Supplementary Information to

A scored human protein-protein interaction network to catalyze genomic interpretation

Li T[1,2,3*], Wernersson R[4,5*], Hansen RB[4*], Horn H[1,2], Mercer JM[1,2], Slodkowicz G[5,6], Workman CT[5], Rigina O[5], Rapacki K[5], Stærfeldt HH[5], Brunak S[7], Jensen TS[4], Lage K[1,2,8]

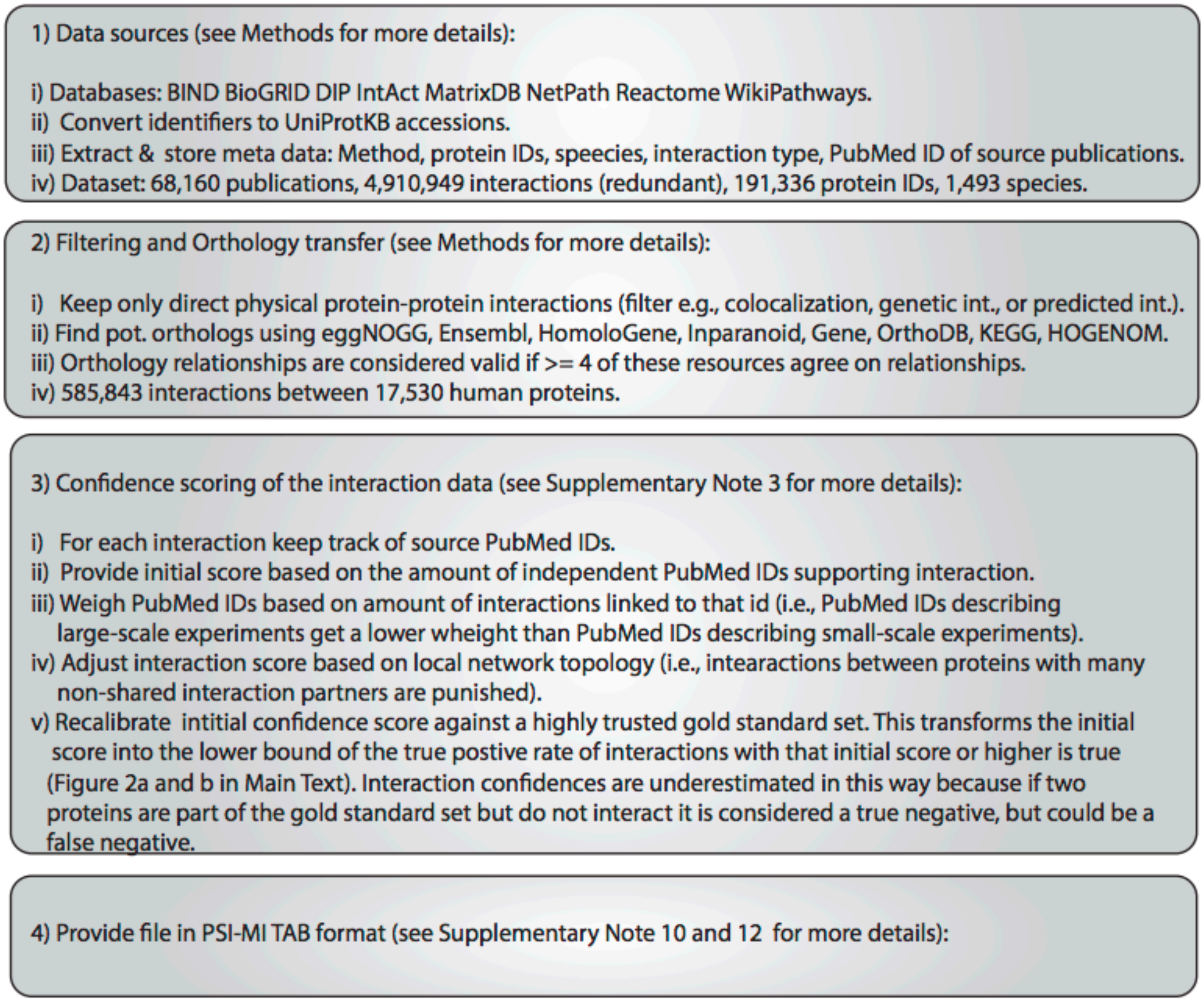
Affiliations:

1. Massachusetts General Hospital, Boston, Massachusetts, USA
2. Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA
3. Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
4. Intomics A/S, Lyngby, Denmark
5. Center For Biological Sequence Analysis, Technical University of Denmark, Lyngby, Denmark
6. EMBL, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
7. Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, Copenhagen, Denmark
8. Harvard Medical School, Boston, Massachusetts, USA

*These authors contributed equally

Correspondence should be addressed to Thomas Skøt Jensen (tsj@intomics.com) and Kasper Lage (lage.kasper@mgh.harvard.edu).

Table of contents:



1) Data sources (see Methods for more details):

i) Databases: BIND BioGRID DIP IntAct MatrixDB NetPath Reactome WikiPathways.
ii) Convert identifiers to UniProtKB accessions.
iii) Extract & store meta data: Method, protein IDs, speecies, interaction type, PubMed ID of source publications.
iv) Dataset: 68,160 publications, 4,910,949 interactions (redundant), 191,336 protein IDs, 1,493 species.

2) Filtering and Orthology transfer (see Methods for more details):

i) Keep only direct physical protein-protein interactions (filter e.g., colocalization, genetic int., or predicted int.).
ii) Find pot. orthologs using eggNOGG, Ensembl, HomoloGene, Inparanoid, Gene, OrthoDB, KEGG, HOGENOM.
iii) Orthology relationships are considered valid if >= 4 of these resources agree on relationships.
iv) 585,843 interactions between 17,530 human proteins.

3) Confidence scoring of the interaction data (see Supplementary Note 3 for more details):

i) For each interaction keep track of source PubMed IDs.
ii) Provide initial score based on the amount of independent PubMed IDs supporting interaction.
iii) Weigh PubMed IDs based on amount of interactions linked to that id (i.e., PubMed IDs describing large-scale experiments get a lower wheight than PubMed IDs describing small-scale experiments).
iv) Adjust interaction score based on local network topology (i.e., intearactions between proteins with many non-shared interaction partners are punished).
v) Recalibrate intitial confidence score against a highly trusted gold standard set. This transforms the initial score into the lower bound of the true postive rate of interactions with that initial score or higher is true (Figure 2a and b in Main Text). Interaction confidences are underestimated in this way because if two proteins are part of the gold standard set but do not interact it is considered a true negative, but could be a false negative.

4) Provide file in PSI-MI TAB format (see Supplementary Note 10 and 12 for more details):

Supplementary Figure 1 | The computational framework leading to InWeb_IM. Details can be seen using the Adobe Zoom Tool and in Text, Figures, and Supplementary Notes as indicated with bold text. InWeb_IM is available from www.lagelab.org and https://www.intomics.com/inbiomap. Moreover, we make the data accessible from a graphical user interface http://apps.broadinstitute.org/genets#InWeb_InBiomap so that it can interactively explored by any individual researcher that wishes to study the interactions of proteins of interest.

