nature | **methods**

An empirical framework for binary interactome mapping

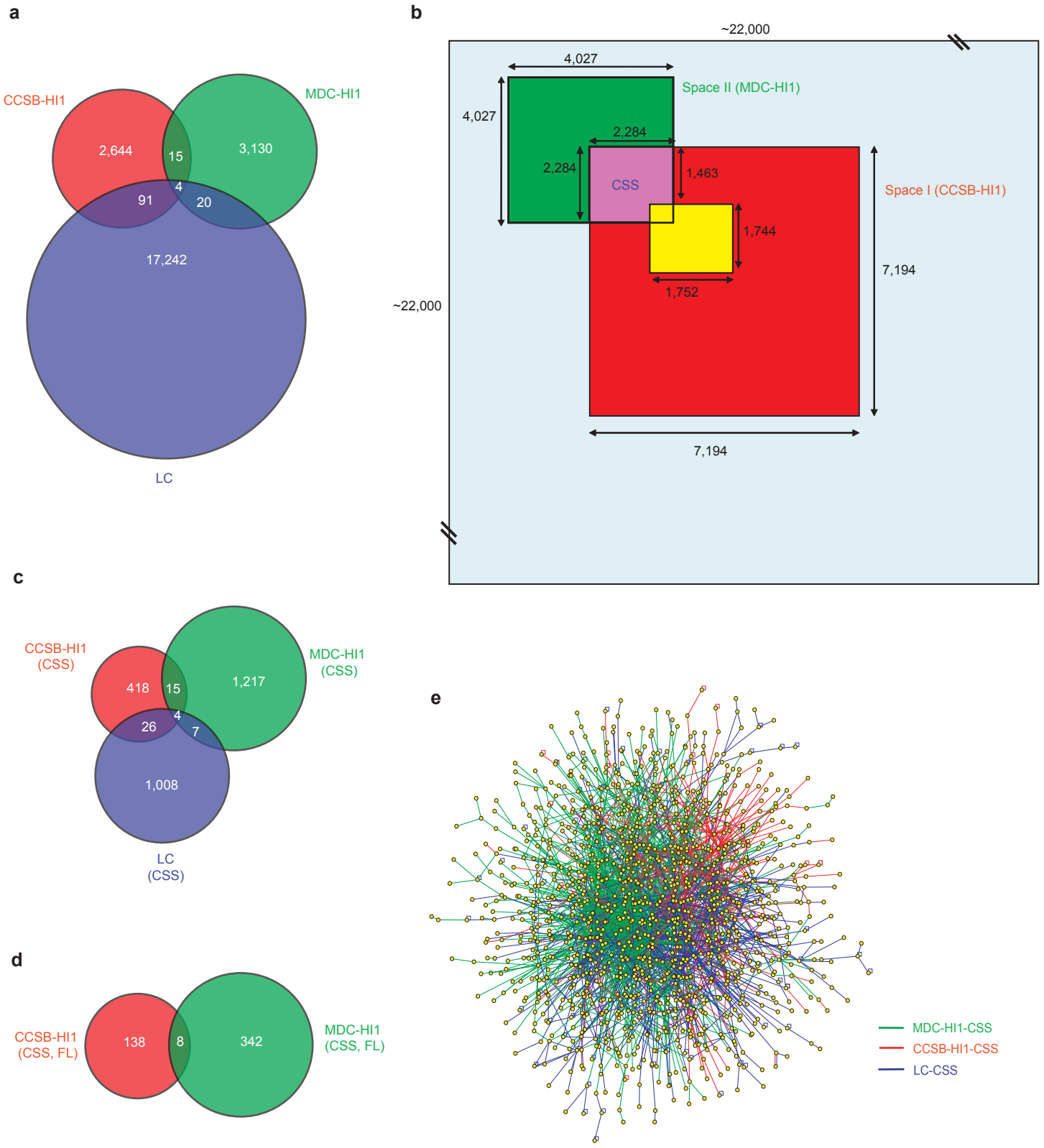
Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amélie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-László Barabási & Marc Vidal

Supplementary figures and text:

Supplementary Figure 1	Overlap and screening completeness of existing human binary interactome maps
Supplementary Table 2	Comparison of the features of the two different technologies, Y2H-CCSB and MAPPIT
Supplementary Table 3	Estimate of various parameters using Monte Carlo simulations based on experimental data and the mixture model of repeat screens
Supplementary Table 5	Calculation of conditional dependence between Y2H-CCSB and MAPPIT
Supplementary Data 1	Consideration of the correlation between pairs scoring positive in Y2H and in MAPPIT assays in the analysis of MAPPIT experiments
Supplementary Data 2	Examination of screening completeness of search spaces and overlap between CCSB-HI1 and MDC-HI1 datasets
Supplementary Data 3	Limitations of previous approaches for estimating data quality and interactome size
Supplementary Data 4	Current status of available human interactome maps
Supplementary Methods	

Note: Supplementary Tables 1, 4 and 6–9 are available on the Nature Methods website.

Supplementary Figure 1
Overlap and screening completeness of existing human binary interactome maps.



(a) Venn diagram depicting the overlap in the number of interactions between CCSB-HI1 (red circle), MDC-HI1 (green circle) and LC (blue circle). **(b)** Screening completeness of various search spaces including Space I (CCSB-HI1), Space II (MDC-HI1), CSS (space common to Space I and Space II). Space depicted in yellow indicates the search space in which repeated screens were performed using the Y2H-CCSB technology. **(c)** Venn diagram depicting the overlap in the number of interactions between CCSB-HI1 (red circle), MDC-HI1 (green circle) and LC (blue circle) in the CSS. **(d)** Venn diagram depicting the overlap between CCSB-HI1 and MDC-HI1 datasets in the subspace of CSS interrogated by full-length proteins in both datasets. **(e)** Network graph of interactions found in the CCSB-HI1 dataset (red lines), the MDC-HI1 dataset (green lines) or LC (blue lines) in the CSS. Proteins are depicted as yellow circles and interactions as lines between them.

Supplementary Table 2 | Comparison of the features of the two different technologies, Y2H-CCSB and MAPPIT.

Name	Transcription factor reconstituted	Host	Vector features	Cellular compartment where interaction occurs	Gateway compatibility	Gene markers used	References
Y2H-CCSB	GAL4 transcription factor	yeast: Mav103 and Mav 203	pDB-dest (bait) and pAD-dest-CYH (prey) vectors: pUC-based, ARSH4, CEN6 (low copy), constitutive moderate ADH1 promoter, N-terminal GAL4-DB and GAL4-AD fusions, *CYH2* marker on prey vector for counter selection of auto-activators	Nucleus	yes	*HIS3, LacZ, URA3*	Vidal et al. 1996
MAPPIT	Type I cytokine receptor signal	mammalian cells: 293T cells	pSEL (bait) vector: pSVsport-based, constitutive low SV40 early promoter, N-terminal chimeric epo/leptin receptor fusion; pMG1 (prey vector): pMET7-based, constitutive strong SRalpha promoter, N-terminal FLAG-gp130 domain fusion	Cell membrane	yes	Luciferase	Eyckerman et al. 2001

Supplementary Table 3 | Estimate of various parameters using Monte Carlo simulations based on experimental data and the mixture model of repeat screens. SFD refers to the systematic false discovery rate, UFD refers to the unsystematic or stochastic false discovery rate, E1 refers to the density of interactions reported in a single screen (ignoring stochastic or unsystematic false positives) and I1 refers to the density of all pairs (true or false positives) reported in a single screen.

