Vast uncharted territories in the human interactome landscape uncovered by a second-generation map

Murat Taşan\textsuperscript{1,2,4*}, Thomas Rolland\textsuperscript{1,5*}, Samuel J. Pevzner\textsuperscript{1,5,6,7*}, Benoît Charloteaux\textsuperscript{1,5*}, Irma Lemmens\textsuperscript{6}, Celia Fontanillo\textsuperscript{9}, Roberto Mosca\textsuperscript{10}, Nidhi Sahni\textsuperscript{1,5}, Song Yi\textsuperscript{1,5}, Atanas Kamburov\textsuperscript{1,5}, Susan Ghiassian\textsuperscript{1,11}, Quan Zhong\textsuperscript{1,5}, Xinping Yang\textsuperscript{1,5}, Lila Ghamsari\textsuperscript{1,5}, Dawit Balcha\textsuperscript{1,5}, Pascal Braun\textsuperscript{1,5}, Martin P. Broly\textsuperscript{1,5}, Anne-Ruxandra Carvunis\textsuperscript{1,5}, Dan Convery-Zupan\textsuperscript{1,5}, Roser Corominas\textsuperscript{12}, Elizabeth Dann\textsuperscript{1,5}, Matija Dreze\textsuperscript{1,5}, Amélie Dricot\textsuperscript{1,5}, Changyu Fan\textsuperscript{1,5}, Eric Franzosa\textsuperscript{1,13}, Fana Gebreab\textsuperscript{1,5}, Bryan J. Gutierrez\textsuperscript{1,5}, Mike Jin\textsuperscript{1,5}, Shuli Kang\textsuperscript{1,5}, Ruth Kiros\textsuperscript{1,5}, Guan Ning Lin\textsuperscript{1,5}, Andrew MacWilliams\textsuperscript{1,5}, Jörg Menche\textsuperscript{1,11}, Ryan R. Murray\textsuperscript{1,5}, Matthew M. Poulin\textsuperscript{1,5}, Xavier Rambout\textsuperscript{1,5,14}, John Rasla\textsuperscript{1,5}, Patrick Reichert\textsuperscript{1,5}, Viviana Romero\textsuperscript{1,5}, Julie M. Sahalie\textsuperscript{1,5}, Annemarie Scholz\textsuperscript{1,5}, Akash A. Shah\textsuperscript{1,5}, Amitabh Sharma\textsuperscript{1,11}, Yun Shen\textsuperscript{1,5}, Stanley Tam\textsuperscript{1,5}, Shelly A. Trigg\textsuperscript{1,5}, Jean-Claude Twizere\textsuperscript{1,5,14}, Kerwin Vega\textsuperscript{1,5}, Jennifer Walsh\textsuperscript{1,5}, Michael E. Cusick\textsuperscript{1,5}, Yu Xia\textsuperscript{1,13}, Albert-László Barabási\textsuperscript{1,11,15}, Lilia Iakoucheva\textsuperscript{1,5}, Patrick Aloy\textsuperscript{1,16}, Javier De Las Rivas\textsuperscript{5}, Jan Tavernier\textsuperscript{8}, Michael A. Calderwood\textsuperscript{1,5}, David E. Hill\textsuperscript{1,5}, Tong Hao\textsuperscript{1,5}, Frederick P. Roth\textsuperscript{1,2,3,4} & Marc Vidal\textsuperscript{1,5}

\textsuperscript{1}Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02215, USA. \textsuperscript{2}Departments of Molecular Genetics and Computer Science, University of Toronto, Toronto, Ontario M5S 3E1, Canada. \textsuperscript{3}Donnelly Centre, University of Toronto, Toronto, Ontario M5S 3E1, Canada. \textsuperscript{4}Samuel Lunenfeld Research Institute, Mt. Sinai Hospital, Toronto, Ontario M5G 1X5, Canada. \textsuperscript{5}Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. \textsuperscript{6}Department of Biomedical Engineering, Boston University, Boston, Massachusetts 02215, USA. \textsuperscript{7}Boston University School of Medicine, Boston, Massachusetts 02118, USA. \textsuperscript{8}Department of Medical Protein Research, VIB, 9000 Ghent, Belgium. \textsuperscript{9}Cancer Research Center (Centro de Investigación del Cancer), University of Salamanca and Consejo Superior de Investigaciones Científicas, Salamanca 37008, Spain. \textsuperscript{10}Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine, Barcelona 08028, Spain. \textsuperscript{11}Center for Complex Networks Research (CCNR) and Department of Physics, Northeastern University, Boston, Massachusetts 02115, USA. \textsuperscript{12}Department of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA. \textsuperscript{13}Department of Bioengineering, McGill University, Montreal, Quebec H3A 0C3, Canada. \textsuperscript{14}Protein Signaling and Interactions Lab, GIGA-R, Université de Liège, 4000 Liège, Belgium. \textsuperscript{15}Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts 02115, USA. \textsuperscript{16}Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain.

* These authors contributed equally to this work and should be considered co-first authors.

\^ These authors contributed equally to this work and should be considered co-second authors.
Complex cellular systems and interactome networks underlie most genotype-phenotype relationships. Just as a high-quality human reference genome sequence revolutionized genotype determination a decade ago, proteome-scale reference maps of protein-protein, protein-nucleic acid, and enzyme-substrate interactions should provide scaffold information necessary to interpret the increasing amount of genotypic data. Here we describe a new empirically-controlled “second-generation” map of the human binary protein-protein interaction network, expanding coverage over similar-quality interactions from the literature by three-fold, altogether bringing us only one order of magnitude away from the complete network. The map uncovers vast uncharted interactome “territories” rich in genetic material related to complex human phenotypes. Interaction “detectedness” correlates with co-functionality, suggesting that the biophysical interactome is biologically heterogeneous. Increased and uniform coverage of the interactome network helps characterize genotype-phenotype associations, particularly for cancer. Thus a reference map of the human interactome network is now within reach.
Complex cellular systems formed by large numbers of interactions amongst genes and gene products, or interactome networks, underlie most cellular functions. Understanding how perturbations of interactome networks relate to inherited and somatic disease susceptibilities will be required for predictive models of genotype-phenotype relationships in human.

A nearly full description of all genotypic variations in the human population is expected to be available soon. Several thousand genes associated with Mendelian disorders have been defined\(^1\), susceptibility loci have been uncovered for a few hundred complex traits\(^2\), and several hundred putative cancer genes have been identified\(^3-5\). With rapidly evolving next-generation sequencing platforms, large numbers of human genome sequences will soon be available from unaffected individuals\(^6\), patients stricken by virtually any disease\(^7\), and diverse tumours\(^8\).

