*D646–D650* *Nucleic Acids Research, 2008, Vol. 36, Database issue* *Published online 26 October 2007*
*doi:10.1093/nar/gkm936*

# CORUM: the comprehensive resource of mammalian protein complexes

**Andreas Ruepp[1],*, Barbara Brauner[1], Irmtraud Dunger-Kaltenbach[1], Goar Frishman[1], Corinna Montrone[1], Michael Stransky[1], Brigitte Waegele[1], Thorsten Schmidt[2], Octave Noubibou Doudieu[1], Volker Stümpflen[1] and H. Werner Mewes[1,2]**

[1]Institute for Bioinformatics (MIPS), German Research Center for Environmental Health, Ingolstaedter Landstraße 1, D-85764 Neuherberg and [2]Technische Universität München, Chair of Genome Oriented Bioinformatics, Center of Life and Food Science, D-85350 Freising-Weihenstephan, Germany

Received September 12, 2007; Revised October 10, 2007; Accepted October 11, 2007

## ABSTRACT

**Protein complexes are key molecular entities that integrate multiple gene products to perform cellular functions. The CORUM (http://mips.gsf.de/genre/proj/corum/index.html) database is a collection of experimentally verified mammalian protein complexes. Information is manually derived by critical reading of the scientific literature from expert annotators. Information about protein complexes includes protein complex names, subunits, literature references as well as the function of the complexes. For functional annotation, we use the FunCat catalogue that enables to organize the protein complex space into biologically meaningful subsets. The database contains more than 1750 protein complexes that are built from 2400 different genes, thus representing 12% of the protein-coding genes in human. A web-based system is available to query, view and download the data. CORUM provides a comprehensive dataset of protein complexes for discoveries in systems biology, analyses of protein networks and protein complex-associated diseases. Comparable to the MIPS reference dataset of protein complexes from yeast, CORUM intends to serve as a reference for mammalian protein complexes.**

## INTRODUCTION

Large community approaches like sequencing of mammalian genomes as well as the characterization of genetic elements within the ENCODE project (1) have revolutionized our knowledge about mammalian genetics. Systems biology approaches to describe cellular processes as suitable models need to integrate these data and to shift from the functional description of the individual gene towards the interaction in cellular networks. In recent years, it has been shown by systematic experiments that the large majority of proteins do not act as isolated entities but form transient or stable interactions with other proteins. Large datasets of protein–protein interaction data can be presented as protein networks, which fulfil higher-level cellular tasks, so-called functional modules. The basic representatives of functional modules are protein complexes, since they display strong and frequent connections within the complex and weak and rare connections to components outside the complex (2). The concerted action of proteins within protein complexes allows cells to acquire novel functionalities, which are beyond the performance of individual proteins. Protein complexes like proteasomes, chaperonins and spliceosomes are central components in vital cellular tasks like protein folding, protein degradation and RNA splicing, respectively.

Most of the analyses concerning eukaryotic protein complexes have been performed with data from yeast. This is due to the availability of high-throughput datasets as well as a manually curated dataset, which received a status as ‘gold standard’ (3,4). In a high-throughput analysis of protein complexes in yeast, 4.9 different subunits per complex were identified on an average (5). Combining information from manual annotation and high-throughput experiments reveals that 2733 (45%) of 6007 protein coding genes in yeast form components of protein complexes (6). This number does not yet include data of two recent large-scale approaches, which aimed to collect a comprehensive set of protein complexes from yeast (5,7). Therefore, it is tempting to speculate, that at least half of the protein-coding genes in the lower eukaryote yeast are involved in the formation of the yeast ‘complexosome’. The absolute number of protein complexes that constitutes the complexosome remains also

*To whom correspondence should be addressed. Tel: +49 89 3187 3189; Fax: 49 89 3187 3585; Email: andreas.ruepp@gsf.de a matter of speculation. Two large-scale approaches identified 491 and 547 protein complexes, respectively, but neither of them covered the manually annotated dataset from MIPS (3) exhaustively, indicating that the final number of protein complexes in yeast extends the current results significantly.

Analyses of protein complexes from yeast have also shown that protein complex subunits have characteristic properties. They exhibit a higher essentiality, are to a large extent co-expressed and evolutionally stronger conserved than other proteins (8,9). It can be assumed, that many of the results which were observed from the analysis of yeast complexes can be adopted for other organisms. However, analysis of protein complexes that are involved in cell-cycle regulation has shown that the transfer of information has limitations, too. Periodically expressed protein complex subunits, that confer the just-in-time assembly of the respective protein complexes, are different in yeast, *Arabidopsis* and human (10).

