifiers to entities. Some others use existing identifiers assigned in external namespaces. We refer to those identifiers as RIDs (Reference-ID). More precisely, a RID is a pair of strings, the first one refers to the reference namespace, and the second string reports the identifier within the namespace (for example HGNC:29665). Most reference sources also provide mappings between internal and external RIDs, such mappings are called cross-references.

In Arena-Idb, RIDs are stored apart from entities, and may be linked to multiple entities, possibly with different entity classes. Interactions are stored as tuples containing the internal identifiers of the interacting biological entities, the names and versions of the original data sources, the tools predicting the interactions (if they are not validated), and the PubmedIDs of the scientific articles reporting them together with supporting sentences from the bibliography.

Identity by sequence: detection of redundant non-coding RNAs by sequence similarity

The first step of the Arena-Idb pipeline integrates sources of non-coding RNA sequences into a non-redundant collection of ncRNA objects. The task is performed by using the Cleanup tool [26], a fast program for removing redundancies from nucleotide sequence databases. Sequences having high grade of identity and overlap, in the same biological biotype, are purged.

Figure 2 shows an input resource providing two ncNRAs with associated sequences s_1 and s_2. The partial collection already contains the ncRNAs having sequences s_1 and s_3. The integration tool recognizes the two ncRNAs having sequences s_1 as the same object, and produces an updated non-redundant collection composed by s_1, s_2, and s_3. The collection of data obtained by merging all the sequence sources is used as base in Arena-Idb for the successive integration steps.

Identity by alias: detection of redundant entities by RIDs comparisons

RIDs in a namespace are designed to be specific of a given object, and cross-references are supposed to help in mapping entities between different namespaces. However, cross-references do not map every namespace to another, and they may introduce inconsistency and ambiguity. As a result, biological entities may share one or several identifiers, making the task of recognizing them as distinct objects a bottleneck on the integration process. In addition, input source may have a lack of information. Mining procedures in Arena-Idb allow deducing missing data. For example, for entities without reported biological classes, Arena-Idb finds out their classes by searching for entities with a similar set of linked RIDs. Arena-Idb follows an order of resource integration corresponding to the amount of information provided by each source (miRTarBase, HMDD, miR2Disease, miRecords, miRandola, circ2Traits, NPInter, miRSponge, starBase, lncACTdb, Psmir, TarBase, Lnc2Cancer, LncRNADisease , lncRNAdb).

The integration procedures are performed by comparing the sets of RIDs associated with them. For every input entity, if the current collection contains an entity with a comparable set of RIDs, then the input entity is matched to it, otherwise the entity is added up to the collection.

Figure 3 shows two input RIDs having the same label that is *microRNA 144* but associated with objects of different class, a ncRNA and a gene. In the current state of Arena-Idb the RID related to *microRNA 144* is mapped to a ncRNA. Therefore, the input ncRNA and the one already in Arena-Idb are recognized as the same object. On the contrary, the input gene does not have a correspondence in Arena-Idb, thus it is added to it, together with its linked RID. Entities of different classes but having same RIDs are real examples of transcripts named with the same label used for their producer genes. Figure 4 shows the import of a cross-reference linking two RIDs, *microRNA 144* and *hsa-mir-144*, that are referred to the same ncRNA object. The current state of Arena-Idb already contains a ncRNA object labelled with *microRNA 144* but missing of the *hsa-mir-144* RID. The identity by aliases approach implemented by Arena-Idb recognizes the equivalence of the two objects, since they have the same label *microRNA 144* in common, and the integration procedure updates, with the additional RID *hsa-mir-144*, the information linked to the ncRNA.

Figure 5 reports a real example of transcripts sharing one or more RIDs, possibly because they are isoforms of the same gene. The input source contains a ncRNA with two RIDs: *HOTAIR* and *ENST00000424518*. The procedure maps the input entity with the ncRNA having a complete match with the set of aliases of the input ncRNA, while the ncRNA associated to *ENST00000453875* partially overlap the set. Figure 5 gives also an example of cross-references. Once entities of an input source are mapped to those already contained in the database, the information regarding interactions and additional cross-references is added to Arena-Idb. As a result, the step unifies the plenty of integrated sources and provides a higher comprehensive view of the currently known information regarding interactions in which ncRNAs are involved.

Finally, during the integration, customized procedures regarding miRNAs and disease names are applied. Arena-Idb adds, to the miRNA entities, additional RIDs that refers to miRNA genes (see `http://www.mirbase.org/help/nomenclature.shtml`). Regarding phenotype entities, in presence of RIDs containing parenthesis, names are split into two or more identifiers. Arena-Idb also defines a set of regular expressions to express all extracted RIDs identifiers (e.g., HGNC:[0-9] refers to HGNC IDs). Since RIDs may lack of reference source names, the integration procedure approximately matches the incomplete RID against a set of regular expressions in order to assign the correct namespace.

