Functional organization of the yeast proteome by systematic analysis of protein complexes

Anne-Claude Gavin*, Markus Bösche*, Roland Krause*, Paola Grandi*, Martina Marzioch*, Andreas Bauer*, Jörg Schultz*, Jens M. Rick*, Anne-Marie Michon*, Cristina-Maria Cruciat*, Marita Remor*, Christian Höfert*, Malgorzata Schelder*, Miro Brajenovic*, Heinz Ruffner*, Alejandro Merino*, Karin Klein*, Manuela Hudak*, David Dickson*, Tatjana Rudi*, Volker Gnau*, Angela Bauch*, Sonja Bastuck*, Bettina Huhse*, Christina Leutwein*, Marie-Anne Heurtier*, Richard R. Copley†, Angela Edelmann*, Erich Querfurth*, Vladimir Rybin*, Gerard Drewes*, Manfred Raida*, Tewis Bouwmeester*, Peer Bork†, Bertrand Seraphin†‡, Bernhard Kuster*, Gitte Neubauer* & Giulio Superti-Furga*†

** Cellzome AG, Meyerhofstrasse 1, 69117 Heidelberg, Germany*
† European Molecular Biology Laboratory, Meyerhofstrasse 1, 69117 Heidelberg, Germany
‡ CGM-CNRS, 91198 Gif sur Yvette Cedex, France

Most cellular processes are carried out by multiprotein complexes. The identification and analysis of their components provides insight into how the ensemble of expressed proteins (proteome) is organized into functional units. We used tandem-affinity purification (TAP) and mass spectrometry in a large-scale approach to characterize multiprotein complexes in *Saccharomyces cerevisiae.* We processed 1,739 genes, including 1,143 human orthologues of relevance to human biology, and purified 589 protein assemblies. Bioinformatic analysis of these assemblies defined 232 distinct multiprotein complexes and proposed new cellular roles for 344 proteins, including 231 proteins with no previous functional annotation. Comparison of yeast and human complexes showed that conservation across species extends from single proteins to their molecular environment. Our analysis provides an outline of the eukaryotic proteome as a network of protein complexes at a level of organization beyond binary interactions. This higher-order map contains fundamental biological information and offers the context for a more reasoned and informed approach to drug discovery.

A formidable challenge of postgenomic biology is to understand how genetic information results in the concerted action of gene products in time and space to generate function. In medicine, this is perhaps best reflected in the numerous disorders based on polygenic traits and the notion that the number of human diseases exceeds the number of genes in the genome[1]. Moreover, the total number of human genes does not differ substantially from the number of genes of the nematode worm *Caenorhabditis elegans*, suggesting that 'complexity' may partly rely on the contextual combination of the gene products[2]. Dissecting the genetic and biochemical circuitry of a cell is a fundamental problem in biology. At the biochemical level, proteins rarely act alone; rather, they interact with other proteins to perform particular cellular tasks[3]. These assemblies represent more than the sum of their parts by having a new 'function'[3].

Our knowledge regarding the identity of the building elements of specific complexes is limited and is based on selected biochemical approaches and genetic analyses. The only comprehensive protein-interaction studies are based on *ex vivo* and *in vitro* systems, such as two-hybrid systems[4–6] and protein chips[7], and need to be integrated with more-physiological approaches. Whenever it has been possible to retrieve and analyse particular cellular protein complexes under physiological conditions, the insight gained from the analysis has been fundamental for the biological understanding of their function, and has often taken the analysis well beyond the limits of genetic analysis[8,9]. Prominent examples are the spliceosome, the cyclosome, the proteasome, the nuclear pore complex and the synaptosome[10–16]. No systematic analysis of protein complexes from the same cell type using the same technique has yet been reported. We have performed a comprehensive analysis of protein complexes of baker's yeast, *S. cerevisiae,* a model system relevant to human biology[17,18].

Large-scale analysis of protein complexes

To systematically purify multiprotein complexes, we developed the strategy depicted in Fig. 1. Gene-specific cassettes containing the TAP tag[19], generated by polymerase chain reaction (PCR), were inserted by homologous recombination at the 3′ end of the genes. We processed 1,739 genes, including 1,143 genes representing eukaryotic orthologues[2]. Orthologues are thought to have evolved by vertical descent from a common ancestor[20] and are presumed to carry out the same function. For comparison, we also targeted a nonorthologous set of 596 genes from chromosomes 1, 2 and 4. To test the tagged genes in the absence of the wild-type allele, we used haploid cells. We generated a library of 1,548 yeast strains, of which 1,167 expressed the tagged proteins to detectable levels (Fig. 1). After growing cells to mid-log phase, assemblies were purified from total cellular lysates by TAP[19]. This technique combines a first high-affinity purification, mild elution using a site-specific protease, and a second affinity purification to obtain protein complexes with high efficiency and specificity[21]. The purified protein assemblies were separated by denaturing gel electrophoresis, individual bands were digested by trypsin, analysed by matrix-assisted laser desorption/ionization–time-of-flight mass spectrometry (MALDI–TOF MS)[9] and identified by database search algorithms. In all, 293 proteins were localized at membranes (integral and peripherally associated)[22]. Because their purification required a separate protocol, only 70 of the membrane-associated proteins were analysed, of which 40 were purified successfully. We analysed the proteins found associated to the 589 (418 orthologues) successfully purified tagged proteins (the 'raw' set of purifications; see Supplementary Information Table S1). This generated 20,946 samples for mass spectrometry and subsequently identified 16,830 proteins. Of these, 1,440 were distinct gene products, representing about 25% of the open reading frames (ORFs) in the genome. The analysis covers proteins of various subcellular compartments, supporting the generality of our approach (Figs 1b, 2a).

