OPEN ACCESS Freely available online

PLoS one

Viral Organization of Human Proteins

Stefan Wuchty[1]*, Geoffrey Siwo[2], Michael T. Ferdig[2]

1 National Center of Biotechnology Information, National Institutes of Health, Bethesda, Maryland, United States of America, **2** Eck Institute for Global Health, Department of Biology, University of Notre Dame, Notre Dame, Indiana, United States of America

Abstract

Although maps of intracellular interactions are increasingly well characterized, little is known about large-scale maps of host-pathogen protein interactions. The investigation of host-pathogen interactions can reveal features of pathogenesis and provide a foundation for the development of drugs and disease prevention strategies. A compilation of experimentally verified interactions between HIV-1 and human proteins and a set of HIV-dependency factors (HDF) allowed insights into the topology and intricate interplay between viral and host proteins on a large scale. We found that targeted and HDF proteins appear predominantly in rich-clubs, groups of human proteins that are strongly intertwined among each other. These assemblies of proteins may serve as an infection gateway, allowing the virus to take control of the human host by reaching protein pathways and diversified cellular functions in a pronounced and focused way. Particular transcription factors and protein kinases facilitate indirect interactions between HDFs and viral proteins. Discerning the entanglement of directly targeted and indirectly interacting proteins may uncover molecular and functional sites that can provide novel perspectives on the progression of HIV infection and highlight new avenues to fight this virus.

Citation: Wuchty S, Siwo G, Ferdig MT (2010) Viral Organization of Human Proteins. PLoS ONE 5(8): e11796. doi:10.1371/journal.pone.0011796

Editor: Diego Di Bernardo, Fondazione Telethon, Italy

Received April 1, 2010; **Accepted** July 2, 2010; **Published** August 25, 2010

Funding: SW was supported by NCBI/NLM/NIH. MTF and GS are supported by NIH grant AI071121. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: wuchtys@ncbi.nlm.nih.gov

Introduction

The determination of webs of protein interactions [1,2,3,4] and protein complexes [5,6,7,8,9] in many different single and multi-cellular organisms progresses at a fast pace, peaking in attempts to determine the human interactome in various ways [10,11,12,13,14]. Although such webs of intracellular interactions are increasingly well characterized, little is known about large-scale maps of protein interactions between cells. Therefore, the investigation of host-pathogen interactions is a crucial step toward a thorough understanding of an organism's pathogenesis, providing an essential foundation for the development of effective therapeutic and prevention strategies to combat diseases. Uetz et al. released the first small map of computationally inferred physical protein interactions between the human host and the Kaposi-Sarcoma associated Herpesvirus (KSHV) as well as the Varicella Zoster-Virus (VZV) [15]. In a different approach, Calderwood et al. [16] experimentally determined a map of physical protein interactions between the Epstein-Barr-Virus and the human host. Recently, Bandyopadhyay et al. [17], identified subnetworks of virus-host proteins that are expressed at different stages of the HIV-infection and Dyer et al. compared experimentally known interactions of different viruses with the human host [18]. Brass et al. utilized a comprehensive large scale-screen of siRNAs to identify HIV dependency factor proteins (HDF). Although these proteins do not directly interact with viral proteins, they play an indirect, yet important role in the infection process of HIV [19].

Here, we pooled experimentally verified interactions between HIV-1 and human proteins, along with a set of HIV-dependency factor proteins (HDF), to investigate the topology of interactions between viral and host proteins on a large scale. We found that targeted and HDF proteins appeared predominantly in rich-clubs, allowing the virus to take control of the human host by reaching protein pathways and diversified cellular functions in a pronounced and focused way. Although HIV-1 does not physically interact with HDFs, we observed that prominent transcription factors and protein kinases establish indirect links between such host and viral proteins, suggesting molecular and functional sites that can be used to systematically hamper the virus.

Results

Rich-clubs of proteins as viral targets

Here, we utilized a compilation of 702 experimentally verified physical protein interactions between 17 HIV-1 and 519 human proteins. In addition, we accounted for 290 HIV dependency factor proteins (HDF) that play a role in the viral infection process [19]. Considering a graphical depiction of the web of all host-viral interactions in Fig. 1A, we observed a single connected subnetwork. Randomizing the human interaction partners of viral proteins, we found that the presence of one connected web is statistically significant ($P<10^{-4}$). In Fig. 1A, we also observed that Tat, Nef and Vpr – viral proteins that predominantly interfere with regulatory host processes – appear to be viral hubs that interfere with the human host interactome on a combinatorial basis. As such, we will show topological features of the human interactome that are potential direct and indirect targets of HIV-1.

In contrast to other protein interaction networks of eukaryotic organisms, such as *S. cerevisiae*, *C. elegans* and *D. melanogaster* [20,21], the human interactome is composed of an oligarchy of highly interacting and intertwined nodes. Such a rich-club phenomenon is quantified by the fraction of edges among nodes that have at