Parameter	Mean	StandardDeviation	95% CI lower-bound	95% CI upper-bound
Assay-sensitivity related				
Assay-sensitivity of Y2H-CCSB (from hsPRS v1 pairs positive in pairwise tests)	17.0%	3.8%	10.3%	25.1%
Assay-sensitivity of Y2H-CCSB (from hsPRS v1 pairs positive in CCSB-HI1)	21.2%	10.5%	9.2%	42.6%
Assay-sensitivity of Y2H-CCSB (from LC-multiple pairs positive in CCSB-HI1)	19.6%	7.9%	13.9%	34.0%
Assay-sensitivity of Y2H-CCSB (combined)	17.9%	4.9%	10.8%	28.3%
Precision-related				
Fraction of hsPRS v1 pairs positive in MAPPIT	21.3%	4.2%	13.6%	30.1%
Fraction of Y2H-supported hsPRS v1 pairs positive in MAPPIT	34.1%	7.1%	20.9%	48.6%
Fraction of hsRRS v1 pairs positive in MAPPIT	2.2%	1.1%	0.6%	4.6%
Fraction of LC pairs positive in MAPPIT	8.0%	1.9%	4.5%	12.1%
Fraction of LCI (Single,Y2H,FL) positive in MAPPIT	9.6%	5.2%	2.1%	22.0%
Precision of LC	24.7%	18.7%	0.0%	72.5%
Fraction of MDC-HI1 positive in MAPPIT	9.9%	2.2%	6.0%	14.6%
Fraction of MDC-HI1 (Same,FL) positive in MAPPIT	31.1%	8.1%	16.7%	47.7%
Precision of MDC-HI1	83.5%	18.6%	41.2%	100.0%
Fraction of CCSB-HI1 positive in MAPPIT	27.4%	3.2%	21.3%	33.9%
Precision of CCSB-HI1	79.4%	15.9%	49.4%	100.0%
Sampling False Discovery rate (UFD_2config) of Y2H-CCSB	11.7%	6.1%	0.0%	19.6%
Systematic False Discovery Rate (SFD_2config) of Y2H-CCSB	13.6%	14.5%	0.0%	45.2%
Sampling-sensitivity related				
p_1config(Y2H-CCSB)	44.8%	22.1%	5.0%	86.0%
p_2config(Y2H-CCSB)	53.1%	10.0%	29.8%	65.6%
e_1config(Y2H-CCSB)	7.13E-05	2.79E-05	4.93E-05	1.33E-04
e_2config(Y2H-CCSB)	1.18E-04	5.42E-05	8.38E-05	2.17E-04
q_1config(Y2H-CCSB)	3.81E-06	1.99E-06	9.80E-09	6.39E-06
q_2config(Y2H-CCSB)	7.61E-06	3.97E-06	1.97E-08	1.28E-05
E1_2config(Y2H-CCSB)	5.73E-05	3.93E-06	5.23E-05	6.52E-05
I1_twoconfig(Y2H-CCSB)	6.49E-05	4.73E-07	6.40E-05	6.60E-05
Interactome-size				
Interactome Size (from Y2H-CCSB, assay sensitivity from pairwise hsPRS v1 expt)	159920	91164	70760	350484
Interactome Size (from Y2H-CCSB, using assay sensitivity from hsPRS v1 pairs positive in CCSB-HI1)	129603	50864	59322	257374
Interactome Size (from Y2H-CCSB, using assay sensitivity from LC-multiple pairs positive in CCSB-HI1)	127958	25200	77253	174006
Interactome Size (from Y2H-CCSB, combined)	130111	32618	73548	199688

Supplementary Table 5 | Calculation of conditional dependence between Y2H-CCSB and MAPPIT.

Universe = All hsPRS v1 pairs
Null hypothesis: Out of all hsPRS v1 pairs, there is no statistically significant correlation between MAPPIT positive pairs and Y2H-CCSB positive pairs

	Y2H-CCSB positive	Y2H-CCSB negative
MAPPIT positive	7	12
MAPPIT negative	8	65

P-value of correlation = 0.01 (significant)
Therefore, the null hypothesis is wrong, i.e., MAPPIT and Y2H-CCSB are correlated/dependent given all hsPRS v1 pairs

Universe = All hsPRS v1 pairs that are Y2H-supported in the original literature ("PRS-Y2H")
Null-hypothesis: Out of all the hsPRS v1 pairs that are Y2H-supported there is no statistically significant correlation between MAPPIT positve pairs and Y2H-CCSB positive pairs

	Y2H-CCSB positive	Y2H-CCSB negative
MAPPIT Positive	5	9
MAPPIT negative	5	23

P-value of correlation = 0.184 (not significant)
Therefore, the null hypothesis is correct, i.e., given the set of Y2H-supported hsPRS v1 pairs, MAPPIT and Y2H-CCSB are independent or not correlated

Thus, reducing the set of hsPRS v1 pairs to the subset supported by Y2H removes the correlation between Y2H-CCSB and MAPPIT

Supplementary Data 1: Consideration of the correlation between pairs scoring positive in Y2H and in MAPPIT assays in the analysis of MAPPIT experiments

In order to compute precision of any given dataset using MAPPIT, we needed to benchmark the performance of that dataset in MAPPIT against the performance of the hsPRS-v1 and hsRRS-v1 pairs. We investigated the conditional dependence of Y2H-CCSB and MAPPIT by computing the overlap between hsPRS-v1 pairs that scored positive in Y2H-CCSB as well as in MAPPIT (**Supplementary Table 5** online). The resulting *P*-value obtained ($P = 0.01$) indicated that there is a bias for pairs scoring positive in Y2H-CCSB to also score positive in MAPPIT. In order to account for this dependence, we chose the subset of hsPRS-v1 pairs supported by Y2H assays (PRS-Y2H) in at least one publication as determined by our recuration of hsPRS-v1-associated publications. Consistent with the dependence between MAPPIT and Y2H-CCSB, a greater fraction (34%) of PRS-Y2H pairs scored positive in MAPPIT compared to 21% of all hsPRS-v1 pairs scoring positive. We asked if, among the PRS-Y2H pairs, there was a further correlation between pairs supported specifically by Y2H-CCSB and found no statistically significant correlation ($P = 0.184$). This indicated that the use of PRS-Y2H pairs as a benchmark to compute precision of other Y2H datasets is sufficient to account for the conditional dependence between Y2H-CCSB and MAPPIT.

Supplementary Data 2: Examination of screening completeness of search spaces and overlap between CCSB-HI1 and MDC-HI1 datasets

We evaluated the screening completeness of datasets by comparing their respective tested spaces. Using version 44.36f of the ENSEMBL human genome annotation (released 03/29/2007) which predicts 22,470 known or novel protein-coding genes, the screening completeness of the tested spaces corresponds to: (1) (7,200 x 7200) / (22,500 x 22,500) = 10% for Space I tested in the generation of the CCSB-HI1 map, (2) (4,000 x 4000) / (22,500 x 22,500) = 3% for Space II (**Supplementary Table 9** online) tested in the generation of the MDC-HI1, and (3) (2,300 x 2,300) / (22,500 x 22,500) = 1% for CSS (**Supplementary Fig. 1** online), or "common subspace", (the set of 2,284 x 2,284 gene loci tested in both Space I and II). Furthermore, only a subspace of 1,463 x 1,463 gene loci was interrogated using full-length clones in both CCSB-HI1 and MDC-HI1 (CSS-FL).