As genome-wide genotypic information accumulates, many fundamental questions pertaining to genotype-phenotype relationships remain unresolved\(^9\). The causal changes and mechanisms that connect genotype to phenotype are still generally unknown, especially for complex trait loci and cancer-associated mutations. Even when identified, it is often unclear how a causal change perturbs the function of the corresponding gene product. Lastly, how such changes perturb interactome networks is largely unknown, which has hindered comprehensive understanding of gene-gene interactions. All in all, to “connect the dots” of the genomic revolution, functions and context must be assigned to hundreds of thousands of genotypic changes.

Globally functionalizing and contextualizing genotype-phenotype relationships requires genome and proteome-scale maps of interactome networks formed by macromolecular interactions such as protein-protein interactions (PPIs), protein-DNA interactions, and post-translational modifiers and their targets\(^9,10\). Although first generation interactome maps have
proposed network-based explanations for some genotype-phenotype relationships, it is difficult to establish how fully these maps represent the complete network and whether their quality is sufficient to derive accurate interpretations (Fig. 1a). There is a dire need for empirically controlled and high-quality “second-generation” reference maps, reminiscent of the high-quality human reference genome sequence that revolutionized genotyping.

The challenges are manifold. Even considering only one splice variant per gene, close to 20,000 protein-coding genes must be handled and ~200 million protein pairs tested to generate a comprehensive binary protein-protein interactome reference map. Accordingly, whether such a comprehensive network could ever be mapped by the collective efforts of small-scale studies remains uncertain. Computational predictions of protein interactions can generate information at proteome scale but are inherently limited by biases in currently available knowledge used to infer such interactome models. Should interactome maps be generated for all individual human tissues using biochemical co-complex association data, or alternatively, might direct binary interaction information on all possible PPIs be preferable or complementary? Finally, even with nearly complete, high-quality reference interactome maps of biophysical interactions in hand, how to evaluate the biological relevance of each interaction in physiological conditions? Here, we begin to address these questions by generating a second-generation map of the human binary interactome and comparing it to alternative network maps.
An uncharted interactome territory in small-scale interaction studies

To investigate whether small-scale studies described in the literature are adequate to qualitatively and comprehensively map the human binary protein-protein interactome network, we assembled data available in 2010 from seven public databases\textsuperscript{15-21} (Lit-2010). After eliminating protein pairs lacking any evidence of direct binary interaction and filtering out evidence from proteome-scale interactome screens\textsuperscript{11,22,23}, ~14,000 putative direct binary PPIs remained (Fig. 1b). Pairs reported in only a single publication and detected by only a single method have higher rates of curation errors than interactions reported in multiple papers and/or with multiple detection methods\textsuperscript{24}, yet represented nearly two thirds of the binary literature set.

We therefore segregated binary interactions with a single piece of evidence (Lit-BS-10) from binary interactions with multiple pieces of evidence (Lit-BM-10), and experimentally tested subsets of both by mammalian PPI trap (MAPPIT)\textsuperscript{25} and yeast two-hybrid (Y2H)\textsuperscript{26} assays. Lit-BS-10 pairs validated at rates that were only slightly higher than the randomly selected protein pairs used as negative control (random reference set; RRS) and considerably lower than Lit-BM-10 pairs (Fig. 1c and Supplementary Notes). Protein pairs from Lit-BS-10 were also found to co-occur in the literature significantly less than Lit-BM-10 pairs as indicated by STRING literature mining scores\textsuperscript{27} (Fig. 1c and Supplementary Notes), suggesting that these pairs were less thoroughly studied even in their corresponding publications. Thus, we restricted Lit-2010 to 4,906 high-quality Lit-BM-10 binary interactions (Supplementary Table 1). This highlights the need for systematic approaches, since estimates of the number of PPIs in the human interactome are close to two orders of magnitude larger than Lit-BM-10\textsuperscript{11,28}. 
The limited number of high-quality PPIs drastically diminishes the ability to prioritize putative disease genes identified in systematic studies. Although cancer driver gene products are well represented in Lit-BM-10, less than 30% of proteins encoded by genes identified in systematic genome-wide cancer screens have any reported interactions in Lit-BM-10 (Fig. 1d). Coverage was similarly low for proteins associated with Mendelian traits or identified in genome-wide association studies (GWAS). Such differences might be explained by the inspection bias inherent to small-scale studies. The ~20,000 proteins of the human proteome are not studied with equal intensity; some are described in hundreds of publications while most have been mentioned only in a few (Fig. 1e). Ranking proteins by the number of publications that mention them, we observed a strong bias of interactions towards highly studied proteins in contrast to a sparsely populated territory for interactions involving less studied proteins. Candidate gene products identified in genome-wide disease screens distribute homogeneously across the publication-ranked interactome space, again highlighting the need for systematic PPI maps to cover this uncharted interactome territory.

A second-generation binary interactome map

The enrichment of interactions detected for highly studied proteins in Lit-BM-10 might indicate that well studied proteins are truly involved in more PPIs. Under this model, more than half of all predicted human proteins would not participate substantially in the interactome network. Alternatively, the sparse territory could be homogenously populated by real PPIs that have been missed so far due to sociological or experimental biases.

To distinguish between these possibilities and to increase the coverage of the interactome space, we generated a new systematic binary interaction map, interrogating a search space four times larger and with deeper sensitivity than in our first proteome-scale attempt (Fig. 2a, Supplementary Fig. 1 and Supplementary Notes). All pairwise
combinations of proteins corresponding to ~13,000 genes (Space II; Supplementary Table 2) were screened twice, and we identified pairs of genes encoding interacting proteins from ~85,000 Y2H positive colonies using next-generation sequencing (Fig. 2b and Supplementary Notes). After quadruplicate retesting, ~14,000 distinct protein pairs were considered verified interactions.

To validate the quality of these PPIs, we compared a sample of ~800 verified interactions both to a positive reference set of ~500 binary interactions from Lit-BM-10 and to our RRS as a negative control. All ~2,000 protein pairs were tested in three binary assays: MAPPIT, a well-based nucleic acid programmable protein array (wNAPPA), and a fluorescence-reconstituting protein-fragment complementation assay (PCA). The verified pairs exhibited validation rates that were statistically indistinguishable from the positive reference set (Fig. 2c and Supplementary Tables 3 and 4), demonstrating the high quality of our PPIs. Our new human interactome dataset covering Space II and first released online in 2011 (“HI-II-11”; Supplementary Table 5) contains 13,944 interactions amongst 4,303 distinct proteins (517 homodimers and 13,427 heterodimers).

To determine the extent to which HI-II-11 contains direct binary interactions, we used three-dimensional co-crystal structures from the Protein Databank (PDB) to distinguish direct contacts from indirect associations. Out of 191 HI-II-11 interactions with such structural information, 189 were in direct contact in the corresponding crystal structure, demonstrating the high precision of HI-II-11 in reporting direct PPIs (99% precision; \( P = 4 \times 10^{-5} \), one-sided binomial test; Supplementary Table 6 and Supplementary Notes).