Sensitivity, specificity and reliability of the approach

Of the 589 purified tagged proteins, 78% presented associated partners, showing that the method is very efficient for the large-scale retrieval and identification of cellular protein complexes. There are several possible reasons why, in the remaining 22%, we were unable to purify and identify interacting proteins. Particular proteins may not form any or sufficiently stable or soluble complexes. In other cases, the 20K (relative molecular mass M_r 20,000) TAP tag may interfere with complex assembly or protein localization and function. Because we used haploid cells, we were able to score for viability. In 18% of the cases when essential genes were tagged, we did not obtain viable strains, confirming that carboxy-terminal tagging can impair protein function. Furthermore, the method may fail to detect transient interactions, low stoichiometric complexes, and/or those interactions occurring only in specific physiological states not present or under-represented in exponentially growing cells. In addition, the size distribution of the

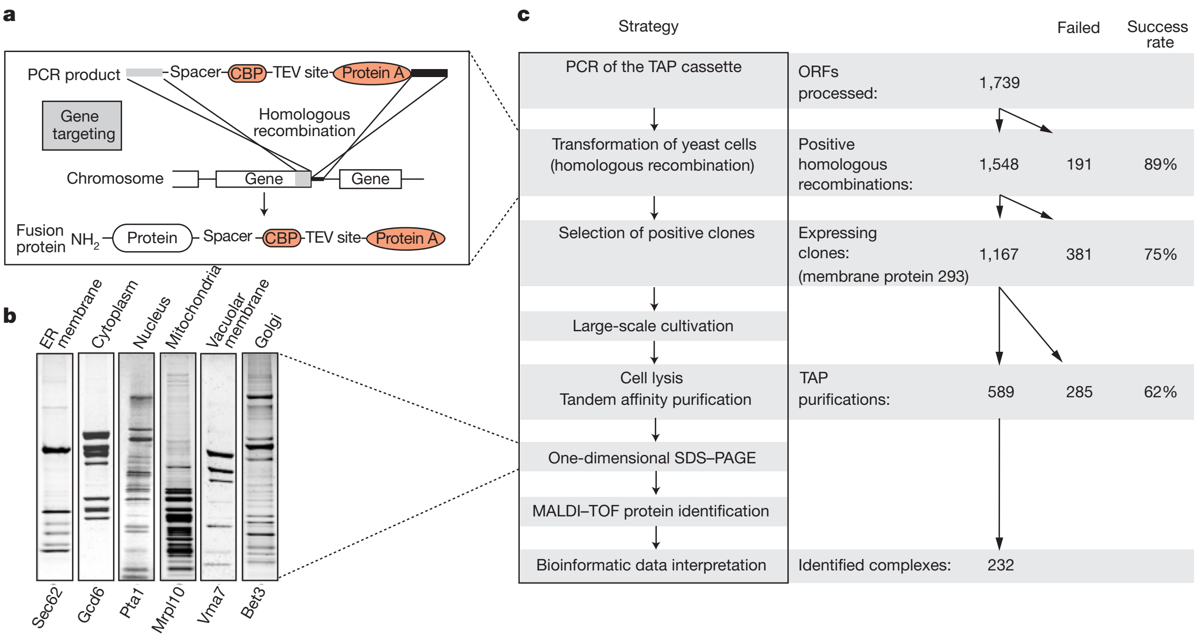

Figure 1 Synopsis of the screen. **a**, Schematic representation of the gene targeting procedure. The TAP cassette is inserted at the C terminus of a given yeast ORF by homologous recombination, generating the TAP-tagged fusion protein. **b**, Examples of TAP complexes purified from different subcellular compartments separated on denaturing protein gels and stained with Coomassie. Tagged proteins are indicated at the bottom. ER, endoplasmic reticulum. **c**, Schematic representation of the sequential steps used for the purification and identification of TAP complexes (left), and the number of experiments and success rate at each step of the procedure (right).

