# PROTEOMICS

# Supporting Information for Proteomics DOI 10.1002/pmic.200600924

Fidel Ramírez, Andreas Schlicker, Yassen Assenov, Thomas Lengauer and

Mario Albrecht

Computational analysis of human protein

interaction networks

 $\ensuremath{\textcircled{}^{\circ}}$  2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

#### **Online Supplement**

### **Computational Analysis of Human Protein Interaction Networks**

Fidel Ramírez, Andreas Schlicker, Yassen Assenov, Thomas Lengauer, and Mario Albrecht

#### **Supplementary Notes**

#### Mapping of gene and protein identifiers

The diverse protein and gene identifiers in the interaction datasets were converted to Entrez Gene IDs using mapping tables from different sources (Tables S1 and S3). Since the mapping of protein to gene identifiers appears counter-intuitive at first glance, it is important to note that several datasets already provided Entrez Gene IDs and the reverse mapping from gene identifiers to protein identifiers would have caused an undesirable expansion of the number of PPIs due to protein fragments and splicing variants with separate protein identifiers in UniProtKB. Therefore, we chose a conservative approach by mapping protein to gene identifiers and removed duplicates. Identifiers referring to non-human genes or proteins were omitted during this process.

#### Quality assessment using recall-precision plots

As an alternative to the LR plots in Figure 3, Figure S4 shows the corresponding recall-precision (RP) plots. Apparently, the precision values obtained by using the combined Y2H datasets CCSB-H11 and MDC as PRS appear much lower than the corresponding values obtained using HPRD-SS as PRS, which is probably due to a considerable rate of false positives in the Y2H screens. In the RP plot using HPRD-SS as PRS, both HiMAP and OPHID datasets have much higher precision and recall than datasets adjacent to them in the RP plot using the combined Y2H datasets as PRS. In the case of HiMAP, it may be biased towards HPRD-SS because a previous release of HPRD was originally used to evaluate the predicted PPIs of HiMAP [1]. Remarkably, the precision of Bioverse-core and HiMAP-core is larger than that of HPRD-SS in the RP plot using the combined Y2H datasets as PRS. These results suggest that predicted PPIs can be quite reliable. The precisions of the two Y2H datasets using HPRD-SS as PRS are significantly different (CCSB-HI1 with 0.389 versus MDC with 0.145).

#### Comparison of the datasets CCSB-HI1 and MDC

The experimental Y2H datasets CCSB-HI1 and MDC share 201 proteins, but they overlap in solely 17 interactions involving 22 proteins (Table S10). The 17 PPIs exhibit a high average *BPscore* (0.810), and 11 of them are present in HPRD-SS, but none in our NRS set. Remarkably, 52 interactions of the 201 shared proteins are present only in the CCSB-HI1 set, of which 9 are also found in HPRD-SS and 1 in our NRS set. In comparison, the same 201 proteins participate in 123 PPIs contained only in the MDC set, where 6 PPIs are in HPRD-SS and 5 in our NRS set. Furthermore, while CCSB-HI1 has an average *BPscore* of 0.464 and contains 250 (16.0%) DDI-validated interactions (3.85%), 11 of which are in the small overlap of both datasets (Tables 2 and S10). Similarly, regarding the interactions of the 201 shared proteins, whereas MDC without CCSB-HI1 achieves only a *BPscore* of 0.446 and 8.89% DDI-validated interactions. Additionally, we analyzed six different subsets of the MDC dataset based on a confidence score defined by the original authors (Table S12) [2]. Regarding the MDC subset that corresponds to interactions with a confidence score equal or greater than 4 (339 interactions), the *BPscore* and LR are similar to that of CCSB-HI1.

#### Comparison of MDC subsets based on the length of protein fragments

We found that most proteins in the overlap of CCSB-HI1 and MDC are full-length proteins (average fragment length is 96.97% of the complete protein length). To further analyze the potential effect of the fragment length on the quality of the MDC dataset, we partitioned this dataset by length (Table S11). Interestingly, the proportion of PPIs contained in the PRS set using HPRD-SS rises with increasing fragment length, while the overlap with the NRS set does not change much. The best LR is obtained when using a relative fragment length of 90% or higher of the complete protein sequence. The fraction of DDI-validated protein interactions increases slightly with larger fragment lengths, but the average *BPscore* is not affected significantly. Regarding the MDC subsets based on the confidence score defined by the original authors (Table S12) [2], an increase of the score is observed with rising average fragment length. However, the increase in fragment length is just slightly larger than the average 78.92% for the whole MDC dataset. In conclusion, the fact that MDC used protein fragments and CCSB-HI1 full-length proteins does not appear to explain the differences of the assessment results between MDC and CCSB-HI1.

