Bioinformatics, 34(2), 2018, 321–322
doi: 10.1093/bioinformatics/btx599
Advance Access Publication Date: 22 September 2017
Applications Note

Databases and ontologies

HoTResDB: host transcriptional response database for viral hemorrhagic fevers

Jonathan Lo[1,†], Deric Zhang[1,†], Emily Speranza[1,2], Jose A. Negron[1] and John H. Connor[1,2,*]

[1]Bioinformatics Program, Boston University, 24 Cummington Mall, Boston, MA 02215, USA and [2]Department of Microbiology, National Emerging Infectious Diseases Laboratories (NEIDL), Boston Univeristy, 620 Albany St, Boston, MA 02218, USA

*To whom correspondence should be addressed
[†]The authors wish it to be known that first two authors contributed equally.

Associate Editor: Janet Kelso

Received on April 6, 2017; revised on August 17, 2017; editorial decision on September 18, 2017; accepted on September 19, 2017

Abstract

Summary: High-throughput screening of the host transcriptional response to various viral infections provides a wealth of data, but utilization of microarray and next generation sequencing (NGS) data for analysis can be difficult. The *Ho*st *T*ranscriptional *R*esponse *D*ata*B*ase (HoTResDB), allows visitors to access already processed microarray and NGS data from non-human primate models of viral hemorrhagic fever to better understand the host transcriptional response.
Availability: HoTResDB is freely available at http://hotresdb.bu.edu
Contact: jhconnor@bu.edu

1 Introduction

The host response to viral infection is an important aspect of viral pathogenesis. In hemorrhagic fevers such as Ebolavirus (EBOV), Marburg virus (MARV) and Lassa virus (LASV) their high levels of pathogenicity can in part be attributed to a hyperactive immune response. High-throughput screening techniques allow for full analysis of the host transcriptional response to various viral infections and provide a wealth of data to researchers to help generate hypotheses. However, utilization of microarray and next generation sequencing (NGS) data for analysis can be difficult. Extrapolation of useful information from raw data requires knowledge of programming and high-level statistics. To help researchers without a background in programming or statistics explore viral hemorrhagic fever data sets, we established a database, the *Ho*st *T*ranscriptional *R*esponse *D*ata*B*ase (HoTResDB), that allows visitors to access processed microarray and NGS data from non-human primates models of viral hemorrhagic fever to better understand the host transcriptional response. Users with gene IDs or Ensembl identifiers of interest can view the data in several graphical formats and download raw data. Additionally, a keyword search is available to facilitate the general search.

2 Description

Using HoTResDB, visitors can explore relationships and trends in data sets of specific genes of interest or browse through the database to find genes of interest. We designed this database to be a resource for the study of the host transcriptional response to viral hemorrhagic fever by giving researchers an easy-to-use interface to explore this data and aid in generating hypotheses.

2.1 Data processing

All currently available data sets are accessible through NCBI. The raw counts tables for the data sets were processed as follows. The Lassa and Marburg RNA-Seq data were processed according to Caballero *et al.* (2014). The two Ebola RNA-Seq studies were processed according to Caballero *et al.* (2016). Finally, the Ebola microarray study was collected from Yen *et al.* (2011) and the data were downloaded from GSE24943. The raw values were normalized in R (http://www.R-project.org/) using the *limma* package (Ritchie *et al.*, 2015) to determine the normalized fluorescence for each probe. The RNA-Seq data was compiled into a single-large table for all studies. The table was read into R and using the edgeR package (Robinson *et al.*, 2010), the data was normalized and the counts per million (CPM) calculated. This was then imported into the database in a counts table format.

2.2 Database design and user interface

Due to the nature of the data being stored, HoTResDB is built as a MySQL relational database. This offers the advantage of easy extensibility and the ability to perform complex, multi-row transactions. The use of a relational database management system also reduces the need for data duplication, which is valuable given the size of gene expression data sets. It runs on a Linux server and is currently hosted using Apache 2.4. For the user interface, HoTResDB uses HTML and PHP files with CGI Python scripts for data interaction and retrieval. The Python scripts were developed with version 2.7 or above. The database is currently separated into seven tables due to the range of data available.

The general search form submits user input to the CGI Python files that handle MySQL database queries. The Python scripts then calculate mean and standard error of the expression values at the various days post-infection as well as the log 2 expression ratio of the data relative to the pre-infection controls. Clustering of heat map data is also done by the Python scripts. The data are then converted into JSON format upon return to the user interface as this data structure is most compatible with the HighCharts API.