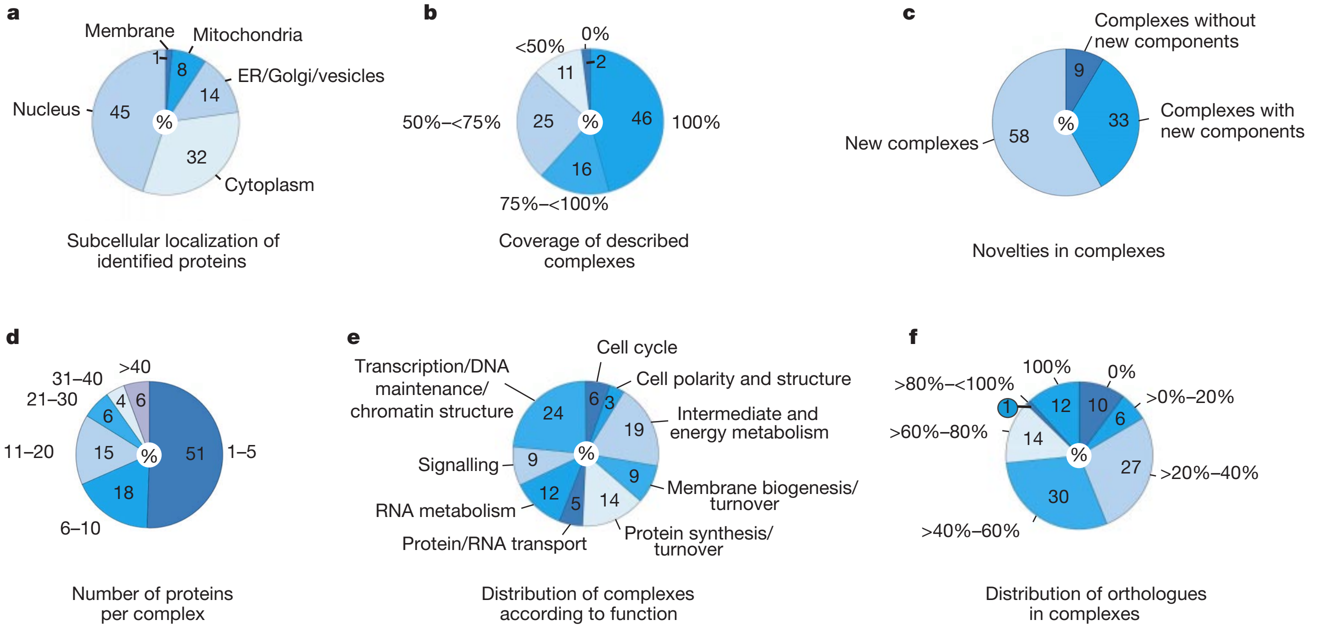

Figure 2 Statistics of proteins and complexes. Numbers inside pie charts represent the percentages of total proteins (**a**) and complexes (**b**–**f**). Outer labels show partitioning of the data according to the chart function.

identified proteins reveals a clear technical bias against proteins below 15K (see Supplementary Information Fig. S1). However, in about 30% of the cases where we failed to purify complexes around a given protein, the protein was detected when other complex components (entry points) were tagged and purified.

To assess the quality of the results obtained for purifications that do contain associated proteins, we compared our data to the literature (see below). We also established reproducibility of the approach by purifying 13 large complexes at least twice. The probability of detecting the same protein in two different purifications from the same entry point is about 70%. Therefore, on average, 30% of all detected associations presented in this study need to be treated with caution. This is particularly the case if complexes were retrieved from only one entry point. This variability may be inherent to the technique (biological samples, purification, mass spectrometry) but also to the large-scale nature of the approach, tuned to high complex coverage and high sensitivity.

To determine the experimental background, we purified mock-transformed control strains, leading to the identification of 17 contaminant proteins, mainly heat-shock and ribosomal proteins (see Supplementary Information Table S2). These are highly expressed proteins[23]. Because these proteins appeared in more than 20 of our purifications (3.5%), we decided to use this cutoff point and pragmatically excluded another 49 proteins present at equal or higher frequencies in our screen from further analysis (see Supplementary Information Table S2). It is prudent to interpret data concerning often-purified proteins just below this cutoff point with caution. Finally, we cannot exclude the occasional artificial interaction generated during cell lysis.

Although stoichiometry was not assessed in this study, we have observed that proteins belonging to the contaminant list generally comprise the weak bands after staining, and typically are identified by fewer tryptic peptides. Moreover, the systematic cutting and analysis of gel slices led to the identification by mass spectrometry even of invisible proteins. The sensitivity of the TAP MS method is high, because we were able to identify proteins present at 15 copies per cell (data not shown). Identified proteins were 6.6K to 559K in size and ranged in pI between 3.9 and 12.4 (see full evaluation in Supplementary Information Fig. S1).

Organization of the purified assemblies into complexes

On the basis of substantial overlaps, we grouped the biochemical purifications obtained with 589 different entry points into a reduced number of biologically meaningful complexes. A total of 245 purifications corresponded to 98 known nonredundant multiprotein complexes present in the yeast protein database (YPD; 60% of the data set)[22]. A further 242 purifications were assembled into 134 new complexes. The remaining 102 proteins showed no detectable association with other proteins when purified directly, or as part of other complexes. The subsequent statistical analysis is based on a list that includes 232 annotated 'TAP complexes' (see Supplementary Information Table S3).

Among the complexes that were assigned to the known YPD complexes, coverage of components was very high (Fig. 2b). Of all 232 TAP complexes, only 9% had no novel element (Fig. 2c). The size of the TAP complexes varied from 2 to 83 components, with an average of 12 components per complex (Fig. 2d). We assigned cellular roles to complexes by computing functional assignments of the individual components according to YPD[22] and by literature mining (Fig. 2e, Supplementary Information Table S3). In general terms, there seemed to be a wide functional distribution of complexes over nine categories (Fig. 2e). Of the 304 proteins with no YPD functional annotation that were identified in our screen, we propose roles for 231 (Supplementary Information Table S3). Moreover, for 113 proteins that had a functional annotation, we discovered a new molecular context (Supplementary Information Tables S3, S4; see also examples below).

Cohesive and dynamic complexes

A particular complex is not necessarily of invariable composition nor are all its building blocks uniquely associated with that specific complex. With several distinct tagged proteins as entry points to purify a complex, core components can be identified and validated, whereas more dynamic, perhaps regulatory components may be present differentially. The dynamics of complex composition are well illustrated by the cellular signalling complexes formed around the protein phosphatase 2A (PP2A; yeast TAP-C151; see Supplementary Information Table S3). Tagging different known PP2A components resulted in the purification of the known trimeric complexes containing Tpd3 (the regulatory A subunit), either of the two catalytic subunits, Pph21 and Pph22, and either of the two regulatory B subunits, Cdc55 and Rts1. The Cdc55-containing complexes were found to additionally contain Zds1 or Zds2, known cell-cycle regulators, revealing preferences among the different complexes and a link to cell-cycle checkpoints. Additional plasticity of the PP2A complexes is apparent by the interaction with three proteins implicated in bud shape and morphogenesis (Lte1, Kel1 and YBL104C). This analysis also shows that the interactions of a signalling enzyme may be sufficiently strong to allow the detection of distinct cellular complexes and thus be diagnostic for a role of this enzyme in different cellular activities.

