An empirical framework for binary interactome mapping

Kavitha Venkatesan[1,2,12,13], Jean-François Rual[1,2,12,13], Alexei Vazquez[1,3,4,13], Ulrich Stelzl[5,6,13], Irma Lemmens[7,13], Tomoko Hirozane-Kishikawa[1,2], Tong Hao[1,2], Martina Zenkner[5], Xiaofeng Xin[8], Kwang-Il Goh[1,3,9], Muhammed A Yildirim[1,2,10], Nicolas Simonis[1,2], Kathrin Heinzmann[1,2,12], Fana Gebreab[1,2], Julie M Sahalie[1,2], Sebiha Cevik[1,2,12], Christophe Simon[1,2,12], Anne-Sophie de Smet[7], Elizabeth Dann[1,2], Alex Smolyar[1,2], Arunachalam Vinayagam[5], Haiyuan Yu[1,2], David Szeto[1,2], Heather Borick[1,2,12], Amélie Dricot[1,2], Niels Klitgord[1,2,12], Ryan R Murray[1,2], Chenwei Lin[1,2], Maciej Lalowski[5,12], Jan Timm[5], Kirstin Rau[5], Charles Boone[8], Pascal Braun[1,2], Michael E Cusick[1,2], Frederick P Roth[1,11], David E Hill[1,2], Jan Tavernier[7], Erich E Wanker[5], Albert-László Barabási[1,3,12] & Marc Vidal[1,2]

© 2009 Nature America, Inc. All rights reserved.

Several attempts have been made to systematically map protein-protein interaction, or 'interactome', networks. However, it remains difficult to assess the quality and coverage of existing data sets. Here we describe a framework that uses an empirically-based approach to rigorously dissect quality parameters of currently available human interactome maps. Our results indicate that high-throughput yeast two-hybrid (HT-Y2H) interactions for human proteins are more precise than literature-curated interactions supported by a single publication, suggesting that HT-Y2H is suitable to map a significant portion of the human interactome. We estimate that the human interactome contains ~130,000 binary interactions, most of which remain to be mapped. Similar to estimates of DNA sequence data quality and genome size early in the Human Genome Project, estimates of protein interaction data quality and interactome size are crucial to establish the magnitude of the task of comprehensive human interactome mapping and to elucidate a path toward this goal.

The protein-protein interactome of an organism is the network formed by all protein-protein interactions that can occur at a range of physiologically relevant protein concentrations. Mapping protein-protein interactions is crucial, albeit not sufficient, for unraveling the dynamic aspects of cellular networks—including when, where and for what purpose protein interactions do occur *in vivo*[1]. Currently available human protein-protein interactome maps have been derived using HT-Y2H[2,3], high-throughput coaffinity purification followed by mass spectrometry[4], curation of published low-throughput experiments[5–10] or computational predictions[11,12]. Despite a few attempts[2,3,13,14], it remains difficult to accurately estimate the quality and coverage of these interactome maps.

Differentiation between sets of protein pairs that can interact (biophysical interactions) and do interact (biological interactions) is only possible with reliable biophysical interactome maps. However, several issues remain unresolved, including what proportion of currently available interactome maps represents true biophysical interactions and what proportion represents artifacts; whether the interactions provided by curated low-throughput experiments are

[1]Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, 1 Jimmy Fund Way, Boston, Massachusetts 02115, USA. [2]Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. [3]Center for Complex Network Research and Department of Physics, University of Notre Dame, 225 Nieuwland Science Hall, Notre Dame, Indiana 46556, USA. [4]The Simons Center for Systems Biology, Institute for Advanced Study, Einstein Drive, Princeton, New Jersey 08540, USA. [5]Max Delbrück Center for Molecular Medicine, Robert-Roessle-Straße 10, D-13125 Berlin, Germany. [6]Otto-Warburg Laboratory, Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany. [7]Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie, and Department of Biochemistry, Faculty of Medicine and Health Sciences, Ghent University, Albert Baertsoenkaai 3, 9000 Ghent, Belgium. [8]Banting and Best Department of Medical Research and Department of Molecular Genetics, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College Street, Toronto, Ontario M5S 3E1, Canada. [9]Department of Physics, Korea University, 1 Anam-dong 5-ga, Seongbuk-gu, Seoul 136-713, Korea. [10]School of Engineering and Applied Sciences, Harvard University, 29 Oxford Street, Cambridge, Massachusetts 02138, USA. [11]Department of Biochemistry and Molecular Pharmacology, Harvard Medical School, 250 Longwood Avenue, Boston, Massachusetts 02115, USA. [12]Present addresses: Novartis Institutes for Biomedical Research, 250 Massachusetts Avenue, Cambridge, Massachusetts 02139, USA (K.V.), Department of Cell Biology, Harvard Medical School, 240 Longwood Avenue, Boston, Massachusetts 02115, USA (J.-F.R.), Centre for Cancer Therapeutics, The Institute of Cancer Research, 15 Cotswold Road, Sutton, SM2 5NG, UK (K.H.), University College Dublin, School of Biomolecular and Biomedical Science, Belfield, Dublin 4, Ireland (S.C.), Genome Exploration Research Group, RIKEN Genomic Sciences Center, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan (C.S.), Department of Biological Sciences, Clemson University, 132 Long Hall, Clemson, South Carolina 29634, USA (H.B.), Bioinformatics Program, Boston University, 24 Cummington Street, Boston, Massachusetts 02215, USA (N.K.), Protein Chemistry/Proteomics/Peptide Synthesis and Array Unit, Biomedicum Helsinki, University of Helsinki, Haartmaninkatu 8, FI-00014 Helsinki, Finland (M.L.) and Center for Complex Network Research and Departments of Physics, Biology and Computer Sciences, Northeastern University, 360 Huntington Avenue, Boston, Massachusetts 02115, USA (A.-L.B.). [13]These authors contributed equally to this work. Correspondence should be addressed to M.V. (marc_vidal@dfci.harvard.edu), A.-L.B. (a.barabasi@neu.edu), E.E.W. (ewanker@mdc-berlin.de) or J.T. (jan.tavernier@ugent.be).