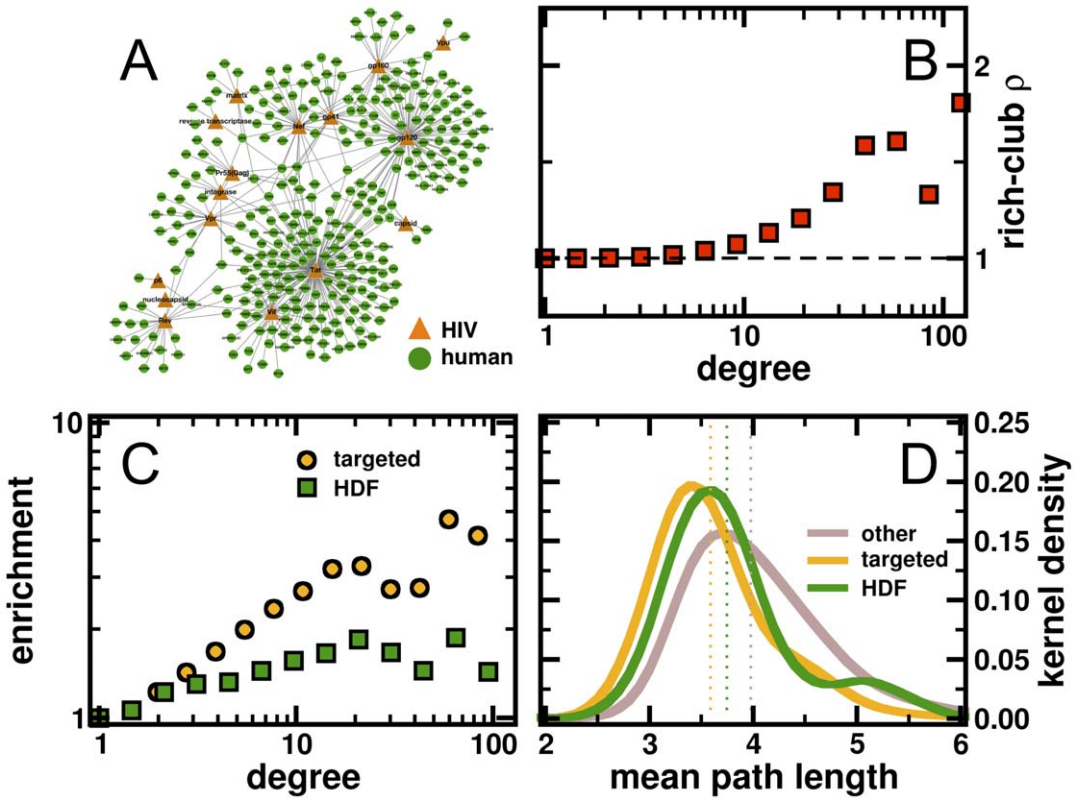

Figure 1. Statistics of targeted and HDF proteins. (**A**) Constructing a bipartite network of interactions between HIV-1 and human proteins, we find one connected subnetwork. Randomizing human proteins, we found that the presence of one connected network is statistically significant ($P<10^{-4}$). (**B**) The mean rich-club coefficient ρ reflects the degree to which proteins with at least a certain number of interaction partners are intertwined among each other. We observed that ρ significantly increased with higher degrees in the human protein interaction network, indicating the presence of an oligarchy (*i.e.* rich clubs) of highly interacting and intertwined proteins. In (**C**) we determined the enrichment of targeted human host proteins in such rich clubs. Specifically, highly connected proteins appeared to be increasingly targeted by the virus. Although weaker HIV dependency factor proteins (HDF) were enriched in rich clubs as well, a difference that is statistically significant compared to the enrichment of targeted proteins (Kolmogorov-Smirnov test, $P<0.01$). (**D**) Utilizing a network of interactions between proteins of the human host, we determined the lengths of shortest paths for each pair of human proteins. Calculating the mean of the shortest path lengths, we found that proteins, which are targeted by the virus, have lowest means (dotted lines). Focusing on HDFs we observed a shift toward longer mean path lengths, a difference that is statistically significant compared to targeted proteins (Student's t-test, $P<0.02$). This observation is reinforced if we account for all remaining human proteins, results that are significantly different in comparison to targeted and HDF proteins, respectively ($P<0.05$).
doi:10.1371/journal.pone.0011796.g001

least a certain number of neighbor's k in the actual and randomized networks. As such, the rich-club coefficient $\rho(k)$ points to the presence of a core of highly intertwined nodes with degree of at least k if $\rho(k)>1$. In the absence of this phenomenon (i.e. $\rho(k)>1$) networks are dominated by many well defined functional communities which are sparsely connected by highly interacting proteins [20]. Collecting pairwise protein interactions in *H. sapiens* from public databases and accounting for phosphorylation events between kinases and other proteins, we assembled a network of 23,752 interactions between 4,075 human proteins that are expressed in the human host cell. In this network, we found a strong rich-club signal among proteins with increasing degree (Fig. 1B). Assuming that such a proteomic feature might be an exploitable target of the virus, we hypothesized that proteins in rich clubs are preferably targeted by the pathogen. In Fig. 1C, we found that there exists an enrichment of targeted proteins in rich clubs. Although weaker, yet significantly different compared to targeted proteins (Kolmogorov-Smirnov test, $P<0.01$), we observed that HIV dependency factor (HDF) proteins are enriched in rich clubs as well. These observations suggest that samples of highly connected and intertwined proteins provide topological features which the virus utilizes as a gateway to seize control of the host cell in a direct and indirect way. As another parameter of centrality, we calculated the mean length of shortest paths from each protein. Focusing on directly targeted host proteins, we found a bell shaped curve (Fig. 1D) at relatively short path lengths. Focusing on HDFs we observed a significant shift toward longer mean path lengths compared to targeted proteins (Student's t-test, $P<0.02$). If we account for all remaining human proteins, we found this trend reinforced, a result that is significantly different in comparison to targeted and HDF proteins, respectively ($P<0.05$).