The fraction of the entire CCSB-HI1 and MDC-HI1 datasets made up of common interactions corresponds to 0.7% and 0.6% respectively. Upon considering only those interactions in the CSS-FL, the fractions of CCSB-HI1 and MDC-HI1 datasets made up of common interactions increase to 5.8% and 2.3% respectively.

Supplementary Data 3: Limitations of previous approaches for estimating data quality and interactome size

We present here a unique conceptual framework, the first one to be based on empirical protein-protein interaction data directly targeted at quality assessment of interactome maps. Previous studies have estimated the precision of existing maps and/or the size of interactomes using (i) analysis of the extent to which interacting proteins share other biological attributes such as co-expression or shared functional annotation[1-5] or (ii)

statistical analysis of various features of existing interactome maps including analysis of the extent of overlap between two maps[6-8]. Our approach addresses various limitations of these studies.

Methods that rely on correlation with other biological attributes to estimate precision of an interactome map assume that our knowledge of functional annotation is complete and unbiased. However, the annotation of most proteomes today is partial and suffers various biases, *e.g.*, classes of proteins being particularly scrutinized because of their involvement in human disease. In this context, interacting proteins not sharing functional annotation should be considered good candidates for novel functional discovery rather than potential false positives. Attributing low confidence to true interactions that poorly correlate with other biological attributes is likely to artificially lower the precision of datasets generated using HT approaches, which are sociologically unbiased and not inherently constrained by pre-existing paradigms governing functional annotation. On the contrary, using functional annotation for evaluating the precision of sociologically biased LC datasets is highly circular since shared functional annotations and physical interactions between protein pairs in the low-throughput literature are inherently dependent on one another[4,9]. Therefore, higher correlation of LC pairs for shared attributes such as co-localization or co-expression cannot be interpreted to reflect higher interaction quality. As a case in point, one of the studies based on analyzing functional correlation of interacting protein pairs[3] suffers from the limitations above. This potentially explains their estimate of greater than 50% false discovery rates for HT-Y2H yeast maps. Our framework overcomes these

limitations by assessing false discovery rates directly using information from protein-protein interaction assays.

Methods based on analyzing the extent of overlap between interactome maps[6-8] are free from the above biases as they rely purely on protein-protein interaction data but current implementations of this approach have specific limitations. Some studies[7,8] used a combination of interactions available in various LC interaction databases as the reference set for their analysis. Most of these interactions are supported by a single publication and, as seen from our MAPPIT experiments (**Fig. 3c**) as well as from our recuration analysis[10], their use as a reference set may not be appropriate given a potentially higher false positive rate than previously anticipated. To limit the potential problems associated with the use of such datasets as reference, we generated a high quality reference set (hsPRS-v1) requiring interactions to be supported by multiple publications and to pass additional recuration (**Fig. 2a**). As Gateway-cloned ORFs[11] are available for every protein involved in our hsPRS-v1, it represents the first positive reference set systematically testable using any binary protein-protein interaction assay[12].

So far, only one previous study attempted to estimate the precision of human HT-Y2H maps[8]. The overlap-based method used in that analysis involves comparison of two interactome datasets to each other and to a reference set[7]. Using this strategy, the precision of the CCSB-HI1 dataset[9] was estimated to be ~10%, corresponding to a false discovery rate of 90%. In striking contrast, we estimated the precision of CCSB-HI1 to be ~80%, corresponding to a false discovery rate of ~20%. We speculate that this discrepancy could be due to various issues. Although the underlying mathematical

method[7] used in their study requires that the two datasets used in the overlap-based estimation be generated using similar or ideally identical assays, their study[8] used the CCSB-HI1 dataset together with computationally predicted interaction datasets[13,14], which are generated using diverse techniques distinct from Y2H. Moreover, another requirement of their method is that the reference set not be biased toward either of the two datasets being compared[7]. We argue that this requirement is not met when comparing a computationally predicted interactome map with a reference set consisting of LC interactions. Indeed, amongst other features, shared functional annotation and/or co-expression were used to predict these human interactome maps[13,14] and, as we discussed above, shared annotation or co-expression, and physical interactions reported in LC databases are inherently dependent on one another. All these limitations together may have led to overestimating false discovery rates for existing HT-Y2H human interactome maps. Our approach to estimate false discovery rates avoids these pitfalls by using two different approaches for our estimate, (i) experimentally testing a random subset of interacting pairs in an independent assay and (ii) comparing overlaps between four homogeneously-derived repeat screens.

Finally, earlier studies failed to consider one or more of the parameters that influence interactome map quality. A recent analysis[15] estimated that the human interactome network contains ~650,000 interactions by scaling up the number of interactions in a given map according to the fraction of the proteome represented in that map. This study does not estimate specific quality parameters (precision, assay sensitivity and sampling sensitivity), which could significantly affect the resulting estimate of interactome size if taken into account. In another scenario, most previous

approaches[1-3,6-8,16] did not distinguish between sampling sensitivity and assay sensitivity when estimating false negative rates. In the sole study that did attempt to distinguish between these two parameters[17], the estimate of sampling sensitivity and assay sensitivity of a given assay is likely to be significantly overestimated. Their analysis is restricted to proteins that are assumed to be detectable in a given assay. This is done considering only the subset of proteins that are involved in at least one interaction as both DB-X fusions ("bait") and AD-Y fusions ("prey") in a given interactome dataset. Out of all the observed interactions involving this "detectable" subset of proteins, they computed the fraction of interactions observed in both configurations ("bait-prey" versus "prey-bait") and assumed that interactions undetected in one of the two configurations are owing to a combination of limited sampling sensitivity and assay sensitivity. This method does not consider those interactions that are detected in neither configuration in a dataset. It is likely that for any assay, a substantial fraction of all true interactions are detected only after multiple screens of an assay (as seen from our repeat screens), or more strikingly never detectable by an assay (as seen from our hsPRS-v1 experiments). Therefore we argue that their estimates of sampling sensitivity and assay sensitivity are likely to be significantly overestimated. For example, although our estimate of assay sensitivity may also be an overestimate (since the hsPRS-v1 pairs are reported in multiple publications, therefore may be more likely to be detectable by current experimental methods than would a random set of true interactions), their approach would have estimated an assay sensitivity of ~48% for Y2H-CCSB, which is significantly higher than the assay sensitivity of ~17% obtained with our approach (**Supplementary Table 3**). While our approach is not perfect, it provides a better estimate of sampling

sensitivity and assay sensitivity based on direct experimental data. In addition, our approach is the first to estimate two different types of false discovery rates, *i.e.*, stochastic and systematic false discovery rates.

