BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 17 2011, pages 2414–2421
doi:10.1093/bioinformatics/btr414

Databases and ontologies
Advance Access publication July 28, 2011

Tissue-specific subnetworks and characteristics of publicly available human protein interaction databases

Tiago J. S. Lopes[1], Martin Schaefer[2], Jason Shoemaker[1], Yukiko Matsuoka[1,3], Jean-Fred Fontaine[2], Gabriele Neumann[4], Miguel A. Andrade-Navarro[2], Yoshihiro Kawaoka[1,4,5] and Hiroaki Kitano[3,6,7,8,*]

[1]JST ERATO KAWAOKA Infection-induced Host Responses Project, Tokyo, Japan, [2]Computational Biology and Data Mining, Max Delbrück Center for Molecular Medicine, Berlin, Germany, [3]The Systems Biology Institute, Tokyo, Japan, [4]Department of Pathobiological Sciences, Influenza Research Institute, University of Wisconsin-Madison, School of Veterinary Medicine, Madison, WI, USA, [5]Institute of Medical Science, Division of Virology, Department of Microbiology and Immunology, University of Tokyo, [6]Sony Computer Science Laboratories, Inc., Tokyo, [7]Open Biology Unit, Okinawa Institute of Science and Technology, Okinawa and [8]Division of Cancer Systems Biology, Cancer Institute, Japanese Foundation for Cancer Research, Tokyo, Japan

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Protein-protein interaction (PPI) databases are widely used tools to study cellular pathways and networks; however, there are several databases available that still do not account for cell type-specific differences. Here, we evaluated the characteristics of six interaction databases, incorporated tissue-specific gene expression information and finally, investigated if the most popular proteins of scientific literature are involved in good quality interactions.

Results: We found that the evaluated databases are comparable in terms of node connectivity (i.e. proteins with few interaction partners also have few interaction partners in other databases), but may differ in the identity of interaction partners. We also observed that the incorporation of tissue-specific expression information significantly altered the interaction landscape and finally, we demonstrated that many of the most intensively studied proteins are engaged in interactions associated with low confidence scores. In summary, interaction databases are valuable research tools but may lead to different predictions on interactions or pathways. The accuracy of predictions can be improved by incorporating datasets on organ- and cell type-specific gene expression, and by obtaining additional interaction evidence for the most 'popular' proteins.

Contact: kitano@sbi.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on April 7, 2011; revised on June 3, 2011; accepted on July 7, 2011

1 INTRODUCTION

Traditionally, studies that assess the cellular metabolism, disease and cancer development, pathogens infections or drug–protein interaction have focused on single genes or proteins. While such studies have created large amounts of data, they typically do not account for the multiple interactions that regulate cellular networks. Recently, high-throughput approaches including yeast two-hybrid screens (Rual *et al.*, 2005), immunoprecipitation studies followed by mass spectrometry analysis (Ewing *et al.*, 2007), transcriptomics (Wilhelm *et al.*, 2008) and metabolomics studies (Braaksma *et al.*, 2011) have become important research tools to identify protein–protein interaction (PPI) partners (Krogan *et al.*, 2006) or cellular factors that are up- or downregulated in response to specific stimuli (Bhattacharya *et al.*, 2004). With the availability of the resulting large datasets, the challenge now lies in the generation of comprehensive and robust interactome maps, ideally capturing all PPIs within a cell and between cells at any given moment in time.

The human proteome is estimated to encompass 130 000–650 000 PPIs (Stumpf *et al.*, 2008; Venkatesan *et al.*, 2009). Of those, only a subset has been described at this point, establishing PPI databases that provide valuable information about the reactions occurring at the proteome level. Previous studies analyzed and compared some of these databases (Mathivanan *et al.*, 2006; Ramirez *et al.*, 2007; von Mering *et al.*, 2002); however, these analyses were based on the significantly smaller datasets available at the time of the analysis, and included only subsets of currently popular PPI databases. Therefore, we analyzed the following four popular PPI databases (Table 1): HPRD [Human Protein Reference Database (Prasad *et al.*, 2009)]; MINT [Molecular INTeraction (Ceol *et al.*, 2010)]; INTACT (Aranda *et al.*, 2010) and BioGRID [Biological General Repository for Interaction Datasets (Breitkreutz *et al.*, 2008)]. In addition, we also included in the comparison a recently published database named HIPPIE (Human Integrated Protein Protein Interaction rEference, http://cbdm.mdc-berlin.de/tools/hippie/) (M.Schaefer *et al.*, submitted for publication). It is assembled through the compilation of several PPI sources, including the previously mentioned databases. Lastly, for the sections of this study not involving network topological characteristics, we also included the STRING database (Search Tool for the Retrieval of INteracting Genes/Proteins) (Jensen *et al.*, 2009), a popular resource that in addition to protein interactions, also contains protein associations from several pathway databases. MINT, HPRD, BIOGRID and

*To whom correspondence should be addressed.

Table 1. Database characteristics

	HPRD	HIPPIE	STRING[a]	MINT	INTACT	BIOGRID
Proteins	9117	11 835	10 546	5206	8310	9057
Interactions	36 239	72 916	144 099	12 579	33 299	37 469
Average degree[b]	8	12	–	4.83	8.01	8.27
Average betweenness[c]	13 528	15 840	–	8009	11 909	13 639
Diameter[d]	14	13	–	12	13	12
Average path length[e]	4.25	3.79	–	4.43	3.96	4.21
Clustering coefficient[f]	0.05	0.05	–	0.03	0.03	0.06

[a] STRING is not a PPI database, thus we did not compute the features that are commonly used for network structure analysis.

