BIOINFORMATICS APPLICATIONS NOTE
Vol. 28 no. 8 2012, pages 1184–1185
doi:10.1093/bioinformatics/bts084

Databases and ontologies

Advance Access publication February 17, 2012

RNA-Seq Atlas—a reference database for gene expression profiling in normal tissue by next-generation sequencing

Markus Krupp[1,*], Jens U. Marquardt[1], Ugur Sahin[2], Peter R. Galle[1], John Castle[2,†] and Andreas Teufel[1,†]

[1]Department of Medicine I and [2]Translational Oncology and Immunology (TRON), Johannes Gutenberg University, 55131 Mainz, Germany

Associate Editor: Alex Bateman

ABSTRACT

Motivation: Next-generation sequencing technology enables an entirely new perspective for clinical research and will speed up personalized medicine. In contrast to microarray-based approaches, RNA-Seq analysis provides a much more comprehensive and unbiased view of gene expression. Although the perspective is clear and the long-term success of this new technology obvious, bioinformatics resources making these data easily available especially to the biomedical research community are still evolving.

Results: We have generated RNA-Seq Atlas, a web-based repository of RNA-Seq gene expression profiles and query tools. The website offers open and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns. To enlarge the scope of the RNA-Seq Atlas, the data were linked to common functional and genetic databases, in particular offering information on the respective gene, signaling pathway analysis and evaluation of biological functions by means of gene ontologies. Additionally, data were linked to several microarray gene profiles, including BioGPS normal tissue profiles and NCI60 cancer cell line expression data. Our data search interface allows an integrative detailed comparison between our RNA-Seq data and the microarray information. This is the first database providing data mining tools and open access to large scale RNA-Seq expression profiles. Its applications will be versatile, as it will be beneficial in identifying tissue specific genes and expression profiles, comparison of gene expression profiles among diverse tissues, but also systems biology approaches linking tissue function to gene expression changes.

Availability and implementation: http://medicalgenomics.org/rna_seq_atlas

Contact: kruppm@uni-mainz.de; teufel@uni-mainz.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 09, 2011; revised on February 1, 2012; accepted on February 13, 2012

1 INTRODUCTION

Over the next years, the availability of next-generation sequencing (NGS) data will offer an entirely new perspective for clinical research and help usher in personalized medicine. So far, several databases offer storage or download of NGS data (Altshuler, 2010; Shumway *et al.*, 2010). However, to access the valuable information of this promising new technique, the user has to manually download the data and be familiar with their analysis to extract the valuable information. This is currently not the case for most biomedical researchers. We therefore created the RNA-Seq Atlas, a database and user interface (UI) providing easy access to NGS data. Currently, RNA-Seq Atlas holds gene expression profiles on eleven human, healthy tissues and can be accessed over an intuitive web interface. To further increase the utility of the RNA-Seq Atlas, the data were linked to multiple microarray gene profiles representing normal and pathological states. Furthermore, various query tools were designed to offer a great variability of individual analysis.

2 DATABASE ORGANIZATION AND ACCESS

2.1 Data sources

The provided genome-wide expression compendium originates from eleven healthy, human tissue samples pooled from multiple donors spanning 32 384 specific transcripts corresponding to 21 399 unique genes (Castle *et al.*, 2010) (ENA ERP000257; ArrayExpress E-MTAB-305). The tissues include adipose, colon, heart, hypothalamus, kidney, liver, lung, ovary, skeletal muscle, spleen and testes. Sequencing was performed on an Illumina GA-II sequencer, generating an average of 50 million reads per tissue, with sequence reads of 36 or 50 nt depending on tissue. The expression level were estimated by mapping and counting reads to single gene sequences derived from the UCSC genome browser, followed by normalization to Reads Per Kilobase of exon model per Million mapped reads values (Mortazavi *et al.*, 2008).

Moreover, to enable an integrative comparison between RNA-Seq and microarray expression profiles we integrated a panel of 84 microarrays from BioGPS (Su *et al.*, 2004; Wu *et al.*, 2009), Normal Tissue Gene Expression Study (Ge *et al.*, 2005) as well as (Ross *et al.*, 2000; Shankavaram *et al.*, 2009) NCI60 into the RNA-Seq Atlas. These gene expression profiles correspond to the equivalent tissues included in the RNA-Seq Atlas and involve >39 000 transcripts from pathological (i.e. cancer) and normal tissues states. Detailed information about data processing and integration can be found in the Supplementary Materials S1 and S2.