Supplementary Data 4: Current status of available human interactome maps

Our results offer a comprehensive picture of the current state and the future potential of interactome mapping. We estimate that 1,872 of the reported 2,754 interactions (68%) in the CCSB-HI1 dataset and 2,282 out of 3,169 (72%) in the MDC-HI1 dataset are true biophysical interactions. Out of the 17,297 LC binary interactions available in the union of the BIND, DIP, HPRD, MINT and MIPS databases, 3,321 out of 15,094 (22%) interactions in the "LC-Single" (LC interactions supported by a single publication) dataset are estimated to be true positives. Taking all three datasets together with the 2,203 interactions in the "LC-Multiple" (LC interactions supported by multiple publications) dataset and given the negligible overlap between the datasets, we estimate that out of 23,220 currently reported human interactions, ~9,700 are genuine (~42%). Thus, the fraction of interactions identified so far represents 5% to 13% of the full interactome, *i.e.,* more than 85% of all interactions remain to be discovered.

SUPPLEMENTARY METHODS

Generation of a binary interaction Positive Reference Set (hsPRS-v1) and Random Reference Set (hsRRS-v1). To generate our human binary interaction Positive Reference Set (hsPRS-v1), we started with 17,297 binary literature-curated (LC) interactions from five curated databases (BIND[18], DIP[19], HPRD[20], MINT[21] and

MIPS[22]), out of which 4,067 pairs were contained in a space corresponding to our human ORFeome v1.1 collection[11] and tested in the Rual *et al.* publication[9] (referred to as LC interactions from "Space I", the HT-Y2H tested space). These LC interactions excluded HT-Y2H interactions reported in the Rual *et al*. or Stelzl *et al*. publications. From these 4,067 pairs, we chose 188 pairs supported by the highest number of publications and curated by the highest number of databases. Systematic recuration of all publications thought to support these 188 protein pairs verified 107 direct binary interactions between human proteins that were supported by multiple publications, 92 of which involved full-length proteins and constituted our hsPRS-v1 (**Supplementary Table 1** online). The remaining 81 LC pairs were annotated as being supported by one or no publication according to our stringent criteria[10]. Proteins involved in the 92 hsPRS-v1 interactions exhibit broad cellular localization, suggesting they represent the currently known interactome obtained from a wide variety of binary assays.

A set of 188 hsRRS-v1 pairs was derived by random selection from all pair-wise combinations in Space I (see above), after removing known biophysical interactions in LC or HT-Y2H datasets (**Fig. 2b** and **Supplementary Table 1** online). These randomly chosen pairs should largely be true negative pairs if the size of the interactome is about 200,000 interactions, the upper bound of interactions estimated from this study.

All 506 full-length open reading frames (ORFs) encoding the proteins to be tested in both hsPRS-v1 and hsRRS-v1 were transferred by Gateway recombinational cloning (Invitrogen) from Gateway Entry clones in the human ORFeome v1.1 collection[11] into MAPPIT and Y2H-CCSB vectors. Thus identical clones are used to test the same pairs in different binary interaction assays.

Description of datasets whose precision was tested by MAPPIT. The three binary interaction datasets are as follows: (i) "CCSB-HI1" dataset[9] with 2,754 HT-Y2H interactions found using Y2H-CCSB in a single sampling of Space I, (ii) the "MDC-HI1" dataset[23] with 3,169 HT-Y2H interactions found using a different Y2H approach (Y2H-MDC) in a single sampling of "Space II" (4,027 x 4,027 gene loci, **Supplementary Table 9** online), and (iii) LC interactions, which corresponds to the set of 4,067 interactions curated from the low-throughput literature in the union of the BIND, DIP, HPRD, MINT and MIPS databases. Importantly, all 761 corresponding full-length ORFs encoding the proteins involved in these pairs selected from the human ORFeome v1.1 resource[11] were transferred individually by Gateway recombinational cloning (Invitrogen) into Gateway compatible MAPPIT Destination vectors[24,25] and processed as described[11].

Y2H-CCSB assays. Y2H-CCSB involves reconstitution of the Gal4 transcription factor in the yeast nucleus to detect interactions and uses various features of stringency, *e.g.,* low-copy plasmids, hybrid protein expression under the control of a constitutive moderate promoter and a scoring system based on multiple reporter genes and counter-selectable markers[26,27]. The specific set of experimental and scoring conditions of Y2H-CCSB evaluated throughout this study is based on a protocol described previously[9] with minor changes. Specifically, a single Y2H-CCSB screen consists of two steps: (i) testing each DB-X bait against mini-pools of AD-Y preys, and (ii) pair-wise retesting of Y2H interactions by mating fresh DB-X bait and AD-Y prey yeast cells. In

each step, scoring of the Y2H readouts is performed as follows. The screen starts with frozen glycerol stocks of DB-X bait and AD-Y prey yeast cells. Mating is performed on solid YEPD plates at 30ºC after which colonies are replica-plated on the following assay plates: Sc-L-T-U, Sc-L-U+CYH, Sc-L-T-H+3AT, Sc-L-H+3AT+CYH and YEPD for a β-galactosidase filter assay. Y2H read-outs are scored four days after replica-plating the cells; the four days include velvet-cleaning (one day after spotting) and three additional days of incubation at 30ºC after velvet-cleaning. To be considered Y2H positives, the cells must grow significantly more on the Sc-L-T-H+3AT than they do on Sc-L-H+3AT+CYH or grow significantly more on the Sc-L-T-U than they do on Sc-L-U+CYH and they must score positive in the β-galactosidase filter assay. If a score cannot be confidently attributed (weak phenotype, weak growth difference between the Sc-L-T+3AT and the Sc-L+3AT+CYH, human error or contamination), the yeast cells are processed through a second round of phenotyping tests. If a score cannot be confidently attributed in the second round, the yeast cells are considered Y2H negatives.

The precision of the CCSB-HI1 dataset generated using this assay implementation was estimated in **Fig. 3c** and the sampling sensitivity of a single Y2H-CCSB screen performed using this assay implementation was estimated in **Fig. 3e**. Finally, as described above, a single Y2H-CCSB screen consists of two independent mating experiments and a protein pair needs to score positive in each of these two experiments in order to be reported positive in the screen. Similarly, in the hsPRS-v1/hsRRS-v1 pair-wise mating experiments, hsPRS-v1 or hsRRS-v1 pairs need to show a positive signal in at least one configuration in both independent pair-wise mating

experiments to be scored positive. Moreover, we assume that pair-wise mating experiments operate at or near full sampling sensitivity since such experiments overcome losses due to pooling, limited selection of positives and sequencing. Therefore, we assume that the fraction of hsPRS-v1 pairs scoring positive in the pair-wise mating experiment in **Fig. 2f** reflects the assay sensitivity of the specific version of Y2H-CCSB described in the above paragraph.

In the alternate Y2H-CCSB protocols shown in the x-axis of **Fig. 2d**, testing of hsPRS-v1 and hsRRS-v1 also starts with frozen glycerol stocks. Mating is performed in liquid YEPD, spotted on Sc-L-T plates and incubated for ~20 hours at 30ºC after which colonies are replica-plated on duplicate sets of the following assay plates: Sc-L-T-U, Sc-L-U+CYH, Sc-L-T+ 2% 5-FOA, Sc-L-T+3AT, Sc-L+3AT+CYH and YEPD for a β-galactosidase filter assay. To be considered Y2H positive in this protocol, a colony needs to show a clear positive signal on at least two independent reporter assays in a cycloheximide-sensitive manner[26], and this result has to be reproducible in the duplicate sets of assay plates. Moreover, to be counted as a positive in the "-URA positive" protocol, a colony needs to show a positive signal on Sc-L-T-U plates.

