bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SEA: The small RNA Expression Atlas

Raza-Ur Rahman[1,2], Abdul Sattar[1,2], Maksims Fiosins[1,2], Abhivyakti Gautam[1], Daniel Sumner Magruder[1,2], Jörn Bethune[1,2], Sumit Madan[3], Juliane Fluck[3], and Stefan Bonn[1,2,4,*]

[1]Laboratory of Computational Systems Biology, German Center for Neurodegenerative Diseases, Göttingen, Germany.

[2]Institute of Medical Systems Biology, Center for Molecular Neurobiology, University Clinic Hamburg-Eppendorf, Hamburg, Germany.

[3]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, Sankt Augustin, Germany.

[4]German Center for Neurodegenerative Diseases, Tübingen, Germany.

*Contact: sbonn@uke.de

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Small RNAs (sRNAs) are important biomolecules that exert vital functions in organismal health and disease, from viruses to plants, animals, and humans. Given the ever-increasing amounts of sRNA deep sequencing data in online repositories and their potential roles in disease therapy and diagnosis, it is important to enable federated sRNA expression querying across samples, organisms, tissues, cell types, and diseases. Here we present the sRNA Expression Atlas (SEA), a web application that allows for the search of known and novel small RNAs across ten organisms using standardized search terms and ontologies. SEA contains re-analyzed sRNA expression information for over 2000 published samples, including many disease datasets and over 700 novel, high-quality predicted miRNAs. We believe that SEA's simple interface and fast search in combination with its detailed interactive reports will enable researchers to better understand the potential function and diagnostic value of sRNAs across tissues, diseases, and organisms.

Availability and Implementation: SEA is implemented in Java, J2EE, Python, R, PHP and JavaScript. It is freely available at http://sea.dzne.de

* To whom correspondence should be addressed.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Introduction

Small RNAs (sRNAs) are a class of short, non-coding RNAs with important biological functions in nearly all aspects of organismal development in health and disease. Especially in diagnostic and therapeutic research sRNAs such as miRNAs and piRNAs received recent attention (Witwer, 2014). Reflecting the importance of sRNAs in biological processes as well as disease diagnosis and therapy is the increasing number of deep sequencing sRNA studies (sRNA-seq). To harvest the true potential of existing data it is important to allow for the querying, visualization, and analysis of sRNA-seq data across organisms, tissues, cell types, and disease states. This would allow researchers, for example, to search for disease-specific sRNA biomarker signatures across all disease entities investigated. Data integration and interoperability require (i) a streamlined analysis workflow to reduce analysis bias between experiments and (ii) also necessitate standardized annotation using ontologies to search and retrieve relevant samples. Here we are presenting the **s**mall-RNA **E**xpression **A**tlas (SEA), a web application that allows for the querying, visualization, and analysis of over 2000 published sRNA-seq expression datasets. SEA automatically downloads and re-analyzes published data using Oasis 2, annotates relevant meta-information using standardized terms, synchronizes sRNA information with other databases, allows for the querying of terms across ontological graphs, and presents quality curated sRNA expression information as interactive web reports (Capece *et al.*, 2015). It currently supports 10 organisms and is continuously updated with novel published sRNA-seq datasets and relevant sRNA information from various online resources.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

2 System Design

SEA stores sRNA expression information as well as deep and standardized meta-data on the samples, analysis workflows, and databases used. Data and meta-data information is normalized using ontologies to allow for standardized search and retrieval across ontological hierarchies (see section 2.3 for details). The following sections will detail the system design of SEA.

2.1 Acquisition of sRNA datasets

SEA acquires raw SRA files of published sRNA-seq datasets and their primary annotation from Gene Expression Omnibus (GEO) and NCBI's Sequence Reads Archive (SRA) repository. GEO makes two databases in SQLite format available for download: GEOmetadb for annotations and SRAdb for SRA sequences. An automated data acquisition pipeline searches for new sRNA data bi-weekly, keeping SEA continuously updated. Novel datasets are downloaded and stored in SEA's raw data repository while corresponding annotations are stored in SEA's annotation database. Raw data is subsequently processed automatically by SEA's sRNA analysis workflow (2.2) while annotations are processed automatically with SEA's annotation workflow (2.3). Processed files and annotations are subsequently semi-automatically curated.

2.2 Data analysis and storage

Following the acquisition of sRNA datasets, the SEA analysis workflow automatically analyzes new files using the Oasis 2.0 API (see biorxiv.org for latest manuscript) (Capece *et al.*, 2015) (https://oasis.dzne.de). The SEA analysis workflow determines data quality and detects and quantifies sRNAs, including the prediction of novel, high-quality miRNAs. Low quality files are flagged automatically and subjected to manual curation. Any files not passing manual curation are removed from SEA. Subsequently, sRNA counts of high-quality samples are stored in the sRNA expression database while corresponding quality information is saved in the data quality repository. SEA also stores expression information of high-quality predicted miRNAs including the ID, organism, chromosomal location, precursor and mature sequences, structure, read counts, prediction scores, and detailed information on the software and its versions used to predict the miRNA. SEA's primary analysis results including per sample quality and expression information can be examined and downloaded as interactive web reports. Detailed information on the primary analysis of sRNAs and predicted miRNAs can be found in the Oasis 2 manuscript (biorxiv.org).

