Contents

Extended Experimental Procedures and Analyses

Extended Experimental Procedures and Analyses

Section 1. Assembly of literature PPI datasets

Literature datasets: We generated two datasets from literature-curated protein-protein interactions. A first dataset was generated in 2010 and used for all experiments, concomitantly with our mapping experiment, and a second dataset was extracted in 2013 to provide an updated version for all computational analyses.

Obtaining the Lit-2010 dataset: The Lit-2010 dataset extracts human protein-protein interactions (PPIs), annotated through December 2010, from seven primary source databases: BIND (Bader et al., 2003), BioGRID (Chatr-aryamontri et al., 2013), DIP (Salwinski et al., 2004), HPRD (Prasad et al., 2009), MINT (Licata et al., 2012), IntAct (Kerrien et al., 2012) and PDB (Berman et al., 2000). For each reported PPI the interacting proteins were mapped to UniProt protein identifiers and then converted to NCBI Entrez gene ID pairs using an ID mapping table downloaded on January 12, 2012 from uniprot.org. Information about the specific publications reporting each interaction was retained and reported interactions that did not have an associated PubMed ID (PMID) were not included in the Lit-2010 dataset.

Assignment of experimental method: We adopted the interaction detection methods in the PSI-MI 2.5 (Proteomics Standards Initiative Molecular Interactions) ontology (Hermjakob et al., 2004; Kerrien et al., 2007). Not all of the source databases precisely follow PSI-MI 2.5, therefore the annotation of each PPI to a PSI-MI method had to handled differently for each database. Briefly:

1. The Biomolecular Interaction Network Database (BIND) has not been updated since 2005 (Bader et al., 2003), before the current PSI-MI ontology was fully implemented (Hermjakob et al., 2004; Kerrien et al., 2007). A subsequent effort (Isserlin et al., 2011) mapped the 2005 version of BIND into PSI-MI 2.5 and from that update we kept only the PPI entries that included an associated PubMed ID and an associated PSI-MI ID.

2. The Biological General Repository for Interaction Datasets (BioGRID) does not follow the full PSI-MI 2.5 ontology, but reports its own system of "experimental evidences codes" to support each reported physical interaction (Chatr-aryamontri et al., 2013). A correspondence table from BioGRID methods to PSI-MI IDs is provided on the BioGRID website and we followed this. Any PPI record that could not be converted to a PSI-MI ID was excluded.
3. The Database of Interacting Proteins (DIP) (Salwinski et al., 2004) follows standard PSI-MI 2.5 ontology, so parsing was straightforward.
4. The Human Protein Reference Database (HPRD) (Prasad et al., 2009) does not follow PSI-MI 2.5 ontology but instead implements a system of three experiment types (yeast 2 hybrid, in vivo, in vitro). When one specific interaction and the associated PubMed ID was also present in another interaction database that follows PSI-MI standards (DIP, MINT, IntAct) the interaction was annotated with the same PSI-MI. For the interactions reported only by HPRD: (i) if the experiment type was "in vivo" or "in vitro" the interaction was discarded as these are not valid experimental interaction detection methods. (ii) if the experiment type provided was "yeast 2 hybrid" the PPI was assigned to MI:0018 "two hybrid".
5. The InterAction (IntAct) database (Kerrien et al., 2012) follows standard PSI-MI 2.5 ontology, so parsing was straightforward.
6. The Molecular INTeraction (MINT) database (Licata et al., 2012) follows standard PSI-MI 2.5 ontology, so parsing was straightforward.
7. The Protein Data Bank (PDB) (Berman et al., 2000) is a protein structure database, not a protein interaction database. PDB does however contain numerous structures of multiprotein complexes, and from these structures binary interactions and indirect associations among the constituent components can be inferred (De Las Rivas and Fontanillo, 2010). A PDB dataset of human PPIs was created following these criteria: (i) all complexes should include only human proteins; (ii) each complex should contain at least two distinct human proteins, obtained by mapping the protein chains to UniProt IDs; (iii) small peptide fragments (annotated as "peptide or with length ≤10 amino acid residues") were not considered as proteins so structures containing these were excluded. Into this PDB set we integrated information about protein-protein interactions provided by the PDBsum database (Laskowski, 2009), which contains detailed information about the specific protein-to-protein interfaces and residue-residue contacts found between the proteins present in each structure. All the binary interfaces found for specific pairs of human proteins were assigned to experimental detection method MI:0114 "x-ray crystallography".

Any interaction not associated with a valid PSI-MI experimental interaction detection method (Hermjakob et al., 2004; Kerrien et al., 2007) was removed. Genetic or protein-DNA interactions reported in some databases (Bader et al., 2003; Chatr-aryamontri et al., 2013; Isserlin et al., 2011) were also removed. Records with the uninformative top-level terms of the PSI-MI experimental interaction detection ontology, "interaction detection method" (MI:0001) and "molecular interaction" (MI:0000) were also discarded. The starting full Lit-2010 dataset comprises 62,163 human protein interactions validated by at least one experimental method and reported in at least one article indexed in PubMed.

Identification of binary interactions: We divided Lit-2010 into the PPIs reported by systematic high-throughput binary human interactome mapping efforts (Rual et al., 2005; Stelzl et al., 2005; Venkatesan et al., 2009) and those detected in small- or medium-scale experiments. A small number of PPIs that had been detected in both systematic and other

studies could appear in both datasets. Removing the PPIs only seen in systematic studies resulted in a dataset of 56,743 human PPIs.

Next we attempted to distinguish binary interactions (direct biophysical contact between two proteins) (Braun, 2012; Braun and Gingras, 2012) from indirect associations (associations between two proteins that are in the same complex, but may or may not directly interact) (Walzthoeni et al., 2013). We evaluated each experimental interaction detection method in the PSI-MI 2.5 and classified them as binary, that is, primarily detects binary interactions, versus indirect, that is, primarily detects association of proteins within a complex (Table S1C). Where an experimental method could be viewed as either, depending on the specific experimental implementation then the method was conservatively classified as indirect. Fewer methods were classified as binary here than in previous (Cusick et al., 2009; Yu et al., 2008) or parallel (Das and Yu, 2012) efforts to ensure the highest confidence binary Lit dataset possible.

After parsing all PPI data from the source databases we obtained a binary human dataset of 13,962 PPIs that contained at least one piece of binary evidence supporting each PPI (there could be other pieces of experimental evidences that were either binary or indirect) and a non-binary dataset containing 42,781 PPIs for which **n**one of the experimental methods are **b**inary (Lit-NB-10).

A paper curated independently by two or more different PPI databases is commonly annotated to different PSI-MI terms, generally to terms of different depth on the same branch of the PSI-MI ontology tree (Turinsky et al., 2010, 2011). If not corrected for, these annotations would count as two or more pieces of evidence for the PPI, when actually there is only one piece of supporting evidence. For example, a yeast two-hybrid experiment might be annotated to the deeper term “two hybrid prey pooling approach” (MI:1112) by one PPI database but to the parent term “two hybrid” (MI:0018) by another database; a co-immunoprecipitation (co-IP) experiment might be annotated to the deeper term “anti-tag co-immunoprecipitation” (MI:0007) by one database but to the parent term “affinity chromatography technology” (MI:0004) by another. To compensate for variability in the annotated methods, when the same paper with the same PMID had different MI terms in two databases, we reassigned the deeper term “up” to the corresponding parent binary or non-binary term on the same PSI-MI branch. In the examples given, the two Y2H annotations collapse to the single ID MI:0018, while the two co-IP annotations collapse to the single ID MI:0004.

The binary human dataset was next separated into “**b**inary **m**ultiple” (Lit-BM-10) (Table S1A), containing all interactions supported by two or more pieces of experimental evidence, at least one of which was binary (4,906 PPIs); versus “**b**inary **s**ingle”, containing all interactions supported by exactly one piece of binary experimental evidence (Lit-BS-10) (9,056 PPIs).