[b] Average degree describes the average number of interactions.

[c] Average betweenness describes the ‘centrality’ of a factor in a network.

[d] Diameter describes the maximal distance between the two most distant nodes in a network.

[e] Average path length describes the average number of steps that connect any two components.

[f] Clustering coefficient describes the tendency of nodes to interact among each other forming groups.

INTACT are manually curated and have thousands of interactions submitted by the community; thus, since they offer original interactions used by other databases, we refer to these four databases as ‘primary resources’. HIPPIE and STRING are composed of interactions taken from primary databases and other sources; hence, we refer to HIPPIE and STRING as ‘derived databases’. In addition, for the purpose of this study we removed all predicted functional associations present in STRING.

Here, we focused on the human subset of interaction databases, and as an improvement over most current analyses, we demonstrated the usefulness of organ or cell type-specific subnetworks. We analyzed these databases for their basic features including protein coverage, number of interactions and neighborhood characteristics (i.e. we compared the number and identity of interactions partners, and asked whether proteins that are a hub in one database occupy a similar position in other databases). Finally, using three databases that assign confidence scores to its interactions, we demonstrated that there is a lack of interaction data with high confidence scores for many intensively studied proteins. Additional experimental evidence for those interactions, either confirming or refuting, would significantly increase the robustness of current PPI databases.

2 METHODS

2.1 Databases

The databases were obtained from their respective websites in the following versions or latest updates: HPRD Release 9; HIPPIE 1.1; STRING 8.3; MINT 15.December.2010; INTACT 21.April.2011; BIOGRID 3.1.76. Before initiating the analysis, the following pre-processing steps were carried out: (i) we removed all redundant interactions, keeping just the interaction with the highest score. (ii) For all protein entries, their database-specific identification tags were converted to a common nomenclature (Entrez Gene IDs). Proteins that did not have a matching ID in Entrez Gene were discarded. Approximately 10% of interactions had to be removed from each database.

In the STRING database, we performed additional pre-processing step: we removed all interactions involving non-human proteins, left only interactions with experimental evidence or obtained from pathway and other interaction databases (i.e. removed interactions derived from co-expression, genomic neighborhood, text mining and other predictive techniques).

2.2 Network and statistical analysis

All interaction databases were converted to an undirected graph and further analyzed using R (version 2.10.1) and the iGraph library (version 0.5.4). From this library, we used routines to find the degree, betweenness, diameter, shortest path, immediate neighbors and clustering coefficient. The other statistical tests (Welch, Wilcoxon, *z*-score) were performed using R with 0.95 confidence interval. Pathway and Gene Ontology enrichment analysis were performed with DAVID (Huang *et al.*, 2008) and ConsensusPath DB (Kamburov *et al.*, 2011) using the default parameters values. For the enrichment analysis tests, we used the list of proteins present in the tissue-specific subnetworks as background.

2.3 Popular genes

The file gene2pubmed from the NCBI public FTP site contains a table with Pubmed IDs and the genes present in this each abstract (sorted by species). This file was used to rank the human genes according to the number of abstracts in which they appear and to select the 10% most popular genes (2911 entries). The file was obtained on April 22, 2010.

2.4 Gene expression data

We obtained an Affymetrix dataset containing the transcription levels of 84 human tissues and cell lines. This dataset is publicly available for query and download from the BioGPS project (Su *et al.*, 2004; Wu *et al.*, 2009).

We obtained the normalized expression data [pre-processed using GCRMA—GeneChip Robust Multiarray Averaging (Gentleman *et al.*, 2004)] and divided our analysis in the following steps: first, we defined that each probe must have an absolute intensity >50 for at least one condition, thus removing any probe not being moderately or strongly expressed in at least one tissue (the original datasets have no specific background level). After this cutoff, 16 704 probes remained from the original dataset of 44 775 probes. With the remaining probes, we converted their Affy_ID to Entrez Gene IDs and in this conversion 3537 probes had no matching ID. In the end, our dataset consisted of 12 956 probes that mapped to 9214 different genes. Finally, we calculated the *z*-score for each probe across all tissues. Using a *z*-score cutoff of 0.1, we determined which genes were moderately to highly expressed in each tissue.

2.5 Protein degree categorization

We classified the proteins into three categories (high-, middle- and low degree) according to their number of interactions. To define the appropriate ranges, we ranked the proteins in decreasing order according to their number of interactions. With this list, we used a procedure which selected two random numbers: the first in the interval [80, 98] (we refer to it as *value1*) and the second in the interval [60, *value1*] (we call it *value2*). Subsequently, we considered high-degree proteins as those that occupied a position among the top *value1*% of the ranked list. Middle-degree proteins were those that occupied a position in the interval [*value2*, *value1*]% of the ranked list and finally, the low-degree proteins were on the [1, value2]% of the list. For a visual explanation of the procedure, please refer to Supplementary Figure S1. To verify the robustness of the results, this procedure was repeated 100 times for each pair of databases being compared and the mean and SDs determined. We used this procedure instead of defining a fixed number of neighbors that a protein should have to belong to each category. The differences in the network sizes would cause the results to be unfairly dependent on the ranges selected.

3 RESULTS

3.1 Database features

Table 1 compares the features of the six databases included in the analysis. The number of proteins (nodes) in these databases ranges from ~5200 to ~12 000. STRING and HIPPIE contain the largest numbers of proteins since they include data from several other databases in addition to their own unique data.