RECEIVED 4 JUNE; ACCEPTED 10 NOVEMBER; PUBLISHED ONLINE 7 DECEMBER 2008; DOI:10.1038/NMETH.1280

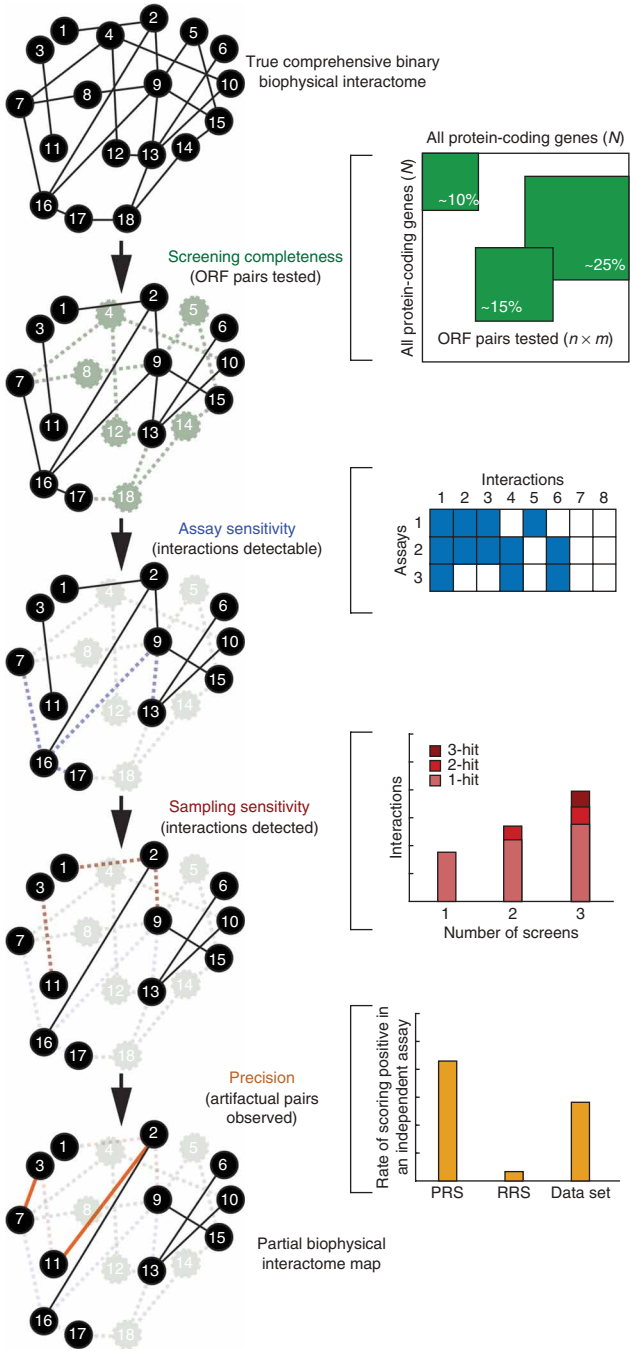

Figure 1 | Conceptual framework for interactome mapping. The concepts of screening completeness (fraction of all pairwise protein combinations tested), assay sensitivity (fraction of all biophysical interactions identifiable by a given assay), sampling sensitivity (fraction of all identifiable interactions that are detected in a single trial) and precision (fraction of pairs reported by a given assay that are true positives) can be estimated independently and combined to empirically estimate the size of binary interactomes. Solid black lines in a given network graph represent true biophysical interactions present in that network; dashed lines represent true biophysical interactions missing from that network; and solid colored lines represent biophysical artifactual pairs present in that network.

© 2009 Nature America, Inc. All rights reserved.

superior in quality to those obtained by high-throughput strategies, as suggested previously[15–17]; and whether the currently available interactome maps represent a significant or a negligible fraction of the human biophysical interactome. Here we provide insights that are crucial for developing a strategy for comprehensive interactome mapping—that is, for estimating the size of the human interactome and thus an endpoint to the project, and for selecting suitable technologies, a realistic timeline and a funding model to achieve this goal.

Previous attempts to assess the quality of interactome maps for humans[13,14,18] or other species[13,15,18–23] relied on measuring either the extent to which interacting proteins share other biological attributes, such as coexpression, or the extent to which different maps of the same interactome share common interactions. Both approaches suffer several inherent limitations. Methods that evaluate the quality of interactions with respect to mRNA coexpression[22,23] are systematically biased against true biological interactions between proteins whose mRNAs are not necessarily correlated, or are even anticorrelated, in expression. Because available annotations for protein function and localization are far from comprehensive, lack of evidence for colocalization of a given pair of proteins does not imply that the interaction observed between these proteins is an artifact. Methods based on measuring the extent of overlap between two interactome maps[13,20,21] require that the corresponding data sets be derived from identical or similar assays. Existing analyses have not always fulfilled this requirement[13]. Most existing methods for quality assessment do not distinguish between the multiple sources of false negatives and false positives associated with any interactome mapping strategy. For instance, interactions missed by a single screen of an assay but identifiable after multiple screens must be distinguished from interactions that would never be identified by that assay even after a saturating number of screens.