Updating the Lit dataset to 2013: To construct Lit-2013 (Figure S1B and Table S1B) we downloaded, on August 5, 2013, the updated curated PPI content of the same seven PPI databases used for Lit-2010. We applied the same protocols for assignment of experimental identification methods as applied for Lit-2010, with the following modifications:

1. Since PDB structures are now annotated with method descriptions and PubMed identifiers, we considered these instead of simply assigning the “x-ray crystallography” experimental detection method (MI:0114) as we did for construction of Lit-2010. Since PSI-MI identifiers are not provided for experiment descriptions in PDB, we manually assigned them as follows:

PDB experiment description	PSI-MI method description
electron microscopy	electron microscopy, MI:0040
x-ray diffraction	x-ray crystallography, MI:0114
neutron diffraction	neutron diffraction, MI:0893
theoretical model	*not considered*
solution nmr	solution state nmr, MI:1103
solution scattering	light scattering, MI:0067

2. PPIs from the recently published Co-Frac study (Havugimana et al., 2012) were classified as systematic interactions and accordingly were separated from Lit-2013 as earlier systematic datasets (Rual et al., 2005; Stelzl et al., 2005; Venkatesan et al., 2009; Yu et al., 2011) had been separated for Lit-2010.
3. For interactions annotated with the uninformative interaction method description "experimental interaction detection" (MI:0045), if another method annotation is found for the same interaction and from the same publication in any of the seven databases, the MI:0045 method annotation is discarded and the alternative method annotation is applied; otherwise it is retained.
4. Between 2010 and 2013, publications reporting proteome-scale mapping of post-translational modifications of proteins (including ubiquitylation, neddylation and sumoylation) have been curated into the BioGRID protein interaction database (Chatr-aryamontri et al., 2013), resulting in thousands of reported protein-protein interactions involving the gene products of the *UBC*, *SUMO1*, *SUMO2*, *SUMO3*, *SUMO4* or *NEDD8* genes as one of the interaction partners. Post-translational protein modifications such as ubiquitylation form a covalent bond between proteins, which does not comply with the conventional definition of physical protein-protein interactions. Moreover, more than half of the human proteome is linked to at least one out of these six proteins, raising questions about the biological specificity of these links. The interaction identification methods provided by source databases cannot discriminate protein modifications from legitimate protein interactions of these six genes. Therefore, we removed from Lit-2013 all interactions involving any of the six gene products, regardless of which database was the source of the interaction.
5. In the PSI-MI formatted dataset downloaded from HPRD experimental methods are not resolved to individual publications. Instead, all method types (yeast 2 hybrid, in vivo, in vitro) available for a particular interaction are ambiguously associated with all publications reporting the interaction. If there is a single experimental type annotation, we assumed that this experiment was used in all provided publications. Otherwise, we collapsed all PubMed identifiers provided by HPRD to a single piece of evidence.
6. We obtained the year of publication for all papers curated into interaction databases through querying the E-Utility application-programming interface of NCBI (http://eutils.ncbi.nlm.nih.gov/entrez/eutils).

Note on Lit-BM-10 and Lit-BM-13: We observed that 575 interactions found in Lit-BM-10 are no longer in Lit-BM-13. Due to the modifications mentioned above and re-annotation by the source databases, 418 (73%) of these pairs were reclassified in Lit-2013 and 157 pairs are not part of the Lit-2013 dataset.

Section 2. Assessing the quality of literature PPI datasets

Assessing the biophysical quality of literature PPI datasets: We measured the overall quality of Lit-2010 binary datasets by testing samples of heterodimers from Lit-BM-10 and Lit-BS-10 as well as all 698 RRS pairs by both Y2H and MAPPIT assays (Figure 1A). The Lit-BM-10 and Lit-BS-10 pairs were mapped to the hORFeome v5.1 collection (Yang et al., 2011) to identify the available ORFs and PPI pairs were selected at random. Overall, 460 pairs from Lit-BM-10 were tested in the Y2H and MAPPIT assays and 153 pairs from Lit-BS-10 were tested in MAPPIT, out of which 99 pairs were also assayed in Y2H (Table S2A). The same 460 pairs randomly selected from Lit-BM-10 were also used in subsequent analyses.

Co-occurrence of Lit-2010 binary pairs in the literature: To obtain the fraction of Lit-BS-10, Lit-BM-10 and RRS protein pairs that are not found to co-occur in scientific publications, we used the "text mining" score from the STRING database (version 9.05) (Franceschini et al., 2013) (Table S1A). The score scales positively with the frequency and closeness within the text of a publication of co-occurrence for pairs of genes found together in scientific publications. STRING assigns a score of zero to pairs that are not found to co-occur. We assumed a score of zero for pairs not listed in STRING at all. To avoid an inflation of the score from pairs cited many times in the same paper or pairs cited in many papers, we calculated the fraction of RRS, Lit-BS-10 and Lit-BM-10 PPIs with a score of zero and at least one (Figure 1A). Homodimers were excluded.

Some of the differences observed between Lit-BS-10 and Lit-BM-10 may be because some Lit-BM-10 are reported in several publications and the same publications reporting the interaction may have been used by STRING to calculate the co-occurrence score. To correct for this potential circularity, we repeated the analysis after excluding from the binary Lit-2010 dataset all interactions that come from two or more publications. While Lit-BS-10 did not change (as per definition), the Lit-BM-10 dataset now contained only 1,992 interactions that come from a single publication and have been detected with two or more methods. We observed a similar trend for this reduced dataset (Figure S1C).

Section 3. Generation of a new systematic binary interaction map

Comparison of first and second-generation interaction mapping approaches: An empirically-controlled framework was introduced to assess the quality of interactome maps (Venkatesan et al., 2009). This framework is based on four parameters: (i) the *completeness*, the fraction of all pairwise protein combinations tested, (ii) the *assay sensitivity*, the fraction of all true biophysical interactions that are identifiable by a given assay, (iii) the *sampling sensitivity*, the fraction of identifiable interactions that are detected in the experiment and (iv) the *precision*, the fraction of reported pairs that are true positives. To generate a more "mature" second-generation map we used three different binary orthogonal assays to assess the precision of our dataset (larger sample of ~800 pairs) against larger reference sets (~500 PRS and ~700 RRS pairs). Altogether, this represents a 15x increase in size of the validation experiment with respect to our previous efforts (Venkatesan et al., 2009). We also increased the coverage of the human interactome by further improving the three first parameters.

To increase completeness, we screened all pairwise combinations of ORFs in our hORFeome v5.1 (http://horfdb.dfci.harvard.edu/) (Table S2B). This search space, Space II, is 3.1 times larger than Space I, the space corresponding to all pairwise combinations of ORFs in our hORFeome v1.1 collection (Rual et al., 2004) that was screened to generate the HI-I-05 map (Rual et al., 2005).

To increase assay sensitivity, we performed the Y2H assay in different strain backgrounds than the ones used in the first generation approach (Rual et al., 2005) changing from MaV103 and MaV203 strains to Y8800 and Y8930 strains. We compared the recovery

of the 92 PRS and RRS pairs described in Venkatesan *et al*. (Venkatesan et al., 2009) in both MaV and Y strains using our original Y2H pipeline, mating on solid media. Without increasing the rate of RRS detection, the PRS recovery increased by 1.3 times (15 positive PRS interactions out of 92 successfully tested in MaV strains, and 19 positive out of 90 successfully tested in Y strains).

We performed two independent screens of the search space (Space II) to increase sampling sensitivity by 1.5 times (Venkatesan et al., 2009). The observed gain in sampling sensitivity was higher than expected (1.6x) given that the two screens combined generated 13,427 heterodimeric PPIs while one single screen generated on average 8,338 PPIs.

Preparation of Open-reading frame clones: Open-reading frame (ORF) clones encoding human proteins were obtained by PCR-based Gateway recombinational cloning following a protocol previously described (Rual et al., 2004). The hORFeome v5.1 (http://horfdb.dfci.harvard.edu/) collection of ~15,500 ORFs (Yang et al., 2011) (Table S2B) representing ~13,000 genes encompasses the hORFeome v1.1 and v3.1 sets (Lamesch et al., 2007; Rual et al., 2004) in addition to ~3,200 novel ORFs transferred from MGC cDNA clones (Gerhard et al., 2004; Strausberg et al., 1999; The MGC Project Team, 2009). Briefly, to generate Entry clones, ORFs were PCR-amplified with KOD HotStart Polymerase (Novagen) with ORF-specific forward and reverse primers containing attB1.1 and attB2.1 recombination sites respectively. PCR products were then transferred into pDONR223 by a Gateway BP reaction, followed by transformation into chemically competent *E. coli* DH5α cells and selection for spectinomycin resistance. The identity of the Entry clones was verified by Sanger end-read sequencing (5' and 3').

Preparation of Y2H bait and prey libraries: ORFs from the hORFeome v5.1 collection were transferred by Gateway recombinational cloning (Invitrogen) into Y2H destination vectors pDEST-DB and pDEST-AD-*CYH2* to generate Gal4 DNA binding domain and Gal4 activation domain hybrid proteins (DB-ORF and AD-ORF, respectively) as described previously (Dreze et al., 2010).

Yeast strains and transformation: Competent yeast strains Y8800, mating type *MAT***a,** and Y8930, mating type *MAT*α, both harboring the genotype *leu2-3,112 trp1-901 his3*Δ*200 ura3-52 gal4*Δ *gal80*Δ *GAL2::ADE2 GAL1::HIS3@LYS2 GAL7::lacZ@MET2cyh2*^{R}, were transformed with individual AD-ORF and DB-ORF constructs respectively and plated onto selective synthetic complete (SC) solid media without tryptophan (SC-Trp) for AD-ORF or without leucine (SC-Leu) for DB-ORF transformants (Dreze et al., 2010).

Auto-activator identification and removal: Prior to Y2H screening, diploid DB-ORF yeast strains were tested for auto-activation of the *GAL1::HIS3* reporter gene in the absence of any AD-ORF plasmid as described (Dreze et al., 2010). Individual DB-ORF yeast strains were mated with the Y8800 yeast strain transformed with empty pDEST-AD-*CYH2* vector. Diploid cells were first selected on solid SC-Leu-Trp media and then transferred onto solid SC media lacking leucine, tryptophan and histidine and containing 1mM 3AT (SC-Leu-Trp-His+3AT). Any DB-ORF yeast strains that grew on SC-Leu-Trp-His+3AT solid media were considered auto-activators and removed from the collection of DB-ORF yeast strains to be screened.