Further, the RNA-Seq Atlas was linked to commonly used and established bioinformatics databases and knowledge repositories. Enabling access to deeper transcriptional information was achieved

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the last two authors should be regarded as joint Last Authors.

by linking the RNA-Seq Atlas data to the NCBI Nucleotide database (Sayers *et al*., 2011). Also, information on corresponding gene symbol, aliases, description, chromosomal location, Entrez ID as well as EnsEMBL ID were assembled from the NCBI Entrez and EnsEMBL databases (Sayers *et al*., 2011; Spudich *et al*., 2007). Additional outgoing links to HGNC (Seal *et al*., 2011), HPRD (Keshava Prasad *et al*., 2009), OMIM (McKusick, 2007), BioGPS (Wu *et al*., 2009), Nextbio (Kupershmidt *et al*.) and GENT (Shin *et al*., 2011) were supported. Finally, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Aoki and Kanehisa, 2005) was accessed to identify gene signaling as well as molecular pathway affiliations; and data on cellular component, biological process and molecular function were collected from the Gene Ontology database (Ashburner *et al*., 2000). Finally, the RNA-Seq Atlas was cross-linked to our liver-specific databases LoMA (Buchkremer *et al*., 2010).

2.2 Data organization and web interface

RNA-Seq Atlas is implemented within a Drupal content management system environment over a Linux-PostgreSQL-Apache-PHP stack. The database organization is founded upon a menu which allows access to the news, data, search and download sections.

The news section keeps users up to date about recent changes and current statistics, whereas the download section give the possibility to download the RNA-Seq Atlas in tab separated text file format. RNA-Seq atlas can be accessed through simple (data section) or advanced (search section) query forms. The advanced query offers four detailed options:

(1) Full text search.

(2) Comparison of specific tissues profiles; also allowing for comparative analysis not only between normal tissue information but also to NCI60 data and thus between normal and tumor tissues.

(3) Explore common (and diverse) gene expression profiles between tissues.

(4) Explore pathway profile; e.g. selecting one or multiple KEGG pathway resulting in a list of involved genes.

Finally, a ‘details’ link provides additional information including: gene symbol, description, aliases, chromosomal location, Entrez ID, Ensembl ID, Gene Ontology, KEGG pathway as well as the expression profile within the normal human tissues and cancer cell lines.

2.3 Future directions

Future directions include an incorporation of more data from healthy and cancer tissue to provide a richer source of comparative transcriptomics and implementation of a Gene Set Enrichment Analysis (GSEA) analysis engine within the RNA-Seq Atlas.

3 CONCLUSION

In this work, we present RNA-Seq Atlas, an easily accessible database and UI, offering access to NGS gene expression profiles. Furthermore, to enhance the bioinformatics integration, the data is linked to a wide variety of commonly used and established databases and knowledge repositories. To further enlarge the very broad scope of RNA-Seq Atlas and to facilitate the analysis of gene expression profiles of several pathological conditions, the data were linked to cancer cell line expression profiles. Finally, the implementation of a wide variety of querying tools allows the user to start individual analysis, enabling for both bioinformaticians and experimental researchers.

Funding: This study was supported by a research grant of the Boehringer Ingelheim Foundation and funding of the core facility bioinformatics of the University Hosiptal of the Johannes Gutenberg University Mainz, Germany.

Conflict of Interest: none declared.

REFERENCES

Altshuler,D. (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Aoki,K.F. and Kanehisa,M. (2005) Using the KEGG database resource. *Curr. Protoc. Bioinformatics*, **Chapter 1**, Unit 1.12.

Ashburner,M. *et al*. (2000) Gene Ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Buchkremer,S. *et al*. (2010) Library of molecular associations: curating the complex molecular basis of liver diseases. *BMC Genomics*, **11**, 189.

Castle,J.C. *et al*. (2010) Digital genome-wide ncRNA expression, including SnoRNAs, across 11 human tissues using polyA-neutral amplification. *PLoS One*, **5**, e11779.

Ge,X. *et al*. (2005) Interpreting expression profiles of cancers by genome-wide survey of breadth of expression in normal tissues. *Genomics*, **86**, 127–141.

Keshava Prasad,T.S. *et al*. (2009) Human Protein Reference Database—2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Kupershmidt,I. *et al*. (2010) Ontology-based meta-analysis of global collections of high-throughput public data. *PLoS One*, **5**, e13066.

McKusick,V.A. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.

Mortazavi,A. *et al*. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Methods*, **5**, 621–628.

Ross,D.T. *et al*. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 227–235.

Sayers,E.W. *et al*. (2011) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **39**, D38–D51.

Seal,R.L. *et al*. (2011) genenames.org: the HGNC resources in 2011. *Nucleic Acids Res.*, **39**, D514–D519.

Shankavaram,U.T. *et al*. (2009) CellMiner: a relational database and query tool for the NCI-60 cancer cell lines. *BMC Genomics*, **10**, 277.

Shin,G. *et al*. (2011) GENT: gene expression database of normal and tumor tissues. *Cancer Inform.*, **10**, 149–157.

Shumway,M. *et al*. (2010) Archiving next generation sequencing data. *Nucleic Acids Res.*, **38**, D870–D871.

Spudich,G. *et al*. (2007) Genome browsing with Ensembl: a practical overview. *Brief. Funct. Genomic. Proteomic.*, **6**, 202–219.

Su,A.I. *et al*. (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA*, **101**, 6062–6067.

Wu,C. *et al*. (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.