The symphony of gene regulation

van Dijk, D.

Publication date
2013

Chapter 3 Inference of Surface Membrane Factors of HIV-1 Infection through Functional Interaction Networks

S. Jaeger[1,3], G. Ertaylan[2], David van Dijk[2], Ulf Leser[1], Peter MA Sloot[2]

[1] Knowledge Management in Bioinformatics, Humboldt-Universität Berlin, Berlin, Germany

[2] Computational Science, University of Amsterdam, Amsterdam, The Netherlands

[3] Algorithmic Computational Biology, Centrum Wiskunde & Informatica, Amsterdam, The Netherlands

PLoS ONE 2010, **5**:10. doi:10.1371/journal.pone.0013139

3.1 Summary

3.1.1 Background

HIV infection affects the populations of T helper cells, dendritic cells and macrophages. Besides, it has a serious impact on the central nervous system. It is yet not clear whether this list is complete and why specifically those cell types are affected. To address this question, we have developed a method to identify cellular surface proteins that permit, mediate or enhance HIV infection in different cell/tissue types in HIV-infected individuals. Receptors associated with HIV infection share common functions and domains, and are involved in similar cellular processes. These properties are exploited by graph theory and a novel gene-ranking algorithm to predict unprecedented surface membrane proteins (SMP) potentially interacting with HIV.

3.1.2 Principal Findings

We compiled a set of SMPs that are known to interact with HIV from the HIV-1 protein interaction network. This set is extended by proteins that have direct interaction and share functional similarity. This resulted in a comprehensive network around the initial SMP set. Using network centrality analysis we predict novel surface membrane factors from the annotated network. We identify 21 surface membrane factors, among which three have confirmed functions in HIV infection, seven have been identified by at least two other studies, and 11 are novel predictions and thus excellent targets for experimental investigation.

3.1.3 Conclusions

Determining to what extent HIV can interact with human SMPs is an important step towards understanding patient specific disease progression. Using graph theory, GeneOntology and a gene-ranking algorithm we generate a set of surface membrane factors that constitutes a well-founded starting point for experimental testing of cell/tissue susceptibility of different HIV strains as well as for cohort studies evaluating patient specific disease progression. In conclusion, our findings constitute the necessary background for future research investigating the role of SMPs during infection with HIV.

3.2 Introduction

One of the important characteristics of Human Immunodeficiency Virus (HIV) is its ability to interact with many cell types and its capacity to alter the function of chemokines that otherwise work in harmony with the immune system. This interaction depends on the phenotype of the virus, the receptor type residing on the cell as well as the chemokines present in the environment. The main factor determining its complex interaction profile is HIV's highly interactive proteome. Structurally, its genome has evolved to interact with many human proteins from

various cellular pathways, as was investigated in Chapter 2. Therefore, each infectious virion consists of viral proteins, such as Tat, Gp120 or Nef, which interact with proteins inside and outside the cell [1-3].

Another contributor to this complex behavior is the high degree of phenotypic variation in the HIV population in-vivo [4]. Interestingly, each transmission event (between individuals) introduces an evolutionary bottleneck since the majority of new infections are usually initiated with a single virus [5].

Typically, HIV infection is thought to originate from the contact of genital epithelia with the infectious virions. It has been suggested that Langerhans cells and resident dendritic cells of stratified squamous epithelia serve as the initial targets of HIV infection [6, 7]. Virions are mobilized to the lymph nodes either via attachment of the HIV Gp120 to the DC-SIGN receptor expressed on dendritic cells (DCs) [6] or by direct infection of DCs within epithelia via CD4 and CCR5 receptors [7]. In the lymph nodes virions are transferred to CD4+ T cells and macrophages.

Moreover, soluble Gp120 binds to Immunoglobulin-E on innate immune system cells, such as basophils, mast cells and monocytes, and induces the secretion of cytokines thereby causing further activation of type-2 T-helper cells (Th2), the primary targets of HIV-1 infection [8]. The system-wide activation of CD4+ T cells results in an increased number of infected cells and high viral reproduction that leads to viral peaks observed in the primary stages of the infection. This translates into virus populations, which essentially are genotypically related cloud(s) of phenotypes (or quasispecies). The infection, which has been ignited with a relatively small number of virions, then spreads to other tissue types harbouring immune system cells, such as CD4+CCR5+CCR3+ microglia and macrophages [9], or hMR+ astrocytes [10], megakaryocytes [11] and monocytes [12].

A puzzling fact is that the cell types that are targets of HIV infection have different receptor expression profiles and do not necessarily harbor main co-receptors CCR5 or CXCR4. For instance, in a clinical study with a heterozygote CCR5-Δ32 (CCR5 delta 32) individual (which gives partial resistance to infection via CCR5 tropic viruses) a wide range of co-receptor usage is observed, suggesting the involvement of other surface membrane factors [13].

Furthermore, binding of HIV to cell surface factors other than CD4 and chemokine receptors does not always permit viral entry but leads to endocytosis of the viral particles. This promotes relocation of the infectious virions, future trans-infection of adjacent cells [14] and leads to the activation of the immune system. Therefore, it is imperative to bear in mind that there are surface membrane factors interacting with HIV proteins, hence affecting the course of infection indirectly.

Another important point regarding surface membrane proteins is that their interactions with HIV-1 proteins are not only restricted to the extracellular environment. Events taking place inside and outside the cell membrane are neither

decoupled processes, nor mutually exclusive. In vitro studies with HIV-1 protein Tat have shown that Tat is able to induce the intrinsic pathway of apoptosis in a number of human cell lines in addition to up-regulating the expression of co-receptor CCR5 and the interleukin-2 (IL-2) in HIV-1-infected cells. Extracellular Tat has also been shown to induce neuronal death by binding to the lipoprotein receptor-related protein (LRP) (see Romani et al. [15] for an extended review).

Although many steps of the virus life cycle have been unraveled and 24 distinct drugs targeted against HIV have been approved, all efforts to achieve an overall eradication of the virus have turned out to be ineffective [16]. However, life expectancy under highly active antiretroviral therapy (HAART) treatment has been extended to 21.5 years [17].

3.2.1 The missing piece of the puzzle

These observations lead to the following questions: What is the extent of surface membrane factors contributing to HIV-1 infection and how do they influence the outcome of the treatment?

HIV exploits the existing signaling and regulatory pathways in its host. The different receptors or surface membrane proteins that are targeted in different cell types are likely to be involved in the same (or closely related) functional pathways, because the range of processes and pathways available to the virus is limited. The complexity in finding the right factors arises from the fact that there are hundreds of surface membrane proteins expressed on a wide variety of cells.

Experimental testing of hundreds of targets from numerous pathways is not feasible. Therefore, we developed a computational approach that generates high quality hypotheses for wet-lab experiments with the aim to identify surface membrane host factors contributing to HIV-1 disease outcome. We adapt a strategy from disease gene discovery that is based on protein interaction, network centrality and functional similarity to receptors that are known to interact with HIV. We infer promising candidates using measures of centrality in the emerging network of proteins. This method reproduces reported factors, such as CCR1, CCBP2 and CD97, but also results in a list of proteins that likely affect the progression of the infection.

3.3 Materials and Methods

We designed a method to identify uncharacterized surface membrane factors interacting with HIV. We employ a ranking strategy based on network centrality that uses documented HIV receptors, human protein interaction data and protein functions. The algorithm is partially adapted from disease gene identification strategies that infer gene-disease associations from similarity networks and their properties. Its underlying principle is based on the assumption that the most central genes or proteins in a specific disease network are likely to be related to the disease [18, 19].

3.3.1 Conceptual design

For identifying novel surface membrane factors we developed a generic framework that infers candidate genes or proteins based on their similarity to a set of reported genes or gene products of interest. The general workflow of this framework, illustrated in Figure 3.1, comprises three steps. First, a seed set is defined by genes/proteins that share specific characteristics of interest that will be later used for growing a functional interaction network. This can be a set of proteins associated with a certain disease, involved in specific pathways, sharing other biological properties or transcripts that are differentially expressed in a condition of interest. In the second step, candidate proteins are extracted based on their functional similarity to the seed set and a domain-specific similarity network is generated by extending this set by all functionally related proteins. The notion of similarity is not necessarily restricted to functional annotation or interaction data but rather can cover any kind of genomic data, such as expression data, SNPs, sequences and phenotypes. Finally, in the last step network centrality analysis is performed to rank those proteins with respect to their relative importance within the network. The most central ones are presumed to be of functional importance for the specific network. Note that for simplicity we only referred to proteins in the description of the framework. However, our method is not restricted to proteins but is also applicable to genes depending on the biological question.

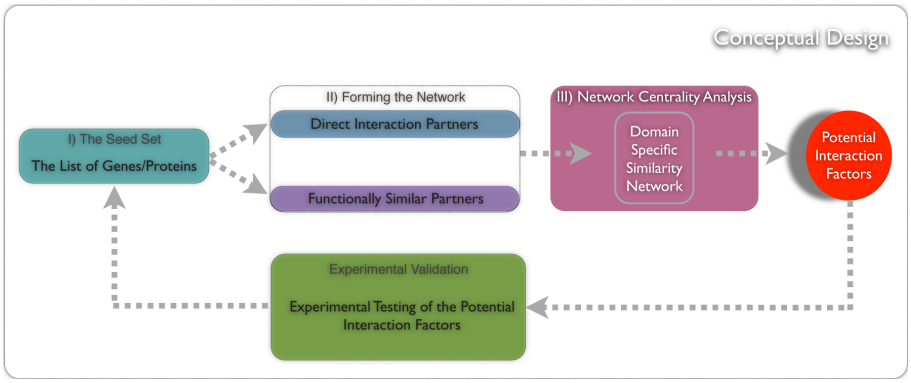

Figure 3.1: Conceptual design of the proposed prediction framework. The method consists of three components. I) Compiling "the seed set" from genes/proteins sharing specific characteristic of interest; II) Forming a network by including direct interaction partners described in database(s) and/or functionally similar partners; III) Network centrality analysis on the domain-specific similarity network to obtain potential interaction factors (PIFs). The final step of such an analysis is the experimental validation of PIFs. Confirmed PIFs then can be included in the seed set and the steps I-III can be repeated for identifying new PIFs.