#### Comparison of HPRD and IntAct by the number of publications and the experimental technique

To analyze the reliability of the protein interactions in the literature-curated datasets further, we subdivided the interactions in HPRD and IntAct by the number of publications reporting them and by the experimental technique (Table S13). The in vivo and in vitro classifications of HPRD obtain similar scores in all assessments, indicating that their reliabilities are similar. It is also apparent that the more publications support a protein interaction, the higher are its scores. Interestingly, PPIs derived from protein arrays have the highest *BPscore* (0.811) and do not overlap with the NRS. However, this could be misleading because all those PPIs come from the same publication [3]. Moreover, as expected, X-ray crystallography returns a very high number of DDI-validated PPIs (85.29%). In contrast, the protein interactions derived from tandem affinity purification (TAP) have the lowest overlap with the combined Y2H datasets (only 1 interaction), and the number of DDIvalidated interactions is the smallest (5.25%). Furthermore, the Y2H interactions contained in HPRD and IntAct have a BPscore and fraction of DDI-validated interactions similar to that of the Y2H dataset CCSB-HI1. Datasets such as HPRD in vivo and in vitro listed in Table S13 have higher BPscore and number of DDI-validated interactions, but lower LR (using the combined CCSB-HI1 and MDC datasets as PRS) than those Y2H interactions in HPRD and IntAct. This might be explained by the idea that Y2H screens can detect interactions not found by other methods [4].

#### Protein interaction predictions based on high-throughput data

PPIs in predicted human datasets are primarily derived from interologs using high-throughput data (Table S2). For instance, the DIP database used by Bioverse and POINT includes a large portion of 80% interologous PPIs detected by high-throughput experiments [5]. A similar portion is contained in MINT so that only 6% PPIs in HomoMINT are derived from small-scale experiments [6]. Our assessments have also shown that predicted datasets such as Sanger derived solely from highthroughput experiments perform similar to Y2H screens. Other predictions such as HomoMINT, OPHID, and POINT that utilized many high-throughput interologs and relatively few from smallscale experiments score only slightly better. The Sanger-core dataset, which is based on interologs reported in more than one publication, achieves higher assessment scores than the Sanger dataset. However, the Sanger-core values of BPscore, DDI-validation, and LR assessment are still similar to those of HomoMINT, OPHID, and POINT. The HiMAP datasets, which do not only rely on interologs, achieve better performance in our assessments. The outstanding scores of Bioverse-core may be due to the inclusion of PPIs from X-ray crystallography and, in contrast to Bioverse, due to the application of a stringent sequence similarity threshold for establishing orthology. Therefore, our results suggest that predictions based on interolog mapping can be as good as the original data used to derive them and even better if appropriate filters and methods are additionally employed.

#### **Supplementary Figure Legends**

**Figure S1.** Dataset comparison using boxplots based on GO biological processes annotated to interacting proteins. The datasets are ordered by the *BPscore* median from left to right. The area of each box is proportional to the size of the corresponding dataset.

**Figure S2.** Dataset comparison using histograms of the *BPscore* distribution. The *BPscore* similarity values based on the biological processes annotated to interacting proteins are binned in 0.1-steps.

**Figure S3.** 2D histograms of the distribution of PPIs according to the biological processes annotated to the interacting proteins in each human interaction dataset. Every dataset is depicted by a triangle matrix whose axes represent top levels of the GO hierarchy. For each matrix cell, a protein interaction density was calculated as the ratio of the number of PPIs assigned to the respective matrix cell divided by the total number of PPIs possibly formed by the proteins annotated in the respective GO categories. The dot color in the histograms reflects the protein interaction density as observed PPIs per 1,000 possible PPIs. The protein interaction density is not shown if the observed number of PPIs is non-significant (p-value  $\geq 0.01$ , using Fisher's exact test as in case of the overlap computation). The numbers along the axes represent the following GO categories: 1: cellular process; 2: cell communication; 3: cell differentiation; 4: cellular physiological process; 5: amino acid and derivative metabolism; 6: cell death; 7: cell motility; 8: electron transport; 9: nucleobase, nucleoside, nucleotide and nucleic acid metabolism; 10: transport; 11: development; 12: physiological process; 13: metabolism; 14: biosynthesis; 15: catabolism; 16: macromolecule metabolism; 17: secretion; 18: regulation of biological process; 19: response to stimulus; 20: behavior.

**Figure S4.** Recall-precision plots using (A) HPRD-SS or (B) the combined Y2H datasets as PRS set. While recall equals the computed true positive rate (TPR =  $|E_i \cap \text{PRS}| / |\text{PRS}|$ ), precision is calculated by the following formula:  $|E_i \cap \text{PRS}| / (|E_i \cap \text{PRS}| + |E_i \cap \text{NRS}|)$ .

**Figure S5.** Plots of the degree distribution together with the exponent  $\gamma$  of the fitted power law for each human interaction dataset.

