Bioinformatics, 31(24), 2015, 4038–4040
doi: 10.1093/bioinformatics/btv503
Advance Access Publication Date: 30 August 2015
Applications Note

OXFORD

Databases and ontologies

RNASeqMetaDB: a database and web server for navigating metadata of publicly available mouse RNA-Seq datasets

Zhengyu Guo[1,2], Boriana Tzvetkova[3], Jennifer M. Bassik[4], Tara Bodziak[4], Brianna M. Wojnar[4], Wei Qiao[1], Md A. Obaida[1], Sacha B. Nelson[3], Bo Hua Hu[4] and Peng Yu[1,2,*]

[1]Department of Electrical and Computer Engineering & [2]TEES-AgriLife Center for Bioinformatics and Genomic Systems Engineering, Texas A&M University, College Station, TX 77843, USA, [3]Department of Biology & Center for Behavioral Genomics, Brandeis University, Waltham, MA 02454, USA and [4]Department of Communicative Disorders and Sciences & Center for Hearing and Deafness, University at Buffalo, Buffalo, NY 14214, USA

*To whom correspondence should be addressed.

Associate Editor: Ivo Hofacker

Received on October 8, 2014; revised on August 1, 2015; accepted on August 23, 2015

Abstract

Summary: Gene targeting is a protocol for introducing a mutation to a specific gene in an organism. Because of the importance of *in vivo* assessment of gene function and modeling of human diseases, this technique has been widely adopted to generate a large number of mutant mouse models. Due to the recent breakthroughs in high-throughput sequencing technologies, RNA-Seq experiments have been performed on many of these mouse models, leading to hundreds of publicly available datasets. To facilitate the reuse of these datasets, we collected the associated metadata and organized them in a database called RNASeqMetaDB. The metadata were manually curated to ensure annotation consistency. We developed a web server to allow easy database navigation and data querying. Users can search the database using multiple parameters like genes, diseases, tissue types, keywords and associated publications in order to find datasets that match their interests. Summary statistics of the metadata are also presented on the web server showing interesting global patterns of RNA-Seq studies.

Availability and implementation: Freely available on the web at http://rnaseqmetadb.ece.tamu.edu.

Contact: pengyu.bio@gmail.com

1 Introduction

Gene targeting (Capecchi, 1989), a powerful technique used to manipulate a specific locus in the genome of an organism, is an indispensable tool for assessing *in vivo* functions of specific gene products. This technique provides great flexibility in manipulating the genome of an organism as it can be used to delete a gene or an exon, to introduce an exogenous gene, or to create point mutations. Moreover, gene targeting not only can introduce permanent mutations, but can also conditionally change targeted genes. Thousands of genetically engineered mice have been generated using this technique. They provide valuble models for studying mechanisms of human diseases.

RNA-Seq (Wang *et al.*, 2009), a high-throughput sequencing (HTS) method for transcriptome analysis, has been successfully used on many of these mouse models, enabling global analyses of specific genomic alterations at a high sequencing depth with a resonable accuracy. As RNA-Seq becomes increasingly popular, hundreds of RNA-Seq datasets have been generated and have been released to the public. These data are currently available from online repositories, such as

Gene Expression Omnibus (GEO) (Barrett *et al.*, 2013), ArrayExpress (Rustici *et al.*, 2013), Sequence Read Archive (Kodama *et al.*, 2012) and European Nucleotide Archive (ENA) (Brunak *et al.*, 2002), whose primary purposes are to store raw and processed HTS data from a wide variety of organisms. However, a data submitter typically provides only limited metadata for each dataset sufficient to get the dataset accepted into the public repository. There is currently no stringent and uniform quality check of submitted metadata. This results in inconsistency and ambiguity in dataset annotation. For example, non-official gene symbols are used in some of the datasets. Other public databases such as InSilico DB (Coletta *et al.*, 2012) also suffer from the same problems, making searching for datasets in these databases inefficient.

The recent HTS data explosion has motivated researchers to create several metadata databases. However, these databases, e.g. CistromeMap (Qin *et al.*, 2012), focus primarily on ChIP-Seq data. To fill the gap for RNA-Seq data, we collected RNA-Seq metadata from all the publicly available datasets that were generated using mouse models mostly with targeted mutations and curated a database called RNASeqMetaDB. Haynes *et al.* (2013) recently suggested that measuring transcription factor binding might not be the best way to decipher transcriptional regulatory networks. Instead, their work showed that gene expression data could be of greater value in revealing functional gene regulatory relations. Therefore, RNASeqMetaDB may be a helpful resource for researchers trying to build gene regulatory networks.

We developed a web server to provide a user-friendly query interface for locating relevant RNA-Seq datasets based on targeted gene names, disease names, tissue types, keywords, publications and accession IDs, etc. An ontological search function is also offered that allows users to find the datasets related to, but not necessarily annotated to, the exact search term. This helps ensure search sensitivity (e.g. see the help page on the website). This database can help biomedical scientists navigate the complex landscape of mouse genetic experiments and can provide rich contexts for these datasets. Using this database, users will be able to find related datasets for further analyses easily. For example, RNA-Seq data can be used to infer splicing isoform functions (Eksi *et al.*, 2013) and information extracted from existing RNA-Seq data can be used as prior knowledge for causal reasoning on biological networks (Chindelevitch *et al.*, 2012). Moreover, RNA-Seq data can be integrated with sequence- and structure-binding preferences of RNA-binding proteins learned with computational methods such as GraphProt (Maticzka *et al.*, 2014), which can increase our understanding of the mechanisms of post-transcriptional regulation.