To investigate whether particular classes of PPIs might be refractory to our assays, we evaluated how the number of residue-residue contacts in three-dimensional structures related to our ability to detect interactions. Among interacting proteins found in PDB complexes, the pairs reported in HI-II-11 tended to involve more residue-residue contacts than those that were
missed (medians of 62 and 44 respectively, \( P = 1.27 \times 10^{-10} \), one-sided Wilcoxon rank-sum test; Fig. 2d and Supplementary Table 7). This result suggests that Y2H sensitivity can be affected by the number of residue-residue contacts and thus presumably by interaction affinity.

PPIs are commonly mediated by protein domains such as those catalogued in the Protein family (Pfam) database\(^{35} \). Observing that HI-II-11 is enriched for proteins containing domains known to be involved in domain-domain interactions\(^{36} \) (Supplementary Fig. 2), we reasoned that the pairing of such domains in interacting proteins provides a strong hypothesis for the mechanism of that interaction. To identify domain pairs mediating HI-II-11 interactions, we looked for pairs of Pfam domains in interacting proteins that appeared more frequently than expected (Fig. 2e and Supplementary Table 8). HI-II-11 is significantly enriched for domain pairs supported by three-dimensional structures\(^{36} \) (\( P = 6 \times 10^{-11} \), Wilcoxon rank sum test), further demonstrating that HI-II-11 is comprised of direct binary interactions. For PPIs involving the 25 most enriched domain pairs, we could identify the structurally supported domain pairs among all possible ones with high accuracy (70% sensitivity and 80% specificity), as exemplified by the calmodulin-binding IQ and EF-hand domains (Fig. 2f).

To compare the growth rates of systematic and literature interactome maps, we updated the literature binary interaction dataset to include information available up to 2013 (Supplementary Fig. 3). The number of high-quality binary literature PPIs has grown at a steady rate to \(~10,000\) interactions, whereas the growth of systematic datasets is punctuated with a small number of large-scale releases (Fig. 2g). The union of previous systematic datasets\(^{11,22,23,30} \) with HI-II-11 provides close to \(~20,000\) binary interactions which, combined with the high-quality binary literature pairs available as of 2013 (“Lit-BM-13”; Supplementary Table 9) provides \(~30,000\) individual PPIs. Collectively, this brings us to only an order of
magnitude from the estimated complete interactome network\textsuperscript{11,28}.

To examine how the sparsely populated territory has evolved, we visualized adjacency matrices for literature-based and systematic maps at regular time points since 1994. Although the sparse territory of the literature map has been slowly populated with interactions, growth in this territory continues to lag behind that in the densely populated one (Fig. 2h and Supplementary Fig. 4). In contrast, current systematic binary maps appear to more uniformly cover the full interactome space. This finding suggests that heterogeneity in literature coverage across the space reflects uneven investigative efforts, and supports the existence of true biophysical interactions across the full human interactome space.

**Apparently sparse territories contain genuine biophysical interactions**

To more deeply explore the heterogeneous coverage of the human interactome, we compared HI-II-11 and Lit-BM-13 to a collection of \(~25,000\) predicted binary PPIs of high-confidence (“PrePPI-HC”)\textsuperscript{13} and a co-fractionation map of \(~14,000\) putative binary interactions (“Co-Frac”)\textsuperscript{37} (Fig. 3a). Although statistically significant \((P < 1 \times 10^{-60}, \text{Fisher’s exact test})\), the overlap of interactions between these four maps is small (Fig. 3b), and the number of interactors, or “degree”, of a protein in one map does not correlate with its degree in the other three maps (Supplementary Fig. 5). Each mapping strategy apparently probes different, perhaps complementary, territories of the underlying true interactome landscape.

To investigate whether these four maps show similar sparse territories as did Lit-BM-13 (Fig. 2h), we generated adjacency matrices for each map with respect to publication frequency (Fig. 3c). Similar to Lit-BM-13, the Co-Frac and PrePPI-HC maps exhibited a strong tendency to report interactions amongst well-studied proteins (Fig. 3c). A potential source of bias for Co-Frac towards well-studied genes arises from the literature-based computational methods used to identify the reported subset of direct PPIs (Supplementary
Similarly, the structural modelling approach used to generate PrePPI-HC relies on existing structures as templates for identifying new interactions and on additional genomic and functional data to enhance prediction specificity. In contrast, HI-II-11 covered the publication-ranked interactome space uniformly.

Because the sensitivity of some interaction-mapping approaches depends on the level of gene expression, we examined interactome maps for expression-related sparse territories. The Co-Frac map shows a strong bias of interactions towards proteins encoded by highly expressed genes (Fig. 3c and Supplementary Fig. 7). The literature map also shows such a bias, likely due to the restricted use of common cell lines in the underlying experiments. In contrast, both HI-II-11 and PrePPI-HC exhibit a uniform interaction density across the full spectrum of expression levels, likely explained by the more uniform expression of proteins tested in Y2H and by the independence of the homology-based predictions from levels of expression.

We next sought more broadly for sources of intrinsic bias that might influence the appearance of sparsely populated territories, such as biological or knowledge-based properties of proteins and their corresponding genes. For example, PrePPI-HC is virtually devoid of interactions between proteins lacking Pfam domains, which is consistent with conserved domains forming the basis of the prediction method used to produce PrePPI-HC (Fig. 3c). HI-II-11 appears to be depleted, but not devoid, of interactions amongst proteins containing predicted transmembrane helices (TMH) (Fig. 3c), consistent with postulated limitations of the Y2H assay\textsuperscript{38}. The Co-Frac network also appears to be depleted in interactions involving TMH-containing proteins, which may be a consequence of membrane-bound proteins being filtered out during biochemical separations\textsuperscript{37}.

In all we examined 21 protein or gene properties, roughly classified as expression-, sequence-, and knowledge-based (Fig. 3d and Supplementary Table 10). We formally
defined a “dense zone” and a “sparse zone” of interactions by determining the location where the imbalance of interaction density is the highest (Supplementary Table 11 and Supplementary Fig. 8). We considered dense and sparse zones only when this imbalance was significantly different and strongly deviating from a uniform distribution of interactions. Besides the dense and sparse zones already described, we uncovered other biases, such as HI-II-11 exhibiting an enrichment of interactions amongst proteins with extensive disordered regions, a bias suggested in earlier Y2H binary datasets\textsuperscript{39}. No property resulted in the existence of dense and sparse zones in all four maps, suggesting that the observed imbalances reflect heterogeneous testing and technical limitations, rather than the existence of territories that are truly devoid of interactions.