**Figure S6.** 2D histograms of the degree distribution for each dataset. The dot colors in the histograms reflect the absolute frequency of two interacting proteins with specific degrees (numbers of interactions). The maximum frequency max=N [X,Y] with degrees X and Y is given above the histogram of each dataset.

**Figure S7.** 2D histograms of the distribution of PPIs according to the length of interacting proteins binned in steps of 50 amino acids. Every human interaction dataset is depicted by a triangle matrix whose axes represent the sequence lengths of interacting proteins. For each matrix cell, a protein interaction density was calculated as the ratio of the number of PPIs assigned to the respective matrix cell divided by the total number of PPIs possibly formed by the proteins members of the bin. The dot color in the histograms reflects the protein interaction density as observed PPIs per 1,000 possible PPIs. The protein interaction density is not shown if the observed number of PPIs is nonsignificant (*p*-value  $\geq 0.01$ , using Fisher's exact test as in case of the overlap computation). The matrix entitled 'MDC fragment length' was derived using the actual lengths of the protein fragments as used in the Y2H screen in contrast to the matrix 'MDC' whose proteins lengths belong to complete protein sequences as in case of all other datasets.

Figure S8. Color version of Figure 3.

Figure S1.



#### Figure S2.





Protein interaction density

Figure S4.



# Figure S5.



#### Figure S6.





Figure S7.



Protein interaction density

Figure S8.



#### **Supplementary Tables**

**Table S1.** Datasets of human protein-protein interactions included in our analysis and their conversion to Entrez Gene IDs. The number of interactions and identifiers of a certain type contained in the original datasets were obtained after removal of duplicates. The number of mappable identifiers is the number of original identifiers for which corresponding Entrez Gene IDs were found in the identifier mapping tables (Table S3). The numbers of final Entrez Gene IDs and interactions refer to the number of unique identifiers and interactions after performing the identifier mapping.

Dataset	Original identifier type	#Original interactions	#Original identifiers	#Mappable identifiers	#Final Entrez Gene IDs	#Final interactions
		Predicted pr	rotein-protein inter	actions		
Bioverse	RefSeq	3218048	36996	16388	7711	233941
Bioverse-core	RefSeq	18327	1753	1481	1263	3266
Himap	Entrez Gene ID	38379	5790	5790	5790	38378
HiMAP-core	Entrez Gene ID	8833	2901	2901	2901	8832
HomoMINT	UniProtKB	10993	4129	4101	4184	10870
OPHID	UniProtKB	26425	4787	4738	4559	28255
POINT	GenInfo Identifier	101783	13047	12982	12058	98528
Sanger	Ensembl (gene)	71806	6231	5788	5923	67518
Sanger-core	Ensembl (gene)	11652	3872	3661	3728	11131
		Experimental Y2	H protein-protein?	interactions		
CCSB-HI1	Entrez Gene ID	2754	1549	1549	1549	2754
MDC	Entrez Gene ID	2124	1124	1124	1124	2033
		Literature-curate	ed protein-protein	interactions		
HPRD-LS	Entrez Gene ID	3151	1983	1983	1983	3151
HPRD-SS	Entrez Gene ID	27955	7686	7686	7686	27955
IntAct	UniProtKB	6734	3484	2977	2988	5809

**Table S2.** Comparison of interolog mapping methods for each predicted dataset regarding data sources of the protein interactions, species of the data sources, and homology detection methods applied. The listed data sources refer to the following studies: Gavin: *S. cerevisiae* TAP purified complexes [7]; Giot: *D. melanogaster* Y2H screen [8]; Ho: *S. cerevisiae* HMS-PCI purified complexes [9]; Ito: *S. cerevisiae* Y2H screen [10]; Li: *C. elegans* Y2H screen [11]; Suzuki: *M. musculus* Y2H screen [12, 13]; Tong: *S. cerevisiae* synthetic genetic array [14]; Uetz: *S. cerevisiae* Y2H screen [15]; von Mering *et al.* grouped the *S. cerevisiae* data from Gavin, Ho, Ito, Uetz, Tong, and added 7,446 predicted interactions derived from gene neighborhood, gene fusion, and co-occurrence of genes [16]; further databases are: DIP for Bioverse [17], DIP for POINT [18], GRID [19], MINT [20], MIPS [21], and PDB [22].