Detection of primary names

A final step of integration is performed to assign a single representative RID, called primary name, to every biological entity. The algorithm extracts subsets of entities belonging to the same biological class and sharing at least one RID. In order to choose the primary names, the algorithm takes into account two properties regarding RIDs. First, it defines the following order of trustiness resources: miRBase,

VEGA, RefSeq, Ensembl, GtRNAdb, piRNABank, snoRNABase, Entrez, and all the other not listed resources have the same preference order. Second, it counts the number of entities that are linked to a given RID. Identifiers with fewer entities are preferred. The described combinatorial approach is hard to solve cause every possible combination of RIDs to entities must be scanned. Since, similar combinatorial problems are well-known in literature, such as the stable marriage problem, we represent entities and RIDs in a bipartite network and apply heuristics to reduce the computational time needed to find a solution for the mapping. Briefly, entities with the fewest number of RIDs linked to them are accounted firstly, and the sets of their RIDs are sorted by the above precedence's list.

Data update

Data update is performed by re-running globally or partially the ETL procedures. More precisely, we can summarize the database population procedure into two main steps. In the first step, semi-automatic ETL procedures (tailored to each input sources) gather data from external primary sources, producing a homogeneous representation of input resources and merge it into a single knowledge base. In the second step, the external interaction sources are parsed and all the interactions among the mates are built. Therefore, a main update of Arena-Idb involves the execution of all the ETL procedures to build the database from scratch. However, updating a single external source only consist of the execution of the scripts related to that source in the first and second phase. Furthermore, the normalization performed by the first ETL phase allows to add new external resources to the system without substantial modification of the overall procedure, the database maintainer can execute only the ETL script related to the new source using the developed ETL as template.

Utility and discussion

The Arena-Idb provides an easy-to-use graphical web interface and graphical visualization to facilitate the retrieval of ncRNAs interactions. The Graphical User Interface (GUI) has been developed as JAVA Web Application in Java Platform Enterprise Edition - Java EE. It uses jQuery/jQuery-UI framework JavaScript on the client layer, Java servlets and JavaServer Pages (jsp) on the server layer. The web application is deployed in a Tomcat web server (`https://tomcat.apache.org`). The Hibernate ORM (Object Relational Mapping, `http://hibernate.org/orm/`) has been adopted to implement the communication between the data layer (MySQL and Neo4j) and the Web Application. It also provides a framework for mapping an object-oriented domain model to relational and graph databases enabling us to handle the data layer as objects in the web pages.

Arena-Idb provides two modes to access to data, Search and Browser. Browser lists in a tabular mode all pairs of interacting entities in Arena-Idb reporting their tuples of information (as described in Data content section). User can browse by RNA-RNA, RNA-gene, RNA-Protein, and RNA-Disease interaction.

The Search mode allows to retrieve ncRNAs using the following criteria: by ncRNA/gene name, by genomic coordinates, and by disease name (see Figure 6). When one starts typing ncRNA/gene name or disease name into the search box,

suggested ncRNA/gene or disease names are displayed in the list box. The end user chooses one of the names associated to the biological entity from the list box. In order to use the search by genomic coordinates the user chooses the number of the chromosome and the starting and ending positions of the desired region in that chromosome. All run queries are listed and can be retrieved in MY SEARCH section.

The results of the search are given as a set of ncRNA cards (see Figure 6). The user can click on the icons in the top of each card to: (i) show in tabular form a detailed page reporting information such as genomic locations, synonyms, sequence, and the list of interactions; (ii) to show interactively the interactions represented graphically as a network; and (iii) to download the interactions in a format compatible with advanced network mining and visualization platforms such as Cytoscape (`http://www.cytoscape.org/`) or as text file in FASTA and TSV format.

Furthermore, Arena-Idb creates a whole network of interactions by merging all the retrieved entities and adding to the network all possible interactions stored in Arena-Idb among them. This can be visualized by clicking on the icon on the top right of query result bar, see Figure 6 (a). The merging can also be done gradually under the guide of the user by adding one at a time interacting entity or type of interaction. The obtained global (merged) network can be downloaded in tabular format as described above.

Networks can be filtered by deselecting entity types (protein, ncRNA, pcRNA, disease, gene and other) and thus removing all nodes of such types and their edges. The edges are associated to scores representing the number of resources reporting such interactions. Scores range from 1 to the total number of integrated resources. A dark grey corresponds to a high score. Clicking on a node, a tooltip window displays all the associated RIDs (name and aliases), while clicking on the arch a tooltip with the score number and the type of interaction is displayed. User can navigate inside the displayed network by zooming in/out. Clicking on a node, Arena-Idb also highlights the node itself and its neighborhood.

Case studies

As an example of Arena-Idb usefulness we describe the case of hsa-mir-4732. Figure 7 shows the interaction network extracted by Arena-Idb searching for hsa-mir-4732. The interaction network indicates that hsa-mir-4732 is related to hsa-miR-449a, hsa-miR-142-3p and hsa-miR-144-3p. Looking at the genomic location of this microRNAs we found that hsa-miR-144, and hsa-mir-4732 are transcribed as a polycistronic gene. Many of the known miRNAs are distributed across chromosomes either individually or in cluster, in which two or more miRNA genes are located within a short distance on the same segment of a chromosome. The miRNA cluster arose through a complex history of duplication and loss of individual members as well as duplication of the entire cluster. Several studies suggest a role of the miRNA-144 cluster in the complex regulation of the expression of genes involved in different diseases and relationships in the hsa-mir-4732 network extracted by Arena-Idb found meaningful evidence in the literature [28, 29, 30, 31, 32, 33].