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In order to reduce bias that could be introduced into the data by using different analysis routines, every sample in SEA has been analyzed by identical analysis workflows using identical databases and annotations. In case of changes in databases or analysis routines, SEA additionally stores versioning information about the software and databases used for an analysis. In addition, SEA contains information about the Geo series accession (GSE) and sample accession (GSM) identifiers along with the sample ID from the Sequence Read Archive (SRA) database (SRR) (Barrett *et al.*, 2013). Given that most meta-data is quite different between experiments we opted to store this expression data and meta-data in a Not Only SQL (NoSql) MongoDB[2] database management system. We optimized search and retrieval times by indexing for the most common queries and most relevant terms.

2.3 Standardized annotation

To allow for the interoperability of data it is important to standardize annotations using ontologies and semantic mapping (Schuurman and Leszczynski, 2008). Ontologies define standard terms, their properties, and the relations between them and dataset terms that are connected to Ontologies are called 'normalized'. The Ontologies and the number of normalized terms in SEA are listed in Table 1.

SEA's sRNA annotation workflow maps free-text GEO annotations to standardized terms in three consecutive steps. In general, GEO data annotations are free text that can be parsed into key-value pairs. In a first fully automated step the annotation workflow extracts key-value relations and stores them in the annotation database. As GEO data information is unstructured and contains very different information, we opted for a NoSql annotation database with an optimized indexing for prototypical questions (see also section 2.2).

The second fully automated step normalizes the extracted keys and values using Ontologies as standard dictionaries. SEA has a list of pre-defined keys, five of which (organism, tissue, disease, cell type, and cell line) can be currently queried for in SEA. Each extracted key is compared to pre-defined keys. For values, the ontologies are used as standard terminology dictionaries. For each pre-defined key, SEA has one of several corresponding ontologies. Each extracted value is searched in the corresponding ontologies and, if the same or a similar term is found, connected with it.

[2] https://www.mongodb.com/

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Automatic annotation is followed by semi-automatic manual curation. For that purpose, we developed an internal curation Web interface using Groovy/Grails[3], which allows browsing and editing of annotations from the annotation database as well as manual normalization of keys and values in annotations, searching among pre-defined keys and corresponding ontologies. Thus, curators examine all keys and values for consistency and update missing or additional information with standardized terms where necessary (e.g. protocols, kit version, lot and batch numbers, publications). At the moment, all SEA annotations are manually curated, a quality standard that we intend to keep for every future SEA entry.

2.4 Querying and visualization

To enable the search across ontological hierarchies we integrated the relevant ontologies into the graph database Neo4j[4] (Figure 1). Graph databases are NoSQL databases which support storage of objects and connections between them, as is the case for ontologies. Following the manual curation (see section 2.3), sample annotations are uploaded to the SEA ontology graph database including all ontological parent terms (having an 'is-a' relation to it). This allows search by ontology terms, as well as by their parents, which are in fact groups of terms (e.g. 'cancer' or 'neurodegenerative disease'). SEA accesses the ontology graph database via the Ontology Lookup Service using a REST interface, supporting complex and compound queries and query auto-completion (Côté *et al.*, 2010).

[3] https://grails.org/
[4] https://neo4j.com/

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3 Results & Conclusions

SEA is designed for the biological or medical end-user that is interested to define where and when an sRNA of interest is expressed. Prototypical questions that can be addressed with SEA are: What is the expression of hsa-miR-488-5p across all human tissues? Is hsa-miR-488-5p expressed higher in adenocarcinomas as compared to other cancer types? Is the tissue-specific expression of hsa-miR-488-5p conserved in mouse? Its unique selling points are the deep and standardized annotation of meta-information, the re-analysis of published data with Oasis 2 to reduce analysis bias, a user-friendly search interface that supports complex queries, and the fast and interactive visualization of analysis results across 10 organisms (Table 2) and various sRNA-species. SEA also contains information on the expression of over 700 high-quality predicted miRNAs, across organisms and tissues. Last but not least, SEA is continuously growing and aims to eventually encompass all sRNA-seq datasets across all organisms deposited in GEO and other repositories. Genome versions will be updated with every major release of SEA. SEA will be backwards compatible in the future by allowing users to choose previous genome versions and annotations. A detailed comparison of SEA to other existing sRNA expression databases highlights that SEA is superior in terms of supported organism, annotations, diseases, and tissues. SEA contains over 2000 samples in its database, which is considerably less than YM500v3 (Chung *et al.*, 2016), which hosts over 8000 cancer samples. It is to be noted, however, that the YM500v3 database only supports cancer datasets and no other disease types (Table 3).