Supplementary Note 1 | Determining the optimal orthology majority voting scheme. To quantitatively test how different parameter choices in our orthology transfer scheme influences the resulting network we repeated the pathway analysis (detailed in the Main Text, Figure 2a, and **Supplementary Note 5**) with different parameters (i.e., requiring that two, three, four, five or six orthology databases needed to agree before transferring data from model organisms to human protein pairs). This analysis shows that requiring at least four orthology methods agree gives the best resulting network (i.e., the best pathway signal and still very good coverage as summarized below):

Number of databases required	**Two**	**Three**	**Four**	**Five**	**Six**
Area under curve (analogous to Figure 2a in the Main Text).	0.80	0.80	0.83	0.81	0.80
Number of interactions without pathways databases	1,607,858	738,775	451,304	303,210	247,809

Supplementary Note 2 | Other networks. There are five other resources with which InWeb_IM was compared. In the case where multiple networks exist for different species only the human network was used. For fair comparison, HGNC symbol and UniProt accession were converted using the mapping table provided by the HUGO Gene Nomenclature Committee (HGNC, http://www.genenames.org/cgi-bin/download). :

Interologous Interaction Database (I2D[1-2], http://ophid.utoronto.ca/) was downloaded on November 3, 2015. It is the most recent version of the network (v2.9, updated on July 10, 2015).

Mentha[3] (http://mentha.uniroma2.it/) was downloaded on November 3, 2015. It is updated on a weekly basis.

iRefIndex[4] (http://irefindex.org/) was downloaded on November 3, 2015. Only binary interactions where both interactors are human proteins and where both interactors had UniProt accessions were used in the analysis; complexes were not included. It is the most recent version of the network (v14.0, updated on April 7, 2015).

Protein Interaction Network Analysis Platform (PINA[5], http://cbg.garvan.unsw.edu.au/pina/) was downloaded on November 3, 2015, and only interactions where both interactors are human proteins were included in the analysis. It is the second and most recent version of the network, updated on May 21, 2014.

High-quality Interactomes (HINT[6], http://hint.yulab.org/) was downloaded on November 3, 2015. Binary, co-complex, high-throughput, and literature-curated interactome for human was pooled to form the resulting network. It was the third and most recent version of the resource (website updated on September 30, 2015).

Supplementary Note 3 | Details to calculating a confidence score for each interaction.

Initial confidence score:

The interaction databases define a ternary relation between PubMed IDs and pairs of UniProt ACs: If an evidence record claims that a publication with PMID p describes an interaction between two proteins that can be mapped to UniProt ACs u and v we write $u \leftarrow p \rightarrow v$. To account for symmetry, we only consider proteins u and v with $u \leq v$ for some total ordering $\leq$. We define the size of the experiment $X(p) = |\{(u,v) \mid u \leq v \wedge u \leftarrow p \rightarrow v\}|$, i.e. the number of different interactions described in the experiment. Note that for pathway databases we do not include the PMIDs that describe the pathway databases themselves.

If a represents a human protein that is orthologous to a protein u, we write $a \sim u$. Furthermore, we define:

- The ternary relation of inferred interactions, written $a \Leftarrow p \Rightarrow b$, iff there exist u and v such that $a \sim u$, $b \sim v$, and $u \leftarrow p \rightarrow v$.
- The binary relation of inferred interactions, written $a \Leftrightarrow b$, iff there exists a p so that $a \Leftarrow p \Rightarrow b$.
- The neighbors of a protein $N(a) = \{b \mid a \Leftrightarrow b\}$.

Note that a human protein is considered to be orthologous with itself, i.e. the relation $a \sim a$ holds. Also note, that we are only interested in interactions $a \Leftrightarrow b$ between human proteins a and b with $a \leq b$ (where $\leq$ is the same total ordering as mentioned before) and $a \neq b$, i.e. we disregard self-interactions.

For an interaction $a \Leftrightarrow b$ let $n = |\{p \mid a \Leftarrow p \Rightarrow b\}|$ be the number of supporting experiments. We denote these experiments by $p_1, \ldots, p_n$ and order them so that $X(p_1) \leq \cdots \leq X(p_n)$. For $i \in \{1, \ldots, n\}$ we let $x_i = 1$ if $X(p_i) \leq 9$, and $x_i = \frac{log(3)}{log(X(p_i)-6)}$ otherwise. Furthermore, we let $D(a \Leftrightarrow b) = (|N(a) \setminus (N(b) \cup \{a,b\})| + 1) \cdot (|N(b) \setminus (N(a) \cup \{a,b\})| + 1)$, and define the initial score of the interaction $a \Leftrightarrow b$ as $I(a \Leftrightarrow b) = D(a \Leftrightarrow b)^{-\tau} \frac{1-\epsilon}{\epsilon} \sum_{i=1}^{n} x_i \epsilon^i$ for suitable parameters ϵ and τ. As the initial score is calibrated to the final score that has a probabilistic interpretation, the exact values are not important. To avoid circularity in the following, we only assign an initial confidence score to interactions with evidence from at least one non-pathway source.

Validating the initial score and providing a probabilistic interpretation:

To validate the initial score, we used it to rank all interactions with evidence from at least one non-pathway source, and plotted the percentage of interactions derived from pathway databases (158,092 interactions extracted from NetPath[7], Reactome[8], and WikiPathways[9], excluding neighboring reactions) found as a function of the percentage of the interactions investigated, starting with those with highest scores (**Figure 1a, Main Text**). The diagonal shows expected performance of a randomized dataset. A strong signal is seen in the first third of the ranked interaction list, indicating that the initial confidence score is capable of prioritizing the interactions most likely to be true.

Probabilistic interpretation of the confidence score:

To give a probabilistic interpretation of the initial confidence score we used the interactions derived from pathway databases as a benchmark set, assuming these interactions are true positives. We ordered the inferred interactions $a_1 \Leftrightarrow b_1, \dots, a_n \Leftrightarrow b_n$that had evidence from at least one non-pathway database such that $I(a_1 \Leftrightarrow b_1) \leq \cdots \leq I(a_n \Leftrightarrow b_n)$, and by using a sliding window approach we then calculated the median score and the benchmark rate for each window (the number of benchmark interactions divided by the window size). A generalized logistic function with lower asymptote 0 and upper asymptote 1 was then fitted to these (median score, benchmark rate) points using the method of least squares (**Figure 1b, Main Text**). The initial scores were then transformed into the final scores using this function and by artificially assigning a final score of 1 to the inferred interactions with evidence from at least one pathway database. Assuming a higher initial score for an interaction implies a higher probability for the interaction to be a true interaction, the final score can then be interpreted as a lower bound on the probability for the interaction being a true positive, since the benchmark set of interactions is assumed to be a subset of the true interactions, and since the benchmark rate must then be lower than the true positive rate.