To confirm that HI-II-11 interactions found in the apparently sparse zones of the three other maps are of high biophysical quality, we assessed MAPPIT validation of HI-II-11 interactions across the interactome space ranked by the various protein properties examined. MAPPIT validation rates of dense and sparse zone pairs were consistent for nearly all properties (Fig. 3e and Supplementary Figure 9), indicating that HI-II-11 interactions throughout the full interactome space are of equivalent biophysical quality.

In summary, current interactome maps contain vast uncharted territories that limit knowledge of the full interactome landscape, whereas HI-II-11 provides high-quality biophysical interactions within these under-explored regions.

**Functional heterogeneity in interactome maps**

One intuitive model is that all biophysical interactions of an organism correspond to functional interactions, so that the biophysical interactome network strictly overlaps with a collapsed version of all functional networks, in all cells and at all stages of development. An alternative model suggests that the biophysical interactome overlaps more loosely with functional
networks, with not all biophysical interactions being biologically relevant\textsuperscript{11}.

To determine whether the four interactome maps contain functionally relevant PPIs, we compared them to functional networks where links between genes represent shared functional annotations such as Gene Ontology (GO) terms or shared phenotypic annotations inferred from mouse orthologs. All four interactome maps showed significant enrichments for the four types of functional annotation examined (Fig. 4a), confirming their overall biological relevance. Thus, our systematic biophysical map generated independently from pre-existing biological information can indeed reveal functional relationships. Segregating HI-II-11 interactions with respect to dense and sparse zones found in other maps, we observed consistent functional enrichments in both dense and sparse zones for all studied protein rankings (Fig. 4b), supporting that HI-II-11 interactions in sparsely populated territories of other maps are biologically relevant.

To investigate which properties of biophysical interactions might explain co-functionality, we categorized interactions according to their “detectedness”, \textit{i.e.} how frequently they have been detected. We partitioned the high-quality Lit-BM-13 interactions between those with two, three, and four or more pieces of evidence. We observed a clear trend between detectedness and MAPPIT recovery rate, where pairs with the highest detectedness were recovered approximately three times more often than those with the lowest detectedness (Fig. 4c). There was a significantly higher functional enrichment for interactions having four or more pieces of evidence than those having two or three pieces of evidence, suggesting a possible correlation between how amenable interactions are to detection and functional relevance. However, this trend may result from a stronger inspection bias for interactions of higher detectedness, artificially increasing their observed functional relevance.

To discriminate between these two possibilities, we partitioned our systematic binary
interaction datasets into sets of increasing detectedness: HI-II-11 interactions found in one of
the two initial screens, those found in both screens (Supplementary Table 5), and the subset
of our first systematic map (HI-I-05)\textsuperscript{22} detected in two different binary assays, MAPPIT and
Y2H (Supplementary Table 4). MAPPIT recovery rate and functional enrichment both
increased with increasing detectedness (Fig. 4c), suggesting a relationship between
detectedness and functional relevance in the human biophysical interactome (Fig. 4d).

Measurements of overlap between literature and systematic datasets have been used
to characterize the overall quality of the latter, with the presumption that small overlaps reflect
poor quality\textsuperscript{40,41}. An alternative explanation is that interactome networks are larger than
originally assumed, with small overlaps resulting when each map detects only a small fraction
of the whole\textsuperscript{42}. We reasoned that dense zones of the literature datasets could be surrogates
for the effective search space. Hence interacting protein pairs found in systematic screens
that reside in literature sparse zones should not be taken into account to measure overlaps.
Restricting the effective search space of both the systematic and literature maps, we
observed a significant increase in overlap for pairs of increasing detectedness (Fig. 4e and
Supplementary Notes). As detectedness increases, the fraction of interactions in the
systematic maps found in the literature maps rises from 2\% to 25\%, suggesting that literature
and systematic approaches converge.

We therefore re-investigated estimates of the human binary interactome size using
reference sets consisting of literature binary interactions with increasing detectedness. Our
estimates vary between \(~600,000 and \(~200,000 interactions, within the range of previous
estimates\textsuperscript{11,28} (Fig. 4f, Supplementary Table 12 and Supplementary Notes). The decrease of
size estimates with reference set detectedness suggests that the interactome is
heterogeneous with respect to detection and functional relevance. This result was robust
when controlling for effective search space and matching detectedness of systematic and
literature datasets. More importantly, these estimates suggest that current systematic and literature datasets are only about an order of magnitude away from the full interactome at every detectedness level.

**Systematic binary interactome networks and human disease**

Topological properties of nodes within interactome networks provide insights into disease\(^{10,43}\), though studies of disease based solely on literature maps might reach potentially biased conclusions because of inspection biases, as identified in the Lit-BM-13 map. The significantly higher connectivity in Lit-BM-13 of cancer-associated gene products from the Sanger Cancer Gene Census (“Cancer Census”)\(^3\) might be because they are more heavily studied. Here we confirm using an unbiased systematic map that cancer proteins do indeed have higher connectivity \((P = 1 \times 10^{-3}, \text{one-sided Wilcoxon test}; \text{Fig. 5a})\).

To investigate the role interactome maps may play in identifying cancer genes, we tested the extent to which products of putative cancer genes identified in GWAS publications can be connected to known cancer-related proteins (Supplementary Table 13). In HI-II-11 proteins encoded by genes in cancer-associated loci interacted with known cancer proteins significantly more frequently than expected by chance, demonstrating how our map can be used to prioritize genes within large loci (Fig. 5b). We found similar results when using a range of linkage disequilibrium thresholds to define loci (Supplementary Fig. 10), and when using the Lit-BM-13 or PrePPI-HC maps (Supplementary Fig. 11).

To further assess how well interactions detected without influence of inspection bias may serve to prioritize cancer-associated genes, we investigated alternative network-free genomic and functional genomic approaches for identifying genes likely to be involved in tumorigenesis. We looked at three candidate gene sets: (i) \(~200\) genes with heightened non-synonymous and deleterious somatic mutation rates in cancers (“Mut”), (ii) \(~350\) human
orthologs of mouse genes associated with tumorigenesis in Sleeping Beauty ("SB") transposon-based mutagenesis screens, and (iii) ~400 human proteins found to interact with proteins from four DNA tumour virus families ("Vir")\textsuperscript{44}. For the proteins present in HI-II-11, the overlap between Mut, SB, and Vir sets and a reference set of 128 Cancer Census genes enables us to evaluate the predictive ability of each approach (Fig. 5c). The frequency of interaction in HI-II-11 between a protein and members of the SB and Vir sets significantly increased predictive ability (SB: $P = 1 \times 10^{-31}$, Vir: $P = 1 \times 10^{-43}$, one-sided paired Wilcoxon rank-sum test; Fig. 5c), arguing for network-guided analyses of systematic cancer gene discovery efforts.