An example of a large, cohesive complex is given by the polyadenylation machinery, which is responsible for the sequential steps necessary for eukaryotic messenger RNA cleavage and polyadenylation[24] (yeast TAP-C162; see Supplementary Information Table S3). Using Pta1 as the entry point, we identified 12 of the 13 known interactors and 7 new components (Fig. 3a). The

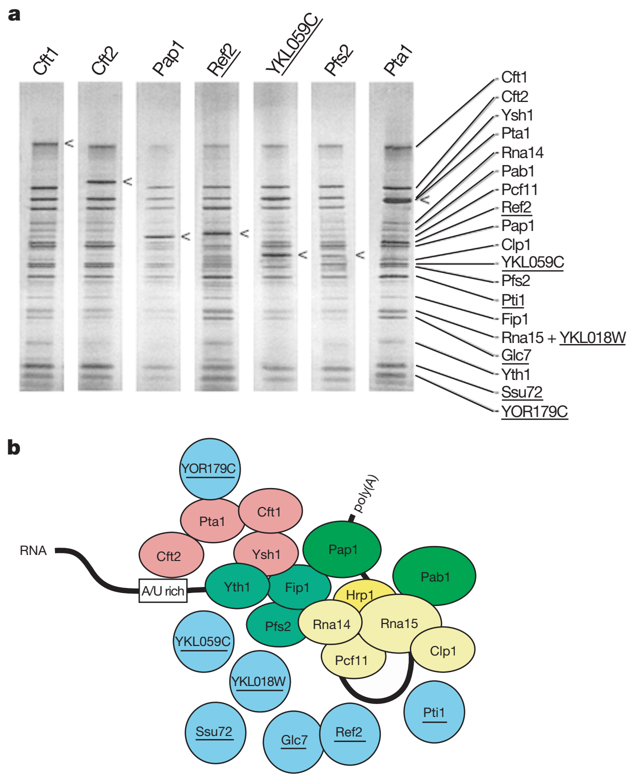

Figure 3 Primary validation of complex composition by 'reverse' purification: the polyadenylation machinery. **a**, A similar band pattern is observed when different components of the polyadenylation machinery complex are used as entry points for affinity purification. Underlined are new components of the polyadenylation machinery complex for which a physical association has not yet been described. The bands of the tagged proteins are indicated by arrowheads. **b**, Proposed model of the polyadenylation machinery.

composition of the complex was validated by extensive ‘reverse’ analysis, including purifications obtained with YKL059C and Ref2, two of the new interactors (Fig. 3a). Four new components—Pti1, YKL018W, YKL059C and YOR179C—had no previous functional annotation, but could be included into a hypothetical model by bioinformatic analysis (Fig. 3b). Ssu72, another of the seven new interactors, had previously been reported to interact with TFIIB, strongly supporting a suspected link between the polyadenylation machinery and transcription. Thus, complexes are often sufficiently strong to show high composition integrity even when purified with different entry points.

A higher-order organization map of the proteome

After assigning individual proteins to protein complexes, we investigated relationships between complexes to understand the integration and coordination of cellular functions. We represented relationship by linking complexes that share components (Fig. 4). By plotting all the relationships, we obtained a network of complexes. Connections in this network not only reflect physical interaction of complexes, but may also represent common regulation, localization, turnover or architecture. Most complexes are linked. The more connected a complex, the more central its position in the network. Complexes composed of at least 50% orthologues are shown as double sized, and complexes are colour-coded according to cellular roles. Several complexes belonging to the same class appear to group, suggesting that sharing of components reflects functional relationships. These relationships are best observed with complexes involved in mRNA metabolism (orange), cell cycle (red), protein synthesis and turnover (light green), and intermediate and energy metabolism (violet). Complexes involved in transport of protein or RNA (pink), in contrast, appear more dispersed and have connections to complexes of all other cellular roles. There are several ‘satellite’ complexes that do not seem to share components. Because this analysis is not exhaustive, we expect more connections for some of these complexes as more are purified and analysed. A software package (available at http://yeast.cellzome.com) allows the navigation of this proteome map at both the protein and complex level. Such a tool is essential to allow for proper data interpretation and for the generation of hypotheses leading to further experimental investigations.

Parallel analysis of human and yeast complexes

Orthologous gene products are thought to be responsible for essential cellular activities. We found that orthologous proteins