Y2H screening: First-pass Y2H screening was carried out essentially as described (Dreze et al., 2010; Rual et al., 2005). Fresh overnight cultures of individual Y8930:DB-ORF yeast strains were mated against Y8800:AD-ORF mini-libraries containing 188 different Y8800:AD-ORF yeast strains. After overnight growth at 30ºC on solid rich medium (YEPD), mated yeast cells were replica-plated onto solid SC-Leu-Trp media to select for diploids. After overnight incubation at 30ºC diploid yeast cells were replica-plated onto SC-Leu-Trp-

His+1mM 3AT solid media to select for activation of the *GAL1::HIS3* reporter gene. In parallel, diploid yeast cells were transferred onto SC-Leu-His+1mM 3AT solid media supplemented with 1mg/l cycloheximide (CHX). All AD-ORF plasmids carry the counter-selectable marker *CYH2,* which allows selection on CHX-containing medium of yeast cells that do not contain any AD-ORF plasmid in order to identify spontaneous DB-ORF auto-activators (Dreze et al., 2010). As described (Dreze et al., 2010), all first-pass screening plates were replica-cleaned after 14-18 hours growth on selective media. Five days post replica-cleaning, genuine positive yeast strains that grew on replica-cleaned SC-Leu-Trp-His+3AT media but not on SC-Leu-His+3AT+CHX media were then processed to determine the identity of the respective bait and prey proteins.

Yeast colony PCR and interaction sequence tag (IST) sequencing: Since each DB-ORF yeast strain was mated against a mini-library of 188 AD-ORF yeast strains in the first-pass screens, it is possible to obtain more than one interaction per mini-library. To account for this we picked three colonies (primary positives) from each growth spot on SC-Leu-Trp-His+3AT media. First-pass positive colonies were processed as described (Dreze et al., 2010) to generate lysates for PCR. One microliter of diluted lysate was used as a template for PCR amplification to generate DB-ORF and AD-ORF PCR amplicons. The individual PCR amplicons were either sequenced individually by end-read sequencing (Rual et al., 2005), or 'stitched' together into a single amplicon and sequenced using the Roche 454FLEX next-generation sequencing technology as described (Yu et al., 2011).

IST identification: Sanger sequencing reads were matched to hORFeome v5.1 database using BLASTN with default parameters except that e-value cutoff was set to 1×10^{-15}. If both DB and AD reads of the same positive colonies passed the cutoff, the best hits of the DB and AD reads were assembled into ISTs. Next generation sequencing of PCR amplicons pooled from ~15,000 yeast colony lysate PCRs was carried out using the Roche 454FLEX platform essentially as described (Yu et al., 2011). 454FLEX reads containing the 82-bp linker sequence plus at least 10 bases of ORF-specific sequences on both side of the linker were retained. For each read, after removal of the linker sequence, two flanking sequences were aligned to hORFeome v5.1 database using BLASTN with an e-value threshold of 1×10^{-3}. The AD versus DB orientation was determined based on the orientation of the linker sequence. The best hits of the DB and AD sequences from the same 454FLEX reads were combined to produce ISTs.

When a DB or AD sequence in an IST from either Sanger or 454FLEX sequencing could not be unambiguously assigned to a single ORF in hORFeome v5.1 due to high sequence similarities between closely related proteins or due to the presence of multiple isoforms for some genes, we provisionally assigned that IST to all possible ORF matches and subjected them all to the next step individually.

Pairwise tests of protein pairs detected in the first-pass screen: The growth phenotype of all first-pass pairs was tested in individual pair-wise tests (Dreze et al., 2010). To ensure unambiguous IST sequence verification of the positive pairs, the pairwise tests were divided into 16 matrices with ORFs having a high degree of sequence similarity assigned to different matrices. We also developed a liquid mating strategy with direct spotting of diploid yeast cells, which reduces costs and labor by eliminating the need for multiple sets of agar plates and the replica-cleaning and replica-plating steps used previously (*Arabidopsis* Interactome Mapping Consortium, 2011; Rual et al., 2005; Yu et al., 2008). For liquid mating, fresh overnight cultures of individual yeast strains were mated in 100µl of YEPD using liquid handling robotics. After overnight incubation at 30ºC, 2µl of mated yeast culture were transferred to 100µl of SC-Leu-Trp media to enrich for diploids. After overnight growth at

30ºC, a 5µl aliquot of the liquid culture was robotically spotted, in quadruplicate, onto both SC-Leu-Trp-His+3AT and SC-Leu-His+3AT+CHX solid media. After an incubation of three days, diploids that gave rise to growth on SC-Leu-Trp-His+3AT in at least three out of four replicates and failed to grow at all on SC-Leu-His+3AT+CHX were classified as pairwise positives. To confirm identity of the pairwise positive interactors, colonies were picked and processed for sequencing using the Roche 454FLEX system as above and aligned to the corresponding ORFs in the respective 16 matrices.

Results of the first-pass screens: We transferred all 15,517 ORFs from the human ORFeome v5.1 collection, corresponding to 12,807 distinct genes, into both AD and DB Y2H vectors. We identified and removed 1,657 auto-activating DB-ORFs, leaving a total of 13,860 DB-ORFs representing 11,221 genes. After first-pass screening the search space (13,860 DB-ORFs against 15,517 AD-ORFs) twice, ~85,000 colonies scored positive for growth on SC-Leu-Trp-His+3AT but not on SC-Leu-His+3AT+CHX. After PCR and sequencing we obtained a total of 35,195 unique ISTs. Due to the inclusion of additional isoforms and paralogs this expanded to a total of 66,468 unique ORF pairs, representing 35,536 unique protein pairs. After quadruplicate pair-wise testing, all positive pairs were picked and verified by DNA sequencing, resulting in a set of 14,687 unique verified interacting protein pairs (Figure 1D).

Homodimer specific screen: Stitched PCR amplicons of homodimeric interacting pairs are more difficult to generate. We tested all 13,860 DB-ORFs against their corresponding AD-ORF fusion protein individually by liquid mating and Sanger sequencing of the individual DB-ORF and AD-ORF amplicons. The 640 interacting pairs identified in the primary screen were verified using the same rigorous pipeline, resulting in 561 homodimeric ORF pairs corresponding to 515 unique protein pairs, 18 of which had been found in the main screen.

Section 4. Validation of binary interactions by orthogonal assays

Preparation of assay reagents: The human ORFs corresponding to the proteins to be tested were transferred by Gateway LR recombinational cloning (Invitrogen) into the appropriate destination vectors for each assay (MAPPIT: pMG1-X and pSEL+2L-X; wNAPPA: pIX-GST and pIX-3xHA; PCA: pF1N and pF2N) and transformed into *E. coli*. After selection of transformants in liquid terrific broth medium containing the appropriate antibiotic selection markers, plasmid DNA was extracted and purified using Qiagen 96 Turbo kits (Qiagen) on a BioRobot 8000 (Qiagen). DNA concentration was measured by PicoGreen (Invitrogen) and normalized as required by each assay.

Mammalian protein-protein interaction trap (MAPPIT): The MAPPIT experiments were performed as described previously (Braun et al., 2009; Eyckerman et al., 2001) with minor modifications. HEK293T cells were grown in 384-well plates and co-transfected with a luciferase reporter and plasmids for both bait and prey fusion proteins. Twenty-four hours post-transfection, cells were either stimulated with ligand (erythropoietin) or left untreated, then incubated for an additional 24 hours before luciferase activity was measured in duplicate. Protein pairs tested in MAPPIT were defined to be valid if:

1. Both bait and prey were successfully cloned into expression vectors, and
2. Bait expression were detected.

For each tested pair the fold-induction value (signal from stimulated cells divided by signal from unstimulated cells) was calculated and compared to the fold-induction value of control wells. Pairs were scored positive at some threshold if the fold-induction value of the bait-prey combination divided by the fold induction value obtained with the irrelevant bait-prey or of the irrelevant prey-bait are both above the indicated threshold.

Well-based nucleic acid programmable protein array (wNAPPA): wNAPPA assay was performed as described previously (Braun et al., 2009; Ramachandran et al., 2008) with minor modifications. Bait and prey fusion proteins were co-expressed using rabbit reticulocyte lysates (Promega) according to the manufacturer's instructions. The expressed protein mix was then incubated for 2 hours on 96-well GST capture plates (GE Amersham) pre-blocked overnight with 5% non-fat dry milk in PBS. Following protein capture, wells were washed and then incubated with 5% non-fat dry milk/PBS. After blocking for 30 minutes, plates were incubated with mouse anti-HA monoclonal antibody (HA-11, Covance; 1:5,000 dilution in 5% non-fat dry milk/PBS) for 1 hour at room temperature. After further washes with 5% non-fat dry milk/PBS, wells were incubated with anti-mouse-HRP-coupled secondary antibody (Amersham; 1:2,000 dilution in 5% non-fat dry milk/PBS) for 1 hour at room temperature before being washed with PBS and developed with ECL reagent (Pierce). Signal was measured at 425nm in a Spectramax plate reader (Molecular Devices).

Protein pairs tested in wNAPPA were defined to be valid if:

1. Both bait and prey were successfully cloned into expression vectors, and
2. At least 10ng of plasmid DNA for both bait and prey were added to the *in vitro* translation reaction.

