Cell

Resource

A Proteome-Scale Map of the Human Interactome Network

Thomas Rolland,[1,2,19] Murat Taşan,[1,3,4,5,19] Benoit Charloteaux,[1,2,19] Samuel J. Pevzner,[1,2,6,7,19] Quan Zhong,[1,2,8,19] Nidhi Sahni,[1,2,19] Song Yi,[1,2,19] Irma Lemmens,[9] Celia Fontanillo,[10] Roberto Mosca,[11] Atanas Kamburov,[1,2] Susan D. Ghiassian,[1,12] Xinping Yang,[1,2] Lila Ghamsari,[1,2] Dawit Balcha,[1,2] Bridget E. Begg,[1,2] Pascal Braun,[1,2] Marc Brehme,[1,2] Martin P. Broly,[1,2] Anne-Ruxandra Carvunis,[1,2] Dan Convery-Zupan,[1,2] Roser Corominas,[13] Jasmin Coulombe-Huntington,[1,14] Elizabeth Dann,[1,2] Matija Dreze,[1,2] Amélie Dricot,[1,2] Changyu Fan,[1,2] Eric Franzosa,[1,14] Fana Gebreab,[1,2] Bryan J. Gutierrez,[1,2] Madeleine F. Hardy,[1,2] Mike Jin,[1,2] Shuli Kang,[13] Ruth Kiros,[1,2] Guan Ning Lin,[13] Katja Luck,[1,2] Andrew MacWilliams,[1,2] Jörg Menche,[1,12] Ryan R. Murray,[1,2] Alexandre Palagi,[1,2] Matthew M. Poulin,[1,2] Xavier Rambout,[1,2,15] John Rasla,[1,2] Patrick Reichert,[1,2] Viviana Romero,[1,2] Elien Ruyssinck,[9] Julie M. Sahalie,[1,2] Annemarie Scholz,[1,2] Akash A. Shah,[1,2] Amitabh Sharma,[1,12] Yun Shen,[1,2] Kerstin Spirohn,[1,2] Stanley Tam,[1,2] Alexander O. Tejeda,[1,2] Shelly A. Trigg,[1,2] Jean-Claude Twizere,[1,2,15] Kerwin Vega,[1,2] Jennifer Walsh,[1,2] Michael E. Cusick,[1,2] Yu Xia,[1,14] Albert-László Barabási,[1,12,16] Lilia M. Iakoucheva,[13] Patrick Aloy,[11,17] Javier De Las Rivas,[10] Jan Tavernier,[9] Michael A. Calderwood,[1,2,20] David E. Hill,[1,2,20] Tong Hao,[1,2,20] Frederick P. Roth,[1,3,4,5,18,*] and Marc Vidal[1,2,*]

[1]Center for Cancer Systems Biology (CCSB) and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[2]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA
[3]Departments of Molecular Genetics and Computer Science, University of Toronto, Toronto, ON M5S 3E1, Canada
[4]Donnelly Centre, University of Toronto, Toronto, ON M5S 3E1, Canada
[5]Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, ON M5G 1X5, Canada
[6]Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA
[7]Boston University School of Medicine, Boston, MA 02118, USA
[8]Department of Biological Sciences, Wright State University, Dayton, OH 45435, USA
[9]Department of Medical Protein Research, VIB, 9000 Ghent, Belgium
[10]Cancer Research Center (Centro de Investigación del Cancer), University of Salamanca and Consejo Superior de Investigaciones Científicas, Salamanca 37008, Spain
[11]Joint IRB-BSC Program in Computational Biology, Institute for Research in Biomedicine (IRB Barcelona), Barcelona 08028, Spain
[12]Center for Complex Network Research (CCNR) and Department of Physics, Northeastern University, Boston, MA 02115, USA
[13]Department of Psychiatry, University of California, San Diego, La Jolla, CA 92093, USA
[14]Department of Bioengineering, McGill University, Montreal, QC H3A 0C3, Canada
[15]Protein Signaling and Interactions Lab, GIGA-R, University of Liege, 4000 Liege, Belgium
[16]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[17]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona 08010, Spain
[18]Canadian Institute for Advanced Research, Toronto M5G 1Z8, Canada
[19]Co-first author
[20]Co-senior author
*Correspondence: fritz.roth@utoronto.ca (F.P.R.), marc_vidal@dfci.harvard.edu (M.V.)
http://dx.doi.org/10.1016/j.cell.2014.10.050

SUMMARY

Just as reference genome sequences revolutionized human genetics, reference maps of interactome networks will be critical to fully understand genotype-phenotype relationships. Here, we describe a systematic map of ~14,000 high-quality human binary protein-protein interactions. At equal quality, this map is ~30% larger than what is available from small-scale studies published in the literature in the last few decades. While currently available information is highly biased and only covers a relatively small portion of the proteome, our systematic map appears strikingly more homogeneous, revealing a "broader" human interactome network than currently appreciated. The map also uncovers significant interconnectivity between known and candidate cancer gene products, providing unbiased evidence for an expanded functional cancer landscape, while demonstrating how high-quality interactome models will help "connect the dots" of the genomic revolution.

INTRODUCTION

Since the release of a high-quality human genome sequence a decade ago (International Human Genome Sequencing Consortium, 2004), our ability to assign genotypes to phenotypes has exploded. Genes have been identified for most Mendelian disorders (Hamosh et al., 2005) and over 100,000 alleles have been implicated in at least one disorder (Stenson et al., 2014). Hundreds of susceptibility loci have been uncovered for numerous complex traits (Hindorff et al., 2009) and the genomes of a few thousand human tumors have been nearly fully sequenced (Chin et al., 2011). This genomic revolution is poised to generate a complete description of all relevant genotypic variations in the human population.

Genomic sequencing will, however, if performed in isolation, leave fundamental questions pertaining to genotype-phenotype relationships unresolved (Vidal et al., 2011). The causal changes that connect genotype to phenotype remain generally unknown, especially for complex trait loci and cancer-associated mutations. Even when identified, it is often unclear how a causal mutation perturbs the function of the corresponding gene or gene product. To "connect the dots" of the genomic revolution, functions and context must be assigned to large numbers of genotypic changes.

Complex cellular systems formed by interactions among genes and gene products, or interactome networks, appear to underlie most cellular functions (Vidal et al., 2011). Thus, a full understanding of genotype-phenotype relationships in human will require mechanistic descriptions of how interactome networks are perturbed as a result of inherited and somatic disease susceptibilities. This, in turn, will require high-quality and extensive genome and proteome-scale maps of macromolecular interactions such as protein-protein interactions (PPIs), protein-nucleic acid interactions, and posttranslational modifiers and their targets.

First-generation human binary PPI interactome maps (Rual et al., 2005; Stelzl et al., 2005) have already provided network-based explanations for some genotype-phenotype relationships, but they remain incomplete and of insufficient quality to derive accurate global interpretations (Figure S1A available online). There is a dire need for empirically-controlled (Venkatesan et al., 2009) high-quality proteome-scale interactome reference maps, reminiscent of the high-quality reference genome sequence that revolutionized human genetics.

The challenges are manifold. Even considering only one splice variant per gene, approximately 20,000 protein-coding genes (Kim et al., 2014; Wilhelm et al., 2014) must be handled and ~200 million protein pairs tested to generate a comprehensive binary reference PPI map. 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 (Zhang et al., 2012) 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 cocomplex association data, or would "context-free" information on direct binary biophysical interaction for all possible PPIs be preferable? To what extent would these approaches be complementary? Even with nearly complete, high-quality reference interactome maps of biophysical interactions, how can the biological relevance of each interaction be evaluated under physiological conditions? Here, we begin to address these questions by generating a proteome-scale map of the human binary interactome and comparing it to alternative network maps.

RESULTS

Vast Uncharted Interactome Zone in Literature

To investigate whether small-scale studies described in the literature are adequate to qualitatively and comprehensively map the human binary PPI network, we assembled all binary pairs identified in such studies and available as of 2013 from seven public databases (Figure S1B, see Extended Experimental Procedures, Section 1). Out of the 33,000 *lit*erature *b*inary pairs extracted, two thirds were reported in only a *s*ingle publication and detected by only a single method (Lit-BS pairs), thus potentially presenting higher rates of curation errors than binary pairs supported by *m*ultiple pieces of evidence (Lit-BM pairs; Tables S1A, S1B, and S1C) (Cusick et al., 2009). Testing representative samples from both of these sets using the mammalian protein-protein interaction trap (MAPPIT) (Eyckerman et al., 2001) and yeast two-hybrid (Y2H) (Dreze et al., 2010) assays, we observed that Lit-BS pairs were recovered at rates that were only slightly higher than the randomly selected protein pairs used as negative control (random reference set; RRS) and significantly lower than Lit-BM pairs (Figure 1A and Table S2A; see Extended Experimental Procedures, Section 2). Lit-BS pairs co-occurred in the literature significantly less often than Lit-BM pairs as indicated by STRING literature mining scores (Figure 1A and Figure S1C; see Extended Experimental Procedures, Section 2) (von Mering et al., 2003), suggesting that these pairs were less thoroughly studied. Therefore, use of binary PPI information from public databases should be restricted to interactions with multiple pieces of evidence in the literature. In 2013, this corresponded to 11,045 high-quality protein pairs (Lit-BM-13), more than an order of magnitude below current estimates of the number of PPIs in the full human interactome (Stumpf et al., 2008; Venkatesan et al., 2009).

The relatively low number of high-quality binary literature PPIs may reflect inspection biases inherent to small-scale studies. Some genes such as *RB1* are described in hundreds of publications while most have been mentioned only in a few (e.g., the unannotated *C11orf21* gene). To investigate the effect of such biases on the current coverage of the human interactome network, we organized the interactome search space by ranking proteins according to the number of publications in which they are mentioned (Figure 1B). Interactions between highly studied proteins formed a striking "dense zone" in contrast to a large sparsely populated zone, or "sparse zone," involving poorly studied proteins. Candidate gene products identified in genome-wide association studies (GWAS) or associated with Mendelian disorders distribute homogeneously across the publication-ranked interactome space (Figure 1B and Figure S1D), demonstrating a need for unbiased systematic PPI mapping to cover this uncharted territory.

A Proteome-wide Binary Interactome Map

Based on literature-curated information, the human interactome appears to be restricted to a narrow dense zone, suggesting that half of the human proteome participates only rarely in the interactome network. Alternatively, the zone that appears sparse in the literature could actually be homogeneously populated by PPIs that have been overlooked due to sociological or experimental biases.

To distinguish between these possibilities and address other fundamental questions outlined above, we generated a new proteome-scale binary interaction map. By acting on all four parameters of our empirically-controlled framework (Venkatesan et al., 2009), we increased the coverage of the human binary interactome with respect to our previous *h*uman *i*nteractome data set obtained by investigating a search space defined by ~7,000 protein-coding genes ("Space *I*") and published in 2005 (HI-I-05)

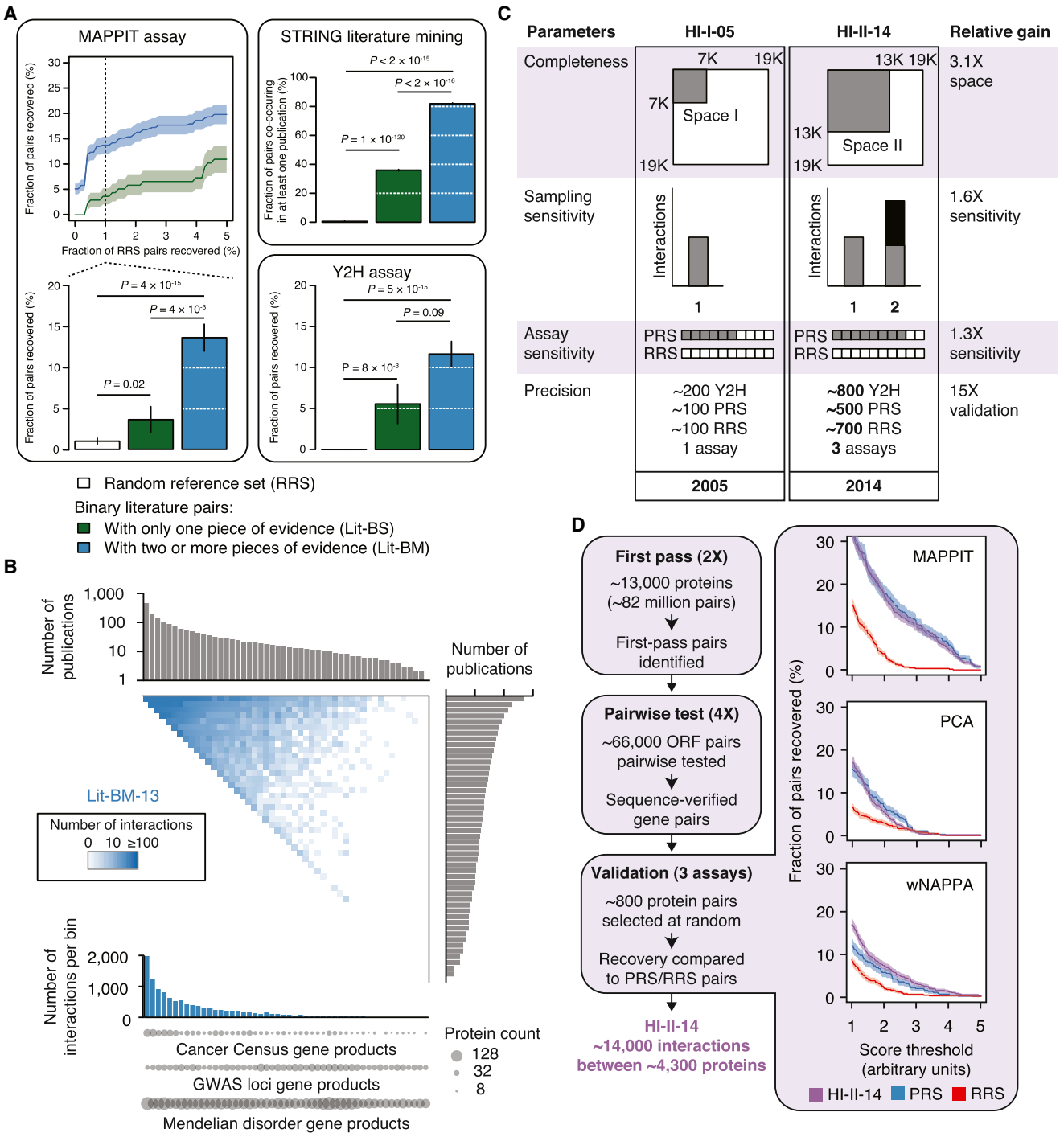


Figure 1. Vast Uncharted Interactome Zone in Literature and Generation of a Systematic Binary Data Set

(A) Validation of binary literature pairs extracted from public databases (Bader et al., 2003; Berman et al., 2000; Chatr-Aryamontri et al., 2013; Kerrien et al., 2012; Licata et al., 2012; Keshava Prasad et al., 2009; Salwinski et al., 2004). Fraction of pairs recovered by MAPPIT at increasing RRS recovery rates (top left) and at 1% RRS recovery rate (bottom left), found to co-occur in the literature as reported in the STRING database (upper right), and recovered by Y2H (lower right). Shading and error bars indicate standard error of the proportion. p values, two-sided Fisher's exact tests. For n values, see Table S6.

