Short Communication

Cataloguing the HIV Type 1 Human Protein Interaction Network

Roger G. Ptak,¹ William Fu,¹ Brigitte E. Sanders-Beer,^{1,*} Jonathan E. Dickerson,²
John W. Pinney,² David L. Robertson,² Mikhail N. Rozanov,³ Kenneth S. Katz,³
Donna R. Maglott,³ Kim D. Pruitt,³ and Carl W. Dieffenbach⁴

Abstract

Although many interactions between HIV-1 and human proteins have been reported in the scientific literature, no publicly accessible source for efficiently reviewing this information was available. Therefore, a project was initiated in an attempt to catalogue all published interactions between HIV-1 and human proteins. HIV-related articles in PubMed were used to develop a database containing names, Entrez GeneIDs, and RefSeq protein accession numbers of interacting proteins. Furthermore, brief descriptions of the interactions, PubMed identification numbers of articles describing the interactions, and keywords for searching the interactions were incorporated. Over 100,000 articles were reviewed, resulting in the identification of 1448 human proteins that interact with HIV-1 comprising 2589 unique HIV-1-to-human protein interactions. Preliminary analysis of the extracted data indicates 32% were direct physical interactions (e.g., binding) and 68% were indirect interactions (e.g., up-regulation through activation of signaling pathways). Interestingly, 37% of human proteins in the database were found to interact with more than one HIV-1 protein. For example, the signaling protein mitogen-activated protein kinase 1 has a surprising range of interactions with 10 different HIV-1 proteins. Moreover, large numbers of interactions were published for the HIV-1 regulatory protein Tat and envelope proteins: 30% and 33% of total interactions identified, respectively. The database is accessible at <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions/> and is cross-linked to other National Center for Biotechnology Information databases and programs via Entrez Gene. This database represents a unique and continuously updated scientific resource for understanding HIV-1 replication and pathogenesis to assist in accelerating the development of effective therapeutic and vaccine interventions.

TWENTY-FIVE YEARS AFTER ITS DISCOVERY, human immunodeficiency virus type 1 (HIV-1) still has a major health and socioeconomic impact, particularly in developing countries. Approximately 33 million people worldwide are infected with HIV-1, with an estimated 2.5 million new infections and 2.1 million AIDS-related deaths occurring each year.¹ The HIV-1 genome expresses 17 processed proteins: six structural proteins, matrix (MA), capsid (CA), nucleocapsid (NC; p7), p6, p2, and p1 derived from viral protease-mediated cleavage of the HIV-1 Gag polyprotein; three enzymes, reverse transcriptase (RT), which has both RT and ribonuclease H (RNase H) enzymatic activities, protease (PR), and integrase (IN) derived from cleavage of the HIV-1 Pol

polyprotein by PR; two envelope glycoproteins, gp120 (SU) and gp41 (TM), derived from cleavage of the HIV-1 Env (gp160) protein by cellular proteases; two regulatory proteins, transactivator (Tat) and regulator of viral protein expression (Rev); and four accessory proteins, negative factor (Nef), viral infectivity factor (Vif), viral protein R (Vpr), and viral protein U (Vpu). Currently, FDA-approved anti-HIV drugs (<http://www.fda.gov/oashi/aids/virals.html>) target only a limited number of proteins: RT, PR, IN, gp41, as well as the cellular protein CCR5, which, along with the CD4 receptor, serves as a co-receptor for HIV-1 entry into host cells.² In addition, no HIV-1 vaccine has shown efficacy, in spite of considerable efforts directed in this area. Vaccines under inves-

¹Southern Research Institute, Frederick, Maryland 21701.

²Faculty of Life Sciences, University of Manchester, Manchester M13 9PL, UK.

³National Center for Biotechnology Information, Bethesda, Maryland 20892.

⁴Division of AIDS, NIAID, NIH, Bethesda, Maryland 20892.

*Current address: BIOQUAL, Inc., Rockville, Maryland 20850.

tigation mainly rely on immune responses to HIV Gag, Pol, and Env, but often include regulatory (Rev and Tat) or accessory proteins (Vpr, Nef, and Vif).³