Raw intensities were then converted to adjusted intensities by log-transforming the raw intensities and subtracting the average of the log-transformed intensities of the two blank wells on each plate. For plate-based normalization each plate was standardized to the mean of the plate and to the standard deviation (SD) adjusted intensities of the tested protein pairs.

Protein complementation assay (PCA): The PCA experiments were performed as described previously (Braun et al., 2009; Nyfeler et al., 2005) except that PCA vector fusions are made with both F1 and F2 fragments fused to the N-terminal regions of both ORFs. Briefly, CHO-K1 cells were co-transfected with bait and prey plasmids and a control plasmid expressing CFP to identify transfected cells. Eighteen hours post-transfection, cells were washed, treated with trypsin and then up to 10,000 cells analysed by fluorescence-activated cell sorting (FACS).

Protein pairs tested in PCA were defined to be valid if:

1. Both bait and prey were successfully cloned into expression vectors, and
2. More than 3,333 cells were successfully analysed by the FACS instrument.

Plate-based controls included gating the YFP and CFP channels of the FACS instrument such that the average signal of two empty vector controls permitted 1% of the sorted cells to exceed both thresholds, *i.e.* a "background" signal of 1% CFP and 1% YFP. The log of the YFP mean signal was the final raw reporter value for each protein pair. This reporter value was then normalized by plate, to control for variability across plates in vector infection rates in the CHO cells, by standardizing to the mean and SD signal for all tested protein pairs on that plate.

Scoring for MAPPIT, wNAPPA and PCA: For all three assays only valid pairs were analysed. In each assay, a quantitative output was used to titrate the "threshold" signal to an acceptable range. The threshold was set such that any pair scoring above that threshold is considered "positive" and the complement of that set called "negative". The recovery rate measured as positive pairs over tested pairs can be viewed as a function of the score threshold. For any analysis requiring recovery rate calculation, we selected the threshold corresponding to a recovery rate of 1% of the RRS pairs in that assay. Precision was calculated as described (Venkatesan et al., 2009).

Validation of verified PPIs in orthogonal binary assays: To assess the quality of the pairwise tested Y2H pairs, 809 heterodimeric pairs were chosen at random and tested in a

randomly assigned orientation, by MAPPIT, PCA and wNAPPA assays (Table S2C) along with 460 PRS and 698 RRS pairs (Figure 1D).

In the verified set, two proteins, CREB3 and LNX2, had a very high number of interactions (degrees of 626 and 614 respectively), but showed a low validation rate (data not shown). To ensure the quality of the overall dataset all interactions with these two proteins were removed and were not considered further for this study. Combining all the remaining interactions from the main screen and interactions from the homodimer experiment, we obtained a total of 13,994 interactions (HI-II-14, Table S2G).

Section 5. Binary PPIs correspond to direct contacts in PDB structures

Direct contact in PDB structures: To investigate direct contacts in structurally defined complexes, we downloaded all available structures of protein complexes in PDB biological units as of January 26, 2013 (Berman et al., 2000). From this dataset we eliminated:

1. Biological units containing non-protein chains,
2. Biological units containing chains that could not be mapped to a UniProt accession using SIFTS (http://www.ebi.ac.uk/pdbe/docs/sifts/),
3. Biological units containing fragments of less than 30 residues, and
4. NMR structures

To identify proteins ("PDB chains") in direct contact in these PDB complexes, we calculated residue-residue distances. Two proteins were considered to directly interact in a complex if at least one of any of the following residue-residue contacts was observed at the given distance:

1. A disulphide bridge, defined as two sulphur atoms of a pair of cysteines at ≤ 3.0Å,
2. A hydrogen bond, defined as any nitrogen-oxygen atom pair with the two atoms at ≤ 3.4Å, or
3. A non-bonded interaction, defined as any pair of nitrogen and oxygen atoms at ≤ 4.0Å or any pair of carbon atoms at ≤ 4.5Å.

Assessment of direct contact in HI-II-14: To identify proteins from the HI-II-14 dataset present or absent from PDB structures, we mapped Entrez Gene IDs to UniProt IDs using the mapping provided by the UniProt mapping service. For entries mapping to multiple UniProt IDs, we selected only the IDs belonging to the "Human Complete Proteome" (ref keyword: 181) as annotated in UniProt and removed the ones belonging to Trembl as they are not curated. For 125 HI-II-14 proteins, the Entrez Gene ID could not be mapped to any UniProt ID. From the 13,944 interactions in HI-II-14, UniProt IDs could be mapped for a total of 13,670 interactions.

A PDB complex was considered to have been tested in our experiment if at least two of the proteins it contains (chains) are in the HI-II-14 map (Table S2D). Similarly, for each tested PDB complex, pairs of proteins were considered to have been tested if both are in HI-II-14, either interacting or not. Pairs of tested proteins were then classified as direct interactions or indirect associations based on the PDB complexes containing them. A pair was considered involved in direct contact if there was at least one structure of a complex where the two proteins are in direct contact. A pair was considered involved in an indirect association if there was no structure were the two proteins are in direct contact. Interactions in HI-II-14 involving proteins found in direct contact in the PDB were true positives (TP) and those involving two proteins found only in indirect association were false positives (FP). Similarly true negatives and false negatives were defined as pairs of proteins present in HI-II-14 but not interacting and found in the PDB only in indirect associations or in direct contact, respectively.

From 928 complexes considered we could collect 996 direct interactions and 148 indirect associations. Of these 1,144 interactions, 743 were considered as effectively tested in the HI-II-14 experiment. Of these 687 corresponding to direct interactions in the protein complexes and 56 to indirect associations (Table S2D). The results along with *P* values from Fisher's exact tests performed on the corresponding confusion matrices are summarized below.

Description	Tested		Identified		Not identified		*P* value
	Direct	Indirect	TP	FP	FN	TN	
All interactions	687	56	189	2	498	54	1×10^{-5}
Homodimers	574	43	153	1	421	42	7×10^{-5}
Heterodimers	113	13	36	1	77	12	0.106

Out of complexes containing pairs of proteins present in HI-II-14, we distinguished between pairs that are or are not in direct contact with each other within these structures. Out of 191 HI-II-14 interactions for which both proteins appeared in the same PDB complex, 189 were in direct contact in the corresponding crystal structure, demonstrating the high precision of HI-II-14 in reporting direct PPIs (99% observed precision compared to an expected rate of 92%; $P = 4 \times 10^{-5}$, one-sided binomial test).

In general, direct interactions inside tested complexes (996) are more frequent than indirect associations (148) at a ratio of ~7:1. This is most likely due to the fact that small complexes and homomers are more represented in the PDB while larger complexes are mainly composed of multiple copies of smaller subcomplexes.

In HI-II-14, direct interactions (687) are oversampled with respect to indirect associations (56) at a ratio of ~12:1. This partly explains why direct interactions are correctly identified. False positives are depleted in HI-II-14, with a log depletion factor of -2.84:

$$E = \log_2 \frac{\frac{FP}{TP + FP}}{\frac{IND}{DIR + IND}} = \log_2 \frac{\frac{2}{189 + 2}}{\frac{56}{687 + 56}} = -2.84$$

This depletion is highly significant, as assessed by the Fisher's exact test on the confusion matrix ($P = 1.01 \times 10^{-5}$).

Detection of PPIs and interaction strength: To investigate whether direct interactions correctly identified in HI-II-14 (true positives) might correspond to stronger interactions inside complexes than those missed in HI-II-14 although the two proteins are present in the map (false positives), we estimated the strength of interaction by counting the number of residue-residue contacts made at the interface of each pair of interacting proteins. For the 687 unique direct interactions inside complexes considered tested in HI-II-14, hydrogen bonds and non-bonded residue-residue contacts were determined based on the method explained above. When several types of contacts satisfied our criteria for the same residue-residue pair, only one was counted. When an interaction was present in more than one complex structure, we kept the maximum number of contacts observed (Table S2E).

Note on "biological units": In these analyses, when available, we used biological units instead of asymmetric units. While asymmetric units can correspond to artifactual crystallographic contacts, biological units are meant to represent biologically relevant multimeric states. Depending on the PDB structure, these biological units are generated either by combining several asymmetric units through crystallographic symmetry operations or by breaking the asymmetric unit into smaller subcomponents.

The separation of biologically relevant interfaces from those arising from crystallization artifacts is a difficult and long-standing issue. Several programs have been developed to address this problem (Henrick and Thornton, 1998; Krissinel and Henrick, 2007; Ponstingl et al., 2003). While these tools usually agree for most of the complex structures, there can be differences for some entries based on the different principles on which every tool operates (biophysical properties of the interfaces, conservation of interfacial residues, etc). Every tool likely presents incorrect predictions to a certain degree. The error rate of PQS (Henrick and Thornton, 1998) and PISA (Krissinel and Henrick, 2007) has been estimated as 17% and 24%, respectively (Bordner and Gorin, 2008) and the error rate of PDB biological units as 15% (Levy, 2007). Another argument in favour of using PDB biological units is that in addition to software-based predictions (usually performed with PISA), they can also be based on information provided by the authors, which we believe is the most accurate to disentangle any oligomerization ambiguities.

