HIVed, a knowledgebase for differentially expressed human genes and proteins during HIV infection, replication and latency

Chen Li1,2,3,*, Sri H. Ramarathinam1,3,*, Jerico Revote4, Georges Khoury5, Jiangning Song1,3,6 & Anthony W. Purcell1,3

Measuring the altered gene expression level and identifying differentially expressed genes/proteins during HIV infection, replication and latency is fundamental for broadening our understanding of the mechanisms of HIV infection and T-cell dysfunction. Such studies are crucial for developing effective strategies for virus eradication from the body. Inspired by the availability and enrichment of gene expression data during HIV infection, replication and latency, in this study, we proposed a novel compendium termed HIVed (HIV expression database; http://hivlatency.erc.monash.edu/) that harbours comprehensive functional annotations of proteins, whose genes have been shown to be dysregulated during HIV infection, replication and latency using different experimental designs and measurements. We manually curated a variety of third-party databases for structural and functional annotations of the protein entries in HIVed. With the goal of benefiting HIV related research, we collected a number of biological annotations for all the entries in HIVed besides their expression profile, including basic protein information, Gene Ontology terms, secondary structure, HIV-1 interaction and pathway information. We hope this comprehensive protein-centric knowledgebase can bridge the gap between the understanding of differentially expressed genes and the functions of their protein products, facilitating the generation of novel hypotheses and treatment strategies to fight against the HIV pandemic.

Over 32 million are infected with HIV and with over 2 million new infections every year; it is still a major public health burden (UNAIDS 2010; http://www.unaids.org/globalreport/Global_report.htm). HIV predominantly infects CD4+ T-cells and leads to their death, impairing T helper immune responses during progressive infection and in the majority of cases leads to immunodeficiency if untreated1–5. While there is no effective vaccine or cure, administration of antiretroviral therapy reduces plasma viral loads in patients and greatly enhances their quality of life6. The therapy is life-long and cessation leads to return of significant viral loads predominantly from latent cellular reservoirs that include macrophages, dendritic cells, and particularly CD4+ T cells7–9. The implementation of effective strategies to clear HIV from body or to develop novel therapeutic interventions is contingent on understanding host cell changes during viral infection and latency. Gene arrays and transcriptomic analyses shed light on how HIV hijacks cell machinery and helped identify over 200 host factors that are crucial for HIV replication10–15. The advent of high-throughput next generation sequencing, proteomic and microarray analysis has generated large sets of data that were instrumental in revealing the role of HIV in modulation of expression of host genes (reviewed in Mehla et al.14).

1Infection and Immunity Program, Biomedicine Discovery Institute, Monash University, Melbourne, VIC 3800, Australia. 2Department of Microbiology, Monash University, Melbourne, VIC 3800, Australia. 3Department of Biochemistry and Molecular Biology, Monash University, VIC 3800, Australia. 4Monash Bioinformatics Platform, Monash University, Melbourne, VIC 3800, Australia. 5Department of Microbiology and Immunology, Peter Doherty Institute for Infection and Immunity, The University of Melbourne, Parkville, VIC 3010, Australia, 6Monash Centre for Data Science, Faculty of Information Technology, Monash University, Melbourne, VIC 3800, Australia. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to A.W.P. (email: Anthony.Purcell@monash.edu)

Received: 02 November 2016
Accepted: 27 February 2017
Published: 30 March 2017
In addition to development of drug interventions to control HIV infection, these studies can also point in the direction of methods to flush out the sources of latent virus. These remarkably stable reservoirs are established within a few days of HIV infection and can persist for years\(^2\) with an estimated half-life of around 43 months\(^7\). Elimination of these latent reservoirs is one of the greatest hurdles in the eradication of the virus\(^8\) and their analysis is crucial to understanding the mechanisms that support latency and to evaluate the effectiveness of agents that can reverse latency. Due to the rarity of latently infected cells in patients, several ex vivo and cellular models have been developed (reviewed in the study of Spina et al.\(^\text{*}\)) to analyse the latent reservoirs. The detailed analysis of these models using high-throughput transcriptomic and genomic approaches has advanced our understanding of viral transcriptional silencing and led to the accumulation of a large and valuable pool of data from disparate sources\(^12,15,20-27\). Studying the mechanisms of action of existing LRAs (Latency-Reversing Agents) could shed light on newer therapeutic interventions by comparing host cell expression data from different cellular models and LRAs.