Here we developed a framework to estimate various quality parameters associated with currently used protein-protein interaction assays, namely screening completeness, assay sensitivity, sampling sensitivity and precision. We generated empirical data to rigorously dissect these quality parameters without relying on correlation with other biological attributes. Combining these parameters provides an estimate of the size of the human binary biophysical interactome and projects a path toward the completion of its mapping.

RESULTS

An interaction mapping framework

To accurately assess the quality of a given interactome map, we need to consider every possible source of false negatives (true interactions missing) and false positives (spurious pairs reported) associated with the assay used to generate the map. Our framework considers four parameters to estimate quality: screening completeness, assay sensitivity, sampling sensitivity and precision (**Fig. 1**).

Screening completeness is the fraction of the total possible space of open reading frame (ORF) pairs that is tested to generate a given interactome map. Because currently available ORF resources[3,24] only allow proteome-wide investigations of one protein isoform per gene, we ignored isoforms encoded by alternatively spliced transcripts here. For example, if we assume that the human genome consists of 22,500 protein-coding genes ($N = 22{,}500 \times 22{,}500/2$ protein pairs), then the screening completeness of the Center for Cancer Systems Biology Human Interactome version 1 (CCSB-HI1) data set[2], a proteome-scale HT-Y2H effort that tested $n = 7{,}000 \times 7{,}000/2$ human protein pairs, is n/N, or $\sim 10\%$.

Assay sensitivity is the fraction of all biophysical interactions that can possibly be identified by an assay conducted under a specific set of experimental conditions. For example, a given HT-Y2H assay may be unable to detect interactions involving specific types of membrane proteins or requiring post-translational modifications that do not occur in yeast cells.

Sampling sensitivity is the fraction of all identifiable interactions that are found in a single trial of an assay conducted under a specific set of experimental conditions. When testing tens, if not hundreds, of millions of protein pairs in any space of pairwise combinations, it might be necessary to sample that space multiple times to report all identifiable interactions.

Lastly, precision is the fraction of observed pairs in an interactome data set that are true positives. False-positive pairs reflect technical artifacts that erroneously score positive in a given assay conducted under a specific set of experimental conditions. We distinguished between two types of artifactual pairs: stochastic false positives, which are observed in only one or a few trials of an assay, and systematic false positives, which are observed in many or all trials.

© 2009 Nature America, Inc. All rights reserved.

Estimation of assay sensitivity

Estimation of the assay parameters described above requires reference sets of positive and negative interacting pairs. To compile a positive reference set (PRS) of high-confidence human binary protein-protein interactions, we started with interactions curated from the literature. From these, we chose 188 pairs present in our human ORFeome v1.1 clone collection[24] that are supported by the greatest number of publications and curated by the highest number of databases. Systematic recuration of all publications thought to support these 188 protein pairs[25] verified 107 direct binary interactions that involve human proteins and are supported by multiple publications. Ninety-two of these interactions involve full-length proteins and constituted our *Homo sapiens* PRS version 1 (hsPRS-v1; **Fig. 2a** and **Supplementary Table 1** online). Proteins involved in the 92 hsPRS-v1 interactions show broad cellular localization (**Fig. 2b**), suggesting that they are representative of the entire human proteome. It is impossible to generate a set of negative interacting pairs with absolute confidence, so we compiled a surrogate random reference set (hsRRS-v1) of 188 protein pairs chosen randomly from the space of all ORFeome v1.1 pairs after excluding known interactions (**Fig. 2c**).

PRS and RRS pairs can be used to experimentally calibrate conditions of an assay to achieve an optimal trade-off between the fraction of PRS and RRS pairs reporting positive[26]. We measured the fraction of hsPRS-v1/hsRRS-v1 pairs scoring positive across a range of experimental and scoring conditions of a stringent version of the Y2H system (Y2H-CCSB)[2] and the mammalian protein-protein interaction trap assay (MAPPIT)[27] (**Supplementary Table 2** online and **Fig. 2d**,**e**).

The results for the hsPRS-v1 and hsRRS-v1 pairs with Y2H-CCSB confirmed that the specific experimental conditions used in generating our first human interactome map, CCSB-HI1 (ref. 2), reflected good assay design. We also derived suitable experimental conditions for the MAPPIT assay. Under these experimental conditions, we estimated the assay sensitivity of Y2H-CCSB and MAPPIT to be 17% and 21%, respectively (**Fig. 2f** and **Supplementary Table 3** online). Using a larger, more recently updated set of ~1,500 literature-curated interactions that are supported by multiple publications, we estimated an assay sensitivity of 20% for Y2H-CCSB, consistent with our hsPRS-v1–based estimate. Y2H-CCSB and MAPPIT recovered partially overlapping sets of hsPRS-v1 interactions. Of the 92 hsPRS-v1 pairs, 27 (29%) were reported by at least one assay, and of those, 7 (26%) were detected by both assays (**Fig. 2f**,**g**). That 20 (74%) of the 27 positive hits are specific to a single assay reflects complementarities between the two assays.

We estimated the false-positive rate (rate at which hsRRS-v1 pairs scored positive) of Y2H-CCSB and MAPPIT to be $<0.5\%$ and 2%, respectively (**Fig. 2f** and **Supplementary Table 4** online). The results of testing hsRRS-v1 pairs by Y2H-CCSB do not permit a direct and reasonable estimate of the false-discovery rate associated with the CCSB-HI1 data set. The millions of pairs tested by Y2H-CCSB in the high-throughput screen leading to the generation of CCSB-HI1 consist mostly of noninteracting pairs, so the number of noninteracting pairs tested in the HT-Y2H screen is orders of magnitude higher than the size of hsRRS-v1. Consequently, small changes in the hsRRS-v1–based estimate of the false-positive rate of Y2H-CCSB can have a large effect on the resulting estimate of the false-discovery rate of CCSB-HI1. Rather than using the Y2H-CCSB experiments on the hsRRS-v1 pairs, we instead used two alternate and independent approaches to estimate the false-discovery rate of our Y2H-CCSB assay: retesting Y2H-CCSB interactions in MAPPIT, and modeling repeated screens of Y2H-CCSB.