An additional example is represented by a circular RNAs (circRNAs), CDR1as. Genome-wide analyses have identified a large number of abundant circRNAs that

represents a recent addition to the growing list of ncRNA classes [34, 35]. CircRNAs can arise from exons (exonic circRNA) or introns (intronic circRNA) and act as miRNA sponges thus playing a role in mediating miRNA targeting. The Figure 8 shows the interaction network extracted by Arena-Idb for CDR1as. In particular it shows a strong relationship with miR-7 (score: 4, sources: lncrnadb) and miR-671 (score: 2, source: lncrnadb), two miRNAs whose activity is affected by CDR1as, as reported by the Kjems laboratory [36].

Conclusion

ncRNAs are crucial for many biological processes. Despite many studies have indicated the importance of ncRNAs in different tissues and diseases, little is known about their biological functions and interactions. New complex interactions among ncRNAs, and between ncRNAs and diseases, have emerged [37, 38, 39]. Research on the functional and clinical role of ncRNAs in molecular biological processes with implications in human diseases has exploded since they were discovered a decade ago, implying a proliferation of online resources to store ncRNAs and their interactions. These databases are frequently developed only in the contest of one or few types of ncRNAs, and they miss in providing an integrated vision of the relationships between different ncRNA classes and other entities. The advantages that Arena-Idb provides to end-users is the availability of a framework for reconstruction of networks of ncRNA interactions with other biological entities and diseases, that can be modelled on-demand and filtered for more specific interactions depending on the users needs. Another important feature is the minimization of ambiguities that in the case of the ncRNAs represents a big problem due to missing effective standards for their nomenclature and heterogeneity of resources used. Moreover, ArenaIdb can be downloaded as a whole system to customize additional resources integration. All together these features make of ArenaIdb an exhaustive and useful reference for user to explore at large any type of interaction and to discover unforeseeable functional role of not yet characterized ncRNAs.

Competing interests
The authors declare that they have no competing interests.

Ethics approval and consent to participate
Not applicable.

Consent to publication
Not applicable.

Acknowledgements
We thank the Fondo Sociale Europeo provided by Regione del Veneto for partially supported this work.

Author's contributions
All authors have read and approved the final manuscript.

Availability of data and materials
Data and materials are available at the web site *http://arenaidb.ba.itb.cnr.it.*

Founding
This work has been partially supported by the following projects: GNCS-INDAM, Fondo Sociale Europeo, and National Research Council Flagship Projects Interomics. This work has been partially supported by the project of the Italian Ministry of Education, Universities and Research (MIUR) "Dipartimenti di Eccellenza 2018-2022". Publication costs have been founded by the Department of Computer Science, University of Verona (Italy), and the Institute for Biomedical Technologies, National Research Council (CNR) (Italy).

Author details
[1]*Department of Computer Science,University of Verona, Strada Le Grazie, Verona, Italy.* [2]*Institute for Biomedical Technologies, National Research Council (CNR), Bari, Italy.*