Translating the general framework into the context of identifying surface membrane factors interacting with HIV-1 implies that proteins, which are related to known HIV receptors through functional similarity or interaction with the same ligand(s), tend to be part of the same pathway and often share the same biological function. Therefore, if a network is built based on documented surface membrane factors that are extended with related genes, yet undiscovered surface proteins should also be central in the resulting network. To study this, we build an enriched HIV receptor network from known HIV receptors and rank all its proteins according to their centrality within the network. Highly ranked proteins are further analyzed to identify potentially novel surface membrane factors.

Below we explain the details of our method for identifying surface membrane factors interacting with HIV-1. One should keep in mind that the framework is neither domain nor disease specific and can be applied for various biological questions other than the one presented in this study.

3.3.2 Data

We use a set of known HIV receptors, their functional annotations and human protein interaction data as a scaffold for building an HIV receptor network. The initial list is compiled by mining the literature and the 'HIV-1, Human Protein Interaction Database' [20]. A receptor is included if it is reported by at least two independent studies. This applies to 16 HIV receptors. However, three of them,

namely Rdc1, Gpr15 and ChemR23, are not documented in the data gathered from protein interaction databases (see below) and thus have not been used in this study. Table 3.1 shows the final list of 13 HIV receptors including protein domain information (InterPro), literature references and their role in HIV infection. The list covers established receptors such as CD4 and DC-SIGN, HIV co-receptors CCR5 and CXCR4 as well as alternative co-receptors CCR2 and CCR3. Only recently reported co-receptors, such as XCR1 [21], have not yet been included since they were not documented by the time the study was conducted. However, we use a list of cell surface proteins that are reported to interact with HIV in a broad sense. Therefore, we do not limit our prediction method to receptors that only permit the entry of HIV into the primary cells.

Table 3.1: Initial set of HIV seed receptors. List of seed HIV receptors, including the receptor type and their functional domains. Receptors are grouped according to their functional domains (see Figure 3.5(a) for the distribution of those domains). A full table including the complete list of references that indicate the association to HIV is provided in Table 3.7.

Receptor	Receptor type	InterPro domains
Ig-like and Other		
CD4	Primary receptor for HIV	Ag_CD4, CD4-extracel, Ig-like, Ig-like_fold, Ig_C2-set, Ig_sub, Ig_V-set_sub
7-TM GPCR and CCR_rcpt		
CCR5	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_5_rcpt
CCR3	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_3_rcpt
CCR2	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_2_rcpt, CC_5_rcpt
CCR8	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_8_rcpt
CCR9	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_9_rcpt
CXCR4	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CXC_4_rcpt
CXCR6	Co-receptor	7TM_GPCR_Rhodpsn, CXC_6_rcpt
CX3CR1	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CX3C_fract_rcpt
7-TM GPCR and Other		
APJ	Alternative co-receptor	7TM_GPCR_Rhodpsn, APJ_rcpt
GPR1	Alternative co-receptor	7TM_GPCR_Rhodpsn, GPR1_rcpt
Integrin-α		
ITGA4	Co-receptor with CD4	Int_alpha_beta-p, Integrin_alpha, Integrin_alpha-2, Integrin_alpha_C
C-type lectin and Other		
DC-SIGN	Receptor for HIV	AntifreezeII, C-type_lectin

Human protein interactions were obtained from the major public protein-protein interaction databases: DIP [22], IntAct [23], BIND [24], Mammalian MIPS [25], HPRD [26], MINT [27] and BioGRID [28]. From each database we retrieved the complete set of available human protein interactions. Table 3.3 provides the number of protein interactions obtained from each database by the time of this study. We integrated the different data sets by mapping the interacting proteins to unique protein identifiers from UniProt [29] or EntrezGene [30] and thus generating one comprehensive protein interaction map for our study. The

integrated protein interaction set comprises 13,494 human proteins and 43,637 unique interactions observed between these proteins. Each protein included in the interaction map is associated with its respective protein domain information [31] and functional Gene Ontology (GO) annotations [32] (also retrieved from UniProt and EntrezGene).

3.3.2.1 HIV receptor network

We generate a specific HIV receptor network using known receptors as seeds (see Table 3.1). We map each seed gene to its protein(s) thus growing a network around them [33]. The network is extended by adding proteins that either directly interact with any seed or that are functionally similar to at least one seed. Functional similarity between two proteins is determined by using a semantic similarity measure proposed by Couto et al. [34]. The formal definition of functional similarity is provided in the Supporting Information at the end of this chapter. In principle, proteins are considered as functionally similar if their semantic similarity to a seed protein is above the threshold of 0.7 (averaged across the three GO subontologies: molecular function, biological process and cellular component). Thereby, we only consider close and significant biological relationships.

Functionally related proteins are integrated into the network through weighted edges to the seeds. Edge weights are assigned by combining a protein interaction and a GO score. The protein interaction score is either 1 if an interaction is documented between a protein and a seed, and 0 otherwise. The GO score ranges between 0 and 1 (see Supporting Information at the end of the chapter) depending on the similarity of the GO annotations between two proteins, whereby 1 indicates functional equality and 0 indicates maximal functional distance. Interactions and functional similarities among all non-seed proteins are also included into the network.

We exploit protein interaction because it strengthens the relationship between (similar) receptors interacting with the same ligand. Human interaction data, however, is still incomplete and will not cover the functional space for our analysis. Therefore, we also integrate functional data to capture cellular surface proteins that show significant functional similarity with the seed receptors. Nevertheless, the functional coverage is still limited and currently only a fraction of the genome is annotated with pathways, functions and phenotypes [19]. Hence, we integrate predicted functions in our framework to functionally enrich proteins that are weakly or not annotated at all.

3.3.3 Functional enrichment

To functionally enrich the HIV network we apply a network-based function prediction method to derive additional annotations. This method compares protein interaction networks across multiple species to detect evolutionarily and functionally conserved subgraphs. This involves the identification of orthologous proteins (using OrthoMCL [35]) and the detection and assembly of conserved interactions. Within each conserved subgraph we infer novel protein functions from

orthology relationships across species and along conserved interactions of neighboring proteins within a species (Jaeger et al. submitted, see [36] for early work). Predicted functions are added to the set of confirmed functions to better characterize proteins that are weakly or not annotated at all. The functional enrichment increases the final cross-validation recovery rate up to 30%.

3.3.4 HIV network centrality analysis

Network centrality analysis is particularly useful for identifying key elements in different biological processes. In general, networks are modeled as mathematical objects called graphs. A graph is an abstract presentation of a set of objects that are connected by links. In the most common sense a graph G = (V;E) consists of a finite set of vertices V and edges E whereas an edge e = (u; v) connects two vertices u and v. Centrality, on the other hand, is formally defined as a function C that determines a numerical value C(v) for every vertex v in a graph that describes its location relative to the other verteces. We are interested in the ranking of vertices of the given graph G, thus we follow the convention that a vertex u is more important than another vertex v if and only if C(u) > C(v) [37].

Different centrality measures have been proposed for analyzing various types of biological networks [37]. Established measures are degree centrality, closeness centrality, betweenness centrality and PageRank centrality.

Here, we chose PageRank [38] to identify the most important factors within the HIV receptor network since the PageRank algorithm assigns numerical scores to each node to determine its relative importance within the network based on the assumption that not all relationships are equally important for determining the centrality of a node. Thus, links to high-scoring nodes contribute more to the PageRank centrality of a node than links to low-scoring nodes.

We used the PageRank centrality measure to discover novel surface membrane factors that are involved in HIV-1 infection. Accordingly, we rank all proteins with respect to their PageRank centrality within the network using the igraph library in R [39]. Clearly, we expect the seed receptors to be highly ranked in the ordered list, since our construction algorithm naturally places them in a central position.

Nevertheless, not all seed receptors are central, and many non-seed proteins are ranked high. We are especially interested in the latter since these are promising candidates for novel surface membrane factors. An appropriate ranking is essential for deciding which factors should be investigated further, e.g. in follow-up experiments.

3.3.5 Validation

We validate our method and the results as follows: First, we use leave-one-out cross-validation to assess the predictive power for finding novel surface membrane HIV factors. Second, we determine the statistical significance of our results by comparing them to a random control set. For cross-validation we remove one seed

receptor from the initial list and try to re-discover this receptor using our method. We build an HIV receptor network from the remaining receptors and rank the proteins according to their centrality within the network. Subsequently, we determine whether the left-out receptor is re-discovered and at which position of the ranked list. We repeat this procedure for each seed receptor and determine the average recovery rate across all receptors.

To determine the statistical significance of the results, we compare them to two random control sets. The first set, Set1, comprises all proteins from the human interaction network as candidates resulting in 13,494 proteins. The second set, Set2, is stricter and contains only proteins with receptor properties, simulating a more informed manual search. To generate this set we use specific GO annotations that imply a receptor activity since there is no general receptor definition indicating whether a protein is a receptor or not. Thus, Set2 is formed by filtering proteins from the interaction data that are annotated with at least one of these specific GO terms. This results in 2,512 candidates – covering 12 out of 13 seed receptors (ITGA4 is missing due to insufficient functional annotation). We randomly draw m samples from each control set, where m corresponds to the average number of proteins within the HIV network and determine whether the known receptors are among the samples. This is repeated 1,000 times and an average recovery rate is calculated which is later compared to the recovery rate from our ranking method.

3.4 Results

We have designed a framework for discovering novel surface membrane factors interacting with HIV-1. To this end, we use protein interaction, protein function, and network centrality analysis to determine yet uncharacterized surface membrane proteins based on their functional similarity and topological closeness to receptors that are known to interact with HIV.

Our strategy is based on the assumption that proteins, which are related to known HIV receptors through functional similarity or direct interaction with the same ligands(s), tend to be part of the same pathway and often share the same biological function. Therefore, an enriched HIV receptor network is built from documented surface membrane factors by populating it with functionally related proteins that either interact directly with or show significant functional similarity to any known factor. Subsequently, all proteins are ranked according to their centrality within the network. The underlying principle of the centrality analysis presumes that the most central proteins in a domain-specific network are likely to be of high functional relevance [40]. Thus, yet undiscovered but prospective surface proteins should also be central in the network. Highly ranked proteins are analyzed further to identify potentially novel surface membrane factors. The key steps of our inference method are illustrated in Figure 3.2.

