Cell

Resource

A Census of Human Soluble Protein Complexes

Pierre C. Havugimana,[1,2,8] G. Traver Hart,[1,2,8] Tamás Nepusz,[4,8] Haixuan Yang,[4,8] Andrei L. Turinsky,[5] Zhihua Li,[6] Peggy I. Wang,[6] Daniel R. Boutz,[6] Vincent Fong,[1] Sadhna Phanse,[1] Mohan Babu,[1] Stephanie A. Craig,[6] Pingzhao Hu,[1] Cuihong Wan,[1] James Vlasblom,[2,5] Vaqaar-un-Nisa Dar,[7] Alexandr Bezginov,[7] Gregory W. Clark,[7] Gabriel C. Wu,[6] Shoshana J. Wodak,[2,3,5] Elisabeth R.M. Tillier,[7] Alberto Paccanaro,[4,*] Edward M. Marcotte,[6,*] and Andrew Emili[1,2,*]

[1]Banting and Best Department of Medical Research, Donnelly Centre for Cellular and Biomolecular Research

[2]Department of Molecular Genetics, Medical Sciences Building

[3]Department of Biochemistry, Medical Sciences Building

University of Toronto, Toronto, Ontario M5S 3E1, Canada

[4]Department of Computer Science, Royal Holloway, University of London, Egham TW20 0EX, UK

[5]Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M5G 1X8, Canada

[6]Center for Systems and Synthetic Biology, Institute for Cellular and Molecular Biology, Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, TX 78712, USA

[7]Campbell Family Institute for Cancer Research, Ontario Cancer Institute, University Health Network, University of Toronto, Toronto, Ontario M5G 1L7, Canada

[8]These authors contributed equally to this work

*Correspondence: alberto.paccanaro@cs.rhul.ac.uk (A.P.), marcotte@icmb.utexas.edu (E.M.M.), andrew.emili@utoronto.ca (A.E.)

http://dx.doi.org/10.1016/j.cell.2012.08.011

SUMMARY

Cellular processes often depend on stable physical associations between proteins. Despite recent progress, knowledge of the composition of human protein complexes remains limited. To close this gap, we applied an integrative global proteomic profiling approach, based on chromatographic separation of cultured human cell extracts into more than one thousand biochemical fractions that were subsequently analyzed by quantitative tandem mass spectrometry, to systematically identify a network of 13,993 high-confidence physical interactions among 3,006 stably associated soluble human proteins. Most of the 622 putative protein complexes we report are linked to core biological processes and encompass both candidate disease genes and unannotated proteins to inform on mechanism. Strikingly, whereas larger multiprotein assemblies tend to be more extensively annotated and evolutionarily conserved, human protein complexes with five or fewer subunits are far more likely to be functionally unannotated or restricted to vertebrates, suggesting more recent functional innovations.

INTRODUCTION

Protein complexes are stable macromolecular assemblies that perform many of the diverse biochemical activities essential to cell homeostasis, growth, and proliferation. Comprehensive characterization of the composition of multiprotein complexes in the subcellular compartments of model organisms like yeast, fly, worm, and bacteria have provided critical mechanistic insights into the global modular organization of conserved biological systems (Hartwell et al., 1999), accelerated functional annotation of uncharacterized proteins via guilt by association (Hu et al., 2009; Oliver, 2000), and facilitated understanding of both evolutionarily conserved and disease-related pathways (Vidal et al., 2011). How the ~20,000 or so proteins encoded by the human genome are partitioned into heteromeric ''protein machines'' remains an important but elusive research question, however, as less than one-fifth of all predicted human open reading frames are currently annotated as encoding subunits of protein complexes in public curation databases (Ruepp et al., 2010).

Loss-of-function mutations in genes encoding the subunits of protein complexes typically give rise to similar phenotypes or, through genetic interaction, amplify the phenotypic effects of other alleles in functionally linked sets of genes. Identifying the membership of protein complexes, therefore, addresses a crucial layer in the hierarchical functional organization of biological systems that links the core biochemistry of a functioning cell to the general physiology of an organism and is fundamental to deciphering the relationship between genotype and phenotype. Although bioinformatics analyses have been used to predict evolutionarily conserved human protein-protein interactions (PPIs) on a large scale (Ramani et al., 2008; Rhodes et al., 2005), most of these associations remain to be verified experimentally.

Affinity purification (AP) of tagged exogenous proteins coupled with tandem mass spectrometry (MS) is an effective method for isolating and characterizing the composition of stably associated human proteins in experiments ranging from dozens to hundreds of different ''baits'' (Behrends et al., 2010; Bouwmeester et al., 2004; Ewing et al., 2007; Hutchins et al., 2010; Jeronimo et al., 2007; Mak et al., 2010; Sardiu et al., 2008; Sowa et al., 2009). Likewise, immunoprecipitation can be used

to systematically isolate endogenous human protein complexes from human cell lines (Malovannaya et al., 2011). Nevertheless, the limited availability of high-quality antibodies or sequence-verified complementary DNA (cDNA) clones suitable for targeted protein complex enrichment precludes scale-up required for the unbiased assessment of the molecular association networks underlying human cells. Hence, despite considerable successes in the comprehensive identification of protein complexes in model organisms (Butland et al., 2005; Gavin et al., 2002, 2006; Guruharsha et al., 2011; Ho et al., 2002; Hu et al., 2009; Krogan et al., 2006; Kühner et al., 2009), clone-based protein purification techniques remain challenging for proteome-scale studies of physical interaction networks in mammalian cells. Conversely, although traditionally used to isolate discrete complexes with specific assayable biochemical properties (e.g., enzymatic activity), classical biochemical fractionation procedures have been used to resolve biological mixtures as a means of ascertaining the collective composition of human protein complexes present in certain subcellular compartments (Ramani et al., 2008; Wessels et al., 2009).

Here, we have combined extensive, scaled-up biochemical fractionation with in-depth, quantitative mass spectrometric profiling and stringent computational filtering to resolve and identify endogenous, soluble, stably associated human protein complexes present in cytoplasmic and nuclear extracts generated from cultured cells. Although the resulting reconstructed high-quality physical interaction network shows strong overlap with existing curated and experimentally derived sets of annotated protein complexes, it contains many predicted subunits and previously unreported complexes with specific functional, evolutionary, and disease-related biological attributes. To our knowledge, this resource represents the largest experimentally derived catalog to date of human protein complexes from cell culture, measured using a single standardized assay, and a reliable first draft reference of the basic physical wiring diagram of a human cell.

RESULTS

High-Throughput Complex Fractionation and Detection by Tandem Mass Spectrometry

To isolate human protein complexes in a sensitive and unbiased manner, we subjected cytoplasmic and nuclear soluble protein extracts isolated from human HeLa S3 and HEK293 cells grown as suspension and adherent cultures, respectively, to extensive complementary biochemical fractionation procedures. These two widely studied laboratory cell lines have been used as models of human cell biology for many decades (Graham et al., 1977; Masters, 2002), providing a rich biological context for interpreting the resulting proteomic data. Stably interacting proteins that cofractionated together were identified subsequently by nanoflow liquid chromatography-tandem mass spectrometry (LC-MS/MS). We optimized our entire experimental pipeline, illustrated schematically in Figure 1A, by using a multipronged strategy to minimize two major confounding issues: limited dynamic range (i.e., preferential detection of high-abundance components) and "chance" coelution (i.e., cofractionation of functionally unrelated proteins).

To address the former concern, we performed extremely deep biochemical fractionations by employing multiple orthogonal separation techniques to better resolve distinct protein complexes. As a primary separation technique, we employed nondenaturing high-performance multibed ion exchange chromatography (IEX-HPLC) by using four different empirically optimized analytical column combinations (see Experimental Procedures) and shallow salt gradients unlikely to perturb nonionic protein associations (Havugimana et al., 2007). In parallel, we applied complementary sucrose gradient centrifugation and isoelectric focusing technologies to capture salt-sensitive protein assemblies. In total, we collected 1,163 different fractions in a total of eight nuclear and five cytosolic extract fractionation experiments (for details see Table S1 available online), which were each subjected to label-free shotgun sequencing (duplicate LC-MS/MS analyses) using highly sensitive ion trap-based mass spectrometers (see Experimental Procedures).

We identified 5,584 distinct human proteins (Figure 1C; estimated theoretical false discovery rate of 1% at both the protein and peptide levels based on a statistical model [Kislinger et al., 2003]; see Experimental Procedures for details). Despite the underrepresentation of membrane proteins in the starting cell extracts, this coverage encompasses about half of the experimentally verified human proteome (Figure S1B) (Nagaraj et al., 2011). This included 989 proteins detected exclusively in nuclear fractions (of which 376 were annotated transcription or chromatin-related factors) and 1,006 with links to human disease (e.g., annotated in a public database like OMIM). Only 1,632 (29%) of the identified proteins had biochemical annotations as subunits of previously reported protein complexes (corresponding to 64% of all existing human protein entries) in the CORUM curation database (Figure S1C; Ruepp et al., 2010). Due to the extensive fractionation, we observed minimal bias in terms of protein abundance beyond that reported for previously annotated complexes or the experimentally defined human proteome (Figure 1D).

Next, to minimize the possibility of chance coelution, rather than simply identifying the proteins present in each fraction, we quantified variation in protein abundance based on the observed patterns of spectral counts recorded across all of the collected fractions to determine the extent to which pairs of proteins coeluted. As shown in Figure 1B, these experimental profiles were highly reproducible (i.e., average Spearman rank correlation coefficients >80% between replicate experiments; Figure S2), even using alternate methods of mass spectrometric quantification (i.e., extracted MS1 peak intensities were largely consistent with spectral counting; Figure S2D). To objectively evaluate the biochemical data, we calculated a stringent summary statistic, termed the coapex score, for each pair of proteins identified by LC-MS/MS by determining the number of fractionation experiments in which the proteins showed maximum (modal) abundance in the same exact peak fraction.

To assess the effectiveness of our cofractionation approach, we performed an initial validation by examining the coelution profiles and coapex scores obtained for a reference set of 20 well-known human protein complexes reported in CORUM. As illustrated by the representative HeLa nuclear extract IEX-HPLC profiles shown in Figure 1B, the subunits of these

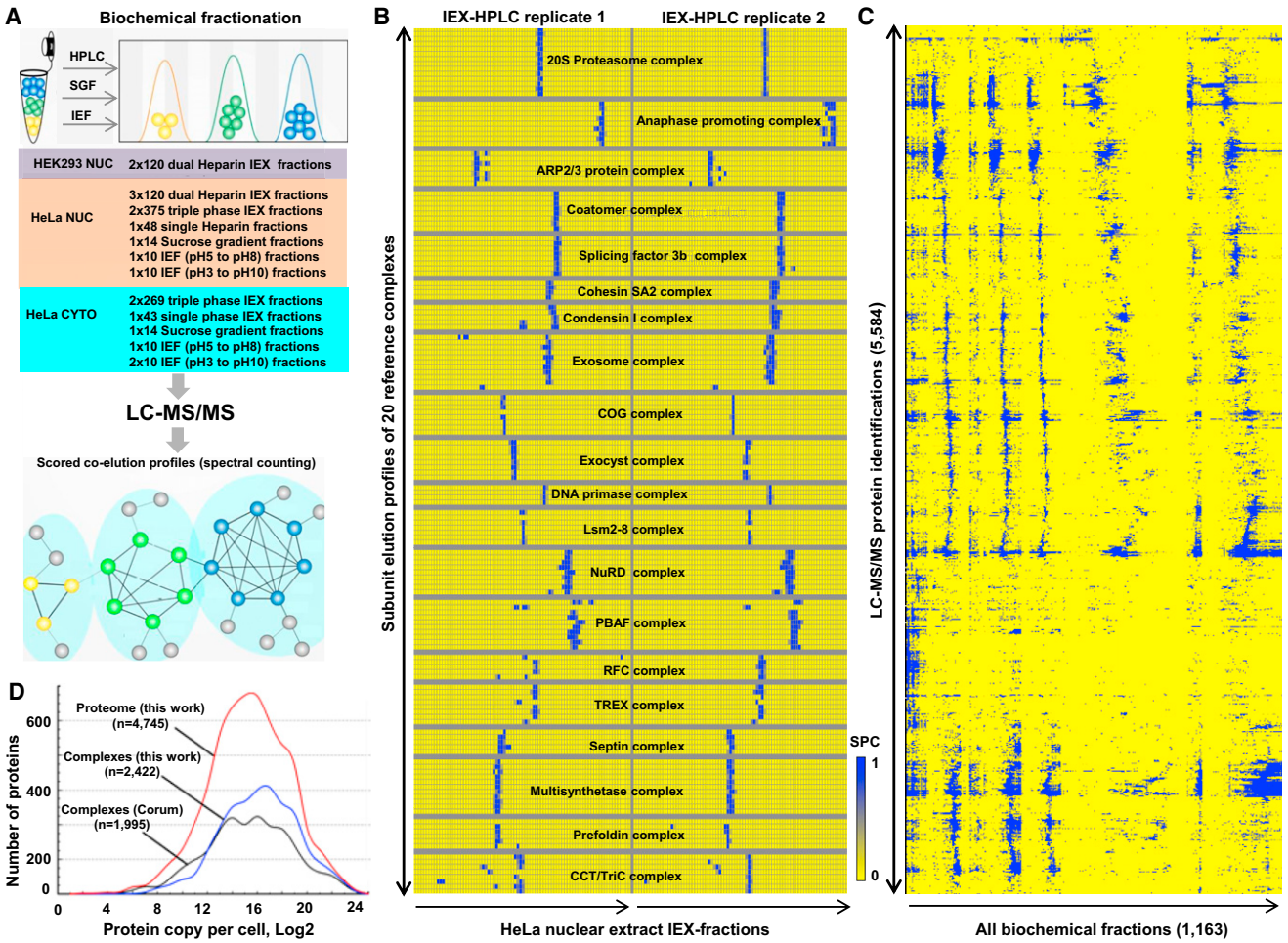


Figure 1. Integrative Cofractionation Strategy Used to Identify Human Soluble Protein Complexes

(A) Cell extracts were extensively fractionated using different biochemical techniques (IEX, ion exchange chromatography; IEF, isoelectric focusing; SGF, sucrose density gradient centrifugation). Coeluting proteins were identified by mass spectrometry, and a coelution network was generated by calculating profile similarity (see Extended Experimental Procedures).

(B) Cofractionation (IEX-HPLC) profiles of annotated subunits of 20 representative human protein complexes from HeLa nuclear extract. Shading indicates normalized spectral counts (SPC). Peak apex and adjacent peaks are shown.

(C) Hierarchical clustering of 5,584 proteins identified by LC-MS/MS.

(D) Protein abundance levels corresponding to components of our identified coeluting proteins (red line), reconstructed complexes (blue), or annotated CORUM complexes (black) estimated from the reported HeLa proteome (Nagaraj et al., 2011).

See also Figure S1 and Table S1.

complexes typically coeluted in the same biochemical fractions. Of the 155 components detected by mass spectrometry, most (85%; 499/585) of the detected subunit pairs of the reference complexes had high coapex similarity scores (i.e., coeluted together in at least two or more experiments), validating the overall efficacy of the fractionation procedures we used to isolate native protein complexes and the general correctness of the protein identification and quantification pipeline.

Reconstruction of a High-Confidence Cocomplex Interaction Network

Despite the consistency in coelution of annotated complex members, certain functionally distinct complexes occasionally exhibited overlapping chromatographic elution profiles (e.g., splicing factor 3b and Coatomer complexes; Figure 2A), presenting a potential source of spurious interactions. Although this artifact was minimized to a certain degree by performing multiple independent fractionation experiments, we used an integrative computational approach to further improve deconvolution (Figure 2B). Because physically interacting cocomplexed proteins often perform related biological functions (Alberts, 1998) and are often evolutionarily coconserved (Hartwell et al., 1999), we devised a machine learning procedure (Figure 2B; see Experimental Procedures for details) to score and select higher-confidence physical interactions based on both the experimentally measured coelution profiles and the existence of additional

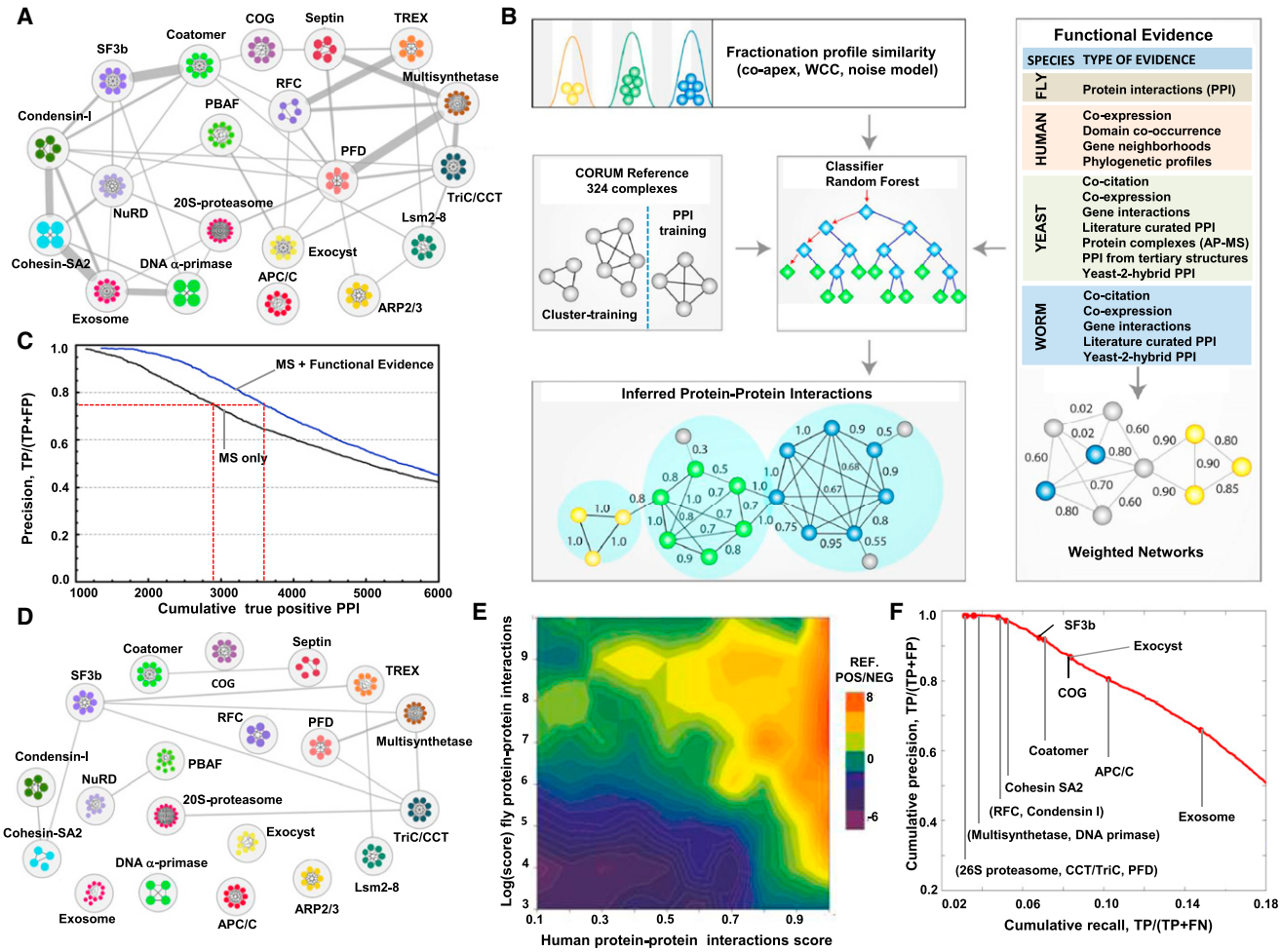


Figure 2. Denoising the Biochemical Coelution Network and Generation of High-Confidence Physical Interactions

(A) Biochemical cofractionation network of 20 reference complexes with coelution coapex scores ≥2. Nodes represent protein subunits (colors reflect complex membership), whereas edges represent interactions (thickness proportional to the number of shared coapexes).