At present, most of these data are hosted on individual websites: Litchi\(^21\), Peachi\(^28\) or in collections\(^29\), which mainly focus on annotating the genome and expression of related genes during HIV infection, replication, and/or latency, with little information on the functional annotations of their protein products. In this study, we implemented a novel knowledgebase, termed HIVed, to provide a comprehensive curation of related human proteins, whose genes have been characterized to be differentially expressed during HIV infection, replication, and latency. This protein-centric database collected experimental data of differentially regulated genes from a variety of studies, creating the comprehensive database that covers a wide range of experimental studies and hence bridges the knowledge annotation gap between known differentially expressed human genes, their corresponding protein products, and functional annotations during the HIV infection, replication, and latency.

**Database construction and utility**

**Data collection.** In order to collect the mainstream experimental studies of gene expression during HIV infection, replication and latency, we searched the literature and extracted eleven published genomic and proteomic high-quality datasets derived from a variety of experimental conditions. Six of the eight studies examine the human gene expression levels during HIV infection and replication\(^22-27\) and two focus on HIV latency. The representative experiments were conducted using different cell lines or tissues, including CD4\(^+\) T-cells\(^23,26,27\) and lymphatic tissue\(^25\), via different experimental techniques, such as transcriptomic profiling/analysis\(^24,26,27\), deep RNA-seq\(^3\), and genome-wide mRNA expression\(^22\). To portray the gene expression level during HIV latency, we selected two large-scale experimental studies\(^26,21\) to incorporate the gene expression data. Mohammadi et al. conducted a number of experiments to reveal the pairwise differential expression of transcripts during HIV latency and the subsequent viral reactivation following treatment of CD4\(^+\) T-cell models using different combinations of six agents, including DMSO, SAHA, CD3, IL7, DISU, and AZA\(^3\). The second selected dataset focused on the transcript regulation during HIV latency (latently infected CD4\(^+\) T-cells vs. uninfected cells) using primary CD4\(^+\) T-cell based models\(^5\). Three proteomic studies indicating the differentially expressed genes mediated by HIV infection using different cell lines (CD4\(^+\) and CD8\(^+\)) were also included\(^30-32\).

We combined all differently expressed genes identified in these experimental studies and further mapped them to the UniProt database\(^33\) to retrieve their protein products for detailed protein functional and structural annotations. Accordingly, a total of 14,318 human genes and their protein products were obtained and documented in our database. To ensure the quality of the curated entries, those genes that could not be mapped to known protein products were excluded from the database. However, in order to provide more comprehensive information, those genes that could not be identified are listed on the ‘Help’ webpage for users reference. Besides the gene expression profiles mediated by HIV, we also investigated whether the deposited entries in HIVed are HIV replication factors. Such host replication factor proteins play a crucial role in assisting HIV infection via their important biological functions in the host\(^12,15,34,35\). Therefore, to facilitate users to identify if a current protein is a replication factor or latency, with little information on the functional annotations of their protein products.

**HIVed construction.** HIVed was constructed using the JavaServer Pages (JSP) technique maintained by the Apache Tomcat\(^*\) web management system (http://tomcat.apache.org/; version: 8.0.32). We employed JavaBeans to facilitate the information transmission between the front-end webpages and the backstage of the database management system. The relational database was generated using the MySQL\(^\text{™}\) workbench (https://www.mysql.com/; version: 5.7.15) and managed by phpMyAdmin\(^\text{®}\) system (https://www.phpmyadmin.net/; version: 4.4.5.4.1). The host web server resides on a Linux operating system (Ubuntu; version: 16.04.1) machine with quad cores and 50 GB storage, allocated and maintained by Australia National eResearch Collaboration Tools
and Resources project (Nectar; https://nectar.org.au/) and Monash eResearch Centre (https://platforms.monash.edu/eresearch/). The interactive user-interface was implemented with the help of JavaScript and other third-party JavaScript plug-ins.

**HIVed utility.** A variety of functionalities, including database search, browse and user submission, have been provided to assist the readers to use HIVed efficiently (Fig. 2). To facilitate the fast navigation of HIVed database, short paths to database browsing, search and new submissions have been provided on the ‘Home’ page (Fig. 2a). A statistics webpage is available to summarize the biological annotations documented in HIVed (Fig. 2b). Top 10 significantly enriched pathways, concluded from a statistical analysis via DAVID platform, were listed on the webpage, where users can click a particular pathway to directly retrieve the entries involved (Fig. 3). For the first-time users of HIVed, we provide a step-by-step document on the ‘Help’ webpage to guide readers to explore HIVed efficiently (Fig. 2c). This page covers the detailed usage of all the functionalities including search, browsing, annotation display and user submission provided in HIVed. In addition, this page offers the link to download the whole HIVed database in the SQL format.