(B) Adjacency matrix showing Lit-BM-13 interactions, with proteins in bins of ~350 and ordered by number of publications along both axes. Upper and right histograms show the median number of publications per bin. The color intensity of each square reflects the total number of interactions between proteins for the corresponding bins. Total number of interactions per bin (lower histogram). Number of products from GWAS loci (Hindorff et al., 2009), Mendelian disease (Hamosh et al., 2005), and Sanger Cancer Gene Census (Cancer Census) (Futreal et al., 2004) genes per bin (circles).

(legend continued on next page)

(Rual et al., 2005) (Figures 1C and 1D; see Extended Experimental Procedures, Section 3). A search space consisting of all pairwise combinations of proteins encoded by ~13,000 genes ("Space II"; Table S2B) was systematically probed, representing a 3.1-fold increase with respect to the HI-I-05 search space. To gain in sensitivity, we performed the Y2H assay in different strain backgrounds that showed increased detection of pairs of a positive reference set (PRS) composed of high-quality pairs from the literature without increasing the detection rate of RRS pairs. To increase our sampling, the entire search space was screened twice independently. Pairs identified in this first pass were subsequently tested pairwise in quadruplicate starting from fresh yeast colonies. To ensure reproducibility, only pairs testing positive at least three times out of the four attempts and with confirmed identity were considered interacting pairs, resulting in ~14,000 distinct interacting protein pairs.

We validated these binary interactions using three binary protein interaction assays that rely on different sets of conditions than the Y2H assay: (1) reconstituting a membrane-bound receptor complex in mammalian cells using MAPPIT, (2) in vitro using the well-based nucleic acid programmable protein array (wNAPPA) assay (Braun et al., 2009; Ramachandran et al., 2008), and (3) reconstituting a fluorescent protein in Chinese hamster ovary cells using a protein-fragment complementation assay (PCA) (Nyfeler et al., 2005) (see Extended Experimental Procedures, Section 4). The Y2H pairs exhibited validation rates that were statistically indistinguishable from a PRS of ~500 Lit-BM interactions while significantly different from an RRS of ~700 pairs with all three orthogonal assays and over a large range of score thresholds (Figure 1D, Tables S2A and S2C), demonstrating the quality of the entire data set. Using three-dimensional cocrystal structures available for protein complexes in the Protein Data Bank (Berman et al., 2000) and for domain-domain interactions (Stein et al., 2011) (Figure S2 and Tables S2D, S2E, and S2F; see Extended Experimental Procedures, Sections 5 and 6), we also demonstrated that our binary interactions reflect direct biophysical contacts, a conclusion in stark contrast to a previous report suggesting that Y2H interactions are inconsistent with structural data (Edwards et al., 2002). Our results also suggested that Y2H sensitivity correlates with the number of residue-residue contacts and thus presumably with interaction affinity. The corresponding *h*uman *i*nteractome data set covering Space *II* and reported in 20*14* (HI-II-14; Table S2G) is the largest experimentally-determined binary interaction map yet reported, with 13,944 interactions among 4,303 distinct proteins.

Overall Biological Significance

To assess the overall functional relevance of HI-II-14, we combined computational analyses with a large-scale experimental approach. We first measured enrichment for shared Gene Ontology (GO) terms and phenotypic annotations and observed that HI-II-14 shows significant enrichments that are similar to those of Lit-BM-13 (Figures 2A and 2B; see Extended Experimental Procedures, Section 7). Second, we measured how much binary interactions from HI-II-14 reflect membership in larger protein complexes as annotated in CORUM (Ruepp et al., 2010) or reported in a cocomplex association map (Woodsmith and Stelzl, 2014). In both cases, we observed a significant enrichment for binary interactions between protein pairs that belong to a common complex ($p < 0.001$; Figure 2B). Third, we performed a similar analysis using tissue-specific mRNA expression data across the 16 human tissues of the Illumina Human Body Map 2.0 project as well as cellular compartment localization annotations from the GO Slim terms. Again, HI-II-14 was enriched for interactions mediated by protein pairs present in at least one common compartment or cell type (Figures 2C and 2D). Finally, we measured the overlap of HI-II-14 with specific biochemical relationships, as represented by kinase-substrate interactions. Both HI-II-14 and Lit-BM-13 contained significantly more PPIs reflecting known kinase-substrate relationships (Hornbeck et al., 2012) than the corresponding degree-controlled randomized networks (Figure 2E). In addition, HI-II-14 tended to connect tyrosine and serine/threonine kinases (Manning et al., 2002) to proteins with tyrosine or serine/threonine phospho-sites (Hornbeck et al., 2012; Olsen et al., 2010), respectively (Figure S3A), pointing to the corresponding interactions being genuine kinase-substrate interactions. In short, our systematic interactome map, which was generated independently from any pre-existing biological information, reveals functional relationships at levels comparable to those seen for the literature-based interaction map.

To further investigate the overall biological relevance of HI-II-14, we used an experimental approach that compares the impact of mutations associated with human disorders to that of common variants with no reported phenotypic consequences on biophysical interactions (Figure 3). Our rationale is that a set of interactions corresponding to genuine functional relationships should more likely be perturbed by disease-associated mutations than by common variants. The following example will illustrate this concept. Mutations R24C and R24H in CDK4 are clearly associated with melanoma by conferring resistance to CDKN2A inhibition (Wölfel et al., 1995), whereas N41S and S52N mutations are of less clear clinical significance (Zhong et al., 2009) and have remained functionally uncharacterized. HI-II-14 contains five CDK4 interactors: two inhibitors (CDKN2C and CDKN2D), two cyclins (CCND1 and CCND3), and HOOK1, a novel interacting partner and a potential phosphorylation target

(C) Improvements from first-generation to second-generation interactome mapping based on an empirically-controlled framework (Venkatesan et al., 2009). Completeness: fraction of all pairwise protein combinations tested; Assay sensitivity: fraction of all true biophysical interactions that are identifiable by a given assay; Sampling sensitivity: fraction of identifiable interactions that are detected in the experiment; Precision: fraction of reported pairs that are true positives. PRS: positive reference set; RRS: random reference set.

(D) Experimental pipeline for identifying high-quality binary protein-protein interactions (left). ORF: open reading frame. Fraction of HI-II-14, PRS, and RRS pairs (right) recovered by MAPPIT, PCA, and wNAPPA at increasing assay stringency. Shading indicates standard error of the proportion. $p > 0.05$ for all assays when comparing PRS and HI-II-14 at 1% RRS, two-sided Fisher's exact tests. For n values, see Table S6.

See also Figures S1 and S2 and Tables S1 and S2.

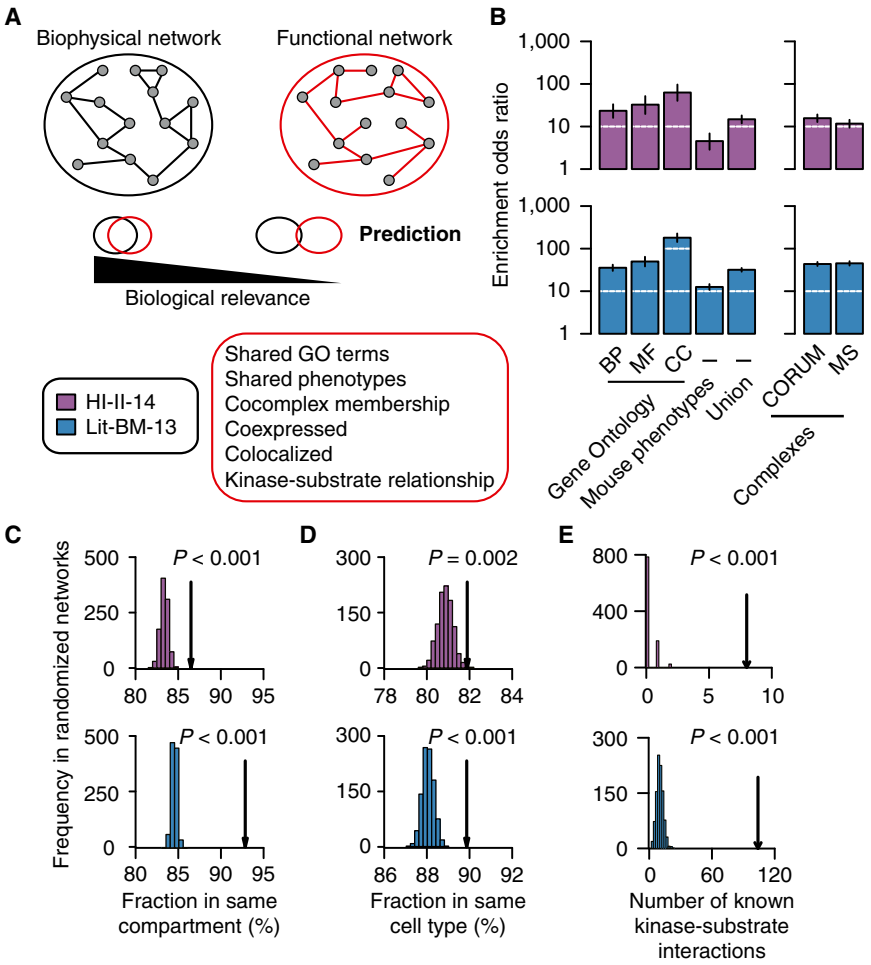


Figure 2. Overall Biological Significance

(A) Schematic of the method to assess biological relevance of binary maps.

(B) Enrichment of binary interactome maps for functional relationships (left) and cocomplex memberships (right). Error bars indicate 95% confidence intervals. BP: Biological process; MF: Molecular function; CC: Cellular component. Mouse phenotypes: Shared phenotypes in mouse models by orthology mapping. MS: Mass-spectrometry-based map. Enrichments: $p \leq 0.05$ for all annotations and maps, two-sided Fisher's exact tests. For n values, see Table S6.

(C) and (D) Fraction of binary interactions between proteins localized in a common cellular compartment and proteins copresent in at least one cell type (arrows) compared to those in 1,000 degree-controlled randomized networks. Empirical p values. For n values, see Table S6.

(E) Number of known kinase-substrate interactions found in binary maps (arrows) compared to those in 1,000 randomized networks. Empirical p values are shown.

See also Figure S3.

of CDK4 (Figure S3B). In agreement with previous reports, the comparative interaction profile shows that R24C and R24H, but not N41S and S52N, specifically perturb CDK4 binding to CDKN2C (Figure 3).

In total, we identified 32 human genes for which: (1) the corresponding gene product is reported to have binary interactors in HI-II-14, (2) germline disease-associated missense mutations have been reported, and (3) common coding missense variants unlikely to be involved in any disease have been identified in the 1000 Genomes Project (1000 Genomes Project Consortium, 2012). To avoid overrepresentation of certain genes, we selected a total of 115 variants, testing up to four disease and four common variants per disease gene for their impact on the ability of the corresponding proteins to interact with known interaction partners (see Extended Experimental Procedures, Section 8). Disease variants were 10-fold more likely to perturb interactions than nondisease variants (Figure 3 and Table S3). Strikingly, more than 55% of the 107 HI-II-14 interactions tested were perturbed by at least one disease-associated variant, and the same trend was observed when considering only mutants with evidence of expression in yeast as indicated by their ability to mediate at least one interaction (Figure S3C). Examples of novel specifically perturbed interactions include AANAT-BHLHE40 and RAD51D-IKZF1 (Figure 3). In the first case, the A129T mutation in AANAT is known to be associated with a delayed sleeping phase syndrome and specifically perturbs an interaction between AANAT and BHLHE40, the product of a gene reported to function in circadian rhythm regulation (Nakashima et al., 2008). In the second case, the breast-cancer-associated RAD51D E233G mutation perturbs interactions with a number of partners, including the known cancer gene product IKZF1 (Futreal et al., 2004).

Altogether these computational and experimental results provide strong evidence that HI-II-14 pairs correspond to biologically relevant interactions and represent a valuable resource to further our understanding of the human interactome and its perturbations in human disease.