Section 6. Calculation of enrichment in PPI-mediating domain pairs

Pfam domains in Pfam 26.0 were assigned to sequences corresponding to the canonical isoforms in UniProt using the pfam_scan.pl script provided by the Pfam service (http://pfam.sanger.ac.uk). Pfam domains in HI-II-14 proteins were then identified by mapping UniProt accessions to Entrez Gene IDs. Domain-domain interactions supported by structural evidence were extracted from an updated version of the 3did database (Stein et al., 2011) based on Pfam 26.0 and the Protein Data Bank as of January 2013. To calculate the enrichment of HI-II-14 proteins in domains involved in domain-domain interactions, we performed a Fisher's exact test comparing the proportion of proteins with a domain in 3did in the network to the proportion of proteins with a domain in 3did in Space II but absent from HI-II-14. To predict the domain pairs involved in protein interactions, we generated all possible domain pairs between pairs of proteins interacting in HI-II-14, and calculated the observed number of times each domain pair appeared. We repeated the process with 1,000 degree-controlled randomized networks. The observed frequency of each domain pair was compared to the frequencies in the randomized networks through z-score normalization (Table S2F). Homomeric interactions were not included in the randomization and thus contributed similarly to domain-domain frequencies both in the observed and the randomized networks. To avoid the calculation of spurious enrichments, only domains found in at least ten proteins were included in the calculations. Also, only domain pairs observed more than twice were considered for subsequent analyses, and domain pairs for which the z-score was not measurable were discarded (*e.g.* in the absence of a domain pair in all randomized networks). The enrichment in known domain pairs mediating interactions was assessed by comparing the z-score distribution of domain pairs supported by structural evidences versus domain pairs with no structural evidence in a two-sided Wilcoxon rank sum test.

Section 7. Calculation of functional enrichment

For each protein-protein interaction (PPI) network, we measured the overlap in number of interactions with each of the functional networks defined above, discarding homodimeric PPIs. For each pairwise comparison, the PPI and functional networks were trimmed to interactions where both proteins were present in both networks and in the hORFeome v5.1 (Space II). Odds ratios are the result of two-sided Fisher's exact tests testing the enrichment in overlapping interactions over the number of non-overlapping interactions in each compared (trimmed) network, where any pair of proteins present in both networks defined the full space

of detectable interactions. All error bars correspond to 95% confidence intervals. Comparison of co-function was performed using Fisher's exact tests.

Section 8. Interaction perturbation by disease and non-disease variants

Disease variants were obtained from the Human Gene Mutation Database (HGMD 2009 V2) (Stenson et al., 2014) whereas common variants were derived from the 1,000 genomes project (1000 Genomes Project Consortium, 2012). Only variants with a minor allele frequency above 1% in the union of populations studied by the 1,000 genomes project were considered common. To avoid an over-representation of genes with many mutations, up to 4 missense disease variants and 4 common variants were selected per gene. To clone these genetic variants, we applied an enhanced site-directed PCR mutagenesis pipeline (Charloteaux et al., 2011; Zhong et al., 2009), using as template the corresponding wild-type clones in our full-length sequence-verified human ORFeome v8.1 collection (Yang et al., 2011). The resulting PCR products were subsequently cloned into pDONR223 via Gateway BP reaction to create the Entry clone plasmid. BP reaction was conducted in vitro followed by transformation of bacterial cells and growth selection for spectinomycin resistance. Single bacterial colonies were picked for PCR amplification of the insert and subjected to next-generation sequencing. Clones with full-length ORFs containing only the desired single nucleotide change were considered as sequence-verified.

To identify interaction perturbations by these variants, we employed the yeast two-hybrid (Y2H) assay as previously described (Charloteaux et al., 2011; Zhong et al., 2009). All disease variants, common variants, and their respective wild-type ORFs were transferred by Gateway LR reaction into the pDEST-DB vector and were subsequently transformed into the yeast strain Y8930. For each HI-II-14 interaction involving one of the wild-type products studied here, all alleles (disease, common, and wild-type) were tested pairwise for interaction with the corresponding partners as described above (Pairwise tests of protein pairs detected in the first-pass screen). In total, we tested 107 PPIs involving the product of 32 distinct disease-associated genes, comparing 67 disease variants to 48 common variants.

For some of the disease variants the observed phenotype may reflect absence of expression or misfolding/destabilization of the corresponding protein. One way to control for this is to repeat the analysis, restricting it to variants for which at least one of the interactions is maintained, indicating that the corresponding mutant protein is expressed, folded and at least partially functional. As shown below, our observations and conclusions hold when only considering variants that conserved at least one interaction of the wild-type allele (25 genes, 94 interactions tested).

Section 9. Identification of dense and sparse zones in PPI maps

Generation of adjacency matrices: For each gene/protein property we ranked all proteins with respect to either increasing or decreasing values of that property. In nearly all cases, the ranking is decreasing, such that the first ordered proteins are those with high values for the property. The reverse orderings are also indicated in tables and figures. Unless otherwise noted, tied values were sorted randomly and missing (or undetectable) values were pushed to the minimum value for sorting. For any given gene/protein property ranking, we can partition the interactome network space into two regions at any property threshold *t*. The first region contains all pairs of proteins where the gene/protein property of both proteins exceeds *t*. The second region consists of the remaining protein pairs.

Measure of the interaction density imbalance: To measure the imbalance of interaction density of a given PPI map for a given threshold, we compared the fraction of

interactions *observed* in the first region to the fraction of PPIs *expected* given the fraction of the total space covered by the first region, assuming a uniform distribution of the PPIs in the space. The difference between these two fractions (expected minus observed), D_t, measures the deviance from expectation of PPI distribution across the two regions at threshold *t* and lies between -1 and 1. Applying this procedure to all possible thresholds *t*, we can find the threshold *t** for which the deviation from expectation is maximal. The difference between the expected and observed ratios at *t**, is our test statistic *D**. Since our statistical procedure is designed to measure increased or decreased interaction density specifically in the *first* region, both increasing and decreasing rankings were considered for all properties, permitting density imbalance measure with respect to the *second* region.

Section 10. Assessing the quality of Co-Frac, PrePPI-HC and Lit-NB-13 datasets

Assessing the functional relevance: Functional enrichment of Co-Frac, PrePPI-HC and Lit-NB-13 were calculated for the entire datasets as explained in section 7 of the Extended Experimental Procedures.

Assessing the biophysical quality in terms of binary PPIs: To evaluate the extent to which Co-Frac, PrePPI-HC and Lit-NB-13 datasets represent reproducible binary protein-protein interactions, we extracted from these datasets all X-Y pairs in a non-symmetrical space defined by two different lists of 1,896 ORFs. 800 pairs of ORFs selected at random from the same space were used as RRS. Pairs from Lit-BM-13 in the same space were also tested for comparison. For Co-Frac, PrePPI-HC and Lit-NB-13 datasets, each X-Y pair was tested in our Y2H assay in both configurations (DB-X *vs* AD-Y and DB-Y *vs* AD-X) as described in section 3 of the Extended Experimental Procedures. A X-Y pair was scored positive if scored positive in at least one configuration. Pairs not scored in any of the two configurations either because of absence of growth in SC-Leu-Trp media or because the DB hybrid was scored as auto-activator, were discarded from the analysis. In total, we successfully tested and scored 787 RRS pairs, 246 pairs from Lit-BM-13, 1,621 pairs from Lit-NB-13, 330 pairs from Co-Frac, and 431 pairs from PrePPI-HC. Co-Frac, PrePPI-HC, and Lit-NB-13 datasets are recovered at a much lower rate than the sample of Lit-BM-13 and are statistically indistinguishable from random pairs.

Conclusion: Given that Co-Frac and PrePPI-HC datasets are enriched in functional relationships and although these enrichments might partly reflect the use of functional relationships during their generation, these results strongly suggest that Co-Frac and PrePPI-HC are more comparable to non-binary rather than to binary datasets.

Section 11. Prioritization of candidate genes in GWAS loci

From GWAS SNPs to loci: 307 distinct cancer-associated SNPs were identified from 75 GWAS publications covering 10 types of cancer (query SNPs; Table S5A). Linkage-disequilibrium (LD) values (as measured by r^2) between 293 of these (14 were not found in the dbSNP database) and all HapMap Phase III SNPs within a 400 kb window (200 kb upstream and 200 kb downstream) were computed for the HapMap Phase III CEU population (http://www.sanger.ac.uk/resources/downloads/human/hapmap3.html).

For each LD threshold considered, the farthest upstream and downstream SNPs associated with each query SNP were identified such that their correlation with the query SNP was above the selected LD threshold. Each such pair of upstream and downstream SNPs defined a genomic range, and all overlapping ranges were merged to give rise to disjoint genomic loci. The numbers of loci considered were 265 and 246 loci for LD threshold (r^2) of 0.9 and 0.7, respectively.

We next identified the overlaps between these loci and UCSC hg19 (GRCh37) transcripts. Loci with zero overlapping transcripts were removed. For each locus with at least one overlapping transcript, the locus boundaries were redefined to extend to the most extreme upstream or downstream coordinates of all overlapping transcripts. Any of these newly-defined loci that overlapped in coordinates were then subjected to a second round of merging, leading ultimately to disjoint gene sets and a reduced number of gene-containing loci. The numbers of these loci are 142 and 153 loci for LD thresholds of 0.9 and 0.7, respectively.

