VExD: A curated resource for human gene expression following viral infection

Phillip J Dexheimer
Division of Biomedical Informatics, Cincinnati Children's Hospital
Department of Biomedical Informatics, University of Cincinnati College of Medicine

Mario Pujato
Division of Biomedical Informatics, Cincinnati Children's Hospital

Krishna Roskin
Division of Biomedical Informatics, Cincinnati Children's Hospital
Division of Immunology, Cincinnati Children's Hospital
Department of Pediatrics, University of Cincinnati College of Medicine

Matthew T Weirauch
Center for Autoimmune Genomics and Etiology, Cincinnati Children's Hospital
Divisions of Biomedical Informatics and Developmental Biology, Cincinnati Children's Hospital
Department of Pediatrics, University of Cincinnati College of Medicine

Abstract

Motivation

Human viruses cause significant mortality, morbidity, and economic disruption worldwide. The human gene expression response to viral infection can yield important insights into the detrimental effects to the host. To date, hundreds of studies have performed genome-scale profiling of the effect of viral infection on human gene expression. However, no resource exists that aggregates human expression results across multiple studies, viruses, and tissue types.

Results

We developed the Virus Expression Database (VExD), a comprehensive curated resource of transcriptomic studies of viral infection in human cells. We have processed all studies within VExD in a uniform manner, allowing users to easily compare human gene expression changes across conditions.

Availability and Implementation

VExD is freely accessible at https://vexd.cchmc.org for all modern web browsers. An Application Programming Interface (API) for VExD is also available. The source code is available at https://github.com/pdexheimer/vexd.

Contact

phillip.dexheimer@cchmc.org, matthew.weirauch@cchmc.org

Introduction

Viral infection can have profound and lasting impacts on human health and society (Vos *et al.*, 2020). Viruses affect host gene expression levels by altering transcriptional mechanisms in infected cells and by initiating the host's immune response. Numerous transcriptomic studies of human viral infection have been deposited in public databases such as GEO (Barrett *et al.*, 2013). Unfortunately, GEO study descriptions are freeform text and require either manual curation or machine learning to search at a large scale (Wang *et al.*, 2019). This task is complicated in virus-based studies due to the inconsistent nomenclature of viruses in the literature (Gibbs, 2003) – for instance, Epstein-Barr Virus can be referred to as Herpes virus 4, *Human gammaherpesvirus 4*, or many abbreviations. Likewise, different studies employ different analysis pipelines to quantify gene expression levels and changes between conditions. Collectively, these issues greatly impede progress on studies of the human response to virus infection. To address this need, we present VExD, the Virus Expression Database (**Figure 1**). VExD contains a manually curated list of 7,903 samples from 289 gene expression studies of infection by 51 distinct viruses. All studies have been subjected to uniform quality control and data processing to maximize comparability across studies.

Identification of samples

We assembled a list of viruses known to infect humans based on annotations in ViralZone (Hulo *et al.*, 2011), UniProt (The UniProt Consortium, 2019), the Virus-Host DB (Mihara *et al.*, 2016), and literature curation studies (Taylor *et al.*, 2001; Woolhouse *et al.*, 2012), as well as common synonyms and abbreviations. We used the species name as defined by the International Committee on Taxonomy of Viruses for the canonical name of each virus.

Using an in-house tool, we scanned GEO for any studies mentioning a virus name or abbreviation in the study or sample description. Only samples annotated as human were considered. To enable standardized processing, we also required that studies use either the Affymetrix microarray or Illumina bulk RNA sequencing platforms. The 22,858 samples identified with the automated scanner were then manually curated to remove false positive results (**Figure 1**, top left). Samples were removed for a variety of reasons, including: 1) wrong platform, 2) use of a virus as a molecular biology tool rather than an infection (e.g., to immortalize cells or as a gene promoter), or 3) the virus name abbreviation identified by the scanner was not used in a virology context (e.g., Measles Virus = MV = Megavolts). The remaining 7,903 samples were annotated with Brenda Tissue Ontology (BTO) terms representing the used cell type or tissue (Gremse *et al.*, 2011) and were loaded into a Mongo database.