A "Broader" Interactome

Unlike literature-curated interactions, HI-II-14 protein pairs are distributed homogeneously across the interactome space (Figure 4A), indicating that sociological biases, and not fundamental biological properties, underlie the existence of a densely populated zone in the literature. Since 1994, the number of high-quality binary literature PPIs has grown roughly linearly to reach ~11,000 interactions in 2013 (Figure 4B), while systematic data sets are punctuated by a few large-scale releases. Although the sparse territory of the literature map gradually gets populated, interaction density in this zone continues to lag behind that of the dense zone (Figure 4B). In terms of proteome coverage, the expansion rate is faster for systematic maps than for literature maps, especially in the sparse territory (Figure 4C and Figure S4A; see Extended Experimental Procedures,

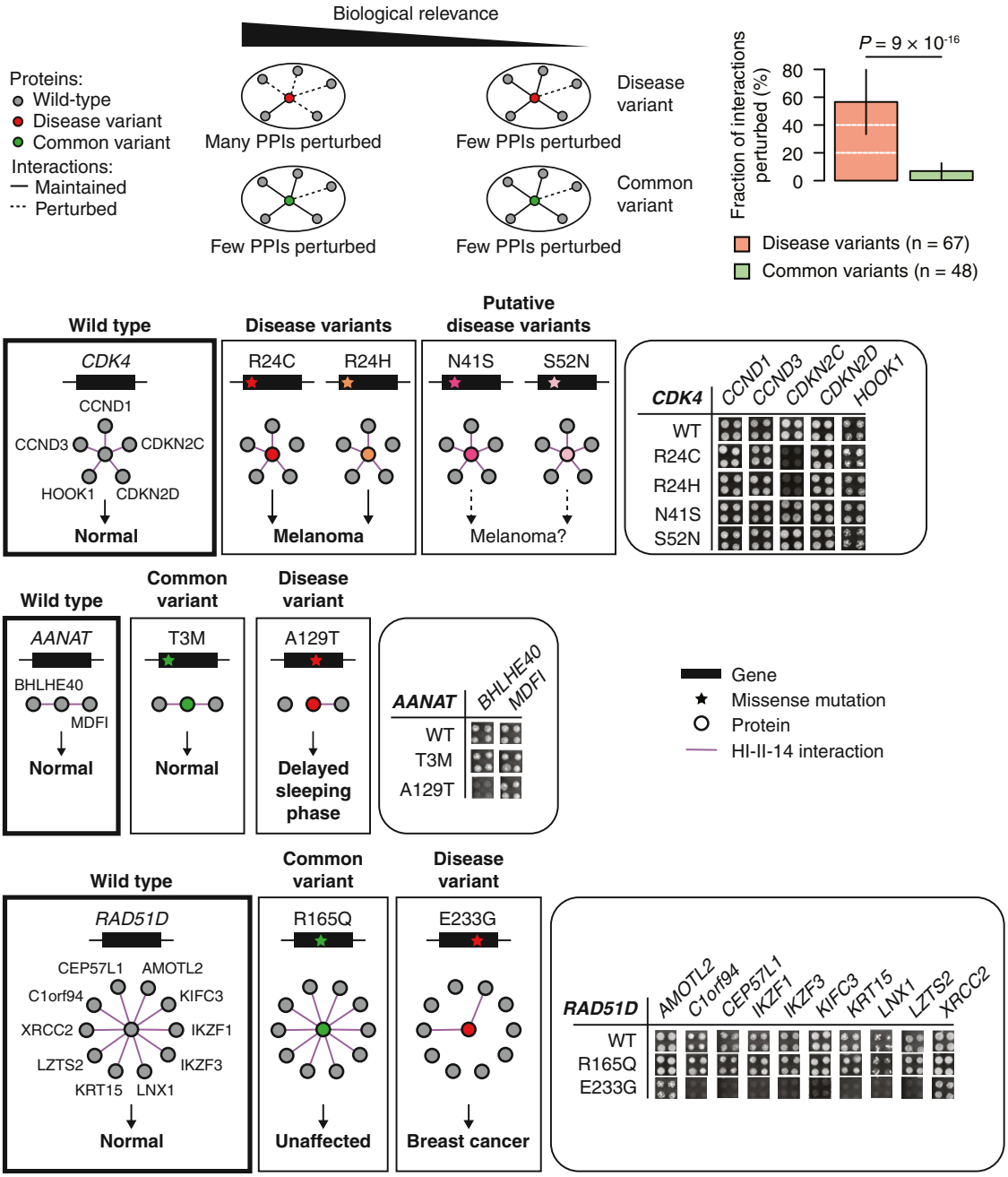


Figure 3. Perturbations of Protein Interactions by Disease and Common Variants

Predicted effect of mutations on PPIs as a function of their biological relevance (top left). Fraction of interactions of the wild-type gene product lost by mutants bearing the disease-associated or common variants (top right, error bars indicate standard error of the proportion). p value, two-sided Fisher's exact test. Comparison of interaction profile of wild-type CDK4, AANAT, and RAD51D to the interaction profile of mutants bearing disease or common variants (bottom). Yeast growth phenotypes on SC-Leu-Trp-His+3AT media in quadruplicate experiments are shown.

See also Figure S3 and Table S3.

Section 9). While Lit-BM-13 provides more information in the dense zone, HI-II-14 reveals interactions for more than 2,000 proteins absent from Lit-BM-13. These observations are likely due to a tendency of the literature map to expand from already connected proteins (Figure 4D).

To more deeply explore the heterogeneous coverage of the human interactome, we compared HI-II-14 and Lit-BM-13 to a collection of ~25,000 *pre*dicted binary *PPI*s of *h*igh-confidence (PrePPI-HC) (Zhang et al., 2012) and a *co-frac*tionation map of ~14,000 potentially binary interactions (Co-Frac) (Havugimana

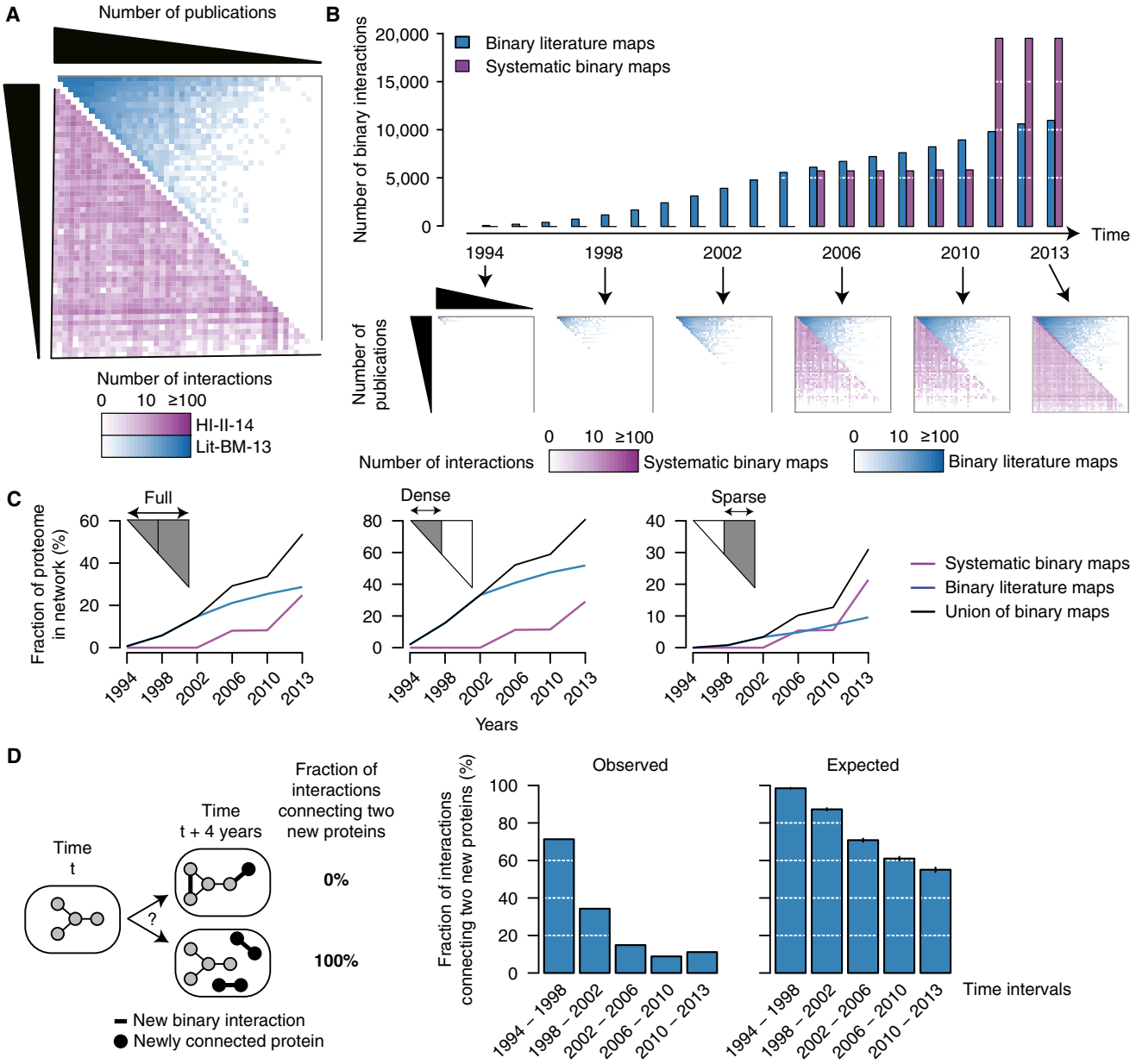


Figure 4. A "Broader" Interactome

(A) Adjacency matrices showing Lit-BM-13 (blue) and HI-II-14 (purple) interactions, with proteins in bins of ~350 and ordered by number of publications along both axes. The color intensity of each square reflects the total number of interactions for the corresponding bins.

(B) Total number of binary interactions in literature and systematic interactome maps over the past 20 years (top), with years reflecting either date of public release of systematic binary data sets or date of publication that resulted in inclusion of interactions in Lit-BM-13. Adjacency matrices (bottom) as in Figure 4A.

(C) Fraction of the human proteome present in binary interactome maps at selected time points since 1994, considering the full interactome space (left) or only dense (middle) and sparse (right) zones of Lit-BM-13 with respect to number of publications.

(D) Fraction of new interactions connecting two proteins that were both absent from the map at the previous time point (four years interval; middle) compared to the average in 1,000 randomized networks (right). Error bars indicate standard deviation.

et al., 2012). We tested the extent to which these two data sets contain binary interactions (see Extended Experimental Procedures, Section 10). Representative samples from both Co-Frac and PrePPI-HC were recovered by Y2H at a much lower rate than a sample of Lit-BM-13 and appeared statistically indistinguishable from random pairs (Figure 5A and Table S4A). A *lit*erature *n*on-*b*inary data set (Lit-NB-13) performed similarly. However, Co-Frac and PrePPI-HC, like Lit-NB-13, were both significantly enriched for functionally relevant relationships. Thus, although these data sets represent potentially valuable

resources, both Co-Frac and PrePPI-HC appear to be more comparable to nonbinary than to binary data sets. Surprisingly, even though PrePPI-HC and Co-Frac systematically surveyed the full human proteome and map different portions of the interactome (Figures S4B), both exhibit a strong tendency to report interactions among well-studied proteins (Figure 5B). This bias is likely due to the integration of functional annotations in the generation of both data sets.

Because coverage might depend on gene expression levels, we also examined interactome maps for expression-related sparse versus dense zones. Co-Frac shows a strong bias toward interactions involving proteins encoded by genes highly expressed in the cell lines used (Figure 5B). This expression-dependent bias is echoed in the literature map, perhaps reflecting a general tendency to study highly expressed proteins. In contrast, both HI-II-14 and PrePPI-HC exhibit a uniform interaction density across the full spectrum of expression levels, likely explained by the standardized expression of proteins tested in Y2H and by the independence of homology-based predictions from expression levels.

We more broadly explored the intrinsic biases that might influence the appearance of sparsely populated zones by examining 21 protein or gene properties, roughly classified as expression-, sequence-, or knowledge-based (Figures 5B and 5C, Tables S4B and S4C; see Extended Experimental Procedures, Section 9). For example, PrePPI-HC is virtually devoid of interactions between proteins lacking Pfam domains, consistent with conserved domains forming the basis of the prediction method. HI-II-14 appears depleted of interactions among proteins containing predicted transmembrane helices, consistent with expected limitations of the Y2H assay (Stagljar and Fields, 2002). Co-Frac is similarly depleted in interactions involving proteins with transmembrane helices, which may result from membrane-bound proteins being filtered out during biochemical fractionations. Compared to HI-II-14, HI-I-05 presented a less homogenous coverage of the space with respect to abundance and knowledge properties, likely reflecting the content of early versions of the hORFeome (Figure S4C). Importantly, no single map appeared unbiased in all 21 examined properties. A combined map presented a slightly increased homogeneity although intrinsic knowledge biases of the three maps using literature-derived evidence were still predominant.

To confirm that HI-II-14 interactions found in the sparse zones of the three other maps are of as high quality as those found in dense zones, we compared MAPPIT validation rates and functional enrichment across these zones for all protein properties examined. MAPPIT validation rates of dense and sparse zone pairs were consistent for nearly all properties (Figures 5D and S4D), indicating that HI-II-14 interactions are of similar biophysical quality throughout the full interactome space. Functional enrichment within the sparse zone was statistically indistinguishable from that of the dense zone (Figures 5D and S4E), demonstrating the functional importance of HI-II-14 biophysical interactions in zones covered sparsely by other types of interactome maps.

Considering all current maps, more than half of the proteome is now known to participate in the interactome network. Our systematic exploration of previously uncharted territories dramatically expands the interactome landscape, suggesting that the human interactome network is broader in scope than previously observed and that the entire proteome may be represented within a fully mapped interactome.