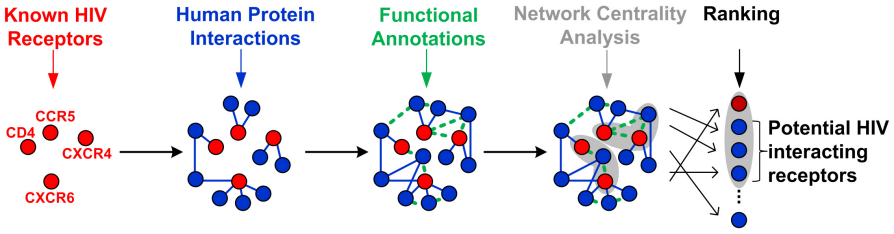

Figure 3.2: Illustration of the key steps in the prediction method. Starting from the seed HIV receptors we add proteins that 1) have direct interaction (blue solid edges) or 2) are functionally similar (green dashed edges) to the known receptors to generate an enriched HIV receptor network. Proteins are ranked according to their centrality within the receptor network. Proteins in shaded areas represent highly central proteins.

In the following subsections, we first evaluate the performance of the prediction method. Subsequently, we investigate the most promising predictions by exploring literature on their functional domains, expression levels and reported clinical evidence.

3.4.1 Cross-validation

Cross-validation is performed on 13 known HIV receptors to evaluate the predictive power of the method. Overall, we achieved a re-discovery rate of 92% (12 out of 13). ITGA4 was not re-discovered by our method, due to its insufficient annotation and low functional similarity to the other 12 receptors.

We studied the recovery rates using interaction data and GO annotation with and without functional enrichment. The comparison shows that the total number of re-discovered receptors is significantly higher when functionally enriched data is employed. Consequently, interaction data in combination with enriched functional annotation are chosen for further analysis.

The same evaluation was performed using random control sets, Set1 and Set2. The random recovery rates are compared to the network-driven recovery rate to assess the statistical significance. We determine the fraction of seed receptors that can be discovered when randomly sampling from the complete protein set (Set1) and a subset including only surface membrane proteins (Set2) (see Methods). On average, we discover 0.69 and 3.4 of the 13 seed receptors when sampling from Set1 and Set2, respectively, which results in random recovery rates of 5.3% and 26.2%. The comparison of recovery rates shows that the network driven recovery rate of 92% is clearly superior to the random recovery rates of 5.3% and 26.2%. The t-test confirms that the observed superiority over the control sets is statistically highly significant (p-value <2.2e-16) and thus underlines the advantage of our network-driven strategy over the random approach.

For the prediction of novel surface membrane factors, we investigate the trade-off between discovering potential candidates vs. false positives by normalizing the recovery rate by the number of proteins considered at each rank. Figure 3.3 compares original and normalized recovery rates across the prioritized protein list. The receptor-per-protein ratio is used to assess the probability to identify new HIV interacting surface proteins. The most significant discovery ratio is 29% (2/7) considering the top 1% proteins. The second best discovery ratio is achieved at 3%, where the probability of rediscovering a known surface membrane factor is 24% (5/21). Note that the probabilities are estimated from the cross-validation on known data and therefore provide lower bounds since all novel findings are counted as false positives.

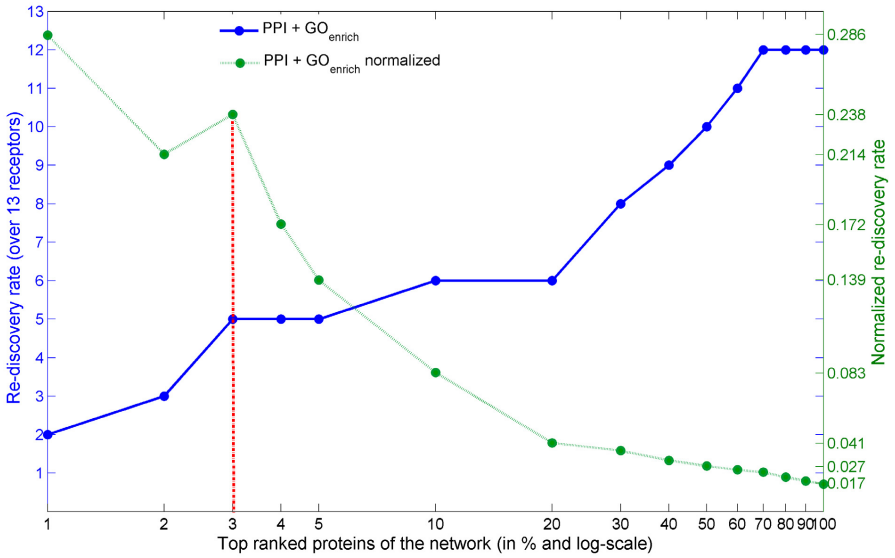

Figure 3.3: Results of the leave-one-out cross-validation over the 13 seed receptors. The average receptor re-discovery rate is determined for the different ranks (in %) of the HIV network. The original re-discovery rate (left y-axis and solid line) is compared to the normalized re-discovery rate (right y-axis and dashed line). The red dashed line indicates the chosen cut-off. The x-axis is in log-scale to focus on the highest ranks (1% to 5%).

We used these receptor-per-protein ratios to define a cut-off to select candidates from the prioritized list. We choose 3% as threshold, since it presents a sensible trade-off between potential candidates and false positives (see above) while yielding a reasonable number of novel candidates. Thus, the top 21 proteins in the ranked list are considered as surface membrane factor candidates.

3.4.2 Predicting novel HIV surface membrane factors

Finally, we consider all 13 known HIV receptors as seeds to build an HIV receptor network with 739 proteins (726 candidates) and 80,000 functional relationships (note that during cross-validation we always removed one of the seeds). We ran the PageRank algorithm and obtained a list of centrality ranked proteins. Seed receptors are removed from the list since they are (by definition) highly ranked. We apply the chosen threshold and consider the first 21 proteins as host factor candidates. Table 3.2 presents the top-ranked candidates including their InterPro domains and cell types.

Receptor	Receptor-specific InterPro domains	Cell types	Association with HIV
7-TM GPCR and Other			
HTR6	Not applicable	Uniform expression[1]	+
HTR1B	5HT1B_rcpt	Uniform expression[1]	?
HTR1E	5HT1F_rcpt	Uniform expression[1]	?
RXFP2	LDL_rcpt_classA_cys-rich_rpt, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcpt	Low expression	?
RXFP1	LDL_rcpt_classA_cys-rich, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcpt	No expression profiles available	?
GPR17	P2_purnocptor	Uniform expression[1]	?
GPR182	G10D_rcpt	Uniform expression[1]	?
NPBWR2	Neuropept_W_rcpt	Uniform expression[1]	−
7-TM GPCR and CCR_rcpt			
CCR1	CC_1_rcpt	High expression: whole blood, monocytes, myeloid, dendritic cell	+
CCBP2	CXC_4_rcpt	Uniform expression[1]	+/−
7-TM GPCR			
DARC	Duffy_cmk_rcpt	High expression: (early) erythroid, endothelial cells	+
Ig-like and Other			
CD2	Ag_CD2, Ig-like_fold, Ig_C2-set, Ig_V-set, T-cell_sdhesion_molc_CD2	High expression: dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood	+
CSF3R	FN_III, Hematopoietin_rcpt_gp130_CS, IgC2-like_lig-bd	High expression: myeloid cells, monocytes and whole blood	+
IL1R1	Ig, Ig-like_fold, Ig_sub, IL1_rcpt_1, IL1R_rcpt	No expression profile available	−
CD79B	Ig-like_fold, Ig_sub, Ig_V-set, Phos_immunorcpt_sig_ITAM	High expression: CD34, endothelial and dendritic cells	+
IL6ST	FN_III, Hematopoietin_rcpt_gp130_CS, Ig-like_fold, IgC2-like_lig-bd	Uniform expression[1]	+
TNFR_Cys_rich_reg and Other			
TNFRSF5	Fas_rcpt	High expression: B lymphoblasts	+
TNFRSF3	TNFR_3_LTBR	High expression: myeloid, monocytes and whole blood	+
Other			
CD97	EGF-type_Asp/Asn_hydroxyl_site, EGF_Ca_bd_2, GPCR_2_CD97, GPCR_2_secretin-like, GPS_dom	High expression: CD34, B lymphoblast, dendritic cells, CD8 and CD4 T-cells, NK, myeloid, monocytes	+
GP1BB	LRR-contain_N, Cys-rich_flank_reg_C	High expression: CD34, monocytes and whole blood	?
GYPB	Glycophorin	High expression: (early) erythroid and endothelial cells	?

Table 3.2: List of inferred surface membrane factors: List of the potential surface membrane proteins that result from our method, including functional domains and cell types. Predictions that are associated with HIV in earlier studies are marked with '+'. '-' indicates predictions with negative evidence. For predictions without literature on interaction the association remains unclear (shown by '?').

[1]Uniform expression in CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8and CD4 T cells, and whole blood.

Figure 3.4 shows the subnetwork from the full HIV receptor network that exhibits only the direct functional relationships between seed receptors and predicted surface membrane factors. The analysis of the known and predicted surface membrane factors regarding their annotated KEGG pathways [41] revealed the involvement of three pathways, namely the chemokine signaling pathway (hsa04062), the hematopoietic cell lineage (hsa04640) and the intestinal immune network for IgA production (hsa04672).