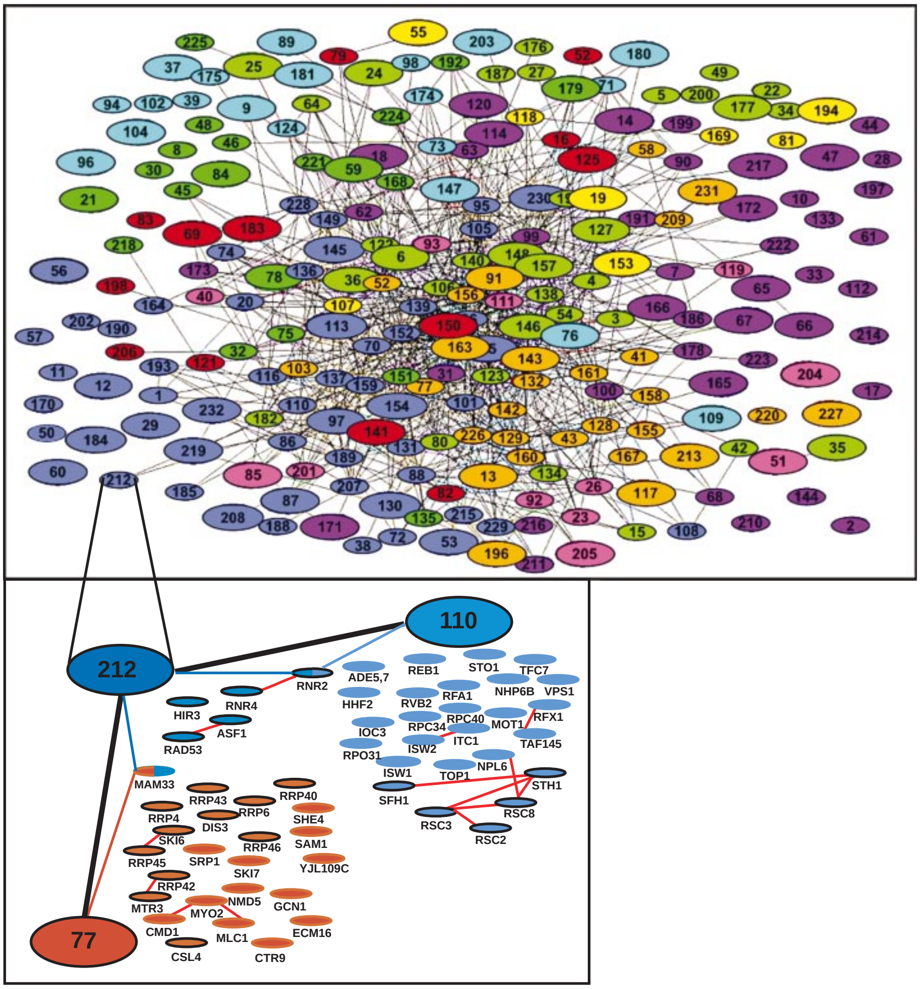

Figure 4 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes (ascribed in Supplementary Information Table S3) are colour coded: red, cell cycle; dark green, signalling; dark blue, transcription, DNA maintenance, chromatin structure; pink, protein and RNA transport; orange, RNA metabolism; light green, protein synthesis and turnover; brown, cell polarity and structure; violet, intermediate and energy metabolism; light blue, membrane biogenesis and traffic. The lower panel is an example of a complex (yeast TAP-C212) linked to two other complexes (yeast TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. Red lines indicate physical interactions as listed in YPD[22].

preferentially interact with complexes enriched with other orthologues (mean 53%; Fig. 2f). In comparison, nonorthologous proteins have a lower propensity for such interaction (mean 31%). Similarly, the likelihood to interact with essential gene products is higher for essential (44%) than for nonessential (17%) proteins. This supports the existence of an 'orthologous proteome' that may represent core functions for the eukaryotic cell[18]. To determine whether the TAP strategy can be applied to retrieve equivalent multiprotein complexes from yeast and human cells, we compared different complexes from distinct subcellular compartments: Arp2/3, a cytoskeleton-associated complex, Ccr4–Not, a nuclear assembly, and Trapp, a Golgi-associated complex. The Arp2/3 complex is a stable multiprotein assembly required for the nucleation of actin filaments in all eukaryotic cells and consists of seven proteins in human and yeast[25]. TAP of Arp2 in yeast (TAP-C153) and ARPC2 in human[25] resulted in the isolation and identification of all known components. This indicates that the TAP approach combined with liquid chromatography coupled to tandem mass spectrometry (LC/MS/MS) is an efficient and sensitive method for the retrieval and characterization of human multiprotein complexes (Fig. 5a).

The yeast Ccr4–Not complex (TAP-C149) is involved in the control of gene expression and consists of eight components[26]. Potential human orthologues have been identified and characterized for yeast Not2, Not3, Not4 and Caf1, but not for yeast Not1 and Ccr4. Moreover, no respective multiprotein complex has yet been described in mammalian cells[27]. TAP of tagged human NOT2 resulted in the identification of a multiprotein complex consisting of human NOT2, CAF1 and CALIF, and two functionally nonannotated gene products, encoded by KIAA1007 and KIAA1194, which we could assign as the orthologues of yeast Not1 and Ccr4, respectively (Fig. 5b). Purification of tagged yeast Ccr4 resulted in the identification of a complex component, Caf40, that has an orthologous counterpart, Rqcd1, also identified in the human complex. These data strongly suggest that the human and yeast Ccr4–Not complexes are comparable in subunit composition.

As a third example we purified and characterized an orthologous human TRAPP (transport protein particle) complex. The yeast complex contains ten subunits that are required for docking of transport vesicles derived from the endoplasmic reticulum to the *cis*-Golgi (yeast TAP-C102)[28]. The human complex had been purified previously as an assembly of about 670K; however, apart from human BET3 and TRS20, none of the other complex subunits had been identified[28]. TAP purification of tagged human BET3 resulted in the identification of a complex consisting of human BET3, MUM2, R32611_2, Sedlin, EHOC-1, PTD009 and KIAA1012, which we assigned as the orthologues of yeast Bet3, Bet5, Trs33, Trs20, Trs130, Trs23 and Trs85, respectively (Fig. 5c). Taken together, these examples show that the analysis of yeast complexes can often predict the composition of the human counterparts. This large-scale yeast proteome analysis could have immediate functional implications for human biology.

Discussion

To assign cellular functions to new, nonannotated gene products, and to understand the context in which proteins operate, several large-scale approaches have been undertaken. These include monitoring of mRNA expression (chips and serial analysis of gene expression (SAGE))[29], loss-of-function approaches combined with subcellular localization screens (in yeast[29,30], RNA-mediated interference in *C. elegans*[31,32], gene trap in mice[33,34]), computational *in silico* methods (protein fusions, gene neighbouring, structural predictions)[35,36], and extensive two-hybrid screens[4–6] and protein chip analysis[7,37]. The TAP/MS-based functional proteomics approach presented here may well constitute the largest analysis of protein complexes to date. We confirmed the expression of 1,440 ORFs as annotated in the genome, of which 59 had been assigned only as hypothetical. A large-scale analysis of yeast proteins performed previously on a crude cell extract identified 1,484 different proteins from exponentially growing cells[38]. Another 714 proteins were detected in our study. This raises the total number of ascertained proteome components to 2,210.