For all databases except STRING, the total number of interactions ranges from ~12 500 to ~73 000 (Table 1). MINT has relatively few proteins and interactions, all of which are covered by one or several of the other databases. In contrast, >140 000 interactions are reported in STRING, which comes close to the number of estimated interactions in the human proteome (Stumpf *et al.*, 2008; Venkatesan *et al.*, 2009). We found that 4361 proteins and 5589 PPIs were reported in at least two different databases, with the largest overlap between STRING and HIPPIE (Supplementary Table S1). Only 1453 proteins and 1619 PPIs are reported in all six databases. These interactions are reported in primary resource databases and are likely to stem from the same portion of literature that was manually curated by the authors (Turinsky *et al.*, 2010).

Next, we compared the average degree and betweenness of the proteins in each database. The average degree (average number of interactions per protein) ranges between 5 and 12, with HIPPIE showing the highest average number of neighbors for each protein (Supplementary Fig. 2A shows the distributions of degree and betweenness in each database). Betweenness, in a broader sense describes the significance of a node (i.e. a protein in a PPI network) for the flow of information between different points in the network. It is calculated as follows:

$$B(v) = \sum \frac{s_{ij}(v)}{s_{ij}}, \qquad \text{with } i \neq j,\ v \neq i \quad \text{and } v \neq j \tag{1}$$

where s_{ij} is the number of shortest paths between the nodes i and j and $s_{ij}(v)$ is the fraction of those shortest paths passing through node v. High betweenness thus indicates that the respective protein has a 'central' position in the network, and that the perturbation of this protein may significantly affect the flow of information through the network. The average betweenness of the analyzed databases are similar (Table 1), with the exception of MINT, which has a slightly lower value. This was expected for all networks since they have similar structure, observed in their clustering coefficients, average degree and path lengths. The majority of proteins in all databases have medium to high betweenness values (defined here as 4.5 to 10.5 on a natural logarithm scale; see Supplementary Fig. 2B), even though the number of interaction partners may be limited for these proteins. This suggests that even proteins with few interaction partners occupy important intermediate positions in a network (Joy *et al.*, 2005).

Finally, several measures of the overall network structure were compared for each database. The 'diameter' of a network defines the *maximal* distance between the two most distant nodes in the network while the average path length (APL) is the mean distance between all protein pairs in the network. As summarized in Table 1, the diameters and APLs of each network are comparable.

These findings collectively show that the databases have a similar network structure, although primary (MINT, INTACT, HPRD) and the derived database (HIPPIE) have a considerable difference in the number of interactions.

3.2 Conserved topological characteristics between databases

After characterizing the basic features of the databases selected for this study, we next assessed their topological characteristics. 'Topology' describes the arrangements in which nodes are connected to each other in a database. Important topological parameters are the number and the identity of interaction partners. Such information is critical for the identification of hubs, which are often targeted for the identification of possible lethal genes (Albert *et al.*, 2000; Coulomb *et al.*, 2005; Jeong *et al.*, 2001), the development of novel drugs (Hase *et al.*, 2009; Yildirim *et al.*, 2007) or network disruption (Quayle *et al.*, 2007).

To this end, we adopted a strategy used for drug target identification and protein essentiality studies in which proteins are grouped into one of three categories based on the number of interactions (Han *et al.*, 2004; Hase *et al.*, 2009; Patil and Nakamura, 2006). We ranked the proteins according to their number of interactions and classified them as high-, middle- or low-degree proteins (Section 2). STRING was excluded from this analysis because it comprises not only protein interactions but also other types of non-physical, protein associations derived from pathway databases, in addition to co-expression of genes and genomic neighborhood.

After categorizing all proteins, we assessed the percentages of proteins that fall into the same or different categories in pair-wise database comparisons. Figure 1 shows that 60–80% of the proteins shared between two databases fall into the same category in both databases. This shows that although the databases differ in the number of proteins and interactions, their shared proteins still have similar connectivity levels.

In our pair-wise comparisons, we matched the smaller database (with fewer interactions; e.g. HPRD) against the larger database (with more interactions; e.g. HIPPIE) (Fig. 1). As a result, most proteins that fall into different categories between the databases shift into a higher degree category (e.g. the protein shifts from 'low degree' to 'middle degree'). However, we observed that when INTACT is matched against HPRD and BIOGRID, ~10% of the proteins that are in the 'middle degree' category in the smaller database (i.e. INTACT) shift to the 'low degree' category in the larger database (i.e. HPRD or BIOGRID) (Fig. 1). Most likely, this is a consequence of the different experimental datasets used in the different databases and we observed that those proteins show enrichment for translational elongation and RNA processing Gene Ontology categories ($P < 0.01$).

Notably, very few proteins changed between the 'high degree' and 'low degree' categories (or vice versa) when comparing databases (Fig. 1). This further supports our notion that the five databases included in this analysis are in fairly good agreement regarding the connectivity of the proteins.

The only exception is the comparison of MINT with HIPPIE and other larger databases, with almost 10% of the proteins falling into the 'low degree' category in MINT, but into the 'high degree' category in HIPPIE. We attribute this finding to the different sizes of databases, with MINT and HIPPIE representing the smallest and largest datasets analyzed (both in terms of numbers of proteins and interactions, Table 1).