Data Quality Control

To make results comparable between 289 studies published over 17 years, all data were analyzed using a common set of gene definitions (Ensembl version 102). RNA-Seq samples were quantified using kallisto (Bray *et al.*, 2016) with the `--bias` parameter to account for sequence-specific biases. Final gene-level quantities were obtained by summing the TPM for all

associated transcripts. Samples with too few usable reads (pseudo-aligned reads < 1.5 million) or poor overall alignment (< 18% reads pseudo-aligned) were removed.

For microarray studies, custom CDFs from the BrainMap project (Dai *et al.*, 2005) that remapped probes onto Ensembl genes were used. All microarrays of a single array type were analyzed together using Robust Multi-chip Analysis (RMA) (Irizarry *et al.*, 2003) as implemented in the Affymetrix Power Tools. Samples were removed if the residual error in the RMA model was too large (`mad_residual_mean` > 0.80) or if the overall signal was too low (`pm_mean` < 65).

For RNA-seq and microarray platforms, studies with fewer than 60% of their samples passing these criteria were removed. In total, 234 of 2,129 (11%) RNA-seq samples and 186 of 4,964 (3.7%) microarray samples were removed by these quality control steps (**Supplemental Table 1**).

Differential Expression Quantification

VExD contains differential expression results between comparable virus infected samples and uninfected controls, allowing users to easily determine the human gene expression changes observed for each virus. Comparable samples are defined as those coming from the same study, using the same tissue or cell type, and the same technological platform, with at least two infected samples and two uninfected control samples. A total of 205 differential expression analyses involving 4,526 infected samples and 1,485 uninfected controls are available in VExD (**Figure 1**). All tests used a two-sided Welch's T test with unequal variance and Benjamini-Hochberg correction.

VExD Website and API

The VExD website was created with the Python Flask framework and is served using the Apache webserver with the mod_wsgi module. Users can search by any combination of virus and tissue type to view overall gene expression levels, or search for individual genes across all studies by symbol, alias, or Ensembl ID (**Figure 1**). Results pages display the virus, cell type, fold change, and adjusted p-value for a given gene in all studies. Download links provide study and differential expression information in a tab-delimited text file format. VExD also supports an Application Programming Interface (API) that can be used to programmatically query the database on the command-line.

Discussion

We present VExD, the Virus Expression Database, a resource containing nearly 8,000 samples in nearly 300 studies of viral infection. To the best of our knowledge, this is the first resource to comprehensively identify and quantify human gene expression levels in the context of virus infection and provide differential expression results from those experiments. We anticipate that VExD will be a valuable resource for investigators in the fields of virology, bioinformatics, genomics, and disease genetics.

Acknowledgements

The authors thank the many labs who have contributed data to GEO. We also thank Carmy Forney, Leah Kottyan, Sayeed Syed, Andrew VonHandorf and Erica DePasquale for constructive feedback on the VExD website and manuscript.

Funding

This work was supported by National Institutes of Health grants [R01 HG010730, R01 NS099068, R01 AI024717, R01 AI148276, U01 AI130830, U01 AI150748, P01 AI150585, R21 HG008186] to M.T.W.

References

Barrett,T. *et al.* (2013) NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Research*, **41**, D991-5.

Bray,N.L. *et al.* (2016) Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.*, **34**, 525–527.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res*, **33**, e175–e175.

Gibbs,A.J. (2003) Viral nomenclature, where next? *Archives of Virology*, **148**, 1645–1653.

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

Hulo,C. *et al.* (2011) ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research*, **39**, D576–D582.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Mihara,T. *et al.* (2016) Linking Virus Genomes with Host Taxonomy. *Viruses*, **8**, 66.

Taylor,L.H. *et al.* (2001) Risk factors for human disease emergence. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **356**, 983–989.

The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research*, **47**, D506–D515.

Vos,T. *et al.* (2020) Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, **396**, 1204–1222.

Wang,Z. *et al.* (2019) Mining data and metadata from the gene expression omnibus. *Biophys Rev*, **11**, 103–110.