For mammalian protein complexes, there is neither a high-throughput dataset nor a comprehensive collection of manually and functionally annotated complexes available. However, there is a growing interest in the analysis of mammalian complexes. Besides the investigation of cell-cycle complexes, the disease relevance of human protein complexes was analysed recently (11). There, the protein complexes were computationally generated from protein–protein interaction data. In order to provide an experimentally validated dataset of mammalian protein complexes, we generated the comprehensive resource of mammalian protein complexes (CORUM) database. In CORUM, a protein complex is defined as a group of two or more proteins that physically interact and form a quaternary structure. Protein complexes can be stable or appear transiently and can be found *in vivo*. The interaction of protein complex members is required to perform together at the same time a cellular function or reaction. CORUM is a protein complex information resource that depicts various features of protein complexes like protein complex composition, biological function, cellular localization and other associated information like disease relevance. CORUM is freely available at http://mips.gsf.de/genre/proj/corum/index.html.

## ANNOTATION OF PROTEIN COMPLEXES

In order to provide a high-quality dataset of mammalian protein complexes, all entries are manually created. Only protein complexes which have been isolated and characterized by reliable experimental evidence are included in CORUM. To be considered for CORUM, a protein complex has to be isolated as one molecule and must not be a construct derived from several experiments. Also, artificial constructs of subcomplexes are not taken into account. Since information from high-throughput experiments contains a significant fraction of false-positive results, this type of data is excluded. References for relevant articles were mainly found in general review articles, cross-references to related protein complexes within analysed literature and comments on referenced articles in UniProt (12). Table 1 shows that the highest fraction of the used articles was published in *The Journal of Biological Chemistry* (23.6%), followed by PNAS (7.1%), *Molecular and Cellular Biology* (6.1%) and *Cell* (5.7%). The vast majority of the used articles are from journals with high-impact factor, which shows that characterization of protein complexes is considered as important information.

**Table 1.** Analysis about the absolute number and fraction of articles from respective scientific journals that were used for the annotation of mammalian protein complexes in CORUM

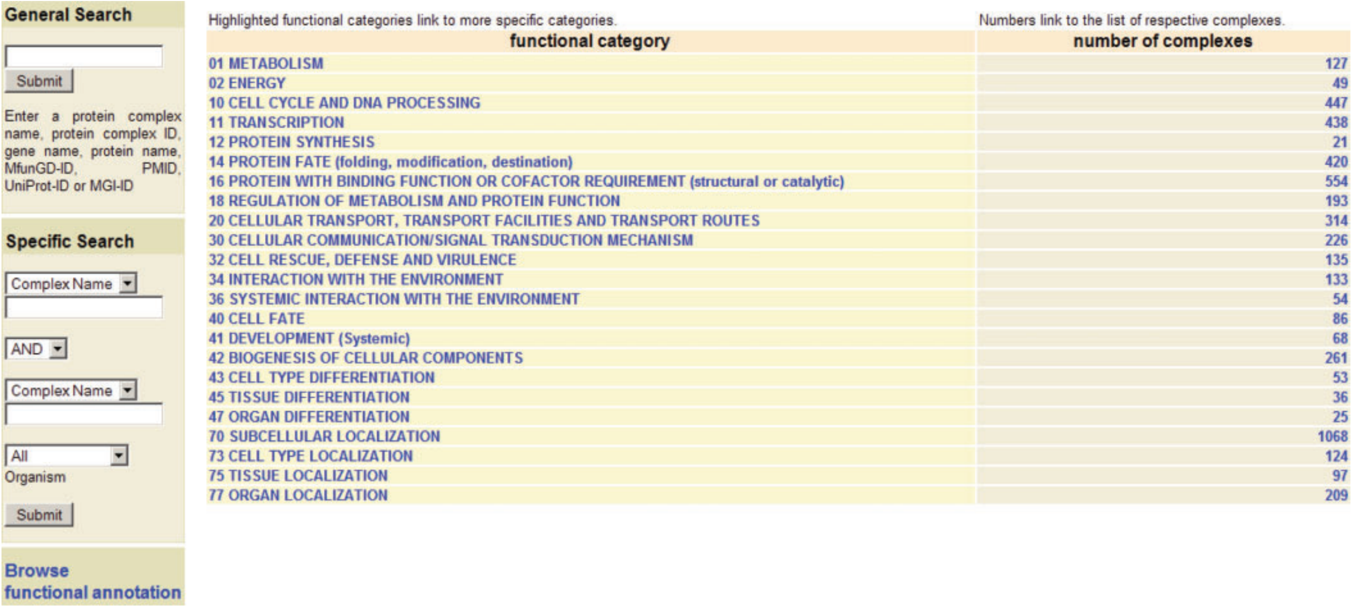
| Journal | Number of articles | Fraction of all articles (%) |
|---|---|---|
| The Journal of Biological Chemistry | 253 | 23.6% |
| PNAS | 76 | 7.1% |
| Molecular and Cellular Biology | 65 | 6.1% |
| Cell | 61 | 5.7% |
| Molecular Cell | 55 | 5.1% |
| The EMBO Journal | 52 | 4.8% |
| Nature | 37 | 3.4% |
| The Journal of Cell Biology | 34 | 3.2% |
| Science | 24 | 2.2% |
| Total | 1073 | 100% |