Interactome Network and Cancer Landscape

Genes associated with the same disease are believed to be preferentially interconnected in interactome networks (Barabási et al., 2011; Vidal et al., 2011). However, in many cases, these observations were made with interactome maps that are composites of diverse evidence, e.g., binary PPIs, cocomplex memberships, and functional associations, a situation further complicated by the uneven quality and sociological biases described above. Using HI-II-14, we revisited this concept for cancer gene products. Our goal was to investigate whether the cancer genomic landscape is limited to the known cancer genes curated in the Sanger Cancer Gene Census (“Cancer Census”) (Futreal et al., 2004), or if, alternatively, it might extend to some of the hundreds of additional candidate genes enriched in somatic mutations uncovered by systematic cancer genome sequencing (“SM genes”) (Chin et al., 2011) and/or identified by functional genomic strategies such as Sleeping Beauty transposon-based screens in mice (“SB genes”) (Copeland and Jenkins, 2010) or global investigations on DNA tumor virus targets (“VT genes”) (Rozenblatt-Rosen et al., 2012).

Given our homogeneous coverage of the space for known (Cancer Census) and candidate (SB, SM, and VT) cancer genes (Figure 6A), we first tested the postulated central role of cancer gene products in biological networks (Barabási et al., 2011) and verified that both sets tend to have more interactions and to be more central in the systematic map than proteins not associated with cancer (Figure 6B). We then examined the interconnectivity of known cancer proteins and showed that Cancer Census gene products interact with each other more frequently than expected by chance, a trend not apparent in HI-I-05 (Figure 6C). We sought to use this topological property as the basis for novel cancer gene discovery in the large lists of cancer candidates from genomic and functional genomic screens.

We examined whether products of candidate cancer genes identified by GWAS (Table S5A) tend to be connected to Cancer Census proteins, and observed significant connectivity in all four maps (Figure S5A; see Extended Experimental Procedures, Section 11). When loci containing a known cancer gene were excluded, only HI-II-14 showed such connectivity, supporting its unique value to identify cancer candidate genes beyond those already well demonstrated (Figures 7A and S5A). In further support of their association with cancer, genes in cancer GWAS loci prioritized by “guilt-by-association” in HI-II-14 tend to correspond to cancer candidates from systematic cancer studies (Figures 7B and 7C). These results suggest that cancer-associated proteins tend to form subnetworks perturbed in tumorigenesis, and that HI-II-14 provides new context to prioritize cancer genes from genome-wide studies.

The following example illustrates the power of our combined approach. C-terminal Binding Protein 2 (CTBP2) is encoded at a locus associated with prostate cancer susceptibility (Thomas et al., 2008) and belongs to both SB and VT gene lists (Mann et al., 2012; Rozenblatt-Rosen et al., 2012). Two Cancer Census genes, *IKZF1* and *FLI1*, encode interacting partners of

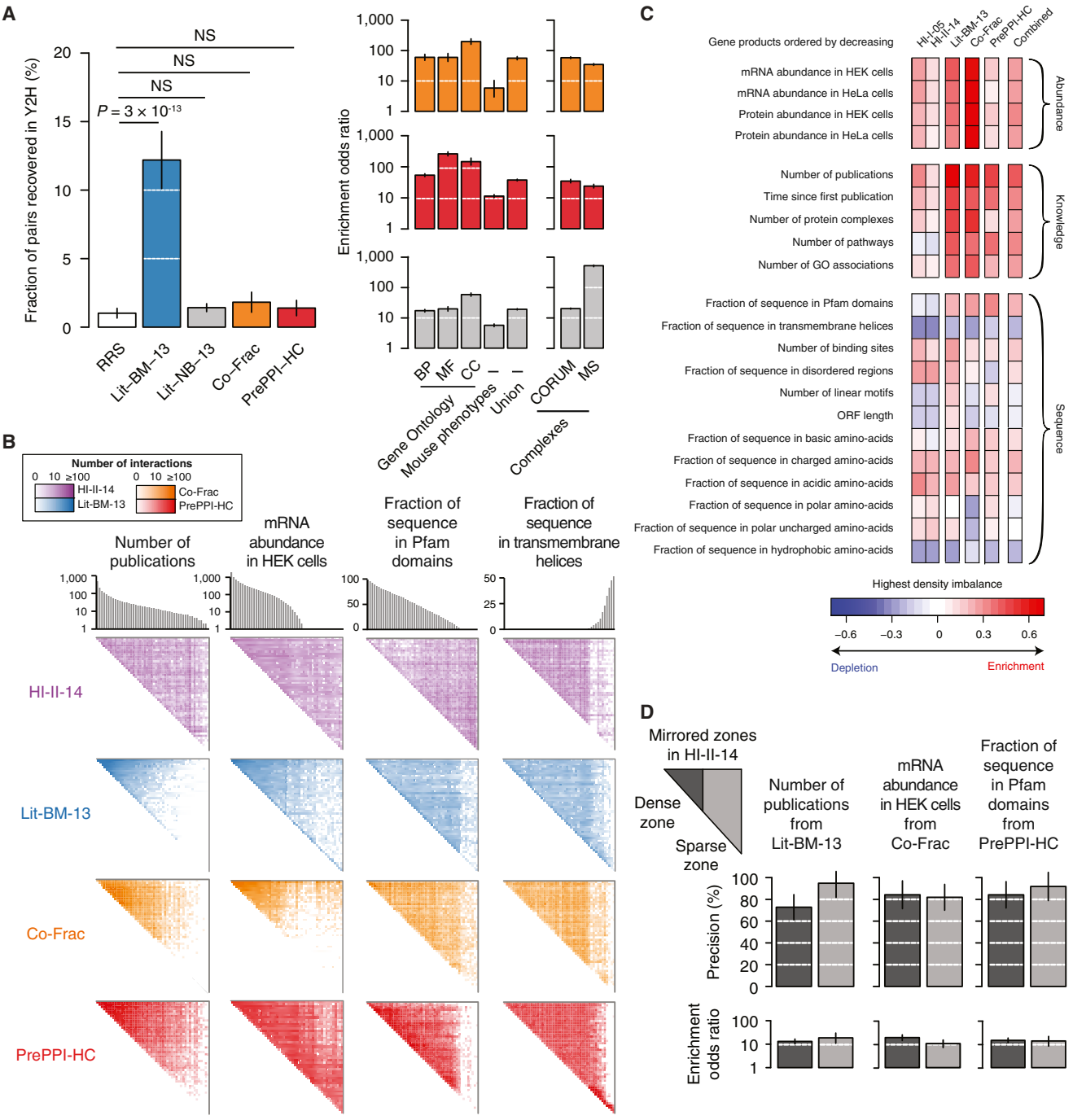


Figure 5. Comparison of Interaction Mapping Approaches

(A) Evaluation of the quality of Co-Frac (orange), PrePPI-HC (red), and pairs from small-scale experiments in the literature with no binary evidence (Lit-NB-13, grey). Fraction of pairs recovered by Y2H as compared to pairs from Lit-BM-13 and pairs of randomly selected proteins (RRS) (left). Error bars indicate standard error of the proportion. Enrichment in functional interactions and cocomplex memberships (right). Legend as in Figure 2B. For n values, see Table S6.

(B) Adjacency matrices for HI-II-14, Lit-BM-13, Co-Frac, and PrePPI-HC maps, with proteins per bins of ~350 and ordered by number of publications, mRNA abundance in HEK cells, fraction of protein sequence covered by Pfam domains, or fraction of protein sequence in transmembrane helices. Figure legend as in Figure 1B.

(C) Highest interaction density imbalances (observed minus expected) in the four maps, the union of all four maps, and our previous binary map (HI-I-05) for 21 protein properties.

(legend continued on next page)

CTBP2 in HI-II-14. These are transcription factors with tumor suppressor (Payne and Dovat, 2011) and proto-oncogene (Kornblau et al., 2011) roles, respectively, in lymphoid tumors. Given its interactions with IKZF1 and FLI1, we investigated the potential role of CTBP2 in lymphoid tumorigenesis. In the Cancer Cell Line Encyclopedia (Barretina et al., 2012), *FLI1* was significantly more often amplified in lymphoid than in other cell lines (Figure 7D), consistent with its proposed proto-oncogenic role in these tumors. In contrast, both *CTBP2* and *IKZF1*, but not *CTBP1*, were deleted significantly more often in lymphoid cancer cell lines. Notably, deletion of *CTBP2* or *IKZF1* and amplification of *FLI1* were mostly nonoverlapping in the different cell lines, suggesting that either event may be sufficient to affect tumorigenesis (Figure S5B). Altogether, these results suggest a role for CTBP2 in suppressing lymphoid tumors by direct repression of FLI1 function, potentially involving IKZF1.

Finally, we assessed how HI-II-14 interactions can be integrated with genomic and functional genomic data sets. Going beyond the "guilt-by-profiling" concept, we also used these gene sets in "guilt-by-association" predictions in a combined model (Figure S6A), which leads to substantially improved cancer gene rankings over those found using either predictive strategy alone (Figures 7E, S6B, and S6C and Table S5B; see Extended Experimental Procedures, Section 12). In contrast, a similar analysis using HI-I-05 interactions showed that its limited size prevented inclusion of any guilt-by-association terms (Figure S6D). Genes significantly mutated in cancer patients from recent TCGA pan-cancer mutation screens (Table S5C) (Lawrence et al., 2014) were enriched among highly ranked predictions from the combined model ($p = 6 \times 10^{-3}$, one-sided Wilcoxon rank test), supporting the validity of our integrated cancer gene predictions. Our top-ranked prediction was the cyclin-dependent kinase 4 (CDK4), a well-known cancer gene product. Four other genes from the Cancer Census list appeared among the top 25 ranked genes. Strikingly, STAT3, which ranked third, was added to the Cancer Census after our training set was established, highlighting the ability of this approach to identify novel cancer gene products.

To characterize the biological processes in which the candidate cancer genes predicted by the combined model are likely to be involved, we identified binary interactions linking them to each other or to Cancer Census proteins in the 12 "pathways of cancer" relevant to cancer development and progression (Table S5D) (Vogelstein et al., 2013). Of our top 100 candidates, 60 mapped to at least one cancer pathway (Figures 7F and S7), twice as many as would be expected from predictions using either the guilt-by-profiling or guilt-by-association approach alone. We propose that many novel cancer candidates can be annotated to specific processes based on their interactions with Cancer Census gene products and known participation in cellular pathways. For example, the candidate protein ID3, a DNA-binding inhibitor, interacts with the two Cancer Census transcription factors TCF12 and TCF3, suggesting a role for ID3 in the regulation of transcription by inhibiting binding of specific transcription factors to DNA (Loveys et al., 1996; Richter et al., 2012). CTBP2, which we identified as a potential suppressor in lymphoid tumors, represents another example (Figures 5E and S7).

In summary, the increased and uniform coverage of HI-II-14 demonstrates that known and candidate cancer gene products are highly connected in the interactome network, which in turn provides unbiased evidence for an expanded functional cancer landscape.

DISCUSSION

By systematically screening half of the interactome space with minimal inspection bias, we more than doubled the number of high-quality binary PPIs available from the literature. Covering zones of the human interactome landscape that have been weakly charted by other approaches, our systematic binary map provides deeper functional context to thousands of proteins, as demonstrated for candidates identified in unbiased cancer genomic screens. Systematic binary mapping therefore stands as a powerful approach to "connect the dots" of the genomic revolution.

Combining high-quality binary pairs from the literature with systematic binary maps, 30,000 high-confidence interactions are now available. It is likely that a large proportion of the human interactome can soon be mapped by taking advantage of the emergence of reference proteome maps (Kim et al., 2014; Wilhelm et al., 2014), a combination of nearly complete clone collections (Yang et al., 2011), rapid improvements in Y2H assay sensitivity, and emerging interaction-mapping technologies that drastically reduce cost (Caufield et al., 2012; Stagljar and Fields, 2002; Yu et al., 2011).

Reference binary interactome maps of increased coverage and quality will be required to interpret condition-specific interactions and to characterize the effects of splicing and genetic variation on interactions (Zhong et al., 2009). While protein-protein interactions represent an important class of interactions between macromolecules, future efforts integrating this information with protein-DNA, protein-RNA, RNA-RNA or protein-metabolite interactions will provide a unified view of the molecular interactions governing cell behavior. Just as a reference genome enabled detailed maps of human genetic variation (1000 Genomes Project Consortium, 2012), completion of a reference interactome network map will enable deeper insight into genotype-phenotype relationships in human.

EXPERIMENTAL PROCEDURES

Extraction of the Literature-Based Data Sets

Human PPIs annotated with tractable publication records were extracted from seven databases through August 2013. Large-scale systematic data sets and

(D) Precision at 1% RRS recovery in the MAPPIT assay (top, error bars indicate standard error of the proportion) and functional enrichment (bottom, union of Gene Ontology and mouse phenotypes based annotations, error bars indicate 95% confidence intervals) of HI-II-14 pairs found in dense and sparse zones mirrored from Lit-BM-13, Co-Frac, and PrePPI-HC. $p > 0.05$ for all pairwise comparisons of dense and sparse zones, two-sided Fisher's exact tests. For n values, see Table S6.

See also Figure S4 and Table S4.