References

1. *Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., Bruford, E.A.: Genenames. org: the hgnc resources in 2015. Nucleic acids research, 1071 (2014)*
2. *Batista, P.J., Chang, H.Y.: Long noncoding RNAs: cellular address codes in development and disease. Cell* **152***(6), 1298–1307 (2013)*
3. *Guttman, M., Rinn, J.L.: Modular regulatory principles of large non-coding RNAs. Nature* **482***(7385), 339–346 (2012)*
4. *Qureshi, I., Mehler, M.: Non-coding rna networks underlying cognitive disorders across the lifespan. Trends in Molecular Medicine* **17***(337-46) (2011)*
5. *Pamudurti, N.R., Bartok, O., Jens, M., Ashwal-Fluss, R., Stottmeister, C., Ruhe, L., Hanan, M., Wyler, E., Perez-Hernandez, D., Ramberger, E., Shenzis, S., Samson, M., Dittmar, G., Landthaler, M., Chekulaeva, M., Rajewsky, N., Kadener, S.: Translation of circrnas. Molecular Cell* **66***(1), 9–21 (2017)*
6. *Russo, F., Di Bella, S., Vannini, F., Berti, G., Scoyni, F., Cook, H.V., Santos, A., Nigita, G., Bonnici, V., Lagan, A., Geraci, F., Pulvirenti, A., Giugno, R., De Masi, F., Belling, K., Jensen, L.J., Brunak, S., Pellegrini, M., Ferro, A.: mirandola 2017: a curated knowledge base of non-invasive biomarkers. Nucleic Acids Research, 854 (2017). doi:10.1093/nar/gkx854*
7. *Wang, J., Cao, Y., Zhang, H., Wang, T., Tian, Q., Lu, X., Lu, X., Kong, X., Liu, Z., Wang, N., Zhang, S., Ma, H., Ning, S., Wang, L.: Nsdna: a manually curated database of experimentally supported ncrnas associated with nervous system diseases. Nucleic Acids Research* **45***(D1), 902–907 (2017). doi:10.1093/nar/gkw1038*
8. *Sardina, D.S., Alaimo, S., Ferro, A., Pulvirenti, A., Giugno, R.: A novel computational method for inferring competing endogenous interactions. Briefings in Bioinformatics, 084 (2016). doi:10.1093/bib/bbw084*
9. *Zarringhalam, K., Tay, Y., Kulkarni, P., Bester, A.C., Pandolfi, P.P., Kulkarni, R.V.: Identification of competing endogenous rnas of the tumor suppressor gene pten: A probabilistic approach. Scientific Reports* **7***(7755) (2017)*
10. *Salmena, L., Poliseno, L., Tay, Y., Kats, L., Pandolfi, P.P.: A cerna hypothesis: the rosetta stone of a hidden rna language? Cell* **146***(3), 353–8 (2011)*
11. *Fan, Y., Siklenka, K., Arora, S., Ribeiro, P., Kimmins, S., Xia, J.: mirnet - dissecting mirna-target interactions and functional associations through network-based visual analysis. Nucl. Acids Res.* **44***(W135141) (2016)*
12. *Russo, F., Di Bella, S., Bonnici, V., Laganà, A., Rainaldi, G., Pellegrini, M., Pulvirenti, A., Giugno, R., Ferro, A.: A knowledge base for the discovery of function, diagnostic potential and drug effects on cellular and extracellular miRNAs. BMC Genomics* **15***(3), 1–7 (2014)*
13. *Cava, C., Colaprico, A., Bertoli, G., Graudenzi, A., Silva, T., Olsen, C., Noushmehr, H., Bontempi, G., Mauri, G., Castiglioni, I.: Spidermir: An r/bioconductor package for integrative analysis with mirna data. Int J Mol Sci* **18***(2) (2017)*
14. *Fiannaca, A., La Rosa, M., La Paglia, L., Messina, A., Urso, A.: Biographdb: a new graphdb collecting heterogeneous data for bioinformatics analysis. In: BIOTECHNO (2016)*
15. *Liu, C., Gao, C., Ma, Z., Cong, R., Zhang, Q., Guo, A.: lncrinter: A database of experimentally validated long non-coding rna interaction. J Genet Genomics* **44***(5), 265–268 (2017)*
16. *Paraskevopoulou, M.D., Vlachos, I.S., Karagkouni, D., Georgakilas, G., Kanellos, I., Vergoulis, T., Zagganas, K., Tsanakas, P., Floros, E., Dalamagas, T., Hatzigeorgiou, A.G.: Diana-lncbase v2: indexing microrna targets on non-coding transcripts. Nucleic Acids Research* **44***(D1), 231–238 (2016). doi:10.1093/nar/gkv1270*
17. *Zhou, K.-R., Liu, S., Sun, W.-J., Zheng, L.-L., Zhou, H., Yang, J.-H., Qu, L.-H.: Chipbase v2.0: decoding transcriptional regulatory networks of non-coding rnas and protein-coding genes from chip-seq data. Nucleic Acids Research* **45***(D1), 43–50 (2017). doi:10.1093/nar/gkw965*
18. *Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., Zhang, Z., Skogerb, G., Chen, L., Lu, H., Zhao, Y., Chen, R.: Npinter: the noncoding rnas and protein related biomacromolecules interaction database. Nucleic Acids Research* **34***(suppl_1), 150–152 (2006). doi:10.1093/nar/gkj025*
19. *Guo, Z., Maki, M., Ding, R., Yang, Y., Zhang, B., Xiong, L.: Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues. Scientific Reports* **4***(5150) (2014)*
20. *Pareja-Tobes, P., Tobes, R., Manrique, M., Pareja, E., Pareja-Tobes, E.: Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv, 016758 (2015)*
21. *Smith, R.N., Aleksic, J., Butano, D., Carr, A., Contrino, S., Hu, F., Lyne, M., Lyne, R., Kalderimis, A., Rutherford, K., et al.: Intermine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics* **28***(23), 3163–3165 (2012)*
22. *Vera, R., Perez-Riverol, Y., Perez, S., Ligeti, B., Kertesz-Farkas, A., Pongor, S.: Jbiowh: an open-source java framework for bioinformatics data integration. Database* **2013** *(2013)*
23. *Fiannaca, A., La Rosa, M., La Paglia, L., Messina, A., Urso, A.: Biographdb: a new graphdb collecting heterogeneous data for bioinformatics analysis. Proceedings of BIOTECHNO (2016)*
24. *Leung, Y.Y., Kuksa, P.P., Amlie-Wolf, A., Valladares, O., Ungar, L.H., Kannan, S., Gregory, B.D., Wang, L.-S.: Dashr: database of small human noncoding rnas. Nucleic Acids Research* **44***(D1), 216–222 (2016). doi:10.1093/nar/gkv1188*
25. *Consortium, T.R.: Rnacentral: a comprehensive database of non-coding rna sequences. Nucleic Acids Research* **45***(D1), 128–134 (2017). doi:10.1093/nar/gkw1008*
26. *Grillo, G., Attimonelli, M., Liuni, S., Pesole, G.: Cleanup: a fast computer program for removing redundancies from nucleotide sequence databases. Comput Appl Biosci.* **12***(1), 1–8 (1996)*
27. *Bonnici, V., Russo, F., Bombieri, N., Pulvirenti, A., Giugno, R.: Comprehensive reconstruction and visualization of non-coding regulatory networks in human. FRONTIERS IN BIOENGINEERING AND BIOTECHNOLOGY* **69***(2), 1–22 (2014)*