To make progress in understanding HIV-1 replication, and the mechanisms of viral pathogenesis, it is important to understand the interactions of HIV-1 proteins with the vast array of human cellular proteins exploited for virus replication.⁴ While these interactions can be direct viral–host cell protein–protein interactions, many are indirect, such as the regulatory interactions that alter expression of a human gene. An in-depth review and comprehension of these interactions enhance insight into HIV pathogenesis on a cellular level and are also essential for focusing efforts on new drug targets and for better understanding vaccine immune responses. Since 1984 the HIV/AIDS research community has published extensively on virus–host interactions, but unlike HIV sequence, resistance, and immunology data (<http://www.hiv.lanl.gov>), no searchable database had been developed to efficiently retrieve information relevant to interactions between HIV-1 and human proteins. To fill this important gap, the Division of AIDS of the National Institute of Allergy and Infectious Diseases initiated the development of an HIV-1 Human Protein Interaction Database that serves as a freely accessible, continuously updated, and comprehensive resource for the scientific community. The goals of this report are to describe the

development of the database, summarize its contents, and present a visualization of the catalogued data demonstrating the complexity of the HIV-1 human protein interaction network.

To facilitate the development of a database describing HIV-1 and human protein interactions, the annotation on the HIV-1 reference sequence (RefSeq nucleotide accession number NC_001802.1) was first improved and expanded. RefSeq⁵ is a database of nucleotide and protein reference sequences annotated and distributed by the National Center for Biotechnology Information (NCBI). The HIV-1 reference sequence was updated in order to provide functional information on the HIV-1 polypeptides and mature peptides and to use standard names. Explicit representation of each mature peptide product supports more specific and accurate cataloguing of the interactions. All references within the database are catalogued using this primary reference sequence and the associated protein accession numbers. Subsequently, literature indexed in PubMed was searched using keywords for each of the HIV-1 proteins [e.g., “HIV and (matrix or p17)”. The article titles and abstracts that were retrieved were reviewed to identify papers describing interactions between HIV-1 and human proteins. Relevant publications were collected and used to manually catalogue interactions into a database containing the (1) protein names,

TABLE 1. SUMMARY OF THE HIV-1 HUMAN PROTEIN INTERACTION DATABASE CONTENTS

<i>HIV-1 protein</i>	<i>Total number of interactions</i>	<i>Number of direct interactions</i>	<i>Number of indirect interactions</i>	<i>Number of interaction descriptions</i>	<i>Number of PMID references</i>
Gag proteins total	208	127	81	330	867
Gag	55	26	29	81	103
Matrix (MA), p17	78	55	23	132	411
Capsid (CA), p24	33	14	19	59	197
Nucleocapsid (NC), p7	25	21	4	29	55
p6	14	11	3	26	98
p2	0	0	0	0	0
p1	3	0	3	3	3
Pol proteins total	207	130	77	268	542
Pol	1	1	0	1	1
Protease (PR)	84	63	21	99	154
Reverse transcriptase (RT)	42	2	40	46	58
Integrase (IN)	80	64	16	122	329
Env proteins total	864	195	669	1737	5508
Env (gp160)	176	57	119	232	681
gp120	532	86	446	1252	4321
gp41	156	52	104	253	506
Signal peptide	0	0	0	0	0
Regulatory proteins total	841	264	577	1826	4476
Tat	775	219	556	1686	4253
Rev	66	45	21	140	223
Accessory proteins total	469	118	351	974	2919
Nef	200	51	149	496	980
Vpu	22	10	12	35	102
Vpr	179	37	142	347	832
Vif	68	20	48	96	1005
All HIV-1 proteins total	2589	834	1755	5135	14,312

FIG. 1. Visualization of the HIV-1 human protein interaction network. The network was visualized with InterView.¹¹ Gray ovals represent HIV-1 proteins. Colored circles represent human proteins and are shown clustering around the HIV-1 protein they interact with. Human proteins that interact with multiple viral proteins are shown toward the center of the image rather than clustered around a specific HIV-1 protein. Colors correspond to human protein categories based on cellular component Gene Ontology (GO) terms.⁹ The number of proteins in each category and percentage of the total 1448 human proteins interacting with HIV-1 are indicated in parentheses. Black lines correspond to direct interactions (e.g., binding) and gray lines to indirect interactions (e.g., downregulation). Circles for specific human proteins of interest discussed in the text (MAPK1, MAPK3, PRKCA, and IFNG) are indicated with dashed red lines. Overlapping circles are a result of the visualization process and do not imply any specific relationship. See supporting online materials for alternative visualizations of the HIV-1 human protein interaction network based on biological process and molecular function GO terms, and for information about specific human proteins that interact with multiple viral proteins.¹⁰