Dataset	Data Sources	Species	Homology Detection Method
Bioverse	DIP, GRID, PDB	50 species	PSI-BLAST (E-value < 1.0)
Himap	Gavin, Giot, Ito, Li, Uetz	fruit fly, worm, yeast	InParanoid
HomoMINT	MINT	15 species	InParanoid, results subsequently filtered by matching domain architecture between human and species orthologs
OPHID	Giot, Li, Suzuki, von Mering, MIPS	fruit fly, worm, yeast	BLASTP reciprocal best-hit (E-value < 10 <sup>-5</sup> ), filtered for hits with length > 50% of human protein sequence
POINT	DIP	fruit fly, worm, yeast, mouse	BLASTP (E-value not given)
Sanger	Gavin, Giot, Ho, Ito, Li, Tong, Uetz, von Mering	fruit fly, worm, yeast	InParanoid

**Table S3.** List of files used to map between different database identifiers.

Source	Web Link	Version	Download Date
NCBI	ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2ref		08 November 2005
NCBI	ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/human.protein.gpff		10 October 2005
Ensembl	http://www.ensembl.org/Multi/martview/by4VYrpPEn.mart	33	27 September 2005
HGNC	http://www.gene.ucl.ac.uk/cgi-bin/nomenclature/gdlw.pl		27 September 2005
IPI	ftp://ftp.ebi.ac.uk/pub/databases/IPI/current/ipi.HUMAN.xrefs.gz	3.10	27 September 2005
UniProtKB	ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/ complete/uniprot_trembl.dat.gz		15 September 2005
UniProtKB	ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/ complete/uniprot_sprot.dat.gz		15 September 2005

# Table S4. List of publications reporting large numbers of PPIs included in HPRD or IntAct.

Publication	Experimental method	HPRD	IntAct
Nakayama <i>et al.</i> , 2002 [23]	Y2H	118	125
Bouwmeester et al., 2004 [24]	Y2H	128	1682
Colland <i>et al.</i> , 2004 [25]	Y2H	706	—
Goehler et al., 2004 [26]	Y2H	154	151
Jin <i>et al.</i> , 2004 [27]	Co-immunoprecipitation	297	—
Lehner <i>et al.</i> , 2004 [28]	Y2H	110	95
Lehner et al., 2004 [29]	Y2H	264	231
Ramachandran et al., 2004 [3]	Protein array	102	109
Barrios-Rodiles et al., 2005 [30]	LUMIER	430	—
Guo et al., 2005 [31]	Far-western blotting	75	—
Rual <i>et al.</i> , 2005 [32]	Y2H	2619	2671
Stelzl et al., 2005 [2]	Y2H	3116	3137
Lim et al., 2006 [33]	Y2H	704	706
Tsang et al., 2006 [34]	Y2H	75	—
Camargo et al., 2007 [35]	Y2H	_	191

**Table S5**. Pairwise overlap of the human interaction datasets. The absolute number of PPIs and proteins shared by two datasets are given in the top and bottom of each table cell, respectively.

	Bioverse	Bioverse core	Himap	HiMAP- core	HomoMINT	OPHID	POINT	Sanger	Sanger- core	CCSB-HI1	MDC	HPRD-LS	HPRD-SS	IntAct	HPRD- Random	Random
Bioverse	233941	3266	5032	2261	1968	3633	6435	5275	1178	117	29	149	4552	627	480	53
Biotelice	7711	1263	3856	2150	2308	2785	5544	3187	2079	793	620	1159	4732	1842	2878	2341
Bioverse-core		3266	674	313	139	949	714	324	240	29	7	23	1088	198	76	0
		1263	876	620	446	881	1047	551	415	139	139	230	1134	533	479	412
НіМАР			38378	8832	1680	1827	2815	2915	1296	59	14	57	1454	225	21	5
			5790	2901	2193	2445	4464	2828	2010	638	505	916	4026	1552	2173	1738
HiMAP-core				8832	765	868	1395	1635	907	48	13	41	767	148	4	1
				2901	1330	1486	2411	1703	1275	368	277	543	2254	979	1098	886
Homo-MINT					10870	5532	7161	6143	1867	64	18	33	806	159	107	3
					4184	3096	3895	3140	2168	566	458	768	2452	1316	1572	1298
OPHID						28255	6581	16477	3486	37	14	447	1301	253	10	3
						4559	4092	3054	2173	575	494	936	2841	1455	1726	1386
POINT							98528	9393	2858	154	39	105	3616	494	573	27
-							12058	4961	3237	1105	839	1602	5830	2440	4592	3638
Sanger								67518	11131	73	22	53	816	191	97	12
g								5923	3728	727	533	958	3130	1448	2229	1811
Sanger-core									11131	52	13	34	530	122	41	2
<u> </u>									3728	538	387	663	2102	1042	1402	1168
CCSB-HI1										2754	17	23	146	46	21	1
										1549	201	395	896	588	568	463
MDC											2033	8	41	16	7	1
											1124	275	685	374	404	321
HPRD-LS												3151	124	1416	15	2
												1983	1339	1316	685	594
HPRD-SS													27955	1543	444	3
													7686	2261	2825	2291
IntAct														5809	47	1
														2988	1062	902
HPRD-Random															30956	10
															6103	1812
Random																30000
																5000