(B) The biochemical data were combined with weighted functional association evidence by using a Random Forest classifier and a training set of reference complexes (CORUM) to filter out spurious connections and to infer a high-confidence interactome. The PPI and predicted clusters were evaluated with independent functional criteria to ensure high quality. Arrows represent data flow, blue diamonds are attributes in the decision tree vector, and green diamonds (leafs) are the final result (positive or negative).

(C) Cumulative precision-prediction rank curves for the LC-MS/MS data alone and after integration with genomic evidence. Incorporation of the functional evidence increased both precision (reduced false positives) and recall (more true positives).

(D) Network of 20 reference complexes after filtering with functional evidence.

(E) Overall correlation (Spearman r = 0.40; n = 11,675) of our scored human PPI with corresponding interaction scores reported for orthologous fly PPI from which validated, high-confidence complexes were derived (Guruharsha et al., 2011). Heatmap shows prediction accuracy (log ratio of CORUM reference positives to negatives), with high-scoring pairs in both studies highly enriched for positives.

(F) Precision-recall curve showing performance obtained after denoising reconstructing withheld reference CORUM complexes highlighted by red dots at the threshold at which half of the protein pairs per complex are recovered.

See also Figure S5 and Table S2.

supporting functional association evidence inferred from correlated evolutionary rates (Tillier and Charlebois, 2009) and functional genomics data sets compiled for *H. sapiens*, *S. cerevisiae*, *D. melanogaster*, and *C. elegans* (see Table S6 for details).

First, for each of the 13 fractionation experiments, we calculated correlation measures between all possible pairs of proteins to capture their tendency to coelute. In addition to the coapex summary statistic, to account for mass spectrometry sampling error, we devised a weighted cross-correlation function to account for slight variation in the protein profiles measured in each experiment. To account for low spectral values, we also generated a Poisson noise model before calculating Pearson correlation scores, deeming the coelution profiles of protein pairs measured with low spectral counts as less predictive of genuine physical interactions (Figure S5). Only protein pairs

with a correlation score of at least 0.5 by at least one of these measures in one or more experiments were considered for further analysis, reducing the total number of pairs from over 15 million initially to the roughly 800,000 pairs with reasonable biochemical evidence.

To improve the assignment of interaction probabilities, we also exploited the predictive power of correlated protein evolutionary rates (Tillier and Charlebois, 2009), messenger RNA (mRNA) coexpression, and domain co-occurrence and, via orthology, fly protein-protein interactions (based on binary yeast two-hybrid assay studies) and extensive physical and functional associations reported previously for yeast and worm (see Experimental Procedures) (Lee et al., 2011). The discriminatory power of the procedure was further improved by penalizing those interactions that lacked independent supporting evidence—and that were thus more likely to correspond to cases of ''chance'' coelution—by integrating evidence from these functional association data (Figure 2B). A feature selection algorithm was used to select the most informative data sets (Table S2) in addition to the biochemical correlation scores, and the resulting features were used to estimate the probability of interaction to protein pairs using a cross-validated random forest classifier.

For training, we used the CORUM curated set of human protein complexes as our base reference, filtered for those complexes annotated as having been observed by biochemical methods. As many CORUM complexes are highly overlapping due to redundancy in existing annotations, we combined complexes sharing subunits (Simpson coefficient >0.5 between complexes). We used half of the resulting 324 nonredundant reference complexes (Table S3) as the training set for cocomplex probability prediction, defining gold standard positive interactions as pairs of proteins in the same complex and inferring gold standard negatives between proteins in different complexes. (The other half of the reference complexes was withheld for subsequent use as an independent training set for cluster optimization, as described below.)

Although the biochemical data were a prerequisite for scoring, the performance curves shown in Figure 2C indicate that the inclusion of the additional functional genomic information substantially increased recall at the same level of precision compared to classifiers based on the profiling data alone. Moreover, the integration of this additional supporting functional evidence removed the bulk of spurious, intercomplex interactions (Figure 2D). Another advantage of our bioinformatic pipeline is that the results of the feature selection algorithm (Table S2) can be explored to examine the impact of each data set. For example, we find generally that sets of smaller biochemical fractionations using different separation techniques, although individually yielding a higher PPI false discovery rate, collectively provided more information on protein complex composition than deeper fractionations using a single separation method.

As an alternate measure of reliability, we compared our scored human protein interactions to a recently reported network of *Drosophila* cocomplex protein interactions (Guruharsha et al., 2011), which had not been used for building the classifier. Strikingly, despite using vastly different experimental methods and scoring schemes, we observed a remarkably good overall correlation (Spearman r = 0.40; n = 11,675 orthologs mapped using Inparanoid). Even after removing interactions supported by alternate *Drosophila* data, high-scoring fly pairs matched high-scoring pairs in our analysis and were strongly enriched for reference-positive cocomplex members (Figure 2E).

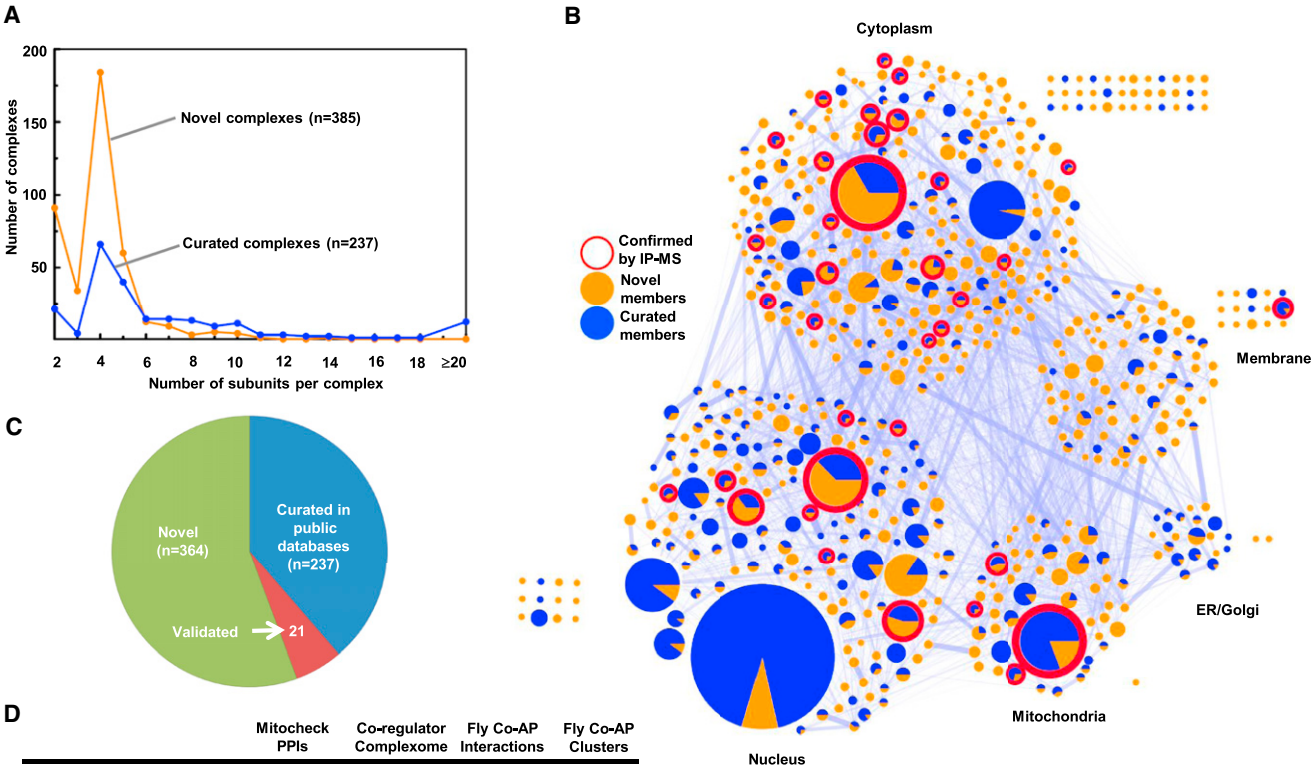
Finally, in order to remove any remaining false positive interactions, we further denoised our cocomplex data set by pruning loosely connected interactions using a computational diffusion procedure calibrated by protein colocalization semantic similarity scores (Pesquita et al., 2009; Yang et al., 2012) to enforce local network topologies more consistent with annotated complexes from the withheld portion of the reference Corum complexes (see Experimental Procedures). Benchmark precision and recall versus the holdout set of known reference complexes (Figure 2F) were significantly higher than those reported for a smaller, recently published set of affinity-purified human protein complexes (Hutchins et al., 2010), validating the reliability of our scoring procedure.

Applying a PPI score threshold of 0.75, which corresponds to an estimated false discovery rate of 21.5% (i.e., well below the ~40% reported for AP/MS-based analyses of protein complexes in model organisms [Gavin et al., 2006; Krogan et al., 2006; Kühner et al., 2009]), we thus derived a high-confidence set of 13,993 cocomplex interactions among 3,006 unique human proteins (Table S2), most of which (8,691 PPI) have not been reported before (i.e., are not publicly annotated). It is worth reiterating that all of these physical interactions were directly supported by the experimental biochemical cofractionation data; the addition of functional data and denoising served only to flag candidates lacking either functional support or topological support within the network (Table S2). The interaction probability scores may be underestimated, however, because the reference ''gold standards'' used for learning are imperfect (Jansen and Gerstein, 2004).

Construction and Validation of Protein Complexes from the Probabilistic Interaction Network

In order to define complex membership, we partitioned the high-confidence probabilistic physical interaction network by using the cluster growth algorithm ClusterONE (Nepusz et al., 2012), which outperformed other clustering methods on the denoised PPI network (Table S5). In total, the clustering predicts 622 discrete putative complexes encompassing 2,634 distinct proteins (Table S3). Complex membership size distribution approximated an inverse power law with a median of four subunits (Figure S4A). The majority (62%; 385/622) of the complexes have not been annotated (i.e., only 237 are currently curated in a public database like CORUM; Figures 3A and 3C). Although the fraction of curated components varies, we also recapitulated 258 previously reported complexes (Figure 3C), including several well-known membrane-associated complexes, such as the coat protein I and II (COPI/II) vesicle transport complexes that shuttle cargo between the Golgi and endoplasmic reticulum. Strikingly, most (67%; 335) of the 500 smaller putative complexes with five or fewer components, including the bulk (74%; 83) of the 112 predicted heterodimers, have never been curated before (Figure 3C).

Both independent experimental validation based on more traditional immunoprecipitation or coaffinity purification methods



	Mitocheck PPIs	Co-regulator Complexome	Fly Co-AP Interactions	Fly Co-AP Clusters
Annotated edges	281	13,663	8,688	8,688
Edges sharing annotation	131	2,873	1,549	536
Fold-enrichment over chance	4	477	9	87
Hypergeometric p-value	*1.69E-52*	*<E-308*	*<E-308*	*<E-308*

Figure 3. Global Validations of the Map of High-Confidence Human Protein Complexes

(A) Complex size distribution of the 622 inferred complexes.

(B) Network of predicted human protein complexes proportioned according to subunit number and displaying existing curations, validation status by AP/MS (Malovannaya et al., 2011), and PPI connectivity (proportioned edge width).

(C) Proportions of annotated complexes in public repositories (CORUM, PINdb, REACTOME, and HPRD) or independently experimentally verified.

(D) Enrichment analysis showing overlap with large-scale APMS data sets generated for human (Hutchins et al., 2010; Malovannaya et al., 2011) and (via orthology) fly (Guruharsha et al., 2011).

See also Table S3.

and orthology mapping support at least 21 of these putative complexes (i.e., not in any reference database) (Table S3; see Supplemental Information for details). For example, Guruharsha et al. (2011) recently reported 299 cocomplex interactions based on pull-down experiments of 43 affinity-tagged human proteins present in 41 of our complexes, of which 143 interactions map precisely to our predicted complexes, representing a 47.8% validation rate (which may be an underestimate, as Guruharsha et al. [2011] do not report human interactions that fall outside the fly interologs examined in their study). Likewise, the results of Malovannaya et al. (2011), who used large-scale immunoprecipitation to isolate native human protein complexes, show excellent agreement to 123 of our complexes (i.e., Benjamini-corrected hypergeometric $p \leq 0.05$), including 42 (34%) of our complexes that are not curated in CORUM (Figure 3B and Table S3). Figure 3D summarizes the highly significant overlap of our inferred complexes with these fully independent data sets, with enrichments ranging from 4- to 477-fold more than chance, thus broadly and systematically validating our network of derived human protein complexes.

By design, insoluble membrane-associated (hydrophobic) protein complexes were largely missed in this study, and the proteins assigned to complexes had a higher average transcript abundance (Figures S2A and S2B). Moreover, in an effort to control the false positive rate, our conservative clustering algorithm, ClusterONE, underweighted small clusters of size 2 or 3 for lack of sufficient association evidence, likely contributing to the prominence of complexes with four subunits in Figure 3A. But we did not observe any significant bias toward negative ($pI \leq 7$) or positive ($pI \geq 7$) charge as compared to complexes curated in CORUM (Figure S4B).

Figure 4 shows the broad functional diversity of the predicted complexes (a navigable map is available online for close visualization of individual clusters and their supporting cocomplex interactions). Consistent with biological expectation (Hartwell et al., 1999; Lage et al., 2007; Oliver, 2000; Vidal et al., 2011),

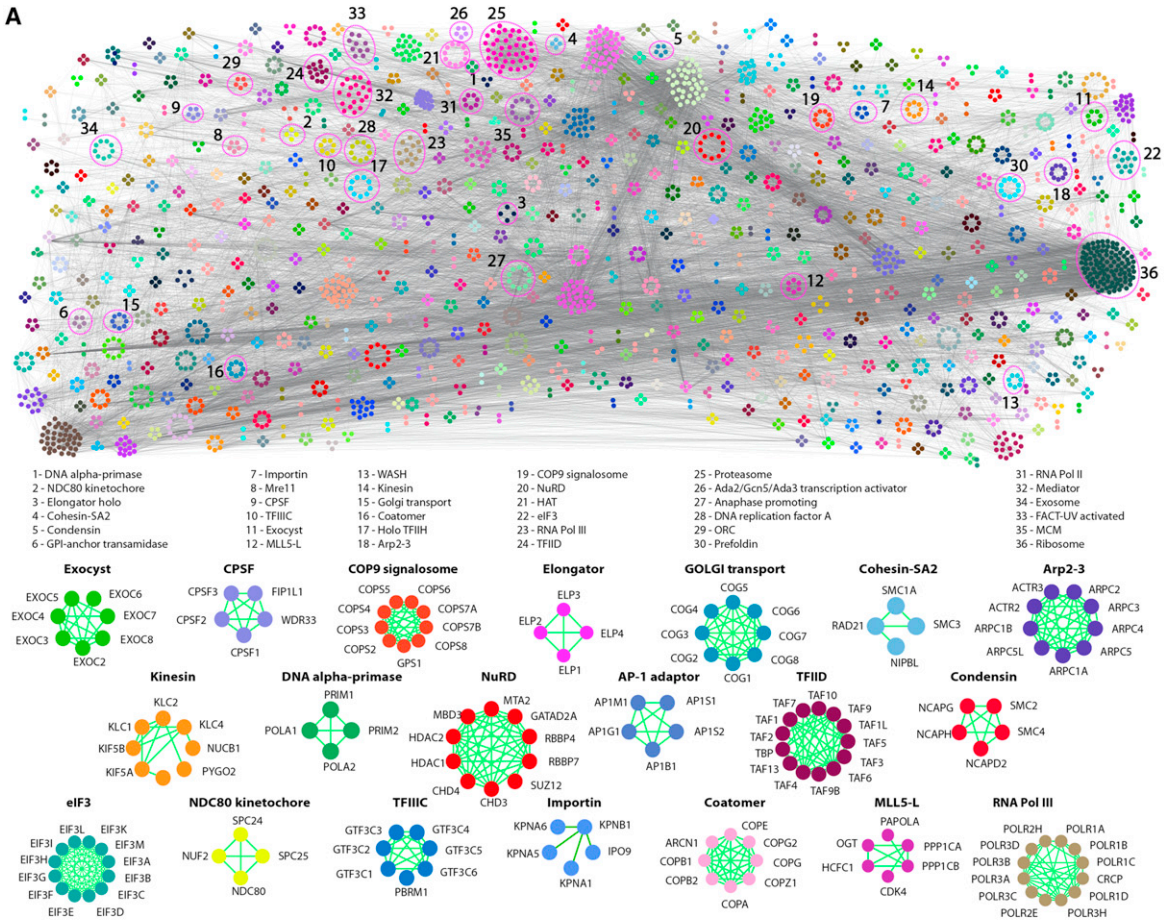


Figure 4. Global Map of High-Confidence Human Protein Complexes

(A) Schematic of the global network of inferred human soluble protein complexes (colored by membership) with representative examples and supporting PPI highlighted.