Woolhouse,M. *et al.* (2012) Human viruses: discovery and emergence. *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, **367**, 2864–2871.

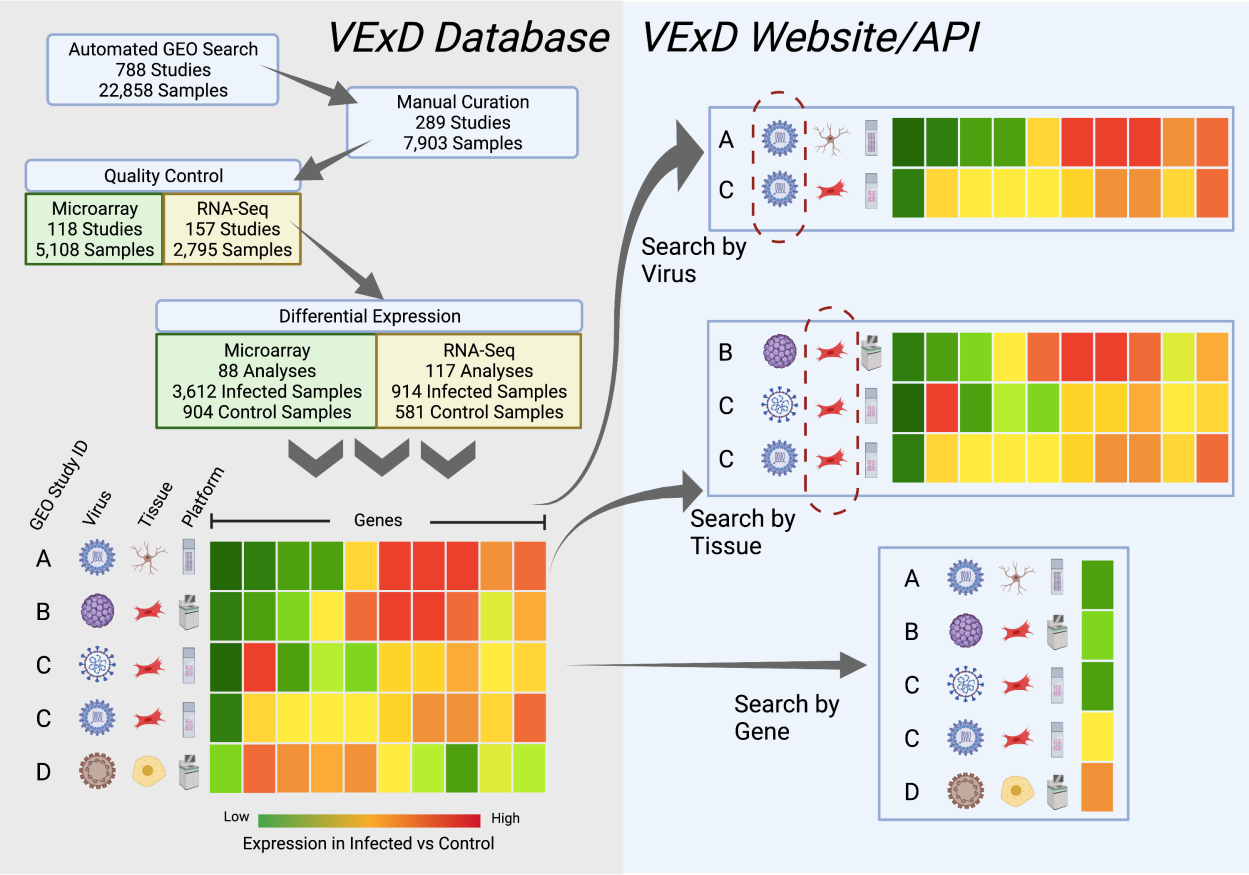

Figure 1. Overview of VExD. The database (left panel) contains a manually curated list of gene expression studies of human virus infection, along with differential human gene expression results for each study. Through the VExD website or API (right panel), users can find studies by searching for a virus or tissue of interest, or examine the behavior of a single gene across all studies. A downloads page provides all data in tab-delimited text files.