	Bioverse	Bioverse- core	Himap	HiMAP- core	HomoMINT	OPHID	POINT	Sanger	Sanger- core	CCSB-HI1	MDC	HPRD-LS	HPRD-SS	IntAct	HPRD- Random	Random
Bioverse	0	0	0	0	0	0	0	0	0	4.62E-55	3.14E-04	1.36E-76	0	0	1.35E-134	1.00
Bioverse-core		0	0	0	1.44E-179	0	0	0	0	3.34E-21	1.79E-02	5.56E-16	0	2.05E-289	2.38E-66	1.00
Himap			0	0	0	0	0	0	0	1.97E-43	4.34E-05	3.47E-39	0	5.38E-230	0.82	1.00
HiMAP-core				0	0	0	0	0	0	6.13E-41	2.93E-06	2.36E-32	0	3.61E-169	0.97	1.00
HomoMINT					0	0	0	0	0	1.02E-68	2.09E-12	1.71E-26	0	3.65E-196	1.65E-75	1.00
OPHID						0	0	0	0	5.69E-20	1.44E-04	0	0	1.76E-267	0.998413	1.00
POINT							0	0	0	2.50E-192	6.47E-32	7.48E-108	0	0	0	0.87
Sanger								0	0	1.29E-43	1.24E-06	3.72E-24	0	5.84E-143	2.03E-12	1.00
Sanger-core									0	2.35E-47	8.13E-07	1.72E-25	0	1.41E-132	6.40E-14	1.00
CCSB-HI1										0	1.34E-11	2.85E-18	3.64E-207	3.31E-45	1.04E-10	0.99
MDC											0	2.53E-04	1.76E-41	4.06E-11	2.06E-02	0.97
HPRD-LS												0	5.78E-160	0	5.11E-06	0.95
HPRD-SS													0	0	0	1.00
IntAct														0	1.69E-27	1.00
HPRD-random															0	1.00
Random																0

**Table S6.** Computed *p*-values using Fisher's exact test to evaluate the overlap sizes of the interaction sets in Table S5.

**Table S7.** Quality assessment using functional GO similarity and structural domain interactions. The consensus sets are ranked by the average *BPscore*. For each set, the percentage of interactions with biological process (BP) terms assigned to both interacting proteins is given next to the percentage of homodimers, the fraction of protein self-interactions. The average information content is calculated from the information content of the BP protein annotations. The percentage of DDI-validated interactions using iPfam relates to the fraction of PPIs in which both proteins have Pfam domain assignments. The rightmost column shows the overlap size of the subset of PPIs with a *BPscore*  $\geq$  0.8 and the subset of all DDI-validated PPIs, relative to the size of the union of both subsets.

		Functional Sin	nilarity using G	0	Domain Interac	tions using iPfam	
Dataset	Average BPscore	GO assignments in percent	Number of homodimers in percent	Average information content	Domain assignments in percent	Number of DDI-validated interactions (%)	Overlap in percent
ConSet6	0.666	82.23	0.00	12.2	96.28	103 (22.1)	41.40
ConSet5	0.614	80.13	0.00	12.1	93.93	277 (18.8)	36.33
ConSet3	0.538	74.19	1.85	12.5	83.45	1585 (17.5)	33.72
ConSet4	0.535	74.53	0.48	12.3	85.74	614 (15.1)	34.75
ConSet2	0.533	76.46	3.17	12.6	86.69	6744 (20.3)	37.47

**Table S8.** Quality assessment using likelihood ratios. The human interaction datasets are ranked by decreasing LR. The ratios TPR, FPR, and LR are computed using PRS and NRS sets, and the PRS set consists either of HPRD-SS or of the combined Y2H datasets. The number of PPIs in the overlap of the respective dataset with the PRS or NRS sets, their average *BPscore*, and the percentage of DDI-validated PPIs are also listed.

	Computed ratios		S	Number in ov	of PPIs erlap	Average Bl PPIs in c	Pscore of overlap	DDI-validated PPIs in overlap in percent		
Dataset	TPR	FPR	LR	PRS	NRS	PRS	NRS	PRS	NRS	
			Use of H	IPRD-SS d	lataset as	PRS set				
ConSet6	0.0033	5.00E-07	6561.2	92	8	0.849	0.251	47.19	12.5	
ConSet5	0.0081	2.80E-06	2878.1	227	45	0.848	0.334	49.77	2.44	
ConSet3	0.0370	3.00E-05	1236.8	1034	477	0.799	0.316	52.81	2.84	
ConSet4	0.0160	1.30E-05	1218.8	455	213	0.833	0.311	52.91	2.13	
ConSet2	0.0960	9.80E-05	987.2	2694	1557	0.774	0.319	52.26	5.06	
		L	lse of com	bined Y2H	l datasets	as PRS set				
ConSet6	0.0013	5.00E-07	2507.8	6	8	0.869	0.251	33.33	12.50	
ConSet5	0.0044	2.80E-06	1560.4	21	45	0.828	0.334	50.00	2.44	
ConSet4	0.0092	1.30E-05	690.7	44	213	0.788	0.311	48.65	2.13	
ConSet3	0.0170	3.00E-05	574.8	82	477	0.805	0.316	55.71	2.84	
ConSet2	0.0310	9.80E-05	322.1	150	1557	0.836	0.319	61.54	5.06	

**Table S9.** Topological network parameters for each human protein interaction dataset. The degree and clustering coefficient distributions are fitted to power laws with exponents  $\gamma$ .