Since some articles contain information about more than one protein complex, the number of articles is lower than the number of annotated complexes.

In order to define community standards for data representation in proteomics to facilitate data comparison, exchange and verification the PSI-MI standard was introduced (13). The CORUM dataset is annotated according to the currently valid PSI-MI 2.5 standard. One rule of PSI-MI annotation is to separate information about molecular interactions, which are described redundantly by different publications. The advantage of this approach is that annotators present the information exactly as described by the authors and do not need to amalgamate the result of different groups, if the experiments show conflicting results. Another advantage is that if one protein complex has been isolated and characterized by different groups, the reproducibility confirms the composition of the protein complex. The drawback of this approach is that it results in a certain extent of redundancy.

Many well-characterized protein complexes are associated with scientific names like ribosome, proteasome or spliceosome in literature. These descriptions are also provided in CORUM, as well as synonyms if they are frequently used in the literature. An example is the eukaryotic chaperonin CCT (chaperonin containing TCP-1), which is also well known as TRiC (TCP-1 ring complex). If there is no name found for a protein complex available, we define one which is usually composed of gene names of the complex, e.g. 'BRCA1-RAD51 complex' or 'Ubiquitin E3 ligase (containing FBXW7, CUL1, SKP1A and RBX1)'.

Another annotated feature is the organism, from which the protein complex originates. The concentration of many research activities towards the biology of humans is



Enter a protein complex name, protein complex ID, gene name, protein name, MfunGD-ID, PMID, UniProt-ID or MGI-ID

Highlighted functional categories link to more specific categories.

Numbers link to the list of respective complexes.

| functional category | number of complexes |
|---|---|
| 01 METABOLISM | 127 |
| 02 ENERGY | 49 |
| 10 CELL CYCLE AND DNA PROCESSING | 447 |
| 11 TRANSCRIPTION | 438 |
| 12 PROTEIN SYNTHESIS | 21 |
| 14 PROTEIN FATE (folding, modification, destination) | 420 |
| 16 PROTEIN WITH BINDING FUNCTION OR COFACTOR REQUIREMENT (structural or catalytic) | 554 |
| 18 REGULATION OF METABOLISM AND PROTEIN FUNCTION | 193 |
| 20 CELLULAR TRANSPORT, TRANSPORT FACILITIES AND TRANSPORT ROUTES | 314 |
| 30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM | 226 |
| 32 CELL RESCUE, DEFENSE AND VIRULENCE | 135 |
| 34 INTERACTION WITH THE ENVIRONMENT | 133 |
| 36 SYSTEMIC INTERACTION WITH THE ENVIRONMENT | 54 |
| 40 CELL FATE | 86 |
| 41 DEVELOPMENT (Systemic) | 68 |
| 42 BIOGENESIS OF CELLULAR COMPONENTS | 261 |
| 43 CELL TYPE DIFFERENTIATION | 53 |
| 45 TISSUE DIFFERENTIATION | 36 |
| 47 ORGAN DIFFERENTIATION | 25 |
| 70 SUBCELLULAR LOCALIZATION | 1068 |
| 73 CELL TYPE LOCALIZATION | 124 |
| 75 TISSUE LOCALIZATION | 97 |
| 77 ORGAN LOCALIZATION | 209 |

**Figure 1.** Overview about the FunCat functional annotation results of the protein complexes in CORUM. The different search options of the database are also shown.

reflected by the high content of human protein complexes in CORUM. The vast majority of all analysed protein complexes in CORUM originates from human (65%), followed by mouse (14%) and rat (14%).