TAP[19] proved invaluable for the purification of complexes from different cellular compartments, including complexes associated with cellular membranes. The approach also allows for the efficient identification of low-abundance proteins that would not be detectable by approaches involving expression proteomics[9,21]. Further,

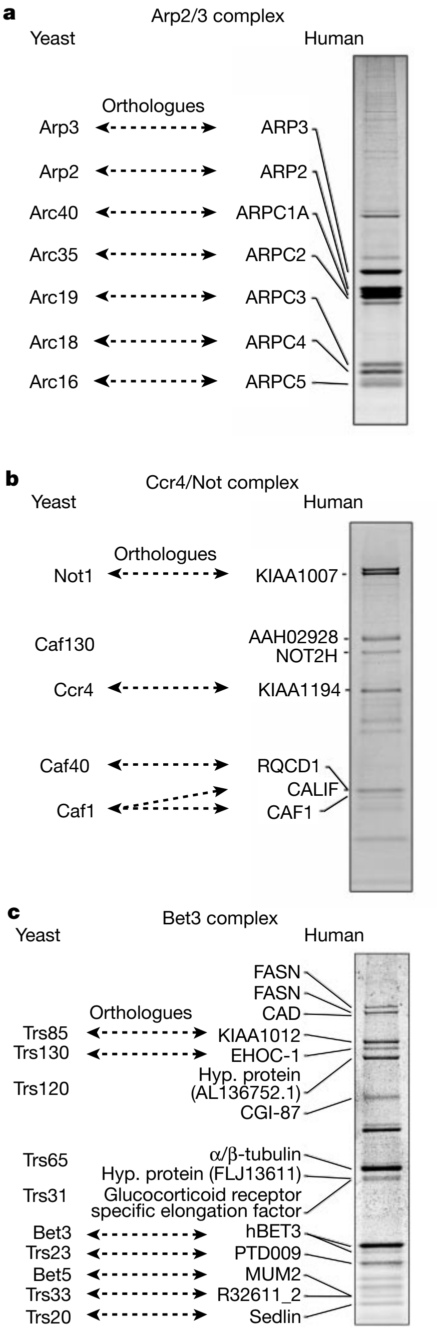

Figure 5 Protein complexes have a similar composition in yeast and human. Comparison of three TAP protein complexes isolated from human and yeast cells. All orthologous pairs are indicated by arrows, demonstrating that the complex composition between yeast and human is largely conserved. Coomassie-stained gels are shown only for the human purifications. **a**, Arp2/3 complex; **b**, Ccr4–Not2 complex; **c**, Trapp complex. Hyp. protein, hypothetical protein.

TAP allows the purification of very large complexes. For example, we were able to purify yeast TAP-C116 (Ino80), a complex reported to have an M_r about 1M–1.5M (ref. 39), and we identified all known and several new interactors.

We sought to gauge the reliability of our data by comparing the experimental results to the literature. Although comparison of our data to the YPD complex database is straightforward (Fig. 2b), it is very difficult to derive meaningful comparative information from the yeast two-hybrid data because the latter produces binary interactions, whereas the TAP/MS method yields complex composition data. When considering all possible interactions between proteins within a complex or a purification, normalized for the proteins that were identified in this study, we find that our data overlap with only 7% of the interactions seen by yeast two-hybrid assays[4–6]. When compared with the YPD protein complexes, this study covers 56%, whereas large-scale yeast two-hybrid approaches cover 10%. There are two reasons for this difference. On one hand, the yeast two-hybrid approaches touch only 35% of the described complexes, compared with 60% in this study. On the other hand, there is a generally lower coverage of individual components within complexes. The figure for the yeast two-hybrid data is 39% compared with 56% in this study. To achieve the respective coverage, the yeast two-hybrid approaches required processing of 95% of all yeast ORFs compared with 25% in our study. Altogether, this illustrates that the two methodologies address different aspects of protein interaction. Comprehensive two-hybrid approaches do not seem to be particularly suited for characterization of protein complexes. This supports the view that complex formation is more than the sum of binary interactions. However, two-hybrid analysis is of exceptional value for the detection of pairwise and transient associations. The success of the TAP/MS approach for the characterization of protein complexes relies on the conditions used for the assembly and retrieval of the complexes. These include maintaining protein concentration, localization and post-translational modifications in a manner that closely approximates normal physiology. Because the TAP/MS method does not provide information on the orientation of complex components, complex characterization and yeast two-hybrid analysis are ideally complementary.

Another outcome of our study is the ease and frequency by which protein complexes can be retrieved from cells. These biophysical properties of protein complexes may suggest cooperative binding. Bridging factors, post-translational modifications, allosteric structural changes, and binding of ions and metabolites can all cooperate to increase the number of short-range interactions between individual proteins in an assembly. Moreover, several proteins with critical regulatory functions are non-globular or intrinsically unstructured[40]. Folding into ordered structures occurs only on binding to other proteins, offering the opportunity of control over the thermodynamics of the binding process. We anticipate that some of the complexes identified in this study will be useful for structural studies.

Although we used only one set of experimental parameters here to grow and maintain cells for the evaluation of complex composition, we will, in the future, systematically modify experimental parameters to evaluate the impact of a changing environment on complex integrity[41]. These studies should help to elucidate the dynamics of complex assembly and disassembly. Moreover, the strains generated can be used for parallel assessment of protein expression levels. Finally, tandem-affinity-purified complexes from this collection may be a starting point to develop protein chips containing physiological protein complexes[7] and for assessment of biochemical activity of proteins within their molecular environment[42].