Dataset	Average number of neighbors	Maximum number of neighbors	$\gamma$ of degree distribution	Network diameter	Average shortest path length	Average clustering coefficient	γ of clustering coefficient distribution
Bioverse	60.24	842	-1.1887	10	3.5035	0.4801	-0.1845
Bioverse-core	4.67	34	-1.7635	17	6.3159	0.5029	0.2160
Himap	13.26	159	-1.7441	18	5.1591	0.4401	-0.0965
HiMAP-core	6.09	44	-1.7982	26	9.3950	0.3156	0.1253
HomoMINT	4.95	68	-2.0799	12	4.9153	0.0650	-0.4486
OPHID	12.39	192	-1.4260	18	4.5375	0.1885	0.0904
POINT	16.26	522	-1.6927	10	3.5284	0.0889	-0.3508
Sanger	22.69	365	-1.4090	10	3.8715	0.2342	0.0039
Sanger-core	5.87	75	-1.8402	20	6.4511	0.1861	0.3704
CCSB-HI1	3.43	129	-1.5637	12	4.3581	0.0626	-0.7932
MDC	3.58	95	-1.5149	12	4.6248	0.0205	-0.8197
HPRD-LS	3.13	213	-1.3000	10	4.4327	0.0602	-1.1700
HPRD-SS	6.78	202	-1.8420	15	4.4627	0.1276	-0.4830
IntAct	3.83	181	-1.4450	18	5.1542	0.1022	-0.7890
Random	11.99	30	_	6	3.6986	0.0026	_

**Table S10.** Comparison of the CCSB-HI1 and MDC datasets using *BPscore*, structural DDI-validation, and the likelihood ratio LR (see supplementary note and legends of Tables 2 and 3). Here, solely the interactions of the 201 proteins shared by CCSB-HI1 and MDC are analyzed. The average (avg.) fragment length in percent refers to the peptide length used in the Y2H experiments of MDC relative to the complete protein sequence.

					Number in ov	Ave BPs of P ove	erage score Pls in erlap	DDI-validated PPIs in overlap in percent		
Dataset	Inter- actions	Average BPscore	DDI-validated interactions in percent	LR	PRS (avg. fragment length %)	NRS (avg. fragment length %)	PRS	NRS	PRS	NRS
Overlap of CCSB-HI1 and MDC	17	0.810	68.75		11 (96.97)	0	0.903	_	72.73	
CCSB-HI1 without MDC	52	0.714	33.33	5134.8	9	1	0.998	_	75.00	l
MDC without CCSB-HI1	123	0.446	8.89	684.6	6 (76.73)	5 (73.90)	0.999	0.437	83.33	_

**Table S11.** Comparison of MDC subsets based on the length of protein fragments used in the Y2H screen (see supplementary note and legends of Tables 2 and 3). The average (avg.) fragment length in percent refers to the peptide length used in experiment relative to the complete protein sequence.

Fragment length	Number of interactions	Avg. fragment length in percent	Average BPscore	DDI-validated PPIs in percent	LR	PRS set overlap (%)	NRS set overlap (%)
All lengths	2033	78.92	0.390	3.85	165.9	41 (2.02)	141 (6.94)
>10%	1963	80.03	0.386	3.86	166.1	39 (1.99)	134 (6.83)
>20%	1879	81.51	0.386	3.97	175.2	39 (2.08)	127 (6.76)
>30%	1679	84.77	0.389	4.51	211.9	39 (2.32)	105 (6.25)
>40%	1447	88.24	0.391	4.95	205.9	35 (2.42)	97 (6.70)
>50%	1313	90.27	0.390	4.93	221.1	31 (2.36)	80 (6.09)
>60%	1079	93.61	0.389	5.39	241.1	30 (2.78)	71 (6.58)
>70%	871	96.69	0.399	6.37	226.5	27 (3.10)	68 (7.81)
>80%	778	98.02	0.399	6.97	261.1	27 (3.47)	59 (7.58)
>90%	637	99.39	0.406	7.64	328.1	23 (3.61)	40 (6.28)
100%	355	100.00	0.423	9.27	242.0	14 (3.94)	33 (9.30)

**Table S12.** Comparison of MDC subsets based on the confidence score defined by the original authors [2] (see supplementary notes and legends of Tables 2 and 3). The score was determined by them using the following six criteria: the interaction had *HIS3*, *URA3*, and *lacZ* reporter activity, the interaction is found in human interaction clusters, the interaction is found in orthologous *D*. *melanogaster* clusters, the interaction is found in orthologous *C*. *elegans* clusters, the interaction is found in orthologous *S*. *cerevisiae* clusters, and the interaction is formed by proteins sharing GO annotations. The confidence score reflects the number of criteria fulfilled.