Precision of existing human interactome data sets

We estimated the precision of the two existing HT-Y2H human interactome data sets, CCSB-HI1 (ref. 2) and Max Delbrück Center for Molecular Medicine Human Interactome version 1 (MDC-HI1; ref. 3), as well as a low-throughput literature-curated human interactome data set[2], by measuring the extent to which a subset of 188 positive pairs chosen randomly from each data set (**Supplementary Table 1**) retested in MAPPIT. To do so, we first benchmarked the performance of each data set in MAPPIT experiments against the false-positive rate of MAPPIT and the false-negative rate of MAPPIT. We estimated these benchmarks by evaluating the fraction of hsRRS-v1 and hsPRS-v1 pairs reporting positive by MAPPIT, respectively. The results with the hsRRS-v1 pairs provided an estimate of MAPPIT's false-positive rate that was sufficiently resolved for estimating false-discovery rates of the various interactome data sets, as the size of the hsRRS-v1 is similar to the size of each of the three different interactome data sets tested. Relative to the proportion of hsPRS-v1 and hsRRS-v1 pairs scoring positive (21% and 2%, respectively), the fractions of pairs that scored positive in the three data sets were 8% for literature curated, 10% for MDC-HI1 and 27% for CCSB-HI1 (**Fig. 3a** and **Supplementary Table 4**).

To adjust the analysis for potential data set biases, we first minimized the effect of differences between the sequences of the clones originally used to report the interactions and sequences of the full-length clones used here. We considered only pairs for which the proteins originally used were described as full length or, whenever identifiable, pairs for which the isoforms originally used were the same ('same') as the ones used here. Second, because the CCSB-HI1 and MDC-HI1 data sets were each described in a single publication, we compared them to the subset of literature-curated interactions also supported by a single publication

Figure 2 | Assay sensitivity and background positive rate of binary interactome mapping assays. (**a**) Method by which hsPRS-v1 interactions were chosen from the interactions available in the curated literature of low-throughput experimentally derived interactions (LC). (**b**) Distribution of cellular location of proteins in the hsPRS-v1 and hsRRS-v1. (**c**) Method by which hsRRS-v1 pairs were chosen from the possible pairs in our human ORFeome v1.1 clone collection[24]. (**d**) Assay sensitivity (fraction of hsPRS-v1 pairs scoring positive) and background positive rate (fraction of hsRRS-v1 pairs scoring positive) of the Y2H-CCSB assay based on varying experimental and scoring conditions, including use of an alternate protocol (**Supplementary Methods**). Because of limited sample size, we did not use the results of testing the hsRRS-v1 pairs to estimate the false-discovery rate of the Y2H-CCSB assay. (**e**) Assay sensitivity and background positive rate of the MAPPIT assay after varying experiment-to-control-ratio (ECR) scores (**Supplementary Methods**). (**f**) Top, assay sensitivity and background positive rate of Y2H-CCSB and MAPPIT under the specific experimental conditions used in the rest of this study (**Supplementary Methods**). For Y2H-CCSB, the fraction of hsPRS-v1 pairs scoring positive in at least one configuration and in both pairwise mating experiments is shown. This condition reflects the assay sensitivity of the specific experimental and scoring conditions of Y2H-CCSB used to generate CCSB-HI1 (ref. 2). Bottom, Venn diagram of hsPRS-v1 pairs scoring positive in the two assays. (**g**) Results of testing each hsPRS-v1 pair and each hsRRS-v1 pair using Y2H-CCSB and MAPPIT. Yellow or blue shaded squares represent protein pairs scored positive by a given assay.

© 2009 Nature America, Inc. All rights reserved.