Viral aspects of pathways

Protein pathways are another level of systems information in which to recognize patterns that reveal how the virus exploits the host cell. This approach relies on the strength of 913 manually curated pathways from the Pathway Interaction Database [22]. Specifically, we tested whether pathways are significantly enriched for genes that are expressed in the human host cell. Applying a Fisher exact test, we found 851 enriched pathways ($P<0.05$, Table S1). Utilizing this set of pathways, we observed that hubs appeared in an increasing number of pathways (inset, Fig. 2A). This observation emphasizes a role of protein hubs being involved in numerous protein pathways and suggests that the pathogens have

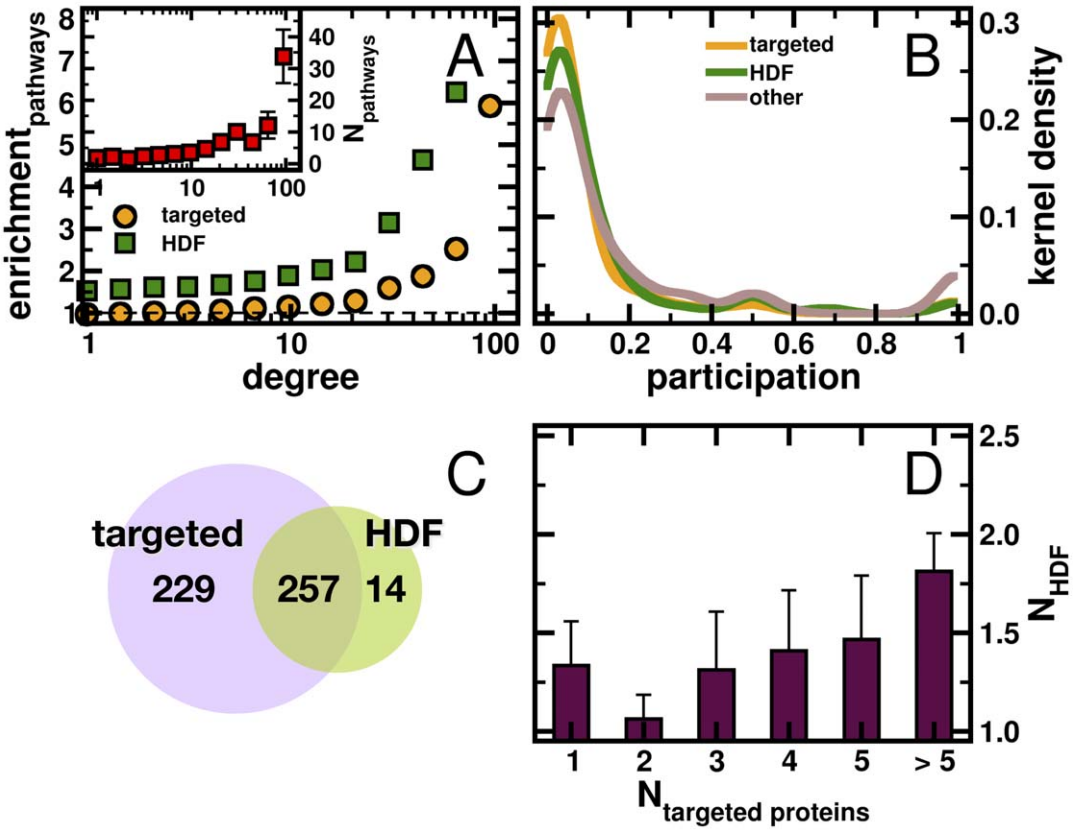

Figure 2. Pathway related characteristics of targeted and HDF proteins. The inset of (**A**) suggests that proteins in rich clubs tend to participate in an elevated number of pathways. Considering proteins that are targeted by HIV, we calculated the total number of pathways such proteins are involved in. In rich clubs, we found a trend toward strong enrichment of pathways that harbor targeted proteins in rich clubs of proteins. Analogously, we observed a similar trend with HIV dependent factor proteins (HDF) that was significantly different from targeted proteins (Kolmogorov-Smirnov test, $P<0.005$). (**B**) A low value of the pathway participation coefficient indicates that the interactions of a protein reach many different pathways and *vice versa*. Considering all targeted proteins, we obtained a maximum around low values and recovered a similar result, when we focused on all HDFs. While the difference of these distributions is insignificant (Kolmogorov-Smirnov test, $P<0.3$), the trend appears diminished if we consider all other human host proteins. In comparison to targeted and HDF proteins, respectively, differences are significant ($P<0.01$), indicating that the placement of targeted and HDF proteins buffers differences in the proteins abilities to reach into many pathways. (**C**) Counting the number of pathways that have at least one protein that is targeted by the virus we found 486, while 271 pathways had at least one HDF. Indicating a significant overlap, 257 pathways involved both targeted and HDF proteins (hypergeometric test, $P<10^{-45}$). (**D**) In these pathways, we determined the corresponding numbers of HDFs and targeted proteins. We found a significant upwarding trend, indicating that pathways that are increasingly targeted by the virus also harbor HDFs (Pearson's $r=0.2$, $P<0.01$).
doi:10.1371/journal.pone.0011796.g002

taken advantage of the host network at the pathway level. Indeed, we found that host proteins that are targeted by the virus appeared in an increasing number of pathways with increasing degree (Fig. 2A). Focusing on HDF proteins, we recovered a similar, yet significantly different trend compared to targeted proteins (Kolmogorov-Smirnov test, $P<0.005$).

A corollary of this result is that the comparably small number of targeted proteins and HDFs would allow the virus to interact with a larger number of pathways than would appear by chance. Out of 851 enriched pathways, all human proteins targeted by the virus were part of 486 pathways, while HDFs touched 271 pathways. Randomizing the sets of targeted and HDF proteins, we found that such numbers are smaller than expected by chance alone ($P<10^{-4}$).