Confidence score	Number of interactions	Avg. fragment length in percent	Average BPscore	DDI-validated PPIs in percent	LR	PRS set overlap (%)	NRS set overlap (%)
≥ 1	2033	78.92	0.390	3.85	165.9	41 (2.02)	141 (6.94)
≥2	1566	79.29	0.405	4.26	209.4	40 (2.55)	109 (6.96)
≥ 3	860	80.25	0.442	4.31	380.4	26 (3.02)	39 (4.53)
≥4	339	80.77	0.469	5.51	466.8	9 (2.65)	11 (3.24)
≥ 5	91	83.28	0.509	4.17	2852.7	5 (5.49)	1 (1.1)
6	4	93.58	0.916	33.33		3 (75)	0 (0)

**Table S13.** Comparison of HPRD and IntAct by the number of publications and the experimental technique (cf. supplementary results and discussion as well as legends of Tables 2 and 3). A subset of protein interactions reported in exactly *n* publications is denoted by '*n* pub.' HPRD classifies experiments into three categories: *in vivo, in vitro,* and *yeast two-hybrid.* The IntAct classification of experimental techniques is based on a controlled vocabulary. Here, we regard only the most common experiment techniques frequently found in IntAct: yeast two-hybrid (Y2H), tandem affinity purification (TAP), co-immunoprecipitation (Co-IP), X-ray crystallography (X-ray), and protein array (PA). All other techniques are labeled 'other'. The datasets are ordered by decreasing *BPscore.* 

Dataset	Number of interactions	Average BPscore	DDI-validated PPIs in percent	LR	PRS set overlap (%)	NRS set overlap (%)
IntAct (PA)	109	0.811	28.99	— 4 (3.67)		0 (0.00)
IntAct (X-ray)	160	0.777	85.29	11702.9	7 (4.38)	2 (1.25)
HPRD (3 pub.)	861	0.668	41.44	3343.7	16 (1.86)	16 (1.86)
HPRD (≥ 4 pub.)	497	0.668	48.31	1308.4	9 (1.81)	23 (4.63)
IntAct (≥ 4 pub.)	93	0.666	41.03	8359.2	5 (5.38)	2 (2.15)
IntAct (other)	605	0.660	31.88	3600.9	14 (2.31)	13 (2.15)
HPRD (2 pub.)	3048	0.654	33.10	1382.1	31 (1.02)	75 (2.46)
IntAct (Co-IP)	920	0.647	23.36	1800.4	14 (1.52)	26 (2.83)
IntAct (3 pub.)	192	0.637	20.25	6130.1	11 (5.73)	6 (3.13)
IntAct (2 pub.)	582	0.637	27.72	1671.8	7 (1.20)	14 (2.41)
HPRD (in vivo)	17417	0.611	22.84	462.2	98 (0.56)	709 (4.07)
HPRD (in vitro)	19616	0.603	25.63	649.7	130 (0.66)	669 (3.41)
HPRD (1 pub.)	26550	0.584	21.23	418.7	144 (0.54)	1150 (4.33)
HPRD (Y2H)	7964	0.554	17.17	1137.9	146 (1.83)	429 (5.39)
IntAct (1 pub.)	4942	0.515	13.64	583.6	37 (0.75)	212 (4.29)
IntAct (Y2H)	2289	0.493	14.69	1005.6	40 (1.75)	133 (5.81)
IntAct (TAP)	1993	0.457	5.25	45.2	1 (0.05)	74 (3.71)

**Table S14.** Presence of human disease-associated proteins in predicted protein interaction datasets. For each dataset, the overall number of disease proteins contained, the overall number T of their interactions, the number D of interactions between solely disease proteins, the respective interaction percentage D divided by T, and the estimated *p*-value for D using Fisher's exact test are listed. Also, the overall number of additional disease protein interactions not present in the experimental datasets CCSB-HI1, MDC, HPRD, and IntAct is given next to both the number of additional disease proteins not present in these experimental datasets either and the number of all of their interactions.