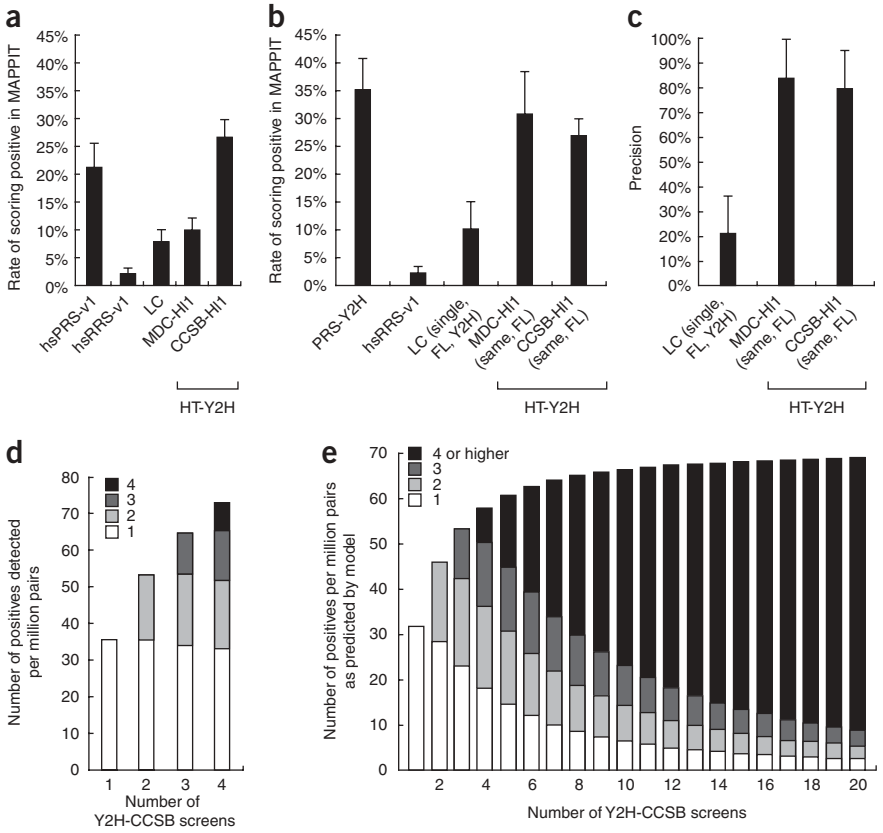

Figure 3 | Precision and sampling sensitivity in interactome data sets. (**a**) Comparison of interactome data sets by comparing the rate of observing a positive by MAPPIT given a positive in the data set. LC, literature-curated interactions. (**b**) Interactome data sets were further compared after removing various biases by considering interactions originally derived using full-length (FL) proteins and using Y2H assays. (**c**) Precision of each tested data set computed by accounting for the rate of detecting hsRRS-v1 pairs and Y2H-supported hsPRS-v1 pairs by MAPPIT in **b**. Error bars represent estimated s.d. of the mean using a Monte Carlo simulation of scores observed in a given assay. (**d**,**e**) Sampling sensitivity and Y2H-CCSB repeat screens. White bars represent protein pairs uncovered in only one screen; progressively darker shades of gray represent protein pairs reported in increasing numbers of screens. (**d**) Total numbers of positive pairs reported after one, two, three or four Y2H-CCSB repeat screens. (**e**) Predicted saturation curve of the number of uncovered interactions against the number of screens for Y2H-CCSB after modeling the data in **d** and assuming a single isoform per gene in the search space.

© 2009 Nature America, Inc. All rights reserved.

('single'), which represents most currently available literature-curated interaction information[25]. Including interactions supported by multiple publications in the literature-curated data set would be circular, as our hsPRS-v1 benchmark was derived from literature-curated interactions supported by multiple publications. Lastly, to account for the moderate bias of MAPPIT in detecting Y2H-supported ('Y2H') interactions, we considered the subset of hsPRS-v1 and literature-curated pairs supported by at least one Y2H experiment in the corresponding curated publications (**Supplementary Data 1** and **Supplementary Table 5** online). Based on these consolidated data sets, 34% of Y2H-supported hsPRS-v1 pairs ('PRS-Y2H') and 2% of hsRRS-v1 pairs scored positive. Relative to this, the fractions of pairs that scored positive in the three subsets of protein pairs were 10% for literature curated (single, full length, Y2H), 31% for MDC-HI1 (same, full length) and 27% for CCSB-HI1 (same, full length; **Fig. 3b**). Thus, the two HT-Y2H data sets performed comparably to the PRS-Y2H pairs in MAPPIT, whereas the literature-curated interactions supported by a single publication performed poorly. Given the fraction of PRS-Y2H pairs and hsRRS-v1 pairs scoring positive by MAPPIT, we computed the precision of each of the three data sets as 25% for literature curated (single, full length, Y2H), 83% for MDC-HI1 (same, full length) and 79% for CCSB-HI1 (same, full length; **Fig. 3c** and **Supplementary Table 3**).

Sampling sensitivity and stochastic false-discovery rate

To estimate sampling sensitivity and the number of screens required to achieve saturation, we repeated four Y2H-CCSB screens ('repeat screens') in a defined search space of 1,822 DB-Xs ('baits' representing 1,744 unique genes) against 1,796 AD-Ys ('preys' representing 1,752 unique genes), representing ~3 million pairwise combinations (**Supplementary Table 6** online). We developed a probabilistic model that considered the search space of 3 million protein pairs to be a mixture of true biophysical interactions and noninteracting pairs. Using a bayesian approach, our model estimated the fraction of all identifiable true biophysical interactions found in one, two, or a saturating number of screens, and the fraction of noninteracting pairs erroneously detected in a screen. In short, our approach estimated distributions of values of the above parameters that fit the experimental results observed in the repeat screens.

Of the 3 million pairwise combinations tested, the four Y2H-CCSB repeat screens together reported 240 protein-protein interactions (**Supplementary Tables 7** and **8** online). Of these interactions, 49% appeared in multiple screens. The total number of new interactions identified after successive screens showed an increasing trend, indicating that more interactions would be found with additional screens (**Fig. 3d**). On the basis of our model, we estimated that the sampling sensitivity per screen is 45% and that after a saturating number of screens, Y2H-CCSB can identify 71 interactions per million pairs tested (**Fig. 3e**). Approximately six screens are needed to reach 90% saturation. Notably, the number of single hits (interactions found in only one of several screens) decreases, whereas the contribution of multiple hits dominates after multiple screens. Adjusting for these repeat screens being done in only one Y2H configuration (bait-prey versus prey-bait), we estimated that after testing both configurations, the sampling sensitivity per screen is 53%, and that after a saturating number of screens, Y2H-CCSB can identify 118 interactions per million pairs tested (**Supplementary Table 3**).

Our model estimated that approximately eight of every million noninteracting pairs tested falsely report positive in Y2H-CCSB. Consequently, our model estimated a stochastic false-discovery rate of 12%, meaning that 12% of the interactions reported in a single

Y2H-CCSB screen correspond to stochastic false positives. Because the MAPPIT experiments (**Fig. 3c**) evaluated the union of systematic and stochastic false positives in a given data set and estimated an overall false-discovery rate of 21% for CCSB-HI1, we deduced a systematic false-discovery rate of 14% (**Supplementary Methods** online).