We directly measured for each protein its corrected frequency of interaction with Cancer Census proteins, directly testing the “guilt-by-association” approach for transfer of function and annotation through PPIs. This method provided vastly improved predictions (SB: $P = 1 \times 10^{-115}$, one-sided paired Wilcoxon rank-sum test; Fig. 5c), and validates the existence of modularity in interactome networks: that proteins sharing similar phenotypes or function tend to directly interact with each other\textsuperscript{45}. Instead of relying solely on Cancer Census annotation, we sought to enhance cancer-gene prioritization by integrating information both about (i) membership of a query protein in candidate lists (“guilt-by-profiling”) and (ii) that of its interacting partners (“guilt-by-association”). We constructed a forward-stepwise logistic regression model to combine these fundamentally node- and edge-centric data types, and found a marked improvement in ability to highly rank Cancer Census proteins ($P = 1 \times 10^{-164}$, one-sided paired Wilcoxon rank-sum test; Fig. 5c and Supplementary Table 14).

Our top-ranked prediction was the cyclin-dependent kinase 4 (CDK4), a gene already in the Cancer Census list. CDK4 is connected in HI-II-11 to its inhibitors CDKN2C and CDKN2D, and to its associated G1/S-specific cyclins CCND1 and CCND3 (Fig. 5d). CDKN2C, CCND1, and CCND3 all have been causally implicated in tumorigenesis, while
CDK4 is frequently mutated in cancerous tissue and serves as a viral target of the E7 protein of human papillomavirus and of the Large T protein of polyomavirus\textsuperscript{44}.

The second ranked and top novel prediction is the coat protein II-coated vesicle (COPII) component SEC24A, which associates with the COPII complex to mediate protein transport from the endoplasmic reticulum (ER) (Fig. 5d). Two neighbours of SEC24A in HI-II-11 are the Cancer Census members Ewing Sarcoma Breakpoint Region 1 (EWSR1) and TRK-Fused Gene (TFG). Both contain a similar regulatory N-terminal region that is frequently present in oncogenic fusion proteins found in several types of sarcoma. Inhibition of TFG prevents protein secretion from the ER due to ER structural failure\textsuperscript{46} and depletion of the \textit{C. elegans} ortholog TFG-1 leads to a decline in COPII levels at ER exit sites\textsuperscript{47}, together arguing that TFG shares a function with SEC24A. MAPK1IP1L, a poorly studied protein found in literature-defined sparse zones of the three other interactome maps, interacts not only with the sarcoma-associated EWSR1 and TFG, but also with SS18L1, another sarcoma-associated Cancer Census gene. Thus, HI-II-11 offers compelling disease-related hypotheses in sparsely populated territories of the interactome.

\textsuperscript{15} Tasan et al.
Discussion

While PPI networks can shed light on complex genotype-phenotype relationships, help to identify relevant disease genes and unravel pathobiology mechanisms, the quality of current maps and their ability to thoroughly cover the human interactome landscape has to be carefully considered. Systematically screening half of the interactome space, independently from any inspection bias, we tripled the number of high-quality binary PPIs available from the literature. Covering regions of the human interactome landscape that are hardly touched by other approaches, the systematic binary map provides functional context to thousands of proteins, as demonstrated for candidates identified in unbiased cancer genetic screens. Systematic binary mapping therefore stands as a powerful tool providing necessary scaffold information to “connect the dots” of the genomic revolution.

How far away is a complete binary interactome map? Combining high-quality binary pairs from the literature and systematic binary maps, 30,000 high-confidence interactions are now available, only one order of magnitude away from what is to be found. A more pertinent question might be whether all interactions are equally functional? In both systematic and literature maps, functional enrichment increases with detectedness, suggesting that detectedness can provide a means to better quantify how functional are individual interactions, to identify functional modules\(^{45}\), and to understand heterogeneity in interactome networks. However, detectedness may in part reflect variation in dissociation constants and interactions with lower dissociation constants are more readily detected and hence better characterized. Development of new, cost-effective systematic experimental approaches will help to complete the mapping of the human interactome and also further assess the relationships between detectedness, affinity and biological functionality.

Reference binary interactome maps of increased coverage and quality are prerequisite to interpret condition-specific interactions\(^{48}\) and characterize the effects of splicing...
and genetic variation on interactions\textsuperscript{49}. Just as a reference genome enabled detailed maps of human genetic variation\textsuperscript{6}, completion of a reference interactome map will enable deeper insight into genotype-phenotype relationships in human.
METHODS SUMMARY

The systematic binary map was generated by screening all pairwise combinations of 15,517 ORFs from the hORFeome v5.1 collection. Interactome mapping was performed as described previously\(^2\) but with the following modifications: (i) the space was screened twice by Y2H to identify primary candidate interacting protein pairs, (ii) candidate interacting protein pairs were identified either by Sanger or by next generation sequencing\(^3\), (iii) candidate interacting protein pairs were verified in four independent retests and only those interactions that scored positive at least three times were retained, (iv) to confirm the quality of the full dataset, a subset of verified interactions were validated in three orthogonal assays MAPPIT, PCA and wNAPPA relative to a high-quality literature derived reference set and a random reference set of 460 and 698 pairs, respectively.

The Lit-2010 and Lit-2013 maps were generated by extracting all human PPIs annotated with tractable publication records from seven databases\(^15\)-\(^21\) through December 2010 and August 2013, respectively. PSI-MI experimental method codes were used to segregate the binary PPIs and the number of pieces of evidence were counted.

All statistical tests were performed using implementations of the R software\(^5\). Empirical \(P\) values were obtained using randomizations detailed in the respective methods. Full details are provided in Supplementary Methods.
FIGURE LEGENDS

Figure 1 | An uncharted interactome territory in small-scale interaction maps. a,
Understanding genotype-phenotype relationships through interactome maps of increasing completeness and quality. b, Flow chart (left) showing the assembly and segregation of protein-protein interactions from the literature. Different pieces of evidence correspond to different papers or methods. Histogram (right) of number of interactions for the full literature and different binary subsets. c, Fraction of Lit-BS-10 and Lit-BM-10 pairs recovered by MAPPIT at increasing RRS recovery rates (upper left). The shading indicates standard error of the proportion. Fraction of Lit-BS-10, Lit-BM-10 or RRS pairs recovered by MAPPIT at 1% RRS recovery rate (lower left), not found to co-occur in the literature as reported in the STRING database (upper right), and recovered by Y2H (lower right). P values are based on two-sided Wilcoxon rank sum tests. d, Fraction of full proteome and putative disease-associated gene products with protein-protein interactions in Lit-2010, Lit-BM-10, and systematic binary screens. Error bars indicate standard error of the proportion. Data sources are cited in Supplementary Information. e, Adjacency matrix showing Lit-BM-10 interactions, with proteins ordered symmetrically along vertical and horizontal axes by number of publications (upper histograms), and grouped in bins of ~350 proteins with median number of publications per bin shown. The colour intensity of each square reflects the total number of interactions between proteins in each bin. Total number of interactions per bin (lower histogram). Number of gene products from GWAS loci, Mendelian disease genes, and Sanger Cancer Gene Census genes per bin (circles).