Providing the scores in standard formats such as PSISCORE

Both the initial scores and the final confidence scores are reported using the PSI-MITAB file format allowing for integration with standards based systems like PSISCORE. This allows the community to easily benefit from the protein-protein interaction scoring scheme that has a probabilistic interpretation, enabling seamless integration with large genetic datasets.

Supplementary Note 4 | Comparing to benchmarking against generating 58 human pull downs

Rationale behind our analysis: To test if our confidence score is correlated with an experimentally derived measure of the confidence of binding between two proteins from an independent dataset of human protein-protein interactions we correlated it with the heavy-to-light (H-L) isotope ratios from a stable isotope labeling in cell culture (SILAC) experiment in human cells. This metric enables the quantitative assessment of the amounts of a protein (e.g., protein X) seen in an immunoprecipitation experiment with a bait compared to a control condition (which is the reason for terming these experiments 'quantitative interaction proteomics'). Importantly, this metric is well suited to compare our confidence score to as it is used widely in the experimental proteomics communities as a direct measure of interaction confidence between proteins in an immunoprecipitation experiment (personal communication Jake Jaffe, Associate Director of the Proteomics Platform of the Broad Institute, and Steve Carr, Director of the Proteomics Platform of the Broad Institute). It was also critical to the choice of dataset for this analysis that it has enough coverage to allow a rigorous correlation of our score and the experimental data. For these reasons we used a recently generated dataset of 58 pull-downs in human cells from Rosenbluh, Mercer et al., *in press* Cell Systems which matched these criteria. Indeed we confirm that our final confidence score and the experimental values are robustly correlated (correlation coefficient of 0.38, C.I. [0.35, 0.42]). We repeated this analysis for the only other two network that assigns scores to the interactions, and found a comparable correlation in iRefIndex (0.41, C.I. [0.38, 0.44]), and a lower correlation in Mentha (0.23, C.I. [0.17, 0.29]), where both correlations are statistically significant. Together these analyses confirm the reliability of our score and show it is significantly correlated with an experimentally derived measure of the confidence of binding between proteins.

Description of the pull down dataset and confirming its quality: Fifty-eight genes involved in colon cancers were chosen as the starting point of a quantitative interaction proteomics experiment based on the stable Isotope labeling by amino acids in cell culture (SILAC) methodology. The bait proteins were tagged with a V5 epitope tag and immunoprecipitated in DLD1 cells to identify 15,189 protein interactions using an anti V5 antibody. The quality of the interaction data was confirmed by developing a method for protein interaction credibility scoring (ICS). For each of the 15,189 PPI identified in draft-PPI, we computed three predictors: (1) Heavy to light ratio, (2) Jaccard similarity coefficient; and (3) Edge betweenness centrality. As true positive protein interactions, we considered 1,149 that were identified in draft-PPI and are also reported as high confidence protein interactions in publicly available PPI databases[10]. We found that using each of these predictors we correctly identified only a subset of the true positive PPIs suggesting that none of predictors alone was suitable for this application. As such, we developed ICS using the Random Forest (RF) binary response classifier, which uses the true positive PPI as a response and computes 5000 different combinations (modules) of these classifiers that correctly enrich true positive PPI. The ICS score is calculated using the number of modules in which a draft-PPI protein interaction scores together with the true positives. We trained the RF model on 70% of the data and used the area under the ROC curve (AUC) on the remaining 30% of the dataset to calculate classification power. The ICS showed high AUC (AUC: 96.9 confidence intervals (95.7%, 98.1%) and 5-fold cross validation showed similar AUC estimates. Furthermore, ICS correctly identified the majority of true positive PPI, demonstrating the reliability of this approach. Importantly, this approach identified well-characterized relationships such as interactions between β-catenin and TCF7L2 or YAP1 and TEAD transcription factors, as well as more recently reported observations between YAP1 and β-catenin. More information about the dataset can be found in (Rosenbluh et al., Cell Systems 2016, Accepted). The quality of the interaction data we use here is further supported by the observation that in **Figure 2e** our comparisons of the dataset and the six different networks leads to AUCs in the range from 0.84 to 0.78.

Supplementary Note 5 | Biological signal of InWeb_IM compared to other resources. The ability for a protein-protein interaction network to reveal structures of biological pathways and discover new proteins functionally associated with a given set of proteins testifies the relevance of the network in understanding a variety of biological processes. We used the Quack algorithm (described in detail on http://www.broadinstitute.org/genets#users/userguide) to examine and compare the performance of each resource to classify pathway membership based on a subset of canonical pathways from the Molecular Signatures Database (MSigDB C2:CP, http://software.broadinstitute.org/gsea/msigdb). The final list of $n = 853$ pathways met a well-established pathway definition (http://www.genome.gov/27530687#al-1), had a reasonable size (extremely large pathways were removed), and were ensured to be non-redundant by analyzing the pairwise Jaccard index of set similarity between pathways. Quack then evaluates how well each network recapitulates the 853 pathways based on machine learning of eighteen topological features in each of the networks and assessing the importance of each of these topological measures in determining which proteins participate in pathways together. We defined the context of each protein to be its first-order interaction partners in the respective networks, and calculated the topological features for both pathway proteins and their contexts. We thus created a 18-dimensional feature vector for each of the proteins specific for each of the 853 pathways. Using the Random Forest algorithm, we trained our classifier on 70% of the data to predict the binary outcome of each of the 30% left-out proteins (to be in an MSigDB gene set or not in an MSigDB gene set). We calculated the area under the receiver operating curve (AUC) for the classifiers in each of the network as a basis for comparing the biological signal in each resource. We assess the performance of each network using both a normalized and a non-normalized dataset where the normalized analysis does not penalize if proteins in a pathway being tested are not covered by data in the network in question. In the normalized analysis InWeb_IM has an AUC of 0.95, Mentha 0.93, I2D 0.91, iRefIndex and PINA 0.89, and HINT 0.88.

If we do take into consideration pathway proteins that are not covered by data in the individual networks (meaning that proteins we are trying to assign to a given pathway, but are not covered by data in the network being tested, will be counted as false negatives), InWeb_IM is 10% better than the next best network with an AUC of 0.86. the next best network is I2D (AUC of 0.78), followed by Mentha (AUC of 0.77), iRefIndex (AUC of 0.74), PINA (AUC of 0.73), and HINT (AUC of 0.63). Additional details available from described in detail on http://www.broadinstitute.org/genets#users/userguide and in Mercer et al., 2016, in preparation.

Supplementary Note 6 | Biological signals of unique and orthology transferred data in InWeb_IM

To further dissect and evaluate the quality of InWeb_IM interactions, we performed Quack on subsets of InWeb_IM. First, we looked at the 343,319 interactions from 13,119 proteins that are uniquely identified in InWeb_IM compared to the other five networks and executed the analysis detailed in **Figure 2d** and **Supplementary Note 5**. The original AUC on this subset is 0.90. For comparison the 242,524 interactions from 16,408 proteins found in InWeb_IM and at least one other network (i.e., shared interactions) lead to an AUC of 0.95, showing that, while the unique interactions have a slightly lower signal, the data is still of high quality.