The MAPPIT experiments showed that existing human HT-Y2H maps have high precision. However, the fraction of CCSB-HI1 and MDC-HI1 interactions common to both maps is small, although statistically significant—only 6% and 2%, respectively ($P = 10^{-18}$; **Supplementary Data 2**, **Supplementary Fig. 1** and **Supplementary Table 9** online). Our results indicate that low sampling sensitivity and differences in assay sensitivity are likely to account for the small overlap.

Estimation of the size of the human interactome

We estimated four important parameters associated with the quality of human binary interactome maps (**Fig. 1**). For the Y2H-CCSB assay evaluated here, we computed the screening completeness of the repeat screens as ~1%; the hsPRS-v1 experiment estimated an assay sensitivity of ~17% (**Fig. 2f**); the model of the repeat screens estimated a sampling sensitivity of ~53% (**Fig. 3e**); and the MAPPIT experiment estimated a precision of ~79% (**Fig. 3c**). We also estimated the variation of these estimates associated with sampling (**Supplementary Table 3**). Integrating these parameters, we predict that the entire human interactome, excluding splice variant complexity, contains 74,000–200,000 binary biophysical interactions (**Table 1**).

Interacting protein pairs and shared functional annotation

A statistically significant fraction ($P < 10^{-3}$), but not all, of the interacting protein pairs in CCSB-HI1 and MDC-HI1 shared functional annotations compared to random expectation (**Fig. 4**). Given the high technical quality of these data sets shown here, interacting pairs that do not share known functional annotations could be promising candidates for biological discovery, particularly true biological interactions that involve proteins currently lacking adequate functional annotations, or they could be true biophysical interactions that do not occur physiologically. We call this latter class 'pseudointeractions', by analogy to pseudogenes. Pseudointeractions could correspond to ancient biological interactions that have evolved to lose physiological relevance and provide interesting insights into the evolution of the interactome.

Table 1 | Sizing the human interactome

Parameter	Test	Result
Average number of interactions detected per screen	Repeat screen experiment	199
Screening completeness of repeat screen search space	Ensembl version 44.36f	1.2%
Assay sensitivity	hsPRS-v1 experiment	17 ± 3.8%
Sampling sensitivity per repeat screen	Repeat screen experiment	53.1 ± 10%
Precision	MAPPIT experiment	79.4 ± 15.9%
Systematic false-discovery rate	MAPPIT and repeat screen experiments	13.6 ± 14.5%
Stochastic false-discovery rate	Repeat screen experiment	11.7 ± 6.1%
Size of the human interactome	Combining all parameters	130,111 ± 32,618 (73,548–199,688)[a]

Estimation of parameters of the binary interactome mapping framework. Average values ± 1 s.d. of each parameter, based on Monte Carlo simulation ($n = 10{,}000$ runs), are indicated after accounting for testing protein pairs in both bait-prey and prey-bait configurations. [a]The range of interactome size reported corresponds to the 95% confidence interval of the size predicted.

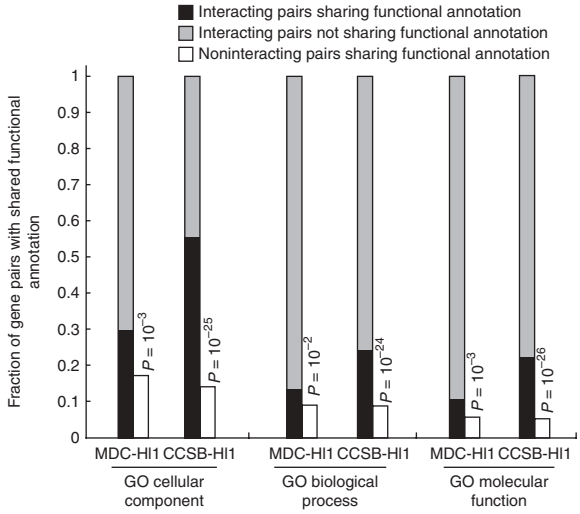

Figure 4 | Analysis of interacting pairs in CCSB-HI1 and MDC-HI1 interactome maps for their ability to share specific Gene Ontology (GO) functional annotations. *P* values indicate the probability of observing shared annotation by chance (compare black bars to white bars), computed using Fisher's exact test. Gray bars reflect the fractions of interacting pairs that do not share specific GO functional annotations. Analysis was done on MDC-HI1 and CCSB-HI1 interactions reported using full-length ORFs.

DISCUSSION

Several previous studies have estimated the precision of existing maps or the size of interactomes[13–15,18,20–23,28]. Our empirical framework addresses limitations of these studies (detailed discussion in **Supplementary Data 3** online). Methods that rely on correlation with other biological attributes to estimate the precision of interactome maps often use as a benchmark literature-curated interaction data sets, which are sociologically biased; assume that knowledge of biological attributes, such as Gene Ontology functional annotation, is complete and unbiased; and are inherently constrained by preexisting paradigms regarding the expectation for interacting protein pairs to share biological attributes. Approaches based on analyzing the extent of overlap between interactome maps[13,20,21] suffer specific limitations in their implementations, such as comparing maps that were not derived using the same assay, or using literature-curated data sets as a reference set, which may not be appropriate given a potentially higher false-positive rate than previously anticipated (**Fig. 3c**)[25]. Earlier studies also did not consider one or more of the parameters that influence interactome map quality—completeness, systematic false-discovery rate, stochastic false-discovery rate, assay sensitivity and sampling sensitivity—which could in turn significantly affect estimates of interactome size. Together, these limitations may have led to overestimated false-discovery rates for HT-Y2H human interactome maps.