(B) Putative complexes with two or more components with human disorder associations annotated in UniProt (UniProt Consortium, 2011), Online Inheritance of Man (OMIM) (Hamosh et al., 2005), or the Genetic Association Database (GAD) (Becker et al., 2004). Inset table shows highly significant interaction overlap

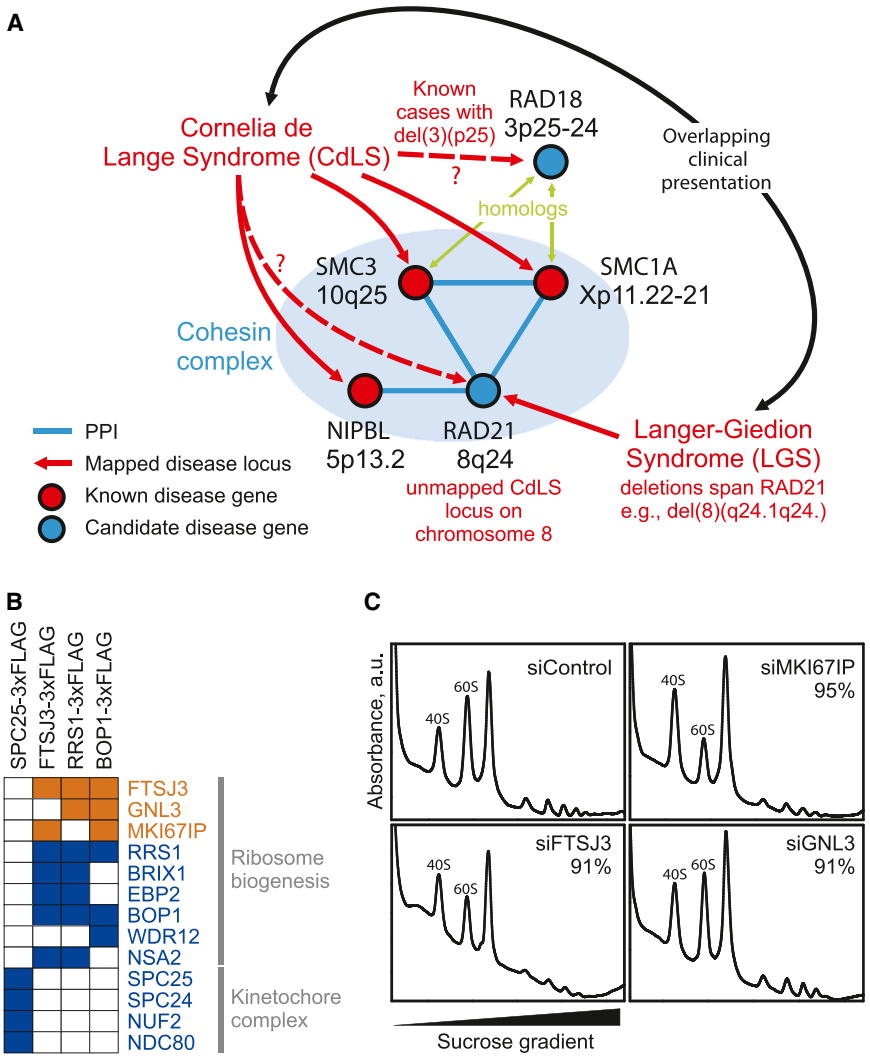


Figure 5. Membership in Complexes Predicts Protein Function and Disease Associations

(A) Three of four proteins mapped to the cohesin complex account for roughly half of cases of the human congenital disorder Cornelia de Lange syndrome (Pié et al., 2010), implicating the fourth component, RAD21, as a candidate disease gene. This association may explain similarities in clinical presentation between CdLS and Langer-Giedion syndrome, as the latter patients routinely harbor RAD21 deletions, e.g., McBrien et al. (2008) and Wuyts et al. (2002).

(B) Confirmation of ribosome biogenesis candidate (orange) associations with annotated components (blue) by AP/MS analysis of tagged proteins (top). Colored squares indicate validation (see Extended Experimental Procedures).

(C) Polysome profiling after siRNA targeting in tissue culture supports functional roles in ribosome biogenesis for three candidate proteins. Knockdown of MKI67IP, FTSJ3, and, to a lesser extent, GNL3, results in 60S ribosomal subunit biogenesis defects manifested by a reduced ratio of free 60S to 40S ribosomal subunits during gradient sedimentation as compared to control. Percentages indicate siRNA knockdown efficiency as measured by qRT-PCR.

the subunits of the complexes were significantly enriched for related biological functions, transcriptional regulatory motifs, and pathological processes (Figure 4B, inset table). Compared to the entire set of identified proteins, the clustered proteins also showed enrichment for posttranslation modifications linked to cellular regulation, like acetylation (Benjamini-corrected $p \leq 10^{-41}$) and phosphorylation ($p \leq 10^{-5}$). Many of the complexes are linked to core cellular processes, such as mRNA splicing ($p \leq 10^{-15}$) or transcription ($p \leq 10^{-5}$), that either are essential in human ($p \leq 10^{-138}$) or that have RNA interference (RNAi)-induced phenotype in cell culture (e.g., cell division arrest, $p \leq 10^{-31}$) or are associated, via orthology, with similar mouse, yeast, or worm mutant phenotypes (Figure 4B, inset table; see Table S4 for details).

Clinical and Biological Implications of the Reconstructed Human Protein Complexes

Consistent with this strong tendency for proteins in the same complex to be affiliated with similar mutational and RNAi phenotypes, subunits of the predicted human protein complexes were much more likely than chance ($p \leq 10^{-46}$) to have links to a documented clinical pathology (Figure 4B; see Table S4 for details), with disease-associated proteins distributed broadly among the complexes (Figures 4B and S4C). Closer examination of the interaction subnetworks—comprising known human disease genes with genes that currently lack annotation or that have not previously been associated with any human disorders (Figure 4B)—highlights the utility of the map.

One such example is shown in Figure 5A, illustrating the case of the human developmental disorder Cornelia de Lange syndrome (CdLS). Mutations in three subunits of the cohesin complex (SMC1A, SMC3, and NIPBL) have been linked to CdLS (Pié et al., 2010), implicating an additional component (RAD21) as a candidate CdLS locus, and are consistent with at least one unmapped CdLS locus residing on chromosome 8 (DeScipio et al., 2005). The link to RAD21 provides a likely

(i.e., shared annotated edges) with phenotypic data sets that reveals that protein subunits of the same predicted human complex tend to exhibit similar disease and genetic associations in human populations (see Extended Experimental Procedures), RNAi phenotypes in cell culture (Neumann et al., 2010), mutational and RNAi phenotypes in other species (via orthology), and shared transcriptional regulatory motifs (Xie et al., 2005).
See also Figure S4C and Table S4.

explanation for the occasional overlap of Langer-Giedion syndrome (LGS) clinical presentation with CdLS, as all LGS patients are at least partially defective for RAD21 (see e.g., McBrien et al., 2008; Wuyts et al., 2002). Similarly, RAD18, a homolog of SMC3 and SMC1A, may play a role in CdLS that is consistent with unmapped CdLS deletions within chromosome 3p25 (DeScipio et al., 2005). Reports coinciding with the preparation of this manuscript confirm that RAD21 mutations do indeed lead to a CdLS-like syndrome (Deardorff et al., 2012), supporting the use of the complex map to prioritize promising candidate genes for human diseases.

Similarly, participation in the same complex suggests shared functions; the map can thus be used to predict new biochemical functions for proteins and other types of functions. We experimentally validated one such case for a ribosome-associated subcomplex containing BOP1, RRS1, GNL3, EBP2, FTSJ3, and MK1671P, and we first confirmed the interactions by affinity tagging/purification and mass spectrometry (Figure 5B). BOP1, EBP2, and the yeast ortholog of RRS1 are known to participate in maturation of the large 60S ribosomal subunit, suggesting that the other factors likewise engage in ribosome assembly, which is consistent with the nucleolar localizations of GNL3, FTSJ3, and MKI67IP. Supporting a role in ribosome biogenesis, short interfering RNA knockdowns of FTSJ3, MKI67IP, and, to a lesser extent, GNL3, perturbed 60S formation in cell culture, decreasing the ratio of free 60S to 40S subunits (Figure 5C). Taken together, these data support roles in ribosome biogenesis for these proteins and confirm the utility of the map for identifying biological functions.

Conservation of Human Protein Complexes

Estimates based on sequence similarity across orthologs indicate that the components of the complexes we detect are generally more ancient and have higher conservation on average than most human proteins (Figure 6A; see Table S3 for details). Using orthology relationships derived from well-established sources and calculating evolutionary rates and ages for all human proteins as a base distribution for gauging the emergence of complexes (see Extended Experimental Procedures), we found that many complexes appear to be quite ancient and slowly evolving (Figure 6B). Strikingly, however, most (60%; 376/622) human complexes likely arose with vertebrates, i.e., orthologs not present in invertebrates or fungi (Table S3). Hence, our analyses suggest a major shift/expansion in the ancestral protein interaction network coincident with the emergence of vertebrates.

Given the availability of experimentally derived networks of fly and yeast protein complexes, we could directly examine the evolutionary conservation of protein complexes across animals by comparing our network of human complexes with the extensive maps of 556 fly protein complexes recently reported for *D. melanogaster* (Guruharsha et al., 2011) and 720 yeast protein complexes documented for *S. cerevisiae* (Babu et al., 2012). Roughly one quarter (24%; 149/622) of the predicted human protein complexes showed statistically significant overlaps with complexes reported for these models (Figure 6B, inset; see Table S3 for details), with half of the subunits having clear orthologs (Figure 6C); the remaining components presumably represent genuine differences or incomplete orthology annotations.

The functional significance of unannotated ancestral human complexes supported by conservation in yeast or fly (Table S3 and Figure 6) warrants further investigations. At least one such complex, a multisubunit transfer RNA (tRNA)-splicing ligase (Popow et al., 2011), was characterized recently. The interaction between DDX1 and C14orf166 was detected at high confidence both in our data set (probability score 0.899) and in the Guruharsha et al. (2011) fly cocomplex data, and the other respective associated complex subunits likewise show significant overlap (Benjamini-corrected p value 1.1×10^{-7}). Additional examples of complex conservation are similarly supported by independent experimental evidence, e.g., such as the matching tissue specificities of the putatively interacting proteins endoplasmin and glucosidase 2β (Figure 6D), which form an uncharacterized complex conserved in both the fly and human maps.

Functional enrichment analysis of ancient complexes in comparison to vertebrate-specific ones also reveals intriguing biological trends. For example, we expected ancient, core cellular functions to be depleted among vertebrate-specific complexes. Consistent with this expectation, we find proteins associated with the ribosome ($p \leq 10^{-67}$, 113 proteins) and RNA polymerase II ($p \leq 10^{-27}$, 45 proteins) to be highly enriched only among conserved complexes. However, we also observe several notable variations from this hypothesis. For example, compared to the genomic background, mitochondrial proteins are more highly enriched among proteins assigned to vertebrate complexes than among those assigned to conserved complexes; 159 vertebrate proteins have a mitochondrial Gene Ontology Biological Process (GO BP) annotation ($p \leq 10^{-31}$) versus only 81 proteins assigned to conserved complexes ($p \leq 10^{-5}$). Similarly, proteins annotated as being part of the splicing apparatus are enriched in both conserved ($p \leq 10^{-33}$; 63 proteins) and vertebrate complexes ($p \leq 10^{-11}$, 43 proteins), which is consistent with an ancient function gaining additional complexity in vertebrates (e.g., increased alternative splicing). Our study therefore offers a unique perspective into the functional conservation and diversification of protein complexes across animals.

Protein Abundance, Ubiquity, and Complex Subunit Stoichiometries

Consistent with the documented origins of the HeLa and HEK293 cells analyzed in this study, the complexes we identified were significantly enriched for epithelial markers ($p \leq 10^{-183}$; UniProt tissue annotations). Explicit comparison of results across the two cell lines used in this study provided little evidence for tissue-specific or cell-type-specific complexes (see Supplemental Information). Most proteins were detected in both cell line fractionations, which is consistent with the similar protein and mRNA expression patterns observed in these cell lines (Figure S1), whereas the few proteins detected uniquely in one cell line or the other did not preferentially assort into tissue-specific complexes (Figure S2). The vast majority of complex components are universally expressed in 11 cancer cell lines (Geiger et al., 2012) (Figure S3A) and show high and largely invariant expression in an mRNA sequencing (mRNA-seq) study of 16 normal human tissues (EBI accession number

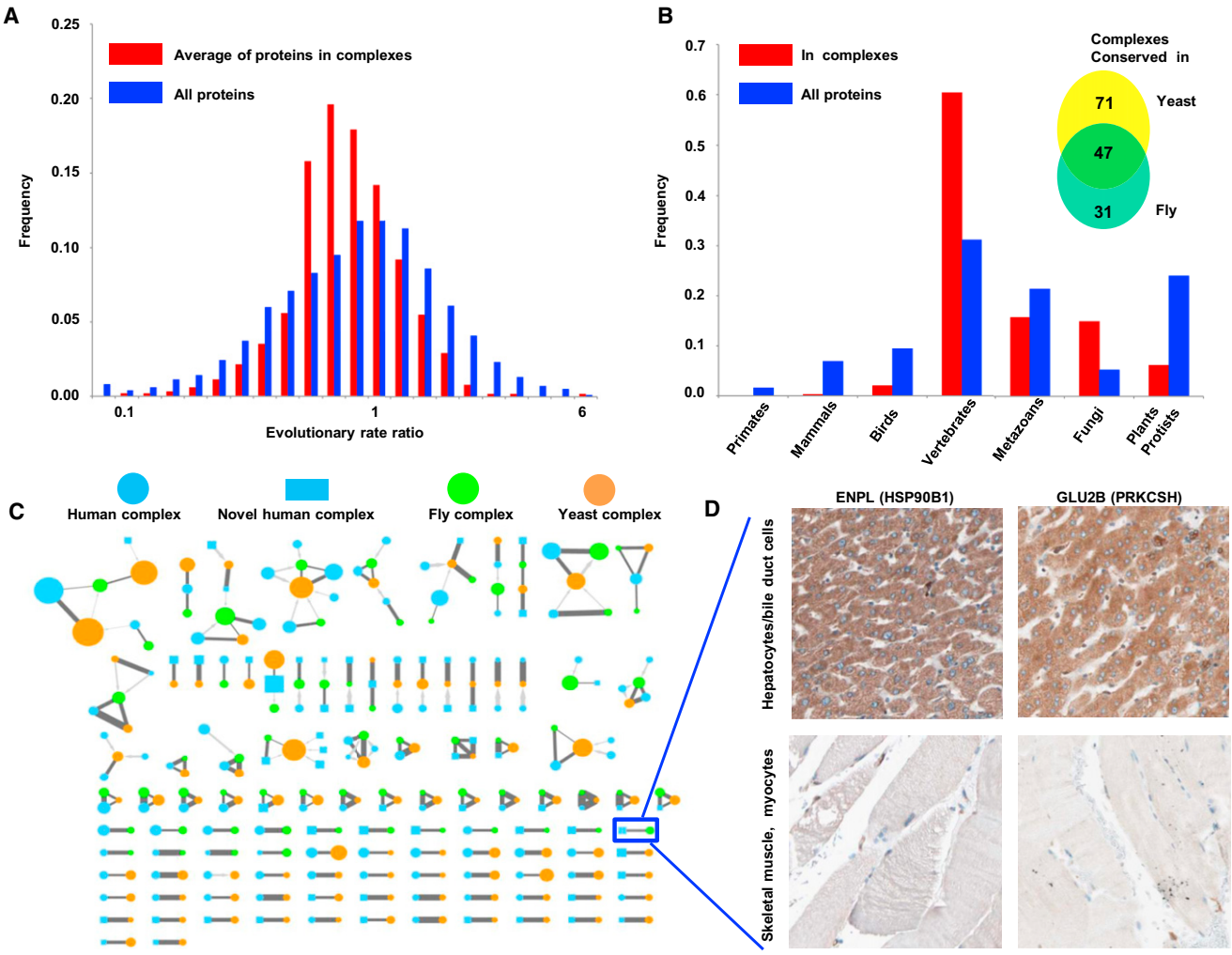


Figure 6. Evolutionary Conservation of Protein Complexes

(A) Components of predicted human complexes—calculated as the average of evolutionary rate ratios—evolved more slowly, as compared to the entire set of expressed proteins (see Extended Experimental Procedures).

(B) Pronounced spike in number of complexes originated with the emergence of vertebrates. x axis shows increasingly inclusive orthologous groups in the phylogeny of eukaryotes.

(C) Human complexes conserved in fly (Guruharsha et al., 2011) and yeast (Babu et al., 2012) (see Table S3 and Extended Experimental Procedures). Nodes represent complexes (human, blue; fly, green; yeast, orange), with size proportional to subunit number. Reciprocal best matches shown as dark gray edges, and nonreciprocal is shown as lighter gray directed edges, with edge thickness proportional to Sorensen-Dice overlap of complex members. Human complexes absent from public databases (putative complexes) are drawn as rectangles, and the remaining are drawn as circles.

(D) Similar tissue-specific expression patterns support a functional association between interacting proteins ENPL and GLU2B, whose orthologs were reported to interact in fly (Guruharsha et al., 2011). Panels show representative antibody staining in normal tissue biopsies collected and reported by the Human Protein Atlas (Uhlen et al., 2010) (www.proteinatlas.org).

See also Figure S3 and Table S3.

E-MTAB-513) (Figure S3B). Indeed, complex subunits are considered near ubiquitous ($p \leq 10^{-11}$; protein information resource [PIR] tissue specificity annotations) and are expressed in the top quartiles of 1,045 of 7,067 neoplastic and normal tissue CGAP EST libraries (1% false discovery rate [FDR]), including normal kidney ($p \leq 10^{-39}$), muscle ($p \leq 10^{-20}$), liver ($p \leq 10^{-12}$), brain ($p \leq 10^{-20}$), vascular ($p \leq 10^{-30}$), bone ($p \leq 10^{-15}$), and embryonic tissue ($p \leq 10^{-31}$). Consistent with this, genes encoding complex subunits also tend to share common upstream transcriptional regulatory motifs ($p \leq 10^{-8}$) (Figure 4B, inset table). Proteins mapped to complexes showed no major bias in abundance over the complete set of human proteins identified by mass spectrometry (Figure 1D).