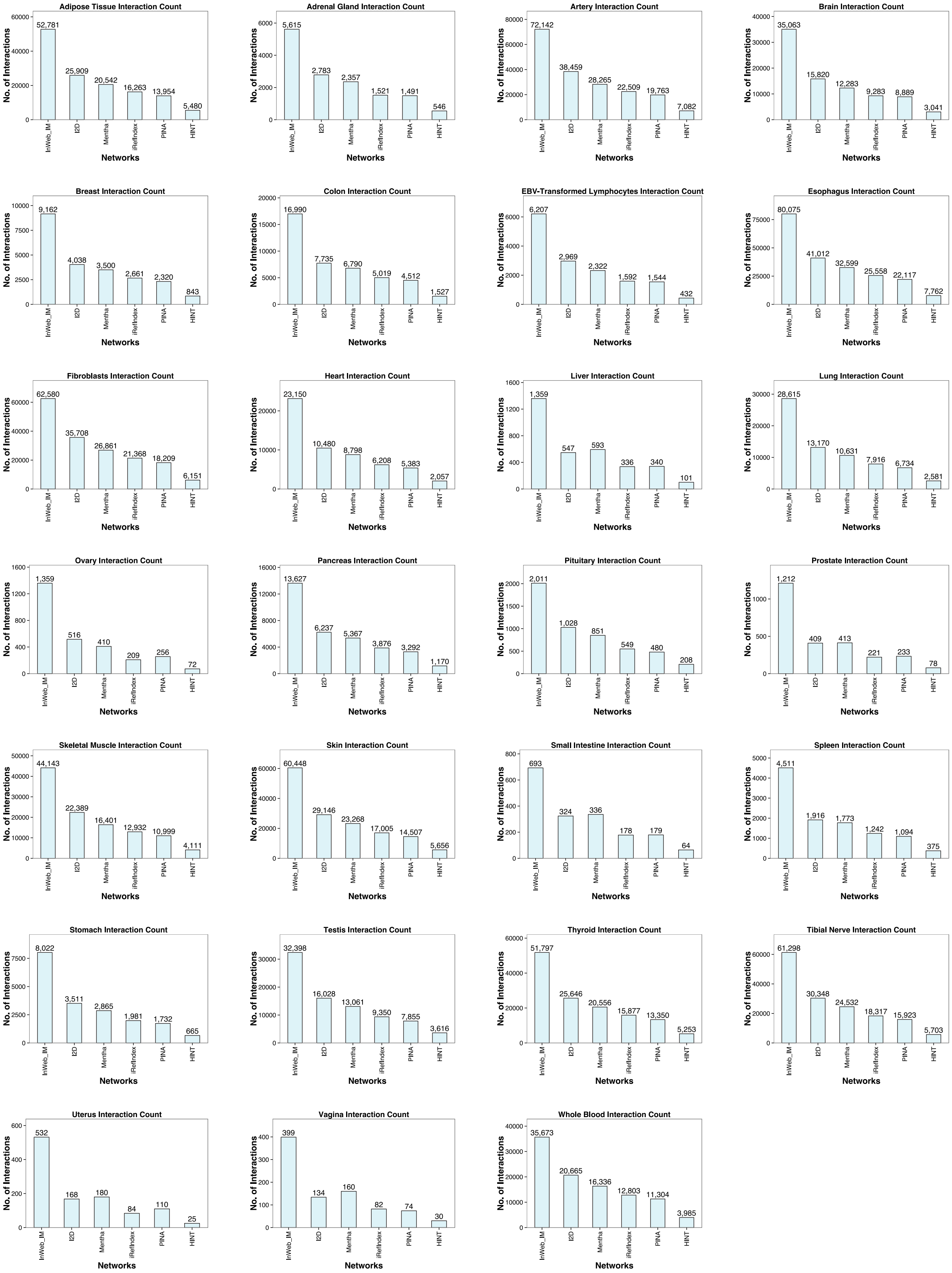
We then looked at the 253,347 interactions from 10,898 proteins which stem from orthology transfer. The AUC of this subset is 0.85 showing that orthology transfer also results in high quality interactions.

The analysis result is summarized in the following table:

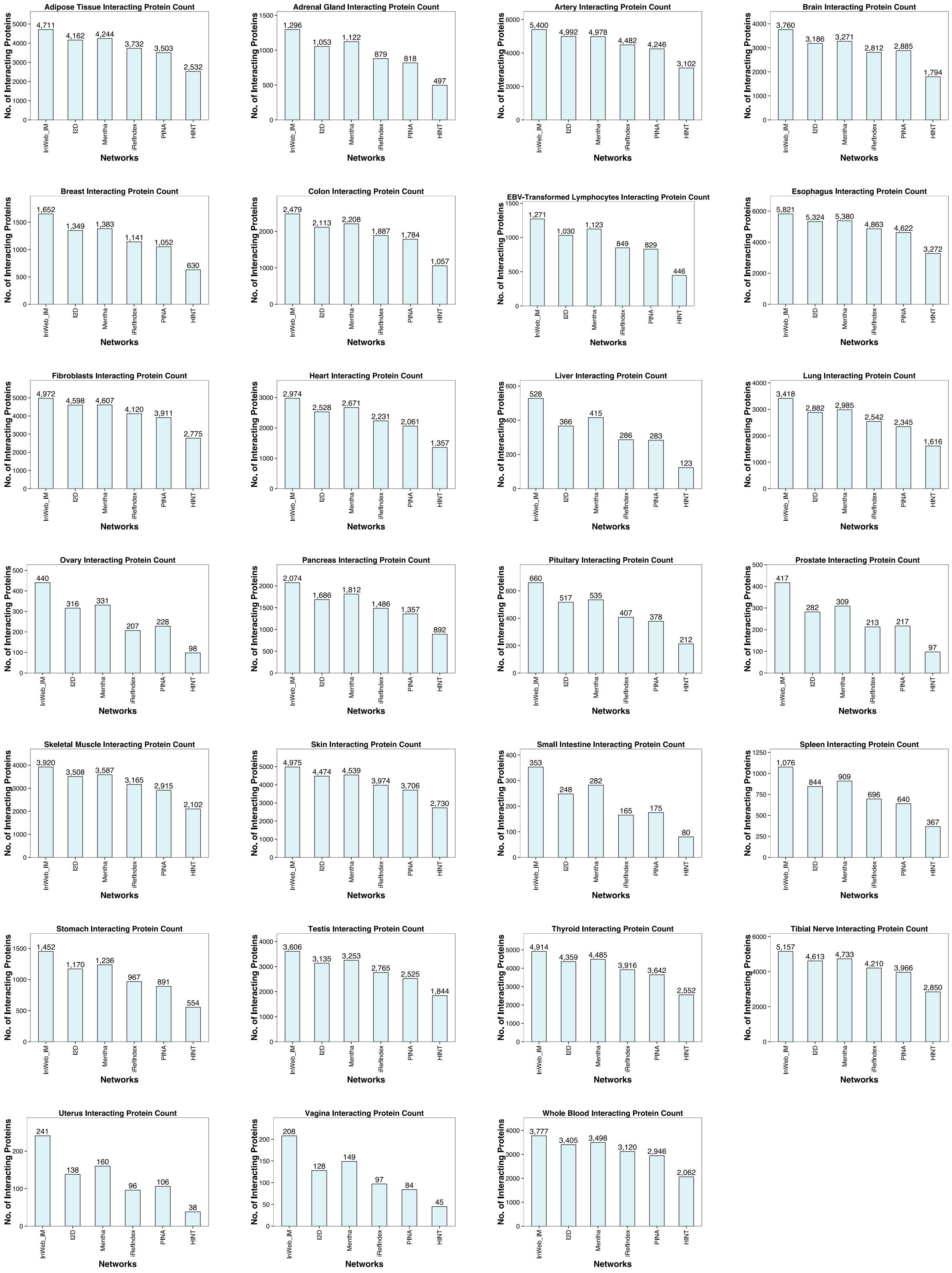
Quack AUC (analogous to Figure 2d in the Main Text)	**Orthology Transfer vs Human**		**Unique vs Shared**		**Entire InWeb_IM**
	Transferred	**Human Only**	**Unique**	**Shared**	
No Normalization	0.85	0.95	0.90	0.95	0.95
# of proteins	10,898	17,187	13,119	16,408	17,530
# of Interactions	253,347	332,496	343,319	242,524	585,843