Our framework overcomes these limitations by considering every possible source of false negatives and false positives, using a high-quality reference set requiring interactions to be supported by multiple publications and to pass additional recuration, assessing

© 2009 Nature America, Inc. All rights reserved.

false-discovery rates directly using information from independent protein-protein interaction assays, and comparing overlaps between four homogeneously derived repeat screens to assess the sampling sensitivity and stochastic false-discovery rate of Y2H-CCSB. Close attention to these parameters will be vital to design the strategy, such as the number of screens and types of assays to use, for future interactome mapping projects.

The hsPRS-v1 and hsRRS-v1 provide hundreds of experimentally testable clone pairs of positive and random reference sets for binary protein-protein interactions. Previous assays typically relied on testing one or a few positive control pairs and a few or no random control pairs. Although our reference sets are a first version and will be improved, they mark a substantial effort toward the standardized calibration of binary interaction mapping assays, an objective that has not been previously achieved systematically.

Although interaction data sets curated from low-throughput literature are commonly perceived to be of better quality than data sets generated with high-throughput technologies[15–17], the results of our MAPPIT experiments indicate that stringent implementations of HT-Y2H assays produce interaction data sets with technical quality at least as good as, if not superior to, literature-curated interactions (**Fig. 3c**). These results substantiate previous computational analyses of human[29] and yeast[30] interactome maps. Large-scale curation of the primary literature is challenging and may have higher error rates than previously anticipated[25]. High-throughput interactome mapping strategies have several advantages over low-throughput strategies: (i) because defined search spaces are used, information about positives (pairs observed to interact) and negatives (pairs not observed to interact) is available; (ii) experiments are standardized and therefore well controlled, comparable and scalable; (iii) cost-efficient strategies can be developed; and (iv) high-throughput strategies are less sociologically biased than low-throughput experiments.

Implementation of our framework can be improved in various ways. The statistical power of the analyses can be increased by testing more PRS interactions, repeatedly screening larger search spaces or using additional independent assays for measuring precision. Our current implementation does not consider multiple splice isoforms per gene, so we are most likely to underestimate the interactome size. Additional modifications to the framework will be required to thoroughly analyze nonbinary complex comembership maps, such as those generated by high-throughput coaffinity purification followed by mass spectrometry[4]. More refined estimates can be made with future enhancements, but the central concepts and overall approach are in place for design and comprehensive evaluation of any interactome mapping assay. Our group recently developed an interaction tool kit consisting of four independent assays to evaluate the quality of any protein interaction data set[26]. Ongoing technological advancements related to assay automation and cost reduction will enable testing of expanded versions of the PRS and thousands (rather than hundreds) of Y2H, literature-curated and other interactions using these assays.

Similar to estimates of the number of protein-coding genes in the human genome, ~14,000–300,000 in the early 1990s[31], empirical sizing of the interactome is crucial to establish the complexity of the network and to estimate how far we are from a complete human interactome map. Assuming one splice isoform per gene, we predict that the size of the human interactome is ~130,000 interactions. This confirms two previous estimates of human interactome size, which ranged from 150,000 to 370,000 interactions[2,13]. Of the ~23,000 currently reported human interactions (a combination of ~17,000 literature-curated interactions and ~6,000 HT-Y2H interactions), our measurements indicate that ~10,000 (~42%) are true binary physical interactions (**Supplementary Data 4** online). Thus, the fraction of interactions identified so far represents ~8% of the full interactome.

With 22,500 protein-coding genes, nearly 250 million protein pairs need to be tested individually, clearly requiring unbiased, systematic and cost-effective high-throughput approaches. Interactome mapping is gradual: six screens are necessary to reach 90% saturation with Y2H-CCSB. No single assay offers complete assay sensitivity. The fraction of protein-protein interactions detectable by the specific version of HT-Y2H used here (Y2H-CCSB) is ~17%. Combining different versions of the Y2H system and using increased expression of both hybrid proteins can increase this proportion to ~40% (data not shown and ref. 26). Still, comprehensive mapping of the interactome will require the development of additional high-throughput versions of MAPPIT and other assays[26].

The potential impact on biology of a complete and reliable biophysical protein interaction map cannot be overestimated. Our results offer a quantitative roadmap in this direction, uncovering both the magnitude of the task ahead as well as the potential roadblocks.

METHODS

Overview. The Y2H-CCSB experiments were done as described[2] with minor changes. MAPPIT experiments were done essentially as described[32]. Mathematical modeling of the repeat screens was done using a bayesian approach. All parameters observed from either the experimental data or the mixture model were used as inputs into a Monte Carlo simulation to calculate the corresponding magnitudes of corresponding numbers reported in the text. Detailed descriptions of all data sets and methods can be found in **Supplementary Methods**.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

We thank members of CCSB and the Vidal, Barabasi, Wanker and Tavernier laboratories and S. Sahasrabuddhe, R. Bell, R. Chettier and C. Wiggins for helpful discussions; E. Smith for help generating **Figure 1**; and Agencourt Biosciences for sequencing assistance. This work was supported by the US National Human Genome Research Institute (2R01HG001715 and 5P50HG004233 to M.V. and F.P.R.), the US National Cancer Institute (5U54CA112952 to J. Nevins, subcontract to M.V.; and 5U01CA105423 to S.H. Orkin, project to M.V.), the US National Institutes of Health (IH U01 A1070499-01 and U56 CA113004 to A.-L.B. and postdoctoral training grant fellowship T32CA09361 to K.V.), the Ellison Foundation (to M.V.), the W.M. Keck Foundation (to M.V.), Dana-Farber Cancer Institute Institute Sponsored Research funds (to M.V.), the US National Science Foundation (ITR DMR-0926737 and IIS-0513650 to A.-L.B.), Deutsches Bundesministerium für Bildung und Forschung (NGFN2, KB-P04T01, KB-P04T03 and 01GR0471 to E.E.W. and U.S.), Deutsche Forschungsgemeinschaft (SFB 577 and SFB618 to E.E.W.), the University of Ghent and the "Fonds Wetenschappelijk Onderzoek–Vlaanderen" (FWO-V) G.0031.06 (GOA12051401 to J. Tavernier) and the National Cancer Institute of Canada (to C.B.). I.L. is a postdoctoral fellow with the FWO-V. M.V. is a "Chercheur Qualifié Honoraire" from the Fonds de la Recherche Scientifique (French Community of Belgium).