The subunits of protein complexes are annotated according to the respective SwissProt entries. In CORUM, only the primary accessions are stored as identifiers. Associated information like gene names and protein names is retrieved via the BioRS sequence retrieval system, enabling up-to-date information from the primary data sources without the need of synchronization.

Other important information besides the identification of the different subunits is the number of individual proteins that are required to assemble the complex. In most cases, the molecular characterization of the protein complex composition is limited to the identification of the subunits. For cases where the stoichiometry of the subunits has been analysed, the information is given in the 'Number of subunits' field (see e.g. complex 960).

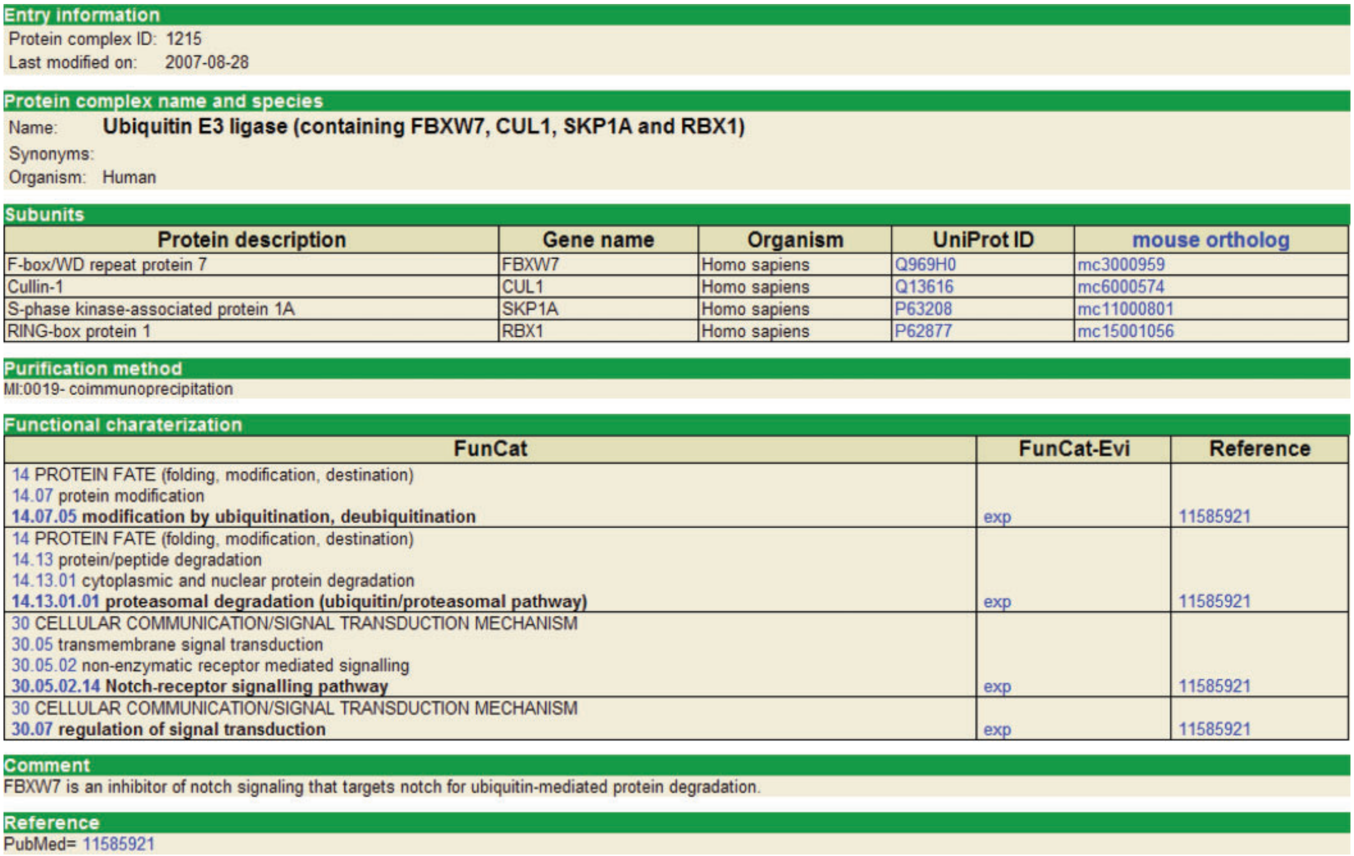
We use the Functional Catalogue (FunCat) annotation scheme for protein and protein complex function characterization (14). The FunCat has been used for manual annotation of model organisms like *Saccharomyces cerevisiae*, *Arabidopsis thaliana* and mouse (15) and was also frequently used for the analysis of protein networks and high-throughput experiments (14). Application of FunCat organizes data in a systematic, computer-readable format. The hierarchical structure of FunCat allows browsing for protein complexes with particular cellular functions or localizations (Figure 1). This reveals subsets, which would otherwise require specialised databases like the PIN database for nuclear protein complexes (16). Examples of such sub datasets are presented on the CORUM home page. In addition, FunCat annotation allows fast access to some statistics of the data. The CORUM dataset contains e.g. far more protein complexes from the nucleus (67% of all complexes with annotated localization) than from the cytoplasm (9%). This might be explained by the complexity of the information processes within the nucleus. However, the data do not necessarily correlate to the situation of living cells but might rather reflect the topics which have been investigated by individual research projects.