Supplementary Note 7 | Details to annotation of genes from 21 tumor types. Network mutation burden (NMB) is an algorithm that tests and quantifies the degree to which cancer driver genes can be accurately classified based on the mutation burden in their first order neighborhood. Briefly, for a given index gene the NMB is formalized into a score that reflects the empirical probability of the observed mutation signal aggregated across its first order biological network, excluding the index gene itself, while normalizing for the number of genes in the network. We confirmed that the NMB accurately calculates the significance level of the mutation burden in the neighborhood of an index gene, based on the fact that the majority of genes fit the null hypothesis and lie on the diagonal in a Q-Q plot. Details of the NMB algorithm can be found here: http://biorxiv.org/content/early/2015/08/25/025445 . For each of the six different protein-protein interaction networks, we tested the extent to which 219 cancer genes in the Cancer5000 Stringent set from Lawrence et al., 2014 (Ref. 11) can be accurately identified amongst a set of genes not related to cancers by using the NMB score as the classifier.

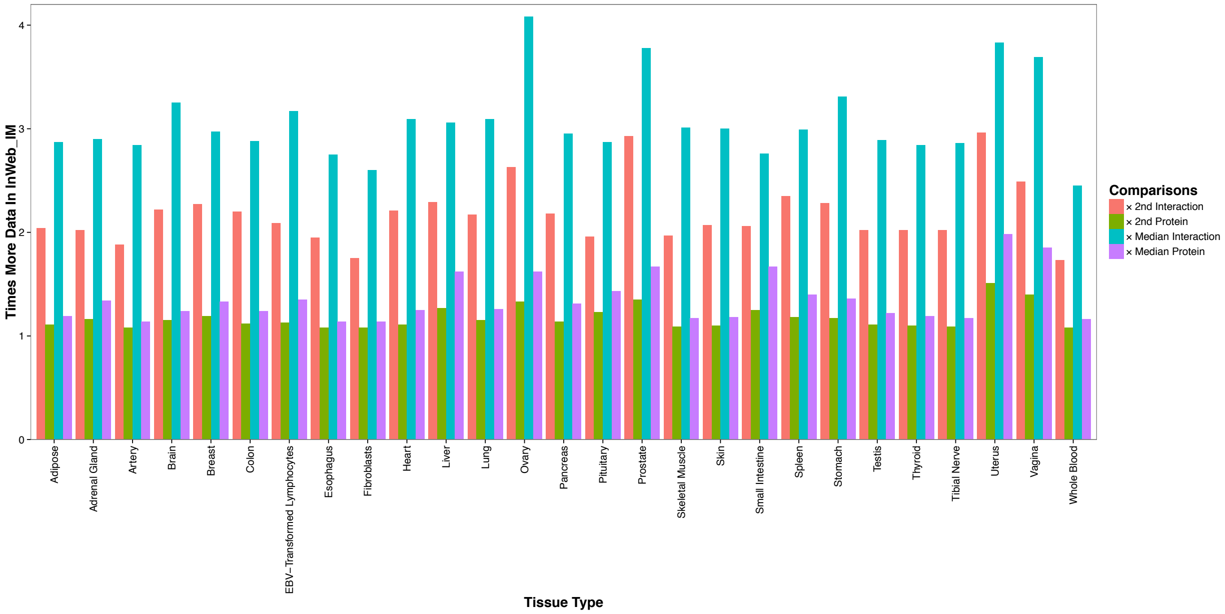
Supplementary Note 8 | Assessing tissue-specific interactions across networks. We downloaded all data on tissue-specific expression quantitative trait loci (eQTL)s from the Genotype-Tissue Expression Project[12] (http://www.gtexportal.org/home/eqtls/) and collapsed eQTLs from tissues with the same higher order label (e.g., all tissues denoted as Brain - X, Brain - Y, … Brain N). This provided 27 tissue-specific sets of genes and for each set we filtered the interaction data in InWeb_IM (by only including interaction data between gene pairs represented in the set in question) to derive 27 tissue-specific InWeb_IM networks. For each tissue-specific InWeb_IM network we counted 1) the amount of physical interactions present in the network and 2) the amount of proteins covered by data. This procedure was repeated for I2D, Mentha, PINA, iRefIndex, and HINT and the tissue-specific networks derived from this procedure were quantitatively compared to each other (**Supplementary Figures 2** and **3**). The tissue-specific InWeb_IMs have between 2 and 3 times more interactions than the median of all networks and 2 times more interactions than the next largest network it is being compared to. InWeb_IM has a comparable amount of proteins covered by tissue-specific interactions in each context (**Supplementary Figure 4**).



Supplementary Figure 2 | Tissue specific interaction counts. Details can be seen by using the Adobe Zoom Tool. Networks are indicated on the x axis and interactions between proteins in which the corresponding genes are involved in a tissue-specific expression quantitative trait locus on the y axis. The analysis is made for 27 tissues as indicated on the top of each panel.