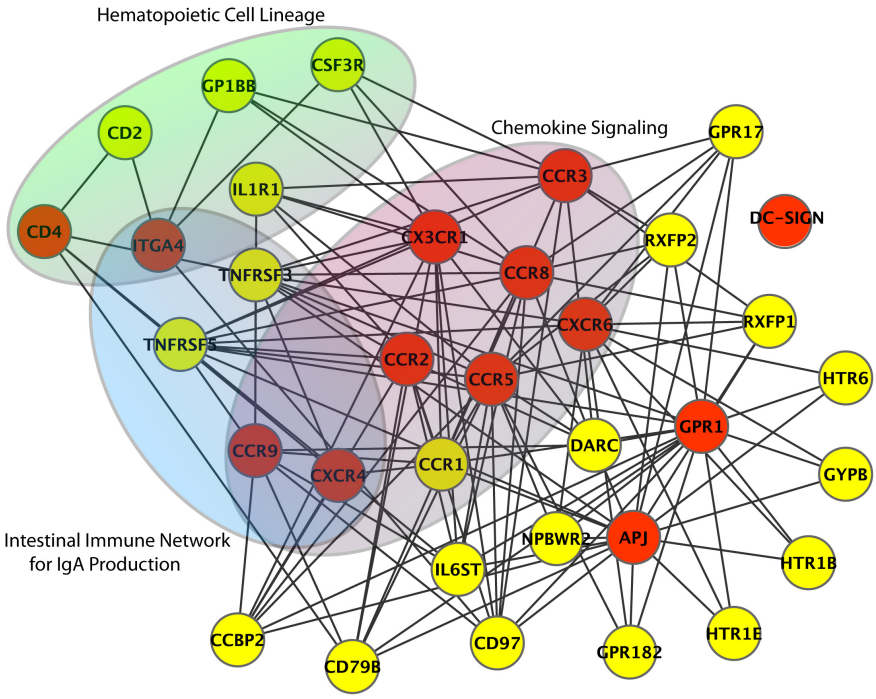

Figure 3.4: Sub-network of the generated HIV receptor network. The sub-network focuses on the functional relationships between the seed receptors (red) and the predicted surface membrane proteins (yellow) within the HIV receptor network. Non-seeds and non-candidate proteins are not shown. Significantly enriched pathways within this sub-network are additionally highlighted.

3.4.3 Support for predictions

We assess the relevancy of the candidates using evidence that supports an association with HIV. We investigate the predictions with respect to functional domains, cell types, expression levels, associated SNPs and chromosomal locations.

Receptor domains: We analyze our predictions by comparing their functional protein domains to the domains of the known seed receptors assuming that overlapping functional domains indicate similar protein properties, e.g. binding the same ligand, and functional similarity [42]. Common protein domains of the seed receptors are:

- G-protein-coupled receptors (GPCR) rhodopsin-like superfamily and 7 transmembrane (7-TM) GPCR rhodopsin-like domains (7-TM GPCR)

- Chemokine receptor domains (CCR rcpt)
- Immunoglobulin and related domains (Ig-like)
- C-type lectin and related domains (C-type lectin like)
- Integrin alpha and related domains (Integrin alpha)

The distribution of the domains among the seed receptors is shown in Figure 3.5(a).

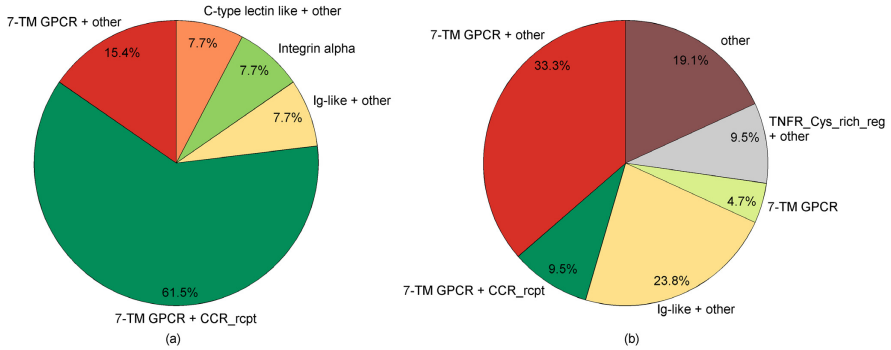

Figure 3.5: Overview on the distribution of protein domains. Distribution of the protein domains for (a) seed receptors and (b) predicted surface membrane factors.

Predicted surface membrane factors are grouped according to their functional domains (see Table 3.2) which results in GPCR with chemokine domains, GPCR without chemokine domains, Ig-like receptors and receptors without any overlapping domains. The respective domain distribution is displayed in Figure 3.5(b). The largest domain overlap is found for 7-TM GPCR rhodopsin-like domains. Half of the predictions have this particular domain, which is also overrepresented in the set of seed receptors (10 of 13, see Figure 3.5(a)). In addition, CCR1 and CCBP2 share a chemokine domain, which is very frequent in the set of initial receptors (8 of 13). Moreover, five predicted surface membrane factors have Ig-like domains that match the primary HIV receptor CD4.

The amount of overlapping functional domains indicates that the functional characteristics of the initial HIV binding receptors are reflected in predicted surface factors. In particular, GPCRs have a broad usage spectrum as co-receptors by primary isolates of HIV [21] and specifically chemokine receptors are known as co-receptors for HIV [43]. Strikingly, CCR1 and CCBP2 share both 7-TM GPCR rhodopsinlike and chemokine domains and are reported as co-receptors of HIV. However, receptors without any overlapping domains might present

unprecedented characteristics that are not documented in the initial set but are reflected in their complementary domain diversity (see Figure 3.5(b)).

Chromosomal locations: Genes with similar properties are sometimes located in the same regions of the human genome. Thus, the genomic location of a gene is often taken into account when new candidate genes are associated with a disease. The reason is that mapping those candidates to a region containing other genes associated to the same disease further supports the association. For example, HIV binding human CC chemokine receptor genes are known to cluster within the 3p21.3 region of the genome [44].

We determine the chromosomal location of the predicted surface proteins and study whether they cluster together with other candidates or known seed factors. The chromosomal location for each seed and prediction retrieved from EntrezGene is shown in Table 3.6: Chromosomal location of known and predicted surface membrane proteins. Similar chromosome regions are colored similarly.. When considering the known receptors there is a group of six chemokine receptors that map to the CCR cluster within 3p21.3, and also two receptors, CCR1 and CCBP2, from the predicted set are associated to this region. However, the remaining ones are located on different chromosomes and do not map together. Only CD97 and DC-SIGN, and GPR17 and CXCR4 are mapped together to 19p13 and 2q21, respectively.

3.5 Discussion

The involvement of co-receptors and surface membrane proteins assisting HIV-1 infection and contributing to viral pathogenesis always has been underestimated [21]. Only a limited number of studies aim to elucidate the role of surface membrane factors interacting with viral proteins even though they are potential amenable drug targets for HIV therapeutics [45, 46].

We predict 21 surface membrane HIV factors that are potentially involved in the different stages of infection influencing the progression of the disease. Remarkably, among these cell surface proteins, three have confirmed functions in HIV infection, seven have been reported by at least two other studies and eleven predictions are novel findings that deserve experimental investigation. It is important to note that the high success rate of our method, as shown using cross-validation, strongly implicates that our predictions can be the missing piece of the puzzle.

3.5.1 Experimentally confirmed predictions

CCR1: The C-C chemokine receptor type 1 is a GPCR that mediates signal transduction and the recruitment of effector immune cells to inflammation sites. It is highly expressed in immune system cells, such as myeloids, monocytes, dendritic cells and whole blood. Independent studies confirmed the usage of CCR1 along with CD4 for the entry of HIV into target cells [21, 47].

CCBP2: The Chemokine-binding protein 2 is another chemokine receptor that is documented to function as alternative co-receptor for HIV [48].

DARC: The Duffy antigen/chemokine receptor belongs to the family of erythrocyte chemokine receptors that bind chemokines. It is highly expressed on red blood cells (RBCs). Several studies demonstrated the binding of HIV-1 to RBCs through DARC enabling RBCs to transmit HIV to peripheral blood mononuclear cells. However, binding HIV to DARC does not permit viral entry but retains the virus viability and mediates trans-infection of HIV-1 from RBCs to susceptible T cells [49, 50]. Recently, He et al. reported that the DARC -46C/C genotype is associated with an increase of 40% in the odds of acquiring HIV-1 in African Americans [50]. However, follow-up studies on different cohorts [51-53] or with correction for population stratification [54] could not establish a significant association of this DARC polymorphism and the increased risk for HIV-1 acquisition or disease progression. Although DARC's association with HIV has been established some questions remain regarding its influence on HIV-1 acquisition and progression.

3.5.2 Prediction with direct and indirect experimental support

CD97: It belongs to the EGF-TM7 family of class II 7-TM molecules and is present on the surface of most activated leukocytes. It is broadly expressed on most hematopetic cells, activated lymphocytes, macrophages, dendritic cells, granulocytes, monocytes and undergoes a rapid up-regulation during T and B cells activation. Recently, CD97 was identified in a large-scale genome RNAi screening as one of six uncharacterized host factors that are required for HIV replication [45] suggesting its crucial postintegration role. Furthermore, Kop et al. [55] showed that CD97 is present on the surface of all human lymphocytes in blood and lymphoid tissue and confirm its up-regulation upon cellular activation. In addition, they demonstrated significant differences in the expression levels between lymphocytes. For instance, T and NK cells possess higher levels of CD97 than B cells and memory CD4+ (but not CD8+) T cells express more CD97 than naive cells. These differences might present the missing factor that is required for active infection of naive T cells in early infection because CD97 is highly expressed inactivated memory CD4+ cells but not in naive subsets. To confirm this hypothesis longitudinal testing of in-vivo expression of CD97 is necessary in patients going through co-receptor switch.

CSF3R: The granulocyte colony-stimulating factor receptor is the receptor for colony stimulating factor 3 (G-CSF), a cytokine that controls the production, differentiation, and function of granulocytes. CSF3R is highly expressed on monocytes and activated T cells [56]. Its ligand modulates cytokine production in monocytes and lymphocytes. In particular, CSF3R is thought to play a role after viral DNA synthesis. The indirect influence on infection and replication in human cells has been demonstrated through the binding of recombinant G-CSF (rG-CSF). rG-CSF is able to activate replication of HIV-1 during hematopoietic stem cell mobilization in HIV-1 infected persons [57] and stimulates viral production through binding to CSF3R that is expressed on HIV-1 chronically infected cell lines [58]. The direct impact of CSF3R on HIV replication has been documented recently [46].