The evidence for applying a functional category is given in a separate field (Figure 2). There are five different evidences which provide information about the underlying rationale why a functional category has been applied. These include different qualities, ranging from experimental evidence (exp) to predicted functions (pred). For all evidences but predicted annotation the underlying PubMed references are provided (Figure 2). Additional information like disease relevance or more detailed information about the cellular function of protein complexes is given in the comment field (Figure 2).

One of the mandatory information for PSI-MI compliant annotation is the experimental method which led to the identification of the protein interaction. For this kind of data the PSI consortium provides a list of methods (http://www.psidev.info/). If several methods were used to isolate a protein complex, all methods are listed. The PubMed reference of the article that describes the isolation of the protein complex is given in the PubMed field (Figure 2).

Concerning the inventory of protein complexes of an organism, the complexosome, important questions are (i) of how many different subunits protein complexes are composed, (ii) the fraction of protein coding genes are devoted by cells to build the complexosome and (iii) how many protein complexes does a cell contain?

(i) In September 2007, the CORUM database contained about 1750 protein complexes. On average, each complex consists of 4.7 different subunits. This is well in line with



**Entry information**

Protein complex ID: 1215
Last modified on: 2007-08-28

**Protein complex name and species**

Name: **Ubiquitin E3 ligase (containing FBXW7, CUL1, SKP1A and RBX1)**
Synonyms:
Organism: Human

**Subunits**

| Protein description | Gene name | Organism | UniProt ID | mouse ortholog |
|---|---|---|---|---|
| F-box/WD repeat protein 7 | FBXW7 | Homo sapiens | Q969H0 | mc3000959 |
| Cullin-1 | CUL1 | Homo sapiens | Q13616 | mc6000574 |
| S-phase kinase-associated protein 1A | SKP1A | Homo sapiens | P63208 | mc11000801 |
| RING-box protein 1 | RBX1 | Homo sapiens | P62877 | mc15001056 |

**Purification method**

MI:0019- coimmunoprecipitation

**Functional charaterization**

| FunCat | FunCat-Evi | Reference |
|---|---|---|
| 14 PROTEIN FATE (folding, modification, destination)<br>14.07 protein modification<br>**14.07.05 modification by ubiquitination, deubiquitination** | exp | 11585921 |
| 14 PROTEIN FATE (folding, modification, destination)<br>14.13 protein/peptide degradation<br>14.13.01 cytoplasmic and nuclear protein degradation<br>**14.13.01.01 proteasomal degradation (ubiquitin/proteasomal pathway)** | exp | 11585921 |
| 30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM<br>30.05 transmembrane signal transduction<br>30.05.02 non-enzymatic receptor mediated signalling<br>**30.05.02.14 Notch-receptor signalling pathway** | exp | 11585921 |
| 30 CELLULAR COMMUNICATION/SIGNAL TRANSDUCTION MECHANISM<br>**30.07 regulation of signal transduction** | exp | 11585921 |

**Comment**

FBXW7 is an inhibitor of notch signaling that targets notch for ubiquitin-mediated protein degradation.