The statistical analysis of the large-scale yeast approach shows a clear tendency of proteins that are part of the set of metazoan orthologues to bind to other proteins of the same set. Moreover, we also observed a propensity to associate among the products of essential genes. Complexes containing orthologues and essential proteins overlap significantly. This recalls the proposition that the products of essential genes are also more likely to represent central components in a protein network[43]. Together, this raises the possibility that orthologue complexes represent the building blocks of a eukaryotic 'core proteome' covering basic cellular functions[18,43]. We believe that a significant number of the yeast complexes described here will have human equivalents and expect that these may form the basis for understanding multifactorial diseases. Through the 'guilt by association' concept, we are able to propose cellular roles for proteins that had no previous functional annotation and new roles for known proteins. Assessment of the physiological molecular context of proteins, as described here, may be one of the most efficient and unambiguous routes towards the assignment of gene identity and function.

Our analysis allowed us to group cellular proteins into about 200 complexes. These complexes are connected to each other by shared components. The network that resulted is a functional description of the eukaryotic proteome at a higher level of organization. Such higher-order maps will bring an increasing quality to our appreciation of biological systems. It is expected that this may provide drug discovery programmes with a molecular context for the choice and evaluation of drug targets. □

Methods

Yeast strain construction and TAP

Yeast strains expressing TAP-tagged ORFs were constructed in a semi-automated way essentially as done previously[19,21]. Cells were cultured at 30 °C in YPD medium, collected during exponential growth, and lysed mechanically with glass beads. Purifications were done as described[19,21]. For the purification of membrane proteins, the detergent concentration was adjusted to 1.5% after lysis.

TAP from human cells

Retroviral transduction vectors were generated by directional cloning of PCR-amplified ORFs into a modified version of a MoMLV-based vector via the Gateway site-specific recombination system (Life Technologies). For NOT2 and ARPC2, the TAP cassette was fused to the amino terminus and for BET3 to the C terminus. In all 3 cases, cell cultures were generated by retrovirus-mediated gene transfer and complexes were purified after cell expansion and cultivation for at least 5 days by a modified TAP protocol[19].

High-throughput protein identification

Purified proteins were concentrated, separated on 4–12% NuPAGE gels (Novex) and stained with colloidal Coomassie blue. Gels were sliced into 1.25-mm bands across the entire separation range of each lane to sample all potential interacting proteins without bias with respect to size and relative abundance. Cut bands were digested with trypsin essentially as described[44]. The resulting tryptic peptide mixtures were analysed by automated MALDI–TOF MS (Voyager DE-STR, Applied Biosystems). Proteins were identified by automated peptide mass fingerprinting using the software tool Knexus (Proteometrics) and an in-house built sequence database of *S. cerevisiae* proteins. Experiments with human orthologues of yeast proteins were subjected to the same cutting and digestion procedure but protein identification was accomplished by automated LC/MS/MS analysis (Ultimate, LC Packings, QTOF2, Micromass) in conjunction with searches of the GenPept database (ftp://ftp.ncbi.nlm.nih.gov/genbank/genpept.fsa.gz) using the software tool Mascot (Matrix Science).

Bioinformatics

Functional and localization information about yeast proteins was retrieved from the YPD released in August 2001. To get a more concise classification for localization and function, YPD classes were merged. For the analysis of assembled complexes from our purifications to the published complexes, we removed redundancy from the YPD data set manually. Protein domain analysis was performed with SMART[45]. PsiBlast[46] was used for homology analysis. All additional analysis software was developed by us, using Perl and Python. For the comparison of purifications to the YPD complexes, each member of a complex and, independently, of the purifications, was considered to be connected to each other. Redundant interactions were merged. For the comparison with yeast two-hybrid assays we counted every described binary interaction included in our purifications and included in the YPD complexes, respectively. For calculation of coverage of complexes, we considered only the YPD complexes that included at least two identified components (from yeast two-hybrid interactions or from complex purification). For the calculation of coverage of complex components, all purifications or yeast two-hybrid pairs with an entry point or bait falling into a described complex were considered.

Received 15 August; accepted 25 October 2001.

1. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405,** 857–865 (2000).
2. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
3. Alberts, B. The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell* **92,** 291–294 (1998).
4. Fromont-Racine, M., Rain, J. C. & Legrain, P. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.* **16,** 277–282 (1997).
5. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623–627 (2000).
6. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA* **98,** 4569–4574 (2001).
7. Zhu, H. *et al.* Global analysis of protein activities using proteome chips. *Science* **293,** 2101–2105 (2001).
8. Blackstock, W. P. & Weir, M. P. Proteomics: quantitative and physical mapping of cellular proteins. *Trends Biotechnol.* **17,** 121–127 (1999).
9. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* **405,** 837–846 (2000).
10. Neubauer, G. *et al.* Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nature Genet.* **20,** 46–50 (1998).
11. Zachariae, W., Shin, T. H., Galova, M., Obermaier, B. & Nasmyth, K. Identification of subunits of the anaphase-promoting complex of *Saccharomyces cerevisiae*. *Science* **274,** 1201–1204 (1996).
12. Varga-Weisz, P. D. *et al.* Chromatin-remodelling factor CHRAC contains the ATPases ISWI and topoisomerase II. *Nature* **388,** 598–602 (1997).
13. Verma, R. *et al.* Proteasomal proteomics: identification of nucleotide-sensitive proteasome-interacting proteins by mass spectrometric analysis of affinity-purified proteasomes. *Mol. Biol. Cell* **11,** 3425–3439 (2000).
14. Neubauer, G. *et al.* Identification of the proteins of the yeast U1 small nuclear ribonucleoprotein complex by mass spectrometry. *Proc. Natl Acad. Sci. USA* **94,** 385–390 (1997).
15. Rout, M. P. *et al.* The yeast nuclear pore complex: composition, architecture, and transport mechanism. *J. Cell Biol.* **148,** 635–651 (2000).
16. Husi, H., Ward, M. A., Choudhary, J. S., Blackstock, W. P. & Grant, S. G. Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. *Nature Neurosci.* **3,** 661–669 (2000).
17. Bassett, D. E. Jr, Boguski, M. S. & Hieter, P. Yeast genes and human disease. *Nature* **379,** 589–590 (1996).
18. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287,** 2204–2215 (2000).
19. Rigaut, G. *et al.* A generic protein purification method for protein complex characterization and proteome exploration. *Nature Biotechnol.* **17,** 1030–1032 (1999).
20. Fitch, W. M. Distinguishing homologous from analogous proteins. *Syst. Zool.* **19,** 99–113 (1970).
21. Puig, O. *et al.* The tandem affinity purification (tap) method: a general procedure of protein complex purification. *Methods* **24,** 218–229 (2001).
22. Costanzo, M. C. *et al.* YPD, PombePD and WormPD: model organism volumes of the BioKnowledge library, an integrated resource for protein information. *Nucleic Acids Res.* **29,** 75–79 (2001).
23. Garrels, J. I. *et al.* Proteome studies of *Saccharomyces cerevisiae*: identification and characterization of abundant proteins. *Electrophoresis* **18,** 1347–1360 (1997).
24. Barabino, S. M. & Keller, W. Last but not least: regulated poly(A) tail formation. *Cell* **99,** 9–11 (1999).
25. Higgs, H. N. & Pollard, T. D. Regulation of actin filament network formation through Arp2/3 complex: activation by a diverse array of proteins. *Annu. Rev. Biochem.* **70,** 649–676 (2001).
26. Liu, H. Y. *et al.* The NOT proteins are part of the CCR4 transcriptional complex and affect gene expression both positively and negatively. *EMBO J.* **17,** 1096–1106 (1998).
27. Albert, T. K. *et al.* Isolation and characterization of human orthologs of yeast CCR4-NOT complex subunits. *Nucleic Acids Res.* **28,** 809–817 (2000).
28. Sacher, M., Barrowman, J., Schieltz, D., Yates, J. R. III & Ferro-Novick, S. Identification and characterization of five new subunits of TRAPP. *Eur. J. Cell Biol.* **79,** 71–80 (2000).
29. Lockhart, D. J. & Winzeler, E. A. Genomics, gene expression and DNA arrays. *Nature* **405,** 827–836 (2000).
30. Ross-Macdonald, P. *et al.* Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature* **402,** 413–418 (1999).
31. Fraser, A. G. *et al.* Functional genomic analysis of *C. elegans* chromosome I by systematic RNA interference. *Nature* **408,** 325–330 (2000).
32. Gonczy, P. *et al.* Functional genomic analysis of cell division in *C. elegans* using RNAi of genes on chromosome III. *Nature* **408,** 331–336 (2000).
33. Friedrich, G. & Soriano, P. Promoter traps in embryonic stem cells: a genetic screen to identify and mutate developmental genes in mice. *Genes Dev.* **5,** 1513–1523 (1991).
34. Leighton, P. A. *et al.* Defining brain wiring patterns and mechanisms through gene trapping in mice. *Nature* **410,** 174–179 (2001).
35. Eisenberg, D., Marcotte, E. M., Xenarios, I. & Yeates, T. O. Protein function in the post-genomic era. *Nature* **405,** 823–826 (2000).
36. Huynen, M., Snel, B., Lathe, W. III & Bork, P. Predicting protein function by genomic context: quantitative evaluation and qualitative inferences. *Genome Res.* **10,** 1204–1210 (2000).
37. Zhu, H. *et al.* Analysis of yeast protein kinases using protein chips. *Nature Genet.* **26,** 283–289 (2000).
38. Washburn, M. P., Wolters, D. & Yates, J. R. III Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnol.* **19,** 242–247 (2001).
39. Shen, X., Mizuguchi, G., Hamiche, A. & Wu, C. A chromatin remodelling complex involved in transcription and DNA processing. *Nature* **406,** 541–544 (2000).
40. Wright, P. E. & Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure–function paradigm. *J. Mol. Biol.* **293,** 321–331 (1999).
41. Ideker, T. *et al.* Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292,** 929–934 (2001).
42. Martzen, M. R. *et al.* A biochemical genomics approach for identifying genes by the activity of their products. *Science* **286,** 1153–1155 (1999).
43. Jeong, H., Mason, S. P., Barabasi, A. L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411,** 41–42 (2001).
44. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68,** 850–858 (1996).
45. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28,** 231–234 (2000).
46. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–3402 (1997).

Supplementary Information accompanies the paper on *Nature*'s website (http://www.nature.com).

Acknowledgements

We thank C. Cohen, S. Artavanis-Tsakonas and F. Kafatos for suggestions and support throughout the work; C. Gässler, C. Kalla, N. Umhey and P. Völkel for help with experiments; K. Herzog and her team for media supply; E. Hurt and H. T. Timmers for sharing unpublished results and for suggestions; and numerous people at the European Molecular Biology Laboratory for help, especially G. Bricard, F. Caspary, O. Puig, G. Stier and members of the Nebreda, Superti-Furga, Klein, Wilm and Bork laboratories.

Correspondence and requests for materials should be addressed to A.C.G. (e-mail: anne-claude.gavin@cellzome.com) or G.S.F. (giulio.superti-furga@cellzome.com). The data are accessible at http://yeast.cellzome.com.