3.3 Neighborhood characteristics of datasets

The topological characteristics of a protein in a database are not only defined by the *number* of interaction partners, but, perhaps even more importantly, by the *identity* of interaction partners. We therefore assessed whether proteins have similar or different

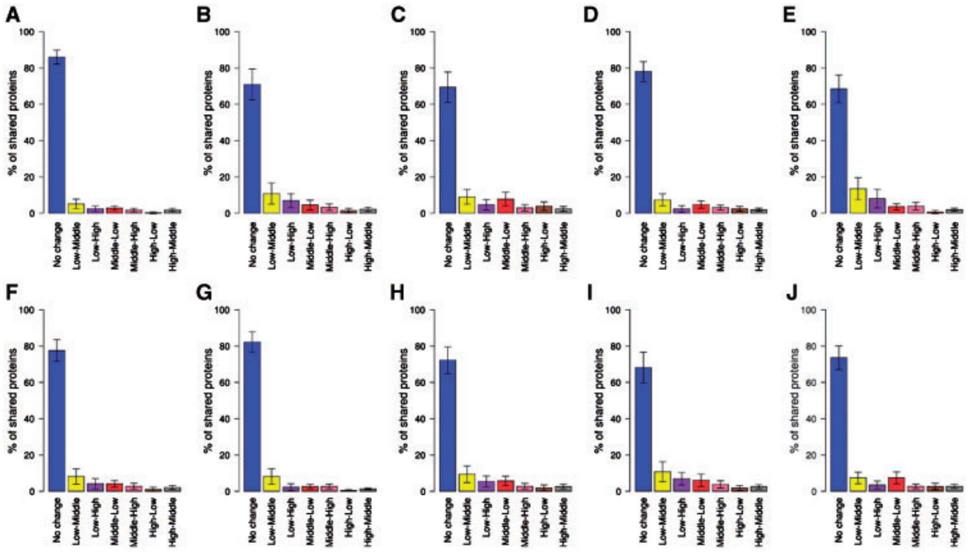

Fig. 1. Proteins were grouped into three categories: low-, middle- and high degree (see Section 2). Then, we assessed the percentages of proteins that fall into the same (or different) categories in pair-wise comparisons of two databases. For most comparisons, 60–80% of proteins fall into the same category in both databases compared. (**A**) HPRD-HIPPIE; (**B**) MINT-HPRD; (**C**) INTACT-HPRD; (**D**) BIOGRID-HPRD; (**E**) MINT-HIPPIE; (**F**) INTACT-HIPPIE; (**G**) BIOGRID-HIPPIE; (**H**) MINT-INTACT; (**I**) MINT-BIOGRID; (**J**) INTACT-BIOGRID.

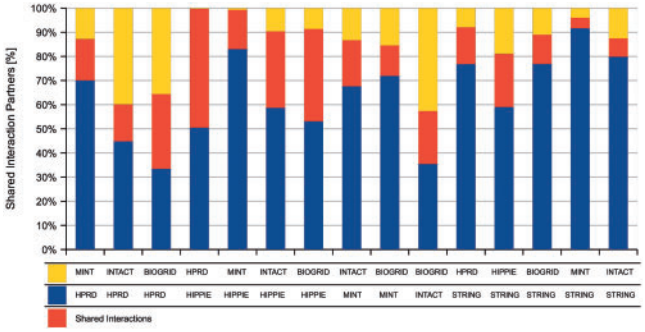

Fig. 2. Shared and exclusive interaction partners in a pair-wise comparison of PPI databases. For proteins shared between two databases, we identified their interaction partners in each of the databases, and then compared the interaction partners. Yellow and blue represent the indicated databases. Shown in red are the interaction partners predicted in both databases.

interaction partners in the databases analyzed. For our analysis, we focused on the ‘shared’ proteins, i.e. those listed in the two databases being compared. For these proteins, we identified their interaction partners in each of the databases, and then compared the interaction partners between the databases (Fig. 2; see also Supplementary Fig. S3 for the absolute numbers).

As expected, the highest percentage of shared neighbors was detected for the comparison of derived resources (STRING and HIPPIE) to primary resources (MINT, BIOGRID and INTACT). However, for comparisons that do not involve the HIPPIE database, no more than 40% of interaction partners are shared. As described earlier, STRING comprises not only protein interactions, but also other functional associations originating, for example, from pathway databases (Jensen *et al.*, 2009; von Mering *et al.*, 2005). This results in a large number of interactions that are not covered by the other databases and transforms the interactions of the other databases into a subset of those reported by STRING.

Collectively, our analysis revealed considerable differences in predicted interaction partners between the databases. These differences likely stem from differences in the size of databases, algorithms used, and differences in the portion of the literature used by primary database curators. Researchers should take these issues into account when attempting to identify critical interaction partners of their protein(s) of interest.

3.4 Quality of interactions of key protein sets

STRING, HIPPIE and MINT assign quality scores to each interaction and this is used to assess the confidence level of an analysis; HIPPIE and MINT calculate the confidence score based on accumulated experimental evidence of protein interactions (M.Schaefer *et al.*, submitted for publication) (Ceol *et al.*, 2010). This stringent approach leads to scores below 0.5 for more than 75% of the interactions reported in these databases (Fig. 3A). STRING calculates its confidence score based on the likelihood that two proteins have a functional association that is as specific as the association between an average pair of proteins present in the same KEGG pathway (Kanehisa *et al.*, 2010; Szklarczyk *et al.*, 2011). In addition, higher scores are assigned to associations supported by several sources of evidence. Consequently, intensively studied interactions are more likely to be supported by higher confidence scores. Indeed, we find that more than 80% of the STRING interactions have scores above the acceptable cut off of 400 (defined by the authors in the program website).