**Reference**

PubMed= 11585921

**Figure 2.** Result page of the Ubiquitin E3 ligase (containing FBXW7, CUL1, SKP1A and RBX1) protein complex from the CORUM database.

data from yeast (see above). The largest protein complex of the dataset, the spliceosome, consists of 143 different proteins. (ii) If all genes of the protein complexes in CORUM are mapped to e.g. the human genome, the dataset covers ~2400 genes. Based on current estimations that the human genome codes for 20 488 protein coding genes (17), the entries from CORUM represent 12% of a mammalian genome. Due to the lack of data it is not possible to give a reliable approximation for the total number of protein coding genes in mammalian organisms involved in complex formation. However, assuming that in mammals like in yeast more than half of the protein coding genes are used for the formation of protein complexes would offer cells an enormous repertoire of building blocks for the development of complexes with novel functionalities. The modular architecture of protein complexes is exemplified by large protein complex families like SNARE complexes (96 members), integrin complexes (69 members) and ubiquitin E3 ligases (54 members) that were annotated in CORUM. These protein complex families originated from the association of protein family members which emerged in evolution as the result of gene duplication and specification events (2). (iii) To date, including data from CORUM, BIND (18) and HPRD (19) the number of non-redundant protein complexes in mammals is well above 2500. Since those are only a part of all protein complexes from literature that have been annotated and novel experimentally characterized protein complexes continuously appear, this number will certainly increase in the future.

## SEARCH OPTIONS

CORUM offers three different possibilities to select suitable protein complexes from the dataset. As a quick start, we offer predefined sets of protein complexes on the home page. The 'Browse protein complexes localized in...' and 'Browse protein complexes involved in...' buttons are linked to selections of protein complexes with a certain cellular localization or function, respectively. The underlying information of the selected complexes is based on the FunCat annotation. Further selections with the same topic can be inspected via the 'more...' link. A comprehensive overview about protein complexes associated with a specific FunCat category is given with the 'Browse functional annotation' link (Figure 1) on the home page. The numbers beside the functional categories show how many protein complexes were annotated with the respective category.

The second search option is the 'General search' which performs simultaneous searches across several attributes (Figure 1). This is especially suited for searches where comprehensiveness rather than specificity is required. A query for 'proteasome' e.g. reveals not only all proteasome complexes but also all complexes that contain a proteasomal subunit.

Finally, the 'Specific search' allows to select individual attributes that were annotated (Figure 1). Additionally, specific searches can be combined by using the logic operators AND, OR and NOT. Searches for gene names and protein names include also the synonyms that were annotated by UniProt.

## AVAILABILITY OF THE DATA

CORUM data is available for download in tab delimited as well as in XML file format. Download versions of the protein complex are regularly updated. Beside the complete dataset, CORUM offers the download of subsets, which were generated by individual searches. Respective buttons for download in PSI-MI 2.5 or tab delimited form are available on the web pages.

## TECHNOLOGY

CORUM is embedded within the MIPS Genome Research Environment (GenRE) (20). This component-oriented multi-tier architecture, based on J2EE technology, ensures scalability and provides consistent data access via Enterprise Java Beans (EJBs). As data exchange format XML is used, thus enabling readability across platforms and systems. The webpage layout is rendered with XSL transformations following the Model-View-Controller design pattern. As data backend, the relational MySQL database system (www.mysql.com) is applied.

## CONCLUSION

Protein complexes are a link between the parts-list of the synthesized gene products and entire cellular systems. The importance of the cellular machines was recognized a decade ago (21) but only in recent years a growing number of bioinformatics articles on this topic appeared. These included the investigation of the properties of the protein members, the evolution of protein complexes, the appearance of protein complexes during the cell cycle and disease relevance of protein complexes (2,8–11). It became clear that a comprehensive understanding of cellular systems is not possible without the knowledge about the cellular machines. Thus, it was stated that 'Identifying all protein complexes in an organism is a major goal of systems biology' (22). With CORUM, we provide a resource of manually annotated protein complexes for systems biology and the investigation of protein-associated diseases.