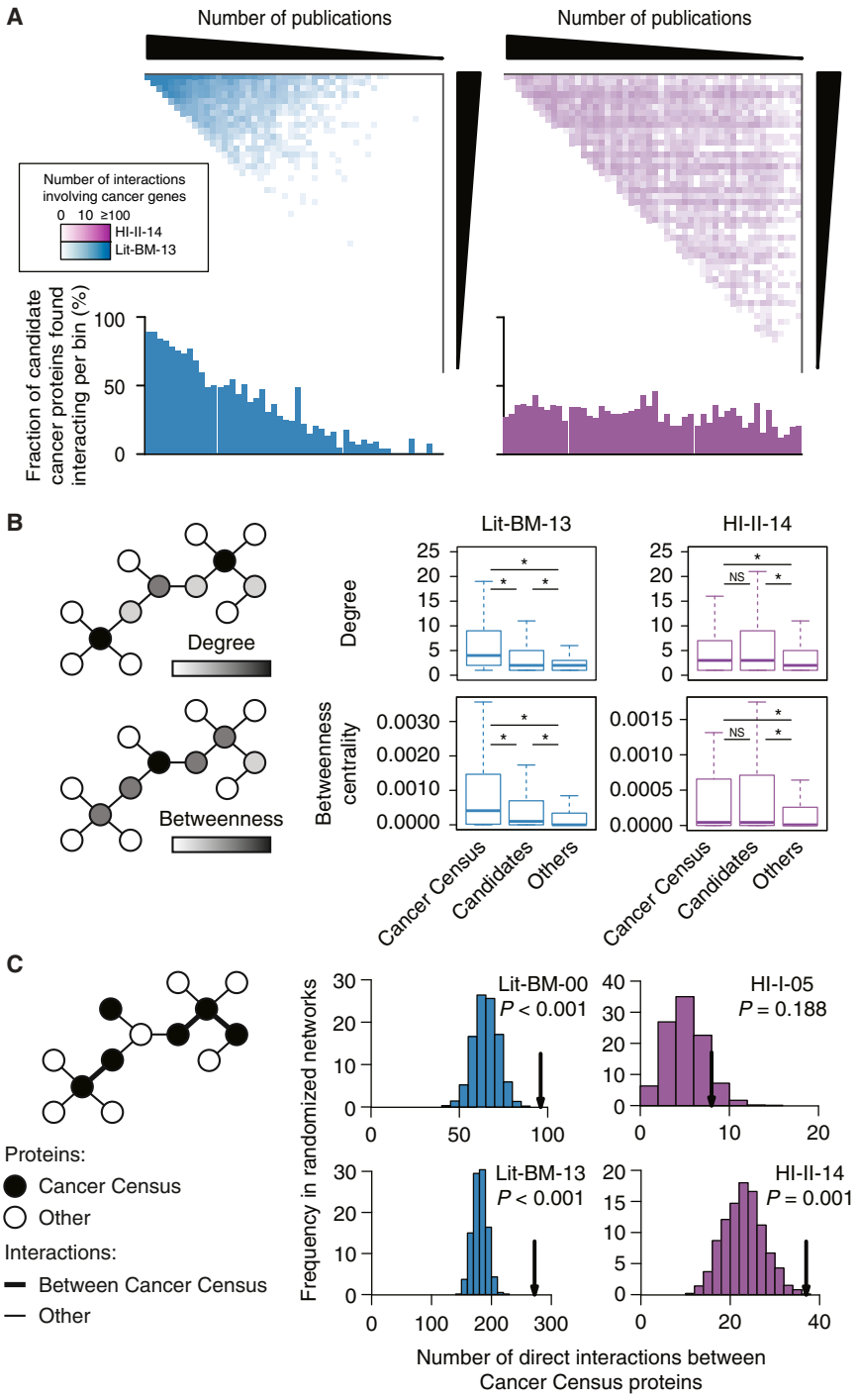


Figure 6. Network Properties of Cancer Gene Products

(A) Adjacency matrices for Lit-BM-13 and HI-II-14 only showing interactions involving the product of a Cancer Census (Futreal et al., 2004) or of a candidate cancer gene. Figure legend as in Figure 1B. Lower histograms show for each bin, the fraction of cancer candidates having at least one interaction.

(B) Distribution of the number of interactions (degree) and normalized number of shortest paths between proteins (betweenness centrality) for products of Cancer Census and of candidate cancer genes in Lit-BM-13 and in HI-II-14 maps as compared to other proteins (right; * for $p < 0.05$, NS for $p > 0.05$, two-sided Wilcoxon rank sum tests). For n values, see Table S6.

(C) Number of interactions between products of Cancer Census genes (arrows) in HI-I-05, HI-II-14, Lit-BM as of 2000 (Lit-BM-00) and as of 2013 (Lit-BM-13), as compared to 1,000 degree-controlled randomized networks. Empirical p values. For n values, see Table S6.

pairs involving products of *UBC*, *SUMO1*, *SUMO2*, *SUMO3*, *SUMO4*, or *NEDD8*, were excluded. The remaining pairs were divided into those having no pieces of binary evidence (Lit-NB) and those with at least one piece of binary evidence based on PSI-MI experimental method codes. Binary pairs were divided between pairs with one and with two or more pieces of evidence (Lit-BS and Lit-BM, respectively). For benchmark experiments in Y2H, MAPPIT, PCA, and wNAPPA, equivalent data sets were extracted similarly in December 2010.

Generation of the Binary Protein-Protein Interaction Map

HI-II-14 was generated by screening all pairwise combinations of 15,517 ORFs from hORFeome v5.1 (Space II) as described previously (Dreze et al., 2010). ORFs encoding first pass pairs were identified either by Sanger sequencing or by Stitch-seq (Yu et al., 2011). HI-II-14 was validated by comparing a subset of 809 interactions to a positive and a random reference set of 460 and 698 protein pairs, respectively, using MAPPIT, PCA, and wNAPPA assays.

Interaction Perturbation by Missense Mutations

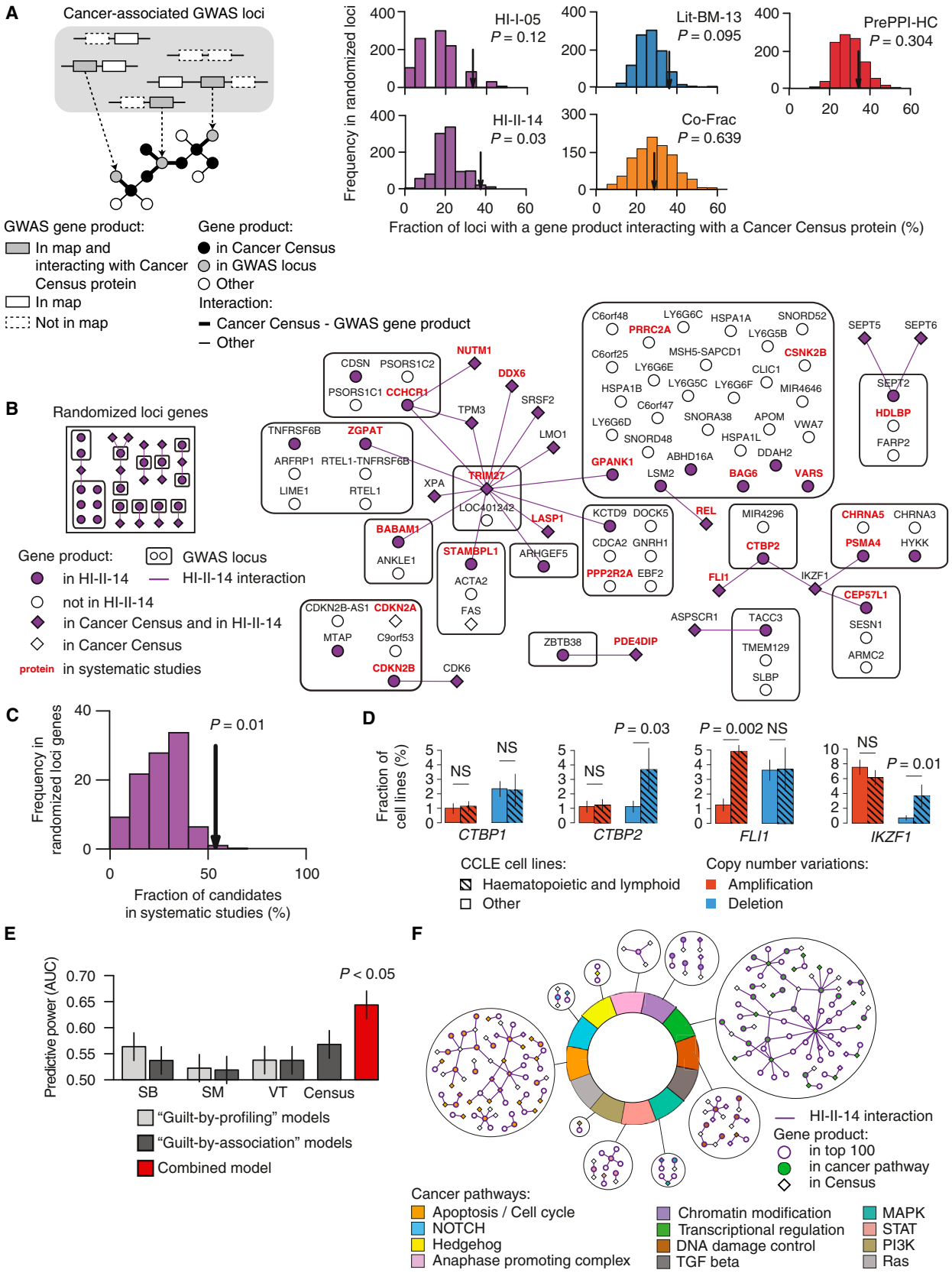
Disease variants were obtained from the Human Gene Mutation Database (HGMD 2009 V2) (Stenson et al., 2014) and common variants were derived from the 1000 Genomes Project (1000 Genomes Project Consortium, 2012). Only variants with a minor allele frequency above 1% were considered common. All successfully cloned disease and common variants were systematically tested for interaction with all interactors of their wild-type counterpart.

Interaction Density Imbalance

For each protein property, we ranked all proteins and, for any property threshold, partitioned the interactome space into a first region containing pairs of proteins both above (or below) the threshold, and a second region containing all remaining pairs. Interaction density imbalance of a given map for a given threshold was calculated as the fraction of PPIs *observed* in the first region minus the fraction of PPIs *expected* assuming a uniform distribution in the space. Dense and sparse zones were defined by identifying the threshold for which the deviation from expectation is maximal.

Measure of Functional Enrichment

For each pairwise comparison, PPI and functional maps were trimmed to pairs where both proteins were present in both maps and restricted to Space II to



(legend on next page)

allow comparison between PPI maps. Functional enrichment odds ratios were calculated using Fisher's exact tests.

GWAS Analysis

307 distinct cancer-associated SNPs were identified from 75 GWAS publications covering 10 types of cancer and 142 distinct loci were identified at a linkage disequilibrium threshold of 0.9. For each map, we calculated the number of loci encoding an interactor of a Cancer Census protein over the number of loci encoding a protein in the PPI map. To assess significance, we measured the corresponding fraction when randomly selecting for each locus the same number of proteins than genes with products in the PPI map.

Cancer Association Scoring System

For each gene, seven features were measured. Three features represent membership in the SB, SM, and VT lists of candidate cancer genes ("guilt-by-profiling" features). The four other features represent its number of interactors in HI-II-14 that are present in these three lists and in the Cancer Census list, normalized by the expected numbers in degree-controlled randomized networks ("guilt-by-association" features). We measured the ability of each feature to prioritize known Cancer Census genes with separate logistic regression models. We combined all seven features in a forward stepwise logistic regression model using the Akaike information criterion to determine the stepwise halting. The final set of features selected was: the SB, SM, and VT guilt-by-profiling and the Cancer Census and SB guilt-by-association features. "Receiver Operating Characteristic" curves were obtained by measuring at decreasing score threshold the fraction of known Cancer Census genes recovered and the corresponding fraction of proteins predicted as candidate cancer genes.

Data Sets

For reference data sets used in this study, see Extended Experimental Procedures, Section 13. All high-quality binary PPIs described in this paper can be accessed at: http://interactome.dfci.harvard.edu/H_sapiens/.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, seven figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2014.10.050.

AUTHOR CONTRIBUTIONS

Computational analyses were performed by T.R., M.T., B.C., S.J.P., C. Fontanillo, R.M., A.K., and S.D.G. with help from A.-R.C., J.C.-H., C. Fan, E.F., M.J., S.K., G.N.L., K.L., J.M., A. Sharma, and Y.S. Experiments were performed by Q.Z., N.S., S.Y., I.L., X.Y., and L.G. with help from D.B., B.E.B., P.B., M.B., M.P.B., D.C.-Z., R.C., E.D., M.D., A.D., F.G., B.J.G., M.F.H., R.K., A.M., R.R.M., A.P., M.M.P., X.R., J.R., P.R., V.R., E.R., J.M.S., A. Scholz, A.A.S., K.S., S.T., A.O.T., S.A.T., J.-C.T., K.V., and J.W. Structural analyses were done by T.R., M.T., and R.M. Extraction of the literature data sets was performed by C. Fontanillo, A.K., Ch.F., M.E.C., and T.H. MAPPIT validation was done by I.L. The adjacency matrix interactome representation was developed by S.J.P. with M.T. Functional enrichment analysis was done by T.R., M.T., and B.C. Interaction perturbation experiments were performed by N.S. and S.Y. with Q.Z. Interactome and proteome coverage analyses were done by T.R. and B.C. Comparison of alternative maps was done by T.R., M.T., B.C., and S.J.P. The density imbalance measure was conceived by M.T. Topological analyses were done by T.R., M.T., B.C., S.J.P., and S.D.G. Cancer-related analyses were done by T.R., M.T., and B.C. The cancer association scoring system was done by M.T. CTBP2 and cancer landscape analyses were done by T.R. Interactome mapping was supervised by D.E.H. and M.V. Principal investigators overseeing primary data management, structural biology, literature recuration and reference set construction, MAPPIT validation, and other computational analyses were T.H., P.A., J.D.L.R., J.T., and F.P.R., respectively. B.C., Y.X., A.-L.B., L.M.I., P.A., J.D.L.R., J.T., M.A.C., D.E.H., T.H., F.P.R., and M.V. designed and/or advised the overall research effort. T.R., M.T., B.C., Q.Z., M.E.C., J.D.L.R., M.A.C., D.E.H., T.H., F.P.R., and M.V. wrote the manuscript with contributions from other coauthors.