Next, we asked whether heavily studied proteins are correspondingly covered by good quality interactions in the PPI databases. To address this question, we selected the 10% most popular human genes/proteins from the literature (that is,

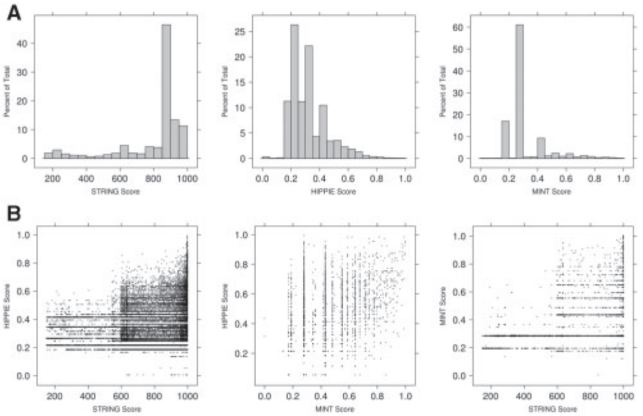

Fig. 3. (**A**) Three databases assign quality scores for protein interactions (HIPPIE, MINT) or functional associations (STRING). MINT and HIPPIE have a stringent quality score based on cumulative evidence from multiple sources and therefore the majority of its interactions have scores below 0.5. STRING on the other hand assigns a high score for proteins that are reported in pathway databases (Szklarczyk *et al.*, 2011). (**B**) Confidence scores of interactions that involve intensively studied proteins. We observed that in general there is no agreement between the database scores, with the exception that among the 31 229 interactions shared between STRING and HIPPIE, 4539 have high confidence score in both databases. In addition, in both comparisons involving STRING, no proteins had high confidence score in MINT or HIPPIE and low confidence score in STRING.

2921 genes/proteins), and ranked them by popularity based on the number of PubMed entries mentioning these genes (Supplementary Table S2); of those, 2790 were present in HIPPIE, 2460 in STRING and 1653 in MINT database.

We performed pair-wise comparisons of the confidence levels of the interactions shared between databases, and that involve the 10% most intensively studied proteins (Fig. 3B; Supplementary Table S3). We observed a lack of agreement between the scores calculated in the databases, i.e. several interactions reported as high confidence in one database are reported as low confidence interactions in the other. In the comparison between STRING and HIPPIE, ~70% of the interactions involving the 10% most studied proteins have a high confidence score in STRING but low confidence score in HIPPIE. On the other hand, we observed that 14% of shared interactions had a score above cut off in both databases. An example is the interaction between TP53 and HMGB1 (Jayaraman *et al.*, 1998), with a score of 0.83 in HIPPIE and 932 in STRING.

As mentioned before, STRING and HIPPIE are derived databases, thus several interactions shared between them were originally reported in MINT. However, each database assign different scores to those interactions, resulting in no correspondence between the scores of different databases. Therefore, to search for tendencies or biases of each scoring scheme, we considered interactions involving at least one popular protein and with conflicting scores between the databases. With these interactions, we created four groups with distinct characteristics (Table 2) and evaluated a sample of 100 interactions (25 from each group), by manually searching experimental evidence supporting these interactions in the scientific literature (Supplementary Table S4).

We observed that a protein association had high confidence score only in STRING (and low scores in the other two databases), and the experimental evidence supporting an association could not be readily identified, reflecting that the scoring scheme used by STRING—assigning a high score to proteins belonging to the same pathway—may be difficult to validate. On the other hand, interactions with high score in MINT or HIPPIE could be confirmed by supporting evidence in one or more publications; although HIPPIE has a very strict scoring scheme: occasionally more than one publication reported an interaction but it still received a low score. Lastly, as part of the iMEX curation guidelines (Orchard *et al.*, 2007), the scoring scheme used by MINT was very accurate: interactions with scores >0.5 could be readily confirmed by manuscripts often containing the identity of both interacting partners in its title and specifically investigating that interaction.

Table 2. Groups of interactions

High score[a]	Low score[b]	Interactions[c]
STRING	HIPPIE	22 177
STRING	HIPPIE and MINT	2225
STRING and HIPPIE	MINT	448
STRING and MINT	HIPPIE	353

[a] High scores considered for STRING, MINT and HIPPIE were values greater than 400, 0.5 and 0.5.

[b] Low scores for STRING, MINT and HIPPIE were values lower than or equal to 400, 0.5 and 0.5.

[c] All interactions included at least one popular protein.

Summarizing, we observed that although there are differences in the calculations of the quality score, interactions that are highly trustable are those that are supported by different experimental systems (especially low-throughput methods), and are manually curated from literature. Ideally, interaction studies should be carried out in different experimental systems to overcome technique-specific bias (Braun *et al.*, 2009; Chen *et al.*, 2010; von Mering *et al.*, 2002).

3.5 Subnetworks based on organ- and cell type-specific expression data

PPI databases are used to address a wide range of questions that span different organisms, cell types, developmental stages and/or phases of the cell cycle. To date, no public PPI database takes these issues into account, with the exception of the HPRD team, which in the long term may also incorporate tissue-specific expression information. Some private companies, e.g. Ingenuity, provide tissue-specific network construction, but as they limit the size of the PPI networks to be on the order of hundreds of nodes, these are not the most suitable tools for the whole network studies. Here, we assessed how the incorporation of organ- and cell-type-specific expression data influence network analysis.

Using a gene expression dataset of 84 human organs and cell types (Su *et al.*, 2004; Wu *et al.*, 2009), we first selected all genes with moderate to high expression levels in each cell type (see Section 2). Next, we evaluated the coverage of each database for the proteins expressed from these genes. STRING and HIPPIE cover ~60% of the organ/cell type-specific proteins, whereas the coverage reaches about 40–50% in the other databases (Supplementary Fig. S4). It is also interesting to note that all databases have a relatively even coverage of all organs and cell types, although the number of genes expressed varies significantly between the different organs/cell types

(Supplementary Fig. S5). For example, ten times more genes are expressed in liver and heart as compared to the ovary; yet, the percent coverage in the PPI databases is comparable for these three organs.