We further hypothesized that the virus tendency to target host pathway hubs effectively mediates the infection while ensuring variety, such that the virus targets numerous distinct pathways with a comparably low number of targeted proteins. As a measure of diversity, we defined the pathway participation coefficient: if a protein predominantly interacts with partners that are members of the same pathway, this measure tends toward 1, while the opposite holds if the interaction partners of the considered protein are distributed among many different pathways. Accounting for all human proteins that are neither targeted nor HDFs, we observed that interactions of a single protein occur in a variety of pathways, as indicated by the maximum around low values of the pathway participation coefficient. In turn, relatively few interactions are confined to a small number of pathways (Fig. 2B). Comparing to the subsets of human proteins that are targeted by the virus, we found a significant reinforcement of the initial diversity signal (Kolmogorov-Smirnov test, $P<0.01$). Similarly, for HDFs we found this signal significantly reinforced ($P<0.01$) as well, confirming that the use of a small subset of host proteins effectively secures a pathogen's reach into a breadth of cellular activities without inundating any particular one. However, we found no significant differences between the corresponding distributions of targeted and HDF proteins, indicating that the placement of targeted and HDF proteins in the network is defined by a cohesive pathway-dependent combination of targeted and HDF proteins. Consequently, we searched for a correlation between the number

of targeted and HDF proteins in pathways. Counting the number of pathways that have at least one protein that is targeted by the virus we found 486. In turn, we found 271 pathways that harbored at least one HDF. In the Venn diagram in Fig. 2C, we observed a significant overlap where 257 pathways involved both targeted and HDF proteins (hypergeometric test, $P<10^{-45}$). In these pathways, we found a significantly upwarding trend between the number of targeted and HDF proteins ($r = 0.2$, $P<0.01$), confirming our hypothesis that pathways which harbor targeted proteins also significantly involve HDFs (Fig. 2D).

Direct and indirect host-virus interactions

Up to this point we considered direct interactions of human host and viral proteins and regarded HIV dependency factors as proteins that are influenced by the virus in some indirect, yet unknown way. However, the integration of information about interactions between proteins can potentially help us to uncover ways viral proteins indirectly interact with HDFs through their host targets. For example, we found protein interactions between viral proteins Tat, Nef and gp120 and TP53 (Fig. 3A). In turn, transcription factor TP53 controls the expression of the HDF protein AKT1 by a protein DNA interaction. Since Tat, Nef and gp120 are connected to AKT1 through TP53, we considered AKT1 indirectly interacting with those three viral proteins. As a consequence, the length of the shortest path of AKT1 to a directly targeted host protein is 1. To determine shortest paths, we considered phosphorylation events between kinases and other proteins as directed in our host network of physical protein-protein interactions. Furthermore, we added experimentally confirmed directed protein-DNA interactions between transcription factors and proteins. Determining shortest paths between each HIV dependency factor protein to a protein that is directly targeted by virus proteins, we found that the majority of HDFs directly interacts with a targeted protein (Fig. 3B; Table S2). We counted the number of viral proteins that target a single host protein and found that the majority of human host proteins are targeted by a single viral protein (1.4 ± 0.7, Fig. 3C). We analogously counted the number of viral proteins that HDF proteins indirectly interact with. In comparison to directly targeted host proteins, the corresponding mean value significantly shifted to higher numbers of viral proteins that indirectly interact with HDF proteins (4.6 ± 3.0), a distribution that is significantly different (Student's t-test, $P<10^{-36}$).