Besides, CSF3R has been linked to the developing congenital neutropenia [59]. This is particularly interesting since DARC has also been associated with benign ethnic neutropenia observed in people of African descent [60]. Thus, we hypothesize that in addition to the genetic predisposition of DARC, CSF3R can account for the observed differences in HIV induced neutropenia.

TNFRSF3: Also known as Lymphotoxin-beta receptor (LT-ßR), is a member of the tumor necrosis factor (TNF) receptor superfamily that participates in the regulation of immune and inflammatory responses by propagating signals that regulate cell survival or death through activation of NF-κB[61]. LT-ßR is expressed on myeloids, dendritic cells and monocytes, which play a critical role in the progression towards AIDS by providing a major source and reservoir of virus when the T cell population is depleted [12,62]. Signaling through LT-ßR via its ligand LT-ß stimulates viral replication within infected monocytes [63].

TNFRSF5 (CD40): CD40 is a type I membrane glycoprotein of TNF receptor superfamily and is expressed on B-lymphocytes. Its ligand CD40L is expressed mainly in activated CD4+ T lymphocytes. The interaction between CD40 and CD40L leads to the activation and differentiation of B-lymphocytes [64].

This mechanism constitutes a non-redundant central role in humoral and cell-mediated immunity. Early studies identified a link between CD40L expression and progression to AIDS [65]. Recently it has been demonstrated that HIV-1 promotes CD4+ T cell infection by inserting CD40L into emerging viral particles and trans-activating B cells in a CD40 dependent manner [66].

CD2: It is typically expressed on T cells and most CD3- Natural Killer (NK) cells. It mediates intracellular adhesion in T lymphocytes and targets cells for lysis in NK cells. CD2 has a pivotal role in activating and inducing latent HIV-1 replication in resting CD4+ T cells through the CD2 pathway [67]. The CD2 pathway is also reported to increase HIV production in-vivo [68]. Moreover, a longitudinal study on 'Highly active antiretroviral therapy' over a three-year period showed a significant increase of CD2 expression on peripheral blood mononuclear cells as well as a slight increase in viral load over the same period [69].

IL6ST (GP130): The Glycoprotein 130 is a transmembrane protein that controls the activity of cytokines, such as IL- 6, IL-11, IL-27 and leukemia inhibitory factor (LIF) [70]. It is expressed in many tissues ranging from gut epithelia to astrocytes and T cell subsets. GP130 was associated with HIV when studying LIF's protective role against vertical transmission of HIV-1 from mother to child [71]. Both are significantly upregulated in lymphoid tissue [72] and found in high concentrations in plasma samples of patients [73] during primary HIV-1 infection.

Moreover, GP130 is involved in differentiation among T-helper cell (Th) subsets. A lack of GP130 in T cell specific conditional gp130 deficient mice models causes the activation of Th2 and regulatory T cell pathways [70]. In the case of HIV infection,

this change in T cell differentiation dynamics may be responsible for various levels of disease progression observed in different individuals. The imbalance of Th subsets is also a strong predictor of pathogenic SIV infection in primate models [74]. Similarly, successful CD4+ T cell restoration was associated with enhanced Th17 CD4+ T cell accumulation when comparing gut associated lymphoid tissue recovery rates from HIV infected individuals [75, 76].

CD79B (B29, IGB): CD79 is a transmembrane protein that forms a complex with the B-cell receptor (BCR) and generates a signal following recognition of an antigen by the BCR. It is expressed almost exclusively on B cells and B-cells neoplasms [76]. It is composed of two distinct chains called CD79A and CD79B. CD79B plays an important role in BCR expression in B cell development [77]. HIV Gp120 is documented to down-regulate CD79B [78] but its underlying mechanism is not yet understood. In theory, down-regulation of CD79B leads to reduced capacity of B-cells to bind antigens and more importantly to a decrease in HIV specific antibody formation [79].

We are aware of the difficulties for implicating HIV-1 strains efficiently using alternative co-receptors for infection of transfected cells. Experimental testing usually requires co-culturing of virus strains showing broad co-receptor usage [13, 21, 80] with appropriate transfected cell lines. However, we believe that this effort is necessary for unraveling potential causes underlying confounding traits of HIV-1 infection.

3.6 Conclusions

We use a systems biology framework that integrates protein interactions, functional annotation and protein domains for inferring surface membrane factors interacting with HIV. The analysis of our predictions confirms that surface membrane proteins, even though they are targeted under different conditions, are likely to be part of the same functional pathways.

We infer ten surface proteins that are involved in a cascade of events in HIV infection. Their involvement ranges from serving as co-receptors for cell entry (CCR1 and CCBP2), mediating transinfection (DARC), activating immune cells (CD97) to inducing viral production from latently infected cells (CSF3R, TNFRSF3 and CD2).

We also present eleven original predictions that are potential HIV interacting factors (see Table 3.2). In particular, the platelet glycoprotein Ib (GPIb) is a surface membrane protein of platelets. Mutations in the GPIb beta subunit are associated with Bernard-Soulier syndrome, which is characterized by thrombocytopenia, circulating giant platelets, and prolonged bleeding time [81]. We speculate that the prolonged interaction of blood platelet expressed GP1BB with HIV might be responsible for thrombocytopenia observed in HIV infection. Furthermore, the relaxin receptors RXFP1 and RXFP2 are expressed on the acrosome of elongated

spermatids [82, 83]. Their intron rich gene organization indicates alternatively spliced variants. This suggests the existence of different protein isoforms that contribute to their diverse expression in-vivo. Their association with HIV might explain the different rates of evolution observed in seminal versus blood plasma of infected patients [84]. Moreover, either one or both receptors might be involved in viral hijacking of the spermatozoa in viral transmission [85].

Several seed receptors, such as CCR5, CCR2 and CX3CR1 [86, 87], have been associated with SNPs that contribute to different disease outcome. Among the 21 predicted factors, except for the controversial -46C/C in DARC, SNPs in CCR1, CCBP2, HTR6, HTR1B, HTR1E, CSF3R, IL1R1, TNFRSF5 are associated with one or more clinical phenotypes but their relation to HIV infection has not been investigated.

Thus, we encourage investigating the SNPs from the predicted surface membrane factors for association with HIV to study their potential effect on HIV infection.

Throughout the chapter we have presented a novel method and its application for identifying surface membrane factors for HIV-1. However, we emphasize that the presented framework is neither domain nor disease specific. More precisely, our approach is only depending on the initial (seed) data that is used to establish characteristic functional similarities. Thus, it can be employed for many biological questions other than the one discussed in this manuscript. Potential further applications include, for instance, clinical genetic studies for determining the downstream components of recently discovered disease genes, or drug-target testing for investigating possible effects/interactions of candidate compounds with proteins other than the intended targets. Note that, regardless of the context, it is crucial to test the novel hypotheses resulted from our algorithm with target-oriented in-vivo experiments to fully understand their impact on the system.

Consequently, in this chapter we started with the HIV-1 human protein interaction network and spatially restricted our focus to surface membrane proteins interacting with HIV. Later, we introduced, validated and applied a novel algorithm for predicting “potential missing links in our protein interaction network” based on their functional similarities with the proteins readily available. Finally we presented promising surface membrane factors that are potentially involved in HIV-1 infection using our algorithm.

3.7 References

[1] Cook JA, August A, Henderson AJ (2002) Recruitment of phosphatidylinositol 3-kinase to cd28 inhibits hiv transcription by a tat-dependent mechanism. J Immunol 169: 254-260.

[2] Piguet V, Trono D (1999) The nef protein of primate lentiviruses. Rev Med Virol 9: 111-120.

[3] Yang B, Akhter S, Chaudhuri A, Kanmogne GD (2009) Hiv-1 gp120 induces cytokine expression, leukocyte adhesion, and transmigration across the blood-brain barrier: modulatory effects of stat1 signaling. Microvasc Res 77: 212-219.

[4] van 't Wout AB, Schuitemaker H, Kootstra NA (2008) Isolation and propagation of hiv-1 on peripheral blood mononuclear cells. Nat Protoc 3: 363-370.

[5] Edwards CTT, Holmes EC, Wilson DJ, Viscidi RP, Abrams EJ, et al. (2006) Population genetic estimation of the loss of genetic diversity during horizontal transmission of hiv-1. BMC Evol Biol 6: 28.

[6] Geijtenbeek TB, Kwon DS, Torensma R, van Vliet SJ, van Duijnhoven GC, et al. (2000) Dc-sign, a dendritic cell-specific hiv-1-binding protein that enhances trans-infection of t cells. Cell 100: 587-597.

[7] Kawamura T, Gulden FO, Sugaya M, McNamara DT, Borris DL, et al. (2003) R5 hiv productively infects langerhans cells, and infection levels are regulated by compound ccr5 polymorphisms. Proc Natl Acad Sci U S A 100: 8401-8406.

[8] Becker Y (2006) Respiratory syncytial virus (rsv) evades the human adaptive immune system by skewing the th1/th2 cytokine balance toward increased levels of th2 cytokines and ige, markers of allergy a review. Virus Genes 33: 235-252.

[9] Blaak H, Boers PHM, Gruters RA, Schuitemaker H, van der Ende ME, et al. (2005) Ccr5, gpr15, and cxcr6 are major coreceptors of human immunodefficiency virus type 2 variants isolated from individuals with and without plasma viremia. J Virol 79: 1686-1700.

[10] Liu Y, Liu H, Kim BO, Gattone VH, Li J, et al. (2004) Cd4-independent infection of astrocytes by human immunodefficiency virus type 1: requirement for the human mannose receptor. J Virol 78: 4120-4133.

[11] Chaipan C, Soilleux EJ, Simpson P, Hofmann H, Gramberg T, et al. (2006) Dc-sign and clec-2 mediate human immunodefficiency virus type 1 capture by platelets. J Virol 80: 8951-8960.

[12] Alexaki A, Liu Y, Wigdahl B (2008) Cellular reservoirs of hiv-1 and their role in viral persistence. Curr HIV Res 6: 388-400.