MAPPIT experiments. MAPPIT detects binary physical interactions by reconstitution of a membrane-bound receptor in mammalian cells[28] and measurement of downstream luciferase reporter activity. MAPPIT experiments were performed essentially as described[29]. MAPPIT results are considered scorable if: (i) LR cloning is successful for both bait and prey constructs; (ii) expression of the bait hybrid protein is sufficient, *i.e.,* when the receptor-bait is able to generate a fold-induction value (mean value of the

ligand-stimulated cells divided by the mean value of the non-stimulated cells) higher than ten when tested in combination with a prey construct (TRIP13) which interacts with the chimeric receptor independently of the bait; and (iii) the bait and the prey proteins are considered to interact specifically, *i.e.*, when the receptor-bait or gp130-prey generate a fold-induction value lower than twenty when tested in combination with an irrelevant prey protein (amino acids 261-708 of SV40 large antigen T) or an irrelevant bait protein (chimeric receptor without bait), respectively. Interactions that failed any of these conditions were counted as “not applicable” (NA). The basic output of MAPPIT is the “Experiment-to-Control Ratio” (ECR), defined as the fold-induction value with bait and prey, divided by the fold-induction value with bait and irrelevant prey, or prey and irrelevant bait. The ECR has to be higher than ten for a trial to be reported positive. The experiment was performed in both configurations (receptor-X vs. gp130-Y and receptor-Y and gp130-X) and in two independent pair-wise trials, giving four distinct outputs (ECR) for each tested pair. Pairs were scored positive if they reported positive in at least one configuration in both pair-wise trials.

It is to be noted that the results of the hsPRS-v1 pairs in MAPPIT primarily serve as a reference *relative* to CCSB-HI1, MDC-HI1 and LC for measuring precision of these datasets. Since the results from the hsPRS-v1 and hsRRS-v1 pairs were used in order to derive suitable experimental conditions (ECR score thresholds) for scoring a positive in MAPPIT, the absolute fraction of the current hsPRS-v1 pairs scoring positive in MAPPIT reflects only gives a gross estimate of the assay sensitivity; a more independent estimate of the assay sensitivity of MAPPIT would require testing a separate set of PRS pairs.

Calculation of precision from MAPPIT experiments. We calculate the precision of the datasets using MAPPIT as follows. For every 100 pairs in each interaction dataset (LC, CCSB-HI1 or MDC-HI1), let I_{obs} be the number of observed positives in MAPPIT for pairs in that dataset. If I_+ is the number of true positives in a particular consolidated (after adjusting for various biases) dataset, f_+ is the false positive rate of MAPPIT (obtained as the fraction of hsRRS-v1 pairs reported positive), and (1-f_-) is the fraction of Y2H-supported hsPRS-v1 pairs that report positive in MAPPIT, then the following equation holds true:

(1.1) $I_{obs} = (100 - I_+) f_+ + I_+ (1 - f_-)$

Substituting the values for I_{obs}, f_+ and f_- above, we can solve for I_+, which represents the number of true positives in the interaction dataset under consideration, and consequently provides an estimate of the precision or percentage of the false positives (*i.e.*, 100 - I_+) in that dataset.

(1.2) $$I_+ = \frac{I_{obs} - 100 f_+}{1 - f_- - f_+}$$

For example, 51/188 CCSB-HI1 pairs CCSB-HI1, 14/42 PRS-Y2H pairs and 3/185 hsRRS-v1 pairs score positive in MAPPIT. Therefore the distributions provided as inputs for the Monte Carlo simulations (see below) in this case are given by I_{obs} = Beta(52,138) , f_- = 1 – Beta(15,29) and f_+ = Beta(4,183).

It is worth noting that in the experiment design and calculations described above, we use MAPPIT to test protein pairs that have been reported positive by a Y2H assay (*e.g.,* CCSB-HI1 interactions). Therefore, to be precise, we should make use of the MAPPIT false positive and negative rates over protein pairs that are reported positive by

the same Y2H assay. Therefore, we restricted the positive reference set, hsPRS-v1, to interactions that were reported positive by Y2H. Similarly, for the random reference set, hsRRS-v1, we would need protein pairs that are true non-interacting pairs that, at the same time, are reported positive by Y2H. The construction of such a negative set is very difficult. Yet, in practice such a set may not be needed. Only the subset of non-interacting protein pairs reporting positive by both Y2H and MAPPIT would affect our estimate of precision. In order for a non-interacting protein pair to report positive in both assays, it almost certainly needs to be a systematic false positive of each assay rather than a stochastic false positive: Stochastic false positives can apply to every non-interacting pair and arise due to technical/human errors while performing an assay once but report negative upon repeated testing, or upon testing in another assay. Systematic false positives, on the other hand, are usually specific non-interacting protein pairs that persistently report positive in an assay. Therefore, among a given set of non-interacting pairs, the subset of pairs that are stochastic false positives of Y2H, or the subset of pairs that are systematic false positives of Y2H but not MAPPIT, would not affect our precision estimate. Among the ~200 Y2H positives tested by MAPPIT, we assume that there is a negligible number of pairs that are systematic false positives of both Y2H and MAPPIT, and under this assumption, the false positive rate estimated using hsRRS-v1 and the resulting estimate of precision are good approximations.

Calculation of average number of interactions reported after *m* 'repeat screens'. To compute the average number of reported interactions after m, m=1,..,M, screens we use as input data the number of interactions N_n reported positive n, n=1,..,M, times in M

screens. Let Pr($l|n,m,M$) be the probability that an interaction is reported l times after the first m screens, given that it was reported positive n times in the M screens. The expected number of reported interactions after m screens is given by

$$(2.1)\quad E_m = \sum_{n=1}^{M} N_n \sum_{l=1}^{\max(n,m)} \Pr(l \mid n,m,M)$$

Furthermore, given an interaction reported positive n times in M screens, the probability that it is reported positive l times after the first m screens is given by the hypergeometric distribution

$$(2.2)\quad \Pr(l \mid n,m,M) = \frac{\binom{n}{l}\binom{M-n}{m-l}}{\binom{M}{m}}$$

Modeling of repeat screens. The sampling sensitivity of an assay, measured as the fraction of interactions detected per screen of that assay, can be modeled using repeat screens. For Y2H assays, one source of screen-to-screen variation is yeast mating, the success rate of which can vary greatly. Bait/prey protein expression fluctuations and other sources of cellular noise, or failure of PCR or sequencing of interaction sequence tags (ISTs), may also result in variability. These unsystematic fluctuations are modeled by a sampling sensitivity parameter p, the probability to report positive in one screen. When repeating a screen M times the probability to obtain n positives is given by the binomial distribution

$$(3.1)\quad \pi_n(p) = \binom{M}{n} p^n (1-p)^{M-n} .$$

Another fluctuation is that different protein pairs may have dissimilar sampling sensitivity. Firstly, we consider a mixture model where we assume that the search space consists of a mixture of two classes of pairs: a fraction *e* of interacting proteins and 1-*e* of non-interacting proteins. When an interacting pair of proteins reports positive, we define it as a true positive (TP), otherwise it is a false negative (FN). When a non-interacting pair reports positive, we define it as a false positive (FP). If we assume a constant sampling sensitivity *p* across all interacting pairs *q* and across all non-interacting pairs, then equation (3.1) gets replaced by

$$\pi_n(e,p,q) = e\binom{M}{n}p^n(1-p)^{M-n} + (1-e)\binom{M}{n}q^n(1-q)^{M-n} \quad (3.2)$$

