ARTICLES

Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*

Nevan J. Krogan[1,2]*†, Gerard Cagney[1,3]*, Haiyuan Yu[4], Gouqing Zhong[1], Xinghua Guo[1], Alexandr Ignatchenko[1], Joyce Li[1], Shuye Pu[5], Nira Datta[1], Aaron P. Tikuisis[1], Thanuja Punna[1], José M. Peregrín-Alvarez[5], Michael Shales[1], Xin Zhang[1], Michael Davey[1], Mark D. Robinson[1], Alberto Paccanaro[4], James E. Bray[1], Anthony Sheung[1], Bryan Beattie[6], Dawn P. Richards[6], Veronica Canadien[6], Atanas Lalev[1], Frank Mena[6], Peter Wong[1], Andrei Starostine[1], Myra M. Canete[1], James Vlasblom[5], Samuel Wu[5], Chris Orsi[5], Sean R. Collins[7], Shamanta Chandran[1], Robin Haw[1], Jennifer J. Rilstone[1], Kiran Gandi[1], Natalie J. Thompson[1], Gabe Musso[1], Peter St Onge[1], Shaun Ghanny[1], Mandy H. Y. Lam[1,2], Gareth Butland[1], Amin M. Altaf-Ul[8], Shigehiko Kanaya[8], Ali Shilatifard[9], Erin O'Shea[10], Jonathan S. Weissman[7], C. James Ingles[1,2], Timothy R. Hughes[1,2], John Parkinson[5], Mark Gerstein[4], Shoshana J. Wodak[5], Andrew Emili[1,2] & Jack F. Greenblatt[1,2]

Identification of protein–protein interactions often provides insight into protein function, and many cellular processes are performed by stable protein complexes. We used tandem affinity purification to process 4,562 different tagged proteins of the yeast *Saccharomyces cerevisiae*. Each preparation was analysed by both matrix-assisted laser desorption/ionization–time of flight mass spectrometry and liquid chromatography tandem mass spectrometry to increase coverage and accuracy. Machine learning was used to integrate the mass spectrometry scores and assign probabilities to the protein–protein interactions. Among 4,087 different proteins identified with high confidence by mass spectrometry from 2,357 successful purifications, our core data set (median precision of 0.69) comprises 7,123 protein–protein interactions involving 2,708 proteins. A Markov clustering algorithm organized these interactions into 547 protein complexes averaging 4.9 subunits per complex, about half of them absent from the MIPS database, as well as 429 additional interactions between pairs of complexes. The data (all of which are available online) will help future studies on individual proteins as well as functional genomics and systems biology.

Elucidation of the budding yeast genome sequence[1] initiated a decade of landmark studies addressing key aspects of yeast cell biology on a system-wide level. These included microarray-based analysis of gene expression[2], screens for various biochemical activities[3,4], identification of protein subcellular locations[5,6], and identifying effects of single and pairwise gene disruptions[7–10]. Other efforts were made to catalogue physical interactions among yeast proteins, primarily using the yeast two-hybrid method[11,12] and direct purification via affinity tags[13,14]; many of these interactions are conserved in other organisms[15]. Data from the yeast protein–protein interaction studies have been non-overlapping to a surprising degree, a fact explained partly by experimental inaccuracy and partly by indications that no single screen has been comprehensive[16].

Proteome-wide purification of protein complexes

Of the various high throughput experimental methods used thus far to identify protein–protein interactions[11–14], tandem affinity purification (TAP) of affinity-tagged proteins expressed from their natural chromosomal locations followed by mass spectrometry[13,17] has provided the best coverage and accuracy[16]. To map more completely the yeast protein interaction network (interactome), *S. cerevisiae* strains were generated with in-frame insertions of TAP tags individually introduced by homologous recombination at the 3′ end of each predicted open reading frame (ORF) (http://www.yeastgenome.org/)[18,19]. Proteins were purified from 4L yeast cultures under native conditions, and the identities of the co-purifying proteins (preys) determined in two complementary ways[17]. Each purified protein preparation was electrophoresed on an SDS polyacrylamide gel, stained with silver, and visible bands removed and identified by trypsin digestion and peptide mass fingerprinting using matrix-assisted laser desorption/ionization–time of flight (MALDI–TOF) mass spectrometry. In parallel, another aliquot of each purified protein preparation was digested in solution and the peptides were separated and sequenced by data-dependent liquid chromatography tandem mass spectrometry (LC-MS/MS)[17,20–22]. Because either mass spectrometry method often fails to

[1]Banting and Best Department of Medical Research, Terrence Donnelly Centre for Cellular and Biomolecular Research, University of Toronto, 160 College St, Toronto, Ontario M5S 3E1, Canada. [2]Department of Medical Genetics and Microbiology, University of Toronto, 1 Kings College Circle, Toronto, Ontario M5S 1A8, Canada. [3]Conway Institute, University College Dublin, Belfield, Dublin 4, Ireland. [4]Department of Molecular Biophysics and Biochemistry, 266 Whitney Avenue, Yale University, PO Box 208114, New Haven, Connecticut 06520, USA. [5]Hospital for Sick Children, 555 University Avenue, Toronto, Ontario M4K 1X8, Canada. [6]Affinium Pharmaceuticals, 100 University Avenue, Toronto, Ontario M5J 1V6, Canada. [7]Howard Hughes Medical Institute, Department of Cellular and Molecular Pharmacology, UCSF, Genentech Hall S472C, 600 16th St, San Francisco, California 94143, USA. [8]Comparative Genomics Laboratory, Nara Institute of Science and Technology 8916-5, Takayama, Ikoma, Nara 630-0101, Japan. [9]Department of Biochemistry, Saint Louis University School of Medicine, 1402 South Grand Boulevard, St Louis, Missouri 63104, USA. [10]Howard Hughes Medical Institute, Department of Molecular and Cellular Biology, Harvard University, 7 Divinity Avenue, Cambridge, Massachusetts 02138, USA. †Present address: Department of Cellular and Molecular Pharmacology, UCSF, San Francisco, California 94143, USA.
*These authors contributed equally to this work.

identify a protein, we used two independent mass spectrometry methods to increase interactome coverage and confidence. Among the attempted purifications of 4,562 different proteins (Supplementary Table S1), including all predicted non-membrane proteins, 2,357 purifications were successful (Supplementary Table S2) in that at least one protein was identified (in 1,613 cases by MALDI–TOF mass spectrometry and in 2,001 cases by LC-MS/MS; Fig. 1a) that was not present in a control preparation from an untagged strain.