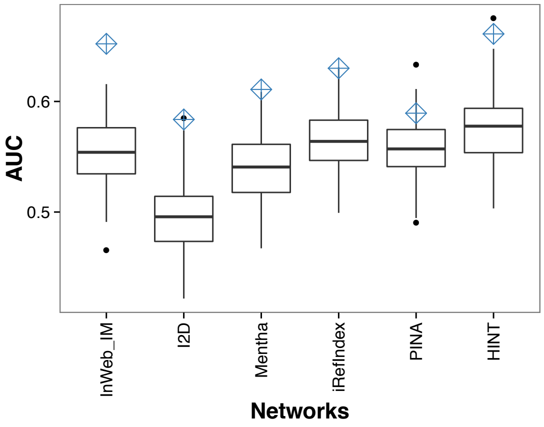


Supplementary Figure 3 | Proteins covered by tissue-specific interactions. Details can be seen using the Adobe Zoom Tool. Networks are indicated on the x axis and proteins that have at least one interaction to another protein with a similar tissue-specific expression quantitative trait locus on the y axis. The analysis is made for 27 tissues as indicated on the top of each panel.



Supplementary Figure 4 | Tissue-specific interactions across networks in InWeb_IM compared to other networks. The x-axis represents the 27 different tissue types from GTEx and the y-axis shows how much data compared to InWeb_IM. Red bars denote the values for the amount interactions in the next-largest network, green bars denote the values for the amount of proteins covered by data in the network with the next highest count, light blue denotes the values for the median amount of interactions across all five comparable networks, and purple denotes the values for the median amount of proteins covered by interactions across all five comparable networks.

Supplementary Note 9 | Annotating autism genes. InWeb_IM has close to 3 times more brain-specific interaction data than the median of the other interaction networks (**Supplementary Figure 2** and **4**). Therefore, we hypothesized that it would be a uniquely enabling property of InWeb_IM to recapitulate pathway relationships between genes involved in psychiatric diseases such as autism. Analogously to the way in which we applied Network Mutation Burden (NMB) to cancer genes we tested and quantified the degree to which known autism genes can be accurately classified based on the association burden in their first order neighborhood. Briefly, for a given index gene the NMB is formalized into a score that reflects the empirical probability of the observed association signal aggregated across its first order biological network, excluding the index gene itself, while normalizing for the number of genes in the network (described in detail in **Supplementary Note 7** and http://biorxiv.org/content/early/2015/08/25/025445). For each of the six different protein-protein interaction networks tested the extent to which 65 bona fide autism genes (defined as those with an FDR <=0.1 in the paper by Sanders et al., 2015[13]) can be accurately identified amongst a set genes not related to autism by using their NMB scores as a classifier. Performance was was determined using standard areas under the receiver operating characteristics curves (AUCs). We determined the statistical significance of the AUCs for each network through permutation tests. Specifically, for each network we created 120 sets of 65 random genes matched to the degree distribution of the 65 autism genes in the network in question and repeatedly the NMB calculation for these random sets (**Supplementary Figure 5**). InWeb_IM has an AUC of 0.65 which is significant at the Adj. P < 0.05 level as the only network. Although one network (HINT) has a marginally higher AUC (0.66) this AUC is not significant at that level because the random AUCs for HINT are inflated (median random AUC = 0.58). This is likely to be because of the smaller amount of interactions in this network which leads to an overestimation of the significance of the NMB scores both for autism genes but also the random controls. A consequence of this is also that the difference between the medium random AUCs and the AUCs observed for the 65 autism genes is smaller in HINT than in InWeb_IM (0.08 versus 0.10).



Supplementary Figure 5 | The AUCs derived from applying NMB in autism. The AUC observed in the network denoted on the x axis is indicated with the blue diamonds (InWeb_IM = 0.65, I2D = 0.58, Mentha = 0.61, iRefIndex = 0.63, PINA = 0.59, HINT = 0.66). The null distribution of AUCs of 120 matched random sets are shown with box whiskers plots. Of these AUCs only the one of InWeb_IM is significant at the Adj. P < 0.05 level.

Supplementary Note 10 | Example of interactions traced to source databases and original publications. We provide InWeb_IM in the PSI-MITAB file format in two versions: “core” and “full”. Both versions use UniProtKB accession numbers as their primary identifier and provide UniProtKB IDs, Ensembl and HUGO identifiers for convenience. In the column for “Interaction detection methods”, the “core” version has the string *psi-mi:"MI:0045"(experimental interaction detection)* if the exact same interaction was reported in at least one of the source databases using this value (or a descendent in the PSI-MI controlled vocabulary). The column contains the string *psi-mi:"MI:0362"(inference)* if the exact same interaction was not reported in any of the source databases (e.g., if the interaction is inferred using orthology). In the column for “Confidence score” we report the confidence scores followed by the initial scores. An example of a line from the “core” version is shown below with each column on a line of its own for readability:

Column	Content
1	uniprotkb:O00429
2	uniprotkb:Q96C03
3	uniprotkb:DNM1L_HUMAN|ensembl:ENSG00000087470|ensembl:ENST000...
4	uniprotkb:MID49_HUMAN|ensembl:ENSG00000177427|ensembl:ENST000...
5	uniprotkb:DNM1L(gene name)|uniprotkb:DNM1L(display_short)
6	uniprotkb:MIEF2(gene name)|uniprotkb:MIEF2(display_short)
7	psi-mi:"MI:0045"(experimental interaction detection)
8	-
9	-
10	taxid:9606(Homo sapiens)
11	taxid:9606(Homo sapiens)
12	-
13	-
14	-
15	0.409|0.693
16	-

Columns 8, 9, 12, 14, and 16 are not used in the “core” version and always contain a dash. Column 13 contains the string *psi-mi:"MI:1106"(pathways database)* if at least one interaction derived from a pathways database was used as evidence for the interaction, and a dash otherwise.

The “full” version consists of the same lines as the “core” version. However, after each line describing an interaction as shown above (“interaction line”), we inserted additional lines reporting the evidence for the interaction (“evidence line”). These two types of lines can be distinguished using column 9, “Identifier of the publication”, which is always a dash for interaction lines and always contains a PubMed reference for evidence lines. The InWeb_IM interactions can therefore be found by scanning column 9 in the preceding lines until a dash is found. For evidence lines, columns 7, 9-14, and 16 contain information extracted from the source databases, in particular the PubMed ID in column 9. The following four tables continue the previous example by each representing a line of evidence for the interaction described in the table above:

Column	Content
1	uniprotkb:O00429
2	uniprotkb:Q96C03
3	uniprotkb:DNM1L_HUMAN|ensembl:ENSG00000087470|ensembl:ENST000...
4	uniprotkb:MID49_HUMAN|ensembl:ENSG00000177427|ensembl:ENST000...
5	uniprotkb:DNM1L(gene name)|uniprotkb:DNM1L(display_short)
6	uniprotkb:MIEF2(gene name)|uniprotkb:MIEF2(display_short)
7	psi-mi:"MI:0096"(pull down)
8	-
9	pubmed:23530241
10	taxid:9606(Homo sapiens)
11	taxid:9606(Homo sapiens)
12	psi-mi:"0407"(direct interaction)
13	psi-mi:"MI:0463"(biogrid)
14	biogrid:853691
15	-
16	-