The pervasiveness of ubiquitously expressed protein complexes argues strongly for broad relevance to basic human cell biology. Although often coexpressed, the subunit stoichiometries of human protein complexes in vivo are largely unknown and have never been systematically measured globally. Because

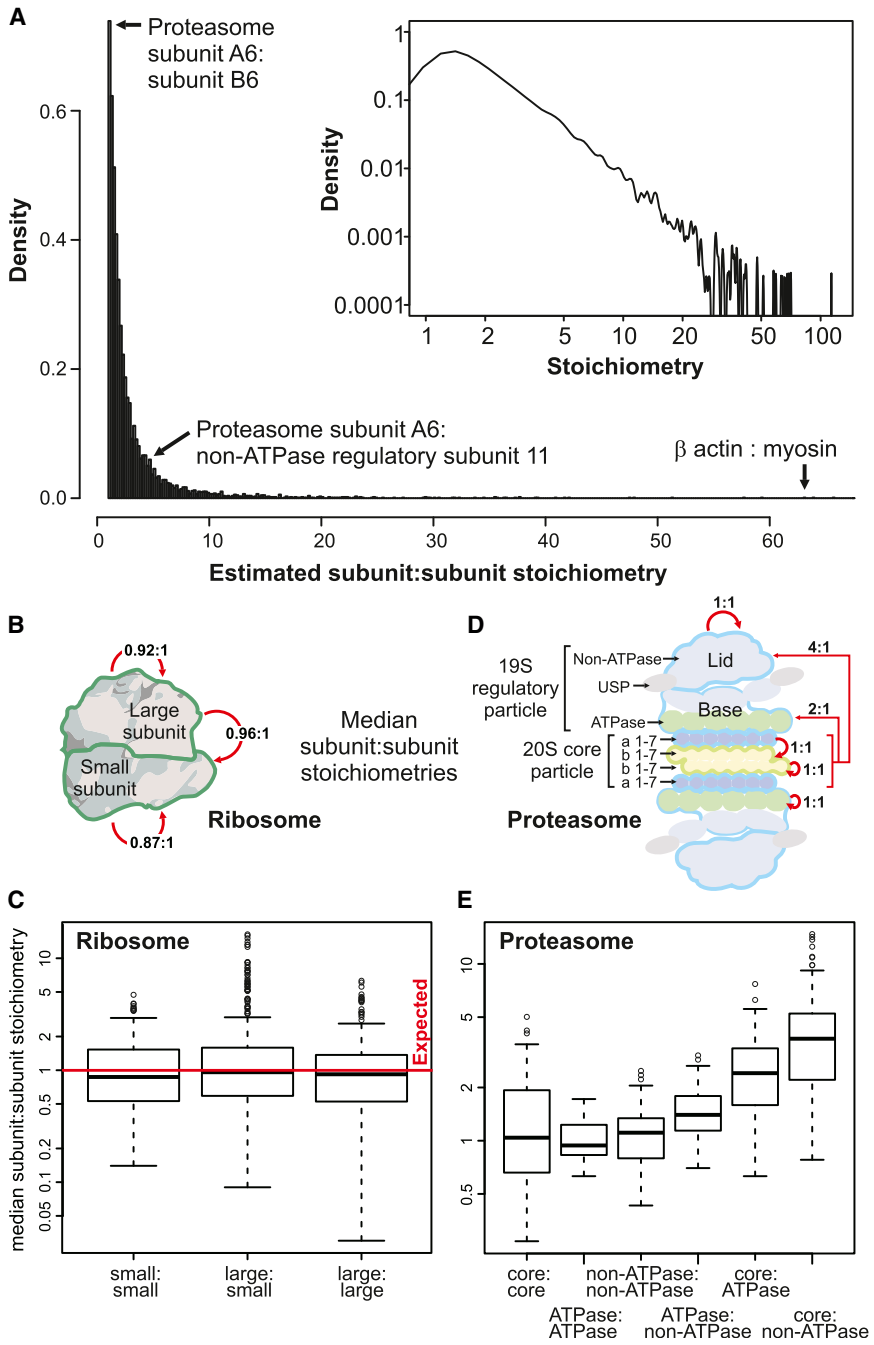


Figure 7. Protein Complex Stoichiometries

(A) Overall distribution of derived intracomplex component stoichiometries.

(B and C) Estimated subunit stoichiometries within and between proteins of the large and small ribosome subunits agree on average with the expected 1:1 ratio. Boxes summarize first quartile, median, and third quartiles, whiskers represent ±1.5 IQR, and circles represent outliers.

(D and E) Estimated protein subunit stoichiometries within and between proteasomal proteins. Intrasubunit stoichiometries within the core, ATPase, or nonATPase regulatory subunits agree well with the expected 1:1 ratio, but stoichiometries observed between these complexes deviate significantly from 1:1 (ATPase:non-ATPase, Mann-Whitney $p \leq 10^{-3}$; core:ATPase, $p \leq 10^{-12}$; core:non-ATPase, $p \leq 10^{-16}$).

See also Table S2.

all reconstructed complexes are supported by the same set of extensive experimental mass spectrometry data, we could estimate subunit stoichiometries based on the ratios of recorded spectral counts after correcting appropriately for protein size and composition (see Extended Experimental Procedures). Although only approximate ratios were inferred and peaked at ~1:1 (Figure 7A), such as between known ribosomal subunits (Figures 7B and 7C), the results highlight intriguing deviations in subunit abundance (Table S2). An example drawn from the proteasome is illustrative: whereas the median stoichiometry of core α and β enzymatic subunits is close to the expected 1:1 ratio, the median of stoichiometries of core to non-ATPase regulatory subunits deviated significantly at ~4:1 (Mann-Whitney $p \leq 10^{-16}$; Figures 7D and 7E). Hence, these data suggest a rich source of information about the physical organization of human proteins.

DISCUSSION

The biochemically based interaction data obtained in this integrative proteomic study have enabled the identification of both 364 previously unannotated protein complexes (i.e., predicted complexes with no statistically significant match to complexes in public databases) encompassing 1,278 human proteins, many of which are linked to human disease, and unexpected components and interactions for well-studied, widely conserved nuclear and cytoplasmic protein machineries, such as ribosome biogenesis, with clear biological implications. Most of the high-confidence protein interactions provided in this resource have not been previously reported in public interaction databases and hence motivate mechanistic investigations of specific biological systems.

Prior to this work, experimental knowledge regarding soluble protein complex membership in human cells has generally been ad hoc or focused on specific subcellular systems. Our relatively unbiased integrative approach, wherein biochemical evidence (cofractionation) of soluble native macromolecules was combined with genomic inferences (imputed functional associations), provides an inclusive snapshot of human protein

complexes present under a standardized cellular context, thus serving as a reference against which future process- or cell-type-specific or dynamic interaction data sets can be compared.

Information gleaned from orthology proved to be an important resource in separating true pairwise interactions from putative false positives and, in turn, could reasonably be expected to bias our results toward conserved complexes. In fact, although we do find conserved complexes as expected, we also find a majority that are not conserved (in fly and yeast) and that seemingly have arisen with vertebrates (i.e., Figure 6B). The slower rate of evolution of the subunits we report for our protein complexes is also a feature of other human PPI networks, such as in CORUM, and thus, our predictions of broad complex conservation, albeit incomplete, are not just artifacts of our methodology.

The fact that we detected little evidence of tissue specificity for most of the derived human protein complexes and few cell-type-specific components likely reflects undersampling by our mass spectrometry procedures, which is a common limitation of LC-MS/MS. At the level of predicted PPI (which are derived from multiple biochemical fractions), differences in the proteomic profiles generated for the two cell lines lie within the variance observed between biological replicates of the same cell line (Figures S1 and S2). Yet it is clear that differential interactomes and the contextual rewiring of PPI networks are major determinants of cell behavior and phenotypes. The complexes we report undoubtedly undergo differential rewiring in response to environmental, physiological, developmental, or disease states. With further refinements to our experimental procedures, our interaction mapping strategy has the potential to interrogate changes in interaction space in a systematic manner in the future.

To enable exploitation of these data by the scientific community, we have generated a dedicated web database of human protein complexes (http://human.med.utoronto.ca) that contains all the data generated in this study in an easily navigated format. These include all of the supporting information for each of the pairwise protein interactions obtained through integration of our cofractionation data with public genomic evidence, a list of the 5,584 proteins detected in each of the 1,163 biochemical fractions collected, and the subunit composition of the 622 putative protein complexes obtained through clustering of our generated high-confidence interaction network. This "first pass" draft of the soluble, stably associated human protein "complexome" provides a glimpse into the global physical molecular organization of human cells, which is likely to be perturbed in pathological states.

EXPERIMENTAL PROCEDURES

Cell Culture and Extract Preparation

HeLa S3 (ATCC#: CCL-2.2) and HEK293 (ATCC#: CRL-1573) soluble nuclear and cytoplasmic protein extracts were prepared by conventional methods (see Extended Experimental Procedures). Prior to fractionation, lysates were treated with 100 units/ml Benzonase (Novagen Inc.) to remove nucleic acids and clarified by centrifugation to remove debris.

Biochemical Fractionation and Proteomic Analysis

We performed weak anion-exchange and mixed-bed ion exchange, both with and without a heparin precolumn to enrich for nucleic-acid-binding proteins. In total, 1,095 chromatography fractions were collected (see Extended Experimental Procedures). Isoelectric focusing was carried out on a MicroRotofor Liquid-Phase IEF cell (Bio-Rad) according to the manufacturer's protocol, with 40 fractions collected across a pH range. Sucrose density gradient centrifugation was performed as previously described (Ramani et al., 2008), with 28 fractions collected.

Proteins were acid precipitated and trypsin digested, and the peptide mixtures were fractionated and sequenced by using nanoflow liquid chromatography-electrospray tandem mass spectrometry. Spectra were collected on an LTQ linear ion trap (ThermoFisher Scientific) (majority) or LTQ Orbitrap Velos hybrid mass spectrometer and searched against a UniProt human target-decoy sequence database by using SEQUEST (Eng et al., 2008) (see Extended Experimental Procedures). The LC-MS/MS identifications were filtered to a 1.0% protein and peptide theoretical FDR.

Bioinformatics Analyses

Protein cofractionation networks were scored by correlation analysis (Pearson correlation, weighted cross-correlation, coapex) based on the protein spectral counts recorded across each set of fractions (see Extended Experimental Procedures). Weighted networks were likewise constructed based on functional evidence reported in HumanNet (Lee et al., 2011), omitting human protein interaction data to minimize circularity that might bias our association predictions. A coevolution network (Tillier and Charlebois, 2009) based on correlated evolutionary rates was built to account for additional associations not covered in HumanNet.

For the machine-learning classifier, we used the fast random forest implementation in Weka (see Extended Experimental Procedures) to integrate all generated networks. Cross-validated decision trees were learned and benchmarked by using independent training and test sets of CORUM reference complexes (Ruepp et al., 2010). We denoised the network by using a diffusion procedure to delete interactions lacking network topology support and by calibrating the diffused interaction scores with Gene Ontology (Cellular Component) normalized semantic similarity scores (see Extended Experimental Procedures).

Clusters were defined by using the ClusterONE algorithm with parameter settings chosen to yield the highest maximum matching ratio (Nepusz et al., 2012) between the predicted complexes and set of cluster-training complexes (see Extended Experimental Procedures).

Stoichiometries calculation is shown in Extended Experimental Procedures.

ACCESSION NUMBERS

The interaction data have been deposited into BioGRID and are also publicly accessible via a dedicated web portal (http://human.med.utoronto.ca/).

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Experimental Procedures, five figures, and six tables and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2012.08.011.

ACKNOWLEDGMENTS

We thank R. Isserlin, Z. Ni, H. Guo, D. Merico and A. Alpert for technical assistance and J. Parkinson, G. Bader, A. Wilde, and J. Greenblatt for critical suggestions. P.C.H. was a recipient of a University of Toronto Open Fellowship, T.N. was supported by the Newton International Fellowship Scheme of the Royal Society, A.E. is an Ontario Research Chair, and S.J.W. is a Canada Research Chair Tier 1. This work was supported by grants from the Biotechnology and Biological Sciences Research Council (BB/F00964X/1 and BB/K004131/1) and the Royal Society (NF080750) to A.P., the Canada Institutes of Health Research (MOP 82940) and the SickKids Foundation to S.J.W., the National Institutes of Health, National Science Foundation, Cancer Prevention Research Institute of Texas, and Welch (F1515) and Packard Foundations to E.M.M., and the Ontario Ministry of Research and Innovation to A.E.

Received: May 26, 2012
Revised: July 30, 2012
Accepted: August 10, 2012
Published: August 30, 2012

REFERENCES

Alberts, B. (1998). The cell as a collection of protein machines: preparing the next generation of molecular biologists. Cell *92*, 291–294.

Babu, M., Vlasblom, J., Pu, S., Guo, X., Graham, C., Bean, B.D.M., Vizeacoumar, F.J., Burston, H.E., Snider, J., Phanse, S., et al. (2012). Interaction Landscape of Membrane Protein Complexes in Saccharomyces cerevisiae. Nature http://dx.doi.org/10.1038/nature11354.

Becker, K.G., Barnes, K.C., Bright, T.J., and Wang, S.A. (2004). The genetic association database. Nat. Genet. *36*, 431–432.

Behrends, C., Sowa, M.E., Gygi, S.P., and Harper, J.W. (2010). Network organization of the human autophagy system. Nature *466*, 68–76.

Bouwmeester, T., Bauch, A., Ruffner, H., Angrand, P.O., Bergamini, G., Croughton, K., Cruciat, C., Eberhard, D., Gagneur, J., Ghidelli, S., et al. (2004). A physical and functional map of the human TNF-alpha/NF-kappa B signal transduction pathway. Nat. Cell Biol. *6*, 97–105.

Butland, G., Peregrín-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., et al. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. Nature *433*, 531–537.

Deardorff, M.A., Wilde, J.J., Albrecht, M., Dickinson, E., Tennstedt, S., Braunholz, D., Mönnich, M., Yan, Y., Xu, W., Gil-Rodríguez, M.C., et al. (2012). RAD21 mutations cause a human cohesinopathy. Am. J. Hum. Genet. *90*, 1014–1027.

DeScipio, C., Kaur, M., Yaeger, D., Innis, J.W., Spinner, N.B., Jackson, L.G., and Krantz, I.D. (2005). Chromosome rearrangements in cornelia de Lange syndrome (CdLS): report of a der(3)t(3;12)(p25.3;p13.3) in two half sibs with features of CdLS and review of reported CdLS cases with chromosome rearrangements. Am. J. Med. Genet. A. *137A*, 276–282.

Eng, J.K., Fischer, B., Grossmann, J., and Maccoss, M.J. (2008). A fast SEQUEST cross correlation algorithm. J. Proteome Res. *7*, 4598–4602.

Ewing, R.M., Chu, P., Elisma, F., Li, H., Taylor, P., Climie, S., McBroom-Cerajewski, L., Robinson, M.D., O'Connor, L., Li, M., et al. (2007). Large-scale mapping of human protein-protein interactions by mass spectrometry. Mol. Syst. Biol. *3*, 89.

Gavin, A.C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature *415*, 141–147.

Gavin, A.C., Aloy, P., Grandi, P., Krause, R., Boesche, M., Marzioch, M., Rau, C., Jensen, L.J., Bastuck, S., Dümpelfeld, B., et al. (2006). Proteome survey reveals modularity of the yeast cell machinery. Nature *440*, 631–636.

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

Graham, F.L., Smiley, J., Russell, W.C., and Nairn, R. (1977). Characteristics of a human cell line transformed by DNA from human adenovirus type 5. J. Gen. Virol. *36*, 59–74.

Guruharsha, K.G., Rual, J.F., Zhai, B., Mintseris, J., Vaidya, P., Vaidya, N., Beekman, C., Wong, C., Rhee, D.Y., Cenaj, O., et al. (2011). A protein complex network of Drosophila melanogaster. Cell *147*, 690–703.

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.

Hartwell, L.H., Hopfield, J.J., Leibler, S., and Murray, A.W. (1999). From molecular to modular cell biology. Nature *402*(6761, Suppl), C47–C52.

Havugimana, P.C., Wong, P., and Emili, A. (2007). Improved proteomic discovery by sample pre-fractionation using dual-column ion-exchange high performance liquid chromatography. J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. *847*, 54–61.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., et al. (2002). Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature *415*, 180–183.

Hu, P., Janga, S.C., Babu, M., Díaz-Mejía, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P., et al. (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. *7*, e96.

Hutchins, J.R., Toyoda, Y., Hegemann, B., Poser, I., Hériché, J.K., Sykora, M.M., Augsburg, M., Hudecz, O., Buschhorn, B.A., Bulkescher, J., et al. (2010). Systematic analysis of human protein complexes identifies chromosome segregation proteins. Science *328*, 593–599.

Jansen, R., and Gerstein, M. (2004). Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. Curr. Opin. Microbiol. *7*, 535–545.

Jeronimo, C., Forget, D., Bouchard, A., Li, Q., Chua, G., Poitras, C., Thérien, C., Bergeron, D., Bourassa, S., Greenblatt, J., et al. (2007). Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme. Mol. Cell *27*, 262–274.

Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003). PRISM, a generic large scale proteomic investigation strategy for mammals. Mol. Cell. Proteomics *2*, 96–106.

Krogan, N.J., Cagney, G., Yu, H., Zhong, G., Guo, X., Ignatchenko, A., Li, J., Pu, S., Datta, N., Tikuisis, A.P., et al. (2006). Global landscape of protein complexes in the yeast Saccharomyces cerevisiae. Nature *440*, 637–643.

Kühner, S., van Noort, V., Betts, M.J., Leo-Macias, A., Batisse, C., Rode, M., Yamada, T., Maier, T., Bader, S., Beltran-Alvarez, P., et al. (2009). Proteome organization in a genome-reduced bacterium. Science *326*, 1235–1240.

Lage, K., Karlberg, E.O., Størling, Z.M., Olason, P.I., Pedersen, A.G., Rigina, O., Hinsby, A.M., Tümer, Z., Pociot, F., Tommerup, N., et al. (2007). A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat. Biotechnol. *25*, 309–316.

Lee, I., Blom, U.M., Wang, P.I., Shim, J.E., and Marcotte, E.M. (2011). Prioritizing candidate disease genes by network-based boosting of genome-wide association data. Genome Res. *21*, 1109–1121.

Mak, A.B., Ni, Z., Hewel, J.A., Chen, G.I., Zhong, G., Karamboulas, K., Blakely, K., Smiley, S., Marcon, E., Roudeva, D., et al. (2010). A lentiviral functional proteomics approach identifies chromatin remodeling complexes important for the induction of pluripotency. Mol. Cell. Proteomics *9*, 811–823.

Malovannaya, A., Lanz, R.B., Jung, S.Y., Bulynko, Y., Le, N.T., Chan, D.W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., et al. (2011). Analysis of the human endogenous coregulator complexome. Cell *145*, 787–799.

Masters, J.R. (2002). HeLa cells 50 years on: the good, the bad and the ugly. Nat. Rev. Cancer *2*, 315–319.

McBrien, J., Crolla, J.A., Huang, S., Kelleher, J., Gleeson, J., and Lynch, S.A. (2008). Further case of microdeletion of 8q24 with phenotype overlapping Langer-Giedion without TRPS1 deletion. Am. J. Med. Genet. A. *146A*, 1587–1592.

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

Nepusz, T., Yu, H., and Paccanaro, A. (2012). Detecting overlapping protein complexes in protein-protein interaction networks. Nat. Methods *9*, 471–472.

Neumann, B., Walter, T., Hériché, J.K., Bulkescher, J., Erfle, H., Conrad, C., Rogers, P., Poser, I., Held, M., Liebel, U., et al. (2010). Phenotypic profiling of the human genome by time-lapse microscopy reveals cell division genes. Nature *464*, 721–727.

Oliver, S. (2000). Guilt-by-association goes global. Nature *403*, 601–603.