In total, 4,087 different yeast proteins were identified as preys with high confidence (≥99%; see Methods) by MALDI–TOF mass spectrometry and/or LC-MS/MS, corresponding to 72% of the predicted yeast proteome (Supplementary Table S3). Smaller proteins with a relative molecular mass (M_r) of 35,000 were less likely to be identified (Fig. 1b), perhaps because they generate fewer peptides suited for identification by mass spectrometry. We were more successful in identifying smaller proteins by LC-MS/MS than by MALDI–TOF mass spectrometry, probably because smaller proteins stain less well with silver or ran off the SDS gels. Our success in protein identification was unrelated to protein essentiality (data not shown) and ranged from 80% for low abundance proteins to over 90% for high abundance proteins (Fig. 1c). Notably, we identified 47% of the proteins not detected by genome-wide western blotting[18], indicating that affinity purification followed by mass spectrometry can be more sensitive. Many hypothetical proteins not detected by western blotting[18] or our mass spectrometry analyses may not be expressed in our standard cell growth conditions. Although our success rates for identifying proteins were 94% and 89% for nuclear and cytosolic proteins, respectively, and at least 70% in most cellular compartments (Fig. 1d), they were lower (61% and 59%, respectively) for the endoplasmic reticulum and vacuole. However, even though we had not tagged or purified most proteins with transmembrane domains, we identified over 70% of the membrane-associated proteins, perhaps because our extraction and purification buffers contained 0.1% Triton X-100. Our identification success rate was lowest (49%) with proteins for which localization was not established[5,6], many of which may not be expressed. We had high success in identifying proteins involved in all biological processes, as defined by gene ontology (GO) nomenclature, or possessing any broadly defined GO molecular function (Fig. 1e, f). We were less successful (each about 65% success) with transporters and proteins of unknown function; many of the latter may not be expressed.

A high-quality data set of protein–protein interactions

Deciding whether any two proteins interact based on our data must encompass results from two purifications (plus repeat purifications, if performed) and integrate reliability scores from all protein identifications by mass spectrometry. Removed from consideration as likely nonspecific contaminants were 44 preys detected in ≥3% of the purifications and nearly all cytoplasmic ribosomal subunits (Supplementary Table S4). Although the cytosolic ribosomes and pre-ribosomes, as well as some associated translation factors, are not represented in the interaction network and protein complexes we subsequently identified, we previously described the interactome for proteins involved in RNA metabolism and ribosome biogenesis[22].

We initially generated an 'intersection data set' of 2,357 protein–protein interactions based only on proteins identified in at least one purification by both MALDI–TOF mass spectrometry and LC-MS/MS with relatively low thresholds (70%) (Supplementary Table S5). This intersection data set containing 1,210 proteins was of reasonable quality but limited in scope (Fig. 2b). Our second approach added to the intersection data set proteins identified either reciprocally or repeatedly by only a single mass spectrometry method