Column	Content
1	uniprotkb:Q8K1M6
2	uniprotkb:Q5NCS9
3	uniprotkb:DNM1L_MOUSE|ensembl:ENSMUSG00000022789|ensembl:ENS...
4	uniprotkb:MID49_MOUSE|ensembl:ENSMUSG00000018599|ensembl:ENS...
5	uniprotkb:Dnm1l(gene name)|uniprotkb:Dnm1l(display_short)
6	uniprotkb:Mief2(gene name)|uniprotkb:Mief2(display_short)
7	-
8	-
9	pubmed:24508339
10	taxid:10090(Mus musculus)
11	taxid:10090(Mus musculus)
12	-
13	psi-mi:"MI:0465"(dip)
14	dip:-
15	-
16	-

Column	Content
1	uniprotkb:O00429
2	uniprotkb:Q96C03
3	uniprotkb:DNM1L_HUMAN|ensembl:ENSG00000087470|ensembl:ENST000...
4	uniprotkb:MID49_HUMAN|ensembl:ENSG00000177427|ensembl:ENST000...
5	uniprotkb:DNM1L(gene name)|uniprotkb:DNM1L(display_short)
6	uniprotkb:MIEF2(gene name)|uniprotkb:MIEF2(display_short)
7	psi-mi:"MI:0018"(two hybrid)
8	-

9	pubmed:21508961
10	taxid:9606(Homo sapiens)
11	taxid:9606(Homo sapiens)
12	psi-mi:"0915"(physical association)
13	psi-mi:"MI:0469"(intact)
14	intact:EBI-8630872
15	-
16	-

Column	**Content**
1	uniprotkb:O00429
2	uniprotkb:Q96C03
3	uniprotkb:DNM1L_HUMAN|ensembl:ENSG00000087470|ensembl:ENST000...
4	uniprotkb:MID49_HUMAN|ensembl:ENSG00000177427|ensembl:ENST000...
5	uniprotkb:DNM1L(gene name)|uniprotkb:DNM1L(display_short)
6	uniprotkb:MIEF2(gene name)|uniprotkb:MIEF2(display_short)
7	psi-mi:"MI:0030"(cross-linking study)
8	-
9	pubmed:21508961
10	taxid:9606(Homo sapiens)
11	taxid:9606(Homo sapiens)
12	psi-mi:"0915"(physical association)
13	psi-mi:"MI:0469"(intact)
14	intact:EBI-8630904
15	-
16	-

Supplementary Note 11 | Discussion of unique features of the score and transparency of InWeb_IM data compared to other networks. The combination of a reliable confidence score with a probabilistic interpretation and complete transparency of the source data for all interactions is unique to InWeb_IM. Two resources (PINA and I2D) do not provide any confidence score or quality-based stratification of the data. HINT separates the interactions crudely into high and low confidence data, and Mentha and iRefIndex provides a score which reflects the weighted sum of the independent publications supporting an interaction which does not have a probabilistic interpretation (**Figure 2c**, **Main Text**). For I2D, the next largest network, it is not possible to trace the publications supporting the interaction data. Importantly, of all these networks only I2D integrate data between organisms to increase coverage of the human protein-protein interaction data.

Supplementary Note 12 | Manual Curation

To assess the accuracy of the curated interactions in InWeb_IM and to demonstrate the accessibility of supporting evidence, we randomly selected 20 human interactions from non-pathway sources, and manually checked whether each interaction is supported by the publications as reported in InWeb_IM. The list of selected interactions is as follows (listing UniProtKB IDs/mnemonics, UniProtKB accession numbers, confidence scores and a concise representation of the supporting evidence):

<table>
<tr><td>POTEF_HUMAN</td><td>ANTR1_HUMAN</td><td>A5A3E0</td><td>Q9H6X2</td><td>0.087</td><td>BioGRID|binary|A5A3E0-Q9H6X2|26186194|9606-9606|9606-9606</td></tr>
<tr><td>SYNE4_HUMAN</td><td>DJC11_HUMAN</td><td>Q8N205</td><td>Q9NVH1</td><td>0.087</td><td>BioGRID|binary|Q8N205-Q9NVH1|26186194|9606-9606|9606-9606</td></tr>
<tr><td>MGRN1_HUMAN</td><td>UB2D4_HUMAN</td><td>O60291</td><td>Q9Y2X8</td><td>0.092</td><td>BioGRID|binary|O60291-Q9Y2X8|19549727|9606-9606|9606-9606
IntAct|binary|O60291-Q9Y2X8|19549727|9606-9606|9606-9606</td></tr>
<tr><td>HSPB1_HUMAN</td><td>ANKR7_HUMAN</td><td>P04792</td><td>Q92527</td><td>0.094</td><td>BioGRID|binary|P04792-Q92527|25277244|9606-9606|9606-9606
IntAct|binary|P04792-Q92527|25277244|9606-9606|9606-9606</td></tr>
<tr><td>PBX1_HUMAN</td><td>FOXC1_HUMAN</td><td>P40424</td><td>Q12948</td><td>0.220</td><td>IntAct|binary|P40424-Q12948|15684392|9606-9606|9606-9606</td></tr>
<tr><td>ANM8_HUMAN</td><td>TR150_HUMAN</td><td>Q9NR22</td><td>Q9Y2W1</td><td>0.087</td><td>BioGRID|binary|Q9NR22-Q9Y2W1|26186194|9606-9606|9606-9606</td></tr>
<tr><td>CBL_HUMAN</td><td>TRIM8_HUMAN</td><td>P22681</td><td>Q9BZR9</td><td>0.093</td><td>BioGRID|binary|P22681-Q9BZR9|22493164|9606-9606|9606-9606
IntAct|binary|P22681-Q9BZR9|22493164|9606-9606|9606-9606</td></tr>
<tr><td>GPC4_HUMAN</td><td>IQCB1_HUMAN</td><td>O75487</td><td>Q15051</td><td>0.090</td><td>BioGRID|binary|O75487-Q15051|21565611|9606-9606|9606-9606
IntAct|spoke|O75487-Q15051|21565611|9606-9606|9606-9606</td></tr>
<tr><td>ENOG_HUMAN</td><td>MAX_HUMAN</td><td>P09104</td><td>P61244</td><td>0.088</td><td>BIND|binary|P09104-P61244|12808131|9606-9606|9606-9606</td></tr>
<tr><td>RFA3_HUMAN</td><td>TAGL2_HUMAN</td><td>P35244</td><td>P37802</td><td>0.090</td><td>BioGRID|binary|P35244-P37802|24332808|9606-9606|9606-9606</td></tr>
<tr><td>NOP56_HUMAN</td><td>EF1A1_HUMAN</td><td>O00567</td><td>P68104</td><td>0.103</td><td>DIP|matrix|Q12460-P02994|11805837|4932-4932|559292-559292
BioGRID|binary|O00567-P68104|12777385|9606-9606|9606-9606</td></tr>
<tr><td>UBC_HUMAN</td><td>PLCH1_HUMAN</td><td>P0CG48</td><td>Q4KWH8</td><td>0.098</td><td>BioGRID|binary|P0CG48-Q4KWH8|21139048|9606-9606|9606-9606
BioGRID|binary|P0CG48-Q4KWH8|25015289|9606-9606|9606-9606</td></tr>
<tr><td>RFIP4_HUMAN</td><td>TS101_HUMAN</td><td>Q86YS3</td><td>Q99816</td><td>0.214</td><td>BioGRID|binary|Q86YS3-Q99816|22348143|9606-9606|9606-9606</td></tr>
<tr><td>CBX7_HUMAN</td><td>HS71A_HUMAN</td><td>O95931</td><td>P0DMV8</td><td>0.097</td><td>DIP|matrix|Q8VDS3-P0DMV8|20543829|10090-9606|10090-9606</td></tr>
</table>