ACKNOWLEDGMENTS

The authors wish to acknowledge past and present members of the Center for Cancer Systems Biology (CCSB) and particularly H. Yu for helpful discussions. This work was supported primarily by NHGRI grant R01/U01HG001715 awarded to M.V., D.E.H., F.P.R., and J.T. and in part by the following grants and agencies: NHGRI P50HG004233 to M.V., F.P.R., and A.-L.B.; NHLBI U01HL098166 subaward to M.V.; NHLBI U01HL108630 subaward to A.-L.B.; NCI U54CA112962 subaward to M.V.; NCI R33CA132073 to M.V.; NIH RC4HG006066 to M.V., D.E.H., and T.H.; NICHD ARRA R01HD065288, R21MH104766, and R01MH105524 to L.M.I.; NIMH R01MH091350 to L.M.I. and T.H.; NSF CCF-1219007 and NSERC RGPIN-2014-03892 to Y.X.; Canada Excellence Research Chair, Krembil Foundation, Ontario Research Fund–Research Excellence Award, Avon Foundation, grant CSI07A09 from Junta de Castilla y Leon (Valladolid, Spain), grant PI12/00624 from Ministerio de Economia 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, grant FWO-V G.0864.10 from the Fund for Scientific Research-Flanders and ERC Advanced Grant N° 340941 to J.T.; EMBO long-term fellowship to A.K.; Institute Sponsored Research funds from the

Figure 7. Interactome Network and Cancer Landscape

(A) Fraction of cancer-related GWAS loci containing at least one gene encoding a protein that interacts with the product of a Cancer Census gene in HI-I-05, HI-II-14, Lit-BM-13, Co-Frac, and PrePPI-HC (arrows) as compared to randomly selected loci genes. GWAS loci already containing a Cancer Census gene are excluded. Empirical p values. For n values, see Table S6.

(B) Network representing products of genes in cancer-associated GWAS loci and their interactions with Cancer Census proteins in HI-II-14 (right), and a representative example of the network obtained for randomized loci genes (left).

(C) Fraction of GWAS loci gene products interacting with a Cancer Census protein also identified in systematic genomic and functional genomic studies (arrow) as compared to the fraction obtained for randomized loci genes (bottom right). Empirical p value.

(D) *CTBP2* and *IKZF1* are deleted in significantly more hematopoietic and lymphoid cancer cell lines than in other cancer cell lines. CCLE, Cancer Cell Line Encyclopedia. Each barplot compares the fraction of cell lines from the 163 hematopoietic and lymphoid (hatched bars) or 717 other (empty bars) cell types where *CTBP1*, *CTBP2*, *FLI* ,or *IKZF1* were found amplified (red) or deleted (blue). Error bars indicate standard error of the proportion. p values, two-sided Fisher's exact tests (NS for $p > 0.05$).

(E) Predictive power of guilt-by-profiling and guilt-by-association models compared to the combined model (Figure S6; see Extended Experimental Procedures, Section 11). AUC: Area under the curve in Figure S6C. Error bars indicate standard error of the proportion. p value, two-sided Wilcoxon rank sum test. SB, Sleeping Beauty transposon-based mouse cancer screen; SM, Somatic mutation screen in cancer tissues; VT, Virus targets.

(F) Binary interactions from HI-II-14 involving the top candidates and Cancer Census gene products in the twelve pathways associated to cancer development and progression.

See also Figures S5, S6, and S7 and Table S5.

Dana-Farber Cancer Institute Strategic Initiative to M.V. I.L. is a postdoctoral fellow with the FWO-V. M.V. is a "Chercheur Qualifié Honoraire" from the Fonds de la Recherche Scientifique (FRS-FNRS, Wallonia-Brussels Federation, Belgium). Since performing the work described, C. Fontanillo has become an employee of Celgene Research SL, part of the Celgene Corporation.

Received: September 17, 2014
Revised: October 21, 2014
Accepted: October 30, 2014
Published: November 20, 2014

REFERENCES

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

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

Barabási, A.-L., Gulbahce, N., and Loscalzo, J. (2011). Network medicine: a network-based approach to human disease. Nat. Rev. Genet. *12*, 56–68.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehár, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

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

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

Caufield, J.H., Sakhawalkar, N., and Uetz, P. (2012). A comparison and optimization of yeast two-hybrid systems. Methods *58*, 317–324.

Chatr-Aryamontri, A., Breitkreutz, B.J., Heinicke, S., Boucher, L., Winter, A., Stark, C., Nixon, J., Ramage, L., Kolas, N., O'Donnell, L., et al. (2013). The BioGRID interaction database: 2013 update. Nucleic Acids Res. *41* (Database issue), D816–D823.

Chin, L., Hahn, W.C., Getz, G., and Meyerson, M. (2011). Making sense of cancer genomic data. Genes Dev. *25*, 534–555.

Copeland, N.G., and Jenkins, N.A. (2010). Harnessing transposons for cancer gene discovery. Nat. Rev. Cancer *10*, 696–706.

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

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

Edwards, A.M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J., and Gerstein, M. (2002). Bridging structural biology and genomics: assessing protein interaction data with known complexes. Trends Genet. *18*, 529–536.

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

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

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33* (Database issue), D514–D517.

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

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

Hornbeck, P.V., Kornhauser, J.M., Tkachev, S., Zhang, B., Skrzypek, E., Murray, B., Latham, V., and Sullivan, M. (2012). PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. Nucleic Acids Res. *40* (Database issue), D261–D270.

International Human Genome Sequencing Consortium (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The IntAct molecular interaction database in 2012. Nucleic Acids Res. *40* (Database issue), D841–D846.

Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., et al. (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. *37* (Database issue), D767–D772.

Kim, M.S., Pinto, S.M., Getnet, D., Nirujogi, R.S., Manda, S.S., Chaerkady, R., Madugundu, A.K., Kelkar, D.S., Isserlin, R., Jain, S., et al. (2014). A draft map of the human proteome. Nature *509*, 575–581.

Kornblau, S.M., Qiu, Y.H., Zhang, N., Singh, N., Faderl, S., Ferrajoli, A., York, H., Qutub, A.A., Coombes, K.R., and Watson, D.K. (2011). Abnormal expression of FLI1 protein is an adverse prognostic factor in acute myeloid leukemia. Blood *118*, 5604–5612.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A.P., Santonico, E., et al. (2012). MINT, the molecular interaction database: 2012 update. Nucleic Acids Res. *40* (Database issue), D857–D861.

Loveys, D.A., Streiff, M.B., and Kato, G.J. (1996). E2A basic-helix-loop-helix transcription factors are negatively regulated by serum growth factors and by the Id3 protein. Nucleic Acids Res. *24*, 2813–2820.

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

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

Nakashima, A., Kawamoto, T., Honda, K.K., Ueshima, T., Noshiro, M., Iwata, T., Fujimoto, K., Kubo, H., Honma, S., Yorioka, N., et al. (2008). DEC1 modulates the circadian phase of clock gene expression. Mol. Cell. Biol. *28*, 4080–4092.

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

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

Payne, K.J., and Dovat, S. (2011). Ikaros and tumor suppression in acute lymphoblastic leukemia. Crit. Rev. Oncog. *16*, 3–12.

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

Richter, J., Schlesner, M., Hoffmann, S., Kreuz, M., Leich, E., Burkhardt, B., Rosolowski, M., Ammerpohl, O., Wagener, R., Bernhart, S.H., et al.; ICGC MMML-Seq Project (2012). Recurrent mutation of the *ID3* gene in Burkitt lymphoma identified by integrated genome, exome and transcriptome sequencing. Nat. Genet. *44*, 1316–1320.

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

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

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., Frishman, G., Montrone, C., and Mewes, H.W. (2010). CORUM: the comprehensive resource of mammalian protein complexes—2009. Nucleic Acids Res. *38* (Database issue), D497–D501.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. *32* (Database issue), D449–D451.

Stagljar, I., and Fields, S. (2002). Analysis of membrane protein interactions using yeast-based technologies. Trends Biochem. Sci. *27*, 559–563.

Stein, A., Céol, A., and Aloy, P. (2011). 3did: identification and classification of domain-based interactions of known three-dimensional structure. Nucleic Acids Res. *39* (Database issue), D718–D723.

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

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

Stumpf, M.P., Thorne, T., de Silva, E., Stewart, R., An, H.J., Lappe, M., and Wiuf, C. (2008). Estimating the size of the human interactome. Proc. Natl. Acad. Sci. USA *105*, 6959–6964.

Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A., et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. Nat. Genet. *40*, 310–315.

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

Vidal, M., Cusick, M.E., and Barabási, A.-L. (2011). Interactome networks and human disease. Cell *144*, 986–998.

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

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. *31*, 258–261.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M.M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., et al. (2014). Mass-spectrometry-based draft of the human proteome. Nature *509*, 582–587.

Wölfel, T., Hauer, M., Schneider, J., Serrano, M., Wölfel, C., Klehmann-Hieb, E., De Plaen, E., Hankeln, T., Meyer zum Büschenfelde, K.H., and Beach, D. (1995). A p16INK4a-insensitive CDK4 mutant targeted by cytolytic T lymphocytes in a human melanoma. Science *269*, 1281–1284.

Woodsmith, J., and Stelzl, U. (2014). Studying post-translational modifications with protein interaction networks. Curr. Opin. Struct. Biol. *24*, 34–44.

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

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

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

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

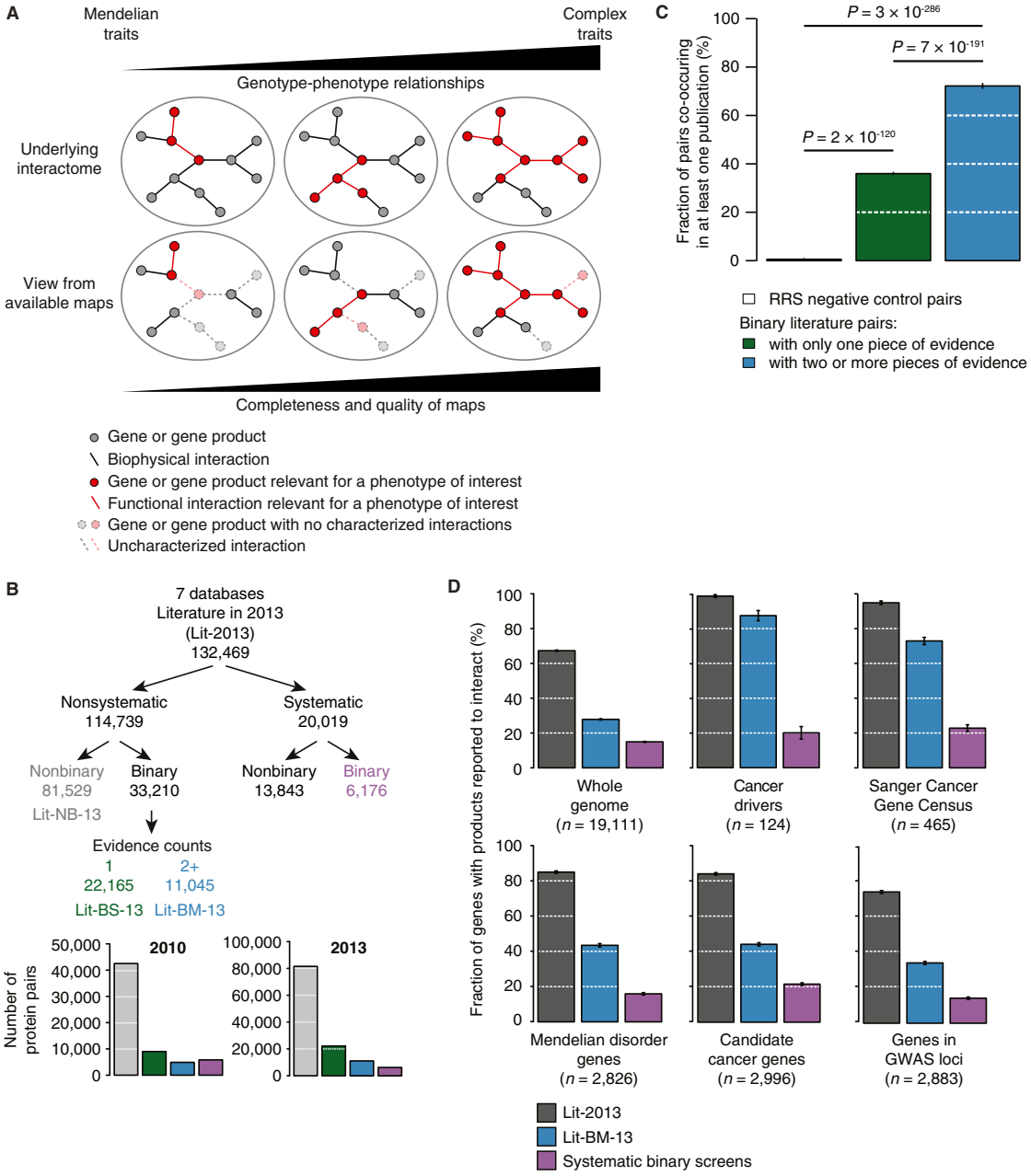


Figure S1. Vast Uncharted Interactome Zone in Literature, Related to Figure 1

(A) Understanding genotype-phenotype relationships through interactome maps. Interactome network maps of increasing completeness and quality are likely required to better understand genotype-phenotype relationships of increasing complexity.