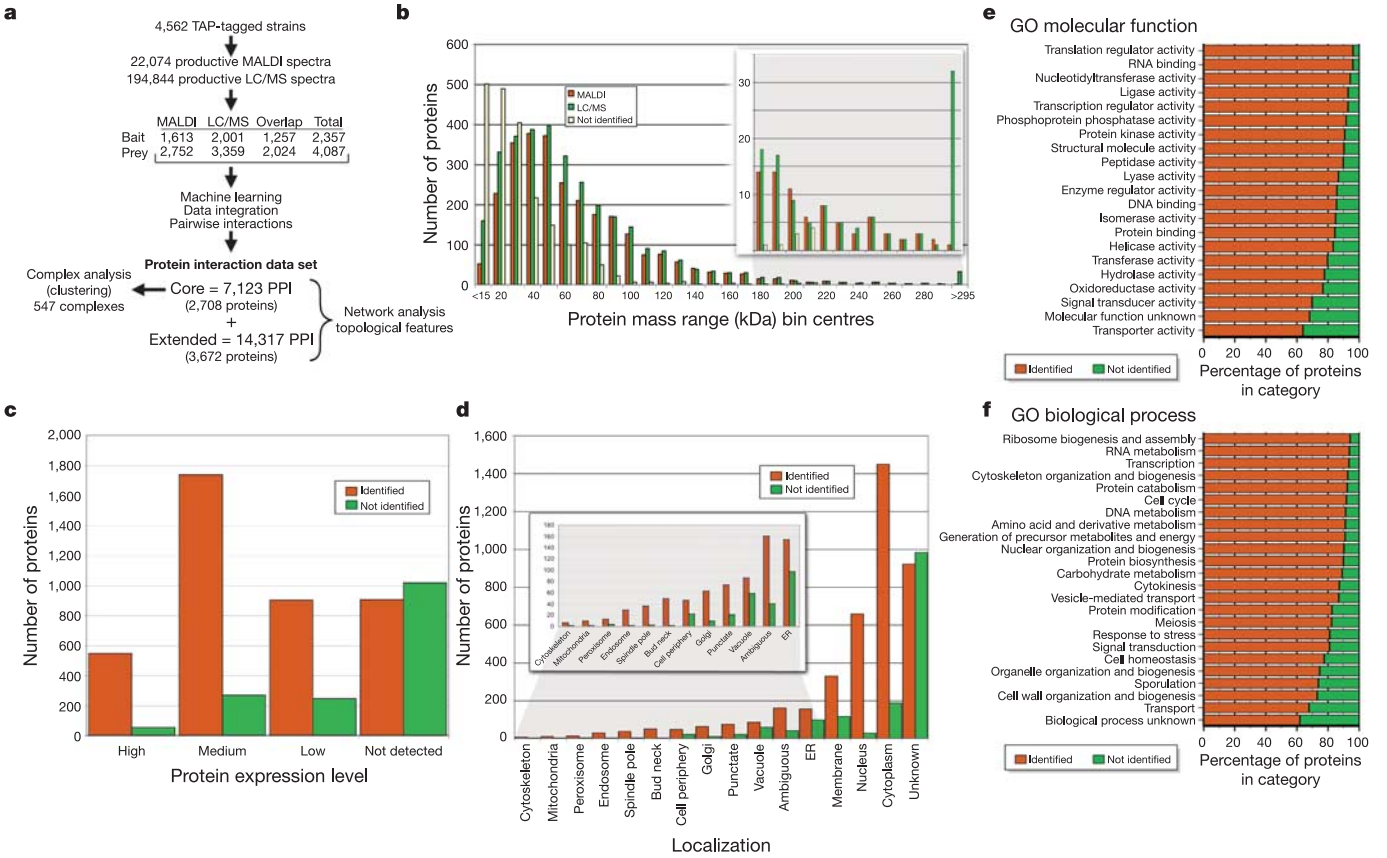

Figure 1 | The yeast interactome encompasses a large proportion of the predicted proteome. **a**, Summary of our screen for protein interactions. PPI, protein–protein interactions. **b–f**, The proportions of proteins identified in the screen as baits or preys are shown in relation to protein mass (**b**), expression level (**c**), intracellular localization (**d**) and annotated GO molecular function (**e**) and GO biological process (**f**).

to generate the 'merged data set'. The merged data set containing 2,186 proteins and 5,496 protein–protein interactions (Supplementary Table S6) had better coverage than the intersection network (Fig. 2b).

To deal objectively with noise in the raw data and improve precision and recall, we used machine learning algorithms with two rounds of learning. All four classifiers were validated by the hold-out method (66% for training and 33% for testing) and ten-times tenfold cross-validation, which gave similar results. Because our objective was to identify protein complexes, we used the hand-curated protein complexes in the MIPS reference database[23] as our training set. Our goal was to assign a probability that each pairwise interaction is true based on experimental reproducibility and mass spectrometry scores from the relevant purifications (see Methods). In the first round of learning, we tested bayesian inference networks and 28 different kinds of decision trees[24], settling on bayesian networks and C4.5-based and boosted stump decision trees as providing the most reliable predictions (Fig. 2a). We then improved performance by using the output of the three methods as input for a second round of learning with a stacking algorithm in which logistic regression was the learner[25]. We used a probability cut-off of 0.273 (average 0.68; median 0.69) to define a 'core' data set of 7,123 protein–protein interactions involving 2,708 proteins (Supplementary Table S7) and a cut-off of 0.101 (average 0.42; median 0.27) for an 'extended' data set of 14,317 protein–protein interactions involving 3,672 proteins (Supplementary Table S8). The interaction probabilities in Supplementary Tables S7 and S8 are likely to be underestimated because the MIPS complexes used as a 'gold standard' are themselves imperfect[26]. We subsequently used the core protein–protein interaction data set to define protein complexes (see below), but the extended data set probably contains at least 1,000 correct interactions (as well as many more false interactions) not present in the core data set.

The complete set of protein–protein interactions and their associated probabilities (Supplementary Table S9) were used to generate a ROC curve with a performance (area under the curve) of 0.95 (Fig. 2b). Predictive sensitivity (true positive rate) or specificity (false positive rate), or both, are superior for our learned data set than for the intersection and merged data sets, each previous high-throughput study of yeast protein–protein interactions[11–14], or a bayesian combination of the data from all these studies[27] (Fig. 2b).

Identification of complexes within the interaction network

In the protein interaction network generated by our core data set of 7,123 protein–protein interactions, the average degree (number of interactions per protein) is 5.26 and the distribution of the number of interactions per protein follows an inverse power law (Fig. 2c), indicating scale-free network topology[28]. These protein–protein interactions could be represented as a weighted graph (not shown) in which individual proteins are nodes and the weight of the arc connecting two nodes is the probability that interaction is correct. Because the 2,357 successful purifications underlying such a graph would represent >50% of the detectably expressed proteome[18], we have typically purified multiple subunits of a given complex. To identify highly connected modules within the global protein–protein interaction network, we used the Markov cluster algorithm, which simulates random walks within graphs[29]. We chose values for the expansion and inflation operators of the Markov cluster procedure that optimized overlap with the hand-curated MIPS complexes[23]. Although the Markov cluster algorithm displays good convergence and robustness, it does not necessarily separate two or more complexes that have shared subunits (for example, RNA polymerases I and III, or chromatin modifying complexes Rpd3C(S) and Rpd3C(L))[30,31].

The Markov cluster procedure identified 547 distinct (non-overlapping) heteromeric protein complexes (Supplementary Table S10), about half of which are not present in MIPS or two previous high-throughput studies of yeast complexes using affinity purification and mass spectrometry (Fig. 3a). New subunits or interacting proteins were identified for most complexes that had been identified previously (Fig. 3a). Overlap of our Markov-cluster-computed complexes with the MIPS complexes was evaluated (see Supplementary Information) by calculating the total precision (measure of the extent to which proteins belonging to one reference MIPS complex are grouped within one of our complexes, and vice versa) and homogeneity (measure of the extent to which proteins from the same MIPS complex are distributed across our complexes, and vice versa) (Fig. 3b). Both precision and homogeneity were higher for the complexes generated in this study—even for the extended set of protein–protein interactions—than for complexes generated by both previous high-throughput studies of yeast complexes, perhaps because the increased number of successful purifications in this study increased the density of connections within most modules. The average number of different proteins per complex is 4.9, but the distribution (Fig. 3c), which follows an inverse power law, is characterized by a large number of small complexes, most often containing only two to four different polypeptides, and a much smaller number of very large complexes.