Pesquita, C., Faria, D., Falcão, A.O., Lord, P., and Couto, F.M. (2009). Semantic similarity in biomedical ontologies. PLoS Comput. Biol. *5*, e1000443.

Pié, J., Gil-Rodríguez, M.C., Ciero, M., López-Viñas, E., Ribate, M.P., Arnedo, M., Deardorff, M.A., Puisac, B., Legarreta, J., de Karam, J.C., et al. (2010). Mutations and variants in the cohesion factor genes NIPBL, SMC1A, and SMC3 in a cohort of 30 unrelated patients with Cornelia de Lange syndrome. Am. J. Med. Genet. A. *152A*, 924–929.

Popow, J., Englert, M., Weitzer, S., Schleiffer, A., Mierzwa, B., Mechtler, K., Trowitzsch, S., Will, C.L., Lührmann, R., Söll, D., and Martinez, J. (2011). HSPC117 is the essential subunit of a human tRNA splicing ligase complex. Science *331*, 760–764.

Ramani, A.K., Li, Z., Hart, G.T., Carlson, M.W., Boutz, D.R., and Marcotte, E.M. (2008). A map of human protein interactions derived from co-expression of human mRNAs and their orthologs. Mol. Syst. Biol. *4*, 180.

Rhodes, D.R., Tomlins, S.A., Varambally, S., Mahavisno, V., Barrette, T., Kalyana-Sundaram, S., Ghosh, D., Pandey, A., and Chinnaiyan, A.M. (2005). Probabilistic model of the human protein-protein interaction network. Nat. Biotechnol. *23*, 951–959.

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.

Sardiu, M.E., Cai, Y., Jin, J., Swanson, S.K., Conaway, R.C., Conaway, J.W., Florens, L., and Washburn, M.P. (2008). Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc. Natl. Acad. Sci. USA *105*, 1454–1459.

Sowa, M.E., Bennett, E.J., Gygi, S.P., and Harper, J.W. (2009). Defining the human deubiquitinating enzyme interaction landscape. Cell *138*, 389–403.

Tillier, E.R., and Charlebois, R.L. (2009). The human protein coevolution network. Genome Res. *19*, 1861–1871.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., et al. (2010). Towards a knowledge-based Human Protein Atlas. Nat. Biotechnol. *28*, 1248–1250.

UniProt Consortium. (2011). Ongoing and future developments at the Universal Protein Resource. Nucleic Acids Res. *39*(Database issue), D214–D219.

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

Wessels, H.J., Vogel, R.O., van den Heuvel, L., Smeitink, J.A., Rodenburg, R.J., Nijtmans, L.G., and Farhoud, M.H. (2009). LC-MS/MS as an alternative for SDS-PAGE in blue native analysis of protein complexes. Proteomics *9*, 4221–4228.

Wuyts, W., Roland, D., Lüdecke, H.J., Wauters, J., Foulon, M., Van Hul, W., and Van Maldergem, L. (2002). Multiple exostoses, mental retardation, hypertrichosis, and brain abnormalities in a boy with a de novo 8q24 submicroscopic interstitial deletion. Am. J. Med. Genet. *113*, 326–332.

Xie, X., Lu, J., Kulbokas, E.J., Golub, T.R., Mootha, V., Lindblad-Toh, K., Lander, E.S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature *434*, 338–345.

Yang, H., Nepusz, T., and Paccanaro, A. (2012). Improving GO semantic similarity measures by exploring the ontology beneath the terms and modelling uncertainty. Bioinformatics *28*, 1383–1389.

EXTENDED EXPERIMENTAL PROCEDURES

Biochemical Fractionation Using Native Chromatography

HPLC Columns, Buffers, and Instrumentation

IEX chromatography columns (weak anion-exchange PolyWAX LP; weak cation-exchange PolyCAT A; mixed-bed PolyCATWAX50/50 columns) were purchased from PolyLC Inc (MD, USA). TSKgel Heparin-5PW affinity column was obtained from Tosoh Bioscience LLC (PA, USA). Our buffer systems (fresh prepared with HPLC grade H_2O) comprised low salt buffer A [10 mM Tris-HCl, pH7.6, 3 mM NaN_3, 0.5 mM DTT, 5%-Glycerol] and high salt Buffer B [Buffer A + 1.5 M NaCl]. We performed all HPLC fractionations using an Agilent 1100 HPLC binary pump system (Agilent Technologies, ON, Canada), essentially described elsewhere (Havugimana et al., 2007). Protein elution was monitored by absorption at 280 nm.

Single-Phase Heparin Fractionation of Nuclear Extract

HeLa nuclear extract (~6.0 mg total proteins) prepared using traditional methods (Dignam et al., 1983) was fractionated on a TSKgel Heparin-5PW affinity column (75 × 7.5 mm id, 10 μm, 1000-A) previously equilibrated with buffer A at a flow rate of 0.5 ml/min. After loading, the bound proteins were eluted from the column with a 50 min gradient from 0 to 50% buffer B (buffer A + 1.5 M NaCl). A 5 min gradient with 50%–100% buffer B was applied to elute tightly bound proteins, with 100% buffer B maintained for an additional 3 min before returning back to 0% B for 7 min to re-equilibrate the column. In total, 48 × 0.75-ml fractions were collected from 0 to 72 min (1.5 min/fraction). Protein was precipitated with 10% TCA overnight at 4°C. The pellet was washed twice with −20°C acetone for 30 min. After air drying, the pellet was dissolved in 50 μl digest solution (50 mM NH_4HCO_3- 50 mM Tris, 1 mM $CaCl_2$). The sample reduction (room temperature, 1 hr) and alkylation (room temperature, 30 min) were respectively performed using 5 mM and 15 mM of Dithiothreitol and Iodoacetamide. Each protein fraction was digested with 1 μg of sequencing grade trypsin (Roche, Mississauga, Canada). After incubation for 18 hr at 30°C with gentle shaking (VWR incubating micro-plate shaker; 300 rpm) samples were dry speed-vac. 20 μl of LC-MS grade buffer (5% Formic Acid in HPLC grade water) were used to solubilise the peptide- digests. 8 μl tryptic peptides aliquot were directly analyzed by LC-MS.

Single-Phase Weak Anion-Exchange Fractionation of HeLa Cytosolic Extract

A total of 2.0-3.0 mg soluble protein in HeLa S3 cytosolic extract were applied to a PolyWAX LP column (200 × 4.6 mm id, 5 μm, 1000-A) equilibrated with buffer A. Elution of bound proteins was achieved through application of a 30 min gradient from 0 to 50% buffer B, with a final 2 min gradient of 50%–100% buffer B applied to elute tightly bound proteins. 100% buffer B was maintained for an additional 2 min before returning back to 0% buffer B in 2 min for re-equilibration of the column for 3 min. A total of 45 × 1.2-ml fractions were collected using a flow rate of 1.2 ml/min. The first and last fractions lacking protein (as judged by UV-absorption at 280 nm) were discarded. The rest of collected fractions were processed as described above.

Dual-Phase Heparin-Mixed-Bed Ion Exchange Fractionation of Nuclear Extracts

To enhance detection of low abundance nuclear proteins by MS, we used an optimized high resolution tandem affinity column coupled online with a mixed-bed ion exchange column to enrich and resolve multi-proteins complexes in nuclear extracts. Typically, 8-10 mg proteins from HeLa or HEK293 nuclear extracts were loaded on a dual TSKgel Heparin-5PW affinity column (75 × 7.5 mm id, 10 μm, 1000-A) coupled in series with PolyCATWAX mixed-bed ion exchange column (200 × 4.6 mm id, 12 μm, 1500-A) mounted to our integrated Agilent 1100 HPLC system (Agilent Technologies, ON, Canada). A 4 hr salt gradient (0.15 - 1.5 M NaCl) in Binding Buffer A was used at 0.25 ml/min to resolve and fractionate proteins into 120 × 0.5-ml time-based fractions for downstream MS protein identification. HeLa nuclear extract was fractionated in duplicates to confirm the reproducibility.

Triple-Phase Ion-Exchange Fractionation of HeLa Nuclear Extracts

As we have shown in our previous work (Havugimana et al., 2006, 2007), tandem weak anion-exchange (WAX) coupled in series to a weak cation-exchange (WCX) offered greater resolution than a single column or WCX-WAX in tandem. To minimize both chance co-elution and bias toward one chromatographic fractionation approach, we used our further semi-preparative optimized and reproducible triple phase IEX-HPLC that comprised our previously optimized columns system preceded with a long weak anion-exchange (250 × 9.4 mm i.d, 12 μm, 1500-A PolyWAX LP → 250 × 9.4-mm i.d, 12 μm, 1500-A PolyWAX LP → 250 × 9.4 mm i.d, 5 μm, 1500-A PolyCAT A) to fractionate 10-12 mg total proteins in HeLa nuclear extracts into 375 × 0.8-ml fractions using elution program consisting of a 10 min gradient with 100% buffer A to allow protein binding followed by a 50 min gradient with 0 to 50% buffer B followed by a10- min gradient with 50 to 100% buffer B, 10 min at 100% buffer B, 10 min with 100 to 0% buffer B, and finally 10- min at 100% buffer A to re-equilibrate the column for the next injection. A flow rate of 4-ml/min was applied in elution gradient program. Collected fractions were analyzed by LC-MS/MS in duplicates.

Triple-Phase Ion-Exchange Fractionation of HeLa Cytosolic Extracts

To identify macromolecular complexes that populate the HeLa cytoplasmic compartment, we scaled down our optimized semi-preparative IEX-HPLC fractionation procedure to enhance protein concentration in each collected fraction. Seven to 9 mg total proteins in HeLa cytoplasmic extract were fractionated on a triple phase IEX-HPLC analytical columns set up (200 × 4.6 mm i.d, 5 μm, 1000-A PolyWAX LP → 200 × 4.6-mm i.d, 5 μm, 1000-A PolyWAX LP → 200 × 4.6 mm i.d, 5 μm, 1000-A PolyCAT A) and resolved into 300 × 0.4-ml fractions using a 2.5 hr gradient elution program (23 min with 100% buffer A; 75 min with 0%–50% buffer B; 3 min with 50%–100% buffer B; 23 min with 100% buffer B; 3 min with 100 to 0% buffer B; 23 min with 100% buffer A) at flow rate of 0.8 ml/min. Both the 19 fractions representing the column flow through and the 12 fractions representing the re-equilibration step

were discarded as no proteins were detected in our short quality control LC-MS/MS analysis. All remaining 269 fractions were analyzed in duplicate by LC-MS/MS.

Biochemical Fractionation Using IEF and Sucrose Gradient Sedimentation

Sample Preparation for Isoelectric Focusing Fractionation

HeLa cells were grown to 70%–80% confluency in 75cm^2 flasks and harvested by mechanical scraping. Cells were washed in ice-cold PBS, pelleted by centrifugation (600xg), and resuspended in lysis buffer [10 mM Tris-HCl (pH 8.0), 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM DTT, and 1x Protease Inhibitor Cocktail Set I (Calbiochem)]. Cells were lysed on ice using a Dounce homogenizer and fractionated into cytosolic and nuclear fractions using a protocol adapted from previous publication (Andersen et al., 2002). Briefly, cells were centrifuged at 1000xg for 5 min (4°C). The supernatant was saved as the cytosolic fraction. The pellet was resuspended in 250 mM sucrose/10 mM $MgCl_2$/1x Protease Inhibitor Cocktail, layered over a sucrose cushion of 880 mM sucrose/0.5 mM $MgCl_2$/1x Protease Inhibitor Cocktail, and centrifuged at 3000xg for 10 min (4°C). The supernatant was discarded and the pellet resuspended in lysis buffer with 5% NP-40 by sonicating water bath (15 min). Following sonication, samples were centrifuged at 3,500xg for 10 min to pellet insoluble material, with the supernatant saved as the nuclear fraction.

IEF Fractionation

Cytosolic and nuclear fractions were further fractionated in solution by isoelectric focusing on a MicroRotofor Liquid-Phase IEF cell (Bio-Rad). Ten fractions per sample were collected across a pH range of either 3-10 or 5-8. Following IEF fractionation, ampholytes were removed by OrgoSol DetergentOUT detergent removal kit (G-Biosciences).

Trypsin Digestion and MS Analysis of IEF Samples

Samples were denatured and reduced in 50% 2,2,2-trifluoroethanol (TFE) and 15 mM DTT at 55°C for 45 min, followed by alkylation with 55 mM iodoacetamide for 30 min at room temperature in the dark. Following alkylation, samples were diluted to 5% TFE in 50 mM Tris-HCl, pH8.0/2 mM $CaCl_2$ and digested with a 1:50 final concentration of Proteomics Grade trypsin (Sigma) for 5 hr at 37°C. Digestion was quenched by addition of 1% formic acid, and the sample volume was reduced to near dry (<20 μl) by speed vac centrifugation. Samples were resuspended in 5% acetonitrile/0.1% formic acid and bound and washed on HyperSep C18 SpinTips (Thermo). Following elution, the sample volume was reduced by speed vac to remove elution buffer. Samples were resuspended in 5% acetonitrile/0.1% formic acid and filtered through Amicon Ultra 10kDa centrifugation filters (Millipore).

Samples were analyzed by LC-MS/MS. Peptides were separated on a Zorbax 300SB-C18 reverse phase column (0.075 × 150 mm, 3.5 μm; Agilent) with an elution gradient of 5%–38% acetonitrile over 230 min followed by 38%–100% over 15 min. Peptides were analyzed by nanoelectrospray ionization onto an LTQ Orbitrap mass spectrometer (Thermo Scientific). Parent mass scans (MS1) were collected at high resolution (100,000) with data dependent ion selection activated for ions of greater than +1 charge. Up to 12 ions per MS1 were selected for CID fragmentation spectrum acquisition (MS2), with ions selected twice within 30 s placed on a dynamic exclusion list for 45 s.

Sucrose Gradient Fractionation of HeLa

Generation of the sucrose density gradient fractions and MS analysis were described elsewhere (Andersen et al., 2002; Ramani et al., 2008). Briefly, they were generated using a 7%–47% continuous sucrose gradient and ultra-high-speed centrifugation of the supernatants from HeLa S3 cell-free extracts. Gradient fractions were analyzed by Mass Spectrometry with LTQ-Orbitrap hybrid mass spectrometer (ThermoFisher), and tandem mass spectra were searched as described below.

LC-MS/MS Separation and Identification of Chromatographic Peptide Fractions

For LC-MS/MS analysis of HPLC protein fractions, samples were overnight 10%-TCA precipitated and neat-cold acetone was used to wash the precipitates. Proteins were then resuspended in 50 μl of trypsin digestion buffer [50 mM Ammonium Bicarbonate, 1 mM $CaCl_2$, 50 mM Tris; pH7.8], subjected to reduction (10 mM DTT, 30 min, 30°C), alkylation (15 mM IAM, 60 min, 30°C in the dark), and digestion (18 hr, 30°C, with gentle agitating) with one μg trypsin sequencing grade (Roche, Mississauga, Canada). The digestion mixture was dried in the Savant Speed Vacuum, and tryptic peptides were re-solubilised in 20 μl of 5% formic acid prior to analysis by LC-MS/MS using a linear ion trap mass spectrometer (LTQ; Thermo Fisher Scientific, CA, USA) or LTQ Orbitrap Velos (Thermo Fisher) coupled online to a nanoflow HPLC System (EASY-nLC; Proxeon, Odense, Denmark) via a nanoelectrospray ion source. Reverse-phase LC-MS/MS using 150-μm i.d × 40 cm in-house packed fused-silica C18 micro-capillary columns (Zorbax XDB-C18, 3.5 μm, Agilent Technologies, Canada) at a flow rate of 500 nl/min were used to resolve peptides mixture in each HPLC fraction. To separate peptides, we used columns with varying between 10-40 cm in length depending on sample complexity in each fractionation experiment. The gradient elution time was adjusted to the length of the column and varied between 2 and 4 hr. For a 2 hr gradient elution, 5 μl of tryptic peptides generated for TCS-HPLC fractions were loaded onto a 20-cm column and eluted with a 0 to 35% solvent B (0.1% formic acid/95% acetonitrile) over 90 min and from 35 to 95% in 15 min. For peptides analyzed on an LTQ ion trap instrument, eluted peptides were directly sprayed into an LTQ ion trap MS instrument via application of a spray voltage of 3.0 kV to a nanospray ion source (Proxeon). The MS was operated in a fully automated data-dependent manner using Xcaliber 2.0 software to acquire one full MS scan (400 - 2,000 m/z) followed by five MS/MS scans selected based on the five most abundant precursor ions and a precursor signal threshold of 1,000 counts. Ion fragmentation was performed in CID mode through application of normalized collision energy of 35%. Ions subjected to MS/MS were excluded from further sequencing for 30 s. For peptide mixtures analyzed on an LTQ-Orbitrap

Velos instrument, peptide samples were directly autosampled onto a 10 cm in-house packed column (75 μm inner diameter) with 3 μm reversed phase beads (Zorbx 80XDB-C18, Agilent). Using a 60 min gradient (5%–35% ACN), peptides were directly electro-sprayed (2.5 kV) into the mass spectrometer. Mass spectrometer was operated in data dependent mode switching automatically between one full scan MS and 10 MS/MS acquisitions. Instrument control was through Tune 2.6.0. and Xcalibur 2.1.0. Full scanMS spectra (400 – 2,000 m/z) were acquired in the Orbitrap analyzer after accumulation to a target value of 10^6 in the linear ion trap (resolution of 60,000 at 400 m/z).Fragmentation was performed in CID mode applying 35% normalized collision energy.

LC-MS/MS Spectra Database Search and Protein Identification

All MS/MS spectra IEF, SGF and IEX experiments acquired during over 9,000 hr of dedicated instrument run time were combined (resulting in > 18,000,000 mass spectra) and rigorously searched against a target-decoy human database downloaded from Universal Protein Resources Database (UniProtKB/Swiss-Prot Release 57.11; comprising 20,328 human proteins supplemented with common contaminants) using the SEQUEST algorithm (V2.7) as previously described (Eng et al., 2008). Static modifications were permitted to allow for the detection of carboxyamidomethylated (+57amu) cysteine. All peptide matches were required to be fully tryptic although one missed cleavage was permitted. The probabilistic STATQUEST model (Kislinger et al., 2003) was used to evaluate and assign confidence scores to all putative matches. Both proteins and peptides were considered positively identified if detected within a 1% false discovery rate cut off (based on empirical target-decoy database search results). The proteomic patterns of the HPLC, IEF and SGF fractions were compared using the CONTRAST software tool (Tabb et al., 2002). We then removed from consideration all proteins that passed our stringent cut off with only a single spectral count across all combined MS runs. Moreover, to ensure a high quality proteomic data set, we confirmed the expression of our LC-MS detected proteins by cross-comparing with previously reported HeLa S3 and HEK293 mRNA deep-sequencing data sets (Morin et al., 2008; Sultan et al., 2008). Additionally, we only kept proteins that were supported by at least two unique peptides in at least one recent comprehensive proteomic study of the HeLa proteome (Selbach et al., 2008; Wiśniewski et al., 2009). This screening procedure resulted in 41,506 unique peptides (supported by ~1.6 million individual mass spectra) matching to 5,584 distinct human proteins. To facilitate cross-mapping between data sets, we used UniProtKB accession numbers as a common identifier and the UniProt ID mapping tool to interconvert different gene and protein identifiers.