Fraction of loci with a gene product interacting with Census proteins: For each map and LD threshold combination, we counted the number of loci that contained at least one gene whose product directly interacted with a Cancer Census protein in the PPI map. The number of such loci divided by the number of loci that contained at least one gene whose product appeared in the PPI map provides the fraction of loci with cancer candidates.

To assess significance of this fraction, we measured the corresponding fraction when randomly selecting the same number of genes in each locus that were present in the PPI map. This method was selected over the alternative of choosing a locus at random in the genome and walking in either direction looking for genes so as to guarantee that the genes in each randomized collection of loci will be in the PPI map.

These fractions and corresponding significance levels were measured both when loci already containing a Cancer Census gene were removed or considered in the analysis (Figure S5A), to measure the ability of interactome maps to capture “novel” information in GWAS loci (Figure 7A). Loci already containing Cancer Census genes were considered in the analyses reported in Figure 7B and 7C and Figure S5A.

Section 12. Calculation of cancer association scores

To prioritize genes associated with cancer, we first collected annotated sets of genes known to be associated with cancer and genes that have been identified via differing technologies or approaches as candidates for association with cancer. For the list of genes known to be associated with cancer we used the Census list (Futreal et al., 2004) of 465 genes implicated in carcinogenesis, which effectively serves as our 'gold standard' reference. Our candidate gene lists come from three distinct sources, SB, SM and VT. These cancer gene sets were first trimmed to only include those genes in heterodimeric interactions in HI-II-14. Each gene in the HI-II-14 map was annotated to indicate membership in these sets. For each gene in our network we counted the number of immediate neighbour genes that were members of each set. For example, CDK4 has five neighbours in HI-II-14, three of which are Census members, one is a VT member, one is an SB member and none of which are SM members.

Since any gene with a greater number of neighbours in the network also naturally has a greater number of neighbours appearing in these sets, we then normalized these neighbour count values with a randomization strategy. The HI-II-14 network was edge-randomized while preserving node degree 10,000 times, and neighbour count values were computed for each gene within each network. Empirical *P* values were then computed for each observed neighbour count *x* with respect to the 10,000 random neighbour counts *X* by counting the number *S* of random counts with *X* equal or above *x* and adding a single pseudocount evenly distributed as $P = (S + 0.5) / 10{,}001$.

This approach generated seven 'features' for each gene: membership in the SB, SM and VT lists and along with the four empirical *P* values describing the frequency neighbours of each gene being in the SB, SM, VT, or Census sets. We measured the ability of each feature

to prioritize the known Census genes with separate logistic regression models, one per feature. Models using the binary features reflect a 'guilt-by-profiling' prediction, while use of the neighbour-count data represents a 'guilt-by-association' paradigm. Performance of all models was measured as average precision which captures the precision found for each Census gene in the final predicted ranked list (Manning et al., 2008).

We combined the guilt-by-profiling and guilt-by-association approaches by allowing any of the seven predictor features to be included in a forward stepwise logistic regression model using the Akaike information criterion (AIC) to determine the stepwise halting. The final set of features selected was: the SB, SM and VT guilt-by-profiling and the Census and SB guilt-by-association neighbour-count scores (Figure S6D).

Variance of the average precision value for each model was estimated by a bootstrap resampling approach with 1,000 iterations. Each bootstrap sample was stratified by the Census membership (the response variable) to preserve the prior probability considered for the fitting procedure for each iteration. Comparisons of the bootstrapped average precision samples between pairs of models were performed using Wilcoxon rank sum tests.

“Receiver Operating Characteristic” curves were obtained for each model by measuring at a decreasing score threshold the fraction of known Cancer Census genes recovered and the corresponding fraction of network proteins predicted as candidate cancer genes. We measured the area under the curve (AUC) and the standard error based on a previously described formula (Hanley and McNeil, 1982). Comparison of the AUC of the combined model to other models was based on simulated normal distributions centred on the observed AUC and with corresponding standard error as standard deviations, then compared for significance using Wilcoxon rank sum tests.

Section 13. Assembly of reference datasets

Positive reference sets (PRS) of human PPIs: A sample of 460 heterodimeric pairs was extracted at random from Lit-BM-10 in space II. This subset of Lit-BM-10 pairs was used as positive reference set in the validation experiments.

Random reference set (RRS) of human PPIs: A RRS of 698 pairs was constructed from random selection of heterodimer protein pairs in space II (~82,000,000 unique pairs), excluding pairs previously shown to interact.

HI-I-05: The HI-I-05 map was extracted from our previous map of the human binary interactome (Rual et al., 2005), containing 2,750 interactions after updating the Entrez Gene ID mapping.

Other systematic screens: Networks corresponding to previous systematic screens based on the Y2H assays were extracted and updated to the most recent Entrez Gene ID mapping. The Stelzl *et al.* dataset comprises 3,038 interactions (Stelzl et al., 2005), the Venkatesan *et al.* dataset comprises 197 interactions (Venkatesan et al., 2009) and the Yu *et al.* dataset comprises 1,167 interactions (Yu et al., 2011).

Co-fractionation and prediction maps: The Co-Frac map was extracted from Havugimana *et al.* (Havugimana et al., 2012), and comprised 13,982 interactions after mapping to Entrez Gene IDs from UniProt IDs. The Zhang *et al.* (Zhang et al., 2012) PrePPI-HC map was obtained via personal communication, corresponding to the PrePPI map at a likelihood ratio (LR) threshold of 15,000, and contained 25,403 interactions after mapping to Entrez Gene IDs from UniProt IDs. We mapped UniProt accession numbers (UniProt ACs) to NCBI Entrez Gene IDs using the mapping table provided by UniProt downloaded on January 12, 2012.

Tissue-specific expression data: RNA-sequencing data of 16 human tissues were obtained from the Illumina Human Body Map 2.0 project. The FPKM data were downloaded from the EMBL-EBI Expression Atlas (GXA) with data set identifier: E-MTAB-513. Ensembl gene IDs were translated to NCBI Gene IDs using the Synergizer service (Berriz and Roth, 2008), and FPKM values for each tissue were averaged across multiple gene IDs for cases where a single Ensembl gene mapped to multiple NCBI Gene IDs.

Protein abundance in normal and cancer tissues: Protein abundance data for normal tissue and cancer tissue was downloaded on October 15, 2014 from the Human Protein Atlas website. For the cancer tissue dataset, the protein abundance was chosen from categories “High”, “Medium”, “Low” or “Not detected” by selecting the category that showed the maximum number of samples.

CORUM co-complex membership: The complete set of human co-complex membership from CORUM was downloaded from the MIPS website in February 2013 (http://mips.helmholtz-muenchen.de/genre/proj/corum).

Gene Ontology (GO) terms: Reference GO-based networks were built as described (Pena-Castillo et al., 2008) and the GO database was downloaded in March 2012. GO term-to-gene associations with IPI (inferred from physical interaction), IEA (inferred from electronic analysis) and ND (not described) were not considered. The three GO networks correspond to the three different GO branches “Biological Process” (BP), “Molecular Function” (MF) and “Cellular Component” (CC). Only those terms that mapped to a maximum of 30 genes, after disallowing the evidence codes listed above, were considered to restrain our analyses to relatively specific terms.

Generic GO slim term mapping: The subset of GO Cellular Component (CC) terms in the Generic GO slim provided by the GO Consortium was downloaded on January 9, 2014. The terms “intracellular", "cellular_component", "ribosome", "cell", and "cell wall" were removed from the analysis, and the following terms were considered: plasma membrane (GO:0005886), nuclear chromosome (GO:0000228), cytoplasmic chromosome (GO:0000229), lysosome (GO:0005764), mitochondrion (GO:0005739), vacuole (GO:0005773), extracellular region (GO:0005576), cytosol (GO:0005829), endoplasmic reticulum (GO:0005783), peroxisome (GO:0005777), proteinaceous extracellular matrix (GO:0005578), extracellular space (GO:0005615), nucleus (GO:0005634), nuclear envelope (GO:0005635), cytoplasm (GO:0005737), nucleoplasm (GO:0005654), chromosome (GO:0005694), nucleolus (GO:0005730), endosome (GO:0005768), Golgi apparatus (GO:0005794), lipid particle (GO:0005811), microtubule organizing center (GO:0005815), cytoskeleton (GO:0005856), cilium (GO:0005929), plastid (GO:0009536), thylakoid (GO:0009579), cytoplasmic membrane-bounded vesicle (GO:0016023), external encapsulating structure (GO:0030312), organelle (GO:0043226). Associations between these terms and human NCBI Gene IDs were acquired from the R/Bioconductor org.Hs.eg.db package (version 2.9.0), with evidence codes "IPI", "IEA", and "ND" excluded from consideration. Each term was thus associated with a set of genes that were then trimmed to the space of HI-II-14 proteins.

Mouse phenotypes: Mouse genes associated with specific phenotypes (Eppig et al., 2012) were downloaded from informatics.jax.org on January 24, 2013 and transferred to human genes using orthology maps provided by the same website. An upper-limit of 20 genes per phenotype was applied (as described above for GO terms) for phenotype terms.

Kinases and phosphorylated proteins: The list of kinases and their annotation was extracted from Manning et al. (Manning et al., 2002). The list of known kinase-substrate relationships and phosphorylated proteins were downloaded from PhosphoSitePlus on April

15, 2014 (Hornbeck et al., 2012). Additional evidence for phosphorylated proteins was extracted from Olsen et al. (Olsen et al., 2010), corresponding to phosphorylation events identified during mitosis.