Proteins in the same complex should have similar function and co-localize to the same subcellular compartment. To evaluate this, we

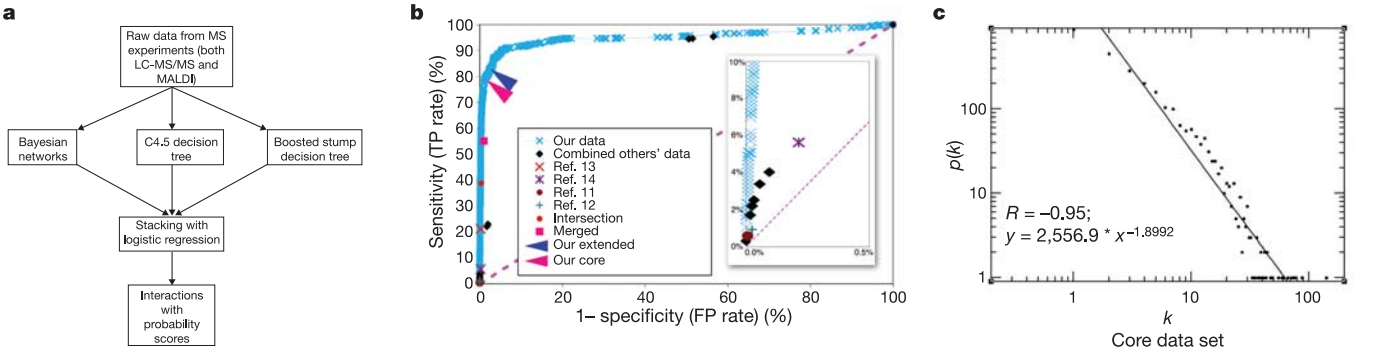

Figure 2 | Machine learning generates a core data set of protein–protein interactions. a, Reliability of observed protein–protein interactions was estimated using probabilistic mass spectra database search scores and measures of experimental reproducibility (see Methods), followed by machine learning. **b**, Precision-sensitivity ROC plot for our protein–protein interaction data set generated by machine learning. Precision/sensitivity values are also shown for the 'intersection' and 'merged' data sets (see text) and for other large-scale affinity tagging[13,14] and two-hybrid[11,12] data sets, and a bayesian networks combination of those data sets[27], all based on comparison to MIPS complexes. FP, false positive; TP, true positive. **c**, Plot of the number of nodes against the number of edges per node demonstrates that the core data set protein–protein interaction network has scale-free properties.