We endeavoured to make the entry search in HIVed database convenient and straightforward. We provided two options for database search, including database ID search (Fig. 4a) and keyword search (Fig. 4b). Apart from UniProt ID, we generated independent IDs for indexing entries in HIVed. An example is available for database ID search by simply clicking the ‘Example’ button. Keyword search options, on the other hand, are flexible in order to meet different user requirements (Fig. 4b). HIVed allows users to search using a number of keywords in terms of the basic gene/protein information and functional annotations. These keywords include protein/gene name, HIV-1 interaction partner name, metabolic/signalling pathway and known protein-drug interaction annotations. Additionally, we provided a search function that allows users to promptly retrieve up- or down-regulated genes across all the experimental datasets integrated. The examples can assist users to search with pre-stored keywords. The search results will be displayed on a separate webpage with links to each detailed entry, as well as gene/protein names (Fig. 4), respectively.

After clicking the link for a particular entry, the extracted protein information, gene expression data and functional annotations will be shown on different panels of a separate webpage (Fig. 5). For each entry, seven panels including protein information, gene expression profile, protein overview, drug–protein interaction, protein secondary structure, HIV-1 interaction, and metabolic/signalling pathway, will be displayed to provide a comprehensive overview. To compare the up- and down-regulated genes across different datasets, the rows have been highlighted with different colours in the ‘Gene Expression Profile’ panel. Two widely used JavaScript plug-ins, Protein Feature View (https://github.com/andreasprlic/proteinfeatureview) and PV (https://github.com/biasmv/pv) were used to enhance the data visualization and promote users’ experiences (Fig. 5d and f). Protein Feature View plug-in offers an integrated view of proteins with their functional sites, domains and secondary structures. The PV plug-in was used to provide the protein secondary structures in a specific but fast way to the users to obtain a general understanding of protein conformational information. Detailed user instructions are available in the online manual that can guide the users to be acquainted with the use of the database.
User submission is another useful feature of the database (Fig. 2d). We would like to encourage immunologists and biochemists to submit their new discoveries related to human gene expression mediated by HIV infection, replication and latency to enhance the coverage and quality of HIVed. Additionally, entries in the database will be updated on a regular basis, by collecting entries that result from up-to-date studies pertinent to HIV latency, replication, and infection. Users can submit a new entry using the Submission module or send us an Email. A link has been provided on the submission page to assist the users to send Emails to database administrator for help. This will enable the database to keep pace with the rapid proliferation of HIV infection, replication and latency studies.

Figure 2. **An overview of convenient functionalities provided by HIVed.** (a) Short paths provided on the home page. (b) A snapshot of the database statistics page. (c) Online documentation for database guidance. (d) User submission page.
In total, 14,318 human genes have been successfully mapped to the UniProt database with detailed functional annotations. To ensure the quality of the annotations stored in the HIVed database, those misregulated genes that could not be mapped to their protein products were disregarded. The statistics for the structural and functional annotations can be found in Table 1. One should bear in mind that the numbers listed in Table 1 should not be interpreted as straightforward biological evidence, due to the potential incompleteness of the current curated dataset in the HIVed database, and should vary upon the subsequential updates of the HIVed. In addition, we were not able to conduct statistical analysis for up- and down-regulated genes, respectively, due to the fact that the raw datasets collected in this study were derived from different tissues and experimental protocols.

HIVed statistics. In total, 14,318 human genes have been successfully mapped to the UniProt database with detailed functional annotations. To ensure the quality of the annotations stored in the HIVed database, those misregulated genes that could not be mapped to their protein products were disregarded. The statistics for the structural and functional annotations can be found in Table 1. One should bear in mind that the numbers listed in Table 1 should not be interpreted as straightforward biological evidence, due to the potential incompleteness of the current curated dataset in the HIVed database, and should vary upon the subsequential updates of the HIVed. In addition, we were not able to conduct statistical analysis for up- and down-regulated genes, respectively, due to the fact that the raw datasets collected in this study were derived from different tissues and experimental protocols.