[13] Gorry PR, Dunfee RL, Meford ME, Kunstman K, Morgan T, et al. (2007) Changes in the v3 region of gp120 contribute to unusually broad coreceptor usage of an hiv-1 isolate from a ccr5 delta32 heterozygote. Virology 362: 163-178.

[14] Dong C, Janas AM, Wang JH, Olson WJ, Wu L (2007) Characterization of human immunodeficiency virus type 1 replication in immature and mature dendritic cells reveals dissociable cis- and trans-infection. J Virol 81: 11352-11362.

[15] Romani B, Engelbrecht S, Glashof RH (2010) Functions of tat: the versatile protein of human immunodefficiency virus type 1. J Gen Virol 91: 1-12.13

[16] Sloot PMA, Coveney P, Ertaylan G, Muller V, Boucher C, Bubak M. (2009) HIV Decision Support: From Molecule to Man. Phil Trans R Soc A 367: 2691-2703.

[17] Fang CT, Chang YY, Hsu HM, Twu SJ, Chen KT, et al. (2007) Life expectancy of patients with newly-diagnosed hiv infection in the era of highly active antiretroviral therapy. QJM 100: 97-105.

[18] Ozgur A, Vu T, Erkan G, Radev DR (2008) Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics 24: i277-i285.

[19] Chen J, Aronow BJ, Jegga AG (2009) Disease candidate gene identification and prioritization using protein interaction networks. BMC Bioinformatics 10: 73.

[20] 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. Nucleic Acids Res 37: D417-D422.

[21] Shimizu N, Tanaka A, Oue A, Mori T, Ohtsuki T, et al. (2009) Broad usage spectrum of g proteincoupled receptors as coreceptors by primary isolates of hiv. AIDS 27: 761-769.

[22] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, et al. (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32: D449-D451.

[23] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, et al. (2007) Intact-open source resource for molecular interaction data. Nucleic Acids Res 35: D561-D565.

[24] Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, et al. (2005) The biomolecular interaction network database and related tools 2005 update. Nucleic Acids Res 33: D418-D424.

[25] Pagel P, Kovac S, Oesterheld M, Brauner B, Dunger-Kaltenbach I, et al. (2005) The MIPS mammalian protein-protein interaction database. Bioinformatics 21: 832-834.

[26] Prasad TSK, Goel R, Kandasamy K, Keerthikumar S, Kumar S, et al. (2009) Human protein reference database - 2009 update. Nucleic Acids Res 37: D767-D772.

[27] Chatr-aryamontri A, Ceol A, Palazzi LM, Nardelli G, Schneider MV, et al. (2007) MINT: the Molecular INTeraction database. Nucleic Acids Res 35: D572-D574.

[28] Breitkreutz BJ, Stark C, Reguly T, Boucher L, Breitkreutz A, et al. (2008) The biogrid interaction database: 2008 update. Nucleic Acids Res 36: D637-D640.

[29] UniProt Consortium (2009) The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res 37: D169-D174.

[30] Maglott D, Ostell J, Pruitt KD, Tatusova T (2007) Entrez gene: gene-centered information at ncbi. Nucleic Acids Res 35: D26-D31.

[31] Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2005) Interpro, progress and status in 2005. Nucleic Acids Res 33: D201-D205.

[32] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25: 25-29.

[33] Goh KI, Cusick ME, Valle D, Childs B, Vidal M, et al. (2007) The human disease network. Proc Natl Acad Sci U S A 104: 8685-8690.

[34] Couto FM, Silva MJ, Pedro Coutinho PM (2007) Measuring semantic similarity between geneontology terms. Data Knowl Eng 61: 137-152.14

[35] Li L, Stoeckert CJ, Roos DS (2003) Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189.

[36] Jaeger S, Leser U (2007) High-precision function prediction using conserved interactions. In: Falter C, Schliep A, Selbig J, Vingron M, Walther D, editors, Proceedings of the German Conference on Bioinformatics, GCB 2007, September 26-28, 2007, Potsdam, Germany. GI, volume 115 of LNI, pp. 146-162.

[37] Koschtzki D, Schreiber F (2008) Centrality analysis methods for biological networks and their application to gene regulatory networks. Gene Regul Syst Bio 2: 193-201.

[38] Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking: Bringing order to the web. Technical Report SIDL-WP-1999-0120, Stanford University.

[39] Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems: 1695.

[40] van Dijk D, Ertaylan G, Boucher CA, Sloot PM (2010) Identifying potential survival strategies of hiv-1 through virus-host protein interaction networks. BMC Syst Biol 4: 96.

[41] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) Kegg for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: D355-D360.

[42] Zhang S, Chen H, Liu K, Sun Z (2009) Inferring protein function by domain context similarities in protein-protein interaction networks. BMC Bioinformatics 10: 395.

[43] Broder CC, Collman RG (1997) Chemokine receptors and hiv. J Leukoc Biol 62: 20-29.

[44] Maho A, Bensimon A, Vassart G, Parmentier M (1999) Mapping of the ccxcr1, cx3cr1, ccbp2 and ccr9 genes to the ccr cluster within the 3p21.3 region of the human genome. Cytogenet Cell Genet. 87: 265-268.

[45] Zhou H, Xu M, Huang Q, Gates AT, Zhang XD, et al. (2008) Genome-scale RNAi screen for host factors required for HIV replication. Cell Host Microbe 4: 495-504.

[46] Dunn SJ, Khan IH, Chan UA, Scearce RL, Melara CL, et al. (2004) Identification of cell surface targets for hiv-1 therapeutics using genetic screens. Virology 321: 260-273.

[47] Utaipat U, Duerr A, Rudolph DL, Yang C, Butera ST, et al. (2002) Coreceptor utilization of hiv type 1 subtype e viral isolates from thai men with hiv type 1-infected and uninfected wives. AIDS Res Hum Retroviruses 18: 1-11.

[48] Neil SJD, Aasa-Chapman MMI, Clapham PR, Nibbs RJ, McKnight A, et al. (2005) The promiscuous cc chemokine receptor d6 is a functional coreceptor for primary isolates of human immunodeficiency virus type 1 (hiv-1) and hiv-2 on astrocytes. J Virol 79: 9618-9624.

[49] Walton RT, Rowland-Jones SL (2008) Hiv and chemokine binding to red blood cells, darc matters. Cell Host Microbe 4: 3-5.

[50] He W, Neil S, Kulkarni H, Wright E, Agan BK, et al. (2008) Duffy antigen receptor for chemokines mediates trans-infection of hiv-1 from red blood cells to target cells and affects hiv-aids susceptibility. Cell Host Microbe 4: 52-62.

[51] Winkler CA, An P, Johnson R, Nelson GW, Kirk G (2009) Expression of Duffy Antigen Receptor for Chemokines (DARC) Has No Effect on HIV-1 Acquisition or Progression to AIDS in African Americans. Cell Host Microbe 5: 411-413.15

[52] Horne KC, Li X, Jacobson LP, Palella F, B D Jamieson and JBM, et al. (2009) Duffy Antigen Polymorphisms Do Not Alter Progression of HIV in African Americans in the MACS Cohort. Cell Host Microbe 5: 415-417.

[53] Julg B, Reddy S, van der Stok M, Kulkarni S, Qi Y, et al. (2009) Lack of Duffy antigen receptor for chemokines: no influence on HIV disease progression in an African treatment-naive population. Cell Host Microbe 5: 413-415.

[54] Walley NM, Julg B, Dickson SP, Fellay J, Ge D, et al. (2009) The duffy antigen receptor for chemokines null promoter variant does not influence hiv-1 acquisition or disease progression. Cell Host Microbe 5: 408-10; author reply 418-9.

[55] Kop EN, Matmati M, Pouwels W, Leclercq G, Tak PP, et al. (2009) Differential expression of cd97 on human lymphocyte subsets and limited effect of cd97 antibodies on allogeneic t-cell stimulation. Immunol Lett 123: 160-168.

[56] Morikawa K, Morikawa S, Nakamura M, Miyawaki T (2002) Characterization of granulocyte colonystimulating factor receptor expressed on human lymphocytes. Br J Haematol 118: 296-304.

[57] Baillou C, Simon A, Leclercq V, Azar N, Rosenzwajg M, et al. (2003) Highly active antiretroviral therapy corrects hematopoiesis in hiv-1 infected patients: interest for peripheral blood stem cellbased gene therapy. AIDS 17: 563-574.

[58] Rapaport R, McLean C, Campbell T (2004) Filgrastim stimulates hiv-1 replication in monocytoid cells.. In: Program Abstr Conf Retrovir Oppor Infect 11th 2004 San Franc Calif.

[59] Beel K, Vandenberghe P (2009) G-csf receptor (csf3r) mutations in x-linked neutropenia evolving to acute myeloid leukemia or myelodysplasia. Haematologica 94: 1449-1452.

[60] Reich D, Nalls MA, Kao WHL, Akylbekova EL, Tandon A, et al. (2009) Reduced neutrophil count in people of african descent is due to a regulatory variant in the duffy antigen receptor for chemokines gene. PLoS Genet 5: e1000360.

[61] Li C, Norris PS, Ni CZ, Havert ML, Chiong EM, et al. (2003) Structurally distinct recognition motifs in lymphotoxin-beta receptor and cd40 for tumor

necrosis factor receptor-associated factor (traf)-mediated signaling. J Biol Chem 278: 50523-50529.

[62] Coleman CM, Wu L (2009) Hiv interactions with monocytes and dendritic cells: viral latency and reservoirs. Retrovirology 6: 51.

[63] Marshall WL, Brinkman BM, Ambrose CM, Pesavento PA, Uglialoro AM, et al. (1999) Signaling through the lymphotoxin-beta receptor stimulates hiv-1 replication alone and in cooperation with soluble or membrane-bound tnf-alpha. J Immunol 162: 6016-6023.

[64] Foy TM, Shepherd DM, Durie FH, Arufo A, Ledbetter JA, et al. (1993) In vivo cd40-gp39 interactions are essential for thymus-dependent humoral immunity. ii. prolonged suppression of the humoral immune response by an antibody to the ligand for cd40, gp39. J Exp Med 178: 1567-1575.