(B) Assembling literature data sets from public databases. Flow chart showing the assembly and segregation of PPIs from the literature as of 2013. Different pieces of evidence correspond to different papers or methods. Histograms showing the number of pairs in the different subsets as of 2010 and 2013.

(C) Evaluation of Lit-BS-10 and Lit-BM-10 quality. Fraction of pairs found to co-occur in the literature based on the "text mining" score from the STRING database when restricting the Lit-BM-10 pairs to those reported in a single paper. Error bars indicate standard error of the proportion. p values based on two-sided Fisher's exact tests. For n values, see Table S6.

(D) Limited number of high-quality protein-protein interactions available for putative disease gene products identified in systematic studies. Fraction of full proteome and putative disease-associated gene products with protein-protein interactions in Lit-2013, Lit-BM-13, and systematic binary screens. Although cancer driver gene products are well represented in Lit-BM-13, less than 30% of proteins encoded by genes identified in systematic genome-wide cancer screens have any reported interactions in Lit-BM-13. Coverage is similarly low for proteins associated with Mendelian traits or identified in genome-wide association studies (GWAS). Error bars indicate standard error of the proportion. Data sources are cited in Extended Experimental Procedures, Section 13.

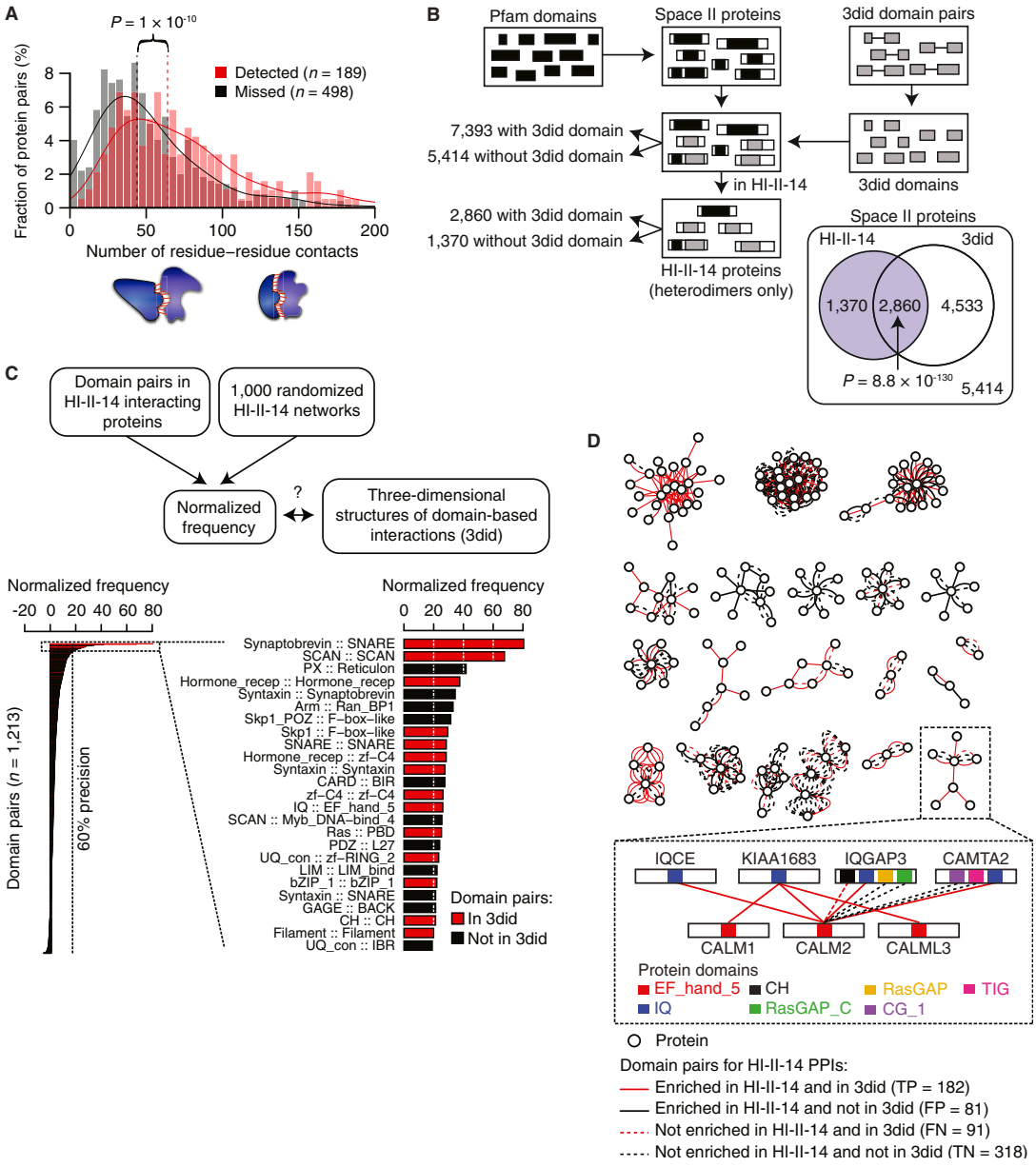


Figure S2. HI-II-14 Binary PPIs Represent Direct Contacts, Related to Figure 1

(A) Fraction of protein pairs in the PDB that HI-II-14 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). Among interacting proteins found in PDB complexes (Table S2D), the pairs reported in HI-II-14 tend to involve more reside-residue contacts than those that were missed (medians of 62 and 44 respectively, Table S2E).

(B) Enrichment in domains mediating protein-protein interactions in HI-II-14. Schematic showing the method to map Pfam domains onto proteins represented in the hORFeome 5.1 (Space II) and to measure the overlap with structurally supported PPI-mediating domains in the 3did database (top). Venn diagram showing for proteins in Space II, the overlap between proteins in HI-II-14 and proteins with a structurally supported PPI-mediating domain in the 3did database. The number of Space II proteins not in HI-II-14 and deprived of a domain in 3did is also mentioned at the bottom. 3did: database of three-dimensional interacting domains. p value based on two-sided Fisher's exact test.

(C) Predicted interacting domain pairs. Schematic of domain pair enrichment scoring method (top). Normalized frequency of all (bottom left; Table S2F) and the top 25 (bottom right) domain pairs. 3did: database of three-dimensional interacting domains.

(D) Examples of interacting domain pairs. Network showing all possible domain combinations between HI-II-14 protein pairs involving the top 25 enriched domain pairs (top). Example of an enriched domain pair corresponding to a known direct interaction (bottom). Pairs of Pfam domains appearing significantly more frequently across interacting proteins than expected by chance tend to correspond to pairs of domains known to mediate PPIs from crystallographic data (70% sensitivity and 80% specificity).

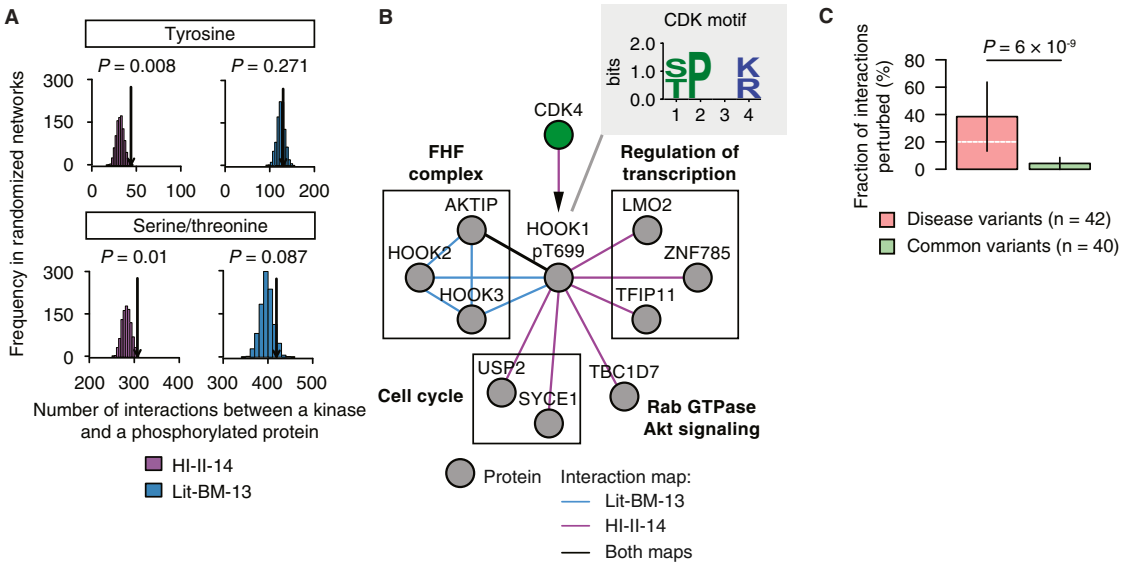


Figure S3. Overall Biological Significance of HI-II-14, Related to Figures 2 and 3

(A) Predicting kinase-substrate relationships. Number of interactions between kinases and phosphorylated proteins (arrows) compared to those in 1,000 degree-controlled randomized networks, classified according to the type of phosphosite. Empirical p values are shown.

(B) Putative phosphorylation of HOOK1 by CDK4. The [S/T]PX[K/R] motif is recognized by the CDK kinase family, and is found at the position where HOOK1 is phosphorylated on a threonine. Phosphorylation of HOOK1 by CDK4 might connect multiple cellular pathways involved in tumorigenesis.

(C) Interaction perturbation by disease and nondisease variants at least partly functional in yeast cells. Fraction of interactions of the wild-type protein lost by mutants bearing the disease-associated or common variants when only considering variants conserving at least one interaction. Error bars indicate standard error of the proportion. p value, two-sided Fisher's exact test.

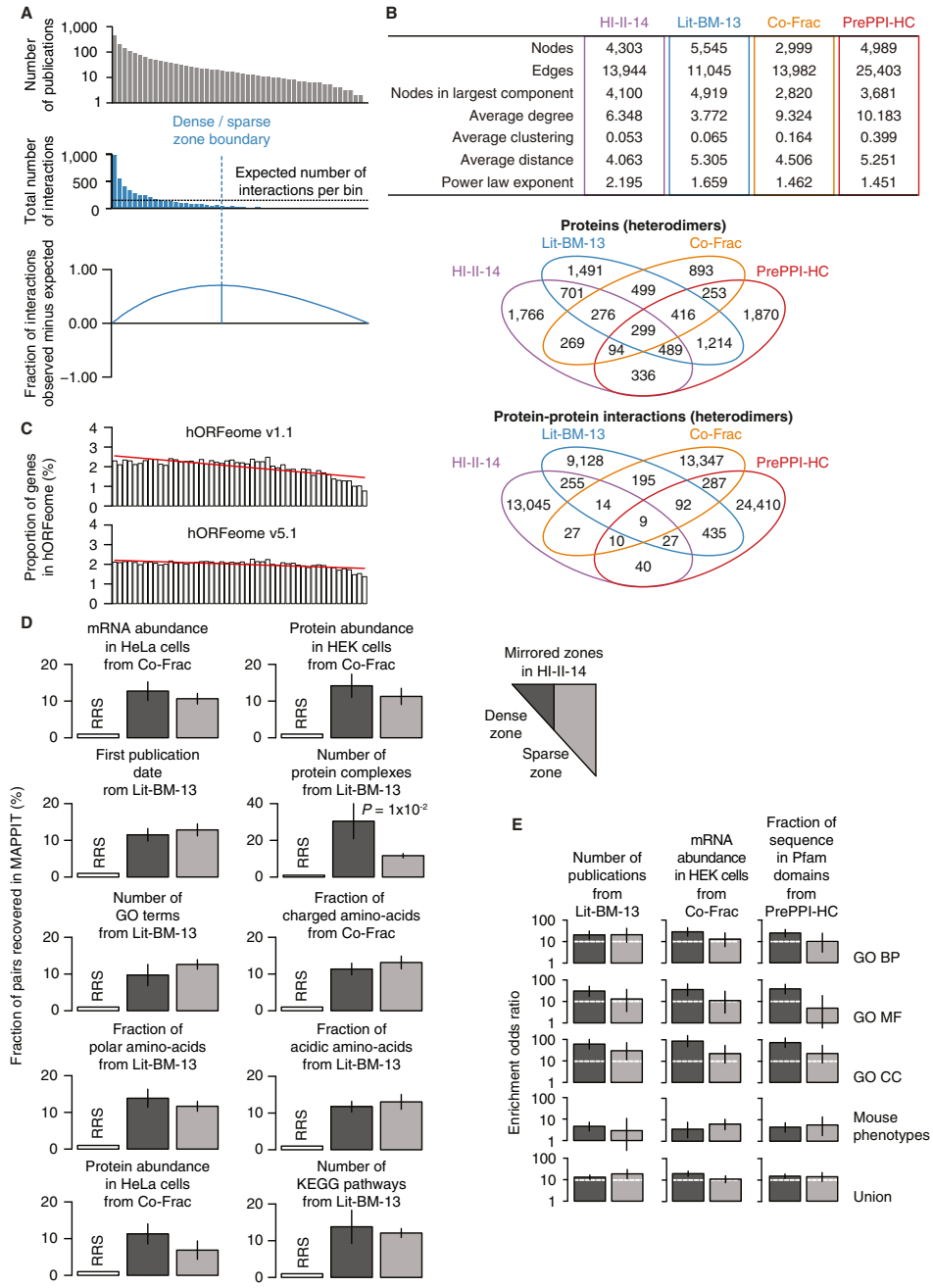


Figure S4. Comparison of Interaction Mapping Approaches, Related to Figure 5

(A) Defining the boundary between dense and sparse zones. Proteins are ordered by number of publications, and grouped in bins of ~350 proteins, with the median number of publications per bin shown (top). Histogram showing total number of interactions per bin (middle). Difference of observed and expected density of interactions along the full interactome landscape (bottom). The vertical line is drawn at the highest difference.