2 Methods

We collected raw annotations of publicly available mouse RNA-Seq datasets from HTS data repositories including GEO, ArrayExpress and ENA. At the time of writing, RNASeqMetaDB contains 306 experiments in total. The following metadata were systematically annotated for each RNA-Seq dataset: gene symbol, genotype, reference (including title, authors, abstract, PubMed ID), disease, tissue type, corresponding author and author's website link. Genotype and disease annotations were manually curated and extracted from the original publications. For consistency, genes, alleles, diseases and tissue types were annotated using official symbols or controlled vocabularies from online resources including Mouse Genome Informatics (MGI) (Blake *et al.*, 2014), Medical Subject Headings (MeSH) (Rogers, 1963) and BRENDA Tissue Ontology (Gremse *et al.*, 2011).

To facilitate querying and data retrieval, we implemented a web server (http://rnaseqmetadb.ece.tamu.edu). All search functionalities are integrated within a single web interface. Users can search for one or multiple datasets using gene symbols, disease names, tissue types or keywords. The search generates a list of matched datasets containing accession IDs, titles, mutated genes, related diseases and tissue types. To support more general database queries, ambiguous keyword search is provided by the server. Users can type one or multiple terms that they are interested in into the Keyword search box. Both the typed words and their synonyms defined by the Experimental Factor Ontology (Malone *et al.*, 2010) will be searched in the database, and then the query results will be displayed. If a more targeted search is needed, users are allowed to use additional terms in the Search text box above the result table to refine the results. This retrieves only the datasets whose titles match these additional terms. When the accession ID of a dataset is clicked, RNASeqMetaDB displays a summary table of all the metadata available for that dataset. The links to other databases and websites like ArrayExpress, GEO, MGI, PubMed and MeSH, and the PIs' lab websites are provided so that users can easily obtain additional information related to the datasets. Users can also bulk download data after registering an account on the website. To keep the data updated, the website allows users to submit requests for adding data entries for newly published RNA-Seq datasets.

3 Results

RNASeqMetaDB permits efficient searching of its database containing comprehensive information for all public RNA-Seq datasets on mice with genotype as a factor. It contains metadata for a total of 306 experiments targeting 298 different genes. These experiments are from 264 different research groups, among which 154 are from the USA and 76 are from Europe. For journals publishing the studies using these datasets, *Nature* ranks at the top with the greatest number of studies, followed by *Proceedings of the National Academy of Sciences of the United States of America* and *Genes and Development*. One interesting observation is that the number of publications on RNA-Seq studies has been increasing exponentially since 2008. This indicates that RNA-Seq experiments on gene-targeted mouse models have become more popular in recent years. Summary statistics of the datasets are available at the database website.

RNASeqMetaDB broadens the use of these datasets by providing well-curated metadata and efficient query functionalities. Scientists can easily find the datasets that they are interested in and retrieve detailed information to enable more comprehensive understanding of the experiments. In the future, we will develop additional functionalities and import more datasets into the database. We believe that RNASeqMetaDB will be a valuable tool for the biomedical research community.

Acknowledgement

The authors thank Youyi Dong and Robert Vethanayagam Robin Swamidoss for their contributions to RNASeqMetaDB.

Funding

This work was supported by startup funding to P.Y. from the ECE department and Texas A&M Engineering Experiment Station/Dwight Look College of Engineering at Texas A&M University and was supported in part by NIDCD R01DC010154.

Conflict of Interest: none declared.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.*, **41**(Database issue), D991–D995.

Blake,J.A. *et al*. (2014) The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse. *Nucleic Acids Res.*, **42**(Database issue), D810–D817.

Brunak,S. *et al*. (2002) Nucleotide sequence database policies. *Science*, **298**, 1333.

Capecchi,M.R. (1989) The new mouse genetics: altering the genome by gene targeting. *Trends Genet.*, **5**, 70–76.

Chindelevitch,L. *et al*. (2012) Causal reasoning on biological networks: interpreting transcriptional changes. *Bioinformatics*, **28**, 1114–1121.

Coletta,A. *et al*. (2012) InSilico DB genomic datasets hub: an efficient starting point for analyzing genome-wide studies in GenePattern, Integrative Genomics Viewer, and R/Bioconductor. *Genome Biol.*, **13**, R104.

Eksi,R. *et al*. (2013) Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data. *PLoS Comput. Biol.*, **9**, e1003314.

Gremse,M. *et al*. (2011) The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.*, **39**(Database issue), D507–D513.

Haynes,B.C. *et al*. (2013) Mapping functional transcription factor networks from gene expression data. *Genome Res.*, **23**, 1319–1328.

Kodama,Y. *et al*. (2012) The Sequence Read Archive: explosive growth of sequencing data. *Nucleic Acids Res.*, **40**(Database issue), D54–D56.

Malone,J. *et al*. (2010) Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, **26**, 1112–1118.

Maticzka,D. *et al*. (2014) GraphProt: modeling binding preferences of RNA-binding proteins. *Genome Biol.*, **15**, R17.

Qin,B. *et al*. (2012) CistromeMap: a knowledgebase and web server for ChIP-Seq and DNase-Seq studies in mouse and human. *Bioinformatics*, **28**, 1411–1412.

Rogers,F.B. (1963) Medical subject headings. *Bull. Med. Lib. Assoc.*, **51**, 114–116.

Rustici,G. *et al*. (2013) ArrayExpress update—trends in database growth and links to data analysis tools. *Nucleic Acids Res.*, **41**(Database issue), D987–D990.

Wang,Z. *et al*. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, **10**, 57–63.