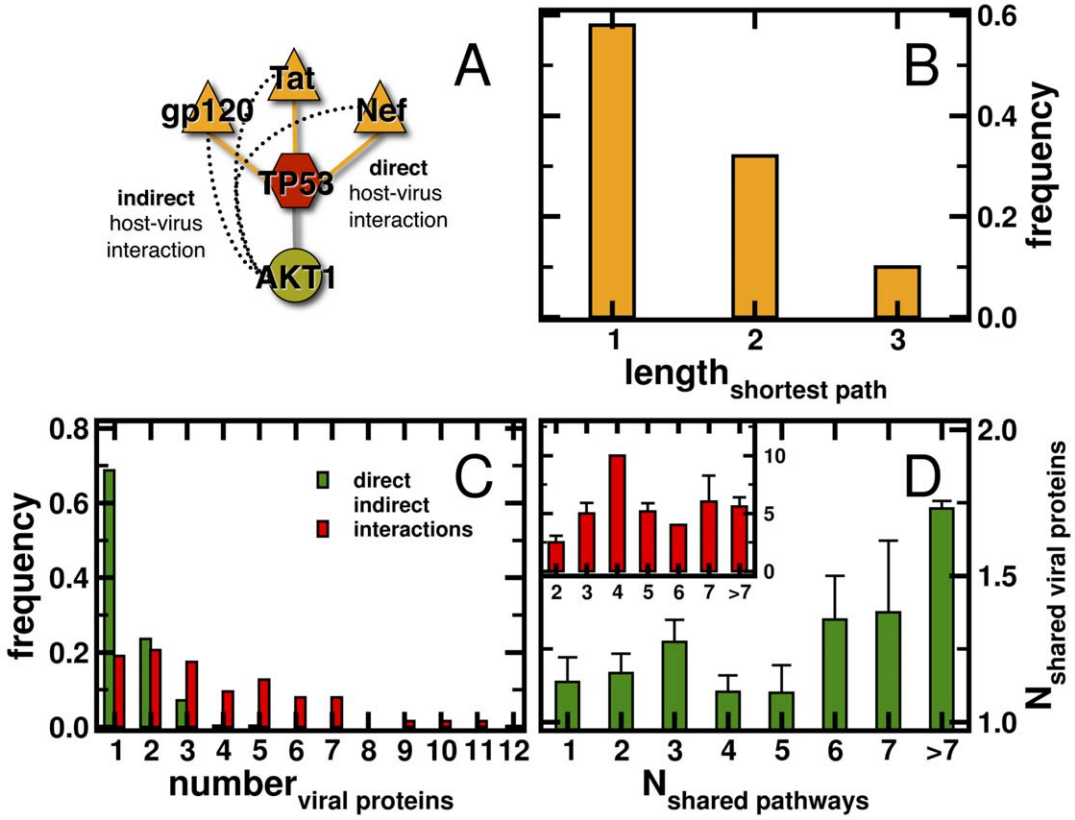

Figure 3. Direct and indirect host pathogen interactions. (**A**) As an example for direct host-virus interactions, we show physical interactions between viral proteins Tat, Nef and gp120 and TP53. In turn, transcription factor TP53 controls the expression of the HDF protein AKT1. Defining indirect host-virus interactions, we considered AKT1 being indirectly linked to Tat, Nef and gp120. (**B**) Calculating shortest paths from each HIV dependency factor (HDF) protein to a targeted protein, we found that the majority of HDFs are interacting with a protein that the virus attacks. (**C**) Counting the number of interacting viral proteins, we found that the majority of directly targeted host proteins binds to one viral protein. The distribution of indirect interactions as previously defined significantly shifted to higher numbers of interacting viral proteins (Student's t-test, $P<10^{-3}$). In (**D**) we connected human proteins if they significantly co-appeared in pathways and constructed a different network, where linked proteins were significantly targeted by the same viral proteins. Comparing links in these networks, we observed a significant correlation between the number of shared viral proteins and pathways where connected host proteins co-appear in (Pearson's $r = 0.47$, $P<0.01$). Analogously, we constructed a network, linking human proteins that significantly shared indirectly interacting viral proteins where we observed a weaker correlation (inset, $r = 0.26$, $P<0.1$).

doi:10.1371/journal.pone.0011796.g003

As for pathway specific aspects, we connected human host proteins if they significantly co-appeared in pathways utilizing a Fisher exact test ($P<0.01$). In a subsequent step, we constructed another network, connecting human host proteins if they significantly shared directly interacting viral proteins ($P<0.01$). Comparing the number of viral proteins and pathways that are shared by the underlying protein links we observed a strong and significant correlation (Fig. 3D; Pearson's $r = 0.47$, $P<0.01$). Such a result suggests that indeed the combinatorial ways the virus interacts with directly targeted proteins reflect the patterns of involvement in certain pathways. Similarly, we connected HDF proteins that significantly share indirectly interacting viral proteins ($P<0.01$) and found a similar, yet weaker correlation between pathway involvement and indirectly targeting viral proteins (inset, Fig. 3D; $r = 0.26$, $P<0.1$).

According to our definition, indirect interactions with a HDF are facilitated through interacting host proteins that are directly targeted by a viral protein. In Fig. 4A and Table S3, we ranked targeted proteins according to the corresponding number of interactions with a HDF, allowing us to observe that transcription factors and kinases such as TP53, RELA or SRC are enriched at the top of the list. Such an observation suggests that the process of seizing control of the host cell goes through well established interaction paths. Utilizing all transcription factors and kinases that facilitate an indirect interaction of viral proteins and HDFs, we show a network of such indirect and direct interactions in Fig. 4B. Specifically, we observed that Tat interacts with prominent transcription factors, including TP53 and NFKB1, as well as kinases, such as SRC and CDK2, which control important HDF proteins such as EGFR and RELA. These observations suggested that these important functional proteins are not only direct targets of the virus but also might serve as a further gateway to the control of downstream factors such as HDFs.

Discussion

HIV-1 invokes intricate processes with a remarkably low number of proteins to take control of the human host cell. Compensating for its low number of proteins, combinations of pathogen proteins give the virus greater access to a broader set of human proteins. In particular, the subtle structure of the human interactome reveals sites that are not only topologically important, but also are targeted by HIV-1 in both direct and indirect ways. Specifically, rich clubs, protein assemblies that are strongly intertwined among each other, provide proteomic sites that are largely targeted by the virus. Although no direct interaction

A

gene	#	TF	kin	gene	#	TF	kin
TP53	20	■		VAV1	6		
SRC	16		■	CDK7	6		■
RELA	16	■		GTF2F2	6		
MAPK1	14		■	NUP214	6		
PRKCA	14		■	CASP1	5		
CREBBP	13			PRKCZ	5		■
MAPK3	12		■	PRKCD	5		■
CDK2	11		■	PRKCE	5		■
PRKCB1	11		■	KPNB1	5		
JUN	11	■		GTF2H1	5		
PRKACA	10			PCAF	5		
PIK3R1	10			EEF1A1	5		
CSNK2A2	10		■	CDC42	5		
CAV1	10			PRKACB	5		■
CEBPB	9	■		NFKB2	5	■	
NFKB1	9	■		PSMD11	5		
LCK	9		■	UBE2I	5		
LYN	8		■	XPO1	5		
PRKDC	8		■	EGR1	5	■	
XRCC6	8			PRKCQ	4		■
POU2F1	7	■		PSMC5	4		
CASP3	7			YWHAH	4		
CDK9	7		■	FAS	4		
CCNH	7			PSIP1	4		
RAN	7			AP2M1	4		