[65] Vanham G, Penne L, Devalck J, Kestens L, Colebunders R, et al. (1999) Decreased cd40 ligand induction in cd4 t cells and dysregulated il-12 production during hiv infection. Clin Exp Immunol 117: 335-342.

[66] Martin G, Roy J, Barat C, Ouellet M, Gilbert C, et al. (2007) Human immunodeficiency virus type 1-associated cd40 ligand transactivates b lymphocytes and promotes infection of cd4+ t cells. J Virol 81: 5872-5881.16

[67] Shen A, Yang HC, Zhou Y, Chase AJ, Boyer JD, et al. (2007) Novel pathway for induction of latent virus from resting cd4(+) t cells in the simian immunodeficiency virus/macaque model of human immunodeficiency virus type 1 latency. J Virol 81: 1660-1670.

[68] Bressler P, Pantaleo G, Demaria A, Fauci AS (1991) Anti-cd2 receptor antibodies activate the hiv long terminal repeat in t lymphocytes. J Immunol 147: 2290-2294.

[69] Wu JQ, Dyer WB, Chrisp J, Belov L, Wang B, et al. (2008) Longitudinal microarray analysis of cell surface antigens on peripheral blood mononuclear cells from hiv+ individuals on highly active antiretroviral therapy. Retrovirology 5: 24.

[70] Fasnacht N, M uller W (2008) Conditional gp130 deficient mouse mutants. Semin Cell Dev Biol 19:379-384.

[71] Patterson BK, Behbahani H, Kabat WJ, Sullivan Y, O'Gorman MR, et al. (2001) Leukemia inhibitory factor inhibits hiv-1 replication and is upregulated in placentae from nontransmitting women. J Clin Invest 107: 287-294.

[72] Tjernlund A, Fleener Z, Behbahani H, Connick E, S onnerborg A, et al. (2003) Suppression of leukemia inhibitor factor in lymphoid tissue in primary HIV infection: absence of HIV replication in gp130-positive cells. AIDS 17: 1303-1310.

[73] Tjernlund A, Barqasho B, Nowak P, Kinloch S, Thorborn D, et al. (2006) Early induction of leukemia inhibitor factor (LIF) in acute HIV-1 infection. AIDS 20: 11-19.

[74] Favre D, Lederer S, Kanwar B, Ma ZM, Proll S, et al. (2009) Critical loss of the balance between th17 and t regulatory cell populations in pathogenic siv infection. PLoS Pathog 5: e1000295.

[75] Macal M, Sankaran S, Chun TW, Reay E, Flamm J, et al. (2008) Effective cd4+ t-cell restoration in gut-associated lymphoid tissue of hiv-infected patients is associated with enhanced th17 cells and polyfunctional hiv-specific t-cell responses. Mucosal Immunol 1: 475-488.

[76] Chu PG, Arber DA (2001) Cd79: a review. Appl Immunohistochem Mol Morphol 9: 97-106.

[77] Dobbs AK, Yang T, Farmer D, Kager L, Parolini O, et al. (2007) Cutting edge: a hypomorphic mutation in igbeta (cd79b) in a patient with immunodeficiency and a leaky defect in b cell development. J Immunol 179: 2055-2059.

[78] Patke CL, Shearer WT (2000) gp120- and tnf-alpha-induced modulation of human b cell function: proliferation, cyclic amp generation, ig production, and b-cell receptor expression. J Allergy Clin Immunol 105: 975-982.

[79] Nance CL, Shearer WT (2002) Sdf-1alpha regulates hiv-1-gp120-induced changes in cd79b surface expression and ig production in activated human b cells. Clin Immunol 105: 208-214.

[80] Shimizu N, Tanaka A, Mori T, Ohtsuki T, Hoque A, et al. (2008) A formylpeptide receptor, fprl1, acts as an efficient coreceptor for primary isolates of human immunodeficiency virus. Retrovirology 5: 52.

[81] Hadjkacem B, Elleuch H, Gargouri J, Gargouri A (2009) Bernard-soulier syndrome: novel nonsense mutation in gpibbeta gene affecting gpib-ix complex expression. Ann Hematol 88: 465-472.

[82] Filonzi M, Cardoso LC, Pimenta MT, Queiroz DBC, Avellar MCW, et al. (2007) Relaxin family peptide receptors Rxfp1 and Rxfp2: mapping of the mRNA and protein distribution in the reproductive tract of the male rat. Reprod Biol Endocrinol 5: 29.17

[83] Gianesello L, Ferlin A, Menegazzo M, Pepe A, Foresta C (2009) Rxfp1 is expressed on the sperm acrosome, and relaxin stimulates the acrosomal reaction of human spermatozoa. Ann N Y Acad Sci 1160: 192-193.

[84] Ghosn J, Viard JP, Katlama C, de Almeida M, Tubiana R, et al. (2004) Evidence of genotypic resistance diversity of archived and circulating viral strains in blood and semen of pre-treated hiv-infected men. AIDS 18: 447-457.

[85] Kern A, Bryant-Greenwood GD (2009) Mechanisms of relaxin receptor (lgr7/rxfp1) expression and function. Ann N Y Acad Sci 1160: 60-66.

[86] Passam AM, Sourvinos G, Krambovitis E, Miyakis S, Stavrianeas N, et al. (2007) Polymorphisms of cx(3)cr1 and cxcr6 receptors in relation to haart therapy of hiv type 1 patients. AIDS Res Hum Retroviruses 23: 1026-1032.

[87] Singh P, Kaur G, Sharma G, Mehra NK (2008) Immunogenetic basis of hiv-1 infection, transmission and disease progression. Vaccine 26: 2966-2980.

3.8 Supplementary Material

Table 3.3: The number of human protein interactions retrieved from each database by the time of our study. The integration of the protein interactions from the different databases results in a protein interaction set with 13,494 human proteins and 43,637 unique interactions between these proteins. Note, that there is an overlap between the databases, thus the numbers do not added up to the final number unique interactions.

PPI Database	Number of human protein interactions
MIPS-MPPI	127
DIP	3045
MINT	3160
BIND	5969
BioGRID	12779
IntAct	14298
HPRD	19215

Table 3.4: Average number of proteins (network size) comprised in each network and the number of seeds that are re-discovered during cross-validation when considering different data for generating the specific HIV receptor network.

HIV Network Type	Average Network Size	Number of recovered receptors
PPI Network	89 (± 6)	2 of 13
PPI+GO Network	418 (± 8)	11 of 13
PPI+ GO_{enrich} Network	726 (± 16)	12 of 13

Table 3.5: List of functional annotation from Gene Ontology that are used to filter for receptor proteins for the control sets.

GO Category	GO Term (GO Id)	Definition
Molecular Function	Receptor activity (GO:0004872)	Combining with an extracellular or intracellular messenger to initiate a change in cell activity.
	Co-receptor activity (GO:0015026)	Combining with an extracellular or intracellular messenger, and in cooperation with a nearby primary receptor, initiating a change in cell activity
Biological Process	Receptor metabolic process (GO:0043112)	The chemical reactions and pathways involving a receptor molecule, a macromolecule that undergoes combination with a hormone, neurotransmitter, drug or intracellular messenger to initiate a change in cell function.
Cellular Compartment	Receptor complex (GO:0043235)	Any protein complex that undergoes combination with a hormone, neurotransmitter, drug or intracellular messenger to initiate a change in cell function.
	Membrane (GO:0016020)	Double layer of lipid molecules that encloses all cells, and, in eukaryotes, many organelles; may be a single or double lipid bilayer; also includes. associated proteins.
	Extracellular space (GO:0005615)	That part of a multicellular organism outside the cells proper, usually taken to be outside the plasma membranes, and occupied by fluid.

Table 3.6: Chromosomal location of known and predicted surface membrane proteins. Similar chromosome regions are colored similarly.

Known factors	Chrom. Location	Predicted factors	Chrom. Location
CXCR4	2q21	CD2	1q13.1
GPR1	2q33.3	DARC	1q21-q22
ITGA4	2q31.3	HTR6	1p36-p35
CCR9	3p21.3	CSFR3	1p35-p34.3
CCR3	3p21.3	IL1R1	2q12
CCR2	3p21.3	GPR17	2q21
CCR5	3p21.31	CCR1	3p21
CX3CR1	3p21|3p21.3	CCBP2	3p21.3
CXCR6	3p21	RXFP1	4q32.1
CCR8	3p22	GYPB	4q28-q31
APJ	11q12	IL6ST	5q11
CD4	12pter-p12	HTR1B	6q13
DC-SIGN	19p13	HTR1E	6q14-q15
		TNFRSF3	12p13
		GPR182	12q13.3
		RXFP2	13q13.1
		CD79B	17q23
		CD97	19p13
		TNFRSF5	20q12-q13.2
		NPBWR2	20q13.3
		GP1BB	22q11.21-q11.23|22q11.21

Table 3.7: List of seed HIV receptors, including receptor name and type, their functional domains and references indicating an association with HIV-1 infection.

Receptor	Receptor type	InterPro domains	References
CD4	Primary receptor for HIV	Ag_CD4, CD4-extracel, Ig-like, Ig-like_fold, Ig_C2-set, Ig_sub, Ig_V-set_sub	[1,2]
CCR5	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_5_rcpt	[1-6]
CCR3	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_3_rcpt	[4,6,7]
CCR2	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_2_rcpt, CC_5_rcpt	[3,4]
CCR8	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_8_rcpt	[8,9]
CCR9	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CC_9_rcpt	[10,11]
CXCR4	Alternative co-receptor with CD4	7TM_GPCR_Rhodpsn, CXC_4_rcpt	[1,4]
CXCR6	Co-receptor	7TM_GPCR_Rhodpsn, CXC_6_rcpt	[12,5]
CX3CR1	Co-receptor with CD4	7TM_GPCR_Rhodpsn, CX3C_fract_rcpt	[12-14]
APJ	Alternative co-receptor	7TM_GPCR_Rhodpsn, APJ_rcpt	[11,15-18]
GPR1	Alternative co-receptor	7TM_GPCR_Rhodpsn, GPR1_rcpt	[19-21]

ITGA4	Co-receptor with CD4	Int_alpha_beta-p, Integrin_alpha, Integrin_alpha-2, Integrin_alpha_C	[22-25]
DC-SIGN	Receptor for HIV	AntifreezeII, C-type_lectin.	[26-29]

Receptors are grouped according to their functional domains.