(2) Entrez GeneIDs (species-specific gene identification numbers from Entrez Gene, NCBI's database for gene-specific information),⁶ (3) RefSeq protein accession numbers, (4) brief descriptions of the interactions, (5) PubMed identification numbers (PMIDs) of articles describing the interactions, and (6) keywords for searching the interactions. Since the articles were from peer-reviewed publications, all identified interactions were incorporated into the database without placing additional judgment on the scientific validity of the report. If two different reports had conflicting conclusions, a comment describing the data ambiguity was incorporated into the interaction descriptions. For the majority of publications the full-text paper was reviewed, but some interactions were catalogued based on abstract only (e.g., when abstracts contained complete descriptions of the interactions or full copies of the articles could not be obtained). Upon completion, the searchable database was uploaded into the HIV-1 protein interactions section of NCBI's Entrez Gene at <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions>. The database is cross-referenced to other NCBI resources⁷ through Entrez Gene, providing a discovery space for searching and understanding the interaction data. Continued cataloguing of interactions and updating of the database to include information from new articles are performed on a regular basis.

The HIV-1 human interaction data can be accessed at NCBI through three mechanisms: (1) Entrez Gene, (2) the dedicated web site, or (3) by file transfer protocol (FTP). Records in Entrez Gene that contain HIV-1 interaction data can be retrieved with the query: "hiv1interactions"[Properties] AND "Homo sapiens"[Organism]. The "Table of Contents" for each gene record provides a link to the section entitled "HIV-1 protein interactions" where data are presented as a table. For human gene records, the table includes a link to the gene record for the HIV-1 protein, the interaction comment, and links to the PubMed citation(s) supporting the comment. On HIV-1 records, the table reports the interaction keyword, provides a link to the corresponding human interactant gene record, and links to the supporting PubMed citation(s). The "Links" menu for each gene record also allows users to readily navigate from Entrez Gene to other NCBI resources to find additional data related to the interacting proteins. For example, gene expression data can be retrieved via the link to "GEO Profiles."⁸ Similarly, all gene records that contain HIV-1 interaction data and that also have gene expression data in the Geo⁸ database can be identified in Entrez Gene using the query: "hiv1interactions"[Properties] AND gene_geo[filter]. Many other queries can be designed to retrieve and study similar subsets of the database. For instance, using specific Gene Ontology⁹ (GO) terms, queries can be built to find all HIV-1 interacting proteins that are located in the cytoplasm (e.g., "hiv1interactions"[Properties] AND cytoplasm[go]) or that are involved in stem cell maintenance (e.g., "hiv1interactions"[Properties] AND "stem cell maintenance"[go]). Furthermore, in addition to Entrez Gene, the HIV-1 human protein interaction database web site lists all interaction types described per HIV-1 gene. Reports include the HIV-1 protein, the interaction type, and the human protein (linked to the gene record). Report pages include a query interface that facilitates accessing gene records that have a specific type of interaction (such as activation) with a specific HIV-1 protein. Results

from these queries return as a list with links to the gene records. The report interface also provides an option to download all, or a subset, of the data in a columnar text format. Finally, the complete dataset can be transferred by FTP from <ftp://ftp.ncbi.nih.gov/gene/GeneRIF/> as the file `hiv_interactions.gz`.

For the entire database, over 100,000 journal abstracts published between 1984 and 2007 were identified by PubMed queries and further reviewed, leading to the identification of 3200 papers describing putative interactions between HIV-1 and human proteins. Table 1 summarizes the HIV-1 protein interactions catalogued from these papers (see supporting online materials for a listing of all interactions).¹⁰ A total of 1448 human proteins that interact with HIV-1, comprising 2589 unique HIV-1-to-human protein interactions, were identified. In addition, 5135 summary descriptions of the interactions were generated, with a total of 14,312 PMID references to the original articles that reported the interactions. Sixty-eight unique keywords (directional from HIV-1 protein to human protein) are associated with these descriptions. Keywords were selected based on the text in the original journal articles by identifying the most important functional keyword used by the authors to describe the interaction. Whenever possible, similar keywords were combined. For example, "downregulates" and "downmodulates" were combined to use the single keyword "downregulates." The most pervasive keywords used in the database are "interacts with," 17.0%; "upregulates," 11.9%; "binds," 11.7%; "activates," 9.7%; "downregulates," 7.8%; "inhibits," 6.9%; "inhibited by," 4.1%; "processed by," 2.6%; "regulated by," 2.1%; and "phosphorylated by," 1.5% (see supporting online materials for a listing of all keywords).¹⁰ While it cannot be excluded that some of these interactions are nonspecific or human-prone errors, nevertheless, 58% of the interactions were confirmed by more than one publication. Collectively, the catalogued interactions provide a unique collection of data generated from the available scientific literature.