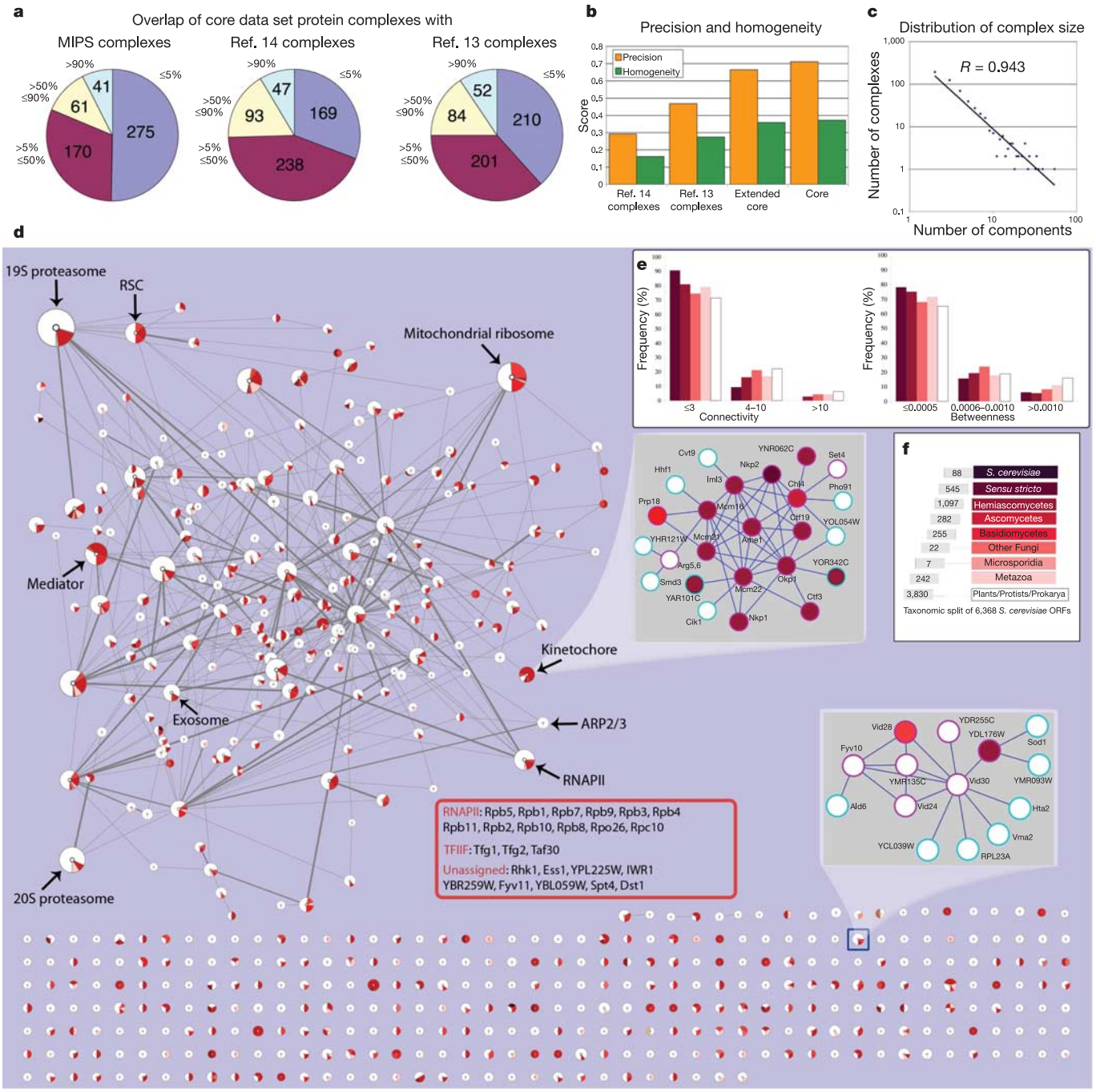

Figure 3 | Organization of the yeast protein–protein interaction network into protein complexes. **a**, Pie charts showing how many of our 547 complexes have the indicated percentages of their subunits appearing in individual MIPS complexes or complexes identified by other affinity-based purification studies[13,14]. **b**, Precision and homogeneity (see text) in comparison to MIPS complexes for three large-scale studies. **c**, The relationship between complex size (number of different subunits) and frequency. **d**, Graphical representation of the complexes. This Cytoscape/GenePro screenshot displays patterns of evolutionary conservation of complex subunits. Each pie chart represents an individual complex, its relative size indicating the number of proteins in the complex. The thicknesses of the 429 edges connecting complexes are proportional to the number of protein–protein interactions between connected nodes. Complexes lacking connections shown at the bottom of this figure have <2 interactions with any other complex. Sector colours (see panel **f**) indicate the proportion of subunits sharing significant sequence similarity to various taxonomic groups (see Methods). Insets provide views of two selected complexes—the kinetochore machinery and a previously uncharacterized, highly conserved fructose-1,6-bisphosphatase-degrading complex (see text for details)—detailing specific interactions between proteins identified within the complex (purple borders) and with other proteins that interact with at least one member of the complex (blue borders). Colours indicate taxonomic similarity. **e**, Relationship between protein frequency in the core data set and degree of connectivity or betweenness as a function of conservation. Colours of the bars indicate the evolutionary grouping. **f**, Colour key indicating the taxonomic groupings (and their phylogenetic relationships). Numbers indicate the total number of ORFs sharing significant sequence similarity with a gene in at least one organism associated with that group and, importantly, not possessing similarity to any gene from more distantly related organisms.

calculated the weighted average of the fraction of proteins in each complex that maps to the same localization categories[5] (see Supplementary Information). Co-localization was better for the complexes in our study than for previous high-throughput studies but, not unexpectedly, less than that for the curated MIPS complexes (Supplementary Fig. S1). We also evaluated the extent of semantic similarity[32] for the GO terms in the 'biological process' category for pairs of interacting proteins within our complexes (Supplementary Fig. S2), and found that semantic similarity was lower for our core data set than for the MIPS complexes or the previous study using TAP tags[13], but higher than for a study using protein overproduction[14]. This might be expected if the previous TAP tag study significantly influenced the semantic classifications in GO.

To analyse and visualize our entire collection of complexes, the highly connected modules identified by Markov clustering for the global core protein–protein interaction network were displayed (http://genepro.ccb.sickkids.ca) using our GenePro plug-in for the Cytoscape software environment[33] (Fig. 3d). Each complex is represented as a pie-chart node, and the complexes are connected by a limited number (429) of high-confidence interactions. Assignment of connecting proteins to a particular module can therefore be arbitrary, and the limited number of connecting proteins could just as well be part of two or more distinct complexes.

The size and colour of each section of a pie-chart node can be made to represent the fraction of the proteins in each complex that maps into a given complex from the hand-curated MIPS complexes (Supplementary Fig. S3). Similar displays can be generated when highlighting instead the subcellular localizations or GO biological process functional annotations of proteins in each complex. Furthermore, the protein–protein interaction details of individual complexes can readily be visualized (see Supplementary Information).

Evolutionary conservation of protein complexes

ORFs encoding each protein were placed into nine distinct evolutionary groups (Fig. 3f) based on their taxonomic profiles (see Methods), and the complexes displayed so as to show the evolutionary conservation of their components (Fig. 3d). Insets highlight the kinetochore complex required for chromosome segregation and a novel, highly conserved complex involved in degradation of fructose-1,6-bisphosphatase. Strong co-evolution was evident for components of some large and essential complexes (for example, 19S and 20S proteasomes involved in protein degradation, the exosome involved in RNA metabolism, and the ARP2/3 complex required for the motility and integrity of cortical actin patches). Conversely, the kinetochore complex, the mediator complex required for regulated transcription, and the RSC complex that remodels chromatin have a