28. *Gao, Z., Liu, R., Liao, J., Yang, M., Pan, E., Yin, L., Pu, Y.: Possible tumor suppressive role of the mir-144/451 cluster in esophageal carcinoma as determined by principal component regression analysis. Molecular medicine reports* **14***(4), 3805–3813 (2016)*
29. *Liu, L., Wang, S., Chen, R., Wu, Y., Zhang, B., Huang, S., Zhang, J., Xiao, F., Wang, M., Liang, Y.: Myc induced mir-144/451 contributes to the acquired imatinib resistance in chronic myelogenous leukemia cell k562. Biochemical and Biophysical Research Communications* **425***(), 368–372 (2012)*
30. *LC, D., JD, A., CO, D.S., Z, Z., X, G., JW, T., et al.: A gata-1- regulated microrna locus essential for erythropoiesis. Proc Natl Acad Sci USA* **105**, *3333–8 (2008)*
31. *Zhang, X., Wang, X., Zhu, H., Zhu, C., Wang, Y., Pu, W.T., Jegga, A.G., Fan, G.-C.: Synergistic effects of the gata-4-mediated mir-144/451 cluster in protection against simulated ischemia/reperfusion-induced cardiomyocyte death. Journal of Molecular and Cellular Cardiology* **49**, *841–850 (2010)*
32. *Wang, X., Zhu, H., Zhang, X., Liu, Y., Chen, J., Medvedovic, M., Li, H., Weiss, M.J., Ren, X., , Fa, G.-C.: Loss of the mir-144/451 cluster impairs ischaemic preconditioning-mediated cardioprotection by targeting rac-1. Cardiovascular Research* **94***(379390) (2012)*
33. *Rasmussen, K.D., Simmini, S., Abreu-Goodger, C., Bartonicek, N., Giacomo, M.D., Bilbao-Cortes, D., Horos, R., Lindern, M.V., Enright, A.J., OCarroll, D.: The mir-144/451 locus is required for erythroid homeostasis. J. Exp. Med.* **207***(7), 1351–1358 (2012)*
34. *Jeck, W.R., Sorrentino, J.A., Wang, K., Slevin, M.K., Burd, C.E., Liu, J., Marzluff, W.F., Sharpless, N.E.: Circular rnas are abundant, conserved, and associated with alu repeats. Rna* **19***(2), 141–157 (2013)*
35. *Salzman, J., Gawad, C., Wang, P.L., Lacayo, N., Brown, P.O.: Circular rnas are the predominant transcript isoform from hundreds of human genes in diverse cell types. PloS one* **7***(2), 30733 (2012)*
36. *Piwecka, M., Glažar, P., Hernandez-Miranda, L.R., Memczak, S., Wolf, S.A., Rybak-Wolf, A., Filipchyk, A., Klironomos, F., Jara, C.A.C., Fenske, P., et al.: Loss of a mammalian circular rna locus causes mirna deregulation and affects brain function. Science* **357***(6357), 8526 (2017)*
37. *Keniry, A., et al.: The h19 lincrna is a developmental reservoir of mir-675 that suppresses growth and igf1r. Nat. Cell Biol.* **14**, *659–665 (2012)*
38. *Emmrich, S.e.t.a.: mir-99a/100 125b tricistrons regulate hematopoietic stem and progenitor cell homeostasis by shifting the balance between tgf and wnt signaling. Genes Dev.* **28**, *858–874 (2014)*
39. *Emmrich, S.e.a.: Lincrnas monc and mir100hg act as oncogenes in acute megakaryoblastic leukemia. Mol. Cancer* **13***(171) (2014)*
40. *Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D.G., et al.: The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome research* **22***(9), 1775–1789 (2012)*
41. *Ashurst, J., Chen, C.-K., Gilbert, J.G., Jekosch, K., Keenan, S., Meidl, P., Searle, S., Stalker, J., Storey, R., Trevanion, S., et al.: The vertebrate genome annotation (vega) database. Nucleic acids research* **33***(suppl_1), 459–465 (2005)*
42. *Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al.: The ensembl genome database project. Nucleic acids research* **30***(1), 38–41 (2002)*
43. *Kozomara, A., Griffiths-Jones, S.: miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic acids research, 1181 (2013)*
44. *Pruitt, K.D., Tatusova, T., Maglott, D.R.: Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research* **33***(suppl_1), 501–504 (2005)*
45. *Chan, P.P., Lowe, T.M.: Gtrnadb: a database of transfer rna genes detected in genomic sequence. Nucleic acids research* **37***(suppl_1), 93–97 (2008)*
46. *Sai Lakshmi, S., Agrawal, S.: pirnabank: a web resource on classified and clustered piwi-interacting rnas. Nucleic acids research* **36***(suppl_1), 173–177 (2007)*
47. *Schriml, L.M., Arze, C., Nadendla, S., Chang, Y.-W.W., Mazaitis, M., Felix, V., Feng, G., Kibbe, W.A.: Disease ontology: a backbone for disease semantic integration. Nucleic acids research* **40***(D1), 940–946 (2012)*
48. *Ghosal, S., Das, S., Sen, R., Basak, P., Chakrabarti, J.: Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. Frontiers in genetics* **4** *(2013)*
49. *Li, Y., Qiu, C., Tu, J., Geng, B., Yang, J., Jiang, T., Cui, Q.: HMDD v2. 0: a database for experimentally supported human microRNA and disease associations. Nucleic acids research, 1023 (2013)*
50. *Ning, S., Zhang, J., Wang, P., Zhi, H., Wang, J., Liu, Y., Gao, Y., Guo, M., Yue, M., Wang, L., et al.: Lnc2cancer: a manually curated database of experimentally supported lncrnas associated with various human cancers. Nucleic acids research, 1094 (2015)*
51. *Wang, P., Ning, S., Zhang, Y., Li, R., Ye, J., Zhao, Z., Zhi, H., Wang, T., Guo, Z., Li, X.: Identification of lncrna-associated competing triplets reveals global patterns and prognostic markers for cancer. Nucleic acids research, 233 (2015)*
52. *Amaral, P.P., Clark, M.B., Gascoigne, D.K., Dinger, M.E., Mattick, J.S.: lncRNAdb: a reference database for long noncoding RNAs. Nucleic acids research* **39***(suppl 1), 146–151 (2011)*
53. *Chen, G., Wang, Z., Wang, D., Qiu, C., Liu, M., Chen, X., Zhang, Q., Yan, G., Cui, Q.: LncRNADisease: a database for long-non-coding RNA-associated diseases. Nucleic acids research* **41***(D1), 983–986 (2013)*
54. *Jiang, Q., Wang, Y., Hao, Y., Juan, L., Teng, M., Zhang, X., Li, M., Wang, G., Liu, Y.: miR2Disease: a manually curated database for microRNA deregulation in human disease. Nucleic acids research* **37***(suppl 1), 98–104 (2009)*
55. *Xiao, F., Zuo, Z., Cai, G., Kang, S., Gao, X., Li, T.: mirecords: an integrated resource for microrna–target interactions. Nucleic acids research* **37***(suppl 1), 105–110 (2009)*
56. *Hsu, S.-D., Lin, F.-M., Wu, W.-Y., Liang, C., Huang, W.-C., Chan, W.-L., Tsai, W.-T., Chen, G.-Z., Lee, C.-J., Chiu, C.-M., et al.: mirtarbase: a database curates experimentally validated microrna–target interactions. Nucleic acids research, 1107 (2010)*
57. *Wang, P., Zhi, H., Zhang, Y., Liu, Y., Zhang, J., Gao, Y., Guo, M., Ning, S., Li, X.: mirsponge: a manually curated database for experimentally supported mirna sponges and cernas. Database* **2015**, *098 (2015)*