More generally we can assume in our mixture model that there are *K* sampling sensitivity classes of interacting protein pairs obtaining

$$\pi_n(e,p,q,K) = \sum_{i=1}^{K} e_i\binom{M}{n}p_i^n(1-p_i)^{M-n} + (1-e)\binom{M}{n}q^n(1-q)^{M-n}\ , \quad (3.3)$$

where

$$e = \sum_{i=1}^{K} e_i\ . \quad (3.4)$$

In principle there could be more than one class of false positives. However, because the rate of false positives *q* happens to be tiny they cannot be differentiated. When *q* is tiny ($qM<<1$), false positives only contribute to protein pairs reported positive once, with a contribution approximated by $(1-e)q(1-q)^{M-1}\approx(1-e)q$. When there are different classes of false positives we obtain the same result, with *q* representing the average stochastic false positive rate and, as anticipated, we cannot differentiate them.

Thus, without loss of generality, we assume a single sampling sensitivity class for non-interacting proteins and interpret *q* as the average stochastic false positive rate.

One HT-Y2H screen consists of testing a search space *S* of protein pairs for protein-protein interactions. *M* HT-Y2H screens consist of repeating the same screen over the search space *M* times. The outcome is N_0 protein pairs never found positive, N_1 pairs found positive once, N_2 pairs found positive twice, …, N_M pairs found positive *M* times, where $\sum_{n=0}^{M} N_n = S$. The likelihood of a particular outcome is given by the multinomial distribution

$$\Pr(N|e,p,q,K) = \frac{S!}{N_0!\cdots N_M!}\left[\pi_n(e,p,q,K)\right]^{N_n}. \quad (3.5)$$

We could proceed by computing the maximum likelihood estimate (MLE) for the model parameters given the data *N*. However, the MLE can bias the parameter estimation. A preferred strategy would be to also consider nearly optimal fits, producing intervals of confidence for the model parameters. Thus, we use a Bayesian approach[30] and compute the posterior distribution

$$\Pr(e,p,q,K \mid N) = \frac{\Pr(N|e,p,q,K)\Pr(e,p,q,K)}{\int dedpdq \Pr(N|e,p,q,K)\Pr(e,p,q,K)}, \quad (3.6)$$

for the model parameters, given the data *N* and a prior distribution Pr(*e*,*p*,*q*,*K*).

For the prior distribution we assume independence between the model parameters *e*, *p*, *q* and *K* and a uniform distribution on the interval [0,1] for *e*, *p* and *q*. The prior distribution for *K* takes into account the model complexity. For a given *K* there are *2K*+1 independent parameters (*e*,*p*,*q*). Because a model with *K*>1 contains models with smaller *K* as particular examples, then the larger the *K* value the better the model fits the data. On the other hand, larger *K* values greatly increase the model complexity. To

account for this increase in model complexity we use the Akaike information theoretical criterion (AIC)[31] and assume

(3.7) $\Pr(K) = Ae^{-\#\text{ independent parameters}} = Ae^{-(2K+1)}$,

where *A* is a normalization constant such that $\sum_{K=1}^{\infty}\Pr(K) = 1$. Following these assumptions for the prior distribution, from (3.7) we obtain

(3.8) $\Pr(e,p,q,K \mid N) = \frac{1}{Z(N)}\Pr(N|e,p,q,K)e^{-2K}$.

where

(3.9) $Z(N) = \sum_{K=1}^{\infty}\int de\int dp\int dqP(N|e,p,q,K)e^{-2K}$

is the partition function. In essence, (3.8) represents a probability distribution in the space of parameters (*e*,*p*,*q*,*K*) given the data *N* and our assumptions for the prior distribution. Thus, the expectation of a certain variable *x*(*e*,*p*,*q*,*K*) is given by

(3.10) $E[x](N) = \sum_{K=1}^{\infty}\int de\int dp\int dq\Pr(e,p,q,K \mid N)x(e,p,q,K)$.

The magnitudes reported in the main text are the expectation of the number of interactions in the search space $e = \sum_{i=1}^{K} e_i$, the fraction of pairs e_i and sampling sensitivity p_i on each class of true interacting proteins, the rate of false positives *q*, and standard deviations. Furthermore, the relative contribution of each model with *K'* classes is given by $E[\delta_{KK'}]$, where $\delta_{KK'} = 1$ when *K*=*K'* and zero otherwise is the Kronecker delta symbol.

The calculation/computation of these averages is quite challenging. The reported results were obtained using the following approximations: (i) Replace the integral over p_i

by a sum with resolution Δp=0.01. (ii) For each K and set of p_i restrict the average over e_i and q to the MLE. (iii) Truncate the sum over K at K=1 or K=2.

The model is run on interaction data at the level of distinct clone pairs. To normalize the resulting estimated parameters to the gene locus level (*i.e.*, to consider parameters assuming one clone per gene), we consider the following. e_i*(clone pair)* represents a *fraction* of pairs, so e_i*(gene pair)* is the same as e_i*(clone pair)*. In theory, p_i*(gene pair)* is potentially greater than p_i*(clone pair)* if there are multiple clones for the bait or prey associated with the interaction geneA-geneB. This is because we can detect the same interaction at the gene level from different combinations at the clone level. Simplistically, if r is the total number of pairs at the clone level divided by the total number of pairs at the gene level, then p_i*(gene pair)* equals r times p_i*(clone pair)*. However, this is valid only under the assumption that for any given interacting gene pair, all interactions at the clone level (*i.e.*, all clone pair combinations) are detectable, which is not true. Thus in general the correction factor between p_i*(clone pair)* and p_i*(gene pair)* will be somewhere between 1 and r. Since, here r (r_{CCSB} = 1.09 and r_{MDC} = 1.22) is not so different from 1, we report p_i*(clone pair)* in Table 1. This approximation does not affect the calculation of the size of the human binary interactome, since we use the value of e_i (which is independent of the number of clones pairs per gene pair) in the calculations.

After preliminary analysis it became clear that two classes of true positives represent a significantly better description of our repeated screen data than just one. Therefore, we extended the sum over K till K=2. The estimated model parameters are

shown in **Supplementary Table 3** online. **Fig. 3e** shows the predicted expected number of positives as a function of the number of screens.

Estimating magnitudes of various parameters and the resulting interactome size. To estimate the human interactome size we need to take into consideration the following factors:

(i) Estimation of the sampling sensitivity "*p*", and the fraction of interactions that would be reported as positive interactions after performing a large number of screens of an assay, "*e*"

This fraction represents the maximum number of interactions expected to be reported positive by a given technology at full sampling sensitivity relative to the search space size and is estimated from the model of the repeated screens. The calculation of "*e*" thus takes into account the sampling sensitivity as measured from the model of the repeated screens. The value of *e* obtained with the model is based on repeated screens performed in a single configuration (DB-X vs. AD-Y) of the search space. To estimate $e_{2config}$, which refers to the density of detectable interactions upon performing the screens in both configurations (DB-Y vs. AD-X), we considered the following.