Disease associations: The disease-associated gene lists we extracted are:

1. A set of 125 well-accepted cancer “drivers” genes described in Vogelstein *et al*. (Vogelstein et al., 2013).
2. A set of 485 cancer-associated genes from the Sanger Cancer Gene Census (Futreal et al., 2004), as of March 15, 2013, called the “Census gene” set.
3. A set of 2,854 genes associated to Mendelian disorders in the OMIM database as of October 24, 2012 (Hamosh et al., 2005)
4. A set of 3,230 genes overlapping tight windows defined by SNPs found in GWA studies and catalogued in the NHGRI catalogue as of October 25, 2012 (Hindorff et al., 2009). The boundaries of each window are defined by the genomic region found to be in tight linkage disequilibrium with the reported SNP: $r^2 \geq 0.99$.
5. A set of 1,359 genes recently implicated in tumorigenesis by *in vivo* transposon mutagenesis screens in mice (Mann et al., 2012; March et al., 2011; Starr et al., 2009; Starr et al., 2011), called the “Sleeping Beauty" (SB) set.
6. A set of 920 genes with elevated rates in cancers of somatic mutations predicted to be deleterious, described in Rozenblatt-Rosen *et al*. (Rozenblatt-Rosen et al., 2012), called the “somatic mutation" (SM) set.
7. A set of 947 genes whose products were found to physically interact with proteins from viruses known to be associated with a variety of cancers (Rozenblatt-Rosen et al., 2012) called the “viral targets" (VT) set.

The last three sets were merged to create a list of 2,995 candidate cancer genes from systematic studies. Note that all these lists were restricted to the same space as the one considered for protein properties.

Cancer pathways: The sets of genes annotated as part of the twelve cancer pathways described in Vogelstein et al. (Vogelstein et al., 2013) were retrieved from the GO database, with the following GO identifiers: APC: GO:0005680; Apoptosis: GO:0006915; Cell cycle: GO:0007049; Chromatin modification: GO:0016568; DNA damage control: GO:0031570; HH: GO:0007224; MAPK: GO:0000165; NOTCH: GO:0007219; PI3K: GO:0014065; Ras: GO:0005099; STAT: GO:0007259; TGF-beta: GO:0007179; Transcriptional regulation: GO:0006355. GO term-to-gene associations with IPI (inferred from physical interaction), IEA (inferred from electronic analysis) and ND (not described) were not considered.

Search space considered: Properties were calculated for the genes in the space defined by the combination of the set of genes for which ORFs were available in our latest human ORFeome collection (hORFeome v7.1) and the set of genes for which ORFs were described in the Consensus CDS database (http://www.ncbi.nlm.nih.gov/CCDS/ downloaded October 2012). When the property is protein-based, values were computed for each available ORF separately and then averaged over each set of ORFs that map to a unique Entrez gene ID (Table S4B).

Protein properties examined:

1. mRNA abundance in HEK cells: mRNA expression values extracted from RNA-seq data (FPKM values) in HEK293 cells (Sultan et al., 2008).
2. mRNA abundance in HeLa cells: mRNA expression values extracted from RNA-seq data (FPKM values) in HeLa cells (Nagaraj et al., 2011).
3. Protein abundance in HEK cells: Protein abundance (iBAQ values) in HEK293 cells measured by mass spectrometry (Geiger et al., 2012).

4. Protein abundance in HeLa cells: Protein abundance (iBAQ values) in HeLa cells measured by mass spectrometry (Geiger et al., 2012).
5. Number of publications: Number of publications per gene calculated from the Gene2pubmed file downloaded from NCBI on November 13, 2012.
6. First publication date: Earliest association between a publication and a gene. The Gene2pubmed file was downloaded from NCBI on November 13, 2012. PubMed IDs were then used to retrieve publication dates from PubMed.
7. Number of protein complexes: Number of non-redundant complexes with which a protein is associated (Havugimana et al., 2012).
8. Number of pathways: Number of pathways with which a gene is associated, as catalogued by the MSigDB database (Liberzon et al., 2011) version 3.0, downloaded on January 21, 2011.
9. Number of GO associations: Number of Gene Ontology (GO) terms (The Gene Ontology Consortium, 2000) with which a gene is associated (using only evidence codes EXP, IDA, IMP, IGI and IEP; R / Bioconductor packages org.Hs.eg.db 2.8.0 and GO.db 2.8.0. GO data for these packages was downloaded in March 2012).
10. Fraction of sequence in Pfam domains: Fraction of residues along ORF length within domains identified via InterProScan (Zdobnov and Apweiler, 2001), using the HMMPfam program with default parameters.
11. Fraction of basic amino acids: Fraction of basic residues along ORF length (R, K and H).
12. Fraction of charged amino acids: Fraction of charged residues along ORF length (R, K, H, D and E).
13. Fraction of acidic amino acids: Fraction of acidic residues along ORF length (D and E).
14. Fraction of polar amino acids: Fraction of polar residues along ORF length (G, S, Y, C, N, T and Q).
15. Fraction of polar and uncharged amino acids: Fraction of polar and uncharged residues along ORF length (S, T, N and Q).
16. Fraction of hydrophobic amino acids: Fraction of hydrophobic residues along ORF length (A, V, I, L, M, F, Y and W).
17. Fraction of sequence in transmembrane helices: Fraction of residues along ORF length that are in transmembrane helices, as predicted by the TMHMM2.0 program (Krogh et al., 2001).
18. Number of binding sites: Number of binding regions along ORF length as defined by the ANCHOR algorithm (Meszaros et al., 2009).
19. Fraction of sequence in disordered regions: Fraction of disordered residues along ORF length as defined by the VSL2 algorithm (Obradovic et al., 2003).
20. Number of linear motifs: Number of Eukaryotic Linear Motifs (Dinkel et al., 2012) along the ORF length.
21. ORF length: Number of residues along the ORF length.

Supplemental References

1000 Genomes Project Consortium (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56-65.

Arabidopsis Interactome Mapping Consortium (2011). Evidence for network evolution in an *Arabidopsis* interactome map. Science *333*, 601-607.

Bader, G.D., Betel, D., and Hogue, C.W. (2003). BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. *31*, 248-250.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235-242.

Berriz, G.F., and Roth, F.P. (2008). The Synergizer service for translating gene, protein and other biological identifiers. Bioinformatics *24*, 2272-2273.

Bordner, A.J., and Gorin, A.A. (2008). Comprehensive inventory of protein complexes in the Protein Data Bank from consistent classification of interfaces. BMC Bioinformatics *9*, 234.

Braun, P. (2012). Interactome mapping for analysis of complex phenotypes: insights from benchmarking binary interaction assays. Proteomics *12*, 1499-1518.

Braun, P., and Gingras, A.C. (2012). History of protein-protein interactions: from egg-white to complex networks. Proteomics *12*, 1478-1498.

Braun, P., Taşan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., Yu, H., Sahalie, J.M., Murray, R.R., Roncari, L., de Smet, A.S.*, et al.* (2009). An experimentally derived confidence score for binary protein-protein interactions. Nat. Methods *6*, 91-97.

Charloteaux, B., Zhong, Q., Dreze, M., Cusick, M.E., Hill, D.E., and Vidal, M. (2011). Protein-protein interactions and networks: forward and reverse edgetics. Methods Mol. Biol. *759*, 197-213.

Chatr-aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L.*, et al.* (2013). The BioGRID interaction database: 2013 update. Nucleic Acids Res. *41*, D816-823.

Cusick, M.E., Yu, H., Smolyar, A., Venkatesan, K., Carvunis, A.-R., Simonis, N., Rual, J.F., Borick, H., Braun, P., Dreze, M.*, et al.* (2009). Literature-curated protein interaction datasets. Nat. Methods *6*, 39-46.

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

De Las Rivas, J., and Fontanillo, C. (2010). Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. PLoS Comput. Biol. *6*, e1000807.

Dinkel, H., Michael, S., Weatheritt, R.J., Davey, N.E., Van Roey, K., Altenberg, B., Toedt, G., Uyar, B., Seiler, M., Budd, A.*, et al.* (2012). ELM—the database of eukaryotic linear motifs. Nucleic Acids Res. *40*, D242-251.

Dreze, M., Monachello, D., Lurin, C., Cusick, M.E., Hill, D.E., Vidal, M., and Braun, P. (2010). High-quality binary interactome mapping. Methods Enzymol. *470*, 281-315.

Eppig, J.T., Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and the Mouse Genome Database Group (2012). The Mouse Genome Database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. Nucleic Acids Res. *40*, D881-886.

Eyckerman, S., Verhee, A., der Heyden, J.V., Lemmens, I., Ostade, X.V., Vandekerckhove, J., and Tavernier, J. (2001). Design and application of a cytokine-receptor-based interaction trap. Nat. Cell Biol. *3*, 1114-1119.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C.*, et al.* (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res. *41*, D808-815.

Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., Rahman, N., and Stratton, M.R. (2004). A census of human cancer genes. Nat. Rev. Cancer *4*, 177-183.

Geiger, T., Wehner, A., Schaab, C., Cox, J., and Mann, M. (2012). Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol. Cell. Proteomics *11*, M111.014050.