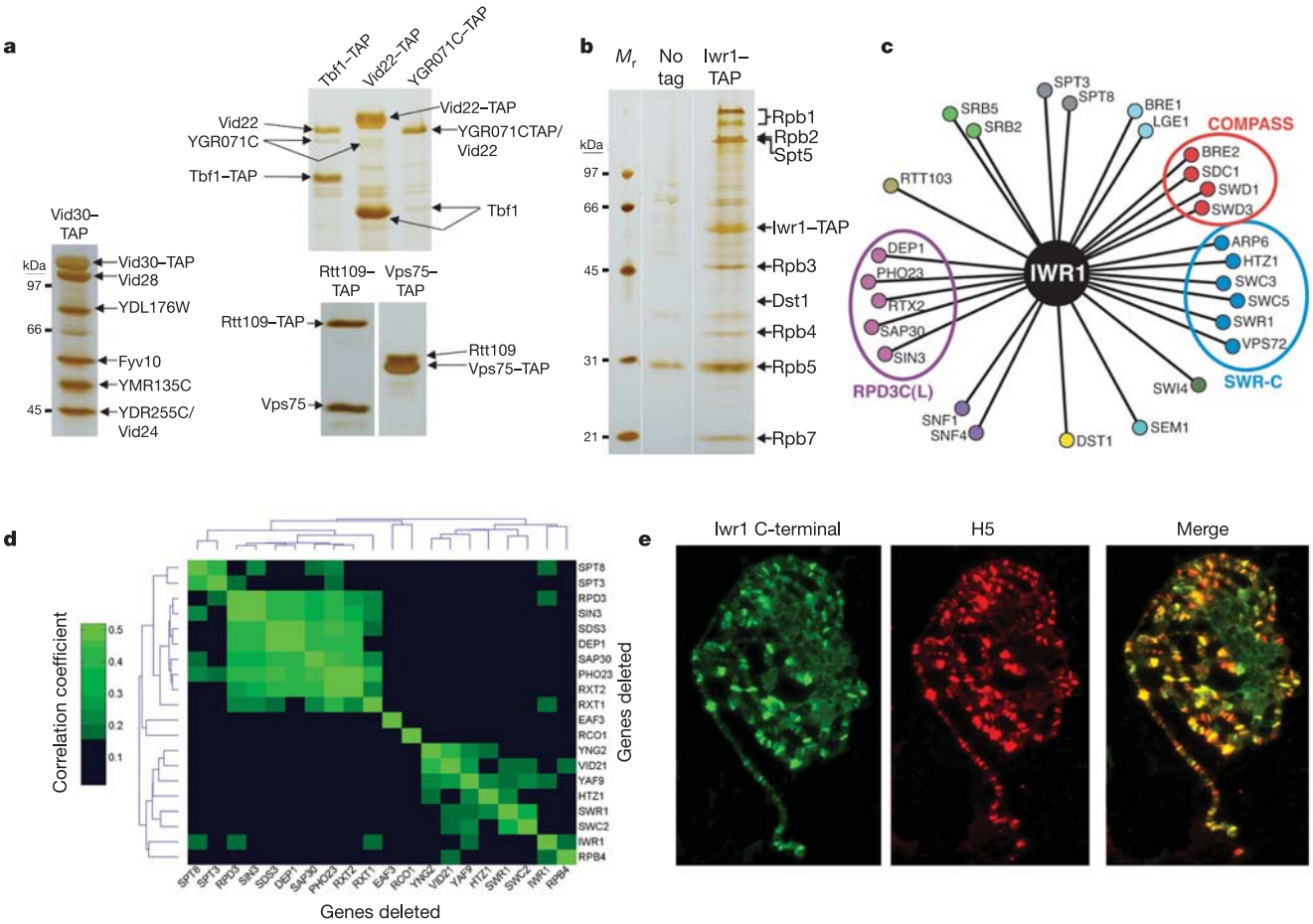

Figure 4 | Characterization of three previously unreported protein complexes and Iwr1, a novel RNAPII-interacting factor. **a**, Identification of three novel complexes by SDS–PAGE, silver staining and mass spectrometry. The same novel complex containing Vid30 was obtained after purification from strains with other tagged subunits (data not shown). **b**, Identification of Iwr1 (interacts with RNAPII). Tagging and purification of unique RNAPII subunits identified YDL115C (Iwr1) as a novel RNAPII-associated factor (Supplementary Fig. S5a). Purification of Iwr1 is shown here. **c**, Genetic interactions of Iwr1 with various transcription factors. Lines connect genes with synthetic lethal/sick genetic interactions. **d**, Microarray analysis on the indicated deletion strains. Pearson correlation coefficients were calculated for the effects on gene expression of each deletion pair and organized by two-dimensional hierarchical clustering. **e**, Antibody generated against the amino-terminal amino acid sequence (DDDDDDDSFASADGE) of the *Drosophila* homologue of Iwr1 (CG10528) and a monoclonal antibody (H5) against RNAPII subunit Rpb1 phosphorylated on S5 of the heptapeptide repeat of its carboxy-terminal domain[48] were used for co-localization studies on polytene chromosomes as previously described[47].

high proportion of fungi-specific subunits. Previous studies have shown that highly connected proteins within a network tend to be more highly conserved[17,34], a consequence of either functional constraints or preferential interaction of new proteins with existing highly connected proteins[28]. For the network as a whole, and consistent with earlier studies, Fig. 3e reveals that the frequency of ORFs with a large number (>10) of connections is proportional to the relative distance of the evolutionary group. 'Betweenness' provides a measure of how 'central' a protein is in a network, typically calculated as the fraction of shortest paths between node pairs passing through a node of interest. Figure 3e shows that highly conserved proteins tend to have higher values of betweenness. Despite these average network properties, the subunits of some complexes (for example, the kinetochore complex) display a high degree of connectedness despite restriction to hemiascomycetes. These findings suggest caution in extrapolating network properties to the properties of individual complexes. We also investigated the relationship between an ORF's essentiality and its conservation, degree of connectivity and betweenness (Supplementary Fig. S4). Consistent with previous studies[17,35], essential genes tend to be more highly conserved, highly connected and central to the network (as defined by betweenness), presumably reflecting their integrating role.

Examples of new protein complexes and interactions

Among the 275 complexes not in MIPS that we identified three are shown in Fig. 4a. One contains Tbf1, Vid22 and YGR071C. Tbf1 binds subtelomeric TTAGGG repeats and insulates adjacent genes from telomeric silencing[36,37], suggesting that this trimeric complex might be involved in this process. Consistent with this, a hypomorphic DAmP allele[10] (3′ untranslated region (UTR) deletion) of the essential *TBF1* gene causes a synthetic growth defect when combined with a deletion of *VID22* (data not shown), suggesting that Tbf1 and Vid22 have a common function. Vid22 and YGR071C are the only yeast proteins containing BED Zinc-finger domains, thought to mediate DNA binding or protein–protein interactions[38], suggesting that each uses its BED domain to interact with Tbf1 or enhance DNA binding by Tbf1. Another novel complex in Fig. 4a contains Vid30 and six other subunits (also see Fig. 3d inset). Five of its subunits (Vid30, Vid28, Vid24, Fyv10, YMR135C) have been genetically linked to proteasome-dependent, catabolite-induced degradation of fructose-1,6-bisphosphatase[39], suggesting that the remaining two subunits (YDL176W, YDR255C), hypothetical proteins of hitherto unknown function, are probably involved in the same process. Vid24 was reported to be in a complex with a M_r of approximately 600,000 (ref. 39), similar to the sum of the apparent M_r values of the subunits of the Vid30-containing complex. The third novel complex contains Rtt109 and Vps75. Because Vps75 is related to nucleosome assembly protein Nap1, and Rtt109 is involved in Ty transposition[40], this complex may be involved in chromatin assembly or function.