BUB3_HUMAN	TCPB_HUMAN	O43684	P78371	0.087	BioGRID|binary|O43684-P78371|26186194|9606-9606|9606-9606
NHRF3_HUMAN	S22A4_HUMAN	Q5T2W1	Q9H015	0.131	BioGRID|binary|Q5T2W1-Q9H015|14531806|9606-9606|9606-9606
PLIN3_HUMAN	MPRD_HUMAN	O60664	P20645	0.385	BIND|binary|O60664-P20645|9590177|9606-9606|9606-9606 BioGRID|binary|O60664-P20645|9590177|9606-9606|9606-9606
RNH2A_HUMAN	ITLN1_HUMAN	O75792	Q8WWA0	0.088	BioGRID|binary|O75792-Q8WWA0|26186194|9606-9606|9606-9606
SAMD1_HUMAN	SH3K1_HUMAN	Q6SPF0	Q96B97	0.095	BioGRID|binary|Q6SPF0-Q96B97|19531213|9606-9606|9606-9606
PROF1_HUMAN	ASB2_HUMAN	P07737	Q96Q27	0.093	BioGRID|binary|P07737-Q96Q27|24337577|9606-9606|9606-9606

We found all 20 interactions as reported in their respective supporting publications. For only one interaction (Q5T2W1/Q9H015), the experiments were performed with mouse proteins, while the rest 19 were correctly labeled as having evidence from experiments with human proteins.

Supplementary Note 13 | A discussion of protein-protein interaction networks versus other functional genomics networks.

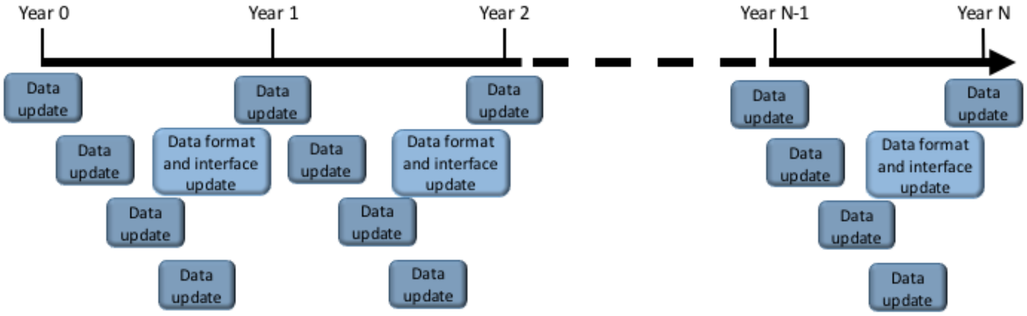
If the objective of an analysis is to define all possible functional associations between a gene of interest and other genes, it can increase coverage to include many types of functional association data. However, such networks also have a more ambiguous interpretation in terms of the molecular biology and biochemistry underlying their gene-gene relationship. Moreover, the point of many network analyses, particularly those where the network data is used to augment and interpret genetic datasets, is to inform specific follow-up experiments and not only to annotate any potential functional association between genes of interest. In these cases it can be an advantage to constrain the network building to physical protein interactions because if an interesting network is identified it will be immediately clear that a follow up protein-protein interaction experiment based on interesting network nodes, is a way to validate and expand the network in question to get added insight into its molecular biology. A more detailed discussion can be seen in Ref. 14.

Supplementary Note 14 | Roadmap for future updates and information about data access

To be maximally useful as a resource to the community we here provide a roadmap for the update of InWeb_IM as well as an overview of the different file formats and resources from where the data can be accessed.

In the past we have updated the network on a quarterly basis and this practice will continue in the future. Upon completion of an InWeb_IM update, the new version will be made available for download for the academic community in standard formats (PSI-MI TAB) – see URLs below. Previous versions of InWeb_IM with build dates and release notes will be available from an archive. The graphical user interfaces from which the data can be accessed will always be updated with the most recent version of the InWeb_IM data (see Supplementary Figure XX for a schematic).

It is not unlikely that new future use cases for InWeb_IM may necessitate update of data formats, user interface and even further improvements of the underlying framework for constructing InWeb_IM. It is therefore also the ambition to continue the development of formats, interface and methods, we expect to make major releases of the InWeb_IM interface on annual basis.



Supplementary Figure 6 | Roadmap for updates of the InWeb_IM data.

Data access:

The raw data can be accessed and downloaded from:

https://www.intomics.com/inbiomap/

This resource includes current version, data archive (all in PSI-MI TAB format), and release notes.

The data can also be accessed from a graphical user interface:

http://apps.broadinstitute.org/genets#InWeb_InBiomap

Supplementary References

1. Brown, K.R. & Jurisica, I. Online predicted human interaction database. *Bioinformatics* **21**, 2076-2082 (2005).

2. Brown, K.R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol* **8**, R95 (2007).

3. Calderone, A., Castagnoli, L. & Cesareni, G. mentha: a resource for browsing integrated protein-interaction networks. *Nat Methods* **10**, 690-691 (2013).

4. Razick, S., Magklaras, G. & Donaldson, I.M. iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics* **9**, 405 (2008).

5. Cowley, M.J. et al. PINA v2.0: mining interactome modules. *Nucleic Acids Res* **40**, D862-865 (2012).

6. Das, J. & Yu, H. HINT: High-quality protein interactomes and their applications in understanding human disease. *BMC Syst Biol* **6**, 92 (2012).

7. Kandasamy, K. et al. NetPath: a public resource of curated signal transduction pathways. *Genome Biol* **11**, R3 (2010).

8. Croft, D. et al. The Reactome pathway knowledgebase. *Nucleic Acids Res* **42**, D472-477 (2014).

9. Kutmon, M. et al. WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Res* (2015).

10. Lage, K. et al. A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat Biotechnol* **25**, 309-316 (2007).

11. Lawrence, M.S. et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501 (2014).

12. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet* **44**, 580-585. (2013).

13. Sanders, S.J. et al. Insights into Autism Spectrum Disorder Genomic Architecture and Biology from 71 Risk Loci. *Neuron* **87**, 1215-1233 (2015).

14. Lage, K. Protein-protein interactions and genetic diseases: The interactome. *Biochim Biophys Acta*. **10**, 1971-1980 (2014).