Dataset	All disease proteins	Interactions of all disease proteins	Interactions between disease proteins	Interaction percentage	<i>p</i> -value	Additional interactions of all disease proteins	Additional disease proteins	All interactions of additional disease proteins
Bioverse	932	57045	4973	8.72	6.20E-140	55254	138	3483
Bioverse-core	260	1015	272	26.80	5.10E-25	514	4	22
Himap	867	11168	1492	13.36	5.70E-87	10716	142	1844
HiMAP-core	450	2800	440	15.71	1.10E-43	2592	47	412
HomoMINT	441	1743	172	9.87	6.50E-06	1522	93	254
OPHID	550	5271	438	8.31	0.1	4548	94	1215
POINT	1276	22610	2198	9.72	1.20E-187	21479	268	3914
Sanger	596	12456	743	5.97	1.20E-187	12290	136	2767
Sanger-core	393	2065	161	7.80	1.40E-02	1952	91	493

#### **Supplementary References**

- [1] Rhodes, D. R., Tomlins, S. A., Varambally, S., Mahavisno, V., et al., Nat Biotechnol 2005, 23, 951-959.
- [2] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., et al., Cell 2005, 122, 957-968.
- [3] Ramachandran, N., Hainsworth, E., Bhullar, B., Eisenstein, S., *et al.*, *Science* 2004, *305*, 86-90.
- [4] Phizicky, E., Bastiaens, P. I., Zhu, H., Snyder, M., Fields, S., *Nature* 2003, *422*, 208-215.
- [5] Salwinski, L., Eisenberg, D., *Curr Opin Struct Biol* 2003, *13*, 377-382.
- [6] Persico, M., Ceol, A., Gavrila, C., Hoffmann, R., et al., BMC Bioinformatics 2005, 6 Suppl 4, S21.
- [7] Gavin, A. C., Bosche, M., Krause, R., Grandi, P., et al., Nature 2002, 415, 141-147.
- [8] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., *et al.*, *Science* 2003, *302*, 1727-1736.
- [9] Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., et al., Nature 2002, 415, 180-183.
- [10] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., et al., Proc Natl Acad Sci US A 2001, 98, 4569-4574.
- [11] Li, S., Armstrong, C. M., Bertin, N., Ge, H., et al., Science 2004, 303, 540-543.
- [12] Suzuki, H., Fukunishi, Y., Kagawa, I., Saito, R., et al., Genome Res 2001, 11, 1758-1765.
- [13] Suzuki, H., Saito, R., Kanamori, M., Kai, C., et al., Genome Res 2003, 13, 1534-1541.
- [14] Tong, A. H., Evangelista, M., Parsons, A. B., Xu, H., et al., Science 2001, 294, 2364-2368.
- [15] Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., et al., Nature 2000, 403, 623-627.
- [16] von Mering, C., Krause, R., Snel, B., Cornell, M., et al., Nature 2002, 417, 399-403.
- [17] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., et al., Nucleic Acids Res 2000, 28, 289-291.
- [18] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., et al., Nucleic Acids Res 2002, 30, 303-305.
- [19] Breitkreutz, B. J., Stark, C., Tyers, M., Genome Biol 2003, 4, R23.
- [20] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., et al., FEBS Lett 2002, 513, 135-140.
- [21] Mewes, H. W., Frishman, D., Mayer, K. F., Munsterkotter, M., et al., Nucleic Acids Res 2006, 34, D169-172.
- [22] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., et al., Nucleic Acids Res 2000, 28, 235-242.
- [23] Nakayama, M., Kikuno, R., Ohara, O., Genome Res 2002, 12, 1773-1784.
- [24] Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P. O., et al., Nat Cell Biol 2004, 6, 97-105.
- [25] Colland, F., Jacq, X., Trouplin, V., Mougin, C., et al., Genome Res 2004, 14, 1324-1332.
- [26] Goehler, H., Lalowski, M., Stelzl, U., Waelter, S., et al., Mol Cell 2004, 15, 853-865.
- [27] Jin, J., Smith, F. D., Stark, C., Wells, C. D., et al., Curr Biol 2004, 14, 1436-1450.
- [28] Lehner, B., Semple, J. I., Brown, S. E., Counsell, D., et al., Genomics 2004, 83, 153-167.
- [29] Lehner, B., Sanderson, C. M., Genome Res 2004, 14, 1315-1323.
- [30] Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., et al., Science 2005, 307, 1621-1625.
- [31] Guo, D., Han, J., Adam, B. L., Colburn, N. H., et al., Biochem Biophys Res Commun 2005, 337, 1308-1318.
- [32] Rual, J. F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., et al., Nature 2005, 437, 1173-1178.
- [33] Lim, J., Hao, T., Shaw, C., Patel, A. J., et al., Cell 2006, 125, 801-814.
- [34] Tsang, H. T., Connell, J. W., Brown, S. E., Thompson, A., et al., Genomics 2006, 88, 333-346.
- [35] Camargo, L. M., Collura, V., Rain, J. C., Mizuguchi, K., et al., Mol Psychiatry 2007, 12, 74-86.