Let *v* be the probability that an interaction is detectable in both configurations, $p_{1config}$ the probability that an interaction is observed in a single configuration in one screen, $p_{2config}$ the probability that an interaction is observed in at least one of the two configurations in one screen, $e_{1config}$ the density of interactions detected after a large number of screens each performed in a single configuration and $e_{2config}$ the density of

interactions detected after a large number of screens each performed in both configurations. Then

(4.1) $p_{2config} = v[1-(1-p_{1config})^2] + (1-v)p_{1config}$

Here the first term on the right hand side of the equation refers to the probability that an interaction is detectable in both configurations times the probability that it is observed in at least one configuration in one screens. The second term refers to the probability that an interactions is detectable in only one configuration times the probability that it is observed in one screen and one configuration. Thus,

(4.2) $p_{2config} = p_{1config} + vp_{1config}(1-p_{1config})$

Similarly, $e_{1config}$ and $e_{2config}$ are related by the equation

(4.3) $e_{1config} = ve_{2config} + \frac{(1-v)e_{2config}}{2}$

The first term on the right hand side refers to the density of interactions detectable by both configurations that are found in at least one configuration upon saturation. The second term refers to the density of interactions detectable in only one configuration that are detected upon saturation. Thus,

(4.4) $e_{2config} = \frac{2e_{1config}}{1+v}$

To compute *v*, we used information from the CCSB-HI1 (Rual *et al.*) screen where protein pairs were tested in both configurations. If *L* was the number of proteins in the Space I and $N_{bothconfig}$ is the number of interactions found in both configurations in this screen, then

(4.5) $N_{bothconfig}(Rual) = v[(\frac{L^2}{2})e_{1config}]p_{1config}^2$

The corresponding density of interactions detectable in both configurations would then be

(4.6) $E_{bothconfig}(Rual) = \frac{N_{bothconfig}(Rual)}{\frac{L^2}{2}} = ve_{1config}p^2$

Thus,

(4.7) $v = \frac{E_{bothconfig}(Rual)}{e_{1config}p^2}$

In CCSB-HI1, 97 interactions where found in both configurations among a search space containing 7194 genes (*i.e.*, 7195*7194/2 gene pairs). Therefore $E_{bothconfig}(CCSB\text{-}HI1)$ equals 3.75E-6. Substituting this into equations 4.7 and 4.4 we can obtain an estimate of $e_{2config}$, which is our starting point for calculating the size of the interactome.

(ii) Estimation of the assay sensitivity of Y2H-CCSB

In order to provide the best possible estimate of assay sensitivity of Y2H-CCSB, we used three measurements and merged them into a single estimate using the inverse variance weighted method, which combines independent estimates of the same variable by weighting each estimate according to the inverse of its variance[32].

1) A Y2H pair-wise experiment can be seen as a saturated screen, as it overcomes the losses due to pooling, limited selection of positives, sequencing and filtering. The proportion of hsPRS-v1 in a pair-wise Y2H test is used as a first measurement of assay sensitivity. We scored 15 out of 92 hsPRS-v1 pairs as positive in an independent pair-wise test.

2) The proportion of hsPRS-v1 pairs detected in our CCSB-HI1 HT screen reflects the overall sensitivity of Y2H-CCSB, therefore we can use this proportion to estimate assay sensitivity, since assay sensitivity = overall sensitivity / sampling sensitivity. We found 9 out of 92 hsPRS-v1 pairs in CCSB-HI1.

3) We considered a set of 1,526 human LC interactions supported by multiple publications from a more recently updated (January 2007[33]) version of LC interaction databases. The proportion of these LC-multiple pairs detected in CCSB-HI1 reflects the overall sensitivity of Y2H-CCSB, therefore we can use this proportion to estimate assay sensitivity, since assay sensitivity = overall sensitivity / sampling sensitivity. We found 149 out of these 1,526 LC-multiple pairs in CCSB-HI1.

We used the resulting beta distributions Beta(16,78), Beta(10,84) and Beta(140,1378) as independent inputs into a Monte Carlo simulation (see below) in order to get three independent estimates of assay sensitivity and three corresponding estimates of interactome size. We then obtained a single combined estimate of the assay sensitivity and interactome size using the inverse-variance weighted method mentioned above. If $SFN_{2config}$ is the number of undetectable interactions relative to the number of true interactions upon testing pairs in both (bait-prey and prey-bait) configurations, then our single combined estimate of assay sensitivity as obtained above corresponds to $(1\text{-}SFN_{2config})$.

(iii) Estimation of stochastic and systematic false discovery rates of Y2H-CCSB

If $FD_{2config}$ is the combined estimate of systematic and stochastic false discovery rates upon testing pairs in both configurations, $SFD_{2config}$ is the systematic false discovery rate

(*i.e.*, number of systematic false positives relative to the number of interactions reported positive upon testing pairs in both configurations) and $UFD_{2config}$ is the stochastic (or unsystematic) false discovery rate (*i.e.*, the number of stochastic false positives relative to the number of reported interactions upon testing pairs in both configurations),

(4.8) $$FD_{2config} = UFD_{2config} + (1 - UFD_{2config})SFD_{2config}$$

Given that MAPPIT as used in our study reflects an assay independent from Y2H-CCSB, both systematic and stochastic false positives in a given dataset should report negative to a similar extent in MAPPIT. Therefore, the false discovery rate measured from the MAPPIT experiments is a combined estimate of systematic and stochastic false discovery rates (*i.e.*, $FD_{2config}$). The model of the repeat screen data estimates the stochastic false positive rate q as the fraction of non-interacting pairs that are reported positive. From this we obtain the stochastic (or unsystematic) false discovery rate, $UFD_{2config}$ as

(4.9) $$UFD_{2config} = \frac{q_{2config}}{e_{2config}\, p_{2config} + q_{2config}}$$

Thus substituting the value of $UFD_{2config}$ from equation 4.9 and the value of $FD_{2config}$ from the MAPPIT experiment results into equation 4.8, we can obtain the value of $SFD2_{config}$.

(iv) Estimation of the size of the human interactome

The size of the entire human proteome search space assuming one splice isoform per gene (*i.e.*, ignoring splice variants). If N is the number of predicted genes in the human genome, the size of the entire search space is given by N^2.

We denote by e_0 the interactome size relative to the search space size. Among those interactions only a fraction $e_0(1\text{-}SFN_{2config})$ are expected to be uncovered by a given technology with assay sensitivity $1\text{-}SFN_{2config}$. On the other hand, among the interactions reported positive after repeated screens using a given assay, only a fraction $e(1\text{-}SFD_{2config})$ is expected to be true interactions, where $SFD_{2config}$ is the systematic false discovery rate of that assay. Because these two expected values should coincide we obtain

(4.10) $$e_0(1\text{-}SFN_{2config}) = e_{2config}\ (1\text{-}SFD_{2config})$$

from which it follows

(4.11) $$e_0 = e_{2config}\ (1\text{-}SFD_{2config})/(1\text{-}SFN_{2config})$$

Substituting the values of $e_{2config}$ (from equation 4.4), $SFN_{2config}$ (from the hsPRS-v1 experiments) and $SFD_{2config}$ (from equation 4.8) into equation 4.11, we compute the expected size of the human interactome.