58. *Liu, C., Bai, B., Skogerbø, G., Cai, L., Deng, W., Zhang, Y., Bu, D., Zhao, Y., Chen, R.: Noncode: an integrated knowledge database of non-coding rnas. Nucleic acids research* **33***(suppl 1), 112–115 (2005)*
59. *Meng, F., Wang, J., Dai, E., Yang, F., Chen, X., Wang, S., Yu, X., Liu, D., Jiang, W.: Psmir: a database of potential associations between small molecules and mirnas. Scientific reports* **6** *(2016)*
60. *Yang, J.-H., Li, J.-H., Shao, P., Zhou, H., Chen, Y.-Q., Qu, L.-H.: starBase: a database for exploring microRNA–mRNA interaction maps from Argonaute CLIP-Seq and Degradome-Seq data. Nucleic acids research* **39***(suppl 1), 202–209 (2011)*
61. *Sethupathy, P., Corda, B., Hatzigeorgiou, A.G.: Tarbase: A comprehensive database of experimentally supported animal microrna targets. Rna* **12***(2), 192–197 (2006)*

Figures

Figure 1 Arena-Idb integration and content overview. *On the left, the Integration schema which shows the type of data extracted from each type of source used and the processes (sequence identity processing and cross-link identity) performed for to obtain the data stored and integrated into Arena-Idb (Content schema shown on the figures right side). The result of the integration process is a comprehensive database collecting information about the objects (genes and their products) and the interactions between ncRNAs and integrated objects.*