Figure 2 | A second-generation binary interactome map. a, Improved empirical parameters from first-generation to second-generation interactome mapping. b, Experimental
pipeline for identifying high-quality binary protein-protein interactions. ORF: Open-reading frame, IST: Interaction Sequence Tag. c, Fraction of HI-II-11, Lit-BM-10, and RRS pairs recovered by MAPPIT (left), PCA (centre), and wNAPPA (right) at increasing assay scores. Shading indicates standard error of the proportion. d, Fraction of protein pairs in the PDB that HI-II-11 detected or missed as a function of number of residue-residue contacts. Vertical dashed lines indicate distribution medians. P value is based on a one-sided Wilcoxon rank sum test. Schematic of residue-residue contact count (bottom). e, Schematic of domain pair enrichment scoring method (top). Normalized frequency of all (left) and the top 25 (right) domain pairs. 3did: database of three-dimensional interacting domains. f, Network showing all possible domain combinations between HI-II-11 protein pairs involving the top 25 enriched domain pairs (left). Example of enriched domain pair corresponding to a known direct interaction (right). g, Total number of binary interactions in binary literature and systematic binary interactome maps over the past 20 years. h, Adjacency matrices showing binary literature and systematic binary maps at selected time points, with proteins ordered by number of publications. Colour intensity reflects the total number of interactions between proteins in each bin of ~350 proteins.

Figure 3 | Apparent sparse zones contain genuine biophysical interactions. a, Network representations of HI-II-11, Lit-BM-13, Co-Frac, and PrePPI-HC interactome maps. b, Topological properties of the four interactome maps. Venn diagrams illustrating the number of shared proteins (upper) and interactions (lower) in the four maps. c, Adjacency matrices for HI-II-11, Lit-BM-13, Co-Frac, and PrePPI-HC maps, with proteins ordered by number of publications, mRNA abundance in HEK cells, fraction of protein sequence covered by Pfam domains, and fraction of protein sequence covered by transmembrane helices. Median values of the protein properties for each bin of ~350 proteins are shown (top histograms).
Colour intensity reflects the total number of interactions between proteins in each bin. Lower histograms show total number of interactions per bin. d, Interaction density imbalances for 21 protein properties in the four maps. e, MAPPII recovery rate of HI-II-11 pairs found in dense and sparse zones of Lit-BM-13, Co-Frac, and PrePPI-HC at 1% RRS recovery. Error bars indicate standard error of the proportion.

**Figure 4 | Functional heterogeneity in interactome maps.** a, Functional enrichment of PPIs in the four interactome maps. Error bars indicate 95% confidence intervals. BP: Biological process, MF: Molecular function, CC: Cellular component. b, Functional enrichment of HI-II-11 interactions found in dense and sparse zones of Lit-BM-13, Co-Frac, and PrePPI-HC. Error bars indicate 95% confidence intervals. c, MAPPII recovery rate of subsets of systematic binary and binary literature interactions of increasing detectedness at 1% RRS rate (left histograms; literature: Lit-BM-13 with 2, 3, and 4 or more pieces of evidence; systematic: HI-II-11 found in one screen, two screens, and HI-I-05 found in two assays). Error bars indicate standard error of the proportion. Functional enrichment of systematic binary and binary literature interactions of increasing detectedness (right histograms). Error bars indicate 95% confidence intervals. *P* values are based on two-sided Fisher's exact tests (* means *P* ≤ 0.05). d, Model for how interaction detectedness may correlate with functionality. e, Fraction of shared interactions between the binary literature and systematic binary maps of increasing detectedness, controlling or not for their effective search spaces, *i.e.* space I and II for the systematic maps and the dense zones for the literature maps. f, Binary human interactome size estimated using literature reference sets of increasing detectedness (top left), or using reference sets and systematic binary maps of matched detectedness and controlling for their effective search spaces (top right). Total
number of high-quality interactions at increasing detectedness in the union of binary literature and systematic binary maps (bottom).

Figure 5 | Systematic binary interactome networks and human disease. a, Average degree of Sanger Cancer Gene Census (Cancer Census) proteins and proteins not associated to any disease in all four maps. \( P \) values are based on Wilcoxon rank sum test. b, Schematic representing interactions between proteins encoded by genes in cancer-associated GWAS loci and known Cancer Census proteins (top left). Fraction of cancer-related GWAS loci containing at least one gene encoding a protein that interacts with a Cancer Census protein in 1,000 randomly selected loci genes (histogram) and the observed value (arrow) for HI-II-11 (top right). Empirical \( P \) value is shown. Network representing proteins encoded by genes in cancer-associated GWAS loci interacting with Cancer Census proteins in HI-II-11, and an example network after randomizing genes (bottom). c, Schematic of the cancer association scoring system (top). Predictive ability of separate and combined models for cancer association scoring (box plot). \( P \) values are based on one-sided paired Wilcoxon rank sum test. d, The 20 proteins with the highest cancer-association scores according to the combined model. For four cancer candidates (bold) their characteristics and their interacting partners in HI-II-11 are indicated. * Cancer Census proteins.
REFERENCES


Supplementary information is available in the online version of the paper.
Acknowledgements: The authors wish to acknowledge members of the Center for Cancer Systems Biology (CCSB) and particularly H. Yu for helpful discussions. This work was supported by the following grants and agencies: NHGRI grants R01HG001715 to M.V., D.E.H., F.P.R. and J.T. and P50HG004233 to M.V., F.P.R. and A.-L.B.; NHLBI grants U01HL098166 subaward to M.V. and U01HL108630 subaward to A.-L.B.; NCI grant U54CA112962 subaward to M.V.; NICHD ARRA R01HD065288 to L.M.I.; NIMH R01MH091350 to L.M.I. and T.H.; NSF Grant CCF-1219007 to Y.X.; Krembil Foundation, Canada Excellence Research Chair, Ontario Research Fund–Research Excellence Award, Canadian Institute for Advanced Research Fellowship to F.P.R.; grant CSI07A09 from Junta de Castilla y Leon (Valladolid, Spain), grant PI12/00624 from Ministerio de Economía y Competitividad (AES 2012, ISCiii, Madrid, Spain,) and grant i-Link0398 from Consejo Superior de Investigaciones Científicas (CSIC, Madrid, Spain) to J.D.L.R.; Spanish Ministerio de Ciencia e Innovación (BIO2010-22073) and the European Commission through the FP7 project SyStemAge grant agreement n:306240 to P.A.; Group-ID Multidisciplinary Research Partnerships of Ghent University and grant FWO-V G.0864.10 from the Fund for Scientific Research Flanders to J.T.; I.L. is a postdoctoral fellow with the FWO-V; Institute Sponsored Research funds from the Dana-Farber Cancer Institute Strategic Initiative to M.V. M.V. is a "Chercheur Qualifié Honoraire" from the Fonds de la Recherche Scientifique (FRS-FNRS, Wallonia-Brussels Federation, Belgium).