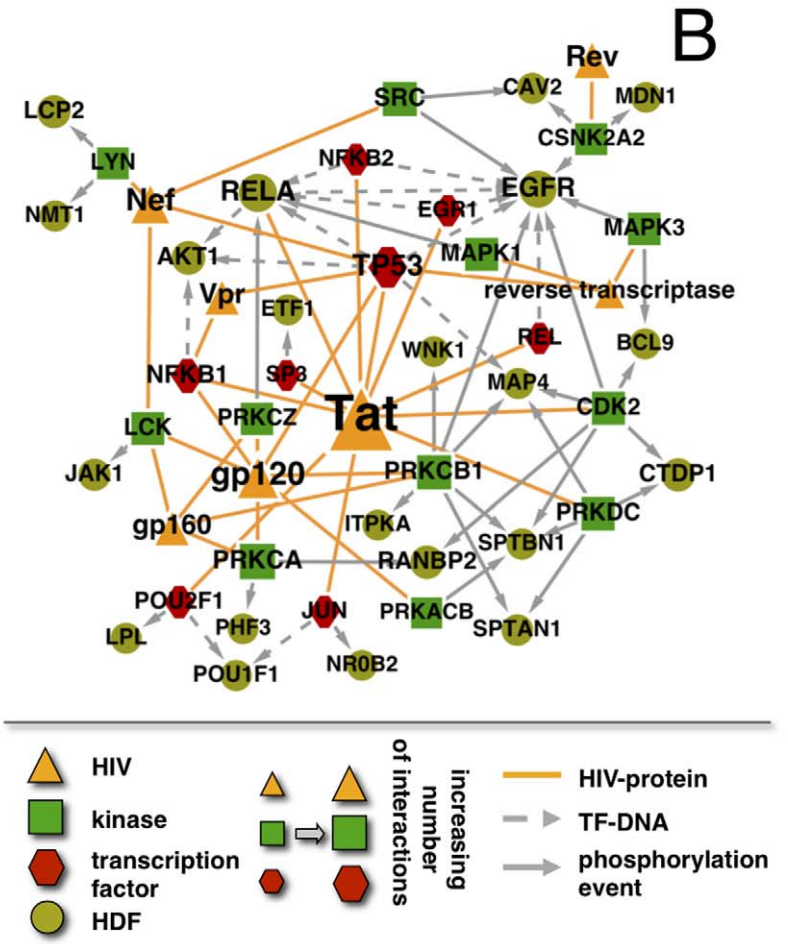

Figure 4. Combinations of viral proteins and map of direct and indirect interactions. In (**A**) we ranked targeted proteins according to the corresponding number of interactions with HIV dependency factor proteins (HDF). Showing the 50 most connected proteins we observed that transcription factors and kinases are enriched at the top of the list. In (**B**), we show a network of transcription factors and kinases that are attacked by viral proteins as well as their interactions with HDFs. Specifically, we observed that Tat largely interacts with prominent transcription factors, such as TP53 and NFKB1 and kinases, such as SRC and CDK2 which control important HDFs such as EGFR and RELA.
doi:10.1371/journal.pone.0011796.g004

targets, HIV dependent factor proteins also provide such a proteomic characteristics that are indirectly exploited by the virus.

Since such proteins are at the intersection of numerous pathways, a large degree of interaction allows the virus to reach into many different functional processes. Subsets of viral proteins reach into the host network to ensure the largest, but focused diversity. Specifically, the use of direct and indirect targets in pathways leads to a cohesive pathway-dependent combination of targeted and HDF proteins. Since we found a strong correlation between the number of shared targeting viral proteins and pathways the underlying host proteins co-appear in, the virus potentially tailored its surface to attack the host cell along well established functional pathways. Recalling that targeted pathways harbor HDF proteins as well, we found a similar yet weaker trend for HDF proteins, suggesting that HDFs may act as downstream mediators of molecular viral information in the underlying pathways.

Utilizing HDFs in an indirect way the virus establishes its control over the host cell, indicating the particular systemic role of proteins that are not directly involved in physical host-pathogen interactions. Such proteins at the interface between the virus and the host are kinases and transcription factors. Such proteins are important mediators of molecular information that allow the virus to effectively utilize them as a gateway to interfere indirectly with a variety of different protein to take control of the human host and ensure the virus' survival. Therefore, untangling the intricate web of indirect and direct interactions is of utmost importance for a thorough understanding of the virus pathogenesis. In the light of these observations, transcription factors and kinases that provide access to proteins in an indirect way seem to be the key players in the subtle molecular strategies a virus employs in order to intercalate a host cell.

Observations that HDF proteins are enriched in rich-clubs, co-appear in many pathways and are largely linked to targeted proteins such as kinases and transcription factors might help to uncover virus dependent factors in systems where information about the interaction interface between a pathogen and a host cell is available. Assuming that the viral take-over of a host cell generally follows similar patterns [18] our results might be tapped to design computational approaches that allow us to predict virus dependable factor proteins in other host pathogen systems. Obviously, the computational prediction of viral dependent proteins offers an efficient and economical way to produce testable hypotheses that can be experimentally investigated further. In addition, the analysis of the entanglement of directly targeted and indirectly interacting proteins may uncover molecular and functional Achilles heels that could be used to systematically hamper viruses [23]. Consequently, defining the web of well defined direct and indirect host-pathogen interactions offers the opportunity to consider viral systems as naturally perturbed biological systems that can be utilized to identify and disentangle relevant pathways in different cellular contexts, ultimately allowing us to eradicate other pathogen driven diseases that plague human kind.