Receptor	**InterPro domains**	**Cell types**	**Association with HIV**
DARC	Duffy_cmk_rcpt.	High expression: (early) erythroid, endothelial cells	+ [30-32]
CCR1	7TM_GPCR_Rhodpsn, CC_1_rcpt	High expression: whole blood, monocytes, myeloid, dendritic cell	+ [33-35]
CCBP2	7TM_GPCR_Rhodpsn, CXC_4_rcpt	Uniform expression*	+ [36-37] - [38]
CD97	EGF-type_Asp/Asn_hydroxyl_site, EGF_Ca_bd_2, GPCR_2_CD97, GPCR_2_secretin-like, GPS_dom	High expression: CD34, B lymphoblast, dendritic cells, CD8 and CD4 T- cells, NK, myeloid, monocytes	+ [39]
GP1BB	LRR-contain_N, Cys-rich_flank_reg_C	High expression: CD34, monocytes and whole blood	?
HTR6	7TM_GPCR_Rhodpsn	Uniform expression	+
HTR1B	5HT1B_rcpt, 7TM_GPCR_Rhodpsn	Uniform expression	?
HTR1E	5HT1F_rcpt, 7TM_GPCR_Rhodpsn	Uniform expression	?
RXFP2	7TM_GPCR_Rhodpsn, LDL_rcpt_classA_cys-rich_rpt, Leu-rich_rpt, LRR-contain_N, Leu-rich_rpt_typical-subtyp, Relaxin_rcpt	Low expression	?
RXFP1	7TM_GPCR_Rhodpsn, LDL_rcpt_classA_cys-rich, Leu-rich_rpt, LRR-contain_N, Leu-	No expression profiles available	?

	rich_rpt_typical-subtyp, Relaxin_rcpt		
GPR17	7TM_GPCR_Rhodpsn, P2_purnocptor	Uniform expression	?
GPR182	7TM_GPCR_Rhodpsn, G10D_rcpt	Uniform expression	?
NPBWR2	7TM_GPCR_Rhodpsn, Neuropept_W_rcpt	Uniform expression	- [40]
GYPB	Glycophorin	High expression: (early) erythroid and endothelial cells	?
CD2	Ag_CD2, Ig-like_fold, Ig_C2-set, Ig_V-set, T-cell_sdhesion_molc_CD2.	High expression: dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood	+ [41-44]
CSF3R	FN_III, Hematopoietin_rcpt_gp130_CS, IgC2-like_lig-bd	High expression: myeloid cells, monocytes and whole blood	+ [45,46]
IL1R1	Ig, Ig-like, Ig-like_fold, Ig_sub, IL1_rcpt_1, IL1R_rcpt	No expression profile available	- [47,48]
CD79B	Ig-like, Ig-like_fold, Ig_sub, Ig_V-set, Phos_immunorcpt_sig_ITAM.	High expression: CD34, endothelial and dendritic cells	+ [42,49]
IL6ST	FN_III, Hematopoietin_rcpt_gp130_CS, Ig-like_fold, IgC2-like_lig-bd	Uniform expression	+ [50-52]
TNFRSF5	Fas_rcpt, TNFR_Cys_rich_reg	High expression: B lymphoblasts	+ [53-55]
TNFRSF3	TNFR_3_LTBR, TNFR_Cys_rich_reg	High expression: myeloid, monocytes and whole blood	+ [56]

(*): CD34, endothelial, B lymphoblasts, dendritic, myeloid, monocytes, NK, CD8 and CD4 T cells, whole blood

(+) Predictions with literature supporting an interaction with HIV are marked as (+).

(-) Indicates predictions with negative evidences.

(?) For predictions without literature on interaction or non-interaction the association remains unclear.

3.8.1 Supporting Information

3.8.1.1 Functional similarity of proteins

We determine the functional similarity between two proteins by analyzing their GO annotations using semantic similarity. We first compute the similarity of two GO terms and extend the measure to determine the functional similarity of two proteins annotated with several GO terms. Note, the functional similarity between two proteins is computed separately for each of the GO subontologies: molecular function (MF), biological process (BP) and cellular component (CC).

Semantic similarity between GO terms:

To compute the semantic similarity between two GO terms we use the approach proposed by Lin [1]. Following Lin's definition, the information content of a GO term *t* is defined as follows:

$$IC(t) = -log\left(\frac{freq(t)}{freq(root)}\right), \quad (1)$$

Where the frequency of a term is defined as the number of times a term or any of its descendants occurs. Thus, less frequent terms and terms with few occurring descendants are considered more informative.

Based on this measure, the semantic similarity between two terms is defined as the ratio of the information content of their most informative common ancestor and the information contents of both concepts [1]. The information content of the most informative common ancestor is given by:

$$shareIC(t_1, t_2) = max\left\{IC(t)|t \in CA(t_1, t_2)\right\}, \quad (2)$$

Where $CA(t_1; t_2)$ is the set of all common ancestors between terms t_1 and t_2. The similarity between two terms is then defined as:

$$sim(t_1, t_2) = \frac{2 * shareIC(t_1, t_2)}{IC(t_1) + IC(t_2)}. \quad (3)$$

Semantic similarity between proteins:

The semantic similarity between proteins is determined based on the similarity of their associated GO terms. Since often proteins are annotated with more than one term, the similarity of a protein *p* to a group *g* of terms is defined as the average similarity of its terms to their most similar terms in *g* [2] (where *t(p)* is the set of terms annotated to protein *p*):

$$Sim(p,g) = \frac{\sum_{t_1 \in t(p)} max\{sim(t_1,t_2)|t_2 \in g\}}{|t(p)|} \quad (4)$$

Finally, the functional GO similarity between two proteins is defined as the average similarity of their GO terms:

$$GO_{Sim}(p_1,p_2) = \frac{Sim(p_1,t(p_2)) + Sim(p_2,t(p_1))}{2}. \quad (5)$$

GO$_{Sim}$ ranges between 0 and 1 depending on the similarity of the GO annotations between two proteins, whereby 1 indicated functional equality and 0 indicates maximal functional distance. The functional similarity of all three GO sub-ontologies is added and then averaged to obtain an overall similarity score for two proteins:

$$GO_{Sim}(p_1,p_2) = \frac{GO_{Sim_{\mathrm{MF}}}(p_1,p_2) + GO_{Sim_{\mathrm{BP}}}(p_1,p_2) + GO_{Sim_{\mathrm{CC}}}(p_1,p_2)}{3}. \quad (6)$$

3.8.1.2 Impact of the functional data on the outcomes of the prediction methods:

We use protein interaction data and functional annotations to generate an HIV specific receptor network. In addition, we assessed the influence of using manually curated and predicted functional annotation on our prediction method by applying it to differently compiled HIV networks.

HIV network types:

First, we only considered proteins that interact directly with any seed receptors when generating the specific HIV receptor network, which will be called PPI network. Next, we integrated proteins that interact directly with any seed and all proteins which are functionally very similar to any seed considering only manually curated functions -PPI-GO network. Third, we consider interaction data in combination with enriched functional annotation (manual curated and predicted function) –PPI-GO$_{enrich}$ network.

Performance comparison of the HIV network types:

We compare the ability of our framework to find novel surface membrane factors within the three different HIV networks by using cross-validation. Leave-one-out cross-validations are performed over the 13 known HIV receptors for the PPI, PPI-GO and PPI-GO$_{enrich}$ networks. For cross-validation, we remove one known HIV receptor from the initial list and try to re-discover this receptor by means of our method. We build an HIV receptor network by considering only the remaining receptors as seeds and rank the proteins according to their centrality within the network. Subsequently, we determine whether the left-out receptor is rediscovered and at which position of the ranked list. We repeat this procedure for each seed and determine an average recovery rate across all receptors and for each network type.

shows the average network size and the number of recovered (hidden) seeds for the three different kinds of HIV-receptor networks.

The seed re-discovery rate is very low when using only protein interaction data. Only two out of 13 receptors can be captured within the generated networks. This rate increases significantly up to 11 and 12 detected receptors when considering additionally functional annotation (PPI-GO) as well as predicted functions (PPI-GO$_{enrich}$), respectively. Two receptors are not covered in the PPI-GO networks, namely DC-SIGN and ITGA4, whereas the latter one is also not detected using the enriched network, most likely due to different ligands and a lower functional similarity to the other seed receptors. In general, the number of re-discovered receptors is relatively low when considering only the top ranked proteins (e.g. x = 5%). However, the recovery rate increases significantly the more proteins of the respective networks are examined (except for PPI), until it converges to the total number of detected seeds. Protein interaction data alone is not sufficient for finding the known receptors, since it captures similar ligands rather then functionally similar receptors. Utilizing interaction and functional annotations allows to generate more complete networks in biological sense. This is reflected in the average network size of the different network types which increases from 89 proteins to 418 and 726 for PPI-GO$_{enrich}$.

Comparing the recovery rates of PPI-GO and PPI-GO$_{enrich}$ across the ranked list clearly shows that 'hidden' receptors are better recovered and more highly ranked within the enriched than in the non-enriched network. However, the superior performance might result from the larger size of the enriched networks, e.g. the number of proteins that is considered at the different x is twice as high for PPI-GO$_{enrich}$ because the networks are in average about two times larger. To ensure that the higher recovery rate is not affected by the larger amount of proteins we normalize the recovery rates by the number of proteins considered at each rank x.

3.9 Additional Material References

[1] Lin D (1998) An information-theoretic defnition of similarity. In: Proceedings of the 15th ICML. Madison WI, pp. 296-304.

[2] Couto FM, Silva MJ, Pedro Coutinho PM (2007) Measuring semantic similarity between gene ontology terms. Data Knowl Eng 61: 137-152.