As far as we are aware SEA is the only sRNA-seq database that supports ontology-based queries, supporting single or combined searches for five pre-defined keys (organism, tissue, disease, cell type, and cell line) across all datasets. However, the SEA database system contains additional (meta)-information including age, gender, developmental stage, genotype as well as technical experimental details such as the sequencing instrument and protocol details (e.g. library kit, RNA extraction procedure). We plan to normalize most of this additional information in future versions of SEA. This will allow users, for example, to query and analyze sRNA expression effects that are introduced by library kit or sequencing platform differences (both of these features can introduce large biases in the detection and expression of sRNAs). Other future developments will include information on sRNA editing, modifications, and mutation events.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In summary, SEA supports interactive result visualization on all levels, from querying and display of sRNA expression information to the mapping and quality information for each of the over 2000 samples. SEA is a fast, flexible, and fully interactive web application for the investigation of sRNA expression across cell lines, tissues, diseases, organisms, and sRNA-species. As such, SEA should be a valuable addition to the landscape of sRNA expression databases.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

ACKNOWLEDGEMENTS

We would like to thank Mariah Snyder, Ashish Rajput, Ting Sun, Vikas Bansal, Michel Edwar Mickael, the DZNE IT, and all of the Oasis users for helpful suggestions.

FUNDING

This work was supported by the DFG (BO4224/4-1), the Network of Centres of Excellence in Neurodegeneration (CoEN) initiative, the Volkswagen Stiftung (Az88705), iMed – the Helmholtz Initiative on Personalized Medicine, and the BMBF grant Integrative Data Semantics in Neurodegeneration (031L0029B, IDSN).

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

REFERENCES

Barrett,T. *et al.* (2013) NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Res.*, **41**, 991–995.

Capece,V. *et al.* (2015) Oasis: online analysis of small RNA deep sequencing data. *Bioinformatics*, **31**, 1–3.

Chung,I.-F. *et al.* (2016) YM500v3: a database for small RNA sequencing in human cancer research. *Nucleic Acids Res.*, **45**, D925–D931.

Côté,R. *et al.* (2010) The Ontology Lookup Service: Bigger and better. *Nucleic Acids Res.*, **38**, 155–160.

Leung,Y.Y. *et al.* (2016) DASHR: Database of Small human noncoding RNAs. *Nucleic Acids Res.*, **44**, D216–D222.

Panwar,B. *et al.* (2017) miRmine: A Database of Human miRNA Expression Profiles. *Bioinformatics*.

Schuurman,N. and Leszczynski,A. (2008) Ontologies for Bioinformatics. *Bioinform. Biol. Insights*, **2**, 187–200.

Vitsios,D.M. *et al.* (2017) Large-scale analysis of microRNA expression, epi-transcriptomic features and biogenesis. *Nucleic Acids Res.*, **45**, 1079–1090.

Witwer,K.W. (2014) Circulating MicroRNA Biomarker Studies: Pitfalls and Potential Solutions. *Clin. Chem.*, **000**.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