Polysome Profiling and Quantitative RT-PCR

HeLa cells were maintained in Dulbecco's Modified Eagle's Media (DMEM) supplemented with 10% fetal calf serum in a humidified 5% CO_2 incubator at 37°C. Cells were transfected with 10 nM ON-TARGETplus SMARTpool siRNA (Thermo Scientific Dharmacon) by using RNAiMAX (Invitrogen) at about 30% confluency. After 48 hr, 100 μg/ml cycloheximide (Sigma) was added into the culture medium and cells were incubated for 5 min in the incubator. Then cells were collected by trypsinization and washed with cold PBS containing 100 μg/ml cycloheximide twice. $1x10^5$ cells were frozen in −80°C for RNA extraction. The remaining cells were lysed in the lysis buffer (20 mM Tris, pH7.4, 100 mM KCl, 10 mM MgCl2, 1% Triton-100, 1 mM DTT, 100 μg/ml cycloheximide, 1x EDTA-free inhibitor tablet) on ice for 5 min. Extracts were clarified by centrifugation at 13,000 rpm for 10 min at 4 deg. The supernatant was loaded onto a linear sucrose gradient (15%–45%) prepared in lysis buffer without Triton. After a 4 hr centrifugation at 36,000 rpm in a Beckman SW40 rotor, the sucrose gradient was fractionated and absorbance at 254 nm was measured (ISCO fractionator). For qRT-PCR, total RNA was extracted by RNeasy Plus Micro (QIAGEN). QuantiTect reverse transcription kit and QuantiFast SYBR Green RT-PCR Kit from QIAGEN were used for qRT-PCR. The primer pairs for each gene in qRT-PCR were as follow: human MKI67IP(rMKI67IP-1: 5′-CCTGTTTGGTGAAAGACTCTTG-3′; rMKI67IP-2: 5′-GCTTTTGTGTTAGTGTCCGATTC-3′), Human GNL3(rGNL3-1: 5′-CATTCGGGTTGGAGTAATTGG-3′; rGNL3-2: 5′-TGTGATCTGTTTGTCCAAGGG-3′), Human DDX18(rDDX18-1: 5′-GATTGTTCAGTATGACCCTCCG-3′; rDDX18-2: 5′-CATGCCCTCTCCCATTTAGG-3′), Human FTSJ3(rFTSJ3-3: 5′-TCTCTGGATA GTGACCTGGATC-3′; rFTSJ3-4: 5′-ACTTCAGTAAGTCGCATACGC-3′), Human GAPDH(GAPDH-Fr: 5′-CTTTGTCAAGCTCATTTC CTGG-3′; GAPDH-Rr: 5′-TCTTCCTCTTGTGCTCTTGC-3′).

Immunoprecipitation Mass Spectrometry

C-terminal 3X-FLAG tagged expression clones of candidate ribosome biogenesis proteins were constructed via Gateway LR cloning (Invitrogen) of human ORF clones from the PlasmidID collection into a modified pcDNA3 vector (Invitrogen) followed by sequence verification. $3x10^6$ HEK293 cells were transfected with 5 μg of DNA of tagged genes and untransfected cells were used as control. FuGene6 (Roche) reagent in DMEM medium with 10% FBS and 1 U/ml of penicillin and streptomycin (Lonza) was used to transfect the cells for 24 hr. Cells were harvested after growing in the same medium with 10 U/ml of penicillin and streptomycin for an additional 24 hr. Cell lysis, FLAG immunoprecipitation (IP) on M2-agarose (Sigma; A2220), immuno-complex elution and digestions were performed according to the method of Dunham et al. (2011). Digested peptide mixtures (9 μl) were loaded onto a reverse phase micro-capillary pre-column (25-mm x 75-μm silica packed with 5-μm Luna C18 stationary phase; Phenomenex) and injected onto a micro-capillary analytical column (100-mm × 75-μm). Peptide separation was performed over 105 min with 5%–95% Acetonitrile (acidified with 0.1% formic acid) via an EASY-nLC system. Eluted peptides were directly sprayed into an Orbitrap Velos mass spectrometer (ThermoFisher Scientific) with collision activated dissociation using a nanospray ion source (Proxeon). 10 MS/MS data-dependent scans were acquired simultaneously with one high resolution (60,000) full scan mass spectrum. An exclusion list was enabled to exclude a maximum of 500 ions for 30 s. Acquired RAW files were extracted from the mass spectrometry data with

the extractms program and submitted for database searching using the SEQUEST search engine against a target-decoy UniProtKB/Swiss-Prot FASTA file. Search parameters were set to allow for one missed cleavage site, one variable modification of +16 for methionine oxidation and one fixed modification of +57 for cysteine carbamidomethylation using precursor ion tolerances of 3 m/z. After searching, peptide and protein hits were filtered using a 20 ppm tolerance for the precursor ion. We required 1% FDR for protein and peptide positive identifications.

Computational Analyses

MS Correlation Measures

Pearson Correlation Coefficient Score. Proteins belonging to the same multi-protein complex should co-elute across a biochemical fractionation, giving rise to similar elution profiles for those proteins. The similarity of elution profiles, represented as vectors containing the observed spectral counts for a protein in each fraction in a single experiment, was initially measured by Pearson correlation coefficient of the normalized elution profiles.

Each fractionation and mass spectrometry series identifies N proteins across M fractions. The raw data matrix is then an N by M matrix A where each $A(i,j)$ represents the number of MS/MS spectra observed to match protein i in fraction j. The normalized data matrix, B, converts numbers of peptides to frequencies, and is calculated as

$$B(i,j) = \frac{A(i,j)}{\sum_i A(i,j)}$$

A protein's normalized elution profile is represented by a row in this matrix, and the Pearson correlation coefficient was measured for each pair of proteins.

While the Pearson correlation coefficient is a good indicator of a co-complex relationship if both proteins are observed at high counts in the matrix, proteins observed at very low counts but found in the same fraction are often perfectly correlated but have poor predictive power (Figure S5).

To circumvent this artifact, we synthetically introduced noise into the raw data matrix and measured the extent to which noise affected the observed correlations and, by extension, the predictive power of correlation as it relates to protein complex membership. The observation of each protein in each fraction is modeled as a Poisson process, with lambda parameter assigned as the maximum likelihood estimate equal to the raw counts of protein i in fraction j (the $A(i,j)$ value). The noise term $1/M$ was added to the maximum likelihood estimate for each cell. The value $1/M$ was chosen on the basis that each protein was represented in the matrix by at least one peptide count, and the background probability for this should be evenly distributed across the M fractions. Thus the noise-added matrix $C = A + 1/M$, a constant. The MS experiment is re-run in silico by drawing randomly from Poisson($C(i,j)$) for each cell, then normalizing as above and calculating the Pearson correlations for each pair of proteins. This process was repeated 1,000 times, and the mean Pearson correlation for each pair was recorded. The noise term has the effect of giving every cell in the matrix a nonzero, albeit small, probability of "discovering" a protein count in that cell. The impact of this discovery on the correlation of that protein's elution profile with other normalized elution profiles is minimal for proteins observed at high counts and maximal for those observed with only one count across all fractions.

Weighted Cross-Correlation

In addition to the noise model correlation scores, a weighted cross correlation score was calculated for each pair of proteins in each experiment. We calculated the similarity of spectra profiles between each pair of proteins using a weighted cross correlation (WCC) approach (de Gelder et al., 2001), which was implemented in the R package wccsom (http://cran.r-project.org/web/packages/wccsom/index.html). The similarity value is between 0 and 1.

There are some advantages of this approach over other similarity measures, such as Pearson correlation coefficient. The WCC approach can take into account the relative shift between spectra profile patterns. In other words, given a protein, we can compare its spectra profile at a point/fraction with the profiles in that neighborhood of the corresponding point/fraction of another protein. Moreover, we can weight the different points in the neighborhood. In our calculation, we considered one point/fraction shift between spectra profile patterns and defined the weights based on a simple triangle function (http://mathworld.wolfram.com/TriangleFunction.html).

Machine Learning Methods

The noise-model correlations and weighted cross correlations of each pair of proteins observed in each of the seven cytoplasmic and eleven nuclear MS fractionation experiments were combined into matrices of protein pairs x 14 (cytoplasmic) or x 22 (nuclear) experimental observations. Missing data, where the pair of proteins were not both observed in a given experiment, were interpreted as zeros.

A gold standard reference set of positive and negative interactions was generated from the CORUM database of curated mammalian protein complexes. Human complexes consisting of 3 or more proteins were identified and filtered for those identified by mass spectrometry and related methods, removing those identified solely by, e.g., two-hybrid approaches, EMSA, and imaging techniques. Highly overlapping complexes (those with Simpson coefficient > 0.5) were merged, resulting in a reference set of 324 complexes comprised of 2,151 proteins. Each complex was then classified as "nuclear" and/or "cytoplasmic" based on the GO

Cellular Component annotation of its constituent proteins, resulting in 198 cytoplasmic and 190 nuclear complexes. These complexes were then randomly split into two groups, one for training pairwise co-complex protein-protein interactions in a machine learning framework and an independent set for optimizing final protein complex predictions from putative PPI. For PPI training, a reference positive interaction was defined as the case when two proteins were annotated to be in the same complex, and a reference negative interaction was defined where both proteins were in the annotated set but never appeared in the same complex. Although the CORUM complexes contain a large number of highly overlapping, redundant complex definitions, merging redundant complexes and reducing the complexes to unique pairwise interactions minimizes this source of bias. To further reduce bias, we omitted the largest complexes from the CORUM reference set (e.g., spliceosome, ribosome), which would otherwise account for a majority of reference PPI. Moreover, although our definition of negative interactions almost certainly contains some actual positives due to incomplete annotations, their effect is necessarily small, as negative interactions greatly outnumber positives. This renders our estimates of accuracy conservative, as some negatives will in fact be mislabeled. Our complete set of reference complexes is listed in Table S3.

The data were subjected to a variety of machine learning algorithms using the Weka suite of tools and assessed for accuracy and coverage. *Naive* Bayes and Logistic Regression classifiers were run using default parameters. Support Vector Machines (SVM) were applied using the SMO engine with a radial basis function kernel. The Random Forest implementation in Weka was too slow to use in an exploratory fashion but the Fast Random Forest re-implementation (http://code.google.com/p/fast-random-forest/) gave a significant performance boost and yielded the best results, as judged by cross-validated recall-precision analysis.

Incorporation of Genomic and Proteomic Evidence

Genomic and proteomic evidence were assembled from the HumanNet functional gene interaction network (Lee et al., 2011). HumanNet integrates a wide array of alternate data types across both human cell lines and model organism experiments into a log likelihood score indicating the strength of evidence suggesting that a given pair of genes operates in the same biological process. We considered only selected lines of evidence from HumanNet, excluding data derived from human experimental and computational prediction of protein-protein interactions, in order to minimize circularity that might bias predictions of PPIs. In all, protein-protein linkages from 17 lines of evidence were individually added to the classifier as independent features, with missing values set to zero. Table S6 lists the data types included in this study.

The nuclear data set thus comprised 41 quantitative features for each protein pair: 11 MS data sets measured by noise-model correlation, and again by weighted cross-correlation; the 17 features from HumanNet; a Co-Evolution score (Clark et al., 2011; Tillier and Charlebois, 2009) measuring correlated evolutionary rates; and a Co-Apex score measuring the number of MS experiments in which both proteins showed maximum (modal) abundance in the same fraction. Likewise, the cytoplasmic data set consisted of 33 features per pair: 14 MS and 19 other.

We used a greedy stepwise feature selection algorithm, implemented in Weka, to rank features and selected only the most informative ones, with the specific goal of choosing the single best correlation metric for each particular MS data set. It was observed that, after the first of the large-scale repeat MS experiments was folded into the classifier, the second repeat added little information and ranked poorly. To rescue these data, we merged the four largest repeats by addition and recalculated the noise model and weighted cross correlation scores for these four data sets. Performing feature selection on these data yielded 22 top-performing, non-duplicated features for the cytoplasmic data and 25 features for the nuclear data (Table S2). Predictions were generated for these sets using the Fast Random Forest classifier in Weka and a combined score was generated for each pair by taking one minus the product of one minus the posterior probability of the pair interacting, as predicted by the classifier. For pairs that appeared in only one data set, that data set's posterior probability was used. Applying the classifier to all pairs which had a correlation measure greater than 0.5 in any one MS data set yielded 817,179 protein pairs, of which 48,915 had posterior probability > = 0.5. Notably, incorporation of the complementary genomic evidence boosted the recall of PPI beyond that from the mass spectrometry evidence alone, across a wide range of predictive precision, e.g., increasing recall by ~20% at a cumulative precision of 0.7. The improvement shown by the final version of the data is shown in the main text in Figure 2C.

Denoising the Inferred Protein-Protein Interactions

We developed a procedure that exploits the network topology and protein co-localization information in order to further reduce the amount of noise in the inferred protein-protein interaction network and to filter it prior to discovering protein complexes.

We first delete the connections in the interaction network for which there is little evidence according to the network topology. The rationale here is that if two proteins belong to the same complex, they should be well connected to each other through many short paths in the graph. Diffusion methods over random graphs have previously been employed to quantify the amount of connectivity existing between two nodes in a graph (Coifman et al., 2005; Paccanaro et al., 2006).

Here we use a multiple-step diffusion which calculates the connectivity between proteins *i* and *j* as the *(i,j)* element of the matrix:

$$e^{\lambda \cdot M} - \lambda \cdot M$$

where *M* is the 5,549 × 5,549 matrix whose entries are the output of the random forest classifiers, and λ is the inverse of the maximal eigenvalue of M. Edges with diffusion values lower than *5E-05* are then deleted from the original graph. We shall indicate this new network with D.

Second, we calibrate the resulting graph using protein co-localization information.

To do this we combine the output of the previous step with the GO-CC (Harris et al., 2004) normalized semantic similarity scores with the assumption that they are independent. The rationale here is that two proteins located in different cellular locations should not interact. The final score for each link is thus given by:

$$1 - (1 - D(i,j)) \cdot \left(1 - \frac{Sim(i,j)}{MS}\right)$$

where *Sim(i,j)* is the maximum of the pairwise similarities between the two groups of GO-CC terms to which protein i and protein j are annotated, and MS is the maximum value among all the semantic similarity scores. In our calculations, for the semantic similarities we used an improved version of the Resnik semantic similarity measure (Resnik, 1999) that we have recently proposed (Yang et al., 2012) and is able to take into account the ontology beneath the GO terms and to model uncertainty.

Note that, among 5,549 proteins, there are 1,790 proteins that are not annotated in GO-CC. Therefore for these proteins we simply used D (output of the first step), as this (second) step cannot be applied to unannotated proteins. When considering the GO-CC annotation we discarded those with evidence codes NR, IEA, and ND.

Scores below a threshold of 0.55 were set to zero. The resulting denoised Protein-Protein Interactions graph contains 13,993 interactions (3,006 proteins) at an estimated 21.5% FDR. The effectiveness of the denoising procedure can be seen in a precision-recall curve for the network after denoising obtained by varying the threshold over the network weights and using as gold standard the CORUM database of curated mammalian protein complexes described earlier (Figure 2F).

Clustering of the Denoised PPI Network to Discover Protein Complexes

Protein complexes appear as densely connected regions within the de-noised interaction network. Because a protein may belong to multiple complexes, these densely connected regions may overlap. To elucidate such overlapping sets in our network, we used an algorithm that we have recently proposed, named ClusterONE (Clustering with Overlapping Neighborhood Expansion) (Nepusz et al., 2012). ClusterONE finds complexes by growing multiple clusters from seed proteins, independently of each other. The growth of a putative complex is governed by a greedy rule that tries to maximize the cohesiveness of the complex. The cohesiveness of a complex *C* is defined as follows:

$$\frac{w_{in}}{w_{in} + w_{out} + p|C|}$$

where W_{in} is the total weight of connections within *C*, W_{out} is the total weight of interactions connecting the complex with the rest of the network and $|C|$ is the size of the complex. *p* is a penalty constant that accounts for the possibility of uncharted connections in the network as it assumes *p* extra external connections for the complex for every protein involved. In each step of the growth process, we add a new adjacent protein to the complex or remove an already added protein in a way that yields the maximal increase in cohesiveness. The growth process stops when it is not possible to increase the cohesiveness further. At this stage, the cluster is declared a protein complex candidate if its density is above a given density threshold *d*, and the growth process restarts from a different seed. The first seed is the protein with the largest total weight on its incident connections (*i.e.*, the protein with the most confident set of interactions), and subsequent seeds are always selected in a similar manner but excluding proteins that have already been added to some protein complex candidate. Because the growth processes are independent of each other, the calculated complexes may overlap. More details on ClusterONE can be found in Nepusz et al. (2012). The algorithm has two main parameters: the penalty *p* and the density threshold *d*. The settings for these parameters were chosen to yield the highest Maximum Matching Ratio (Nepusz et al., 2012) on the cluster-training complex subset (see above). These were p = 2.9 and *d* = 0.4 and used to derive the final set of complexes.

To evaluate the overlap of the predicted complexes with the CORUM complexes, we calculated: (1) the number of CORUM complexes matching at least one predicted complex by a matching score greater than 0.25 (matching score = size of intersection squared, divided by the product of the two complexes sizes, as defined by (Bader and Hogue, 2003), (2) the Maximum Matching Ratio, (Nepusz et al., 2012), calculated by matching each predicted complex to at most one reference complex and vice versa, while maximizing the total matching score between them (with the theoretical maximum of 1.0 considered as a perfect match), (3) geometric accuracy as defined by (Brohée and van Helden, 2006) (square-root of the product of positive predictive value and clustering-wise sensitivity). The predicted complexes showed better correspondence with the CORUM catalog of reference human protein complexes than the results of other popular methods, including MCODE, MCL, CMC and RNSC (see Table S5). Applying ClusterONE to our denoised network, we obtained a set of 771 complexes. We then further filtered this set using the same procedure that we had applied to the CORUM set, which combined complexes sharing subunits (Simpson coefficient >0.5 between complexes). This produced our final set of 622 protein complexes.