Gerhard, D.S., Wagner, L., Feingold, E.A., Shenmen, C.M., Grouse, L.H., Schuler, G., Klein, S.L., Old, S., Rasooly, R., Good, P.*, et al.* (2004). The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC). Genome Res. *14*, 2121-2127.

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514-517.

Hanley, J.A., and McNeil, B.J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology *143*, 29-36.

Havugimana, P.C., Hart, G.T., Nepusz, T., Yang, H., Turinsky, A.L., Li, Z., Wang, P.I., Boutz, D.R., Fong, V., Phanse, S.*, et al.* (2012). A census of human soluble protein complexes. Cell *150*, 1068-1081.

Henrick, K., and Thornton, J.M. (1998). PQS: a protein quaternary structure file server. Trends Biochem. Sci. *23*, 358-361.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C.*, et al.* (2004). The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. Nat. Biotechnol. *22*, 177-183.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc. Natl. Acad. Sci. U.S.A. *106*, 9362-9367.

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40*, D261-270.

Isserlin, R., El-Badrawi, R.A., and Bader, G.D. (2011). The Biomolecular Interaction Network Database in PSI-MI 2.5. Database *2011*, baq037.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U.*, et al.* (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. *40*, D841-846.

Kerrien, S., Orchard, S., Montecchi-Palazzi, L., Aranda, B., Quinn, A.F., Vinod, N., Bader, G.D., Xenarios, I., Wojcik, J., Sherman, D.*, et al.* (2007). Broadening the horizon – level 2.5 of the HUPO-PSI format for molecular interactions. BMC Biol. *5*, 44.

Krissinel, E., and Henrick, K. (2007). Inference of macromolecular assemblies from crystalline state. J. Mol. Biol. *372*, 774-797.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. *305*, 567-580.

Lamesch, P., Li, N., Milstein, S., Fan, C., Hao, T., Szabo, G., Hu, Z., Venkatesan, K., Bethel, G., Martin, P.*, et al.* (2007). hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics *89*, 307-315.

Laskowski, R.A. (2009). PDBsum new things. Nucleic Acids Res. *37*, D355-359.

Levy, E.D. (2007). PiQSi: protein quaternary structure investigation. Structure *15*, 1364-1367.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739-1740.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E.*, et al.* (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40*, D857-861.

Mann, K.M., Ward, J.M., Yew, C.C., Kovochich, A., Dawson, D.W., Black, M.A., Brett, B.T., Sheetz, T.E., Dupuy, A.J., Australian Pancreatic Cancer Genome Initiative*, et al.* (2012). Sleeping Beauty mutagenesis reveals cooperating mutations and pathways in pancreatic adenocarcinoma. Proc. Natl. Acad. Sci. U.S.A. *109*, 5934-5941.

Manning, C.D., Raghavan, P., and Schütze, H. (2008). Introduction to information retrieval (New York: Cambridge University Press).

Manning, G., Whyte, D.B., Martinez, R., Hunter, T., and Sudarsanam, S. (2002). The protein kinase complement of the human genome. Science *298*, 1912-1934.

March, H.N., Rust, A.G., Wright, N.A., ten Hoeve, J., de Ridder, J., Eldridge, M., van der Weyden, L., Berns, A., Gadiot, J., Uren, A.*, et al.* (2011). Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. Nat. Genet. *43*, 1202-1209.

Meszaros, B., Simon, I., and Dosztanyi, Z. (2009). Prediction of protein binding regions in disordered proteins. PLoS Comput. Biol. *5*, e1000376.

Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011). Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. *7*, 548.

Nyfeler, B., Michnick, S.W., and Hauri, H.P. (2005). Capturing protein interactions in the secretory pathway of living cells. Proc. Natl. Acad. Sci. U.S.A. *102*, 6350-6355.

Obradovic, Z., Peng, K., Vucetic, S., Radivojac, P., Brown, C.J., and Dunker, A.K. (2003). Predicting intrinsic disorder from amino acid sequence. Proteins *53 Suppl 6*, 566-572.

Olsen, J.V., Vermeulen, M., Santamaria, A., Kumar, C., Miller, M.L., Jensen, L.J., Gnad, F., Cox, J., Jensen, T.S., Nigg, E.A.*, et al.* (2010). Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. Sci. Signal. *3*, ra3.

Pena-Castillo, L., Taşan, M., Myers, C.L., Lee, H., Joshi, T., Zhang, C., Guan, Y., Leone, M., Pagnani, A., Kim, W.K.*, et al.* (2008). A critical assessment of *Mus musculus* gene function prediction using integrated genomic evidence. Genome Biol. *9 Suppl 1*, S2.

Ponstingl, H., Kabir, T., and Thornton, J.M. (2003). Automatic inference of protein quaternary structure from crystals. J. Appl. Crystallogr. *36*, 1116-1122.

Prasad, T.S.K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A.*, et al.* (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37*, D767-772.

Ramachandran, N., Raphael, J.V., Hainsworth, E., Demirkan, G., Fuentes, M.G., Rolfs, A., Hu, Y., and LaBaer, J. (2008). Next-generation high-density self-assembling functional protein arrays. Nat. Methods *5*, 535-538.

Rozenblatt-Rosen, O., Deo, R.C., Padi, M., Adelmant, G., Calderwood, M.A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M.*, et al.* (2012). Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature *487*, 491-495.

Rual, J.-F., Hirozane-Kishikawa, T., Hao, T., Bertin, N., Li, S., Dricot, A., Li, N., Rosenberg, J., Lamesch, P., Vidalain, P.O.*, et al.* (2004). Human ORFeome version 1.1: a platform for reverse proteomics. Genome Res. *14*, 2128-2135.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G.F., Gibbons, F.D., Dreze, M., Ayivi-Guedehoussou, N.*, et al.* (2005). Towards a proteome-scale map of the human protein-protein interaction network. Nature *437*, 1173-1178.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32*, D449-451.

Starr, T.K., Allaei, R., Silverstein, K.A., Staggs, R.A., Sarver, A.L., Bergemann, T.L., Gupta, M., O'Sullivan, M.G., Matise, I., Dupuy, A.J.*, et al.* (2009). A transposon-based genetic screen in mice identifies genes altered in colorectal cancer. Science *323*, 1747-1750.

Starr, T.K., Scott, P.M., Marsh, B.M., Zhao, L., Than, B.L., O'Sullivan, M.G., Sarver, A.L., Dupuy, A.J., Largaespada, D.A., and Cormier, R.T. (2011). A Sleeping Beauty transposon-mediated screen identifies murine susceptibility genes for adenomatous polyposis coli (*Apc*)-dependent intestinal tumorigenesis. Proc. Natl. Acad. Sci. U.S.A. *108*, 5765-5770.

Stein, A., Ceol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *39*, D718-723.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F.H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S.*, et al.* (2005). A human protein-protein interaction network: a resource for annotating the proteome. Cell *122*, 957-968.

Stenson, P.D., Mort, M., Ball, E.V., Shaw, K., Phillips, A., and Cooper, D.N. (2014). The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. Hum. Genet. *133*, 1-9.

Strausberg, R.L., Feingold, E.A., Klausner, R.D., and Collins, F.S. (1999). The mammalian gene collection. Science *286*, 455-457.

Sultan, M., Schulz, M.H., Richard, H., Magen, A., Klingenhoff, A., Scherf, M., Seifert, M., Borodina, T., Soldatov, A., Parkhomchuk, D.*, et al.* (2008). A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. Science *321*, 956-960.

The Gene Ontology Consortium (2000). Gene Ontology: tool for the unification of biology. Nat. Genet. *25*, 25-29.

The MGC Project Team (2009). The completion of the Mammalian Gene Collection (MGC). Genome Res. *19*, 2324-2333.

Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M., and Wodak, S.J. (2010). Literature curation of protein interactions: measuring agreement across major public databases. Database *2010*, baq026.

Turinsky, A.L., Razick, S., Turner, B., Donaldson, I.M., and Wodak, S.J. (2011). Interaction databases on the same page. Nat. Biotechnol. *29*, 391-393.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.I.*, et al.* (2009). An empirical framework for binary interactome mapping. Nat. Methods *6*, 83-90.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., Jr., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546-1558.

Walzthoeni, T., Leitner, A., Stengel, F., and Aebersold, R. (2013). Mass spectrometry supported determination of protein complex structure. Curr. Opin. Struct. Biol. *23*, 252-260.

Yang, X., Boehm, J.S., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., Green, T.M.*, et al.* (2011). A public genome-scale lentiviral expression library of human ORFs. Nat. Methods *8*, 659-661.

Yu, H., Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N.*, et al.* (2008). High-quality binary protein interaction map of the yeast interactome network. Science *322*, 104-110.

Yu, H., Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X.*, et al.* (2011). Next-generation sequencing to generate interactome datasets. Nat. Methods *8*, 478-480.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847-848.

Zhang, Q.C., Petrey, D., Deng, L., Qiang, L., Shi, Y., Thu, C.A., Bisikirska, B., Lefebvre, C., Accili, D., Hunter, T.*, et al.* (2012). Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature *490*, 556-560.

Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., Yu, H., Venkatesan, K., Mou, D.*, et al.* (2009). Edgetic perturbation models of human inherited disorders. Mol. Syst. Biol. *5*, 321.