To create organ/cell type-specific PPI networks, we then identified in the PPI database interactions for which both partners are expressed in the same organ/cell type (while eliminating interactions between proteins that are expressed in different organs/cell types). Each organ/cell type subnetwork was then built from the resulting dataset and we included 570 housekeeping proteins that are believed to be expressed in all tissues (Eisenberg and Levanon, 2003). As expected, the resulting organ/cell type-specific subnetworks possess significantly fewer interactions than the original PPI databases (between 1% and 25%) (Supplementary Fig. S6). In addition, these subnetworks are considerably more fragmented than the parent networks, resulting in several smaller connected components (Supplementary Fig. S7). We observed significant differences between the numbers of interactions for organ/cell type-specific subnetworks, which strongly correlated with the number of genes expressed in the respective organ/cell type (Supplementary Fig. S8). For example, more than 6000 different genes are expressed in BDCA dendritic cells (DCs), resulting in a subnetworks that retained 20% of the interactions found in the respective parental PPI databases. In contrast, fewer than 700 genes are expressed in ovary or skin, which reduced the specific subnetworks to just 0.4% of interactions reported in the parental networks (Supplementary Fig. S6).

To assess the potential value of organ/cell type-specific subnetworks, we analyzed the interaction of cellular proteins with two medically relevant human viruses, hepatitis C virus (HCV) and human immunodeficiency virus (HIV). First, we obtained a list of 481 human proteins that interact with HCV proteins (de Chassey *et al.*, 2008) and compared these to the HIPPIE subnetwork created for liver. The HIPPIE database was chosen because it contains a relatively large number of interactions and covers most of the other databases; we focused on the liver subnetwork because of the relevance of this organ in HCV infection (Patrick, 1999).

From the original list of 481 HCV interactors, 98 proteins were present in the liver-specific subnetwork and they interacted with 394 different host proteins (Supplementary Table S5). Comparing the pathway membership of these 492 proteins (interactors and neighbors) with proteins specifically expressed in the liver as a background set, we observed appreciable enrichment in complement and coagulation cascades (*P*-value: 0.04), apoptosis (*P*-value: 2.94e-4), Chemokine signaling pathway (*P*-value: 0.0009) and focal adhesion (*P*-value: 1.03e-7). In contrast, when we used the complete HIPPIE database, 372 of 481 HCV interactors mapped to the database and were involved in 8489 interactions with 3317 different proteins. Using the same analysis that we used for the subnetwork analysis, the HCV interactors and their neighbors fell into many different categories, and no specific pathways or Gene Ontology categories were significantly enriched, making it very difficult to identify critical pathways for the HCV pathogenesis. Hence, organ/cell type-specific subnetworks may aid in the identification of nodes that are critical in specific biological processes.

As a second example of subnetwork analysis, we studied the interaction of HIV with host cells. From the HIV-1 Human Protein Database (Ptak *et al.*, 2008), we obtained a dataset of 1432 host proteins that interact with viral proteins. Next, we created subnetworks containing housekeeping genes and genes expressed in BDCA DCs, CD14+ monocytes and CD4+ T-cells (all datasets were derived from the HIPPIE database). These datasets were chosen since these cell types play critical roles in HIV infections (Dragic *et al.*, 1996; McDonald *et al.*, 2003; Zhu *et al.*, 2002).

From the original list of 1432 cellular proteins that interact with HIV proteins, 72 were exclusively found in the DC subnetwork and had 55 neighbors not present in the other two subnetworks. According to the pathway databases, these proteins are present in the systemic lupus erythematosus pathway (*P*-value: 0.001) and in the B-cell receptor signaling pathway (*P*-value: 0.01). In contrast, 65 cellular HIV interactors were restricted to the CD14+ monocyte subnetwork (interacting with 31 exclusive neighbors), and showed an enrichment for the apoptosis pathway (*P*-value: 0.08), focal adhesion (*P*-value: 0.007) and Fc gamma R-mediated phagocytosis (*P*-value: 0.04). Finally, 58 cellular HIV interactors (and 39 neighbors) were only detected in the CD4+ T-cell subnetwork, with an enrichment for T-cell receptor signaling (*P*-value: 6.8e-5) and primary immunodeficiency pathway (*P*-value: 0.05). These analyses demonstrate cell type-specific interactions between HIV and cellular proteins that may be critical for the infection process. The complete list of cell-specific HIV interactors and neighbors is available in Supplementary Table S6.

4 DISCUSSION

In this study, we compared six widely used public PPI databases for their basic characteristics, neighborhood features and overlap with the other databases analyzed. In addition, we demonstrated that predictions could be significantly improved by the analysis of cell/tissue-specific subnetworks, and by obtaining additional experimental verification for the interaction partners of the most intensively studied genes from the literature.

The six databases compared here have different levels of coverage, in regard to both the number of proteins and the number of PPIs. Nonetheless, they assign similar topological positions to particular proteins within the network; hence, proteins with few or many interaction partners in one database are likely to have few or many interaction partners in the other databases analyzed. However, the identity of these interaction partners may differ between the databases, resulting in great uncertainty in model building. These differences reflect the differences in the algorithms, portion of literature curated by the different groups (Turinsky *et al.*, 2010) and the experimental techniques used to build the databases.

Many PPI datasets are generated by expressing the two proteins of interest in one cell (for example, in the yeast two-hybrid system). In such *in vitro* assays, proteins may be co-expressed and interact, but in reality their expression may be dependent on cell type, different experimental stages and/or during different phases of the cell cycle/organism development. As a result, the currently available PPI databases are believed to contain a significant percentage of false positive entries (Deane *et al.*, 2002). To address this weakness, PPI databases could be combined with the increasing number of transcriptomics or proteomics datasets that assess the expression of genes or proteins in a specific organ, cell type, developmental or cell cycle stage. We here provide two examples that demonstrate the potential of this approach.