## ACKNOWLEDGEMENT

This work was supported by 031U212C BFAM (BMBF) to HWM. Funding to pay the Open Access publication charges for this article was provided by German Research Center for Environmental Health.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Birney,E., Stamatoyannopoulos,J.A., Dutta,A., Guigo,R., Gingeras,T.R., Margulies,E.H., Weng,Z., Snyder,M., Dermitzakis,E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
2. Pereira-Leal,J.B., Levy,E.D. and Teichmann,S.A. (2006) The origins and evolution of functional modules: lessons from protein complexes. *Philos. T. Roy. Soc. Bi.*, **361**, 507–517.
3. Guldener,U., Munsterkotter,M., Oesterheld,M., Pagel,P., Ruepp,A., Mewes,H.W. and Stumpflen,V. (2006) MPact: the MIPS protein interaction resource on yeast. *Nucleic Acids Res.*, **34**, D436–D441.
4. Yu,H., Luscombe,N.M., Lu,H.X., Zhu,X., Xia,Y., Han,J.D., Bertin,N., Chung,S., Vidal,M. *et al.* (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.*, **14**, 1107–1118.
5. Krogan,N.J., Cagney,G., Yu,H., Zhong,G., Guo,X., Ignatchenko,A., Li,J., Pu,S., Datta,N. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.
6. Guldener,U., Munsterkotter,M., Kastenmuller,G., Strack,N., Van Helden,J., Lemer,C., Richelles,J., Wodak,S.J., Garcia-Martinez,J. *et al.* (2005) CYGD: the comprehensive yeast genome database. *Nucleic Acids Res.*, **33**, D364–D368.
7. Gavin,A.C., Aloy,P., Grandi,P., Krause,R., Boesche,M., Marzioch,M., Rau,C., Jensen,L.J., Bastuck,S. *et al.* (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.
8. Fraser,H.B. (2005) Modularity and evolutionary constraint on proteins. *Nat. Genet.*, **37**, 351–352.
9. Kim,P.M., Lu,L.J., Xia,Y. and Gerstein,M.B. (2006) Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, **314**, 1938–1941.
10. Jensen,L.J., Jensen,T.S., de,L.U., Brunak,S. and Bork,P. (2006) Co-evolution of transcriptional and post-translational cell-cycle regulation. *Nature*, **443**, 594–597.
11. Lage,K., Karlberg,E.O., Storling,Z.M., Olason,P.I., Pedersen,A.G., Rigina,O., Hinsby,A.M., Tumer,Z., Pociot,F. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nat. Biotechnol.*, **25**, 309–316.
12. The UniProt Consortium, (2007) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **35**, D193–D197.
13. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U. *et al.* (2004) The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat. Biotechnol.*, **22**, 177–183.
14. Ruepp,A., Zollner,A., Maier,D., Albermann,K., Hani,J., Mokrejs,M., Tetko,I., Guldener,U., Mannhaupt,G. *et al.* (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes. *Nucleic Acids Res.*, **32**, 5539–5545.
15. Ruepp,A., Doudieu,O.N., van den,O.J., Brauner,B., Dunger-Kaltenbach,I., Fobo,G., Frishman,G., Montrone,C., Skornia,C. *et al.* (2006) The Mouse Functional Genome Database (MfunGD): functional annotation of proteins in the light of their cellular context. *Nucleic Acids Res.*, **34**, D568–D571.
16. Luc,P.V. and Tempst,P. (2004) PINdb: a database of nuclear protein complexes from human and yeast. *Bioinformatics*, **20**, 1413–1415.
17. Pennisi,E. (2007) Genetics. Working the (gene count) numbers: finally, a firm answer? *Science*, **316**, 1113.
18. Bader,G.D., Betel,D. and Hogue,C.W. (2003) BIND: the biomolecular Interaction network database. *Nucleic Acids Res.*, **31**, 248–250.
19. Mishra,G.R., Suresh,M., Kumaran,K., Kannabiran,N., Suresh,S., Bala,P., Shivakumar,K., Anuradha,N., Reddy,R. *et al.* (2006) Human protein reference database–2006 update. *Nucleic Acids Res.*, **34**, D411–D414.
20. Mewes,H.W., Frishman,D., Mayer,K.F., Munsterkotter,M., Noubibou,O., Pagel,P., Rattei,T., Oesterheld,M., Ruepp,A. *et al.* (2006) MIPS: analysis and annotation of proteins from whole genomes in 2005. *Nucleic Acids Res.*, **34**, D169–D172.
21. Alberts,B. (1998) The cell as a collection of protein machines: preparing the next generation of molecular biologists. *Cell*, **92**, 291–294.
22. Hart,G.T., Lee,I. and Marcotte,E.R. (2007) A high-accuracy consensus map of yeast protein complexes reveals modular nature of gene essentiality. *BMC Bioinformatics*, **8**, 236.