Materials and Methods

Human HIV Protein Interactions

We utilized a compilation of 702 experimentally obtained protein interactions between the human host and HIV-1, accounting for interactions that have been found in vital cells in the human immune system such as helper T cells, macrophages and dendritic cells [24].

Protein Interaction and Pathway Data

Collecting pairwise protein interactions in *H. sapiens* from public databases [12,25,26] we obtained a network of 9,888 proteins embedded in 69,194 physical interactions.

As a reliable source of experimentally confirmed protein-DNA interactions, we used 6,669 interactions between 2,822 transcription factors and structural genes from the TRED database [27]. As for phosphorylation events between kinases and other proteins we found 5,462 interactions between 1,707 human proteins utilizing networKIN [28,29] and phosphoELM database [30]. As a source of reliable human protein pathway information we utilized 913 annotated pathways from the Pathway Interaction Database [22].

Rich-Club Coefficient

The so-called rich-club phenomenon is quantitatively defined by the rich-club coefficient $\Phi(k)$ [20]. Denoting by $E_{\geq k}$ the number of edges among the nodes $N_{\geq k}$ which have at least k interaction partners, the rich-club coefficient is expressed as $\Phi(k)=\frac{2E_{\geq k}}{N_{\geq k}(N_{\geq k}-1)}$, where $\frac{N_{\geq k}(N_{\geq k}-1)}{2}$ represents the maximally possible number of edges among $N_{\geq k}$ nodes. An appropriate choice for normalizing the rich-club coefficient is provided by the ratio $\rho(k)=\Phi(k)/\Phi_r(k)$, where $\Phi_r(k)$ is the rich-club coefficient of a random network with the same degree distribution $P(k)$. In order to have a reasonably large ensemble, we repeated the randomization process 10,000 times. Binning nodes according to their degrees k we obtained a degree dependent mean value of the rich-club coefficient by averaging over all ρ's in each bin. A ratio larger than one, $\rho<1$, is the actual evidence for the presence of a rich-club phenomenon, an increase in the interconnectivity of large degree nodes compared to the random case. This process is well displayed by the presence of an oligarchy of highly interacting nodes that are well connected among each other. A ratio $\rho<1$ points to a lack of interconnectivity among large degree nodes that are separated in distinguishable modules.

Enrichment

Each rich club where each protein has at least k interactions $N_{\geq k}$ is represented as a subset of all proteins N in the underlying network, $N_{\geq k}\subseteq N$. In order to obtain an estimate if proteins with a feature a are overrepresented in a rich-club, we calculated the corresponding fraction $f_{a,\geq k}=N_{a,\geq k}/N_{\geq k}$ in the underlying rich club $N_{\geq k}$. As a null hypothesis, we assumed that the feature a is randomly distributed among human proteins. Determining the randomized fraction of such proteins $f_{r,a,\geq k}$, we defined $E_{a,\geq k}=f_{a,\geq k}/f_{r,a,\geq k}$ as the enrichment of proteins that have feature a in a rich club. Averaging E over 10,000 randomizations rich clubs are enriched with feature a if $E>1$ and *vice versa*.

Pathway Participation Coefficient

For each protein that is part of at least one pathway, we defined the pathway participation coefficient of a protein i, as $P_i=\sum_{s=1}^{N}\left(n_{i,s}\Big/\sum_{s=1}^{N}n_{i,s}\right)^2$ where $n_{i,s}$ is the number of links protein i has to proteins in pathway s out of all N pathways. If a protein predominantly interacts with partners that are members of the same pathway, P tends to *1* while the opposite holds if the interaction partners are distributed among many different pathways.

Significance of Attacked Pathways

Determining the significance of pathways that are enriched with proteins expressed in a human T-cell, we formed a 2×2

contingency table by determining α expressed proteins and the remainder of β proteins in a given pathway. While γ is the number of expressed proteins and δ is the number of remaining proteins in all the other pathways we calculated the probability of obtaining any such set of values randomly by $p* = \frac{\binom{\alpha+\beta}{\alpha}\binom{\gamma+\delta}{\gamma}}{\binom{N}{\alpha+\gamma}}$, where $N=\alpha+\beta+\gamma+\delta$. In order to investigate the two tails of the underlying distribution we constructed all possible contingency tables by keeping the sum of rows and columns constant. The P-value to reject the null hypothesis being the independence of rows and columns in the contingency table is the sum of the probabilities p_i, of all contingency tables i where $p_i \leq p^*, P=\sum_{p_i \leq p*} p_i$ [31].

Significance of Links between Proteins

We applied a hypergeometric distribution to model the probability of obtaining a number of shared features of proteins v and w at or above the observed number by chance. Considering a total of T proteins, we defined the significance that proteins v and w share similar features as

$$P = \sum_{i=|\Gamma(v)\cap\Gamma(w)|}^{\min(|\Gamma(v)|,|\Gamma(w)|)} \frac{\binom{|\Gamma(v)|}{i}\binom{T-|\Gamma(v)|}{|\Gamma(w)|-i}}{\binom{T}{|\Gamma(w)|}}$$

where $G(x)$ represents the feature of protein x.

Kernel Density Function

A simple way to analyze a series of values $x = x_1, \ldots, x_n$ would be a histogram. However, if the number of observations is low the significance of a histogram is rather limited. Therefore, we defined the kernel density approximation, a smoothing operation that allows the estimation of a putative probability density function of data points around a certain point x as $f(x)=n^{-1}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right)$. $K(y)$ is the kernel function, satisfying $\int_{-\infty}^{\infty} K(y)dy=1$, and h is a smoothing parameter. In particular, we chose the Gaussian as kernel function $K(y)=\frac{1}{\sqrt{2\pi}}e^{-y^2/2}$.