In one example, we show that the host cellular interaction partners of HCV proteins are not enriched for particular Gene Ontology categories or pathways in an analysis based on the entire HIPPIE database; in contrast, three KEGG pathways (apoptosis, focal adhesion, complement and coagulation cascades) are highly enriched when the HIPPIE database was analyzed in combination with a liver-specific gene expression dataset. Regulation of apoptosis may play a critical role in HCV infection to establish chronic or persistent infections (Bantel and Schulze-Osthoff, 2003). Activation of the complement and coagulation pathways has been described for HCV infections (Ueda *et al.*, 1993), and it was verified that hepatic inflammation can be reduced by administering CD55, a regulator of the complement pathway (Chang *et al.*, 2009). However, the significance of proteins involved in focal adhesion for HCV infections is currently not known, which may be addressed in further investigations. This example demonstrates how the generation of subnetworks may help in the prioritization of pathways for future studies.

In the second example, we show that each cell type subnetwork has exclusive proteins that interact with HIV. Among the exclusive proteins from each cell type are some representing critical processes studied and validated experimentally. Apoptosis induced by HIV proteins was reported to be a critical aspect of its pathogenicity (Castedo *et al.*, 2002; Rasola *et al.*, 2001; Zheng *et al.*, 2007). Cases of patients with concomitant systemic lupus erythematosus and HIV have been reported (Calza *et al.*, 2003; Gould and Tikly, 2004), and the interplay between autoimmune diseases and retroviruses is an active topic of research (Balada *et al.*, 2010). In addition, the association between HIV-infection and the downregulation of Fc-gammaR-mediated phagocytosis in HIV-infected macrophages was observed (Kedzierska *et al.*, 2002).

Some studies have generated subnetworks to address medical questions. In one example, subnetworks from normal and cancer cells have been established to identify PPIs that are characteristic of cancer development and could be targeted to 'rewire' these cells (Quayle *et al.*, 2007). In the context of a metabolic study, the creation of tissue-specific subnetworks helped to elucidate post-transcriptional regulation of genes from 10 different tissues that are involved in metabolic diseases (Shlomi *et al.*, 2008). Collectively, these and our own analyses demonstrate that cell/tissue-specific subnetworks can be used to increment the biological relevance of PPI datasets.

Our analysis also revealed that current databases possess many interactions that are characterized by low confidence scores, a finding that is of particular concern for intensively studied proteins. While it is not feasible to verify all predicted interactions with different techniques, we suggest here focusing PPI evaluation efforts on the verification of low confidence interactions of selected proteins widely used in research models but lacking high confidence interactions. Toward this goal, we created a priority list of interactions that include highly investigated proteins such as TP53 (described earlier), MAPK1 (mitogen-activated protein kinase 1), BCL2 (B-cell CLL/lymphoma 2) or TNF (tumor necrosis factor F), among many others. Additional experimental data confirming or revealing new interactions of these 'key players' with their predicted cellular interaction partners will push PPI databases a step closer to becoming a reliable, daily-use tool for researchers, in the same way sequence analysis and protein structure databases already are.

ACKNOWLEDGEMENTS

We are grateful to the skillful and critical reading of the manuscript by Amie Eisfeld-Fenney and fruitful discussions with Takeshi Hase.

Funding: Japanese Science and Technology Agency (JST), Project ERATO Kawaoka, by the German Ministry of Education and Research (BMBF) [grant number 01GS08170]; the Helmholtz Alliance in Systems Biology (Germany).

Conflict of Interest: none declared.

REFERENCES

Albert,R. *et al.* (2000) Error and attack tolerance of complex networks. *Nature*, **406**, 378–382.

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

Balada,E. *et al.* (2010) Implication of human endogenous retroviruses in the development of autoimmune diseases. *Int. Rev. Immunol.*, **29**, 351–370.

Bantel,H. and Schulze-Osthoff,K. (2003) Apoptosis in hepatitis C virus infection. *Cell Death Differ.*, **10**, S48–S58.

Bhattacharya,B. *et al.* (2004) Gene expression in human embryonic stem cell lines: unique molecular signature. *Blood*, **103**, 2956–2964.

Braaksma,M. *et al.* (2011) Metabolomics as a tool for target identification in strain improvement: the influence of phenotype definition. *Microbiology*, **157**, 147–159.

Braun,P. *et al.* (2009) An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods*, **6**, 91–97.

Breitkreutz,B.J. *et al.* (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res.*, **36**, D637–D640.

Calza,L. *et al.* (2003) Systemic and discoid lupus erythematosus in HIV-infected patients treated with highly active antiretroviral therapy. *Int. J. STD AIDS*, **14**, 356–359.

Castedo,M. *et al.* (2002) Sequential involvement of Cdk1, mTOR and p53 in apoptosis induced by the HIV-1 envelope. *EMBO J.*, **21**, 4070–4080.

Ceol,A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Chang,M.-L. *et al.* (2009) Hepatic inflammation mediated by hepatitis C virus core protein is ameliorated by blocking complement activation. *BMC Med. Genomics*, **2**, 51.

Chen,Y.C. *et al.* (2010) Exhaustive benchmarking of the yeast two-hybrid system. *Nat. Methods*, **7**, 667–668.

Coulomb,S. *et al.* (2005) Gene essentiality and the topology of protein interaction networks. *Proc. Biol. Sci.*, **272**, 1721–1725.

de Chassey,B. *et al.* (2008) Hepatitis C virus infection protein network. *Mol. Syst. Biol.*, **4**, 230.