Enrichment Analysis of Protein Pairs with Shared Annotations

To evaluate interacting and co-complexed protein pairs, we collected the following large-scale sets of protein-protein interactions: 1,991 co-complex interactions related to chromosome segregation (Hutchins et al., 2010); 17,775 "co-regulator" interactions identified through affinity purification and mass spectrometry-based methods (Malovannaya et al., 2011); and 209,913 interactions from

a *D. melanogaster* co-complex interaction network (Guruharsha et al., 2011). In addition, we collected the following sets of gene annotations: three available sets of 1,023, 3,563, and 114,477 human disease-gene associations (Becker et al., 2004; Hamosh et al., 2005; UniProt Consortium, 2011), 2,065 gene-mitotic phenotype associations (Hutchins et al., 2010; Neumann et al., 2010), curated sets of 74,250 mouse, 86,383 yeast, and 27,065 worm gene-phenotype associations assembled in (McGary et al., 2010), upstream transcription factor regulatory motifs for 265,270 genes (Xie et al., 2005), and a set of 869 essential genes collected from (Amsterdam et al., 2004; Blake et al., 2011; Harborth et al., 2001; Kittler et al., 2004; Silva et al., 2008).

We tested whether protein interaction partners are enriched for having common functional or phenotypic associations. That is, are protein pairs which are predicted to interact significantly more likely to share annotations? For each annotation set, we calculated the total number of protein pairs sharing annotations in the space of all possible pairs formed from the background set of annotated proteins detectable through our experimental procedures. We compared this "expected" fraction of pairs with shared annotations with the "observed" fraction of interaction partners with shared annotations. To measure the significance of the observed fraction, we obtained a *p*-value from the following hypergeometric test:

$$p(x \geq k) = \sum_{x=k}^{\min(n,m)} \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}, \quad (1)$$

where *N* is the number of possible annotated pairs, *m* is the number of possible pairs with shared annotation, *n* is the number of annotated interaction partners, and *k* is the number of interaction partners with shared annotation. In the case of testing for essentiality enrichment, we used the complete set of possible proteins pairs. Enrichments were additionally confirmed (data not shown) with two empirical *p*-values by calculating shared annotation fractions from 10,000 random trials, in which we (1) drew random protein partners from the background protein set and (2) shuffled the protein labels on the predicted protein interaction map. Lastly, we repeated the analysis for protein edges implied in our predicted protein cluster sets.

Tissue/Cell Line Specificity of Protein Complexes

It is important to note that HeLa cells were sampled in our profiling pipeline much more deeply than HEK, for which only nuclear fractionations were performed. Nevertheless, we examined the abundance of the interacting human proteins in HEK293 and HeLa cell lines on the basis of publicly available next-gen RNA sequencing data for both HeLa versus HEK293, well aware of the fact that mRNA expression levels may not necessarily reflect protein abundance. Considering all IEX MS experiments, HeLa proteins are discovered at slightly higher rates than those expressed at the same level in HEK293 (Figure S2B). We can clearly distinguish the few proteins that show differential tissue expression e.g., unique to one cell line. Among proteins assigned to complexes, only 82 show HeLa-specific expression and 11 HEK-specific expression (i.e., difference in Log2 (fpkm) expression > 2), yet these proteins show no preferential assortment into tissue-specific complexes.

The distribution of potentially tissue-specific proteins in complexes may reflect possible false positives arising from our analysis but is readily explained as a consequence of the false negative rate of protein detection, due to under-sampling by LC-MS. Hence, we directly examined the reproducibility of our fractionation/mass spectrometry data across biological replicates of the two cell lines, comparing MS1 intensities versus MS2 spectral counting as alternate methods of quantification. Moreover, it is worth noting that we find no evidence for stronger sampling biases in either proteome beyond what is to be expected for mass spectrometry in general. At the level of predicted PPI (which are derived from multiple biochemical fractions), we find that differences in the proteomic measurements generated for the two cell lines (again, in which HeLa was sampled far more extensively, particularly with regards to cytoplasmic extracts) lie within the variance actually observed between biological replicates of the same cell line (Figures S1 and S2).

The conclusion that the complexes we report are likely ubiquitous is supported by the expression of protein complex subunits across different tissues. For example, the Mann group surveyed the proteomes of 11 cancer cell lines; proteins in our complexes are generally found in all 11 lines (Figure S3A). Moreover, across 16 healthy human tissues for which RNA-seq data is available (EBI accession number E-MTAB-513), we find our complexed proteins to be highly and invariantly expressed (Figure S3B). Across 17,927 confirmed protein-coding genes detected in any of the 16 tissues, the median standard deviation of gene expression is 1.30, while for the 11,325 genes detected in all 16 tissues (63% of the total) it is 0.90. The standard deviation of genes we assign to protein complexes is 0.73; among these proteins, 91% are detected in all 16 healthy tissues. Thus the protein complexes described here exhibit largely invariant expression across the tissues sampled in the RNA-seq study.

Enrichment Analysis of Protein Clusters with Particular Phenotype Associations

We tested whether predicted protein clusters are enriched for particular human, mouse, or worm gene-phenotype associations. The significance of members of a cluster sharing a particular phenotype was determined by the hypergeometric probability, as above, where *N* is the number of annotated proteins in the background protein set, *m* is the number of proteins annotated with the queried phenotype, *n* is the number of annotated proteins in the cluster, and *k* is the number of proteins in the cluster annotated by the queried phenotype.

Cross-Validations with Curated Complexes in Public Databases and Independent Studies

We compared our network of complexes to curated complexes in 5 public databases, including CORUM (Ruepp et al., 2010), REACTOME (Haw et al., 2011), PINdb (Luc and Tempst, 2004), and HPRD (Prasad et al., 2009) databases, and specified complexes

within the GO cellular component category (Ashburner et al., 2000) to assess the agreement between our complexes and the literature. Statistically significant overlap between complexes was evaluated using the Fisher's exact test for hypergeometric distribution and the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct for multiple testing (estimated false discovery rate ≤ 0.05), with a minimum of 2 shared subunits. Next, we validated putative new complexes (i.e., not curated in the above public repositories) through comparison with recently published independent co-affinity purification data (Guruharsha et al., 2011; Malovannaya et al., 2011). In particular, we accessed the recent human protein interaction results of Guruharsha et al. (2011). This group performed affinity-tag pull-down experiments for human proteins present in 41 of our complexes. Overall, of the 299 relevant human bait-prey interactions reported, 143 likewise occur within our complexes, representing a 47.8% validation rate. This agreement is comparable to the 63.8% validation rate they claim for their own complex predictions, and is probably an underestimate because they don't report all the proteins actually detected by mass spectrometry, but rather only human proteins with orthologs in their initial *Drosophila* PPI network. The matched clusters are reported in Table S3.

We also compared our complexes with the results of Malovannaya et al. (2011), which verified a total of 127 of our complexes (i.e., clusters show a Simpson matching coefficient > 0.5 between studies), including 42 (33%) of our complexes that are not curated in CORUM. These matched complexes are listed in Table S3. Taken together, these analyses represent a nearly 40% validation rate and strongly argue for the high fidelity of the mapped complexes.

Conservation of Complexes across Model Organisms

To examine to what extent human protein complexes identified in this study have known counterparts in yeast and fly, we considered the set of 720 multi-protein complexes in *S. cerevisiae* identified in a recent study (Babu et al., 2012) and the 556 complexes recently derived for *D. melanogaster* (Guruharsha et al., 2011). Both sets of complexes were identified using AP/MS techniques. Briefly, human complexes were converted into an ortholog representation by mapping, whenever possible, the components of each complex to their orthologs in yeast and fly, respectively. Using the ortholog representation of individual complexes, we then searched for the most statistically significant match between this representation and all known complexes from the corresponding organism. The process was also repeated in the opposite direction, mapping model-organism complexes onto the human collection in order to identify reciprocally best matches. Statistical significance was established using the Fisher's exact test for hypergeometric distribution and the Benjamini-Hochberg method (Benjamini and Hochberg, 1995) to correct for multiple testing (estimated false discovery rate ≤ 0.05). Orthology relationships for human, yeast and fruit fly were derived from two well established sources: the InParanoid 7.0 (Ostlund et al., 2010) and Ensembl Compara (Vilella et al., 2009). The latter includes both the current Ensembl release 64 (ftp://ftp.ensembl.org/pub/release-64/mysql/ensembl_compara_64/) and Ensembl Genomes release 11 (ftp://ftp.ensemblgenomes.org/pub/pan_ensembl/release-11/mysql/ensembl_compara_pan_homology_11_64/). The Ensembl IDs from Compara were mapped using BioMart Perl API (http://www.biomart.org/martservice.html). In addition, we extended the human-to-yeast orthology map by matching human and yeast genes that share a common fly ortholog.

Coevolution

For the calculation of coevolution scores, we used the program MatrixMatchMaker (MMM) (Clark et al., 2011; Tillier and Charlebois, 2009). Orthologous protein sequence clusters were obtained from the OMA Database (Schneider et al., 2007) to obtain 204,689 eukaryotic groups that span 96 species, of which 20,800 contained human orthologs. The groups containing a human protein and at least 10 orthologous sequences were aligned using MAFFT (Katoh et al., 2005) and distance matrices were obtained by using protdist from PHYLIP (Felsenstein, 2005) with the PMB distance matrix (Veerassamy et al., 2003) to correct for multiple substitutions. We ran MMM in an all-by-all manner with a selected tolerance of 0.1 (10%) and chose to use taxon information such that only sequences from the same species could be matched.

Relative Evolutionary Rate

An average matrix was obtained by averaging the distance matrix entries over all of the OMA groups' matrices. We used the average matrix to compute the relative rate of an OMA group's evolution, as the ratio of its rate (average distance to the human ortholog) over the average matrix's rate for the same subset of species pairs. Values greater than 1 are proteins that are evolving faster than average, whereas values less than one indicate more slowly evolving proteins.

Evolutionary Age

The distribution of species present in the OMA orthologous groups determined the ancestral node in the phylogenetic tree of all eukaryotic species. The evolutionary distance from the human sequence to this last common ancestral node was then calculated and, in the case of complexes, averaged over the proteins in the complex. This gives an approximate evolutionary origin of the human orthologs.

Interaction Database and PPI Orthology

All OMA proteins were assigned ROGiDs based on their amino acid sequence. These IDs were then used to identify the known physical (or inferred by the author) protein-protein interactions from the iRefIndex database (Razick et al., 2008), which combines protein interaction data from multiple public databases: BIND, BioGRID, CORUM, DIP, HPRD, IntAct, MINT, MPact, MPPI and OPHID. Human protein interaction data were also downloaded from most of these public databases and some other online available resources independently. These databases / resources included BioGRID (Stark et al., 2011), DIP (Salwinski et al., 2004), MINT (Ceol et al., 2010), HPRD (Prasad et al., 2009), INTACT (Aranda et al., 2010), NCBI Gene (http://www.ncbi.nlm.nih.gov/gene/), CORUM (Ruepp et al., 2010) and the Human interactome database (Rual et al., 2005). Orthology of the PPIs was then determined using the species distribution of the OMA groups.

Approximation of Subunit Stoichiometries

The relative stoichiometries of interacting proteins were approximated from their associated mass spectral MS/MS counts as follows: For each pair of interacting proteins, we considered all biochemical fractions in which both proteins were observed, and calculated relative stoichiometries for interacting protein pairs observed together in at least 10 fractions. Their relative stoichiometry was estimated as the median (across the fractions) of the ratios of their MS/MS spectral counts divided by their expected ratios of spectral counts given the proteins' differences in numbers of potential tryptic peptides. This was calculated as:

$$Stoichiometry = median\left(\frac{c_{1,j}/e_1}{c_{2,j}/e_2}\right)$$

where $c_{1,j}$ and $c_{2,j}$ are the spectral counts of protein 1 and protein 2 in fraction j (out of n), and e_1 and e_2 are the numbers of potential tryptic peptides for proteins 1 and 2, respectively, calculated using the same parameters as in the initial identification of proteins from the raw mass spectrometry data (e.g., considering up to one missing tryptic cleavage and employing the same spectral lookup database). Stoichiometries estimated by this approach between ribosomal subunits and between core proteasomal subunits were consistent with the expected 1:1 ratios, as shown in Figure 7.

Evaluating Potential Bias

We evaluated our final complexes for possible biases toward hydrophobic or low abundant proteins, underrepresented organelles, and complex size–considerations that address some of the technical limitations of our approach. By design, insoluble membrane-associated (hydrophobic) protein complexes were largely missed in this study. Consistent with other proteomics studies, our data are biased toward highly expressed genes (Figure S2B). Our protein complexes are preferentially enriched for water-soluble nuclear and cytosolic proteins (Benjamini-corrected $p \leq 10^{-52}$ and $p \leq 10^{-12}$, respectively), which nevertheless cover a wide spectrum of biological functions (as judged by enrichment for diverse functional annotation terms).

We also compared both the isoelectric point (pI) and subunit memberships of our predicted protein complexes versus those reported in the CORUM database. To this end, we first minimized the inflated number of redundant protein complexes in CORUM by merging complexes with similar annotated subunit compositions but reported by different authors. We then integrated protein complexes with Simpson coefficients > 0.5 to deduce a consolidated non-redundant set of 734 curated protein complexes ranging from 2 to 142 (spliceosome) annotated protein subunits per complex. As shown in Figure S4B, we do not observe significant bias toward negatively ($pI \leq 7$) or positively ($pI \geq 7$) charged protein complexes in our data set as compared to CORUM.

Our clustering strategy, ClusterONE, underweights small clusters of size 2 or 3 in an effort to control the false positive rate, resulting in a peak of clusters at size = 4 subunits as evident in Figure 3A in the main text. Despite this apparent bias, ClusterONE outperformed the competing clustering algorithms we tested against the independent benchmark set of reference complexes, as detailed above and summarized in Table S5. In practice, we find that most competing algorithms yield an exceptionally large number of small clusters, for which it is difficult to establish meaningful measures of accuracy. Nevertheless, although our informatic approach yields complexes with a biased size distribution, overall our complexes show demonstrably good performance against the reference sets noted in the text.

SUPPLEMENTAL REFERENCES

Amsterdam, A., Nissen, R.M., Sun, Z., Swindell, E.C., Farrington, S., and Hopkins, N. (2004). Identification of 315 genes essential for early zebrafish development. Proc. Natl. Acad. Sci. USA *101*, 12792–12797.

Andersen, J.S., Lyon, C.E., Fox, A.H., Leung, A.K., Lam, Y.W., Steen, H., Mann, M., and Lamond, A.I. (2002). Directed proteomic analysis of the human nucleolus. Curr. Biol. *12*, 1–11.

Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., Bridge, A., Derow, C., Feuermann, M., Ghanbarian, A.T., Kerrien, S., Khadake, J., et al. (2010). The IntAct molecular interaction database in 2010. Nucleic Acids Res. *38* (*Database issue*), D525–D531.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics *4*, 2.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc., B *57*, 289–300.

Blake, J.A., Bult, C.J., Kadin, J.A., Richardson, J.E., and Eppig, J.T.; Mouse Genome Database Group (2011). The Mouse Genome Database (MGD): premier model organism resource for mammalian genomics and genetics. Nucleic Acids Res. *39* (*Database issue*), D842–D848.

Brohée, S., and van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. BMC Bioinformatics *7*, 488.

Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L., and Cesareni, G. (2010). MINT, the molecular interaction database: 2009 update. Nucleic Acids Res. *38* (*Database issue*), D532–D539.

Clark, G.W., Dar, V.U., Bezginov, A., Yang, J.M., Charlebois, R.L., and Tillier, E.R. (2011). Using coevolution to predict protein-protein interactions. Methods Mol. Biol. *781*, 237–256.

Coifman, R.R., Lafon, S., Lee, A.B., Maggioni, M., Nadler, B., Warner, F., and Zucker, S.W. (2005). Geometric diffusions as a tool for harmonic analysis and structure definition of data: diffusion maps. Proc. Natl. Acad. Sci. USA *102*, 7426–7431.

de Gelder, R., Wehrens, R., and Hageman, J.A. (2001). A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. J. Comput. Chem. *22*, 273–289.

Dignam, J.D., Lebovitz, R.M., and Roeder, R.G. (1983). Accurate transcription initiation by RNA polymerase II in a soluble extract from isolated mammalian nuclei. Nucleic Acids Res. *11*, 1475–1489.

Dunham, W.H., Larsen, B., Tate, S., Badillo, B.G., Goudreault, M., Tehami, Y., Kislinger, T., and Gingras, A.C. (2011). A cost-benefit analysis of multidimensional fractionation of affinity purification-mass spectrometry samples. Proteomics *11*, 2603–2612.

Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Department of Genomic Sciences, University of Washington.

Harborth, J., Elbashir, S.M., Bechert, K., Tuschl, T., and Weber, K. (2001). Identification of essential genes in cultured mammalian cells using small interfering RNAs. J. Cell Sci. *114*, 4557–4565.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., et al.; Gene Ontology Consortium (2004). The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. *32* (Database issue), D258–D261.

Havugimana, P.C., Wong, P., and Emili, A. (2006). Enhanced proteomic analysis by HPLC prefractionation. In Handbook of Pharmaceutical Biotechnology, S.C. Gad, ed. (Hoboken, NJ: John Wiley & Sons), pp. 1491–1501.

Haw, R.A., Croft, D., Yung, C.K., Ndegwa, N., D'Eustachio, P., Hermjakob, H., and Stein, L.D. (2011). The Reactome BioMart. Database (Oxford) *2011*, bar031.

Katoh, K., Kuma, K., Miyata, T., and Toh, H. (2005). Improvement in the accuracy of multiple sequence alignment program MAFFT. Genome Inform *16*, 22–33.

Kittler, R., Putz, G., Pelletier, L., Poser, I., Heninger, A.K., Drechsel, D., Fischer, S., Konstantinova, I., Habermann, B., Grabner, H., et al. (2004). An endoribonuclease-prepared siRNA screen in human cells identifies genes essential for cell division. Nature *432*, 1036–1040.

Luc, P.V., and Tempst, P. (2004). PINdb: a database of nuclear protein complexes from human and yeast. Bioinformatics *20*, 1413–1415.

McGary, K.L., Park, T.J., Woods, J.O., Cha, H.J., Wallingford, J.B., and Marcotte, E.M. (2010). Systematic discovery of nonobvious human disease models through orthologous phenotypes. Proc. Natl. Acad. Sci. USA *107*, 6544–6549.

Morin, R., Bainbridge, M., Fejes, A., Hirst, M., Krzywinski, M., Pugh, T., McDonald, H., Varhol, R., Jones, S., and Marra, M. (2008). Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing. Biotechniques *45*, 81–94.

Ostlund, G., Schmitt, T., Forslund, K., Köstler, T., Messina, D.N., Roopra, S., Frings, O., and Sonnhammer, E.L. (2010). InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. *38* (*Database issue*), D196–D203.

Paccanaro, A., Casbon, J.A., and Saqi, M.A. (2006). Spectral clustering of protein sequences. Nucleic Acids Res. *34*, 1571–1580.

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol. Biol. *577*, 67–79.