To demonstrate the complexity of the HIV-1 human protein interaction network, the catalogued data were visualized using InterView¹¹ and Gene Ontology⁹ (GO) terms (Fig. 1). The Gene Ontology is a set of three structured controlled ontologies that describes gene products in terms of their as-

TABLE 2. NUMBER OF INTERACTING HIV-1 PROTEINS PER CELLULAR PROTEIN

<i>Number of human proteins</i>	<i>Interacting with N HIV-1 proteins</i>	<i>Percent of total</i>
919	N = 1	63.47
219	N = 2	15.12
172	N = 3	11.88
55	N = 4	3.80
40	N = 5	2.76
18	N = 6	1.24
17	N = 7	1.17
4	N = 8	0.28
3	N = 9	0.21
1	N = 10	0.07
1448	Total	100

sociated cellular component (Fig. 1), biological process, or molecular function in a species-independent manner. GO terms were collected from Entrez Gene for each HIV-interacting human protein (see supporting online materials for alternative visualizations based on biological processes and molecular function GO terms, and for distribution of interactions based on GO terms).¹⁰ Proteins without GO terms in the three ontologies were annotated as unknown by using the root ontology term (cellular component, 14%; biological process, 9.5%; and molecular function, 7.8%). These visualizations reveal the extent to which HIV-1 interacts with diverse human proteins and demonstrate there are many examples of the virus interacting with human proteins that are part of the same functional category. Interestingly, the majority of interactions, 68%, are indirect (e.g., altered expression of a human protein), while only 32% are direct physical interactions (e.g., binding) (Table 1). In addition, 529 (37%) of the human proteins in the database were found to interact with more than one HIV-1 protein. For example, the signaling protein mitogen-activated protein kinase 1 (MAPK1) has a surprising range of interactions with 10 different HIV-1 proteins. MAPK1 is a member of the MAP kinase family involved in a wide variety of cellular processes such as proliferation, differentiation, transcription regulation, and development. Thus, it is likely that MAPK1 is intimately involved in many steps of the HIV-1 replication cycle. Similarly, mitogen-activated protein kinase 3 (MAPK3), protein kinase C- α (PRKCA), and interferon- γ (IFNG) have been described as interacting with nine different HIV-1 proteins each, suggesting these proteins also play important roles in HIV-1 replication and pathogenesis.

Table 2 summarizes the distribution and number of interacting HIV-1 proteins per cellular protein (see supporting online materials for a comprehensive listing of the specific HIV-1 proteins interacting with each cellular protein).¹⁰ Moreover, large numbers of interactions were published for the HIV-1 regulatory protein Tat, as well as for the envelope proteins: 30% and 33% of the total interactions identified, respectively. Of particular note is the cataloguing of 273 different nuclear proteins that interact with Tat, including over 40 transcription factors and regulators. Similarly, 219 extracellular and plasma membrane proteins were identified as interacting with HIV-1 gp120 and 67 were identified as interacting with gp41, including over 70 cellular receptors, integrins, and adhesion molecules. Overall, the database catalogues a wealth of information that can be mined to understand better the breadth of the HIV-1 human protein interaction network.

In conclusion, the HIV-1 human protein interaction network forms the basis of a detailed map for tracking the cellular interactions that drive HIV-1 replication and pathogenesis. Integration of the database into NCBI's online resources⁷ (<http://www.ncbi.nlm.nih.gov/Database/index.html>) has significantly expanded the gene-specific information reported and provides a discovery space to the scientific community for researching and better understanding the impact of HIV-1 on the cell. Furthermore, efforts are underway to incorporate these data into Canada's Biomolecular Object Network Databank (BOND) (<http://bond.unleashedinformatics.com>), a database cataloguing the interactions between all cellular proteins, which will provide a powerful tool for advancing the available knowledge of how HIV-1 replicates in the context of the whole cell and for broadening our understanding of how these interac-

tions control HIV-1 replication and mediate pathogenic effects on infected cells. This set of known interactions between HIV-1 and human proteins is the first step toward new strategies for inhibiting HIV-1 replication (e.g., targeting of nonessential host-cell proteins).¹² As an example of how this resource has already had a significant positive impact on HIV research, Brass *et al.*¹³ recently published a study describing the use of an RNAi screen that identified a set of 273 cellular proteins involved in HIV-1 replication, which they termed HIV-dependency factors (HDFs).¹⁴ The HIV-1 Human Protein Interaction Database was used extensively by the authors to support the categorization and analysis of the HDFs identified through this research. Similarly, a subset of the HIV-1 Human Protein Interaction Database that has already been uploaded into the Biomolecular Interaction Network Database¹⁵ (BIND; predecessor to BOND) supported the work of Dyer *et al.*,¹⁶ who performed a broad analysis of human proteins interacting with viruses and other pathogens. The future integration of the human protein interaction network,¹⁷ expression data,^{8,18} cell-type, disease stage, and other factors will permit better insight into our understanding of the HIV-1 human protein interaction network and will help accelerate the development of more effective antiviral and vaccine interventions.