All output graphs are downloadable in several formats: PNG, JPEG, PDF and SVG, as a built-in function of the HighCharts API. This helps the user find a compatible format for their intended use of the graphs (i.e. PowerPoints or a printed image). In the current version, the graphs generated by the database are not recommended to be used as a rigorous analytical tool, but are meant to be a visualization tool for existing data sets.

3 Using HoTResDB

To initiate a search on HoTResDB, the user must select one or more viruses, strains and data types from the available options. The search form requires the user to input gene identifiers in the appropriate box. An option to upload a plain text file containing a list of genes separated by whitespace is available for user convenience. By default, the search only includes fatal outcomes from infection. However, the user can choose to include survivor data if available in data sets of interest. Additionally, for first time users, the sidebar provides an example search, which automatically performs a basic search of one gene in a single data set.

3.1 Viewing search results

Once collected, the data can be viewed in various different formats. Data, if any are found matching the query, are organized into several tabs on the search page: ‘Search Results’, ‘Expression Data’, ‘Counts Graphs’, ‘Fold Change Graphs’, and ‘Heat Maps’. These sub-tabs are organized in such a way that each successive tab is an increased level of processing from the one previous. For example, ‘Search Results’ shows the genes that matched the search parameters, ‘Expression Data’ shows the count values from those genes, ‘Counts Graphs’ show line graphs of those counts, ‘Fold Change Graphs’ show the log fold change of the counts relative to pre-infection and ‘Heat Maps’ shows the clustered fold changes.

3.1 Browsing the database

Users can also browse genes in the database using the browse database function. Using this feature, users are able to look for any human gene ID, macaque gene ID, gene description or gene ontology (GO) term that contains their term of interest. In addition, the browse function offers the ability to further narrow search parameters using Boolean search operators, which allow users to search multiple parameters simultaneously or concurrently.

4 Conclusions

Investigators with specific gene sets or gene categories of interest looking for a convenient method for visualizing translational data in NHPs infected with hemorrhagic fever viruses will be able find use in HoTResDB. The database is useful for exploring questions regarding the effect hemorrhagic fever virus infection has on the expression of certain groups of genes. For example, the following question could be addressed: ‘How does transcription of interferon-stimulated genes differ in response to infection by different strains or species of filoviruses?’ Through using HoTResDB, investigators will be able to find an expression pattern of interest that may lead to new hypotheses.

The database will aid in generating hypotheses on the elucidation of hemorrhagic fever virus host transcriptional response mechanisms. Users would be able to gain insight into potential virus–host interactions through visualizing gene expression data in specific gene sets. This can help to drive hypotheses about how the virus and host interact during infection and how the host responds to highly pathogenic viruses.

HoTResDB is also unique in that it provides a look at the temporal progression of the host transcriptional response to infection. This type of data is typically difficult to obtain due to the challenge of repeatedly sampling infected subjects to measure their gene expression over time, especially in individuals infected with VHFs.

Acknowledgements

We would like to give special thanks to the members of the Connor Lab and Muhlberger Lab at Boston University, BUMC, Department of Microbiology, National Emerging Infectious Disease Laboratories (NEIDL) and Dr. Arthur J. Goff and group at USAMRIID for being testers. We give special thanks to Dr. Gary Benson and Dan Lancour for their assistance on the development of the database structure and for their help during the BF768 Course at Boston University.

Funding

This work was supported by the National Institutes of Health (NIAID) [R01-AI1096159, R21-AI121933 to J.H.C.]; the National Science Foundation Graduate Research Fellowship [DEG-1247312 to E.S.]; and the Boston University Graduate Scholarship to J.L. and D.Z. The NIH played no role in the manuscript.

Conflict of Interest: none declared.

References

Caballero,I.S. *et al.* (2014) Lassa and Marburg viruses elicit distinct host transcriptional responses early after infection. *BMC Genomics*, **15**, 960.

Caballero,I.S. *et al.* (2016) In vivo Ebola virus infection leads to a strong innate response in circulating immune cells. *BMC Genomics*, **17**, 707.

Ritchie,M.E. *et al.* (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.*, **43**, e47.

Robinson,M.D. *et al.* (2010) edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.

Yen,J.Y. *et al.* (2011) Therapeutics of Ebola hemorrhagic fever: whole-genome transcriptional analysis of successful disease mitigation. *J. Infectious Dis*, S1043-S1052.