Figure 2 Arena-Idb integration process: identity by sequences. *The Figure gives an example of integration performed by sequence identity recognition. Two miRNAs, identified by miRBase symbols, are integrated into a partial state of Arena-Idb that contains two ncRNA, identified by their Ensembl IDs. The sequences of one of the two miRBase miRNAs is recognized in the partial state thus the miRBase symbol is added up to the list of aliases assigned to the miRNA object. Instead, no compatible sequences are found for the other miRBase miRNA. The results of the integration is a collection of three ncRNAs.*

Figure 3 Arena-Idb integration process: identity by aliases, example 1. *The Figure gives an example of integration regarding the addition of a ncRNA and a gene into a partial state of Arena-Idb that contains a ncRNA. The input ncRNA is labelled with a HGNC symbol that equals the identifier assigned to the ncRNA present in the partial state. Instead, there is no identifier that can match the gene symbol. The input information also report a biological production of the ncRNA from the given gene. The gene is added to the partial knowledge, the two ncRNAs are matched, and the biological relation is flushed.*

Figure 4 Arena-Idb integration process: indetity by aliases, example 2 *The Figure gives an example of integration that does not add any new object to the current knowledge, instead it extends the set of aliases linked with the existing object. The input ncRNA has no assigned sequence, thus a recognition by sequences is not available. The HGNC symbol is used to recognize the identity of the two ncRNAs, and the miRBase identifier is added to the list of the aliases linked with the ncRNA.*

Figure 5 Arena-Idb integration process: identity by aliases, example 3 *The Figure reports an example of integration of a ncRNA and a protein into a partial state of Arena-Idb. The ncRNA is labelled with two identifiers, a symbol and an Ensembl ID, and the protein is labelled only with a symbol. A interaction between the two objects is reported. In this case, two ncRNAs are already present in the partial knowledge. They are alternative transcripts of the HOTAIR gene, thus they are labelled with the symbol $HOTAIR$. However, the two ncRNAs can be distinguished by the specific Ensembl identifier linked with them. The integration procedure recognizes the identity of the input ncRNA with one of the two already present in the partial state by means of the Ensembl identifier. On the contrary, the input protein is directly mapped to a protein already in the partial state since no alias ambiguity arises. Finally, the biological interaction is flushed to the final knowledge base.*

Tables

Figure 6 Arena-Idb Search web interface. *Search is performed by ncRNA or gene name, by genomic coordinates, or by disease name. Here the search is performed by looking at 28861072 to 28861966 positions in chromosome 17. Results are memorized on query named 'chr17-q'. Arena-Idb returns 11 entities retrieved in the desired chromosome location. Each card shows the subnetwork of the retrieved entity. User can click on the buttons in top right (a) of the query result frame to visualize a global network obtained by merging all possible interactions among nodes in the retrieved subnetworks.*

Figure 7 Arena-Idb interaction network visualization of hsa-mir-4732. *The Figure shows the network visualization interface displaying the interactions regarding the miRNA primary transcript $hsa\text{-}mir\text{-}4732$. The transcript has a total of 9 interactions, three with other miRNAs and 6 with proteins. In the network visualization mode different colours are used to represent diverse biological entities (i.e. nodes): Protein (fuchsia), ncRNA (cyan), pcRNA (green), disease (purple), gene (orange), and other (pink). Arena-Idb also contains information about the generic miRNA $hsa\text{-}miR\text{-}4732$ (not in its primary transcript form), its 3' and 5' transcription and the corresponding gene. The box on the left side of the image shows the right selection of the transcript regarding the network example from the search interface.*

Figure 8 Arena-Idb interaction network visualization of CDR1as. *In [36] authors show that CDR1as causes miRNAs deregulation and affects brain function, in particular miR-7 and miR-671. Arena-Idb is able to retrieve and easy visualizes the strong cited relations, suggesting also further entities to investigate.*

Table 1 Overview of the major classes of ncRNAs: classification and functional characterization.

Symbol	Name	Size	Function
miRNAs	microRNAs	$18 - 24$ nt	They act as negative control of gene expression by silencing or catalysing mRNA destabilization.
snoRNAs	Small nucleolar RNAs	70 nt	Conserved nuclear RNA in Cajal bodies or nucleoli where they either function in the modification of snRNA or participate in the processing of rRNA ribosome subunit maturation.
snRNAs	Small nuclear RNAs	$100 - 300$ nt	RNA localized in the eukaryotic cell nucleus. They are part of splicesome multisubunit complex which assembles on RNA and carriers out RNA splicing. The snRNAs are classified in different type according of their role.
siRNAs	Small-interfering RNAs	$20 - 25$ nt	siRNA derived from much longer double stranded RNA (dsRNA) precursor by DICER ribonucleases and play a substantial role in genetic and epigenetic regulatory.
ceRNAs	Competitive endogenous RNAs	> 200 nt	ceRNAs are transcripts that can crosstalk through their ability to compete for mRNA binding and they act to sequester miRNAs.
circRNAs	Circular RNAs	> 200 nt	circRNAs arise from exons or intronics and may be also translate into protein. Exonic circRNAs are very stable in cell and have specific roles in cellular physiology.
piRNAs	PIWI-interacting RNAs	$25 - 35$ nt	piRNAs show specific expression in germ cells. Recent studies suggest that piRNA represents adaptive control mechanisms that protect genomics architectures again transposable elements (TE). Most piRNA are derived from genomic piRNA clusters.
lincRNAs	Long intergenic non-coding RNAs	> 200 nt	Perform various regulatory roles, but the majority remain functionally uncharacterized and typically low abundance and poor evolutionary conservation.
lncRNAs	Long non-coding	> 200 nt	lncRNAs are transcripts that lack RNAs apparent protein coding and are largely heterogeneous and functionally uncharacterized. The increasing evidence began to suggest that they play critical regulatory roles in manu human disease.