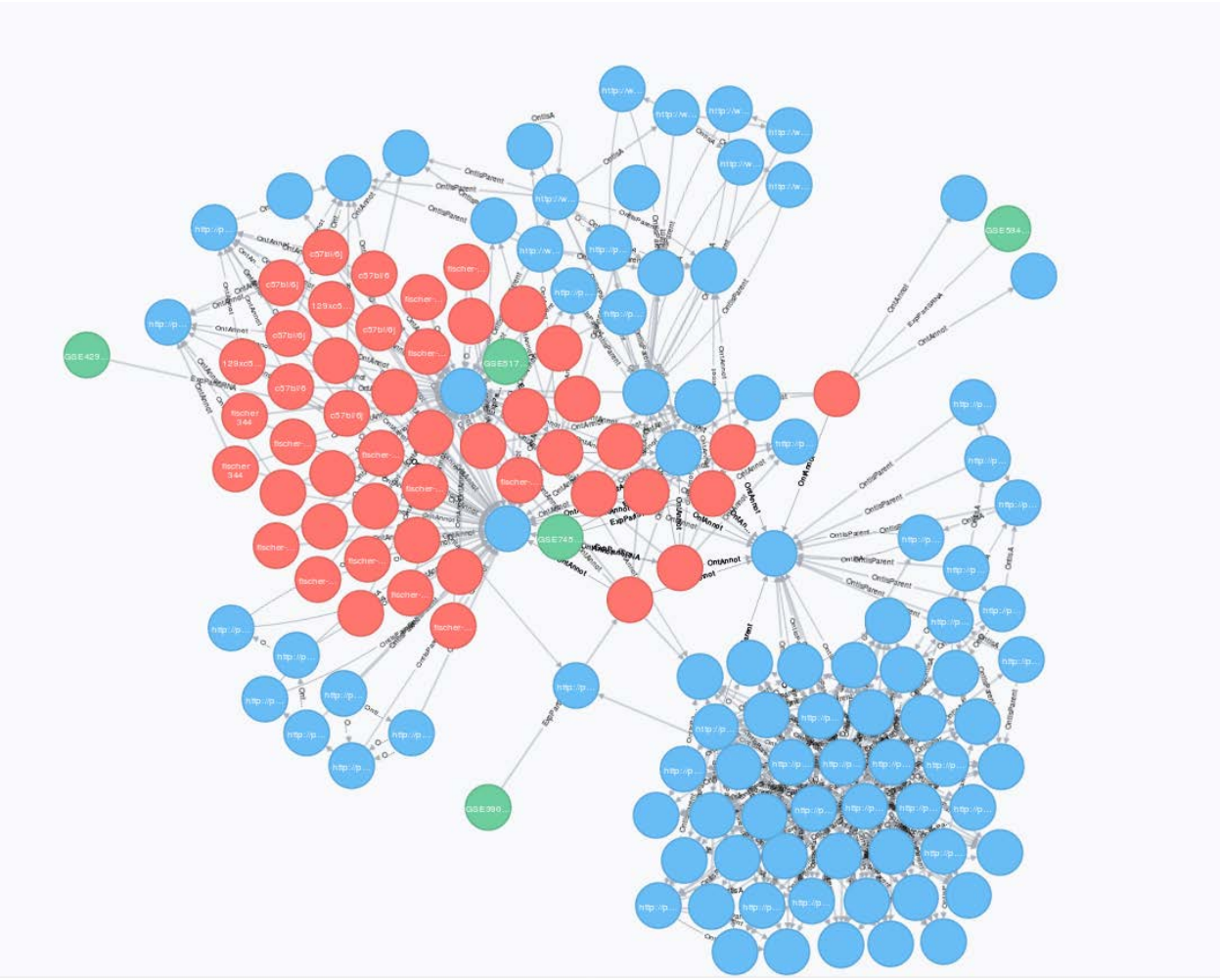

Fig.1. Objects in the SEA graph database (Neo4j). A fragment of the SEA graph database is visualized, where green nodes represent datasets, red nodes represent samples and blue nodes represent ontology terms. Grey edges represent 'is a' relations between the different datasets, samples, and ontology terms.

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1. SEA keys and used ontologies (as of April, 21st 2017).

Key	Ontology(s)	# Annotations	# Terms
Organism	NCBI Taxonomy[5]	2105	10
Tissue	BRENDA tissue / enzyme source[6]	1595	86
Disease	Human Disease Ontology[7]	791	68
Cell type	Cell Ontology[8]	517	57
Cell line	Cell Line Ontology[9]	39	12
	Experimental Factor Ontology[10]	253	55

[5] https://www.ncbi.nlm.nih.gov/taxonomy
[6] http://www.brenda-enzymes.info/
[7] http://www.disease-ontology.org/
[8] http://obofoundry.org/ontology/cl.html
[9] http://www.clo-ontology.org/
[10] http://www.ebi.ac.uk/efo/

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 2. Supported SEA organisms and their corresponding genome versions.

Organism	genome-version	genome-date
Bos taurus	UMD3.1	2009-11
Caenorhabditis elegans	WBcel235	2012-12
Danio rerio	GRCz10	2014-09
Drosophila melanogaster	BDGP6	2014-07
Mus musculus	GRCm38	2012-01
Gallus gallus	Galgal4	2011-11
Rattus norvegicus	Rnor_6.0	2014-07
Homo sapiens	GRCh38	2013-12
Sus scrofa	Sscrofa10.2	2011-08
Anopheles gambiae	AgamP4	2006-02

bioRxiv preprint doi: https://doi.org/10.1101/133199; this version posted August 4, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Feature	**SEA**	**miRmine[1]**	**DASHR[2]**	**miratlas[3]**	**YM500v3[4]**
Organisms	10	1	1	2	1
sRNA types	5	1	5	1	5
Samples	>2000	304	187	461	>8000*
Novel miRNAs	+	-	-	-	-
Ontology search[#]	+	-	-	-	-

Table 3. Comparison of sRNA expression databases. This table includes recent sRNA expression databases and a list of features we deem relevant. *Supports mainly cancer-related datasets. [#]Use of ontological graphs for the annotation and querying of samples. [1](Panwar et al., 2017), [2](Leung et al., 2016), [3](Vitsios et al., 2017), [4](Chung et al., 2016)