The interactome size computed here is likely to be an under-estimate because (i) our measurement of Y2H-CCSB assay sensitivity may be an overestimate given that our hsPRS-v1 pairs are composed of interactions found in multiple experiments in the original literature, (ii) we ignore splice variant complexity.

Monte Carlo Simulations of all reported parameters. We perform Monte Carlo simulations to compute the distribution of different magnitudes which are indirectly estimated from the experimental data, *i.e.*, precision of a given assay such as Y2H-CCSB, sampling sensitivity of Y2H-CCSB, assay sensitivity of a given assay, or the size of the human interactome.

The distribution of the probability p_i that an interaction from dataset *i* [*i*=hsPRS-v1, PRS-Y2H, hsRRS-v1, CCSB-HI1, MDC-HI1, MDC-HI1 (Same, FL), LC, LC (Single, Y2H, FL)] is reported positive in a given assay (Y2H-CCSB or MAPPIT) is estimated from the observation of n_i positive pairs among N_i tested pairs. In the following, we omit the index *i* and the results apply equally to pairs from any dataset. The likelihood to observe these *n* positives after testing *N* pairs is given by a binomial probability distribution with probability *p* as:

(5.1) $$P(n;N,p) = \binom{N}{n} p^n (1-p)^{N-n}$$

Furthermore we assume a uniform prior distribution for *p* in the interval between zero and one. Given the likelihood from the binomial distribution and prior *p*, from the Bayes theorem we obtain the posterior distribution as

(5.2) $$P(p;n,N) = \frac{1.\binom{N}{n} p^n (1-p)^{N-n}}{1.\int_0^1 \binom{N}{n} p^n (1-p)^{N-n} dp}$$

where the numerator is the product of the probability density of the prior and the likelihood given by the binomial, and the integral in the denominator normalized the distribution so that the area under the curve equals unity. Simplifying this expression we obtain that *p* follows a beta distribution *beta*(*p*;*n*+1,*N*-*n*+1) since the probability density function of the beta distribution is given by

(5.3) $$P(p;\alpha_1,\alpha_2) = \frac{p^{\alpha_1-1}(1-p)^{\alpha_2-1}}{\int_0^1 p^{\alpha_1-1}(1-p)^{\alpha_2-1} dp}$$

In a given iteration of the simulation, we use the beta distribution above to obtain a value for the probability of hsPRS-v1 or hsRRS-v1 pairs scoring positive in Y2H-CCSB and the probability of hsPRS-v1, PRS-Y2H, hsRRS-v1, LC, LC (Single, Y2H, FL), MDC-HI1, MDC-HI1 (Same, FL) or CCSB-HI1 pairs scoring positive in MAPPIT. These simulated values are used to compute precision of LC, MDC-HI1 or CCSB-HI1 pairs. We generate 10,000 random values for the parameters *p(PRS-Y2H)*, *p(hsRRS-v1)* and *p(CCSB-HI1)* from the beta distribution. For each value of *p(PRS-Y2H)*, *p(hsRRS-v1)* and *p(CCSB-HI1)*, we compute *precision(CCSB-HI1)* using the equation 1.1 (as described above) and using all the 10,000 *precision(CCSB-HI1)* values we obtain the histogram of *precision(CCSB-HI1)*, mean, empirical standard deviation and empirical 95% confidence intervals. These values are reported in **Supplementary Table 3**.

The distribution of parameters characterizing sampling sensitivity and stochastic-false positive rates are obtained from the analysis of the repeat screens data as described above in Supplementary Methods from which 10,000 values are sampled at random in the Monte Carlo simulations. In each iteration of the simulation, the value of all these parameters are then combined according to equation 4.11 in order to compute a value for the size of the human interactome.

REFERENCES

1. Deane, C.M., Salwinski, L., Xenarios, I. & Eisenberg, D. Protein interactions: two methods for assessment of the reliability of high throughput observations. *Mol. Cell. Proteomics* **1**, 349-356 (2002).
2. Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919-923 (2003).
3. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399-403 (2002).
4. Futschik, M.E., Chaurasia, G. & Herzel, H. Comparison of human protein-protein interaction maps. *Bioinformatics* **23**, 605-611 (2007).

5. Patil, A. & Nakamura, H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics* **6**, 100 (2005).
6. Grigoriev, A. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* **31**, 4157-4161 (2003).
7. D'Haeseleer, P. & Church, G.M. Estimating and improving protein interaction error rates. *Proc. IEEE Comput. Syst. Bioinform. Conf.*, 216-223 (2004).
8. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006).
9. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173-1178 (2005).
10. Cusick, M.E. *et al.* Literature-curated protein interaction datasets. *Nat. Meth.* **In press**.
11. Rual, J.F. *et al.* Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.* **14**, 2128-2135 (2004).
12. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Meth.* **In press**.
13. Lehner, B. & Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol.* **5**, R63 (2004).
14. Rhodes, D.R. *et al.* Probabilistic model of the human protein-protein interaction network. *Nat. Biotechnol.* **23**, 951-959 (2005).
15. Stumpf, M.P. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959-6964 (2008).
16. Chiang, T., Scholtens, D., Sarkar, D., Gentleman, R. & Huber, W. Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol.* **8**, R186 (2007).
17. Huang, H., Jedynak, B.M. & Bader, J.S. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* **3**, e214 (2007).
18. Alfarano, C. *et al.* The Biomolecular Interaction Network Database (BIND) and related tools 2005 update. *Nucleic Acids Res.* **33**, D418-424 (2005).
19. Xenarios, I. *et al.* DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Res.* **30**, 303-305 (2002).
20. Mishra, G.R. *et al.* Human protein reference database–2006 update. *Nucleic Acids Res.* **34**, D411-414 (2006).
21. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572-574 (2007).
22. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832-834 (2005).
23. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957-968 (2005).
24. Walhout, A.J. *et al.* GATEWAY recombinational cloning: application to the cloning of large numbers of open reading frames or ORFeomes. *Methods Enzymol.* **328**, 575-592 (2000).
25. Hartley, J.L., Temple, G.F. & Brasch, M.A. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10**, 1788-1795. (2000).
26. Vidalain, P.O., Boxem, M., Ge, H., Li, S. & Vidal, M. Increasing specificity in high-throughput yeast two-hybrid experiments. *Methods* **32**, 363-370 (2004).
27. Vidal, M., Braun, P., Chen, E., Boeke, J.D. & Harlow, E. Genetic characterization of a mammalian protein-protein interaction domain by using a yeast reverse two-hybrid system. *Proc. Natl. Acad. Sci. USA* **93**, 10321-10326 (1996).
28. Eyckerman, S. *et al.* Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.* **3**, 1114-1119 (2001).

29. Lemmens, I., Lievens, S., Eyckerman, S. & Tavernier, J. Reverse MAPPIT detects disruptors of protein-protein interactions in human cells. *Nat. Protoc.* **1**, 92-97 (2006).
30. Hastie, T., Tibshirani, R. & Friedman, J. The elements of statistical learning. (Springer, New York; 2001).
31. Akaike, H. A new look at the statistical model identification. *IEEE Trans. Aut. Cont.* **AC-19**, 716-723 (1974).
32. Hasselblad, V. Meta-analysis of environmental health data. *Sci. Total Environ.* **160-161**, 545-558 (1995).
33. Simonis, N. *et al.* Empirically-controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. *Nat. Meth.* **In press**.