AUTHOR CONTRIBUTIONS

K.V., J.-F.R., A. Vazquez, U.S., I.L., J. Tavernier, E.E.W., A.-L.B. and M.V. conceived the project. K.V., J.-F.R., A. Vazquez, U.S. and I.L. coordinated the experiments and data analyses. J.-F.R., U.S., T.H.-K., M.Z., X.X., K.H., F.G., J.M.S., P.B., H.Y.,

© 2009 Nature America, Inc. All rights reserved.

S.C., C.S., E.D., J. Timm, K.R. and C.B. conducted the Y2H experiments. J.-F.R., T.H.-K. and C.S. conducted the high-throughput ORF cloning for MAPPIT experiments. I.L. and A.-S.d.S. conducted the MAPPIT experiments. K.V., A. Vazquez, T.H., K.-I.G., M.A.Y., A. Vinayagam, N.S., N.K., C.L., M.L. and F.P.R. conducted the computational and statistical analyses. M.E.C., A.S., H.B., J.-F.R. and K.V. conducted the literature-curated interaction recuration analyses. D.S., A.D. and R.R.M. provided laboratory support. K.V., J.-F.R., A. Vazquez, U.S., I.L., M.E.C., D.E.H., J. Tavernier, E.E.W., A.-L.B. and M.V. wrote the manuscript. D.E.H, J. Tavernier, E.E.W., A.-L.B. and M.V. codirected the project.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Vidal, M. Interactome modeling. *FEBS Lett.* **579**, 1834–1838 (2005).
2. Rual, J.F. *et al.* Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **437**, 1173–1178 (2005).
3. Stelzl, U. *et al.* A human protein-protein interaction network: a resource for annotating the proteome. *Cell* **122**, 957–968 (2005).
4. Ewing, R.M. *et al.* Large-scale mapping of human protein-protein interactions by mass spectrometry. *Mol. Syst. Biol.* **3**, 89 (2007).
5. Peri, S. *et al.* Human protein reference database as a discovery resource for proteomics. *Nucleic Acids Res.* **32**, D497–D501 (2004).
6. Zanzoni, A. *et al.* MINT: a Molecular INTeraction database. *FEBS Lett.* **513**, 135–140 (2002).
7. Bader, G.D. *et al.* BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res.* **29**, 242–245 (2001).
8. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452–D455 (2004).
9. 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).
10. Mewes, H.W. *et al.* MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.* **30**, 31–34 (2002).
11. Ramani, A.K., Bunescu, R.C., Mooney, R.J. & Marcotte, E.M. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.* **6**, R40 (2005).
12. Lehner, B. & Fraser, A.G. A first-draft human protein-interaction map. *Genome Biol.* **5**, R63 (2004).
13. Hart, G.T., Ramani, A.K. & Marcotte, E.M. How complete are current yeast and human protein-interaction networks? *Genome Biol.* **7**, 120 (2006).
14. Futschik, M.E., Chaurasia, G. & Herzel, H. Comparison of human protein-protein interaction maps. *Bioinformatics* **23**, 605–611 (2007).
15. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein-protein interactions. *Nature* **417**, 399–403 (2002).
16. Reguly, T. *et al.* Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J. Biol.* **5**, 11 (2006).
17. Gandhi, T.K. *et al.* Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* **38**, 285–293 (2006).
18. Patil, A. & Nakamura, H. Filtering high-throughput protein-protein interaction data using a combination of genomic features. *BMC Bioinformatics* **6**, 100 (2005).
19. 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).
20. D'Haeseleer, P. & Church, G.M. Estimating and improving protein interaction error rates. *Proc. IEEE Comput. Syst. Bioinform. Conf.* 216–223 (2004).
21. Grigoriev, A. On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res.* **31**, 4157–4161 (2003).
22. 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).
23. Sprinzak, E., Sattath, S. & Margalit, H. How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919–923 (2003).
24. Rual, J.F. *et al.* Human ORFeome version 1.1: a platform for reverse proteomics. *Genome Res.* **14**, 2128–2135 (2004).
25. Cusick, M.E. *et al.* Literature-curated protein interaction datasets. *Nat. Methods* (in the press).
26. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* advance online publication, doi:10.1038/nmeth.1281 (7 December 2008).
27. Eyckerman, S. *et al.* Design and application of a cytokine-receptor-based interaction trap. *Nat. Cell Biol.* **3**, 1114–1119 (2001).
28. Stumpf, M.P. *et al.* Estimating the size of the human interactome. *Proc. Natl. Acad. Sci. USA* **105**, 6959–6964 (2008).
29. Ramírez, F., Schlicker, A., Assenov, Y., Lengauer, T. & Albrecht, M. Computational analysis of human protein interaction networks. *Proteomics* **7**, 2541–2552 (2007).
30. Collins, S.R. *et al.* Towards a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics* **6**, 439–450 (2007).
31. Fields, C., Adams, M.D., White, O. & Venter, J.C. How many genes in the human genome? *Nat. Genet.* **7**, 345–346 (1994).
32. 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).

© 2009 Nature America, Inc. All rights reserved.