Acknowledgments

We thank Joel Gillman for providing database support. R.G.P., W.F., and B.E.S.-B. were funded in whole with federal funds from the NIH, NIAID under Contracts N01-AI-05415 and N01-AI-70042, and wish to thank Dr. Roger Miller, Project Officer, DAIDS. J.E.D. was supported by a Wellcome Trust studentship and J.W.P. by a BBSRC project grant (BB/C515412/1). M.N.R., K.S.K., D.R.M., and K.D.P. were supported by the Intramural Research Program of the NIH, National Library of Medicine. R.G.P. and W.F. catalogued the interactions and curated the database with support from B.E.S.-B.; J.E.D., J.W.P., and D.L.R. carried out the Gene Ontology analysis and visualization of the interactions; M.N.R. and R.G.P. generated the improved NC_001802.1 RefSeq annotation; K.S.K., D.R.M., and K.D.P. generated the open access database resource at NCBI and incorporated the data into Entrez Gene; C.W.D. provided the original concept for the database and support for the project through NIAID/DAIDS.

Disclosure Statement

No competing financial interests exist.

References

1. UNAIDS: 2007 AIDS epidemic update. December 2007. Available from http://data.unaids.org/pub/EPISlides/2007/2007_epiupdate_en.pdf.
2. Emmelkamp JM and Rockstroh JK: CCR5 antagonists: Comparison of efficacy, side effects, pharmacokinetics and interactions—review of the literature. *Eur J Med Res* 2007;12:409–417.
3. Hurwitz JL, Zhan X, Brown SA, *et al.*: HIV-1 vaccine development: Tackling virus diversity with a multi-envelope cocktail. *Front Biosci* 2008;13:609–620.
4. Uetz P, Dong YA, Zeretzke C, *et al.*: Herpesviral protein networks and their interaction with the human proteome. *Science* 2006;311:239–242.

5. Pruitt KD, Tatusova T, and Maglott DR: NCBI reference sequences (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* 2007;35(Database issue):D61–65.
6. Maglott D, Ostell J, Pruitt KD, and Tatusova T: Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Res* 2007;35(Database issue):D26–31.
7. Wheeler DL, Barrett T, Benson DA, *et al.*: Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 2008;36(Database issue):D13–21.
8. Barrett T, Troup DB, Wilhite SE, *et al.*: NCBI GEO: Mining tens of millions of expression profiles—database and tools update. *Nucleic Acids Res* 2007;35(Database issue):D760–765.
9. Ashburner M, Ball CA, Blake JA, *et al.*: Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet* 2000;25:25–29.
10. Supporting online materials are available from the HIV-1 Human Protein Interaction Database web site at <http://www.ncbi.nlm.nih.gov/RefSeq/HIVInteractions>.
11. Holden BJ, Pinney JW, Lovell SC, Amoutzias GD, and Robertson DL: An exploration of alternative visualisations of the basic helix-loop-helix protein interaction network. *BMC Bioinformatics* 2007;8:289.
12. Westby M and van der Ryst E: CCR5 antagonists: Host-targeted antivirals for the treatment of HIV infection. *Antiviral Chem Chemother* 2005;16:339–354.
13. Brass AL, Dykxhoorn DM, Benita Y, *et al.*: Identification of host proteins required for HIV infection through a functional genomic screen. *Science* 2008;319:921–926.
14. Cohen J: HIV gets by with a lot of help from human host. *Science* 2008;319:143–144.
15. Gilbert D: Biomolecular interaction network database. *Brief Bioinform* 2005;6:194–198.
16. Dyer MD, Murali TM, and Sobral BW: The landscape of human proteins interacting with viruses and other pathogens. *PLoS Pathog* 2008;4:e32.
17. Rual JF, Venkatesan K, Hao T, *et al.*: Towards a proteome-scale map of the human protein-protein interaction network. *Nature* 2005;437:1173–1178.
18. Bandyopadhyay S, Kelley R, and Ideker T: Discovering regulated networks during HIV-1 latency and reactivation. *Pacific Symp Biocomput* 2006;11:354–366.

Address reprint requests to:

Roger G. Ptak

Department of Infectious Disease Research

Southern Research Institute

431 Aviation Way

Frederick, Maryland 21701

E-mail: ptak@southernresearch.org