Our systematic characterization of complexes by TAP and mass spectrometry has often led to the identification of new components of established protein complexes (Fig. 3a)[41–43]. Figure 4 highlights Iwr1 (YDL115C), which co-purifies with RNA polymerase II (RNAPII) along with general initiation factor TFIIF and transcription elongation factors Spt4/Spt5 and Dst1 (TFIIS) (Figs 4b and 3d (inset); see also Supplementary Fig. S5a). We used synthetic genetic array (SGA) technology[9] in a quantified, high-density E-MAP format[10] to systematically identify synthetic genetic interactions for *iwr1Δ* with deletions of the elongation factor gene *DST1*, the SWR complex that assembles the variant histone Htz1 into chromatin[44], an Rpd3-containing histone deacetylase complex (Rpd3(L)) that mediates promoter-specific transcriptional repression[30,31], the histone H3 K4 methyltransferase complex (COMPASS), the activity of which is linked to elongation by RNAPII[45], and other transcription-related genes (Fig. 4c). Moreover, DNA microarray analyses of the effects on gene expression of deletions of *IWR1* and other genes involved in transcription by RNAPII, followed by clustering of the genes according to the similarity of their effects on gene expression, revealed that deletion of *IWR1* is most similar in its effects on mRNA levels to deletion of *RPB4* (Fig. 4d), a subunit of RNAPII with multiple roles in transcription[46]. We also made use of the fact that Iwr1 is highly conserved (Supplementary Fig. S5b), with a homologue, CG10528, in *Drosophila melanogaster*. Fig. 4e shows that *Drosophila* Iwr1 partly co-localizes with phosphorylated, actively transcribing RNAPII on polytene chromosomes, suggesting that Iwr1 is an evolutionarily conserved transcription factor.

Conclusions

We have described the interactome and protein complexes underlying most of the yeast proteome. Our results comprise 7,123 protein–protein interactions for 2,708 proteins in the core data set. Greater coverage and accuracy were achieved compared with previous high-throughput studies of yeast protein–protein interactions as a consequence of four aspects of our approach: first, unlike a previous study using affinity purification and mass spectrometry[14], we avoided potential artefacts caused by protein overproduction; second, we were able to ensure greater data consistency and reproducibility by systematically tagging and purifying both interacting partners for each protein–protein interaction; third, we enhanced coverage and reproducibility, especially for proteins of lower abundance, by using two independent methods of sample preparation and complementary mass spectrometry procedures for protein identification (in effect, up to four spectra were available for statistically evaluating the validity of each PPI); and finally, we used rigorous computational procedures to assign confidence values to our predictions. It is important to note, however, that our data represent a 'snapshot' of protein–protein interactions and complexes in a particular yeast strain subjected to particular growth conditions.

Both the quality of the mass spectrometry spectra used for protein identification and the approximate stoichiometry of the interacting protein partners can be evaluated by accessing our publicly available comprehensive database (http://tap.med.utoronto.ca/) that reports gel images, protein identifications, protein–protein interactions and supporting mass spectrometry data (Supplementary Information and Supplementary Fig. S6). Soon to be linked to our database will be thousands of sites of post-translational modification tentatively identified during our LC-MS/MS analyses (manuscript in preparation). The protein interactions and assemblies we identified provide entry points for studies on individual gene products, many of which are evolutionarily conserved, as well as 'systems biology' approaches to cell physiology in yeast and other eukaryotic organisms.

METHODS

Experimental procedures and mass spectrometry. Proteins were tagged, purified and prepared for mass spectrometry as previously described[43]. Gel images, mass spectra and confidence scores for protein identification by mass spectrometry are found in our database (http://tap.med.utoronto.ca/). Confidence scores for protein identification by LC-MS/MS were calculated as described previously[43]. After processing 72 database searches for each spectrum, a score of 1.25, corresponding to 99% confidence (A.P.T. and N.J.K, unpublished data), was used as a cut-off for protein identification by MALDI–TOF mass spectrometry. Synthetic genetic interactions and effects of deletion mutations on gene expression were identified as described previously[30]. *Drosophila* polytene chromosomes were stained with dIwr1 anti-peptide antibody and H5 monoclonal antibody as previously described[47].

Identification of protein complexes. Details of the methods for identification of protein complexes and calculating their overlaps with various data sets are described in Supplementary Information.

Protein property analysis. We used previously published yeast protein localization data[5,6], and yeast protein properties were obtained from the SGD (http://www.yeastgenome.org/) and GO (http://www.geneontology.org) databases. Proteins expressed at high, medium or low levels have expression log values of >4, 3–4, or <3, respectively[18].

Phylogenetic analysis. For each *S. cerevisiae* sequence a BLAST and TBLASTX

search was performed against each of the different organism data sets, including predicted ORFs from fully sequenced genomes, expressed sequence tag consensus sequences (obtained from http://www.partigenedb.org) and some raw genomic sequences. Using a BLAST bit score cut-off of 50, a taxonomic profile for each ORF was obtained by identifying sequences sharing significant similarity to at least one organism from each group. An ORF is said to be specific to each group only if it has a match to an organism within that group and not to any organism deemed to be more distantly related. Values of betweenness were calculated using the software Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

Received 20 December 2005; accepted 23 February 2006.
Published online 22 March 2006.