Table 2 List of the database resources with related information extracted and used in Arena-Idb platform. Legend: BI= Basic Information; S=Sequences; CR=Cross references;); ncRNAs (non-coding RNA); pcRNAs (protein coding RNA); G=Gene; Ps=Pseudogene; D=Disease; P=Protein, GO=Ontology, I=Interactions (NN:ncRNA-ncRNA, NM:ncRNA-pcRNA, NG:ncRNA-Gene, NS:ncRNA-Pseudogene, ND:ncRNA-Disease, NO:ncRNA-Others).

Database	Biological Entities extracted	Annotated Information	Description
HGNC [1]	ncRNA, pcRNA, G, D	BI, CR	A curated collection of approved Human Gene Nomenclature
Genecode [40]	ncRNA, pcRNA, G, PS	BI, S	Reference gene annotation and experimental validation for human and mouse.
VEGA/Havana [41]	ncRNA	BI, S	A repository for gene model produced by the manual annotation.
Ensembl [42]	ncRNA	BI, S, CR	Genome browser database for vertebrate with annotate gene.
miRBase[43]	ncRNA	BI, S	Database of of published miRNA sequences and annotation.
RefSeq [44]	ncRNA	BI, S	Collection of integrated, non-redundant and well annotated set of transcript and genomic data.
GtRNAdb [45]	ncRNA	BI, S	Genomic tRNA database.
piRNAbank [46]	ncRNA	BI, S	Resource on classified and clustered piRNAs.
Disease Ontology [47]	D, GO	CR	Database of standardized ontology of human disease.
Circ2Traits [48]	ncRNA, pcRNA, G, D	NN, NM, NG, ND	A comprehensive database of human circRNAs associated with diseases and traits.
HMDD [49]	ncRNA, G, D	NG, ND	A collection of experimentally supported human miRNAs and disease associations.
Lnc2Cancer [50]	ncRNA, D	CR, ND	A manually curated database of experimentally lncRNAs associated with cancer.
LncActDB [51]	ncRNA, D	NN, NG, ND	Database containing a list of lncRNA and mRNA with regulatory roles.
LncRNAdb [52]	ncRNA, G, P	NN, NG, NP	A database of functional lncRNAs.
LncRNADisease [53]	ncRNA, D	NP, ND	A curated DB of lncRNA with diseases.
Mir2diseases [54]	ncRNA, G, D	NG, ND	A manually curated database for miRNA deregulation in human diseases.
MiRandola [6]	ncRNA, D	ND	Collection of extracellular circulating miRNAs and their deregulation in human disease.
miRecords [55]	ncRNA, G	NG	A collection of validate miRNA target interaction with the exclusion of predicted interactions.
miRTarBase [56]	ncRNA, G	NG	A database of experimentally validate miRNA target interactions.
mirSponge [57]	ncRNA, pcRNA, G, Ps, D	NN, NM, NG, NP, ND	Manually curated database of miRNA spanges and ceRNAs.
NONCODE [58]	ncRNA	CR	A database of ncRNA with integrated only the Cross-References.
NPInter [18]	ncRNA, P	NN, NP	Database of experimentally verified interaction between ncRNA and other biomolecules.
PSMIR [59]	ncRNA	NO	A database of potential associations between small molecules and miRNAs.
StarBase [60]	ncRNA, G, P, Ps	NN, NG, NS, NP	A database of miRNA-mRNA interactions.
TarBase [61]	ncRNA, Gene	NG	A database of curated experimentally validate miRNA targets.

Table 3 List of the number of biotypes with alias present in Arena-Idb and the number of their interactions.

Name of Biotype	Total
ncRNA	170.919
pcRNA	4.987
Gene	51.599
Pseudogene	16.754
Protein	2.019
Disease	844
Other-Small molecule	1.309

Table 4 Number of interactions between different biological classes in the Arena-Idb platform.

Interactions type	Total
ncRNA-ncRNA	285.346
ncRNA-pcRNA	455.041
ncRNA-Gene	3.124.380
ncRNA-Pseudogene	24.589
ncRNA-Protein	126.702
ncRNA-Disease	64.278
ncRNA-other	150.535