Deane,C.M. *et al.* (2002) Protein interactions. *Mol. Cell. Proteomics*, **1**, 349–356.

Dragic,T. *et al.* (1996) HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature*, **381**, 667–673.

Eisenberg,E. and Levanon,E.Y. (2003) Human housekeeping genes are compact. *Trends Genet.*, **19**, 362–365.

Ewing,R.M. *et al.* (2007) Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.*, **3**, 89.

Gentleman,R. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Gould,T. and Tikly,M. (2004) Systemic lupus erythematosus in a patient with human immunodeficiency virus infection – challenges in diagnosis and management. *Clin. Rheumatol.*, **23**, 166–169.

Han,J.D. *et al.* (2004) Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, **430**, 88–93.

Hase,T. *et al.* (2009) Structure of protein interaction networks and their implications on drug design. *PLoS Comput. Biol.*, **5**, e1000550.

Huang,D.W. *et al.* (2008) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protocols*, **4**, 44–57.

Jayaraman,L. *et al.* (1998) High mobility group protein-1 (HMG-1) is a unique activator of p53. *Genes Dev.*, **12**, 462–472.

Jensen,L.J. *et al.* (2009) STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, **37**, D412–D416.

Jeong,H. *et al.* (2001) Lethality and centrality in protein networks. *Nature*, **411**, 41–42.

Joy,M.P. *et al.* (2005) High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.*, **2005**, 96–103.

Kamburov,A. *et al.* (2011) ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res.*, **39**, D712–D717.

Kanehisa,M. *et al.* (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res.*, **38**, D355–D360.

Kedzierska,K. *et al.* (2002) HIV-1 Down-modulates γ signaling chain of FcγR in human macrophages: a possible mechanism for inhibition of phagocytosis. *J. Immunol.*, **168**, 2895–2903.

Krogan,N.J. *et al.* (2006) Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. *Nature*, **440**, 637–643.

Mathivanan,S. *et al.* (2006) An evaluation of human protein-protein interaction data in the public domain. *BMC Bioinformatics*, **7** (Suppl. 5), S19.

McDonald,D. *et al.* (2003) Recruitment of HIV and its receptors to dendritic cell-T cell junctions. *Science*, **300**, 1295–1297.

Orchard,S. *et al.* (2007) Submit your interaction data the IMEx way: a step by step guide to trouble-free deposition. *Proteomics*, **7** (Suppl. 1), 28–34.

Patil,A. and Nakamura,H. (2006) Disordered domains and high surface charge confer hubs with the ability to interact with multiple proteins in interaction networks. *FEBS Lett.*, **580**, 2041–2045.

Patrick,M. (1999) Hepatitis C: the clinical spectrum of the disease. *J. Hepatol.*, **31**, 9–16.

Prasad,T.S. *et al.* (2009) Human Protein Reference Database and human proteinpedia as discovery tools for systems biology. *Methods Mol. Biol.*, **577**, 67–79.

Ptak,R.G. *et al.* (2008) Cataloguing the HIV type 1 human protein interaction network. *AIDS Res. Hum. Retroviruses*, **24**, 1497–1502.

Quayle,A.P. *et al.* (2007) Perturbation of interaction networks for application to cancer therapy. *Cancer Inform.*, **5**, 45–65.

Ramirez,F. *et al.* (2007) Computational analysis of human protein interaction networks. *Proteomics*, **7**, 2541–2552.

Rasola,A. *et al.* (2001) Apoptosis enhancement by the HIV-1 Nef protein. *J. Immunol.*, **166**, 81–88.

Rual,J.F. *et al.* (2005) Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, **437**, 1173–1178.

Shlomi,T. *et al.* (2008) Network-based prediction of human tissue-specific metabolism. *Nat. Biotechnol.*, **26**, 1003–1010.

Stumpf,M.P.H. *et al.* (2008) Estimating the size of the human interactome. *Proc. Natl Acad. Sci. USA*, **105**, 6959–6964.

Su,A.I. *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl Acad. Sci. USA*, **101**, 6062–6067.

Szklarczyk,D. *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Res.*, **39**, D561–D568.

Turinsky,A.L. *et al.* (2010) Literature curation of protein interactions: measuring agreement across major public databases. *Database*, **2010**, Available at http://database.oxfordjournals.org/content/2010/baq026.long.

Ueda,K. *et al.* (1993) The association between hepatitis C virus infection and in vitro activation of the complement system. *Ann. Clin. Biochem.*, **30** (Pt 6), 565–569.

Venkatesan,K. *et al.* (2009) An empirical framework for binary interactome mapping. *Nat. Methods*, **6**, 83–90.

von Mering,C. *et al.* (2002) Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, **417**, 399–403.

von Mering,C. *et al.* (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.*, **33**, D433–D437.

Wilhelm,B.T. *et al.* (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

Wu,C. *et al.* (2009) BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol.*, **10**, R130.

Yildirim,M.A. *et al.* (2007) Drug-target network. *Nat. Biotechnol.*, **25**, 1119–1126.

Zheng,L. *et al.* (2007) HIV Tat protein increases Bcl-2 expression in monocytes which inhibits monocyte apoptosis induced by tumor necrosis factor-alpha-related apoptosis-induced ligand. *Intervirology*, **50**, 224–228.

Zhu,T. *et al.* (2002) Evidence for human immunodeficiency virus type 1 replication in vivo in CD14+ monocytes and its potential role as a source of virus in patients on highly active antiretroviral therapy. *J. Virol.*, **76**, 707–716.