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546, 563–567 (1996).
2. Hughes, T. R. *et al.* Functional discovery via a compendium of expression profiles. *Cell* **102,** 109–126 (2000).
3. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286,** 1153–1155 (1999).
4. Zhu, H. & Snyder, M. Protein chip technology. *Curr. Opin. Chem. Biol.* **7,** 55–63 (2003).
5. Huh, W. K. *et al.* Global analysis of protein localization in budding yeast. *Nature* **425,** 686–691 (2003).
6. Kumar, A. *et al.* Subcellular localization of the yeast proteome. *Genes Dev.* **16,** 707–719 (2002).
7. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402,** 413–418 (1999).
8. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285,** 901–906 (1999).
9. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294,** 2364–2368 (2001).
10. Schuldiner, M. *et al.* Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* **123,** 507–519 (2005).
11. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
12. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
13. Gavin, A. C. *et al.* Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature* **415,** 141–147 (2002).
14. Ho, Y. *et al.* Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415,** 180–183 (2002).
15. Xia, Y. *et al.* Analyzing cellular biochemistry in terms of molecular networks. *Annu. Rev. Biochem.* **73,** 1051–1087 (2004).
16. von Mering, C. *et al.* Comparative assessment of large-scale data sets of protein–protein interactions. *Nature* **417,** 399–403 (2002).
17. Butland, G. *et al.* Interaction network containing conserved and essential protein complexes in *Escherichia coli*. *Nature* **433,** 531–537 (2005).
18. Ghaemmaghami, S. *et al.* Global analysis of protein expression in yeast. *Nature* **425,** 737–741 (2003).
19. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
20. Link, A. J. *et al.* Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnol.* **17,** 676–682 (1999).
21. McCormack, A. L. *et al.* Direct analysis and identification of proteins in mixtures by LC/MS/MS and database searching at the low-femtomole level. *Anal. Chem.* **69,** 767–776 (1997).
22. Krogan, N. J. *et al.* High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell* **13,** 225–239 (2004).
23. Mewes, H. W. *et al.* MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res.* **32,** D41–D44 (2004).
24. Mitchell, T. *Machine Learning* (McGraw Hill, 1997).
25. Wolpert, D. H. Stacked generalization. *Neural Netw.* **5,** 241–259 (1992).
26. Jansen, R. & Gerstein, M. Analyzing protein function on a genomic scale: the importance of gold-standard positives and negatives for network prediction. *Curr. Opin. Microbiol.* **7,** 535–545 (2004).
27. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302,** 449–453 (2003).
28. Barabasi, A. L. & Albert, R. Emergence of scaling in random networks. *Science* **286,** 509–512 (1999).
29. Enright, A. J., Van Dongen, S. & Ouzounis, C. A. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* **30,** 1575–1584 (2002).
30. Keogh, M. C. *et al.* Cotranscriptional Set2 methylation of Histone H3 lysine 36 recruits a repressive Rpd3 complex. *Cell* **123,** 593–605 (2005).
31. Carrozza, M. J. *et al.* Histone H3 methylation by Set2 directs deacetylation of coding regions by Rpd3S to suppress spurious intragenic transcription. *Cell* **123,** 581–592 (2005).
32. Lord, P. W., Stevens, R. D., Brass, A. & Goble, C. A. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* **19,** 1275–1283 (2003).
33. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
34. Fraser, H. B., Wall, D. P. & Hirsh, A. E. A simple dependence between protein evolution rate and the number of protein-protein interactions. *BMC Evol. Biol.* **3,** 11 (2003).
35. Joy, M. P., Brock, A., Ingber, D. E. & Huang, S. High-betweenness proteins in the yeast protein interaction network. *J. Biomed. Biotechnol.* **2005,** 96–103 (2005).
36. Fourel, G., Revardel, E., Koering, C. E. & Gilson, E. Cohabitation of insulators and silencing elements in yeast subtelomeric regions. *EMBO J.* **18,** 2522–2537 (1999).
37. Brigati, C., Kurtz, S., Balderes, D., Vidali, G. & Shore, D. An essential yeast gene encoding a TTAGGG repeat-binding protein. *Mol. Cell. Biol.* **13,** 1306–1314 (1993).
38. Aravind, L. The BED finger, a novel DNA-binding domain in chromatin-boundary-element-binding proteins and transposases. *Trends Biochem. Sci.* **25,** 421–423 (2000).
39. Regelmann, J. *et al.* Catabolite degradation of fructose-1,6-bisphosphatase in the yeast *Saccharomyces cerevisiae*: a genome-wide screen identifies eight novel GID genes and indicates the existence of two degradation pathways. *Mol. Biol. Cell* **14,** 1652–1663 (2003).
40. Scholes, D. T., Banerjee, M., Bowen, B. & Curcio, M. J. Multiple regulators of Ty1 transposition in *Saccharomyces cerevisiae* have conserved roles in genome maintenance. *Genetics* **159,** 1449–1465 (2001).
41. Krogan, N. J. & Greenblatt, J. F. Characterization of a six-subunit holo-elongator complex required for the regulated expression of a group of genes in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.* **21,** 8203–8212 (2001).
42. Krogan, N. J. *et al.* Proteasome involvement in the repair of DNA double-strand breaks. *Mol. Cell* **16,** 1027–1034 (2004).
43. Krogan, N. J. *et al.* RNA polymerase II elongation factors of *Saccharomyces cerevisiae*: a targeted proteomics approach. *Mol. Cell. Biol.* **22,** 6979–6992 (2002).
44. Korber, P. & Horz, W. SWRred not shaken; mixing the histones. *Cell* **117,** 5–7 (2004).
45. Hampsey, M. & Reinberg, D. Tails of intrigue: phosphorylation of RNA polymerase II mediates histone methylation. *Cell* **113,** 429–432 (2003).
46. Sampath, V. & Sadhale, P. Rpb4 and Rpb7: a sub-complex integral to multi-subunit RNA polymerases performs a multitude of functions. *IUBMB Life* **57,** 93–102 (2005).
47. Eissenberg, J. C. *et al.* dELL is an essential RNA polymerase II elongation factor with a general role in development. *Proc. Natl Acad. Sci. USA* **99,** 9894–9899 (2002).
48. Allison, L. A., Moyle, M., Shales, M. & Ingles, C. J. Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell* **42,** 599–610 (1985).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank M. Chow, N. Mohammad, C. Chung and V. Fong for their assistance with the creation of the web resources. We are grateful to J. van Helden and S. Brohée for sharing information on their comparison of clustering methods before publication. This research was supported by grants from Genome Canada and the Ontario Genomics Institute (to J.F.G. and A.E.), the Canadian Institutes of Health Research (to A.E., N.J.K., J.F.G., S.J.W., S.P. and C.J.I.), the National Cancer Institute of Canada with funds from the Canadian Cancer Society (to J.F.G.), the Howard Hughes Medical Institute (to J.S.W. and E.O.), the McLaughlin Centre for Molecular Medicine (to S.J.W. and S.P.), the Hospital for Sick Children (to J.M.P.-A.), the National Sciences and Engineering Research Council (to N.J.K., T.R.H. and A.E.) and the National Institutes of Health (to A.S., M.G., A.P. and H.Y.).

Author Information Protein interaction information from this paper has been provided to the BioGRID database (http://thebiogrid.org), as well as the International Molecular Interaction Exchange consortium (IMEx, http://imex.sf.net) consisting of BIND, DIP, IntAct, MINT and Mpact (MIPS). Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F.G (jack.Greenblatt@utoronto.ca) or A.E. (Andrew.emili@utoronto.ca).