Razick, S., Magklaras, G., and Donaldson, I.M. (2008). iRefIndex: a consolidated protein interaction database with provenance. BMC Bioinformatics *9*, 405.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. *11*, 95–130.

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

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

Schneider, A., Dessimoz, C., and Gonnet, G.H. (2007). OMA Browser—exploring orthologous relations across 352 complete genomes. Bioinformatics *23*, 2180–2182.

Selbach, M., Schwanhäusser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58–63.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617–620.

Stark, C., Breitkreutz, B.J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X., Shi, X., et al. (2011). The BioGRID Interaction Database: 2011 update. Nucleic Acids Res. *39* (*Database issue*), D698–D704.

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

Tabb, D.L., McDonald, W.H., and Yates, J.R., III (2002). DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. J. Proteome Res. *1*, 21–26.

Veerassamy, S., Smith, A., and Tillier, E.R. (2003). A transition probability model for amino acid substitutions from blocks. J. Comput. Biol. *10*, 997–1010.

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., and Birney, E. (2009). EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. *19*, 327–335.

Wiśniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

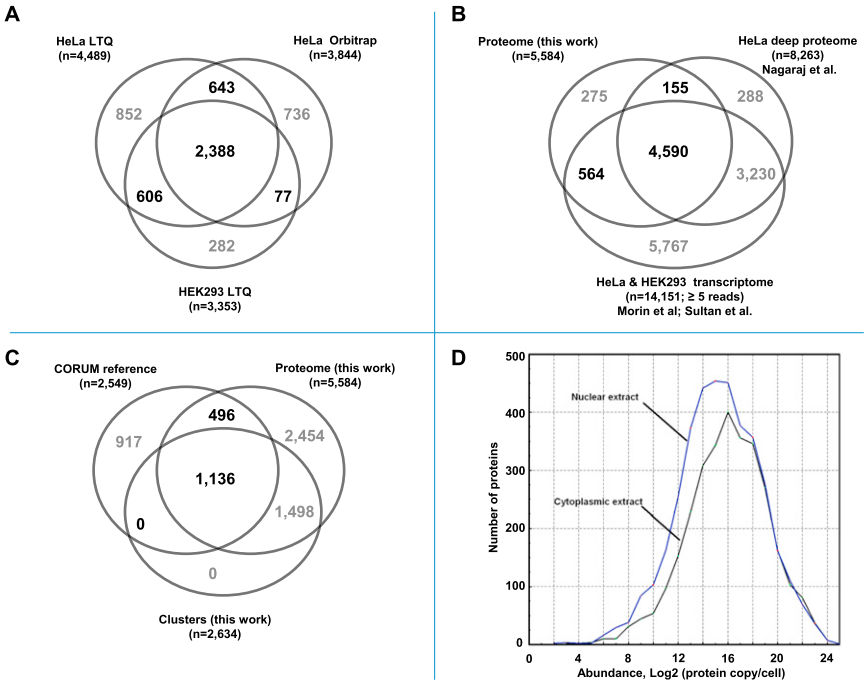


Figure S1. Assessment of LC-MS/MS Protein Detection Bias, Related to Figure 1 and Table S1

(A) Approximately 5% of proteins are unique to HEK cells (most likely to technical variations or sampling).

(B) Approximately 95% of the proteins identified in this study are supported by mRNA cognate transcript/or proteomic data produced with high resolution mass spectrometer (Nagaraj et al., 2011; Morin et al., 2008; Sultan et al., 2008).

(C) Proteins identified in this study covered 64% of the proteins present in the CORUM reference database.

(D) Deep fractionation allows to enrich and identify low abundance nuclear proteins by LC-MS/MS. Proteins abundances were estimated from recent study of HeLa Proteome by Mann group (Nagaraj et al., 2011).

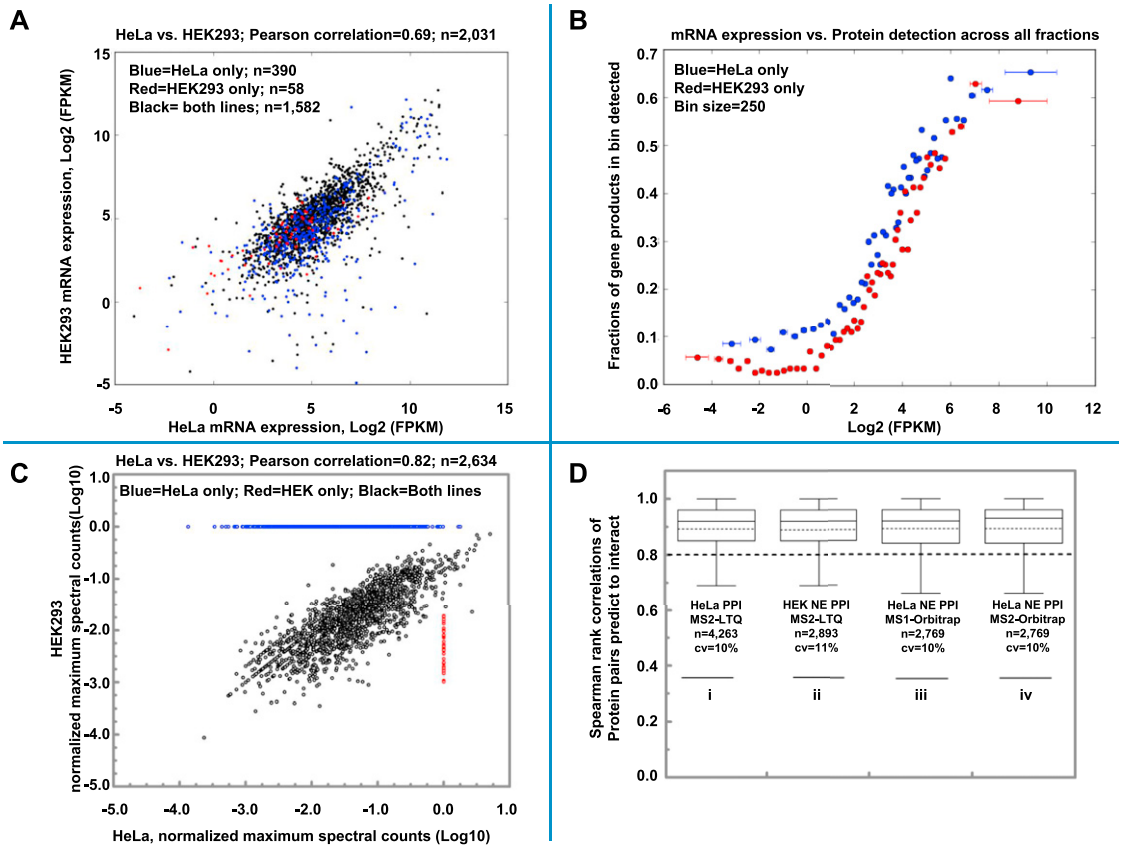


Figure S2. Comparison of HeLa and HEK Protein Profiles, Related to Figures 1 and 4 and Tables S1, S2, and S3

(A) Few proteins detected preferentially in HeLa or HEK293 cells have proportionally higher relative mRNA transcript levels in one of the two cell lines (Morin et al., 2008; Sultan et al., 2008); most show consistent transcript levels in both cell lines. Proteins detected in both cell lines are represented as black dots, and those detected only in HeLa or HEK cells are shown in blue or red, respectively.

(B) Gene products expressed in HeLa (blue) and HEK293 (red) cells (Morin et al., 2008; Sultan et al., 2008) were rank-ordered by mRNA-seq abundance level (log2(fpkm)) and binned (bin size = 250). For each bin, the fraction of gene products detected across all IEX fractionation experiments is plotted against the mean (+/− s.d.) expression of genes in the bin. Higher detection rate of HeLa proteins is consistent with deeper sampling of this cell line in our experiments.

(C) Positive correlation (r = 0.82) between HeLa (blue) and HEK (red) proteins assigned in our 622 complexes (2,634 proteins). For each protein in our set of 622 complexes, we retrieved its maximum spectral count across our 1,163 fraction and divided it by its length (i.e., number of amino acids). We then plotted the HEK versus HeLa after logarithmic transformation of the normalized spectral counts. Observed differences in protein detection, particular in HeLa, is mostly due to the protein detected in HeLa cytoplasmic extract.

(D) Box-and-whiskers quartile plots showing the high consistency (profile correlation > 0.8) of the co-fractionation data using different measures of protein abundance (MS2 spectral counts versus MS1 peptide intensities). Data reproducibility was calculated using the Spearman rank correlation coefficients of replicate profiles. Horizontal solid lines mark the minimum, first quartile, median, third quartile and maximum spearman correlation values; black dashed lines mark mean Spearman correlations. High-scoring interacting protein pairs show reproducible HeLa and HEK293 co-elution profiles measured on a linear ion-trap (i and ii, MS2 spectral counts for HeLa and HEK293, respectively) or a high precision Orbitrap instrument (iv, MS2 spectral counts; iii, MS1 peptide intensities based on MaxQuant).

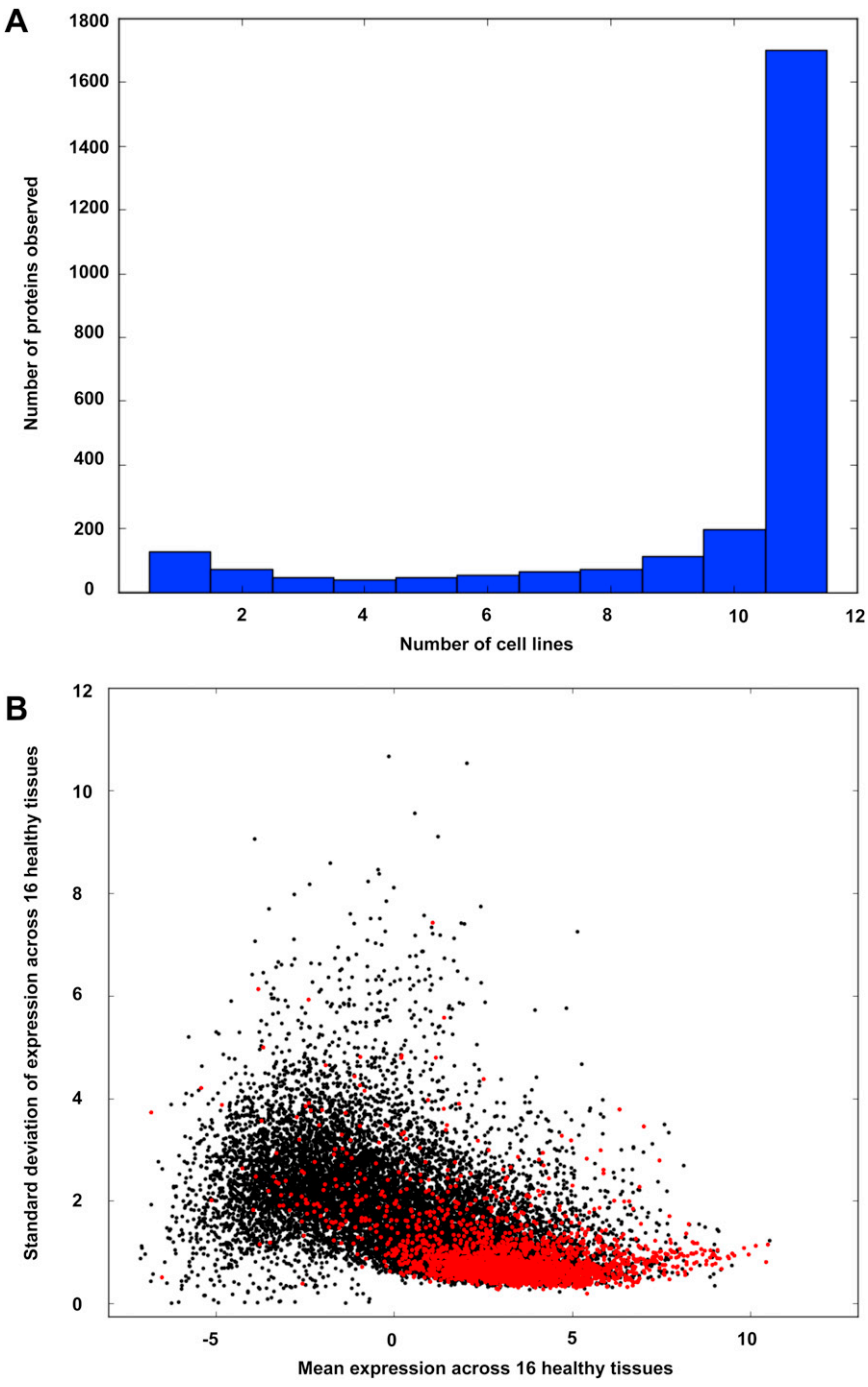


Figure S3. Tissue Expression of Proteins in Complexes, Related to Figure 6

(A) Histogram of number of cancer cell lines in which proteins assigned to our complexes were observed. Data from Mann group proteomic survey of 11 cancer cell lines (Geiger et al., 2012).

(B) Expression levels of RefSeq protein-coding genes across 16 healthy human tissues measured using the Illumina BodyMap 2.0 RNA-seq data (EBI accession E-MTAB-513). Here, mean expression (log2(fpkm)) across all tissues in which a gene product is observed is plotted against the standard deviation of expression: black, all genes; red, subunits assigned to protein complexes in this study. High mean and low variability of expression among protein complex components implies ubiquitous expression.

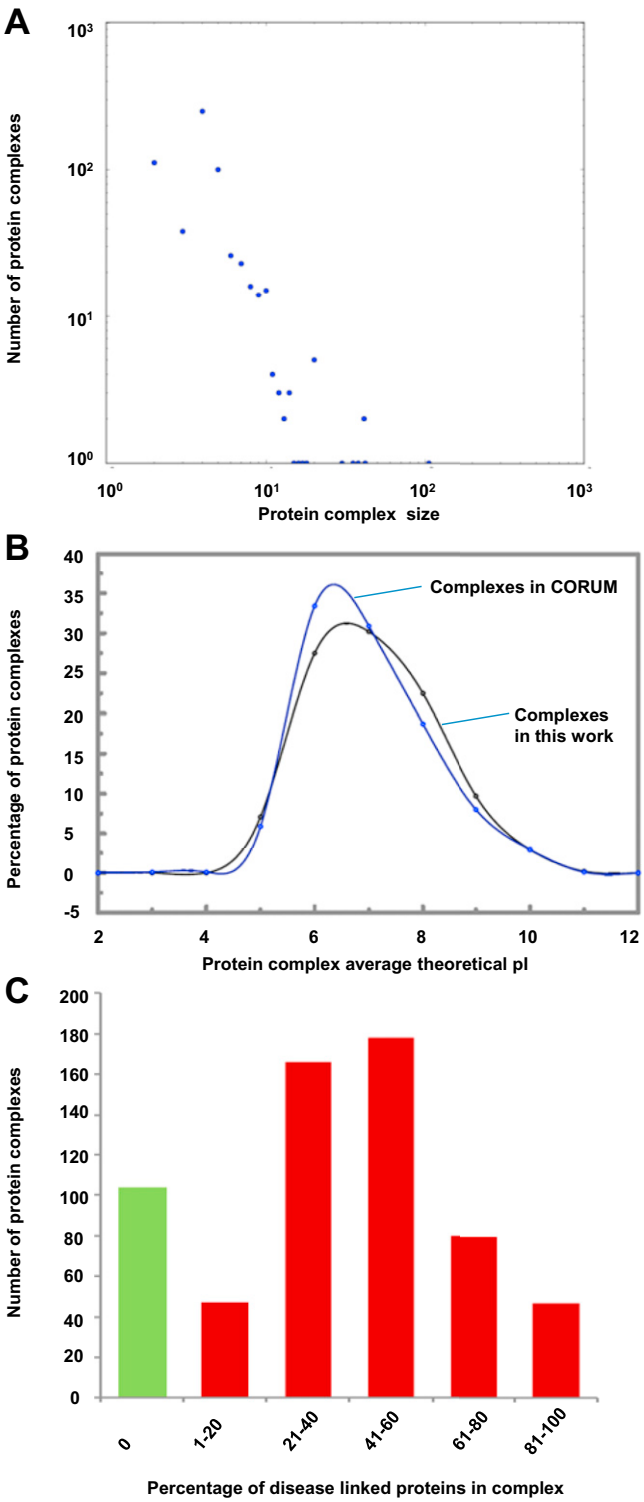


Figure S4. Physical and Biological Properties of our Predicted Human Protein Complexes, Related to Figure 4 and Table S3

(A) Size distribution of mapped protein complexes. The frequency distribution of the number of proteins per complex approximates an inverse power law.

(B) Evaluating bias in complexes. Theoretical pI for each individual protein was calculated using the open source "Compute pI/Mw" tool from the ExPASy (http://web.expasy.org/compute_pi/). To estimate the pI of the protein complex, theoretical pI for individual proteins in complex were averaged and rounded to integer values. Blue; complexes in CORUM reference. Black; complexes derived in this study.

(C) Distribution of annotated disease-associated proteins that are present in our compendium of 622 protein complexes.

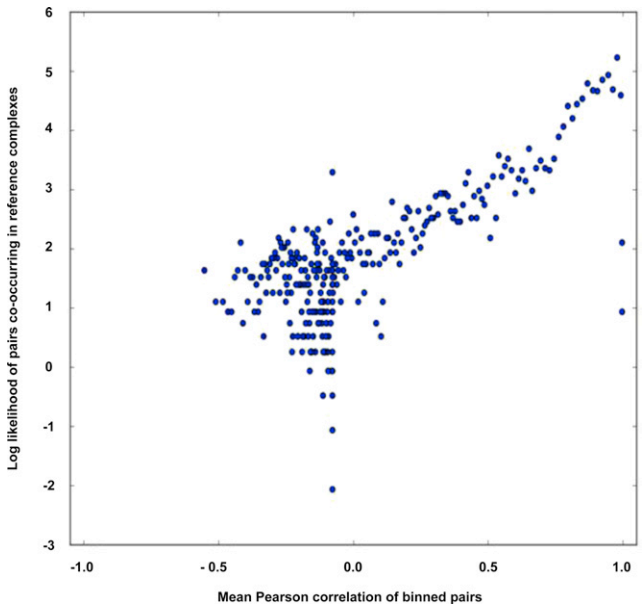


Figure S5. Pearson Correlation between Elution Profiles Breaks Down at High Correlations, Related to Figure 2B

Data from the cytoplasmic fraction of the sucrose gradient MS experiment were analyzed by ranking pairs according to the Pearson correlation coefficient of the normalized elution profiles (*x* axis), binning, and calculating for each bin the log likelihood of containing reference set co-complex protein pairs (*y* axis). Correlation coefficient is predictive of LLS score but breaks down as correlation approaches unity. This drop-off is caused by low-count proteins showing perfect correlations, and was compensated through the use of a Poisson weighted correlation test.