Author Information The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.A.C. (michael_calderwood@dfci.harvard.edu). D.E.H. (david_hill@dfci.harvard.edu), T.H. (tong_hao@dfci.harvard.edu), F.P.R. (fritz.roth@utoronto.ca) and M.V. (marc_vidal@dfci.harvard.edu).
Genotype-phenotype relationships

Gene or gene product

Biophysical interaction

Functional interaction relevant for a phenotype of interest

Underlying interactome

Completeness and quality of maps

Mendelian traits

Complex traits

View from available maps

- Gene or gene product
- Biophysical interaction
- Functional interaction relevant for a phenotype of interest
- Gene or gene product with no characterized interactions
- Uncharacterized interaction
1b

7 public databases (Lit-2010) (PPIs with valid evidence)

62,163

- Non-systematic
  - 56,743
    - Non-binary
      - 42,781
    - Binary
      - 13,962

- Systematic
  - 5,871
    - Binary
      - 5,871

- Pieces of evidence
  - 1
    - 9,056
  - 2
    - 2,646
  - 3
    - 1,013
  - 4+
    - 1,247

Lit-BS-10  Lit-BM-10

Number of interactions (thousands)

- Lit-2010
- Lit-BS-10
- Lit-BM-10
- Systematic
1c

**MAPPIT assay**

Fraction of RRS pairs recovered (%)

**STRING literature mining**

Fraction of pairs not found to co-occur (%)

**Y2H assay**

Binary literature pairs:
- Lit-BS-10
- Lit-BM-10

P = 5 × 10⁻¹⁵

P = 8 × 10⁻³
Whole proteome (n = 19,111)

Well-accepted cancer driver (n = 124)

Sanger Cancer Gene Census (n = 465)

Mendelian disorder gene (n = 2,826)

Systematic cancer screen gene (n = 2,996)

GWAS loci gene for any trait (n = 2,883)

Lit-2010

Lit-BM-10

Systematic binary screens
### Parameters

<table>
<thead>
<tr>
<th>Completeness</th>
<th>Sampling sensitivity</th>
<th>Assay sensitivity</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>7K 19K</td>
<td>1</td>
<td>PRS RRS</td>
<td>92 PRS 92 RRS 1 assay</td>
</tr>
<tr>
<td>13K 19K</td>
<td></td>
<td>PRS RRS</td>
<td>~500 PRS ~700 RRS 3 assays</td>
</tr>
</tbody>
</table>

### Methods

<table>
<thead>
<tr>
<th>Interactions</th>
<th>Interactions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1 2</td>
</tr>
</tbody>
</table>

### Relative gain

- 3.1X space
- 1.6X sensitivity
- 1.3X sensitivity
- > 5X reference sets

### Assay Completions

- 2005: 92 PRS 92 RRS 1 assay
- 2011: ~500 PRS ~700 RRS 3 assays
Tasan et al.

2b

Screen (2X)
~13,000 proteins
hORFeome 5.1
~82,000,000
protein pairs tested
~35,000 IST pairs
identified

Verification
~66,000 ORF pairs
4x pairwise retested
Sequence verification
of 3x and 4x positives
~15,000 gene pairs
verified

Validation
~800 protein pairs
selected at random
Tested in 3 binary
orthogonal assays
Recovery compared
to Lit/RRS pairs

HI-II-11
~14,000 interactions
between ~4,300 proteins
Tasan et al.

2c

![Graphs showing MAPPIT, PCA, and wNAPPA](image)

Fraction of pairs recovered (%) vs. Score threshold (arbitrary units)

- **MAPPIT**: HI-II-11, Lit-BM-10, RRS
- **PCA**: HI-II-11, Lit-BM-10, RRS
- **wNAPPA**: HI-II-11, Lit-BM-10, RRS
Figure 2d

- Fraction of protein pairs (%)
  - $P = 1.3 \times 10^{-10}$
  - Detected (189)
  - Missed (498)

Tasan et al.
Domain pairs in HI-II-11 interacting proteins

1,000 randomized HI-II-11 networks

Normalized frequency

Three-dimensional structures of domain-based interactions

Domain pairs (n = 1,213)

60% precision

Normalized frequency

Normalized frequency

Domain pairs:

- In 3did
- Not in 3did

Tasan et al.
Possible domain pairs for HI-II-11 interactions:
- Predicted and in 3did (TP = 182)
- Predicted and not in 3did (FP = 81)
- Not predicted and in 3did (FN = 91)
- Not predicted and not in 3did (TN = 318)
2g

- Binary literature (multiple pieces of evidence)
- Systematic binary screens

Number of binary interactions (thousands)

Time

Tasan et al.

2h

Space ordered by number of publications


Number of interactions

0 10 ≥100

Systematic binary maps
Binary literature maps
### 3b

<table>
<thead>
<tr>
<th></th>
<th>HI-II-11</th>
<th>Lit-BM-13</th>
<th>Co-Frac</th>
<th>PrePPI-HC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>4,303</td>
<td>5,545</td>
<td>2,999</td>
<td>4,989</td>
</tr>
<tr>
<td>Edges</td>
<td>13,944</td>
<td>11,045</td>
<td>13,982</td>
<td>25,403</td>
</tr>
<tr>
<td>Nodes in largest component</td>
<td>4,100</td>
<td>4,919</td>
<td>2,820</td>
<td>3,681</td>
</tr>
<tr>
<td>Mean degree</td>
<td>6.348</td>
<td>3.772</td>
<td>9.324</td>
<td>10.183</td>
</tr>
<tr>
<td>Mean clustering</td>
<td>0.053</td>
<td>0.065</td>
<td>0.164</td>
<td>0.399</td>
</tr>
<tr>
<td>Mean distance</td>
<td>4.063</td>
<td>5.305</td>
<td>4.506</td>
<td>5.251</td>
</tr>
<tr>
<td>Power law exponent</td>
<td>2.195</td>
<td>1.659</td>
<td>1.462</td>
<td>1.451</td>
</tr>
</tbody>
</table>

#### Proteins (heterodimers)

- HI-II-11: 1,491, 701, 1,766, 269, 94, 336
- Co-Frac: 893, 253, 1,870
- PrePPI-HC: 253, 1,214, 1,870

#### Protein-protein interactions (heterodimers)

- HI-II-11: 9,128, 255, 13,045, 27, 14, 10, 27, 40
- Lit-BM-13: 13,347, 195, 9, 27, 435
- Co-Frac: 287, 287, 287
- PrePPI-HC: 435, 287, 287
3c