(B) Comparison of the four interactome network maps. Topological properties of the four interactome maps (top). Venn diagrams illustrate the number of shared proteins (middle) and interactions (bottom) between the four maps. Although statistically significant ($p < 1 \times 10^{-60}$, two-sided Fisher's exact test), the overlap of interactions between these four maps is small.

(C) Bias of hORFeome 1.1 toward well-studied genes. Proportion of hORFeome 1.1 (top) and 5.1 (bottom) genes present in bins of ~350 genes ordered by the number of publications per gene.

(D) Genuine biophysical interactions in apparent sparse zones. MAPPIT recovery rate of HI-II-14 pairs found in dense and sparse zones of other maps at 1% RRS recovery. Error bars indicate standard error of the proportion. p values based on two-sided Fisher's exact tests. For n values, see Table S6.

(E) Genuine functional interactions in apparent sparse zones. Functional enrichment of HI-II-14 pairs found in dense and sparse zones mirrored from Lit-BM-13, Co-Frac, and PrePPI-HC. Error bars indicate 95% confidence intervals. $p > 0.05$ for all pairwise comparisons of dense and sparse zones except for the Lit-BM-13 zones defined by the number of protein complexes, two-sided Fisher's exact tests. For n values, see Table S6.

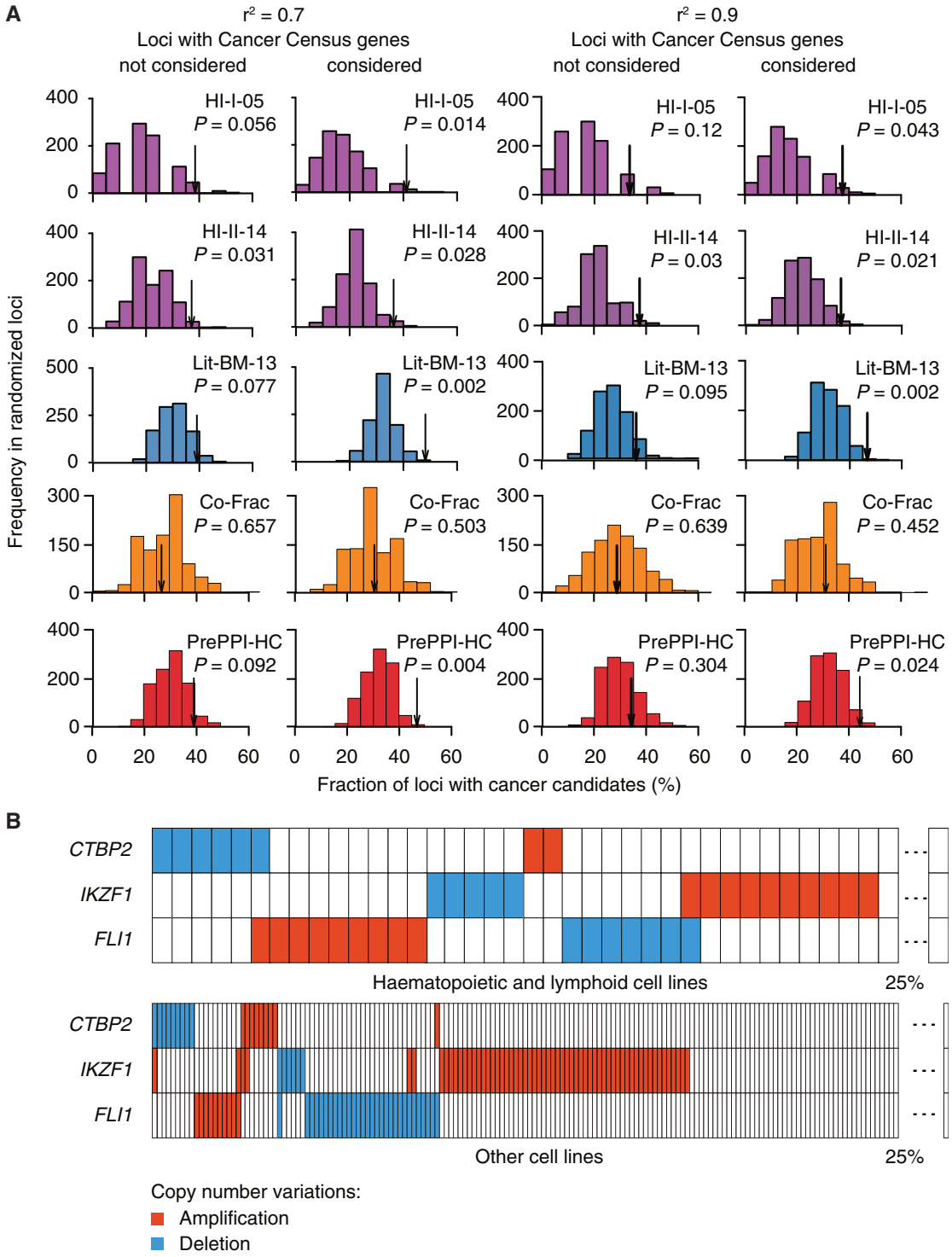


Figure S5. Prioritization of Cancer Candidates in GWAS Loci Using Interactome Maps, Related to Figure 7

(A) Robustness of cancer GWAS loci gene prioritization in interactome maps. Fraction of cancer-related GWAS loci encoding at least one protein that interacts with a Cancer Census protein (arrows) with respect to 1,000 sets of randomly selected loci genes (histograms) in HI-I-05, HI-II-14, Lit-BM-13, Co-Frac, and PrePPI-HC at different linkage-disequilibrium thresholds (r2). Empirical p values are shown. For n values, Table S6.

(B) *CTBP2*, *IKZF1*, and *FLI1* copy number variants in cancer cell lines. Copy number variants found in 25% of 163 hematopoietic and lymphoid and 717 other cell lines from CCLE are shown.

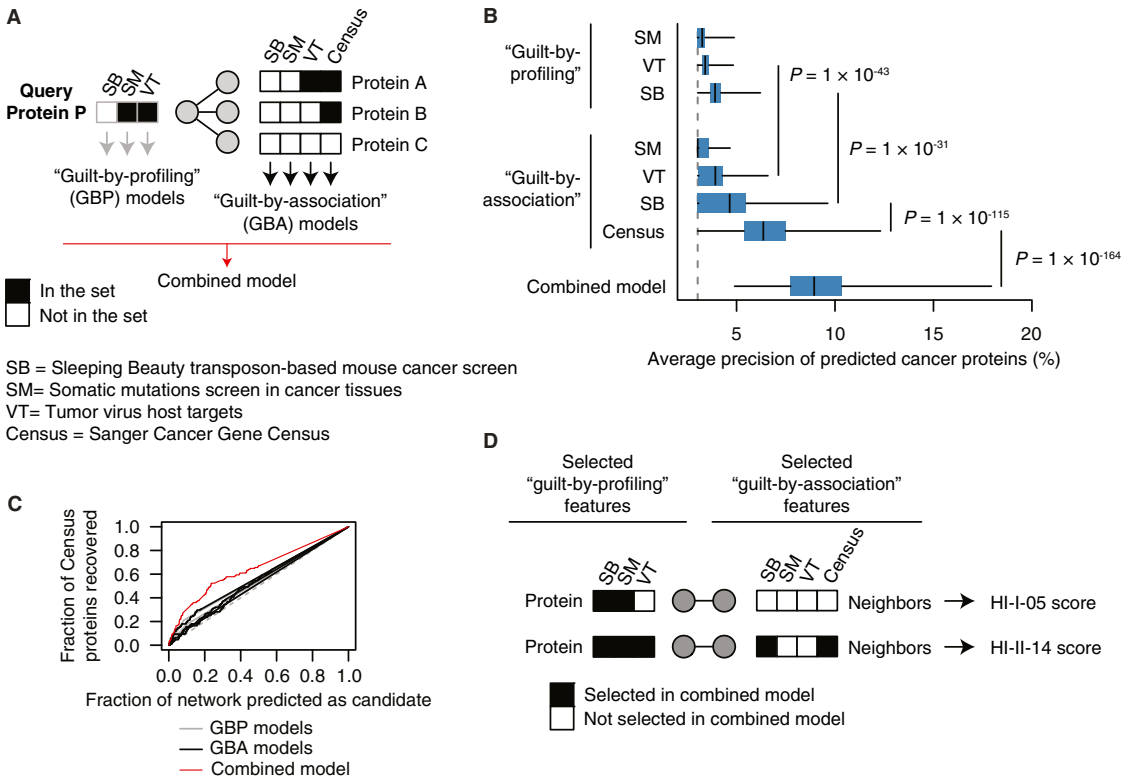


Figure S6. Prediction of Cancer Candidate Genes by Combining Guilt-by-Profiling and Guilt-by-Association Approaches, Related to Figure 7

(A) Cancer-association scoring system. SB, Potential candidate from Sleeping Beauty transposon-based mouse cancer screen; SM, Potential candidate from Somatic mutation screen in cancer tissues; VT, Potential candidate from virus target screens.

(B) Precision of regression models. Predictive ability of separate and combined models for cancer association scoring. p values are based on one-sided paired Wilcoxon rank sum tests.

(C) Predictive power of regression models. Receiver Operating Characteristic curve showing the fraction of Census proteins recovered at increasing score threshold compared to the fraction of proteins of the network predicted as candidates by each prediction model.

(D) Features selected in combined regression models using HI-I-05 or HI-II-14 interactions.

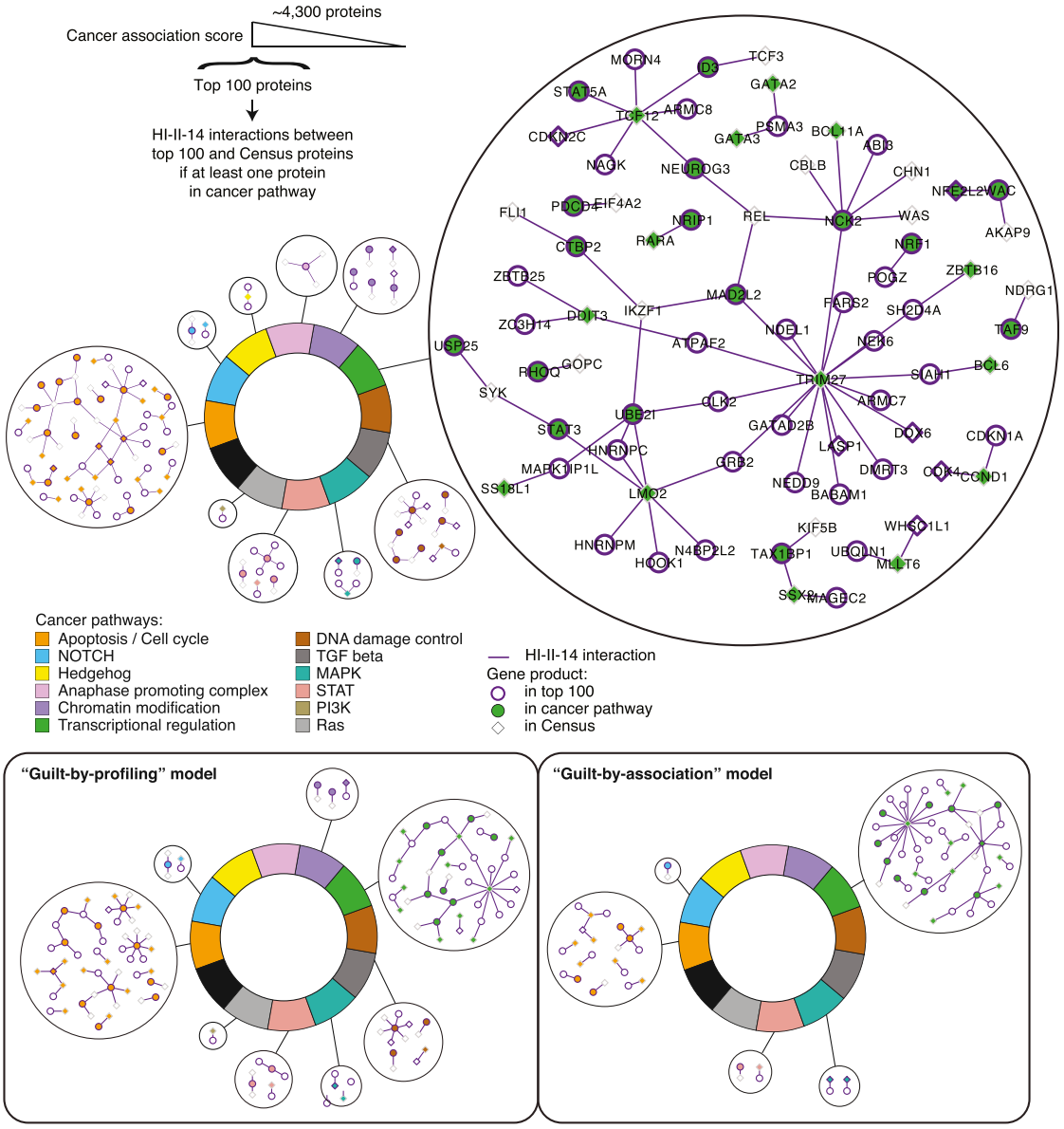


Figure S7. Expanded Cancer Landscape, Related to Figure 7

Top regression model candidates in the twelve pathways associated to cancer development and progression. Binary interactions from HI-II-14 involving the top candidates and Cancer Census gene products in the twelve pathways associated to cancer development and progression. Similar representations are given when using only guilt-by-profiling (bottom left) or guilt-by-association (bottom right) predictions.