Comparison with other existing gene expression databases for HIV infection, replication and latency. As aforementioned, current databases mainly focus on the gene expression data derived from
different experimental settings, such as Litchi21, Peachi28 and GuavaH29. These online resources provide useful information regarding the dysregulated human genes and graphical annotations showing the gene expression levels at a single time point and/or during a time course. Among these databases, GuavaH provides a variety of useful gene annotations in the perspective of HIV acquisition, viral load, durable control and HIV gene variation. It is very important to note that none of these databases focuses on the products of the dysregulated genes and their functions. Differentiating from these databases, HIVed is featured as a protein-centric knowledgebase with a special focus on the protein product of the dysregulated human genes during HIV infection, replication and latency, by combining and comparing the mainstream gene expression datasets across different experimental protocols. HIVed will enable the virologists and immunologists to easily map the human genes disturbed (or perturbed) by HIV to their functional annotations and generate hypotheses for HIV cure research.

Figure 5. A series of screenshots showing the detailed structural and functional annotations of a particular entry in the HIVed database. (a) Overview of the webpage. (b) Protein basic information panel. (c) Gene expression panel. (d) Overview of protein functional site/domain using the Protein Feature View plug-in. (e) Drug-protein interaction panel. (f) Protein secondary structure panel with detailed portrait via the PV plug-in. (g) HIV-1 protein interaction panel. (h) Signalling/metabolic pathway panel.
Conclusion

In this study, we propose an open-access HIV expression database, HIVed, to provide proteome-wide annotations of differentially expressed genes and their protein products during HIV infection, replication and latency, by systematically integrating a number of experimental studies across annotations protocols. By integrating a variety of biological annotations including protein basic description, Gene Ontology terms, protein confirmation, HIV-1 interaction and pathway annotations, HIVed can be seen as a comprehensive protein-centric database that allows users to input and look up genes/proteins of interest. While we acknowledge the disparate sources from which the data has been collected, comparison of such data has been difficult in the past and HIVed aims to display unbiased data reported by authors of individual studies. We anticipate this database will greatly benefit and facilitate the functional annotation and generation of novel hypotheses related to HIV-perturbed host genes and proteins with an ability to display common threads across various HIV latency and infection conditions and measurements.

References

1. Rosenberg, E. S. et al. Immune control of HIV-1 after early treatment of acute infection. Nature 407, 523–526, doi: 10.1038/35035103 (2000).
2. Douek, D. C. et al. HIV preferentially infects HIV-specific CD4+ T cells. Nature 417, 95–98, doi: 10.1038/417095a (2002).
3. Alimonti, J. B., Ball, T. B. & Fowke, K. R. Mechanisms of CD4+ T lymphocyte cell death in human immunodeficiency virus infection and AIDS. The journal of general virology 84, 1649–1661, doi: 10.1099/vir.0.19110-0 (2003).
4. Pantaleo, G. & Fauci, A. S. Immunopathogenesis of HIV infection. Annual review of microbiology 50, 825–854, doi: 10.1146/annurev.micro.50.1.825 (1996).
5. Moir, S., Chun, T. W. & Fauci, A. S. Pathogenic mechanisms of HIV disease. Annual review of pathology 6, 223–248, doi: 10.1146/annurev.pathol.011110-130254 (2011).
6. Oguntibeju, O. O. Quality of life of people living with HIV and AIDS and antiretroviral therapy. HIV AIDS (Auckl) 4, 117–124, doi: 10.1016/j.hivma.2012.08.010 (2012).
7. Zhou, H. et al. Genome-scale RNAi screen for host factors required for HIV replication. Cell host & microbe 0100-x (2012).
8. Zhou, H. et al. Precise Quantitation of the Latent HIV-1 Reservoir: Implications for Eradication Strategies. Proceedings of the National Academy of Sciences 110, 825–854, doi: 10.1146/annurev.micro.223–012818 (2015).
9. Tyagi, M. & Bukrinsky, M. Human Immunodeficiency Virus (HIV) Latency: The Major Hurdle in HIV Eradication. PLoS pathogens 9, e1003834, doi: 10.1371/journal.ppat.1003834 (2013).
10. Immune response of CD4+ T cells during primary HIV-1 infection. Journal of virology 87, 9148–9158, doi: 10.1128/JVI.00252-11 (2011).
11. Sherrill-Mix, S., Ocwieja, K. E. & Bushman, F. D. Gene activity in primary T cells infected with HIV89.6: intron retention and induction of genomic repeats. Retrovirology 12, 79, doi: 10.1186/s12977-015-0205-0 (2015).
12. Chang, S. T. et al. Next-generation sequencing reveals HIV-1-mediated suppression of T cell activation and RNA processing and regulation of noncoding RNA expression in a CD4+ T cell line. mBio 2, doi: 10.1128/mBio.00134-11 (2011).
13. Sherrill-Mix, S., Ocwieja, K. E. & Bushman, F. D. Gene activity in primary T cells infected with HIV89.6: intron retention and induction of genomic repeats. Retrovirology 12, 79, doi: 10.1186/s12977-015-0205-0 (2015).
14. Gelet, A. et al. Distinct transcriptional profiles in ex vivo CD4+ and CD8+ T cells are established early in human immunodeficiency virus type 1 infection and are characterized by a chronic interferon response as well as extensive transcriptional changes in CD8+ T cells. Journal of virology 81, 3477–3486, doi: 10.1128/JVI.01522-06 (2007).