<table>
<thead>
<tr>
<th>Number of interactions</th>
<th>mRNA abundance in HEK cells</th>
<th>Fraction of sequence in Pfam domains</th>
<th>Fraction of sequence in transmembrane helices</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1,000</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>10</td>
<td>100</td>
<td>10</td>
<td>25</td>
</tr>
<tr>
<td>≥100</td>
<td>&lt;1,000</td>
<td>&lt;100</td>
<td>&lt;50</td>
</tr>
</tbody>
</table>

Tasan et al. 41
<table>
<thead>
<tr>
<th>mRNA abundance in HEK cells</th>
<th>mRNA abundance in HeLa cells</th>
<th>Protein abundance in HEK cells</th>
<th>Protein abundance in HeLa cells</th>
</tr>
</thead>
<tbody>
<tr>
<td>First publication date</td>
<td>Number of publications</td>
<td>Number of protein complexes</td>
<td>Number of pathways</td>
</tr>
<tr>
<td>Number of GO associations</td>
<td>Fraction of sequence in Pfam domains</td>
<td>Fraction of basic amino-acids</td>
<td>Fraction of charged amino-acids</td>
</tr>
<tr>
<td>Fraction of acidic amino-acids</td>
<td>Fraction of polar amino-acids</td>
<td>Fraction of polar and uncharged amino-acids</td>
<td>Fraction of hydrophobic amino-acids</td>
</tr>
<tr>
<td>Number of binding sites</td>
<td>Fraction of sequence in transmembrane helices</td>
<td>Number of linear motifs</td>
<td>ORF length</td>
</tr>
</tbody>
</table>

Abundance

Knowledge

Sequence

Enrichment
Balanced representation
Depletion
Tasan et al.

3e

Mirrored zones in HI-II-11

Dense zone

Sparse zone

Fraction of pairs recovered in MAPPIT (%)

Number of publications (mirrored from Lit-BM-13)

mRNA abundance in HEK cells (mirrored from Co-Frac)

Fraction of sequence in Pfam domains (mirrored from PrePPI-HC)
Tasan et al.

4b

Mirrored zones in HI-II-11

Dense zone

Sparse zone

Number of publications (mirrored from Lit-BM-13)

mRNA abundance in HEK cells (mirrored from Co-Frac)

Fraction of sequence in Pfam domains (mirrored from PrePPI-HC)

1,000

100

10

1

1,000

100

10

1

1,000

100

10

1

1,000

100

10

1

1,000

100

10

1

1,000

100

10

1
“True” biophysical interaction network

Pseudo-interactions

Functional interactions

Detectedness

Fraction of interactions

Tasan et al.
Detectedness

Lit-BM-13 map
(pieces of evidence)

Systematic map
HI-II-11 in 1 screen
HI-II-11 in 2 screens
HI-I-05 in 2 assays

Effective search space:  Not controlled  Controlled

Tasan et al.
Effective search space

- Lit-BM-13 map (pieces of evidence)
- Systematic map

Detectedness

- Binary interactome size (thousands)

Controlled for systematic map:

- ≥2
- ≥3
- ≥4

Controlled for both maps:

- ≥1
- 2
- 2

<table>
<thead>
<tr>
<th>HI-II-11</th>
<th>HI-I-05</th>
</tr>
</thead>
<tbody>
<tr>
<td>(screens)</td>
<td>(assays)</td>
</tr>
</tbody>
</table>

Controlled for systematic map

- HI-II-11

Controlled for both maps

- HI-II-11
- HI-I-05

Tasan et al.
5a

Proteins not in OMIM or Cancer Census

Knowledge based & Systematic

\[ P = 1 \times 10^{-7} \]

\[ P = 1 \times 10^{-31} \]

\[ P = 0.2 \]

\[ P = 0.01 \]

Cancer Census proteins
Proteins not in OMIM or Cancer Census
5b

Cancer-associated GWAS loci

Sanger Cancer Gene Census

- Black circles: Gene product in HI-II-11
- White circles: Interaction in HI-II-11

Graph showing the frequency in random networks of loci with cancer candidates.

- Frequency in random networks (y-axis)
- Fraction of loci with cancer candidates (x-axis)
- P = 0.02

Gene in GWAS locus whose product:
- Purple circle: Is present in HI-II-11
- White circle: Is absent from HI-II-11
- Red and white: GWAS locus
- Diamond: Protein in Sanger Cancer Gene Census
- Purple line: HI-II-11 interaction

Randomized genes

Gene product in HI-II-11

Interaction in HI-II-11

Protein in Sanger Cancer Gene Census

HI-II-11 interaction

Graph illustrates the relationship between cancer-associated GWAS loci and genes in the Sanger Cancer Gene Census.
Query node data only
“guilt-by-profiling”
Hi-II-11 neighbor data
“guilt-by-association”

Protein P

SB  Mut  Vir

Protein A
Protein B
Protein C

In the set
Not in the set

Correct for frequencies across random networks

Cancer association score

SB = Sleeping Beauty transposon-based mouse cancer screen
Mut = Somatic mutations screen in cancer tissues
Vir = Tumor virus host targets
Census = Sanger Cancer Gene Census

Average precision of predicted cancer proteins (%)

"Guilt-by-
profiling"

"Guilt-by-
association"

Combined model
Top 20 ranked proteins

<table>
<thead>
<tr>
<th>Protein</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDK4*</td>
<td>43</td>
</tr>
<tr>
<td>SEC24A</td>
<td>35</td>
</tr>
<tr>
<td>STAT3</td>
<td>33</td>
</tr>
<tr>
<td>GRB2</td>
<td>32</td>
</tr>
<tr>
<td>CDKN2C*</td>
<td>31</td>
</tr>
<tr>
<td>MAPK1IP1L</td>
<td>27</td>
</tr>
<tr>
<td>SH2D4A</td>
<td>25</td>
</tr>
<tr>
<td>HNRNPM</td>
<td>23</td>
</tr>
<tr>
<td>ATPAF2</td>
<td>22</td>
</tr>
<tr>
<td>DDX6*</td>
<td>21</td>
</tr>
<tr>
<td>NAGK</td>
<td>20</td>
</tr>
<tr>
<td>NFE2L2*</td>
<td>20</td>
</tr>
<tr>
<td>CDKAL1</td>
<td>20</td>
</tr>
<tr>
<td>AP2B1</td>
<td>19</td>
</tr>
<tr>
<td>ARMC8</td>
<td>17</td>
</tr>
<tr>
<td>PSME4</td>
<td>17</td>
</tr>
<tr>
<td>KRAS*</td>
<td>16</td>
</tr>
<tr>
<td>WAC</td>
<td>16</td>
</tr>
<tr>
<td>TAX1BP1</td>
<td>15</td>
</tr>
<tr>
<td>NEUROG3</td>
<td>15</td>
</tr>
</tbody>
</table>

In the set

Not in the set