Supporting Information

Table S1 List of 851 pathways that are enriched in the human host cell ($P<0.05$).
Found at: doi:10.1371/journal.pone.0011796.s001 (0.49 MB XLS)

Table S2 79 HIV dependant factor proteins (HDF), their attacked proteins they are connected to and targeting viral proteins.
Found at: doi:10.1371/journal.pone.0011796.s002 (0.03 MB XLS)

Table S3 Shows targeted genes that appear in shortest paths to HDFs (#: number of appearances in paths, TF: transcription factors, kin: kinases).
Found at: doi:10.1371/journal.pone.0011796.s003 (0.03 MB XLS)

Author Contributions

Conceived and designed the experiments: SW MTF. Performed the experiments: SW. Analyzed the data: SW GHS MTF. Wrote the paper: SW GHS MTF.

References

1. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, et al. (2006) Comprehensive curation and analysis of global interaction networks in Saccharomyces cerevisiae. J Biol 5: 11.
2. Rain JC, Selig L, De Reuse H, Battaglia V, Reverdy C, et al. (2001) The protein-protein interaction map of Helicobacter pylori. Nature 409: 211–215.
3. Li S, Armstrong CM, Bertin N, Ge H, Milstein S, et al. (2004) A map of the interactome network of the metazoan C. elegans. Science 303: 540–543.
4. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, et al. (2003) A protein interaction map of Drosophila melanogaster. Science 302: 1727–1736.
5. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, et al. (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature 440: 637–643.
6. Gavin AC, Aloy P, Grandi P, Krause R, Boesche M, et al. (2006) Proteome survey reveals modularity of the yeast cell machinery. Nature 440: 631–636.
7. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, et al. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415: 141–147.
8. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–183.
9. Sprinzak E, Altuvia Y, Margalit H (2006) Characterization and prediction of protein-protein interactions within and between complexes. Proc Natl Acad Sci U S A 103: 14718–14723.
10. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol 6: R40.
11. Lehner B, Fraser AG (2004) A first-draft human protein-interaction map. Genome Biol 5: R63.
12. Gandhi TK, Zhong J, Mathivanan S, Karthick L, Chandrika KN, et al. (2006) Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. Nat Genet 38: 285–293.
13. Rhodes DR, Tomlins SA, Varambally S, Mahavisno V, Barrette T, et al. (2005) Probabilistic model of the human protein-protein interaction network. Nat Biotechnol 23: 951–959.
14. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, et al. (2005) A human protein-protein interaction network: a resource for annotating the proteome. Cell 122: 957–968.
15. Uetz P, Dong YA, Zeretzke C, Atzler C, Baiker A, et al. (2006) Herpesviral protein networks and their interaction with the human proteome. Science 311: 239–242.
16. Calderwood MA, Venkatesan K, Xing L, Chase MR, Vazquez A, et al. (2007) Epstein-Barr virus and virus human protein interaction maps. Proc Natl Acad Sci U S A 104: 7606–7611.
17. Bandyopadhyay S, Kelley R, Ideker T (2006) Discovering regulated networks during HIV-1 latency and reactivation. Pac Symp Biocomput. pp 354–366.
18. Dyer MD, Murali TM, Sobral BW (2008) The landscape of human proteins interacting with viruses and other pathogens. PLoS Pathog 4: e32.
19. Brass AL, Dykxhoorn DM, Benita Y, Yan N, Engelman A, et al. (2008) Identification of host proteins required for HIV infection through a functional genomic screen. Science 319: 921–926.
20. Colizza V, Flamini A, Serano M, Vespignani A (2006) Detecting rich-club ordering in complex networks. Nat Phys 2: 110–115.
21. Wuchty S (2007) Evolutionary conservation of motif constituents within the yeast protein interaction network. PLoS One 2: e335.
22. Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, et al. (2009) PID: The Pathway Interaction Database. Nucl Acids Res 37: D674–679.
23. Tan SL, Ganji G, Paeper B, Proll S, Katze MG (2007) Systems biology and the host response to viral infection. Nat Biotechnol 25: 1383–1389.
24. Fu W, Sanders-Beer BE, Katz KS, Maglott DR, Pruitt KD, et al. (2009) Human immunodeficiency virus type 1, human protein interaction database at NCBI. Nucl Acids Res 37: D417–422.
25. Joshi-Tope G, Gillespie M, Vastrik I, D'Eustachio P, Schmidt E, et al. (2005) Reactome: a knowledgebase of biological pathways. Nucleic Acids Res 33: D428–432.
26. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, et al. (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Res 38: D525–531.
27. Jiang C, Xuan Z, Zhao F, Zhang MQ (2007) TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res 35: D137–140.

28. Linding R, Jensen LJ, Ostheimer GJ, van Vugt MA, Jorgensen C, et al. (2007) Systematic discovery of in vivo phosphorylation networks. Cell 129: 1415–1426.
29. Linding R, Jensen LJ, Pasculescu A, Olhovsky M, Colwill K, et al. (2008) NetworKIN: a resource for exploring cellular phosphorylation networks. Nucleic Acids Res 36: D695–699.
30. Diella F, Gould CM, Chica C, Via A, Gibson TJ (2008) Phospho.ELM: a database of phosphorylation sites–update 2008. Nucleic Acids Res 36: D240–244.
31. Francesconi M, Remondini D, Neretti N, Sedivy JM, Cooper LN, et al. (2008) Reconstructing networks of pathways via significance analysis of their intersections. BMC Bioinformatics 9 Suppl 4: S9.