Scientific Reports | 7:45509 | DOI: 10.1038/srep45509
32. Navare, A. T. et al. Quantitative proteomic analysis of HIV-1 infected CD4+ T cells reveals an early host response in important biological pathways: protein synthesis, cell proliferation, and T-cell activation. *Virology* **429**, 37–46, doi: 10.1016/j.virol.2012.03.026 (2012).

33. Uniprot Consortium, T. U. UniProt: a hub for protein information. *Nucleic acids research* **43**, D204–212, doi: 10.1093/nar/gku989 (2015).

34. Schoggins, J. W. & Rice, C. M. Interferon-stimulated genes and their antiviral effector functions. *Current opinion in virology* **1**, 519–525, doi: 10.1016/j.coviro.2011.10.008 (2011).

35. Lu, J. et al. The IFITM proteins inhibit HIV-1 infection. *Journal of virology* **85**, 2126–2137, doi: 10.1128/JVI.01531-10 (2011).

36. Liu, L. et al. A whole genome screen for HIV restriction factors. *Retrovirology* **8**, 94, doi: 10.1186/1742-4690-8-94 (2011).

37. Smith, A. J. et al. Host genes associated with HIV-1 replication in lymphatic tissue. *J Immunol* **185**, 5417–5424, doi: 10.4049/jimmunol.1002197 (2010).

38. Berman, H. M. et al. The Protein Data Bank, *Nucleic acids research* **28**, 235–242 (2000).

39. Law, V. et al. DrugBank 4.0: shedding new light on drug metabolism. *Nucleic acids research* **42**, D1091–1097, doi: 10.1093/nar/gkt1068 (2014).

40. Ako-Adjei, D. et al. HIV-1, human interaction database: current status and new features. *Nucleic acids research* **43**, D566–570, doi: 10.1093/nar/gku1204 (2015).

41. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic acids research* **44**, D457–462, doi: 10.1093/nar/gkv1070 (2016).

42. Chatr-Aryamontri, A. et al. The BioGRID interaction database: 2015 update. *Nucleic acids research* **43**, D470–478, doi: 10.1093/nar/gku1204 (2015).

43. Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L. P., Marcit-Houben, M. & Gabaldon, T. PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome. *Nucleic acids research* **42**, D897–902, doi: 10.1093/nar/gkt177 (2014).

44. Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols* **4**, 44–57, doi: 10.1038/nprot.2008.211 (2009).

45. Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research* **37**, 1–13, doi: 10.1093/nar/gkn923 (2009).

**Acknowledgements**

The authors would like to thank all the authors whose experimental data have been included in the database and the curators of several public databases, for making this work possible. This work was supported by Australian National Health and Medical Research Council (NHMRC) (ID 1084283). A.W.P. is supported by Australian NHMRC Senior Research Fellowship (1044215). S.H.R. was supported by an Australian Postgraduate Award.

**Author Contributions**

A.W.P., C.L. and S.H.R. conceived and designed the project; C.L. and S.H.R. performed data collection, constructed database and drafted the manuscript; J.R. and C.L. contributed to the server construction and maintenance. G.K. and J.S. provided critical feedback for the database and participated in the discussion for data analysis. All authors revised and approved the manuscript.

**Additional Information**

**Competing Interests:** The authors declare no competing financial interests.

**How to cite this article:** Li, C. et al. HIVed, a knowledgebase for differentially expressed human genes and proteins during HIV infection, replication and latency. *Sci. Rep.* **7**, 45509; doi: 10.1038/srep45509 (2017).

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

© The Author(s) 2017